appliedbiosystems

Nothing has changed, **except everything**

Sanger sequencing and fragment analysis like you have never experienced before. Same workflow, same trusted technology, now with an innovative all-in-one cartridge that takes setup time from hours to minutes. Introducing the **Applied Biosystems™ SeqStudio™ Genetic Analyzer**.

Find out more at **thermofisher.com/seqstudio**

ThermoFisher SCIENTIFIC

AJHG

The American Journal of Human Genetics

CellPress
www.cell.com

Best of AJHG 2016 and 2017

AJHGBO 2017

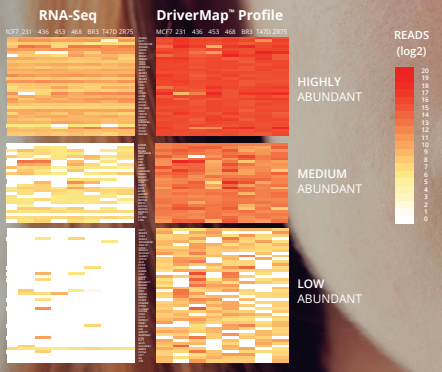
CELLECTA

DriverMap™ Targeted RNA Expression Profiling Solution

Faster. Simpler. More Sensitive than targeted RNA Seq or Microarrays. Illumina-NGS Compatible.

Profile all 19,000 human protein-coding genes in a single tube.

- Globin depletion or mRNA enrichment steps not required
- Human-specific profiling without interference from non-human background cells
- Available as a service or in a kit



At ASHG, Cellecta will be in Booth 1009.

Join us on **Wednesday, October 18** for a lunch **workshop**
Development of Novel Functional Genomic Tools for Identifying Candidate Disease Genes and Biomarkers
Meet our experts

Gus Frangou, M.Sc., D.Phil.
Chief Scientist, Cellecta, Inc.

Development of Novel Synergistic Functional Genomics Technologies to Understand Complex Diseases

Wednesday, October 18, 2017
12:30 - 1:00 pm Hilton Orlando
Lake George Room, Lobby Level

Lester Kobzik, M.D.
Professor of Pathology, Brigham Women's Hospital, Harvard Medical School

Assessment of Immune Status and Biomarker Discovery Using Blood Transcriptomics

Wednesday, October 18, 2017
1:00 - 1:30 pm Hilton Orlando
Lake George Room, Lobby Level

Paul Diehl, Ph.D.
Chief Operating Officer, Cellecta, Inc

CRISPR screens with pooled sgRNA libraries serve as useful tools to Identify genes responsible for key biological responses

Friday, October 20, 2017, 1:15 pm
Hilton Orlando (Agilent workshop, separate registration required)
Lake Down Room, Lobby Level

Register for our Wednesday lunch workshop at www.cellecta.com/ashg2017

Who we are

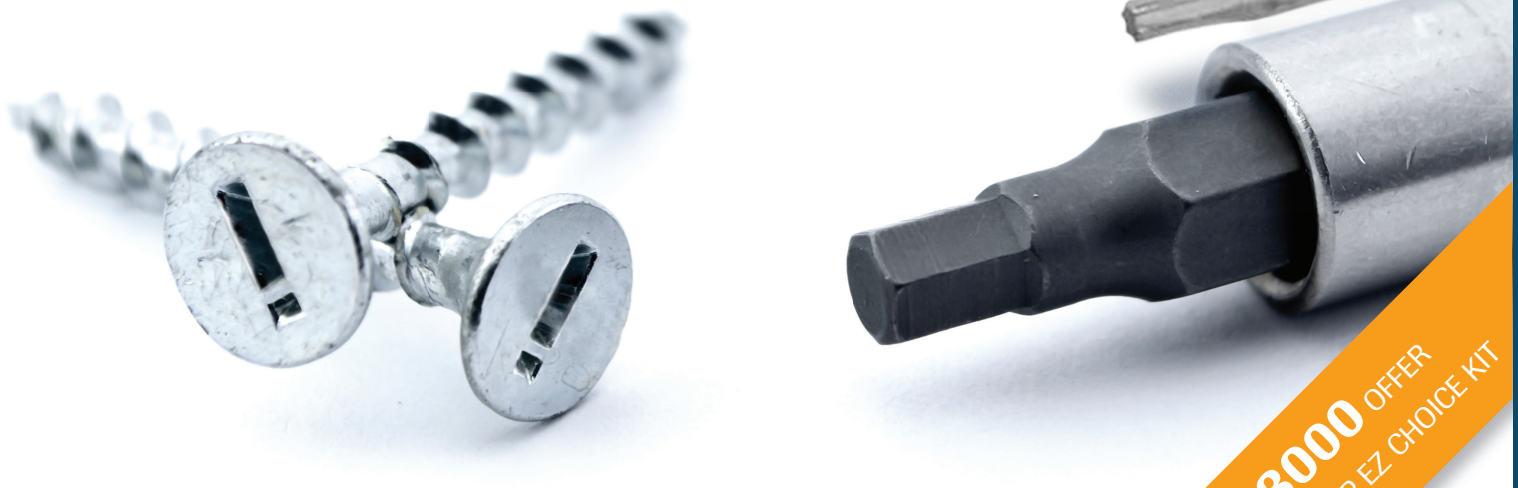
Cellecta is a leading provider of genomic products and services. Our functional genomics portfolio includes gene knockout and knockdown screens, custom and genome-wide CRISPR and RNAi libraries, construct services, cell engineering, NGS kits and targeted expression profiling products and services.

We can help your discovery efforts.
www.cellecta.com info@cellecta.com +1 877-938-3910 or +1 650-938-3910

CELLECTA

Roche

YOUR CUSTOM SCIENCE NEEDS A CUSTOM TOOL



SeqCap EZ Choice Custom Target Enrichment

SPECIAL **$3000** OFFER
4-REACTION SEQCAP EZ CHOICE KIT

Design and order your custom capture today for only $3K

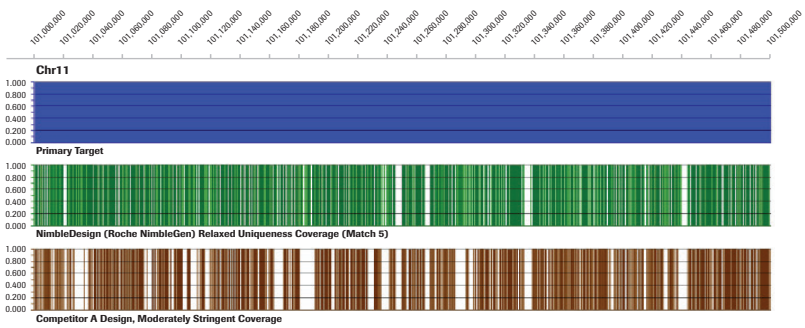
Now is a great time to design and evaluate your custom capture for NGS. Now until April 30th, 2018, US customers can evaluate a 4-reaction pack (before multiplexing) of SeqCap EZ Choice human design for only $3000*! Target any content of 0 - 7 Mb.

SeqCap EZ empowers you.

- Increase coverage and uniformity by deploying up to 2.1 million probes
- Reduce expense by fully supported multiplexing
- Combine Roche SeqCap EZ Choice with KAPA HyperPrep or KAPA HyperPlus library prep kits in an optimized HyperCap Workflow
- Easily scale from pilot studies to high-throughput production
- Increase throughput with help from our Automation Support Team
- Design your capture online or by working directly with our Design Team

Start your customized SeqCap EZ Choice Design today!

Visit **go.roche.com/seqcapeval** to start generating your design, request panel-design assistance, or obtain a quotation for your $3K 4-reaction pack. Learn more about the advantages of SeqCap Target Enrichment at **go.roche.com/te**.



Design comparison between SeqCap EZ Choice Enrichment System and Competitor A Design Software. A 3.51 Mb region on Chromosome 18 was targeted to compare probe coverage across design platforms. Roche NimbleDesign probes covered 79.3% of the target region (2.79 Mb of target) compared to 50.2% coverage (1.77 Mb of target) by Supplier A's design software. Note: Probe design used standard design parameters for both NimbleDesign and Supplier A's software. Coordinate files from each design were uploaded to SignalMap software to obtain coverage file illustrated above. About 1.05 Mb of coverage space was unique to the NimbleDesign platform and not covered by Supplier A's software.

*Visit **go.roche.com/seqcapeval** today!*

*Offer limited to first-time, US-based SeqCap EZ users only. Valid on human capture targets of 0 - 7 Mb. Purchase order must be received by 4/30/2018. Contact your Roche representative for pricing on larger designs (7-200 Mb) or on captures for non-human samples.

LIFE IS BUSY ENOUGH

Your patients with urea cycle disorders (UCDs)—and their families—don't have time for complicated dosing. RAVICTI® (glycerol phenylbutyrate) Oral Liquid offers 24/7 ammonia control with easy administration for busy lives.[1]

RAVICTI IS APPROVED FOR PATIENTS 2 MONTHS OF AGE AND OLDER WITH UCDs.[1]

- The only FDA-approved oral liquid nitrogen scavenger therapy[1]
- Nearly tasteless and odorless[2]
- No pill or powder preparation[1,2]
- Taken with meals via oral dosing syringe[1]

Consider switching your patients now.[1]

Visit ASHG booth #322 and RavASHG.com

RAVICTI is indicated for the chronic management of patients with UCDs ≥2 months of age who cannot be managed by diet and supplementation alone. It must be used with dietary protein restriction.[1]

RAVICTI is not indicated for the treatment of acute hyperammonemia or for NAGS deficiency, and is contraindicated in patients <2 months of age.[1]

- Of the 17 pediatric patients 2 months to <2 years of age in 3 open-label studies, 7 patients (41%) reported a total of 11 hyperammonemic crises.[1]
- Of the 26 pediatric patients 6 to 17 years of age in both 12-month studies of RAVICTI, 5 patients (19%) reported a total of 5 hyperammonemic crises.[1]
- Of the 51 adult patients in the 12-month study of RAVICTI, 7 patients (14%) reported a total of 10 hyperammonemic crises.[1]

Please review the Brief Summary for RAVICTI on the following page. Visit RavASHG.com to download a copy of the RAVICTI Full Prescribing Information and Medication Guide.

Abbreviation: ASHG, the American Society of Human Genetics.

References: 1. RAVICTI [package insert]. Lake Forest, IL: Horizon Pharma USA, Inc.; 2017. **2.** Diaz GA, Krivitzky LS, Mokhtarani M, et al. *Hepatology*. 2013;57(6):2171-2179. doi:10.1002/hep.26058.

HORIZON PHARMA

RAVICTI® (glycerol phenylbutyrate) Oral Liquid

Rethink potential.

RAVICTI® (glycerol phenylbutyrate) Oral Liquid

BRIEF SUMMARY

INDICATIONS AND USAGE

RAVICTI is indicated for use as a nitrogen-binding agent for chronic management of patients 2 months of age and older with urea cycle disorders (UCDs) who cannot be managed by diet and/or supplementation alone. RAVICTI must be used with dietary protein restriction and, in some cases, supplements (eg, essential amino acids, arginine, citrulline, protein-free calorie supplements).

LIMITATIONS OF USE

- RAVICTI is not indicated for the treatment of acute hyperammonemia in patients with UCDs because rapidly acting interventions are essential to reduce plasma ammonia levels.
- The safety and efficacy of RAVICTI for the treatment of *N*-acetylglutamate synthase (NAGS) deficiency has not been established.

DETAILED IMPORTANT SAFETY INFORMATION

CONTRAINDICATIONS

- *Patients less than 2 months of age:* Children less than 2 months of age may have immature pancreatic exocrine function, which could impair hydrolysis of RAVICTI, leading to impaired absorption of phenylbutyrate and hyperammonemia.
- *Patients with known hypersensitivity to phenylbutyrate:* Reactions include wheezing, dyspnea, coughing, hypotension, flushing, nausea, and rash.

WARNINGS AND PRECAUTIONS

- *Neurotoxicity:* Phenylacetate (PAA), the major metabolite of RAVICTI, may be toxic at levels of 500 µg/mL or greater. Reduce RAVICTI dosage if symptoms of neurotoxicity, including vomiting, nausea, headache, somnolence, or confusion, are present in the absence of high ammonia or other intercurrent illnesses.
- *Reduced phenylbutyrate absorption in pancreatic insufficiency or intestinal malabsorption:* Low or absent pancreatic enzymes or intestinal disease resulting in fat malabsorption may result in reduced or absent digestion of RAVICTI and/or absorption of phenylbutyrate and reduced control of plasma ammonia. Monitor ammonia levels closely.

USE IN SPECIFIC POPULATIONS

- *Pregnancy:* RAVICTI should be used with caution in patients who are pregnant or planning to become pregnant. Based on animal data, RAVICTI may cause fetal harm. A voluntary patient registry monitors pregnancy outcomes in women exposed to RAVICTI. For more information regarding the registry program, visit www.ucdregistry.com or call 1-855-823-2595.
- *Nursing mothers:* Breastfeeding is not recommended during treatment with RAVICTI. There are no data on the presence of RAVICTI in human milk, the effects on the breastfed infant, nor the effects on milk production.

ADVERSE REACTIONS

- In ≥10% of adult patients: diarrhea, flatulence, and headache occurred during 4-week treatment (n=44) with RAVICTI; nausea, vomiting, diarrhea, decreased appetite, dizziness, headache, and fatigue occurred during 12-month treatment (n=51) with RAVICTI.
- In ≥10% of pediatric patients ages 2 to 17 years: upper abdominal pain, rash, nausea, vomiting, diarrhea, decreased appetite, and headache occurred during 12-month treatment (n=26) with RAVICTI.
- In ≥10% of pediatric patients ages 2 months to less than 2 years: neutropenia, vomiting, diarrhea, pyrexia, hypophagia, cough, nasal congestion, rhinorrhea, rash, and papule occurred during 12-month treatment (n=6) with RAVICTI.

DRUG INTERACTIONS

- Corticosteroids, valproic acid, or haloperidol may increase plasma ammonia level. Monitor ammonia levels closely.
- Probenecid may affect renal excretion of metabolites of RAVICTI, including phenylacetylglutamine (PAGN) and PAA.
- CYP3A4 substrates with narrow therapeutic index (eg, alfentanil, quinidine, cyclosporine): RAVICTI may decrease exposure to the concomitant drug.
- Midazolam: Use of RAVICTI decreased exposure of midazolam with concomitant use.

You are encouraged to report negative side effects of prescription drugs to the FDA. Visit www.fda.gov/medwatch or call 1-800-FDA-1088.

Visit RavASHG.com to download a copy of the RAVICTI Full Prescribing Information and Medication Guide.

Echo® Acoustic LIQUID HANDLING
for **SEQUENCING**

Visit Us @ **ASHG 2017**

Booth 507

Reduce Library Prep Costs

100-Fold

Echo® Liquid Handlers enable library preparation in low microliter volumes for a range of sequencing methods. Dramatically reduce reagent costs, conserve samples, and eliminate steps – all while improving library quality.

Echo acoustic liquid handling allows...

- 100-fold reduction of library prep reaction volumes
- 30-fold reduction of sample pooling turnaround time
- Increased sample throughput
- Automation of workflow to easily prepare thousands of samples
- Improved accuracy of results

Comparison of Liquid Handling Methods*

	Manual Pipetting	Echo® Liquid Handler
Amount of DNA	50 ng	0.06 – 2.0 ng
DNA volume (Rxn)	25 µL	200 nL
Library prep volume (Rxn)	25 µL	300 nL
Total volume	50 µL	0.5 µL
Reactions per kit	96	9600
Cost per reaction	$72.91	**$0.73**

For more information, visit www.labcyte.com/**sequencing.**

** Low-Cost, High-Throughput Sequencing of DNA Assemblies Using a Highly Multiplexed Nextera Process. Shapland et al. ACS Synth. Biol., 2015*

LABCYTE™
The Future of Science is Sound

@LabcyteInc info-us@labcyte.com

MOVE ASIDE, ROVER.

Biomarker detection just got a lot more sensitive.

The new Quanterix SR-Plex™ Ultra-Sensitive Biomarker Detection System features multiplexed detection of analytes and 1,000 times the sensitivity of ELISA, changing the biomarker detection game one molecule at a time.

Discover the SR-Plex at
go.quanterix.com/sr-plex

Quanterix™
The Science of Precision Health

Introducing the Illumina Bio-Rad Single-Cell Sequencing Solution. Access high-resolution insights into gene expression in a single, comprehensive workflow.

Single-cell RNA-Seq delivers higher resolution of gene regulation for a deeper view of cell function, disease progression, and identification of therapeutic targets in research, compared to RNA-Seq. Developed by the industry leaders in sequencing and Droplet Digital™ technologies, our robust, scalable, and user-friendly workflow allows transcriptome profiling of hundreds to tens of thousands of single cells.

Gain insights into your research.
Learn more at **bio-rad.com/ddSEQsinglecell**

For Research Use Only. Not for use in diagnostic procedures.

illumina® | BIO-RAD

GE Healthcare

Accelerate your drug discovery research at every step

GE Healthcare provides innovative tools and analytics to support therapeutic development from target identification, all the way through bioprocessing. Confidently move through the drug development pipeline while reducing costs and maximizing productivity throughout development and biomanufacturing. Learn more at **gelifesciences.com.**

imagination at work

Foreword

We are pleased to bring you the latest installment of the *Best of AJHG* reprint collection from Cell Press, which gives us a chance to reflect upon the science that engaged and influenced *AJHG* readers in late 2016 and early 2017. This collection includes nine of the most accessed research articles, which span a range of topics, as well as the most highly accessed Commentary. To select the articles, we used the number of requests for PDF and full-text HTML versions of a given article. We acknowledge that no single measurement can truly be indicative of "the best" research papers over a given period of time. This is especially true when sufficient time has not passed to allow one to fully appreciate the relative impact of a discovery. Nonetheless, we hope you agree that it is still informative to look back at our scientific community's interests in what has been published in *AJHG* over the past year.

In this collection, you will see a range of the exciting topics that have widely captured the attention and enthusiasm of our readers, including genetic ancestry, translational genomics, the genetic bases of disease, and genetic risk prediction. We are especially pleased to note that included in this Best of collection are the two winners of the 2017 Cotterman Award, which recognizes the best papers published in the Journal for which the first author was either a pre- or post-doctoral trainee and an ASHG member.

As always, we welcome your submissions and look forward to working with you to bring your best work to the attention of the human genetics community.

We hope that you will enjoy reading this special collection and that you will visit www.cell.com/AJHG to check out the latest findings that we have had the privilege to publish. Also be sure to visit www.cell.com to find other high-quality papers published in the full collection of Cell Press journals. Please feel free to contact us at ajhg@ajhg.net to tell us about your latest work or to provide feedback. We look forward to working with you in 2018 and beyond!

Finally, we are grateful for the generosity of our sponsors, who helped make this reprint collection possible.

AJHG

David L. Nelson, PhD
Editor

Sara B. Cullinan, PhD
Deputy Editor

Sarah Ratzel, PhD
Scientific Editor

For more information about the Best Of Series, please contact:
Jonathan Christison | Program Director | jchristison@cell.com | 617-397-2893

eppendorf

Exceptional Sample Safety

Eppendorf CryoCube® F740-series ULT –85 °C Freezer

Advancements designed for rapid recovery times and maximum temperature uniformity make the new Eppendorf CryoCube F740i ULT –85 °C Freezer a secure harbor for your samples while dramatically reducing power consumption and noise output.

> Automatic vacuum release port allows quick and easy re-entry
> Broad, flat gaskets keep the cold inside and minimize frost buildup
> Whisper-quiet operation for a comfortable lab environment
> Voltage inverter provides protection from in-line power fluctuations

www.eppendorf.com • 800-645-3050

ASHG2018

SAN DIEGO • OCTOBER 16–20, 2018

SHARING DISCOVERIES. **SHAPING** OUR FUTURE.

Abstract Deadline
June 7, 2018

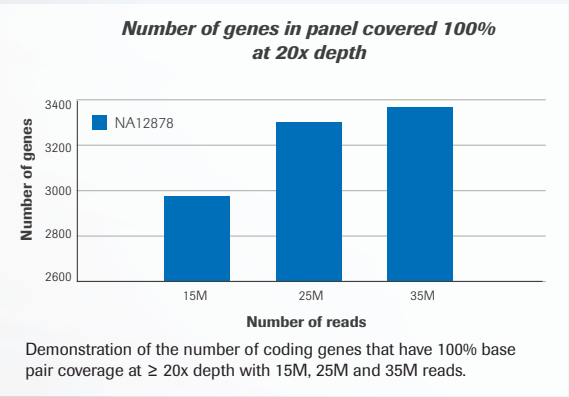
Invited Proposals and Workshops accepted through
December 14, 2017

ashg.org/2018meeting

INHERITED DISEASE RESEARCH FOR THE NEXT GENERATION

Target next-generation sequencing efforts on the exonic regions of over 4,100 medically relevant genes. Choose the optimized hybridization-based SeqCap EZ Inherited Disease Panel to achieve highly uniform sequencing coverage for genes classified as pathogenic from OMIM, as well as other medically relevant content identified by scientific collaborators.

- Use a single 11.8 Mb panel before NGS to comprehensively replace many commonly run NGS disease research panels and PCR-based single-gene assays
- Eliminate or decrease the amount of Sanger sequencing necessary for full gene coverage
- Increase lab efficiency and decrease sequencing costs with uniform panel coverage



Demonstration of the number of coding genes that have 100% base pair coverage at ≥ 20x depth with 15M, 25M and 35M reads.

Request a free sample today

Evaluate the SeqCap EZ Inherited Disease Panel in your lab with a complimentary sample* of this optimized panel.

Visit **go.roche.com/IDPsample** to learn more or request your free panel today.

For Research Use Only. Not for use in diagnostic procedures.
*Free sample offer available only to US-based medical researchers evaluating for commercial use.

AJHG

Best of 2016 and 2017

(continued)


Asuragen®
Genetics
Bringing clarity
to complex gene targets.
While highly repetitive regions of the genome pose significant hurdles to reliable genetic analysis, their implications in disease are becoming clear.
Asuragen is committed to providing clinical researchers with the kits needed to investigate the most compelling and challenging targets, including genes implicated in Alzheimer's, ALS, FTD, Myotonic Dystrophy and Fragile X Syndrome.
Our clinical research kits overcome target complexity so you can focus on what matters most — scientific breakthroughs that drive better outcomes.
AmplideX® PCR/CE FMR1[1,2] | C9orf72[2] | TOMM40[2] | DMPK[3]
Xpansion Interpreter® for AGG Interruptions in FMR1[4]
Visit Asuragen at the American Society of Human Genetics Annual Meeting
Booth # 824 | Oct 18-21, 2017 | Orange County Convention Center | Orlando, FL
Sensitive | Accurate | Complete
www.asuragen.com/ComplexityMadeClear
complexity made CLEAR
[1]CE-IVD. [2]RUO. [3]In development. [4]Only available from the Asuragen Clinical Laboratory.

Whole-Exome Sequencing Identifies Loci Associated with Blood Cell Traits and Reveals a Role for Alternative *GFI1B* Splice Variants in Human Hematopoiesis

Linda M. Polfus, Rajiv K. Khajuria, Ursula M. Schick, Nathan Pankratz, Raha Pazoki, Jennifer A. Brody, Ming-Huei Chen, Paul L. Auer, James S. Floyd, Jie Huang, Leslie Lange, Frank J.A. van Rooij, Richard A. Gibbs, Ginger Metcalf, Donna Muzny, Narayanan Veeraraghavan, Klaudia Walter, Lu Chen, Lisa Yanek, Lewis C. Becker, Gina M. Peloso, Aoi Wakabayashi, Mart Kals, Andres Metspalu, Tõnu Esko, Keolu Fox, Robert Wallace, Nora Franceshini, Nena Matijevic, Kenneth M. Rice, Traci M. Bartz, Leo-Pekka Lyytikäinen, Mika Kähönen, Terho Lehtimäki, Olli T. Raitakari, Ruifang Li-Gao, Dennis O. Mook-Kanamori, Guillaume Lettre, Cornelia M. van Duijn, Oscar H. Franco, Stephen S. Rich, Fernando Rivadeneira, Albert Hofman, André G. Uitterlinden, James G. Wilson, Bruce M. Psaty, Nicole Soranzo, Abbas Dehghan, Eric Boerwinkle, Xiaoling Zhang, Andrew D. Johnson, Christopher J. O'Donnell, Jill M. Johnsen, Alexander P. Reiner, Santhi K. Ganesh, and Vijay G. Sankaran

Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project

Paul L. Auer, Alex P. Reiner, Gao Wang, Hyun Min Kang, Goncalo R. Abecasis, David Altshuler, Michael J. Bamshad, Deborah A. Nickerson, Russell P. Tracy, Stephen S. Rich, NHLBI GO Exome Sequencing Project, and Suzanne M. Leal

Cotterman Award Winners

Modeling the Mutational and Phenotypic Landscapes of Pelizaeus-Merzbacher Disease with Human iPSC-Derived Oligodendrocytes

Zachary S. Nevin, Daniel C. Factor, Robert T. Karl, Panagiotis Douvaras, Jeremy Laukka, Martha S. Windrem, Steven A. Goldman, Valentina Fossati, Grace M. Hobson, and Paul J. Tesar

Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations

Alicia R. Martin, Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, and Eimear E. Kenny

2018 Membership Opens in Late Fall

Enjoying the journal? There's never been a better time to join the worldwide community of human genetics specialists by becoming a member of ASHG. ASHG membership provides career and leadership opportunities, networking and collaboration, and valuable savings.

Learn more: www.ashg.org/membership

International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases

Kym M. Boycott,[1,*] Ana Rath,[2] Jessica X. Chong,[3] Taila Hartley,[1] Fowzan S. Alkuraya,[4,5] Gareth Baynam,[6] Anthony J. Brookes,[7] Michael Brudno,[8] Angel Carracedo,[9] Johan T. den Dunnen,[10] Stephanie O.M. Dyke,[11] Xavier Estivill,[12,13] Jack Goldblatt,[6] Catherine Gonthier,[2] Stephen C. Groft,[14] Ivo Gut,[15] Ada Hamosh,[16] Philip Hieter,[17] Sophie Höhn,[2] Matthew E. Hurles,[18] Petra Kaufmann,[19] Bartha M. Knoppers,[11] Jeffrey P. Krischer,[20] Milan Macek, Jr.,[21] Gert Matthijs,[22] Annie Olry,[2] Samantha Parker,[23] Justin Paschall,[18] Anthony A. Philippakis,[24] Heidi L. Rehm,[24] Peter N. Robinson,[25,26] Pak-Chung Sham,[27] Rumen Stefanov,[28] Domenica Taruscio,[29] Divya Unni,[2] Megan R. Vanstone,[1] Feng Zhang,[30,31] Han Brunner,[32,33] Michael J. Bamshad,[3,34] and Hanns Lochmüller[35]

Provision of a molecularly confirmed diagnosis in a timely manner for children and adults with rare genetic diseases shortens their "diagnostic odyssey," improves disease management, and fosters genetic counseling with respect to recurrence risks while assuring reproductive choices. In a general clinical genetics setting, the current diagnostic rate is approximately 50%, but for those who do not receive a molecular diagnosis after the initial genetics evaluation, that rate is much lower. Diagnostic success for these more challenging affected individuals depends to a large extent on progress in the discovery of genes associated with, and mechanisms underlying, rare diseases. Thus, continued research is required for moving toward a more complete catalog of disease-related genes and variants. The International Rare Diseases Research Consortium (IRDiRC) was established in 2011 to bring together researchers and organizations invested in rare disease research to develop a means of achieving molecular diagnosis for all rare diseases. Here, we review the current and future bottlenecks to gene discovery and suggest strategies for enabling progress in this regard. Each successful discovery will define potential diagnostic, preventive, and therapeutic opportunities for the corresponding rare disease, enabling precision medicine for this patient population.

Introduction

Rare diseases, though individually rare, are collectively common. A rare disease is defined as one that affects fewer than 200,000 people in the US[1] or less than 1 in 2,000 people in Europe.[2]A substantive number of rare diseases are due to altered functions of single genes. Cumulatively, these rare genetic diseases (RGDs), also termed Mendelian or monogenic diseases, affect at least 1 in 50 individuals in the European-derived general population.[3] Our understanding of the number of RGDs that exist is incomplete but is estimated to be well over 7,000 according to current medical and genetic evidence[4] (also see Orphanet in the Web Resources). Despite their often chronic and progressive nature, long-term complications can be

[1]Children's Hospital of Eastern Ontario Research Institute, University of Ottawa, Ottawa, ON K1H 8L1, Canada; [2]Orphanet, Institut National de la Santé et de la Recherche Médicale US14, 75014 Paris, France; [3]Department of Pediatrics, University of Washington, Seattle, WA 98195, USA; [4]Department of Genetics, King Faisal Research Center, Riyadh 11211, Saudi Arabia; [5]Saudi Human Genome Program, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia; [6]Genetic Services of Western Australia, Perth, WA 6008, Australia; [7]Department of Genetics, University of Leicester, Leicester LE1 7RH, UK; [8]Department of Computer Science, University of Toronto, Toronto M5S 1A1, Canada; [9]Genomic Medicine Group, Galician Foundation of Genomic Medicine and University of Santiago de Compostela, 15782 Santiago de Compostela, Spain; [10]Departments of Human Genetics and Clinical Genetics, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, the Netherlands; [11]Centre of Genomics and Policy, Department of Human Genetics, Faculty of Medicine, McGill University, Montreal, QC H3A 1A4, Canada; [12]Experimental Division, Sidra Medical and Research Center, PO Box 26999, Doha, Qatar; [13]Genetics Unit, Dexeus Woman's Health, 08028 Barcelona, Spain; [14]National Center for Advancing Translational Sciences, National Institutes of Health, Bethesda, MD 20892-4874, USA; [15]Centre Nacional d'Anàlisi Genòmica, Center for Genomic Regulation, Barcelona Institute of Science and Technology, Universitat Pompeu Fabra, 08028 Barcelona, Spain; [16]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21286, USA; [17]Michael Smith Laboratories, Department of Medical Genetics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; [18]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK; [19]Office of Rare Diseases Research, National Center for Advancing Translational Sciences, National Institutes of Health, Bethesda, MD 20892-4874, USA; [20]University of South Florida Health Informatics Institute, Tampa, FL 33620, USA; [21]Department of Biology and Medical Genetics, Second Faculty of Medicine, Charles University and University Hospital Motol, 150 06 Prague 5, Czech Republic; [22]Center for Human Genetics, University of Leuven, 3000 Leuven, Belgium; [23]Lysogene, 92 200 Neuilly-sur-Seine, France; [24]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [25]Institut für Medizinische Genetik und Humangenetik, Charité Universitätsmdizin Berlin, 13353 Berlin, Germany; [26]Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA; [27]Centre for Genomic Sciences, University of Hong Kong, Hong Kong, China; [28]Department of Social Medicine and Public Health, Faculty of Public Health, Medical University of Plovdiv, Plovdiv 4002, Bulgaria; [29]National Centre for Rare Diseases, Istituto Superiore di Sanità, Rome 299-00161, Italy; [30]WuXi AppTec, Waigaoqiao Free Trade Zone, Shanghai 200131, China; [31]WuXi NextCODE, Cambridge, MA 02142, USA; [32]Department of Human Genetics, Radboud University Medical Center, 6525 GA Nijmegen, the Netherlands; [33]Maastricht University Medical Center, Department of Clinical Genetics, 6229 GT Maastricht, the Netherlands; [34]Division of Genetic Medicine, Seattle Children's Hospital, Seattle, WA 98105, USA; [35]John Walton Muscular Dystrophy Research Centre, MRC Centre for Neuromuscular Diseases, Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne NE1 3BZ, UK
*Correspondence: kboycott@cheo.on.ca
http://dx.doi.org/10.1016/j.ajhg.2017.04.003.

lessened or delayed for some RGDs if they are diagnosed early (e.g., via newborn screening) and optimally managed by standard and/or targeted therapies. In addition, a definitive molecular diagnosis can obviate the need for further diagnostic investigations, facilitate appropriate access to healthcare resources, reduce prognostic uncertainty, provide accurate recurrence-risk counseling, foster reproductive choices in affected families, and impart psychosocial benefits to the patient and their family. Importantly, understanding the underlying genetic etiology and linking a RGD to a causative biological pathway is leading to highly effective targeted therapies for some severe, previously only symptomatically treatable RGDs (e.g., ivacaftor for class III *CFTR* [MIM: 602421] pathogenic variants).[5] Ultimately, successful deployment of precision medicine will be directly related to diagnostic success for patients with RGDs.

Current Understanding of Phenotypic and Genetic Diversity of RGDs

Knowledge of the phenotypic and genetic diversity of RGDs is steadily increasing; however, substantial gaps remain. Establishing the number of RGDs is challenging for several reasons, not the least of which is distinguishing between novel and known diseases to objectively segment a continuum of pathologies into discrete disease entities. Two international databases curate clinical and genetic data for the community: Online Mendelian Inheritance in Man (OMIM)[4] and Orphanet.[6] OMIM has continuously provided curation and classification of Mendelian disease since it began as *Mendelian Inheritance in Man*, first published by Dr. V. McKusick in 1966; OMIM has been online and searchable since 1987. OMIM mines the biomedical literature and, according to expert review, curates significant new information on genes and genetic phenotypes into separate gene and phenotype entries. OMIM numbers for Mendelian diseases are incorporated into the biomedical literature across many disciplines of medicine. OMIM emphasizes gene-phenotype relationships by cataloging the same or similar phenotypes caused by pathogenic variants in different genes as distinct entities; genetic heterogeneity is displayed through the associated Phenotypic Series. A recent analysis of OMIM (data downloaded September 5, 2016) recognized 3,209 unique genes associated with 4,550 monogenic rare diseases.

Orphanet (see Web Resources) has maintained an inventory of both genetic and other rare diseases since 1997. Within Orphanet, a rare disease is defined as a recognizable and homogeneous clinical presentation, whatever the cause or the number of genes related to it. Disorders are organized in a multi-hierarchical classification and can be further subdivided into subtypes, of which genetic subtypes are included. Orphanet performs a literature survey and curates the published literature of newly discovered genes or new gene-disease relations. As a result, a semantic relation is assigned to couple the gene and disease in the database. As of September 14, 2016, Orphanet documented 3,654 unique genes associated with 3,551 rare diseases.

The discrepancy in the number of rare diseases with monogenic etiology documented in each of the two databases (4,550 for OMIM and 3,551 for Orphanet) can be attributed to the way each database is structured; OMIM categorizes rare diseases on the basis of genetic etiology, whereas Orphanet groups by clinically recognizable diseases and can include more than one OMIM entry when the same disease is caused by variants in more than one gene. Recently, the Clinical Genome Resource (ClinGen)[7] has begun defining the strength of evidence for published gene-disease associations. The evidence levels are scored according to semiquantitative frameworks, and the scores are posted on ClinGen's website along with the scoring sheets that structure the evidence and sources. These scores will also soon be posted on OMIM. As ClinGen grows, it will enable a clear delineation between those genes for which gene-disease causality is substantiated and those claims that will require further evidence for implication.

Although substantial progress has been made toward identifying the genetic basis of rare diseases, the underlying etiologies for approximately half remain undiscovered. Beginning in the mid-1980s, and for the following two decades, the primary approach to gene discovery was a combination of linkage analysis, positional cloning, and sequencing of candidate or regionally selected genes, most of which was hypothesis driven. The subsequent introduction of next-generation sequencing (NGS) strategies to identify genes associated with disease, primarily based on whole-exome sequencing (WES), in 2009 accelerated the pace of discovery by enabling hypothesis-free approaches. Today, WES is routinely used as the primary technological approach to discovering disease-gene associations (Figure 1). Its favor over whole-genome sequencing (WGS) has primarily been due to its significantly lower cost and that the majority of pathogenic variants continue to be within the protein-coding portion of the genome. Without a doubt, as the cost of WGS decreases, clinicians and researchers will transition to its use given its more even coverage, its ability to identify structural variation, and the opportunity it provides to uncover non-exomic variants.

Our analysis of OMIM documented an average of 259 "novel" RGD discoveries per year from 2012 to 2015 (Figure 1), comprising 157 new disease-gene discoveries (here defined as pathogenic variants in a gene that had not been previously associated with disease) and 102 new disease-gene relations each year (defined as pathogenic variants in a gene previously associated with a different disease; data not shown).[8] Orphanet documents an average of 281 novel RGD discoveries per year over the same time period: 160 new disease-gene discoveries and 121 new disease-gene relations (Figure 2). Orphanet and OMIM report essentially

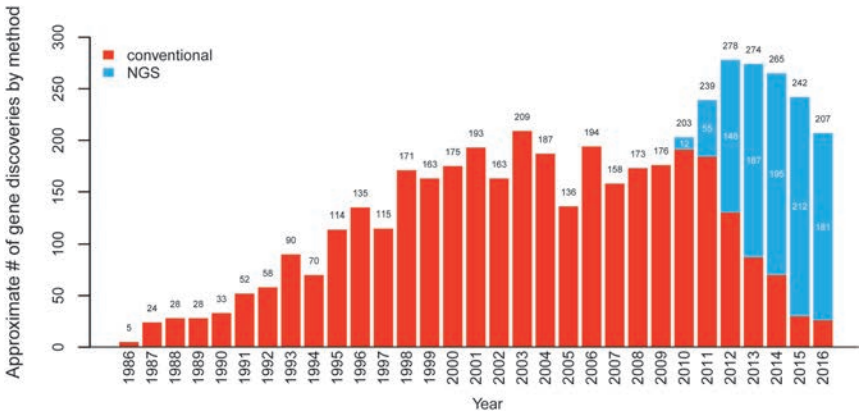


Figure 1. Approximate Number of Gene Discoveries Made by WES and WGS versus Conventional Approaches since 2010 according to OMIM Data
Since the introduction of WES and WGS in 2010, the pace of the discovery of genes underlying RGDs per year has increased, and the proportion of discoveries made by WES or WGS (blue) or by conventional approaches (red) has steadily increased. Since 2013, WES and WGS have discovered nearly three times as many genes as conventional approaches, but the rate of discovery appears to be declining. Adapted from Chong et al.[8]

the same number of new disease-gene discoveries (average of 160 and 157, respectively, over the same time period), but more disease-gene relations have been reported by Orphanet (121 versus 102 for OMIM). In a manual review of randomly selected discrepancies between OMIM and Orphanet, this is most likely attributable to differences in the process of curation; OMIM is more likely to decide that the publication reports a phenotypic expansion of an already explained RGD than a new disease-gene relation. Nevertheless, the data from OMIM and Orphanet both show that a significant proportion of RGD discoveries are new diseases associated with pathogenic variants in previously known genes (gene-disease relations): 38 and 43%, respectively. This is an interesting trend in comparison with a recent analysis of all of OMIM's data, which demonstrated that nearly 25% of all genes associated with Mendelian disease underlie two or more clinically distinct disorders.[8]

Since the introduction of WES, many RGDs that were previously intractable to conventional gene-discovery approaches, largely because they were associated with a substantially reduced reproductive fitness, have been found to be caused by de novo pathogenic variants or to exhibit high allelic or locus heterogeneity. These RGDs are enriched with highly recognizable clinical presentations; are often associated with early age of onset, severe phenotype, and/or clear laboratory and/or medical imaging features; and are caused by highly penetrant pathogenic, protein-coding genomic variants (i.e., in legacy terminology, "mutations"). In addition, these RGDs are usually autosomal, X-linked recessive, or de novo dominant, rendering them relatively more accessible and amenable to current discovery strategies relying on WES; these RGDs represent the sweet spot of WES-based approaches. Both OMIM and Orphanet data (Figures 1 and 2) show a trend toward a decreasing number of discoveries per year; whether this trend is real or will continue will require analysis of data from future years. However, what is clear is that recognized bottlenecks must be addressed if the current pace of discoveries is to be maintained, or even accelerated, after the more straightforward RGDs have been solved.

The International Rare Diseases Research Consortium

The International Rare Diseases Research Consortium (IRDiRC) was established in 2011 to bring together researchers and organizations invested in rare disease research. Three IRDiRC Scientific Committees (Diagnostics, Interdisciplinary, and Therapies) and representation from three patient-advocacy groups (two from the US [National Organization for Rare Disorders (NORD) and Genetic Alliance] and one from Europe [Rare Diseases Europe-EURORDIS]), advise the Consortium Assembly (formerly the Executive Committee), which includes public research funders and private-sector members from pharma and biotech from 42 member institutions. Each has committed at least $10,000,000 USD to rare disease research within their jurisdiction (Figure 3; data accessed January 11, 2017). Currently, rare disease research coordinated under the umbrella of IRDiRC totals more than $2,000,000,000 USD. IRDiRC aims to facilitate the understanding of all rare genetic diseases.

The focus of the Diagnostics and Interdisciplinary Committees, and their associated working groups and task forces, has been identifying current and future bottlenecks to RGD discovery and suggesting strategies by which international cooperation can address them. We anticipate that several shortcomings of the present-day discovery pipeline will need to be addressed if we are to continue to make important RGD discoveries at the current pace, or even accelerate it. These include the collection and analysis of clinical and genomic data, data discovery and sharing, genetic and functional support for the establishment of disease causality, and the presence of disease mechanisms that are intractable to our current analytical and genomics-based approaches, as summarized in Table 1.

Strategies for Enabling the Diagnosis of All RGDs

The coming years will see an expanding need for large-scale infrastructure, resources, and tools for completing the grand challenge: understanding the molecular pathogenesis of all RGDs. Over the past few years, our committees, working groups, and task forces have identified specific areas of high

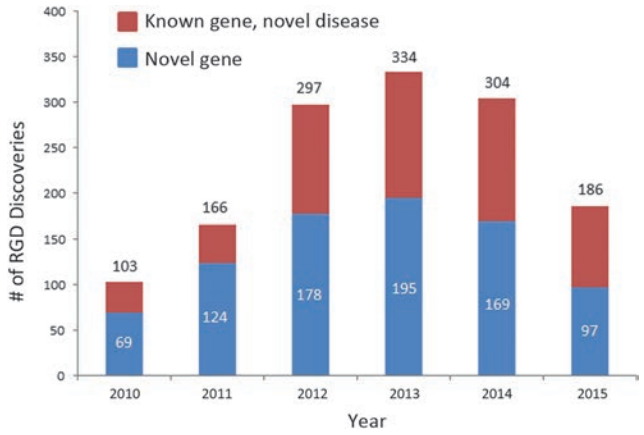


Figure 2. Approximate Number of Novel Gene-Phenotype Discoveries from 2010 to 2015 according to Ophanet Data
Since 2010, the proportion of discoveries that are new disease-gene relations each year (known genes associated with a new disease) has steadily increased. Since 2013, the rate of discovery of both novel genes and new disease-gene relations appears to be declining.

priority to facilitate the achievement of this goal. To this end, the IRDiRC has developed a quality indicator, "IRDiRC Recognized Resources,"[9] on the basis of specific criteria to highlight key resources (e.g., platforms, tools, standards, and guidelines), which, if used more broadly, would accelerate the pace of discoveries.

Ontologies, Terminologies, and Nosologies for Exchanging Clinical Data

Understanding how genomic alterations result in different disease-related phenotypes is fundamental to human health research. In this endeavor, if careful phenotypic characterization is lacking, having genomic data, even from large numbers of individuals, is of limited value. Although we have made large strides toward enabling the sharing of genotype data, standards are not widely used for the exchange of phenotypic data. For undiagnosed RGDs, the situation is even more problematic because only a few individuals in the world might have the same undiagnosed condition. Currently, numerous ontologies, terminologies, and nosologies are used, reflecting the disparate needs and practices of different communities involved in translational research and patient care in many fields of medicine.

The IRDiRC recognizes phenotype ontologies, terminologies, and disease nosologies as critical for RGD research. The Human Phenotype Ontology (HPO)[10,11] has been recognized as a useful annotation of phenotypic abnormalities of RGDs, with the understanding that other resources might be suitable in certain situations, and is being used by RGD databases such as PhenomeCentral,[12] DECIPHER,[13] the UK10K Project,[14] and many others. The HPO has been incorporated into the United Medical Language System (UMLS), which will allow interoperability with an even larger range of medical informatics resources. The HPO is more than a clinical terminology; all terms are set in a hierarchical structure, and it is designed to allow computational analysis of clinical findings for differential diagnostics,[15] as well as RGD phenotypic stratification prior to WES analysis in both the clinical[16] and discovery settings.[17] A key area for ontological development is increasing the granularity and coverage of the HPO across some less well-covered rare-disease domains. Additionally, enabling a means of making longitudinal assessments (onset and temporality), utilizing phenotype negation (the patient does not have phenotype X), and making quantitative specifications (e.g., levels of abnormality of laboratory results) will be important.

To bridge the compatibility gap between various systems and the lack of terminology specific enough for RGDs, the newly established International Consortium for Human Phenotype Terminologies (ICHPT) has worked to provide the community with phenotype terminology standards and definitions for the more often used phenotype terms for database interoperability, in particular to allow the linking of phenotype and genotype databases for RGDs. The ICHPT was created with input from members of several groups, including Orphanet (under the EuroGentest project; see Web Resources), HPO,[18] and OMIM (Robinson et al., 2014, Am. Soc. Hum. Genet., abstract). The outcome of this effort is a set of >2,300 terms that should be present in any terminology through one of its synonyms. These terms have already been mapped to a few of the major terminologies, including HPO,[11] PhenoDB,[19] Orphanet, Elements of Morphology,[20] POSSUM, SNOMED, MeSH, and MedDRA, facilitating cross-compatibility between systems. Where ontologies contain more detailed terms at a finer level of granularity, these terms will map "up" to the broader aligned terms. The IRDiRC recognizes and encourages the ICHPT as the minimal set of standard terms to be used for sharing phenotypic data.

Two complementary rare-disease nosologies exist, the Orphanet Rare Disease Ontology (ORDO)[21] and OMIM.[4] ORDO is a structured vocabulary for rare diseases and is derived from the Orphanet database; it captures relationships between diseases, genes, and other relevant features to form a useful resource for the computational analysis of rare diseases. It integrates nosologies (classifications of rare diseases), relationships (gene-disease relations and epidemiological data), and connections with other terminologies (MeSH, UMLS, and MedDRA), databases (OMIM, UniProtKB, HGNC, Ensembl, Reactome, IUPHAR, and Geneatlas), or classifications (e.g., International Statistical Classification of Diseases and Related Health Problems-10 [ICD-10]). It should be noted that ICD-10 contains only ~500 unique rare-disease classification codes. This deficiency is now being overcome by the development of a

Figure 3. Map of the IRDiRC
The IRDiRC was formally launched in 2011 and currently includes member institutions from Asia, the Middle East, Australasia, Europe, and North America. The current cumulative commitment from the 42 member institutions from both the public and private sectors is estimated at more than $2,000,000,000 USD.

hierarchical rare-disease classification and coding (Orpha numbers) scheme by Orphanet, which will become the basis for inclusion of the majority of known rare diseases into ICD. Orpha numbers are now increasingly used by European healthcare systems for informatics tracing of RGDs, and their introduction is fostered by National Action Plans and Strategies for Rare Diseases and recommended by the European Commission expert group on rare diseases.[22]

OMIM has also played a central role in the naming and classification of Mendelian diseases by defining recognizable patterns of features and highlighting those that allow one condition to be distinguished from another. In general, OMIM creates separate phenotype entries on the basis of molecular etiology, that is, genetic heterogeneity. OMIM's clinical synopsis for each phenotype includes only those features that have been reported in individuals with mutations in the disease-associated gene. Each OMIM phenotype is assigned a unique and stable identifier (MIM number) that is used in the aforementioned databases and in the biomedical literature. The IRDiRC strongly supports the continued interoperability between the rare-disease nosologies ORDO and OMIM, both of which are recognized for rare-disease classification.

Standards, Tools, and Resources to Facilitate Genomic Data Analyses

Our ability to analyze, annotate, and ultimately share genomic datasets is fundamental to the RGD research agenda. Currently, tools and methods for analysis and annotation are not standardized and lack interoperability; as a result, the sharing of outputs from large genomic datasets is hampered. Pipelines for analyzing DNA sequences still have much room for improvement in terms of sequence alignment, variant calling, and functional annotation and prediction, especially for more complex variation such as insertions, deletions, and the wide spectrum of structural variants,[23] calling for a harmonized approach. This observation is supported by recent data suggesting that the limited yield of WES as reported in the literature, at least in the context of certain recessive diseases, is mostly accounted for by our limited ability to correctly call variants.[24] An example of such a platform has been developed by the RD-Connect EU project for research and diagnosis, together with the EURenOmics and NeurOmics RGD research projects. Furthermore, existing tools will need to be made interoperable and widely adopted, and their curation and updates should be duly coordinated.

Genomic data analyses for RGD discovery are also challenged by the identification of rare variants to be prioritized for further interpretation. Investigators studying the causes of RGDs are relying heavily on WES datasets compiled by consortia, such as the Exome Aggregation Consortium (ExAC; 60,000 exomes) and the NHLBI Exome Sequencing Project (ESP; 6,500 exomes), that investigate different diseases as reference datasets for analyses, and this is proving useful in decreasing the number of variants to a manageable number for certain populations. However, many of these first comparative exome datasets have been generated from populations of Western European and North American origin. This limits pathogenic variant discovery, especially from populations that have been sparsely assessed, if sampled at all. The 1000 Genomes Project has made significant contributions to our understanding of the architecture of the human genome as a large heterogeneous population dataset. Most recently, gnomAD has aggregated 15,000 genomes and 120,000 exomes, including data from the 1000 Genomes Project and the ExAC and ESP exome datasets. Increasing such population datasets and generating and sharing datasets from populations with little to no representation in existing repositories that can be used by the RGD research community, as well as others investigating human health, will be of great importance in the future. The Global Alliance for Genomics and Health (GA4GH) is active in this space and is committed to enabling responsible and effective sharing of genomic and clinical data through a federated ecosystem approach; we support these efforts and their application to RGDs.[25] For example, the Beacon Network, a demonstration project of GA4GH, is a global search engine for genetic

Table 1. Factors Contributing to Bottlenecks in the Gene-Discovery Pipeline

Clinical data	• non-specific clinical presentations (e.g., developmental delay and hypotonia) • ultra-rare and unrecognized genetic diseases • lack of ontology encompassing the complete spectrum of human phenotypes • insufficient utilization of ontologies or 3D facial-gestalt analysis in phenotyping • inconsistent multidisciplinary approaches to patient evaluation • inability to account for and compare age-specific disease presentations
Genomic data	• technical limitations of WES (e.g., copy-number variants and structural variation are not captured well) • lack of standardized technical and informatics approaches • incompleteness of population-specific control datasets
Data discovery and sharing	• lack of a widely adopted data-sharing framework • lack of common data-sharing standards • lack of a systematic way to record data-use conditions • lack of a privacy-preserving linkage system for each research participant
Genetic evidence	• siloed datasets • lack of and use of data-sharing infrastructure
Functional evidence	• lack of standardized and moderate-throughput analyses of variant impact • lack of biological insight into the function of most human genes
Novel disease mechanisms	• lack of expertise in the analysis of non-coding variants • other mechanisms including tissue-specific mosaicism, methylation, and di- or oligogenic inheritance

variation and connects 60 databases representing every inhabited continent, enabling global discovery of genetic variation.

Ethical Standards to Enable Data Discovery and Sharing

The RGD research community is acutely and universally aware of the need for data discovery and sharing.[26] Given the challenge ahead of us to understand and be able to diagnose RGDs of ever increasing rarity, the ability to share clinical and genetic data maximally has become of central importance. In this regard, the IRDiRC is collaborating with the Human Variome Project (HVP) and GA4GH to tackle major ethical, legal, and social issues and agree on standards for international data to break down existing hurdles. The IRDiRC has recognized the Framework for Responsible Sharing of Genomic and Health-Related Data[27] as a resource on the basis of international adherence to Article 27 of the UN Declaration of Human Rights, which holds that everyone has a right "to share in scientific advancement and its benefits" and "to the protection of the moral and material interests resulting from any scientific … production of which [a person] is the author."[28] Recently, recommendations and models for "Data Transfer Agreements" have been published with the "IRDiRC recognized" label.[29]

The IRDIRC-HVP-GA4GH collaboration is paving the way for international recognition of common data-sharing standards. Several critical areas of data-sharing governance are currently the focus of collaborative efforts. First, the collaboration developed a "tiered" consent policy that is dependent on the context of data collection and use (clinical or research) and on the level of risk that the shared data will be identified; this policy is currently in use by the Matchmaker Exchange[30,31](MME; see below). Two related initiatives, namely the Consent Codes[32] model and the Automatable Discovery and Access Matrix (ADA-M), seek to enable systematized representation of consent-, legal-, and institutional-based permissions and restrictions associated with research and clinical records to facilitate streamlined and appropriate discovery, sharing, and use of extant datasets. This will also help to better standardize consent-form clauses, thereby guiding best practices in both research and ethics review committees. Just as consent practices need to become interoperable so as to enable greater data sharing, so too do data-access mechanisms. Efforts are currently underway to produce a new model that would facilitate data access (registered access) and use interactions with initiatives such as MME by authorizing users through a standard online authentication and attestation process. Registered access will address different categories of potential data users (researchers, clinical care professionals, and patients), as well as different levels of data depending on their identifiability and sensitivity. Additional IRDiRC-GA4GH collaboration is underway to develop a privacy-preserving linkage system that would link data from the same individual across multiple projects while also respecting privacy. Policy for recognizing ethics review to encourage streamlined and coherent ethics review for international projects and consortia is also available. Over time, such efforts will harmonize local ethical, legal, and social policies and procedures for efficient and responsible international sharing and analysis of genomic and clinical data.

Genetic Evidence to Support Gene Discovery

Reports from several large-scale collaborative research initiatives, including the FORGE Canada Consortium,[33] US Centers for Mendelian Genomics,[8] and UK Deciphering of Developmental Disorders study,[34] indicate that under very select circumstances (including ascertainment of multiple, thoroughly phenotyped families with the same condition), the "solve rate" for RGDs is often >50%. Reports focusing on disease-causing variants in known

disease-related genes in over 9,000 cases from various clinical diagnostic settings indicate an overall success rate of ~30%.[35–39] These latter cohorts have demonstrated that a substantial fraction (25%–30%) of clinical diagnostic success depends on recent progress in the discovery of genes underlying disease. This observation in combination with the higher solve rate in the research setting suggests that the unsolved fractions of these clinical cohorts contain many discoveries.

Case-Based Matching for Gene Discovery. The discovery of disease-gene associations requires confirmation of pathogenic genomic variation in multiple unrelated individuals affected by the same rare disease. Our collective experience suggests that it takes approximately 2–3 years to identify an additional unrelated individual with likely pathogenic mutations in the same gene after publication of a single patient or family. Thus, a central challenge is to efficiently identify additional and unrelated persons with pathogenic variant(s) in the same gene and an overlapping phenotype. It is difficult to gauge the number of such single surviving candidate genes (containing deleterious-appearing genetic variation that remains after multiple filtering steps with segregation data and pathway and/or model-organism support from existing literature) that remain unpublished and/or in inaccessible "silos" worldwide, but we estimate it to be more than 1,000.

To address this challenge, several collaborative initiatives have developed platforms for genotype- and phenotype-driven matching algorithms[12,13,40–52]; however, a connection between these existing solutions has been lacking. Very recently, the IRDiRC Diagnostics Scientific Committee, in collaboration with each participating data-sharing service, Can-SHARE, and the GA4GH, has contributed to launching a federated platform termed the MME.[53] This platform facilitates the identification of unsolved patients and families with similar phenotypic and genotypic profiles through a standardized application programming interface (API) and standard operating procedures.[40] The MME enables searches of multiple databases at once, circumventing the need to separately search all services by depositing data in each one. Under this initial API, each server can treat any description arbitrarily: the level of similarity required (on either the genotype or phenotype level) before a match is triggered is left to the discretion of each service. The launch of the MME is a major step forward, and currently PhenomeCentral,[12] GeneMatcher,[41] DECIPHER,[13] MyGene2,[54] *matchbox*, and Patient Archive, representing data from more than 20,000 unrelated RGD patients, are connected to one another. However, truly optimizing this type of case-based matching and enable RGD discovery on a global scale will require improvement of international data sharing, optimization, financial support, and scaling up of such infrastructure, operating procedures, and algorithms.

Functional Evidence to Support Gene Discovery

Integration of Genomic Data into Systems Biology. Parallel to the enormous advances in gene identification through WES, other large-scale -omics approaches have been developed (e.g., proteomics, transcriptomics, and metabolomics) to aid RGD discovery and facilitate the validation of variants of unknown significance. For instance, changes in protein levels or function help to identify the disease-causing variant if more than one plausible gene has been identified through WES. Data integration across different -omics datasets on population or individual patient levels will also be required for understanding the importance of disease-modifying variants in conditions with high phenotypic variability or incomplete penetrance and for assisting the development of diagnostics and therapeutic biomarkers and will play an increasing role in developing targeted therapies. For example, RD-Connect is establishing a platform where genomic data on rare disease patients are combined with other -omics data and standardized phenotypes.[55] Such initiatives need to be increased in number and made sustainable.

Model Systems to Facilitate Gene Discovery. Model-systems research (in humans, yeast, flies, worms, zebrafish, mice, and other organisms) will continue to be critical in determining the functional consequences of genomic variants in candidate disease-related genes and in discovering and validating new drug targets, candidate drugs, and other therapeutic strategies. The pace of allele discovery is outstripping our ability to understand the biological consequences of individual mutations on gene, pathway, and network function. There is an opportunity for the next generation of disease modeling to address this gap in an efficient, cost-effective, and generalizable manner with higher throughput. Improved infrastructure is required for (1) allowing clinician scientists who have discovered a disease-causing variant to be exposed to the full range of experimental tools available to them, (2) allowing experts in a variety of model organisms to apply their skills on pertinent questions of biological and clinical interest, and (3) creating efficiencies so that studies are not duplicated and existing models are utilized to their full potential. Linking clinician scientists and basic researchers early and providing seed funds for collaborative experiments would be the ultimate goals of such an effort.

One approach to accelerating collaborations between clinicians and basic researchers is to proactively identify collaborative "matches" and to provide seed funding to ignite collaborative research projects. In Canada, a national infrastructure, the "Rare Diseases: Models and Mechanisms" network, has been established to link clinicians and basic researchers as soon as disease-related genes are discovered.[56] The network is in its second year of its 3 year funding cycle and has been successful in catalyzing collaborative links for over 40 clinician and basic-scientist matches. An alternative approach is through an "enabling"

scheme, in which national funding agencies allow investigators to jointly apply for supplemental funding to existing grants. In the US, for example, administrative supplements to "R" and "P" grants are not uncommon; indeed, this model has been used by the NIH Undiagnosed Disease Program to seed research on candidate genes discovered by that effort.[57] An integrated international virtual network allowing clinician scientists to discover relevant researchers might also be a complementary and intermediate approach.

It will also be important to stimulate the establishment and validation of novel phenotyping pipelines that have correlates in other organisms by emphasizing disease relevance, pathophysiological pathways, and high efficiency. This will accelerate the evaluation of genomic variants and candidate genes, drug and drug-target testing using disease-relevant output measures, and fundamental understanding of disease mechanisms and pathologies. Phenotyping pipelines can, in some cases, assess disease traits that resemble hallmarks of the human disorder in an obvious manner (e.g., malformations, behaviorial features, or other findings). If sufficiently specific (i.e., unique), such phenotypes can validate the relevance of a disease model. The Monarch Initiative has been working in this realm since 2009 and acts as an integrative data and analytic platform that connects phenotypes and genotypes across species. Alternatively, phenotyping pipelines can assess traits that are not linked to the disease of interest in an obvious manner but that do result from the same molecular defects underlying the disease phenotype in humans and thus represent orthologous phenotypes ("phenologs").[58] In addition, it will be important to develop and validate novel efficient and disease-relevant test paradigms and phenotypes that can be cross-compared between species (parallel phenotyping). Such validated disease-relevant phenotypes across organisms could provide the required output measures for overcoming current bottlenecks, such as the validation of alleles and disease-related genes, at a scale that is urgently required in the post-genome-sequence era.

Novel Disease Mechanisms

Progress toward the discovery of the genetic basis of every RGD has been substantial over the past several years. Yet, there remain a non-trivial number of well-known rare diseases (e.g., Hallerman-Streiff syndrome, Dubowitz syndrome, VACTERL, Gomez-Lopez-Hernandez syndrome, Aicardi syndrome, and PHACE syndrome) for which, despite multiple groups' efforts to use WES and, in some cases, WGS, the causal genetic mechanism remains elusive. The reasons that such discovery efforts fail are myriad and most likely include both technical limitations (e.g., annotation errors, missed coding and non-coding variation, and structural variation) and complex biology (e.g., extreme locus heterogeneity, tissue-specific somatic mosaicism, unusual modes of inheritance, intrafamilial allelic or locus heterogeneity, and causal synonymous variants). Approaches that overcome these barriers to RGD discovery are few in number. Moreover, the rare genetic conditions for which the genetic mechanism has yet to be identified are likely enriched with those that will not be solved easily by existing WES-based approaches. Identifying the molecular basis of conditions intractable to existing approaches requires broader and innovative application of existing discovery strategies (e.g., WGS, RNA sequencing of affected cells or tissues, and deep sequencing of tissues derived from the three major embryonic lineages); improvement of computational and statistical models for variant identification, annotation, functional prediction, and prioritization—particularly for variants in non-coding regions;[59] and development of strategies for discovering causal genetic mechanisms. Also, temporally focused, multidisciplinary assessments that take advantage of cumulative expert clinician experience and precision phenotyping centered around single patients, such as the Undiagnosed Diseases Network International,[60] are part of a suite of approaches to supporting the discovery of rare-disease mechanisms. The development and application of these strategies will further leverage investments that support genetic and functional approaches for the discovery of underlying genetic mechanisms.

Critical Next Steps

Achieving the IRDiRC's goal of a means of diagnosing all RGDs will require the discovery of the genetic mechanism underlying every disorder. This challenge—producing a complete catalog of the phenotypic characteristics of all RGDs and their corresponding causal variants, developing successful approaches to discovering the underlying etiology of RGDs caused by non-traditional modes of inheritance, and establishing tools and resources to translate this new knowledge into patient care (e.g., harmonization and adoption of international guidelines for the clinical application of NGS-based approaches)—is significant. This grand challenge can be achieved only with significant international cooperation and engagement of all relevant stakeholders at a scale the community has never seen before. Efforts to engage the research community, such as the IRDiRC and GA4GH, are of critical importance, and international coordination and funding of activities will be necessary. Improving translation and reimbursement strategies for clinical genome-wide analysis of patients with rare diseases will be essential; this is particularly important for avoiding the large number of pathogenic variants identified in known genes in research projects focused on discovery and reallocating research funding to the generation and validation of novel insights. Engaging clinical laboratories, researchers, and the patient community to share their data will be critical.

We must also recognize that as more and more genes are discovered to be associated with human disease and appropriate analytical tests are

established, a significant challenge in RGD diagnosis will remain: that of interpreting a growing numbers of variants of uncertain significance. DNA diagnostics for RGD is primarily based on shared knowledge about genes, genomic variation, and phenotypes. Currently, diagnostic data are collected through a multitude of approaches by many different diagnostic laboratories and are stored in a wide variety of server systems and databases, which generally lack federated connections, i.e., "silos." Local solutions need to be developed and implemented for storing data on genetic variants and their associated phenotypes in an easy and reproducible way with common standards and terminologies. In addition, these local systems need to be connected worldwide to form a "genetics knowledge web." Making this type of sharing part of the normal standard of care will require community engagement. Integrating existing platforms that store clinical genetic and phenotype data (e.g., ClinVar,[61] Leiden Open Variation Database [LOVD],[62] and DECIPHER[13]), linking different types of data (e.g., array and sequencing), and encompassing small (single-nucleotide) to large (deletion, duplication, inversion, etc.) variants will be essential. These challenges are further compounded by the rate and impact of false-positive causative variant assignments[63] that exist in such databases, so ultimately the curation of this knowledge by relevant experts will be the key to diagnostic precision. Variant classification as pathogenic or benign will rely heavily on the same tools that are critically needed for RGD discovery, specifically the availability of population-specific disease and control databases for a diverse range of populations, the use of orthogonal assays such as metabolomics, transcriptomics, or proteomics to clarify functional effect, and the systematic screening of mutations in disease-related genes in tractable models or cell systems. Clearly, the task of assigning pathogenicity to individual variants is mission critical to informed patient care.

Achieving a means of diagnosing all RGDs will be of great importance for patients and families. It will allow genetic counseling, better prognostication, and identification of specific health risks to the individual and will prevent unnecessary or harmful diagnostic interventions and treatments. Ultimately, such insights can be applied to genome-wide sequencing in newborns for both diagnosis and screening.[64] In an increasing number of patients, effective drug treatment is available once the exact diagnosis (e.g., lysosomal-storage disorders or congenital myasthenic syndromes) has been established.[65] In addition, this aim will allow more patients to participate in research cohorts for clinical trials that require a definite molecular and phenotypic diagnosis, providing potential benefit from new drugs or interventions being developed by academia and the private sector.[66] In our view, the understanding of all RGDs will be the cornerstone of precision medicine; the power of genomics to explain these rare diseases with concomitant fundamental insights into biological processes will rapidly transform medical care for these patients and their families.

Acknowledgments

We thank the present and former staff of the IRDiRC Scientific Secretariat, Ségolène Aymé, Lilian Lau, Anneliene Jonker, Antonia Mills, Barbara Cagniard, and members of the scientific committees, working groups, and task forces. We also thank the former chairs of the IRDiRC Therapies Scientific Committee (Yann Le Cam and Josep Torrent i Farnell) for helpful discussions. We thank Rachel Thompson, Emma Heslop, Victoria Hedley, and Lena Dolman for their support. Orphanet was supported in part by funding from AFM-Téléthon and the Joint action "677024/RD-ACTION," which received funding from the European Union's Health Programme (2014-2020). T.H. and M.R.V. were supported by the Care4Rare Canada Consortium, funded by Genome Canada, the Canadian Institutes of Health Research, the Ontario Genomics Institute, Ontario Research Fund, Genome Quebec, and Children's Hospital of Eastern Ontario Foundation. J.X.C. and M.J.B. were supported by the University of Washington Center for Mendelian Genomics though the National Human Genome Research Institute (NHGRI) and National Heart, Lung, and Blood Institute (grant U54HG006493) to Drs. Debbie Nickerson, Michael Bamshad, and Suzanne Leal. A.H. was supported by OMIM through NHGRI grant 1U41HG006627 and the Baylor-Hopkins Center for Mendelian Genomics through NHGRI grant 1U54HG006542 to Drs. David Valle and James Lupski. M.M. was supported by funding from 00064203, 16-30880A, 15-34904A, OPPK CZ.2.16/3.1.00/24022, and NF-CZ11-PDP-3-003-2014. H.L.R. was supported by UM1HG008900. H.L. was supported by funding from the Medical Research Council (reference G1002274, grant 98482), the Wellcome Trust (reference 201064/Z/16/Z), and the European Union Seventh Framework Programme (FP7/2007-2013) under grants 305444 (RD-Connect) and 305121 (NeurOmics).

Web Resources

1000 Genomes, http://www.1000genomes.org
Can-SHARE, http://www.p3g.org/resources/can-share
ClinGen, https://www.clinicalgenome.org
ClinVar, http://www.ncbi.nlm.nih.gov/clinvar
DECIPHER, https://decipher.sanger.ac.uk
EURenOmics, http://eurenomics.eu
Exome Aggregation Consortium (ExAC) Browser, http://exac.broadinstitute.org
GeneMatcher, https://genematcher.org
Genomics England, https://www.genomicsengland.co.uk
Genotype to Mendelian Phenotype (Geno2MP), http://geno2mp.gs.washington.edu
Global Alliance for Genomics and Health (GA4GH), http://genomicsandhealth.org
gnomAD, http://gnomAD.broadinstitute.org
Human Phenotype Ontology (HPO), http://www.human-phenotype-ontology.org
Human Variome Project, http://www.humanvariomeproject.org/
International Consortium of Human Phenotype Ontologies (ICHPT), http://www.irdirc.org/ichpt
International Rare Diseases Research Consortium (IRDiRC), http://www.irdirc.org
Leiden Open Variation Database (LOVD), http://www.lovd.nl/3.0/home
Matchbox, https://seqr.broadinstitute.org

Matchmaker Exchange (MME), http://www.matchmakerexchange.org
Medical Subject Headings (MeSH), http://www.ncbi.nlm.nih.gov/mesh
Medical Dictionary for Regulatory Activities (MedDRA), https://www.meddra.org
Monarch Initiative, https://monarchinitiative.org
MyGene2, http://mygene2.org
NeurOmics, http://rd-neuromics.eu
NHLBI Exome Sequencing Project (ESP) Exome Variant Server, http://evs.gs.washington.edu/EVS
OMIM, http://omim.org
Orphanet, http://www.orpha.net
Orphanet Rare Disease Ontology, http://bioportal.bioontology.org/ontologies/ORDO
Orphanet RD-Action, http://www.rd-action.eu
Patient Archive, http://patientarchive.org
PhenoDB, https://phenodb.org
PhenomeCentral, https://www.phenomecentral.org
POSSUM, http://www.possum.net.au
Rare Diseases Models and Mechanisms Network (RDMM), http://rare-diseases-catalyst-network.ca
RD-Connect, https://platform.rd-connect.eu
SNOMED CT, http://www.ihtsdo.org/snomed-ct
UK10K, http://www.uk10k.org

References

1. United States Congress. (2002). Rare Diseases Act of 2002. https://www.gpo.gov/fdsys/pkg/PLAW-107publ280/html/PLAW-107publ280.htm.
2. The European Parliament and the Council of the European Union (1999). Regulation (EC) No 141/2000 of the European parliament and of the council of 16 December 1999 on orphan medicinal products. http://ec.europa.eu/health//sites/health/files/files/eudralex/vol-1/reg_2000_141_cons-2009-07/reg_2000_141_cons-2009-07_en.pdf.
3. Ana Rath, ed. (2014). Prevalence and incidence of rare diseases: Bibliographic data. In Orphanet Report Series: Rare Diseases collection. http://www.orpha.net/orphacom/cahiers/docs/GB/Prevalence_of_rare_diseases_by_alphabetical_list.pdf.
4. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). Nucleic Acids Res. *43*, D789–D798.
5. Ramsey, B.W., Davies, J., McElvaney, N.G., Tullis, E., Bell, S.C., Dřevínek, P., Griese, M., McKone, E.F., Wainwright, C.E., Konstan, M.W., et al.; VX08-770-102 Study Group (2011). N. Engl. J. Med. *365*, 1663–1672.
6. Rath, A., Olry, A., Dhombres, F., Brandt, M.M., Urbero, B., and Ayme, S. (2012). Hum. Mutat. *33*, 803–808.
7. Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al.; ClinGen (2015). N. Engl. J. Med. *372*, 2235–2242.
8. Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., McMillin, M.J., Wiszniewski, W., Gambin, T., et al.; Centers for Mendelian Genomics (2015). Am. J. Hum. Genet. *97*, 199–215.
9. Lochmüller, H., Le Cam, Y., Jonker, A.H., Lau, L.P., Baynam, G., Kaufmann, P., Lasko, P., Dawkins, H.J., Austin, C.P., and Boycott, K.M. (2017). Eur. J. Hum. Genet. *25*, 162–165.
10. Köhler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C., Brown, D.L., Brudno, M., Campbell, J., et al. (2014). Nucleic Acids Res. *42*, D966–D974.
11. Robinson, P.N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). Am. J. Hum. Genet. *83*, 610–615.
12. Buske, O.J., Girdea, M., Dumitriu, S., Gallinger, B., Hartley, T., Trang, H., Misyura, A., Friedman, T., Beaulieu, C., Bone, W.P., et al. (2015). Hum. Mutat. *36*, 931–940.
13. Chatzimichali, E.A., Brent, S., Hutton, B., Perrett, D., Wright, C.F., Bevan, A.P., Hurles, M.E., Firth, H.V., and Swaminathan, G.J. (2015). Hum. Mutat. *36*, 941–949.
14. Peplow, M. (2016). BMJ *353*, i1757.
15. Köhler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dölken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P.N. (2009). Am. J. Hum. Genet. *85*, 457–464.
16. Zemojtel, T., Köhler, S., Mackenroth, L., Jäger, M., Hecht, J., Krawitz, P., Graul-Neumann, L., Doelken, S., Ehmke, N., Spielmann, M., et al. (2014). Sci. Transl. Med. *6*, 252ra123.
17. Robinson, P.N., Köhler, S., Oellrich, A., Wang, K., Mungall, C.J., Lewis, S.E., Washington, N., Bauer, S., Seelow, D., Krawitz, P., et al.; Sanger Mouse Genetics Project (2014). Genome Res. *24*, 340–348.
18. Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., et al. (2017). Nucleic Acids Res. *45* (D1), D865–D876.
19. Hamosh, A., Sobreira, N., Hoover-Fong, J., Sutton, V.R., Boehm, C., Schiettecatte, F., and Valle, D. (2013). Hum. Mutat. *34*, 566–571.
20. Allanson, J.E., Biesecker, L.G., Carey, J.C., and Hennekam, R.C. (2009). Am. J. Med. Genet. A. *149A*, 2–5.
21. Vasant, D., Chanas, L., Malone, J., Hanauer, M., Olry, A., Jupp, S., Robinson, P.N., Parkinson, H., and Rath, A. (2014). ORDO: An Ontology Connecting Rare Disease, Epidemiology and Genetic Data. Proceedings of the 22nd Annual International Conference on Intelligent Systems for Molecular Biology. https://www.researchgate.net/publication/281824026_ORDO_An_Ontology_Connecting_Rare_Disease_Epidemiology_and_Genetic_Data.
22. European Commission (2009) National plans or strategies for rare diseases. http://ec.europa.eu/health/rare_diseases/national_plans/detailed_en.
23. Alioto, T.S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M.D., Hovig, E., Heisler, L.E., Beck, T.A., Simpson, J.T., Tonon, L., et al. (2015). Nat. Commun. *6*, 10001.
24. Shamseldin, H.E., Maddirevula, S., Faqeih, E., Ibrahim, N., Hashem, M., Shaheen, R., and Alkuraya, F.S. (2016). Genet. Med. Published October 6, 2016. http://dx.doi.org/10.1038/gim.2016.155.
25. Global Alliance for Genomics and Health (2016). Science *352*, 1278–1280.
26. Brookes, A.J., and Robinson, P.N. (2015). Nat. Rev. Genet. *16*, 702–715.
27. Knoppers, B.M. (2014). HUGO J. *8*, 3.
28. United Nations (1948). Universal declaration of human rights. http://www.un.org/en/universal-declaration-human-rights/index.html.
29. Mascalzoni, D., Dove, E.S., Rubinstein, Y., Dawkins, H.J., Kole, A., McCormack, P., Woods, S., Riess, O., Schaefer, F., Lochmüller, H., et al. (2015). Eur. J. Hum. Genet. *23*, 721–728.
30. The Matchmaker Exchange (2016) Matchmaker Exchange Tiered Informed Consent Proposal. http://www.matchmakerexchange.org/assets/files/MatchmakerExchangeTieredInformedConsentProposal.pdf.
31. Dyke, S.O.M., Knoppers, B.M.K., Hamosh, A., Hurles, M., Firth, H., Brudno,

M., Boycott, K.M., Philippakis, A., and Rehm, H. (2017). Hum. Mutat., in press.
32. Dyke, S.O., Philippakis, A.A., Rambla De Argila, J., Paltoo, D.N., Luetkemeier, E.S., Knoppers, B.M., Brookes, A.J., Spalding, J.D., Thompson, M., Roos, M., et al. (2016). PLoS Genet. *12*, e1005772.
33. Beaulieu, C.L., Majewski, J., Schwartzentruber, J., Samuels, M.E., Fernandez, B.A., Bernier, F.P., Brudno, M., Knoppers, B., Marcadier, J., Dyment, D., et al.; FORGE Canada Consortium (2014). Am. J. Hum. Genet. *94*, 809–817.
34. Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., van Kogelenberg, M., King, D.A., Ambridge, K., Barrett, D.M., Bayzetinova, T., et al.; DDD study (2015). Lancet *385*, 1305–1314.
35. Yang, Y., Muzny, D.M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C., et al. (2014). JAMA *312*, 1870–1879.
36. Lee, H., Deignan, J.L., Dorrani, N., Strom, S.P., Kantarci, S., Quintero-Rivera, F., Das, K., Toy, T., Harry, B., Yourshaw, M., et al. (2014). JAMA *312*, 1880–1887.
37. Saudi Mendeliome Group (2015). Genome Biol. *16*, 134.
38. Farwell, K.D., Shahmirzadi, L., El-Khechen, D., Powis, Z., Chao, E.C., Tippin Davis, B., Baxter, R.M., Zeng, W., Mroske, C., Parra, M.C., et al. (2015). Genet. Med. *17*, 578–586.
39. Retterer, K., Juusola, J., Cho, M.T., Vitazka, P., Millan, F., Gibellini, F., Vertino-Bell, A., Smaoui, N., Neidich, J., Monaghan, K.G., et al. (2016). Genet. Med. *18*, 696–704.
40. Buske, O.J., Schiettecatte, F., Hutton, B., Dumitriu, S., Misyura, A., Huang, L., Hartley, T., Girdea, M., Sobreira, N., Mungall, C., and Brudno, M. (2015). Hum. Mutat. *36*, 922–927.
41. Sobreira, N., Schiettecatte, F., Valle, D., and Hamosh, A. (2015). Hum. Mutat. *36*, 928–930.
42. Gonzalez, M., Falk, M.J., Gai, X., Postrel, R., Schüle, R., and Zuchner, S. (2015). Hum. Mutat. *36*, 950–956.
43. Lancaster, O., Beck, T., Atlan, D., Swertz, M., Thangavelu, D., Veal, C., Dalgleish, R., and Brookes, A.J. (2015). Hum. Mutat. *36*, 957–964.
44. Lambertson, K.F., Damiani, S.A., Might, M., Shelton, R., and Terry, S.F. (2015). Hum. Mutat. *36*, 965–973.
45. Kirkpatrick, B.E., Riggs, E.R., Azzariti, D.R., Miller, V.R., Ledbetter, D.H., Miller, D.T., Rehm, H., Martin, C.L., Faucett, W.A.; and ClinGen Resource (2015). Hum. Mutat. *36*, 974–978.
46. Mungall, C.J., Washington, N.L., Nguyen-Xuan, J., Condit, C., Smedley, D., Köhler, S., Groza, T., Shefchek, K., Hochheiser, H., Robinson, P.N., et al. (2015). Hum. Mutat. *36*, 979–984.
47. Brownstein, C.A., Holm, I.A., Ramoni, R., Goldstein, D.B.; and Members of the Undiagnosed Diseases Network (2015). Hum. Mutat. *36*, 985–988.
48. Krawitz, P., Buske, O., Zhu, N., Brudno, M., and Robinson, P.N. (2015). Hum. Mutat. *36*, 989–997.
49. Akle, S., Chun, S., Jordan, D.M., and Cassa, C.A. (2015). Hum. Mutat. *36*, 998–1003.
50. Jurgens, J., Sobreira, N., Modaff, P., Reiser, C.A., Seo, S.H., Seong, M.W., Park, S.S., Kim, O.H., Cho, T.J., and Pauli, R.M. (2015). Hum. Mutat. *36*, 1004–1008.
51. Au, P.Y., You, J., Caluseriu, O., Schwartzentruber, J., Majewski, J., Bernier, F.P., Ferguson, M., Valle, D., Parboosingh, J.S., Sobreira, N., et al. (2015). Hum. Mutat. *36*, 1009–1014.
52. Loucks, C.M., Parboosingh, J.S., Shaheen, R., Bernier, F.P., McLeod, D.R., Seidahmed, M.Z., Puffenberger, E.G., Ober, C., Hegele, R.A., Boycott, K.M., et al. (2015). Hum. Mutat. *36*, 1015–1019.
53. Philippakis, A.A., Azzariti, D.R., Beltran, S., Brookes, A.J., Brownstein, C.A., Brudno, M., Brunner, H.G., Buske, O.J., Carey, K., Doll, C., et al. (2015). Hum. Mutat. *36*, 915–921.
54. Chong, J.X., Yu, J.H., Lorentzen, P., Park, K.M., Jamal, S.M., Tabor, H.K., Rauch, A., Saenz, M.S., Boltshauser, E., Patterson, K.E., et al. (2016). Genet. Med. *18*, 788–795.
55. Thompson, R., Johnston, L., Taruscio, D., Monaco, L., Béroud, C., Gut, I.G., Hansson, M.G., 't Hoen, P.B., Patrinos, G.P., Dawkins, H., et al. (2014). J. Gen. Intern. Med. *29* (*Suppl 3*), S780–S787.
56. Foley, K.E. (2015). Nat. Med. *21*, 1242–1243.
57. Gahl, W.A., Wise, A.L., and Ashley, E.A. (2015). JAMA *314*, 1797–1798.
58. McGary, K.L., Park, T.J., Woods, J.O., Cha, H.J., Wallingford, J.B., and Marcotte, E.M. (2010). Proc. Natl. Acad. Sci. USA *107*, 6544–6549.
59. Smedley, D., Schubach, M., Jacobsen, J.O., Köhler, S., Zemojtel, T., Spielmann, M., Jäger, M., Hochheiser, H., Washington, N.L., McMurry, J.A., et al. (2016). Am. J. Hum. Genet. *99*, 595–606.
60. Taruscio, D., Groft, S.C., Cederroth, H., Melegh, B., Lasko, P., Kosaki, K., Baynam, G., McCray, A., and Gahl, W.A. (2015). Mol. Genet. Metab. *116*, 223–225.
61. Harrison, S.M., Riggs, E.R., Maglott, D.R., Lee, J.M., Azzariti, D.R., Niehaus, A., Ramos, E.M., Martin, C.L., Landrum, M.J., and Rehm, H.L. (2016). Curr. Protoc. Hum. Genet. *89*, 8.16.1–8.16.23.
62. Fokkema, I.F., Taschner, P.E., Schaafsma, G.C., Celli, J., Laros, J.F., and den Dunnen, J.T. (2011). Hum. Mutat. *32*, 557–563.
63. Dorschner, M.O., Amendola, L.M., Turner, E.H., Robertson, P.D., Shirts, B.H., Gallego, C.J., Bennett, R.L., Jones, K.L., Tokita, M.J., Bennett, J.T., et al.; National Heart, Lung, and Blood Institute Grand Opportunity Exome Sequencing Project (2013). Am. J. Hum. Genet. *93*, 631–640.
64. Berg, J.S., Agrawal, P.B., Bailey, D.B. Jr., Beggs, A.H., Brenner, S.E., Brower, A.M., Cakici, J.A., Ceyhan-Birsoy, O., Chan, K., Chen, F., et al. (2017). Pediatrics *139*, e20162252.
65. Evangelista, T., Hanna, M., and Lochmüller, H. (2015). J Neuromuscul Dis *2* (*Suppl 2*), S21–S29.
66. Bladen, C.L., Rafferty, K., Straub, V., Monges, S., Moresco, A., Dawkins, H., Roy, A., Chamova, T., Guergueltcheva, V., Korngut, L., et al. (2013). Hum. Mutat. *34*, 1449–1457.

Introducing Sneak Peek, a preview of Cell Press papers under review

Exciting science needs to be shared, and fast. That's why we're introducing Cell Press Sneak Peek, a preview of the papers under review at our primary research journals. Sneak Peek makes papers discoverable earlier in the publication process—so authors can surface their research quickly and readers can build on their work.

Go on, satisfy your curiosity!
Visit www.cell.com/sneakpeek

CellPress

Sneak Peek
A PREVIEW OF PAPERS UNDER REVIEW

InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines

Quan Li[1,4] and Kai Wang[1,2,3,*]

In 2015, the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) published updated standards and guidelines for the clinical interpretation of sequence variants with respect to human diseases on the basis of 28 criteria. However, variability between individual interpreters can be extensive because of reasons such as the different understandings of these guidelines and the lack of standard algorithms for implementing them, yet computational tools for semi-automated variant interpretation are not available. To address these problems, we propose a suite of methods for implementing these criteria and have developed a tool called InterVar to help human reviewers interpret the clinical significance of variants. InterVar can take a pre-annotated or VCF file as input and generate automated interpretation on 18 criteria. Furthermore, we have developed a companion web server, wInterVar, to enable user-friendly variant interpretation with an automated interpretation step and a manual adjustment step. These tools are especially useful for addressing severe congenital or very early-onset developmental disorders with high penetrance. Using results from a few published sequencing studies, we demonstrate the utility of InterVar in significantly reducing the time to interpret the clinical significance of sequence variants.

Introduction

With the continued development and deployment of massively parallel next-generation sequencing (NGS) technologies, clinical and molecular laboratories are now rapidly adopting NGS in genetic testing and human genetics research. Although it is becoming easier and more affordable for individual laboratories to generate NGS data, the major hurdle in utilizing these data lies in how to interpret the genotype-phenotype relationships, especially in genomic medicine settings.[1,2] The process of identifying disease-causing or disease-contributing variants among the thousands of genetic variants within an individual's genome generally involves a number of steps, such as variant annotation, variant filtering, in silico prediction, and clinical interpretation by human experts.[3] Each of these steps can involve the use of specific computational and bioinformatics tools.

Several tools and databases have been developed to assist laboratories and clinicians with understanding the functional significance of genetic variants with respect to their potential effects on genes and diseases. They generally fall into several categories. First, a number of annotation tools, such as ANNOVAR,[4,5] VAAST,[6] SeattleSeq,[7] SNPeff,[8] and VEP,[9] can predict how genetic variants affect transcript structure or coding sequences. They can classify variants into intronic, intergenic, splice, and exonic variants, and for exonic variants, they can compute how amino acid sequences are affected. Second, for coding variants, a variety of tools can predict whether the variant is deleterious to protein function or structure by using evolutionary information, context within the protein sequence, and biochemical properties. These in silico methods include individual scoring systems, such as SIFT,[10] PolyPhen-2,[11] CADD,[12] FATHMM,[13] and MutationTaster,[14] as well as meta-predictors, such as Condel[15] and MetaSVM.[16] Many have a similar theoretical basis, but they also have known limitations, such as moderate accuracy, low specificity, and over-prediction.[17,18] Third and finally, public disease-specific and gene-specific databases, such as the Human Gene Mutation Database (HGMD),[19] ClinVar,[20] and various locus-specific databases,[21] can document functionally or clinically validated genetic variants that are pathogenic for particular diseases. The HGMD is a comprehensive collection of germline mutations in nuclear genes that underlie, or are associated with, human inherited disease and is compiled primarily from the published literature.[19] ClinVar[20] archives the clinical significance of variants reported directly from submitters. However, these databases often contain variants that are incorrectly classified without a primary review of evidence, and they sometimes have contradictory records on the assessment of pathogenicity. The NIH began the ClinGen initiative[22] to build an authoritative central resource that defines the clinical relevance of genes and variants for use in precision medicine and research. To improve the accuracy of variant interpretations, ClinGen uses a ranking system to denote the quality associated with each submission to ClinVar. Despite the existence of a variety of resources, a more systematic way to evaluate the pathogenicity of genetic variants observed in sequencing studies is needed to facilitate clinical evaluation of variants and to enable the precise implementation of genomic medicine.

