THE FOUNDATION OF PRECISION MEDICINE: INTEGRATION OF ELECTRONIC HEALTH RECORDS WITH GENOMICS THROUGH BASIC, CLINICAL, AND TRANSLATIONAL RESEARCH

EDITED BY: Mariza de Andrade, Helena Kuivaniemi and Marylyn D. Ritchie

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88919-872-6
DOI 10.3389/978-2-88919-872-6

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

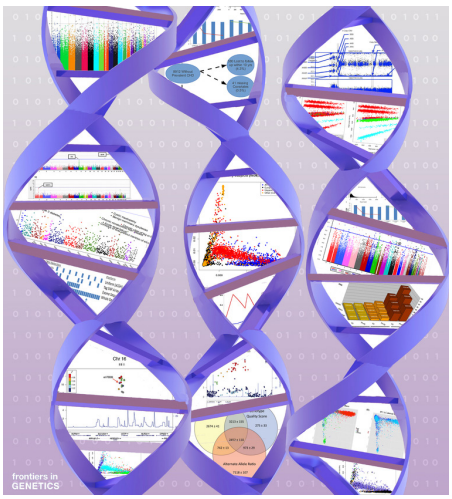
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

THE FOUNDATION OF PRECISION MEDICINE: INTEGRATION OF ELECTRONIC HEALTH RECORDS WITH GENOMICS THROUGH BASIC, CLINICAL, AND TRANSLATIONAL RESEARCH

Topic Editors:
Mariza de Andrade, Mayo Clinic, USA
Helena Kuivaniemi, Stellenbosch University, South Africa
Marylyn D. Ritchie, The Pennsylvania State University and Institute of Biomedical and Translational Informatics, Geisinger Health System, USA



Examples of the diversity of analyses that can be performed when linking electronic health records with genomics data. Each image in the DNA helices represents an image in an article part of this eBook.
Figure by Adam Hardebeck

This eBook contains the 19 articles that were part of a Special Topic in Frontiers in Genetics entitled "Genetics Research in Electronic Health Records Linked to DNA Biobanks". The Special Issue was published on-line in 2014-2015 and contained papers representing the diverse research ongoing in the integration of electronic health records (EHR) with genomics through basic, clinical, and translational research.

We have divided the eBook into four Chapters. Chapter 1 describes the Electronic Medical Records and Genomics (eMERGE) network and its contribution to genomics. It highlights methodological questions related to large data sets such as imputation and population stratification. Chapter 2 describes the results of genetic studies on different diseases for which all the phenotypic information was extracted from the EHR with highly specific ePhenotyping algorithms. Chapter 3 focuses on more complex analyses of the genome including copy number variants (CNV), pleiotropy combined with phenome-wide association studies (PheWAS), and epistasis (gene-gene interactions). Chapter 4 discusses the use of genetic data together with EHR-derived clinical data in clinical settings, and how to return genetic results to patients and providers. It also contains a comprehensive review on genetic risk scores. We have

included mostly Original Research Articles in the eBook, but also Reviews and Methods papers on the relevant topics of analyzing and integrating genomic data.

The release of this eBook is timely, since several countries are launching Precision Medicine initiatives. Precision Medicine is a new concept in patient care taking into account individual variability in genetic, environmental and lifestyle factors, when treating diseases or trying to prevent them from developing. It has become an important focus for biomedical, clinical and translational informatics. The papers presented in this eBook are well positioned to educate the readers about Precision Medicine and to demonstrate the potential study designs, methods, strategies, and applications where this type of research can be performed successfully. The ultimate goal is to improve diagnostics and provide better, more targeted care to the patient.

We would like to thank the Editorial Staff of the Frontiers in Genetics for their patience and support and for making this eBook possible.

Mariza de Andrade, PhD
Helena Kuivaniemi, MD, PhD
Marylyn D. Ritchie, PhD

Citation: de Andrade, M., Kuivaniemi, H., Ritchie, M. D., eds. (2016). The Foundation of Precision Medicine: Integration of Electronic Health Records with Genomics Through Basic, Clinical, and Translational Research. Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-872-6

Table of Contents

CHAPTER 4: Use of genetic data together with EHR-derived clinical data in clinical settings, and how to return genetic results to patients and providers

frontiers in Genetics

EDITORIAL
published: 17 March 2015
doi: 10.3389/fgene.2015.00104

The foundation of precision medicine: integration of electronic health records with genomics through basic, clinical, and translational research

Marylyn D. Ritchie[1,2]*, *Mariza de Andrade*[3] and *Helena Kuivaniemi*[4,5]

[1] Biochemistry and Molecular Biology, Center for Systems Genomics, The Pennsylvania State University, University Park, PA, USA, [2] Institute of Biomedical and Translational Informatics, Geisinger Health System, Danville, PA, USA, [3] Division of Biomedical Statistics and Informatics, Department of Health Science Research, Mayo Clinic, Rochester, MN, USA, [4] The Sigfried and Janet Weis Center for Research, Geisinger Health System, Danville, PA, USA, [5] Department of Surgery, Temple University School of Medicine, Philadelphia, PA, USA

Keywords: electronic health records, precision medicine, genomic medicine, EHR, genomics

The members of the Genomics Workgroup in the Electronic Medical Records and Genomics (eMERGE) network (Gottesman et al., 2013) led the development of a Special Topic in Frontiers in Genetics titled "Genetics Research in Electronic Health Records Linked to DNA Biobanks[1]." The goal was to publish papers representing the diverse research ongoing in the integration of electronic health records (EHR) with genomics through basic, clinical, and translational research. The special topic with its 18 papers is extremely timely given the recent announcement of the Precision Medicine initiative by the White House[2], which includes the potential to build a biobank of 1 million Americans with rich, phenotypic data—likely from EHR. eMERGE has, therefore, served as an excellent test case for how a 1 million person project might work across several medical centers, EHR systems, and genetic datasets.

The first group of papers (Almoguera et al., 2014; Crawford et al., 2014; Crosslin et al., 2014; Verma et al., 2014) belonging to this special issue presents the eMERGE network and its contribution to genomics. The paper by Crawford et al. (2014) describes the initial goal of eMERGE network that was to explore the utility of EHRs in genomics and whether the phenotypes identified through algorithms using EHRs combined with the genome-wide genotypes could lead to fruitful results. The beginning of the network included individual genotype datasets that were later combined to form the merged eMERGE datasets and the combination with phenotypes from EHRs has led to new genomic discoveries. All of these steps subsequently lead to new goals that have included next generation sequencing and clinical practice. The second paper (Verma et al., 2014) introduces the new challenges involved in merging genotype data from different eMERGE sites. Since genotypes at different sites were derived from different genotyping platforms it was impossible to create a single merged data file based on raw genotype data alone. The solution was first to impute each site separately using the same software and pipeline, and then merge the imputed genotype data sets to form a combined dataset. The authors used two different imputation software packages and describe the challenges involved in using diverse ethnic populations and different genotype platforms, which lead to a complete pipeline that not only performs imputation but also ensures appropriate quality control for merging genotype data sets. The final eMERGE imputed data set is

[1] Available online at: http://journal.frontiersin.org/ResearchTopic/2198
[2] Available online at: WH.GOV/PRECISION-MEDICINE

OPEN ACCESS

Edited and reviewed by:
Anthony Gean Comuzzie, Texas Biomedical Research Institute, USA

***Correspondence:**
Marylyn D. Ritchie, mdr23@psu.edu

Specialty section:
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics

Received: *17 February 2015*
Accepted: *27 February 2015*
Published: *17 March 2015*

Citation:
Ritchie MD, de Andrade M and Kuivaniemi H (2015) The foundation of precision medicine: integration of electronic health records with genomics through basic, clinical, and translational research. Front. Genet. 6:104. doi: 10.3389/fgene.2015.00104

a valuable resource for genomic discovery by using the clinical data generated by the EHRs and will be available in dbGaP soon. The third paper (Crosslin et al., 2014) discusses the issues of population stratification and genotype platform bias. Principal components analysis (PCA) is commonly used to control for population stratification; however other factors such as local genomic variation, multiple study sites and multiple genotyping platforms may also increase the correlation patterns in the PCA. In this paper Crosslin et al. (2014) provided an alternative approach to PCA by deriving components from subject loadings determined by the 1000 Genomes reference sample that avoid the bias introduced by site and genotype platform effects. This alternative approach was applied successfully in the eMERGE genome-wide association study (GWAS) for venous thromboembolism in African Americans. The fourth paper in this group by Almoguera et al. (2014) evaluated the utility of large imputed genotype data sets to identify subjects with *TPMT* defective alleles. They used around 87,000 samples from the biobank at the Children's Hospital of Philadelphia. For 12 samples also Sanger sequencing data were available allowing comparison between the imputed and observed genotypes. The concordance rate between the non-carriers of the risk alleles was 98.88%; however the sensitivity of imputation for homozygous carriers was ∼80%. The authors recommend using imputation of *TMPT* alleles as a first step to screen individuals at risk.

The papers of group 2 (Kullo et al., 2014a; Mitchell et al., 2014; Namjou et al., 2013; Parihar et al., 2014; Ye et al., 2014) describe different applications of the EHR derived phenotypes. The first paper (Namjou et al., 2013) investigated whether the common variants in the genes *FTO*, *MC4R* and *TMEM18* associated with BMI in adults are also associated in pediatric population in the eMERGE network. First they used a linear regression model with the dependent variable BMI, adjusted for age, sex, and PC by cohort; and then meta-analyzed the results using a weighted z-score approach. They not only reproduced the findings for the pediatric cohorts but also identified a novel locus at *COL6A5*. The second paper (Mitchell et al., 2014) described the issues when using cases generated from Stroke Genetics Network (SiGN) and using genotyped controls from eMERGE leading to recommendations regarding the controls selection, population stratification, imputation, and association analysis. The third paper by Kullo et al. (2014a) performed a two-stage association study to identify variants associated with peripheral arterial disease. The first stage was a GWAS adjusted for age and sex in subjects of European ancestry. In the second stage the top 48 SNPs were replicated in new set of cases and controls. One single nucleotide polymorphism (SNP) in the *ATXN2-SH2B3* gene was significant where this SNP is in high LD with a missense variant in *SH2B3*, a gene that is related to immune and inflammatory response pathways and vascular homeostasis, indicating a pleiotropic effect. The fourth paper (Parihar et al., 2014) carried out a GWAS for lipid-related phenotypes derived from the EHR using the Metabochip array. These phenotypes consist of laboratory, anthropomorphic and demographic data on a cohort of extremely obese subjects. They replicated 12 of 21 previously identified lipid-associated SNPs demonstrating the validity of using phenotype data available from the EHR and the usefulness of the Metabochip array. The fifth paper (Ye et al., 2014) performed GWAS to identify genetic variants associated with diseases caused by *Staphylococcus aureus* infection. They used different approaches to identify the genetic susceptibility from single SNP, gene set and pathway. No SNPs or genes were found to be genome-wide significant leaving with the speculation that multiple genes contribute to the severity of the infection.

The third group of papers (Connolly et al., 2014; Cronin et al., 2014; Namjou et al., 2014; Patel et al., 2014; Sun et al., 2014) in this special issue focused on more complex analyses of the genome including copy number variants (CNV), pleiotropy combined with phenome-wide association studies (PheWAS), and epistasis (gene-gene interactions). The first paper by Namjou et al. (2014) describes the first PheWAS in a pediatric cohort based on 4268 samples and 2476 sSNPs selected from previously published GWAS studies. A total of 539 EMR-derived phenotypes were explored. The authors identified a number of known associations which serve as a positive control as well as several novel associations including *NDFIP1* associated with mental retardation and *PLCL1* associated with developmental delays and speech disorder. The second paper by Cronin et al. (2014) is another PheWAS, focused on one specific gene, *FTO*, in 10,487 individuals from the eMERGE network and another 13,711 individuals from the Vanderbilt biobank BioVU. They identified highly significant associations between *FTO* and obesity, type II diabetes, and sleep apnea, all of which are expected for variants in this gene. A novel association was identified between *FTO* and fibrocystic breast disease. The third paper by Sun et al. (2014) is a review of methods to filter genome-wide SNP data to explore epistasis models effectively. There are a number of challenges with the search for epistasis in genome-wide data including the computational complexity of exploring that many different combinations of variables which can exceed computational feasibility as well as the magnitude of the multiple testing incurred by testing the genome in exhaustive interaction analyses. The authors discuss two different filtering approaches, namely using statistical effects or biological prior knowledge. Strengths and weakness of these different strategies are described as well as additional resources for consideration before a genome-wide epistasis analysis is initiated. The fourth paper by Connolly et al. (2014) is a review on recent research in the area of CNV including successful applications in rare and common diseases. Methods for identifying CNVs from array-based genotyping data and sequencing data are described. Finally, how CNVs might be evaluated and used with medical records is discussed. The fifth paper of this group is by Patel et al. (2014) and describes quality control processes for whole exome sequencing data, specifically using Mendelian errors as a filtering strategy to minimize errors. The group developed the Cincinnati Analytical Suite for Sequencing Informatics (CASSI) to store sequencing files, metadata, and others. Their data cleaning process can be used to improve the signal-to-noise ratio and improve the identification of candidate disease causative variants.

The fourth group of papers (Goldstein et al., 2014; Kullo et al., 2014b; Schrodi et al., 2014; Sleiman et al., 2014) belonging to this special issue discusses the use of genetic data together with EHR-derived clinical data in clinical settings. The first one of these papers (Sleiman et al., 2014) used imputed GWAS data to study

two loss-of-function variants in the *PCSK9* gene. The study of 8028 genotyped biobank participants with extensive laboratory data from the EHR demonstrated that EHR-linked biobanks are a rich resource for exploring functional aspects of genetic variants. The second paper (Schrodi et al., 2014) is a review article about genetic-based prediction by Schrodi et al. (2014) and it provides a comprehensive discussion about disease prediction using both genetic and clinical data, again highlighting the usefulness of available EHR-linked genetic data on large cohorts. As the title of their article reveals, predicting who is at risk for a given disease has turned out to be a difficult task. Currently the most promising results can be found in cancer genomics, population screening of rare Mendelian diseases, and pharmacogenetics. Developing prediction models for common complex diseases such as type 2 diabetes mellitus, stroke and inflammatory arthritis has been more challenging and the results have been disappointing. This was also evident in the third paper of this group (Goldstein et al., 2014) in which coronary heart disease was investigated in the NIH-funded Atherosclerosis Risk in Communities (ARIC) cohort. The authors combined a genetic risk score derived from 45 SNPs with a clinical risk score, but received only minimal improvement in discrimination and calibration statistics of the risk score. Schrodi et al. (2014) conclude their review article with a positive note pointing out that in the near future we can rely on having access to additional genome-wide data which might help in refining the risk prediction. These data will include whole genome and whole exome sequence data, and other omics data such as information on DNA methylation, histone modification, and the transcriptomes of different tissues. Additional advances leading to more refined phenotyping, and development of new, more robust computational approaches will contribute to improved accuracy in risk estimates. The last paper in the fourth group (Kullo et al., 2014b) deals with the key questions about returning results to patients and providers. The authors are from the eMERGE network and point out that one of the mandates of the network is to come up with the best practices for implementing genomic medicine. The goal is to have the clinically relevant genetic results in the EHR so that they are easily available for the practicing physician to be used at point-of-care. These results could be individual risk genotypes or combined risk scores. Each of the eMERGE network sites is carrying out a feasibility projects, e.g., the group at Icahn School of Medicine at Mount Sinai is using *APOL1* variants in African Americans to predict chronic kidney disease and investigators at Vanderbilt University have chosen 14 actionable pharmacogenetic variants to be returned to the EHR.

Precision medicine (See Footnote 2) is an important focus for biomedical, clinical and translational informatics in the current era. The manuscripts presented in this special topic are well positioned to educate and demonstrate the potential study designs, methods, strategies, and applications where this type of research can be performed successfully. The ultimate goal is to improve diagnostics and provide better, more targeted care to the patient.