To standardize the clinical interpretation of genetic variants, the American College of Medical Genetics and Genomics (ACMG) recommended standards for the

[1]Zilkha Neurogenetic Institute, University of Southern California, Los Angeles, CA 90089, USA; [2]Institute for Genomic Medicine, Columbia University, New York, NY 10032, USA; [3]Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA
[4]Present address: Faculty of Medicine, Memorial University of Newfoundland, St. John's, NL A1B 3V6, Canada
*Correspondence: kw2701@cumc.columbia.edu
http://dx.doi.org/10.1016/j.ajhg.2017.01.004.

interpretation of sequence variations and offered a decision-tree algorithm for variant interpretation in 2000 and 2007.[23,24] With the rapid development and adoption of NGS, variant interpretation has become more complex, and new challenges in the clinical interpretation of Mendelian and complex diseases have emerged. To address these challenges and to provide more concrete guidelines, the ACMG and the Association for Molecular Pathology (AMP) published updated guidelines for the interpretation of sequence variants in May of 2015.[25] This new report describes updated standards and guidelines for classifying sequence variants by using criteria informed by expert opinion and experience. To better describe the causality of variants identified in genes associated with Mendelian diseases, the ACMG and AMP recommend a widely used five-tiered categorization system—pathogenic, likely pathogenic, uncertain significance, likely benign, and benign—for classifying variants. The system uses a total of 28 criteria based on different sources of data, such as population data, in silico data, functional data, and segregation data. The ACMG and AMP also propose a set of scoring rules, which combine criteria to give the five-tier classification system for genetic variants.

Although the ACMG-AMP guidelines were developed to enable consistent and reliable interpretation of genetic variants, application of the ACMG-AMP criteria still involves some discrepancies between intra- and inter-laboratory settings. Some efforts have been taken to reduce inter-laboratory inconsistencies,[26] but >66% of variant classifications are still discordant in inter-laboratory classifications. There could be several reasons for the discordances. For many clinical labs, implementing the variant scoring rules into a standardized workflow is difficult with available informatics tools. For example, the ACMG and AMP recommend using 28 criteria during the interpretation process; however, gathering information on each of the criteria is quite complicated and might not be easily accomplished by individual interpreters or might not be reproducible by the same interpreter at different times. Furthermore, the ACMG and AMP provide only general guidelines on how to assess each criterion but do not offer specific algorithms for implementing these guidelines (for example, which databases to use); different researchers might prefer to use different algorithms, making the results less reproducible between different human interpreters. Finally, although a variety of databases (such as ClinVar and the 1000 Genomes Project) or in silico tools (such as SIFT and PolyPhen-2) are available online and easily accessible to the average user, there is a lack of tools that combine all of these databases together to offer a one-stop shop for human interpreters to derive a final score for genetic variants. Addressing these challenges will require easy-to-use yet automated computational tools and web services that can generate versioned and reproducible criteria for every variant and help human interpreters quickly understand the clinical significance of genetic variants. In this study, we present such a tool, InterVar (Clinical Interpretation of Genetic Variants), to fill these unmet needs on the basis of the 2015 ACMG-AMP guidelines and user-supplied domain knowledge.

Material and Methods

Generation of Variant Annotation

The required input for InterVar is a simple tab-delimited file including a list of variants that are already annotated with a set of required information, such as amino acid changes and allele frequency. Users can generate this input file themselves by using an in-house variant analysis workflow; alternatively, InterVar can take a VCF file, call the ANNOVAR software (a powerful and widely used annotation tool), and generate the required input data. The following is an example command line for running ANNOVAR: "perl table_annovar.pl input.vcf humandb/ -buildver hg19 -remove -out output -protocol refGene,esp6500siv2_all, 1000g2015aug_all,avsnp144,dbnsfp30a,clinvar_20160302,exac03, dbscsnv11,dbnsfp31a_interpro,rmsk,ensGene,knownGene -operation g,f,f,f,f,f,f,f,f,r,g,g -nastring. -vcfinput." The description for these databases is given below: "esp6500siv2_all" is a database for allele frequency in the NHLBI Exome Sequencing Project (ESP6500), "refGene" is a database for gene annotation from RefSeq, "1000 g2015aug_all" is a database for alternative allele frequency (AAF) in the 1000 Genomes Project[27] (version August 2015), "exac03" is a database for AAF in the Exome Aggregation Consortium (ExAC) Browser[28] (version 0.3), "dbnsfp30a" is a database for various functional deleteriousness prediction scores from dbNSFP[29,30] (version 3.0a), "clinvar_20160302" is for the variants reported in ClinVar[20] (version 20160302), "avsnp144" is for the ANNOVAR-compiled dbSNP (version 144), "ensGene" is for gene annotation from Ensembl, "knownGene" is for gene annotation from UCSC Known Genes, "dbnsfp31a_interpro" is a database of the domain information from dbNSFP[29,30] and InterPro[31] (which integrates information about protein families, domains, and functional sites), "dbscsnv11" is a database for predicting the splicing impact by Ada Boost and Random Forest,[32] and "rmsk" is a database on the repeat masking track from the UCSC Genome Browser. These databases might be updated to new versions when they become available.

Criteria and Scoring System

Based on the 2015 ACMG-AMP guidelines, the criteria fall into two sets: pathogenic or likely pathogenic (P/LP) and benign or likely benign (B/LB), whereas "uncertain significance" is assigned to variants for which the criteria for P/LP and B/LB are contradictory or not met. There are a total of 28 criteria: the 16 criteria for the P/LP criterion are very strong (PVS1), strong (PS1–PS4), moderate (PM1–PM6), or supporting (PP1–PP5); whereas the 12 criteria for the B/LB criterion are stand-alone (BA1), strong (BS1–BS4), or supporting (BP1–BP7). If a criterion is positive, InterVar will assign 1; otherwise, InterVar will assign 0. For these 28 criteria, InterVar can automatically generate predictions on 18 (PVS1, PS1, PS4, PM1, PM2, PM4, PM5, PP2, PP3, PP5, BA1, BS1, BS2, BP1, BP3, BP4, BP6, and BP7) according to the current annotation datasets, yet the rest (PS2, PS3, PM3, PM6, PP1, PP4, BS3, BS4, BP2, and BP5) require user input in the manual adjustment step. Below, we describe the details on how to assign these criteria from various sources of annotation information.

PVS1 by Automated Scoring

The null variants include nonsense variants, frameshift indels, and canonical splice variants, which often lead to loss of function (LOF). From ANNOVAR annotations, these LOF variants are represented as frameshift indel, stop-gain, stop-loss, and splicing variants in canonical transcripts. We first filtered ClinVar (version 20160302) by taking those variants shown in MedGen and then removing common variants (allele frequencies > 5%) and variants with conflicting annotations. The variants in ClinVar were annotated by ANNOVAR with RefGene definitions, and we identified 1,988 genes harboring at least one LOF variant that is "pathogenic" in ClinVar. Recently, the ExAC analyzed high-quality exome (protein-coding region) DNA sequence data for 60,706 individuals and identified 3,230 genes as LOF intolerant.[28] We combined these two gene sets from ClinVar and the ExAC Browser and generated 4,807 genes as our final LOF-intolerant gene list. Null variants in the canonical transcripts for these 4,807 genes were assigned a PVS1 of 1. However, on the basis of the canonical rules for nonsense-mediated mRNA decay,[33] we did not consider nonsense variants that are downstream of or within 50 nucleotides of the final exon-junction complex.

PS1 and PM5 by Automated Scoring

Generally speaking, if one missense variant is pathogenic, then a different nucleotide change that results in the same amino acid alteration should also be pathogenic for PS1. However, if a different nucleotide change results in a different amino acid change, then it suggests moderate evidence of pathogenicity by PM5. We first filtered ClinVar (subject to the same data-cleaning procedure described above), picked out all missense variants annotated as pathogenic, and stored the amino acid changes in an InterVar-specific database. We also inferred the splicing impact of these exonic missense variants by ANNOVAR from the "dbscsnv11" database to assess the possibility that they act through splicing disruption rather than amino acid changes. If a variant supplied by the user results in the same amino acid change, the PS1 value will be assigned as 1. However, if a variant supplied by the user results in a different amino acid change, then PM5 will be assigned as 1.

PS2 and PM6 by Manual Scoring

The de novo status of the variants gives strong support for the pathogenic status for PS2 if both maternity and paternity can be confirmed; if maternity or paternity is not confirmed, then moderate evidence of pathogenicity should be applied to PM6. Because InterVar cannot directly annotate the de novo status of the user's input variants, PS2 and PM6 are treated as user-supplied values in the second step (manual adjustment) of InterVar.

PS3 and BS3 by Manual Scoring

If in vitro or in vivo functional studies are supportive of a damaging effect on the gene or gene product, PS3 should be assigned as 1. If in vitro or in vivo functional studies show no damaging effect on protein function or splicing, BS3 should be assigned as 1. InterVar does not have the information on functional studies, so by default these values are 0 and can be overridden by users. In the future, we might establish a database with validated genetic variants that are known to affect the function of genes or gene products.

BA1, BS1, BS2, PS4, and PM2 by Automated Scoring

The AAFs in control populations are useful for scoring the pathogenicity of variants, given that frequently occurring variants in the population are unlikely to cause rare diseases. We retrieved information on disease prevalence from OrphaNet and translated OrphaNet identifiers into OMIM identifiers. Here, we used three datasets to assess the variant frequency: the NHLBI Exome Sequencing Project (ESP6500), 1000 Genomes Project, and ExAC Browser. If any of the AAFs in any database is >5%, BA1 will be assigned as 1. If the AAF in the ExAC Browser is great than expected for the disorder caused by mutations in the corresponding gene, BS1 will be assigned as 1 (here, we set a default cutoff as 1% for rare disease, but users can specify their own cutoff in the configuration file of InterVar). If a variant is observed in a healthy adult in the 1000 Genomes Project as a homozygote (for diseases defined as recessive in OMIM) or as a heterozygote otherwise, then BS2 will be applied. We manually removed known major adult-onset disorders from consideration. We did not use the ExAC Browser or ESP6500 here because these datasets can contain variants from individuals with various diseases.

Variants that are absent or are present at extremely low frequencies in a large control cohort could represent moderate evidence for pathogenicity. If a variant that is responsible for dominant diseases is absent in all control subjects from ESP6500, 1000 Genomes Project, and the ExAC Browser, PM2 will be applied. If the variant causes recessive diseases and has a very low frequency with AAF < 0.5%, then PM2 can also be applied. Information on the gene-disease relationship, such as dominance or recessiveness, is obtained from OMIM.

In some cases, pathogenic variants have a significantly higher frequency in affected subjects than in control subjects. To handle these variants, we also cataloged all variants with an odds ratio (OR) > 5.0 from GWASdb[34] version 2. For these variants, PS4 will be applied. For some rare variants where case-control studies might not reach statistical significance, PS4 also can be downgraded to a moderate level during the manual adjustment step.

PM1 by Automated Scoring

Many protein domains play essential roles for protein function, so missense variants in these domains tend to be pathogenic. The domain information can be inferred from dbNSFP by ANNOVAR through the "dbnsfp31a_interpro" database. We first annotated all ClinVar variants (subject to the same data-cleaning procedure described above) with protein-domain information and then compiled a list in which domains contained only pathogenic or likely pathogenic variants without benign or common (allele frequency > 5%) variants. This list is provided within the InterVar package and will be updated regularly. If the user's input variants are located in these domains, then PM1 will be applied.

PM3 and BP2 by Manual Scoring

The pathogenicity of a variant also needs to be evaluated on the basis of whether variants with known pathogenicity exist in *cis* or *trans* with it. InterVar does not know the *cis*/*trans* status for variants, so this needs to be provided by users in the second step (manual adjustment) of InterVar. For two heterozygous variants that are present in a gene associated with recessive disorders, if one is pathogenic and the other is located in *tran*s, then moderate evidence of PM3 will be applied. If more than two variants are observed in *trans*, then moderate evidence for pathogenicity can be upgraded to strong. If the variants are present in a gene associated with dominant diseases, yet one variant is pathogenic and the other is located in *trans*, then supporting evidence of benign status will be applied to BP2 for the other variant. Regardless of models of disease inheritance, for two variants, if one is pathogenic and the other is observed in *cis*, then BP2 will be applied for the other variant.

PM4 and BP3 by Automated Scoring

Indels and stop losses can change the length of proteins and disrupt protein function. We annotated the repeat region by using

the "rmsk" database from the UCSC Genome Browser. This database was created by the RepeatMasker program, which screens DNA sequences for interspersed repeats and low-complexity DNA sequences. When the variants are "non-frameshift insertion," "non-frameshift deletion" in the non-repeat region, or stop-loss variants, PM4 will be applied. If the variants are "non-frameshift insertion" or "non-frameshift deletion" in the repeat region, BP3 will be applied.

PP1 and BS4 by Manual Scoring

Familial segregation of a variant with a disease is an important sign for linking the variant to the disease. If segregation is found in multiple affected family members and if this gene is definitively known to be associated with this disease, then PP1 will be applied. When there is a lack of segregation in affected members of a family, then the benign supporting evidence of BS4 will be applied. Because InterVar does not know the information on segregation, this piece of evidence can be provided by users in the second step (manual adjustment) of InterVar.

PP2 and BP1 by Automated Scoring

For many genes, the spectrum or distribution of pathogenic and benign variants can be informative for the pathogenicity status. For a given gene, if the missense variants are common causes of the disorder and the gene also has very few benign variants, then a missense variant in this gene can be supporting evidence for pathogenicity, and PP2 will be applied. However, if the truncating variants are major causes of the disease, then a missense variant in this gene can be supporting evidence for a benign status, and BP1 will be applied.

We annotated all variants in ClinVar (subject to the same data-cleaning procedure described above). For a given gene, if most of the pathogenic variants (>80% and at least one variant) are missense, and if a small proportion (<10% and less than one variant) of missense variants are benign, then for missense variants, PP2 will be assigned as 1. The treatment for BP1 is similar to that for PP2, but we assess whether most of pathogenic variants (>80% and at least one variant) are truncating variants. The truncating variants are defined as stop-gain, stop-loss, frameshift indel, or those disrupting splice sites. If the user's variants are missense in this gene, BP1 will be assigned as 1.

PP3 and BP4 by Automated Scoring

When multiple pieces of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.), then the supporting pathogenic evidence of PP3 will be assigned as 1. In comparison, when multiple pieces of computational evidence suggest no impact on the gene or gene product, then supporting benign evidence of BP4 will be assigned as 1. All sets of in silico results must agree when PP3 or BP4 is assigned.

These multiple pieces of computational evidence can be provided by ANNOVAR from the "dbnsfp30a" database, where the MetaSVM score[16] is used for deleteriousness prediction and GERP++ is used for evolutionary conservation. The splicing impacts can be inferred by ANNOVAR from the "dbscsnv11" database. For the evidence of PP3 and BP4, we set the cutoff to 0.0 for MetaSVM scores (greater scores indicate more likely deleterious effects), 2.0 for GERP++_RS (smaller scores indicate less conservation), and 0.6 for adaptive boosting (ADA) and random forest (RF) scores of dbscSNV as splicing impact (larger scores indicate more likely splice altering).

PP4 by Manual Scoring

For a given gene, if the individual's phenotype or family history is highly specific to the disorder associated with the gene, then it is supporting evidence for pathogenicity; in such a case, PP4 should be applied. This information needs to be provided by the user in the second step (manual adjustment) of InterVar.

PP5 and BP6 by Automated Scoring

If a reputable source has already reported a variant as pathogenic but the evidence is not provided for independent evaluation, then PP5 will be applied. When a reputable source has already reported a benign variant but without detailed evidence, then BP6 will be applied. In InterVar, we used the ClinVar dataset (subject to the same data-cleaning procedure described above) to perform this analysis by default, but users can select to use HGMD or other proprietary databases for this analysis.

BP5 by Manual Scoring

If a disease has an alternate molecular basis (caused by more than one gene) and if a variant is observed in a gene related to the disease, then it will be supporting evidence for a benign status, and BP5 will be assigned as 1. Note that this criterion is stronger for a gene associated with a dominant disorder than for a gene associated with a recessive disorder. Because of the multiple exceptions for this criterion, as described before,[25] users can adjust this criterion by using their own knowledge in the manual adjustment step.

BP7 by Automated Scoring

If a synonymous (silent) variant has no effect on splicing and if the nucleotide position is not highly conserved, then we can classify this variant as likely benign and assign BP7 as 1. The prediction on the effect on splicing can be extracted by ANNOVAR with the "dbscSNV" database. Both scores dbscSNV_RF_SCORE and dbscSNV_ADA_SCORE should be <0.6 when the variant is predicted to have no impact on splicing. The conservation information is retrieved from the "dbnsfp30a" database, where a GERP++ score > 2 indicates that the nucleotide is highly conserved.

InterVar and wInterVar

InterVar is a command-line-driven software written in Python and can be used as a standalone application on a variety of operating systems—including Windows, Linux, and MacOS—where Python is installed. The source code of InterVar is available from GitHub (see Web Resources).

InterVar takes either pre-annotated files in tab-delimited formats or unannotated input files in VCF format or ANNOVAR input format, where each line corresponds to one genetic variant. If the input files are unannotated, InterVar will call ANNOVAR to generate necessary annotations. Users can also use software tools other than ANNOVAR to generate pre-annotated files. The execution of InterVar mainly consists of two major steps: (1) automatically interpreting the variant by using the criteria outlined above and (2) manually adjusting specific criteria to re-interpret the clinical significance. However, users can also specify their own evidence file for a subset of the criteria and import it into InterVar by using the argument "-evidence_file" so that one single step is sufficient to generate the final results. In the output, on the basis of all 28 pieces of criteria that are either automatically generated or manually supplied by the user, each variant will be assigned as pathogenic, likely pathogenic, uncertain significance, likely benign, or benign by rules specified in the 2015 ACMG-AMP guidelines.[25]

We also developed a web server called wInterVar, which offers a graphical user interface for InterVar (see Web Resources). Users can directly input their missense variants into wInterVar by chromosomal position, by dbSNP identifier, or by gene name with the

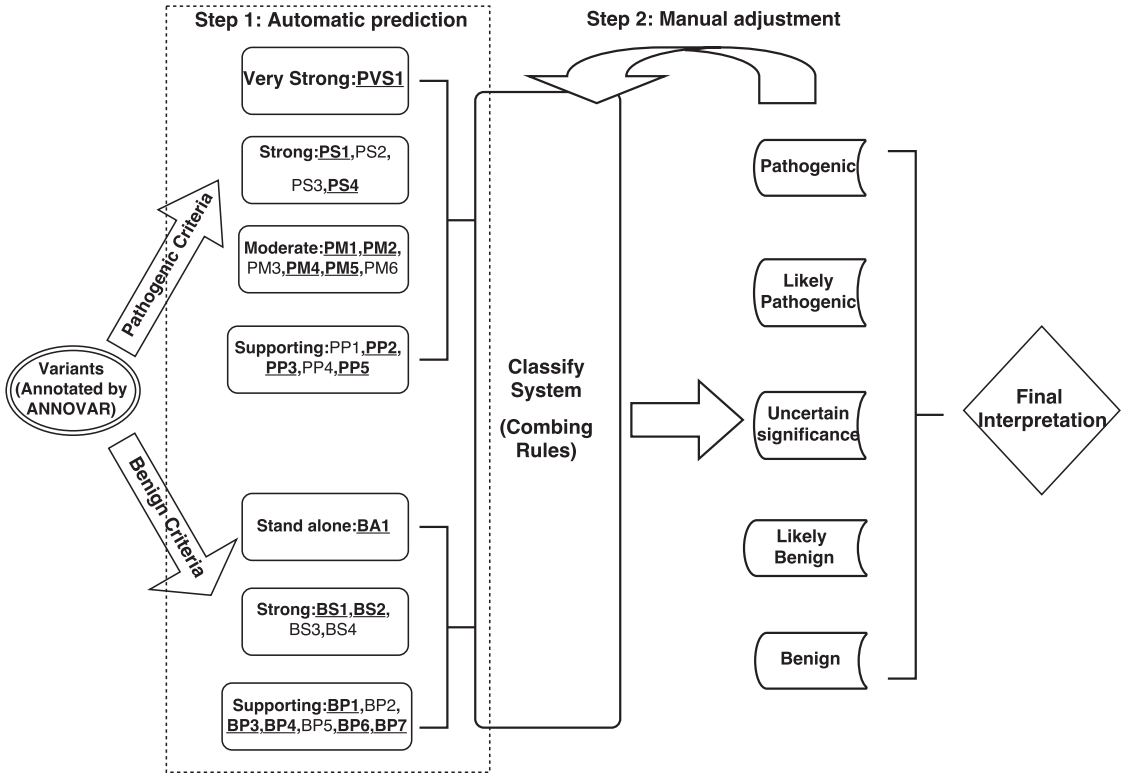


Figure 1. Flowchart of the Two-Step Procedure of InterVar
Underlined and bold fonts denote automated criteria.

nucleic acid change. The wInterVar server will provide full details on the variants, including all automatically generated criteria, most of the supportive evidence, and sub-population information. Users then have the ability to manually adjust these criteria and resubmit to the server to perform re-interpretation. We scanned all exons, and for each position we generated all three possible nucleotide changes. If the mutation was non-synonymous, we kept it in our database. The human genome contains approximately 80,000,000 non-synonymous variants, and we precomputed the 18 criteria for all of them. Therefore, the execution of wInterVar is very fast, typically less than 1 s to obtain the result on a variant. However, the wInterVar server cannot process other types of variants (such as indels), and the user will need to rely on InterVar instead.

Results

Summary of the Interpretation Procedure

A flowchart for InterVar is given in Figure 1. InterVar mainly consists of two major steps: (1) automated scoring on each of the 18 pieces of criteria and (2) manual review and adjustment on specific criteria to arrive at a final interpretation. During the first step, InterVar calls an annotation software, such as ANNOVAR,[5] to obtain necessary annotation information on variants and then uses its own internal annotation database to supplement additional annotations. Using these annotations on variants and genes, InterVar performs a preliminary interpretation of the variant and presents all relevant evidence for manual review. Currently, 18 pieces of criteria can be automatically generated and used in the first step. During the second step, the user can manually adjust each of the criteria on the basis of prior information (such as a variant's de novo status) or his or her own domain knowledge to reach a final interpretation. We emphasize here that automated scoring is based on default parameters and that users are advised to examine detailed evidence and use prior knowledge on ethnicity and/or disease to perform manual adjustments. A detailed explanation of these 28 criteria is given in Figure 2.

For example, consider missense variant chr12: 52,093,447T>C (GRCh37 coordinate) in exon 7 of *SCN8A* (MIM: 600702), which causes early infantile epileptic encephalopathy type 13 (MIM: 614558). We recently reported this variant as a de novo mutation in a 4-year-old female who, at 5 months of age, exhibited symptoms of epilepsy that progressed to a severe condition with very little movement, including the inability to sit or walk on her own.[35] We illustrate the scoring logic for this variant. This variant is located in a protein domain called the ion transport domain. This domain does not have any benign variants in public databases compiled by InterVar, so we assigned PM1 as 1. In addition, this variant is not present in the 1000 Genomes Project, ExAC Browser, or ESP6500, so PM2 was assigned as 1. For *SCN8A*, all known pathogenic variants are missense, so PP2 was

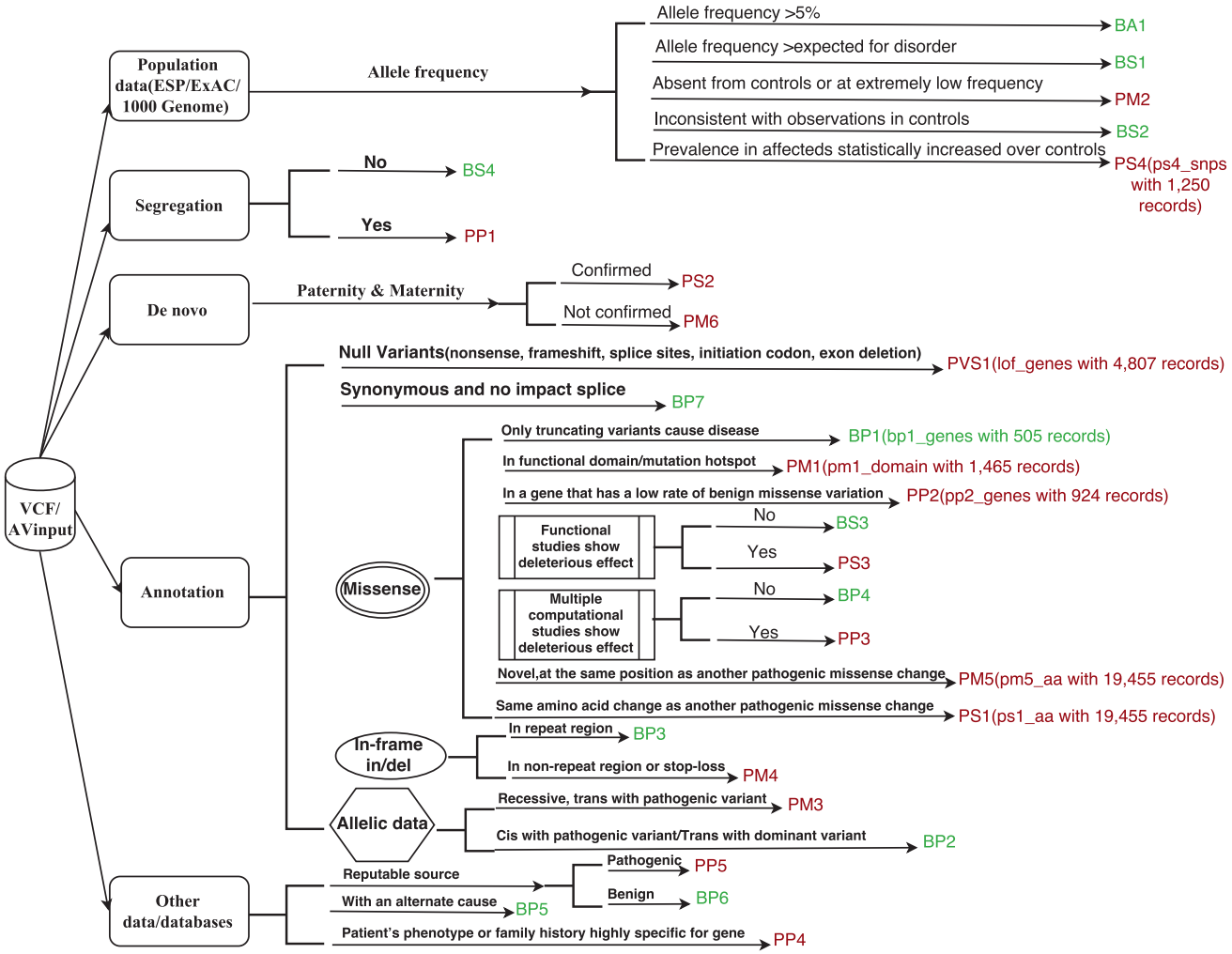


Figure 2. Illustration of the 28 Criteria from the 2015 ACMG-AMP Guidelines
For some criteria, the name of the internal database and its size are denoted within parentheses.

assigned as 1. According the 2015 ACMG-AMP rules, the variant falls into the class of "uncertain significance." In the second step, if we manually adjust the criteria by providing de novo information as PS2 = 1, then the clinical significance will change to "likely pathogenic" on the basis of "1 strong (PS1–PS4) and 1–2 moderate (PM1–PM6)." This procedure illustrates how to use automated interpretation and manual adjustment to derive a final interpretation for genetic variants.

Interpretation of De Novo Variants in Neurodevelopmental Disorders

We compiled a dataset of 9,305 de novo variants from 12 published trio-based exome sequencing studies on autism spectrum disorders,[36,37] developmental disorders,[38] schizophrenia,[39–42] epileptic encephalopathies,[43] and intellectual disability.[44–47] Among them, 8,346 variants were detected from affected subjects (n = 6,515), and 959 were detected from control subjects (n = 900). Among these 8,346 variants from affected subjects, 4,526 were non-synonymous, resulting in coding sequence changes in 3,462 genes, whereas 616 non-synonymous variants were present in 592 genes from control subjects.

We next performed automated variant interpretation by InterVar on all of these variants by using default options in the program and setting expected prevalence for these disorders as 1% (Table 1). Given that each published exome sequencing study used Sanger sequencing to validate the de novo status of the variants, we assigned PM6 as 1, indicating that these variants were assumed to be de novo without confirmed paternity or maternity. Among these variants, 4,459 (53.4%) and 493 (51.4%) were interpreted as having uncertain significance in affected and control subjects, respectively. Among affected subjects, 430 (5.1%) and 1,666 (20.0%) variants were interpreted as pathogenic and likely pathogenic, respectively. Among control subjects, 10 (1.0%) and 206 (21.5%) variants were interpreted as pathogenic and likely pathogenic, respectively.

We next combined variants with a benign or likely benign interpretation as one category (B/LB) and those with pathogenic or likely pathogenic as another category

Table 1. Illustration of Automated Interpretation of De Novo Variants from Individuals with Several Different Diseases and Control Subjects

Interpretation	DD	SCZ	ASD	EE	ID	Affected Subjects	Control Subjects
Benign	7	3	52	0	0	62	0
Likely benign	288	241	1,085	59	56	1,729	250
Uncertain significance	819	466	2,869	180	125	4,459	493
Likely pathogenic	339	199	967	81	80	1,666	206
Pathogenic	125	26	226	17	36	430	10
Total	1,578	935	5,199	337	297	8,346	959
Benign and likely benign	295	244	1,137	59	56	1,791	250
Pathogenic and likely pathogenic	464	225	1,193	98	116	2,096	216
p value (compared to control subjects)[a]	4.71E−7	0.65	0.06	0.00061	2.07E−6	0.0022	–
OR and 95% CI	0.55 (0.44–0.69)	0.94 (0.72–1.21)	0.82 (0.67–1.00)	0.52 (0.35–0.75)	0.42 (0.29–0.60)	0.74 (0.61–0.90)	–

Abbreviations are as follows: DDD, developmental disorder; SCZ, schizophrenia; ASD, autism spectrum disorder; EE, epileptic encephalopathy; ID, intellectual disability; OR, odds ratio; and CI, confidence interval.
[a]p values were calculated with a two-sided Fisher's exact test.

(P/LP) and compared their frequency between affected and control subjects. (Please note that we do not have access to individual-level data, so our analysis below focused on comparing detected variants between affected and control subjects.) Using Fisher's exact test, we detected a strong enrichment of P/LP variants among de novo variants in affected subjects ($p = 0.0022$) on the basis of automated interpretation. This result confirms that de novo variants that might be pathogenic are more prevalent in subjects with neurodevelopmental disorders than in control subjects. Please note that this analysis leveraged results only from automated interpretation (step 1) and did not account for manual adjustment (step 2) based on additional domain knowledge of the variants, genes, phenotypes, or diseases.

In comparison, we also predicted the pathogenicity of these variants by using SIFT and PolyPhen-2 scores on a subset of the variants for which the scores were available (Table 2). SIFT predicted 2,242 (26.8%) of 8,346 variants as deleterious (SIFT < 0.05 as the cutoff) for the subjects with neurodevelopmental disorders and predicted 283 (29.5%) of 959 variants as deleterious for control subjects. PolyPhen-2 predicted 3,157 (37.8%) of 8,346 variants as probably damaging or possibly damaging (PolyPhen-2_HDIV > 0.453 as the cutoff) for affected subjects and predicted 403 (42.0%) of 959 variants as probably damaging or possibly damaging for control subjects. Comparing affected and control subjects (Table 2), we did not observe a strong enrichment of P/LP variants with these two methods ($p = 0.64$ for SIFT and $p = 0.08$ for PolyPhen-2_HDIV). These results demonstrate that in silico predictions alone might not be sufficient to identify P/LP variants in exome sequencing studies.

Comparative Analysis on ClinVar

Although variant databases such as HGMD, ClinVar, and OMIM have been very helpful for cataloging genetic variants known to be associated with human diseases, they also have known limitations, e.g., that a portion of benign variants are incorrectly classified as pathogenic variants.[48,49] For example, Dorschner et al.[50] manually examined primary literature for 239 unique variants reported as pathogenic in HGMD and confirmed that only 7.5% are actually pathogenic from the original publication. The discrepancy in variant clinical significance between HGMD and clinical labs also highlights the lack of standards in interpreting a variant as pathogenic or likely pathogenic in the literature. Similarly, Bell et al.[51] found that 27% of the pathogenic variants cited in the literature are common polymorphisms or misannotated, underscoring the need for better mutation databases. Interestingly, we recently sequenced a personal genome and identified two variants reported as pathogenic in ClinVar, but manual examination of the cited publication indicated that neither was reported as pathogenic in the original publication.[52] This problem has been increasingly recognized in recent years,[48] suggesting that "known" pathogenic variants in various databases should not be taken at face value and instead deserve more detailed re-examination. Here, we analyzed the entire ClinVar dataset and compared their annotations with the automated interpretation (step 1) by InterVar to assess the concordance rates and examine sources of discordance. We recognized that because InterVar compiled some of its internal databases from ClinVar, its interpretation might be slightly biased toward being more similar to that of ClinVar.

We retrieved ClinVar version 2016-03-02 and selected all non-conflicting nonsynonymous variants categorized as

Table 2. Analysis of De Novo Variants by SIFT and PolyPhen-2

	SIFT		PolyPhen-2	
Interpretation	**Affected Subjects**	**Control Subjects**	**Affected Subjects**	**Control Subjects**
Benign or tolerated	2,608 (31.2%)	343 (35.7%)	1,426 (17.1%)	214 (22.3%)
Deleterious, probably damaging, or possibly damaging	2,242 (26.8%)	283 (29.5%)	3,157 (37.8%)	403 (42.0%)
Unknown	3,496 (42.0%)	333 (34.8%)	3763 (45.1%)	342 (35.7%)
Total	8,346	959	8,346	959
p value (compared to control subjects)[a]	0.64		0.08	

[a]p values were calculated with a two-sided Fisher's exact test.

one of the following: (1) benign or likely benign and (2) pathogenic or likely pathogenic. We then re-interpreted these variants by using the automated interpretation function in InterVar (Table 3). For the benign category in ClinVar, InterVar also classified 4,898 (80.6%) variants as benign or likely benign, suggesting that InterVar is largely consistent with ClinVar on this category of variants. However, for variants in the pathogenic category, InterVar and ClinVar have large differences. In fact, InterVar classified only 2,058 (13.9%) variants in the category as likely pathogenic yet none as pathogenic. Obviously, we acknowledge that all of these interpretations by InterVar were based on only 18 pieces of criteria in step 1, and none of them were subjected to manual examination; yet, additional information such as familial segregation, family history, and de novo status could move some variants with uncertain significance into a more deleterious category (likely pathogenic or pathogenic).

Given the differences between ClinVar annotation and InterVar prediction, we performed a more detailed analysis on the 513 (3.5%) variants that were classified as pathogenic by ClinVar but predicted as benign or likely benign by InterVar. First, we plotted the distribution of the maximum AAF of these variants in three databases (1000 Genomes Project, ExAC Browser, and NHLBI ESP6500; Figure 3). From this analysis, we found that there were >10% variants with AAF > 0.01 and 5% variants with AAF > 0.1. Clearly, >10% variants might be merely genetic polymorphisms that were incorrectly cataloged as pathogenic in ClinVar. Nevertheless, we also confirmed that in ClinVar, more than half of the pathogenic or likely pathogenic variants were very rare with an AAF < 0.0001, and >85% pathogenic variants had an AAF < 0.001, which fits our expectations. For manual examination of these variants, the cutoff of disease prevalence could be essential for assigning benign criteria such as BS1.

Analysis on Previously Reported Clinically Actionable Variants

Clinical exome and genome sequencing are likely to uncover "incidental findings" that are unrelated to the indication for ordering the sequencing tests but are of clinical significance.[53] The ACMG has recommended returning incidental findings from a minimum set of 56 actionable genes,[53] but many researchers have used an expanded list of genes selected according to domain knowledge. Several studies have examined incidental findings from large-scale genome or exome sequencing projects, so here we investigated how InterVar classifies clinically actionable genetic variants reported in previous studies.

Amendola et al.[54] previously examined exome sequencing data on 4,300 European Americans and 2,203 African Americans as part of NHLBI ESP6500 and reported 616 variants in 112 actionable genes (Table 4). These 616 variants were classified as actionable and pathogenic on the basis of HGMD annotations. Amendola et al. re-classified these 616 variants by using their own classification criteria, such as rules based on allele frequency, segregation, de novo status, function data, etc. They found only 70 (11.4%) as pathogenic or likely pathogenic, yet most of them (66.4%) were classified as variants of uncertain significance. Automated prediction (step 1) from InterVar classified only 33 (5.4%) variants as pathogenic or likely pathogenic, whereas most of the variants (43.2%) were classified as benign or likely benign. Please note that during variant classification, Amendola et al. leveraged information such as segregation and de novo status, but we did not have access to these pieces of information. Therefore, the number of pathogenic variants classified by InterVar in step 2 (manual adjustment) could increase significantly given additional information. Nevertheless, these results already suggest that the interpretation of InterVar is consistent with the manual interpretation by Amendola et al., who concluded that the vast majority of variants annotated as pathogenic in HGMD are probably not really pathogenic. This analysis confirms that incorrect classification of the pathogenic variant, even in ACMG actionable genes, represents a substantial issue when HGMD is the only criterion used for variant interpretation.

Comparative Analysis with CLINVITAE

CLINVITAE (see Web Resources) is a database of clinically observed genetic variants aggregated from public sources and is operated and made freely available by INVITAE. Although the vast majority of the variants were collected

Table 3. Illustration of Automated Interpretation of Pathogenic and Benign Variants Annotated in ClinVar

	ClinVar	
InterVar (Automated Interpretation)	**Pathogenic or Likely Pathogenic**	**Benign or Likely Benign**
Benign	65 (0.4%)	1,505 (24.8%)
Likely benign	448 (3.0%)	3,393 (55.9%)
Uncertain significance	12,207 (82.6%)	1,173 (19.3%)
Likely pathogenic	2,058 (13.9%)	0 (0%)
Pathogenic	0 (0%)	0 (0%)
Sum of five tiers	14,778	6,071
Benign and likely benign	513 (3.5%)	4,898 (80.6%)
Pathogenic and likely pathogenic	2,058 (13.9%)	0 (0%)

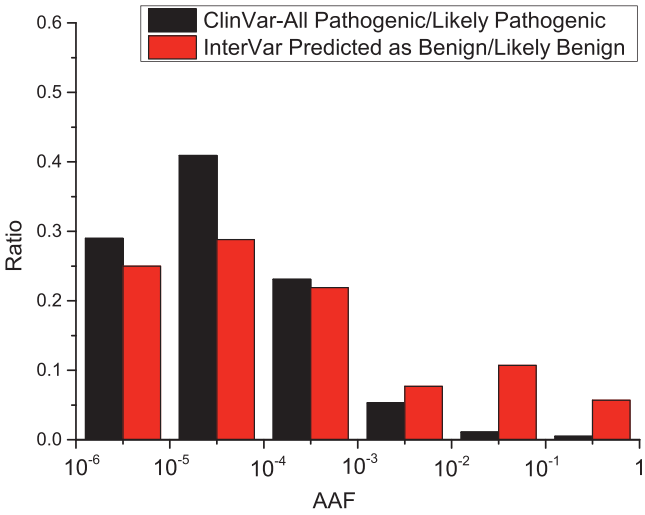


Figure 3. AAF Distribution of Pathogenic or Likely Pathogenic ClinVar Variants Predicted to Be Benign or Likely Benign by InterVar and All Pathogenic or Likely Pathogenic ClinVar Variants

from public databases, 11,696 variants were detected and classified by the INVITAE team. Unlike ClinVar and HGMD, which compile information from diverse sources, CLINVITAE potentially represents a more homogeneous collection of variants interpreted by a consistent set of institution-specific rules. Among these 11,696 variants, 5,405 (46.2%) and 717 (6.1%) were classified as benign or likely benign and pathogenic or likely pathogenic, respectively. Among them, 4,226 (36.1%) benign or likely benign variants were also classified as benign or likely benign by InterVar, whereas only 227 (1.9%) pathogenic or likely pathogenic variants were classified as pathogenic or likely pathogenic by InterVar (Table 5). This analysis again demonstrates that the concordance between automated interpretation of InterVar and expert-compiled classification is higher for benign or likely benign variants than for pathogenic or likely pathogenic variants.

wIntervar: Web Version of InterVar to Facilitate Manual Interpretation

wInterVar (see Web Resources) is a web implementation of InterVar so that users can use an online web server to perform interpretation on individual variants without using command-line tools. The wInterVar server has two steps for assessing and adjusting the clinical significance of variants: users first input a variant to obtain pre-computed, automated interpretation (Figure 4A). After reviewing the results of automated interpretation, users can then click the "adjust" button to perform the manual adjustment step by selecting and de-selecting appropriate criteria according to additional information and domain knowledge. The wInterVar server will then perform the final interpretation with the two-step procedure (Figure 4B).

We assessed the speed of InterVar and wInterVar. Using a machine with 16 GB of memory and two Intel Xeon X5650 (2.67 GHz) CPUs, the InterVar pipeline takes approximately 40 min to annotate 3,000,000 variants from a whole genome. The runtime can be greatly reduced to <5 min (~0.1 ms per variant) if an existing ANNOVAR annotation file is already available. For the wInterVar server, all annotation results for all possible non-synonymous variants were already pre-computed and imported into MongoDB, a NoSQL database system. Therefore, users can quickly search specific variants and receive an almost immediate response (<1 s for a variant). In addition, users can manually adjust the criteria and re-submit to wInterVar to obtain the final interpretation with an almost immediate response.

Discussion

In this article, we have presented two computational tools, InterVar and wInterVar, for performing evidence-based clinical interpretation of genetic variants according to the 2015 ACMG-AMP guidelines. To the best of our knowledge, we are not aware of software tools that are freely available to the academic community and perform similar functionalities. We wish to emphasize that although InterVar is a computational tool, it requires human input to derive accurate results with a two-step design: in the first step, InterVar performs automated interpretation with preliminary results, yet in the second step, InterVar takes additional information provided by human experts to adjust the criteria and provide a final interpretation. The two-step procedure allows InterVar to leverage automated information retrieval as much as possible, yet also allows additional input by human experts, to obtain the most accurate interpretations for genetic variants.

We applied InterVar to annotate and interpret de novo variants in subjects with neurodevelopmental disease and control subjects and observed a strong enrichment of pathogenic or likely pathogenic variants in affected subjects. In comparison, simple deleteriousness prediction algorithms such as SIFT and PolyPhen-2 failed to

Table 4. Interpretation of 616 HGMD-Classified Pathogenic Variants from NHLBI ESP6500

Clinical Significance	InterVar (Automated Interpretation)	ESP6500 Team (Manual Interpretation)	Concordant
Benign	5	0	0
Likely benign	261	137	77
Likely pathogenic	30	38	2
Pathogenic	3	32	0
Uncertain significance	317	409	234
Sum of five tiers	616	616	313
Benign or likely benign	266	137	79
Pathogenic or likely pathogenic	33	70	6

Table 5. Comparison of Variant Interpretation by CLINVITAE and Automated Interpretation by InterVar

Clinical Significance	InterVar (Automated Interpretation)	CLINVITAE	Concordant
Benign	242	2,407	230
Likely benign	6,593	2,998	2,428
Likely pathogenic	286	106	11
Pathogenic	137	611	132
Uncertain significance	4,438	5,574	3,047
Sum of five tiers	11,696	11,696	5,848
Benign or likely benign	6,835	5,405	4,226
Pathogenic or likely pathogenic	423	717	227

differentiate affected from control subjects. This observation suggests that one should compile multiple sources of criteria (in this case, up to 28 criteria), including deleteriousness prediction algorithms, to assess the potential pathogenicity of genetic variants rather than rely on deleteriousness prediction algorithms only.

Currently, a number of public databases, such ClinVar and HGMD, document the clinical significance of genetic variants, which are mostly provided by submitters or manually compiled from scientific literature. Because different submitters or different authors can have very different criteria to assess the pathogenicity of genetic variants, the quality of entries in these databases can be highly heterogeneous. As a result, it is expected that a proportion of pathogenic variants in these databases might simply be false positives that are misclassified.[48–51] Several studies have demonstrated that after manual re-interpretation, many of the pathogenic variants are indeed benign or have uncertain significance.[55–57] Our results in the current study further support the observation that a very large proportion of documented pathogenic or likely pathogenic variants are indeed polymorphisms segregating in the population and are unlikely to contribute significantly to disease risk. These observations further support the importance of efforts, such as ClinGen, to compile high-quality, gold-standard datasets with confidence scores to be used by the community for more accurate interpretation of genetic variants.

InterVar has several limitations that we wish to discuss here. First, InterVar needs a variant knowledgebase for accurate interpretation, so some variants in some genes might be more accurately interpreted than others. For example, well-studied genes tend to have more entries in clinical databases and are more likely to be interpreted accurately. Second, InterVar is designed to interpret genetic variants that are likely to cause Mendelian diseases or are highly penetrant for Mendelian diseases (OR $>$ 5) and cannot handle alleles that increase susceptibility to common and complex traits. Therefore, we caution that the current interpretation is appropriate only for Mendelian diseases or Mendelian forms of complex diseases. Third, although we provide a set of default databases to help implement 18 of the 2015 ACMG-AMP criteria, it is expected that different users or groups might want to use their own versions of these criteria. Therefore, we designed InterVar to be highly flexible in taking user-supplied annotations for each of the criteria to accommodate a variety of users with different needs.

Another issue we wish to emphasize is that the 2015 ACMG-AMP guidelines use 28 criteria with equal weights. One underlying rationale might be that it is difficult to quantify the contribution of each criterion given the complexity of interpreting genetic evidence.[25] Another potential reason is that equal weighting is intuitively easier to understand and implement by clinicians and researchers. However, it is expected that different types of criteria might have different contributions and weights for the classification of the pathogenicity or quantitative prediction of pathogenicity. If we can accumulate very large datasets of true positives and true negatives, it is possible to apply machine-learning approaches in the future for more accurate prediction and quantitative assessment of pathogenicity for genetic variants.

One important caveat that we wish to stress is that InterVar is better suited to addressing the variant-interpretation problem for severe congenital or very early-onset developmental disorders with nearly 100% penetrance, but it might work less well for late-onset or recessive diseases. For example, amyotrophic lateral sclerosis (ALS) is a fatal, progressive neurodegenerative disease, and the non-canonical IκB kinase family member TANK binding kinase 1 (*TBK1* [MIM: 604834]) was recently identified as an ALS-related gene in whole-exome sequencing of 2,874 ALS individuals and 6,405 control individuals.[58] InterVar classified all *TBK1* variants reported in the study as benign or having uncertain significance. Another example is *TREM2* (MIM:

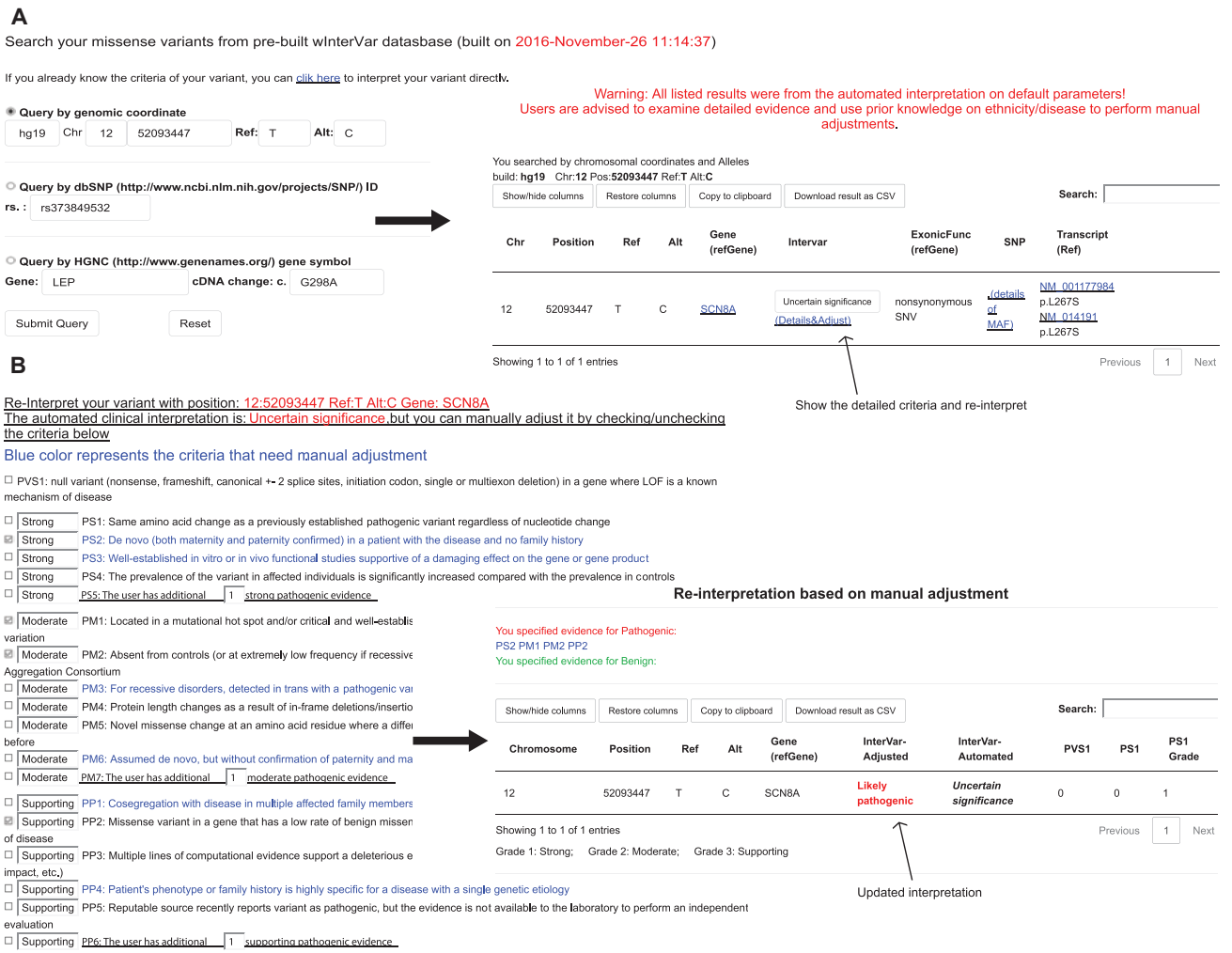


Figure 4. Illustration of wInterVar
(A) Automatic interpretation of genetic variants, which can be entered by several means.
(B) Once users click "adjust," the full list of criteria is shown for manual adjustment, after which the final results are given.

605086), associated with Alzheimer disease, from a recent sequencing study on a heterogeneous population of 1,092 affected and 1,107 control subjects.[59] Rare variants in *TREM2* (especially SNP rs75932628, which has the strongest association) were reported in their study. However, none of these variants were predicted to be pathogenic by InterVar. One main reason is that databases such as the ExAC Browser and ESP6500 were used in compiling the criteria, but they are technically not appropriate control databases because they are actually composed of many adult individuals with diseases. In comparison, the 1000 Genome Project is probably a more appropriate source of general control subjects, but its sample size is too small to enable adequate evaluation of rare variants. In any case, when databases such as the ExAC Browser and ESP6500 are used, it could be tricky to assign BS1 and BS2 to adult-onset or late-onset disorders, and some user-specific adjustments might be necessary for these diseases.

In summary, we have developed InterVar, a computational tool, and wInterVar, a web server, for the clinical interpretation of genetic variants according to the 2015 ACMG-AMP guidelines. InterVar can automatically generate the preliminary interpretations for 18 criteria and then allow manual adjustment of additional criteria to arrive at the final interpretation. InterVar can be easily used by researchers and clinicians and will greatly facilitate our understanding of the functional consequences of genetic variants in human diseases.

Acknowledgments

The authors thank Dr. Fan Xia (Baylor College of Medicine) and Dr. Rong Mao (ARUP Laboratories) for reading the manuscripts and offering valuable suggestions on the web server. We thank three anonymous reviewers for their valuable comments, which helped improve the manuscript substantially. We also want to thank members of the K.W. lab for testing the InterVar and wInterVar tools and providing feedback. This study was supported by NIH grants HG006465 and MH108728. K.W. was previously a board member and stock holder of Tute Genomics, a bioinformatics software company.

Received: April 27, 2016
Accepted: December 30, 2016
Published: January 26, 2017

Web Resources

1000 Genomes Project, http://www.1000genomes.org/
ANNOVAR, http://annovar.openbioinformatics.org/
ClinVar, https://www.ncbi.nlm.nih.gov/clinvar/
CLINVITAE, http://clinvitae.invitae.com/
dbNSFP, https://sites.google.com/site/jpopgen/dbNSFP
dbscSNV, https://sites.google.com/site/jpopgen/dbNSFP
dbSNP, http://www.ncbi.nlm.nih.gov/SNP
Ensembl, http://www.ensembl.org/
Exome Aggregation Consortium (ExAC) Browser, http://exac.broadinstitute.org
GERP++, http://mendel.stanford.edu/SidowLab/downloads/gerp/
GWASdb, http://jjwanglab.org/gwasdb
HGMD, http://www.hgmd.org
InterVar, https://github.com/WGLab/InterVar
MedGen, https://www.ncbi.nlm.nih.gov/medgen/
NHLBI Exome Sequencing Project (ESP) Exome Variant Server, http://evs.gs.washington.edu/EVS/
OMIM, http://omim.org/
OrphaNet, http://www.orpha.net/
PolyPhen-2, http://genetics.bwh.harvard.edu/pph2
RefSeq, http://www.ncbi.nlm.nih.gov/refseq
RepeatMasker, http://www.repeatmasker.org/
SIFT, http://sift.jcvi.org/
UCSC Genome Browser, http://genome.ucsc.edu
wIntervar, http://wintervar.wglab.org/

References

1. McPherson, J.D. (2009). Next-generation gap. Nat. Methods *6* (11, Suppl), S2–S5.
2. Lyon, G.J., and Wang, K. (2012). Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress. Genome Med. *4*, 58.
3. Quintáns, B., Ordóñez-Ugalde, A., Cacheiro, P., Carracedo, A., and Sobrido, M.J. (2014). Medical genomics: The intricate path from genetic variant identification to clinical interpretation. Appl. Transl. Genomics *3*, 60–67.
4. Chang, X., and Wang, K. (2012). wANNOVAR: annotating genetic variants for personal genomes via the web. J. Med. Genet. *49*, 433–436.
5. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.
6. Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L.B., and Reese, M.G. (2011). A probabilistic disease-gene finder for personal genomes. Genome Res. *21*, 1529–1542.
7. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature *461*, 272–276.
8. Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92.
9. McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics *26*, 2069–2070.
10. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081.
11. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
12. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.
13. Shihab, H.A., Gough, J., Cooper, D.N., Day, I.N., and Gaunt, T.R. (2013). Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics *29*, 1504–1510.
14. Schwarz, J.M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods *7*, 575–576.
15. González-Pérez, A., and López-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am. J. Hum. Genet. *88*, 440–449.
16. Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., and Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum. Mol. Genet. *24*, 2125–2137.
17. Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. Hum. Mutat. *32*, 358–368.
18. Thompson, B.A., Greenblatt, M.S., Vallee, M.P., Herkert, J.C., Tessereau, C., Young, E.L., Adzhubey, I.A., Li, B., Bell, R., Feng, B., et al. (2013). Calibration of multiple in silico tools for predicting pathogenicity of mismatch repair gene missense substitutions. Hum. Mutat. *34*, 255–265.
19. Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum. Genet. *133*, 1–9.
20. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44* (D1), D862–D868.
21. Horaitis, O., Talbot, C.C., Jr., Phommarinh, M., Phillips, K.M., and Cotton, R.G. (2007). A database of locus-specific databases. Nat. Genet. *39*, 425.
22. Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al.; ClinGen (2015). ClinGen–the Clinical Genome Resource. N. Engl. J. Med. *372*, 2235–2242.
23. Kazazian, H.H., Boehm, C.D., and Seltzer, W.K. (2000). ACMG recommendations for standards for interpretation of sequence variations. Genet. Med. *2*, 302–303.

24. Richards, C.S., Bale, S., Bellissimo, D.B., Das, S., Grody, W.W., Hegde, M.R., Lyon, E., Ward, B.E.; and Molecular Subcommittee of the ACMG Laboratory Quality Assurance Committee (2008). ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genet. Med. *10*, 294–300.
25. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al.; ACMG Laboratory Quality Assurance Committee (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–424.
26. Amendola, L.M., Jarvik, G.P., Leo, M.C., McLaughlin, H.M., Akkari, Y., Amaral, M.D., Berg, J.S., Biswas, S., Bowling, K.M., Conlin, L.K., et al. (2016). Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. Am. J. Hum. Genet. *98*, 1067–1076.
27. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
28. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
29. Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum. Mutat. *37*, 235–241.
30. Liu, X., Jian, X., and Boerwinkle, E. (2011). dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum. Mutat. *32*, 894–899.
31. Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., et al. (2012). InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. *40*, D306–D312.
32. Jian, X., Boerwinkle, E., and Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. *42*, 13534–13544.
33. Lewis, B.P., Green, R.E., and Brenner, S.E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc. Natl. Acad. Sci. USA *100*, 189–192.
34. Li, M.J., Liu, Z., Wang, P., Wong, M.P., Nelson, M.R., Kocher, J.P., Yeager, M., Sham, P.C., Chanock, S.J., Xia, Z., and Wang, J. (2016). GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res. *44* (D1), D869–D876.
35. Malcolmson, J., Kleyner, R., Tegay, D., Adams, W., Ward, K., Coppinger, J., Nelson, L., Meisler, M.H., Wang, K., Robison, R., and Lyon, G.J. (2016). SCN8A mutation in a child presenting with seizures and developmental delays. Cold Spring Harb Mol Case Stud *2*, a001073.
36. De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al.; DDD Study; Homozygosity Mapping Collaborative for Autism; and UK10K Consortium (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209–215.
37. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.
38. Deciphering Developmental Disorders, S.; and Deciphering Developmental Disorders Study (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature *519*, 223–228.
39. Girard, S.L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., Dionne-Laporte, A., Spiegelman, D., Henrion, E., Diallo, O., et al. (2011). Increased exonic de novo mutation rate in individuals with schizophrenia. Nat. Genet. *43*, 860–863.
40. Xu, B., Ionita-Laza, I., Roos, J.L., Boone, B., Woodrick, S., Sun, Y., Levy, S., Gogos, J.A., and Karayiorgou, M. (2012). De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. Nat. Genet. *44*, 1365–1369.
41. Gulsuner, S., Walsh, T., Watts, A.C., Lee, M.K., Thornton, A.M., Casadei, S., Rippey, C., Shahin, H., Nimgaonkar, V.L., Go, R.C., et al.; Consortium on the Genetics of Schizophrenia (COGS); and PAARTNERS Study Group (2013). Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. Cell *154*, 518–529.
42. Fromer, M., Pocklington, A.J., Kavanagh, D.H., Williams, H.J., Dwyer, S., Gormley, P., Georgieva, L., Rees, E., Palta, P., Ruderfer, D.M., et al. (2014). De novo mutations in schizophrenia implicate synaptic networks. Nature *506*, 179–184.
43. Allen, A.S., Berkovic, S.F., Cossette, P., Delanty, N., Dlugos, D., Eichler, E.E., Epstein, M.P., Glauser, T., Goldstein, D.B., Han, Y., et al.; Epi4K Consortium; and Epilepsy Phenome/Genome Project (2013). De novo mutations in epileptic encephalopathies. Nature *501*, 217–221.
44. Hamdan, F.F., Srour, M., Capo-Chichi, J.M., Daoud, H., Nassif, C., Patry, L., Massicotte, C., Ambalavanan, A., Spiegelman, D., Diallo, O., et al. (2014). De novo mutations in moderate or severe intellectual disability. PLoS Genet. *10*, e1004772.
45. Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet *380*, 1674–1682.
46. de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N. Engl. J. Med. *367*, 1921–1929.
47. Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature *511*, 344–347.
48. (2016). Improving databases for human variation. Nat. Methods *13*, 103.
49. MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. Nature *508*, 469–476.
50. Dorschner, M.O., Amendola, L.M., Turner, E.H., Robertson, P.D., Shirts, B.H., Gallego, C.J., Bennett, R.L., Jones, K.L., Tokita, M.J., Bennett, J.T., et al.; National Heart, Lung, and Blood

Institute Grand Opportunity Exome Sequencing Project (2013). Actionable, pathogenic incidental findings in 1,000 participants' exomes. Am. J. Hum. Genet. *93*, 631–640.
51. Bell, C.J., Dinwiddie, D.L., Miller, N.A., Hateley, S.L., Ganusova, E.E., Mudge, J., Langley, R.J., Zhang, L., Lee, C.C., Schilkey, F.D., et al. (2011). Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci. Transl. Med. *3*, 65ra4.
52. Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., Fu, A., Li, Q., Li, N., Gong, S., et al. (2016). Long-read sequencing and de novo assembly of a Chinese genome. Nat. Commun. *7*, 12065.
53. Green, R.C., Berg, J.S., Grody, W.W., Kalia, S.S., Korf, B.R., Martin, C.L., McGuire, A.L., Nussbaum, R.L., O'Daniel, J.M., Ormond, K.E., et al.; American College of Medical Genetics and Genomics (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet. Med. *15*, 565–574.
54. Amendola, L.M., Dorschner, M.O., Robertson, P.D., Salama, J.S., Hart, R., Shirts, B.H., Murray, M.L., Tokita, M.J., Gallego, C.J., Kim, D.S., et al. (2015). Actionable exomic incidental findings in 6503 participants: challenges of variant classification. Genome Res. *25*, 305–315.
55. Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E.V., Mort, M., Phillips, A.D., Shaw, K., Stenson, P.D., Cooper, D.N., Tyler-Smith, C.; and 1000 Genomes Project Consortium (2012). Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. Am. J. Hum. Genet. *91*, 1022–1032.
56. Shearer, A.E., Eppsteiner, R.W., Booth, K.T., Ephraim, S.S., Gurrola, J., 2nd, Simpson, A., Black-Ziegelbein, E.A., Joshi, S., Ravi, H., Giuffre, A.C., et al. (2014). Utilizing ethnic-specific differences in minor allele frequency to recategorize reported pathogenic deafness variants. Am. J. Hum. Genet. *95*, 445–453.
57. Tabor, H.K., Auer, P.L., Jamal, S.M., Chong, J.X., Yu, J.H., Gordon, A.S., Graubert, T.A., O'Donnell, C.J., Rich, S.S., Nickerson, D.A., Bamshad, M.J.; and NHLBI Exome Sequencing Project (2014). Pathogenic variants for Mendelian and complex traits in exomes of 6,517 European and African Americans: implications for the return of incidental results. Am. J. Hum. Genet. *95*, 183–193.
58. Cirulli, E.T., Lasseigne, B.N., Petrovski, S., Sapp, P.C., Dion, P.A., Leblond, C.S., Couthouis, J., Lu, Y.F., Wang, Q., Krueger, B.J., et al.; FALS Sequencing Consortium (2015). Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. Science *347*, 1436–1441.
59. Guerreiro, R., Wojtas, A., Bras, J., Carrasquillo, M., Rogaeva, E., Majounie, E., Cruchaga, C., Sassi, C., Kauwe, J.S.K., Younkin, S., et al.; Alzheimer Genetic Analysis Group (2013). TREM2 variants in Alzheimer's disease. N. Engl. J. Med. *368*, 117–127.

MARRVEL: Integration of Human and Model Organism Genetic Resources to Facilitate Functional Annotation of the Human Genome

Julia Wang,[1,2,11] Rami Al-Ouran,[3,4,11] Yanhui Hu,[5,11] Seon-Young Kim,[3,6,11] Ying-Wooi Wan,[3,4,7] Michael F. Wangler,[1,3,4,6] Shinya Yamamoto,[1,3,6] Hsiao-Tuan Chao,[3,4,8] Aram Comjean,[5] Stephanie E. Mohr,[5] UDN, Norbert Perrimon,[5,9] Zhandong Liu,[3,4,*] and Hugo J. Bellen[1,3,6,10,*]

One major challenge encountered with interpreting human genetic variants is the limited understanding of the functional impact of genetic alterations on biological processes. Furthermore, there remains an unmet demand for an efficient survey of the wealth of information on human homologs in model organisms across numerous databases. To efficiently assess the large volume of publically available information, it is important to provide a concise summary of the most relevant information in a rapid user-friendly format. To this end, we created MARRVEL (model organism aggregated resources for rare variant exploration). MARRVEL is a publicly available website that integrates information from six human genetic databases and seven model organism databases. For any given variant or gene, MARRVEL displays information from OMIM, ExAC, ClinVar, Geno2MP, DGV, and DECIPHER. Importantly, it curates model organism-specific databases to concurrently display a concise summary regarding the human gene homologs in budding and fission yeast, worm, fly, fish, mouse, and rat on a single webpage. Experiment-based information on tissue expression, protein subcellular localization, biological process, and molecular function for the human gene and homologs in the seven model organisms are arranged into a concise output. Hence, rather than visiting multiple separate databases for variant and gene analysis, users can obtain important information by searching once through MARRVEL. Altogether, MARRVEL dramatically improves efficiency and accessibility to data collection and facilitates analysis of human genes and variants by cross-disciplinary integration of 18 million records available in public databases to facilitate clinical diagnosis and basic research.

Introduction

One major challenge encountered with interpreting human genetic variants is the limited understanding of the functional impact of genetic alterations on biological processes. Traditional variant interpretation methodology relies on restricting clinical interpretation to known Mendelian diseases and employing in silico prediction algorithms. For most genes, few variants have reliable and validated clinical significance designation, resulting in difficulties in differentiating between benign and pathogenic variants or determining whether variants in a candidate gene are causative.[1] The wealth of available biological information across multiple model organisms could aid in the interpretation of variants such as known molecular functions of the candidate gene. However, there are major barriers to search for biological data in specific model organism databases due to the intricacies of evaluating orthologs and navigating seven different websites' different organization, different approaches, and different use of gene or protein identifiers (Figure S1). This limits the efficiency of incorporating known model organism data into analysis of candidate genes.

Therefore, there is an unmet demand for resources to facilitate rapid curation of available human gene and variant information, to determine conservation, and to gather relevant information on homologous genes in model organisms. Furthermore, such data compilation is relevant to evaluating the consequences of human genetic variation in model organisms.[2] To provide a concise and user-friendly curation of pertinent and publicly available knowledge, we created MARRVEL (model organism aggregated resources for rare variant exploration). MARRVEL is an open-access resource that synthesizes genetic and model organism information from several public databases into a single user-friendly website (Figure 1).

The major impetus for developing MARRVEL arose from growing efforts to analyze the potential pathogenicity of genetic alterations in genes that are either not previously associated with human genetic disease or associated with different clinical features. A wide range of efforts for the discovery of disease-causing variants include the research consortiums for rare (e.g., Center for Mendelian Genomics[3] and Undiagnosed Diseases Network[4]) or common (CHARGE consortium[3]) diseases, clinical genetics laboratories, large-scale sequencing projects,[5,6] and collaborations between

[1]Program in Developmental Biology, Baylor College of Medicine (BCM), Houston, TX 77030, USA; [2]Medical Scientist Training Program, BCM, Houston, TX 77030, USA; [3]Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX 77030, USA; [4]Department of Pediatrics, BCM, Houston, TX 77030, USA; [5]Drosophila RNAi Screening Center, Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA; [6]Department of Molecular and Human Genetics, BCM, Houston, TX 77030, USA; [7]Department of Obstetrics and Gynecology, BCM, Houston, TX 77030, USA; [8]Department of Pediatrics, Section of Child Neurology, BCM, Houston, TX 77030, USA; [9]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA; [10]Howard Hughes Medical Institute, BCM, Houston, TX 77030, USA
[11]These authors contributed equally to this work
*Correspondence: zhandonl@bcm.edu (Z.L.), hbellen@bcm.edu (H.J.B.)
http://dx.doi.org/10.1016/j.ajhg.2017.04.010.

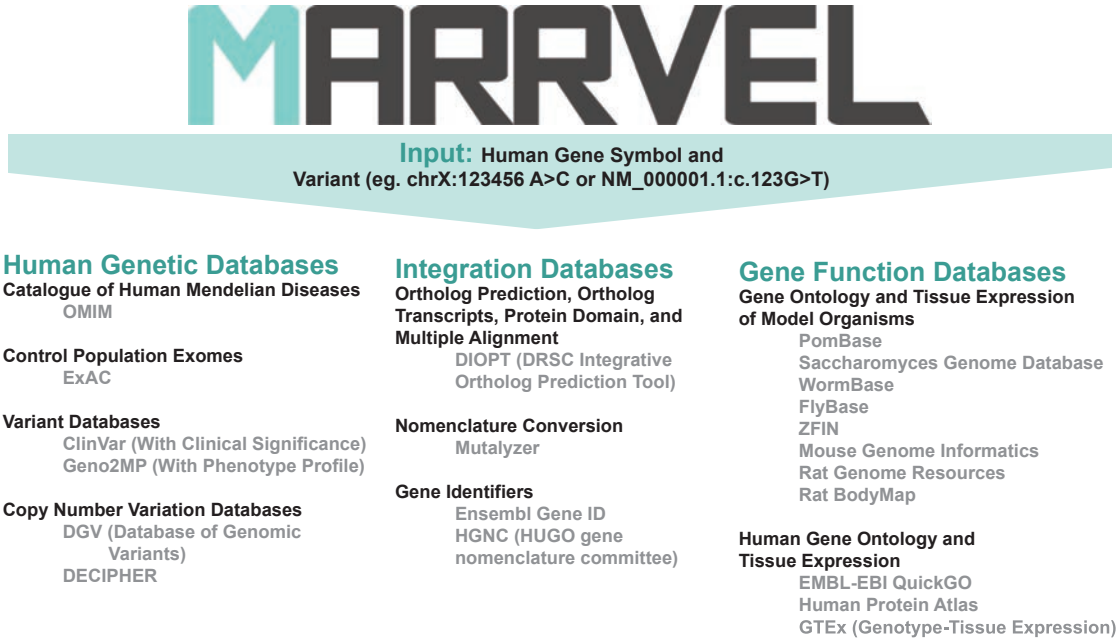


Figure 1. Overall Structure of MARRVEL
MARRVEL integrates 21 different databases to facilitate human gene and variant analysis for further study in model organisms. Human genetic databases are selected to provide data on disease association, statistics on variants found in a gene of interest, and exact matches with a variant of interest. Integration databases are important to the overall structure of MARRVEL due to the complicated structures and connections between each database that require homology prediction, specific gene identifiers, and nomenclature. Gene function databases are selected to provide a concise summary of what is known about a gene of interest across organisms.

human geneticists and model organism researchers.[7] Together, these research efforts generate growing numbers of large human genomic datasets that require the development of resources and tools to facilitate efficient data analysis.

For example, the Undiagnosed Diseases Network[4] combines the expertise of clinicians, sequencing centers (e.g., whole-exome, whole-genome, RNA-seq), metabolomics laboratories, and model organism scientists (fruit fly, zebrafish, and mouse) to diagnose individuals with rare disorders that eluded traditional diagnostic modalities. Many of these cases are predicted to have a primary genetic cause but the suspected causative variant may not be in disease-associated genes. When candidate pathogenic gene variants are identified, model organism data available for predicted orthologs of the human gene are an invaluable resource for interpreting the biological significance of the genetic alterations. However, this model organism-based resource is underutilized due to limited accessibility by non-model organism researchers. Currently, researchers need to visit and navigate separate model organism-specific databases (e.g., FlyBase,[8] MGI,[9] ZFIN[10]) that utilize distinct genotype and phenotype nomenclature as well as data organization. Moreover, in the study of genes or variants linked to human diseases, model organisms provide powerful platforms for mechanistic studies. Hence, a user-friendly open-access web-based resource to curate and synthesize current knowledge and resources from model organisms and human genomics databases is invaluable.[11–13]

Material and Methods

Human Genetics Databases

Human genetics data are extracted from Online Mendelian Inheritance in Man (OMIM),[14] Exome Aggregation Consortium (ExAC),[15] Genotype to Mendelian Phenotype (Geno2MP), ClinVar,[16] Database of Genomic Variants (DGV),[17] and DECIPHER (database of genomic variation and phenotype in humans using Ensembl resources).[18]

We display the human gene description, gene-phenotype relationships, and reported alleles from OMIM. Next, control population gene summary from the ExAC[15] database is displayed. ExAC is a public collection of more than 60,000 exomes that have been selected against individuals with severe early-onset Mendelian phenotypes.[15] When MARRVEL is primarily applied to early-onset pediatric phenotypes and used to evaluate candidate genes for Mendelian disease, the ExAC data can be considered as a "control" dataset. We will refer to this data as "control" throughout the paper though it should be noted these samples should not be considered similarly for adult neurodegenerative phenotypes, for example. Within the control population gene summary, we include the pLI (the probability of being loss-of-function [LoF] intolerant) score of a gene, which assesses the probability that a gene is extremely intolerant to loss of function variants (nonsense, splice acceptor, and splice donor variants) caused by single-nucleotide changes.[15]

We next display data from the Geno2MP database. Geno2MP is a database sponsored by the University of Washington Center for Mendelian Genetics displaying variants from Mendelian gene discovery projects and provide phenotype information for individuals with specific genotypes, including affected and unaffected family members.

Next, we extract data from ClinVar[16] containing more than 255,000 unique variants annotated with clinical significance and review status (i.e., level of evidence). When a user searches for a gene and variant, MARRVEL displays all ClinVar variants reported in the gene of interest, summarizes the number of variants in each category of clinical significance, and highlights any variant(s) that match the location of the variant of interest. We provide both a high-level summary of the variants in terms of its reported clinical significance as well as a table with details for each reported variant. In addition, any alleles that overlap with the location of the variant of interest is highlighted in blue.

We then display data from the Database of Genomic Variants (DGV)[17] database, which contains a large collection of structural variants from more than 54,000 individuals. The database includes samples of reportedly healthy individuals, at the time of ascertainment, from up to 72 different studies. Using the DGV database, we report all copy-number variants (CNVs) that overlap the input gene. If a CNV containing the gene of interest exists, we display the frequency, type of CNV, and publications associated with the CNV.

Finally, we display additional CNV information from the DECIPHER[18] database based on the variant coordinate that includes common variants from the control population. Due to data display restrictions, we are able to provide the users only with control population data from DECIPHER.

Gene Function Databases

Biological and genetic features of human genes and their putative orthologous genes, including tissue expression pattern and Gene Ontology (GO) terms, are extracted from the following model organism databases: *Saccharomyces* Genome Database (SGD)[19] for the budding yeast *Saccharomyces cerevisiae*, PomBase[20] for the fission yeast *Schizosaccharomyces pombe*, WormBase[21] for the nematode worm *Caenorhabditis elegans*, FlyBase[8] for the fruit fly *Drosophila melanogaster*, ZFIN[10] for the zebrafish *Danio rerio*, Mouse Genome Informatics (MGI)[9] for mouse *Mus musculus*, and Rat Genome Database[22] and Bodymap[23] for rat *Rattus norvegicus*. For humans, we extract GO terms from QuickGO[24] and tissue expression data from GTEx[25] and Protein Atlas.[26] To identify the putative orthologs of the human gene, we incorporate information from DIOPT (*Drosophila* RNAi Screening Center [DRSC] Integrative Ortholog Prediction Tool),[27] an online tool integrating 14 ortholog prediction tools to provide a homology score for each predicted ortholog pair. Additionally, DIOPT is used to display a multiple protein alignment that is generated with MAFFT and human gene functional domains.[27]

Data Processing

MARRVEL search allows three types of inputs: a single HUGO gene symbol,[28] a single human variant, or a combination of both. The human variant input can be in the format conforming to HGVS nomenclature[29] or in the genomic variant format [chromosome number]:[genomic coordinate] [Reference nucleotide]>[Alternate nucleotide]), for example 6:99365567T>C. If the variant is input in HGVS nomenclature format, then the Mutalyzer Position Converter tool[30] is used to transform the variant input into genomic coordinate, as variants stored in our database follow the genomic variant format.

If the input to MARRVEL includes both variant and gene symbol, data from OMIM[14] are retrieved using the OMIM API and gene summary table is extracted from the ExAC website in real time. Variant data from the ExAC[15] and Geno2MP databases are retrieved from our MySQL[31] database as explained in the following section. Regarding ClinVar[16] alleles, MARRVEL searches by the gene symbol and reports all alleles that overlap with the input gene. MARRVEL also provides a summary on clinical significance from these alleles. MARRVEL displays DGV[17] copy-number variants based on the genes that are encompassed by the copy-number variants. Variant data from DECIPHER[18] are retrieved from our MySQL database based on the chromosomal location.

If the input includes only a gene symbol, MARRVEL retrieves the gene summary table from the ExAC website. For Geno2MP, it shows all variants overlapping the gene in the database and its heterozygote count, homozygote count, and their sum. For DGV, it shows all CNV regions overlapping the gene. DECIPHER data are not retrieved since it does not provide report data associated with genes.

When the input includes only a variant, MARRVEL first searches the ExAC database to retrieve the variant information. It then shows gene-related information such as OMIM, orthologs and their functions, and protein alignment of the first gene the ExAC database matches.

For any combination of gene and variant input, the gene function table includes the following columns: the orthologous genes column, the DIOPT[27] score column, the tissue expression column, and the associated GO terms' columns. The orthologous genes column displays the putative orthologs predicted using DIOPT with a link to each organism database as well as a PubMed link. The PubMed link is generated from the NCBI[32] gene page's sub-link "Related articles in PubMed - See all citations in PubMed" under the "Bibliography" section of the NCBI gene page. By default, the gene function table shows only the putative orthologs with the best DIOPT score. All predicted orthologs can also be displayed by deselecting the check-box for this option at the top of the gene function table. The tissue expression column displays the tissue expression data for human and six model organisms. Expression data shown in the table list the names of tissues that highly express the gene of interest. For humans, there is an option to show all tissues with high protein expression levels from Protein Atlas[26] and a bar graph of mRNA expression data from GTEx,[25] including tissue names and its median value of RPKM. The mouse and zebrafish expression show only tissues expressed in wild-type. For fly, the tissues with high expression levels are displayed.

MARRVEL also provides human gene protein domain information and protein alignments for the gene and its homolog genes, which is extracted from DIOPT.

Server and Data Storing

MARRVEL is hosted on Amazon Web Service (AWS) EC2. We extracted data from the databases either by the database's API or by downloading and storing files publicly available into a MySQL database. Multiple protein alignment and domain information from DIOPT are stored using AWS S3. Human variant data from the ExAC and Geno2MP databases were extracted by downloading and processing their respective VCF files and storing them in our database while ExAC gene summary data is pulled on demand from the ExAC website. Human copy-number variation data were extracted from the DGV[17] and DECIPHER[18] databases by downloading the databases' tab delimited files and storing them in our database as well. MARRVEL's usage of DECIPHER data adheres to the DECIPHER Data Access Agreement. ClinVar[16] data were pulled from the ClinVar website and stored in the

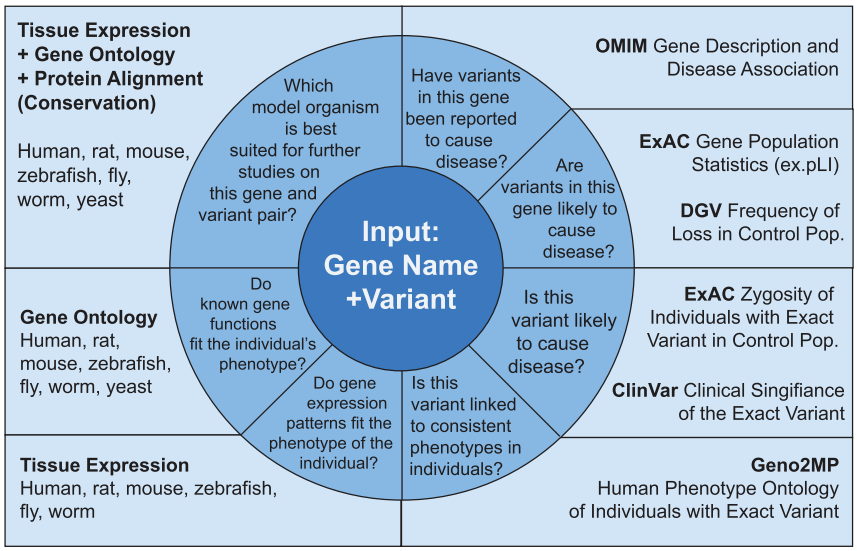


Figure 2. Example of an Approach for Variant Analysis using MARRVEL
An example of how MARRVEL output can be used to analyze human genes and variants is illustrated by a question asked by the user in the inner ring and the answer that can be found in MARRVEL's output in the outer boxes. We start at the noon position and advance clockwise.

MySQL database. MARRVEL retrieves updated data from ClinVar bi-weekly.

Additional gene function data were obtained by accessing and extracting data from the DIOPT website.[27] MARRVEL uses DIOPT's ortholog prediction, protein alignment, and domain information. Human tissue expression data were obtained from Protein Atlas[26] website API and median values from GTEx[25] downloadable files. Human GO terms are directly downloaded from the QuickGO[24] web pages. Rat expression data and GO terms are from RGD downloadable files.[22] Mouse expression data and GO terms are from MGI website.[9] Zebrafish data are from ZFIN downloadable files.[10] Fly expression data and GO terms are pulled from the FlyBase website.[8] *S. cerevisiae* data and *S. pombe* data are from SGD[19] and PomBase.[20]

MARRVEL's interface is implemented using the Twitter bootstrap framework v.4.0.0, jQuery v.2.2.0, and Angular JS v.1.6.1. The server backend was implemented using the Node.js framework v.6.7.0.

For the exact database versions, please see Tables S1 and S2.

Results

MARRVEL Integrates Data from Human and Model Organism Databases

MARRVEL builds upon and complements existing tools by integrating population genetics, model organism functional data, multiple protein alignments, and other information into one web- and mobile device-friendly site (Figure 1). The simple interface at MARRVEL allows entry of a human gene or variant to begin the survey with the results falling into two main categories. First, MARRVEL aggregates information from widely used human genetic databases (ExAC, Geno2MP, ClinVar, DGV, DECIPHER-control population), including sources of control and disease population data, to facilitate gene variant analysis. Second, MARRVEL displays a concise summary of available information for putative orthologs across yeast, worm, fly, fish, mouse, and rat (see Material and Methods). For genes that are not previously associated with human disease, there is often limited in vivo human functional data. However, there is often a wealth of model organisms data that can be used to infer the human gene function. By integrating biological and biochemical data across multiple model organisms, we provide links between human disease and gene function through a comprehensive overview of publicly available data. In total, MARRVEL integrates variants from 115,000 control individuals, 12.3 million variants, 6.95 million genotype-phenotype relationships, and 20,683 GO terms used to describe 235,928 model organism homolog-human gene pairs.

MARRVEL Facilitates Human Genomic Analysis

MARRVEL collects a wide range of data that can be used for multiple purposes for users from all fields. Here, we present just one of many ways that MARRVEL assists in human gene and variant analysis (Figure 2). In our approach, MARRVEL is used downstream of initial Whole Exome/Genome Sequencing bioinformatics analysis that results in a short list of candidate variants for a given individual's phenotype. MARRVEL first extracts key data from public human databases for gene-based analysis. We first display results from OMIM (Online Mendelian Inheritance in Man).[14] If the gene is documented at OMIM to be associated with a disease and the individual's phenotype is consistent, then the variant is likely causative. However, there is the caveat that in some cases the variant may be benign and this does not exclude the possibility that genetic alterations in other genes may also result in similar clinical phenotype. If a unique variant is in a disease-associated gene but the phenotypes are inconsistent with previously reported phenotypes, then this suggests a possible phenotypic expansion. If there are no known diseases or phenotypes associated with the gene in OMIM, then this may represent a potential disease-association for the gene.

The next set of data is used to assess whether variants in a specific gene is potentially pathogenic. The pLI score from ExAC expresses the probability that a gene is intolerant to loss-of-function alterations. For CNVs, the data that we collect are the deletion/duplications in the control population that contains the gene of interest. We obtain datasets from DGV and DECIPHER. The data obtained by DECIPHER is restricted to CNVs found in control population. DGV (Database of Genomic Variants) contains

copy-number variations from a large number of non-disease individuals (control populations from many published cohorts). A high frequency of deletions in the gene of interest in this population suggests that the gene tolerates haploinsufficiency. Similarly, a high frequency of duplications in the gene of interest in DGV suggests that gain of one copy is likely tolerated, depending on the specific location of duplications. DECIPHER similarly provides copy-number variations for a control population.

Next, MARRVEL displays the presence or absence of the variant of interest in ExAC, displayed as an estimate of allele frequency in a large cohort of individuals without early-onset disease. For candidate gene variants in individuals with early-onset disease and a proposed dominant mode of inheritance, the presence of the same variant in ExAC decreases the likelihood that the variant is pathogenic especially if the disease is early onset. However, ExAC does include data from populations known to be affected by adult-onset diseases, including schizophrenia and cardiovascular diseases. In contrast, if the variant is absent in ExAC, then the variant may be a potential candidate for further analysis. For candidate gene variants with a proposed recessive mode of inheritance, the presence of individuals homozygous for the variant of interest in ExAC suggests that different gene variants may need to be considered in evaluating disease pathogenesis.

ClinVar[16] is a valuable resource for researchers and clinicians to deposit gene variants and associated phenotypes. It contains more than 255,000 unique variants that are annotated with clinical significance and review status (i.e., level of evidence). When a user searches for a gene and variant, MARRVEL displays all ClinVar variants reported in the gene of interest, summarizes the number of variants in each category of clinical significance, and highlights the variant(s) that match the location of the variant of interest. If the variant of interest is documented in ClinVar as "benign" or "likely benign" with review status of "criteria provided, multiple submitters, no conflicts," then the variant is unlikely pathogenic. However, if the variant is designated as "risk factor," "likely pathogenic," or "uncertain significance" and with review status such as "no assertion criteria provided" or "single submitter," then the variant should remain a pathogenic candidate.