Acknowledgments

The eMERGE Network was initiated and funded by NHGRI, in conjunction with additional funding from NIGMS through the following grants: U01HG004438 (Johns Hopkins University); U01HG004424 (The Broad Institute); U01HG004438 (CIDR); U01HG004610 and U01HG006375 (Group Health Cooperative/University of Washington); U01HG004608 (Marshfield Clinic); U01HG006389 (Essentia Institute of Rural Health); U01HG04599 and U01HG006379 (Mayo Clinic); U01HG004609 and U01HG006388 (Northwestern University); U01HG04603 and U01HG006378 (Vanderbilt University); U01HG006385 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG006382 (Geisinger Clinic); U01HG006380 (Icahn School of Medicine at Mount Sinai); U01HG006830 (Children's Hospital of Philadelphia); and U01HG006828 (Cincinnati Children's Hospital/Boston Children's Hospital).

References

Almoguera, B., Vazquez, L., Connolly, J. J., Bradfield, J., Sleiman, P., Keating, B., et al. (2014). Imputation of TPMT defective alleles for the identification of patients with high-risk phenotypes. *Front. Genet*. 5:96. doi: 10.3389/fgene.2014.00096

Connolly, J. J., Glessner, J. T., Almoguera, B., Crosslin, D. R., Jarvik, G. P., Sleiman, P. M., et al. (2014). Copy number variation analysis in the context of electronic medical records and large-scale genomics consortium efforts. *Front. Genet*. 5:51. doi: 10.3389/fgene.2014.00051

Crawford, D. C., Crosslin, D. R., Tromp, G., Kullo, I. J., Kuivaniemi, H., Hayes, M. G., et al. (2014). eMERGEing progress in genomics—the first seven years. *Front. Genet*. 5:184. doi: 10.3389/fgene.2014.00184

Cronin, R. M., Field, J. R., Bradford, Y., Shaffer, C. M., Carroll, R. J., Mosley, J. D., et al. (2014). Phenome Wide Association Studies demonstrating pleiotropy of genetic variants within FTO with and without adjustment for body mass index. *Front. Genet*. 5:250. doi: 10.3389/fgene.2014.00250

Crosslin, D. R., Tromp, G., Burt, A., Kim, D. S., Verma, S. S., Lucas, A. M., et al. (2014). Controlling for population structure and genotyping platform bias in the eMERGE multi-institutional biobank linked to Electronic Health Records. *Front. Genet*. 5:352. doi: 10.3389/fgene.2014.00352

Goldstein, B. A., Knowles, J. W., Salfati, E., Ioannidis, J. P., and Assimes, T. L. (2014). Simple, standardized incorporation of genetic risk into non-genetic risk prediction tools for complex traits: coronary heart disease as an example. *Front. Genet*. 5:254. doi: 10.3389/fgene.2014.00254

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., et al. (2013). The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* 15, 761–771. doi: 10.1038/gim.2013.72

Kullo, I. J., Haddad, R. A., Prows, C. A., Holm, I., Sanderson, S. C., Garrison, N. A., et al. (2014b). Return of Genomic results in the genomic medicine projects of the eMERGE network. *Front. Genet*. 5:50. doi: 10.3389/fgene.2014.00050

Kullo, I., Shameer, K., Jouni, H., Lesnick, T. G., Pathak, J., Chute, C. G., et al. (2014a). The ATXN2-SH2B3 locus is associated with peripheral arterial disease: an electronic medical record-based genome-wide association study. *Front. Genet*. 5:166. doi: 10.3389/fgene.2014.00166

Mitchell, B. D., Fornage, M., McArdle, P. F., Cheng, Y.-C., Pulit, S., Wong, Q., et al. (2014). Using previously genotyped controls in genome-wide association studies (GWAS): application to the Stroke Genetics Network (SiGN). *Front. Genet*. 5:95. doi: 10.3389/fgene.2014.00095

Namjou, B., Keddache, M., Marsolo, K., Wagner, M., Lingren, T., Cobb, B., et al. (2013). EMR-linked GWAS study: Investigation of variation

landscape of loci for Body Mass Index in children. *Front. Genet*. 4:268. doi: 10.3389/fgene.2013.00268

Namjou, B., Marsolo, K., Carroll, R., Denny, J., Ritchie, M. D., Setia, S., et al. (2014). Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts. *Front. Genet*. 5:401. doi: 10.3389/fgene.2014.00401

Parihar, A., Wood, G. C., Chu, X., Jin, Q., Argyropoulos, G., Still, C. D., et al. (2014). Extension of GWAS results for lipid-related phenotypes to extreme obesity using electronic health record (EHR) data and the Metabochip. *Front. Genet*. 5:222. doi: 10.3389/fgene.2014.00222

Patel, Z. H., Kottyan, L. C., Lazaro, S., Williams, M. S., Ledbetter, D. H., Tromp, G., et al. (2014). The struggle to find reliable results in exome sequencing data: Filtering out Mendelian errors. *Front. Genet*. 5:16. doi: 10.3389/fgene.2014.00016

Schrodi, S. J., Mukherjee, S., Shan, Y., Tromp, G., Sninsky, J. J., Callear, A. P., et al. (2014). Genetic-based prediction of disease traits: prediction is very difficult, especially about the future. *Front. Genet*. 5:162. doi: 10.3389/fgene.2014.00162

Sleiman, P., Bradfield, J., Mentch, F., Almoguera, B., Connolly, J., and Hakonarson, H. (2014). Assessing the functional consequence of loss of function variants using electronic medical record and large-scale genomics consortium efforts. *Front. Genet*. 5:105. doi: 10.3389/fgene.2014.00105

Sun, X., Lu, Q., Mukheerjee, S., Crane, P., Elston, R. C., and Ritchie, M. D. (2014). Analysis pipeline for the epistasis search – statistical versus biological filtering. *Front. Genet*. 5:106. doi: 10.3389/fgene.2014.00106

Verma, S. S., de Andrade, M., Tromp, G., Kuivaniemi, H., Pugh, E., Namjou, B., et al. (2014). Imputation and quality control steps for combining multiple genome-wide datasets. *Front. Genet*. 5:370. doi: 10.3389/fgene.2014.00370

Ye, Z., Vasco, D. A., Carter, T., Brilliant, M., Schrodi, S. J., and Shukla, S. K. (2014). Genome wide association study of SNP-, gene-, and pathway-based approaches to identify genes influencing susceptibility to *Staphylococcus aureus* infections. *Front. Genet*. 5:125. doi: 10.3389/fgene.2014.00125

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in GENETICS

REVIEW ARTICLE
published: 17 June 2014
doi: 10.3389/fgene.2014.00184

eMERGEing progress in genomics—the first seven years

Dana C. Crawford[1,2]*, David R. Crosslin[3,4], Gerard Tromp[5], Iftikhar J. Kullo[6], Helena Kuivaniemi[5], M. Geoffrey Hayes[7], Joshua C. Denny[8,9], William S. Bush[1,8], Jonathan L. Haines[10,11], Dan M. Roden[9,12], Catherine A. McCarty[13], Gail P. Jarvik[3,4] and Marylyn D. Ritchie[14,15]

[1] Center for Human Genetics Research, Vanderbilt University, Nashville, TN, USA
[2] Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN, USA
[3] Medical Genetics, Department of Medicine, School of Medicine, University of Washington, Seattle, WA, USA
[4] Department of Genome Sciences, University of Washington, Seattle, WA, USA
[5] The Sigfried and Janet Weis Center for Research, Geisinger Health System, Danville, PA, USA
[6] Division of Cardiovascular Diseases and the Gonda Vascular Center, Mayo Clinic, Rochester, MN, USA
[7] Division of Endocrinology, Metabolism, and Molecular Medicine, Department of Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA
[8] Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA
[9] Department of Medicine, Vanderbilt University, Nashville, TN, USA
[10] Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA
[11] Institute for Computational Biology, Case Western Reserve University, Cleveland, OH, USA
[12] Department of Pharmacology, Vanderbilt University, Nashville, TN, USA
[13] Essentia Institute of Rural Health, Duluth, MN, USA
[14] Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA
[15] Center for Systems Genomics, Pennsylvania State University, University Park, PA, USA

Edited by:
Mariza De Andrade, Mayo Clinic, USA

Reviewed by:
Alexis C. Frazier-Wood, University of Alabama at Birmingham, USA
Yiran Guo, Children's Hospital of Philadelphia, USA

***Correspondence:**
Dana C. Crawford, Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall, Nashville, TN 37232-0700, USA
e-mail: crawford@chgr.mc.vanderbilt.edu

The electronic MEdical Records & GEnomics (eMERGE) network was established in 2007 by the National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) in part to explore the utility of electronic medical records (EMRs) in genome science. The initial focus was on discovery primarily using the genome-wide association paradigm, but more recently, the network has begun evaluating mechanisms to implement new genomic information coupled to clinical decision support into EMRs. Herein, we describe this evolution including the development of the individual and merged eMERGE genomic datasets, the contribution the network has made toward genomic discovery and human health, and the steps taken toward the next generation genotype-phenotype association studies and clinical implementation.

Keywords: biobanks, genome-wide association studies, pharmacogenomics, electronic medical records

INTRODUCTION

Revolutions in genotyping technology (Ragoussis, 2009) and computational power coupled with the creation of public scientific resources such as The Human Genome Project (2001; Venter et al., 2001), The International HapMap Project (2003; The International HapMap Consortium 2005), and most recently the 1000 Genomes Project (2012), have accelerated genomic discovery, most commonly through genome-wide association studies (GWAS). As of late March 2014, the National Human Genome Research Institute (NHGRI) GWAS catalog listed 1201 publications with 3961 SNPs associated with approximately 571 human diseases and traits at a significance threshold of 5.0×10^{-8} (Welter et al., 2014) (https://www.genome.gov/26525384)

The majority of genomic discoveries published to date have been from case-control or cohort epidemiologic studies that collected specific health-related data and DNA samples. These traditional epidemiologic collections already exist and are primed for genomic discovery studies (Willett et al., 2007), making them ideal for large-scale GWAS. Also, although currently under-utilized in genomic discovery, many of the cohorts have collected exposure data that can be interrogated for gene-environment interaction studies (Manolio et al., 2006; Thomas, 2010). However, a major disadvantage of accessing existing epidemiologic cohorts for genomic discoveries is limited representation of diverse racial/ethnic groups (Rosenberg et al., 2010) and of children (Collins and Manolio, 2007). Also, the existing health-related data can be limiting, especially for cohorts or case-controls collections designed with very specific disease outcomes for study such as cancers or cardiovascular disease. Finally, establishing and maintaining an on-going cohort study can pose significant cost burden (Rukovets, 2013).

The disadvantages of accessing existing case-control and cohort studies coupled with the continued need for genotype-phenotype data for genomic discoveries led to the consideration of alternative study designs and data sources such as biorepositories linked to electronic medical records (EMRs). In addition for the potential for large sample sizes of diverse groups, biobanks linked to EMRs make possible the study of many different outcomes and traits, many of which may not be routinely collected by traditional epidemiologic cohorts. And, in this burgeoning era of

precision or personalized medicine, biobanks in clinical settings offer unprecedented opportunities to quickly translate research findings to improvements in patient care.

In recognition of the potential for EMR-linked biobanks to genomic discovery and personalized medicine, NHGRI established the electronic MEdical Records & GEnomics (eMERGE) network. The eMERGE network began in 2007 with a Coordinating Center (Vanderbilt University) and five study sites: Group Health/University of Washington, Marshfield Clinic, Mayo Clinic, Northwestern University, and Vanderbilt University (McCarty et al., 2011). The network expanded to include new adult study sites (The Icahn School of Medicine at Mount Sinai and Geisinger Health System) in 2011 as well as pediatric study sites in 2012 (Children's Hospital of Philadelphia and Boston Children's Hospital/Cincinnati Children's Hospital Medical Center) (Gottesman et al., 2013). The major goals of eMERGE I (McCarty et al., 2011) have evolved with experience, and the major activities of the Genomics Work Group of the eMERGE II network are outlined in **Figure 1**. Here we review from the perspective of the eMERGE Genomics Work Group the contributions the network has made toward genomic discovery since 2007. We also foreshadow the eMERGE network's contributions to the second generation of genotype-phenotype associations as well as implementation of genomic medicine.

eMERGE GENOMIC RESOURCES

The first few years of the eMERGE network required data generation both at the phenotype and genotype levels (McCarty et al., 2011; Gottesman et al., 2013). In the first phase of the eMERGE network, each study site proposed an outcome or trait for phenotype algorithm development and selection of DNA samples for genotyping. Since EMR data are generated for the purposes of clinical care, a necessary step to identifying populations of interest was to create and validate algorithms that queried data elements from the EMR to find phenotypes of interest (Kho et al., 2011; Newton et al., 2013). Typically, these algorithms involved Boolean combinations of billing codes, medication exposures, laboratory, and test results, and/or natural language processing. All algorithms and their validation results in the eMERGE network are available on PheKB (www.phekb.org).

After validation of phenotype algorithms by blinded review, typically by physicians, matching case, and control samples were genotyped. All DNA samples were genotyped using either the Illumina 660-Quad (primarily for participants of European ancestry) or the Illumina 1M (primarily for participants of African ancestry) at either the Broad Institute Center for Genotyping and Analysis or the Center for Inherited Disease Research (CIDR). The eMERGE Coordinating Center established a pipeline to process each study site's data for quality control, data cleaning, and eventual Database of Genotypes and Phenotypes (dbGaP) (Mailman et al., 2007) documentation and deposition (Turner et al., 2011a). The initial round of phenotyping and genotyping resulted in the generation of GWAS-level data on 19,637 samples, of which 18,663 passed quality control metrics. The phenotypes and samples sizes available from these eMERGE phase I efforts included cataracts/HDL-C (2642 cases and 1322 controls; led by Marshfield Clinic), dementia (1241 cases and 2043 controls; led by Group Health Cooperative/University of Washington), electrocardiographic traits (3034 individuals; led by Vanderbilt University), peripheral artery disease (1641 cases and 1604; controls led by Mayo Clinic), and type 2 diabetes (2706 cases and 1496 controls; led by Northwestern University).

During phase I of the eMERGE network, high-density genotyping had matured such that many large cohorts and biorepositories linked to EMRs had existing GWAS-level data. This included expanded genotype datasets at some eMERGE I sites and as such, no new high density genome-wide genotyping was performed in eMERGE phase II. All existing and new study sites in eMERGE II offered existing data on a variety of genotyping platforms and genetic ancestries. With the inclusion of the eMERGE phase I data, a total of 60,766 (47,507 adult and 13,259 pediatric) samples with GWAS-level genotypes or other large-scale data [such as Metabochip (Voight et al., 2012)] generated by either Illumina or Affymetrix arrays are available for study in eMERGE phase II. As detailed in a separate manuscript (Verma et al., in press), pooling and merging of these data required imputation and extensive quality control. The current eMERGE phase II merged dataset (version 2) available for analysis includes 51,038 samples linked to EMRs imputed to >36 million SNPs using the 1000 Genomes Project cosmopolitan reference panel ($n = 1092$) and IMPUTE2 (Verma et al., in press).