Geno2MP (Genotype to Mendelian Phenotype Browser) provides gene variant and phenotype correlation. Geno2MP provides cursory phenotypes for each sequenced individual in an affected population, as well as their unaffected relatives (if available), with human phenotype ontology (HPO) profiles. If a variant of interest is also present in an affected individual in Geno2MP, then the HPO terms would allow determination if similar biological systems are affected (i.e., potential phenotypic similarity). When both the genotype and phenotype of an affected individual in Geno2MP are consistent with the variant or gene of interest and the variant is not identified in an unaffected relative, then the variant is a pathogenic candidate.

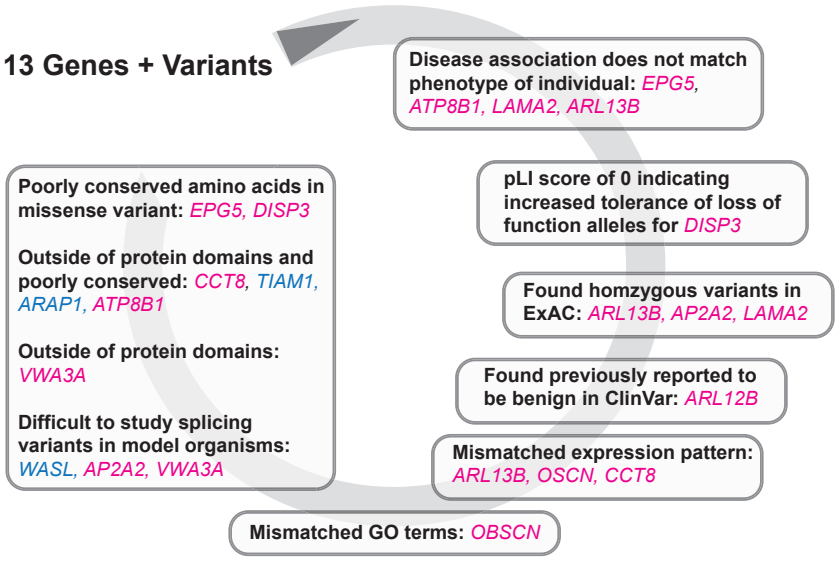
MARRVEL Curates Gene Function Data in Humans and Model Organisms

MARRVEL summarizes human and model organism data relevant to gene function in three main steps. The first step compares expression patterns in specific organs or tissues across human and model organisms (except for yeasts). For human expression data, the source of the data in MARRVEL is protein levels from the Protein Atlas.[26] In addition, GTEx provides quantitative expression data, RPKM (reads per kilobase per million mapped reads), of each gene in 53 human tissues. Model organism tissue expression data are obtained from individual model organism databases. Detailed information about the data sources can be found in Table S2. The tissue expression data serve at least three purposes. First, genes for which the pattern of expression is similar in humans and model organisms (e.g., expressed in comparable tissues) might be more likely to be informative in the context of human variant analysis. Second, display of the human tissue data allows for quick assessment of gene expression in the tissues affected in the individual under study. Third, expression patterns in human and model organism tissues can be used to design tissue-specific studies in model systems. One caveat to note is that the expression of a gene does not indicate necessity of the gene product in a specific tissue. In addition, reported developmental expression patterns often cover only specific stages and therefore may not provide valuable information. Moreover, many genes are only transiently expressed or their expression is induced only under specific environmental or physiological conditions. Finally, expression of many genes is below detection levels of current techniques.

The second step compares GO terms across human and model organisms for biological process, molecular function, and cellular component.[33] GO terms are useful to quickly compare biological and molecular functions of the gene across species. In many cases, a gene may be well studied in one or more model systems but not in others. Data from the model organisms can be compared to provide insight into the degree of conservation, reveal possible disease mechanisms, and assist in the selection of one or more specific model organism for further mechanistic study.

The third step examines the conservation of specific amino acids and protein domains among orthologs based on multiple alignments of the human protein sequence and putative orthologs in model organisms (Figure S2). The alignment provides information on conservation of the amino acid or functional domain affected by the missense variant. MARRVEL also lists functional domains present in the human protein, highlighted in the multiple alignment. These steps further assist in determining whether there is evolutionary selection against variation at the residues analogous to the human variant of interest and in the selection of model organisms for pursuing further study of disease mechanism.

A



B

OGDHL *Chr 10:50946295 G>A*

MARRVEL output	Useful Data
OMIM	No OMIM phenotype association
ExAC/ClinVar/Geno2MP	Not found
Gene Ontology	Microtubule and mitochondrion association
Expression Pattern	Highly expressed in human cerebellum
Multiple protein alignment	Highly conserved amino acid from yeast to human, located in the catalytic domain

Figure 3. Example of Variant Analysis using MARRVEL

We re-analyzed a previously published case by Yoon et al.[13] by following our strategy outlined in Figure 2.

(A) Magenta genes and variants are eliminated based on multiple criteria. Blue genes and variants are eliminated with only one criteria and users may consider further analysis. The arrangement of the chart is reflective of Figure 2 which explains our strategy of analysis.

(B) *OGDHL* is the most likely candidate out of the 13 genes and variants to cause the individual's phenotypes based on MARRVEL data. For more details on the genes and variants, see Table S3.

Case Example of How MARRVEL Facilitates Gene and Variant Analysis

Here, we provide an example of how MARRVEL displays information useful for variant prioritization in an output to facilitate analysis of genes and variants. We describe a specific case for which MARRVEL can be used to facilitate manual analysis of putative human disease-causing variants in an individual (Figure 3).

Yoon et al.[13] recently successfully performed functional studies in *Drosophila* to demonstrate the pathogenicity of a gene variant found in a proband with a neurodegenerative phenotype. Previous analysis performed by an expert identified a homozygous variant in *OGDHL* (HGNC:25590) as the likely cause of the proband's phenotype.[34] Yoon et al.[13] subsequently showed that loss of *Ogdh*, the *Drosophila* homolog of *OGDHL*, exhibits a neurodegenerative phenotype consistent with the proband's neurological disorder. As an example of how MARRVEL can be used to conduct variant analysis, we obtained a list of 13 candidate variants for this case[13] and conducted a variant analysis using output from MARRVEL to determine whether we were able to come to a similar conclusion.

The example presented is a 15-year-old girl with developmental delay, microcephaly, ataxia, motor impairment, hypotonia, language impairments, brain abnormalities, and hypoplasia of the corpus callosum. She was identified in a consanguineous family in a Turkish brain malformation cohort.[34] The proband, her unaffected parents, and an unaffected sibling received whole-exome sequencing. After filtering for variants that were both unique to the proband and rare in the population (at least <0.01 minor allele frequency), variants in 13 different genes remained. Subsequent steps illustrate how MARRVEL can be incorporated downstream of whole-exome or -genome analysis pipelines.

We manually filtered and analyze Yoon et al.'s list of 13 candidates[13] through MARRVEL's synopsis of publicly available databases including OMIM and ExAC; tissue expression patterns; and the location of the amino acid change relative to known functional domains. Table S3 shows the manual analysis of the MARRVEL output of these 13 genes in comparison to the original analysis by an expert (see Table S2 in Yoon et al.[13]). The first step is to examine any existing phenotypic associations with the gene. Of the 13 genes, 4—*EPG5* (MIM: 615068, HGNC:29331), *ATP8B1* (MIM: 602397, HGNC:3706), *ARL13B* (MIM: 608922, HGNC:25419), and *LAMA2* (MIM: 156225, HGNC:6482)—have a disease association that is partially or completely inconsistent with the individual's phenotype. Although phenotypic expansion may still be possible, our current strategy defers that possibility until all other candidate genes are ruled out.

Based on the ExAC data, a gene suspected to have a de novo variant, *DISP3* (MIM: 611251, HGNC:29251), has a pLI score of 0, indicating a high tolerance of loss-of-function variants. In addition, there are three candidate genes—*ARL13B*, *AP2A2* (MIM: 607242, HGNC:562), and *LAMA2*—in the individual that are either homozygous or compound heterozygous variants. For these three variants the same homozygous mutations are listed in ExAC,

suggesting that these variants are unlikely to result in early-onset developmental disorders. The *ARL13B* variant was also reported in ClinVar as benign and likely benign by multiple submitters. Through curation of human genomics information, the list of 13 candidate genes was narrowed to 6 remaining genes.

Model organism gene expression data and biological function GO terms were analyzed next. For three genes—*ARL13B*, *OBSCN* (MIM: 608616, HGNC:15719), and *CCT8* (HGNC:1623)—the tissue expression pattern did not match the nervous system involvement in the individual of interest. Additionally, for *OBSCN* the GO terms across model organisms exclusively focuses on muscle development, structure, and function, making it less likely to be involved in nervous system-related pathology.

Further analysis of the missense variants revealed that the affected amino acid residues in *EPG5* and *DISP3* are poorly conserved amino acids across model organisms. In addition, the variant in *ARL13B* affects a site outside of the coding regions or protein domains, and variants in *CCT8*, *TIAM1* (MIM: 600687, HGNC:11805), *ARAP1*, and *ATP8B1* are poorly conserved and encode residues located outside of protein domains. These variants are therefore less likely to disrupt protein function. Splicing variants such as those found in *WASL* (MIM: 605056, HGNC: 12735) are difficult to study in model organisms and should be pursued in alternative approaches such as quantitative measure of mRNA in human samples.

Altogether, the homozygous variant in *OGDHL* emerged as the best pathogenic candidate for further study based on the human and model organism output from MARRVEL. In the model organism output from MARRVEL, three lines of information suggested that *OGDHL* is a promising candidate for further study in model organisms (Figure 3B). (1) Although the gene had not been functionally studied in vertebrate or *Drosophila*, the gene has been linked to mitochondria function in *C. elegans* and yeast, consistent with some of the neurodegenerative phenotypes. (2) Expression data in human and model organisms suggest that the gene is highly expressed in the affected tissue (brain). (3) The amino acid is highly conserved throughout evolution and is located in a highly conserved stretch of the protein. Indeed, Yoon et al.[13] showed that flies with a null allele of *Ogdh* exhibit neurodegenerative phenotypes consistent with a neurological disorder. Importantly, the variant found in the individual corresponds to a severe loss-of-function allele based on gene humanization and rescue experiments in *Drosophila*,[8] indicating that *OGDHL* is the likely candidate responsible for the neurological phenotype. In summary, MARRVEL displays information that provides input for the prioritization of potentially disease-causing variants for functional validation (Figure 3).

We recognize that there are multiple approaches to the analysis of possible disease-causing variants. Above, we provided one example of using reanalysis of published data for how MARRVEL can be applied to the downstream analysis of sequencing data for determination of candidate disease genes. If inheritance pattern is unclear, then multiple parallel analyses should be performed assuming that the variant could result in either dominant or recessive phenotypes. Furthermore, the variant interpretation can be evaluated for possible functional consequences including gain of function, haploinsufficiency, and dominant negative.

Discussion

In summary, MARRVEL affords an efficient aggregation of information from multiple human genomics and model organism databases to allow for rapid view and assessment of candidate genes and variants. OMIM provides fundamental information about disease association for the gene of interest. ExAC provides a powerful resource for examining the allele frequency of rare variants and can be used to prioritize the frequency of a coding variant and potential pathogenicity.[15] Geno2MP and ClinVar provide unique sources of phenotypic and interpretation data for a variant of interest. DGV and DECIPHER control population provide publically available data, copy-number variants in apparently healthy individuals, which complements the data from ExAC, Geno2MP, and ClinVar. MARRVEL displays all of this information in a concise format providing highly integrated, convenient, and fast access (Figure S1). For potential genes in which disruption may cause disease, there is often limited in vivo human gene functional data; however, there is a wealth of information in model organisms that can be used to develop meaningful hypotheses regarding human gene function and to inform the likelihood that a variant causes or contributes to a disease phenotype. For example, in the case of *OGDHL*, integrating human and model organism data in MARRVEL allows us to aggregate all the information needed to prioritize this gene and variant to be tested experimentally. One key benefit of MARRVEL is allowing the data from model organism databases to be reviewed in a concise format. In MARRVEL, key biological and genetic features of putative orthologous genes, including tissue expression pattern and Gene Ontology (GO) terms, are extracted from model organism databases. MARRVEL displays all the relevant information normally assessed in a manual analysis pipeline described in Figure 2.

Several bioinformatics tools exist for aggregating available data to increase efficiency of variant analysis. For example, GeneCards is an aggregation of human gene-centric data. MARRVEL and GeneCards have a number of overlapping datasets. However, MARRVEL places more emphasis on human variant data (ExAC, ClinVar, etc.) and has a much broader range of data from model organisms. Combined Annotation Dependent Depletion (CADD)[35] and PolyPhen[36] focus on predicting the pathogenicity of an amino acid change. These two tools incorporate a combination of homology, structural, and machine learning analysis to predict whether

or not a single amino acid change is likely to disrupt protein function.[37,38] However, there are cases where additional population frequency data and model organism phenotypic data are needed to improve variant interpretation.[35,39]

The Monarch Initiative[40] addresses the challenge of annotating the human genome by gathering data on known phenotypes in other organisms (phenotype-centric) to assist in variant analysis whereas MARRVEL provides a gene-centric toolkit including non-vertebrate model organisms and protein alignments. Although most bioinformatics tools and strategies are useful guides, combining multiple resources often provides a better view of the variant and higher predictive value when analyzing variants and genes.

Clinical genetics labs and research sequencing centers have access to well-established variant analysis and annotation pipelines such as Exomiser/PHenotypic Interpretation of Variants in Exomes (PHIVE),[41] ANNOVAR,[42] and Codified Genomics (see Web Resources) that utilize existing tools to analyze entire sets of sequencing data. These require familiarity with bioinformatics data processing and access to these resources. By contrast, clinicians and model organism researchers often have access only to variants reported in clinical sequencing reports and in the literature. Furthermore, the majority of clinicians and model organism researchers lack training in bioinformatics data analysis. In the absence of an integrated pipeline, it is difficult for clinicians and basic scientists to efficiently obtain information on candidate disease variants, as the information needed is spread across various databases and tools for variant analysis (Figure S1).

Despite an increasing interest to utilize model organism data in human genetic analysis pipelines such as in the Monarch Initiative and Exomiser/PHIVE, the current major focus is on matching phenotypic information.[41,43] Although the similarities between human and model organism mutant phenotype can be informative, this approach may miss numerous opportunities in which the protein functions are part of conserved pathways among organisms when the orthologous phenotypes are not obviously analogous.[44] For example, a yeast model for angiogenesis[44] and a worm model for breast cancer[44] revealed molecular pathways that contribute to these disorders based on the "phenology" concept. Therefore, we adapted a gene-centric rather than phenotypic-centric approach to study gene function by integrating model organism and human data in a single aggregated web-based resource.

Many model organism databases, such as FlyBase,[8] WormBase,[21] and ZFIN,[10] are comprehensive and contain a monumental amount of data accumulated over numerous decades.[45] However, the extremely valuable information in these databases is not easily accessible by those outside the field. Importantly, there is a barrier to search specific model organism databases due to the intricacies of evaluating orthologs and navigating different websites and the different use of gene or protein identifiers (Figure S1). MARRVEL organizes this information across multiple species in a clear and concise way and also provides the best predicted orthologs.

In recent years, whole-exome or -genome sequencing has increasingly been used to assist in the diagnosis of human diseases.[46] Meaningful analysis and interpretation of the sequencing results require a team of dedicated experts. Current bioinformatics pipelines are efficient at identifying previously reported pathogenic variants in known human genes in which disruption causes disease. By filtering out previously identified benign variants, as well as those appearing at a high frequency in control populations, the number of potentially disease-causing variants can be narrowed down significantly. Further analysis to identify variants to functionally test in model organisms will benefit from a survey of currently available model organism data. In conclusion, MARRVEL is a flexible web resource that provides a useful and accessible tool for efficiently matching an input against 18 million records of human variants and genes as well as model organism homologs. MARRVEL provides a step toward the overarching goal of integrating model organism databases with human gene-centric user interfaces[41] to improve the accessibility and evaluation of data typically used by experts fluent in specific data formats and software. Our future goals for MARRVEL include continuing to integrate additional human genomics and model organism resources as they become publicly available to ensure that MARRVEL remains a valuable and up-to-date analytical resource.

Supplemental Data

Supplemental Data include three figures, three tables, and Supplemental Acknowledgments and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2017.04.010.

Consortia

Members of the Undiagnosed Diseases Network are Christopher J. Adams, David R. Adams, Mercedes E. Alejandro, Patrick Allard, Euan A. Ashley, Mashid S. Azamian, Carlos A. Bacino, Ashok Balasubramanyam, Hayk Barseghyan, Alan H. Beggs, Hugo J. Bellen, Jonathan A. Bernstein, Anna Bican, David P. Bick, Camille L. Birch, Braden E. Boone, Lauren C. Briere, Donna M. Brown, Matthew Brush, Elizabeth A. Burke, Lindsay C. Burrage, Katherine R. Chao, Gary D. Clark, Joy D. Cogan, Cynthia M. Cooper, William J. Craigen, Mariska Davids, Jyoti G. Dayal, Esteban C. Dell'Angelica, Shweta U. Dhar, Katrina M. Dipple, Laurel A. Donnell-Fink, Naghmeh Dorrani, Daniel C. Dorset, David D. Draper, Annika M. Dries, David J. Eckstein, Lisa T. Emrick, Christine M. Eng, Cecilia Esteves, Tyra Estwick, Paul G. Fisher, Trevor S. Frisby, Kate Frost, William A. Gahl, Valerie Gartner, Rena A. Godfrey, Mitchell Goheen, Gretchen A. Golas, David B. Goldstein, Mary G. Gordon, Sarah E. Gould, Jean-Philippe F. Gourdine, Brett H. Graham, Catherine A. Groden, Andrea L. Gropman, Mary E. Hackbarth, Melissa Haendel, Rizwan Hamid, Neil A. Hanchard, Lori H. Handley, Isabel Hardee, Matthew R. Herzog, Ingrid A. Holm, Ellen M. Howerton, Howard J. Jacob, Mahim Jain, Yong-hui Jiang, Jean M. Johnston, Angela L. Jones, Alanna E. Koehler, David M. Koeller, Isaac S. Kohane, Jennefer N. Kohler, Donna M. Krasnewich, Elizabeth L. Krieg, Joel B. Krier, Jennifer E. Kyle, Seema R. Lalani, Lea Latham, Yvonne L. Latour, C. Christopher Lau, Jozef Lazar, Brendan H. Lee,

Hane Lee, Paul R. Lee, Shawn E. Levy, Denise J. Levy, Richard A. Lewis, Adam P. Liebendorfer, Sharyn A. Lincoln, Carson R. Loomis, Joseph Loscalzo, Richard L. Maas, Ellen F. Macnamara, Calum A. MacRae, Valerie V. Maduro, May Christine V. Malicdan, Laura A. Mamounas, Teri A. Manolio, Thomas C. Markello, Paul Mazur, Alexandra J. McCarty, Allyn McConkie-Rosell, Alexa T. McCray, Thomas O. Metz, Matthew Might, Paolo M. Moretti, John J. Mulvihill, Jennifer L. Murphy, Donna M. Muzny, Michele E. Nehrebecky, Stan F. Nelson, J. Scott Newberry, John H. Newman, Sarah K. Nicholas, Donna Novacic, Jordan S. Orange, J. Carl Pallais, Christina G.S. Palmer, Jeanette C. Papp, Loren D.M. Pena, John A. Phillips III, Jennifer E. Posey, John H. Postlethwait, Lorraine Potocki, Barbara N. Pusey, Rachel B. Ramoni, Amy K. Robertson, Lance H. Rodan, Jill A. Rosenfeld, Sarah Sadozai, Katherine E. Schaffer, Kelly Schoch, Molly C. Schroeder, Daryl A. Scott, Prashant Sharma, Vandana Shashi, Edwin K. Silverman, Janet S. Sinsheimer, Ariane G. Soldatos, Rebecca C. Spillmann, Kimberly Splinter, Joan M. Stoler, Nicholas Stong, Kimberly A. Strong, Jennifer A. Sullivan, David A. Sweetser, Sara P. Thomas, Cynthia J. Tifft, Nathanial J. Tolman, Camilo Toro, Alyssa A. Tran, Zaheer M. Valivullah, Eric Vilain, Daryl M. Waggott, Colleen E. Wahl, Nicole M. Walley, Chris A. Walsh, Michael F. Wangler, Mike Warburton, Patricia A. Ward, Katrina M. Waters, Bobbie-Jo M. Webb-Robertson, Alec A. Weech, Monte Westerfield, Matthew T. Wheeler, Anastasia L. Wise, Lynne A. Wolfe, Elizabeth A. Worthey, Shinya Yamamoto, Yaping Yang, Guoyun Yu, and Patricia A. Zornio.

Acknowledgments

We thank the Undiagnosed Diseases Network Model Organism Working Group and Coordinating Center, Jim Lupski, Zeynep Akdemir, Ender Karaca, John Seavitt, George Eisenhoffer, Swathi Arur, Grzegorz Ira, and Karen Schulze for providing input in the design of MARRVEL. We thank Wan Hee Yoon for providing input in the manuscript. This work was supported by the NINDS (1U54NS093793-01) to the Model Organisms Screening Center of the UDN and NIH/ORIP (1R24 OD022005-01). J.W. is supported by The Robert and Janice McNair Foundation McNair MD/PhD Student Scholar Program and Baylor College of Medicine Medical Scientist Training Program. H.J.B. and N.P. are Investigators of the Howard Hughes Medical Institute. S.-Y.K., S.Y., M.F.W., and H.J.B. are supported by NIH (3U54NS093793-02S1). H.J.B. is supported by NIH (R01 GM067858). Z.L., R.A.-O., and W.-W.W. are supported by NSF (DMS 1263932), CPRIT (RP170387), NIH (R01 GM120033), Houston Endowment, Huffington Foundation, Belfer Foundation, and T T Chao Family Foundation. N.P., Y.H., A.C., and S.E.M. are supported by NIH NIGMS (R01 GM067761, NIGMS R01 GM084947), NIH (R24 RR032668, R24 OD021997 to N.P., P.I.), and Dana Farber/Harvard Cancer Center (NCI Cancer Center Support Grant # NIH 5 P30 CA06516 to S.E.M.). H.-T.C. is supported by the Pediatric Neurology Basic Neuroscience Research Track residency training program at Baylor College of Medicine. S.Y. is supported by the Texas Children's Hospital (NRI Fellowship) and Alzheimer's Association (New Investigator Research Grant NIRG-15-364099). M.F.W. is supported by NIH (U01HG007709). S.Y. and M.F.W. are supported by the Simons Foundation (#368479 SFARI Functional Screen of Autism-Associated Variants Award).

Received: February 22, 2017
Accepted: April 18, 2017
Published: May 11, 2017

Web Resources

Angular JS v.1.6.1, https://angularjs.org/
Bootstrap v.4.0.0, v4-alpha.getbootstrap.com
ClinVar, https://www.ncbi.nlm.nih.gov/clinvar/
Codified Genomics, http://codifiedgenomics.com/
Database of Genomic Variants (DGV), http://dgv.tcag.ca/dgv/app/home
DECIPHER, http://decipher.sanger.ac.uk/
DIOPT, http://www.flyrnai.org/diopt
Ensembl GRCh37 Rest API, http://grch37.rest.ensembl.org
ExAC Browser, http://exac.broadinstitute.org/
FlyBase, http://flybase.org/
Geno2MP (March 2017 accessed), http://geno2mp.gs.washington.edu/Geno2MP/#/
HUGO Gene Nomenclature Committee, http://www.genenames.org/
jQuery v.2.2.0, https://jquery.com/
MARRVEL, http://marrvel.org/
Mouse Genome Informatics, http://www.informatics.jax.org/
Mutalyzer, https://mutalyzer.nl/index
Node.js framework v.6.7.0, https://nodejs.org/en/
OMIM, http://www.omim.org/
PomBase, https://www.pombase.org/
QuickGO, https://www.ebi.ac.uk/QuickGO/
Saccharomyces Genome Database, http://www.yeastgenome.org/
The Human Protein Atlas, http://www.proteinatlas.org/
Undiagnosed Diseases Network, https://undiagnosed.hms.harvard.edu/
WormBase, http://www.wormbase.org/
ZFIN, http://zfin.org

References

1. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al.; ACMG Laboratory Quality Assurance Committee (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–424.
2. Bellen, H.J., and Yamamoto, S. (2015). Morgan's legacy: fruit flies and the functional annotation of conserved genes. Cell *163*, 12–14.
3. Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., McMillin, M.J., Wiszniewski, W., Gambin, T., et al.; Centers for Mendelian Genomics (2015). The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. Am. J. Hum. Genet. *97*, 199–215.
4. Gahl, W.A., Mulvihill, J.J., Toro, C., Markello, T.C., Wise, A.L., Ramoni, R.B., Adams, D.R., Tifft, C.J.; and UDN (2016). The NIH Undiagnosed Diseases Program and Network: applications to modern medicine. Mol. Genet. Metab. *117*, 393–400.
5. O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., Mackenzie, A.P., Ng, S.B., Baker, C., et al. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat. Genet. *43*, 585–589.
6. Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., van Kogelenberg, M., King, D.A., Ambridge, K., Barrett, D.M., Bayzetinova, T., et al.; DDD study (2015). Genetic

diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet *385*, 1305–1314.

7. Ramocki, M.B., and Zoghbi, H.Y. (2008). Failure of neuronal homeostasis results in common neuropsychiatric phenotypes. Nature *455*, 912–918.
8. Marygold, S.J., Crosby, M.A., Goodman, J.L.; and FlyBase Consortium (2016). Using FlyBase, a database of *Drosophila* genes and genomes. Methods Mol. Biol. *1478*, 1–31.
9. Eppig, J.T., Smith, C.L., Blake, J.A., Ringwald, M., Kadin, J.A., Richardson, J.E., and Bult, C.J. (2017). Mouse genome informatics (MGI): resources for mining mouse genetic, genomic, and biological data in support of primary and translational research. Methods Mol. Biol. *1488*, 47–73.
10. Ruzicka, L., Bradford, Y.M., Frazer, K., Howe, D.G., Paddock, H., Ramachandran, S., Singer, A., Toro, S., Van Slyke, C.E., Eagle, A.E., et al. (2015). ZFIN, the zebrafish model organism database: updates and new directions. Genesis *53*, 498–509.
11. Chao, H.-T., Davids, M., Burke, E., Pappas, J.G., Rosenfeld, J.A., McCarty, A.J., Davis, T., Wolfe, L., Toro, C., Tifft, C., et al. (2017). A syndromic neurodevelopmental disorder caused by de novo variants in EBF3. Am. J. Hum. Genet. *100*, 128–137.
12. Chen, K., Ho, T.S.-Y., Lin, G., Tan, K.L., Rasband, M.N., Bellen, H.J., Ackermann, E., Guo, S., Booten, S., Alvarado, L., et al. (2016). Loss of Frataxin activates the iron/sphingolipid/PDK1/Mef2 pathway in mammals. eLife *5*, 43–44.
13. Yoon, W.H., Sandoval, H., Nagarkar-Jaiswal, S., Jaiswal, M., Yamamoto, S., Haelterman, N.A., Putluri, N., Putluri, V., Sreekumar, A., Tos, T., et al. (2017). Loss of Nardilysin, a mitochondrial co-chaperone for α-Ketoglutarate Dehydrogenase, promotes mTORC1 activation and neurodegeneration. Neuron *93*, 115–131.
14. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. *43*, D789–D798.
15. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
16. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44* (D1), D862–D868.
17. MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. *42*, D986–D992.
18. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: database of chromosomal imbalance and phenotype in humans using Ensembl resources. Am. J. Hum. Genet. *84*, 524–533.
19. Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., et al. (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. *40*, D700–D705.
20. Wood, V., Harris, M.A., McDowall, M.D., Rutherford, K., Vaughan, B.W., Staines, D.M., Aslett, M., Lock, A., Bähler, J., Kersey, P.J., and Oliver, S.G. (2012). PomBase: a comprehensive online resource for fission yeast. Nucleic Acids Res. *40*, D695–D699.
21. Howe, K.L., Bolt, B.J., Cain, S., Chan, J., Chen, W.J., Davis, P., Done, J., Down, T., Gao, S., Grove, C., et al. (2016). WormBase 2016: expanding to enable helminth genomic research. Nucleic Acids Res. *44* (D1), D774–D780.
22. Shimoyama, M., De Pons, J., Hayman, G.T., Laulederkind, S.J.F., Liu, W., Nigam, R., Petri, V., Smith, J.R., Tutaj, M., Wang, S.-J., et al. (2015). The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. Nucleic Acids Res. *43*, D743–D750.
23. Yu, Y., Fuscoe, J.C., Zhao, C., Guo, C., Jia, M., Qing, T., Bannon, D.I., Lancashire, L., Bao, W., Du, T., et al. (2014). A rat RNA-seq transcriptomic BodyMap across 11 organs and 4 developmental stages. Nat. Commun. *5*, 3230.
24. Huntley, R.P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M.J., and O'Donovan, C. (2015). The GOA database: gene ontology annotation updates for 2015. Nucleic Acids Res. *43*, D1057–D1063.
25. Carithers, L.J., Ardlie, K., Barcus, M., Branton, P.A., Britton, A., Buia, S.A., Compton, C.C., DeLuca, D.S., Peter-Demchok, J., Gelfand, E.T., et al.; GTEx Consortium (2015). A novel approach to high-quality postmortem tissue procurement: the GTEx Project. Biopreserv. Biobank. *13*, 311–319.
26. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419.
27. Hu, Y., Flockhart, I., Vinayagam, A., Bergwitz, C., Berger, B., Perrimon, N., Mohr, S.E., McKusick, V., Hamosh, A., Scott, A., et al. (2011). An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics *12*, 357.
28. Yates, B., Braschi, B., Gray, K.A., Seal, R.L., Tweedie, S., and Bruford, E.A. (2017). Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res. *45* (D1), D619–D625.
29. den Dunnen, J.T. (2017). Describing sequence variants using HGVS nomenclature. Methods Mol. Biol. *1492*, 243–251.
30. Wildeman, M., van Ophuizen, E., den Dunnen, J.T., and Taschner, P.E.M. (2008). Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum. Mutat. *29*, 6–13.
31. Axmark, D., and Widenius, M. (2002). MySQL Reference Manual, P. DuBois, ed. (O'Reilly & Assoc.).
32. NCBI Resource Coordinators (2016). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. *44* (D1), D7–D19.
33. Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. Nucleic Acids Res. *43*, D1049–D1056.
34. Karaca, E., Harel, T., Pehlivan, D., Jhangiani, S.N., Gambin, T., Coban Akdemir, Z., Gonzaga-Jauregui, C., Erdin, S., Bayram, Y., Campbell, I.M., et al. (2015). Genes that affect brain structure and function identified by rare variant analyses of Mendelian neurologic disease. Neuron *88*, 499–513.
35. Miosge, L.A., Field, M.A., Sontani, Y., Cho, V., Johnson, S., Palkova, A., Balakishnan, B., Liang, R., Zhang, Y., Lyon, S., et al. (2015). Comparison of predicted and actual consequences of missense mutations. Proc. Natl. Acad. Sci. USA *112*, E5189–E5198.
36. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010).

A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

37. Katsonis, P., Koire, A., Wilson, S.J., Hsu, T.-K., Lua, R.C., Wilkins, A.D., and Lichtarge, O. (2014). Single nucleotide variations: biological impact and theoretical interpretation. Protein Sci. *23*, 1650–1666.
38. Katsonis, P., and Lichtarge, O. (2014). A formal perturbation equation between genotype and phenotype determines the evolutionary action of protein-coding variations on fitness. Genome Res. *24*, 2050–2058.
39. Sun, S., Yang, F., Tan, G., Costanzo, M., Oughtred, R., Hirschman, J., Theesfeld, C.L., Bansal, P., Sahni, N., Yi, S., et al. (2016). An extended set of yeast-based functional assays accurately identifies human disease mutations. Genome Res. *26*, 670–680.
40. Mungall, C.J., McMurry, J.A., Köhler, S., Balhoff, J.P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., et al. (2017). The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res. *45* (D1), D712–D722.
41. Robinson, P.N., Köhler, S., Oellrich, A., Wang, K., Mungall, C.J., Lewis, S.E., Washington, N., Bauer, S., Seelow, D., Krawitz, P., et al.; Sanger Mouse Genetics Project (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. *24*, 340–348.
42. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.
43. Mungall, C., McMurry, J., Koehler, S., Balhoff, J., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., et al. (2016). The Monarch Initiative: Insights across species reveal human disease mechanisms. bioRxiv. http://dx.doi.org/10.1101/055756.
44. McGary, K.L., Park, T.J., Woods, J.O., Cha, H.J., Wallingford, J.B., and Marcotte, E.M. (2010). Systematic discovery of nonobvious human disease models through orthologous phenotypes. Proc. Natl. Acad. Sci. USA *107*, 6544–6549.
45. Kaiser, J., Lock, A., Harris, M.A., Nurse, P., Wood, V., Kaiser, J., Bond, M., Holthaus, S.-M., Tammen, I., Tear, G., et al. (2016). BIOMEDICAL RESOURCES. Funding for key data resources in jeopardy. Science *351*, 14–14.
46. Yang, Y., Muzny, D.M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C., et al. (2014). Molecular findings among patients referred for clinical whole-exome sequencing. JAMA *312*, 1870–1879.

Chad Genetic Diversity Reveals an African History Marked by Multiple Holocene Eurasian Migrations

Marc Haber,[1,*] Massimo Mezzavilla,[1,2] Anders Bergström,[1] Javier Prado-Martinez,[1] Pille Hallast,[1,3] Riyadh Saif-Ali,[4] Molham Al-Habori,[4] George Dedoussis,[5] Eleftheria Zeggini,[1] Jason Blue-Smith,[6,10] R. Spencer Wells,[7] Yali Xue,[1] Pierre A. Zalloua,[8,9] and Chris Tyler-Smith[1,*]

Understanding human genetic diversity in Africa is important for interpreting the evolution of all humans, yet vast regions in Africa, such as Chad, remain genetically poorly investigated. Here, we use genotype data from 480 samples from Chad, the Near East, and southern Europe, as well as whole-genome sequencing from 19 of them, to show that many populations today derive their genomes from ancient African-Eurasian admixtures. We found evidence of early Eurasian backflow to Africa in people speaking the unclassified isolate Laal language in southern Chad and estimate from linkage-disequilibrium decay that this occurred 4,750–7,200 years ago. It brought to Africa a Y chromosome lineage (R1b-V88) whose closest relatives are widespread in present-day Eurasia; we estimate from sequence data that the Chad R1b-V88 Y chromosomes coalesced 5,700–7,300 years ago. This migration could thus have originated among Near Eastern farmers during the African Humid Period. We also found that the previously documented Eurasian backflow into Africa, which occurred ~3,000 years ago and was thought to be mostly limited to East Africa, had a more westward impact affecting populations in northern Chad, such as the Toubou, who have 20%–30% Eurasian ancestry today. We observed a decline in heterozygosity in admixed Africans and found that the Eurasian admixture can bias inferences on their coalescent history and confound genetic signals from adaptation and archaic introgression.

Introduction

African genetic diversity is still incompletely understood, and vast regions in Africa remain genetically undocumented. Chad, for example, makes up ~5% of Africa's surface area, and its central location, connecting sub-Saharan Africa with North and East Africa, positions it to play an important role as a crossroad or barrier to human migrations. However, Chad has been little studied at a whole-genome level, and its position within African genetic diversity is not well known. With 200 ethnic groups and more than 120 indigenous languages and dialects, Chad has extensive ethnolinguistic diversity.[1] It has been suggested that this diversity can be attributed to Lake Chad, which has attracted human populations to its fertile surroundings since prehistoric times, especially after the progressive desiccation of the Sahara starting ~7,000 years ago (ya).[2,3]

Important questions about Africa's ethnic diversity are the relationships among the different groups and the relationships between cultural groups and existing genetic structures. In the present study, we analyzed four Chadian populations with different ethnicities, languages, and modes of subsistence. Our samples are likely to capture recent genetic signals of migration and mixing and also have the potential to show ancestral genomic relationships that are shared among Chadians and other populations. An additional major question relates to the prehistoric Eurasian migrations to Africa: what was the extent of these migrations, how have they affected African genetic diversity, and what present-day populations harbor genetic signals from the ancient migrating Eurasians? We have previously reported evidence of gene flow from the Near East to East Africa ~3,000 ya, as well as subsequent selection in Ethiopians on non-African-derived alleles related to light skin pigmentation.[4] A recent attempt to quantify the extent of such backflow into Africa more generally, by using ancient DNA (aDNA), suggested that the impact of the Eurasian migration was mostly limited to East Africa.[5] However, previous studies using mitochondrial DNA and the Y chromosome in populations from the Chad Basin found some with an East African[6] or Mediterranean and Eurasian influence,[7,8] and analysis based on genome-wide data[9] found a non-African component (suggested to be from East Africa) in central Sahelian populations. Thus, studying diverse Chadian populations on a whole-genome level presents an opportunity to shed more light on the history of African-Eurasian mixtures, including whether or not selection after admixture is a widespread phenomenon in Africa and how the historical events in Chad are related to events that have occurred elsewhere in Africa and the Near East.

In this work, we present a genetic dataset of 480 Chadian, Near Eastern, and European individuals genotyped at 2.5 million SNPs, in addition to high-coverage

[1]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK; [2]Institute for Maternal and Child Health, IRCCS Burlo Garofolo, University of Trieste, 34137 Trieste, Italy; [3]Institute of Molecular and Cell Biology, University of Tartu, Tartu 51010, Estonia; [4]Department of Biochemistry and Molecular Biology, Faculty of Medicine and Health Sciences, Sana'a University, Sana'a 19065, Yemen; [5]Department of Nutrition and Dietetics, Harokopio University Athens, Athens 17671, Greece; [6]National Geographic Society, Washington, DC 20036, USA; [7]Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712, USA; [8]Lebanese American University, Chouran, Beirut 1102 2801, Lebanon; [9]Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
[10]Present address: Karius, Inc., 1505A Adams Drive, Menlo Park, CA 94025, USA
*Correspondence: mh25@sanger.ac.uk (M.H.), cts@sanger.ac.uk (C.T.-S.)
http://dx.doi.org/10.1016/j.ajhg.2016.10.012.

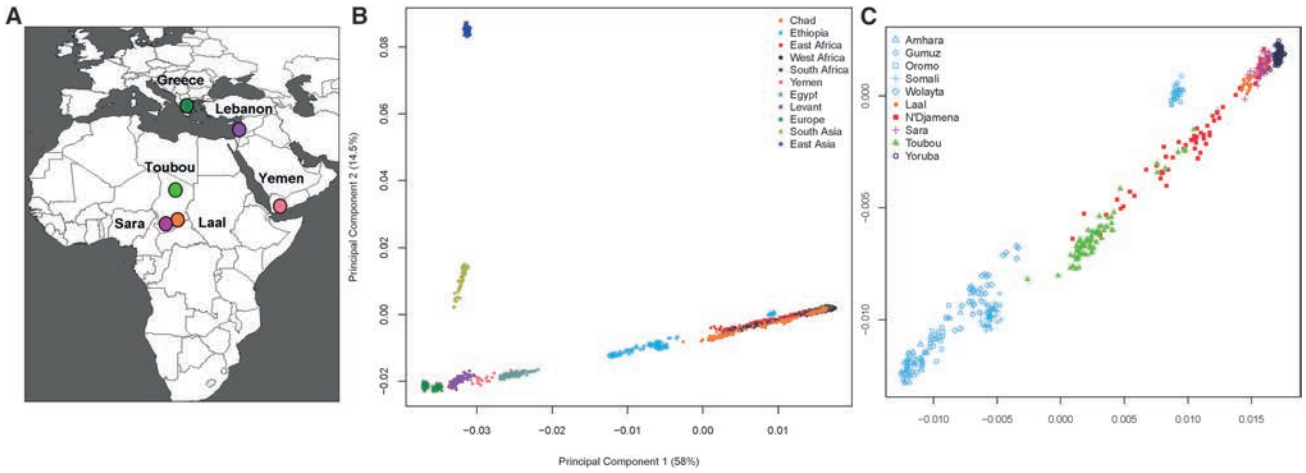


Figure 1. Population Locations and Genetic Structure
(A) The map shows the location of newly genotyped or sequenced populations.
(B) PCA of worldwide populations shows that Near Easterners and East Africans are intermediate to Eurasians and sub-Saharan Africans on PC1. Chad populations are close to sub-Saharan Africans and have some samples drawn toward Ethiopians.
(C) Magnification of the African PCA shows different affinities of the Chad populations to other Africans: the Toubou cluster close to Ethiopians, whereas the Sara and Laal speakers are close to the Yoruba. The mixed samples from N'Djamena, the capital, are intermediate to the Toubou, Sara, and Laal speakers.

whole-genome sequences from 19 of these individuals. From Chad, we studied (1) the Toubou, who are nomads from northern Chad and speak a Nilo-Saharan language; (2) the Sara, who are a sedentary population from southern Chad and also speak a Nilo-Saharan language; (3) the Laal speakers, a population of just ~750 individuals who speak an unclassified language isolate and live in southern Chad; and (4) an urban population from the capital city of N'Djamena. In addition to the Chadians, we included Greek, Lebanese, and Yemen samples whose location and history suggest that they might be informative about early African-Eurasian migrations. We used this dataset to advance our understanding of human genetic diversity in Africa and neighboring regions by focusing on population migration and mixing and how the admixture process has shaped present-day genetic variation.

Subjects and Methods

Samples and Data

Samples were collected from Chad (238), Lebanon (126), Greece (96), and Yemen (20) (Figure 1A); details can be found in Table S1. All samples (except for those from Greece) were genotyped with the Illumina HumanOmni2.5-8 BeadChip, which covers ~2.5 million SNPs. Greek genotype information for the 2.5 million sites was extracted from sequence data (E.Z., unpublished data) and merged with array data from other populations. In addition, 19 samples (Chad [11], Greece [4], and Lebanon [4]) were whole-genome sequenced at >30× depth with Illumina HiSeq X Ten or HiSeq 2500 technology. Genotyping and sequencing were completed at the Wellcome Trust Sanger Institute. Informed consent was obtained from the studied subjects, and the use of the samples in genetic studies was approved by the Human Materials and Data Management Committee at the Wellcome Trust Sanger Institute (approval numbers 09/056 and 14/072) and by the institutional review board (number SMPZ121307-02) of the Lebanese American University.

The genotyping data were merged with data from the African Genome Variation Project,[10] the 1000 Genomes Project,[11] and Pagani et al.,[12] resulting in a combined dataset of ~1.1 million SNPs in 2,453 samples. Analyses including ancient genomes involved merging the panel described above with the Haak et al. dataset,[13] resulting in ~90,000 SNPs in common. Comparative whole-genome sequences were obtained from Pagani et al.[12] and Complete Genomics.[14]

Genotype data were processed with PLINK:[15] the SNP genotype success rate required was set to 99%, whereas SNPs with a minor allele frequency < 0.001 or Hardy-Weinberg p value < 0.000001 were removed. Genotypes from sequence data were called with SAMtools v.1.2[16] and BCFtools v.1.2 with the command "samtools mpileup -q 20 -Q 20 -C 50 | bcftools call -c -V indels." Concordance with array genotypes had a rate of 0.999. Phasing was carried out with SHAPEIT[17] with 1000 Genomes Project phase 3 haplotypes[18] as a reference panel.

Population Structure and Gene Flow

Principal components were computed with EIGENSOFT v.4.2.[19] Effective population size and rates of gene flow were inferred by the multiple sequentially Markovian coalescent (MSMC) approach[20] with four high-coverage phased genomes from each population. We assumed a generation time of 30 years and a mutation rate of 1.25×10^{-8} mutations per nucleotide per generation. Admixture masks to identify African and Eurasian segments within mixed high-coverage genomes were generated with PCAdmix[21] including two ancestral populations based on the 1000 Genomes Project YRI (Yoruba in Ibadan, Nigeria) and CEU (Utah residents with northern and western European ancestry from the CEPH collection) populations. 1 cM windows with a posterior probability of >0.9 for the most likely ancestral state were collected and used for creating African and Eurasian masks.

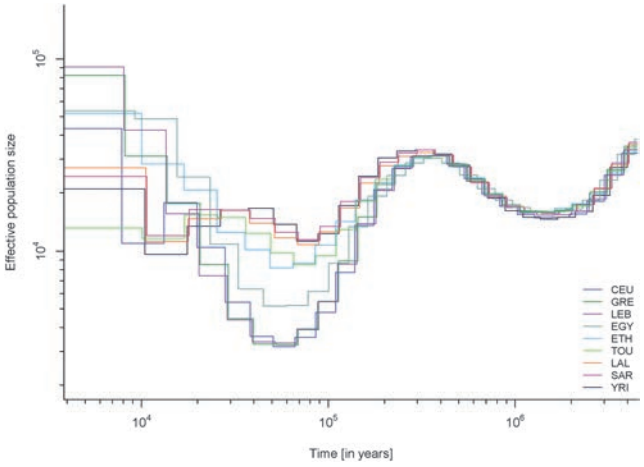


Figure 2. Population-Size Estimates from Whole-Genome Sequences
Population size was inferred by MSMC analysis with four haplotypes from each population. Eurasian populations had a distinctive bottleneck at the time of their exodus from Africa ~60,000 ya. Compared to other Africans, admixed Africans (from a Eurasian gene flow), such as Egyptians, Ethiopians, and the Toubou, also showed a decline in population size during the same period.

Phylogenetic analysis of whole Y chromosome sequences was carried out as described in Bergström et al.[22] Internal node ages were estimated with the rho-statistic[23] and converted to units of years by application of a Y chromosome mutation rate of 0.76×10^{-9} (95% confidence interval [CI] = 0.67×10^{-9} to 0.86×10^{-9}) mutations per site per year.[24] Additionally, Y chromosome haplogroups were defined to the highest resolution possible with 636 SNPs from the array data that overlapped International Society of Genetic Genealogy (ISOGG) markers.

Admixture Analysis

Population-mixture signals and proportions were tested with qp3Pop, qpDstat, and qpF4Ratio from the ADMIXTOOLS package.[25] Admixture proportions were additionally estimated with ADMIXTURE v.1.3.0.[26] ALDER[27] and MALDER[28] were used to date the time of admixture with all pairs of African-Eurasian populations as references. Significant results with a p value < 0.01 were collected and plotted.

Measure of Heterozygosity and Simulations

Heterozygosity on a per-individual basis was estimated with VCFtools v.0.1.13[29] for ~2.17 Gb of the uniquely mappable genome.[11] Heterozygosity was also estimated after correction for recent inbreeding via the removal of long runs of homozygosity (>2 Mb). We investigated the effect of gene flow on the observed heterozygosity by using individual-based forward-time simulations implemented in SimuPOP v.1.1.7.[30]

Selection after Admixture

Evidence for positive selection was tested with the population branch statistic (PBS)[31] with correction for the long-term effective population size.[32] We constructed a tree with the Toubou population branching from the Laal speakers and the Chinese Han as an outgroup. We collected values above the 95th percentile of the PBS distribution and looked for variants previously reported under putative selection in Europeans.

Results

Genetic Structure in Chad Indicates a Complex Admixture History

We performed an initial exploration of our dataset by using principal-component analysis (PCA).[19] The first component (PC1) captured the genetic differentiation between Africans and Eurasians (Figure 1B). Populations such as the Near Easterners and North and East Africans fell between the Europeans and sub-Saharan Africans. The Chadian groups lay near the sub-Saharan Africans: the Sara and Laal speakers clustered tightly with sub-Saharan Africans, such as the Yoruba, whereas the Toubou were somewhat more distant and appeared drawn toward East Africans, such as the Ethiopians. Samples collected from the capital of Chad, N'Djamena, appeared in a central position between the Toubou cluster and the Sara and Laal cluster (Figure 1C). Many individuals from N'Djamena have not reported their ethnicity or have reported a mixed ethnic origin. Therefore, recent mixture could be responsible for their position on the PCA.

We further investigated the genetic variation in Chad by estimating changes in the effective populations size (N_e) over time via the MSMC approach.[20] Eurasians and Africans diverged around 60,000–80,000 ya and subsequently had different patterns of population-size changes: in particular, compared with Africans, the Eurasian population experienced a sharp decrease in size ~60,000 ya.[20] We observed this expected pattern in most populations in our dataset (Figure 2), but a few stood out: (1) Egyptians had a population bottleneck that was much more pronounced than that of other Africans but not as sharp as that of Eurasian populations; and (2) the Toubou and Ethiopians shared a very similar pattern during the bottleneck: they were close to other Africans but had a somewhat sharper decrease in population size (Figure 2). We would not expect such different fluctuations in population sizes at 60,000 ya in populations who shared a common origin during this period. For example, all Eurasians trace their origin to a population who exited Africa ~60,000 ya, and this is reflected in indistinguishable N_e patterns during this period,[20,33] which we also observed in the CEU, Greeks, and Lebanese (Figure 2), as expected. A shared pattern of N_e in ancient times was also observed in the Sara, Laal speakers, and other Africans, such as the Yoruba. We suggest that the deviation from the expected N_e pattern in the Toubou is related to extensive admixture history with Eurasians, like the Eurasian admixture seen in Ethiopians, and we explore this possibility directly with admixture tests below.

Multiple Eurasian Admixtures in Africa after 6,000 ya

We have previously reported massive gene flow ~3,000 ya from Eurasians to Ethiopian populations.[4] Here, we reassess the presence of Eurasian ancestry in Africa by using f_3 statistics[25] in the form of f_3(X; Eurasian, Yoruba), where

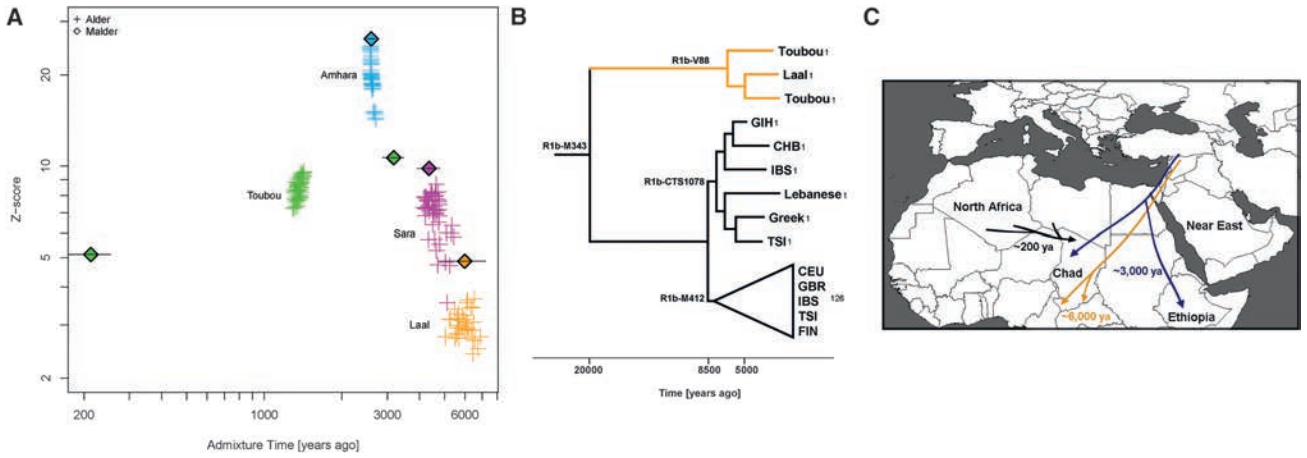


Figure 3. Timing of the Eurasian Admixture in Africa
(A) Crosses represent significant admixture events in the history of the Toubou, Amhara, Sara, and Laal speakers. Time of admixture is estimated from LD by ALDER with all pairs of African-Eurasian populations in our dataset as references. MALDER extends ALDER inference by detecting multiple mixture events, such as in the case of the Toubou population (shown here in green lozenges).
(B) A maximum-likelihood tree shows the males belonging to haplogroup R1b in the 1000 Genomes Project and the R1b males in our dataset. The number of samples is shown on each branch tip. We estimate that the Chadian R1b emerged 5,700–7,300 ya, whereas most European R1b haplogroups emerged 7,300–9,400 ya. The African and Eurasian lineages coalesced 17,900–23,000 ya.
(C) Putative sources and times of admixture of the Eurasian ancestry in Chad and East Africa.

a negative value with a Z score < -4 indicates that X is a mixture of Africans and Eurasians. We found, as expected, that most Ethiopians are a mixture of Africans and Eurasians. An exception is the Gumuz population, where f_3(Gumuz; Eurasian, Yoruba) is always positive. The Gumuz language belongs to the Nilo-Saharan family, which could have isolated the Gumuz from the Afro-Asiatic-speaking Ethiopians. However, we found that the Toubou in Chad, who also speak a Nilo-Saharan language, are a mixture of Africans and Eurasians, making f_3(Toubou; Eurasian, Yoruba) always significantly negative. This suggests that the impact of Eurasian migrations today extends beyond East Africa and the Afro-Asiatic-speaking populations. We did not detect significant (Z score < -4) Eurasian admixture in the Sara (Nilo-Saharan language family) or the Laal speakers (unclassified language) with the use of f_3 statistics (lowest Z score for the Sara was >-2.9; for the Laal speakers, Z scores were all positive). However, this statistic loses sensitivity with small mixture proportions and post-admixture drift,[27] so positive values from the f_3 statistics do not necessarily reflect a complete absence of admixture. We thus further tested for admixture by using ALDER and MALDER, which assess admixture-induced linkage disequilibrium (LD) and can detect small mixture proportions from a substantially diverged reference possibly missed by the f_3 statistic. ALDER detected admixture in the Toubou, Sara, and Laal speakers (Table S2). MALDER, which has the potential to determine whether or not the admixture LD in the population is best represented as the result of one or multiple mixtures, showed that two mixture events had occurred in the Toubou (Figure 3A; Table S3). The first event occurred 2,850–3,500 ya (Z score = 11), a time close to the date of mixture in East Africans 2,500–2,700 ya (Z score = 26). The second mixture event occurred much more recently at 170–260 ya (Z score = 5). In southern Chad, we detected mixture events that were more ancient than those in the north. Mixture occurred 3,900–4,800 ya (Z score = 10) in the Sara and 4,750–7,200 ya (Z score = 5) in the Laal speakers (Figure 3A). These time estimates overlap, and we interpret them as signals from the same admixture event, whose time in the distant past was estimated more reliably in the Laal speakers because they carry more Eurasian ancestry (1.25%–4.5%) than the Sara (0.3%–2%) (see estimates of admixture proportions below), even though the Sara have smaller standard errors because of their larger sample size. In particular, we suggest that the Eurasian mixture event in the Sara and Laal speakers is independent of the mixture event in East Africans and the Toubou for two reasons: (1) admixture LD showed that the events in southern Chad preceded the events in East Africa by 2,000–4,500 years, and (2) we found in Chad a Eurasian Y chromosome lineage (Y haplogroup R1b-V88) that had penetrated all Chadian populations examined but was absent or rare from the Ethiopians examined (Table S4; Figure S1). From whole Y chromosome sequences (Figure S2), we estimate that the Chadian R1b-V88 chromosomes sampled emerged 5,700–7,300 ya (Figure 3B), a time comparable to the Laal speaker admixture dates (4,750–7,200 ya) estimated from genome-wide LD-decay patterns.

The Sources of Eurasian Backflow into Chad and East Africa Are Correlated

Previous studies have suggested that the Eurasian backflow into East Africa came from a population related to early Neolithic farmers.[5] We wanted to know whether the Eurasian ancestry we found in the Toubou, which we

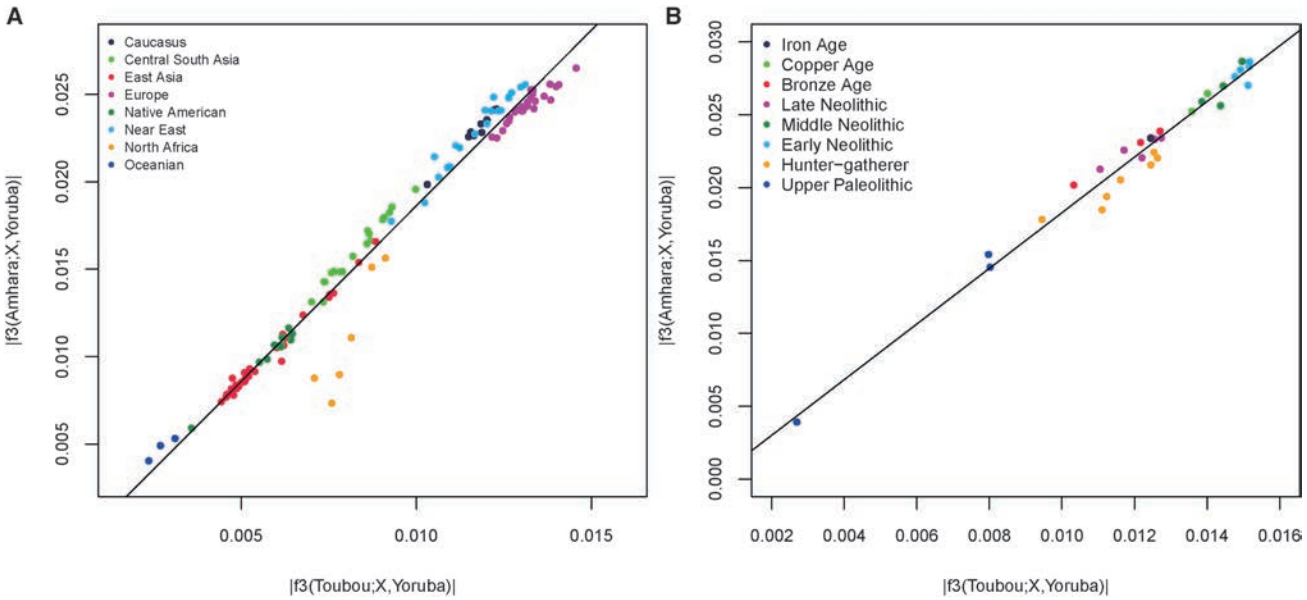


Figure 4. Sources of the Eurasian Ancestry in Chad and Ethiopia
The plot shows significant Eurasian sources for the Toubou and Amhara according to a three-population test (Z score < -4). An increase in the absolute value of the f_3 statistic implies an increase in the genetic affinity of the Eurasian population X to the Toubou and Amhara. (A) With the exception of North Africans, who showed increased affinity to the Toubou, present-day populations showed correlated affinity to both the Toubou and Amhara. Among modern populations, Sardinians showed the highest genetic affinity to both the Toubou and Amhara.
(B) Ancient Eurasians also showed correlated affinity to both the Toubou and Amhara; the early Neolithic LBK (Linearbandkeramik, or Linear Pottery) population (~5,000 BCE) had the highest affinity.

attribute to a mixture close in time to the date of mixture in East Africans, can be traced to the same source populations that influenced Ethiopia. We performed the tests f_3(Toubou; Yoruba, X) and f_3(Amhara; Yoruba, X), where X is a present-day non-sub-Saharan African population in our dataset and is related to one that contributed ancestry to the Toubou and Amhara (Z score < -4) (Table S5). We then looked at the correlation of the f_3 statistic values between the two tests (Figure 4A). We found that the Eurasian source populations for the Amhara and Toubou were highly correlated ($r = 0.98$; 95% CI = 0.98–0.99; p value $< 2.2 \times 10^{-16}$) and that the most significant result was for present-day Sardinians. Exceptions to this correlation were the North African populations (Tunisians, Mozabite, Algerians, and Saharawi), who appeared to have contributed more ancestry to the Toubou than to the Amhara. We repeated the tests by using published ancient genomes (Table S6) and also found a high correlation of the Eurasian sources for the Amhara and Toubou ($r = 0.98$; 95% CI = 0.97–0.99; p value $< 2.2 \times 10^{-16}$); early Neolithic farmers were the most significant contributors, as reported previously[5] (Figure 4B). When we substituted the Amhara with other Ethiopians (Wolayta and Oromo), we found similar results (data not shown). In a parallel comparison, we checked whether the sources of the African ancestry in different Near Eastern populations were also correlated. We tested f_3(Lebanese; British, X) and f_3(Yemeni; British, X) and found a lower correlation of the f_3 values ($r = 0.62$; 95% CI = 0.32–0.80), suggesting a more complicated history of gene flow from genetically different Africans to different populations in the Near East.

We next quantified the proportion of African-Eurasian mixture in the study populations by using two methods: (1) ADMIXTURE[26] supervised with $K = 2$ and the British and Yoruba as ancestral populations and (2) the f_4 ratio $\alpha = f_4$(British, chimp; X, Yoruba)/f_4(British, chimp; early Neolithic farmer, Yoruba), where X is one of the populations in our dataset (Figure S3). The results from the two tests were highly correlated ($r = 0.998$; 95% CI = 0.996–0.999; p value $< 2.2 \times 10^{-16}$). Eurasian ancestry was estimated at 26%–30% in the Toubou, 0.3%–2% in the Sara, and 1.2%–4.5% in the Laal speakers. Eurasian ancestry in Ethiopians ranged from 11%–12% in the Gumuz to 53%–57% in the Amhara. African ancestry in the Near East ranged from 7%–14% (Yemen) to 0.7%–5% (Lebanese Christians).

Eurasian Gene Flow Shaped the Genomes of Admixed Africans

Our results from the PCA and MSMC analysis showed a deviation of the admixed populations from the patterns observed in unadmixed (or less admixed) populations in the same geographical region. The MSMC analysis, in particular, showed that admixed Africans had patterns indicative of a decline in heterozygosity (increased bottleneck ~60,000 ya), somewhat similarly to Eurasians. We tested whole-genome heterozygosity in these populations and found that it decreased in admixed Africans according

to their Eurasian ancestry (Figure S4A). This decrease was not related to recent inbreeding, given that removing segments with long runs of homozygosity did not change the overall pattern. Our simulations suggest that decay in heterozygosity is expected after gene flow from a population with diversity comparable to that of Eurasians (Figures S4B and S4C). We further investigated heterozygosity in admixed Africans by assessing heterozygosity of the different ancestral segments in the Toubou genome. We found that admixed African-Eurasian segments had more heterozygosity (1.23 hets/kb) than segments of the genome where African-African haplotypes were present (1.19 hets/kb) (Figure S5). However, the Toubou genome segments with complete Eurasian ancestry (Eurasian-Eurasian) had considerably lower heterozygosity (~0.96 hets/kb; Figure S5), leading to the genome-wide pattern of decay in heterozygosity observed in Africans with Eurasian ancestry (Figure S4).

We wanted to understand the consequence of admixture on the models that use the density of heterozygous sites to infer the demographic history of populations. We first tested whether the coalescent history estimated by MSMC was affected by a small proportion of mixture, such as the African mixture found in Greeks and Lebanese (ranging from 0% to 5%). We tested the Greek, Lebanese, CEU, and CHB (Han Chinese in Beijing, China) split times from the Yoruba and found that all populations split from the Yoruba ~70,000–80,000 ya, implying that the low proportions of African admixture in the Greeks and Lebanese did not detectably affect the estimates of relative cross-coalescence rate (Figure S6A). We next tested the Toubou, who have ~30% Eurasian ancestry. The Toubou appeared to split from Eurasians ~30,000–40,000 ya, a time more recent than expected considering the African-Eurasian split 60,000–80,000 ya[20] (Figure S6B). We tested other Africans in our dataset and found that the Sara, Laal speakers, and Yoruba split from Eurasians, as expected, ~70,000–80,000 ya (Figure S6B). We then tested directly whether the Eurasian ancestry affected the relative cross-coalescence rate between the Toubou and Eurasians by masking some of the Eurasian ancestry in the Toubou. We used PCAdmix[21] to estimate the ancestry along each chromosome and then used the identified Eurasian segments as a negative mask in our analysis. The split times between the Toubou and Lebanese, for example, increased by ~15,000 years (Figure S6B), shifting the split date toward the expected African-Eurasian split time.

We found that, in addition to influencing the relative cross-coalescence rate, admixture can also inflate putative signals of positive selection. For example, using the PBS[31] to detect recent positive selection that occurred in the Toubou after their divergence from the Yoruba, we found signals of selection on *MCM6* (MIM: 601806) rs4988235, a variant associated with the lactase-persistence phenotype. This SNP was previously found to be under strong positive selection in Europeans, where it was probably advantageous to individuals living in pastoralist societies.[34] The frequency of this variant in the Toubou is 2%, and it is absent from the sub-Saharan African and other Chadic samples (the Sara and Laal speakers) examined here. Although this SNP appears to be a candidate for selection, we suggest that it has probably drifted neutrally in the Toubou after the Eurasian gene flow: the Toubou have ~30% Eurasian ancestry from a population similar to the Greeks, who have 13% derived alleles at rs4988235, suggesting an expectation of ~3.9% of the derived allele simply from admixture. We similarly found in the Toubou signals at *HERC2* (MIM: 605837) rs1129038 a major contributor to blue eye color in Europeans[35] (Toubou derived allele frequency [DAF] = 0.014; Greek DAF = 0.33; Yoruba, Sara, and Laal DAF = 0), as well as a signal at *SLC24A5* (MIM: 609802) rs1834640, a major contributor to pigmentation[36] (Toubou DAF = 0.19; Greek DAF = 0.99; Yoruba, Sara, and Laal DAF = 0–0.04).

In addition to introducing to African populations genes that were positively selected in Europe, the recent African-Eurasian admixture carried Neanderthal alleles to Central and East Africa. Neanderthals are closer to the Amhara than to the Yoruba: D(Neanderthal, chimp, Amhara, Yoruba) = 0.0094. Neanderthals are also closer to the Toubou than to the Yoruba: D(Neanderthal, chimp, Toubou, Yoruba) = 0.0041. On the other hand, we found that Neanderthals are closer to Europeans than to Near Easterners: D(Neanderthal, chimp; French, Yemen) = 0.0056 and D(Neanderthal, chimp; French, Palestinian) = 0.0040.

We estimated the archaic ancestry proportions by using the ratio

$$\alpha = \frac{f4(\text{Altai, chimp; } X, \text{Yoruba})}{f4(\text{Altai, chimp; Vindija, Yoruba})}$$

and found that Neanderthal ancestry was ~0.5% in the Toubou and~1% in the Amhara. We then computed the correlation between the Neanderthal ancestry proportions and the Eurasian and African ancestry proportions we identified. Neanderthal ancestry in admixed Africans and Near Easterners was highly correlated with their Eurasian ancestry ($r = 0.98$; p value $< 2.2 \times 10^{-16}$) and inversely correlated with their African ancestry (Figure 5).

Discussion

We have generated an extensive set of genotyping and high-coverage whole-genome sequencing data to study the genetic history of Chad and neighboring populations. We found substantial genetic differences between the ethnic groups inhabiting Chad today and suggest that multiple ancient Eurasian migrations played a major role in shaping the genetic diversity of the region (Figure 3C). Here, we discuss these migrations and how the mixed ancestry can confound proper interpretation of the evolutionary processes that occurred in their history and therefore needs to be thoroughly accounted for in the study of African genetic diversity.

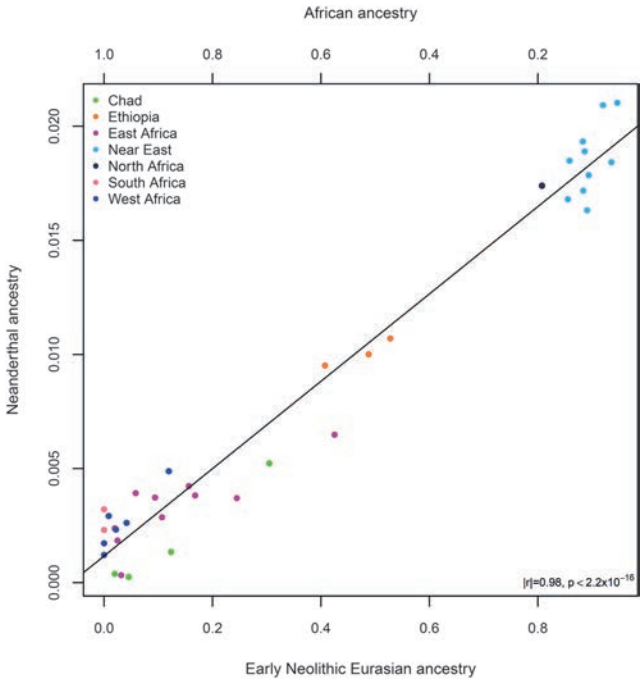


Figure 5. Neanderthal Ancestry Correlation with the African-Eurasian Admixture

Neanderthal ancestry is not expected in Africa, yet today many Africans carry Neanderthal-derived alleles. The plot shows that the Neanderthal ancestry proportion in Africans is correlated with gene flow from Eurasians. For example, knowing that today Eurasians carry ~2% of Neanderthal ancestry, we observed that East Africans (Ethiopians) had ~1% Neanderthal ancestry and ~50% Eurasian ancestry. Correspondingly, Near Easterners showed a decline in Neanderthal ancestry proportional to their levels of African ancestry.

We detected the earliest Eurasian migrations to Africa in the Laal-speaking people, an isolated language group of fewer than 800 speakers who inhabit southern Chad. We estimate that mixture occurred 4,750–7,200 ya, thus after the Neolithic transition in the Near East, a period characterized by exponential growth in human population size. Environmental changes during this period (which possibly triggered the Neolithic transition) also facilitated human migrations. The African Humid Period, for example, was a humid phase across North Africa that peaked 6,000–9,000 ya[37] and biogeographically connected Africa to Eurasia, facilitating human movement across these regions.[38] In Chad, we found a Y chromosome lineage (R1b-V88) that we estimate emerged during the same period 5,700–7,300 ya (Figure 3B). The closest related Y chromosome groups today are widespread in Eurasia and have been previously associated with human expansions to Europe.[39,40] We estimate that the Eurasian R1b lineages initially diverged 7,300–9,400 ya, at the time of the Neolithic expansions. However, we found that the African and Eurasian R1b lineages diverged 17,900–23,000 ya, suggesting that genetic structure was already established between the groups who expanded to Europe and Africa. R1b-V88 was previously found in Central and West Africa and was associated with a mid-Holocene migration of Afro-asiatic speakers through the central Sahara into the Lake Chad Basin.[8] In the populations we examined, we found R1b in the Toubou and Sara, who speak Nilo-Saharan languages, and also in the Laal people, who speak an unclassified language. This suggests that R1b penetrated Africa independently of the Afro-asiatic language spread or passed to other groups through admixture.

In addition to the early Eurasian migration to Africa ~6,000 ya, a second migration ~3,000 ya affected the Toubou population in northern Chad but had no detectable genetic impact on other Chadian populations. This migration appears to be associated with the previously reported Eurasian backflow into East Africa, given that the source populations and dates of mixture are similar. Occurring at the start of the Iron Age, these migrations could have been facilitated by advances in warfare and transportation technology in the Near East. It is uncertain why the impact of this migration in Chad affected only the Toubou. The African ancestral component in the Toubou is best represented by the Laal-speaking population, suggesting that the African-Eurasian mixture probably occurred in Chad. However, ethnolinguistic barriers could have already been established at this time between the Chad groups, preventing a widespread dissemination of the Eurasian ancestry. The Toubou, despite their Islamic faith, do not show the genetic admixture detected in many Near Eastern and North African populations around 1,100 ya,[41] suggesting conversion without population mixing at this time. They did, however, receive additional Eurasian ancestry in the past 200 years from a source represented by North African populations such as Tunisians, Mozabite, Algerians, and Sahrawi (Figure 3C). This recent interaction could have been promoted by the nomadic lifestyle of the present-day Toubou and a shared Muslim religion with North Africans. Unsurprisingly, we also detected a likely mixing of Chad populations in the sample from the capital, which could be even more recent.

Eurasian backflow into Africa thus appears to have been a recurrent event in the history of many Africans, given its considerable impact on their genomes. Although population mixture in general is a process that increases genetic diversity, we observed a decrease in heterozygosity in the admixed Africans. Our simulations showed that these results are expected after mixture at these proportions with the Eurasians who suffered a significant bottleneck at the time of their exodus from Africa ~60,000 ya. Consequently, we found that mixture can complicate interpretation of the coalescent history inferred from models that use the density of heterozygous sites in their implementations. In addition, we detected in admixed Africans an inflation of positive-selection signals on alleles associated with adult lactose tolerance and pigmentation in Europeans, but we suggest that these alleles have drifted neutrally in Africans after admixture. Furthermore, we detected Neanderthal ancestry in admixed Africans and found it to be proportional to their Eurasian ancestry.

Similarly, in admixed Near Easterners, we found a decrease in Neanderthal ancestry proportional to the gene flow they have received from Africans. Although a higher genetic affinity of Neanderthals to Europeans than to Near Easterners was previously interpreted as additional Neanderthal admixture in the history of Europeans,[42] we propose that a more parsimonious explanation for these observations is that African-Eurasian mixtures both introduced Neanderthal ancestry to Africa and "diluted" the Neanderthal ancestry in the Near East.

It is important to note that in this work we inevitably invoke Occam's razor to support the simplest model consistent with our data; the history of the populations studied here, including the time and sources of the Eurasian admixture in Africa, could be more complex. aDNA from Chad and neighboring regions remains a challenge given the poor DNA preservation in hot climates, but future successful efforts in aDNA research could provide additional insights and reveal additional complexities not considered by the modern-DNA-based models favored here.[43]

Our study has shown that human genetic diversity in Africa is still incompletely understood and that ancient admixture adds to its complexity. This work highlights the importance of exploring underrepresented populations, such as those from Chad, in genetic studies to improve our understanding of the demographic processes that shaped genetic variation in Africa and globally.

Accession Numbers

Whole-genome sequencing and SNP genotyping data are available through the European Genome-phenome Archive (EGA) under accession numbers EGA: EGAD00001002742 (sequences from Chad and Lebanon), EGAD00001001440 (sequences from Greece), and EGAS00001001231 (SNP data from Chad, Lebanon, and Yemen).

Supplemental Data

Supplemental Data include six figures and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.10.012.

Acknowledgments

We thank all sample donors for making this work possible and the Wellcome Trust Sanger Institute pipelines for generating genotype and sequence data. We also thank Andrea Massaia for comments on analyzing Y chromosome haplogroups from array data and David Soria-Hernanz for comments on the ethnic groups and languages in Chad. M.H., M.M., A.B., J.P.-M., E.Z., Y.X., and C.T.-S. were supported by the Wellcome Trust (098051). P.H. was supported by Estonian Research Council grant PUT1036. J.B.-S. is a full-time employee of Karius, Inc., and R.S.W. is founder and CEO of Insitome.

Received: August 15, 2016
Accepted: October 24, 2016
Published: November 23, 2016

Web Resources

European Genome-phenome Archive (EGA), https://www.ebi.ac.uk/ega/home
ISOGG Y-DNA Haplotype Tree v.11.201, http://www.isogg.org/tree/

References

1. Central Intelligence Agency. (2014). The World Factbook 2014. https://www.cia.gov/library/publications/the-world-factbook/index.html.
2. Kröpelin, S., Verschuren, D., Lézine, A.M., Eggermont, H., Cocquyt, C., Francus, P., Cazet, J.P., Fagot, M., Rumes, B., Russell, J.M., et al. (2008). Climate-driven ecosystem succession in the Sahara: the past 6000 years. Science *320*, 765–768.
3. Löhr, D. (2009). Lake Chad and the migratory routes to Borno: A linguistic trail. In Migrations and Spatial Mobility in the Lake Chad Basin, XIIIth Mega-Chad Conference, H. Tourneaux, ed. (Editions de l'IRD), pp. 665–681.
4. Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Gallego Romero, I., Ayub, Q., Mehdi, S.Q., Thomas, M.G., Luiselli, D., et al. (2012). Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. Am. J. Hum. Genet. *91*, 83–96.
5. Gallego Llorente, M., Jones, E.R., Eriksson, A., Siska, V., Arthur, K.W., Arthur, J.W., Curtis, M.C., Stock, J.T., Coltorti, M., Pieruccini, P., et al. (2015). Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. Science *350*, 820–822.
6. Cerný, V., Fernandes, V., Costa, M.D., Hájek, M., Mulligan, C.J., and Pereira, L. (2009). Migration of Chadic speaking pastoralists within Africa based on population structure of Chad Basin and phylogeography of mitochondrial L3f haplogroup. BMC Evol. Biol. *9*, 63.
7. Cerezo, M., Černý, V., Carracedo, Á., and Salas, A. (2011). New insights into the Lake Chad Basin population structure revealed by high-throughput genotyping of mitochondrial DNA coding SNPs. PLoS ONE *6*, e18682.
8. Cruciani, F., Trombetta, B., Sellitto, D., Massaia, A., Destro-Bisol, G., Watson, E., Beraud Colomb, E., Dugoujon, J.M., Moral, P., and Scozzari, R. (2010). Human Y chromosome haplogroup R-V88: a paternal genetic record of early mid Holocene trans-Saharan connections and the spread of Chadic languages. Eur. J. Hum. Genet. *18*, 800–807.
9. Triska, P., Soares, P., Patin, E., Fernandes, V., Cerny, V., and Pereira, L. (2015). Extensive Admixture and Selective Pressure Across the Sahel Belt. Genome Biol. Evol. *7*, 3484–3495.
10. Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A., et al. (2015). The African Genome Variation Project shapes medical genetics in Africa. Nature *517*, 327–332.
11. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A.; and 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
12. Pagani, L., Schiffels, S., Gurdasani, D., Danecek, P., Scally, A., Chen, Y., Xue, Y., Haber, M., Ekong, R., Oljira, T., et al. (2015). Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. Am. J. Hum. Genet. *96*, 986–991.

13. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. Nature *522*, 207–211.
14. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science *327*, 78–81.
15. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
16. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.; and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
17. Delaneau, O., Zagury, J.F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods *10*, 5–6.
18. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
19. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.
20. Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. Nat. Genet. *46*, 919–925.
21. Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J.G., and Bustamante, C.D. (2012). PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Hum. Biol. *84*, 343–364.
22. Bergström, A., Nagle, N., Chen, Y., McCarthy, S., Pollard, M.O., Ayub, Q., Wilcox, S., Wilcox, L., van Oorschot, R.A., McAllister, P., et al. (2016). Deep roots for Aboriginal Australian Y chromosomes. Curr. Biol. *26*, 809–813.
23. Forster, P., Harding, R., Torroni, A., and Bandelt, H.J. (1996). Origin and evolution of Native American mtDNA variation: a reappraisal. Am. J. Hum. Genet. *59*, 935–945.
24. Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., Bondarev, A.A., Johnson, P.L., Aximu-Petri, A., Prüfer, K., de Filippo, C., et al. (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. Nature *514*, 445–449.
25. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. Genetics *192*, 1065–1093.
26. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.
27. Loh, P.R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K., Reich, D., and Berger, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. Genetics *193*, 1233–1254.
28. Pickrell, J.K., Patterson, N., Loh, P.R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., and Reich, D. (2014). Ancient west Eurasian ancestry in southern and eastern Africa. Proc. Natl. Acad. Sci. USA *111*, 2632–2637.
29. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.
30. Peng, B., and Kimmel, M. (2005). simuPOP: a forward-time population genetics simulation environment. Bioinformatics *21*, 3686–3687.
31. Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science *329*, 75–78.
32. Ayub, Q., Mezzavilla, M., Pagani, L., Haber, M., Mohyuddin, A., Khaliq, S., Mehdi, S.Q., and Tyler-Smith, C. (2015). The Kalash genetic isolate: ancient divergence, drift, and selection. Am. J. Hum. Genet. *96*, 775–783.
33. Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496.
34. Enattah, N.S., Sahi, T., Savilahti, E., Terwilliger, J.D., Peltonen, L., and Järvelä, I. (2002). Identification of a variant associated with adult-type hypolactasia. Nat. Genet. *30*, 233–237.
35. Sturm, R.A., Duffy, D.L., Zhao, Z.Z., Leite, F.P., Stark, M.S., Hayward, N.K., Martin, N.G., and Montgomery, G.W. (2008). A single SNP in an evolutionary conserved region within intron 86 of the *HERC2* gene determines human blue-brown eye color. Am. J. Hum. Genet. *82*, 424–431.
36. Stokowski, R.P., Pant, P.V., Dadd, T., Fereday, A., Hinds, D.A., Jarman, C., Filsell, W., Ginger, R.S., Green, M.R., van der Ouderaa, F.J., and Cox, D.R. (2007). A genomewide association study of skin pigmentation in a South Asian population. Am. J. Hum. Genet. *81*, 1119–1132.
37. deMenocal, P.B., and Tierney, J.E. (2012). Green Sahara: African Humid Periods paced by Earth's orbital changes. Nature Education Knowledge *3*, 12.
38. Larrasoaña, J.C., Roberts, A.P., and Rohling, E.J. (2013). Dynamics of green Sahara periods and their role in hominin evolution. PLoS ONE *8*, e76514.
39. Balaresque, P., Bowden, G.R., Adams, S.M., Leung, H.Y., King, T.E., Rosser, Z.H., Goodwin, J., Moisan, J.P., Richard, C., Millward, A., et al. (2010). A predominantly neolithic origin for European paternal lineages. PLoS Biol. *8*, e1000285.
40. Batini, C., Hallast, P., Zadik, D., Delser, P.M., Benazzo, A., Ghirotto, S., Arroyo-Pardo, E., Cavalleri, G.L., de Knijff, P., Dupuy, B.M., et al. (2015). Large-scale recent expansion of European patrilineages shown by population resequencing. Nat. Commun. *6*, 7152.
41. Haber, M., Gauguier, D., Youhanna, S., Patterson, N., Moorjani, P., Botigué, L.R., Platt, D.E., Matisoo-Smith, E., Soria-Hernanz, D.F., Wells, R.S., et al. (2013). Genome-wide diversity in the levant reveals recent structuring by culture. PLoS Genet. *9*, e1003316.
42. Rodriguez-Flores, J.L., Fakhro, K., Agosto-Perez, F., Ramstetter, M.D., Arbiza, L., Vincent, T.L., Robay, A., Malek, J.A., Suhre, K., Chouchane, L., et al. (2016). Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations. Genome Res. *26*, 151–162.
43. Haber, M., Mezzavilla, M., Xue, Y., and Tyler-Smith, C. (2016). Ancient DNA and the rewriting of human history: be sparing with Occam's razor. Genome Biol. *17*, 1.

Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits

Nicholas Mancuso,[1,*] Huwenbo Shi,[2] Pagé Goddard,[3] Gleb Kichaev,[2] Alexander Gusev,[4,5,6,8] and Bogdan Pasaniuc[1,2,7,8,*]

Although genome-wide association studies (GWASs) have identified thousands of risk loci for many complex traits and diseases, the causal variants and genes at these loci remain largely unknown. Here, we introduce a method for estimating the local genetic correlation between gene expression and a complex trait and utilize it to estimate the genetic correlation due to predicted expression between pairs of traits. We integrated gene expression measurements from 45 expression panels with summary GWAS data to perform 30 multi-tissue transcriptome-wide association studies (TWASs). We identified 1,196 genes whose expression is associated with these traits; of these, 168 reside more than 0.5 Mb away from any previously reported GWAS significant variant. We then used our approach to find 43 pairs of traits with significant genetic correlation at the level of predicted expression; of these, eight were not found through genetic correlation at the SNP level. Finally, we used bi-directional regression to find evidence that BMI causally influences triglyceride levels and that triglyceride levels causally influence low-density lipoprotein. Together, our results provide insight into the role of gene expression in the susceptibility of complex traits and diseases.

Introduction

Although genome-wide association studies (GWASs) have identified tens of thousands of common genetic variants associated with many complex traits,[1] with some notable exceptions,[2,3] the causal variants and genes at these loci remain unknown. Multiple lines of evidence have shown that GWAS risk variants co-localize with genetic variants that regulate expression—i.e., expression quantitative trait loci (eQTLs).[4] This suggests that a substantial proportion of GWAS risk variants influence complex traits by regulating expression levels of their target genes.[4–7] Analyses of genotype, phenotype, and gene expression measurements from multiple tissues in the same set of individuals can directly investigate this plausible chain of causality. However, doing so is challenging because of cost and tissue availability; therefore, GWAS and eQTL datasets remain largely independent (i.e., no overlapping subjects).[8,9] Recent work has shown that one way to integrate GWAS and eQTL data is to predict gene expression levels for GWAS samples and then test for association between the predicted expression and traits.[10–12] This approach, referred to as transcriptome-wide association study (TWAS), can increase power over GWAS when the causal mechanism includes genetic variants that regulate the expression of susceptibility genes. TWAS benefits from a lower multiple-testing burden by probing several thousands of genes, whereas GWAS probes several million SNPs. Although TWAS can also be performed with measured gene expression levels directly, using predicted gene expression has several benefits. First, expression measurements are usually not available in GWAS data. Second, predicted gene expression removes environmental noise by focusing on the genetically regulated component, which can increase statistical power. Third, using the predicted expression to test for association can eliminate potential confounding from reverse causation, where traits affect gene expression levels.[10,11] However, compared with GWAS, TWAS is underpowered when risk is not mediated through expression or when expression data are not available in the right tissue.

In this work, we introduce methods for estimating the genetic correlation between gene expression and a complex trait from summary GWAS and eQTL data. We utilize the local (*cis*) genetic variation near a gene (i.e., ± 0.5 Mb around the transcription start site [TSS]) to estimate the correlation in the genetic effects between gene expression and the trait. We show that under this framework, TWAS can be viewed as a test for non-zero genetic covariance between expression and a trait from summary association data. In addition to identifying susceptibility genes, the predicted expression can also be used for estimating the genome-wide genetic correlation between pairs of complex traits at the level of predicted expression. This is analogous to computing genome-wide genetic correlation between complex traits,[13] whereby correlations are determined over predicted gene expression effects rather than SNP effects, and

[1]Department of Pathology & Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90024, USA; [2]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA 90024, USA; [3]Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, Los Angeles, CA 90024, USA; [4]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [6]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [7]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90024, USA
[8]These authors contributed equally to this work
*Correspondence: nmancuso@mednet.ucla.edu (N.M.), bpasaniuc@mednet.ucla.edu (B.P.)
http://dx.doi.org/10.1016/j.ajhg.2017.01.031.

can give insights into the component of genetic correlation mediated through expression. We demonstrate through extensive simulations that our approach is approximately unbiased and well calibrated under the null and slightly conservative when true correlation is near the boundaries. Finally, we utilize estimated effects of predicted expression within a bi-directional regression approach[14] to investigate putative causal direction for pairs of complex traits that are genetically correlated.

We analyze summary statistics from 30 GWASs spanning 2.3 million phenotype measurements[15–28] jointly with 45 expression panels[8,29–34] sampled from more than 35 tissues to gain insight into the role of expression in the etiology of complex traits. First, we test each gene-tissue pair across 45 panels to perform a multi-tissue TWAS for each of the 30 traits to identify 1,196 gene associations. For example, at four independent loci, we find 11 genes that do not overlap a genome-wide significant SNP for educational years. Notably, all four loci were replicated in a recent, larger GWAS for educational years.[35] Second, we identify 43 pairs of traits showing a genome-wide-significant genetic correlation at the level of predicted expression. Overall, the predicted-expression correlation was highly concordant with SNP-level genetic correlation from cross-trait linkage disequilibrium (LD) score regression, which suggests that a large component of genetic correlation between complex traits is driven by local regulation of gene expression. Finally, we use our bi-directional analysis to provide evidence of putative causal effects between pairs of these traits. Overall, our results shed light on shared biological mechanisms responsible for susceptibility to disease and complex traits, as well as potential downstream effects between traits.

Material and Methods

Datasets

We used summary association statistics from 30 large-scale (n = 20,000 subjects) GWASs, including various anthropometric[15,27,28] (body mass index [BMI], femoral neck bone mineral density [BMD], forearm BMD, lumbar spine BMD, and height), hematopoietic[23,25,26] (hemoglobin, HbA_{1c}, mean cell hemoglobin [MCH], MCH concentration, mean cell volume, number of platelets, packed cell volume, and red blood cell count), immune-related[17,19] (Crohn disease [OMIM: 266600], inflammatory bowel disease [OMIM: 266600], ulcerative colitis [OMIM: 266600], and rheumatoid arthritis [OMIM: 180300]), metabolic[16,20,22,24] (age of menarche, fasting glucose, fasting insulin, high-density lipoprotein [HDL], HOMA-B, HOMA-IR, low-density lipoprotein [LDL], triglycerides [TG], type 2 diabetes [OMIM: 125853], and total cholesterol [TC] levels), neurological[18] (schizophrenia [OMIM: 181500]), and social[21] (college and educational attainment) phenotypes (see Table S1). We removed SNPs that were strand ambiguous or had a minor allele frequency (MAF) $\leq$ 1% (see Table S1).

Gene expression data from RNA sequencing data were obtained from the CommonMind Consortium[29] (brain, n = 613), the Genotype-Tissue Expression Project[8] (GTEx; 41 tissues; see Table S2 for sample size per tissue), and the Metabolic Syndrome in Men study[31,32] (adipose, n = 563). Expression microarray data were obtained from the Netherlands Twins Registry[34] (NTR; blood, n = 1,247), and the Young Finns Study[30,33] (YFS; blood, n = 1,264).

Performing TWAS with GWAS Summary Statistics

We estimated SNP heritability for observed expression levels partitioned into *cis*-h_g^2 (1 Mb region surrounding the TSS) and *trans*-h_g^2 (rest of genome) components. We used the AI-REML algorithm implemented in Genome-wide Complex Trait Analysis (GCTA),[36] which allows estimates to fall outside of the (0, 1) boundaries to maintain unbiasedness. To control for confounding, we included batch variables and the top 20 principal components estimated from genome-wide SNPs. Genes with significant *cis*-heritability in expression data were used for prediction (*cis*-h_g^2 $p < 0.05$ in a likelihood ratio test between the *cis*-only and joint models). The average number of genes with significant *cis*-h_g^2 across expression studies was 816 (min = 70 genes from GTEx small intestine samples; max = 3,704 genes from the YFS).

We performed 45 TWASs for each of the 30 GWASs;[11] for each trait, we used Bonferroni correction for all gene-tissue pairs tested (see Table S2). In brief, we estimated the strength of association between the predicted expression of a gene and a complex trait (z_{TWAS}) as a function of the vector of GWAS summary *Z* scores at a given *cis*-locus, $\boldsymbol{z}_T^{'}$ (i.e., vector of SNP association Wald statistics), and the LD-adjusted weight vector learned from the gene expression data, $\boldsymbol{w}_{GE}$, as

$$z_{TWAS} = \frac{\boldsymbol{w}_{GE}^{'}\boldsymbol{z}_T}{\sqrt{\mathrm{var}(\boldsymbol{w}_{GE}^{'}\boldsymbol{z}_T)}} = \frac{\boldsymbol{w}_{GE}^{'}\boldsymbol{z}_T}{\sqrt{\boldsymbol{w}_{GE}^{'}\boldsymbol{V}\boldsymbol{w}_{GE}}},$$

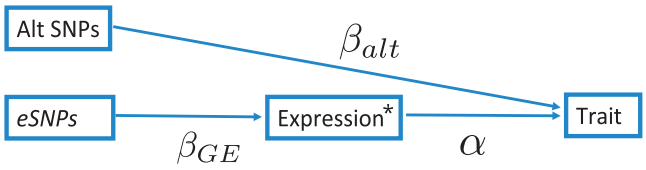
where $\boldsymbol{V}$ is a covariance matrix across SNPs at the locus (i.e., LD). We estimated $\boldsymbol{w}_{GE}$ by using GBLUP[37] from eQTL data and computed z_{TWAS} by using GWAS summary data for all 30 traits and the ~36,000 gene expression measurements across all studies. We removed all loci in the human leukocyte antigen (HLA) region as a result of complex LD patterns.

Estimating the Proportion of Trait Variance Explained by Predicted Expression

We use the LD score regression[38,39] approach described in Guseve et al.[11] to quantify the heritability explained by predicted expression for a complex trait (denoted here as h_{GE}^2). The expected χ^2 statistic under a polygenic trait is $E[\chi^2] = 1 + (N_T\ell/M)h_{GE}^2 + N_T a$, where N_T is the number of individuals in the GWAS, *M* is the number of genes, ℓ is the LD score, and *a* is the effect of population structure. We estimate ℓ for each gene by predicting expression for 503 European samples in 1000 Genomes[40] by using the GBLUP weights (see above) and then computing sample correlation. For each trait, we perform LD score regression by using z_{TWAS}^2 (which follows a χ^2 distribution asymptotically) to infer h_{GE}^2. We estimate heritability for each expression study separately to account for varying sample sizes and repeated gene measurements.

Estimating Genetic Correlation of Expression and Complex Traits from Summary Data

Let expression and traits be modeled as a linear function of the genotypes in a ~1 Mb locus flanking the gene: $\boldsymbol{\gamma}_{GE} = \boldsymbol{X}\boldsymbol{\beta}_{GE} + \boldsymbol{\epsilon}_{GE}$ and $\boldsymbol{\gamma}_T = \boldsymbol{X}\boldsymbol{\beta}_T + \boldsymbol{\epsilon}_T$, where $\boldsymbol{X}$ is the standardized genotype matrix, $\boldsymbol{\beta}_{GE}$ and $\boldsymbol{\beta}_T$ are the standardized effects for expression and traits,



$$\rho_g = cor([\beta_{GE} \times \alpha; \beta_{alt}]_{T_1}, [\beta_{GE} \times \alpha; \beta_{alt}]_{T_2})$$

$$\rho_{GE} = cor(\alpha_{T_1}, \alpha_{T_2})$$

Figure 1. Causal Diagram Illustrating the Genetic Component of a Trait

The total effect of SNPs on a trait can be partitioned into components that are mediated through *cis*-regulated (i.e., predicted, indicated by an asterisk) gene expression ($\beta_{GE} \times \alpha$) or through alternative pathways (β_{alt}). In contrast to ρ_g, which quantifies the correlation of the total SNP effects between two traits ($\beta_{GE} \times \alpha$; β_{alt}), ρ_{GE} focuses exclusively on the effects of *cis*-regulated gene expression (α).

respectively, and ϵ_{GE} and ϵ_T are the environmental noise for expression and traits, respectively. The local covariance between expression and complex traits is

$$\begin{aligned}\mathrm{cov}(\boldsymbol{y}_{GE}, \boldsymbol{y}_T) &= \mathrm{cov}(\boldsymbol{X}\boldsymbol{\beta}_{GE} + \boldsymbol{\epsilon}_{GE}, \boldsymbol{X}\boldsymbol{\beta}_T + \boldsymbol{\epsilon}_T) \\ &= \boldsymbol{\beta}_{GE}^{'}\mathrm{cov}(\boldsymbol{X}, \boldsymbol{X})\boldsymbol{\beta}_T + \mathrm{cov}(\boldsymbol{\epsilon}_{GE}, \boldsymbol{\epsilon}_T) \\ &= \boldsymbol{\beta}_{GE}^{'}\boldsymbol{V}\boldsymbol{\beta}_T + \mathrm{cov}(\boldsymbol{\epsilon}_{GE}, \boldsymbol{\epsilon}_T),\end{aligned}$$

where $\boldsymbol{V}$ is the LD matrix. If no individuals are shared between studies, then $\mathrm{cov}(\boldsymbol{\epsilon}_{GE}, \boldsymbol{\epsilon}_T) = 0$ (as in eQTL studies and GWASs). The local genetic correlation between expression and traits can be computed as

$$\rho_{g,local} = \frac{\boldsymbol{\beta}_{GE}^{'}\boldsymbol{V}\boldsymbol{\beta}_T}{\sqrt{h^2_{g,local}(GE)}\sqrt{h^2_{g,local}(T)}},$$

where $h^2_{g,local}(GE)$ and $h^2_{g,local}(T)$ are the local SNP heritability[41] for expression and traits, respectively, estimated at the locus. However, this requires knowledge of the true effect sizes. Given association statistics $\boldsymbol{z}_T$, we estimate an LD-adjusted effect size as $\widehat{\boldsymbol{\beta}}_T = \frac{1}{\sqrt{N_T}}\boldsymbol{V}^{-1}\boldsymbol{z}_T$. Hence, an estimate of the local genetic covariance[42] is given by

$$\widehat{\boldsymbol{\beta}}_{GE}^{'}\boldsymbol{V}\widehat{\boldsymbol{\beta}}_T = \frac{1}{\sqrt{N_{GE}}\sqrt{N_T}}(\boldsymbol{z}_{GE}^{'}\boldsymbol{V}^{-1})\boldsymbol{V}(\boldsymbol{V}^{-1}\boldsymbol{z}_T) = \widehat{\boldsymbol{b}}_{GE}^{'}\boldsymbol{V}^{-1}\widehat{\boldsymbol{b}}_T,$$

where $\widehat{\boldsymbol{b}}_{GE}$ and $\widehat{\boldsymbol{b}}_T$ are the marginal (i.e., LD-unadjusted) standardized effect-size estimates.[41,43] It follows that

$$\begin{aligned}\frac{1}{\sqrt{N_T}}z_{TWAS} &= \frac{1}{\sqrt{N_T}}\frac{\widehat{\boldsymbol{\beta}}_{GE}^{'}\boldsymbol{z}_T}{\sqrt{\mathrm{var}\left(\widehat{\boldsymbol{\beta}}_{GE}^{'}\boldsymbol{z}_T\right)}} = \frac{\widehat{\boldsymbol{b}}_{GE}^{'}\boldsymbol{V}^{-1}\widehat{\boldsymbol{b}}_T}{\sqrt{h^2_{g,local}(GE)}} \\ &= \rho_{g,local}\sqrt{h^2_{g,local}(T)}.\end{aligned}$$

We standardize this estimate to obtain our final local genetic correlation estimate as

$$\widehat{\rho}_{g,local} = \frac{z_{TWAS}}{\sqrt{N_T \times h^2_{g,local}(T)}}.$$

In practice, we use the variance explained by the local index SNP (i.e., smallest p value) as a proxy for $h^2_{g,local}(T)$.

Genetic Correlation between Traits at the Level of Predicted Expression

Consider a simple model where the genetic component of a trait can be decomposed into genetic effects that are mediated through *cis*-gene expressions of k genes plus genetic effects not mediated through expression at other loci in the genome:

$$\boldsymbol{y}_T = \sum_{i=1}^{k}(\boldsymbol{X}_i\boldsymbol{\beta}_{\mathbf{GE}_i})\alpha_i + \boldsymbol{X}_{\mathbf{alt}}\boldsymbol{\beta}_{\mathbf{alt}} + \boldsymbol{\epsilon}_T,$$

where $\boldsymbol{X}_i$ is a vector of genotypes at the *cis*-locus of gene i, $\boldsymbol{\beta}_{GE_i}$ is the casual eQTL effect vector for gene i, α_i is the direct effect of gene expression on a trait, and $\boldsymbol{X}_{\mathbf{alt}}$ and $\boldsymbol{\beta}_{\mathbf{alt}}$ refer to the genotype and causal effects, respectively, of variants not mediated through expression. We define the genome-wide genetic correlation at the level of expression between two complex traits as the correlation across the gene effects: $\rho_{GE} = \mathrm{cor}(\alpha_{T_1}, \alpha_{T_2})$. In practice, we do not know α, but we can estimate it as

$$\widehat{\alpha} = \frac{\mathrm{cov}(\boldsymbol{X}\boldsymbol{\beta}_{GE}, \boldsymbol{y}_T)}{\mathrm{var}(\boldsymbol{X}\boldsymbol{\beta}_{GE})} = \frac{\boldsymbol{\beta}_{GE}^{'}\boldsymbol{V}\boldsymbol{\beta}_T}{h^2_{g,local}(GE)} = \widehat{\rho}_{g,local}\Big/\frac{\sqrt{h^2_{g,local}(GE)}}{\sqrt{h^2_{g,local}(\boldsymbol{y}_T)}}$$

to obtain an estimate of expression correlation by using predicted expression ($\widehat{\rho}_{GE}$). In practice, we use the standardized estimates of $\widehat{\alpha}$, which are proportional to $\widehat{\rho}_{g,local}$. Unlike SNP-based genetic correlation (ρ_g), which captures genetic correlation across all common variants in the genome, ρ_{GE} captures only the component of genetic correlation driven by *cis* genetic effects on expression (see Figure 1). For instance, a pair of traits with highly correlated effects in *cis*-regions but weakly correlated effects in *trans*-regions will result in $\rho_{GE} > \rho_g$. In the absence of large *trans*-eQTL effects, we expect $\rho_{GE} \approx \rho_g$. Furthermore, because ρ_{GE} accounts for only the shared effect from predicted expression, any genetic effect on a trait not driven through expression in the measured eQTL data will not be represented in ρ_{GE}. We test for significance by assuming $\widehat{\rho}_{GE}\sqrt{(M-2)/(1-\widehat{\rho}^2_{GE})} \sim t(M-2)$, where M is the number of genes and t is the t distribution with $M-2$ degrees of freedom. This procedure requires the effects of M genes on the trait to be independent, which could be violated in practice; hence, we compute $\widehat{\rho}_{GE}$ by using one gene per 1 Mb locus.

Estimating Putative Casual Relationships between Pairs of Traits

To glean insight into the underlying causal relationship between pairs of traits, we perform a bi-directional regression[14] and estimate two different values of ρ_{GE} by varying gene sets. Before describing the approach, we first review several causal models that explain non-zero ρ_{GE} between two traits (see Figure 2). Models A and B depict causal relationships in which the effects of a gene set are mediated by one trait on the other. We can formally state model A (without loss of generality for B). Let trait 1 (T_1) be defined as $\boldsymbol{y}_{T_1} = \boldsymbol{G}_{T_1}\boldsymbol{\beta}_{T_1} + \boldsymbol{\epsilon}_{T_1}$, where $\boldsymbol{G}_{T_1}$ denotes the matrix of predicted expression at the causal genes, $\boldsymbol{\beta}_{T_1}$ is the effect size, and $\boldsymbol{\epsilon}_{T_1}$ is environmental noise. We define trait 2 (T_2) as

$$\boldsymbol{y}_{T_2} = \boldsymbol{y}_{T_1}\gamma_{T_1} + \boldsymbol{G}_{T_2}\boldsymbol{\beta}_{T_2} + \boldsymbol{\epsilon}_{T_2} = \boldsymbol{G}_{T_1}\boldsymbol{\beta}_{T_1}\gamma_{T_1} + \boldsymbol{G}_{T_2}\boldsymbol{\beta}_{T_2} + \boldsymbol{\epsilon}_{T_2'},$$

where γ_{T_1} is the causal effect of T_1 on T_2, $\boldsymbol{G}_{T_2}$ and $\boldsymbol{\beta}_{T_2}$ are the remaining causal genes and their effects, respectively, for T_2, and $\boldsymbol{\epsilon}_{T_2'}$ is the combined environment component. Under model A, the causal gene set for T_1 will have a non-zero effect on T_2 (i.e.,

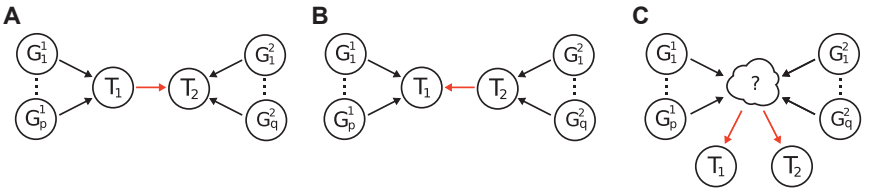


Figure 2. Illustration of Several Causal Models That Explain Expression Correlation for Traits 1 and 2 Given Their Causal Gene Sets

(Model A) Trait 1 directly influences trait 2. In this case, the effect of genes $G_1^1, \ldots, G_p^1$ on trait 2 is mediated by trait 1, which implies $\{G_i^1\}_{i=1}^{p} \subsetneq \{G_i^2\}_{i=1}^{q}$.
(Model B) Trait 2 directly influences trait 1. Similarly, the effect of genes $G_1^2, \ldots, G_q^2$ on trait 1 is mediated by trait 2, which implies $\{G_i^2\}_{i=1}^{q} \subsetneq \{G_i^1\}_{i=1}^{p}$.
(Model C) Traits 1 and 2 are influenced independently through an unobserved trait or traits.

$\gamma_{T_1} \neq 0$); however, if T_1 does not cause T_2, this effect will be zero given that unrelated genes have no downstream effect. Bi-directional regression provides a test to distinguish between models A and B by regressing estimated effect sizes for gene sets under model A (i.e., $\beta_{T_1} \sim \beta_{T_1}\gamma_{T_1}$) and comparing to estimates under model B (i.e., $\beta_{T_2} \sim \beta_{T_2}\gamma_{T_2}$). Because the causal gene sets for each trait are unknown, we use their identified susceptibility genes as a proxy. We estimate ρ_{GE} by conditioning on the gene set for trait i and denote its value as $\rho_{j|i}$. We repeat this procedure by ascertaining the gene set for trait j to obtain $\rho_{i|j}$. We perform a Welch's t test[44] to determine whether estimates of $\rho_{i|j}$ and $\rho_{j|i}$ are significantly different, thus providing evidence consistent with a causal direction. To minimize spurious results, we require at least ten genes for estimation in each conditional test. This approach mirrors bi-directional regression analyses of estimated SNP effects on two complex traits.[45,46] We stress that although a bi-directional approach is capable of rejecting model A in favor of model B (or vice versa), it cannot rule out model C, in which a shared pathway (or set of pathways) drives both traits independently (see Figure 2).

Simulation Framework

We simulate gene expression levels by using real genotype data measured in 503 European individuals from the 1000 Genomes Project.[40] Given a gene locus, we generate expression levels under the linear model $\boldsymbol{E} = \boldsymbol{X}\boldsymbol{w} + \epsilon$, where $\boldsymbol{E}$ is a gene expression vector of length N, X is the $N \times 2$ mean-centered and variance-standardized genotype matrix over two randomly selected SNPs in the locus, $\boldsymbol{w}$ is the causal effect, and ϵ is the environmental noise. We sample effect sizes $\boldsymbol{w}_i \sim \mathrm{N}(0, [h_g^2/2])$ for $i = 1$ and 2 and noise from a normal distribution to yield $h_g^2 = 0.1$ (consistent with what we observe in real gene expression data). We consider only SNPs with a MAF ≥ 0.01 and Hardy-Weinberg equilibrium deviation $p \geq 1 \times 10^{-5}$. We simulate a complex trait as a linear function of predicted gene expression for $k = 100$ genes, given by $\boldsymbol{y} = \sum_{i=1}^{k}(\boldsymbol{X}_i\boldsymbol{w}_i)\alpha_i + \epsilon$, where $\boldsymbol{X}_i\boldsymbol{w}_i$ is the predicted expression of the i^{th} gene with effect sizes $\alpha_i \sim \mathrm{N}(0, h_{GE}^2/k)$. For simulations involving ρ_{GE}, we simulate the two traits $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ by using the same process, except effects for the i^{th} gene are drawn from a bivariate normal distribution:

$$\begin{bmatrix} \alpha_{i,1} \\ \alpha_{i,2} \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{\alpha,1}^2 & \rho_{GE}\sigma_{\alpha,1}\sigma_{\alpha,2} \\ \rho_{GE}\sigma_{\alpha,1}\sigma_{\alpha,2} & \sigma_{\alpha,2}^2 \end{bmatrix} \right),$$

where $\sigma_{\alpha,*}^2 = (h_{GE,*}^2)/k$. Lastly, we perform an association scan on $\boldsymbol{y}$ by using all SNPs at each gene locus to obtain SNP-level Z scores $\boldsymbol{z}_T$.

Results

Accurate Estimation of Expression-Trait Genetic Correlation in Simulations

To validate our statistical framework for estimating $\rho_{g,local}$, we used real genotype data to perform simulations under various architectures (see Material and Methods). In brief, we simulated gene expression for 100 independent gene loci, which we then used to simulate a complex trait. Using our approach, we performed a GWAS and estimated $\rho_{g,local}$ from TWAS summary statistics (see Material and Methods). We observed unbiased estimates for $\rho_{g,local}$ both when causal variants were typed and when they were masked from the data (see Figure S1). Estimated values of $\rho_{g,local}$ were highly correlated with their true values ($r = 0.73$; $p < 2.2 \times 10^{-16}$), which indicates that using weights inferred from GBLUP maintains moderate power levels. This slight loss in power extended to h_{GE}^2 estimates, which quantify the total effect of predicted expression on a trait ($r = 0.74$; $p < 6.7 \times 10^{-12}$; see Table S3). As eQTL datasets increase in sample size, and predictive models become more accurate, we expect this attenuation bias to decrease.

We next performed extensive simulations to validate our procedure for estimating genetic correlation due to predicted expression (ρ_{GE}) between pairs of traits. We simulated genetically correlated complex traits from predicted expression by sampling effects from a bivariate normal distribution with correlation ρ_{GE} (see Material and Methods). We first estimated $\rho_{g,local}$ for each gene-trait pair, which served as input for estimating ρ_{GE}. Overall, we observed our estimator to be approximately unbiased, with conservative estimates for ρ_{GE} when its underlying value was near the boundaries (see Figure 3). Importantly, estimates were relatively unbiased when causal variants were untyped in the data. Our method appropriately accounted for LD among variants, resulting in a large improvement over the naive SNP correlation approach (which simply correlates the Z scores by ignoring LD). We also assessed our approach for testing for deviations from $\rho_{GE} = 0$ and found estimates consistent with the null distribution with $\lambda_{GC} = 0.97$ (Jack-knife 95% CI = [0.86, 1.08]; see Figure S2). To measure how sensitive our approach is to estimates of $h_{g,local}^2(GE)$ at each gene, we repeated simulations by using variance explained by the top eQTL as a proxy for local heritability. Although estimates were highly similar ($r = 0.99$; $p < 6.6 \times 10^{-7}$), our approach produced estimates closer to the ground truth (see Figure S3).

TWAS Identifies 1,196 Genes Associated with 30 Complex Traits and Diseases

We integrated GWAS summary data of 30 complex traits with gene expression to identify 1,196 susceptibility genes (i.e., genes with at least one significant trait association),

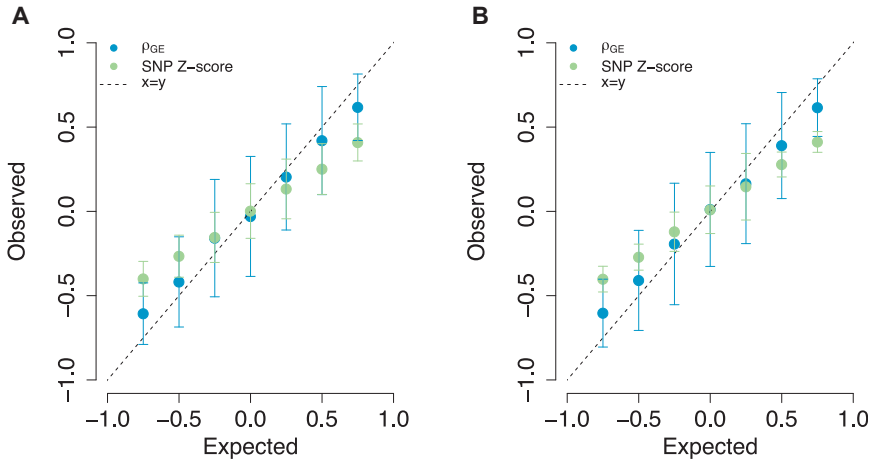


Figure 3. Simulation Results for $\widehat{\rho}_{GE}$ and Correlation of SNP *Z* Scores
Each point represents the mean estimate over 100 simulations. Error bars represent the 95% confidence interval estimated by the mean SE across simulations. The dotted line represents the identity line.
(A) Causal SNPs for gene expression are typed in the data.
(B) Causal SNPs are untyped.

comprising 5,490 total associations (after Bonferroni correction; see Material and Methods). Of these associations, we observed 1,789 distinct gene-trait pairs, of which 783 were found in anthropometric traits, 423 in metabolic traits, 215 in immune-related traits, 213 in hematopoietic traits, 137 in neurological traits (e.g., schizophrenia), and 18 in social traits (see Tables 1, S4, and S5). For example, the 137 susceptibility genes found for schizophrenia included *SNX19* (e.g., GTEx cerebellum; $p < 2.2 \times 10^{-8}$) and *NMRAL1* (e.g., GTEx skeletal muscle; $p < 9.7 \times 10^{-7}$); this is consistent with a previously reported study[12] that used different methods and expression data (see Table S6). We did not find susceptibility genes for forearm BMD, HOMA-B, or MCH concentration, consistent with low GWAS signal for these traits (see Table 1). Indeed, the number of GWAS risk loci strongly correlated with the number of identified susceptibility genes ($r = 0.99$; $p < 2.2 \times 10^{-16}$). Using the PANTHER database,[47] we explored putative molecular function and pathways enriched with identified susceptibility genes but were underpowered to detect molecular function for most individual traits (see Appendix A).

Next, we quantified the overlap of susceptibility genes and GWAS signals. Of the 1,789 identified gene-trait pairs, 168 (9%) were not proximal (more than 0.5 Mb from the TSS) to any genome-wide-significant SNP for that respective trait (see Table 2). This measure was robust to increases in window size, such that 140 (8%) gene-trait pairs did not overlap a genome-wide-significant SNP within 1 Mb of the TSS. We observed increased SNP association statistics at these genes (mean $\chi^2 = 6.5$; see Figure S4), which suggests that GWASs with an increased sample size will discover genome-wide-significant SNPs nearby. We tested this hypothesis by assessing the new TWAS loci for educational years[21] ($n = 126{,}599$) in a recent, much larger GWAS for educational years[35] ($n = 293{,}723$). All four independent loci contained a genome-wide-significant SNP in the larger GWAS (see Table S7). Of the 1,526 GWAS risk loci, 1,405 (92%) overlapped at least one eGene (i.e., a gene with heritable expression levels in at least one of the considered expression panels), and 551 (36%) overlapped at least one susceptibility gene (see Table 1). Focusing on the 1,621 TWAS associations that overlapped a genome-wide-significant SNP, we observed 1,350 (83%) genes that were not the closest, suggesting that the traditional heuristic of prioritizing genes closest to GWAS SNPs is typically not supported by evidence from eQTL data[48] (see Figure S5). This is also supported by the mean χ^2 association statistics for genes closest to index SNPs ($\chi^2 = 43.9$) and the top association ($\chi^2 = 72.9$; see Figure S6). In addition, lead GWAS SNPs typically have a weaker eQTL effect for the proximal gene than for the TWAS-implicated gene in 1,088 of 1,350 TWAS associations. This result, consistent with earlier reports,[11,12] highlights the importance of utilizing the entire locus and estimates of LD to prioritize genes.

Although GWAS SNPs provide the majority of the power in this approach, the flexibility of TWASs to leverage allelic heterogeneity provides a significant gain.[11] We found 219 instances across 19 traits where association signal was stronger (20% higher χ^2 statistics on average) in TWASs than in GWASs. For example, predicted expression in *CCDC88B* (OMIM: 611205; a gene involved in T cell maturation and inflammation[49]) exhibited strong association with Crohn disease ($p_{TWAS} = 6.32 \times 10^{-8}$), whereas the index SNP (i.e., top overlapping GWAS SNP) at site rs11231774 was only suggestive ($p_{GWAS} = 2.47 \times 10^{-6}$). This effect was most dramatic for height, such that 108 susceptibility genes had a stronger signal than GWAS index SNPs. We observed that the χ^2 statistics for predicted expression in *CRELD1* (OMIM: 607170; $p_{TWAS} = 1.55 \times 10^{-10}$) were 2.6× higher than those for the index SNP rs1473183 ($p_{GWAS} = 6.33 \times 10^{-5}$).

Recent work[50] applied a similar approach[12] that used summary eQTLs from blood and GWAS data to identify 71 genes for 28 complex traits.[50] Of the investigated traits, 12 overlapped those in our study. Overall, whereas that study reported 63 genes for these traits, we identified 564 genes. Surprisingly, despite using independent methods and expression data, we replicated 40 out of 51 associations for genes assayed in both studies (see Table S8). This increase in power can be attributed to two reasons. First, we integrated many more expression panels sampled from many tissues, leading to many more genes for the assay. Second, we used a method that jointly tests the entire locus rather than the index SNPs. We have shown

Table 1. Summary of GWAS and TWAS Results

		Number of GWASs				Number of Susceptibility Genes	
Trait	Abbreviation	Loci	Loci with an eGene	Loci with a Single Susceptibility Gene	Loci with at Least One Susceptibility Gene	Genes Overlapping GWASs	Genes Not Overlapping GWASs
Age at menarche	AM	70	60	14	19	34	9
Body mass index	BMI	76	60	10	18	44	11
College	COL	5	5	2	2	1	4
Crohn disease	CD	50	48	4	17	65	5
Educational years	EY	7	4	2	2	2	11
Fasting glucose	FG	12	11	2	5	8	1
Fasting insulin	FI	0	0	0	0	0	1
Femoral neck bone mineral density	FN	20	20	2	2	2	1
Forearm bone mineral density	FA	3	3	0	0	0	0
Hemoglobin	HB	22	21	2	5	22	3
HbA_{1c}	–	10	10	0	1	4	0
Height	–	482	454	94	225	669	52
High-density lipoprotein	HDL	100	95	11	29	98	4
HOMA-B	–	4	3	0	0	0	0
HOMA-IR	–	0	0	0	0	0	1
Inflammatory bowel disease	IBD	63	59	12	23	70	11
Low-density lipoprotein	LDL	75	72	8	25	84	3
Lumbar spine	LS	24	23	2	3	4	0
Mean cell hemoglobin concentration	MCHC	5	3	0	0	0	0
Mean cell hemoglobin	MCH	35	31	5	17	46	7
Mean cell volume	MCV	43	40	8	20	49	1
Number of platelets	PLT	35	34	6	13	30	8
Packed cell volume	PCV	14	13	1	3	5	1
Red blood cell count	RBC	25	21	3	10	35	2
Rheumatoid arthritis	RA	44	41	7	13	30	5
Schizophrenia	SCZ	95	74	15	31	113	24
Total cholesterol	TC	88	85	13	40	117	0
Triglycerides	TG	70	67	4	18	59	1
Type 2 diabetes	T2D	12	12	0	1	3	0
Ulcerative colitis	UC	37	36	5	9	27	2
Total		1,526	1,405	232	551	1,621	168

The first four numeric columns summarize GWAS risk loci. The last two numeric columns summarize identified TWAS susceptibility genes. The majority (92%) of GWAS risk loci overlap at least one eGene, of which 40% contain at least one susceptibility gene. We report 168 (9%) identified gene-trait pairs that do not overlap a GWAS variant, providing risk loci for follow up.

that many identified susceptibility genes contain signals of allelic heterogeneity; therefore, using individual SNPs will decrease power.

Genes Associated with Multiple Traits

We investigated the degree of pleiotropic susceptibility genes (i.e., genes associated with more than one trait) in our data and found 380 (32%) genes associated with multiple traits (see Figure S7). For example, *IKZF3* (OMIM: 606221) displayed strong associations with Crohn disease (NTR; $p = 1.6 \times 10^{-9}$), HDL levels (NTR; $p = 6.6 \times 10^{-15}$), inflammatory bowel disease (NTR; $p = 7.9 \times 10^{-16}$), rheumatoid arthritis (NTR; $p = 6.0 \times 10^{-8}$), and ulcerative colitis (NTR; $p = 9.2 \times 10^{-10}$). Indeed, *IKZF3* has been

Table 2. Susceptibility Genes That Do Not Overlap a Genome-wide Significant SNP within 0.5 Mb of the Transcription Start and End Sites for Each Trait

Trait	Genes
AM	*CCDC65*, *COG6*, *INO80E*, *NUCKS1*, *PMS2P5*, *RAB7L1*, *SLC26A9*, *STAG3L2*, and *TMEM180*
BMI	*CDK5RAP3*, *CERCAM*, *DHRS11*, *GGNBP2*, *INO80E*, *RP11-6N17.10*, *RP11-6N17.9*, *SLC27A4*, *STAG3L1*, *TUBA1C*, and *URM1*
CD	*CCDC88B*, *CISD1*, *PPP1R14B*, *RIT1*, and *SMIM19*
COL	*ABCB9*, *AC091729.9*, *AFF3*, and *RNF123*
EY	*ABCB9*, *EIF3CL*, *MIR4721*, *MPHOSPH9*, *NFATC2IP*, *RP11-1348G14.4*, *SDCCAG8*, *SH2B1*, *STK24*, *SULT1A1*, and *TUFM*
FG	*MAPRE3*
FI	*KNOP1*
FN	*FGFRL1*
HB	*CCDC117*, *UBE2Q2*, and *WNT3*
HDL	*HRAS*, *KNOP1*, *RETSAT*, and *TYRO3*
HEIGHT	*ARL17A*, *ATF1*, *ATP5J2*, *C20orf194*, *C9orf156*, *CCDC116*, *CNIH4*, *COX6B1*, *CRELD1*, *CRHR1*, *DAB2IP*, *DESI1*, *DLG5*, *DUS3L*, *ECHDC2*, *FAM35A*, *FUCA2*, *H2AFJ*, *HIBADH*, *INO80E*, *IQGAP1*, *KANSL1*, *LBX2-AS1*, *LRRC37A2*, *MAPT*, *MAT2A*, *MED4*, *MEGF9*, *MGMT*, *MORC2-AS1*, *MSRB2*, *P4HTM*, *PHF19*, *PLEKHA1*, *PSMD5*, *PSMD5-AS1*, *RP11-173M1.8*, *RP11-455F5.3*, *RP11-4O1.2*, *RP11-67A1.2*, *RP13-39P12.3*, *RP4-612B15.3*, *RRN3*, *SFTPD*, *SH3YL1*, *SUSD1*, *TMEM128*, *UBE2L3*, *UTP18*, *WDR60*, *YPEL3*, and *YWHAB*
HOMA-IR	*KNOP1*
IBD	*ADCY3*, *CCDC88B*, *FAM189B*, *GBA*, *GBAP1*, *HCN3*, *PPP1R14B*, *RMI2*, *SATB2*, *TMEM180*, *ZFP90*
LDL	*DHRS13*, *ERAL1*, and *WDR25*
MCH	*AP003419.16*, *GSTP1*, *PABPC4*, *PTPRCAP*, *RP11-69E11.4*, *RP1-18D14.7*, and *RPS6KB2*
MCV	*COX4I2*
PCV	*PLEKHH2*
PLT	*ACTR1A*, *BAZ2A*, *CCDC17*, *IPP*, *MUTYH*, *PRIM1*, *TESK2*, and *TMEM180*
RA	*METTL21B*, *RNF40*, *RPS26*, *SLC26A10*, and *SUOX*
RBC	*COX4I2* and *FBXL20*
SCZ	*ALMS1P*, *ARL14EP*, *CAD*, *CBR3*, *CEBPZ*, *CORO7*, *CPNE7*, *DND1*, *EMB*, *ENDOG*, *EPN2*, *GRAP*, *IK*, *NMRAL1*, *NRBP1*, *PCNX*, *PFDN1*, *PRR12*, *PRRG2*, *RNF112*, *RP11-135L13.4*, *SEPT10*, *SRA1*, and *TMCO6*
TG	*L3MBTL3*
UC	*SATB2* and *TNPO3*

For details on individual genes, expression studies, and association statistics, see Table S4. Genome-wide significance: $p < 5 \times 10^{-8}$.

shown to influence lymphocyte development and differentiation.[51,52] These traits are known to have a strong autoimmune component;[53] hence, association with predicted *IKZF3* expression levels is consistent with a model where *cis*-regulated variation in *IKZF3* product levels contributes to risk. Similarly, we observed three susceptibility genes shared between educational years (EY) and height (see Figure 4): *ABCB9* (OMIM: 605453; GTEx heart left ventricle; $p_{height} = 1.38 \times 10^{-15}$; $p_{EY} = 1.28 \times 10^{-6}$), *BTN2A3P* (OMIM: 613592; GTEx subcutaneous adipose; $p_{height} = 3.82 \times 10^{-12}$; $p_{EY} = 1.90 \times 10^{-7}$), and *MPHOSPH9* (OMIM: 605501; GTEx thyroid; $p_{height} = 5.84 \times 10^{-18}$; $p_{EY} = 1.30 \times 10^{-6}$). Although not direct evidence of co-localization of educational years and height at these loci, this result is consistent with a recent study[13] that reported a non-zero genetic correlation between height and educational years ($\hat{\rho}_g = 0.13$; $p = 3.82 \times 10^{-6}$).

The Effect of *cis* Expression on Traits Is Consistent across Tissues

Having established the importance of individual predicted gene expression levels for these traits, we next estimated the amount of trait variance explained by predicted expression by using all examined genes, including those not significantly associated, and an LD score regression approach (see Material and Methods). We found 108 tissue-trait pairs across 17 traits and 33 tissues where the cumulative effect of all measured genes on the trait was significantly greater ($p < 0.05/45$) than for the significant-only set (see Table S9). For example, in height we estimated $h^2_{GE} = 0.07$ (Jack-knife SE = 0.02; $p = 5.6 \times 10^{-4}$) by using all 3,733 measured genes in YFS and $h^2_{GE} = 0.015$ (Jack-knife SE = 6.9; p = 0.03) by using only the 169 YFS susceptibility genes ($p_{all>sig} = 5.6 \times 10^{-3}$). This suggests that height has additional susceptibility genes, which we are underpowered to detect. Strikingly, the predicted expression from all

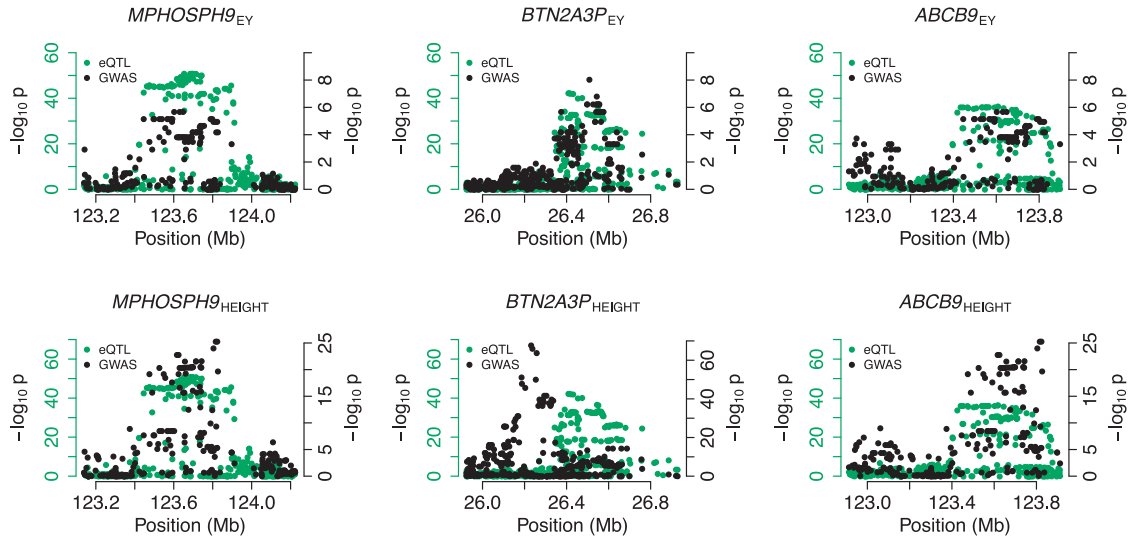


Figure 4. Susceptibility Genes Shared for Educational Years and Height
We indicate $-\log_{10}$ p values for eQTLs in green and trait-specific GWASs in black on separate axes to simplify illustration.

YFS genes accounts for 12% of SNP heritability measured in height.[54] However, for most trait-tissue pairs, we did not observe a significant difference at our given sample sizes. Indeed, we measured a significant association between expression-study sample size and number of eGenes ($r = 0.73$; SE = 0.10; $p = 1.3 \times 10^{-8}$), which indicates that smaller studies lack power to find eGenes and thus underestimate the total h^2_{GE}.

We next asked whether any tissues are burdened with increased levels of risk for a given trait. To test this hypothesis, we examined the difference between estimated trait variance explained per gene and the average. Our results did not suggest tissue-specific enrichment at the current sample sizes (see Table S10). We observed a significant correlation between gene expression sample size and tissue enrichment estimates ($p = 62.4 \times 10^{-6}$). One explanation for this relationship is that the number of eGenes identified per study increases with sample size, which increases h^2_{GE} estimates. Given no observable difference in tissue-specific risk, we expect local estimates of genetic correlation to be highly similar across tissues. When estimating $\rho_{g,local}$, we observed consistent effect-size estimates in both sign and magnitude estimates across tissues (mean tissue-tissue $r = 0.82$; see Figure 5). These results are compatible with earlier work that found that *cis* effects on expression are largely consistent across tissues.[55] To obtain a meta-estimate of local genetic correlation for gene-trait pairs with measurements in multiple tissues, we used the mean genetic correlation across all expression panels in all of the following analyses.

Genetic Correlation between Traits at the Level of Predicted Expression

To evaluate the shared contribution of predicted expression on pairs of traits, we used nominally significant ($p < 0.05$) genes to compute the genome-wide genetic correlation at levels of predicted expression (see Material and Methods). For 435 distinct pairs, we discovered 43 significant expression correlations, 22 of which had previously reported non-zero genetic correlations[13] (see Figure 6 and Table 3). For example, age of menarche and BMI had $\hat{\rho}_{GE} = -0.32$ (95% CI = [−0.32, −0.21]; $p = 7.97 \times 10^{-8}$). This negative correlation is consistent with estimates published in

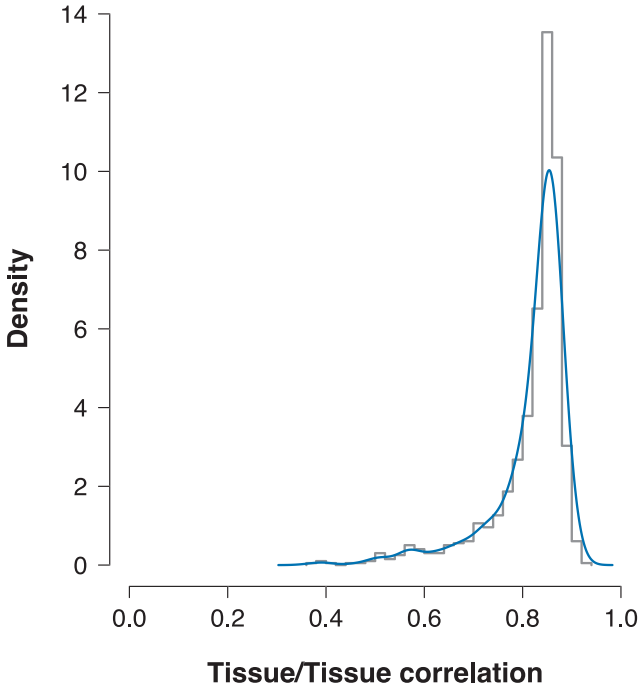


Figure 5. Histogram and Density Estimate for Correlation of $\rho_{g,local}$ across Tissues
We computed the correlation across pairs of different tissues by using local estimates of genetic correlation between expression and traits. Most tissues exhibited a high correlation over the underlying gene effects on traits with an estimated mean of $r = 0.82$.

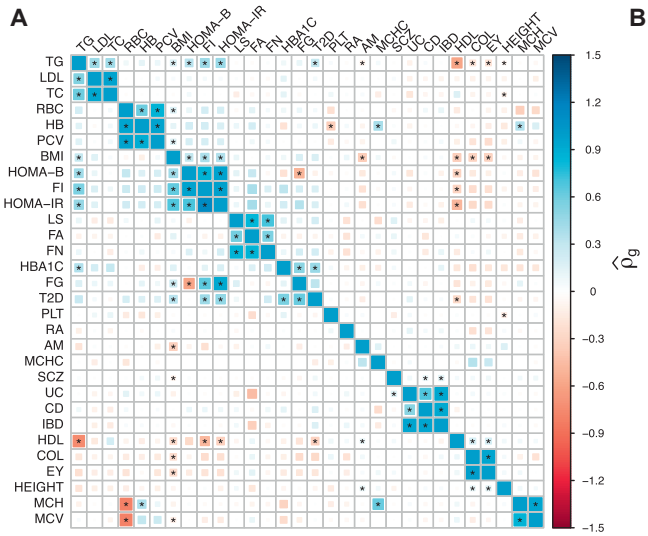


Figure 6. Estimates of Genetic Correlation $\widehat{\rho}_g$ Obtained from LD Scores versus Estimates of Expression Correlation $\widehat{\rho}_{GE}$ from Nominally Significant TWAS Results
(A) Correlation matrix for 30 traits. The lower triangle contains $\widehat{\rho}_{GE}$, and the upper triangle contains $\widehat{\rho}_g$ estimates. Correlation estimates that are significantly nonzero ($p < 0.05/435$) are marked with an asterisk (*). The strength and direction of correlation are indicated by size and color. We found 43 significantly correlated traits by using predicted expression and 62 by using genome-wide SNPs.
(B) Linear relationship between estimates of $\widehat{\rho}_{GE}$ and $\widehat{\rho}_g$. We indicate whether individual estimates were significant in either approach by color. Non-significant trait pairs are reduced in size for visibility.

epidemiological studies,[56] in addition to studies probing genetic correlation across complex traits.[13] To determine whether estimates were sensitive to changes in scale, we recomputed $\widehat{\rho}_{GE}$ by using the top eQTL as a proxy for local heritability of gene expression and observed similar results ($r = 0.99$; $p = 2.2 \times 10^{-16}$; see Figure S8). Results were also robust to increasing window size for gene pruning, such that there was no significant difference in estimates between 2 and 4 Mb windows ($r_{2Mb} = 0.99$; $r_{4Mb} = 0.98$). Using estimates of $\widehat{\rho}_{GE}$, we clustered traits and observed groups forming naturally in the trait-trait matrix (see Figure 6). Interestingly, BMI clustered with insulin-related traits (HOMA-B, HOMA-IR, and fasting insulin). Our estimates were highly consistent with the results of LD score regression (see Figure 6 and Table S11). Out of 435 pairs of traits, 35 demonstrated significance for $\widehat{\rho}_{GE}$ and $\widehat{\rho}_g$, whereas 8 and 27 were exclusive to $\widehat{\rho}_{GE}$ and $\widehat{\rho}_g$, respectively. Given the high degree of concordance between estimates, we tested for significant differences and found four insulin-related pairs of traits and three blood-related pairs with more extreme values for $\widehat{\rho}_{GE}$ (see Table S11). Differences for these pairs of traits can be partially explained by overconfident standard errors for $\widehat{\rho}_{GE}$ (see Table S12). Overall, we found $\widehat{\rho}_{GE}$ to explain most of the variation in $\widehat{\rho}_g$ ($r^2 = 0.72$). We compared this to the naive approach of computing the correlation of SNP Z scores across susceptibility gene loci and observed a much smaller proportion of variance explained in $\widehat{\rho}_g$ ($r^2 = 0.46$). This reinforces that, compared to the naive approach, our method incorporates LD to aggregate signal.

Bi-directional Regression Suggests Putative Causal Relationships

Given pairs of traits with significant estimates of ρ_{GE}, we aimed to distinguish among possible causal explanations by performing bi-directional regression analyses (see Material and Methods). To empirically validate our approach, we regressed HDL, LDL, and TG with TC. TC is the direct consequence of summing over TG, HDL, and LDL levels, so we expected to observe higher signal for $\rho_{TC\,|\,lipid}$ than for $\rho_{lipid\,|\,TC}$. Of these three, we found evidence that TG influences TC ($p = 2.34 \times 10^{-3}$). We observed consistent, but not significant, evidence for the effects of LDL on TC ($p = 0.07$) and HDL on TC ($p = 0.55$; see Figure 7). These results suggest that point estimates from the bi-directional approach favor the correct model but might not have adequate power required for significance.

We tested the 43 pairs of traits identified above (see Table 3) while ascertaining susceptibility genes and observed asymmetric effects at $p < 0.05$ for BMI-TG and LDL-TG (see Figure 8 and Table 4). For example, in the bi-directional analysis on BMI and TG, we observed a significant effect for $\rho_{TG\,|\,BMI} = 0.62$ (95% CI = [0.27, 0.83]; $p = 2.06 \times 10^{-3}$). By contrast, the reverse analysis estimate overlapped 0 at $\rho_{BMI\,|\,TG} = -0.04$ (95% CI = [−0.49, 0.42]; $p = 0.86$). Individual estimates for $\rho_{TG\,|\,BMI}$ and $\rho_{BMI\,|\,TG}$ were significantly different ($p = 0.01$, Welch's t test), which is consistent with a model where BMI directly influences TG levels. In practice, we used susceptibility genes found through a TWAS ($p \sim 1 \times 10^{-6}$), but this could be too strict an inclusion threshold for genes for which we lack power to detect. We conducted analyses with weaker thresholds and observed similar results (see Tables S13 and S14). Our results reinforce previous estimates of putative causal effects where BMI influences TG levels.[45,57]

Discussion

In this work, we described an approach to estimate the local genetic covariance and correlation between gene expression and complex traits by using GWAS summary data. We also introduced a method of estimating genome-wide genetic correlation between complex traits at the level of predicted expression. Using simulations, we demonstrated that both approaches are relatively unbiased under realistic

Table 3. Pairs of Traits with Significant Estimates of ρ_{GE}

		All Nominally Significant Genes			
Trait 1	**Trait 2**	$\widehat{\rho}_{GE}$	**95% CI**		**M**
AM	BMI	−0.33	−0.43	−0.21	257
BMI	COL	−0.31	−0.44	−0.18	190
BMI	EY	−0.31	−0.43	−0.18	210
BMI	FI	0.39	0.25	0.51	164
BMI	HDL	−0.34	−0.45	−0.23	256
BMI	HOMA-B	0.31	0.17	0.44	168
BMI	HOMA-IR	0.36	0.22	0.49	162
BMI	TG	0.29	0.17	0.41	233
CD	IBD	0.93	0.91	0.94	366
CD	UC	0.51	0.41	0.60	218
COL	EY	0.95	0.94	0.96	363
FA	FN	0.57	0.44	0.67	149
FA	LS	0.60	0.49	0.69	170
FG	FI	0.65	0.53	0.74	133
FG	HOMA-B	−0.60	−0.70	−0.47	125
FG	HOMA-IR	0.92	0.89	0.94	136
FI	HDL	−0.31	−0.44	−0.17	168
FI	HOMA-B	0.97	0.96	0.98	243
FI	HOMA-IR	0.99	0.99	0.99	383
FI	TG	0.57	0.45	0.66	152
FN	LS	0.86	0.83	0.89	264
HB	MCH	0.37	0.23	0.50	156
HB	MCHC	0.40	0.23	0.55	105
HB	PCV	0.97	0.96	0.97	338
HB	PLT	−0.36	−0.49	−0.20	141
HB	RBC	0.95	0.94	0.96	260
HbA_{1c}	T2D	0.46	0.30	0.59	110
HbA_{1c}	TG	0.37	0.21	0.50	137
HDL	HOMA-IR	−0.32	−0.46	−0.18	159
HDL	T2D	−0.32	−0.45	−0.19	186
HDL	TG	−0.74	−0.79	−0.69	274
HOMA-B	HOMA-IR	0.97	0.96	0.98	227
HOMA-B	TG	0.43	0.27	0.56	127
HOMA-IR	TG	0.48	0.34	0.60	138
IBD	UC	0.96	0.95	0.96	415
LDL	TC	0.97	0.96	0.97	452
LDL	TG	0.54	0.44	0.63	231
MCH	MCHC	0.63	0.51	0.72	127
MCH	MCV	0.96	0.95	0.97	320
MCH	RBC	−0.81	−0.85	−0.76	207
MCV	RBC	−0.80	−0.85	−0.75	208
PCV	RBC	0.96	0.95	0.97	278
TC	TG	0.61	0.53	0.68	248

Estimates were computed with M pruned genes that were nominally significant ($p < 0.05$) in both traits.

scenarios. We used GWAS summary statistics from 30 complex traits and diseases jointly with expression data collected across 45 expression panels to identify 1,196 susceptibility genes for complex traits. Interestingly, susceptibility genes that were identified for educational years and not proximal to a genome-wide significant SNP were validated in a much larger GWAS.[35] We leveraged estimates of local genetic correlation between gene expression and traits to compute ρ_{GE} for 435 trait pairs. This quantified the shared effect of predicted expression levels between two complex traits. To provide evidence of possible causal direction, we adapted a recently proposed causality test[45] to operate at the level of predicted gene expression. Our results suggest that TG influences LDL and that BMI influences TG. As more GWAS and eQTL summary results become publicly available, we expect additional studies to integrate cross-trait information to make inferences about mechanistic bases for complex traits. Indeed, recent work has combined chromatin phenotypes with alternatively spliced introns and total gene expression (the latter of which overlaps expression used in this study) to identify regulatory mechanisms for schizophrenia.[58]

Under the assumption that gene expression mediates the effect of genetics on complex traits, testing for association between predicted gene expression and traits is equivalent to a two-sample Mendelian randomization test for a causal effect of expression on a trait.[59,60] This test for causality is valid if SNPs do not exhibit pleiotropic effects, which is difficult to prove; therefore, TWAS associations do not provide direct evidence of causal relationships between gene expression and complex traits but rather reflect associations between expression levels and traits. This set of assumptions extends to our bi-directional approach to inferring causal direction. A bi-directional regression is capable of distinguishing between directions of effect but cannot rule out pleiotropy. Therefore, our results show consistency with a putative causal mechanism and should not be interpreted as direct proof of causality.

We conclude with several caveats. First, we note that using estimates of genetic correlation to find susceptibility genes could still be biased as a result of confounding. The expression weights used for TWASs could tag variants that are causal through other genes or non-genic mechanisms. In principle, this can be partially remedied by joint testing of multiple genes and a trait. In this work, we combined

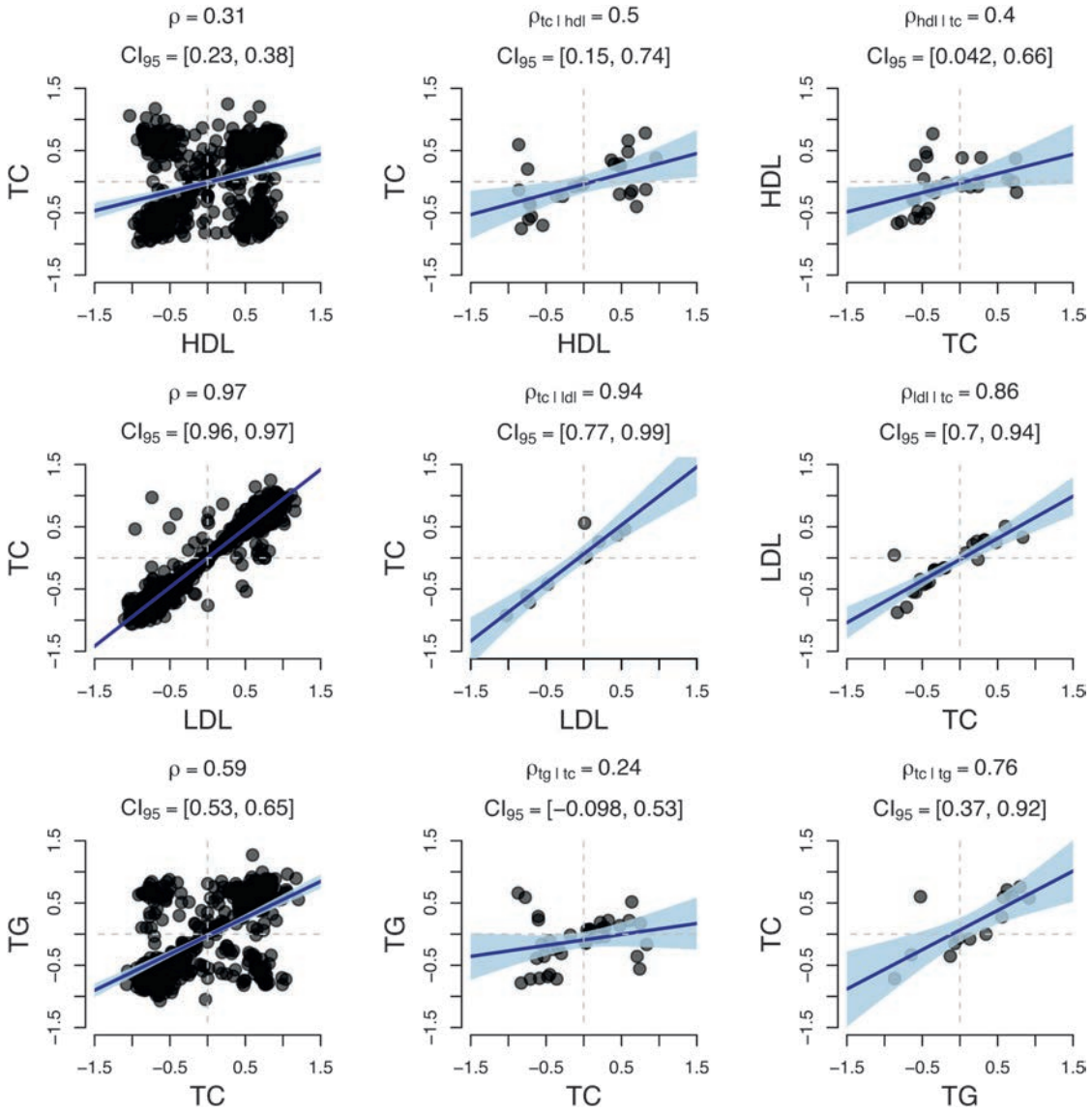


Figure 7. Estimates of Expression Correlation ρ_{GE} between TC and HDL, LDL, and TG
(Left column) Estimates of ρ_{GE} with the use of nominally significant genes ($p < 0.05$).
(Middle column) We repeated the analysis by using only susceptibility genes found in the x axis trait but not found in the y axis trait.
(Right right) Same analysis as in the middle column but with the other trait's susceptibility genes.
All three analyses resulted in stronger point estimates for $\rho_{TC\ |lipid}$ when conditioning on HDL, LDL, and TG genes than for $\rho_{lipid\ |TC}$; however, significance was observed only for $\rho_{TC|TG}$ ($p = 2.34 \times 10^{-3}$). Shaded regions indicate the estimated 95% confidence interval for the regression line.

estimates across tissues by taking the mean effect to compute the genetic correlation between traits and expression. This approach is unbiased but could be inefficient. Recent work[61] has described a random-effect model that combines estimates across tissues to increase power. Finally, our method of estimating correlation between traits by using the genetically predicted component of gene expression makes several simplifying assumptions. First, we remedied the non-independence of genes by sampling single genes within a 1 Mb region, an approach that has been used previously.[46] However, a more powerful approach could take correlations across genes into account. Second, we limited predictive models to the local (or *cis*) effects on gene expression, which ignores distal (or *trans*) effects that regulate gene expression. Although the predictive accuracy of models for gene expression used in this study can account for most of the variation due to genetics,[11] we believe that incorporating additional sources of genomic information (e.g., functional priors on SNP effects[39,62,63]) could make additional refinement possible.

Appendix A: Pathway Analysis

We used the PANTHER database[47] to explore putative molecular function and pathways enriched with identified

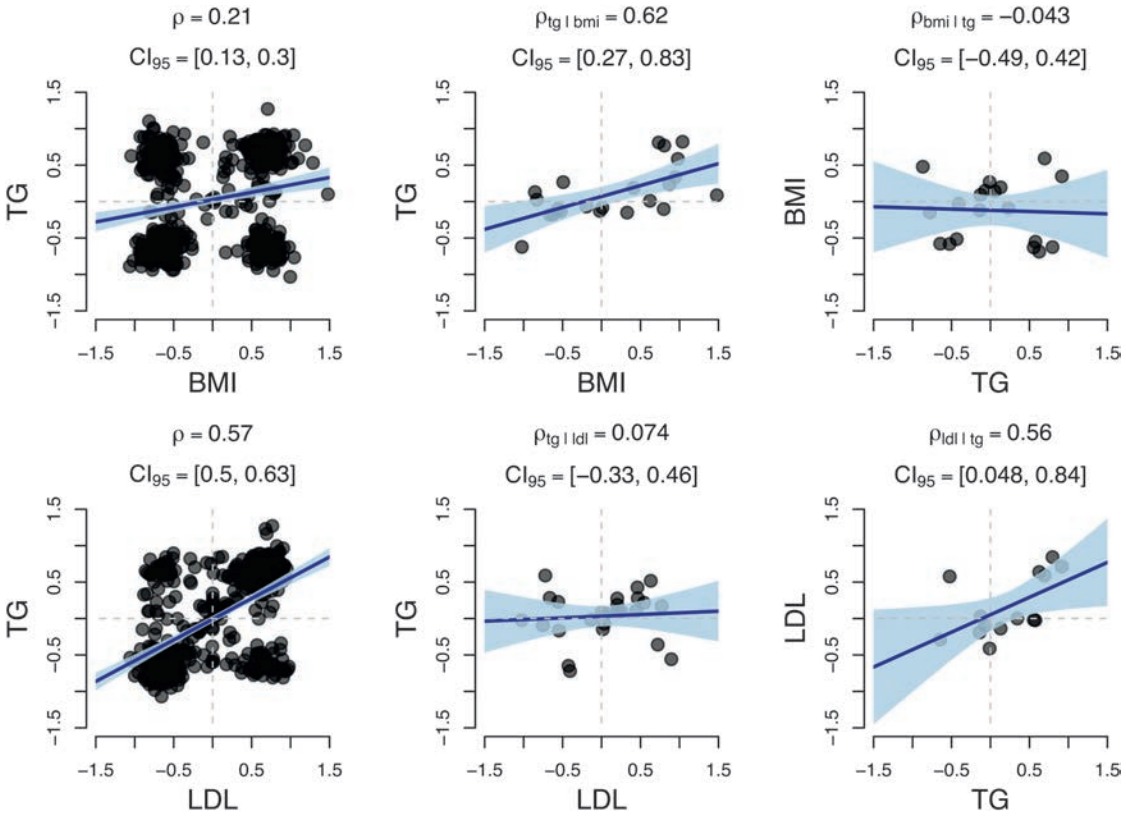


Figure 8. Estimates of $\hat{\rho}_{GE}$ for TG with BMI and for TG with LDL
We present results for pairs of traits that displayed a significant difference ($p < 0.05$, Welch's t test) in their conditional estimates. These results are consistent with a causal model where BMI influences TG and TG influences LDL. Shaded regions indicate the estimated 95% confidence interval for the regression line.

susceptibility genes. Using all susceptibility genes across all traits, we found 13 significantly enriched categories, of which seven were related to binding functions. Catalytic activity exhibited the strongest enrichment at 1.3× (GO: 0003824; $p = 5.17 \times 10^{-9}$; see Figure S9). We next focused on individual traits (see Figure S10); however, most individually tested gene sets did not indicate significant enrichment, except for height, LDL, and TC. For example, height had a significant enrichment of genes with catalytic activity (1.31×; $p = 4.77 \times 10^{-4}$). We next looked at biological processes and found TWAS genes enriched at 1.2× for metabolic processes (GO: 0008152; $p = 7.29 \times 10^{-11}$) and 1.57× cellular catabolic processes (GO: 0044248; $p = 2.51 \times 10^{-2}$; see Figures S11 and S12). Enrichment was most pronounced in susceptibility genes specific to height (1.3×; $p = 1.03 \times 10^{-6}$).

Supplemental Data

Supplemental Data include 12 figures and 14 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2017.01.031.

Acknowledgments

We would like to thank Valerie Arboleda, Robert Brown, Kathy Burch, and Malika Kumar for helpful discussions and feedback. We also thank Dr. Nicole Soranzo for sharing summary data for the

Table 4. Bi-directional Estimates of Genome-wide Genetic Correlation at the Level of Predicted Expression

		Results when Ascertaining for Trait 1				Results when Ascertaining for Trait 2				Test for Difference		
Trait 1	**Trait 2**	$\hat{\rho}_{GE}$	**SE**	**p**	***M***	$\hat{\rho}_{GE}$	**SE**	**p**	***M***	***t***	**p**	***~M***
BMI	TG	0.62	0.10	2.06×10^{-3}	22	−0.04	0.22	8.62×10^{-1}	19	2.74	1.12×10^{-2}	25
LDL	TG	0.07	0.19	7.25×10^{-1}	25	0.56	0.13	3.55×10^{-2}	14	−2.17	3.69×10^{-2}	36
TC	TG	0.24	0.14	1.63×10^{-1}	36	0.76	0.08	1.79×10^{-3}	14	−3.22	2.34×10^{-3}	47

We denote the number of ascertained genes used in the test as *M*. We tested for a difference as a *t* statistic, where $t = \frac{|\rho_{GE,1} - \rho_{GE,2}|}{\sqrt{SE_1^2 + SE_2^2}} \sim t(df)$ and df is the approximate degrees of freedom determine by the Welch-Satterthwaite equation.

platelet traits. This research was funded in part by NIH awards GM105857, GM053275, and HG009120. G.K. is supported by the Biomedical Big Data Training Program (NIH-NCI T32CA201160). CMC data were generated as part of the CommonMind Consortium, supported by funding from Takeda Pharmaceuticals, F. Hoffman-La Roche, and NIH grants R01MH085542, R01MH093725, P50MH066392, P50MH080405, R01MH097276, RO1-MH-075916, P50M096891, P50MH084053S1, R37MH057881, R37MH057881S1, HHSN271201300031C, AG02219, AG05138, and MH06692. Brain tissue for the study was obtained from the following brain bank collections: the Mount Sinai NIH Brain and Tissue Repository, the University of Pennsylvania Alzheimer Disease Core Center, the University of Pittsburgh NeuroBioBank and Brain and Tissue Repositories, and the National Institute of Mental Health (NIMH) Human Brain Collection Core. CommonMind Consortium leadership includes Pamela Sklar, Joseph Buxbaum (Icahn School of Medicine at Mount Sinai), Bernie Devlin, David Lewis (University of Pittsburgh), Raquel Gur, Chang-Gyu Hahn (University of Pennsylvania), Keisuke Hirai, Hiroyoshi Toyoshiba (Takeda Pharmaceuticals), Enrico Domenici, Laurent Essioux (F. Hoffman-La Roche), Lara Mangravite, Mette Peters (Sage Bionetworks), Thomas Lehner, Barbara Lipska (NIMH).

Received: August 31, 2016
Accepted: January 23, 2017
Published: February 23, 2017

Web Resources

CommonMind Consortium, https://www.synapse.org
FUSION software package, http://gusevlab.org/projects/fusion/
GCTA, http://cnsgenomics.com/software/gcta/
Gene Ontology, http://www.geneontology.org/
GTEx Portal, http://www.gtexportal.org/home/
OMIM, http://www.omim.org
PLINK, https://www.cog-genomics.org/plink2/
RhoGE software, https://github.com/bogdanlab/RHOGE

References

1. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.
2. Claussnitzer, M., Dankel, S.N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., et al. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N. Engl. J. Med. *373*, 895–907.
3. Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C.C., Patsopoulos, N.A., Moutsianas, L., Dilthey, A., Su, Z., Freeman, C., Hunt, S.E., et al.; International Multiple Sclerosis Genetics Consortium; and Wellcome Trust Case Control Consortium 2 (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature *476*, 214–219.
4. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. *6*, e1000888.
5. Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. Nature *452*, 423–428.
6. Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. *6*, e1000895.
7. Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. *16*, 197–212.
8. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al.; GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.
9. Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al.; Geuvadis Consortium (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.
10. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J., Im, H.K.; and GTEx Consortium (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.
11. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.
12. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., and Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. *48*, 481–487, advance online publication.
13. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., Duncan, L., Perry, J.R., Patterson, N., Robinson, E.B., et al.; ReproGen Consortium; Psychiatric Genomics Consortium; and Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3 (2015). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.
14. Davey Smith, G., and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum. Mol. Genet. *23* (R1), R89–R98.
15. Zheng, H.F., Forgetta, V., Hsu, Y.H., Estrada, K., Rosello-Diez, A., Leo, P.J., Dahia, C.L., Park-Min, K.H., Tobias, J.H., Kooperberg, C., et al.; AOGC Consortium; and UK10K Consortium (2015). Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature *526*, 112–117.
16. Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., et al.; Wellcome Trust Case Control Consortium; Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; and DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. *44*, 981–990.
17. Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., et al.;

International Multiple Sclerosis Genetics Consortium; and International IBD Genetics Consortium (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. *47*, 979–986.

18. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
19. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., et al.; RACI consortium; and GARNET consortium (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature *506*, 376–381.
20. Perry, J.R.B., Day, F., Elks, C.E., Sulem, P., Thompson, D.J., Ferreira, T., He, C., Chasman, D.I., Esko, T., Thorleifsson, G., et al.; Australian Ovarian Cancer Study; GENICA Network; kConFab; LifeLines Cohort Study; InterAct Consortium; and Early Growth Genetics (EGG) Consortium (2014). Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. Nature *514*, 92–97.
21. Rietveld, C.A., Medland, S.E., Derringer, J., Yang, J., Esko, T., Martin, N.W., Westra, H.-J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., et al.; LifeLines Cohort Study (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science *340*, 1467–1471.
22. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al.; Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283.
23. Soranzo, N., Sanna, S., Wheeler, E., Gieger, C., Radke, D., Dupuis, J., Bouatia-Naji, N., Langenberg, C., Prokopenko, I., Stolerman, E., et al.; WTCCC (2010). Common variants at 10 genomic loci influence hemoglobin A_1(C) levels via glycemic and nonglycemic pathways. Diabetes *59*, 3229–3239.
24. Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A.U., Wheeler, E., Glazer, N.L., Bouatia-Naji, N., Gloyn, A.L., et al.; DIAGRAM Consortium; GIANT Consortium; Global BPgen Consortium; Anders Hamsten on behalf of Procardis Consortium; and MAGIC investigators (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat. Genet. *42*, 105–116.
25. Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A.H., Labrune, Y., et al. (2011). New gene functions in megakaryopoiesis and platelet formation. Nature *480*, 201–208.
26. van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D.S., Elling, U., Allayee, H., Li, X., et al. (2012). Seventy-five genetic loci influencing the human red blood cell. Nature *492*, 369–375.
27. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al.; LifeLines Cohort Study; ADIPOGen Consortium; AGEN-BMI Working Group; CARDIOGRAMplusC4D Consortium; CKDGen Consortium; GLGC; ICBP; MAGIC Investigators; MuTHER Consortium; MIGen Consortium; PAGE Consortium; ReproGen Consortium; GENIE Consortium; and International Endogene Consortium (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature *518*, 197–206.
28. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al.; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; and LifeLines Cohort Study (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.
29. Fromer, M., Roussos, P., Sieberts, S.K., Johnson, J.S., Kavanagh, D.H., Perumal, T.M., Ruderfer, D.M., Oh, E.C., Topol, A., Shah, H.R., et al. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat. Neurosci. *19*, 1442–1453.
30. Raitakari, O.T., Juonala, M., Rönnemaa, T., Keltikangas-Järvinen, L., Räsänen, L., Pietikäinen, M., Hutri-Kähönen, N., Taittonen, L., Jokinen, E., Marniemi, J., et al. (2008). Cohort profile: the cardiovascular risk in Young Finns Study. Int. J. Epidemiol. *37*, 1220–1226.
31. Stancáková, A., Civelek, M., Saleem, N.K., Soininen, P., Kangas, A.J., Cederberg, H., Paananen, J., Pihlajamäki, J., Bonnycastle, L.L., Morken, M.A., et al. (2012). Hyperglycemia and a common variant of GCKR are associated with the levels of eight amino acids in 9,369 Finnish men. Diabetes *61*, 1895–1902.
32. Stancáková, A., Javorský, M., Kuulasmaa, T., Haffner, S.M., Kuusisto, J., and Laakso, M. (2009). Changes in insulin sensitivity and insulin release in relation to glycemia and glucose tolerance in 6,414 Finnish men. Diabetes *58*, 1212–1221.
33. Nuotio, J., Oikonen, M., Magnussen, C.G., Jokinen, E., Laitinen, T., Hutri-Kähönen, N., Kähönen, M., Lehtimäki, T., Taittonen, L., Tossavainen, P., et al. (2014). Cardiovascular risk factors in 2011 and secular trends since 2007: the Cardiovascular Risk in Young Finns Study. Scand. J. Public Health *42*, 563–571.
34. Wright, F.A., Sullivan, P.F., Brooks, A.I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y.-H., et al. (2014). Heritability and genomics of gene expression in peripheral blood. Nat. Genet. *46*, 430–437.
35. Okbay, A., Beauchamp, J.P., Fontana, M.A., Lee, J.J., Pers, T.H., Rietveld, C.A., Turley, P., Chen, G.-B., Emilsson, V., Meddens, S.F.W., et al.; LifeLines Cohort Study (2016). Genome-wide association study identifies 74 loci associated with educational attainment. Nature *533*, 539–542.
36. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
37. de Los Campos, G., Vazquez, A.I., Fernando, R., Klimentidis, Y.C., and Sorensen, D. (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. *9*, e1003608.
38. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M.; and Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
39. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al.; ReproGen Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; and RACI Consortium (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.
40. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean,

G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
41. Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. Am. J. Hum. Genet. *99*, 139–153.
42. Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2016). Local genetic correlation gives insights into the shared genetic architecture of complex traits. bioRxiv. http://dx.doi.org/10.1101/092668.
43. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Madden, P.A.F., Heath, A.C., Martin, N.G., Montgomery, G.W., Weedon, M.N., Loos, R.J., et al.; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; and DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet. *44*, 369–375, S1–S3.
44. Welch, B.L. (1947). The generalisation of student's problems when several different population variances are involved. Biometrika *34*, 28–35.
45. Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y., and Hinds, D.A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. Nat. Genet. *48*, 709–717.
46. Do, R., Willer, C.J., Schmidt, E.M., Sengupta, S., Gao, C., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., et al. (2013). Common variants associated with plasma triglycerides and risk for coronary artery disease. Nat. Genet. *45*, 1345–1352.
47. Mi, H., Muruganujan, A., Casagrande, J.T., and Thomas, P.D. (2013). Large-scale gene function analysis with the PANTHER classification system. Nat. Protoc. *8*, 1551–1566.
48. Won, H., de la Torre-Ubieta, L., Stein, J.L., Parikshak, N.N., Huang, J., Opland, C.K., Gandal, M.J., Sutton, G.J., Hormozdiari, F., Lu, D., et al. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. Nature *538*, 523–527.
49. Kennedy, J.M., Fodil, N., Torre, S., Bongfen, S.E., Olivier, J.-F., Leung, V., Langlais, D., Meunier, C., Berghout, J., Langat, P., et al. (2014). CCDC88B is a novel regulator of maturation and effector functions of T cells during pathological inflammation. J. Exp. Med. *211*, 2519–2535.
50. Pavlides, J.M.W., Zhu, Z., Gratten, J., McRae, A.F., Wray, N.R., and Yang, J. (2016). Predicting gene targets from integrative analyses of summary data from GWAS and eQTL studies for 28 human complex traits. Genome Med. *8*, 84.
51. Hosokawa, Y., Maeda, Y., Takahashi, E.-i., Suzuki, M., and Seto, M. (1999). Human aiolos, an ikaros-related zinc finger DNA binding protein: cDNA cloning, tissue expression pattern, and chromosomal mapping. Genomics *61*, 326–329.
52. Quintana, F.J., Jin, H., Burns, E.J., Nadeau, M., Yeste, A., Kumar, D., Rangachari, M., Zhu, C., Xiao, S., Seavitt, J., et al. (2012). Aiolos promotes TH17 differentiation by directly silencing Il2 expression. Nat. Immunol. *13*, 770–777.
53. Farh, K.K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J.H., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337–343.
54. Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Lee, S.H., Robinson, M.R., Perry, J.R.B., Nolte, I.M., van Vliet-Ostaptchouk, J.V., et al.; LifeLines Cohort Study (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. *47*, 1114–1120.
55. Gutierrez-Arcelus, M., Ongen, H., Lappalainen, T., Montgomery, S.B., Buil, A., Yurovsky, A., Bryois, J., Padioleau, I., Romano, L., Planchon, A., et al. (2015). Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. PLoS Genet. *11*, e1004958.
56. Parsons, T.J., Power, C., Logan, S., and Summerbell, C.D. (1999). Childhood predictors of adult obesity: a systematic review. Int. J. Obes. Relat. Metab. Disord. *23* (*Suppl 8*), S1–S107.
57. Fall, T., Hägg, S., Mägi, R., Ploner, A., Fischer, K., Horikoshi, M., Sarin, A.-P., Thorleifsson, G., Ladenvall, C., Kals, M., et al.; European Network for Genetic and Genomic Epidemiology (ENGAGE) consortium (2013). The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis. PLoS Med. *10*, e1001474.
58. Gusev, A., Mancuso, N., Finucane, H.K., Reshef, Y., Song, L., Safi, A., Oh, E., McCaroll, S., Neale, B., Ophoff, R., et al. (2016). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. bioRxiv. http://dx.doi.org/10.1101/067355.
59. Pickrell, J. (2015). Fulfilling the promise of Mendelian randomization. bioRxiv. http://dx.doi.org/10.1101/018150.
60. Smith, G.D., and Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int. J. Epidemiol. *32*, 1–22.
61. Wang, J., Gamazon, E.R., Pierce, B.L., Stranger, B.E., Im, H.K., Gibbons, R.D., Cox, N.J., Nicolae, D.L., and Chen, L.S. (2016). Imputing Gene Expression in Uncollected Tissues Within and Beyond GTEx. Am. J. Hum. Genet. *98*, 697–708.
62. Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.
63. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. *10*, e1004722.

AJHG

AJHG offers unparalleled international readership

AJHG publishes some of the most-read articles in the field, with more than 2 million papers downloaded globally in 2016 alone.

Here are just a few of the most-read recent papers*:

- International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases
- Systematic Computational Identification of Variants That Activate Exonic and Intronic Cryptic Splice Sites
- Mutations in Epigenetic Regulation Genes Are a Major Cause of Overgrowth with Intellectual Disability
- The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States
- CHARGE and Kabuki Syndromes: Gene-Specific DNA Methylation Signatures Identify Epigenetic Mechanisms Linking These Clinically Overlapping Conditions

*Most-downloaded AJHG articles from the last 30 days
(according to cell.com May 15, 2017)

Did you know?

ASHG member benefits

- Free online access to *AJHG*
- No page or color-figure charges

CellPress

ASHG

Mutations in Epigenetic Regulation Genes Are a Major Cause of Overgrowth with Intellectual Disability

Katrina Tatton-Brown,[1,2] Chey Loveday,[1] Shawn Yost,[1] Matthew Clarke,[1] Emma Ramsay,[1] Anna Zachariou,[1] Anna Elliott,[1] Harriet Wylie,[1] Anna Ardissone,[3] Olaf Rittinger,[4] Fiona Stewart,[5] I. Karen Temple,[6,7] Trevor Cole,[8] Childhood Overgrowth Collaboration, Shazia Mahamdallie,[1] Sheila Seal,[1] Elise Ruark,[1] and Nazneen Rahman[1,9,10,*]

To explore the genetic architecture of human overgrowth syndromes and human growth control, we performed experimental and bioinformatic analyses of 710 individuals with overgrowth (height and/or head circumference $\geq +2$ SD) and intellectual disability (OGID). We identified a causal mutation in 1 of 14 genes in 50% (353/710). This includes *HIST1H1E*, encoding histone H1.4, which has not been associated with a developmental disorder previously. The pathogenic *HIST1H1E* mutations are predicted to result in a product that is less effective in neutralizing negatively charged linker DNA because it has a reduced net charge, and in DNA binding and protein-protein interactions because key residues are truncated. Functional network analyses demonstrated that epigenetic regulation is a prominent biological process dysregulated in individuals with OGID. Mutations in six epigenetic regulation genes—*NSD1*, *EZH2*, *DNMT3A*, *CHD8*, *HIST1H1E*, and *EED*—accounted for 44% of individuals (311/710). There was significant overlap between the 14 genes involved in OGID and 611 genes in regions identified in GWASs to be associated with height ($p = 6.84 \times 10^{-8}$), suggesting that a common variation impacting function of genes involved in OGID influences height at a population level. Increased cellular growth is a hallmark of cancer and there was striking overlap between the genes involved in OGID and 260 somatically mutated cancer driver genes ($p = 1.75 \times 10^{-14}$). However, the mutation spectra of genes involved in OGID and cancer differ, suggesting complex genotype-phenotype relationships. These data reveal insights into the genetic control of human growth and demonstrate that exome sequencing in OGID has a high diagnostic yield.

Introduction

Human growth control, at the organismal and cellular level, is a complex process essential for health and dysregulated in many developmental disorders and cancers. The mechanistic control of cell size and proliferation has been studied, by diverse approaches, in many different species.[1,2] However, the control of overall size of an organism has been relatively understudied and is still poorly understood. The study of human growth disorders therefore not only improves diagnosis and management of human disease, it also offers an opportunity to enhance knowledge about the fundamental processes governing control of human size.

Human overgrowth syndromes are a nebulous group of conditions defined as having height and/or head circumference ≥ 2 SD above the mean, together with additional phenotypic abnormalities, the most common of which is intellectual disability.[3] Overgrowth syndromes usually occur sporadically within a family and can be caused by several different mechanisms, including gene mutations, imprinting disruption, and chromosome dosage abnormalities.[3,4]

Single-gene disorders associated with overgrowth and intellectual disability (OGID) are well recognized; Sotos syndrome (MIM: 117550) and Weaver syndrome (MIM: 277590) are prototypic examples, due to *NSD1* (MIM: 606681) and *EZH2* (MIM: 601573) mutations, respectively (see GeneReviews by Tatton-Brown et al. in Web Resources).[5] OGID syndromes have been increasingly identified over the last decade.[3,4] The advent of next-generation sequencing has been the foremost reason for this progress and has allowed elucidation of the genetic causes of clinically established syndromes and the delineation of new syndromes.[5–12]

Despite these advances, many individuals with OGID remain without a genetic diagnosis. In addition, the relative contribution of the different genes to OGID is unknown. To better characterize the genetic landscape of OGID, we have here studied 710 affected individuals including 323 parent-proband trios (Table S1).

[1]Division of Genetics and Epidemiology, Institute of Cancer Research, 15 Cotswold Road, London SM2 5NG, UK; [2]South West Thames Regional Genetics Service, St George's University Hospitals NHS Foundation Trust, London SW17 0QT, UK; [3]Child Neurology Unit, Foundation IRCCS C Besta Neurological Institute, Milan 20133, Italy; [4]Landeskrankenanstalten Salzburg, Kinderklinik Department of Pediatrics, Klinische Genetik, Salzburg 5020, Austria; [5]Northern Ireland Regional Genetics Service, Belfast City Hospital, Belfast BT9 7AB, Northern Ireland; [6]Human Development and Health Academic Unit, Faculty of Medicine, University of Southampton, Southampton SO17 1BJ, UK; [7]Wessex Clinical Genetics Service, University Hospital Southampton NHS Trust, Southampton SO16 6YD, UK; [8]West Midlands Regional Genetics Service, Birmingham Women's Hospital NHS Foundation Trust and University of Birmingham, Birmingham Health Partners, Birmingham B15 2TG, UK; [9]Cancer Genetics Unit, Royal Marsden NHS Foundation Trust, London SW3 6JJ, UK
[10]Twitter: @rahman_nazneen
*Correspondence: rahmanlab@icr.ac.uk
http://dx.doi.org/10.1016/j.ajhg.2017.03.010.

Subjects and Methods

Subjects

We recruited participants through the Childhood Overgrowth (COG) Study, which began recruitment in 2005, approved by the London Multicenter Ethics Committee (05/MRE02/17). Informed consent was obtained from all participants and/or parents, as appropriate. Individuals were eligible for this study if they had height and/or head circumference at least two standard deviations above the mean ($\geq +2$ SD, UK90 growth data)[13] at some point in childhood, together with intellectual disability. We have termed this condition OGID (overgrowth + intellectual disability). Overgrowth phenotypes that are not associated with intellectual disability, such as Beckwith Wiedemann syndome (MIM: 130650) or Marfan syndrome (MIM: 154700), were not included. Regional or asymmetric overgrowth phenotypes (e.g., hemihypertrophy) in the absence of increased height or head circumference were not included.

710 individuals with OGID were included. 97% (693) were recruited to the study from clinical genetics departments. For 323 individuals, samples from both parents were also available and included. 205 probands had both height and head circumference $\geq +2$ SD, termed "head+height" in Table S1. 138 had height $\geq +2$ SD with OFC <2 SD, termed "height only" and 109 had OFC $\geq +2$ SD and height <2 SD, termed "head only." For the remaining 258 individuals, the child was recruited to the study because they had overgrowth, but measurements for both height and head were not provided. The overgrowth category is termed "unspecified" for these case subjects in Table S1. Intellectual disability was classified by the referring clinician as severe (77 case subjects), moderate (228 case subjects), or mild (229 case subjects). The referrer did not state the severity of the OGID for 176 individuals (termed "unspecified" in Table S1).

Control Data

We used the Exome Aggregation Consortium (ExAC) data v.3 accessed on 13/11/2015 (excluding the TCGA samples)[14] and the ICR1000 UK exome series[15] as reference data. We generated and analyzed the ICR1000 UK exome series data using the same sequencing and analysis pipeline described for the OGID samples.