New to eMERGE phase II is the eMERGE-PGx project, which involves the targeted sequencing of 84 pharmacogenes identified by the Pharmacogenomics Research Network (PGRN) using DNA capture and contemporary sequencing technologies (known as PGRN-Seq) (Rasmussen-Torvik et al., in press). For this effort, each eMERGE II study site is enrolling ∼1000 patients as a pilot study of pharmacogenetic sequencing in clinical practice. Enrollment and sequencing is on-going, and the anticipated network-wide sample size is 9000. All variants annotated through this effort will be available in summary data form via the eMERGE on-line resource "*S*equence, *P*henotype, and p*H*armacogenomics *IN*tegration e*X*change" or "SPHINX" (www.emergesphinx.org). The eMERGE-PGx project will help establish best practices for implementing personalized medicine including exploring and establishing guidelines for returning results to physicians and patients (Kullo et al., 2014). These data will also contribute toward the catalog of rare and less common variants and couple them to EMR data which may increase their clinical utility.

eMERGE GENOMIC DISCOVERIES

It was recognized early in the phenotype and genotype data generation phase of eMERGE I that large sample sizes are needed to have sufficient statistical power for genetic association studies. Indeed, initial GWAS of single eMERGE study site datasets demonstrated that known genotype-phenotype associations such as *SCN10A* and PR duration (Chambers et al., 2010; Holm et al., 2010; Pfeufer et al., 2010) could be replicated albeit at a significance threshold above 5.0×10^{-8} (Denny et al., 2010b). While this exercise of replication demonstrated that EMR-derived phenotypes could be used in genotype-phenotype studies, genomic discovery of new associations would require larger sample sizes.

To achieve this goal, the eMERGE network employed several strategies, including (1) pooled analysis across the network, (2)

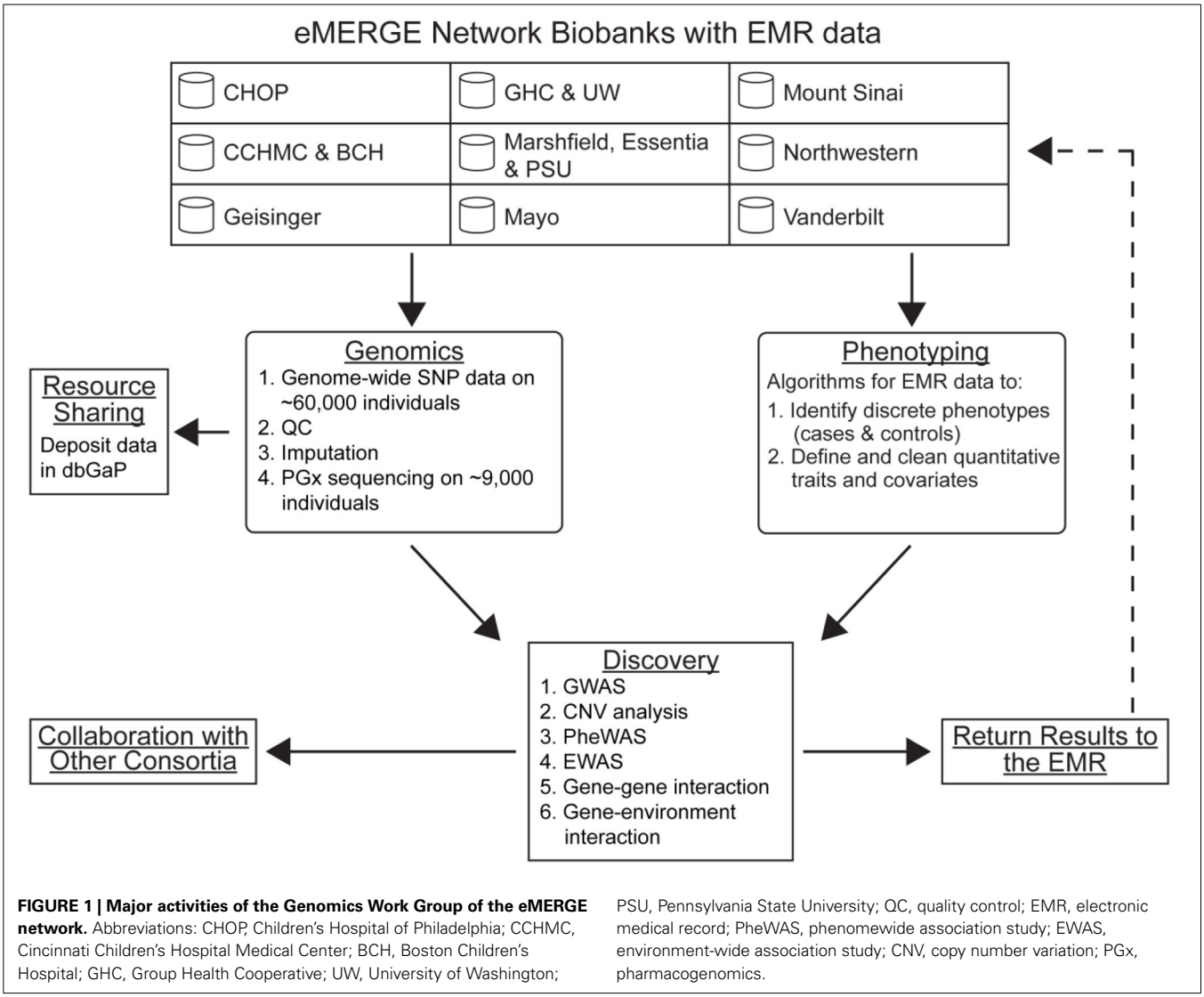


FIGURE 1 | Major activities of the Genomics Work Group of the eMERGE network. Abbreviations: CHOP, Children's Hospital of Philadelphia; CCHMC, Cincinnati Children's Hospital Medical Center; BCH, Boston Children's Hospital; GHC, Group Health Cooperative; UW, University of Washington; PSU, Pennsylvania State University; QC, quality control; EMR, electronic medical record; PheWAS, phenomewide association study; EWAS, environment-wide association study; CNV, copy number variation; PGx, pharmacogenomics.

meta-analysis within and with outside consortia, and (3) generation of new phenotype and genotype data for new studies. In the first strategy, each eMERGE study site deployed not only the phenotype used to select study subjects for the genotype-phenotype association studies of the site's primary phenotype, but also the phenotype algorithms designed by other sites to identify additional cases and controls with existing GWAS-level genotyping for these secondary phenotypes, This strategy was successful and identified >15,000 additional samples with existing GWAS-level data to be repurposed for other phenotypes. This effort to share and deploy phenotype algorithms across sites enabled network-wide genomic discoveries for a variety of quantitative traits (**Table 1**) and facilitated data sharing for meta-analysis efforts outside of the eMERGE network for complex diseases such as late onset Alzheimer's disease (Naj et al., 2011) and electrocardiographic traits (Jeff et al., in press).

Implicit in the eMERGE data sharing strategy is the concept that phenotype algorithms are portable across different study sites with different EMRs software systems as well as different health care practices and cultures (Kho et al., 2011). Also, it was assumed that each study site could reuse data collected for a specific phenotype or trait to conduct studies for other unrelated phenotypes without introducing substantial biases. For example, in the type 2 diabetes (T2D) association study, there was considerable heterogeneity in the proportion of type 2 diabetes cases at each site, as well the odds ratio estimates for the index T2D SNP within each site's cohort, but when combined across the sites the odds ratio was indistinguishable from those using larger purposely-collected T2D case-control collections (Kho et al., 2012). These data suggest that potential study heterogeneity was magnified or measurable at the single study level but dampened at the larger network-wide level of analysis.

To further test the boundaries of these assumptions and early observations, eMERGE undertook a network-wide study of hypothyroidism, a new phenotype not related to any of the study site-specific phenotypes. The phenotype algorithm was developed at the Vanderbilt University study site and deployed and evaluated by all eMERGE study sites, like other eMERGE phenotypes.

Table 1 | eMERGE and genomic discovery.

Phenotype	Nearest gene (rs number)	Genetic effect size	*P*	Study design (Population)	Sample size	References
Alzheimer's Disease, late onset	*BIN1* (rs7561528)	*OR* = 1.17 (95% CI: 1.13, 1.22)	4.2 ×10–14	Consortium meta-analysis, replication (EA)	8309 cases 7366 controls	Naj et al., 2011
	CD2AP (rs9349407)	*OR* = 1.11 (95% CI: 1.07, 1.15)	8.6 ×10–9	Consortium meta-analysis, discovery + replication (EA)	18,762 cases 29,827 controls	
	CD33 (rs3865444)	*OR* = 0.91 (95% CI: 0.88, 0.93)	1.6 ×10–9	Consortium meta-analysis, discovery + replication (EA)	18,762 cases 29,827 controls	
	CLU (rs1532278)	*OR* = 0.89 (95% CI: 0.85, 0.93)	1.9 ×10–8	Consortium joint-analysis, replication (EA)	8309 cases 7366 controls	
	CR1 (rs6701713)	*OR* = 1.16 (95% CI: 1.11, 1.22)	4.6 ×10–10	Consortium meta-analysis, replication (EA)	8309 cases 7366 controls	
	EPHA1 (rs11767557)	*OR* = 0.90 (95% CI: 0.86, 0.93)	6.0 ×10–10	Consortium meta-analysis, discovery + replication (EA)	18,762 cases 35,597 controls	
	MS4A4A (rs4938933)	*OR* = 0.88 (95% CI: 0.85, 0.92)	1.7 ×10–9	Consortium meta-analysis, discovery + replication (EA)	8309 cases 7366 controls	
	PICALM (rs561655)	*OR* = 0.87 (95% CI: 0.84, 0.91)	7.0 ×10–11	Consortium meta-analysis, replication (EA)	8309 cases 7366 controls	
Erythrocyte sedimentation rate	*C1orf63* (rs1043879)	$\beta = -0.09$	2 ×10–9	eMERGE joint analysis, discovery + replication (EA)	7607 individuals	Kullo et al., 2011
	CR1 (rs650877)	$\beta = -0.18$	3 ×10–26	eMERGE joint analysis, discovery + replication (EA)	7607 individuals	
	CRIL (rs7527798)	$\beta = 0.10$	2 ×10–9	eMERGE joint analysis, discovery + replication (EA)	7607 individuals	
	TMEM50A (rs25547372)	$\beta = -0.10$	2. × 10–13	eMERGE joint analysis, discovery + replication (EA)	7607 individuals	
	TMEM57 (rs25631242)	$\beta = -0.10$	1 ×10–12	eMERGE joint analysis, discovery + replication (EA)	7607 individuals	
	TMEM57 (rs25641524)	$\beta = -0.10$	5 ×10–13	eMERGE joint analysis, discovery + replication (EA)	7607 individuals	
HDL-C	*CETP* (rs3764261)	$\beta = 2.25$ (*SE* = 0.21)	1.22 ×10–25	eMERGE analysis, replication (EA)	3740 individuals	Turner et al., 2011b
	LIPC (rs11855284)	$\beta = 2.00$ (*SE* = 0.26)	3.92 ×10–14	eMERGE analysis, replication (EA)	3740 individuals	

(Continued)

Table 1 | Continued

Phenotype	Nearest gene (rs number)	Genetic effect size	*P*	Study design (Population)	Sample size	References
Hypothyroidism	*FOXE1* (rs7850258)	*OR* = 0.74 (95% CI: 0.67, 0.82)	3.96 ×10–9	eMERGE joint analysis, discovery (EA)	1317 case 5053 controls	Denny et al., 2011
LDL-C	*APOE* (rs7412)	β = −20.0 mg/dl (95% CI: −25.9, −14.1)	6.3 ×10–11	eMERGE joint analysis, discovery (AA)	618 individuals	Rasmussen-Torvik et al., 2012
Monocyte count	*CCBP2* (rs2228467)	β = 0.32	2.39 ×10–8	eMERGE joint analysis, discovery (EA)	11,014 individuals	Crosslin et al., 2013
	IRF8 (rs424971)	β = −0.25	6.32 ×10–18	eMERGE joint analysis, discovery (EA)	11,014 individuals	
	ITGA4 (rs2124440)	β = −0.22	1.35 ×10–14	eMERGE joint analysis, replication (EA)	11,014 individuals	
	RPN1 (rs2712381)	β = −0.22	4.52 ×10–14	eMERGE joint analysis, replication (EA)	11,014 individuals	
PheWAS	*EXOC2* (rs12210050)	*OR* = 1.32 (95% CI: 1.20, 1.45)	1.9 ×10–8	eMERGE pooled analysis, discovery for actinic keratosis (EA)	13,835 individuals	Denny et al., 2013
	IRF4 (rs12203592)	*OR* = 1.69 (95% CI: 1.53, 1.86)	4.1 ×10–26	eMERGE pooled analysis, discovery for actinic keratosis (EA)	13,835 individuals	
	IRF4 (rs12203592)	*OR* = 1.50 (95% CI: 1.36, 1.64)	3.8 ×10–17	eMERGE pooled analysis, discovery for non-melanoma skin cancer (EA)	13,835 individuals	
	NM37 (rs16861990)	*OR* = 3.71 (95% CI: 2.57, 5.34)	2.0 ×10–12	eMERGE pooled analysis, discovery for hypercoagulable state (EA)	13,835 individuals	
	TYR (rs1847134)	*OR* = 1.28 (95% CI: 1.18, 1.38)	2.6 ×10–10	eMERGE pooled analysis, discovery for non-melanoma skin cancer (EA)	13,835 individuals	
Platelets	*ARHGEF3* (rs1354034)	β = −0.19	9.0 ×10–34	eMERGE pooled analysis, discovery for mean platelet volume (EA)	6291 individuals	Shameer et al., 2014
	ARHGEF3 (rs1354034)	β = 7.97	6.0 ×10–24	eMERGE pooled analysis, discovery for platelet counts (EA)	13,424 individuals	
	BET1L (rs11602954)	β = −6.46	5.0 ×10–12	eMERGE pooled analysis, discovery for platelet counts (EA)	13,424 individuals	

(Continued)

Table 1 | Continued

Phenotype	Nearest gene (rs number)	Genetic effect size	*P*	Study design (Population)	Sample size	References
	DNM3 (rs2180748)	β = 0.09	2.0 ×10–8	eMERGE pooled analysis, discovery for mean platelet volume (EA)	6291 individuals	
	FLJ36031-PIK3CG (rs342240)	β = −0.15	5.0 ×10–22	eMERGE pooled analysis, discovery for mean platelet volume (EA)	6291 individuals	
	HBS1L-MYB (rs4895441)	β = −5.42	9.0 ×10–10	eMERGE pooled analysis, discovery for platelet counts (EA)	13,424 individuals	
	JMJD1C (rs4379723)	β = 0.13	3.0 ×10–16	eMERGE pooled analysis, discovery for mean platelet volume (EA)	6291 individuals	
	NFE2 (rs10506328)	β = −0.09	2.0 ×10–9	eMERGE pooled analysis, discovery for mean platelet volume (EA)	6291 individuals	
	RCL1 (rs423955)	β = 4.94	1.0 ×10–9	eMERGE pooled analysis, discovery for platelet counts (EA)	13,424 individuals	
	SH2B3 (rs3184504)	β = −5.33	5.0 ×10–11	eMERGE pooled analysis, discovery for platelet counts (EA)	13,424 individuals	
	TAOK1 (rs9900280)	β = 0.10	1.0 ×10–10	eMERGE pooled analysis, discovery for mean platelet volume (EA)	6291 individuals	
	TMCC2 (rs9660992)	β = 0.11	3.0 ×10–13	eMERGE pooled analysis, discovery for mean platelet volume (EA)	6291 individuals	
	WDR66 (rs7961894)	β = −0.31	6.0 ×10–38	eMERGE pooled analysis, discovery for mean platelet volume (EA)	6291 individuals	
QRS duration	*SCN5a* (rs1805126)	β = −1.0	1.45 ×10–8	eMERGE pooled analysis, replication (EA)	5272 individuals	Ritchie et al., 2013
Red blood cell traits	*G6PD* (rs1050828)	β = −0.20 (*SE* = 0.03)	4.0 ×10–13	eMERGE pooled analysis, discovery + replication for RBC count (AA)	2315 individuals	Ding et al., 2013
	G6PD (rs1050828)	β = 2.46 (*SE* = 0.32)	1.0 ×10–14	eMERGE pooled analysis, discovery + replication for mean corpuscular volume (AA)	2315 individuals	

(Continued)

Table 1 | Continued

Phenotype	Nearest gene (rs number)	Genetic effect size	*P*	Study design (Population)	Sample size	References
	G6PD (rs1050828)	$\beta = 0.72$ ($SE = 0.12$)	9.0 ×10–9	eMERGE pooled analysis, discovery + replication for mean corpuscular hemoglobin (AA)	2315 individuals	
	ITFG3 (rs9924561)	$\beta = -3.57$ ($SE = 0.32$)	5.0 ×10–29	eMERGE pooled analysis, discovery + replication for mean cell volume (AA)	2315 individuals	
	ITFG3 (rs9924561)	$\beta = -1.56$ ($SE = 0.12$)	8.0 ×10–36	eMERGE pooled analysis, discovery + replication for mean corpuscular hemoglobin (AA)	2315 individuals	
	ITFG3 (rs9924561)	$\beta = -0.47$ ($SE = 0.06$)	4.0 ×10–13	eMERGE pooled analysis, discovery + replication for mean corpuscular hemoglobin concentration (AA)	2315 individuals	
	(rs7120391)	$\beta = 0.30$ ($SE = 0.05$)	5.0 ×10–9	eMERGE pooled analysis, discovery + replication for mean corpuscular hemoglobin concentration (AA)	2315 individuals	
Red blood cell traits	*CDT1* (rs837763)	−0.06	2.0 ×10–8	eMERGE pooled analysis, discovery + replication for mean corpuscular hemoglobin concentration (EA)	12,486 individuals	Ding et al., 2012
	PTPLAD1/ C15orf44 (rs8035639)	0.13	8.0 ×10–9	eMERGE pooled analysis, discovery + replication for mean corpuscular hemoglobin (EA)	12,486 individuals	
	THRB (rs9310736)	0.35	6.0 ×10–9	eMERGE pooled analysis, discovery + replication for mean corpuscular volume (EA)	12,486 individuals	
	(rs9937239)	0.06	2.0 ×10–8	eMERGE pooled analysis, discovery + replication for mean corpuscular hemoglobin concentration (EA)	12,486 individuals	
Type 2 diabetes	*TCF7L2* (rs7903146)	$OR = 1.41$	2.98 ×10–10	eMERGE meta-analysis, replication (EA)	2413 cases 2392 controls	Kho et al., 2012

(Continued)

Table 1 | Continued

Phenotype	Nearest gene (rs number)	Genetic effect size	*P*	Study design (Population)	Sample size	References
White blood cell count	*DARC* (rs12075)	$\beta = 1.28$ ($SE = 0.12$)	4.92 ×10–24	eMERGE joint analysis, discovery (AA)	361 individuals	Crosslin et al., 2012
White blood cell count	*GSDMA* (rs3859192)	$\beta = 0.14$ ($SE = 0.02$)	1.75 ×10–12	eMERGE joint analysis, discovery (EA)	13,562 individuals	Crosslin et al., 2012
	MED24 (rs9916158)	$\beta = -0.13$ ($SE = 0.02$)	4.92 ×10–10	eMERGE joint analysis, discovery (EA)	13,562 individuals	
	PSMD3 (rs4065321)	$\beta = 0.14$ ($SE = 0.02$)	3.47 ×10–11	eMERGE joint analysis, discovery (EA)	13,562 individuals	

The eMERGE network has conducted or contributed data toward genome-wide association studies. For each study with genome-wide significant results ($p < 5 \times 10^{-8}$), we list the primary phenotype, the nearest genes associated, the index rs number, the reported genetic effect size, the p-value, the study design, the population, the sample size, and the reference. Abbreviations: AA, African American; EA, European American; β, beta; CI, confidence interval; OR, odds ratio; SE, standard error.

Despite potential differences in billing and coding practices across study sites, a total of 1317 cases and 5053 controls were identified with average weighted positive predictive values of 92.4 and 98.5, respectively (Denny et al., 2011). The subsequent GWAS identified common genetic variants near *FOXE1* associated with European American cases, and the findings were replicated in an independent dataset from the Mayo Genome Consortia as well as externally in the literature (Eriksson et al., 2012). These studies illustrate that existing genotype data linked to EMR data can be reused for other genomic discovery studies, a potentially cost-effective strategy. However, further study is needed to determine the extent of biases that were introduced in the generation of these data that may impact the widespread adoption of this strategy across a range of phenotypes available in the EMR.

As evident in the *FOXE1*/hypothyroidism example, existing genotype data linked to EMR data enable the relatively rapid identification of cases and controls for traditional GWAS where one disease or trait is studied. These data have also enabled the study of pleiotropy, whereby a genetic variant influences or impacts multiple phenotypes or traits (Stearns, 2010; Solovieff et al., 2013). In one popular approach, known as phenome-wide association studies or PheWAS, a GWAS-identified variant is interrogated for other associations throughout the available phenome. PheWAS has been performed in both epidemiologic (Pendergrass et al., 2013a) and EMR-based datasets such as eMERGE (Denny et al., 2010a, 2013). Collectively, these and other data (Sivakumaran et al., 2011) suggest that pleiotropy among GWAS-identified variants is not uncommon. PheWAS conducted in the EMR setting can reveal novel genotype-phenotype pleiotropic relationships not possible in traditional epidemiologic cohorts. For example, a recent PheWAS in the eMERGE participants of European ancestry revealed a potential association between actinic keratosis and *IRF4* rs12203592 (Denny et al., 2013) (**Table 1**), a GWAS-identified variant previously associated with hair color, eye color, and non-melanoma skin cancer (Han et al., 2008; Eriksson et al., 2010; Zhang et al., 2013).

Much like its contributions toward the study of pleiotropy, the eMERGE network is beginning to make substantial contributions to understudied or burgeoning areas of interest in genomic discovery such as the study of pediatric populations and diverse racial/ethnic groups. Indeed, with the addition of the pediatric study sites, eMERGE II boasts one of the largest collections of pediatric DNA samples linked to EMRs for genomic discovery (Gottesman et al., 2013). The current version (2) of the merged, imputed eMERGE II dataset includes >12,000 pediatric samples linked to EMRs. As of March 15, 2014, fewer than 5% of the GWAS annotated by the NHGRI GWAS Catalog (Welter et al., 2014) mention children as a study population, highlighting the tremendous opportunity for genomic discovery in this cohort. To calibrate the eMERGE II datasets, a site-specific investigation was recently performed for body mass index (BMI) *z*-scores using BMI extracted from the pediatric EMRs and calculated using the Centers for Disease Control and Prevention (CDC) growth charts (Namjou et al., 2013). Similar to epidemiologic datasets (Frayling et al., 2007; Meyre et al., 2009; Scherag et al., 2010), this EMR-based study demonstrated that adult GWAS-identified obesity variants such as those in *FTO* were also relevant for children of European-descent (Namjou et al., 2013). Genomic discovery using GWAS in pediatric populations is currently underway in eMERGE II for complex phenotypes such as autism and asthma.

In the past several years, most GWAS have included individuals of European ancestry (Rosenberg et al., 2010). Indeed, only approximately 10% of the GWAS annotated in the NHGRI GWAS Catalog include populations of African ancestry (https://www.genome.gov/26525384). The eMERGE network is significantly poised to contribute to GWA studies for populations of non-European ancestry given that several study sites (notably Northwestern University, Vanderbilt University, and The Icahn School of Medicine at Mount Sinai) include participants of

African ancestry. eMERGE I has already contributed genome-wide associated variants (at a threshold of $p < 10^{-5}$) in participants of African ancestry to the NHGRI GWAS Catalog for LDL-C (Rasmussen-Torvik et al., 2012), red blood cell traits (Ding et al., 2013), white blood cell traits (Crosslin et al., 2012), type 2 diabetes (Kho et al., 2012), and electrocardiographic traits (Jeff et al., 2013). As an extension of GWAS, eMERGE investigators have also begun fine-mapping GWAS-identified regions to identify the best index variant in African ancestry populations as well as exploring alternative genomic discovery methods such as admixture mapping to identify potentially novel or population-specific associations (Jeff et al., 2014).

Beyond conventional GWAS, the eMERGE network has also led efforts to identify genetic (G × G) and environmental (G × E) modifiers of common, complex phenotypes. In an early example, eMERGE investigators used extrinsic biological knowledge via the Biofilter algorithm (Bush et al., 2009) to prioritize genetic variants for SNP-SNP modeling to identify gene-gene interactions relevant for HDL-C (Turner et al., 2011b). The extrinsic biological knowledge approach has also been recently implemented for both G × G and G × E tests of association for cataracts, with the latter including only environmental variables known to be associated with the eye disease (Pendergrass et al., 2013b,c). Finally, eMERGE investigators have implemented environmental-wide association studies (EWAS) to identify and prioritize environmental factors important for type 2 diabetes (Hall et al., 2014), a relatively new approach to identify all possible environmental variables that may be relevant for G × E studies for the disease of interest.

eMERGE SECOND GENERATION GWAS

The majority of GWAS described to date for the eMERGE network represent data and efforts from phase I of the network's existence. Phase II analyses of larger, more diverse sample sizes are on-going (Gottesman et al., 2013). As documented and described in an accompanying article (Verma et al., in press), eMERGE II network datasets include single site datasets, a network-wide merged genotyped dataset, single site imputed datasets, and a network-wide merged imputed dataset; the merged set includes >36 million SNPs for samples from >50,000 individuals linked to EMRs. Imputation of the X-chromosome is underway, and future eMERGE II analyses will include this chromosome. Network-wide efforts are also underway to annotate copy number variants (Connolly et al., 2014) as well as to annotate and identify potentially deleterious null variants. Site-specific efforts are also underway to collect or extract additional standardized environmental data for GxE studies using the PhenX Toolkit (Hamilton et al., 2011; McCarty et al., 2014). Efforts are underway to develop analytical approaches for repeated measures data characteristic of the EMR, to conduct mapping studies for populations with three-way admixture events, and to incorporate phenotyping uncertainty when balancing sample size/power and misclassification (McDavid et al., 2013). With >36 million SNPs, large sample sizes, and phenotypically dense EMRs, eMERGE II and beyond promises to continue genomic discovery in the second generation of GWAS.

ACKNOWLEDGMENTS

The eMERGE Network is funded by NHGRI, with additional funding from NIGMS through the following grants: U01HG04599 and U01HG006379 to Mayo Clinic; U01HG004610 and U01HG006375 to Group Health Cooperative; U01HG004608 to Marshfield Clinic; U01HG006389 to Essentia Institute of Rural Health; U01HG004609 and U01HG006388 to Northwestern University; U01HG04603 and U01HG006378 to Vanderbilt University; U01HG006385 to the Coordinating Center; U01HG006382 to Geisinger Clinic; U01HG006380 to Mount Sinai School of Medicine; U01HG006830 to The Children's Hospital of Philadelphia; and U01HG006828 to Cincinnati Children's Hospital and Boston Children's Hospital.

REFERENCES

An integrated map of genetic variation from 1092 human genomes. (2012). *Nature* 491, 56–65. doi: 10.1038/nature11632

Bush, W. S., Dudek, S. M., and Ritchie, M. D. (2009). Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac. Symp. Biocomput*. 368–379.

Chambers, J. C., Zhao, J., Terracciano, C. M. N., Bezzina, C. R., Zhang, W., Kaba, R., et al. (2010). Genetic variation in SCN10A influences cardiac conduction. *Nat. Genet*. 42, 149–152. doi: 10.1038/ng.516

Collins, F. S., and Manolio, T. A. (2007). Merging and emerging cohorts: necessary but not sufficient. *Nature* 445:259. doi: 10.1038/445259a

Connolly, J. J., Glessner, J. T., Almoguera, B., Crosslin, D. R., and Jarvik, G. P., Sleiman, P. M. et al. (2014). Copy number variation analysis in the context of electronic medical records and large-scale genomics consortium efforts. *Front. Genet*. 5:51. doi: 10.3389/fgene.2014.00051

Crosslin, D., McDavid, A., Weston, N., Nelson, S., Zheng, X., Hart, E., et al. (2012). Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. *Hum. Genet*. 131, 639–652. doi: 10.1007/s00439-011-1103-9

Crosslin, D. R., McDavid, A., Weston, N., Zheng, X., Hart, E., de Andrade, M., et al. (2013). Genetic variation associated with circulating monocyte count in the eMERGE Network. *Hum. Mol. Genet*. 22, 2119–2127. doi: 10.1093/hmg/ddt010

Denny, J. C., Crawford, D. C., Ritchie, M. D., Bielinski, S. J., Basford, M. A., Bradford, Y., et al. (2011). Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am. J. Hum. Genet*. 89, 529–542. doi: 10.1016/j.ajhg.2011.09.008

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotech*. 31, 1102–1111. doi: 10.1038/nbt.2749

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., et al. (2010a). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* 26, 1205–1210. doi: 10.1093/bioinformatics/btq126

Denny, J. C., Ritchie, M. D., Crawford, D. C., Schildcrout, J. S., Ramirez, A. H., Pulley, J. M., et al. (2010b). Identification of genomic predictors of atrioventricular conduction. *Circulation* 122, 2016–2021. doi: 10.1161/CIRCULATIONAHA.110.948828

Ding, K., Shameer, K., Jouni, H., Masys, D. R., Jarvik, G. P., Kho, A. N., et al. (2012). Genetic loci implicated in erythroid differentiation and cell cycle regulation are associated with red blood cell traits. *Mayo Clin. Proc*. 87, 461–474. doi: 10.1016/j.mayocp.2012.01.016

Ding, K., de Andrade, M., Manolio, T. A., Crawford, D. C., Rasmussen-Torvik, L. J., Ritchie, M. D., et al. (2013). Genetic variants that confer resistance to malaria are associated with red blood cell traits in African-Americans: an electronic medical record-based genome-wide association study. *G3: Genes Genomes Genetics* 3, 1061–1068. doi: 10.1534/g3.113.006452

Eriksson, N., Macpherson, J. M., Tung, J. Y., Hon, L. S., Naughton, B., Saxonov, S., et al. (2010). Web-based, participant-driven studies yield novel genetic

associations for common traits. *PLoS Genet*. 6:e1000993. doi: 10.1371/journal.pgen.1000993

Eriksson, N., Tung, J. Y., Kiefer, A. K., Hinds, D. A., Francke, U., Mountain, J. L., et al. (2012). Novel associations for hypothyroidism include known autoimmune risk loci. *PLoS ONE* 7:e34442. doi: 10.1371/journal.pone.0034442

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316, 889–894. doi: 10.1126/science.1141634

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., et al. (2013). The electronic Medical Records and Genomics (eMERGE) network: past, present, and future. *Genet. Med*. 15, 761–771. doi: 10.1038/gim.2013.72

Hall, M. A., Dudek, S. M., Goodloe, R., Crawford, D. C., Pendergrass, S. A., Peissig, P., et al. (2014). Environment-wide association study (EWAS) for type 2 diabetes in the marshfield personalized medicine research project biobank. *Pac. Symp. Biocomput*. 200–211.

Hamilton, C. M., Strader, L. C., Pratt, J. G., Maiese, D., Hendershot, T., Kwok, R. K., et al. (2011). The PhenX Toolkit: get the most from your measures. *Am. J. Epidemiol.* 174, 253–260. doi: 10.1093/aje/kwr193

Han, J., Kraft, P., Nan, H., Guo, Q., Chen, C., Qureshi, A., et al. (2008). A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet*. 4:e1000074. doi: 10.1371/journal.pgen.1000074

Holm, H., Gudbjartsson, D. F., Arnar, D. O., Thorleifsson, G., Thorgeirsson, G., Stefansdottir, H., et al. (2010). Several common variants modulate heart rate, PR interval and QRS duration. *Nat. Genet*. 42, 117–122. doi: 10.1038/ng.511

Initial sequencing and analysis of the human genome. (2001). *Nature* 409, 860–921. doi: 10.1038/35057062

Jeff, J. M., Armstrong, L. L., Ritchie, M. D., Denny, J. C., Kho, A. N., Basford, M. A., et al. (2014). Admixture mapping and subsequent fine-mapping suggests a biologically relevant and novel association on chromosome 11 for type 2 diabetes in African Americans. *PLoS ONE* 9:e86931. doi: 10.1371/journal.pone.0086931

Jeff, J. M., Brown-Gentry, K., Goodloe, R., Ritchie, M. D., Denny, J. C., Kho, A. N., et al. (in press). Replication of SCN5A associations with electrocardiographic traits in African Americans from clinical and epidemiologic studies. *Lect. Notes Comp. Sci*.