Targeted Gene Analyses

We previously reported mutations in *NSD1*, *EZH2*, *DNMT3A* (MIM: 602769), and *PPP2R5D* (MIM: 601646) in 198 case subjects. The relevant references are in Table S1. Intragenic mutations in these genes were detected with Sanger sequencing. *NSD1* is unusual among the 14 OGID genes included in this study in being prone to deletion by a 2 Mb 5q35 microdeletion, mediated by flanking low-copy repeats.[16] We used MLPA to identify 5q35 microdeletions encompassing *NSD1*.[17] *NSD1* MLPA is also capable of detecting exon CNVs that account for ~5% of *NSD1* mutations.[17] Microdeletions and exon CNVs in the other genes were not sought, but are unlikely to be a major contributor because the surrounding sequence architecture and/or mechanism of pathogenicity make it much less likely that such events will cause OGID.

Exome Sequencing

We performed exome sequencing in all probands in whom no mutation had been identified by targeted gene analyses and in parental samples where available. We performed exome sequencing using the Nextera Rapid Capture Exome Kit (Illumina). We prepared libraries from 50 ng genomic DNA using the Nextera DNA Sample Preparation Kit (Illumina). On average 33M reads mapped to the pull-down and 86% of targeted bases had $\geq 15\times$ coverage. The captured libraries were PCR amplified using the supplied paired-end PCR primers. Exome sequencing in 57 samples was performed before the Nextera Exome Kit was available using the TruSeq Exome Enrichment Kit, which includes the 14 genes involved in OGID. When converting our exome pipeline from TruSeq to Nextera, we undertook in-house evaluation and validation to ensure that the performance was equivalent. Sequencing was performed on an llumina HiSeq 2000 or HiSeq 2500 (high output mode) using v3 chemistry and generating 2×101 bp reads.

Variant Calling

We used the OpEx v1.0 pipeline to perform variant calling.[18] We converted raw data to FASTQs using CASAVA v.1.8.2 with default settings. The OpEx v1.0 pipeline uses Stampy[19] to map to the human reference genome, Picard to flag duplicates, Platypus[20] to call variants, and CAVA[21] to provide consistent annotation of variants with the HGVS-compliant CSN (Clinical Sequencing Notation) standard v1.0.[21] The transcript information for variant annotation for the 14 relevant genes are given in Table 1.

Variant Prioritization and Validation

We excluded variants with MAF $> 0.5\%$ in either the Exome Aggregation Consortium (ExAC) and/or the ICR1000 UK exome series. For the de novo analyses, we identified and validated any high-quality (as defined by OpEx[18]) variant in the child that was not present in either parent. We evaluated and validated all rare variants identified in the 14 genes.

We confirmed all small variants in Table S1 that were called in exomes via Sanger sequencing of M13-tagged PCR products generated from genomic DNA. We performed PCR using the QIAGEN Multiplex PCR Kit according to the manufacturer's instructions. We sequenced PCR products using M13 sequencing primers, the BigDye Terminator Cycle Sequencing Kit, and an ABI 3730 Genetic Analyzer (Applied Biosystems). We analyzed sequences using Mutation Surveyor software v.3.20 (SoftGenetics)

Table 1. Gene and Transcript Information for 14 Genes Involved in OGID

Gene	MIM Number	HGNC ID	Ensembl Transcript	RefSeq Transcript
AKT3	611223	HGNC:393	ENST00000366539	NM_005465
BRWD3	300553	HGNC:17342	ENST00000373275	NM_153252
CHD8	610528	HGNC:20153	ENST00000399982	NM_001170629
DNMT3A	602769	HGNC:2978	ENST00000264709	NM_175629
EED	605984	HGNC:3188	ENST00000263360	NM_003797
EZH2	601573	HGNC:3527	ENST00000320356	NM_004456
GPC3	300037	HGNC:4451	ENST00000370818	NM_004484
HIST1H1E	142220	HGNC:4718	ENST00000304218	NM_005321
MTOR	601231	HGNC:3942	ENST00000361445	NM_004958
NFIX	164005	HGNC:7788	ENST00000360105	NM_002501
NSD1	606681	HGNC:14234	ENST00000439151	NM_022455
PIK3CA	171834	HGNC:8975	ENST00000263967	NM_006218
PPP2R5D	601646	HGNC:9312	ENST00000485511	NM_006245
PTEN	601728	HGNC:9588	ENST00000371953	NM_000314

and verified the outputs by manual inspection by two individuals, independently.

Pathogenic Mutation Determination

Apart from *HIST1H1E* (MIM: 142220), we considered a variant in the other 13 genes to be pathogenic if it fulfilled one or more of the following criteria. (1) It was a de novo mutation in a gene for which such de novo mutations were already proven to cause OGID. (2) The inheritance was unknown, because parental samples were unavailable, but it had been previously identified as a pathogenic de novo mutation in OGID. (3) It was a protein-truncating variant ([PTV] frameshifting indels, stop-gain, or essential splice-site variants) in a gene in which truncating mutations have been proven to be pathogenic. (4) There was clear evidence from the literature that it was pathogenic. The evidence for *HIST1H1E* mutations being pathogenic is provided in the Results.

HIST1H1E Statistical Analyses

We used the methods described in the DDD study[22] to calculate the probability of identifying four de novo frameshift mutations in *HIST1H1E* using the gene-specific mutation rates from Samocha et al.[23] The frameshift mutation rate in *HIST1H1E* (4.18×10^{-7}) was multiplied by twice the number of case subjects in this study (710) in order to get the expected number of frameshift mutations. We calculated the probability of observing four or more de novo frameshift mutations in *HIST1H1E* given the expected number of frameshift mutations via the ppois function in R.

We modeled the significance of mutation clustering in *HIST1H1E* under a binomial distribution where the probability of observing a mutation in a 12 bp region, which comprises 1.8% of the coding sequence, was 0.018.

Protein Net Charge Calculation

We obtained wild-type *HIST1H1E* cDNA (frame 1) sequence from Ensembl (ENST00000304218.5). We generated the *HIST1H1E* cDNA sequences edited with OGID mutations (frame 2). We used the variant c.430delG to generate the other possible alternative reading frame in *HIST1H1E* (frame 3). We translated the cDNA sequences using the Translate Tool at ExPASy. We calculated the net charge of the carboxy-terminal domain, from p.Lys110 onward, at neutral pH using the Peptide Property Calculator at the Innovagen website.

Functional Network Analyses

We performed functional enrichment analysis using g:Profiler (v.r1665_e85_eg32).[24] We used the 14 genes in Table 1 as our query set. We looked for enrichment among Gene Ontology molecular function terms and KEGG pathway gene sets, requiring the size of the functional category to be between 5 and 500 genes and using the Benjamini-Hochberg false discovery rate as the significance threshold. The FDR q values presented are the Benjamini-Hochberg critical values.

Phenotypic Analyses

We tested for significant difference in the diagnostic yields between different phenotypic groups using the prop.test function in R. We calculated the significance of association between an individual having macrocephaly and their mutation status (either a mutation in a PI3K/AKT pathway gene or a mutation in an epigenetic regulation gene) using a Fisher's exact test, which we implemented with the

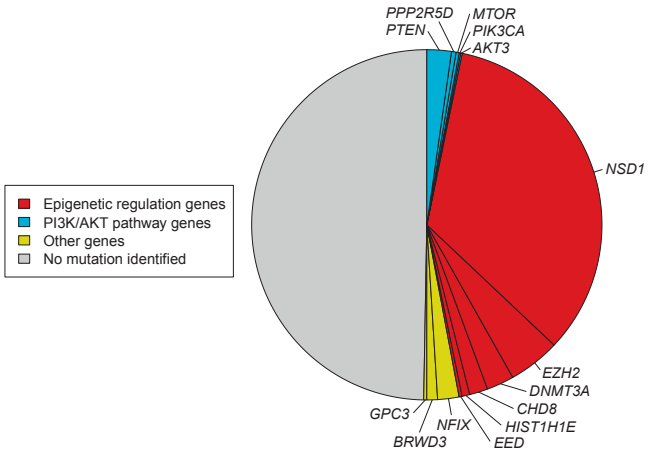


Figure 1. Causal Mutation Identified in 50% of OGID Probands
Proportion of pathogenic mutations identified in 710 individuals with OGID. Epigenetic regulation genes (red), including *NSD1* which is the predominant gene, constitute the major gene set. PI3K/AKT pathway genes (blue) also significantly contribute to OGID.

fisher.test function in R. We calculated the significance of association between an individual having macrocephaly in the absence of increased height and their mutation status, and the significance of association between an individual having increased height in the absence of macrocephaly and their mutation status in the same way. We tested for significant difference in the proportion of individuals with mild intellectual disability for those with a mutation in a PI3K/AKT pathway OGID gene and those with a mutation in an epigenetic regulation OGID gene using the prop.test function in R.

Height GWAS Gene and Cancer Driver Gene Comparisons

We obtained the list of 611 genes located in regions associated with human height through GWASs from Table S1 of Wood et al.[25] We obtained a list of 260 somatically mutated cancer genes from Table S2 of Lawrence et al.[26] and the somatic mutations from the tumor portal website.

We calculated the probability of seeing the observed overlap of the OGID gene set with the GWAS gene set under a hypergeometric probability distribution assuming a total hypothetical size of 20,000 protein-coding genes in the exome using the phyper function in R. We calculated the probability of seeing the observed overlap of OGID gene set with the cancer driver gene set in the same way.

Results

Contribution of Gene Mutations to OGID

Using exome or targeted gene analyses, we identified a pathogenic mutation in one of 14 genes in 357 individuals with OGID, giving a diagnostic yield of 50% (Figure 1). By far the most common cause was a mutation in *NSD1* (240 cases, 34%), followed by *EZH2* (34, 4.8%), *DNMT3A* (18, 2.5%), *PTEN* (MIM: 601728) (16, 2.3%), *NFIX* (MIM: 164005) (14, 2.0%), *CHD8* (MIM: 610528) (12, 1.7%), *BRWD3* (MIM: 300553) (7, 1.0%), *HIST1H1E* (5, 0.7%), *PPP2R5D* (3, 0.4%), (2 cases each) *EED* (MIM: 605984), *GPC3* (MIM: 300037), and *MTOR* (MIM: 601231), and (1 case each) *AKT3* (MIM: 611223) and *PIK3CA* (MIM: 171834) (Table S1). Among the 323 parent-proband trios, we identified a cause in 191 (59%) of which 179 were de novo mutations and 12 were inherited.

Our data allow confirmation that *EED* mutations cause OGID. Two case reports of individuals with a characteristic phenotype that includes overgrowth have been published.[10,27] We here present two additional cases with a de novo *EED* mutation. The individuals have the same facial phenotype to each other and to previously reported case subjects, with long, narrow palpebral fissures, telecanthus, and retrognathia. Notably, EED is a direct binding partner of EZH2,[28] which has an established role in causing OGID.[29] Some role in overgrowth was either known, or has been proposed, for the remainder of these, apart from *HIST1H1E* (see GeneReviews by Eng in Web Resources).[6,9,10,12,29–35]

HIST1H1E Mutations Cause OGID

We present here data showing that certain *HIST1H1E* mutations cause OGID. Through exome sequencing we identified five unrelated probands—COG0405, COG0412, COG0552, COG1739, and COG1832—with heterozygous *HIST1H1E* protein truncating variants (PTVs) (Figure 2, Tables 1 and S1). In four probands the PTV had arisen de novo. Parental samples were not available for the fifth child, but she carried the same mutation as one of the children with a de novo mutation. The detection of four de novo *HIST1H1E* mutations in 710 individuals is highly unlikely to have occurred by chance, as determined from gene-specific de novo mutation rates ($p = 5.17 \times 10^{-15}$). None of the mutations are present in the ExAC dataset, nor in 11,677 exomes analyzed in-house with similar pipelines. These results strongly support *HIST1H1E* mutations as a cause of OGID.

HIST1H1E encodes histone H1.4. In humans, H1.4 is one of 11 H1 linker histones that mediate the formation of higher-order chromatin structures and regulate the accessibility of regulatory proteins, chromatin remodelling factors, and histone-modifying enzymes to their target sites.[36,37] The five mutations we identified cluster significantly ($p = 2.0 \times 10^{-9}$) to a 12-bp region in the carboxy-terminal domain (CTD) that is involved in chromatin binding and protein-protein interactions (Figure 2A).[36] PTVs in the intronless histones have been shown to evade nonsense-mediated mRNA decay.[38] Thus the OGID-causing mutations are predicted to generate a truncated product.

The CTD of linker histones regulate higher-order chromatin structure through neutralization of negatively

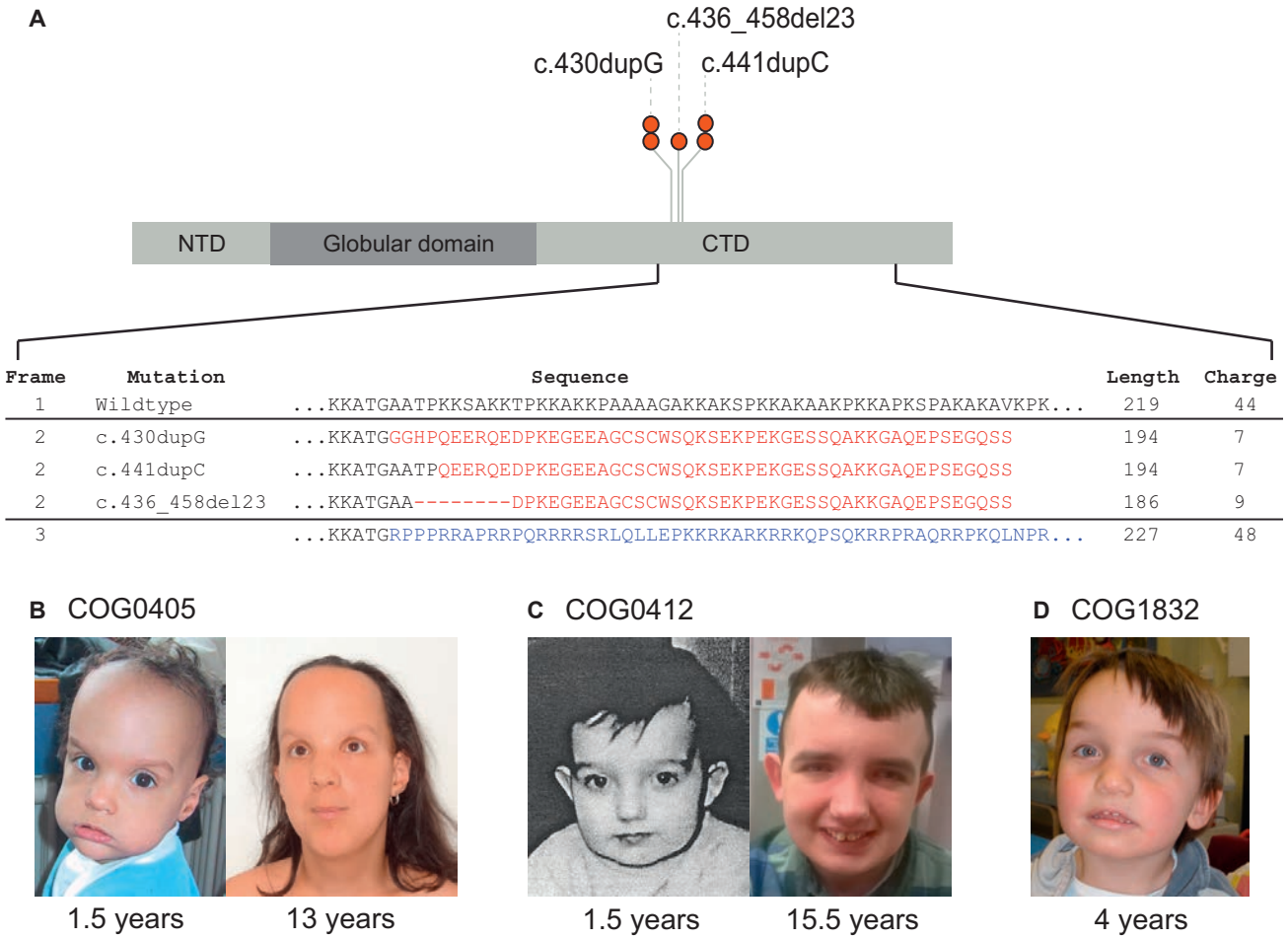


Figure 2. *HIST1H1E* Mutations Cause OGID
(A) *HIST1H1E* mutations cluster within 12 bp region in the carboxy-terminal domain (CTD) and have a similar predicted impact on protein function. The three different frameshift mutations generate the same open reading frame (frame 2), which is predicted to reduce the length and net charge (at pH 7) of the CTD compared to the wild-type (frame 1). The other possible alternate reading frame (frame 3) increases the protein length and net charge. Abbreviations: CTD, carboxy-terminal domain; NTD, amino-terminal domain.
(B–D) Facial images of three individuals with *HIST1H1E* mutations showing full cheeks and high hairline.

charged linker DNA.[36] The pathogenic *HIST1H1E* mutations all result in the same shift in the reading frame and are predicted to generate similar truncated proteins, with a reduced net charge of 7–9 (compared to 44 for the wild-type protein) (Figure 2A). The mutant protein is thus likely to be less effective in neutralizing negatively charged linker DNA. Moreover, the truncation of the C-terminus likely impedes DNA binding and protein-protein interactions. It is also noteworthy that the other possible alteration in reading frame would reduce neither the net charge nor the length of the protein (Figure 2A). Taken together, these data suggest that specific *HIST1H1E* mutations, restricted in position and type, cause human overgrowth.

HIST1H1E Clinical Phenotype

Individuals with *HIST1H1E* mutations had similar facial appearance in childhood with full cheeks, high hairline, and telecanthus (Figures 2B–2D). Height, head circumference, and degree of intellectual disability were variable, as were the additional clinical features. It is currently unclear whether these additional features are *HIST1H1E* associations or coincidental findings. Individual case descriptions are below.

COG0405, a female individual, was born at term with a weight of 3.58 kg (+0.1 SD) and a length of 53 cm (+1.5 SD). She was floppy in the neonatal period. A brain MRI scan at 4 months demonstrated mild ventricular dilatation but no other abnormalities. Her bone age at chronological age of 7 months was advanced to 18–24 months. By 19 months, her length was 87 cm (+2.0 SD) with a weight of 13.4 kg (+1.8 SD) and she had developed a strabismus. At 13 years of age, the individual was noted to have normal growth with a height of 150.8 cm (−0.6 SD), a head circumference of 55.8 cm (−0.5 SD), and a weight of 48.85 kg (+0.4 SD). She has developed a severe kyphoscoliosis for which she required surgery and has a mild intellectual disability.

COG0412, a male individual, was born at 1 week after term following an uncomplicated pregnancy and delivery. He weighed 4.75 kg (+2.4 SD). In the neonatal period he was noted to be floppy; he had poor feeding and undescended testes. At 1.5 years he was very tall at 105 cm (+8.3 SD) with a weight of 18.8 kg (+4.6 SD) and a head circumference of 52.5 cm (+2.6 SD). He was reported to have multiple nevi and redundant skin on the palms of his hands. He had a moderate intellectual disability and no behavioral issues at that time. When he was reviewed at 15.5 years, he was no longer tall with a height of 166.5 cm (−0.6 SD). His head circumference was 58.7 cm (+1.4 SD). By this age he had developed an anxiety disorder that was refractory to medical treatment. He had also developed phobias. In addition, he had major dental problems with crumbling teeth and he had dry, flaky nails.

COG0552, a female individual, was born at term with a weight of 4.79 kg (+2.5 SD) and length of 57 cm (+3.6 SD). She was floppy in the neonatal period with poor feeding. She developed no new medical problems in childhood. At the age of 4.2 years she was reported to be delayed in her development. She had a height of 108 cm (+1.2 SD), head circumference of 55 cm (+3.2 SD), and weight of 24 kg (+2.7 SD).

COG1739, a female individual, was initially thought clinically to have Weaver syndrome. She was born at 37 weeks after an uncomplicated pregnancy and labor with a weight of 3.25 kg (+0.8 SD), length of 49 cm (+0.7 SD), and head circumference of 37 cm (+3.3 SD). She was hypoglycemic and hypertonic in the neonatal period, and was also noted to have camptodactyly. At 1.9 years she was diagnosed with a moderate intellectual disability and had a height of 85 cm (mean), head circumference of 51 cm (+1.8 SD), and weight of 12 kg (−0.3 SD).

COG1832, a male individual, was born at 1 week after term weighing 3.74 kg (+0.4 SD). The pregnancy had been complicated by exposure to chicken pox. At birth, COG1832 was noted to have talipes equinovarus and later in the neonatal period was diagnosed with delayed visual maturation. A brain MRI scan showed a slender corpus callosum and unusual ventricular outline, possibly indicative of a periventricular leukomalacia. At 8.5 years, height was 133.2 cm (+0.5 SD) with a weight of 33 kg (+1.2 SD). The head circumference at 6.3 years was 59 cm (+3.7 SD). He has limited speech but with verbal comprehension markedly ahead of this ability to express himself. He has left amblyopia and astigmatism. His hearing is normal. He suffers from constipation. At times his behavior is challenging.

Functional Network Analyses

To investigate the biological processes abrogated by OGID pathogenic mutations, we performed functional enrichment analysis using the GO molecular function terms and KEGG pathway gene sets in g:Profiler.[24] The chromatin binding (FDR q value = 1.58×10^{-6}) and PI3K/AKT signaling pathway (FDR q value = 6.80×10^{-5}) gene sets were significantly enriched.

Six genes—*NSD1*, *EZH2*, *DNMT3A*, *EED*, *CHD8*, and *HIST1H1E*—were in the chromatin binding gene set. All encode proteins involved in epigenetic regulation (Figure 3A). NSD1 is a histone methyltransferase that catalyzes methylation of H3K36, and to lesser extent H4K20, and is primarily associated with transcriptional activation.[39] EZH2 and EED are key components of the polycomb repressive complex 2 (PRC2), which catalyzes methylation of H3K27, resulting in transcriptional repression of target genes.[28] DNMT3A is a DNA methyltransferase crucial for the establishment of new methylation marks during early embryogenesis and the sex-dependent methylation of imprinted genes.[40,41] *CHD8* encodes an ATP-dependent chromatin remodeler that binds to methylated H3K4, a key histone modification at active promoters.[35] As noted above, H1.4 binds to linker DNA between nucleosomes and has key roles in chromatin compaction and regulation of gene expression.[37] Together, mutations in these six genes accounted for 311 (44%) of our series. Disruption of epigenetic regulation is therefore a prominent molecular mechanism underlying OGID (Figure 1).

Five of the genes—*PTEN*, *AKT3*, *PIK3CA* (which encodes p110α, the catalytic domain of the heterodimeric PI3K lipid kinase), *MTOR*, and *PPP2R5D* (which encodes B56δ a regulatory subunit of the heterotrimeric PP2A protein phosphatase)—are in the PI3K/AKT pathway, which plays a key role in the regulation of growth (Figure 3B). Activation of the PI3K/AKT pathway results in cellular growth promotion through increased cell metabolism, cell survival, cell turnover, and protein synthesis.[42] Together mutations in these genes made only a minor contribution to our OGID series (23 case subjects, 3.2%). In part this is because individuals with mutations in these genes are more often diagnosed with other conditions, such as Cowden syndrome (MIM: 158350), megalencephaly-capillary malformation syndrome (MIM: 602501), or regional overgrowth (see GeneReviews by Eng in Web Resources).[34]

The remaining three genes—*NFIX*, *GPC3*, and *BRWD3*—encode a transcription factor, a proteoglycan, and a bromodomain-containing protein, respectively[6,31,32] (23 case subjects, 3.2%). There is currently no clear functional link between these genes and the other genes we report here. However, it is possible that *BRWD3* mutations also cause overgrowth through epigenetic regulation dysfunction, as there are data suggesting it is involved in histone H3.3 regulation.[43]

Phenotype Analyses

There was enrichment of mutations in individuals with both increased height and head circumference, compared to individuals in whom only one growth parameter was increased, as would be expected. Specifically the diagnostic yield in individuals with both macrocephaly and increased height was 59% (120/205), significantly higher than the

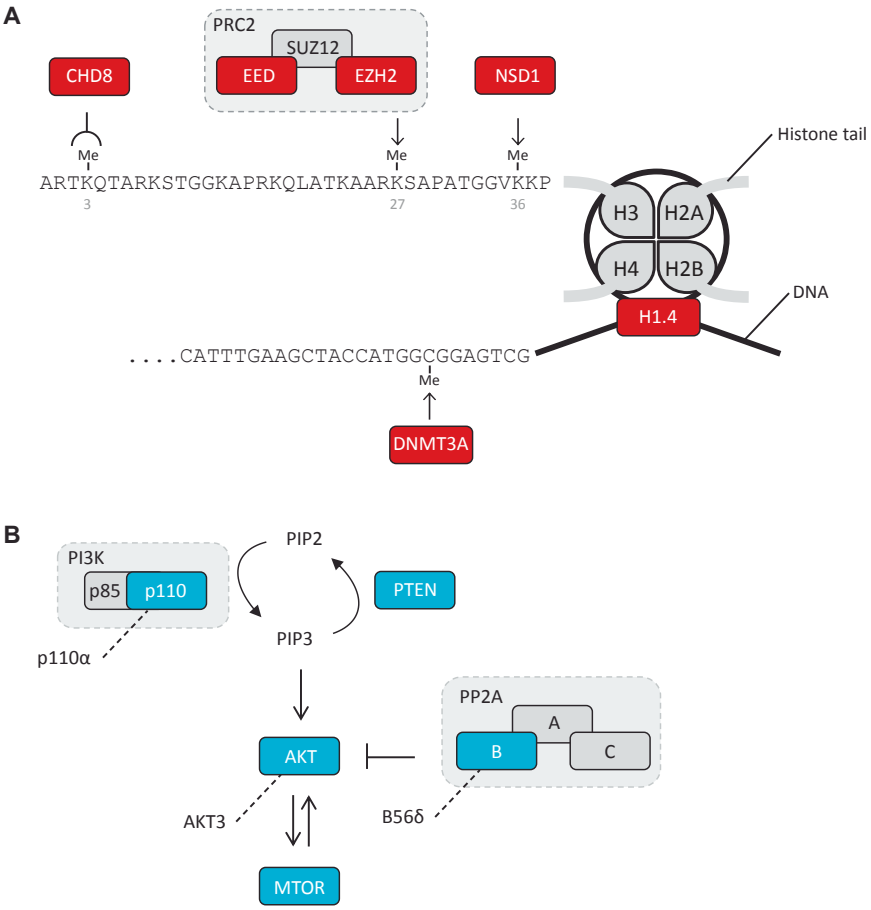


Figure 3. Schematic of Key Biological Processes Impacted in OGID
(A) Epigenetic regulation. NSD1, EED, and EZH2 directly methylate specific histone tail lysine residues. DNMT3A is a de novo DNA methyltransferase and CHD8 is a chromatin remodeling complex protein that binds methylated lysine 4 of histone H3. H1.4 (encoded by *HIST1H1E*) stabilizes higher-order chromatin structures. (B) All OGID mutations are predicted to lead to reduced function PI3K/AKT pathway. The PI3K/AKT pathway positively regulates growth. AKT3, MTOR, and p110α (encoded by *PIK3CA*) are pathway activators. PTEN and B56δ (encoded by *PPP2R5D*) are pathway suppressors. OGID mutations in *AKT3*, *MTOR*, and *PIK3CA* are activating, whereas OGID mutations in *PTEN* and *PPP2R5D* are inactivating.

diagnostic yields in individuals with only macrocephaly (43%, 47/109, $p = 0.006$) or only increased height (45%, 62/138, $p = 0.009$). There was no significant difference between the diagnostic yields in individuals with only macrocephaly and in those with only increased height ($p = 0.146$). There was also no significant difference between the diagnostic yield in individuals with unspecified growth parameters (50%, 130/258) and any other group.

To further explore the phenotypic spectrum of OGID, we compared the growth and intellectual disability severity of the individuals due to mutations in the epigenetic regulation genes and those involved in the PI3K/AKT pathway, using case subjects for which the relevant phenotypic information was available (217 individuals with complete growth data and 263 individuals with intellectual disability severity information) (Figure 4). Macrocephaly (i.e., head circumference $\geq$2 SD above the mean) occurred more frequently in individuals with PI3K/AKT pathway gene mutations; all 17 had macrocephaly, compared with 140/200 individuals with OGID due to epigenetic regulation gene mutations ($p = 4.1 \times 10^{-3}$; Figure 4A). Furthermore, 9/17 of the PI3K/AKT pathway case subjects had macrocephaly without increased height compared with 32/200 of the epigenetic regulation pathway cases ($p = 1.0 \times 10^{-3}$; Figure 4A). The remaining 60/200 had increased height without macrocephaly, a combination not present in OGID due to PI3K/AKT pathway gene mutations ($p = 4.1 \times 10^{-3}$; Figure 4A).

Varying severity of intellectual disability was a feature of both groups, but mild intellectual disability was more common in OGID due to PI3K/AKT pathway gene mutations (14/20) than OGID due to epigenetic regulation gene mutations (101/243; $p = 0.01$) (Figure 4B).

The risk of childhood cancer is one the most controversial areas of OGID management. 8/710 OGID-affected individuals in this study developed cancer in childhood (Table S1). This includes 4/357 with an identified genetic cause, three of whom had an *EZH2* mutation. COG1724 developed neuroblastoma at 46 months, COG0285 developed T cell non-hodgkins lymphoma at 13 years, and COG1521 was diagnosed with both neuroblastoma and acute lymphoblastic leukemia at 13 months. The childhood cancer incidence for *EZH2* mutation carriers in this study was thus 9% (3/34). The remaining child had an *NSD1* microdeletion and T cell non-hodgkins lymphoma. This information will be useful in family discussions about childhood cancer risk, particularly in relation to surveillance strategies, which are generally of unproven benefit and can be associated with appreciable false positive rates.[44]

Height GWAS Loci Comparative Analyses

We next explored the overlap between the 14 genes and 611 genes implicated through genome-wide association studies (GWASs) to be involved in the control of human height.[25] There was significant overlap; six genes involved in OGID were also present in height GWAS regions ($p = 6.8 \times 10^{-8}$) (Figure S1). The overlap is primarily through the epigenetic regulation genes, all of which (except *EED*) were represented in height GWAS regions. Two separate intronic SNPs in each of *NSD1* and *DNMT3A* were independently associated with height in the GWAS and

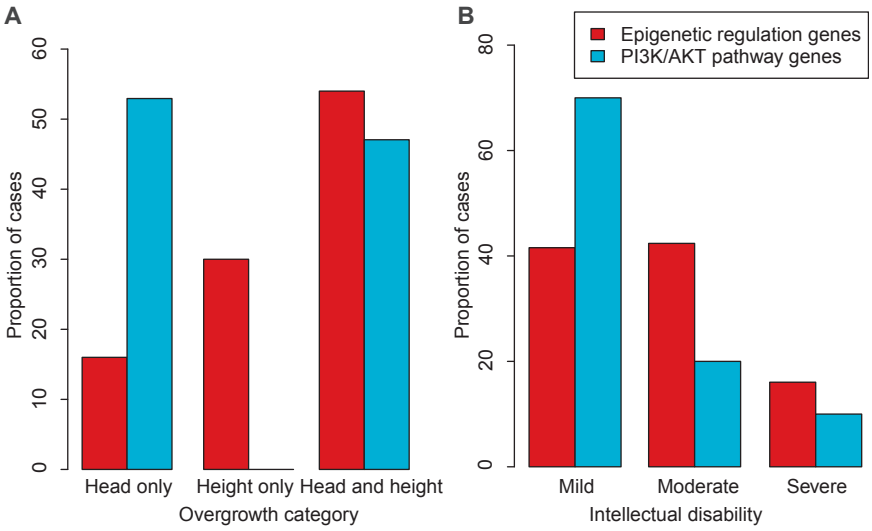


Figure 4. Phenotypic Differences between OGID due to Mutations in Epigenetic Regulation Genes Compared to PI3K/AKT Pathway Genes

Comparison of the distribution of (A) overgrowth categories and (B) degree of intellectual disability in case subjects with epigenetic regulation gene mutations (red) compared with PI3K/AKT pathway gene mutations (blue).

there were no other genes within the linkage disequilibrium (LD) blocks of association. This strongly suggests that NSD1 and DNMT3A functional impact underlie the height association in these regions (Figure S1). Single SNPs in intron 5 of *CHD8*, intron 9 of *MTOR*, 1 kb downstream of *HIST1H1E*, and 48 kb upstream of *EZH2* were also associated with height.[25] For *HIST1H1E* and *EZH2*, there were no other genes in the LD block of association. For *MTOR* the variant associated with a *cis*-eQTL affecting *MTOR* expression, though the association was better accounted for by an upstream variant (rs2295080) in the *MTOR* promoter region that was in LD with the height SNP (LD $r^2 = 0.85$).[25] Although the causal SNPs and mechanisms of association are not fully elucidated, these data suggest that common variation in some genes involved in OGID also influence height at a population level.

Cancer Somatic Driver Mutation Comparative Analyses

Dysregulated cellular growth is a hallmark of cancer, and certain human conditions are associated with both overgrowth and increased cancer risk (see GeneReviews by Eng in Web Resources).[45] We therefore next sought to investigate the overlap between the 14 genes and 260 somatically mutated cancer driver genes reported by Lawrence et al.[26] There was significant overlap; 8/14 genes involved in OGID were somatically mutated in a diverse range of cancers (*NSD1*, *EZH2*, *DNMT3A*, *PTEN*, *CHD8*, *HIST1H1E*, *MTOR*, *PIK3CA*; $p = 1.7 \times 10^{-14}$). For the PI3K/AKT pathway genes, the mutation spectra are similar in OGID and cancer.[34] By contrast, for the epigenetic regulation genes, the mutation spectra in OGID and cancer have substantial, distinctive differences.

Somatic mutations in *HIST1H1E*, *EZH2*, and *DNMT3A* occur in hematological malignancies.[26,46–50] *HIST1H1E* and *EZH2* mutations are each present in ~20% of B cell lymphomas.[48,49] Somatic *HIST1H1E* mutations are nonsynonymous mutations throughout the gene and do not include the clustered PTVs that cause OGID (Figure 5). *EZH2* mutations in B cell lymphomas are often activating nonsynonymous mutations in the SET domain, the majority of which target a single amino acid, p.Tyr646.[48] Nonsynonymous mutations at this residue have not been detected in OGID and are not present in ExAC, perhaps suggesting that germline *EZH2* mutations altering p.Tyr646 are not compatible with life (Figure 5). Inactivating *EZH2* mutations are present in myeloid malignancies and in T-ALL.[46–48] A proportion of these latter mutations overlap with *EZH2* mutations in OGID.

DNMT3A is one of the most frequently mutated genes in AML and mutations also occur less frequently in other hematological malignancies.[26,50] The majority target a single residue, p.Arg882, with the remainder being nonsynonymous variants and PTVs scattered through the gene. Mutations at p.Arg882 have not thus far been reported in OGID (Figure 5). Protein modeling suggests that the somatic mutations primarily impact DNA binding, whereas the mutations in OGID are more likely to impact histone binding.[12]

Somatic *NSD1* mutations are seen in ~10% of head and neck squamous cell carcinomas[26,51] and somatic *CHD8* mutations are present in ~3% of glioblastoma multiforme (GBM).[26] For these cancers the mutation pattern is similar to that observed in OGID, with PTVs being the most frequent mutation type (Figure 5).[30] Interestingly, Lawrence et al. found *NSD1* and *CHD8* to each be significant in their pan-cancer analysis, present in 2% of cancers.[26] However, the pan-cancer mutation spectra for each gene was different to that observed in OGID, with most being nonsynonymous mutations scattered throughout the gene (Figure 5).

Discussion

We present here the largest genetic study of overgrowth and intellectual disability performed to date, including 710 affected individuals and 636 parents. We show that OGID is a highly heterogeneous condition, involving at least 14 genes. Perturbation of epigenetic regulation is a prominent mechanism causing OGID and can be caused by mutations in at least six different genes. *NSD1* mutation is by far the most frequent cause of OGID, accounting for

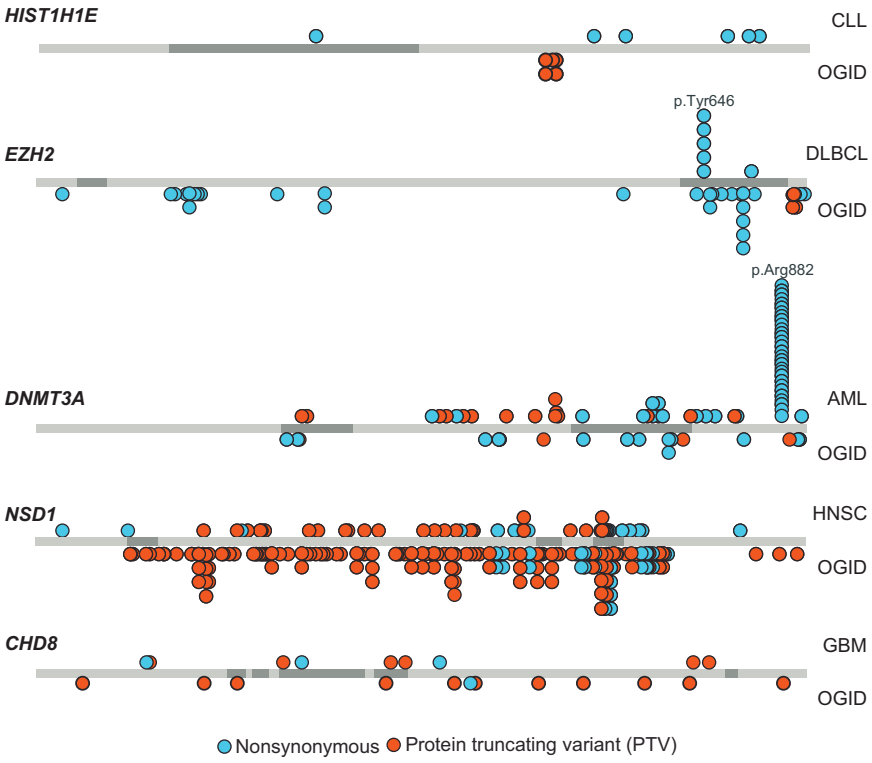


Figure 5. Mutations in Epigenetic Regulation Genes in OGID and Cancers
Protein schematics showing the position of mutations in *HIST1H1E*, *EZH2*, *DNMT3A*, *NSD1*, and *CHD8* in OGID (below the gene) and specific cancers (above the gene). The somatic cancer driver mutations are from Lawrence et al.[26] Abbreviations are as follows: AML, acute myeloid leukemia; CLL, chronic lymphocytic leukemia; DLBCL, diffuse large B-cell lymphoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; OGID, overgrowth-intellectual disability.

240 (34%) of our series. Notably, *NSD1* is within a 2 Mb region flanked by low-copy repeats that mediate a microdeletion, which is one of the commonest causes of Sotos syndrome[16] and was present in 29 individuals. Furthermore, exon deletions or duplications (exon CNVs) are reported in ~5% of case subjects[17] and were present in 9 individuals. We analyzed *NSD1* for these types of mutations, using MLPA, as they are not robustly identifiable in our exome data. We did not examine the other genes for microdeletions or exon CNVs. However, they are not known to be a major contributor to pathogenic mutations in the other genes. Even after excluding microdeletions and exon CNVs, *NSD1* is still the most common cause of OGID, accounting for 202 (28%) of our series.

The comparative analyses of genes involved in OGID with GWAS height loci and with cancer driver genes highlight intriguing similarities and differences. Our data strongly suggest that common variation impacting epigenetic regulation of gene function influences height at a population level. Further investigation of these GWAS loci would be of considerable interest, particularly in relation to advancing knowledge on how, and why, epigenetic regulation dysfunction impacts human growth.

Several genes involved in OGID are somatically mutated in a diverse range of cancers, but the spectra of mutations, particularly in the epigenetic regulation genes, is different in OGID and cancer. The underlying reasons for these differences will be complex and may include embryonic lethality of certain oncogenic mutations when they occur in the germline. Integration of germline and somatic mutational data in future research will be useful, and will likely advance functional and mechanistic understanding of the genes.

One of the most striking results of this study is the high diagnostic yield of genetic testing in OGID; a genetic cause was identified in 50% (357/710) of case subjects. This is likely to be an underestimate as we have been conservative in attributing pathogenicity to OGID gene variants and additional OGID genes almost certainly exist. Indeed, among the 132 trios in whom a definitive cause was not found, a de novo mutation possibly associated with their phenotype was present in 28; for example, two had de novo nonsynonymous variants in *XRN1*.

The diagnostic yield in our OGID series is higher than exome-sequencing studies in other phenotypes that include intellectual disability, which ranged from 13% to 35%.[22,52–56] The studies are not directly comparable, as most other exome studies included case subjects in which prior genetic testing was negative. Our study recruitment started prior to the discovery and clinical testing of most of the genes we report here, which allows us to provide a much better estimate of the overall contribution of rare gene mutations to this phenotype.

Given the high success rate, strong consideration should be given to using exome sequencing as a first-line diagnostic test in OGID. Height and head circumference can be easily measured and intellectual disability is readily diagnosable. Therefore, implementation of exome sequencing in OGID should be straightforward. Gene testing would provide important diagnostic and recurrence risk information to many families. Furthermore, it would increase genotype-phenotype data, which are urgently required to improve prognostic information. Of equal importance, exome sequencing in OGID would lead to the identification of new genes and new mutations in known genes. In turn, this will stimulate and facilitate scientific research, enhancing knowledge of basic biological processes controlling growth and the diverse pathologies in which human growth control is dysfunctional.

Supplemental Data

Supplemental Data include one figure, one table, and Supplemental Note and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2017.03.010.

Acknowledgments

We thank the families for their participation and the clinicians that recruited them. The full list of collaborators is in the Supplemental Note. We are grateful to Margaret Warren-Perry for assistance in recruitment. We are grateful to Sandra Hanks, Silvana Powell, Imran Uddin, and Ann Strydom for technical and administrative support and Tara Mills for assistance with the GWAS analyses. We acknowledge support from the NIHR RM/ICR Biomedical Research Centre and Wessex NIHR clinical research network. K.T.-B. is supported by funding from the Child Growth Foundation (GR01/13). This work was supported by Wellcome Trust Award 100210/Z/12/Z.

Received: December 7, 2016
Accepted: March 24, 2017
Published: April 27, 2017

Web Resources

ExAC Browser, http://exac.broadinstitute.org/
ExPASy Translate Tool, http://web.expasy.org/translate/
g:Profiler, http://biit.cs.ut.ee/gprofiler/
GenBank, http://www.ncbi.nlm.nih.gov/genbank/
GeneReviews, Eng, C. (1993). PTEN hamartoma tumor syndrome. https://www.ncbi.nlm.nih.gov/books/NBK1488/
GeneReviews, Tatton-Brown, K., Cole, T.R.P., and Rahman, N. (1993). Sotos syndrome. https://www.ncbi.nlm.nih.gov/books/NBK1479/
ICR1000 UK Exome Series, http://www.icr.ac.uk/icr1000exomes
OMIM, http://www.omim.org/
OpEx NGS Pipeline, http://www.icr.ac.uk/opex
Picard, http://broadinstitute.github.io/picard/
Protein calculator, http://pepcalc.com/protein-calculator.php
TumorPortal, http://www.tumorportal.org/

References

1. Stocker, H., and Hafen, E. (2000). Genetic control of cell size. Curr. Opin. Genet. Dev. *10*, 529–535.
2. Saucedo, L.J., and Edgar, B.A. (2002). Why size matters: altering cell size. Curr. Opin. Genet. Dev. *12*, 565–571.
3. Tatton-Brown, K., and Weksberg, R. (2013). Molecular mechanisms of childhood overgrowth. Am. J. Med. Genet. C. Semin. Med. Genet. *163C*, 71–75.
4. Edmondson, A.C., and Kalish, J.M. (2015). Overgrowth syndromes. J. Pediatr. Genet. *4*, 136–143.
5. Tatton-Brown, K., Murray, A., Hanks, S., Douglas, J., Armstrong, R., Banka, S., Bird, L.M., Clericuzio, C.L., Cormier-Daire, V., Cushing, T., et al.; Childhood Overgrowth Consortium (2013). Weaver syndrome and EZH2 mutations: Clarifying the clinical phenotype. Am. J. Med. Genet. A. *161A*, 2972–2980.
6. Malan, V., Rajan, D., Thomas, S., Shaw, A.C., Louis Dit Picard, H., Layet, V., Till, M., van Haeringen, A., Mortier, G., Nampoothiri, S., et al. (2010). Distinct effects of allelic NFIX mutations on nonsense-mediated mRNA decay engender either a Sotos-like or a Marshall-Smith syndrome. Am. J. Hum. Genet. *87*, 189–198.
7. Gibson, W.T., Hood, R.L., Zhan, S.H., Bulman, D.E., Fejes, A.P., Moore, R., Mungall, A.J., Eydoux, P., Babul-Hirji, R., An, J., et al.; FORGE Canada Consortium (2012). Mutations in EZH2 cause Weaver syndrome. Am. J. Hum. Genet. *90*, 110–118.
8. Cordeddu, V., Redeker, B., Stellacci, E., Jongejan, A., Fragale, A., Bradley, T.E., Anselmi, M., Ciolfi, A., Cecchetti, S., Muto, V., et al. (2014). Mutations in ZBTB20 cause Primrose syndrome. Nat. Genet. *46*, 815–817.
9. Loveday, C., Tatton-Brown, K., Clarke, M., Westwood, I., Renwick, A., Ramsay, E., Nemeth, A., Campbell, J., Joss, S., Gardner, M., et al.; Childhood Overgrowth Collaboration (2015). Mutations in the PP2A regulatory subunit B family genes PPP2R5B, PPP2R5C and PPP2R5D cause human overgrowth. Hum. Mol. Genet. *24*, 4775–4779.
10. Cohen, A.S., and Gibson, W.T. (2016). EED-associated overgrowth in a second male patient. J. Hum. Genet. *61*, 831–834.
11. Baynam, G., Overkov, A., Davis, M., Mina, K., Schofield, L., Allcock, R., Laing, N., Cook, M., Dawkins, H., and Goldblatt, J. (2015). A germline MTOR mutation in Aboriginal Australian siblings with intellectual disability, dysmorphism, macrocephaly, and small thoraces. Am. J. Med. Genet. A. *167*, 1659–1667.
12. Tatton-Brown, K., Seal, S., Ruark, E., Harmer, J., Ramsay, E., Del Vecchio Duarte, S., Zachariou, A., Hanks, S., O'Brien, E., Aksglaede, L., et al.; Childhood Overgrowth Consortium (2014). Mutations in the DNA methyltransferase gene DNMT3A cause an overgrowth syndrome with intellectual disability. Nat. Genet. *46*, 385–388.
13. Freeman, J.V., Cole, T.J., Chinn, S., Jones, P.R., White, E.M., and Preece, M.A. (1995). Cross sectional stature and weight reference curves for the UK, 1990. Arch. Dis. Child. *73*, 17–24.
14. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
15. Ruark, E., Münz, M., Renwick, A., Clarke, M., Ramsay, E., Hanks, S., Mahamdallie, S., Elliott, A., Seal, S., Strydom, A., et al. (2015). The ICR1000 UK exome series: a resource of gene variation in an outbred population. F1000Res. *4*, 883.
16. Kurotaki, N., Stankiewicz, P., Wakui, K., Niikawa, N., and Lupski, J.R. (2005). Sotos syndrome common deletion is mediated by directly oriented subunits within inverted Sos-REP low-copy repeats. Hum. Mol. Genet. *14*, 535–542.
17. Douglas, J., Tatton-Brown, K., Coleman, K., Guerrero, S., Berg, J., Cole, T.R., Fitzpatrick, D., Gillerot, Y., Hughes, H.E., Pilz, D., et al. (2005). Partial NSD1 deletions cause 5% of Sotos syndrome and are readily identifiable by multiplex ligation dependent probe amplification. J. Med. Genet. *42*, e56.
18. Ruark, E., Münz, M., Clarke, M., Renwick, A., Ramsay, E., Elliott, A., Seal, S., Lunter, G., and Rahman, N. (2016). OpEx - a validated, automated pipeline optimised for clinical exome sequence analysis. Sci. Rep. *6*, 31029.
19. Lunter, G., and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. *21*, 936–939.
20. Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S.R., Wilkie, A.O., McVean, G., Lunter, G.; and WGS500

Consortium (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat. Genet. *46*, 912–918.
21. Münz, M., Ruark, E., Renwick, A., Ramsay, E., Clarke, M., Mahamdallie, S., Cloke, V., Seal, S., Strydom, A., Lunter, G., and Rahman, N. (2015). CSN and CAVA: variant annotation tools for rapid, robust next-generation sequencing analysis in the clinical setting. Genome Med. *7*, 76.
22. Deciphering Developmental Disorders Study (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature *519*, 223–228.
23. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat. Genet. *46*, 944–950.
24. Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., and Vilo, J. (2016). g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. *44* (W1), W83–W89.
25. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al.; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; and LifeLines Cohort Study (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.
26. Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501.
27. Cohen, A.S., Tuysuz, B., Shen, Y., Bhalla, S.K., Jones, S.J., and Gibson, W.T. (2015). A novel mutation in EED associated with overgrowth. J. Hum. Genet. *60*, 339–342.
28. Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. (2002). Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science *298*, 1039–1043.
29. Tatton-Brown, K., Hanks, S., Ruark, E., Zachariou, A., Duarte, Sdel.V., Ramsay, E., Snape, K., Murray, A., Perdeaux, E.R., Seal, S., et al.; Childhood Overgrowth Collaboration (2011). Germline mutations in the oncogene EZH2 cause Weaver syndrome and increased human height. Oncotarget *2*, 1127–1133.
30. Tatton-Brown, K., Douglas, J., Coleman, K., Baujat, G., Cole, T.R., Das, S., Horn, D., Hughes, H.E., Temple, I.K., Faravelli, F., et al.; Childhood Overgrowth Collaboration (2005). Genotype-phenotype associations in Sotos syndrome: an analysis of 266 individuals with NSD1 aberrations. Am. J. Hum. Genet. *77*, 193–204.
31. Field, M., Tarpey, P.S., Smith, R., Edkins, S., O'Meara, S., Stevens, C., Tofts, C., Teague, J., Butler, A., Dicks, E., et al. (2007). Mutations in the BRWD3 gene cause X-linked mental retardation associated with macrocephaly. Am. J. Hum. Genet. *81*, 367–374.
32. Cottereau, E., Mortemousque, I., Moizard, M.P., Bürglen, L., Lacombe, D., Gilbert-Dussardier, B., Sigaudy, S., Boute, O., David, A., Faivre, L., et al. (2013). Phenotypic spectrum of Simpson-Golabi-Behmel syndrome in a series of 42 cases with a mutation in GPC3 and review of the literature. Am. J. Med. Genet. C. Semin. Med. Genet. *163C*, 92–105.
33. Saxena, A., and Sampson, J.R. (2014). Phenotypes associated with inherited and developmental somatic mutations in genes encoding mTOR pathway components. Semin. Cell Dev. Biol. *36*, 140–146.
34. Mirzaa, G., Timms, A.E., Conti, V., Boyle, E.A., Girisha, K.M., Martin, B., Kircher, M., Olds, C., Juusola, J., Collins, S., et al. (2016). PIK3CA-associated developmental disorders exhibit distinct classes of mutations with variable expression and tissue distribution. JCI insight *1*.
35. Barnard, R.A., Pomaville, M.B., and O'Roak, B.J. (2015). Mutations and modeling of the chromatin Remodeler CHD8 define an emerging autism etiology. Front. Neurosci. *9*, 477.
36. Harshman, S.W., Young, N.L., Parthun, M.R., and Freitas, M.A. (2013). H1 histones: current perspectives and challenges. Nucleic Acids Res. *41*, 9593–9609.
37. Kalashnikova, A.A., Rogge, R.A., and Hansen, J.C. (2016). Linker histone H1 and protein-protein interactions. Biochim. Biophys. Acta *1859*, 455–461.
38. Maquat, L.E., and Li, X. (2001). Mammalian heat shock p70 and histone H4 transcripts, which derive from naturally intronless genes, are immune to nonsense-mediated decay. RNA *7*, 445–456.
39. Qiao, Q., Li, Y., Chen, Z., Wang, M., Reinberg, D., and Xu, R.M. (2011). The structure of NSD1 reveals an autoregulatory mechanism underlying histone H3K36 methylation. J. Biol. Chem. *286*, 8361–8368.
40. Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247–257.
41. Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, N., Li, E., and Sasaki, H. (2004). Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature *429*, 900–903.
42. Engelman, J.A., Luo, J., and Cantley, L.C. (2006). The evolution of phosphatidylinositol 3-kinases as regulators of growth and metabolism. Nat. Rev. Genet. *7*, 606–619.
43. Chen, W.Y., Shih, H.T., Liu, K.Y., Shih, Z.S., Chen, L.K., Tsai, T.H., Chen, M.J., Liu, H., Tan, B.C., Chen, C.Y., et al. (2015). Intellectual disability-associated dBRWD3 regulates gene expression through inhibition of HIRA/YEM-mediated chromatin deposition of histone H3.3. EMBO Rep. *16*, 528–538.
44. Katanoda, K. (2016). Neuroblastoma mass screening–what can we learn from it? J. Epidemiol. *26*, 163–165.
45. Lapunzina, P. (2005). Risk of tumorigenesis in overgrowth syndromes: a comprehensive review. Am. J. Med. Genet. C. Semin. Med. Genet. *137C*, 53–71.
46. Ernst, T., Chase, A.J., Score, J., Hidalgo-Curtis, C.E., Bryant, C., Jones, A.V., Waghorn, K., Zoi, K., Ross, F.M., Reiter, A., et al. (2010). Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. Nat. Genet. *42*, 722–726.
47. Ntziachristos, P., Tsirigos, A., Van Vlierberghe, P., Nedjic, J., Trimarchi, T., Flaherty, M.S., Ferres-Marco, D., da Ros, V., Tang, Z., Siegle, J., et al. (2012). Genetic inactivation of the polycomb repressive complex 2 in T cell acute lymphoblastic leukemia. Nat. Med. *18*, 298–301.
48. Bödör, C., Grossmann, V., Popov, N., Okosun, J., O'Riain, C., Tan, K., Marzec, J., Araf, S., Wang, J., Lee, A.M., et al. (2013). EZH2 mutations are frequent and represent an early event in follicular lymphoma. Blood *122*, 3165–3168.
49. Okosun, J., Bödör, C., Wang, J., Araf, S., Yang, C.Y., Pan, C., Boller, S., Cittaro, D., Bozek, M., Iqbal, S., et al. (2014).

Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. Nat. Genet. *46*, 176–181.
50. Yang, L., Rau, R., and Goodell, M.A. (2015). DNMT3A in haematological malignancies. Nat. Rev. Cancer *15*, 152–165.
51. Cancer Genome Atlas Network (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature *517*, 576–582.
52. Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet *380*, 1674–1682.
53. de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N. Engl. J. Med. *367*, 1921–1929.
54. Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature *511*, 344–347.
55. Vissers, L.E., Gilissen, C., and Veltman, J.A. (2016). Genetic studies in intellectual disability and related disorders. Nat. Rev. Genet. *17*, 9–18.
56. Martinez, F., Caro-Llopis, A., Rosello, M., Oltra, S., Mayo, S., Monfort, S., and Orellana, C. (2017). High diagnostic yield of syndromic intellectual disability by targeted next-generation sequencing. J. Med. Genet. *54*, 87–92.

Whole-Exome Sequencing Identifies Loci Associated with Blood Cell Traits and Reveals a Role for Alternative *GFI1B* Splice Variants in Human Hematopoiesis

Linda M. Polfus,[1,38] Rajiv K. Khajuria,[2,3,4,38] Ursula M. Schick,[5,38] Nathan Pankratz,[6] Raha Pazoki,[7] Jennifer A. Brody,[8] Ming-Huei Chen,[9] Paul L. Auer,[10] James S. Floyd,[8] Jie Huang,[11] Leslie Lange,[12] Frank J.A. van Rooij,[7] Richard A. Gibbs,[13] Ginger Metcalf,[13] Donna Muzny,[13] Narayanan Veeraraghavan,[13] Klaudia Walter,[11] Lu Chen,[11,14] Lisa Yanek,[15]

(Author list continued on next page)

Circulating blood cell counts and indices are important indicators of hematopoietic function and a number of clinical parameters, such as blood oxygen-carrying capacity, inflammation, and hemostasis. By performing whole-exome sequence association analyses of hematologic quantitative traits in 15,459 community-dwelling individuals, followed by in silico replication in up to 52,024 independent samples, we identified two previously undescribed coding variants associated with lower platelet count: a common missense variant in *CPS1* (rs1047891, MAF = 0.33, discovery + replication $p = 6.38 \times 10^{-10}$) and a rare synonymous variant in *GFI1B* (rs150813342, MAF = 0.009, discovery + replication $p = 1.79 \times 10^{-27}$). By performing CRISPR/Cas9 genome editing in hematopoietic cell lines and follow-up targeted knockdown experiments in primary human hematopoietic stem and progenitor cells, we demonstrate an alternative splicing mechanism by which the *GFI1B* rs150813342 variant suppresses formation of a GFI1B isoform that preferentially promotes megakaryocyte differentiation and platelet production. These results demonstrate how unbiased studies of natural variation in blood cell traits can provide insight into the regulation of human hematopoiesis.

Human genetic studies have provided important insights into hematopoiesis. Genome-wide association studies (GWASs) performed in large, population-based samples have identified associations of genomic regions and common genetic (usually non-coding) variants with inter-individual differences in blood cell traits[1–5], though the causal DNA variants and their functional mechanisms often remain elusive. Whole-exome and targeted sequencing approaches have been used to identify rare, sometimes private, loss (or gain)-of-function coding variants segregating within families with hematologic traits at the extremes of the phenotypic distribution[6–12]. As of yet, whole-exome sequencing has not been applied to large population-based cohorts well-phenotyped for hematologic traits to identify rare, functional variation with moderate-to-large phenotypic effects and to provide new biologic insight.

To this end, we performed exome sequencing in 15,459 unrelated European ancestry (EU) and African American (AA) individuals enrolled in six population-based cohort studies (see Supplemental Note). Replication of significant findings was performed in up to 52,024 additional samples via a combination of whole-exome-based or genome-based sequencing, genotyping, and imputation (Supplemental Note). Our a priori hypothesis was that systematic evaluation of coding variation detected by exome sequence analysis in samples unselected for blood cell traits would identify low-frequency variants influencing hematologic traits and could provide functional insights into hematopoiesis. We analyzed platelet count and 12 other blood cell traits (Table S1). The means of the traits were as expected in a sample of unselected healthy individuals from the population (Table S1). Association results from single-variant and from gene-based burden and sequence kernel association tests (SKATs) meeting our a priori significance thresholds in either EU, AA, or trans-ethnic discovery meta-analyses are summarized for both previously known and novel (which we define as those not reported in the available literature) loci in Table 1 and Tables S2–S5 and described further in the Supplemental Note. Lambda values showed no significant inflation (Table S6).

[1]Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [2]Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA,; [3]Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [4]Berlin-Brandenburg School for Regenerative Therapies, Charité Universitätsmedizin Berlin, Berlin 13353, Germany; [5]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [6]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN 55454, USA; [7]Department of Epidemiology, Erasmus University Medical Center, Rotterdam 3000, the Netherlands; [8]Cardiovascular Health Research Unit and Department of Medicine, University of Washington, Seattle, WA 98195, USA; [9]Department of Neurology, School of Medicine, Boston University, Boston, MA 02118, USA; [10]School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI 53205, USA; [11]Human Genetics, Wellcome Trust Sanger Institute, Hinxton CB10 1HH, UK; [12]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA; [13]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA; [14]Department of Haematology, University of Cambridge, Cambridge CB2 0AH, UK; [15]GeneSTAR Research Program, Division of General Internal Medicine, Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA; [16]Center for Human Genetic Research,

(Affiliations continued on next page)

Lewis C. Becker,[15] Gina M. Peloso,[16] Aoi Wakabayashi,[2,3] Mart Kals,[17] Andres Metspalu,[17] Tõnu Esko,[17] Keolu Fox,[18] Robert Wallace,[19] Nora Franceshini,[20] Nena Matijevic,[21] Kenneth M. Rice,[8] Traci M. Bartz,[8] Leo-Pekka Lyytikäinen,[22] Mika Kähönen,[23] Terho Lehtimäki,[22] Olli T. Raitakari,[24] Ruifang Li-Gao,[25] Dennis O. Mook-Kanamori,[25,26] Guillaume Lettre,[27] Cornelia M. van Duijn,[28] Oscar H. Franco,[7] Stephen S. Rich,[29] Fernando Rivadeneira,[28] Albert Hofman,[28] André G. Uitterlinden,[28] James G. Wilson,[30] Bruce M. Psaty,[8,31] Nicole Soranzo,[11,14] Abbas Dehghan,[7] Eric Boerwinkle,[1] Xiaoling Zhang,[32] Andrew D. Johnson,[33] Christopher J. O'Donnell,[34] Jill M. Johnsen,[35] Alexander P. Reiner,[36,39] Santhi K. Ganesh,[37,39] and Vijay G. Sankaran[2,3,39,*]

Four gene-based associations were discovered for red blood cell (RBC) traits (*ACTN4*, *MMACHC*, *MYOM2*, and *MRPL43*). Trans-ethnic discovery meta-analyses are summarized for both previously identified loci, which we verify in this study, and previously unreported loci. A summary of these findings, and driving variants, are provided in the Supplemental Note and Table S3. None of these gene-based SKAT or burden findings could be replicated in independent samples. Nonetheless, a few of the individual rare variants driving the gene-based associations in the discovery sample showed suggestive evidence of association in the replication sample (Supplemental Note and Table S3**)**.

Among the three single-variant associations we identified (Table 1), two coding variants were associated with lower platelet count in our discovery sample: *CPS1* rs1047891, a common missense variant encoding p.Thr1412Asn (EU + AA minor-allele frequency [MAF] = 0.33, EU + AA $p = 5.7 \times 10^{-8}$) and *GFI1B* rs150813342, a rare synonymous variant encoding p.Phe192 and located in alternatively spliced exon 5 (EU MAF = 0.009, EU $p= 4.7 \times 10^{-8}$; EU + AA MAF = 0.008, EU + AA $p = 2.64 \times 10^{-8}$). One single-nucleotide variant (SNV) result (rs9656446; EU + AA MAF = 0.03, EU + AA $p = 1.48 \times 10^{-7}$) associated with basophils in trans-ethnic analyses was in the ATP/GTP binding protein-like 3 (*AGBL3*) gene. However, the allele frequencies in the discovery sample differed by ethnicity (EU MAF = 0.001 and AA MAF = 0.08), and replication in samples of EU ethnicity from the UK10K project was not significant (EU p = 0.71). In our combined replication sample, we replicated the associations of *CPS1* rs1047891 (EU + AA MAF = 0.328, EU + AA $p = 1.02 \times 10^{-4}$) and *GFI1B* rs150813342 (EU + AA $p = 5.71 \times 10^{-21}$) with lower platelet counts. In the combined discovery and replication samples, the p values for *CPS1* rs1047891 and *GFI1B* rs150813342 were 6.38×10^{-10} and 1.79×10^{-27}, respectively. A Manhattan plot for single-variant associations with platelet count and quantile-quantile (Q-Q) plots are shown in Figures S1–S3. Forest plots of the discovery cohorts for the two replicated findings (*GFI1B* rs150813342 and *CPS1* rs1047891) are provided in Figures S4 and S5, as well as regional plots calculating linkage disequilibrium of SNVs in the gene with respect to index SNVs (Figures S6 and S7).

AGBL3 is a metallocarboxypeptidase involved in processing tubulins of the blood cell cytoskeleton. *CPS1* encodes carbamoyl-phosphate synthase 1, a mitochondrial enzyme involved in the urea cycle. The *CPS1* variant (or its LD proxies) has been associated with various cardiometabolic traits, including high-density lipoprotein (HDL) cholesterol, homocysteine, fibrinogen, serum metabolite levels, and kidney function.[13–17] *GFI1B* is a known transcriptional repressor and a key regulator of platelet and red blood cell development. There was no evidence that either *CPS1* rs1047891 or *GFI1B* rs150813342 were significantly associated with other hematologic traits assessed in the discovery sample (Tables S7A and S7B). Moreover, neither *GFI1B* rs150813342 nor *CPS1* rs1047891 was associated with mean platelet volume, platelet aggregation, or expression of platelet surface markers, though these analyses were limited to much smaller numbers of individuals (Supplemental Note, Tables S8 and S10). However, a decrease in

Massachusetts General Hospital, Boston, MA 02114, USA; [17]Estonian Genome Center, University of Tartu, Tartu, 51010, Estonia; [18]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA; [19]College of Public Health, the University of Iowa, Iowa City, IA 52242, USA; [20]Department of Medicine, School of Medicine, University of North Carolina, Chapel Hill, NC 27599, USA; [21]Department of Surgery, University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [22]Department of Clinical Chemistry, Fimlab Laboratories and University of Tampere School of Medicine, Tampere 33520, Finland; [23]Department of Clinical Physiology, Tampere University Hospital and University of Tampere School of Medicine, Tampere 33521, Finland; [24]Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital and Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku 20520, Finland; [25]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden RC 2300, the Netherlands; [26]Epidemiology Section, Department of Biostatistics, Epidemiology, and Scientific Computing Department, King Faisal Specialist Hospital and Research Centre, Riyadh 11211 Saudi Arabia; [27]Montreal Heart Institute and Université de Montréal, Montreal, QC H1T 1C8, Canada; [28]Department of Internal Medicine, Erasmus University Medical Center, Rotterdam 3000, the Netherlands; [29]Center for Public Health Genomics, University of Virginia, Charlottesville VA 22908, USA; [30]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson MS 39216, USA; [31]Group Health Research Institute, Group Health Cooperative, Seattle, WA 98101, USA; [32]Departments of Medicine and Biostatistics, Schools of Medicine and Public Health, Boston University, Boston, MA 02118, USA; [33]Cardiovascular Epidemiology and Human Genomics Branch, Framingham Heart Study, National Heart, Lung, and Blood Institute, Framingham, MA 01702, USA; [34]Framingham Heart Study, National Heart, Lung, and Blood Institute, Framingham, MA 01702, USA; [35]Bloodworks Northwest, Seattle, WA 98102, USA; [36]Women's Health Initiative Clinical Coordinating Center, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; [37]Division of Cardiovascular Medicine, Departments of Internal Medicine and Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA
[38]These authors contributed equally to this work
[39]These authors contributed equally to this work
*Correspondence: sankaran@broadinstitute.org (V.G.S.)
http://dx.doi.org/10.1016/j.ajhg.2016.06.016.

Table 1. Single-Variant Association Findings

Trait	Discovery Ethnicity	Gene	SNP Chromosome Position, rs Number, and Function	Discovery p Value	Replication p Value	Discovery MAF	Replication MAF	Discovery Beta Coefficient (SE)	Replication Z Score[a]	Discovery N	Replication N
PLT	EU + AA	*GFI1B*	chr9: 135864513, rs150813342, synonymous	2.64×10^{-8}	5.71×10^{-21}	0.008	0.007	−0.402 (0.07)	−9.40	13,744	48,099[b]
PLT	EU + AA	*CPS1*	chr2: 211540507, rs1047891, missense	5.73×10^{-8}	1.02×10^{-4}	0.328	0.313	−0.07 (0.013)	−3.89	13,744	48,394[b]
BASO	EU + AA	*AGBL3*	chr7: 134717656, rs9656446, synonymous	1.48×10^{-7}	0.71	0.031[c]	0.002	0.271 (0.051)	−0.05 (0.13)	6,877	6,699[d]

AA, African American individuals; BASO, basophil count; EU, European ancestry individuals; MAF, minor-allele frequency; PLT, platelet count.
[a]Z score is reported from N-weighted replication meta-analyses, where more than one replication cohort was available; otherwise, beta coefficient and SE are reported.
[b]UK10K project samples and imputed EU, Cardiovascular Health Study (CHS), and Atherosclerosis Risk in Communities (ARIC) study samples.
[c]EU MAF = 0.001; AA MAF = 0.078; EU + AA MAF = 0.031.
[d]UK10K project samples and imputed EU samples.

the median fluorescence intensity of large, platelet-marker positive ($CD41^+CD61^+$) events[18] was detected by flow cytometry in *GFI1B* variant carriers even after adjustment for circulating platelet count ($p < 0.0001$), which could reflect a decrease in circulating platelet aggregates or a skewing of a platelet subpopulation with regards to platelet-surface-marker expression or size (see Supplemental Note).

We conducted bioinformatic and functional analyses to understand the impact of the *GFI1B* exon 5 synonymous variant and the *CPS1* rs1047891 variant (p.Thr1412Asn) on gene and protein function. The *CPS1* p.Thr1412Asn amino acid substitution is predicted to be benign and tolerated by SIFT and PolyPhen. Moreover, according to the GTEx Portal database, there is no evidence of an expression quantitative trait loci (eQTL) effect for rs1047891. Nonetheless, the CPS1 p.Thr1412Asn missense substitution is located within a region critical for N-acetyl-glutamate binding and has been reported to result in 20%–30% higher enzymatic activity[19] and to influence vascular function.[15]

We initially assessed the association of rs150813342 with *GFI1B* expression by using Affymetrix GeneChip Human Exon 1.0 ST Array data on whole-blood RNA available from 881 Framingham Heart Study participants.[20] There was no evidence for association of the rs150813342 genotype with expression of any *GFI1B* exon, though statistical power is likely limited by the low frequency of the rs150813342 variant allele, which was present in only 7 of the 881 individuals. According to SPANR,[21] rs150813342 had a predicted effect on splicing (difference in the percentage of transcripts with the exon spliced in [dPSI] score of −4.6). rs150813342 was predicted to disrupt a putative exon splicing enhancer (ESE) in exon 5 that contains a consensus SRSF1 binding motif.[22] To functionally evaluate the impact of this variant on GFI1B transcript splicing in a relevant cell type, we used CRISPR/Cas9 genome editing to create multiple independent isogenic K562 hematopoietic cell lines harboring the *GFI1B* synonymous single-nucleotide change (Figure 1A). These cell lines were homozygous for the variant and exhibited inclusion of less than 30% of exon 5 relative to other surrounding exons in the *GFI1B* mRNA (Figure 1B). Semi-quantitative RT-PCR showed that the presence of the synonymous variant resulted in reduced formation of the *GFI1B* isoform containing exon 5 (herein referred to as the long isoform), as well as preferential formation of the isoform lacking exon 5 (herein referred to as the short isoform) (Figures 1C and 1D). No other isoforms or intron inclusion events were detected (Figure 1C, Figure S8).

Although *GFI1B* has been implicated in both RBC and platelet production (erythropoiesis and megakaryopoiesis, respectively),[23–25] only a role for the short isoform in erythroid cells has been suggested previously.[26] We next assessed the effect of the altered splicing of *GFI1B* on lineage-specific hematopoietic differentiation. We chemically induced differentiation of the isogenic K562 cell lines with either hemin (to promote erythroid differentiation) or phorbol 12-myristate 13-acetate (PMA, to promote megakaryocytic differentiation) (Figure 2A). Although erythroid

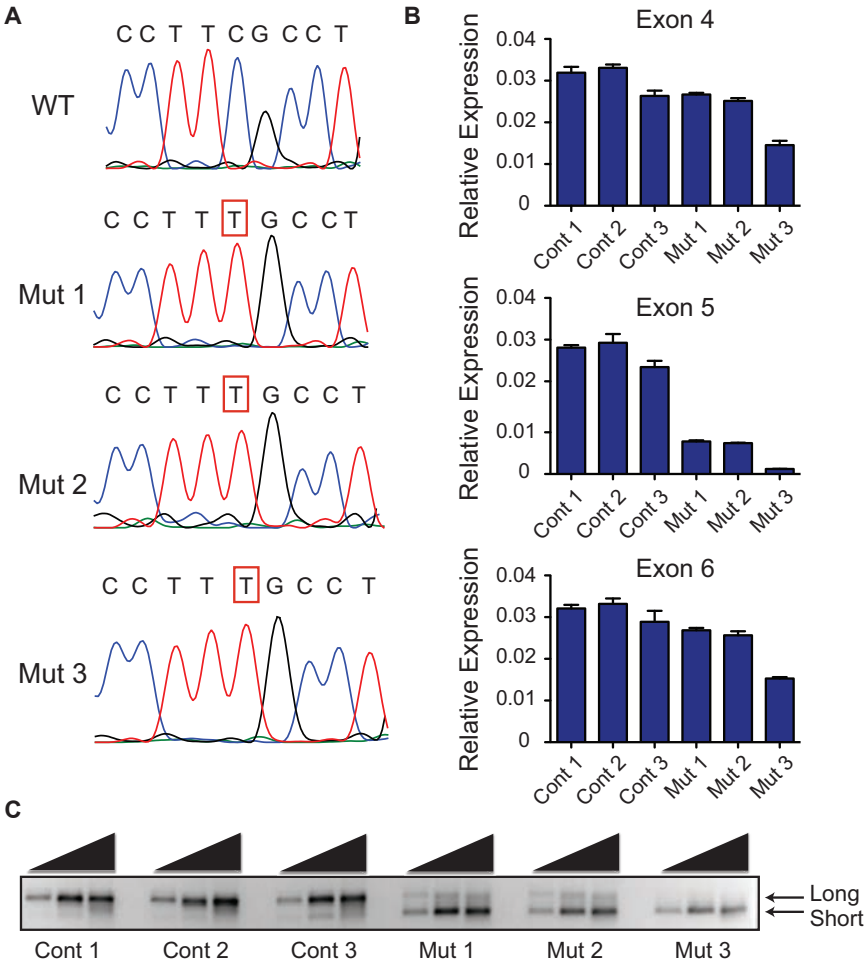


Figure 1. The Variant rs150813342 Results in Reduced Formation of the Long *GFI1B* Isoform and Preferential Formation of the Short Isoform

(A) Chromatograms of the sequence surrounding the altered nucleotide in *GFI1B* exon 5 showing the wild-type (WT) sequence and sequences of isogenic hematopoietic K562 cell mutant clones (Mut 1, Mut 2, and Mut 3) harboring the C>T single-nucleotide variant (SNV) generated via CRISPR/Cas9 mediated homologous repair.

(B) qRT-PCR of *GFI1B* exons 4, 5, and 6 measured from isogenic control (Cont) and mutant K562 cell clones showing inclusion of less than 30% of *GFI1B* exon 5 relative to the surrounding exons in *GFI1B* mRNA from mutant clones (n = 3 per group). Error bars show SD.

(C) Semi-quantitative RT-PCR with *GFI1B* exon 4 forward and exon 6 reverse primers with progressively increasing cycle numbers (26, 28, and 30 cycles) demonstrates reduced formation of the long *GFI1B* isoform and preferential formation of the short isoform, as well as no other intermediate isoforms in the clones harboring the SNV.

(D) rs150813342 is predicted to disrupt a putative exon splicing enhancer (ESE) in exon 5 that contains a consensus SRSF1 binding motif. Disruption of this binding motif results in reduced inclusion of exon 5 and preferential formation of the short isoform. The promotion of alternative splicing by SRSF1 through the spliceosome complex is indicated by an arrow to a light blue circle. Forward (F) and reverse (R) PCR primers of the respective exon are indicated.

differentiation appeared to proceed normally, as assessed morphologically (Figure 2B), and with expression of the surface marker GYPA (CD235a) (Figure 2C) and terminal erythroid marker genes (Figure 2D), megakaryocyte differentiation appeared severely impaired; the cells retained an immature blast-like morphology and failed to upregulate the surface marker of megakaryocyte differentiation, CD41a (encoded by *ITGA2B*), and mRNAs whose expression is characteristic of terminal megakaryopoiesis (Figures 2B–2D, Figure S9). The megakaryocyte genes *PPBP*, *SELP*, and *PF4* were downregulated by an average of 8.6-, 6.7-, and 41.1-fold, respectively, in the isogenic clones ($p = 0.0001$, 0.0013, and 0.0459, respectively) versus in the controls (Figure 2D). These results suggest that the long isoform of GFI1B is necessary for normal megakaryocyte differentiation.

To confirm a preferential role for this long GFI1B isoform in megakaryocyte differentiation, we identified two independent short hairpin RNAs (shRNAs) that specifically targeted *GFI1B* exon 5, which would thereby selectively downregulate the long but not the short isoform. We utilized

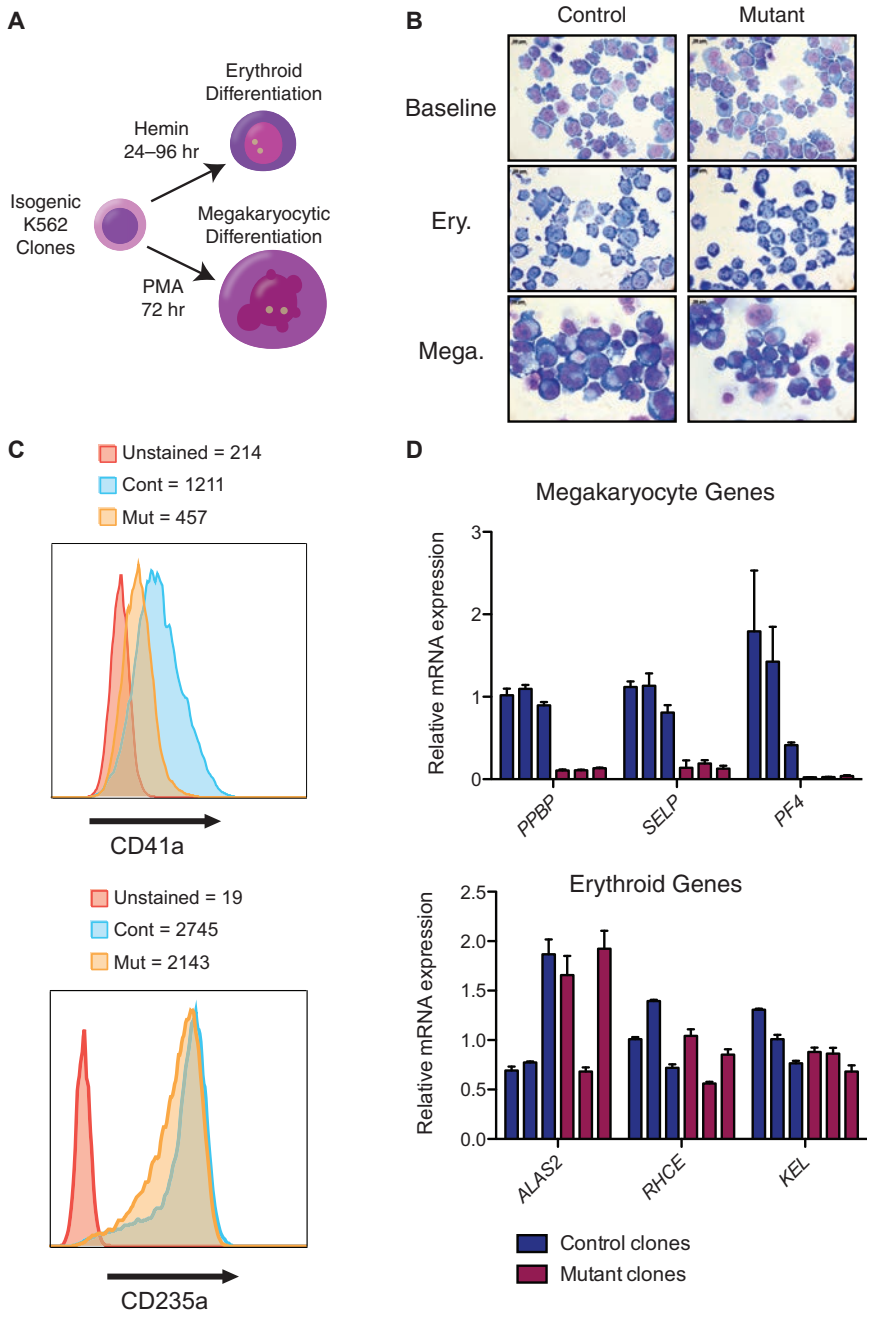


Figure 2. Impaired Megakaryopoiesis and Retained Erythropoiesis in K562 Cells Harboring the rs150813342 SNV in *GFI1B* Exon 5

(A) Scheme of phorbol 12-myristate 13-acetate (PMA)-induced megakaryocytic differentiation and hemin-induced erythroid differentiation of the hematopoietic K562 cell models.

(B) Representative May-Grünwald-Giemsa-stained cytospin images of 72 hr PMA-induced and 96 hr hemin-induced isogenic control and mutant clones showing megakaryocytic differentiation that appears severely impaired, with the cells retaining an immature blast-like morphology in the mutant clones, whereas the erythroid differentiation appears unaffected.

(C) Representative flow cytometry analysis of the megakaryocyte marker CD41a and the erythroid marker CD235a further confirmed the impaired megakaryopoiesis and the retained erythropoiesis as shown by the histogram plots with the mean fluorescence intensity (MFI) for each marker in unstained cells, control, and mutant clones, respectively.

(D) Gene expression analysis by qRT-PCR of the megakaryocyte markers *PPBP*, *SELP*, and *PF4* after 72 hr of PMA-induced differentiation and of the erythroid markers *ALAS2*, *RHCE*, and *KEL* after 24 hr of hemin-induced differentiation (n = 3 per group). Error bars show SD.

lentiviral-mediated shRNA delivery in primary human adult mobilized peripheral-blood hematopoietic stem and progenitor cells (HSPCs), which are capable of differentiation toward the erythroid and megakaryocyte lineages under appropriate culture conditions.[27] We observed a knockdown efficiency of the *GFI1B* long isoform by ~50% for both shRNAs, whereas the short isoform levels increased conversely (Figures 3A and 3B), which resulted in a 1.5- to 1.8-fold reduction in the formation of $CD41a^{+}$ megakaryocytic cells (relative to lineage-marker negative cells) in HSPCs undergoing differentiation (Figure 3C). In contrast, $CD235a^{+}$ erythroid cells appeared to be present in comparable percentages and numbers (Figure 3C). Moreover, whereas numerous morphologically mature erythroblasts could be readily visualized in both groups, fewer mature megakaryocytic cells were seen with knockdown of the long isoform than in the controls (Figure 3D, Figure S10). Overall cell growth appeared comparable between the knockdown and control cells (Figure S10). These findings are in line with our exome-sequence association findings, in which no significant effect was seen on circulating RBC levels.

GFI1B private, loss-of-function mutations (nonsense, frameshift) in the DNA-binding fifth and sixth zinc (Zn)-finger domains have recently been identified in families with an autosomal-dominant form of Gray Platelet syndrome (GPS) or related forms of thrombocytopenia, which are characterized by dysmegakaryopoiesis, thrombocytopenia, large platelets, and platelet α-granule deficiency (MIM: 187900)[28,29]. The truncating *GFI1B* mutations reported in GPS appear to have a dominant-negative effect and inhibit transcriptional activity of the *GFI1B* wild-type form. Our population study extends the allelic spectrum of naturally occurring *GFI1B* coding sequence variants associated with a lower circulating platelet count to include a more frequent, synonymous change that alters

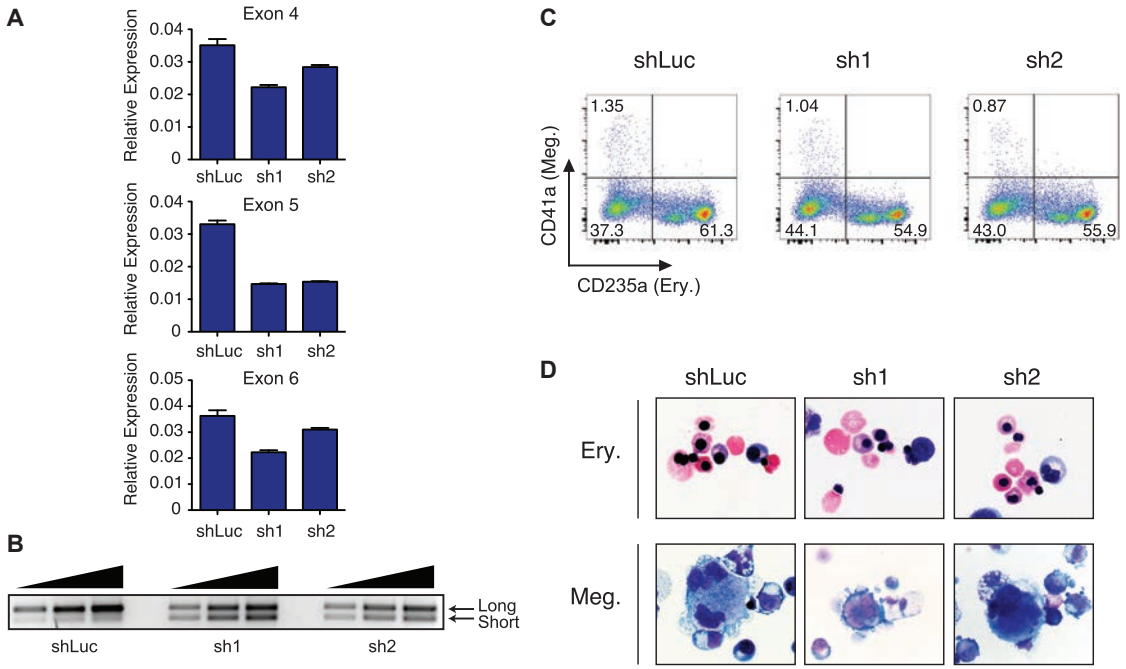


Figure 3. The Long *GFI1B* Isoform is Critical for Megakaryopoiesis in a Human Primary Cell Model
(A) qRT-PCR of *GFI1B* exons 4, 5, and 6 on day 4 after infection showing the identification of two short hairpin RNAs (shRNAs) that specifically target *GFI1B* exon 5 and thereby selectively downregulate the long isoform by ~50%, but not the short isoform (n = 3 per group). Error bars show SD.
(B) Semi-quantitative RT-PCR with *GFI1B* exon 4 forward and exon 6 reverse primers with progressively increasing cycle numbers (26, 28, and 30 cycles) demonstrates reduced formation of the long *GFI1B* isoform and increased formation of the short isoform, as well as no other intermediate isoforms in cells with targeted knockdown of GFI1B exon 5.
(C) Representative flow cytometry analysis of thrombopoietin (TPO)- and erythropoietin (EPO)-stimulated primary human hematopoietic stem and progenitor cells on day 11 of differentiation with assessment of $CD41a^{+}$ megakaryocytic (Meg) cells and $CD235a^{+}$ erythroid (Ery) cells.
(D) Representative May-Grünwald-Giemsa-stained cytospin images of megakaryocytic cells (from day 7 of differentiation) and erythroid cells (from day 13 of differentiation) showing immature megakaryocyte morphology in cells with knockdown of the long *GFI1B* isoform, in comparison with the control. In contrast, maturation of erythroblasts appears unaffected.

an exonic splicing enhancer, resulting in the skipping of exon 5, containing the first and second Zn-finger domains. Heterozygous carriers of the synonymous exon 5 variant in *GFI1B* have an average platelet count that is reduced by 25,000 to 30,000 platelets per microliter, which would be a clinically detectable effect. We also provide additional support for distinct roles of GFI1B long- and short-isoforms, which are differentially expressed at various stages of differentiation during normal hematopoiesis.[23,30] The long GFI1B isoform is expressed in HSPCs and lineage-committed myeloid, erythroid, and megakaryocytic progenitors. The abnormalities in megakaryocyte maturation with reduced formation of the *GFI1B* long isoform in the isogenic K562 cell clones containing the rs150813342 variant and in primary HSPCs with targeted suppression of the long isoform are consistent with an essential role for the GFI1B long isoform in megakaryopoiesis and platelet production. This finding is also congruent with prior work showing that the GFI1B short isoform is required for erythropoiesis[26] and provides insight into how these different splice variants function in distinct aspects of human hematopoiesis.