Jeff, J. M., Ritchie, M. D., Denny, J. C., Kho, A. N., Ramirez, A. H., Crosslin, D., et al. (2013). Generalization of variants identified by genome-wide association studies for electrocardiographic traits in African Americans. *Ann. Hum. Genet*. 77, 321–332. doi: 10.1111/ahg.12023

Kho, A. N., Hayes, M. G., Rasmussen-Torvik, L., Pacheco, J. A., Thompson, W. K., Armstrong, L. L., et al. (2012). Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J. Am. Med. Info. Assoc*. 19, 212–218. doi: 10.1136/amiajnl-2011-000439

Kho, A. N., Pacheco, J. A., Peissig, P. L., Rasmussen, L., Newton, K. M., Weston, N., et al. (2011). Electronic Medical Records for Genetic Research: results of the eMERGE consortium. *Sci. Trans. Med*. 3, 79re1. doi: 10.1126/scitranslmed.3001807

Kullo, I. J., Ding, K., Shameer, K., McCarty, C. A., Jarvik, G. P., Denny, J. C., et al. (2011). Complement receptor 1 gene variants are associated wtih erythrocyte sedimentation rate. *Am. J. Hum. Genet*. 89, 131–138. doi: 10.1016/j.ajhg.2011.05.019

Kullo, I. J., Haddad, R., Prows, C. A., Holm, I., Sanderson, S. C., Garrison, N. A., et al. (2014). Return of results in the genomic medicine projects of the eMERGE network. *Front. Genet*. 5:50. doi: 10.3389/fgene.2014.00050

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., et al. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet*. 39, 1181–1186. doi: 10.1038/ng1007-1181

Manolio, T. A., Bailey-Wilson, J. E., and Collins, F. S. (2006). Genes, environment and the value of prospective cohort studies. *Nat. Rev. Genet*. 7, 812–820. doi: 10.1038/nrg1919

McCarty, C., Berg, R., Rottscheit, C., Waudby, C., Kitchner, T., Brilliant, M., et al. (2014). Validation of PhenX measures in the personalized medicine research project for use in gene/environment studies. *BMC Med. Genomics* 7:3. doi: 10.1186/1755-8794-7-3

McCarty, C., Chisholm, R., Chute, C., Kullo, I., Jarvik, G., Larson, E., et al. (2011). The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics* 4:13. doi: 10.1186/1755-8794-4-13

McDavid, A., Crane, P. K., Newton, K. M., Crosslin, D. R., McCormick, W., Weston, N., et al. (2013). Enhancing the power of genetic association studies through the use of silver standard cases derived from electronic medical records. *PLoS ONE* 8:e63481. doi: 10.1371/journal.pone.0063481

Meyre, D., Delplanque, J., Chevre, J. C., Lecoeur, C., Lobbens, S., Gallina, S., et al. (2009). Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat. Genet*. 41, 157–159. doi: 10.1038/ng.301

Naj, A. C., Jun, G., Beecham, G. W., Wang, L. S., Vardarajan, B. N., Buros, J., et al. (2011). Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat. Genet*. 43, 436–441. doi: 10.1038/ng.801

Namjou, B., Keddache, M., Marsolo, K., Wagner, M., Lingren, T., Cobb, B., et al. (2013). EMR-linked GWAS study: investigation of variation landscape of loci for body mass index in children. *Front. Genet*. 4:268. doi: 10.3389/fgene.2013.00268

Newton, K. M., Peissig, P. L., Kho, A. N., Bielinski, S. J., Berg, R. L., Choudhary, V., et al. (2013). Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J. Am. Med. Info. Assoc*. 20, e147–e154. doi: 10.1136/amiajnl-2012-000896

Pendergrass, S. A., Brown-Gentry, K., Dudek, S., Frase, A., Torstenson, E. S., Goodloe, R., et al. (2013a). Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet*. 9:e1003087. doi: 10.1371/journal.pgen.1003087

Pendergrass, S. A., Frase, A., Wallace, J., Wolfe, D., Katiyar, N., Moore, C., et al. (2013b). Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development. *BioData Mining* 6:25. doi: 10.1186/1756-0381-6-25

Pendergrass, S. A., Verma, S. S., Holzinger, E. R., Moore, C. B., Wallace, J., Dudek, S. M., et al. (2013c). Next-generation analysis of cataracts: determining knowledge driven gene-gene interactions using Biofilter, and gene-environment interactions using the PhenX Toolkit. *Pac. Symp. Biocomput*. 18, 147–158.

Pfeufer, A., van Noord, C., Marciante, K. D., Arking, D. E., Larson, M. G., Smith, A. V., et al. (2010). Genome-wide association study of PR interval. *Nat. Genet*. 42, 153–159. doi: 10.1038/ng.517

Ragoussis, J. (2009). Genotyping technologies for genetic research. *Annu. Rev. Genomics Hum. Genet*. 10, 117–133. doi: 10.1146/annurev-genom-082908-150116

Rasmussen-Torvik, L., Stallings, S. C., Gordon, A. S., Almoguera, B., Basford, M. A., Bielinski, S. J., et al. (in press). Design and anticipated outcomes of the eMERGE-PGx project: a multi-center pilot for pre-emptive pharmacogenomics in electronic health records systems. *Front. Genet.*

Rasmussen-Torvik, L. J., Pacheco, J. A., Wilke, R. A., Thompson, W. K., Ritchie, M. D., Kho, A. N., et al. (2012). High density GWAS for LDL cholesterol in African Americans using electronic medical records reveals a strong protective variant in APOE. *Clin. Transl. Sci*. 5, 394–399. doi: 10.1111/j.1752-8062.2012.00446.x

Ritchie, M. D., Denny, J. C., Zuvich, R. L., Crawford, D. C., Schildcrout, J. S., Bastarache, L., et al. (2013). Genome- and Phenome-Wide Analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* 127, 1377–1385. doi: 10.1161/CIRCULATIONAHA.112.000604

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. doi: 10.1038/nrg2760

Rukovets, O. (2013). Framingham Heart Study loses 40 percent of funding due to sequestration. *Neurol. Today* 13, 15–18.

Scherag, A., Dina, C., Hinney, A., Vatin, V., Scherag, S., Vogel, C. I. G., et al. (2010). Two new loci for body-weight regulation identified in a joint analysis of Genome-Wide Association Studies for early-onset extreme obesity in French and German study groups. *PLoS Genet*. 6:e1000916. doi: 10.1371/journal.pgen.1000916

Shameer, K., Denny, J., Ding, K., Jouni, H., Crosslin, D., Andrade, M., et al. (2014). A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum. Genet*. 133, 95–109. doi: 10.1007/s00439-013-1355-7

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., et al. (2011). Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet*. 89, 607–618.

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet*. 14, 483–495. doi: 10.1038/nrg3461

Stearns, F. W. (2010). One hundred years of Pleiotropy: a retrospective. *Genetics* 186, 767–773. doi: 10.1534/genetics.110.122549

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature* 437, 1299–1320. doi: 10.1038/nature04226

The International HapMap Project. (2003). *Nature* 426, 789–796.

Thomas, D. (2010). Gene–environment-wide association studies: emerging approaches. *Nat. Rev. Genet*. 11, 259–272. doi: 10.1038/Lnrg2764

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., et al. (2011a). Quality control procedures for genome-wide association studies. *Curr. Protoc. Hum. Genet*. 68, 1–19. doi: 10.1002/0471142905.hg0119s68

Turner, S. D., Berg, R. L., Linneman, J. G., Peissig, P. L., Crawford, D. C., Denny, J. C., et al. (2011b). Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing hdl cholesterol level in two independent EMR-linked biobanks. *PLoS ONE* 6:e19586. doi: 10.1371/journal.pone.0019586

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science* 291, 1304–1351. doi: 10.1126/science.1058040

Verma, S. S., de Andrade, M., Tromp, G. C., Kuivaniemi, H. S., Pugh, E., Namjou-Khales, B., et al. (in press). Imputation and QC for combining multiple Genome-Wide Datasets. *Front. Genet*.

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., et al. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet*. 8:e1002793. doi: 10.1371/journal.pgen.1002793

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*. 42, D1001–D1006. doi: 10.1093/nar/gkt1229

Willett, W. C., Blot, W. J., Colditz, G. A., Folsom, A. R., Henderson, B. E., and Stampfer, M. J. (2007). Merging and emerging cohorts: not worth the wait. *Nature* 445, 257–258. doi: 10.1038/445257a

Zhang, M., Song, F., Liang, L., Nan, H., Zhang, J., Liu, H., et al. (2013). Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans. *Hum. Mol. Genet*. 22, 2948–2959. doi: 10.1093/hmg/ddt142

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 March 2014; paper pending published: 23 April 2014; accepted: 30 May 2014; published online: 17 June 2014.

*Citation: Crawford DC, Crosslin DR, Tromp G, Kullo IJ, Kuivaniemi H, Hayes MG, Denny JC, Bush WS, Haines JL, Roden DM, McCarty CA, Jarvik GP and Ritchie MD (2014) eMERGEing progress in genomics—the first seven years. Front. Genet. **5**:184. doi: 10.3389/fgene.2014.00184*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 11 December 2014
doi: 10.3389/fgene.2014.00370

Imputation and quality control steps for combining multiple genome-wide datasets

Shefali S. Verma[1], Mariza de Andrade[2], Gerard Tromp[3], Helena Kuivaniemi[3], Elizabeth Pugh[4], Bahram Namjou-Khales[5], Shubhabrata Mukherjee[6], Gail P. Jarvik[6], Leah C. Kottyan[5], Amber Burt[6], Yuki Bradford[1], Gretta D. Armstrong[1], Kimberly Derr[3], Dana C. Crawford[7,8], Jonathan L. Haines[8], Rongling Li[9], David Crosslin[6] and Marylyn D. Ritchie[1]*

[1] Department of Biochemistry and Molecular Biology, Center for Systems Genomics, The Pennsylvania State University, Pennsylvania, PA, USA
[2] Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[3] The Sigfried and Janet Weis Center for Research, Geisinger Health System, Danville, PA, USA
[4] Center for Inherited Disease Research, John Hopkins University, Baltimore, MD, USA
[5] Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[6] Department of Medicine, University of Washington, Seattle, WA, USA
[7] Center for Human Genetics Research, Vanderbilt University, Nashville, TN, USA
[8] Department of Epidemiology and Biostatistics, Case Western University, Cleveland, OH, USA
[9] Division of Genomic Medicine, National Human Genome Research Institute, Bethesda, MD, USA

Edited by:
Robert Klein, Ichan School of Medicine at Mt. Sinai, USA

Reviewed by:
Robert Klein, Ichan School of Medicine at Mt. Sinai, USA
Semanti Mukherjee, Feinstein Institute for Medical Research North-Shore LIJ, USA

***Correspondence:**
Marylyn D. Ritchie, Biochemistry and Molecular Biology,Center for Systems Genomics, The Pennsylvania State University, 512 Wartik Laboratory, University Park, PA 16802, USA
e-mail: mdr23@psu.edu

The electronic MEdical Records and GEnomics (eMERGE) network brings together DNA biobanks linked to electronic health records (EHRs) from multiple institutions. Approximately 51,000 DNA samples from distinct individuals have been genotyped using genome-wide SNP arrays across the nine sites of the network. The eMERGE Coordinating Center and the Genomics Workgroup developed a pipeline to impute and merge genomic data across the different SNP arrays to maximize sample size and power to detect associations with a variety of clinical endpoints. The 1000 Genomes cosmopolitan reference panel was used for imputation. Imputation results were evaluated using the following metrics: accuracy of imputation, allelic R^2 (estimated correlation between the imputed and true genotypes), and the relationship between allelic R^2 and minor allele frequency. Computation time and memory resources required by two different software packages (BEAGLE and IMPUTE2) were also evaluated. A number of challenges were encountered due to the complexity of using two different imputation software packages, multiple ancestral populations, and many different genotyping platforms. We present lessons learned and describe the pipeline implemented here to impute and merge genomic data sets. The eMERGE imputed dataset will serve as a valuable resource for discovery, leveraging the clinical data that can be mined from the EHR.

Keywords: imputation, genome-wide association, eMERGE, electronic health records

INTRODUCTION

Imputation methods are widely used for inferring unobserved genotypes in a genotypic dataset using haplotypes from a more densely genotyped reference dataset (Browning, 2008; Howie et al., 2009, 2011, 2012; Li et al., 2009). This process is particularly important when combining or performing meta-analysis on data generated using multiple different genotyping platforms. Imputation allows for the utilization of a reference dataset and a genotyping backbone, identifying what the unobserved likely

Abbreviations: AA, African American descent; ACT, Group Health Illumina Human Omni Express genotyped subject dataset; AffyA6, Affymetrix Genome-Wide Human SNP Array 6.0; BCH, Boston Children's Hospital, eMERGE network site; BEAGLE, BEAGLE Genetic Analysis Software Package; CCHMC, Cincinnati Children's Hospital Medical Center, eMERGE network site; CHOP, The Children's Hospital of Philadelphia, eMERGE network site; DDR3, an abbreviation for double data rate type three synchronous dynamic random access memory in computing systems; EA, European American descent; EHRs, Electronic Health Records; eMERGE, The Electronic Medical Records and Genomics (eMERGE) Network is a national consortium organized by NHGRI; GB, A unit of computer memory or data storage capacity equal to 1024 megabytes; GHz, When measuring the speed of microprocessors, a GHz represents 1 billion cycles per second; HA, Hispanic American descent; HapMap, The HapMap is a catalog of common genetic variants that occur in human beings; IBD, identical by descent (IBD); IMPUTE2, IMPUTE version 2 (also known as IMPUTE2) is a genotype imputation and haplotype phasing program; kB, kilobyte is a multiple of the unit byte for digital information, 1024 bytes; kbp, kbp stands for kilobase pairs, a unit of length equal to 1000 base pairs in deoxyribonucleic acid or 1000 nitrogenous bases in ribonucleic acid; KING, software making use of high-throughput SNP data for determining family relationship inference and pedigree error checking and other uses; LD, linkage disequilibrium (LD) SNPs in the genome that can represent broader genomic regions; LIFTOVER, software tool that converts genome coordinates and genome annotation files between assemblies; MAF, minor allele frequency (MAF) refers to the frequency at which the least common allele occurs in a given population; MB, unit of computer memory or data storage capacity equal to 1,048,576 bytes (1024 kilobytes or 220) bytes; NHGRI, National Human Genome Research Institute; NWIGM, Northwest Institute of Genetic Medicine (NWIGM): Group Health Illumina 660W-Quad BeadChip genotyped subject dataset; PCA, Principal Component Analysis; Pos, chromosome position of a SNP; SHAPEIT2, version 2 of the haplotype inference software; UCSC, The University of California, Santa Cruz (UCSC).

SNPs are using patterns of linkage disequilibrium (LD) amongst surrounding markers. Multiple imputation software packages and algorithms have been developed for imputing SNPs (Browning, 2008; Browning and Browning, 2009; Li et al., 2010; Delaneau et al., 2013) (Howie et al., 2012, 2009). Although each method has clear strengths and limitations, a single "best-practice" imputation software package has not yet emerged as each tool will have different assumptions, benefits and weaknesses.