In summary, whole-exome sequence association analysis performed in over 15,000 samples discovered SNVs associated with a lower platelet count in community-dwelling individuals, including a common variant in *CPS1* and a rare, synonymous variant in *GFI1B*. Follow-up genome editing and targeted knockdown experiments identified a mechanism by which alternative splicing associated with the *GFI1B* rs150813342 variant allele suppresses formation of a specific GFI1B long isoform that is required for lineage-specific megakaryocyte differentiation, while being dispensable for erythropoiesis. Functional studies coupled with an association finding demonstrated a previously unappreciated splicing-based mechanism for lineage-specific blood cell production, providing important insights into human hematopoiesis. Genes regulated by the long GFI1B isoform could provide additional understanding of downstream transcriptional events and molecular pathways required for megakaryocyte specification and platelet production. These findings hold promise for the development of therapeutics for altering platelet count without adverse effects on other blood lineages. Further characterization of the role of GFI1B isoforms could have clinical or therapeutic implications for disorders of platelet and other blood cell

production or function, as well as for the prospect of improving the manufacture of ex vivo cell therapies.[31–33]

Supplemental Data

Supplemental Data include a Supplemental Note, ten figures, nine tables, and Supplemental Acknowledgments and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.06.016.

Acknowledgments

The work described in this manuscript was supported in part by NIH grants HL122684 to S.K.G. and DK103794 and HL120791 to V.G.S. N.S. is supported by the Wellcome Trust (grant codes WT098051 and WT091310), the EU 7th Framework Programme (EPIGENESYS grant code 257082 and BLUEPRINT grant code HEALTH-F5-2011-282510) and the NIH Research (NIHR) Blood and Transplant Research Unit (BTRU) in Donor Health and Genomics at the University of Cambridge in partnership with NHS Blood and Transplant (NHSBT).

Received: March 14, 2016
Accepted: June 20, 2016
Published: August 4, 2016

Web Resources:

OMIM, http://www.omim.org/

References

1. Okada, Y., and Kamatani, Y. (2012). Common genetic factors for hematological traits in humans. J. Hum. Genet. *57*, 161–169.
2. Ganesh, S.K., Zakai, N.A., van Rooij, F.J., Soranzo, N., Smith, A.V., Nalls, M.A., Chen, M.H., Kottgen, A., Glazer, N.L., Dehghan, A., et al. (2009). Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nat. Genet. *41*, 1191–1198.
3. Soranzo, N., Spector, T.D., Mangino, M., Kühnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., et al. (2009). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat. Genet. *41*, 1182–1190.
4. Auer, P.L., Teumer, A., Schick, U., O'Shaughnessy, A., Lo, K.S., Chami, N., Carlson, C., de Denus, S., Dubé, M.P., Haessler, J., et al. (2014). Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. Nat. Genet. *46*, 629–634.
5. Ulirsch, J.C., Nandakumar, S.K., Wang, L., Giani, F.C., Zhang, X., Rogov, P., Melnikov, A., McDonel, P., Do, R., Mikkelsen, T.S., and Sankaran, V.G. (2016). Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. Cell *165*, 1530–1545.
6. Shiohara, M., Shigemura, T., Saito, S., Tanaka, M., Yanagisawa, R., Sakashita, K., Asada, H., Ishii, E., Koike, K., Chin, M., et al. (2009). Ela2 mutations and clinical manifestations in familial congenital neutropenia. J. Pediatr. Hematol. Oncol. *31*, 319–324.
7. Minelli, A., Maserati, E., Rossi, G., Bernardo, M.E., De Stefano, P., Cecchini, M.P., Valli, R., Albano, V., Pierani, P., Leszl, A., et al. (2004). Familial platelet disorder with propensity to acute myelogenous leukemia: genetic heterogeneity and progression to leukemia via acquisition of clonal chromosome anomalies. Genes Chromosomes Cancer *40*, 165–171.
8. Sankaran, V.G., and Gallagher, P.G. (2013). Applications of high-throughput DNA sequencing to benign hematology. Blood *122*, 3575–3582.
9. Albers, C.A., Cvejic, A., Favier, R., Bouwmans, E.E., Alessi, M.C., Bertone, P., Jordan, G., Kettleborough, R.N., Kiddle, G., Kostadima, M., et al. (2011). Exome sequencing identifies NBEAL2 as the causative gene for gray platelet syndrome. Nat. Genet. *43*, 735–737.
10. Johnsen, J.M., Nickerson, D.A., and Reiner, A.P. (2013). Massively parallel sequencing: the new frontier of hematologic genomics. Blood *122*, 3268–3275.
11. Giani, F.C., Fiorini, C., Wakabayashi, A., Ludwig, L.S., Salem, R.M., Jobaliya, C.D., Regan, S.N., Ulirsch, J.C., Liang, G., Steinberg-Shemer, O., et al. (2016). Targeted Application of Human Genetic Variation Can Improve Red Blood Cell Production from Stem Cells. Cell Stem Cell *18*, 73–78.
12. Sankaran, V.G., Ghazvinian, R., Do, R., Thiru, P., Vergilio, J.A., Beggs, A.H., Sieff, C.A., Orkin, S.H., Nathan, D.G., Lander, E.S., and Gazda, H.T. (2012). Exome sequencing identifies GATA1 mutations resulting in Diamond-Blackfan anemia. J. Clin. Invest. *122*, 2439–2443.
13. Köttgen, A., Pattaro, C., Böger, C.A., Fuchsberger, C., Olden, M., Glazer, N.L., Parsa, A., Gao, X., Yang, Q., Smith, A.V., et al. (2010). New loci associated with kidney function and chronic kidney disease. Nat. Genet. *42*, 376–384.
14. Paré, G., Chasman, D.I., Parker, A.N., Zee, R.R., Mälarstig, A., Seedorf, U., Collins, R., Watkins, H., Hamsten, A., Miletich, J.P., and Ridker, P.M. (2009). Novel associations of CPS1, MUT, NOX4, and DPEP1 with plasma homocysteine in a healthy population: a genome-wide evaluation of 13 974 participants in the Women's Genome Health Study. Circ Cardiovasc Genet *2*, 142–150.
15. Summar, M.L., Gainer, J.V., Pretorius, M., Malave, H., Harris, S., Hall, L.D., Weisberg, A., Vaughan, D.E., Christman, B.W., and Brown, N.J. (2004). Relationship between carbamoylphosphate synthetase genotype and systemic vascular function. Hypertension *43*, 186–191.
16. Pearson, D.L., Dawling, S., Walsh, W.F., Haines, J.L., Christman, B.W., Bazyk, A., Scott, N., and Summar, M.L. (2001). Neonatal pulmonary hypertension–urea-cycle intermediates, nitric oxide production, and carbamoyl-phosphate synthetase function. N. Engl. J. Med. *344*, 1832–1838.
17. Sabater-Lleal, M., Huang, J., Chasman, D., Naitza, S., Dehghan, A., Johnson, A.D., Teumer, A., Reiner, A.P., Folkersen, L., Basu, S., et al.; VTE Consortium; STROKE Consortium; Wellcome Trust Case Control Consortium 2 (WTCCC2); C4D Consortium; CARDIoGRAM Consortium (2013). Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. Circulation *128*, 1310–1324.
18. Catellier, D.J., Aleksic, N., Folsom, A.R., and Boerwinkle, E. (2008). Atherosclerosis Risk in Communities (ARIC) Carotid MRI flow cytometry study of monocyte and platelet markers: intraindividual variability and reliability. Clin. Chem. *54*, 1363–1371.
19. Summar, M.L., Hall, L., Christman, B., Barr, F., Smith, H., Kallianpur, A., Brown, N., Yadav, M., Willis, A., Eeds, A., et al.

(2004). Environmentally determined genetic expression: clinical correlates with molecular variants of carbamyl phosphate synthetase I. Mol. Genet. Metab. *81* (*Suppl 1*), S12–S19.
20. Zhang, X., Joehanes, R., Chen, B.H., Huan, T., Ying, S., Munson, P.J., Johnson, A.D., Levy, D., and O'Donnell, C.J. (2015). Identification of common genetic variants controlling transcript isoform variation in human whole blood. Nat. Genet. *47*, 345–352.
21. Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.
22. Cho, S., Hoang, A., Sinha, R., Zhong, X.Y., Fu, X.D., Krainer, A.R., and Ghosh, G. (2011). Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly. Proc. Natl. Acad. Sci. USA *108*, 8233–8238.
23. Foudi, A., Kramer, D.J., Qin, J., Ye, D., Behlich, A.S., Mordecai, S., Preffer, F.I., Amzallag, A., Ramaswamy, S., Hochedlinger, K., et al. (2014). Distinct, strict requirements for Gfi-1b in adult bone marrow red cell and platelet generation. J. Exp. Med. *211*, 909–927.
24. Randrianarison-Huetz, V., Laurent, B., Bardet, V., Blobe, G.C., Huetz, F., and Duménil, D. (2010). Gfi-1B controls human erythroid and megakaryocytic differentiation by regulating TGF-beta signaling at the bipotent erythro-megakaryocytic progenitor stage. Blood *115*, 2784–2795.
25. Saleque, S., Cameron, S., and Orkin, S.H. (2002). The zinc-finger proto-oncogene Gfi-1b is essential for development of the erythroid and megakaryocytic lineages. Genes Dev. *16*, 301–306.
26. Laurent, B., Randrianarison-Huetz, V., Frisan, E., Andrieu-Soler, C., Soler, E., Fontenay, M., Dusanter-Fourt, I., and Duménil, D. (2012). A short Gfi-1B isoform controls erythroid differentiation by recruiting the LSD1-CoREST complex through the dimethylation of its SNAG domain. J. Cell Sci. *125*, 993–1002.
27. Ludwig, L.S., Gazda, H.T., Eng, J.C., Eichhorn, S.W., Thiru, P., Ghazvinian, R., George, T.I., Gotlib, J.R., Beggs, A.H., Sieff, C.A., et al. (2014). Altered translation of GATA1 in Diamond-Blackfan anemia. Nat. Med. *20*, 748–753.
28. Stevenson, W.S., Morel-Kopp, M.C., Chen, Q., Liang, H.P., Bromhead, C.J., Wright, S., Turakulov, R., Ng, A.P., Roberts, A.W., Bahlo, M., and Ward, C.M. (2013). GFI1B mutation causes a bleeding disorder with abnormal platelet function. J. Thromb. Haemost. *11*, 2039–2047.
29. Monteferrario, D., Bolar, N.A., Marneth, A.E., Hebeda, K.M., Bergevoet, S.M., Veenstra, H., Laros-van Gorkom, B.A., MacKenzie, M.A., Khandanpour, C., Botezatu, L., et al. (2014). A dominant-negative GFI1B mutation in the gray platelet syndrome. N. Engl. J. Med. *370*, 245–253.
30. Chen, L., Kostadima, M., Martens, J.H., Canu, G., Garcia, S.P., Turro, E., Downes, K., Macaulay, I.C., Bielczyk-Maczynska, E., Coe, S., et al.; BRIDGE Consortium (2014). Transcriptional diversity during lineage commitment of human blood progenitors. Science *345*, 1251033.
31. Vassen, L., Khandanpour, C., Ebeling, P., van der Reijden, B.A., Jansen, J.H., Mahlmann, S., Dührsen, U., and Möröy, T. (2009). Growth factor independent 1b (Gfi1b) and a new splice variant of Gfi1b are highly expressed in patients with acute and chronic leukemia. Int. J. Hematol. *89*, 422–430.
32. Koldehoff, M., Zakrzewski, J.L., Beelen, D.W., and Elmaagacli, A.H. (2013). Additive antileukemia effects by GFI1B- and BCR-ABL-specific siRNA in advanced phase chronic myeloid leukemic cells. Cancer Gene Ther. *20*, 421–427.
33. Thon, J.N., Medvetz, D.A., Karlsson, S.M., and Italiano, J.E., Jr. (2015). Road blocks in making platelets for transfusion. J. Thromb. Haemost. *13* (*Suppl 1*), S55–S62.

Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project

Paul L. Auer,[1,2] Alex P. Reiner,[2,3] Gao Wang,[4,5] Hyun Min Kang,[6] Goncalo R. Abecasis,[6] David Altshuler,[7,8] Michael J. Bamshad,[9,10] Deborah A. Nickerson,[9] Russell P. Tracy,[11,12] Stephen S. Rich,[13] NHLBI GO Exome Sequencing Project, and Suzanne M. Leal[4,*]

Massively parallel whole-genome sequencing (WGS) data have ushered in a new era in human genetics. These data are now being used to understand the role of rare variants in complex traits and to advance the goals of precision medicine. The technological and computing advances that have enabled us to generate WGS data on thousands of individuals have also outpaced our ability to perform analyses in scientifically and statistically rigorous and thoughtful ways. The past several years have witnessed the application of whole-exome sequencing (WES) to complex traits and diseases. From our analysis of NHLBI Exome Sequencing Project (ESP) data, not only have a number of important disease and complex trait association findings emerged, but our collective experience offers some valuable lessons for WGS initiatives. These include caveats associated with generating automated pipelines for quality control and analysis of rare variants; the importance of studying minority populations; sample size requirements and efficient study designs for identifying rare-variant associations; and the significance of incidental findings in population-based genetic research. With the ESP as an example, we offer guidance and a framework on how to conduct a large-scale association study in the era of WGS.

Introduction

Early in 2015, President Obama used his State of the Union address to champion the pursuit of "precision medicine," i.e., utilizing genetic and molecular techniques to individually tailor treatments and preventive measures for chronic diseases. The head of the US NIH and National Cancer Institute swiftly followed suit, describing the over-arching goals and plans for the Precision Medicine Initiative (PMI),[1] which may eventually include whole-genome sequencing of a longitudinal cohort of 1 million or more Americans. To support the PMI at the National Heart, Lung, and Blood Institute (NHLBI), the Trans-Omics for Precision Medicine Program (TOPMed) will use WGS data along with molecular, environmental, and clinical data to investigate the etiology of heart, lung, blood, and sleep disorders. As the PMI and TOPMed programs are launched, many tens of thousands of whole-genome sequences will be generated, providing researchers with access to genetic data on a scale of unprecedented size and complexity.

Our ability to thoughtfully analyze and interpret large-scale WGS data will be a significant scientific and computational challenge. However, there are a number of recently completed, large-scale, population-based sequencing studies that offer empirical guidance. One such study, the NHLBI Exome Sequencing Project (ESP), was launched in 2009. Funded through the American Reinvestment and Recovery Act (ARRA), the ESP was conceived to identify rare, putatively functional, protein-coding variants associated with heart-, lung-, and blood-related diseases and traits. The ESP generated high read depth data on both European Americans (EAs) and African Americans (AAs) and was used to study genetic associations with more than 70 traits. Many of the analytic and logistical challenges we encountered in ESP provide a useful starting point for thinking about WGS studies. Here we describe the major findings and methodological advances from the ESP and the implications they have for the design and analysis of future large-scale sequencing projects.

Material and Methods

Study Design

The original design of the ESP was focused on several phenotypes of high public health significance, as defined by the NHLBI Strategic Plan. Given the cost of deep sequencing at that time, only modest samples sizes were affordable. To enhance statistical power and enrich for variants with strong effects, the ESP employed two selection strategies for many of the subsets: sampling the extremes of quantitative traits and selection of individuals with early age at onset of disease. Several large population-based cohort and case-control

[1]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI 53205, USA; [2]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; [3]Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA 98195, USA; [4]Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA; [5]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA; [6]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA; [7]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [8]Vertex Pharmaceuticals, Boston, MA 02210, USA; [9]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA; [10]Department of Pediatrics, University of Washington, Seattle, WA 98195, USA; [11]Department of Pathology, University of Vermont, Colchester, VT 05405, USA; [12]Department of Biochemistry, University of Vermont, Burlington, VT 05405, USA; [13]Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA
*Correspondence: sleal@bcm.edu
http://dx.doi.org/10.1016/j.ajhg.2016.08.012.

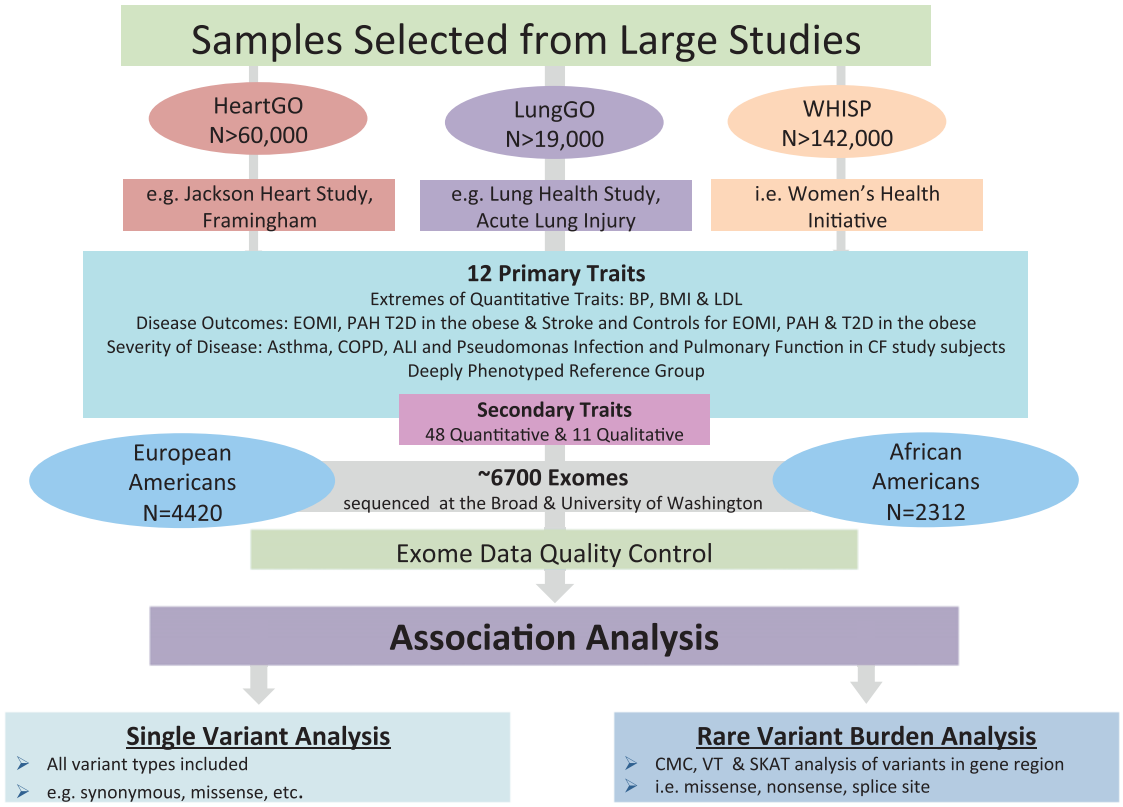


Figure 1. Schematic of the Work Flow for Sample Selection and Data Analysis in ESP
Primary traits were selected from large, population-based studies with widely available data on secondary traits. Both European and African American samples were selected for sequencing. Association analyses were conducted using both genes and single variants as units of analysis.

studies (comprising >220,000 individuals) with detailed phenotype information and available DNA were used to select the 7,034 individuals (4,405 EAs, 2,954 AAs, and 35 of other ancestry) in the ESP (Figure 1).

Primary clinical disease-related phenotypes included acute lung injury (ALI), asthma (MIM: 600807), chronic obstructive pulmonary disease (COPD [MIM: 606963]), early-onset myocardial infarction (EOMI [MIM: 608446]), ischemic stroke (MIM: 601367), type 2 diabetes (T2D [MIM: 125853]) with obesity as a co-morbidity, and pulmonary arterial hypertension-systemic sclerosis (PAH-ssa [MIM: 178600]). Several quantitative cardiovascular risk factors were studied in ESP, including blood pressure (BP), body mass index (BMI), and low-density lipoprotein (LDL), by selecting individuals with either extremely high or low trait values. In addition to the 12 primary traits, many of the ESP participants had data on up to 59 secondary phenotypes, including 48 quantitative biochemical, anthropometric, and subclinical measures of cardiovascular, blood, lung, and kidney disease/function. Detailed descriptions of the sample selection criteria, phenotype definitions, and contributing studies can be found in the Supplemental Data. All participants provided informed consent and the study was approved by the Institutional Review Board of each participating study.

Data Generation and Quality Control

The ESP generated exome-sequence data on 7,034 individuals that had been previously recruited through several large, NHLBI-funded cohort studies and deeply phenotyped on traits of public health importance. After rigorous quality control (QC), data were available on 4,392 EAs and 2,307 AAs. Exome sequencing of the DNA samples was performed at the Broad Institute of Harvard and MIT (n = 3,199) and the University of Washington (n = 3,893). Sequencing was performed to an average read depth of ~90×. Reads were aligned to the human reference sequence (hg19) using the Burrows-Wheeler Alignment tool (BWA[2]) and the resulting binary alignment map (BAM)[3] files were used to call single-nucleotide variants (SNVs) across all samples, i.e., multi-sample calling.

Although we report a 90× mean read-depth, coverage is very unbalanced across the exome and our goal was to obtain at least a 20× read depth for 80% of the exome. As illustrated in Figure S2, there are many regions with low read depth (e.g., <10). Variant sites in these regions would not be accurately genotyped without multi-sample calling. Additionally, multi-sample calling has clear advantages over single-sample calling for variant filtering and for creating a "squared off" call-set where genotypes are called at the same variant sites across all individuals. For these reasons, future uses of the ESP data that seek to combine with other datasets for association testing should consider re-analyzing the BAM files with multi-sample calling.

To identify potentially false-positive variant sites, a support vector machine classifier was used to separate likely true-positive from false-positive variant sites.[4] Sites deemed false positive were excluded from further analyses.

Multidimensional scaling (MDS) was performed in order to validate self-reported EA and AA ancestry.[5] Exomes were screened for

cryptic relatedness and sample duplicates using KING software.[6] Both cryptic and intentionally related and duplicate samples were uncovered; duplicate samples (n = 52) were included as part of the QC process. We found that including intentionally duplicated samples significantly helped us calibrate our QC procedures, although intentional duplicates represent a minimum marginal additional cost (0.74%).

QC that is too stringent can lead to a loss of power if causal variants are removed. On the other hand, inclusion of an excess of false-positive variant sites or incorrect genotypes can increase type I errors as well as reduce power. With this in mind, we sought to maximize the concordance rates of known duplicate samples and transition-transversion ratios while minimizing the amount of data removed during QC. As with the generation and analysis of genotype array data,[7] implementation of standardized protocols for the generation and QC of sequence data will be important for future studies.

A complete description of the exome-sequencing and variant calling protocols has been described in detail in the Supplementary Methods of Fu et al.[8] Details of the variant and sample-level quality control that were implemented in ESP are comprehensively described in Crosby et al.[9] The final, cleaned dataset that was used for association analysis is referred to as the ESP6800.

Phenotype QC

We removed duplicate pairs and first- to third-degree relative pairs by retaining only the sample with the higher call rate. For each phenotype, we removed gross outliers by visual inspection and implausible values (e.g., BMI > 90). We also winsorised trait values to the 0.05% and 99.5%, i.e., trait values greater than the 99.5 quantile or less than the 0.005 quantile were truncated to the 99.5 and 0.005 quantiles, respectively. If necessary, quantitative traits were log-transformed for normality without winsorization. On the final set of samples that were used in the association analysis, we ran principal-component analyses, stratified by genetic ancestry. This was done using the MDS option in PLINK.[5]

Covariates

For each trait we used a model selection procedure to select covariates to be included in the association tests. All regressions included the first two ancestry-specific principal components. Other possible covariates were selected from the following list: age, age^2, sex, BMI, smoking, and an indicator variable representing the capture-array and primary phenotype group for each sample.

Variant-Level Association Testing

We ran per-variant analyses to assess whether any individual variants were associated with an increase or decrease in the quantitative trait (or an increase or decrease in the odds for qualitative traits). Within each genetic ancestry group and for each di-allelic variant with at least 10 observed minor alleles in at least 30 samples, we tested for association between genotype and phenotype with a linear regression model. p values were obtained empirically with an adaptive permutation procedure. For computational efficiency, we also ran linear regression for the qualitative traits. Because our p values were obtained empirically, the tests were still statistically valid.[10] In the autosome, we assumed an additive genetic model as described above. On the X chromosome, we assumed a dominance model in order to have consistent results across both males and females.

Results were meta-analyzed across genetic ancestries if there were at least 10 minor alleles and at least 30 observations present in both ancestry groups. For quantitative traits, meta-analysis was performed using the inverse-variance weighted technique; for qualitative traits, meta-analysis was performed using the sample-size weighted technique.[11]

Gene-Level Association Testing

We ran three different types of gene-level tests: Combined Multivariate Collapsing (CMC),[12] Variable Threshold (VT),[10] and Sequence Kernel Association Test (SKAT).[13] Only missense, nonsense, and splice-variants were considered for inclusion in the gene-level tests. We annotated the variants using the Seattle-Seq annotation server v.134, with the hg19 build of the human reference genome and the NCBI full genes (NM, XM) gene model option.

We noticed that missing genotype data can cause aggregate tests to have increased type I and type II errors. We therefore removed those variant sites missing greater than 10% of their data and imputed missing genotypes to the mean value.[14]

In general, aggregate rare variant association tests suffer from a loss of power when non-causal variants are included in the unit of analysis. For this reason, we a priori excluded synonymous variants, even though a fraction of them may be causal.[15] Although there are a number of tools available for predicting the impact of nonsynonymous variants (e.g., PolyPhen-2[16] and CADD[17]), there are no gold standard approaches; thus, we chose not to include these predictions in our rare variant association testing pipeline.

For the CMC tests, we considered only variants with a within ancestry minor allele frequency ≤ 0.01 that was calculated from the entire ESP6800 call-set. Furthermore, we considered only genes for which the cumulative MAF ≥ 0.005. p values were obtained empirically with an adaptive permutation procedure. p values were meta-analyzed across ancestries if the cumulative MAF ≥ 0.005 in both genetic ancestry groups. Meta-analysis was performed using the sample-size weighted technique.

For the VT tests, we considered only variants with a within ancestry minor allele frequency ≤ 0.05 that was calculated from the entire ESP6800 call-set. For the VT, we considered only genes for which the cumulative MAF ≥ 0.005 at the MAF cutoff that attained the maximum test statistic. p values were meta-analyzed across ancestries if the cumulative MAF ≥ 0.005 in both genetic ancestry groups. Meta-analysis was performed using the sample-size weighted technique.

For the SKAT tests, we considered only variants with a within ancestry minor allele frequency ≤ 0.05 that was calculated from the entire ESP6800 call-set. p values were meta-analyzed across ancestries if the cumulative MAF ≥ 0.005 in both genetic ancestry groups. Meta-analysis was performed using Fisher's Product Method. All association testing was conducted using the Variant Association Tools software.[18]

Imputation

In addition to the direct analyses of the exome-sequence data, ESP investigators utilized imputation in additional samples that were derived from some of the same parent NHLBI cohorts, who were genotyped (but not sequenced). Genotype imputation (in silico genotyping) is a statistical technique for predicting genotypes at variants that are not directly measured.[19] Genotype imputation utilizes a set of reference samples that have been densely genotyped to identify segments of haplotypes that are shared with

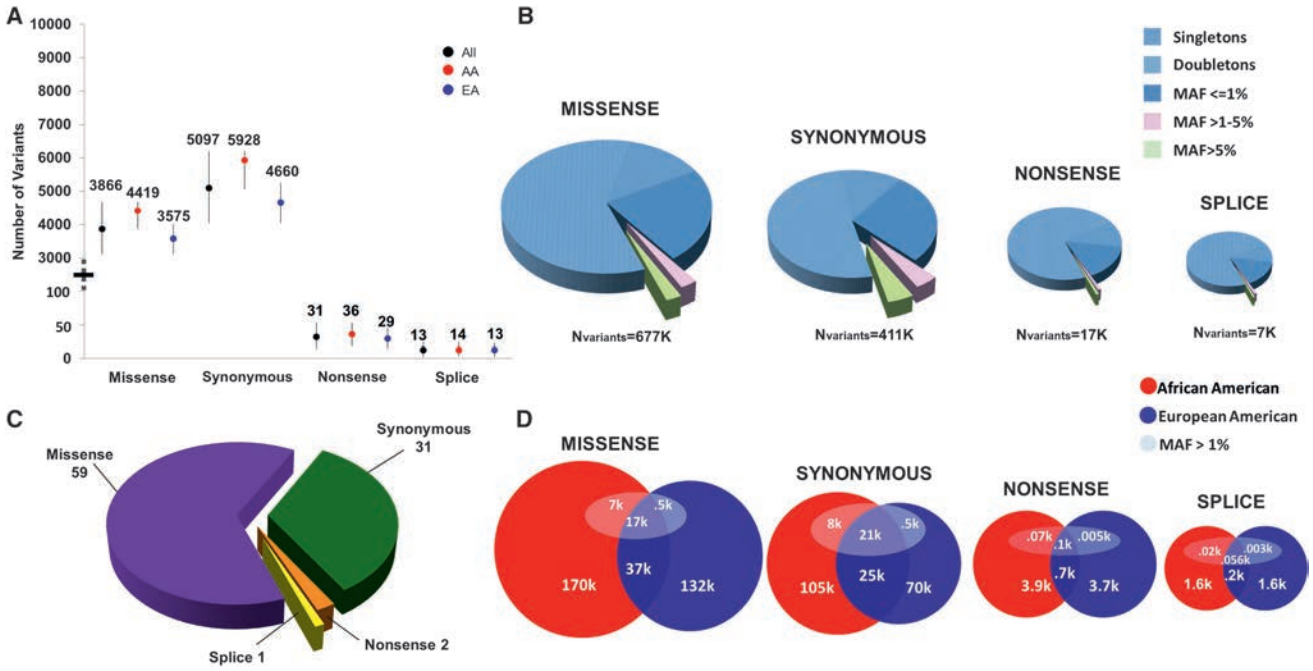


Figure 2. Coding Variants Observed in the NHLBI-ESP

(A) The average number of missense, synonymous, nonsense, and splice site variants per study subject for 2,307 African Americans and 4,392 European Americans and all study subjects (n = 6,699) for the intersect of all four targets. The vertical lines display the smallest and largest number of variants of each type observed per person.

(B) The number of missense, synonymous, nonsense, and splice sites observed for NHLBI-ESP (n = 6,699) study subjects. Represented in each pie chart is the number of singletons, doubletons, and variant sites with an MAF of ≤1%, >1%–5%, and >5%.

(C) The average number of unique missense, synonymous, nonsense, and splice site variants per individual. The variants are not only exclusive to the NHLB-ESP but also are not observed in either dbSNP or 1000 Genomes.

(D) Comparison of the number of coding variant sites observed in AAs and EAs. The number of missense, synonymous, nonsense, and splice site variants that are unique to each population are observed in both populations and have a MAF of ≥1%. The numbers displayed are exclusive to one category. In order to fairly compare the number of variant sites in African Americans and European Americans, equal numbers of African Americans (n = 2,312) and European Americans (n = 2,312) were studied.

the study or "target" population. Prior to ESP, 1,692 AAs and 471 EAs from ESP had been genotyped on the Affymetrix 6.0 array. Using these 4,336 haplotypes as a reference panel, we imputed coding variants from the ESP into ~13,000 AA samples with Affymetrix 6.0 GWAS data. The imputation was performed in several stages throughout the course of the project; details of the imputation at each stage have been previously reported.[20–23]

Power Simulations

Using a simplistic model assigning every variant in the exome the same effect size, we estimated the sample sizes necessary to detect an association at exome-wide significance. We simulated EA samples using parameters adapted from recently published demographic models[24] as input to the forward-time simulator Variant Simulation Tools.[25] DNA sequences of ~1.2 million haplotypes were simulated for all coding regions of CCDS genes on hg19 in the presence of purifying selection.[26] A binary disease with a prevalence of 1% was simulated, assuming all rare variants in the gene have an odds ratio of 1.5. Quantitative traits were simulated similarly, with an effect size of 0.5σ, where σ is the standard deviation of the trait. We evaluated sample size requirements for the CMC and burden of rare variants (BRV)[14] fixed effect tests, as well as the random effects SKAT method at a significance level of 2.5×10^{-6} assuming all rare variants in a gene are causal. A binary search method was used to obtain empirical sample size estimates for 80% power to detect associations.[27]

Results

Most Coding Variation Is Rare and Population Specific

A total of 1,788,563 variant sites were observed in ESP, classified as missense (677,277), synonymous (410,554), nonsense (16,538), and splice (7,049) coding variant sites (Figure 2A). Rare (MAF < 1%) variants comprised the majority within all variant classes: 95.28% missense, 91.11% synonymous, 98.28% nonsense, and 98.25% splice. The majority of coding variants were singletons: 59% missense, 51% synonymous, 71% nonsense, and 72% splice (Figure 2B). Even with a preponderance of singletons, the average number of unique coding variants per individual is <100 (Figure 2C). For all classes of coding variant sites, AAs had, on average, a greater number of variant sites than EAs (Figure 2D). Although a different allelic architecture was observed for EAs and AAs, there was overlap of variant sites: synonymous 20.1% (95% CI 19.95%–20.28%), missense 14.8% (95% CI 14.67%–14.90%), nonsense 9.7% (CI 9.04%–10.31%), and splice 8.2% (CI 7.33%–9.18%). For variant sites that were exclusive to one population with a MAF > 1%, a larger proportion was unique to AAs (missense [$p < 2.2 \times 10^{-16}$], synonymous [$p < 2.2 \times 10^{-16}$], nonsense [$p = 7.3 \times 10^{-16}$], and

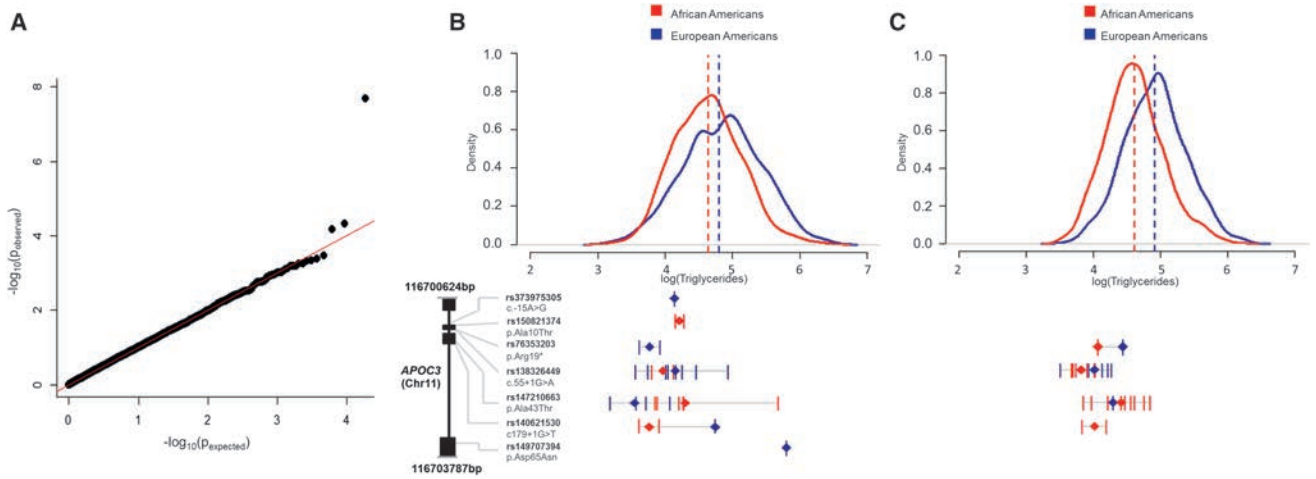


Figure 3. Triglyceride Rare Variant Association Analysis and Association of Rare Variants in *APOC3*

(A) QQ plot of the meta-analysis for African Americans and European Americans of rare variant burden analysis of triglyceride levels. Base 10 –log values of the observed p values are displayed versus their expected values. Rare variant association analysis was performed separately for African Americans (n = 1,654) and European Americans (n = 2,074) using the CMC analyzing those variant sites with a MAF $\leq$ 0.01.

(B) Distribution of triglyceride levels for NHLBI-ESP study subjects and triglyceride levels for individuals with an *APOC3* variant. The quantitative trait distribution of triglycerides after natural log transformation for African Americans and European Americans who are study subjects in the NHLBI-ESP. For the 27 individuals (8 African American and 19 European American) who are heterozygous for one of the 7 coding variants (3 splice, 1 stop-gain, and 3 missense), a tick represents their triglyceride levels after natural log transformation. For each variant site a diamond (red for African Americans and blue for European Americans) represents the average triglyceride levels for carriers of that variant.

(C) Distribution of triglyceride levels for study subjects from the Women's Health Initiative (WHI) and triglyceride levels for individuals with an APOC3 variant. The quantitative trait distribution of triglycerides after natural log transformation for African Americans (n = 1,820) and European Americans (n = 1,643) who are study subjects from the WHI. The DNA samples from the study subjects were genotyped on the exome chip. Of the seven variants that were observed in NHLBI-ESP, four were represented on the exome chip.

splice [$p = 1.5 \times 10^{-5}$]) compared to those distinct to EAs (Figure 2D). For variants outside of coding regions, both the UK10K and 1000 Genomes projects report that most common genetic variants are shared across the world and that most rare variants are specific to closely related populations.[28,29]

Non-synonymous coding variants showed evidence of evolutionary constraint, consistent with purifying selection of deleterious alleles.[30] A modified Out-of-Africa demographic model with accelerated population growth beginning approximately 5,000 years ago demonstrated that the observed excess of rare variation is attributable largely to explosive population growth,[30] with 73% of protein-coding variants in the ESP estimated to have arisen in the past 5,000–10,000 years.[8] This increased mutational load has led to increased allelic and genetic heterogeneity of traits.[8] For disease gene mapping, these results suggest that the complexity we observe in many traits is due, in part, to recent explosive population growth. The implications for association testing are clear: most variants are very rare and testing them individually for association will be under-powered.

Rare-Variant Associations, Imputation, and Replication

We did not observe any systematic inflation of significance from association testing (Figure S1). Ancestry-specific results were examined individually, as well as meta-analyzed for both single variant and aggregate rare variant analyses using a sample-size weighted approach.[11] We noticed that many of our association results were not concordant between EAs and AAs. Though this may be due to false positives within each ancestry group, it may also be due to population-specific allelic architectures where the same rare variants have different effect sizes or simply do not underlie complex trait etiology across major ancestry groups.

The design of ESP was structured around identifying rare variants with large effects. A trade off of this design was that ESP was under-powered to detect common variant associations of modest effects. Indeed, for most traits, we did not identify novel associations for common variants, but rather replicated many known hits. For instance, coding variants in *APOE* (MIM: 107741) and *PCSK9* (MIM: 607786) were associated with LDL cholesterol levels.[31] However, we did report a common missense variant in *PDE4DIP* (MIM: 608117) that was not in linkage disequilibrium with any tag SNPs on any commercial GWAS array and was associated with risk for ischemic stroke.[32]

We identified several trait associations where a burden of rare variants within a gene accounted for the association.[9,31,33,34] Of note, we identified multiple nonsynonymous variants in *APOC3* (MIM: 107720) associated with lower triglyceride levels in both EAs and AAs (Figure 3).[9] A burden of multiple, rare nonsynonymous variants were

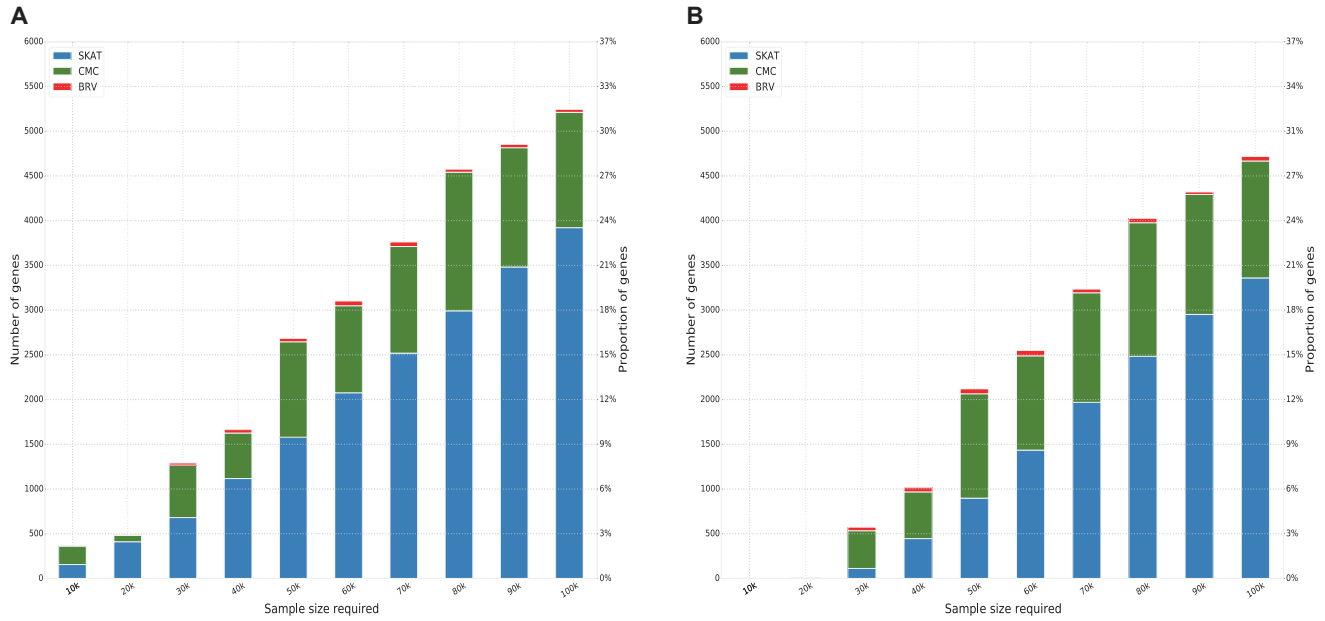


Figure 4. An Analysis of Statistical Power to Detect Associations across the Exome
(A) Sample sizes necessary to detect associations for a binary trait across the exome.
(B) Sample sizes for a quantitative trait.
Results from the SKAT, CMC, and BRV rare-variant association tests are shown in blue, green, and red, respectively.

also found in *DCTN4* (MIM: 614758) to be associated with time to first psuedomonas infection in cystic fibrosis (MIM: 219700).[34] The extremes of LDL cholesterol were combined with other non-LDL extremes to identify a burden of rare and low-frequency variants in *PNPLA5* (MIM: 611589) associated with higher LDL cholesterol levels;[31] rare, nonsynonymous variants in *LDLR* (MIM: 606945) and *APOA5* (MIM: 606368) were shown to be associated with risk of early-onset myocardial infarction.[33] However, very few of the ~70 traits analyzed in ESP resulted in exome-wide significant associations (Figure S1).

In order to replicate significant findings and to follow up in larger populations those associations that did not attain exome-wide significance, the ESP contributed to the development of the Exomechip, a custom genotyping array of rare variant content. Specifically, loss-of-function mutations in *APOC3* and *NPC1L1* (MIM: 608010) were included on the Exomechip and facilitated discovery and replication of associations between these variants and coronary heart disease.[9,35]

In addition to the Exomechip, we utilized genotype imputation to increase sample sizes and enhance our power to detect associations. We created a custom imputation reference panel using samples with both ESP and GWAS data. This reference panel out-performed imputation using 1000 Genomes data as a reference panel[22] and led to the discovery of multiple associations with hematologic and anthropometric traits.[20,21,23,36] Of note, none of these associations were exome-wide significant in the ESP data alone. Only after augmenting our samples size through genotype imputation were we able to identify these associations.

Power to Detect Associations

For the analysis of individual variants, the power to detect an association is affected by disease prevalence, allele frequency, effect size, sample size, and significance (alpha) level.[37] For example, for a disease with a 1% prevalence, a sample size of 40,000 case subjects and 40,000 control subjects is required to have 80% power to detect an association with $\alpha = 5 \times 10^{-8}$ for a variant with MAF = 0.5% and an odds ratio (OR) = 1.5. In addition to these parameters, when aggregating rare variants into a larger unit of analysis (e.g., a gene), association tests are affected by the allelic architecture within a gene (the number of variant sites, their cumulative MAF, the direction and size of their effects, and the proportion of non-causal variants). Thus, for any given trait, the power to detect an aggregate rare variant association varies across genes.

The power to detect aggregate rare-variant associations varies considerably across the exome, with most genes requiring >100,000 samples in order to robustly allow detection of an association (Figure 4). We detect associations with 30% of genes using a sample size of 100,000. Of particular relevance to ESP is that under these idealized conditions, only 1.25% of genes have sufficient power to allow detection of association with 10,000 samples.

Our simulations suggest that even when implementing aggregate rare-variant association methods, even larger sample sizes than those used in large-scale common-variant GWASs will be required to detect associations of modest effects using rare-variant association methods. Due to the differences in allelic architecture between AAs and EAs, sample size calculations for gene-level associations will need to be distinct between these two major ethnic groups as well.

In addition to our simulations with aggregate rare-variant association tests, we sought to quantifiably estimate the additional power obtained from the extreme trait design compared to random ascertainment. If 10,000 samples were selected from the extremes of the quantitative trait values for a cohort of 220,000 individuals, which is the equivalent size of ESP, an association could be detected at an exome-wide significance of 2.5×10^{-6} using the aggregate rare-variant association test for ~80% of the genes in the exome where the effect size is 0.35 σ per each missense, nonsense, and splice site rare variant, with MAF $< 1\%$. This represents a significant increase in power compared to analyzing 10,000 randomly ascertained samples, where an association could be detected for only $< 1.0\%$ of genes. We also found that the increase in power is most significant as the QT threshold goes from 1% to 5%, and for quantitative trait thresholds greater than 5%, although the sample sizes are larger, the power gains are marginal (Figure S3). Thus, if the underlying cohort is large enough to permit extreme sampling beyond the 5^{th} and 95^{th} percentiles in sufficient numbers, and the focus of the study is on a single quantitative trait, we recommend an extreme trait sampling design to boost power for detecting associations.

Discussion

Based on the experiences of the ESP and several similar recent WES projects, data generation will not represent a major technical hurdle for future sequencing-based studies of rare-variant associations. Nonetheless, as throughput continues to increase with decreasing sequencing costs, the data-management, variant-calling, QC, and analysis of these data will continue to pose challenges to the scientific community. Through the efforts of hundreds of investigators, the ESP helped establish best practices to turn terabytes of raw sequence data into genetic discoveries for complex traits and diseases.

There were a number of issues involved in having two separate sequencing facilities process and sequence the DNA for this project. The main advantages were competition and innovation: both sequencing centers were actively involved in optimizing their capture and sequencing protocols that led to improvements in coverage and data quality. However, differences between centers due to capture re-agents and analysis strategies created batch effects that we had to control for in the downstream analyses. Although the use of joint-calling over all samples mitigated some of these effects, in retrospect, the analysis pipeline would have been benefitted from uniform alignment and processing strategies, e.g., use of the same capture array at the two centers. Our experience highlights the importance of good experimental design; for instance, balancing case and control subjects across sequencing centers. In order to control type I and II errors, our genetic association analysis incorporated the study design, by including dummy variables to represent different sequencing centers, capture re-agents, and the source and ascertainment of samples.

With improvements in capture re-agents and consistency in coverage across the exome, we anticipate that future WES projects may be able to successfully sequence at lower depth. Specifically, an average depth per variant of about 25× appears to be a sweet spot where variant sites are covered at about the same depth as invariant sites (see Figure S2). This is slightly less than the recommended read depth of 30×, which is used for whole-genome sequencing and which produces substantially more even coverage than exome sequencing. Indeed, one or our rationale for choosing 90× read depth was to provide 80% coverage of the target bases with at least 20× read depth.

As more sequence data continues to be generated, studies will inevitably encounter a situation similar to ESP, where most observed variants are rare and population specific. In order to detect phenotypic associations from these data, the ESP pioneered several approaches for increasing statistical power. Samples were drawn from extremes of continuous traits and early-onset cases of complex diseases, special statistical methods were used to leverage the extreme trait data, rare variants were aggregated into larger units of analysis (i.e., genes), and through imputation and genotyping, large sample sizes were utilized for both discovery and replication.

Importantly, the ESP data represent the largest single collection of AA exomes to date. The AA exomes in the ESP generated multiple discoveries that would have been impossible to detect in US populations of European ancestry.[20,23] Indeed, AAs had on average a greater number of variant sites than EAs and a larger proportion of rare variants were exclusive to AAs compared to EAs. Although AAs are traditionally an under-studied population in human genetics, the ESP showed that the genetic diversity in AA genomes can be harnessed for uncovering rare-variant associations. In particular, variants in *APOC3* were identified in both EAs and AAs and were associated with lower triglyceride levels in both ethnicities. Our work with imputation of the ESP AA sequence data also identified several variants that were monomorphic in EAs but reached exome-wide significance in AAs.

The sample selection of ESP presented distinct challenges: (1) heterogeneity between cohorts introduced noise that could not be entirely mitigated through phenotype harmonization; and (2) a simpler design focused on fewer traits with larger sample sizes may have yielded more discoveries. Nonetheless, our findings from AAs point to a major advantage of a heterogeneous sample of multiple ancestry groups, focused on a number of different traits. By including data from large, deeply phenotyped, US cohort studies, we were able to scan 71 different traits for genetic associations in two major ancestry groups. And though it complicated our analyses of both primary and secondary traits, we were able to sample from the extremes of multiple large cohorts, providing us with much more extreme trait values than if we

had sampled from a single cohort. Consequently when results are based on an extreme sampling design, caution should be used in generalizing the results to the larger population, because they may be different from the extremes in systematic and unanticipated ways.

Incidental Findings

Prior to the ESP, no large-scale study had generated sufficient quantities of protein-coding sequence variants to enable the estimation of the number of medically actionable genetic variants per individual. Both an initial (based on 500 EA and AA participants) and a final (based on 6,503 participants) analysis of the ESP data provided robust estimates of the carrier frequency of adults with high-penetrance actionable or likely pathogenic variants.[38,39] The ESP data also provided estimates of carrier burden for complex traits such as age-related macular degeneration and drug response.[40] These studies from ESP demonstrated the many challenges in variant classification and association with heterogeneous human disease burden. Future sequencing studies will need to grapple with these challenges as more populations are studied and the data expand outside of protein-coding regions.

Statistical Considerations with Association Testing

Due to its unprecedented scale and unique study design, the ESP prompted development of statistical methods and software to handle both extreme trait sampling and rare variant association testing. Quantitative trait data that have been generated by an extreme trait design (as in the ESP) are not normally distributed and should not be analyzed with standard methods. Modifications to likelihood-based approaches (such as conditional likelihood) can overcome the inherent bias in such a design.[41] Secondary traits (traits that were not used as the basis of sample selection) from an extreme trait design likewise require special consideration.[42]

Using the ESP Data

The ESP data are available to investigators through the NIH database of genotypes and phenotypes (dbGaP). As a cautionary note, the ESP variant data should not be used as convenience controls for rare variant association testing. Doing so may significantly inflate association signals. If the ESP data or any other data from sequence-based studies are to be combined with sequence data from other projects, we recommend recalling all genotypes from the underlying BAM files using multi-sample calling, in order to avoid batch effects.

In our analytic pipeline for analyzing the ESP data, we removed related individuals in order to satisfy the assumption that the observations are independent in our regression framework. This required removal of only a few individuals and at the time, family-based methods for rare-variant association analysis using linear mixed models were not available. The advantages of mixed models are that all individuals can be retained in the analysis, and type I inflations due to either relatedness or cryptic relatedness are avoided.

Implications for Future Studies

As new resources emerge (e.g., data from the Encyclopedia of DNA Elements [ENCODE] and Roadmap projects) for interpreting DNA variation outside of coding regions, and as sequencing costs continue to decline,[43] sequencing-based studies will not be limited to the exome. However, the exome offers a natural unit of analysis (i.e., a gene) for aggregate rare-variant association methods. It is unclear how best to aggregate association signals outside of coding regions. Will the genetic effects in enhancers, promoters, and other elements related to gene regulation be detectable by the same methods that were used for the exome? For future WGS studies to uncover aggregate signals outside of coding regions, methods development in this area will be crucial. In addition, many of the known loci that underlie complex diseases are located in regulatory elements outside of coding regions.[44] Our experiences in the ESP confirm that for many traits and diseases, collaboration with other sequencing consortia (e.g., CHARGE, T2DGENES) will be necessary to accumulate the tens of thousands of samples required to detect associations with low-frequency and rare variants.[26,45] Imputation will also play an important role in future studies. The advantages of using a study-specific ESP reference panel for imputing rare variants appear to generalize to the whole genome.[46,47] With larger reference panels for imputation such as from the Haplotype Reference Consortium (HRC),[48] the ability to impute rare variants across the genome with higher accuracy will continue to improve; current estimates suggest the HRC panel can impute variants to 0.1% MAF.

Just as in GWASs of common variants, the human genetics community is coalescing around the notion that the path to discovering insights into the biological mechanisms that underlie complex diseases is through data sharing and large-scale consortia. With the assistance of the NCBI's database of genotypes and phenotypes (dbGaP), the ESP was a pioneer in data sharing and rapid analysis of large-scale sequence data. As new data continue to be generated for the study of complex trait genetics, this model of large-scale collaboration and data sharing should be emulated.

Supplemental Data

Supplemental Data include Supplemental Notes containing descriptions of the sample selection criteria, phenotype definitions, and contributing studies, three figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.08.012.

Consortia

The NHLBI GO Exome Sequencing Project consists of the following individuals. From BroadGO: Stacey B. Gabriel, David M. Altshuler, Gonçalo R. Abecasis, Hooman Allayee, Sharon Cresci, Mark J. Daly, Paul I. W. de Bakker, Mark A. DePristo, Ron Do, Peter Donnelly,

Deborah N. Farlow, Tim Fennell, Kiran Garimella, Stanley L. Hazen, Youna Hu, Daniel M. Jordan, Goo Jun, Sekar Kathiresan, Hyun Min Kang, Adam Kiezun, Guillaume Lettre, Bingshan Li, Mingyao Li, Christopher H. Newton-Cheh, Sandosh Padmanabhan, Gina Peloso, Sara Pulit, Daniel J. Rader, David Reich, Muredach P. Reilly, Manuel A. Rivas, Steve Schwartz, Laura Scott, David S. Siscovick, John A. Spertus, Nathan O. Stitziel, Nina Stoletzki, Shamil R. Sunyaev, Benjamin F. Voight, and Cristen J. Willer. From HeartGO: Stephen S. Rich, Ermeg Akylbekova, Larry D. Atwood, Christie M. Ballantyne, Maja Barbalic, R. Graham Barr, Emelia J. Benjamin, Joshua Bis, Eric Boerwinkle, Donald W. Bowden, Jennifer Brody, Matthew Budoff, Greg Burke, Sarah Buxbaum, Jeff Carr, Donna T. Chen, Ida Y. Chen, Wei-Min Chen, Pat Concannon, Jacy Crosby, L. Adrienne Cupples, Ralph D'Agostino, Anita L. DeStefano, Albert Dreisbach, Josée Dupuis, J. Peter Durda, Jaclyn Ellis, Aaron R. Folsom, Myriam Fornage, Caroline S. Fox, Ervin Fox, Vincent Funari, Santhi K. Ganesh, Julius Gardin, David Goff, Ora Gordon, Wayne Grody, Myron Gross, Xiuqing Guo, Ira M. Hall, Nancy L. Heard-Costa, Susan R. Heckbert, Nicholas Heintz, David M. Herrington, DeMarc Hickson, Jie Huang, Shih-Jen Hwang, David R. Jacobs, Nancy S. Jenny, Andrew D. Johnson, Craig W. Johnson, Steven Kawut, Richard Kronmal, Raluca Kurz, Ethan M. Lange, Leslie A. Lange, Martin G. Larson, Mark Lawson, Cora E. Lewis, Daniel Levy, Dalin Li, Honghuang Lin, Chunyu Liu, Jiankang Liu, Kiang Liu, Xiaoming Liu, Yongmei Liu, William T. Longstreth, Cay Loria, Thomas Lumley, Kathryn Lunetta, Aaron J. Mackey, Rachel Mackey, Ani Manichaikul, Taylor Maxwell, Barbara McKnight, James B. Meigs, Alanna C. Morrison, Solomon K. Musani, Josyf C. Mychaleckyj, Jennifer A. Nettleton, Kari North, Christopher J. O'Donnell, Daniel O'Leary, Frank S. Ong, Walter Palmas, James S. Pankow, Nathan D. Pankratz, Shom Paul, Marco Perez, Sharina D. Person, Joseph Polak, Wendy S. Post, Bruce M. Psaty, Aaron R. Quinlan, Leslie J. Raffel, Vasan S. Ramachandran, Alexander P. Reiner, Kenneth Rice, Jerome I. Rotter, Jill P. Sanders, Pamela Schreiner, Sudha Seshadri, Steve Shea, Stephen Sidney, Kevin Silverstein, David S. Siscovick, Nicholas L. Smith, Nona Sotoodehnia, Asoke Srinivasan, Herman A. Taylor, Kent Taylor, Fridtjof Thomas, Russell P. Tracy, Michael Y. Tsai, Kelly A. Volcik, Chrstina L Wassel, Karol Watson, Gina Wei, Wendy White, Kerri L. Wiggins, Jemma B. Wilk, O. Dale Williams, Gregory Wilson, James G. Wilson, Phillip Wolf, and Neil A. Zakai. From ISGS and SWISS: John Hardy, James F. Meschia, Michael Nalls, Stephen S. Rich, Andrew Singleton, and Brad Worrall. From LungGO: Michael J. Bamshad, Kathleen C. Barnes, Ibrahim Abdulhamid, Frank Accurso, Ran Anbar, Terri Beaty, Abigail Bigham, Phillip Black, Eugene Bleecker, Kati Buckingham, Anne Marie Cairns, Wei-Min Chen, Daniel Caplan, Barbara Chatfield, Aaron Chidekel, Michael Cho, David C. Christiani, James D. Crapo, Julia Crouch, Denise Daley, Anthony Dang, Hong Dang, Alicia De Paula, Joan DeCelie-Germana, Allen Dozor, Mitch Drumm, Maynard Dyson, Julia Emerson, Mary J. Emond, Thomas Ferkol, Robert Fink, Cassandra Foster, Deborah Froh, Li Gao, William Gershan, Ronald L. Gibson, Elizabeth Godwin, Magdalen Gondor, Hector Gutierrez, Nadia N. Hansel, Paul M. Hassoun, Peter Hiatt, John E. Hokanson, Michelle Howenstine, Laura K. Hummer, Seema M. Jamal, Jamshed Kanga, Yoonhee Kim, Michael R. Knowles, Michael Konstan, Thomas Lahiri, Nan Laird, Christoph Lange, Lin Lin, Xihong Lin, Tin L. Louie, David Lynch, Barry Make, Thomas R. Martin, Steve C. Mathai, Rasika A. Mathias, John McNamara, Sharon McNamara, Deborah Meyers, Susan Millard, Peter Mogayzel, Richard Moss, Tanda Murray, Dennis Nielson, Blakeslee Noyes, Wanda O'Neal, David Orenstein, Brian O'Sullivan, Rhonda Pace, Peter Pare, H. Worth Parker, Mary Ann Passero, Elizabeth Perkett, Adrienne Prestridge, Nicholas M. Rafaels, Bonnie Ramsey, Elizabeth Regan, Clement Ren, George Retsch-Bogart, Michael Rock, Antony Rosen, Margaret Rosenfeld, Ingo Ruczinski, Andrew Sanford, David Schaeffer, Cindy Sell, Daniel Sheehan, Edwin K. Silverman, Don Sin, Terry Spencer, Jackie Stonebraker, Holly K. Tabor, Laurie Varlotta, Candelaria I. Vergara, Robert Weiss, Fred Wigley, Robert A. Wise, Fred A. Wright, Mark M. Wurfel, Robert Zanni, and Fei Zou. From SeattleGO: Deborah A. Nickerson, Mark J. Rieder, Phil Green, Jay Shendure, Joshua M. Akey, Michael J. Bamshad, Kristine L. Bucasas, Carlos D. Bustamante, David R. Crosslin, Evan E. Eichler, P. Keolu Fox, Wenqing Fu, Adam Gordon, Simon Gravel, Gail P. Jarvik, Jill M. Johnsen, Mengyuan Kan, Eimear E. Kenny, Jeffrey M. Kidd, Fremiet Lara-Garduno, Suzanne M. Leal, Dajiang J. Liu, Sean McGee, Timothy D. O'Connor, Bryan Paeper, Peggy D. Robertson, Joshua D. Smith, Jeffrey C. Staples, Jacob A. Tennessen, Emily H. Turner, Gao Wang, and Qian Yi. From WHISP: Rebecca Jackson, Kari North, Ulrike Peters, Christopher S. Carlson, Garnet Anderson, Hoda Anton-Culver, Themistocles L. Assimes, Paul L. Auer, Shirley Beresford, Chris Bizon, Henry Black, Robert Brunner, Robert Brzyski, Dale Burwen, Bette Caan, Cara L. Carty, Rowan Chlebowski, Steven Cummings, J. David Curb, Charles B. Eaton, Leslie Ford, Nora Franceschini, Stephanie M. Fullerton, Margery Gass, Nancy Geller, Gerardo Heiss, Barbara V. Howard, Li Hsu, Carolyn M. Hutter, John Ioannidis, Shuo Jiao, Karen C. Johnson, Charles Kooperberg, Lewis Kuller, Andrea LaCroix, Kamakshi Lakshminarayan, Dorothy Lane, Ethan M. Lange, Leslie A. Lange, Norman Lasser, Erin LeBlanc, Cora E. Lewis, Kuo-Ping Li, Marian Limacher, Dan-Yu Lin, Benjamin A. Logsdon, Shari Ludlam, JoAnn E. Manson, Karen Margolis, Lisa Martin, Joan McGowan, Keri L. Monda, Jane Morley Kotchen, Lauren Nathan, Judith Ockene, Mary Jo O'Sullivan, Lawrence S. Phillips, Ross L. Prentice, Alexander P. Reiner, John Robbins, Jennifer G. Robinson, Jacques E. Rossouw, Haleh Sangi-Haghpeykar, Gloria E. Sarto, Sally Shumaker, Michael S. Simon, Marcia L. Stefanick, Evan Stein, Hua Tang, Kira C. Taylor, Cynthia A. Thomson, Timothy A. Thornton, Linda Van Horn, Mara Vitolins, Jean Wactawski-Wende, Robert Wallace, Sylvia Wassertheil-Smoller, and Donglin Zeng. From NHLBI GO ESP Project Team: Deborah Applebaum-Bowden, Michael Feolo, Weiniu Gan, Dina N. Paltoo, Jacques E. Rossouw, Phyliss Sholinsky, and Anne Sturcke.

Acknowledgments

We acknowledge the support of the National Heart, Lung and Blood Institute (NHLBI), the contributions of the research institutions that participated in this study, and the study investigators, field staff, and study participants who created this resource for biomedical research. Funding for the GO (Grand Opportunity) Exome Sequencing Project was provided by NHLBI grants RC2 HL-103010 (Heart GO), RC2 HL-102923 (Lung GO), and RC2 HL-102924 (WHISP). The exome sequencing was supported by NHLBI grants RC2 HL-102925 (Broad GO) and RC2 HL-102926 (Seattle GO).

Received: February 26, 2016
Accepted: August 8, 2016
Published: September 22, 2016

Web Resources

Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE), http://www.chargeconsortium.com

dbGaP, http://www.ncbi.nlm.nih.gov/gap

ENCODE, https://www.encodeproject.org/

Exome Chip Design, http://genome.sph.umich.edu/wiki/Exome_Chip_Design

NHLBI Exome Sequencing Project, https://esp.gs.washington.edu/drupal/

OMIM, http://www.omim.org/

Precision Medicine Initiative (PMI), NIH, https://www.nih.gov/precision-medicine-initiative-cohort-program

Roadmap, http://www.roadmapepigenomics.org/

SeattleSeq Annotation Server, http://snp.gs.washington.edu/SeattleSeqAnnotation134/

T2D-GENES Consortium, https://t2d-genes.sph.umich.edu

The Haplotype Reference Consortium, http://www.haplotype-reference-consortium.org/home

Trans-Omics for Precision Medicine (TOPMed) Program, https://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed

References

1. Collins, F.S., and Varmus, H. (2015). A new initiative on precision medicine. N. Engl. J. Med. *372*, 793–795.
2. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
3. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
4. Jun, G., Wing, M.K., Abecasis, G.R., and Kang, H.M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Res. *25*, 918–925.
5. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
6. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.
7. Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
8. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al.; NHLBI Exome Sequencing Project (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216–220.
9. Crosby, J., Peloso, G.M., Auer, P.L., Crosslin, D.R., Stitziel, N.O., Lange, L.A., Lu, Y., Tang, Z.Z., Zhang, H., Hindy, G., et al.; TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute (2014). Loss-of-function mutations in APOC3, triglycerides, and coronary disease. N. Engl. J. Med. *371*, 22–31.
10. Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. *86*, 832–838.
11. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191.
12. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. *83*, 311–321.
13. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.
14. Auer, P.L., Wang, G., and Leal, S.M. (2013). Testing for rare variant associations in the presence of missing data. Genet. Epidemiol. *37*, 529–538.
15. Scheidecker, S., Etard, C., Haren, L., Stoetzel, C., Hull, S., Arno, G., Plagnol, V., Drunat, S., Passemard, S., Toutain, A., et al. (2015). Mutations in TUBGCP4 alter microtubule organization via the γ-tubulin ring complex in autosomal-recessive microcephaly with chorioretinopathy. Am. J. Hum. Genet. *96*, 666–674.
16. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
17. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.
18. Wang, G.T., Peng, B., and Leal, S.M. (2014). Variant association tools for quality control and analysis of large-scale sequence and genotyping array data. Am. J. Hum. Genet. *94*, 770–783.
19. Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nat. Rev. Genet. *11*, 499–511.
20. Auer, P.L., Johnsen, J.M., Johnson, A.D., Logsdon, B.A., Lange, L.A., Nalls, M.A., Zhang, G., Franceschini, N., Fox, K., Lange, E.M., et al. (2012). Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. Am. J. Hum. Genet. *91*, 794–808.
21. Du, M., Auer, P.L., Jiao, S., Haessler, J., Altshuler, D., Boerwinkle, E., Carlson, C.S., Carty, C.L., Chen, Y.D., Curtis, K., et al.; National Heart, Lung, and Blood Institute (NHLBI) GO Exome Sequencing Project (2014). Whole-exome imputation of sequence variants identified two novel alleles associated with adult body height in African Americans. Hum. Mol. Genet. *23*, 6607–6615.
22. Duan, Q., Liu, E.Y., Auer, P.L., Zhang, G., Lange, E.M., Jun, G., Bizon, C., Jiao, S., Buyske, S., Franceschini, N., et al. (2013). Imputation of coding variants in African Americans: better performance using data from the exome sequencing project. Bioinformatics *29*, 2744–2749.
23. Johnsen, J.M., Auer, P.L., Morrison, A.C., Jiao, S., Wei, P., Haessler, J., Fox, K., McGee, S.R., Smith, J.D., Carlson, C.S., et al.; NHLBI Exome Sequencing Project (2013). Common and rare von Willebrand factor (VWF) coding variants, VWF levels, and factor VIII levels in African Americans: the NHLBI Exome Sequencing Project. Blood *122*, 590–597.
24. Gazave, E., Ma, L., Chang, D., Coventry, A., Gao, F., Muzny, D., Boerwinkle, E., Gibbs, R.A., Sing, C.F., Clark, A.G., and Keinan, A. (2014). Neutral genomic regions refine models of recent rapid human population growth. Proc. Natl. Acad. Sci. USA *111*, 757–762.

25. Peng, B. (2015). Reproducible simulations of realistic samples for next-generation sequencing studies using Variant Simulation Tools. Genet. Epidemiol. *39*, 45–52.
26. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. Proc. Natl. Acad. Sci. USA *106*, 3871–3876.
27. Wang, G.T., Li, B., Santos-Cortez, R.P., Peng, B., and Leal, S.M. (2014). Power analysis and sample size estimation for sequence-based association studies. Bioinformatics *30*, 2377–2378.
28. Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., Lawson, D., et al.; UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.
29. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
30. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al.; Broad GO; Seattle GO; NHLBI Exome Sequencing Project (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.
31. Lange, L.A., Hu, Y., Zhang, H., Xue, C., Schmidt, E.M., Tang, Z.Z., Bizon, C., Lange, E.M., Smith, J.D., Turner, E.H., et al.; NHLBI Grand Opportunity Exome Sequencing Project (2014). Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. Am. J. Hum. Genet. *94*, 233–245.
32. Auer, P.L., Nalls, M., Meschia, J.F., Worrall, B.B., Longstreth, W.T., Jr., Seshadri, S., Kooperberg, C., Burger, K.M., Carlson, C.S., Carty, C.L., et al.; National Heart, Lung, and Blood Institute Exome Sequencing Project (2015). Rare and coding region genetic variants associated with risk of ischemic stroke: The NHLBI Exome Sequence Project. JAMA Neurol. *72*, 781–788.
33. Do, R., Stitziel, N.O., Won, H.H., Jørgensen, A.B., Duga, S., Angelica Merlini, P., Kiezun, A., Farrall, M., Goel, A., Zuk, O., et al.; NHLBI Exome Sequencing Project (2015). Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. Nature *518*, 102–106.
34. Emond, M.J., Louie, T., Emerson, J., Zhao, W., Mathias, R.A., Knowles, M.R., Wright, F.A., Rieder, M.J., Tabor, H.K., Nickerson, D.A., et al.; National Heart, Lung, and Blood Institute (NHLBI) GO Exome Sequencing Project; Lung GO (2012). Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis. Nat. Genet. *44*, 886–889.
35. Stitziel, N.O., Won, H.H., Morrison, A.C., Peloso, G.M., Do, R., Lange, L.A., Fontanillas, P., Gupta, N., Duga, S., Goel, A., et al.; Myocardial Infarction Genetics Consortium Investigators (2014). Inactivating mutations in NPC1L1 and protection from coronary heart disease. N. Engl. J. Med. *371*, 2072–2082.
36. Naik, R.P., Derebail, V.K., Grams, M.E., Franceschini, N., Auer, P.L., Peloso, G.M., Young, B.A., Lettre, G., Peralta, C.A., Katz, R., et al. (2014). Association of sickle cell trait with chronic kidney disease and albuminuria in African Americans. JAMA *312*, 2115–2125.
37. Purcell, S., Cherny, S.S., and Sham, P.C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics *19*, 149–150.
38. Amendola, L.M., Dorschner, M.O., Robertson, P.D., Salama, J.S., Hart, R., Shirts, B.H., Murray, M.L., Tokita, M.J., Gallego, C.J., Kim, D.S., et al. (2015). Actionable exomic incidental findings in 6503 participants: challenges of variant classification. Genome Res. *25*, 305–315.
39. Dorschner, M.O., Amendola, L.M., Turner, E.H., Robertson, P.D., Shirts, B.H., Gallego, C.J., Bennett, R.L., Jones, K.L., Tokita, M.J., Bennett, J.T., et al.; National Heart, Lung, and Blood Institute Grand Opportunity Exome Sequencing Project (2013). Actionable, pathogenic incidental findings in 1,000 participants' exomes. Am. J. Hum. Genet. *93*, 631–640.
40. Tabor, H.K., Auer, P.L., Jamal, S.M., Chong, J.X., Yu, J.H., Gordon, A.S., Graubert, T.A., O'Donnell, C.J., Rich, S.S., Nickerson, D.A., and Bamshad, M.J.; NHLBI Exome Sequencing Project (2014). Pathogenic variants for Mendelian and complex traits in exomes of 6,517 European and African Americans: implications for the return of incidental results. Am. J. Hum. Genet. *95*, 183–193.
41. Lin, D.Y., Zeng, D., and Tang, Z.Z. (2013). Quantitative trait analysis in sequencing studies under trait-dependent sampling. Proc. Natl. Acad. Sci. USA *110*, 12247–12252.
42. Liu, D.J., and Leal, S.M. (2012). A unified method for detecting secondary trait associations with rare variants: application to sequence data. PLoS Genet. *8*, e1003075.
43. Hayden, E.C. (2014). Technology: The $1,000 genome. Nature *507*, 294–295.
44. Consortium, E.P.; ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
45. Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., et al. (2012). Exome sequencing and the genetic basis of complex traits. Nat. Genet. *44*, 623–630.
46. Deelen, P., Menelaou, A., van Leeuwen, E.M., Kanterakis, A., van Dijk, F., Medina-Gomez, C., Francioli, L.C., Hottenga, J.J., Karssen, L.C., Estrada, K., et al.; Genome of Netherlands Consortium (2014). Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. Eur. J. Hum. Genet. *22*, 1321–1326.
47. Pistis, G., Porcu, E., Vrieze, S.I., Sidore, C., Steri, M., Danjou, F., Busonero, F., Mulas, A., Zoledziewska, M., Maschio, A., et al. (2015). Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. Eur. J. Hum. Genet. *23*, 975–983.
48. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al.; Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. http://dx.doi.org/10.1038/ng.3643.

Modeling the Mutational and Phenotypic Landscapes of Pelizaeus-Merzbacher Disease with Human iPSC-Derived Oligodendrocytes

Zachary S. Nevin,[1] Daniel C. Factor,[1] Robert T. Karl,[1] Panagiotis Douvaras,[2] Jeremy Laukka,[3] Martha S. Windrem,[4] Steven A. Goldman,[4,5] Valentina Fossati,[2] Grace M. Hobson,[6,7,8] and Paul J. Tesar[1,*]

Pelizaeus-Merzbacher disease (PMD) is a pediatric disease of myelin in the central nervous system and manifests with a wide spectrum of clinical severities. Although PMD is a rare monogenic disease, hundreds of mutations in the X-linked myelin gene proteolipid protein 1 (*PLP1*) have been identified in humans. Attempts to identify a common pathogenic process underlying PMD have been complicated by an incomplete understanding of PLP1 dysfunction and limited access to primary human oligodendrocytes. To address this, we generated panels of human induced pluripotent stem cells (hiPSCs) and hiPSC-derived oligodendrocytes from 12 individuals with mutations spanning the genetic and clinical diversity of PMD—including point mutations and duplication, triplication, and deletion of *PLP1*—and developed an in vitro platform for molecular and cellular characterization of all 12 mutations simultaneously. We identified individual and shared defects in *PLP1* mRNA expression and splicing, oligodendrocyte progenitor development, and oligodendrocyte morphology and capacity for myelination. These observations enabled classification of PMD subgroups by cell-intrinsic phenotypes and identified a subset of mutations for targeted testing of small-molecule modulators of the endoplasmic reticulum stress response, which improved both morphologic and myelination defects. Collectively, these data provide insights into the pathogeneses of a variety of *PLP1* mutations and suggest that disparate etiologies of PMD could require specific treatment approaches for subsets of individuals. More broadly, this study demonstrates the versatility of a hiPSC-based panel spanning the mutational heterogeneity within a single disease and establishes a widely applicable platform for genotype-phenotype correlation and drug screening in any human myelin disorder.

Introduction

Leukodystrophies are a set of rare genetic disorders characterized by developmental delay and motor impairment due to deficits in myelin, also called "white matter," in the central nervous system (CNS).[1,2] Myelin is a highly structured membrane that ensheathes neuron axons to provide ancillary support and promote proper coordination of electric impulses. Most leukodystrophies have an onset in early childhood, and many are fatal. Although individuals are routinely diagnosed on the basis of symptoms and genetic testing, most leukodystrophies are still poorly understood, and treatment options are largely limited to palliative symptom management.[3–5]

Pelizaeus-Merzbacher disease (PMD [MIM: 312080]) is an X-linked leukodystrophy that affects approximately 1,000 children in the United States[6–9] (also see GeneReviews in Web Resources). First described in 1885,[10,11] PMD was mapped to proteolipid protein 1 (*PLP1* [MIM: 300401]) in the late 1980s.[12–15] *PLP1* and its splice isoform, *DM20*, are predominantly expressed by oligodendrocytes—the myelin-producing cell of the CNS—and their progenitors (oligodendrocyte progenitor cells [OPCs]), respectively.[16] PLP1 is the major protein component of myelin and has been found to compose as much as 50% of myelin's total protein content.[17] PLP1's amino acid composition is 100% conserved among humans, rats, and mice,[17,18] there is only a single variant in dogs (p.Val161Ile), and mutations that cause PMD-like symptoms have been identified in each of these species.[19–23] PLP1's strong cell-type specificity, abundance in myelin, and inter-species conservation all suggest that it fills an indispensable role in oligodendrocyte and myelin biology. However, PLP1's function in oligodendrocytes and myelin has only recently begun to be elucidated[24–27] and is still very much in question. As a result, there is no current consensus on the pathogenic processes by which *PLP1* mutations cause PMD.

Although PMD is a monogenic disease, affected individuals present with a surprising spectrum of onset, disability, and mortality, which have been grouped into three categories. The common "classic" form manifests as a constellation of hypotonia, nystagmus, and/or motor delay in early childhood and the development of progressive spasticity, ataxia, and/or choreiform movements through adolescence and early adulthood.[28–31] Some individuals live into their seventh decade, but many develop fatal complications of hypotonia and spasticity by their late 20s. In the more severe "connatal" form, symptoms arise early in infancy and

[1]Department of Genetics and Genome Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH 44106, USA; [2]New York Stem Cell Foundation Research Institute, New York, NY 10032, USA; [3]Departments of Neurology and Neuroscience, College of Medicine and Life Science, University of Toledo, Toledo, OH 43614, USA; [4]Center for Translational Neuromedicine, University of Rochester Medical Center, Rochester, NY 14642, USA; [5]Center for Neuroscience, Faculty of Medicine and Health Sciences, University of Copenhagen, 2200 Copenhagen, Denmark; [6]Nemours Biomedical Research, Alfred I. duPont Hospital for Children, Wilmington, DE 19803, USA; [7]Department of Biological Sciences, University of Delaware, Newark, DE 19716, USA; [8]Department of Pediatrics, Jefferson Medical College, Thomas Jefferson University, Philadelphia, PA 19107, USA
*Correspondence: paul.tesar@case.edu
http://dx.doi.org/10.1016/j.ajhg.2017.03.005.

are typically fatal within the first few years of life. Lastly, a few males and most of the exceedingly rare females who present with PMD can develop mild, late-onset spasticity in the legs or assorted mild peripheral neuropathies with minimal CNS deficits[32] (also see GeneReviews in Web Resources).

This significant clinical heterogeneity has been attributed to hundreds of different mutations of *PLP1*. A majority of PMD cases (70%) are caused by duplications of the *PLP1* locus and manifest with classic PMD of mild to moderate severity.[7] Rare triplications (<1%) cause severe connatal disease, whereas full gene deletions (1%–2%) are associated with mild, late-onset symptoms, often termed "null syndrome"[18,33–36] (also see GeneReviews in Web Resources). Additionally, over 200 unique point mutations have been identified in individuals (25%) presenting with PMD across the entire range of severity.[18] Point mutations and indels have been found throughout *PLP1*'s coding sequence, splice sites, and introns. *PLP1* has one splice isoform, *DM20*, created by a cryptic splice signal in exon 3 and exclusion of the latter 105 nucleotides in that exon (exon 3b).[37] In the oligodendrocyte lineage, *DM20* is the first isoform expressed in developing OPCs, whereas expression and upregulation of the full-length *PLP1* occur coincidently with the maturation of OPCs to oligodendrocytes. Of note, mutations in the *PLP1*-specific region of exon 3 often manifest as mild PMD. However, apart from this observation, there are no clear correlations between mutation locus and disease severity (see GeneReviews in Web Resources).

This surfeit and variety of human mutations suggest that multiple pathogenic processes could be responsible for the diverse manifestations of PMD. In prior literature, five possible molecular defects have been ascribed to certain *PLP1* mutations: reduced expression, overexpression,[38] direct disruption of protein functional domains,[26] protein mistrafficking,[25,39,40] and protein misfolding leading to endoplasmic reticulum (ER) stress.[38,41–43] The occurrence of these defects individually or in combination most likely accounts for much of the clinical heterogeneity observed in PMD. However, because prior studies have largely focused on mutations one at a time, it is difficult to ascribe any findings to a mutation apart from that in which it was originally observed.

Replicating the efforts of the past 30 years of PMD research for each new *PLP1* mutation is a daunting proposition if left to traditional cellular approaches. PLP1 trafficking and membrane dynamics can be modeled to an extent with the use of immortalized cells, but the myelin sheath is a highly specialized membrane that cannot truly be recapitulated apart from oligodendrocytes. Access to primary human oligodendrocytes is severely lacking, however, because brain biopsies are implicitly dangerous and, in the case of developmental myelin disorders, the relevant stages of PMD pathogenesis have already occurred by the time a clinical diagnosis is made, let alone autopsy. As a result, animal models have proven indispensable for in vivo studies of myelin development but would be prohibitively expensive and time consuming on the scale necessary to span the genetic diversity of PMD.

Instead of attempting to adapt surrogate systems to model PMD heterogeneity, the advent of human induced pluripotent stem cell (hiPSC) and cell-fate reengineering technologies now provide us with robust methods for generating oligodendrocytes for large-scale studies directly in disease-relevant human-derived cells.[44,45] In the current study, we developed a hiPSC-based platform to efficiently model and functionally assess point mutations and duplication, triplication, and deletion of *PLP1* across 12 individuals with PMD. We utilized these hiPSCs to generate OPCs and oligodendrocytes from all 12 individuals in parallel for comparative molecular and cellular assessments. These studies establish a framework for classifying PMD subgroups on the basis of defects observed in disease-relevant cells, inform personalized therapeutics testing, and demonstrate the power of using hiPSC panels to model heterogeneity in a monogenic disease.

Material and Methods

Generation of hiPSCs

Skin fibroblast samples, de-identified except for mutation and clinical severity, were obtained from Coriell (PMD6), James Garbern, and G.M.H. Before receipt, fibroblasts had been isolated from skin biopsies, cultured for one to seven passages, and frozen. Upon receipt, samples were assigned arbitrary identifications (PMD1–PMD12) according to statutes for exempt human subjects research outlined by the institutional review board of Case Western Reserve University. In preparation for hiPSC generation, fibroblasts were thawed, expanded for two passages, and tested mycoplasma free.

hiPSCs were generated by standard approaches, including either a floxed polycistronic lentivirus (hSTEMCCA) encoding the pluripotency factors OCT3/4, SOX2, KLF4, and c-MYC (PMD1–PMD4, PMD6, and PMD8–PMD10)[46] or non-integrating episomal vectors encoding OCT3/4, SOX2, KLF4, L-MYC, LIN28, and a p53 shRNA (PMD5, PMD7, PMD11, and PMD12).[47] At least three independent hiPSC colonies were selected for clonal expansion according to colony morphology and OCT3/4 immunocytochemistry. Clones were subsequently split 1:6 every 4 or 5 days until they expanded sufficiently for the collection of DNA and RNA and freeze-down stocks.

Two independent clonal lines derived from each PMD sample were ultimately selected for further characterization and inclusion in these studies. Seven control human pluripotent lines (designated "NC") were also included: three approved human embryonic stem cell (hESC) lines from the NIH hESC Registry (NC1, "H1" NIHhESC-10-0043; NC2, "H7" NIHhESC-10-0061; NC3, "H9" NIHhESC-10-0062)[48] and four in-house-derived hiPSC lines from healthy donors (NC4–NC7).

DNA Isolation and Analyses

Genomic DNA was extracted from each pluripotent cell line either one or two passages before initiation of the oligodendrocyte differentiation protocol (see below) with the DNeasy Blood & Tissue Kit (69504, QIAGEN). Isolation was performed according to the manufacturer's protocol for cultured cells (July 2006 Handbook).

For Sanger sequencing of the *PLP1* coding sequence, individual PCR primer pairs were designed to encompass each exon of *PLP1* with NCBI Primer-BLAST (Table S1). Each exon was amplified with KAPA HiFi HotStart ReadyMix (07958935001, Roche) at the manufacturer's suggested reaction concentrations and cycling conditions: annealing temperature of 61°C, extension time of 15 s, and 30 cycles. PCR products were purified with the QIAquick PCR Purification Kit (28104, QIAGEN) and sequenced at the Case Western Reserve University Genomics Core facility. Reported *PLP1* mutations were validated in PMD samples, and all sequences were compared against the consensus human sequence (UCSC Genome Browser hg19).

High-density whole-genome SNP genotyping was performed with the Illumina Infinium Omni5 DNA Analysis BeadChip. Log R values were adjusted with the genomic.wave.pl program within PennCNV.[49] Ordered log R values of every coordinate were plotted for visualization of any large-scale copy-number variations present in each line.

For clonal confirmation and future disambiguation, cell-line identity was established and confirmed by short-tandem-repeat-based DNA fingerprinting in fibroblasts and derived hiPSC lines, as well as in control pluripotent lines (Cell Line Genetics).

RNA Isolation, RNA-Seq, and Expression and Splicing Analyses

Pluripotent cells were collected for RNA isolation simultaneously with initial passaging for oligodendrocyte differentiation (see below). 500,000 cells were pelleted at 1,200 rpm for 4 min, resuspended in 1 mL TRIzol (15596026, ThermoFisher), and immediately frozen at −80°C. Total RNA was isolated with the QIAGEN RNeasy Mini Kit (74104, QIAGEN), including minor modifications to the beginning of the manufacturer's protocol ("Purification of Total RNA from Animal Cells using Spin Technology," Fourth Edition, June 2012 RNeasy Mini Handbook). In brief, cells in TRIzol were thawed on ice, vortexed until homogeneous, and incubated at room temperature for 5 min. 200 μL chloroform was added and mixed vigorously. The sample was transferred to PhaseLock gel tubes (2302830, 5 Prime), incubated at room temperature for 3 min, and then centrifuged at 12,000 × *g* for 15 min at 4°C. The aqueous phase was collected, and 1.5 volumes of 100% ethanol was added and mixed thoroughly. The sample was then transferred to an RNeasy Mini column, and the remainder of the protocol, including the recommended DNase digest, was followed as written.