In the electronic MEdical Records and GEnomics (eMERGE) network (Gottesman et al., 2013) funded by the National Human Genome Research Institute (NHGRI), multiple genotyping platforms were have been used to generate genome-wide genotype data for thousands of patient samples and a variety of phenotypes extracted from electronic health records (EHR). To allow for either meta-analysis across the eMERGE sites or a combined mega-analysis whereby all of the eMERGE datasets are combined in a single analysis, imputation is essential to fill in the missing genotypes caused by using disparate genotyping platforms. The eMERGE Coordinating Center (CC) at the Pennsylvania State University performed genotype imputations for the eMERGE Phase-II project data [which includes all samples from eMERGE-I (McCarty et al., 2011; Zuvich et al., 2011), and eMERGE-II (Gottesman et al., 2013; Overby et al., 2013) using two different imputation pipelines: (1) BEAGLE (Browning and Browning, 2009) version 3.3.1 for phasing and imputation, and (2) SHAPEIT2 (version r2.644) (Delaneau et al., 2013) for phasing in combination with IMPUTE2 (version 2.3.0) software (Howie et al., 2012) for imputation. Imputation was performed for all autosomes, with a cosmopolitan reference panel selected from the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2012). BEAGLE used the October 2011 release and IMPUTE2 used the March 2012 release based on the timing of when imputations were performed. We did not perform X-chromosome imputations as part of this paper but the imputation of the X chromosome for these datasets is currently in progress. In these imputations, 1000 Genomes cosmopolitan reference panel was selected whereby 1092 samples from multiple race, ethnicity and ancestry groups were included in the reference panel. Using a cosmopolitan reference panel is advantageous when imputing data based on multiple ancestry or mixed-ancestry groups (Howie et al., 2011), as is the case in eMERGE datasets. To maximize our use of computational resources and allow for high quality imputations, the CC imputed the data as they were submitted to the CC, in datasets by site and genotyping platform, using the cosmopolitan panel from the 1000 Genomes.

Imputed data from all eMERGE sites were merged based on the set of intersecting SNPs present in all datasets. For the merging process, datasets that were not genotyped on dense, genome wide platform, and the datasets with fewer than 100 samples were not included as these sets routinely showed much lower quality imputation results (See Materials and Methods; additional data not shown). For example, genotyping panels containing markers in only some regions of the genome [such as the Illumina MetaboChip (Voight et al., 2012)] do not provide a suitable backbone for high quality genome-wide imputation. We looked at the quality of imputation in each of these datasets by the estimated imputation "info" score (See Results). Additionally, for datasets with very small sample size and/or not genotyped densely, median info score was close to 0 (For e.g., CHOP Illumina OmniExpress dataset with only 32 samples had median info score of 0.007), so we excluded these datasets from the merged data. After imputation and merging of the datasets, quality control procedures were implemented to create high quality, analysis-ready data set for genome-wide association studies.

Here we describe the imputation pipelines implemented using BEAGLE and SHAPEIT2/IMPUTE2; provide results of the two imputation pipelines; and describe the quality control procedures after merging multiple imputed datasets. Numerous lessons were learned along the way for each of these imputation pipelines and we share all of the challenges encountered in the project. The imputation and quality control procedures resulted in unique and comprehensive a dataset of over 50,000 samples with genotypes imputed to the 1000 Genomes reference panel, all linked to de-identified EHR to allow for a vast array of genotype-phenotype association studies.

MATERIALS AND METHODS

STUDY DATA

The eMERGE network consists of seven adult sites and two pediatric sites, each with DNA databanks linked to EHR. Each site in the network has a set of at least 3000 samples that have been genotyped on one or more genotyping platforms (Gottesman et al., 2013). **Table 1** provides a summary of the number of samples from each site and the genotyping platforms used. Previous studies have shown that the quality of input genotype data does not affect imputation quality in a significant manner (Southam et al., 2011), but nevertheless we selected the genomic data sets for the current imputation study that had all undergone the pre-processing recommended by the eMERGE CC to eliminate samples and SNPs with call rates less than 99–95% depending on the coverage of genotyping for each platform (Zuvich et al., 2011). Minor allele frequency (MAF) threshold of 5% was also applied. This ensures that only high quality data were considered for imputation and downstream analyses.

Several eMERGE sites genotyped duplicate samples on multiple different genotyping platforms, for quality control purposes. A total of 56,890 samples were submitted to the eMERGE CC for imputation, out of which 53,200 samples were unique. All of these samples were genotyped and deposited to CC at different times, so imputation was performed as the datasets arrived. This resulted in imputing some datasets with fewer than 100 samples. When the dataset was less than 100 samples, we included the 1000 Genomes dataset with the study data during phasing. We imputed all samples; however, for the purpose of merging the data, we only merged high quality datasets (defined by having masked analysis concordance rate greater than 80%; described in more detail below). We included only one sample from pairs of duplicates; specifically the sample genotyped on the higher density genotyping platform. Our final merged dataset contains 51,035 samples. Samples that had low quality due to either of the following two reasons were not included:

Table 1 | Sample summary across all eMERGE datasets.

Sample set	Genotyping platform	Samples for imputations	Samples in merged data set
ADULT DNA SAMPLES			
eMERGE-I 1M	Illumina 1M	2634	2634
eMERGE-I 660	Illumina 660	18663	16029
Geisinger OMNI	Illumina HumanOmni Express	3111	3111
Geisinger Metabochip	Metabochip	918	0
Mayo Clinic	Illumina Human 610, 550, and 660W Quad-v1	3149	3118
Mt. Sinai AA	Affymetrix 6.0	863	863
Mt. Sinai EA	Affymetrix 6.0	700	700
Mt. Sinai HA	Affymetrix 6.0	1212	1212
Mt. Sinai OMNI_AA	Illumina HumanOmni Express	3515	3515
Northwestern University	Illumina HumanOmni Express 12v1_C	3030	2951
Vanderbilt University	Illumina HumanOmni Express 12v1_C	3565	3461
Group Health/ACT	Illumina HumanOmni Express	398	398
Group Health/NWIGM	Illumina 660W-Quad Beadchip	341	333
Marshfield Clinic	Affymetrix/Illumina 660	500	500
Total for adult DNA samples		42,599	38,824
PEDIATRIC DNA SAMPLES			
CCHMC	610/660W/AffyA6/OMNI1/OMNI5	5558	4322
BCH	Affymetrix Axiom	1038	1038
CHOP	550/610/Beadchip/AffyA6/AffyAxiom/OmniExpress	7695	6850
Total for pediatric DNA samples		14,291	12,210
Total		56,890	51,035

"Samples for imputations" column contain number of samples that were obtained by coordinating center at different time points. "Samples in merged dataset" contain number of unique samples that were used in merged dataset. For the samples that were genotyped on multiple platforms, sample on platform with high genotype efficiency was used in merged dataset.

1. Samples not genotyped on dense, genome-wide genotyping platform (e.g., the MetaboChip).
2. Sample size of the dataset on the specific platform for phasing was fewer than 100 (as recommended in SHAPEIT2 guidelines).

A small number of samples were also genotyped for two SNPs (rs1799945 and rs1800562) using commercially available 5′-nuclease assays (TaqMan® Assay; Life Technologies). Genotyping reactions were carried out in 10 μl volumes in an ABI 7500 Fast Real-Time PCR System (Life Technologies). The genotypes were called using ABI 7500 software version 2.0.4 (Life Technologies). These data were used to evaluate the concordance of imputed genotypes with TaqMan generated genotypes.

PRE-IMPUTATION DATA PROCESSING

The quality of imputation relies on the quality of the reference panel as well as the quality of the study data. To ensure high data quality, there are a number of steps that were taken before imputation begins. At the start of the BEAGLE imputation, the GENEVA HAPO European Ancestry Project Imputation Report (Geneva_Guidelines[1]) by Sarah Nelson through GENEVA (Gene-Environment Association Studies) was used as a guide and a starting point for implementation of the eMERGE imputation pipeline. GENEVA is an NIH-funded consortium of sixteen genome-wide association studies (GWAS) and this guide served as the basis to begin the eMERGE Phase-II imputation process.

CONVERTING REFERENCE PANEL AND STUDY DATA TO THE SAME GENOME BUILD

The genotype data were initially accessed from binary PLINK files (Purcell et al., 2007). All SNP names and locations for the genotypic data being imputed had to be specified based on the same genome build, as well as the same genome build of the reference genome. The Genome Reference Consortium Human build 37 (GRCh37 or build 37) is the reference genome used in our study (2010). Some eMERGE sites had their data in build 37, while others were still in build 36. Any datasets that were not in build 37 were first converted from build 36 to build 37 using the Batch Coordinate Conversion program liftOver (Karolchik et al., 2011) via the following steps:

1. SNPs with indeterminate mappings were removed (either unknown chromosome and/or unknown position) in build 37.
2. SNP names were updated.

[1]Available at: https://www.genevastudy.org/sites/www/content/files/data cleaning/imputation/Lowe_Eur_1000G_imputation_final.pdf

3. The chromosome positions were updated.
4. The base pair positions were updated.

The program liftOver is a tool developed by the Genome Bioinformatics team at the University of California, Santa Cruz (UCSC) to convert genome coordinates and genome annotation files between assemblies. This process ensures that all study data from eMERGE sites and the 1000 Genomes reference data are referring to SNPs by the same alleles and genome location.

CHECKING STRAND

Study and reference data allele calls must be on the same strand for proper imputation, however the strand could vary from study site to study site due to genotyping platform and calling algorithm. High quality imputation is exceptionally reliant upon the study and reference data allele calls to be on the same physical strand of DNA in respect to the human genome reference sequences ("reference"). Datasets could have different notations depending on the genotyping platform and the calling algorithm. For example, Genome Studio will allow the user to create genotype files using different orientations. In addition, some users may use custom genotype callers—not provided by the genotyping chip manufacturer. For example, some platforms use the forward strand of the human genome assembly and some use Illumina's TOP alleles, and some use Illumina's AB alleles (Illumina TechNote[2]). To identify the SNPs requiring a strand flip to convert the forward allele to the "+" strand of the human genome reference assembly so that all sites were consistent in terms of the same strand, we used the BEAGLE strand check utility for BEAGLE imputations and the SHAPEIT2 strand check for IMPUTE2 imputations even though IMPUTE2 automatically addresses ambiguous strand alignments by comparing allele labels. During strand check, alleles are changed to their complementary alleles (C-G and A-T) based on three criteria: (a) the observed alleles, (b) minor allele frequencies (MAF), and (c) linkage-disequilibrium (LD) pattern within 100-SNP windows. SNPs where MAF and LD patterns are inconsistent and also cannot be resolved by flipping, those SNPs are discarded from the dataset. Before phasing, we subset the data by chromosomes and also flipped strand for the SNPs to align the dataset with "+" strand so that it corresponds to reference panel strand correctly.

PHASING

Haplotype phasing is the next step after ensuring that all data was using the same strand, identifying alleles co-localized on the same chromosome. BEAGLE performs phasing jointly with imputations. "Pre-phasing" indicates that a computational step is implemented prior to imputation where haplotype phase is estimated for all of the alleles. We utilized a pre-phasing approach because it helps to make the process of imputation faster, and the phased data can be used for any future imputation of the data. Improved reference panels will be introduced over time, and thus having the data saved pre-phased for imputation can speed up later imputation of the data. Phasing the data can introduce some error to the imputations, because of any haplotype uncertainty (Howie et al., 2012).

For IMPUTE2 imputations, following "best practices" guidelines in the IMPUTE2 documentation (Howie et al., 2009) (Impute2, 2.3.0) we first phased the study data with the SHAPEIT2 haplotype estimation tool (Howie et al., 2012). We were able to reduce general runtime through using multiple computational processing cores via the "—thread" argument. A general example of the command line syntax used to run the SHAPEIT2 program on chromosome 22 using the "—thread" argument is shown below:

```
shapeit2 --input-ped StudyData_chr22.ped
StudyData_chr22.map \
--input-map genetic_map_chr22_combined_
b37.txt \
--output-max StudyData_chr22.haps
StudyData_chr22.sample \
--thread 2 --output-log shapeit_chr22.log
```

IMPUTATION USING BEAGLE

To expedite the imputation using BEAGLE, we divided each chromosome into segments including 30,000 SNPs each (referred to as SNPlets), following one of several recommendations in the BEAGLE documentation (Browning and Browning, 2009) for imputing large data sets. A buffer region of 700 SNPs was added to each end of every SNPlet to account for the degradation in imputation quality that may occur at the ends of imputed segments. An illustration of this segmentation is shown in **Figure 1**. Partitioning was implemented by dividing the ".markers" files created at the end of the strand check into 1 ".markers" file for each SNPlet of 30,000 SNPs and a 700 SNP buffer region on either end. In each results file, the data for all SNPs in the buffer regions were removed such that each imputed SNP had results from only one segment. The SNP annotation and quality metrics file accompanying these data indicate to which segment each SNP was assigned.

Below is an example of the command line syntax used to run BEAGLE on the first segment of chromosome 22. The "phased=" argument corresponds to the 1000 Genome Project reference panel input file; the "excludemarkers=" argument points to a combined list of SNPs that are either (1) triallelic SNPs or (2) have reference MAF < 0.005. The "unphased=" argument points to a BEAGLE-formatted input file:

```
java -Djava.io.tmpdir=/scratch/tmp
-Xmx4700m -jar BEAGLE.jar \
unphased= chr22_mod.bgl \
phased= chr22.filt_mod.bgl \
markers= chr22_*.markers \
excludemarkers= allchr22_snpsexclude.txt \
lowmem=true verbose=true missing=0
out=out_chr22set1
```

IMPUTATION USING IMPUTE2

To perform imputation with IMPUTE2 on our phased data, we divided each chromosome into base pair regions of approximately

[2]Available at: http://res.illumina.com/documents/products/technotes/technote_topbot.pdf

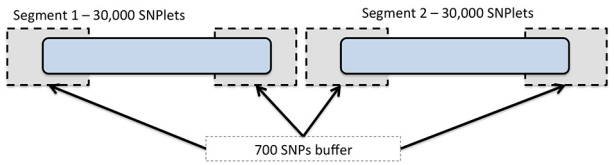


FIGURE 1 | Chromosome segmentation strategy for genome-wide imputation with BEAGLE. Each chromosome was divided into SNPlets which included 30,000 SNPs with a buffer of 700 SNPs at each end.

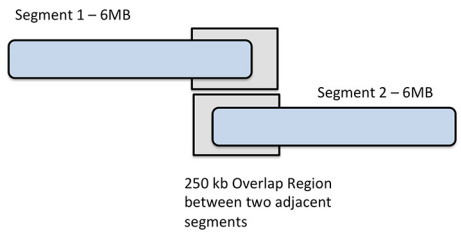


FIGURE 2 | Chromosome segmentation strategy for imputation with IMPUTE2. Each chromosome was divided into 6 MB segments with 250 kbp overlap between them.

6 Mb in size, beginning at the first imputation target, as displayed in **Figure 2**. As a result, we partitioned 22 autosomes into 441 segments, ranging from only 7 segments on chromosome 21, to the largest number of segments (39) on chromosome 2. It is of interest to note that there were 36 segments on chromosome 1. It was beneficial to use this process of breaking the genotypic data into smaller regions because IMPUTE2 has been reported to have improved accuracy over smaller genomic regions and also separating data into segments helps allows for the parallelization of jobs over a multi-core compute cluster. Segments either overlapping the centromere or at the terminal ends of chromosomes were merged into the segment immediately upstream.

IMPUTE2 labels SNPs by the panels in which they have been genotyped. Each label denotes a specific functional role. SNPs that have genotype data only in the reference panel are labeled *Type 0* or *Type 1* (for phased and unphased reference panels, respectively), whereas SNPs that have genotypes in the study dataset are labeled *Type 2*. These are considered SNPs for the imputation basis. *Type 2* SNPs dictate which reference panel haplotypes should be "copied" for each individual; then, the reference panel alleles at *Type 0/1* SNPs are used to fill in the missing genotypes of the individual.

As recommended by the IMPUTE2 guidelines, we ensured that each base pair region that was imputed contained at least some observed (type-2) SNPs. To utilize type-2 SNPs for estimating haplotype structure, a buffer region on both sides of segments is required. 250 kb buffer region is default for IMPUTE2 so we used the default buffer size of 250 kb for eMERGE imputations. By default, IMPUTE2 flanks imputation segments with a 250 kb buffer, where type-2 SNPs are used to estimate haplotype structure. We used the default buffer size of 250 kb for imputations.

An example of the command line syntax we used to run first 6 MB segment (pre-phased) for chromosome 22 by IMPUTE2 (version 2) is shown below:

```
impute2 -use_prephased_g -m genetic_map_
chr22_combined_b37.txt \
-h ALL_1000G_phase1interim_jun2011_chr22_
impute.hap.gz \
-l ALL_1000G_phase1interim_jun2011_chr22_
impute.legend.gz \
-int 16000001 2.1e+07 -buffer 500 -allow_
large_regions \
-known_haps_g StudyData_chr22.haps \
-filt_rules_l Study_data.maf<0.001
-align_by_maf_g \
-o StudyData_chr22.set1.gprobs \
-i StudyData_chr22.set1.metrics -verbose
```

RESULTS

COMPUTATION TIME AND MEMORY USAGE

Imputation jobs were run in parallel across several high-performance computing clusters; specialized systems were chosen depending on the memory and processor requirements of the software and the size of the datasets. **Figure 3** shows the workflow of both imputation methods using the different software and how the performance results and computational requirements differed for each. Each job required between 4 and 24 GB RAM and from 4 to 80 CPUs (cores). The number of jobs submitted to be run in parallel also ranged from a few 100 to several 1000 at a time according to the sample size of each data set. **Table 2A** provides information on one of the computing clusters that were used to perform these extensive imputations by the eMERGE CC. **Table 2B** lists maximum time and memory from each of the datasets that was required to run both imputation and phasing. One thing to note here is that according to available sources at the time of running specific job, different CPU cores were utilized.

The largest variance for computing resource requirements was in the computational time required on the same cluster computing systems for the two different pipelines. Previous studies have compared both BEAGLE and IMPUTE2 programs based on the quality and imputation times (Pei et al., 2008; Howie et al., 2009, 2011; Nothnagel et al., 2009). Our work similarly showed that IMPUTE2 ran much faster than BEAGLE. For BEAGLE imputations, SNPlet runtimes varied between 40 and 200 h, on average using 6 GB of memory for each job for a total of 1.64×10^6 CPU hours.

In summary SHAPEIT2 and IMPUTE2 processing, took only 13 h on an average for phasing using 10 GB memory with a maximum of 16 CPUs (4 cluster computing nodes where each node had 4 CPUs). Similarly imputations on average could be completed in 4.5 h of time using 24 CPUs (across multiple cluster nodes). For processing the final merged set, approximately 80 CPUs were required. The total computational time required for the SHAPEIT2 and IMPUTE2 processing was less than 600 CPU hours. Using the pre-phasing approach, imputation time was decreased by more than 10–fold with the unfortunate side-effect of utilizing intensive memory.

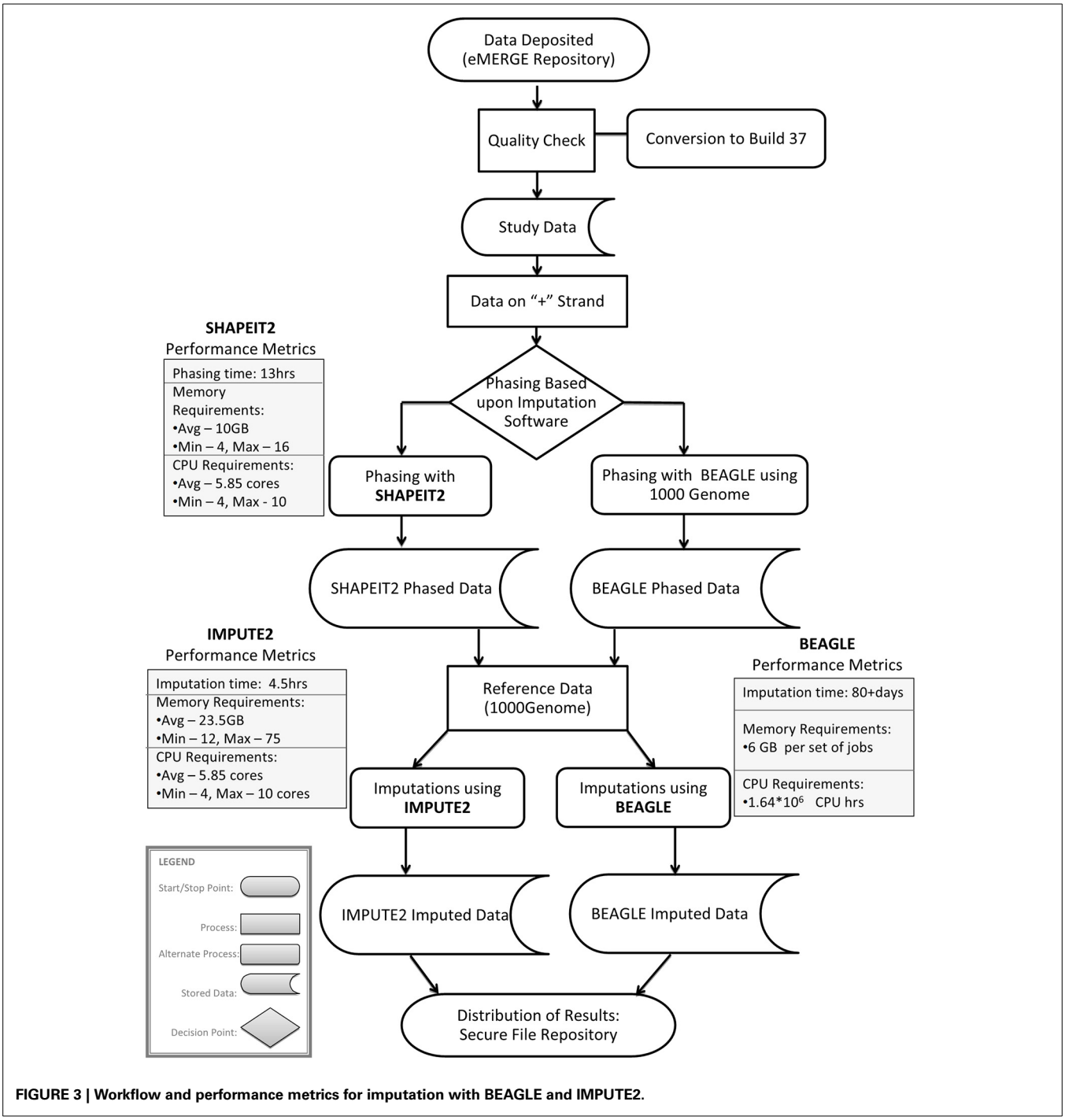


FIGURE 3 | Workflow and performance metrics for imputation with BEAGLE and IMPUTE2.

COMPARISION OF BEAGLE AND IMPUTE2

BEAGLE and IMPUTE2 methods have been compared extensively by previous studies of a single ancestry (i.e., European or African) and using a cosmopolitan reference panel (Browning and Browning, 2009; Howie et al., 2009; Nothnagel et al., 2009; Jostins et al., 2011; 1000 Genomes Project Consortium et al., 2012). Initially, we planned to perform a direct comparison of the two imputation programs. We found that the resource requirements to do that were prohibitive, since the 1000 Genomes reference was updated in between our BEAGLE runs and our IMPUTE2 runs. This update presented a conundrum since the update includes a large number of InDels and our proposed downstream analyses would be improved by using the updated reference set. Due to BEAGLE's compute intensive implementation we did not have the compute resources or the time to repeat the imputation with the new reference dataset. Similarly repeating the IMPUTE2 runs using the old reference, even though it was much faster than BEAGLE was prohibitive

Table 2 | Computational resources used for conducting the imputations.

(A) Penn State Lion XG: systems specifications

Component	Server	Quantity	Processor	Number of processor cores	Memory (GB)
Login Node	Dell PowerEdge R620	1	Intel Xeon E5-2670 2.6 GHz	16	64
Compute Node	Dell M620	48	Intel Xeon E5-2665 2.4 GHz	16	128
Compute Node	Dell M620	48	Intel Xeon E5-2665 2.4 GHz	16	64
Compute Node	HP BL460c Gen8	96	Intel Xeon E5-2665 2.4 GHz	16	64

(B) Phasing and imputation time and RAM required for each dataset

Site_name	#Samples	Phasing time (maximum seconds)	Phasing RAM	Imputation_time (maximum seconds)	Imputation RAM
eMerge-I-1M	2634	199984	4	7728	20
eMerge-I-660	18663	17263	4	35517	75
BCH	1038	22963	4	2380	12
Geisinger_Metabochip	918	55002	8	6011	24
Geisinger_OMNI	3111	2010	8	7397	20
Grouphealth_ACT	398	62141	8	1330	16
Grouphealth_NWIGM	341	8063	6	730	20
Mayo	6307	5627	6	21172	16
MtSinai_EA	700	130737	16	3617	30
MtSinai_AA	863	33832	10	2308	12
MtSinai_HA	1212	50884	10	3120	12
MtSinai_OMNI	3515	52000	16	13276	12
NU	3030	311211	32	6089	24
Vanderbilt	3565	19392	12	10583	20
Total CCHMC	4322	82450	12	4310	28
Total CHOP	6850	74501	12	7768	30

in terms of compute time and storage space with a dataset of 55,000 samples. Therefore, we will provide only anecdotal differences that we observed between IMPUTE2 and BEAGLE. For more complete, direct comparisons of the two approaches, we direct the reader to some of the earlier studies mentioned above.

In our study dataset, we observed that IMPUTE2 is substantially faster than BEAGLE but they both achieved comparatively equal accuracy with a large reference panel, such as the 1000 Genomes. Our BEAGLE imputations were only performed for adult data, so to look at the frequency of high quality markers, we compare the counts to adult only data in IMPUTE2. We observed that 8,899,961 SNPs passed allelic R^2 filter of 0.7 in BEAGLE imputations whereas for same data using IMPUTE2 imputations, 12,504,941 SNPs passed info score filter of 0.7. Lastly, we also observed that in BEAGLE imputed data at MAF = 0.05, there were SNPs with Allelic R^2 value less than 0.6 whereas with IMPUTE2 imputed data all SNPs with MAF = 0.05, were above info score value greater than 0.6.

Keeping the huge computational advantage of IMPUTE2 as well as quality of imputation in mind, especially when dealing with the imputation of over 50,000 samples, we used IMPUTE2 for further imputations and analyses. Thus, in the remainder of the paper, we will describe the output and quality metrics that we observed for IMPUTE2 in the eMERGE dataset.

MASKED ANALYSIS

One of the greatest challenges with imputation is knowing how well it is working. A common strategy used to evaluate this is called "masked analysis." In a masked analysis, a subset of SNPs that were actually genotyped in the study sample are removed, those SNPs are then imputed as though they were not genotyped, and then the imputed SNPs are compared to their original genotypes. The results of the imputation are contrasted with the original genotypic data, showing the degree of concordance between the original genotypic data and the imputed data after masking. This gives a good sense of how accurate the imputations are with respect to that set of SNPs. An additional way the results of masking and imputation are evaluated is to compare the allelic dosage of the original genotypic data with that of the allelic dosage in the imputed data. If there are three genotypes AA, AB, and BB, the allelic dosage for each individual can be represented as probabilities (P) of each of three genotypes via $2*P(AA) + 1*P(AB) + 0*P(BB)$ to obtain the expected allelic dosage from the original genotypic data and the observed allelic dosage for the masked and imputed genotype for each SNP. The correlation between the expected allelic dosages and the observed allelic dosages over all individuals can then be calculated at each

masked SNP. This correlation metric is an exact variant of the imputation R^2 metrics of MACH (Li et al., 2010) and BEAGLE, which corresponded with the IMPUTE "info" score which is calculated automatically as part of IMPUTE2. Here Type 2 SNPs are removed from imputation, and then imputed, and contrasted with imputation input. Thus, metric files from IMPUTE2 provide information from these masked SNP tests, including concordance and correlation metrics, and an "info" metric for having treated a Type2 the SNP as Type 0.

Overall concordance is vigorously impacted by the MAF and we say so on the grounds that for SNPs with MAF $< 5\%$ by simply allocating imputed genotypes to the major homozygous state would result in $>90\%$ concordance. Thus, there is an inclination of high concordance values at low MAF SNPs, where major homozygotes are prone to be imputed "correctly" just by chance. We observed approximately 99% average concordance in masked SNPs grouped by MAF.

ORTHOGONAL GENOTYPING ANALYSIS

As another imputation quality check, we compared the genotypes generated in the imputation with those genotyped on orthogonal genotyping platforms. Two SNPs, rs1800562 and rs1799945 were genotyped using TaqMan by the genotyping facility at Geisinger Health System. The concordance between the TaqMan genotype and the imputed data was 98.9 and 98.3% for rs1800562 and rs1799945, respectively. These are very similar to results observed in the Marshfield Clinic PMRP where an orthogonal platform was used (Verma et al., 2014).

MERGING OF IMPUTED DATASETS

Prior to imputation, we explored the option of combining the raw genotype data based on overlapping SNPs from the multiple GWAS platforms. Unfortunately the number of overlapping SNPs was minimal (only 37,978 SNPs). This was not sufficient for imputation. Thus, we imputed each dataset based on site

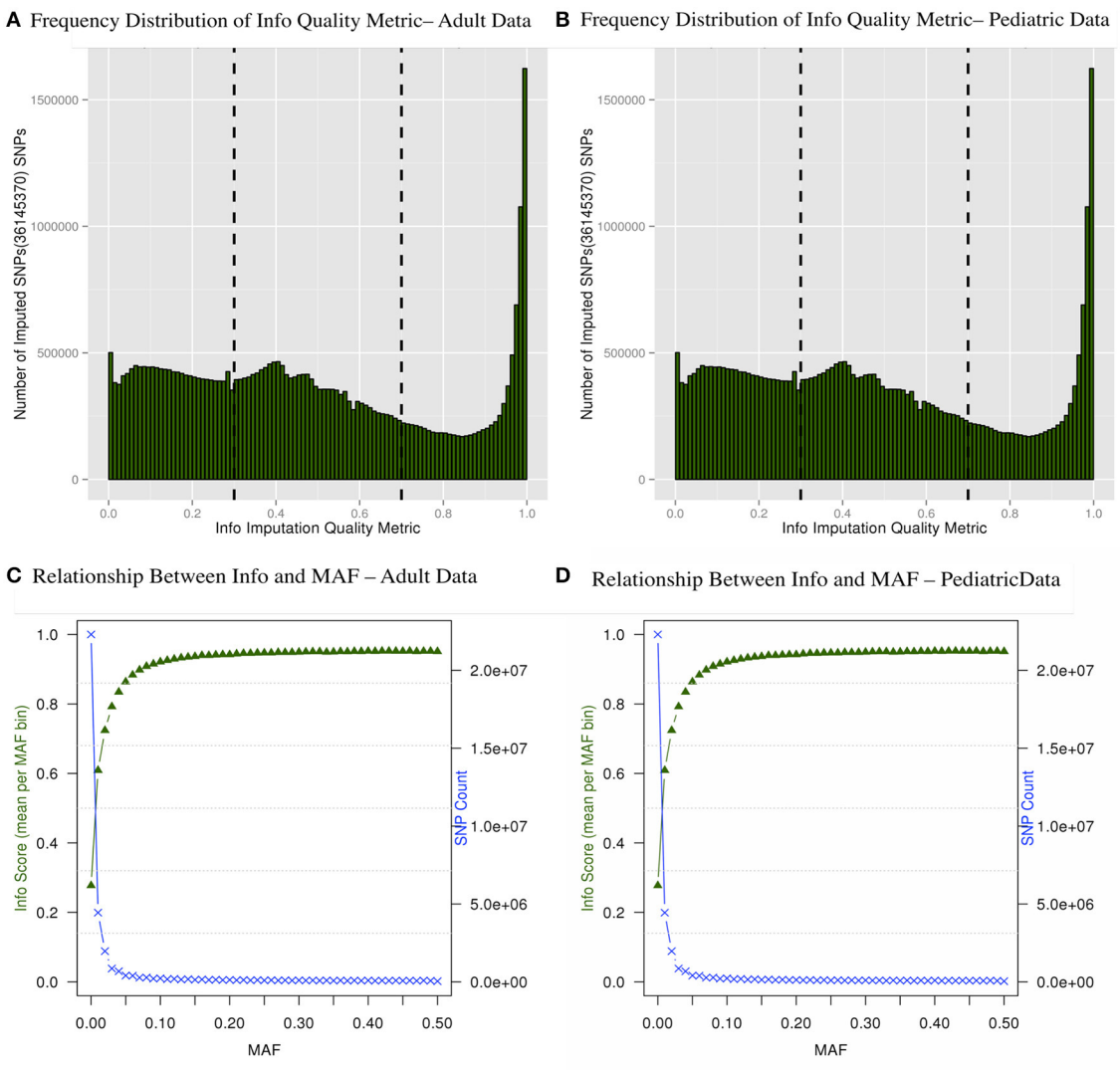


FIGURE 4 | Frequency distribution of "info" quality metric (A,B) and relationship between the "info" score and MAF are shown (C,D). The secondary axis indicates the count of SNPs in each MAF bin (0.01 intervals).