For generation of the cDNA library for RNA sequencing (RNA-seq), 1 μg of each sample was rRNA depleted (Ribo-Zero Gold rRNA Removal Kit, MRZG12324, Illumina), fragmented, and indexed with the TruSeq Stranded mRNA Library Preparation Kit (RS-122-2103, Illumina) per the manufacturer's protocol. 100 bp paired-end reads were generated for each sample with an Illumina HiSeq 2500 at the Case Western Reserve University Genomics Core facility. Output FASTQ files were aligned to the hg19 genome with TopHat (version 2.0.8) with default settings.[50] Data were normalized, and fragments per kilobase per million reads (FPKM) were calculated for known RefSeq genes with Cufflinks (version 2.0.2).[51] With the "heatmap.2" function of the gplots R package, Pearson's correlation distance was calculated for comparing transcriptome similarities between individual cell lines. For analysis of *PLP1* mRNA splicing, aligned BAM files were loaded into the Integrated Genomics Viewer (version 2.3.68 [97]) and visualized with the "sashimi plot" function.

OPC and Oligodendrocyte Differentiation

Differentiation of OPCs and oligodendrocytes was performed on two independently derived hiPSC clones for each of the 12 PMD samples (n = 2 biological replicates per mutation) and three hESC and four hiPSC normal controls (n = 7 biological replicates identical in *PLP1* sequence). The entire panel of 31 lines was differentiated simultaneously with two technical replicates per line.

OPCs and oligodendrocytes were generated from pluripotent cells according to a pre-publication version of the Douvaras et al. protocol.[44,52] Minor variations from the published protocol are noted here.

Immediately before differentiation, cells were incubated in 10 μM Y-27632, dissociated with collagenase and dispase, and plated at 200,000 cells per 9.5 cm^2 Matrigel-coated well in mTeSR1 medium and 10 μM Y-27632. Cells were cultured for 2 days, during which mTeSR1 was changed each day. Differentiation (protocol day 0) began on the third day after passaging and was conducted as follows:

Days 0–7: cells underwent daily complete media changes of DMEM/F12 (11320-033, GIBCO), 1× high-insulin N2 supplement (N2; AR009, R&D Systems), 10 μM SB431542 (04-0010, Stemgent), 250 nM LDN189193 (04-0074, Stemgent), 100 nM all-trans retinoic acid (RA; R2625, Sigma), and 5 U/mL penicillin and streptomycin (PenStrep).

Days 8–11: cells underwent daily complete medium changes of DMEM/F12, 1× low-insulin N2 (17502048, Life Technologies), 100 nM RA, 1 μM smoothened agonist (SAG; 566660, EMD Millipore), and 5 U/mL PenStrep.

Days 12–19: on day 12, cells were manually lifted with a cell scraper, broken into approximately 10- to 50-cell clusters, and plated into ultra-low attachment plates (3471, Corning) for promoting the formation of free-floating neurospheres. Every other day, spheres underwent two-thirds medium changes of DMEM/F12, 1× N2, 1× B27 supplement without vitamin A (B27; 12587010, Life Technologies), 100 nM RA, 1 μM SAG, and 5 U/mL PenStrep.

Days 20–30: every other day, cells underwent two-thirds medium changes of DMEM/F12, 1× N2, 1× B27, 10 ng/mL platelet-derived growth factor (PDGF; 221-AA, R&D Systems), 10 ng/mL insulin-like growth factor 1 (IGF; 291-G1, R&D Systems), 5 ng/mL hepatocyte growth factor (HGF; 294-HG, R&D Systems), 10 ng/mL neurotrophin-3 (NT3; GF031, EMD Millipore), 60 ng/mL 3,3′,5-triiodothronine (T3; ST2877, Sigma), 100 ng/mL biotin (4639, Sigma), 1 μM cyclic-AMP (cAMP; D0260, Sigma), 25 μg/mL insulin (I9278, Sigma), and 5 U/mL PenStrep.

Days 30–60: on day 30, neurospheres were plated onto 0.1 mg/mL poly-L-ornithine- (P3655, Sigma) and 10 μg/mL laminin-coated (L2020, Sigma) 9.5 cm^2 plates and allowed to attach. Every other day, cells underwent two-thirds medium changes of DMEM/F12, 1× N2, 1× B27, 10 ng/mL PDGF, 10 ng/mL IGF, 5 ng/mL HGF, 10 ng/mL NT3, 60 ng/mL T3, 100 ng/mL biotin, 1 μM cAMP, 25 μg/mL insulin, and 5 U/mL PenStrep.

After Day 60, cultures could either be maintained in PDGF-containing medium for the promotion of OPC proliferation or be transitioned to PDGF-, IGF-, or HGF-free medium for further differentiation of oligodendrocytes.

For oligodendrocyte phenotyping of the entire panel of PMD and control cell lines in Figure 4, on day 96, cultures were dissociated with Accutase (AT-104, Innovative Cell Technologies) for 40 min and split 1:6 into 96-well poly-D-lysine VisiPlates (1450-605 PerkinElmer) pre-incubated for 1 hr with 10 μg/mL laminin. Every other day, cells underwent two-thirds medium changes of DMEM/F12, 1× N2, 1× B27, 10 ng/mL NT3, 60 ng/mL T3, 100 ng/mL biotin, 1 μM cAMP, 25 μg/mL insulin, and 5 U/mL PenStrep.

For oligodendrocyte differentiation of NC2, PMD2, and PMD10 for time-lapse imaging (Movies S1, S2, and S3), small-molecule testing (Figure 5), and neuron co-culture (Figure 6), differentiation was repeated as above through day 60. After day 60, cultures were transitioned to two-thirds medium changes of DMEM/F12, 1× N2, 1× B27, 60 ng/mL T3, 100 ng/mL biotin, 1 μM cAMP, 25 μg/mL insulin, 10 mM HEPES sodium salt (H3784, Sigma), 20 μg/mL L-ascorbic acid (A4544, Sigma), and 5 U/mL PenStrep every third day.

Flow Cytometry

OPC differentiation cultures were incubated in pre-warmed Accutase for 40 min at 37°C until cells lifted off the plate. Lifted cultures were diluted with DMEM/F12 supplemented with 1% bovine serum albumin (BSA; 15260-037, GIBCO) and gently pipetted with a 1,000 μL capacity tip for dissociating the cells that had grown out from neurospheres without shearing the OPCs. The spheres themselves remained intact and were removed and discarded. Remaining single cells were counted and centrifuged at 200 × *g* for 5 min at room temperature. Cells were resuspended at up to 10 × 10^6 cells per 100 μL in DMEM/F12 supplemented with 5% donkey serum and PE-conjugated anti-PDGFRA (CD140a, 1:50; 556002, BD Biosciences) and incubated on ice for 45 min. Cells were washed three times with DMEM/F12 supplemented with 1% BSA, centrifuged at 200 × *g* for 5 min at room temperature, resuspended in 100 μL DMEM/F12 supplemented with 5% donkey serum and APC-conjugated anti-A2B5 (1:11; 130093582, Miltenyi), and incubated on ice for 45 min. Cells were washed three times with DMEM/F12 and 1% BSA; centrifuged at 200 × *g* for 5 min at room temperature; resuspended in 500 μL DMEM/F12, 1× N2, 1× B27, 10 ng/mL PDGF, 10 ng/mL IGF, 5 ng/mL HGF, 10 ng/mL NT3, 60 ng/mL T3, 100 ng/mL biotin, 1 μM cAMP, 25 μg/mL insulin, and 5 U/mL PenStrep; and filtered through a cell strainer. Cells were flowed with a 100 μm nozzle on a FACSAria (BD Biosciences) at 25 psi and 1,500–1,800 events per second. 10,000 events were recorded. Data were analyzed with WinList 3D (version 7.0). Initial debris and doublet gates were set with unstained NC2-derived cultures and validated against unstained NC6-derived cultures. Gates were set on the basis of side scatter (SSC-A) or forward scatter (FSC-A) for distinguishing live cells from dead cells and debris and then side-scatter width (SSC-W) or side-scatter height (SSC-H) for excluding cell doublets. A bona fide, $CD140a^+$ OPC population was initially gated on the basis of immunostained NC2-derived cultures, validated against the remaining six immunostained control cultures, and only then used for evaluating the number of derived OPCs in each PMD culture.

Small Molecules

GSK2656157 (5046510001, EMD Millipore) 10 mM stock solution in DMSO was prepared, aliquoted, and stored at −20°C. Guanabenz (0885, Tocris Bioscience) 20 mM stock solution in DMSO was prepared, aliquoted, and stored at −20°C. Small molecules were warmed to 37°C for 20 min before being added to pre-warmed medium. Frozen aliquots were thawed no more than twice before being discarded. During treatments, every 3 days, cells underwent a two-thirds medium change that included 100 nM GSK2656157, 1 μM GSK2656157, or 2.5 μM guanabenz.

Oligodendrocyte and Dorsal Root Ganglion Neuron Co-culture

24-well VisiPlates (1450-605, PerkinElmer) were pre-incubated with 125 μL rat tail collagen and allowed to air dry for 72 hr. 50,000 dorsal root ganglion neurons (DRGs)[53] were plated in 100 μL of M1 medium (MEM [11095-080, GIBCO], 10% FBS, and 2% glucose [G7528, Sigma]) with 100 ng/mL NGF (556-NG, R&D Systems) and 5 U/mL PenStrep in the center of the well; bubbles were added with a p100 pipette; and DRGs were allowed to attach for 24 hr at 37°C. The next day, wells were flooded with 500 μL E2F medium (MEM, 2% glucose, 1× N2 supplement, 245 ng/mL FDU [F0503, Sigma], and 245 ng/mL uridine [U3003, Sigma]) with 100 ng/mL NGF and 5 U/mL PenStrep. Medium was changed every second or third day thereafter (M1, NGF, and PenStrep on days 3, 7, 11, 14, and 17; E2F, NGF, and PenStrep on days 5 and 9). On day 20, one-half of a 9.5 cm^2 well of differentiating OPC cultures was seeded onto DRGs in DMEM/F12, 1× N2, 1× B27, and 5 U/mL PenStrep. Medium was changed every other day for 20 days and then fixed for immunostaining.

Live-Cell Imaging

Cultures of NC2, PMD2, and PMD10 OPCs were differentiated to oligodendrocytes over the course of 20 days as specified above. On day 20, cultures were dissociated and plated onto 0.1 mg/mL poly-L-ornithine- and 10 μg/mL laminin-coated 4 cm^2 plates at low density (split 1:6 by surface area) to permit observation of individual cells. The next day, cultures were transferred to a humidity-, temperature-, and CO_2-controlled chamber on an inverted microscope (DMI6000, Leica) for phase imaging. Regions of interest were identified manually and marked with Leica Application Suite X software, after which they were automatically imaged every 10 min for the next 60 hr. Static image series were stitched into a movie with Microsoft Windows Movie Maker (version 2012).

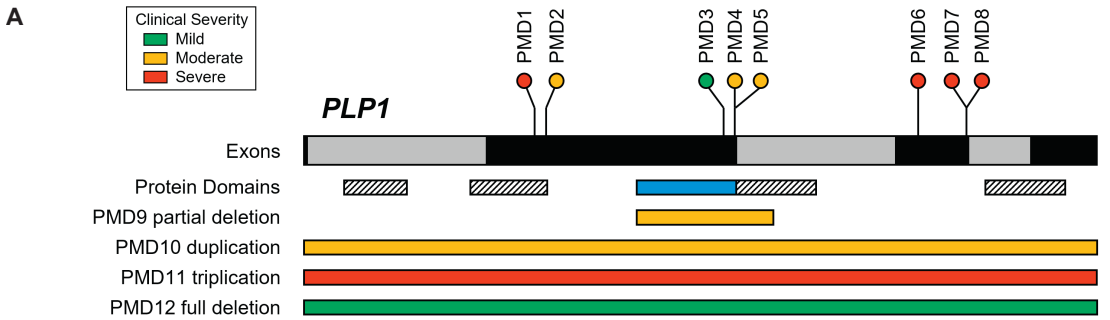
Immunocytochemistry

Cultures for immunocytochemistry were initially fixed with 4% ice-cold paraformaldehyde for 15 min.

O4 immunostaining was performed on live cells prior to fixation. O4 antibody supernatant was added to cultures and incubated for 30 min at 37°C. Wells were washed three times with room-temperature DMEM/F12 and then immediately fixed. Additional immunostaining was performed as below.

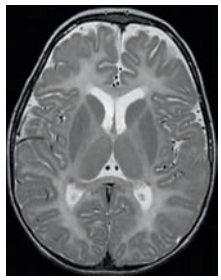
Monolayer cell cultures (e.g., hiPSCs and OPCs) were permeabilized with 0.2% Triton X for 10 min at room temperature, blocked in 10% donkey serum in PBS for 1 hr at room temperature, incubated in primary antibody in blocking buffer for 1 hr at room temperature (typically) or overnight at 4°C (for PLP1 [AA3] antibody only), washed three times with PBS, incubated in secondary antibody in blocking buffer for 45 min at room temperature, washed three times (with DAPI, when used, in the first wash to stain nuclei), and imaged with an Operetta High-Content Imaging System with Harmony Analysis Software (HH12000000, PerkinElmer) and standard fluorescence settings.

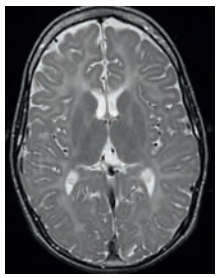
DRG co-cultures to be immunostained for PLP1 were permeabilized with 10% Triton X for 30 min at room temperature, washed three times with PBS, blocked in 5% donkey serum and 0.1% Triton X in PBS for 1 hr at room temperature, incubated in PLP1 antibody in blocking buffer overnight at 4°C, and then incubated in neurofilament (NF) and SOX10 antibodies in blocking buffer for 1 hr at room temperature, washed three times with PBS, incubated in secondary antibodies in blocking buffer for 45 min at room temperature, washed three times, and imaged with an inverted fluorescence microscope (Leica DM IL LED), 12-bit monochrome camera (QIC-F-M-12-C, QImaging), and QCapture Pro imaging software (version 6.0.0.605, QImaging).



B

Sample	*PLP1* Mutation	Amino Acid	FDS	Severity	Reprog.	Clones
PMD1	c.242T>G	p.Leu81Arg	4	Sev	Lentiviral	CWRU 133 CWRU 134
PMD2	c.254T>G	p.Leu85Arg	10	Mod	Lentiviral	CWRU 111 CWRU 113
PMD3	c.441A>T	p.Gly147=		Mild	Lentiviral	CWRU 129 CWRU 130
PMD4	c.453G>T	p.Lys151Asn		Mod	Lentiviral	CWRU 120 CWRU 122
PMD5	c.453+750 G>A	p.=	11	Mod	Episomal	CWRU 251 CWRU 252
PMD6	c.646C>T	p.Pro216Ser		Sev	Lentiviral	CWRU 72 CWRU 75
PMD7	c.762+3G>T	p.Phe233_Leu254del	3	Sev	Episomal	CWRU 261 CWRU 264
PMD8	c.762+3G>T	p.Phe233_Leu254del	3	Sev	Lentiviral	CWRU 211 CWRU 212
PMD9	c.352_499del	p.Thr118_Val166del		Mod	Lentiviral	CWRU 235 CWRU 239
PMD10	Duplication	2 copies	9	Mod	Lentiviral	CWRU 227 CWRU 231
PMD11	Triplication	3 copies	2	Sev	Episomal	CWRU 270 CWRU 271
PMD12	Deletion	No *PLP1*	18	Mild	Episomal	CWRU 272 CWRU 273





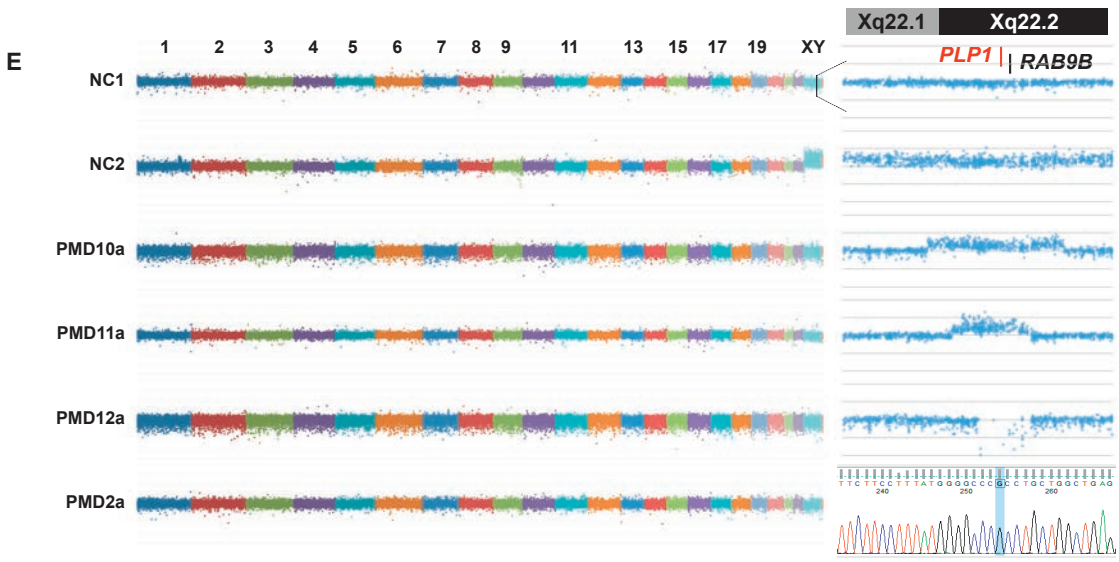


Figure 1. Genetic Characterization of a PMD hiPSC Panel

(A) A schematic of *PLP1* and the 12 mutations included in this study. Full-length *PLP1* consists of seven exons (black and gray bars), whereas its splice isoform, *DM20*, results from exclusion of the *PLP1*-specific domain (blue bar). Both isoforms contain four putative transmembrane domains (striped bars). The locations of individual mutations are indicated as lollipop plots (point mutations) or bars (partial deletion and copy-number variants). Relative clinical severities are indicated by color (green, mild; yellow, moderate; red, severe).

(B) Individuals were selected for this study with the intent of maximizing genetic and phenotypic diversity. Clinical severities had been previously assessed and reported by functional disability score (FDS) and/or clinical impression. Skin fibroblasts were reprogrammed to

(legend continued on next page)

DRG co-cultures to be immunostained for myelin basic protein (MBP) were washed three times in PBS, permeabilized with 100% ice-cold methanol for 30 min at −20°C, washed three times with PBS, blocked in 5% donkey serum and 0.1% Triton X in PBS for 1 hr at room temperature, incubated in MBP, NF, and SOX10 antibodies in 2% donkey serum and 0.1% saponin overnight at 4°C, washed three times with PBS, incubated in secondary antibodies in 10% donkey serum and 0.1% Triton X for 1 hr at room temperature, washed three times, and imaged as above.

Primary antibodies included mouse-anti-O4 (1:10 unconcentrated supernatant, generously provided by Bruce Trapp, Robert Miller, and Wendy Macklin), OCT3/4 (400 ng/mL; SC-5279, Santa Cruz), Pax6 (6.67 μg/mL; PRB-278P, Covance), SOX1 (1 μg/mL; AF3369, R&D Systems), OLIG2 (1:1,000; AB9610, Millipore), NKX2.2 (1:100; 74.5A5, Developmental Studies Hybridoma Bank), rat-anti-PLP1 (1:100; AA3, generously provided by Bruce Trapp), rat-anti-MBP (1:100; AB7349, Abcam), goat-anti-SOX10 (2 μg/mL; AF2864, R&D Systems), mouse-anti-pan-axonal NF (1:1,000; SMI311, Covance), mouse-anti-pan-neuronal NF (500 ng/mL; SMI312, Covance), and DAPI (1 μg/mL; D8417, Sigma).

All secondary antibodies were Alexa-Fluor-conjugated secondary antibodies (Life Technologies) used at a dilution of 1:500.

Results

Assembly, Generation, and Characterization of a Panel of PMD-Derived hiPSCs

The goal of this study was to establish a platform for assessing the developmental, cellular, and molecular defects caused by PMD-relevant *PLP1* mutations in human-derived oligodendrocytes. In order to capture the genetic and phenotypic heterogeneity found in PMD, we selected samples for inclusion according to three criteria: type of mutation, distribution of point mutations throughout *PLP1*, and reported clinical severity. Before being included in this study, all individuals had been diagnosed with PMD clinically and had *PLP1* mutations confirmed by genetic testing. We obtained primary fibroblasts de-identified except for their mutation and clinical severity impression (mild, moderate, or severe) or functional disability score (ranging from 1 [most severe] to 32 [least severe]).[54] Our panel ultimately consisted of 12 lines with various *PLP1* mutations (Figures 1A and 1B and Table S2) and seven normal control lines (Table S3).

For two individuals, PMD2 and PMD10, axial T2-weighted MRI taken when they were ages 4 and 12 years, respectively, was also available (Figures 1C and 1D). Both children exhibited diffusely increased signal intensity in white-matter structures and atrophy of the subcortical white matter. The gross reduction in white matter, particularly in PMD2, resulted in moderate enlargement of the lateral ventricles. Such MRI is highly representative of children with moderate to severe PMD and demonstrates both the ambiguity and convergence of clinical presentations across people with disparate *PLP1* mutations.[54–56]

To generate a renewable source of PMD-derived cells, we reprogrammed fibroblasts from all 12 individuals to hiPSCs (see Material and Methods). Two independently derived hiPSC lines per individual were ultimately selected for rigorous characterization. These 24 PMD lines (PMD1–PMD12, A and B), along with four hiPSC lines from healthy individuals (NC4–NC7) and three normal human embryonic stem cell (hESC) lines (NC1–NC3), constitute the "panel" of 31 pluripotent stem cell lines used throughout the following experiments.

Initially, we rigorously characterized each line in the panel to ensure its identity, pluripotency, and genomic integrity. To first validate the *PLP1* point mutations reported for each PMD line and to confirm that no additional mutations were present in any PMD or control line, we Sanger sequenced all seven exons of *PLP1* for each line in the panel. PMD7 and PMD8 were found to contain a nonpathogenic synonymous SNP (c.609T>C [p.Asp203=]) that is common in the general population (rs1126707, C = 22.6%).[57] All other lines conformed to the consensus human sequence (GenBank: NM_000533.4).[58]

Because the process of hiPSC reprogramming can occasionally induce chromosomal defects, copy-number variation in each cell line was evaluated at fine resolution with a high-density SNP microarray. Comparison of the relative copy number of each SNP confirmed the absence of any gross chromosomal duplications or deletions in all lines (Figures 1E and S1). Furthermore, this resolution allowed delineation of the relative sizes and boundaries of the *PLP1* locus duplication, triplication, and deletion in PMD10, PMD11, and PMD12, respectively (Figure 1E, right).

To confirm that our cell lines had been completely reprogrammed and retained their pluripotent properties throughout subsequent expansion and characterization, one passage before oligodendrocyte differentiation, we isolated, sequenced, and compared RNA from each cell line against the RNA profiles of primary fibroblasts from which control hiPSCs had been derived. Hierarchical clustering

hiPSCs with either a lentiviral or episomal construct. Two independently derived clonal hiPSC lines were selected for characterization and inclusion in subsequent studies. *PLP1* mutations were confirmed by Sanger sequencing of each hiPSC line.

(C) Taken at age 4 years, T2-weighted MRI of PMD2 demonstrates increased signal intensity throughout white-matter structures and enlargement of the lateral ventricles.

(D) Taken at age 10 years, T2-weighted MRI of PMD10 demonstrates increased signal intensity and distinct atrophy of white-matter structures.

(E) Plots demonstrating the gross genomic integrity of derived pluripotent lines. Relative copy number was calculated for each SNP in a high-density SNP microarray and plotted as a normalized log R ratio. (Left) Plots of every 100th SNP, arranged by ranked genomic coordinate and colored by chromosome. NC1 (male) and NC2 (female) demonstrate the relative enrichment of the X chromosome in NC2. (Top right) Plots of each SNP within a 2 Mb region surrounding *PLP1* on the X chromosome, arranged by relative genomic coordinate. For PMD10, PMD11, and PMD12, the SNP array delineates the region of chromosome X duplication, triplication, and deletion, respectively. (Bottom right) A Sanger sequencing trace showing the T-to-G substitution found in PMD2.

demonstrated close association of all pluripotent lines, and there was no significant distinction between PMD and control lines, hiPSCs and hESCs, or out-grouping of both PMD-derived lines from any single individual (Figure 2B). All pluripotent lines also showed robust and consistent expression of canonical pluripotency markers whose expression correlates with complete reprogramming and acquisition of pluripotent identity (Figure 2C).[59] Using this rigorous pipeline, we generated and characterized a diverse panel of hiPSCs to provide a cellular resource for the study of PMD.

Assessment of *PLP1* Transcript Dynamics and Defects in hiPSCs

Endogenous expression of PLP1 and DM20 is restricted to oligodendrocytes and OPCs, respectively. As a result, studies on the effects of specific human mutations on protein structure and expression have previously been limited to post mortem tissue or transgenic overexpression. However, although the protein is not translated, *DM20* mRNA is robustly transcribed in pluripotent stem cells.[60] Serendipitously, this provides an opportunity for rapid assessment of specific transcript defects without the protracted differentiation of oligodendrocytes. To begin to characterize the effects of mutations in our panel, we used our RNA-seq dataset to interrogate *DM20* mRNA expression and splicing directly in hESCs and hiPSCs (Figure 2A).

Comparison of mRNA transcript levels between control and PMD-derived pluripotent lines provides a glimpse into the effects of copy-number variations at the *PLP1* locus. Control cultures, both hiPSC and hESC, displayed an average *DM20* expression level of 18.1 ± 4.5 FPKM (Figure 2D). Similarly, all point mutations (PMD1–PMD8) and the partial deletion (PMD9) showed no significant differences in transcript levels (average FPKMs of 18.1 ± 2.5 and 15.5 ± 1.2, respectively; Figure 2D). However, PMD10 and PMD11 expressed 2- and 3-fold more *DM20* than controls (32.7 ± 5.5 and 55.1 ± 1.8, respectively), consistent with their respective duplication and triplication of the *PLP1* locus (Figure 2D). As expected, the deletion (PMD12) showed no expression of *DM20* (Figure 2D).

In addition to permitting quantification of expression levels, the presence of *DM20* mRNA in pluripotent cells also allows for identification of mutation-specific splicing defects. In all controls, exon 3 terminated at the internal *DM20* splice site (exon 3a), indicating that only the shorter, OPC-specific isoform is present in pluripotent stem cells (Figure 2E). This preempts appreciation of any splicing defects in PMD3–PMD5, whose mutations fall in exon 3b and intron 3 (Figure S2). Additionally, PMD1, PMD2, PMD6, and PMD10–PMD12 showed no defects or alternative splicing (Figure S2) but would not particularly be expected to, considering the nature of their exon and copy-number mutations. However, splicing analysis in PMD7 and PMD8, brothers with an intronic mutation outside the canonical splice site, presented with complete skipping of the exon preceding their *PLP1* mutation (Figure 2E), confirming prior analyses from autopsy tissues.[61] Lastly, PMD9, a partial deletion spanning exons 3b, intron 3, and part of exon 4, demonstrated the expected in-frame deletion of the proximal portion of exon 4, and no additional splicing defects were found (Figure 2E).

Assessment of OPC Production in PMD Cultures

We next wanted to determine whether we could use our genetically diverse panel of PMD-derived hiPSCs to garner insights into the clinical variability of PMD in disease-relevant cells (Figure 3A). Our pluripotent stem cell panel was differentiated to OPCs over a 90 day time course according to a protocol that recapitulates in vivo neurodevelopmental transitions including patterning of neuroectoderm, ventralization, OPC specification, and OPC proliferation (Figure 3B).[52] We performed differentiation and all subsequent experiments in parallel for all 12 PMD samples (in duplicate with two independently derived hiPSC lines per individual) and seven control samples to enable direct comparison of data and minimize the influence of variables, such as reagent lots, ambient conditions, or handling.

Throughout the process of differentiation, cultures were immunostained for markers of key stages in the development of oligodendrocytes. Day 6 immunostaining for the neural lineage transcription factors PAX6 and SOX1 demonstrated efficient induction across both control (95% ± 1.9%) and PMD (94% ± 1.4%) cultures (Figures 3C and 3D). Day 12 immunostaining for the early glial lineage transcription factors OLIG2 and NKX2.2 also showed strong induction consistent across control (73% ± 9.2%) and PMD (77% ± 2.8%) cultures (Figures 3C and 3E). These data demonstrate that despite the presence of *PLP1* mRNA transcripts in hiPSCs, early neurodevelopmental differentiation appears to be unaffected by mutations in *PLP1*.

We maintained cultures an additional 11 weeks to allow OPC specification, at which point we quantified the cultures by flow cytometry for the percentage of platelet-derived growth factor receptor alpha $(PDGFRA)^{+}$ OPCs in each culture.[62] The average proportion of $PDGFRA^{+}$ OPCs was significantly lower in PMD cultures (24% ± 3.7%) than in controls (49% ± 3.3%) (Figure 3F). However, although the proportion of OPCs was generally consistent between both hiPSC clones derived from any given individual, there was substantial variability between cultures derived from separate people (Figure 3G). OPC numbers were strikingly reduced in a majority of PMD cultures (PMD2–PMD4, PMD6–PMD8, and PMD10–PMD12), whereas only four cultures (PMD1, PMD5, PMD9, and PMD11) contained OPCs at proportions comparable to those in controls. Intriguingly, PMD3, from a mildly affected individual with a synonymous substitution, demonstrated the greatest depletion of OPCs. Similarly, PMD12, from a mildly affected individual with a *PLP1* deletion, also displayed poor OPC numbers, whereas more severely affected individuals possessing duplication and triplication of *PLP1* (PMD10 and PMD11) trended toward normal numbers of OPCs.

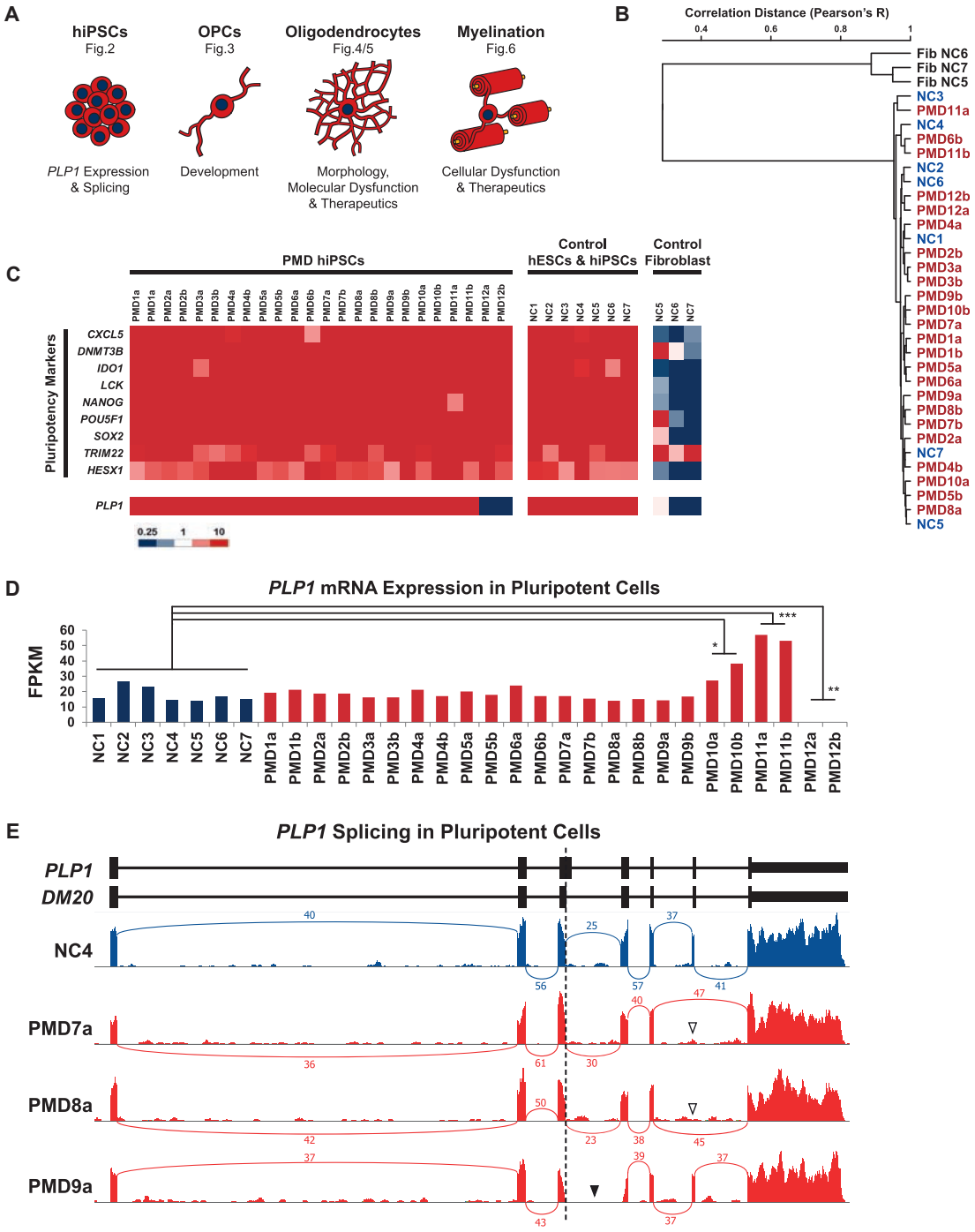


Figure 2. RNA Characterization of PMD hiPSCs and *PLP1* Transcript Defects

(A) A schematic of the major cell types derived in this study, the stereotypic morphologic appearance of each cell type when cultured in vitro, some of the insights these cells can provide, and the figure(s) in which they feature.

(B) Dendrogram depicting hierarchical clustering analysis of stranded RNA-seq. RNA was isolated from each pluripotent line one passage before initiation of the OPC differentiation protocol and compared against RNA isolated from primary fibroblasts corresponding to NC5–NC7.

(C) A heatmap depicting the FPKM of canonical pluripotency genes and *PLP1* across all PMD hiPSCs, normal controls, and primary fibroblasts corresponding to NC5–NC7.

(D) A bar graph comparing levels of *PLP1* mRNA expression in hiPSCs between various PMD cultures and controls (*$p = 0.0134$, **$p = 0.0017$, ***$p < 0.0001$).

(E) Sashimi plots of RNA-seq transcripts aligning to *PLP1* (hg19) quantify *PLP1* and *DM20* mRNA splicing events. Numeric labels indicate the number of quality-filtered transcripts (sequencing depth) that span the indicated exon-exon junction. Exclusion of the distal portion of *PLP1* exon 3 (dotted vertical line) in control transcripts indicates that *DM20* is the solely expressed isoform in pluripotent cells. PMD7 and PMD8 demonstrate skipping of exon 6 (white arrowheads). The exon 3–4 junction cannot be annotated in PMD9 (black arrowhead) because of its partial deletion, which spans the *PLP1*-specific region of exon 3 and proximal portion of exon 4.

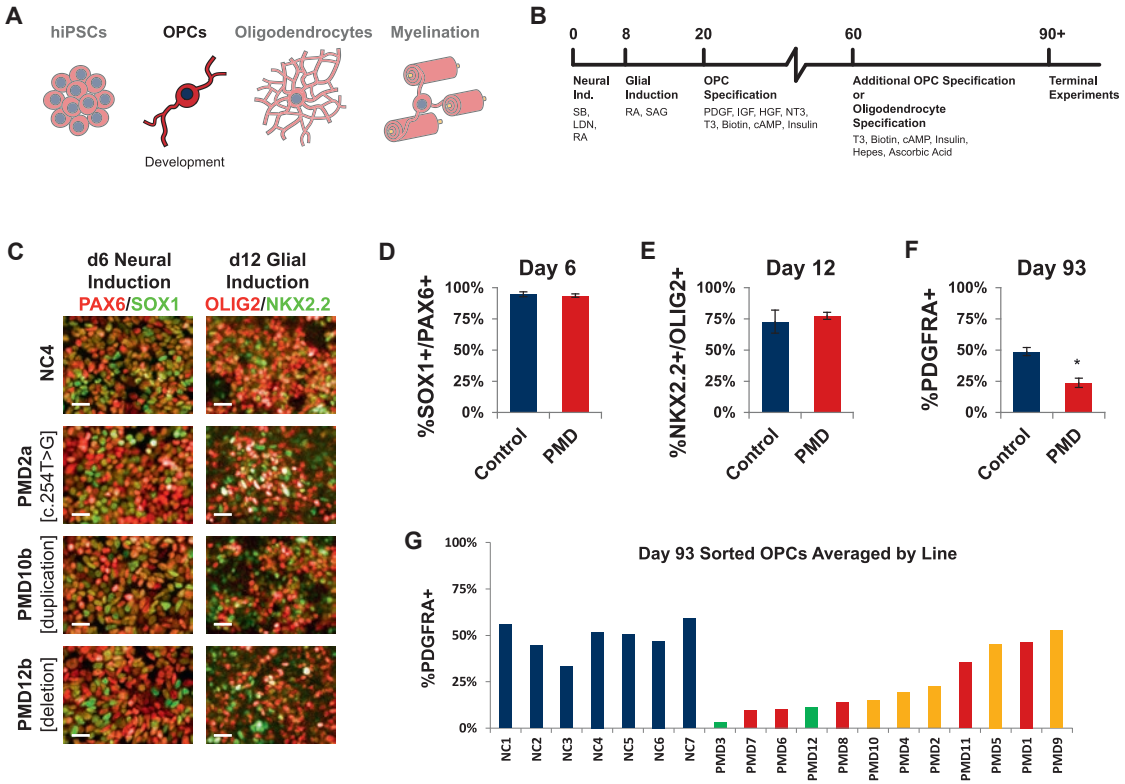


Figure 3. Differentiation to OPCs Demonstrates PMD Variability

(A) Schematic of the experimental stage, OPC morphology, and insights presented in this figure.

(B) Overview of the timeline, small molecules, and growth factors used for generating OPCs and oligodendrocytes.

(C) Representative immunofluorescence images comparing stage-specific transcription factors in PMD and control cultures on day 6 (immunostained for neural lineage markers PAX6 and SOX1) and day 12 (immunostained for glial lineage markers OLIG2 and NKX2.2) of the differentiation protocol. Scale bars represent 25 μm.

(D) Quantification of the percentage of $PAX6^{+}SOX1^{+}$ cells as of 6 days in culture. Shown here are the averages of all controls (n = 7) versus all PMD lines (n = 24, including n = 2 biologic replicates per PMD line). Two independently differentiated wells per line were immunostained (n = 2 technical replicates). Error bars indicate standard error of the mean. No significant difference was found between control and PMD lines.

(E) Quantification of the percentage of $OLIG2^{+}NKX2.2^{+}$ cells as of 12 days in culture. Shown here are the averages of all controls (n = 7) versus all PMD lines (n = 24, including n = 2 biologic replicates per PMD line). Two independently differentiated wells per line were immunostained (n = 2 technical replicates). Error bars indicate standard error of the mean. No significant difference was found between control and PMD lines.

(F) Day 93 cultures were immunostained for the OPC-specific marker PDGFRA and counted by flow cytometry. Shown here are the averages of all controls (n = 7) versus all PMD lines (n = 24, including n = 2 biologic replicates per PMD line). One well per line was counted (n = 1 technical replicate). Error bars indicate standard error of the mean (*p = 0.0016).

(G) The same results from (F), plotted here as individual controls versus the average of both hiPSC lines derived from a given PMD sample. PMD results are rank ordered by average number of OPCs and colored according to clinical severity (green, mild; yellow, moderate; red, severe).

Classes of Cellular and Molecular Defects in PMD Oligodendrocytes

We next wanted to determine whether specific defects would manifest as the OPCs matured into oligodendrocytes (Figure 4A). We induced PMD and control OPCs to mature into pre-myelinating oligodendrocytes and assessed cell morphology by immunostaining for O4 antigen (an early oligodendrocyte-specific surface sulfatide) and PLP1 (with a C-terminal antibody that defines a more mature oligodendrocyte stage). A typical wild-type oligodendrocyte generated in cell culture has a readily identifiable morphology consisting of a round, central cell body with multiple branching processes that extend symmetrically outward and lend the oligodendrocyte a spider-in-a-web-like appearance (Figure 4B, "controls").

We made two major observations in these cultures. First, despite OPC deficits, most PMD cultures (except PMD5 and PMD11) were still capable of producing oligodendrocytes. However, all PMD-derived oligodendrocytes were noticeably defective (Figure 4B). In order to elucidate the defects in these cells, we developed a machine-learning-based algorithm by using PerkinElmer Harmony High-Content Imaging and Analysis software to trace and measure oligodendrocyte processes and identify branch points (Figure 4C). In PMD oligodendrocytes, total process length was significantly lower than in controls (Figure 4D). Although PMD and control oligodendrocytes extended a similar number

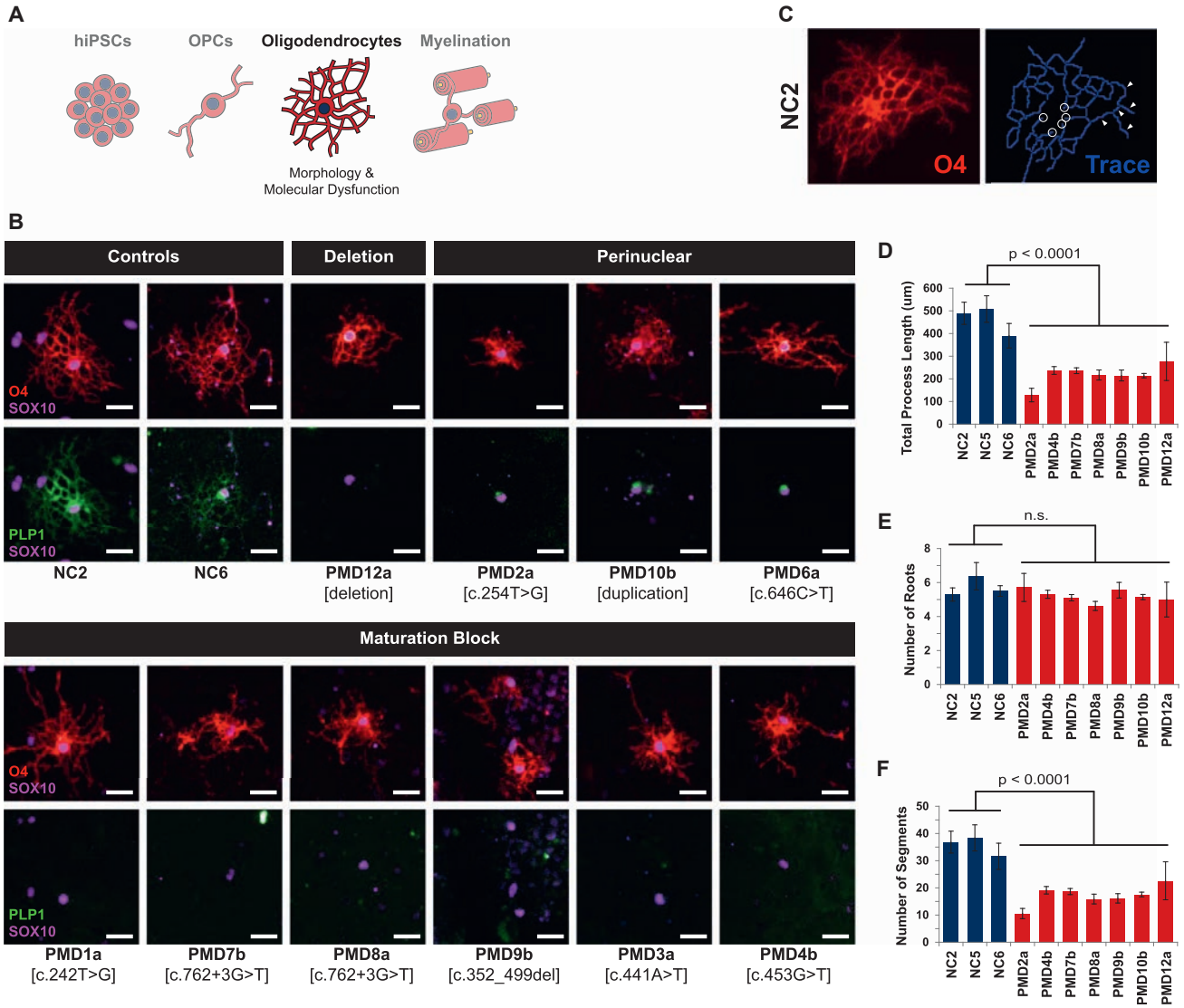


Figure 4. Classifications of Oligodendrocyte Phenotypes

(A) Schematic of the experimental stage, oligodendrocyte morphology, and insights presented in this figure.

(B) OPCs from one line per PMD sample (n = 1 biologic replicate) were differentiated to oligodendrocytes (n = 2 technical replicates) and immunostained for oligodendrocyte markers O4, SOX10, and PLP1. Shown here are representative images from each culture. PMD5 and PMD11 failed to produce any $O4^+$ cells. PMD2, PMD6, and PMD10 demonstrated perinuclear retention of PLP1. The remaining lines produced $O4^+$ cells but failed to produce a PLP1 signal. Scale bars represent 25 μm.

(C) A representative immunofluorescence image of an $O4^+$ oligodendrocyte from the NC2 control line (left) and a trace of its processes (right) generated by an oligodendrocyte identification, tracing, and quantification algorithm derived with PerkinElmer Harmony software. White circles highlight examples of "roots" where processes contact the cell body. White arrows indicate examples of individual "segments" between process branch points.

(D) The processes of $O4^+$ oligodendrocytes were traced and measured by a machine-learning algorithm derived in house. Total process length was calculated as an average across individually measured oligodendrocytes. Error bars indicate standard error of the mean.

(E) Junctions of the oligodendrocyte cell body and extending processes were identified and counted by the tracing algorithm. The total number of roots was calculated as an average across individually traced oligodendrocytes. No significant difference was found between control and PMD lines. Error bars indicate standard error of the mean.

(F) Segments, defined as a linear portion of process between any two intersections (branches) in the trace, were identified and counted by the tracing algorithm. The total number of segments was calculated as an average across individually traced oligodendrocytes. Error bars indicate standard error of the mean.

of primary processes from the cell body (Figure 4E), the number of distal process branches was severely reduced (Figure 4F). Interestingly, oligodendrocytes of PMD12, the full *PLP1* deletion with a mild phenotype, demonstrated both the highest average and widest range of total process length and branches (Figures 4B, 4D, and 4F). Collectively, these data suggest that PMD oligodendrocytes suffer either a PLP1-induced defect of process extension and branching or a non-specific arrest of maturation as a result of general disruption of cellular homeostasis. Time-lapse imaging of maturing PMD2 and PMD10 cultures captured oligodendrocytes producing short processes that failed to extend

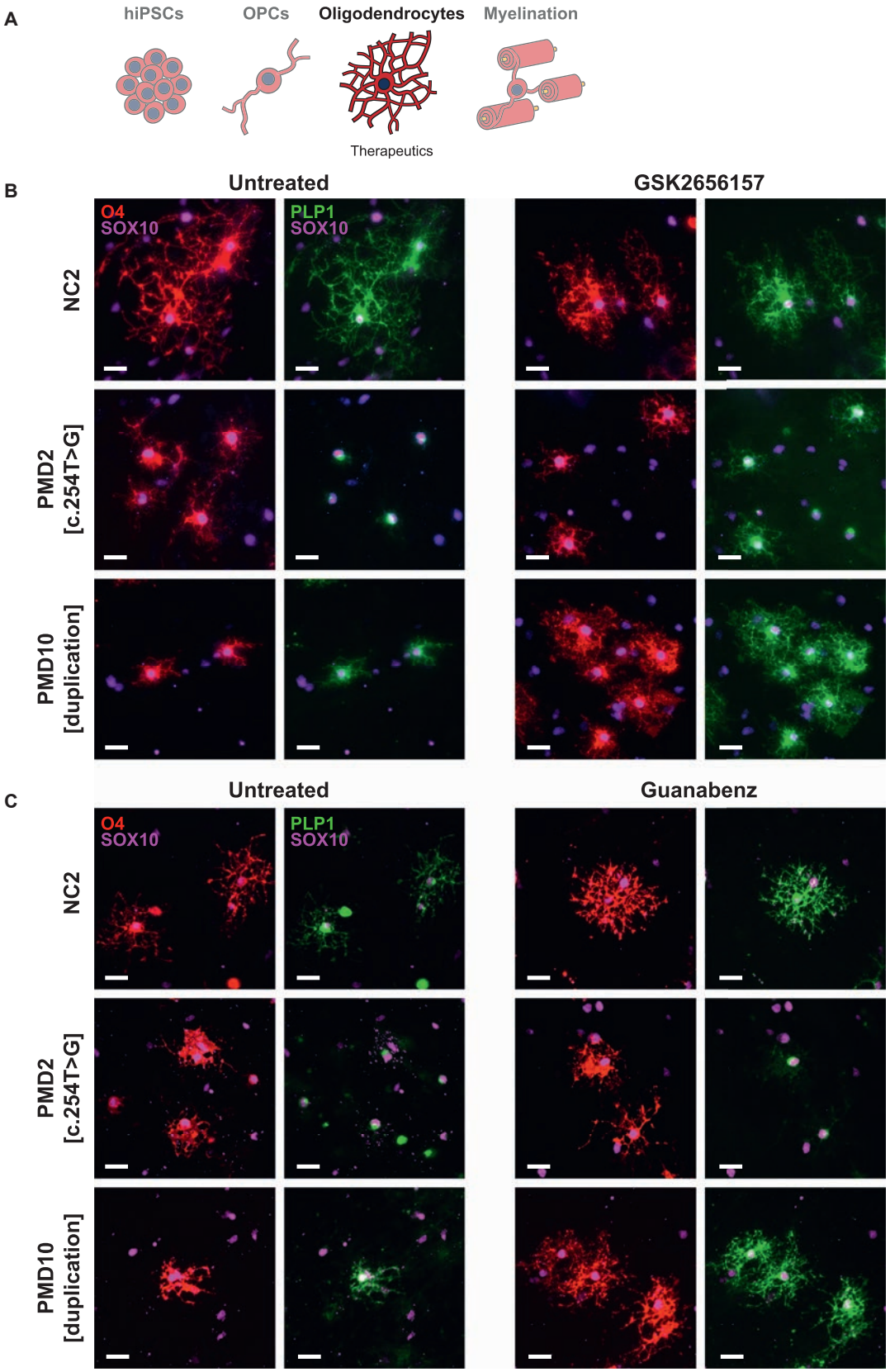


Figure 5. Modulation of the ER Stress Response Improves PLP1 Perinuclear Retention
(A) Schematic of the experimental stage, oligodendrocyte morphology, and insights presented in this figure.
(B) Representative images of oligodendrocytes after 28 days of treatment with 1 μM GSK2656157 (n = 2 technical replicates) and immunostaining for O4, SOX10, and PLP1. Note the rescue of PLP1 distribution in treated PMD2-derived oligodendrocytes and the improvement of oligodendrocyte morphology in PMD10. Scale bars represent 25 μm.

(legend continued on next page)

or branch distally and ultimately resulted in cell death (Movies S1, S2, and S3).

In addition to these shared cellular defects, two molecular defects were observed in subsets of the PMD hiPSC-derived oligodendrocytes. First, in the majority of the cultures (PMD1, PMD3, PMD4, PMD7–PMD9, and PMD12), $O4^+$ oligodendrocytes were present, but PLP1 expression was not detectable (Figure 4B). This was expected for PMD12, a complete *PLP1* deletion, but not the additional cultures that failed to mature to a $PLP1^+$ stage. Interestingly, although the individual with the *PLP1* deletion has a mild phenotype, other cultures that failed to express PLP1 were derived from individuals presenting with some of the most clinically severe presentations of the panel (Figure 1B). Second, in PMD2, PMD6, and PMD10, cells matured to a $PLP1^+$ stage; however, PLP1 signal was completely restricted to the perinuclear region of the cell body, and no signal was evident in the processes (Figure 4B).

Mobilization of PLP1 into Oligodendrocyte Processes by Small-Molecule Modulators of ER Stress Pathways

Perinuclear retention of a misfolded protein is a hallmark of ER stress. Our comparative hiPSC panel identified that only a subgroup of PMD cultures (PMD2, PMD6, and PMD10) exhibited perinuclear retention of PLP1, so we selected PMD2 and PMD10, from individuals with genetically distinct point and duplication mutations, to explore strategies for modulating the ER stress response (Figure 5A).

We tested two small molecules that specifically inhibit or enhance the ER stress response. GSK2656157 is a recently described inhibitor of protein kinase R-like ER kinase (PERK), which senses misfolded proteins and initiates a response to ER stress.[63] Guanabenz is an inhibitor of the protein phosphatase 1 regulatory subunit GADD34, which allows normal cellular functions to recommence once the stressor has been resolved.[64]

We assessed the effects of GSK2656157 and guanabenz on oligodendrocyte cultures from control NC2 and *PLP1* mutants PMD2 and PMD10. PMD10 oligodendrocytes demonstrated remarkable restoration of cell morphology under both conditions (Figures 5B and 5C). However, when treated with GSK2656157, PMD2 showed modest mobilization of PLP1 into cell processes but had no response to guanabenz (Figures 5B and 5C). Neither small molecule caused appreciable changes in the morphology of control NC2 cells.

Modulation of ER Stress Phenotypes in Oligodendrocyte-Neuron Co-Cultures

Oligodendrocytes in vivo do not exist in a state of homogeneous, monolayer culture, and although exceptional for identifying cell-intrinsic deficits, phenotyping in this system does not capture defects of myelination. In order to create a more physiologically relevant model of the defects caused by *PLP1* mutations, we adapted a protocol for co-culturing human oligodendrocytes on rat dorsal root ganglion neurons (Figure 6A)[53] to assess oligodendrocyte maturation, axonal tracking, and ensheathment (in vitro "myelination"). In these conditions, control oligodendrocytes extend processes that search out and travel along individual neuron axons (Figure 6B), forming long linear tracts as opposed to the branching, web-like morphology seen in monoculture. PLP1 signal is present throughout the cell, including the cell body, processes, and tracts. Meanwhile, MBP, a structural protein in the myelin sheath, is restricted to the cell body and tracts, identifying regions in the early stages of myelination.

NC2, PMD2, and PMD10 oligodendrocytes were seeded onto neurons in basal medium supplemented with GSK2656157 or guanabenz. Similar to monocultures, untreated PMD2 oligodendrocytes showed prominent PLP1 perinuclear retention and no PLP1 immunofluorescence in the processes (Figure 6C, top). Meanwhile, MBP immunofluorescence delineated the entirety of the oligodendrocytes, including extensive, matted processes, but showed no tracking with neurons (Figure 6C, bottom). Interestingly, GSK2656157-treated PMD2 oligodendrocytes did not show mobilization of PLP1 into the processes, as they had in monoculture (Figure 6D, top). Despite this, the process matting seen in MBP largely resolved, and short lengths of tracking were clearly visible (Figure 6D, bottom). On the other hand, guanabenz promoted a degree of PLP1 mobilization into oligodendrocyte processes (Figure 6E, top) but did not resolve the MBP matting to the same extent as GSK2656157 (Figure 6E, bottom).

Untreated PMD10 oligodendrocytes recapitulated the PLP1 perinuclear retention observed in monoculture (Figure 6F, top), but MBP immunofluorescence was present throughout the processes instead of being restricted to defined tracts (Figure 6F, bottom). Where present, tracts were also shorter than for controls. Treatment with GSK2656157 completely restored PLP1 mobilization into processes (Figure 6G, top) but did not improve MBP signal or tracking over untreated conditions (Figure 6G, bottom). Guanabenz, however, drastically increased PLP1 mobilization into processes, rescued MBP fluorescence intensity, and increased the prevalence of tracts, although they were still less prevalent and shorter than for controls (Figure 6H).

Discussion

Historically, PMD has been a challenging disease to parse. Affected individuals present with symptoms across

(C) Representative images of oligodendrocytes after 35 days of treatment with 2.5 μM guanabenz (n = 2 technical replicates) and immunostaining for O4, SOX10, and PLP1. Note the improvement of cell morphology in PMD10-derived, but not PMD2-derived, oligodendrocytes. Also, compared with those in Figure 4B, untreated PMD10 cells here demonstrate PLP1 diffusing throughout the cell body in addition to intense perinuclear signal. This appears to be due to a longer period of culture between passaging and immunostaining, which possibly allowed the cells a degree of recovery from the added extrinsic stress of passaging. Scale bars represent 25 μm.

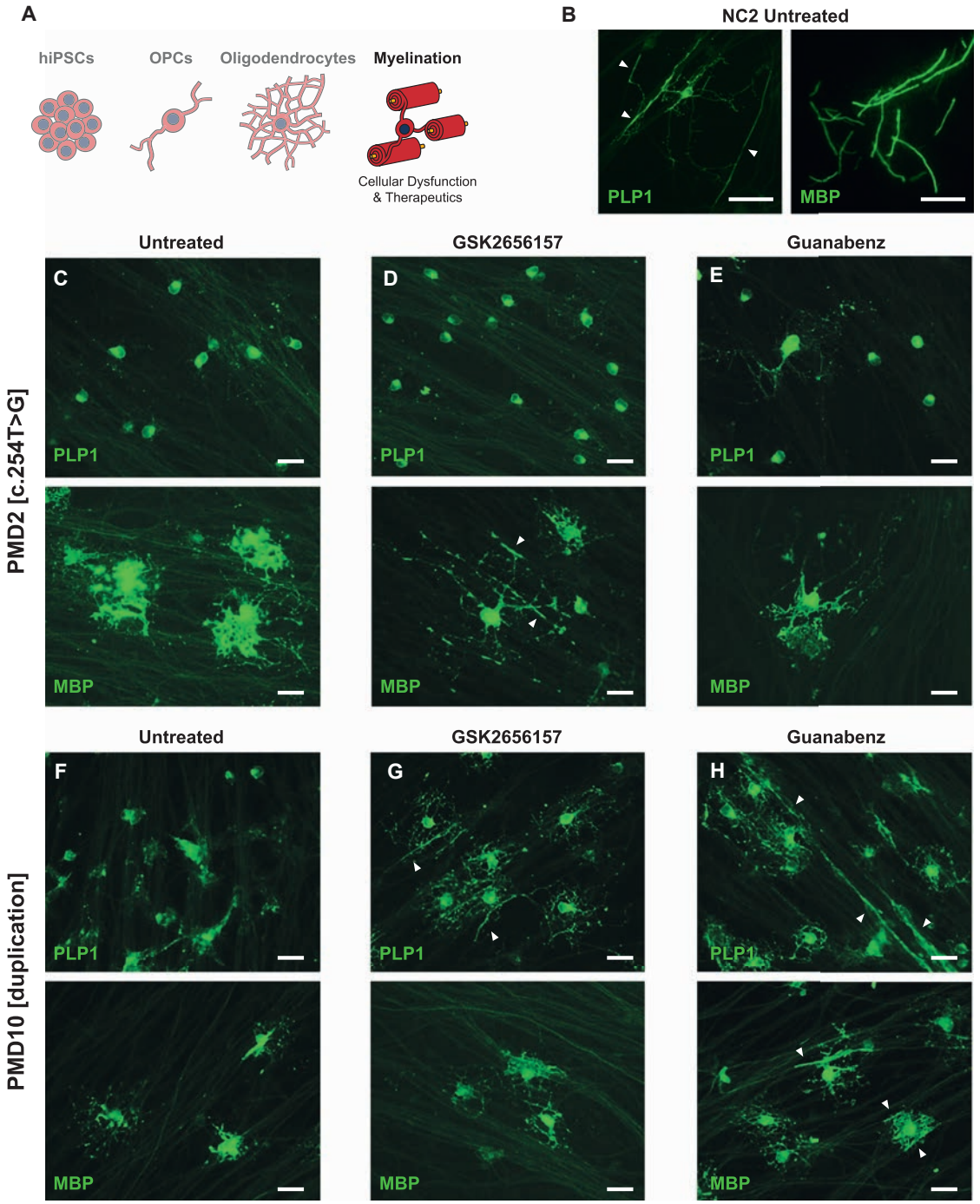


Figure 6. Modulation of the ER Stress Response in Oligodendrocytes Co-cultured with DRG Neurons

(A) Schematic of the experimental stage, oligodendrocyte morphology, and insights presented in this figure.

(B) Representative images of untreated NC2 oligodendrocytes demonstrating the two distinct morphologies appreciable in co-culture by immunostaining with PLP1 versus MBP. PLP1 signal marks the oligodendrocyte cell body, processes, and early tracts extending along neurons (white arrowheads). MBP signal is restricted to the tracts and is indicative of early axonal ensheathment. Scale bars represent 50 μm.

(C–E) Representative PLP1 (top) and MBP (bottom) immunofluorescence images of PMD2 oligodendrocytes co-cultured with neurons and untreated (C) or treated with 100 nM GSK2656157 (D) or 2.5 μM guanabenz (E) (n = 2 technical replicates). Note the rescue of MBP-positive cell processes' morphology when treated with GSK26157. Scale bars represent 25 μm.

(F–H) Representative PLP1 (top) and MBP (bottom) immunofluorescence images of PMD10 oligodendrocytes co-cultured with neurons and untreated (F) or treated with 100 nM GSK2656157 (G) or 2.5 μM guanabenz (H) (n = 2 technical replicates). Note the rescue of PLP1 distribution in treated oligodendrocytes (G and H). White arrowheads depict tracts of myelin along neuronal axons (G and H). Scale bars represent 25 μm.

a spectrum of disease severity, and there are no direct links between an individual's unique *PLP1* mutation and the etiology or course of their disease. The inaccessibility of primary oligodendrocytes severely limits direct studies in humans, and it would be prohibitively expensive and time consuming to model the hundreds of different PMD-linked *PLP1* mutations in animals. In this study, by taking advantage of natural mutational diversity and recent advancements in hiPSC-based disease modeling, we demonstrated the feasibility of generating, differentiating, and assessing a panel of hiPSCs that capture the full spectrum of PMD's clinical and genetic heterogeneity. We characterized hiPSCs, OPCs, and oligodendrocytes from 12 individuals with PMD and identified shared and individual defects spanning *PLP1* expression, *PLP1* splicing, OPC production, oligodendrocyte morphology, and response to small-molecule therapeutics.

Although PLP1 is restricted to the oligodendrocyte lineage, *DM20* mRNA is robustly transcribed in pluripotent cells,[60] providing an opportunity for rapid assessment of mutation-specific transcript defects in a scalable, homogeneous population of cells. *DM20*, the OPC-specific isoform of *PLP1* and the sole isoform expressed in hiPSCs, limited our ability to interpret the effects of mutations in the *PLP1*-specific region of exon 3b. However, we were interested to discover that, although neither the point mutations nor the partial deletion showed any effect on *DM20* expression levels, the duplication and triplication showed 2- and 3-fold higher expression, respectively, than did controls. Given the fact that PLP1 in a normal hemizygous individual constitutes as much as 50% of myelin's total protein, we speculate that this linear relationship between copy number and mRNA expression could lead to excess protein abundance in oligodendrocytes of individuals with supernumerary copies of *PLP1*. This linear trend has not been reported previously in animals.[65] However, those studies were performed on whole brain tissue, wherein OPCs are but a small fraction of the total cell population. Alternately, when protein is actually translated in OPCs and oligodendrocytes, it is possible that PLP1's overabundance could trigger a degree of feedback regulation that we cannot appreciate in hiPSCs.

The presence of *DM20* mRNA also enables analysis of splicing and structural defects in hiPSCs. Importantly, whereas protein levels vary by cell type and can be inferred only from hiPSC studies, splicing defects have a direct and immutable impact on protein structure across cell types. We confirmed a prior report of *PLP1* exon 6 skipping caused by a mutation of the +3 nucleotide of intron 5.[61] Skipping of exon 6 causes an in-frame deletion of 22 amino acids, including part of PLP1's putative fourth transmembrane domain. Next, we demonstrated that our partial deletion causes an in-frame fusion of exon 3a to the distal portion of exon 4 with no additional mis-splicing or decrease in expression. Similarly, we found that a mutation at the +750 nucleotide of intron 3 does not appear to affect splicing or expression in hiPSCs. Collectively, mRNA analyses in hiPSCs provide a rapid and minimally invasive means of identifying convergent structural and expression defects caused by disparate *PLP1* mutations. As more mutations are characterized, this approach could eventually allow subgrouping of mutations by mRNA defect and aid in predicting individuals' prognoses.

Differentiating hiPSCs to OPCs provides important insight into cell-intrinsic pathologic consequences of disparate *PLP1* mutations. Although we anticipated that certain mutations could lead to an OPC defect, we were surprised that two-thirds of the PMD cultures were severely depleted of OPCs. On the basis of prior imaging and pathology studies, PMD has traditionally been considered a disease of myelin production and structure and thus predominantly a defect of oligodendrocytes.[17] Our findings contradict this generalization and suggest that for at least a subset of individuals with PMD, a precedent defect at the OPC stage would limit subsequent oligodendrocyte production and thus preempt myelination. Of particular note, mutations within the *PLP1*-specific region of exon 3b (e.g., PMD3 and PMD4) would not be expected to manifest in OPCs, given the predicted expression of the *DM20* splice isoform. Yet here, PMD3 cultures were some of the most severely depleted in the entire panel. This apparent paradox would have been difficult to appreciate without the ability to directly compare against the spectrum of other PMD-derived cultures. It is important to acknowledge that this single time point is not sufficient for discerning whether relative reductions in OPCs are due to a block of proliferation, failure of migration out of neurospheres, premature cell maturation and loss of the PDGFRA marker, or outright cell death. Nonetheless, our collective results suggest that prevalent, OPC-intrinsic defects strongly contribute to phenotypic variability. Most importantly, these findings suggest that future therapeutic development could necessitate earlier intervention than previously appreciated.

On the basis of simultaneous comparative phenotyping of the 12 individuals in our panel, we identified three distinct classes of cellular defects in PMD-derived oligodendrocytes: failure to produce oligodendrocytes, failure to produce $PLP1^+$ oligodendrocytes, and perinuclear retention of PLP1. A point mutation in intron 3 and a *PLP1* triplication were the only cultures that did not generate any $O4^+$ or $PLP1^+$ oligodendrocytes, even though these cultures had previously generated robust numbers of OPCs. This could be due to either a block in the maturation of OPCs to oligodendrocytes or rapid death of newly formed oligodendrocytes as PLP1 production is upregulated. Of the remaining cultures, oligodendrocytes were either $O4^+PLP1^-$ or $O4^+PLP1^+$, but the PLP1 signal was completely retained perinuclearly. Perinuclear localization is a hallmark of protein misfolding and ER retention and has previously been demonstrated in vitro in two human oligodendrocyte lines with *PLP1* point mutations, one of which corresponds with our PMD6 mutation.[41] Additionally, across both of these categories, all oligodendrocytes presented with severe

defects of process extension and branching. The function of PLP1 has not been fully established. However, PLP1 has been implicated in both OPC migration[26] and as a bridge between the membrane and cytoskeleton.[66] It is possible that, in the absence of PLP1, cell processes become insensitive to stimuli that would normally trigger them to extend. Similarly, observed deficits in distal branching could be secondary to failed process extension or result from loss of PLP1's structural cytoskeletal support. Identification of a mutation where cells extend processes that are of normal length but completely unbranched could elucidate this further. Ultimately, the fact that an individual's mutation and clinical presentation alone were not predictive of their cellular phenotype highlights the power of this platform to categorize disparate *PLP1* mutations, enabling the development of PMD-subgroup-specific prognoses and therapeutic plans.

The genetic breakdown of our perinuclear retention cohort was intriguing because it contained a substitution of a proline in an extracellular loop, a substitution of a leucine in a transmembrane domain, and a full *PLP1* duplication. Prior studies in animal and human models have implicated ER stress as a pathogenic result of particular *PLP1* mutations,[41,67–71] but the larger PMD community has struggled to leverage these findings to treat the general PMD population. We suspect that this is due to intrinsic differences in the types of mutations being targeted. To address this, we treated our duplication and transmembrane point mutations with two small molecules that modulate ER stress in two completely opposite manners. GSK2656157, a newly described PERK inhibitor, targets the initiation of the pathway, averting the negative downstream effects of a continuous ER stress response, particularly apoptosis. At the other end, guanabenz, a GADD34 inhibitor, deliberately prolongs the stress response, maintaining expression of molecular chaperones to promote clearance of misfolded protein aggregates. The point mutation and duplication responded variably to these two approaches. In the case of the duplication, both inhibiting and enhancing the ER stress pathway had a drastic positive effect, relieving the stress sufficiently for oligodendrocytes to reestablish normal morphologies. The point mutation, however, contains a gain-of-function mutation that shows only partial response to inhibition of ER stress and no response to enhanced refolding. It is possible that this mutation is simply refractory to refolding. However, because PLP1 was observed to mobilize into the processes with GSK2656157 treatment yet the cells did not fully recover morphologically, it is more likely that the position of this mutation disrupts either membrane integration or a functional domain of PLP1. Further characterization and titration of GSK2656157 will provide a foundation for assessment of this and other small-molecule therapeutics for personalized applications in the future.

The neuron-oligodendrocyte co-culture system provides a model of the endogenous structural environment that influences oligodendrocyte and myelin biology in the brain, providing additional insight into the nuances of PMD pathogenesis. Using this in vitro system, we demonstrated that individual PMD phenotypes can be modulated to restore oligodendrocyte morphology and axonal tracking. As opposed to our initial monoculture system, the co-culture system revealed more appreciable differences between PMD oligodendrocytes' responses to GSK2656157 and guanabenz. It was interesting to observe that untreated cells' morphologies were improved in this system, presumably as a result of supportive factors released by the neurons that were absent from our media conditions and less frequent passaging and the extrinsic stress that entails. The transmembrane point mutation again demonstrated improvement of cell morphology when the ER stress response was completely shut down by GSK2656157 but nonetheless could not associate with neurons. This further confirms that ER stress exacerbates the pathogenesis of this mutation, but persistence of the mutation itself prevents true functional restoration by GSK2656157 alone. On the other hand, the duplication showed a dramatic response when the ER stress response was prolonged by guanabenz. Presumably, whereas inhibition of ER stress leaves misfolded protein stuck in the ER, the prolonged action of chaperones and other protective molecules triggered by ER stress promotes folding and mobilization of PLP1 to its ultimate destination in the emerging myelin sheath.

Since their first report a decade ago, hiPSCs have been transformed into a multitude of different cell types, providing invaluable insights into human health and disease. hiPSC technologies are a particular boon to the study of PMD and other leukodystrophies, obviating many of the challenges that have previously limited our ability to model and investigate oligodendrocyte dysfunction. Using hiPSC technologies, we can now generate the entire oligodendrocyte lineage in the laboratory, model the full progression of disease pathology, and observe pathogenesis in real time. Importantly, many of the defects we report here could not have been predicted by individuals' clinical histories or mutations alone. However, characterization of these defects across all samples in parallel enabled identification of distinct subclasses of cellular and molecular pathogeneses that now link disparate *PLP1* mutations. We hope this work will serve as a foundation for the assessment of oligodendrocyte dysfunction throughout the greater community of genetic myelin diseases.

Accession Numbers

The accession number for the raw RNA-seq datasets reported in this paper is GEO: GSE96049.

Supplemental Data

Supplemental Data include two figures, three tables, and three movies and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2017.03.005.

Conflicts of Interest

P.J.T. is on the scientific advisory board of Cell Line Genetics.

Acknowledgments

This research was supported by grants from the Pelizaeus-Merzbacher Disease Foundation (P.J.T. and G.M.H.), NIH grants R01NS093357 (P.J.T. and M.W.) and R01NS058978 (G.M.H.), the New York Stem Cell Foundation (P.J.T. and V.F.), NIH predoctoral training grants T32GM007250 and F30HD084167 (Z.S.N.), and the Adelson Medical Research and Novo Nordisk Foundations (S.A.G. and M.S.W.). P.D. is a NYSCF-Druckenmiller fellow. Additional support was provided by the Cytometry & Imaging Microscopy and Genomics core facilities of the Case Comprehensive Cancer Center (P30CA043703). We are grateful to the late James Garbern for providing PMD samples; Leslie Cooperman, Elizabeth Shick, and William Qu for technical assistance; and Peter Scacheri, Anthony Wynshaw-Boris, Nancy Bass, Marius Wernig, and the Tesar Lab for discussion and comments on the manuscript.

Received: December 7, 2016
Accepted: March 9, 2017
Published: March 30, 2017

Web Resources

Cufflinks, http://cole-trapnell-lab.github.io/cufflinks/
Gene Expression Omnibus (GEO), https://www.ncbi.nlm.nih.gov/geo/
GeneReviews, Hobson, G.M., and Kamholz, J. (1993). PLP1-Related Disorders, https://www.ncbi.nlm.nih.gov/books/NBK1182/
IGV, http://software.broadinstitute.org/software/igv/
OMIM, http://www.omim.org
PennCNV, http://penncnv.openbioinformatics.org/en/latest/
R, http://www.R-project.org
RefSeq, https://www.ncbi.nlm.nih.gov/refseq/
TopHat, https://ccb.jhu.edu/software/tophat/index.shtml
UCSC Genome Browser, https://genome.ucsc.edu/

References

1. Parikh, S., Bernard, G., Leventer, R.J., van der Knaap, M.S., van Hove, J., Pizzino, A., McNeill, N.H., Helman, G., Simons, C., Schmidt, J.L., et al.; GLIA Consortium (2015). A clinical approach to the diagnosis of patients with leukodystrophies and genetic leukoencephelopathies. Mol. Genet. Metab. *114*, 501–515.
2. Kevelam, S.H., Steenweg, M.E., Srivastava, S., Helman, G., Naidu, S., Schiffmann, R., Blaser, S., Vanderver, A., Wolf, N.I., and van der Knaap, M.S. (2016). Update on Leukodystrophies: A Historical Perspective and Adapted Definition. Neuropediatrics *47*, 349–354.
3. Helman, G., Van Haren, K., Escolar, M.L., and Vanderver, A. (2015). Emerging treatments for pediatric leukodystrophies. Pediatr. Clin. North Am. *62*, 649–666.
4. Helman, G., Van Haren, K., Bonkowsky, J.L., Bernard, G., Pizzino, A., Braverman, N., Suhr, D., Patterson, M.C., Ali Fatemi, S., Leonard, J., et al.; GLIA Consortium (2015). Disease specific therapies in leukodystrophies and leukoencephalopathies. Mol. Genet. Metab. *114*, 527–536.
5. Van Haren, K., Bonkowsky, J.L., Bernard, G., Murphy, J.L., Pizzino, A., Helman, G., Suhr, D., Waggoner, J., Hobson, D., Vanderver, A., Patterson, M.C.; and GLIA Consortium (2015). Consensus statement on preventive and symptomatic care of leukodystrophy patients. Mol. Genet. Metab. *114*, 516–526.
6. Hobson, G.M., and Garbern, J.Y. (2012). Pelizaeus-Merzbacher disease, Pelizaeus-Merzbacher-like disease 1, and related hypomyelinating disorders. Semin. Neurol. *32*, 62–67.
7. Inoue, K. (2005). PLP1-related inherited dysmyelinating disorders: Pelizaeus-Merzbacher disease and spastic paraplegia type 2. Neurogenetics *6*, 1–16.
8. Bonkowsky, J.L., Nelson, C., Kingston, J.L., Filloux, F.M., Mundorff, M.B., and Srivastava, R. (2010). The burden of inherited leukodystrophies in children. Neurology *75*, 718–725.
9. Numata, Y., Gotoh, L., Iwaki, A., Kurosawa, K., Takanashi, J., Deguchi, K., Yamamoto, T., Osaka, H., and Inoue, K. (2014). Epidemiological, clinical, and genetic landscapes of hypomyelinating leukodystrophies. J. Neurol. *261*, 752–758.
10. Pelizaeus, F. (1885). Uber eine eigenthümliche Form Spastischer Lähmung mit Cerebralerschinungen auf hereditärer Grundlage (Multiple Sklerose). Arch. Psychiatr. Nervenkr. *16*, 698–710.
11. Merzbacher, L. (1910). Eine eigenarige familiär-hereditare Erkrankungsform (Aplasia axialis extra-corticalis congenita). Z. Neurol. Psychiatr. *3*, 1–138.
12. Willard, H.F., and Riordan, J.R. (1985). Assignment of the gene for myelin proteolipid protein to the X chromosome: implications for X-linked myelin disorders. Science *230*, 940–942.
13. Dautigny, A., Mattei, M.G., Morello, D., Alliel, P.M., Pham-Dinh, D., Amar, L., Arnaud, D., Simon, D., Mattei, J.F., Guenet, J.L., et al. (1986). The structural gene coding for myelin-associated proteolipid protein is mutated in jimpy mice. Nature *321*, 867–869.
14. Koeppen, A.H., Ronca, N.A., Greenfield, E.A., and Hans, M.B. (1987). Defective biosynthesis of proteolipid protein in Pelizaeus-Merzbacher disease. Ann. Neurol. *21*, 159–170.
15. Gencic, S., Abuelo, D., Ambler, M., and Hudson, L.D. (1989). Pelizaeus-Merzbacher disease: an X-linked neurologic disorder of myelin metabolism with a novel mutation in the gene encoding proteolipid protein. Am. J. Hum. Genet. *45*, 435–442.
16. LeVine, S.M., Wong, D., and Macklin, W.B. (1990). Developmental expression of proteolipid protein and DM20 mRNAs and proteins in the rat brain. Dev. Neurosci. *12*, 235–250.
17. Baumann, N., and Pham-Dinh, D. (2001). Biology of oligodendrocyte and myelin in the mammalian central nervous system. Physiol. Rev. *81*, 871–927.
18. Garbern, J.Y. (2007). Pelizaeus-Merzbacher disease: Genetic and cellular pathogenesis. Cell. Mol. Life Sci. *64*, 50–65.
19. Phillips, R.J. (1954). Jimpy, a new totally sexlinked gene in the house mouse. Z. Indukt. Abstamm. Vererbungsl. *86*, 322–326.
20. Eicher, E.M., and Hoppe, P.C. (1973). Use of chimeras to transmit lethal genes in the mouse and to demonstrate allelism of the two X-linked male lethal genes jp and msd. J. Exp. Zool. *183*, 181–184.
21. Duncan, I.D., Hammang, J.P., and Trapp, B.D. (1987). Abnormal compact myelin in the myelin-deficient rat: absence of proteolipid protein correlates with a defect in the intraperiod line. Proc. Natl. Acad. Sci. USA *84*, 6287–6291.

22. Griffiths, I.R., Duncan, I.D., McCulloch, M., and Harvey, M.J. (1981). Shaking pups: a disorder of central myelination in the Spaniel dog. Part 1. Clinical, genetic and light-microscopical observations. J. Neurol. Sci. *50*, 423–433.
23. Griffiths, I.R., Scott, I., McCulloch, M.C., Barrie, J.A., McPhilemy, K., and Cattanach, B.M. (1990). Rumpshaker mouse: a new X-linked mutation affecting myelination: evidence for a defect in PLP expression. J. Neurocytol. *19*, 273–283.
24. Griffiths, I., Klugmann, M., Anderson, T., Thomson, C., Vouyiouklis, D., and Nave, K.A. (1998). Current concepts of PLP and its role in the nervous system. Microsc. Res. Tech. *41*, 344–358.
25. Hüttemann, M., Zhang, Z., Mullins, C., Bessert, D., Lee, I., Nave, K.A., Appikatla, S., and Skoff, R.P. (2009). Different proteolipid protein mutants exhibit unique metabolic defects. ASN Neuro *1*, 1.
26. Harlow, D.E., Saul, K.E., Komuro, H., and Macklin, W.B. (2015). Myelin Proteolipid Protein Complexes with αv Integrin and AMPA Receptors In Vivo and Regulates AMPA-Dependent Oligodendrocyte Progenitor Cell Migration through the Modulation of Cell-Surface GluR2 Expression. J. Neurosci. *35*, 12018–12032.
27. Laukka, J.J., Kamholz, J., Bessert, D., and Skoff, R.P. (2016). Novel pathologic findings in patients with Pelizaeus-Merzbacher disease. Neurosci. Lett. *627*, 222–232.
28. Boulloche, J., and Aicardi, J. (1986). Pelizaeus-Merzbacher disease: clinical and nosological study. J. Child Neurol. *1*, 233–239.
29. Hodes, M.E., Pratt, V.M., and Dlouhy, S.R. (1993). Genetics of Pelizaeus-Merzbacher disease. Dev. Neurosci. *15*, 383–394.
30. Cailloux, F., Gauthier-Barichard, F., Mimault, C., Isabelle, V., Courtois, V., Giraud, G., Dastugue, B., Boespflug-Tanguy, O.; and Clinical European Network on Brain Dysmyelinating Disease (2000). Genotype-phenotype correlation in inherited brain myelination defects due to proteolipid protein gene mutations. Eur. J. Hum. Genet. *8*, 837–845.
31. Hudson, L.D. (2003). Pelizaeus-Merzbacher disease and spastic paraplegia type 2: two faces of myelin loss from mutations in the same gene. J. Child Neurol. *18*, 616–624.
32. Hurst, S., Garbern, J., Trepanier, A., and Gow, A. (2006). Quantifying the carrier female phenotype in Pelizaeus-Merzbacher disease. Genet. Med. *8*, 371–378.
33. Klugmann, M., Schwab, M.H., Pühlhofer, A., Schneider, A., Zimmermann, F., Griffiths, I.R., and Nave, K.A. (1997). Assembly of CNS myelin in the absence of proteolipid protein. Neuron *18*, 59–70.
34. Inoue, K., Osaka, H., Thurston, V.C., Clarke, J.T., Yoneyama, A., Rosenbarker, L., Bird, T.D., Hodes, M.E., Shaffer, L.G., and Lupski, J.R. (2002). Genomic rearrangements resulting in PLP1 deletion occur by nonhomologous end joining and cause different dysmyelinating phenotypes in males and females. Am. J. Hum. Genet. *71*, 838–853.
35. Gao, Q., Thurston, V.C., Vance, G.H., Dlouhy, S.R., and Hodes, M.E. (2005). Genetic diagnosis of PLP gene duplications/deletions in patients with Pelizaeus-Merzbacher disease. Clin. Genet. *68*, 466–467.
36. Harlow, D.E., Saul, K.E., Culp, C.M., Vesely, E.M., and Macklin, W.B. (2014). Expression of proteolipid protein gene in spinal cord stem cells and early oligodendrocyte progenitor cells is dispensable for normal cell migration and myelination. J. Neurosci. *34*, 1333–1343.
37. Nave, K.A., Lai, C., Bloom, F.E., and Milner, R.J. (1987). Splice site selection in the proteolipid protein (PLP) gene transcript and primary structure of the DM-20 protein of central nervous system myelin. Proc. Natl. Acad. Sci. USA *84*, 5665–5669.
38. Karim, S.A., Barrie, J.A., McCulloch, M.C., Montague, P., Edgar, J.M., Kirkham, D., Anderson, T.J., Nave, K.A., Griffiths, I.R., and McLaughlin, M. (2007). PLP overexpression perturbs myelin protein composition and myelination in a mouse model of Pelizaeus-Merzbacher disease. Glia *55*, 341–351.
39. Simons, M., Kramer, E.M., Macchi, P., Rathke-Hartlieb, S., Trotter, J., Nave, K.A., and Schulz, J.B. (2002). Overexpression of the myelin proteolipid protein leads to accumulation of cholesterol and proteolipid protein in endosomes/lysosomes: implications for Pelizaeus-Merzbacher disease. J. Cell Biol. *157*, 327–336.
40. Gow, A., Friedrich, V.L., Jr., and Lazzarini, R.A. (1994). Many naturally occurring mutations of myelin proteolipid protein impair its intracellular transport. J. Neurosci. Res. *37*, 574–583.
41. Numasawa-Kuroiwa, Y., Okada, Y., Shibata, S., Kishi, N., Akamatsu, W., Shoji, M., Nakanishi, A., Oyama, M., Osaka, H., Inoue, K., et al. (2014). Involvement of ER stress in dysmyelination of Pelizaeus-Merzbacher Disease with PLP1 missense mutations shown by iPSC-derived oligodendrocytes. Stem Cell Reports *2*, 648–661.
42. Southwood, C.M., Garbern, J., Jiang, W., and Gow, A. (2002). The unfolded protein response modulates disease severity in Pelizaeus-Merzbacher disease. Neuron *36*, 585–596.
43. Gow, A., and Sharma, R. (2003). The unfolded protein response in protein aggregating diseases. Neuromolecular Med. *4*, 73–94.
44. Douvaras, P., Wang, J., Zimmer, M., Hanchuk, S., O'Bara, M.A., Sadiq, S., Sim, F.J., Goldman, J., and Fossati, V. (2014). Efficient generation of myelinating oligodendrocytes from primary progressive multiple sclerosis patients by induced pluripotent stem cells. Stem Cell Reports *3*, 250–259.
45. Wang, S., Bates, J., Li, X., Schanz, S., Chandler-Militello, D., Levine, C., Maherali, N., Studer, L., Hochedlinger, K., Windrem, M., and Goldman, S.A. (2013). Human iPSC-derived oligodendrocyte progenitor cells can myelinate and rescue a mouse model of congenital hypomyelination. Cell Stem Cell *12*, 252–264.
46. Somers, A., Jean, J.C., Sommer, C.A., Omari, A., Ford, C.C., Mills, J.A., Ying, L., Sommer, A.G., Jean, J.M., Smith, B.W., et al. (2010). Generation of transgene-free lung disease-specific human induced pluripotent stem cells using a single excisable lentiviral stem cell cassette. Stem Cells *28*, 1728–1740.
47. Okita, K., Matsumura, Y., Sato, Y., Okada, A., Morizane, A., Okamoto, S., Hong, H., Nakagawa, M., Tanabe, K., Tezuka, K., et al. (2011). A more efficient method to generate integration-free human iPS cells. Nat. Methods *8*, 409–412.
48. Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. (1998). Embryonic stem cell lines derived from human blastocysts. Science *282*, 1145–1147.
49. Diskin, S.J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J.M., and Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res. *36*, e126.
50. Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

51. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.
52. Douvaras, P., and Fossati, V. (2015). Generation and isolation of oligodendrocyte progenitor cells from human pluripotent stem cells. Nat. Protoc. *10*, 1143–1154.
53. Chan, J.R., Watkins, T.A., Cosgaya, J.M., Zhang, C., Chen, L., Reichardt, L.F., Shooter, E.M., and Barres, B.A. (2004). NGF controls axonal receptivity to myelination by Schwann cells or oligodendrocytes. Neuron *43*, 183–191.
54. Laukka, J.J., Stanley, J.A., Garbern, J.Y., Trepanier, A., Hobson, G., Lafleur, T., Gow, A., and Kamholz, J. (2013). Neuroradiologic correlates of clinical disability and progression in the X-linked leukodystrophy Pelizaeus-Merzbacher disease. J. Neurol. Sci. *335*, 75–81.
55. Laukka, J.J., Makki, M.I., Lafleur, T., Stanley, J., Kamholz, J., and Garbern, J.Y. (2014). Diffusion tensor imaging of patients with proteolipid protein 1 gene mutations. J. Neurosci. Res. *92*, 1723–1732.
56. Sumida, K., Inoue, K., Takanashi, J., Sasaki, M., Watanabe, K., Suzuki, M., Kurahashi, H., Omata, T., Tanaka, M., Yokochi, K., et al. (2016). The magnetic resonance imaging spectrum of Pelizaeus-Merzbacher disease: A multicenter study of 19 patients. Brain Dev. *38*, 571–580.
57. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.
58. Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., et al. (2014). RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. *42*, D756–D763.
59. Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z.D., Ziller, M., Croft, G.F., Amoroso, M.W., Oakley, D.H., et al. (2011). Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. Cell *144*, 439–452.
60. Shimojima, K., Inoue, T., Imai, Y., Arai, Y., Komoike, Y., Sugawara, M., Fujita, T., Ideguchi, H., Yasumoto, S., Kanno, H., et al. (2012). Reduced PLP1 expression in induced pluripotent stem cells derived from a Pelizaeus-Merzbacher disease patient with a partial PLP1 duplication. J. Hum. Genet. *57*, 580–586.
61. Hobson, G.M., Davis, A.P., Stowell, N.C., Kolodny, E.H., Sistermans, E.A., de Coo, I.F., Funanage, V.L., and Marks, H.G. (2000). Mutations in noncoding regions of the proteolipid protein gene in Pelizaeus-Merzbacher disease. Neurology *55*, 1089–1096.
62. Sim, F.J., McClain, C.R., Schanz, S.J., Protack, T.L., Windrem, M.S., and Goldman, S.A. (2011). CD140a identifies a population of highly myelinogenic, migration-competent and efficiently engrafting human oligodendrocyte progenitor cells. Nat. Biotechnol. *29*, 934–941.
63. Axten, J.M., Romeril, S.P., Shu, A., Ralph, J., Medina, J.R., Feng, Y., Li, W.H., Grant, S.W., Heerding, D.A., Minthorn, E., et al. (2013). Discovery of GSK2656157: An Optimized PERK Inhibitor Selected for Preclinical Development. ACS Med. Chem. Lett. *4*, 964–968.
64. Tsaytler, P., Harding, H.P., Ron, D., and Bertolotti, A. (2011). Selective inhibition of a regulatory subunit of protein phosphatase 1 restores proteostasis. Science *332*, 91–94.
65. Clark, K., Sakowski, L., Sperle, K., Banser, L., Landel, C.P., Bessert, D.A., Skoff, R.P., and Hobson, G.M. (2013). Gait abnormalities and progressive myelin degeneration in a new murine model of Pelizaeus-Merzbacher disease with tandem genomic duplication. J. Neurosci. *33*, 11788–11799.
66. Yamazaki, R., Ishibashi, T., Baba, H., and Yamaguchi, Y. (2016). Knockdown of Unconventional Myosin ID Expression Induced Morphological Change in Oligodendrocytes. ASN Neuro *8*, 8.
67. Numata, Y., Morimura, T., Nakamura, S., Hirano, E., Kure, S., Goto, Y.I., and Inoue, K. (2013). Depletion of molecular chaperones from the endoplasmic reticulum and fragmentation of the Golgi apparatus associated with pathogenesis in Pelizaeus-Merzbacher disease. J. Biol. Chem. *288*, 7451–7466.
68. Roboti, P., Swanton, E., and High, S. (2009). Differences in endoplasmic-reticulum quality control determine the cellular response to disease-associated mutants of proteolipid protein. J. Cell Sci. *122*, 3942–3953.
69. Dhaunchak, A.S., and Nave, K.A. (2007). A common mechanism of PLP/DM20 misfolding causes cysteine-mediated endoplasmic reticulum retention in oligodendrocytes and Pelizaeus-Merzbacher disease. Proc. Natl. Acad. Sci. USA *104*, 17813–17818.
70. Gow, A., Southwood, C.M., and Lazzarini, R.A. (1998). Disrupted proteolipid protein trafficking results in oligodendrocyte apoptosis in an animal model of Pelizaeus-Merzbacher disease. J. Cell Biol. *140*, 925–934.
71. Gow, A., and Lazzarini, R.A. (1996). A cellular mechanism governing the severity of Pelizaeus-Merzbacher disease. Nat. Genet. *13*, 422–428.

Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations

Alicia R. Martin,[1,2,3,4] Christopher R. Gignoux,[4] Raymond K. Walters,[1,2,3] Genevieve L. Wojcik,[4] Benjamin M. Neale,[1,2,3] Simon Gravel,[5,6] Mark J. Daly,[1,2,3] Carlos D. Bustamante,[4] and Eimear E. Kenny[7,8,9,10,*]

The vast majority of genome-wide association studies (GWASs) are performed in Europeans, and their transferability to other populations is dependent on many factors (e.g., linkage disequilibrium, allele frequencies, genetic architecture). As medical genomics studies become increasingly large and diverse, gaining insights into population history and consequently the transferability of disease risk measurement is critical. Here, we disentangle recent population history in the widely used 1000 Genomes Project reference panel, with an emphasis on populations underrepresented in medical studies. To examine the transferability of single-ancestry GWASs, we used published summary statistics to calculate polygenic risk scores for eight well-studied phenotypes. We identify directional inconsistencies in all scores; for example, height is predicted to decrease with genetic distance from Europeans, despite robust anthropological evidence that West Africans are as tall as Europeans on average. To gain deeper quantitative insights into GWAS transferability, we developed a complex trait coalescent-based simulation framework considering effects of polygenicity, causal allele frequency divergence, and heritability. As expected, correlations between true and inferred risk are typically highest in the population from which summary statistics were derived. We demonstrate that scores inferred from European GWASs are biased by genetic drift in other populations even when choosing the same causal variants and that biases in any direction are possible and unpredictable. This work cautions that summarizing findings from large-scale GWASs may have limited portability to other populations using standard approaches and highlights the need for generalized risk prediction methods and the inclusion of more diverse individuals in medical genomics.

Introduction

The majority of genome-wide association studies (GWASs) have been performed in populations of European descent.[1–4] An open question in medical genomics is the degree to which these results transfer to new populations. GWASs have yielded tens of thousands of common genetic variants significantly associated with human medical and evolutionary phenotypes, most of which have replicated in other ethnic groups.[5–7] However, GWASs are optimally powered to discover common variant associations, and the European bias in GWASs results in associated SNPs with higher minor allele frequencies on average compared to other populations. The predictive power of GWAS findings and genetic diagnostic accuracy in non-Europeans are therefore limited by population differences in allele frequencies and linkage disequilibrium structure. For example, a previous study showed that the accuracy of breeding values and genomic prediction decays approximately linearly with increasing divergence between the discovery and target population.[8] Additionally, multiple individuals with African ancestry have received false positive misdiagnoses of hypertrophic cardiomyopathy that would have been prevented with the inclusion of even small numbers of African Americans in these studies.[9] Further, a previous study finding that 96% of GWAS participants are of European descent[1] has recently been updated; although the non-European proportion of GWAS participants has increased to nearly 20%, this is primarily driven by Asian individuals, and the proportion of individuals with African and Hispanic/Latino ancestry in GWASs has remained essentially unchanged.[4]

As GWAS sample sizes grow to hundreds of thousands of samples, they also become better powered to detect rare variant associations.[10–12] Large-scale sequencing studies have demonstrated that rare variants show stronger geographic clustering than common variants.[13–15] Rare, disease-associated variants are therefore expected to track with recent population demography and/or be population restricted.[14,16–18] As the next era of GWASs expands to evaluate the disease-associated role of rare variants, it is not only scientifically imperative to include multi-ethnic populations, it is also likely that such studies will encounter increasing genetic heterogeneity in very large study populations. A comprehensive understanding of the genetic diversity and demographic history of multi-ethnic populations is critical for appropriate applications of GWASs and ultimately for ensuring that genetics does not contribute to or enhance health disparities.[4]

[1]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA; [2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [3]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [4]Department of Genetics, Stanford University, Stanford, CA 94305, USA; [5]Department of Human Genetics, McGill University, Montreal, QC H3A 0G1, Canada; [6]McGill University and Genome Quebec Innovation Centre, Montreal, QC H3A 0G1, Canada; [7]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [8]The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [9]Center of Statistical Genetics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [10]Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
*Correspondence: eimear.kenny@mssm.edu
http://dx.doi.org/10.1016/j.ajhg.2017.03.004.

The most recent release of the 1000 Genomes Project (phase 3) provides one of the largest global reference panels of whole-genome sequencing data, enabling a broad survey of human genetic variation.[19] The depth and breadth of diversity queried facilitates a deep understanding of the evolutionary forces (e.g., selection and drift) shaping existing genetic variation in present-day populations that contribute to adaptation and disease.[20–25] Studies of admixed populations have been particularly fruitful in identifying genetic adaptations and risk for diseases that are stratified across diverged ancestral origins.[26–31] Admixture patterns became especially complex during the peopling of the Americas, with extensive recent admixture spanning multiple continents. Processes shaping structure in these admixed populations include sex-biased migration and admixture, isolation-by-distance, differential drift in mainland versus island populations, and variable admixture timing.[14,32,33]

Standard GWAS strategies approach population structure as a nuisance factor. A typical stepwise procedure first detects dimensions of global population structure in each individual, using principal-component analysis (PCA) or other methods,[34–37] and often excludes "outlier" individuals from the analysis and/or corrects for inflation arising from population structure in the statistical model for association. Such strategies reduce false positives in test statistics, but can also reduce power for association in heterogeneous populations and are less likely to work for rare variant association.[38,39] Recent methodological advances have leveraged patterns of global and local ancestry for improved association power,[27,40,41] fine-mapping,[42] and genome assembly.[43] At the same time, population genetic studies have demonstrated the presence of fine-scale sub-continental structure in the African, Native American, and European components of populations from the Americas.[44–47] If trait-associated variants follow the same patterns of demography, then we expect that modeling sub-continental ancestry may enable their improved detection in admixed populations.

The dawn of the GWAS era saw limited success in identifying genome-wide significant loci associated with disease, and a major endeavor to better understand the genetic architecture of complex traits emerged. The peaks that met genome-wide significance typically did not explain a significant fraction of the phenotypic variance, and a major goal to estimate how many more signals remained yet to be discovered arose; this objective ushered in a wave of methodological development in heritability, linear mixed models, and polygenic risk prediction, as discussed and reviewed extensively elsewhere.[11,48–56] Numerous complex traits have been studied with cohort sizes in the hundreds of thousands, and yet in each case there are many more signals that improve prediction accuracy than meet genome-wide significance.[48,57–59] For example, including only genome-wide significant loci in the prediction of schizophrenia explains <3% of the phenotypic variance, whereas loci meeting the significance threshold that optimally balances signal versus noise (in this case, $p \leq 0.1$) in the meta-analysis explains considerably more (>18%) of the phenotypic variance.[11] Because the prediction accuracy, which is usually measured via prediction R^2, Nagelkerke's R^2, or receiver operator curve AUC, of polygenic risk scores is currently low for most traits,[56] genetic risk prediction is not clinically viable at present, but polygenic risk scores have nonetheless repeatedly proven valuable in research contexts across a multitude of complex traits[11,48,60–65] and will become increasingly useful as GWAS sample sizes grow.[59] Additionally, several methodological advancements to the standard approach have recently been undertaken.[58,66–68]

In this study, we explore the impact of population diversity on the landscape of variation underlying human traits. We infer demographic history for the global populations in the 1000 Genomes Project, focusing particularly on admixed populations from the Americas, which are underrepresented in medical genetic studies.[4] We disentangle local ancestry to infer the ancestral origins of these populations. We link this work to ongoing efforts to improve study design and disease variant discovery by quantifying biases in clinical databases and GWASs in diverse and admixed populations. These biases have a striking impact on genetic risk prediction; for example, a previous study calculated polygenic risk scores for schizophrenia in East Asians and Africans based on GWAS summary statistics derived from a European cohort and found that prediction accuracy was reduced by more than 50% in non-European populations.[67] To disentangle the role of demography on polygenic risk prediction derived from single-ancestry GWASs, we designed a coalescent-based simulation framework reflecting modern human population history and show that polygenic risk scores derived from European GWASs are biased when applied to diverged populations. Specifically, we identify reduced variance in risk prediction with increasing divergence from Europe reflecting decreased overall variance explained, and demonstrate that an enrichment of low-frequency risk and high-frequency protective alleles contribute to an overall protective shift in European inferred risk on average across traits. Our results highlight the need for the inclusion of more diverse populations in GWASs as well as genetic risk prediction methods improving transferability across populations.

Material and Methods

Ancestry Deconvolution

We used the phased haplotypes from the 1000 Genomes consortium. We phased reference haplotypes from 43 Native American samples from Mao et al.[69] inferred to have >0.99 Native ancestry in ADMIXTURE using SHAPEIT2 (v.2.r778),[70] then merged the haplotypes using scripts made publicly available. These combined phased haplotypes were used as input to the PopPhased version of RFMix v.1.5.4[71] with the following flags: -w 0.2, -e 1, -n 5, --use-reference-panels-in-EM, --forward-backward EM. The node size of 5 was selected to reduce bias in

random forests resulting from unbalanced reference panel sizes (AFR panel N = 504, EUR panel N = 503, and NAT panel N = 43). We used the default minimum window size of 0.2 cM to enable model comparisons with previously inferred models using *Tracts*.[72] We used 1 EM iteration to improve the local ancestry calls without substantially increasing computational complexity. We used the reference panel in the EM to take better advantage of the Native American ancestry tracts from the Hispanic/Latinos in the EM given the small NAT reference panel. We set the LWK, MSL, GWD, YRI, and ESN as reference African populations, the CEU, GBR, FIN, IBS, and TSI as reference European populations, and the samples from Mao et al.[69] with inferred >0.99 Native ancestry as reference Native American populations, as in Abecasis et al.[73]

Ancestry-Specific PCA

We performed ancestry-specific PCA, as described in Moreno-Estrada et al.[32] The resulting matrix is not necessarily orthogonalized, so we subsequently performed singular value decomposition in python 2.7 using numpy. There were a small number of major outliers, as seen previously.[32] There was one outlier (ASW individual NA20314) when analyzing the African tracts, which was expected as this individual has no African ancestry. There were eight outliers (PUR HG00731, PUR HG00732, ACB HG01880, ACB HG01882, PEL HG01944, ACB HG02497, ASW NA20320, ASW NA20321) when analyzing the European tracts. Some of these individuals had minimal European ancestry, had South or East Asian ancestry misclassified as European ancestry resulting from a limited 3-way ancestry reference panel, or were unexpected outliers. As described in the PCAmask manual, a handful of major outliers sometimes occur. As AS-PCA is an iterative procedure, we therefore removed the major outliers for each sub-continental analysis and orthogonalized the matrix on this subset.

Tracts

The RFMix output was collapsed into haploid bed files, and "UNK" or unknown ancestry was assigned where the posterior probability of a given ancestry was <0.90. These collapsed haploid tracts were used to infer admixture timings, quantities, and proportions for the ACB and PEL (new to phase 3) using *Tracts*.[72] Because the ACB have a very small proportion of Native American ancestry, we fit three 2-way models of admixture, including one model of single- and two models of double-pulse admixture events, using *Tracts*. In both of the double-pulse admixture models, the model includes an early mixture of African and European ancestry followed by another later pulse of either European or African ancestry. We randomized starting parameters and fit each model 100 times and compared the log-likelihoods of the model fits. The single-pulse and double-pulse model with a second wave of African admixture provided the best fits and reached similar log-likelihoods, with the latter showing a slight improvement in fit.

We next assessed the fit of nine different models in *Tracts* for the PEL,[72] including several two-pulse and three-pulse models. Ordering the populations as NAT, EUR, and AFR, we tested the following models: ppp_ppp, ppp_pxp, ppp_xxp, ppx_xxp, ppx_xxp_ppx, ppx_xxp_pxx, ppx_xxp_pxp, ppx_xxp_xpx, and ppx_xxp_xxp, where the order of each letter corresponds with the order of populations given above, an underscore indicates a distinct migration event with the first event corresponding with the most generations before present, p corresponds with a pulse of the ordered ancestries, and x corresponds with no input from the ordered ancestries. We tested all nine models preliminarily three times, and for all models that converged and were within the top three models, we subsequently fit each model with 100 starting parameter randomizations.

Imputation Accuracy

Imputation accuracy was calculated using a leave-one-out internal validation approach. Two array designs were compared for this analysis: Illumina OmniExpress and Affymetrix Axiom World Array LAT. Sites from these array designs were subset from chromosome 9 of the 1000 Genomes Project Phase 3 release for admixed populations. After fixing these sites, each individual was imputed using the rest of the dataset as a reference panel.

Overall imputation accuracy was binned by minor allele frequency (0.5%–1%, 1%–2%, 2%–3%, 3%–4%, 4%–5%, 5%–10%, 10%–20%, 20%–30%, 30%–40%, 40%–50%) comparing the genotyped true alleles to the imputed dosages. A second round of analyses stratified the imputation by local ancestry diplotype, which was estimated as described earlier. Within each ancestral diplotype (AFR_AFR, AFR_NAT, AFR_EUR, EUR_EUR, EUR_NAT, NAT_NAT), imputation accuracy was again estimated within MAF bins.

Empirical Polygenic Risk Score Inferences

In the most standard approach, genetic risk scores for a target cohort are generated using genome-wide summary statistics from a discovery GWAS with a set of SNPs common to both studies. From this starting set of SNPs, a further reduced set of pruned, approximately independent SNPs are then identified through a greedy clumping algorithm. Typically, progressively larger sets of SNPs defined by a range of p value thresholds (e.g., $p < 5 \times 10^{-8}$, 1×10^{-5}, 1×10^{-4}, 1×10^{-3}, 0.01, etc.) are evaluated to identify the best model balancing the signal to noise ratio to maximize phenotypic variance explained.[57,58] Once the optimal significance threshold and the final set of pruned, approximately independent set of SNPs have been selected, a polygenic risk score for each individual in a target sample is computed as the sum of the count of risk alleles weighted by the effect size (e.g., log odds ratio).

To compute polygenic risk scores in the 1000 Genomes samples using summary statistics from previous GWASs, we first filtered to biallelic SNPs and removed ambiguous AT/GC SNPs from the integrated 1000 Genome call set. To get relatively independent associations when multiple significant p value associations are in the same region in a GWAS (i.e., in LD), we performed clumping in plink using the --clump flag for all variants with MAF $\geq$ 0.01,[74] which uses a greedy algorithm ordering SNPs by p value, then selectively removes SNPs within close proximity and LD in ascending p value order (i.e., starting with the most significant SNP). As a population cohort with similar LD patterns to the study sets, we used European 1000 Genomes samples (CEU, GBR, FIN, IBS, and TSI). To compute the polygenic risk scores, we considered all SNPs with p values $\leq 1 \times 10^{-2}$ in the GWAS, a window size of 250 kb, and an R^2 threshold of 0.5 in Europeans to group SNPs. After obtaining the most significant, approximately independent signals (Table S4), we computed polygenic scores using the --score flag in plink.[74]

Polygenic Risk Score Simulations

We simulated genotypes in a coalescent framework with msprime v.1.3[75] for chromosome 20 incorporating a recombination map of GRCh37 and an assumed mutation rate of 2×10^{-8} mutations / (base pair * generation). We used a demographic model previously

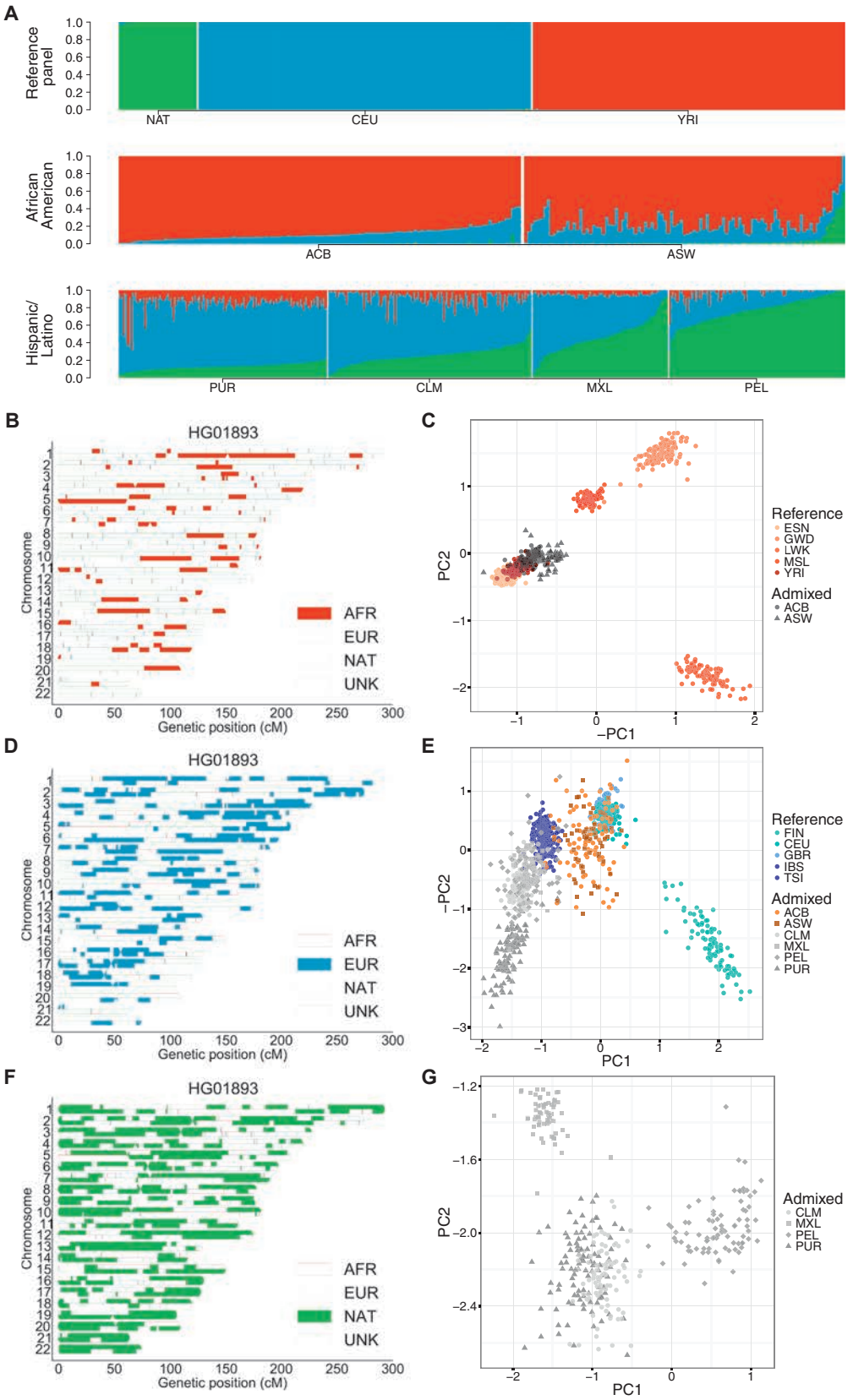


Figure 1. Sub-continental Diversity and Origins of African, European, and Native American Components of Recently Admixed American Populations

(A) ADMIXTURE analysis at K = 3 focusing on admixed Americas samples, with the NAT,[69] CEU, and YRI as reference populations.

(legend continued on next page)

inferred using 1000 Genomes sequencing data[14] to simulate individuals that reflect European, East Asian, and African population histories. We focus on these populations as the demography has previously been modeled and this avoids the challenges of simulating the geographically heterogeneous[47] and sex-biased process of admixture in the Americas.[76] To imitate a GWAS with European sample bias and evaluate polygenic risk scores in other populations, we simulated 200,000 European, 200,000 East Asian, and 200,000 African individuals. Next, we assigned "true" causal effect sizes to m evenly spaced alleles. Specifically, we randomly assigned effect sizes as

$$\beta \sim N\left(0, \frac{h^2}{m}\right)$$

where the normal distribution is specified by the mean and standard deviation (as in python's numpy package). For all other non-causal sites, the effect size is zero. We then define X as

$$X = \sum_{i=1}^{m} g_i \beta_i$$

where g_i are the genotype states (i.e., 0, 1, or 2). To handle varying allele frequencies and potential weak LD between causal sites, to ensure a neutral model with random true polygenic risks with respect to allele frequencies, and to obtain the total desired variance, we normalize X as

$$Z_X = \frac{X - \mu_X}{\sigma_X}.$$

We then compute the true polygenic risk score as

$$G = \sqrt{h^2} * Z_X$$

such that the total variance of the scores is h^2. We also simulated environmental noise and standardize to ensure equal variance between normalized genetic and environmental effects before, defining the environmental effect E as

$$\varepsilon = N\left(0, 1 - h^2\right)$$

$$Z_\varepsilon = \frac{\varepsilon - \mu_\varepsilon}{\sigma_\varepsilon}$$

$$E = \sqrt{1 - h^2} * Z_\varepsilon$$

such that the total variance of the environmental effect is $1 - h^2$. We then define the total liability as

$$L = \sqrt{h^2} * Z_X + \sqrt{1 - h^2} * Z_\varepsilon$$

$$= G + E.$$

We assigned 10,000 European individuals at the most extreme end of the liability threshold "case" status assuming a prevalence of 5%. We randomly assigned 10,000 different European individuals "control" status. We ran a GWAS with these 10,000 European case subjects and 10,000 European control subjects, computing Fisher's exact test for all sites with MAF > 0.01. As before for empirical polygenic risk score calculations from real GWAS summary statistics, we clumped these SNPs into LD blocks for all sites with $p \leq 1 \times 10^{-2}$, and $R^2 \leq 0.5$ in Europeans within a window size of 250 kb. We used these SNPs to compute inferred polygenic risk scores as before, summing the product of the log odds ratio and genotype for the true polygenic risk in a cohort of 10,000 simulated European, African, and East Asian individuals (all not included in the simulated GWAS). We compared the true versus inferred polygenic risk scores for these individuals across varying complexities (m = 200, 500, 1,000) and heritabilites (h^2 = 0.33, 0.50, 0.67).

Results

Genetic Diversity within and between Populations in the Americas

We first assessed the overall diversity at the global and sub-continental level of the 1000 Genomes Project (phase 3) populations[19] using a likelihood model via ADMIXTURE[77] and PCA[78] (Figures S1 and S2). The six populations from the Americas demonstrate considerable continental admixture, with genetic ancestry primarily from Europe, Africa, and the Americas, recapitulating previously observed population structure.[19] To quantify continental genetic diversity in these populations, we repeated the analysis using YRI, CEU, and NAT[69] samples as reference panels (population labels and abbreviations in Table S1). We observed widely varying continental admixture contributions in the six populations from the Americas at K = 3 (Figure 1A and Table S2). For example, when compared to the ASW, the ACB have a higher proportion of African ancestry ($\mu = 0.88$, 95% CI = [0.87–0.89] versus $\mu = 0.76$, 95% CI = [0.73–0.78]; two-sided t test $p = 3.0 \times 10^{-13}$) and a smaller proportion of EUR and NAT ancestry. The PEL have more NAT ancestry than all of the other AMR populations ($\mu = 0.77$, 95% CI = [0.75–0.80] versus CLM: $\mu = 0.26$, 95% CI = [0.24, 0.27], $p = 2.9 \times 10^{-95}$; PUR: $\mu = 0.13$, 95% CI = [0.12, 0.13], $p = 4.8 \times 10^{-93}$; and MXL: $\mu = 0.47$, 95% CI = [0.43, 0.50], $p = 1.7 \times 10^{-28}$) ascertained in 1000 Genomes.

We explored the origin of the subcontinental-level ancestry from recently admixed individuals by identifying local ancestry tracts[26,32,71,79] (Material and Methods, Figure S3). As proxy sources of populations for the recent admixture, we used EUR and AFR continental samples from the 1000 Genomes Project as well as NAT samples genotyped previously.[69] Concordance between global ancestry estimates inferred using ADMIXTURE at K = 5 and RFMix was typically high (Pearson's correlation ≥98%, see Figure S4). Using *Tracts*,[72] we modeled

(B, D, and F) Local ancestry karyograms for representative PEL individual HG01893 with (B) African, (D) European, and (F) Native American components shown.
(C, E, and G) Ancestry-specific PCA applied to admixed haploid genomes as well as ancestrally homogeneous continental reference populations from 1000 Genomes (where possible) for (C) African tracts, (E) European tracts, and (G) Native American tracts. A small number of admixed samples that constituted major outliers from the ancestry-specific PCA analysis were removed, including (C) one ASW sample (NA20314) and (E) eight samples, including three ACB, two ASW, one PEL, and two PUR samples.

the length distribution of the AFR, EUR, and NAT tracts to infer that admixing began ~12 and ~8 generations ago in the PEL and ACB populations, respectively (Figure S5), consistent with previous estimates from other populations from the Americas.[44,72,32]

We further investigated the subcontinental ancestry of admixed populations from the Americas one ancestry at a time using a version of PCA modified to handle highly masked data (ancestry-specific or AS-PCA) as implemented in PCAmask.[32] Example ancestry tracts in a PEL individual subset to AFR, EUR, and NAT components are shown in Figures 1B, 1D, and 1F, respectively. Consistent with previous observations, the inferred European tracts in Hispanic/Latino populations most closely resemble southern European IBS and TSI populations with some additional drift[32] (Figure 1E). The European tracts of the PUR are more differentiated compared to the CLM, MXL, and PEL populations, consistent with sex bias (Figure S6 and Table S3) and excess drift from founder effects in this island population.[32] In contrast to the southern European tracts from the Hispanic/Latino populations, the African descent populations in the Americas have European admixture that more closely resembles the northwestern CEU and GBR European populations. The clusters are less distinct, owing to lower overall fractions of European ancestry, but the European components of the Hispanic/Latino and African American populations are significantly different (Wilcoxon rank sum test $p = 2.4 \times 10^{-60}$).

The ability to localize aggregated ancestral genomic tracts enables insights into the evolutionary origins of admixed populations. To disentangle whether the considerable Native American ancestry in the ASW individuals arose from recent admixture with Hispanic/Latino individuals or recent admixture with indigenous Native American populations, we queried the European tracts. We find that the European tracts of all ASW individuals with considerable Native American ancestry are well within the ASW cluster and project closer in Euclidean distance with AS-PC1 and AS-PC2 to northwestern Europe than the European tracts from Hispanic/Latino samples ($p = 1.15 \times 10^{-3}$), providing support for the latter hypothesis and providing regional nuance to previous findings.[44]

We also investigated the African origin of the admixed AFR/AMR populations (ACB and ASW), as well as the Native American origin of the Hispanic/Latino populations (CLM, MXL, PEL, and PUR). The African tracts of ancestry from the AFR/AMR populations project closer to the YRI and ESN of Nigeria than the GWD, MSL, and LWK populations (Figure 1C). This is consistent with slave records and previous genome-wide analyses of African Americans indicating that most sharing occurred in West and West-Central Africa.[80–82] There are subtle differences between the African origins of the ACB and ASW populations (e.g., difference in distance from YRI on AS-PC1 and AS-PC2 $p = 6.4 \times 10^{-6}$), likely due either to mild island founder effects in the ACB samples or differences in African source populations for enslaved Africans who remained in Barbados versus those who were brought to the USA. The Native tracts of ancestry from the AMR populations first separate the southernmost PEL populations from the CLM, MXL, and PUR on AS-PC1, then separate the northernmost MXL from the CLM and PUR on AS-PC2, consistent with a north-south cline of divergence among indigenous Native American ancestry (Figure 1G).[32,83]

Impact of Continental and Sub-continental Diversity on Disease Variant Mapping

To investigate the role of ancestry in phenotype interpretation from genetic data, we assessed diversity across populations and local ancestries for recently admixed populations across the whole genome and sites from two reference databases: the GWAS catalog and ClinVar pathogenic and likely pathogenic sites. We recapitulate results showing that there is less variation across the genome (both genome-wide and on the Affymetrix 6.0 GWAS array sites used in local ancestry calling) in out-of-Africa versus African populations, but that GWAS variants are more polymorphic in European and Hispanic/Latino populations (Figures S7A, S7B, S8A, and S8B). We use a normalized measure of the minor allele frequency, an indicator of the amount of diversity captured in a population, to obtain a background coverage of each population, as done previously (e.g., Figure S4 from Auton et al.[19]). We show that the Affymetrix 6.0 array has a slight European bias (Figures S5A and S6A). We compared the site frequency spectrum of variants across the genome versus at GWAS catalog sites and identify elevated allele frequencies at GWAS catalog loci, particularly in populations with more European ancestry (e.g., the EUR, AMR, and SAS super populations, Figures S5C and S5D). We further compared heterozygosity (estimated here as $2pq$) and the site frequency spectrum in recently admixed populations across diploid and haploid local ancestry tracts, respectively. Sites in the GWAS catalog and ClinVar are more and less common than genome-wide variants, respectively (Figure 2). Whereas heterozygosity across the whole genome is highest in African ancestry tracts, it is consistently the greatest in European ancestry tracts across these databases (Figures 2, S8C, and S8D), reflecting a strong bias toward European study participants.[1–4,19,84] These results highlight imbalances in genome interpretability across local ancestry tracts in recently admixed populations and the utility of analyzing these variants jointly with these ancestry tracts over genome-wide ancestry estimates alone.

We also assessed imputation accuracy across the 3-way admixed populations from the Americas (CLM, MXL, PEL, PUR) for two arrays: the Illumina OmniExpress and the Affymetrix Axiom World Array LAT. Imputation accuracy was estimated as the correlation (r^2) between the original genotypes and the imputed dosages. For both array designs, imputation accuracy across all minor allele frequency (MAF) bins was highest for populations with the largest proportion of European ancestry (PUR) and

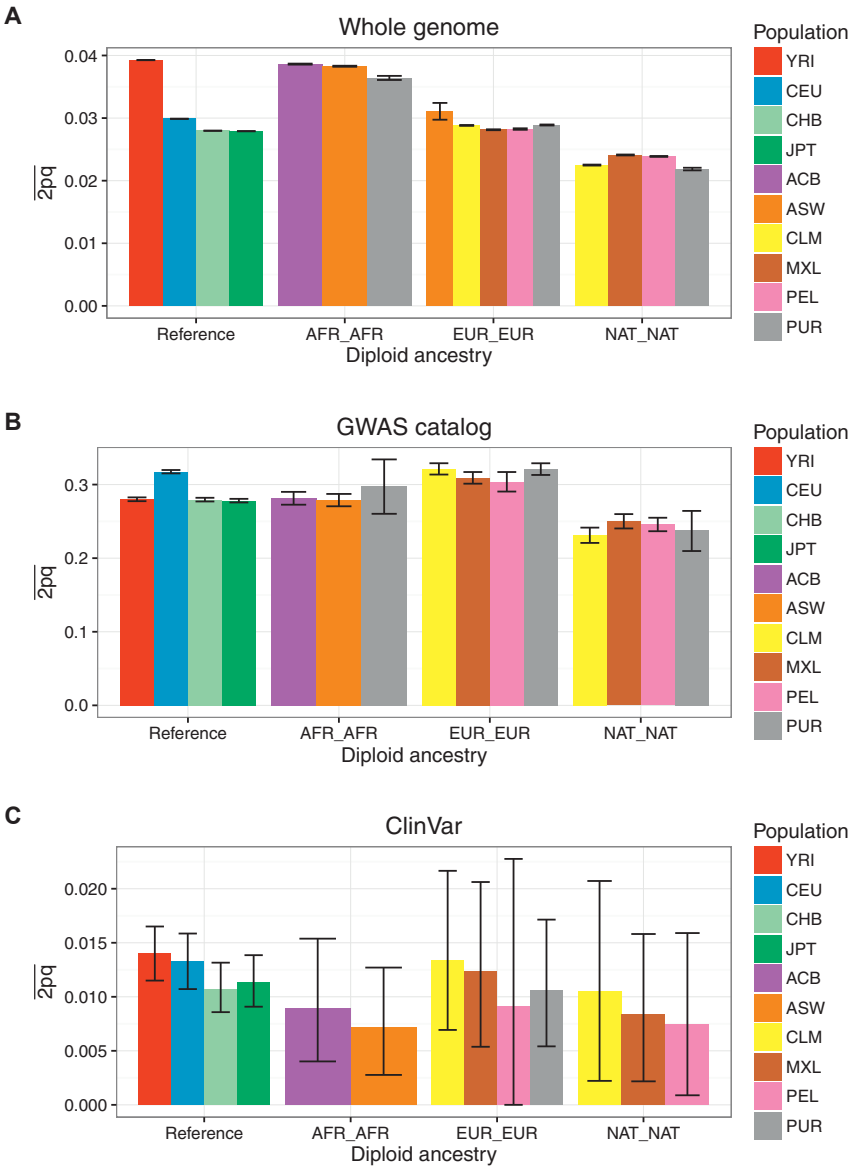


Figure 2. Heterozygosity by Continental and Diploid Local Ancestry

Heterozygosity, estimated here as 2pq, is calculated in admixed populations stratified by diploid local ancestry in (A) the whole genome, (B) sites from the GWAS catalog, and (C) sites from ClinVar classified as "pathogenic" or "likely pathogenic." The mean and 95% confidence intervals were calculated by bootstrapping 1,000 times. Populations not shown in a given panel have too few diploid ancestry tracts overlapping sites to calculate heterozygosity.

lowest for populations with the largest proportion of Native American ancestry (PEL, Figures S9A and S9B). We also stratified imputation accuracy by local ancestry tract diplotype within the Americas. Consistently, tracts with at least one Native American ancestry tract had lower imputation accuracy when compared to tracts with only European and/or African ancestry (Figures 3 and S10).

Transferability of GWAS Findings across Populations

To quantify the transferability of European-biased genetic studies to other populations, we next used published GWAS summary statistics to infer polygenic risk scores[48] across populations for well-studied traits, including height,[10] waist-hip ratio,[85] schizophrenia,[11] type II diabetes,[86,87] and asthma[88] (Figures 4A–4D and S11, Material and Methods). Most of these summary statistics are derived from studies with primarily European cohorts, although GWASs of type II diabetes have been performed in both European-specific cohorts as well as across multi-ethnic cohorts. We identify clear directional inconsistencies in these inferred scores. For example, although the height summary statistics show the expected southern/northern cline of increasing European height (FIN, CEU, and GBR populations have significantly higher polygenic risk scores than IBS and TSI, $p = 1.5 \times 10^{-75}$, Figure S9A), polygenic scores for height across super populations show biased predictions; the African populations sampled are genetically predicted to be considerably shorter than all Europeans and minimally taller than East Asians (Figure 4A), which contradicts empirical observations (with the exception of some indigenous pygmy/pygmoid populations).[89,90] Additionally, polygenic risk scores for schizophrenia, while at a similar prevalence across populations where it has been well studied[91] and sharing significant genetic risk across populations,[92] shows considerably decreased scores in Africans compared to all other populations (Figure 4B). Lastly, the relative order of polygenic risk scores computed for type II diabetes across populations differs depending on whether the summary statistics are derived from a European-specific (Figure 4C) or multi-ethnic (Figure 4D) cohort.

Ancestry-Specific Biases in Polygenic Risk Score Estimates

We performed coalescent simulations to determine how GWAS signals discovered in one ancestral case/control cohort (i.e., "single-ancestry" GWAS) are expected to impact polygenic risk score estimates in other populations under neutrality using summary statistics (for details, see Material and Methods). In brief, we simulated variants according to a previously published demographic model inferred from Africans, East Asians, and Europeans.[14] We

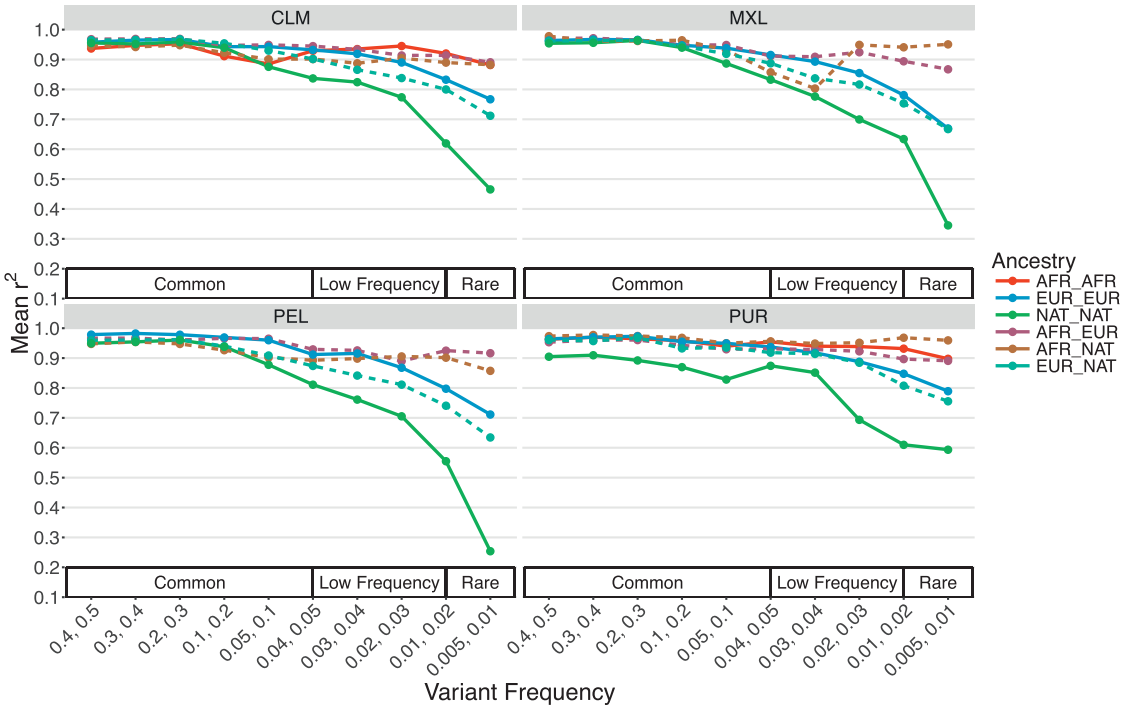


Figure 3. Imputation Accuracy by Local Ancestry in the Americas
Accuracy was assessed via a leave-one-out strategy, stratified by diploid local ancestry on chromosome 9 for the Illumina OmniExpress genotyping array. Dashed lines indicate heterozygous diploid ancestry, and solid lines show homozygous diploid ancestry.

specified "causal" alleles and effect sizes randomly, such that each causal variant has evolved neutrally and has a mean effect of zero with the standard deviation equal to the global heritability divided by number of causal variants. We computed the true polygenic risk for each individual as the product of the estimated effect sizes and genotypes, then standardized the scores across all individuals. We calculated the total liability as the sum of the genetic and random environmental contributions, then identified 10,000 European case subjects with the most extreme liabilities and 10,000 other European control subjects. We computed Fisher's exact tests with this European case-control cohort, then quantified inferred polygenic risk scores as the sum of the product of genotypes and log odds ratios for 10,000 samples per population not included in the GWAS.

In our simulations and consistent with realistic coalescent models, most variants are rare and population specific; "causal" variants are sampled from the global site frequency spectrum, resulting in subtle differences in true polygenic risk across populations (Figures S12, 5A, and 5B). We mirrored standard practices for performing a GWAS and computing polygenic risk scores (see above and Material and Methods). While causal variants in our simulations are drawn from the global site frequency spectrum and are therefore mostly rare, inferred scores are derived specifically from common variants that are typically much more common in the study population than elsewhere (here Europeans with case/control MAF $\geq$ 0.01). Consequently, while the distribution of mean true polygenic risk across simulation runs for each population are not significantly different (Figure 5A), the inferred risk is less than zero in Europeans ($p = 1.9 \times 10^{-54}$, 95% CI = [−84.3, −67.4]), slightly less than zero in East Asians ($p = 5.9 \times 10^{-5}$, 95% CI = [−19.1, −6.6]), and not significantly different from zero in Africans (Figure 5B); the variance in inferred risk scores, a proxy for the fraction of heritable variation explained, also decreases with this trend. Specifically, when $h^2 = 0.67$ and $m = 1{,}000$ causal markers, we find that the true and inferred polygenic risk scores in the EUR population are significantly correlated (i.e., non-zero, mean $\rho = 0.59$, $p < 1 \times 10^{-200}$), but the correlations in EAS and AFR populations are significantly less than in EUR ($\rho = 0.35$ and $p = 1.5 \times 10^{-48}$, $\rho = 0.22$ and $p < 1 \times 10^{-200}$, respectively). Because of allele frequency differences, number of SNPs, and inferred effect size differences along the frequency spectrum, the scale is orders of magnitude different between the true and inferred raw, unstandardized scores, cautioning that while they are informative on a relative scale (Figures 5C and S11), their absolute scale should not be over interpreted. The inferred risk difference between populations is driven by the increased power to detect minor risk alleles rather than protective alleles in the study population,[93] given the differential selection of case and control subjects in the liability threshold model. We demonstrate this empirically in these neutral simulations within the European population (Figure S14A), indicating that this phenomenon occurs even in the absence of population structure and when case and control cohort sizes are equal.

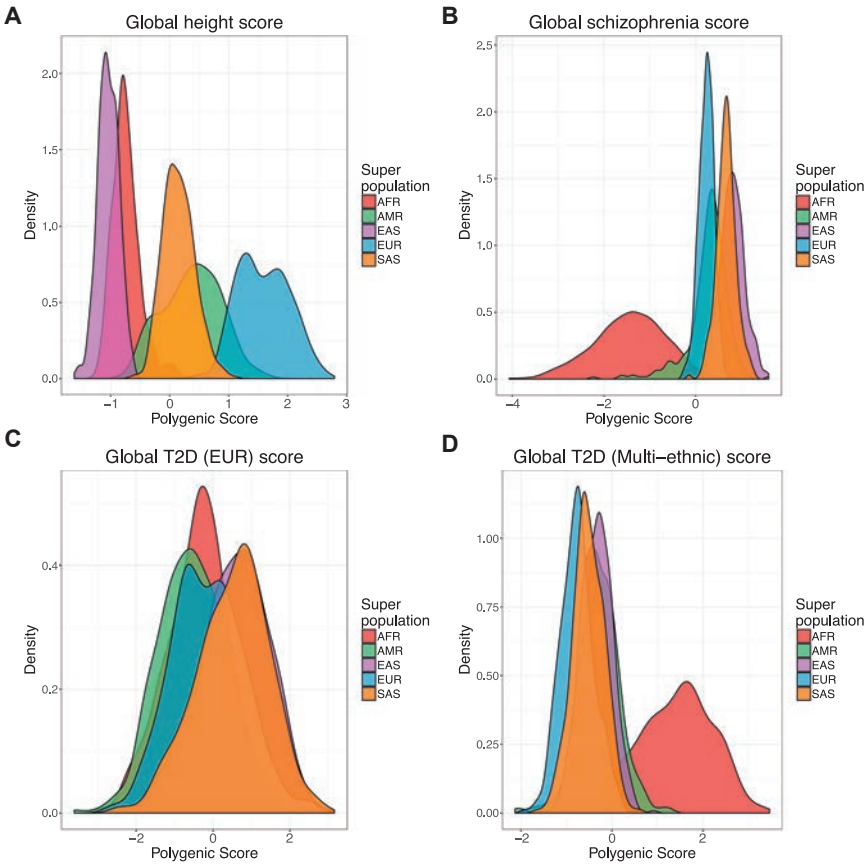


Figure 4. Biased Genetic Discoveries Influence Disease Risk Inferences
Inferred and standardized polygenic risk scores across all individuals and colored by population for (A) height based on summary statistics from Wood et al.,[10] (B) schizophrenia based on summary statistics from the Schizophrenia Working Group of the Psychiatric Genomics Consortium,[11] (C) type II diabetes summary statistics derived from a European cohort from Gaulton et al.,[86] and (D) type II diabetes summary statistics derived from a multi-ethnic cohort from Mahajan et al.[87]

We find that the correlation between true and inferred polygenic risk is generally low (Figures 5C and S13), consistent with limited variance explained by polygenic risk scores from GWASs of these cohort sizes for height (e.g., ~10% of variance explained for a cohort of size 183,727[63]) and schizophrenia (e.g., ~7% variance explained for a cohort of size 36,989 case subjects and 113,075 control subjects[11]). Low correlations in our simulations are most likely because common tag variants are a poor proxy for rare causal variants. As expected, correlations between true and inferred risk within populations are typically highest in the European population (i.e., the population in which variants were discovered, Figures 5A and S13). To quantify the differential prediction accuracy of polygenic risk scores across populations, we also evaluate the log odds ratio of being a case subject compared to a control subject across deciles of inferred polygenic risk in each population. We identify greater power to discern between case and control subjects in the EUR discovery population relative to the AFR and EAS populations (i.e., more heritable variation explained, as evidenced by a steeper slope) (Figure S14B). Across all populations, the mean Spearman correlations between true and inferred polygenic risk increase with increasing heritability while the standard deviations of these correlations significantly decrease ($p = 0.05$); however, there is considerable within-population heterogeneity resulting in high variation in scores across all populations. We find that in these neutral simulations, a polygenic risk score bias in essentially any direction is possible even when choosing the exact same causal variants and heritability and varying only fixed effect size (i.e., inferred polygenic risk in Europeans can be higher, lower, or intermediate compared to true risk relative to East Asians or Africans, Figures S12 and 5B).

Discussion

To date, GWASs have been performed opportunistically in primarily single-ancestry European cohorts, and an open question remains about their biomedical relevance for disease associations in other ancestries. As studies gain power by increasing sample sizes, effect size estimates become more precise and novel associations at lower frequencies are feasible. However, rare variants are largely population-private, and their effects are unlikely to transfer to new populations. Because linkage disequilibrium and allele frequencies vary across ancestries, effect size estimates from diverse cohorts are typically more precise than from single-ancestry cohorts (and often tempered),[5] and the resolution of causal variant fine-mapping is considerably improved.[87] Across a range of genetic architectures, diverse cohorts provide the opportunity to reduce false positives. At the Mendelian end of the spectrum, for example, disentangling risk variants with incomplete penetrance from benign false positives and localizing functional effects in genes is much more feasible with large diverse population cohorts than with single-ancestry analyses.[94] Multiple false positive reports of pathogenic variants causing hypertrophic cardiomyopathy, a disease with relatively simple genomic architecture, have been returned to individuals of African descent or unspecified ancestry that would have been prevented if even a small number of African American samples were included in control cohorts.[9] At the highly complex end of the polygenicity spectrum, we and others have shown that the utility of polygenic risk inferences and

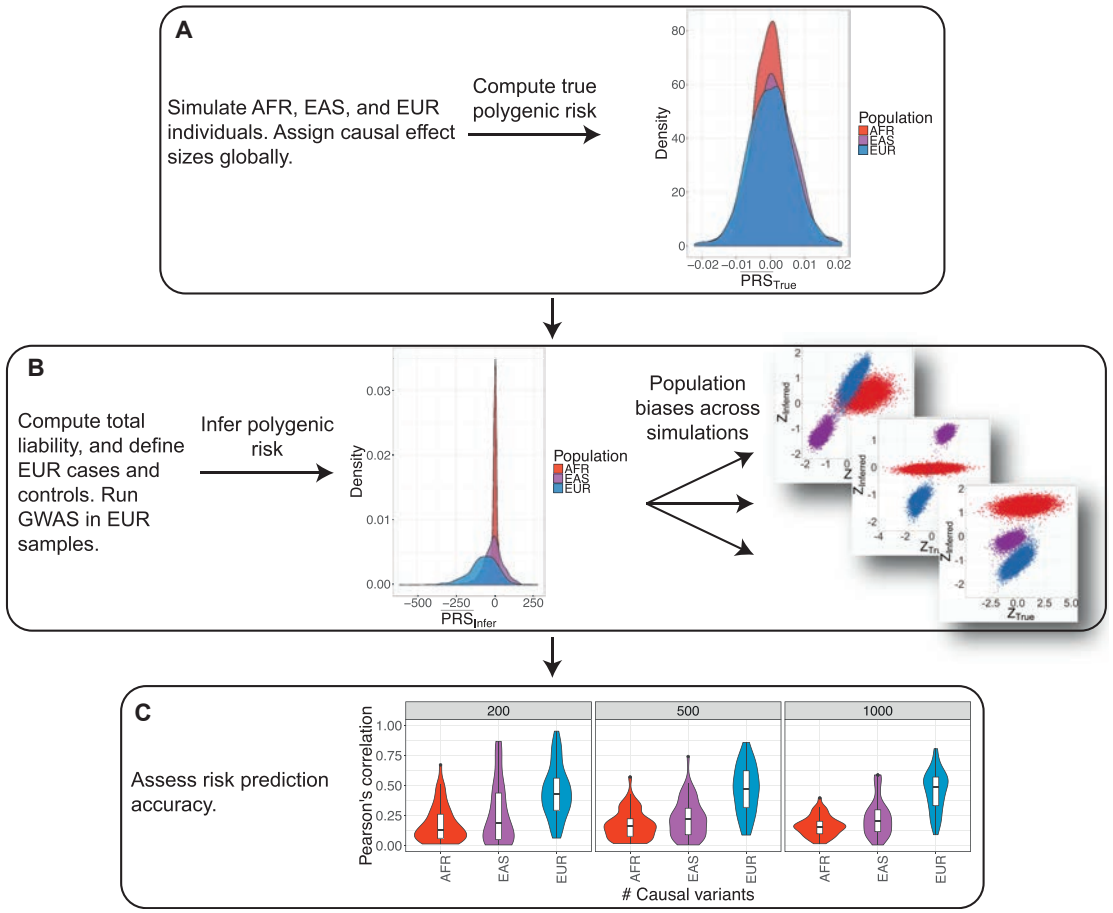


Figure 5. Coalescent Simulation Framework to Generate True and Inferred Polygenic Risk Scores
Results of true and inferred polygenic risk scores, as well as their correlation, were computed via GWAS summary statistics from 10,000 simulated EUR case and control subjects modeling European, East Asian, and African population history (demographic parameters are from Gravel et al.[14]).
(A) The distribution of mean true, unstandardized polygenic risk scores for each population across 500 simulations with $m = 1{,}000$ causal variants and $h^2 = 0.67$.
(B) The distribution of mean inferred, unstandardized polygenic risk for the same simulation parameters as in (A) (center) and standardized true versus inferred polygenic risk scores for three different coalescent simulation replicates showing 10,000 randomly drawn samples from each population not included as case or control subjects (right).
(C) Violin plots show Pearson's correlation across 50 iterations per parameter set between true and inferred polygenic risk scores across differing genetic architectures, including $m = 200$, 500, and 1,000 causal variants and $h^2 = 0.67$.

the heritable phenotypic variance explained in diverse populations is improved with more diverse cohorts.[92,95]

Standard single-ancestry GWASs typically apply linear mixed model approaches and/or incorporate principal components as covariates to control for confounding from population structure with primarily European-descent cohorts.[1–3] A key concern when including multiple diverse populations in a GWAS is that there is increasing likelihood of identifying false positive variants associated with disease that are driven by allele frequency differences across ancestries. However, previous studies have analyzed association data for diverse ancestries and replicated findings across ethnicities, assuaging these concerns.[6,87] In this study, we show that this ancestry stratification is not continuous along the genome: long tracts of ancestrally diverse populations present in admixed samples from the Americas are easily and accurately detected. Querying population substructure within these tracts recapitulates expected trends, e.g., European ancestry in African Americans primarily descends from northern Europeans in contrast to European ancestry from Hispanic/Latinos, which primarily descends from southern Europeans, as seen previously.[44] Additionally, population substructure follows a north-south cline in the Native component of Hispanic/Latinos, and the African component of admixed African descent populations in the Americas most closely resembles reference populations from Nigeria (notwithstanding the limited set of African populations from the 1000 Genomes Project). Admixture mapping has been successful at large sample sizes for identifying ancestry-specific genetic risk factors for disease.[30] Given the level of accuracy and subcontinental resolution attained with local ancestry tracts in admixed populations, we emphasize the utility

of a unified framework to jointly analyze genetic associations with local ancestry simultaneously.[40]

The transferability of GWASs is aided by the inclusion of diverse populations.[96] We have shown that European discovery biases in GWASs are recapitulated in local ancestry tracts in admixed samples. We have quantified GWAS study biases in ancestral populations and shown that GWAS variants are at lower frequency specifically within African and Native tracts and higher frequency in European tracts in admixed American populations. Imputation accuracy is also stratified across diverged ancestries, including across local ancestries in admixed populations. With decreased imputation accuracy especially on Native American tracts, there is decreased power for potential ancestry-specific associations. This differentially limits conclusions for GWASs in an admixed population in a two-pronged manner: the ability to capture variation and the power to estimate associations.

As GWASs scale to sample sizes on the order of hundreds of thousands to millions, genetic risk prediction accuracy at the individual level improves.[59] However, we show that the utility of polygenic risk scores computed using GWAS summary statistics are dependent on genetic similarity to the discovery cohort. Best linear unbiased prediction (BLUP) methods have been proposed to improve risk scores, but they require access to raw genetic data typically from very large datasets, are also dependent on LD structure in the study population, and offer only modest improvements in prediction accuracy.[52] Furthermore, polygenic risk scores (PRSs) contain a mix of true positives (which have the bias described above) and false positives in the training GWAS. False positives, being chance statistical fluctuations, do not have the same allele frequency bias and therefore unfortunately play an outsized role in applying a PRS in a new population.

We have demonstrated that polygenic risk scores computed via current standard methods with summary statistics from a single-ancestry discovery cohort have numerous problems: differences in polygenic risk scores across populations are significant but not supported by epidemiological or anthropometric studies of the same traits, and directionality biases in polygenic risk scores across populations are unpredictable. Our coalescent simulations recapitulate these results and show that across replicates (i.e., traits, and thus not necessarily within a single trait), cross-population prediction accuracy is diminished with increasing divergence from the discovery cohort. These simulations provide further insight into directional inconsistencies in inferred polygenic risk scores with the same demographic model across replicate simulations, indicating that different traits are likely to suffer from biases that cannot be adjusted, e.g., using prinicipal components alone. Directional selection is expected to bias polygenic risk inferences even more. Because biases arise from genetic drift alone, we recommend (1) avoiding interpretations from polygenic risk score differences extrapolated across populations, as these are likely confounded by latent population structure that is not properly corrected for with current standard methods, (2) mean-centering polygenic risk scores for each population, and (3) computing polygenic risk scores in populations with similar demographic histories as the study sample to ensure maximal predictive power. Further, additional methods that account for local ancestry in genetic risk prediction to incorporate different ancestral linkage disequilibrium and allele frequencies are needed. This study demonstrates the utility of disentangling ancestry tracts in recently admixed populations for inferring recent demographic history and identifying ancestry-stratified analytical biases; we also motivate the need to include more ancestrally diverse cohorts in GWASs to ensure that health disparities arising from genetic risk prediction do not become pervasive in individuals of admixed and non-European descent.

Supplemental Data

Supplemental Data include 14 figures and 4 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2017.03.004.

Conflicts of Interest

C.D.B. is a member of the scientific advisory boards for Liberty Biosecurity, Personalis, 23andMe Roots into the Future, Ancestry.com, IdentifyGenomics, and Etalon and is a founder of CDB Consulting. C.R.G. owns stock in 23andMe. M.J.D. is a member the scientific advisory board for Ancestry.com. E.E.K. and C.R.G. are members of the scientific advisory board for Encompass Biosciences. E.E.K. consults for Illumina. B.M.N. is a member of the scientific advisory board for Deep Genomics.

Acknowledgments

We thank Suyash Shringarpure, Brian Maples, Andres Moreno-Estrada, Danny Park, Noah Zaitlen, Alexander Gusev, and Alkes Price for helpful discussions/feedback. We thank Verneri Antilla for providing GWAS summary statistics. We thank Jerome Kelleher for several conversations about msprime, providing example scripts, and implementing new simulation capabilities. This work was supported by funds from several grants: the National Human Genome Research Institute under award numbers U01HG009080 (E.E.K., C.D.B., C.R.G.), U01HG007419 (C.D.B., C.R.G., G.L.W.), U01HG007417 (E.E.K.), U01HG005208 (M.J.D.), T32HG000044 (C.R.G.), and R01GM083606 (C.D.B.), the National Institute of General Medical Sciences under award number T32GM007790 (A.R.M.) at the National Institute of Health, the National Institute for Mental Health 5U01MH094432-02 (R.G.W., M.J.D.), the Directorate of Mathematical and Physical Sciences award 1201234 (S.G., C.D.B.) at the National Science Foundation, the Canadian Institutes of Health Research through the Canada Research Chair program and operating grant MOP-136855 (S.G.), and a Sloan Research Fellowship (S.G.).

Received: November 23, 2016
Accepted: March 10, 2017
Published: March 30, 2017

Web Resources

ancestry_pipeline, https://github.com/armartin/ancestry_pipeline/
Local ancestry calls, https://personal.broadinstitute.org/armartin/tgp_admixture/
msprime, https://github.com/jeromekelleher/msprime
PCAmask, https://sites.google.com/site/pcamask/dowload
Phased 1000 Genomes haplotypes, ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/supporting/shapeit2_scaffolds/wgs_gt_scaffolds/
Tracts, https://github.com/sgravel/tracts

References

1. Need, A.C., and Goldstein, D.B. (2009). Next generation disparities in human genomics: concerns and remedies. Trends Genet. *25*, 489–494.
2. Bustamante, C.D., Burchard, E.G., and De la Vega, F.M. (2011). Genomics for the world. Nature *475*, 163–165.
3. Petrovski, S., and Goldstein, D.B. (2016). Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. Genome Biol. *17*, 157.
4. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164.
5. Carlson, C.S., Matise, T.C., North, K.E., Haiman, C.A., Fesinmeyer, M.D., Buyske, S., Schumacher, F.R., Peters, U., Franceschini, N., Ritchie, M.D., et al.; PAGE Consortium (2013). Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. PLoS Biol. *11*, e1001661.
6. Waters, K.M., Stram, D.O., Hassanein, M.T., Le Marchand, L., Wilkens, L.R., Maskarinec, G., Monroe, K.R., Kolonel, L.N., Altshuler, D., Henderson, B.E., and Haiman, C.A. (2010). Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. PLoS Genet. *6*, 6.
7. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.
8. Scutari, M., Mackay, I., and Balding, D. (2016). Using genetic distance to infer the accuracy of genomic prediction. PLoS Genet. *12*, e1006288.
9. Manrai, A.K., Funke, B.H., Rehm, H.L., Olesen, M.S., Maron, B.A., Szolovits, P., Margulies, D.M., Loscalzo, J., and Kohane, I.S. (2016). Genetic misdiagnoses and the potential for health disparities. N. Engl. J. Med. *375*, 655–665.
10. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al.; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; and LifeLines Cohort Study (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.
11. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
12. Muñoz, M., Pong-Wong, R., Canela-Xandri, O., Rawlik, K., Haley, C.S., and Tenesa, A. (2016). Evaluating the contribution of genetics and familial shared environment to common disease using the UK Biobank. Nat. Genet. *48*, 980–983.
13. Mathieson, I., and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nat. Genet. *44*, 243–246.
14. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Bustamante, C.D.; and 1000 Genomes Project (2011). Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA *108*, 11983–11988.
15. Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., Lawson, D., et al.; UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.
16. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within Europe. Nature *456*, 98–101.
17. Do, R., Kathiresan, S., and Abecasis, G.R. (2012). Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum. Mol. Genet. *21* (R1), R1–R9.
18. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
19. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
20. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al.; Broad GO; Seattle GO; and NHLBI Exome Sequencing Project (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.
21. Grossman, S.R., Shlyakhter, I., Karlsson, E.K., Byrne, E.H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., et al. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. Science *327*, 883–886.
22. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al.; 1000 Genomes Project Consortium (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823–828.
23. Lohmueller, K.E., Indap, A.R., Schmidt, S., Boyko, A.R., Hernandez, R.D., Hubisz, M.J., Sninsky, J.J., White, T.J., Sunyaev, S.R., Nielsen, R., et al. (2008). Proportionally more deleterious genetic variation in European than in African populations. Nature *451*, 994–997.
24. Fu, W., Gittelman, R.M., Bamshad, M.J., and Akey, J.M. (2014). Characteristics of neutral and deleterious protein-coding variation among individuals and populations. Am. J. Hum. Genet. *95*, 421–436.
25. Simons, Y.B., Turchin, M.C., Pritchard, J.K., and Sella, G. (2014). The deleterious mutation load is insensitive to recent population history. Nat. Genet. *46*, 220–224.
26. Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. *5*, e1000519.

27. Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G.K., Tandon, A., Kao, W.H.L., Ruczinski, I., Fornage, M., Siscovick, D.S., Zhu, X., et al. (2011). Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. PLoS Genet. *7*, e1001371.
28. Fejerman, L., Chen, G.K., Eng, C., Huntsman, S., Hu, D., Williams, A., Pasaniuc, B., John, E.M., Via, M., Gignoux, C., et al. (2012). Admixture mapping identifies a locus on 6q25 associated with breast cancer risk in US Latinas. Hum. Mol. Genet. *21*, 1907–1917.
29. Fejerman, L., Ahmadiyeh, N., Hu, D., Huntsman, S., Beckman, K.B., Caswell, J.L., Tsung, K., John, E.M., Torres-Mejia, G., Carvajal-Carmona, L., et al.; COLUMBUS Consortium (2014). Genome-wide association study of breast cancer in Latinas identifies novel protective variants on 6q25. Nat. Commun. *5*, 5260.
30. Freedman, M.L., Haiman, C.A., Patterson, N., McDonald, G.J., Tandon, A., Waliszewska, A., Penney, K., Steen, R.G., Ardlie, K., John, E.M., et al. (2006). Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc. Natl. Acad. Sci. USA *103*, 14068–14073.
31. Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S., Myers, S., Tandon, A., Spencer, C., et al. (2011). Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. Am. J. Hum. Genet. *89*, 368–381.
32. Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J.L., Byrnes, J.K., Gignoux, C.R., Ortiz-Tello, P.A., Martínez, R.J., Hedges, D.J., Morris, R.W., et al. (2013). Reconstructing the population genetic history of the Caribbean. PLoS Genet. *9*, e1003925.
33. Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C.D., and Ostrer, H. (2010). Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proc. Natl. Acad. Sci. USA *107* (*Suppl 2*), 8954–8961.
34. Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.
35. Tang, H., Peng, J., Wang, P., and Risch, N.J. (2005). Estimation of individual admixture: analytical and study design considerations. Genet. Epidemiol. *28*, 289–301.
36. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.
37. Price, A.L., Zaitlen, N.A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. *11*, 459–463.
38. Mathieson, I., and McVean, G. (2014). Demography and the age of rare variants. PLoS Genet. *10*, e1004528.
39. O'Connor, T.D., Fu, W., Mychaleckyj, J.C., Logsdon, B., Auer, P., Carlson, C.S., Leal, S.M., Smith, J.D., Rieder, M.J., Bamshad, M.J., et al.; NHLBI GO Exome Sequencing Project; and ESP Population Genetics and Statistical Analysis Working Group, Emily Turner (2015). Rare variation facilitates inferences of fine-scale population structure in humans. Mol. Biol. Evol. *32*, 653–660.
40. Szulc, P., Bogdan, M., Frommlet, F., and Tang, H. (2016). Joint genotype- and ancestry-based genome-wide association studies in admixed populations. bioRxiv. http://dx.doi.org/10.1101/062554.
41. Conomos, M.P., Reiner, A.P., Weir, B.S., and Thornton, T.A. (2016). Model-free estimation of recent genetic relatedness. Am. J. Hum. Genet. *98*, 127–148.
42. Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E., and Halperin, E. (2010). Leveraging genetic variability across populations for the identification of causal variants. Am. J. Hum. Genet. *86*, 23–33.
43. Genovese, G., Handsaker, R.E., Li, H., Kenny, E.E., and McCarroll, S.A. (2013). Mapping the human reference genome's missing sequence by three-way admixture in Latino genomes. Am. J. Hum. Genet. *93*, 411–421.
44. Baharian, S., Barakatt, M., Gignoux, C.R., Shringarpure, S., Errington, J., Blot, W.J., Bustamante, C.D., Kenny, E.E., Williams, S.M., Aldrich, M.C., and Gravel, S. (2016). The great migration and African-American genomic diversity. PLoS Genet. *12*, e1006059.
45. Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., Mesa, N., et al. (2012). Reconstructing Native American population history. Nature *488*, 370–374.
46. Ruiz-Linares, A., Adhikari, K., Acuña-Alonzo, V., Quinto-Sanchez, M., Jaramillo, C., Arias, W., Fuentes, M., Pizarro, M., Everardo, P., de Avila, F., et al. (2014). Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. PLoS Genet. *10*, e1004572.
47. Moreno-Estrada, A., Gignoux, C.R., Fernández-López, J.C., Zakharia, F., Sikora, M., Contreras, A.V., Acuña-Alonzo, V., Sandoval, K., Eng, C., Romero-Hidalgo, S., et al. (2014). Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science *344*, 1280–1285.
48. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., Sklar, P.; and International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.
49. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
50. Yang, J., Weedon, M.N., Purcell, S., Lettre, G., Estrada, K., Willer, C.J., Smith, A.V., Ingelsson, E., O'Connell, J.R., Mangino, M., et al.; GIANT Consortium (2011). Genomic inflation factors under polygenic inheritance. Eur. J. Hum. Genet. *19*, 807–812.
51. Wray, N.R., Goddard, M.E., and Visscher, P.M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. *17*, 1520–1528.
52. Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., and Visscher, P.M. (2013). Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. *14*, 507–515.
53. Wray, N.R., Lee, S.H., Mehta, D., Vinkhuyzen, A.A., Dudbridge, F., and Middeldorp, C.M. (2014). Research review: polygenic methods and their application to psychiatric traits. J. Child Psychol. Psychiatry *55*, 1068–1087.
54. Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat. Rev. Genet. *17*, 392–406.
55. Dudbridge, F. (2016). Polygenic epidemiology. Genet. Epidemiol. *40*, 268–272.

56. So, H.C., and Sham, P.C. (2017). Exploring the predictive power of polygenic scores derived from genome-wide association studies: a study of 10 complex traits. Bioinformatics *33*, 886–892.
57. Euesden, J., Lewis, C.M., and O'Reilly, P.F. (2015). PRSice: polygenic risk score software. Bioinformatics *31*, 1466–1468.
58. Shi, J., Park, J.H., Duan, J., Berndt, S.T., Moy, W., Yu, K., Song, L., Wheeler, W., Hua, X., Silverman, D., et al.; MGS (Molecular Genetics of Schizophrenia) GWAS Consortium; GECCO (The Genetics and Epidemiology of Colorectal Cancer Consortium); GAME-ON/TRICL (Transdisciplinary Research in Cancer of the Lung) GWAS Consortium; PRACTICAL (PRostate cancer AssoCiation group To Investigate Cancer Associated aLterations) Consortium; PanScan Consortium; and GAME-ON/ELLIPSE Consortium (2016). Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. PLoS Genet. *12*, e1006493.
59. Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. PLoS Genet. *9*, e1003348.
60. Pharoah, P.D., Antoniou, A.C., Easton, D.F., and Ponder, B.A. (2008). Polygenes, risk prediction, and targeted prevention of breast cancer. N. Engl. J. Med. *358*, 2796–2803.
61. Evans, D.M., Visscher, P.M., and Wray, N.R. (2009). Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum. Mol. Genet. *18*, 3525–3531.
62. Okbay, A., Beauchamp, J.P., Fontana, M.A., Lee, J.J., Pers, T.H., Rietveld, C.A., Turley, P., Chen, G.B., Emilsson, V., Meddens, S.F., et al.; LifeLines Cohort Study (2016). Genome-wide association study identifies 74 loci associated with educational attainment. Nature *533*, 539–542.
63. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.
64. Bush, W.S., Sawcer, S.J., de Jager, P.L., Oksenberg, J.R., McCauley, J.L., Pericak-Vance, M.A., Haines, J.L.; and International Multiple Sclerosis Genetics Consortium (IMSGC) (2010). Evidence for polygenic susceptibility to multiple sclerosis–the shape of things to come. Am. J. Hum. Genet. *86*, 621–625.
65. Stahl, E.A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B.F., Kraft, P., Chen, R., Kallberg, H.J., Kurreeman, F.A., et al.; Diabetes Genetics Replication and Meta-analysis Consortium; and Myocardial Infarction Genetics Consortium (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat. Genet. *44*, 483–489.
66. Maier, R., Moser, G., Chen, G.B., Ripke, S., Coryell, W., Potash, J.B., Scheftner, W.A., Shi, J., Weissman, M.M., Hultman, C.M., et al.; Cross-Disorder Working Group of the Psychiatric Genomics Consortium (2015). Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am. J. Hum. Genet. *96*, 283–294.
67. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.R., Bhatia, G., Do, R., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. *97*, 576–592.
68. Chen, H., Hey, J., and Slatkin, M. (2015). A hidden Markov model for investigating recent positive selection through haplotype structure. Theor. Popul. Biol. *99*, 18–30.
69. Mao, X., Bigham, A.W., Mei, R., Gutierrez, G., Weiss, K.M., Brutsaert, T.D., Leon-Velarde, F., Moore, L.G., Vargas, E., McKeigue, P.M., et al. (2007). A genomewide admixture mapping panel for Hispanic/Latino populations. Am. J. Hum. Genet. *80*, 1171–1178.
70. O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I., et al. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. *10*, e1004234.
71. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.
72. Gravel, S. (2012). Population genetics models of local ancestry. Genetics *191*, 607–619.
73. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A.; and 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
74. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
75. Kelleher, J., Etheridge, A.M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput. Biol. *12*, e1004842.
76. Mathias, R.A., Taub, M.A., Gignoux, C.R., Fu, W., Musharoff, S., O'Connor, T.D., Vergara, C., Torgerson, D.G., Pino-Yanes, M., Shringarpure, S.S., et al.; CAAPA (2016). A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. Nat. Commun. *7*, 12522.
77. Shringarpure, S.S., Bustamante, C.D., Lange, K.L., and Alexander, D.H. (2016). Efficient analysis of large datasets and sex bias with ADMIXTURE. bioRxiv. http://dx.doi.org/10.1101/039347.
78. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
79. Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., et al. (2012). Fast and accurate inference of local ancestry in Latino populations. Bioinformatics *28*, 1359–1367.
80. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O., et al. (2009). The genetic structure and history of Africans and African Americans. Science *324*, 1035–1044.
81. Zakharia, F., Basu, A., Absher, D., Assimes, T.L., Go, A.S., Hlatky, M.A., Iribarren, C., Knowles, J.W., Li, J., Narasimhan, B., et al. (2009). Characterizing the admixed African ancestry of African Americans. Genome Biol. *10*, R141.
82. Schroeder, H., Ávila-Arcos, M.C., Malaspinas, A.S., Poznik, G.D., Sandoval-Velasco, M., Carpenter, M.L., Moreno-Mayar, J.V., Sikora, M., Johnson, P.L., Allentoft, M.E., et al. (2015).

Genome-wide ancestry of 17th-century enslaved Africans from the Caribbean. Proc. Natl. Acad. Sci. USA *112*, 3669–3673.

83. Gravel, S., Zakharia, F., Moreno-Estrada, A., Byrnes, J.K., Muzzio, M., Rodriguez-Flores, J.L., Kenny, E.E., Gignoux, C.R., Maples, B.K., Guiblet, W., et al.; 1000 Genomes Project (2013). Reconstructing Native American migrations from whole-genome and whole-exome data. PLoS Genet. *9*, e1004023.
84. Kessler, M.D., Yerges-Armstrong, L., Taub, M.A., Shetty, A.C., Maloney, K., Jeng, L.J.B., Ruczinski, I., Levin, A.M., Williams, L.K., Beaty, T.H., et al.; Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) (2016). Challenges and disparities in the application of personalized genomic medicine to populations with African ancestry. Nat. Commun. *7*, 12521.
85. Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., Ferreira, T., Locke, A.E., Mägi, R., Strawbridge, R.J., Pers, T.H., Fischer, K., Justice, A.E., et al.; ADIPOGen Consortium; CARDIOGRAMplusC4D Consortium; CKDGen Consortium; GEFOS Consortium; GENIE Consortium; GLGC; ICBP; International Endogene Consortium; LifeLines Cohort Study; MAGIC Investigators; MuTHER Consortium; PAGE Consortium; and ReproGen Consortium (2015). New genetic loci link adipose and insulin biology to body fat distribution. Nature *518*, 187–196.
86. Gaulton, K.J., Ferreira, T., Lee, Y., Raimondo, A., Mägi, R., Reschen, M.E., Mahajan, A., Locke, A., Rayner, N.W., Robertson, N., et al.; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2015). Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. Nat. Genet. *47*, 1415–1425.
87. Mahajan, A., Go, M.J., Zhang, W., Below, J.E., Gaulton, K.J., Ferreira, T., Horikoshi, M., Johnson, A.D., Ng, M.C., Prokopenko, I., et al.; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium; Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; Mexican American Type 2 Diabetes (MAT2D) Consortium; and Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples (T2D-GENES) Consortium (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat. Genet. *46*, 234–244.
88. Moffatt, M.F., Gut, I.G., Demenais, F., Strachan, D.P., Bouzigon, E., Heath, S., von Mutius, E., Farrall, M., Lathrop, M., Cookson, W.O.; and GABRIEL Consortium (2010). A large-scale, consortium-based genomewide association study of asthma. N. Engl. J. Med. *363*, 1211–1221.
89. N'Diaye, A., Chen, G.K., Palmer, C.D., Ge, B., Tayo, B., Mathias, R.A., Ding, J., Nalls, M.A., Adeyemo, A., Adoue, V., et al. (2011). Identification, replication, and fine-mapping of loci associated with adult height in individuals of african ancestry. PLoS Genet. *7*, e1002298.
90. Gustafsson, A., and Lindenfors, P. (2004). Human size evolution: no evolutionary allometric relationship between male and female stature. J. Hum. Evol. *47*, 253–266.
91. Whiteford, H.A., Degenhardt, L., Rehm, J., Baxter, A.J., Ferrari, A.J., Erskine, H.E., Charlson, F.J., Norman, R.E., Flaxman, A.D., Johns, N., et al. (2013). Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. Lancet *382*, 1575–1586.
92. de Candia, T.R., Lee, S.H., Yang, J., Browning, B.L., Gejman, P.V., Levinson, D.F., Mowry, B.J., Hewitt, J.K., Goddard, M.E., O'Donovan, M.C., et al.; International Schizophrenia Consortium; and Molecular Genetics of Schizophrenia Collaboration (2013). Additive genetic variation in schizophrenia risk is shared by populations of African and European descent. Am. J. Hum. Genet. *93*, 463–470.
93. Chan, Y., Lim, E.T., Sandholm, N., Wang, S.R., McKnight, A.J., Ripke, S., Daly, M.J., Neale, B.M., Salem, R.M., Hirschhorn, J.N.; DIAGRAM Consortium; GENIE Consortium; GIANT Consortium; IIBDGC Consortium; and PGC Consortium (2014). An excess of risk-increasing low-frequency variants can be a signal of polygenic inheritance in complex diseases. Am. J. Hum. Genet. *94*, 437–452.
94. Minikel, E.V., Vallabh, S.M., Lek, M., Estrada, K., Samocha, K.E., Sathirapongsasuti, J.F., McLean, C.Y., Tung, J.Y., Yu, L.P., Gambetti, P., et al.; Exome Aggregation Consortium (ExAC) (2016). Quantifying prion disease penetrance using large population control cohorts. Sci. Transl. Med. *8*, 322ra9.
95. Li, Y.R., and Keating, B.J. (2014). Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Med. *6*, 91.
96. Rosenberg, N.A., Huang, L., Jewett, E.M., Szpiech, Z.A., Jankovic, I., and Boehnke, M. (2010). Genome-wide association studies in diverse populations. Nat. Rev. Genet. *11*, 356–366.

With Cell Press Webinars, our network is your network!

Cell Press Webinars give you access to hot topics in emerging research and the application of new technology.

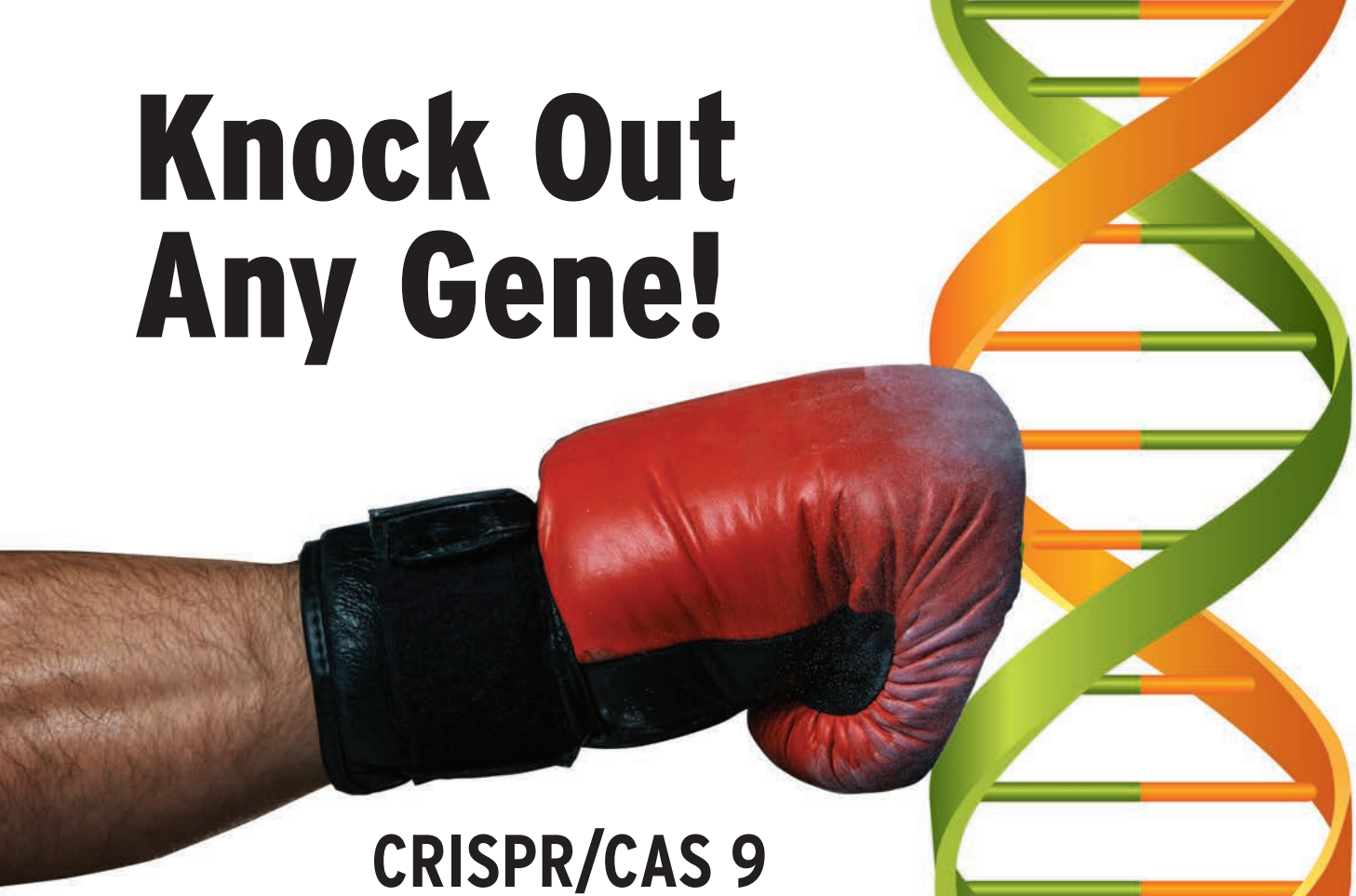
Watch essential, need-to-know webcasts via live streaming or on demand from the comfort and convenience of your office, lab, or home.

- Need-to-know topics, editorially curated
- World-class presenters, experts in their field
- Moderated by Cell Press editors

Tap into our network today!
Visit www.cell.com/webinars

CellPress
Webinars

Knock Out Any Gene!



CRISPR/CAS 9 Genome Editing Kits

OriGene's system for genome disruption and gene replacement delivers pre-designed plasmids and all the vectors needed to knock out any human or mouse gene. Knockout is as simple as 1-2-3:

1. Search the gene symbol on origene.com and order
2. Follow the simple protocol for transfection and Puro selection
3. Validate the knockout

Scan to view CRISPR/Cas-9 Video

Kit Components Include:

- 2 guide RNA vectors to ensure efficient cleavage
- Donor vector with predesigned homologous arms
- Knockin GFP-Puro for selection
- Scramble gRNA as negative control included

Come to OriGene, the trusted molecular biology expert, for your CRISPR needs.

ORIGENE

www.origene.com/CRISPR-CAS9

Enhancing human research since 1999

Whole Genome/Exome Sequencing

Targeted Sequencing

Single-Cell RNA-Seq

SNP Genotyping

Synthetic Libraries

CLIA-Compliant Sanger Sequencing

genewiz.com

NEW! Amplicon-EZ

Delivers fast, cost-effective, ultra-deep sequencing of PCR products/amplicons via next generation sequencing.

Experience the GENEWIZ Difference

- Superior data quality
- Starting at just $50 a sample
- Complete sample-to-answer workflows
- Interactive reports
- Results in as few as three days!

Visit web.genewiz.com/amplicon-ez

GENEWIZ®
Solid science. Superior service.

CONTACT GENEWIZ:
Ph.D. experts available to discuss your project:

+1(908) 222-0711 ext. 1
NGS@genewiz.com
genewiz.com

appliedbiosystems

Nothing has changed, **except everything**

Sanger sequencing and fragment analysis like you have never experienced before. Same workflow, same trusted technology, now with an innovative all-in-one cartridge that takes setup time from hours to minutes. Introducing the **Applied Biosystems™ SeqStudio™ Genetic Analyzer**.

Find out more at **thermofisher.com/seqstudio**

ThermoFisher SCIENTIFIC