Table 3 | Number and proportion of SNPs dropped and remaining at different genotyping call rate threshold after merged data is filtered at info score >0.7.

Threshold	SNPs dropped at threshold	Proportion of SNPs dropped at threshold	SNPs remaining at threshold	Proportion of SNPs remaining at threshold
ADULT DNA SAMPLES				
0.95	1650764	0.1494	9400761	0.8506
0.98	3609986	0.3267	7441539	0.6733
0.99	5619475	0.5085	5432050	0.4915
PEDIATRIC DNA SAMPLES				
0.95	2165777	0.1275	14810393	0.8724
0.98	4983111	0.2935	11993059	0.7065
0.99	7692022	0.4531	9284148	0.5469

Table 4 | Number and proportion of samples dropped and remaining at different sample call rate threshold after merged data is filtered at info score >0.7 and marker call rate 99%.

Threshold	SNPs dropped at threshold	Proportion of dropped at threshold	SNPs remaining at threshold	Proportion of SNPs remaining at threshold
ADULT DNA SAMPLES				
0.95	5	0.0001	38823	0.9999
0.98	57	0.0015	38771	0.9985
0.99	4632	0.1193	34196	0.8807
PEDIATRIC DNA SAMPLES				
0.95	10	0.0008	12200	0.9991
0.98	79	0.0647	12131	0.9935
0.99	497	0.0407	11713	0.9592

and platform individually. After imputing each study dataset, we attempted to merge all of the imputed datasets together to generate a mega-analysis ready dataset (combining all eMERGE sites together). The imputed data from all eMERGE sites studying adult-onset diseases were merged into one dataset and all pediatric data were merged into a second set. Future directions include combining adult and pediatric data. Imputed datasets were merged based on the set of intersecting markers [only markers that were of high quality in all of the imputed data were combined (i.e., info score >0.7)]. Duplicate samples were removed, whereby the highest quality version of the sample was maintained. For example, if a sample was genotyped on two platforms with different call rates, we kept the result from the platform with the higher call rate. Additionally, the low quality data were omitted from the final version of the merged data. Low quality of imputation was determined by assessing the masked concordance rates calculated from IMPUTE2. Notably, most data that were not genotyped on a dense, genome-wide platform (such as MetaboChip or Illumina HumanHap 550 Duo BeadChip) had masked concordance rates <80% (Nelson et al., 2013). The lower concordance was probably due to a lack of a uniform backbone or imputation basis to use for construction of the LD patterns for imputation. As such, those datasets were not included in merged dataset. Finally, as recommended in both the SHAPEIT2 and IMPUTE2 guidelines (Impute2[3]), small sample size datasets (<100 samples) did not achieve high quality imputations; thus, we excluded them from the merged data.

To merge all of the datasets together, we implemented a script that takes IMPUTE2-formatted input files and cross-matches them based on SNP position and alleles, rather than the marker label (as sometimes marker labels are not shared). For each matching position, allele1, and allele2, which are found in all inputs, the output is given the most common label from among the inputs. The script detects cases where there are different SNP labels for the same Chr:Pos and alleles and resolves it by treating these as equivalent markers which will be joined into one output line, using the marker label which has larger rs#. For cases where there are more than one position for the same SNP label, the script will then drop both of them.

Imputation results have multiple columns of information. The first five columns relate to Chromosome, SNP ID, base pair location, and the two SNP alleles, where the first allele indicated is assigned "allele A," and the second is assigned "allele B." The following three columns represent the genotype probabilities of the three-genotype classes (AA, AB, and BB) for each individual sample; a simulated example shown here:

Table 5 | MAF distribution for all SNPs after applying info score (0.7) and marker call rate filter (99%).

Threshold	SNPs dropped at threshold	Proportion of SNPs dropped at threshold	SNPs remaining at threshold	Proportion of SNPs remaining at threshold
ADULT DNA SAMPLES				
0.05	2803753	5.1615e-01	2628296	0.4838
0.01	995223	1.8321e-01	4436826	0.8168
0.005	466779	8.5930e-02	4965270	0.9141
0.001	13979	2.5734e-03	5418070	0.997
0.0005	624	1.1487e-04	5431425	0.9998
0.0001	1	1.8409e-07	5432048	0.9999
PEDIATRIC DNA SAMPLES				
Threshold	**#SNPs dropped at threshold**	**Proportion of SNPs dropped at threshold**	**#SNPs remaining at threshold**	**Proportion of SNPs remaining at threshold**
0.05	6523370	7.3938e-01	2299322	0.2606
0.01	4631783	5.2498e-01	4190909	0.4750
0.005	3141053	3.5601e-01	5681639	0.6440
0.001	240254	2.7231e-02	8582438	0.9728
0.0005	30674	3.4767e-03	8792018	0.9965
0.0001	19	2.1535e-06	8822673	0.9999

[3] Available at: https://mathgen.stats.ox.ac.uk/impute/impute_v2.html

CHR	SNP_ID	POSITION	allele A	allele B	Sample1_AA	Sample1_AB	Sample1_BB
22	rs149201999	16050408	T	C	0.251	0.501	0.248
22	rs146752890	16050612	C	G	0.302	0.495	0.203
22	rs139377059	16050678	C	T	0.252	0.501	0.247
22	rs188945759	16050984	C	G	1	0	0

Imputed genotype files contain three types of IMPUTE2 SNPs: Type 0 (imputation target); Type 2 (imputation basis); and Type 3 (study only). Accompanying information metrics files provide information on what type of SNP each SNP within the dataset was. Note there are no sample identifiers in the probabilities files, consequently it is important to utilize sample information documents provided to adjust imputed probabilities to sample data. Merged "info" or quality metrics file contains following information:

1. "snp_id" is always "---"which is how it often appears in the input files, and "rs_id" and "position" match the genotype output file.
2. "type" is the numeric minimum of the observed input values.
3. The other columns are all simple (equally weighted) averages of the input values, except that any −1 inputs are ignored (for example. the average of 0.5, −1, 0.3 is 0.4, ignoring the −1).
4. There is also a special case for the "exp_freq_a1" column for inputs which have alleles reversed compared to the first input (allele1 is not always major or minor allele); in that case the value is subtracted from 1.0 before going into the average so that we always report frequencies for minor allele in merged dataset.

IMPUTED DATA STATISTICS FOR IMPUTE2

There are multiple results from imputation that should be evaluated before proceeding with association analyses for imputed SNPs. For instance, it is critical to consider the uncertainty of the imputed genotypes **Figure 4A** shows the distributions of the information (reported as "info score") metrics for all variants in the adult imputed datasets and **Figure 4B** demonstrates the relationship between MAF and imputation quality for all variants in the pediatric imputed datasets by showing average "info" scores plotted in all variants grouped by MAF (bin sizes of 0.1). Although the total number of imputed variants for the adult and pediatric datasets is very similar, it was notable that there were comparatively more markers with low info score in the pediatric dataset. One potential reason for this discrepancy could be due to a much large number of genotyping platforms in the imputations

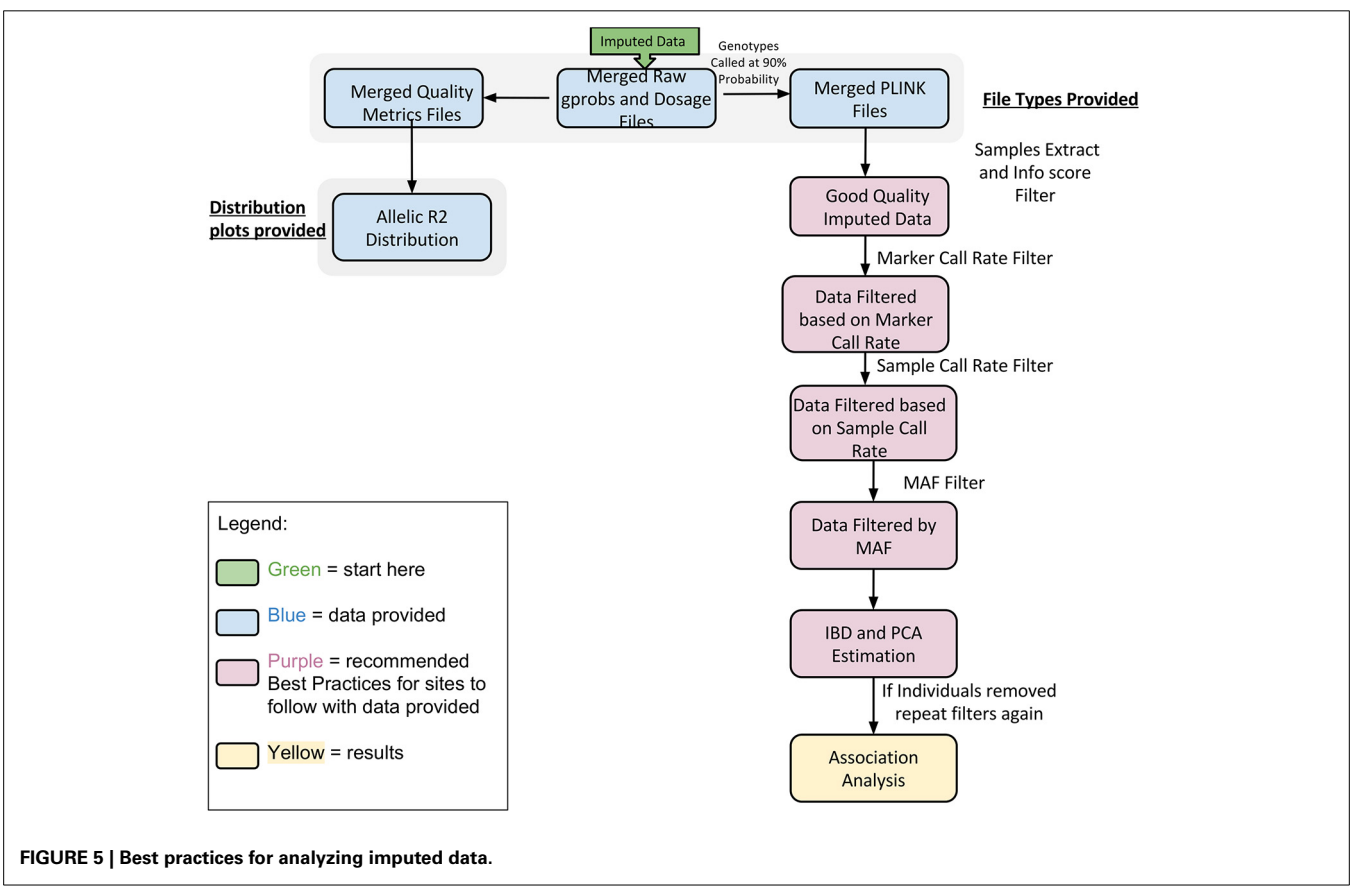


FIGURE 5 | Best practices for analyzing imputed data.

of the pediatric datasets. While the average "info" scores with MAF < 0.05 fall lower than an info score of 0.8 as demonstrated in **Figures 4C,D** for adults and pediatric data respectively, within higher MAF bins, the average "info" scores increase to approximately 0.9. This metric demonstrates that variants imputed to have low MAF in the study samples are likely to have low MAF in the reference panel. We attempted to not include any monomorphic SNPs in the imputed dataset, our inclusion criteria was to include any imputed SNP that had at least one copy of minor allele. So the reason that we see a lot of SNPs with low "info" score is mostly due to the chosen imputation target and not any procedural error. Although there is no consensus in filtering the imputed datasets based on uncertainty of imputation, we used a variant level filter (info score >0.7) (Lin et al., 2010; Southam et al., 2011) for the downstream analyses. This is a conservative threshold, whereby we are balancing the quantity of lost data with data quality. Other studies may choose to be more liberal (info score >0.3) or even more conservative (info score >0.9).

QUALITY CONTROL PROCEDURE

We performed downstream analysis of the complete, imputed merged dataset to take into account the uncertainty of the imputed genotypes. We filtered data based on info score of 0.7 after looking at the distribution of markers at all possible info scores. Because of the potential for genotyping errors in SNPs and samples with low call rates, it is essential to investigate the distribution of call rates by marker and by sample and the overlap of the two. **Table 3** shows, for each marker call rate threshold, the number of SNPs dropped and the proportion of the total SNP count. **Table 4** shows the sample call rate after filtering the markers with <99% call rate. At this point, we have not excluded any samples from the merged data based on sample call rates but it is very important to keep that in mind for any further analyses with these data.

We have also investigated the distribution of SNPs at different MAF thresholds. We expected that imputing using the 1000 Genomes reference panel will result in a high proportion of low frequency variants. **Table 5** shows the number of SNPs below and above each threshold. This summary table can be used for deciding what MAF threshold to use for association analyses. Based on power calculations, one can determine at what MAF the dataset is sufficiently powered. Subsequently, the MAF threshold can be used as a filter for analysis. We have also illustrated MAF as a filter after using a SNP call rate filter of 99%. As expected, the greater majority of the dataset consists of variants with MAF $< 5\%$.

In **Figure 5** we summarize all of the "Best Practices" steps and measures for imputed data prior to using the data in any further analyses. We provided a final quality control (QC) dataset filtered

Table 6 | SNP summary for samples from adults participants of the eMERGE network.

Chromosome	Imputation output	Filter at info score 0.7
1	2992265	857604
2	3292685	934303
3	2751021	807422
4	2725555	804138
5	2519463	725837
6	2414293	760529
7	2205621	633948
8	2174126	625724
9	1645320	479658
10	1874401	572475
11	1885432	553047
12	1818431	531244
13	1367340	414471
14	1251729	365975
15	1125278	312685
16	1204600	325272
17	1039660	276340
18	1083944	312821
19	810927	224571
20	851007	242258
21	515507	149262
22	491574	141941
Totals	38040179	11051525

"Imputation output" lists number of SNPs as result of imputation and "Filter at info score 0.7" lists number of SNPs passing info score threshold.

Table 7 | SNP summary for samples from pediatric participants of the eMERGE network.

Chromosome	Imputation output	Filter at info score 0.7
1	2992265	1323149
2	3292686	1363591
3	2751022	1234814
4	2725555	1264290
5	2519464	1158692
6	2414294	1157434
7	2205622	990787
8	2174126	994888
9	1645320	740663
10	1874401	869202
11	1885432	851466
12	1818431	826132
13	1367340	640670
14	1251729	562490
15	1125278	491576
16	1204601	496824
17	1039661	411860
18	1083944	490988
19	810927	311239
20	851007	369387
21	515507	225258
22	491574	200770
Totals	38040186	16976170

"Imputation output" lists number of SNPs as result of imputation and "Filter at info score 0.7" lists number of SNPs passing info score threshold.

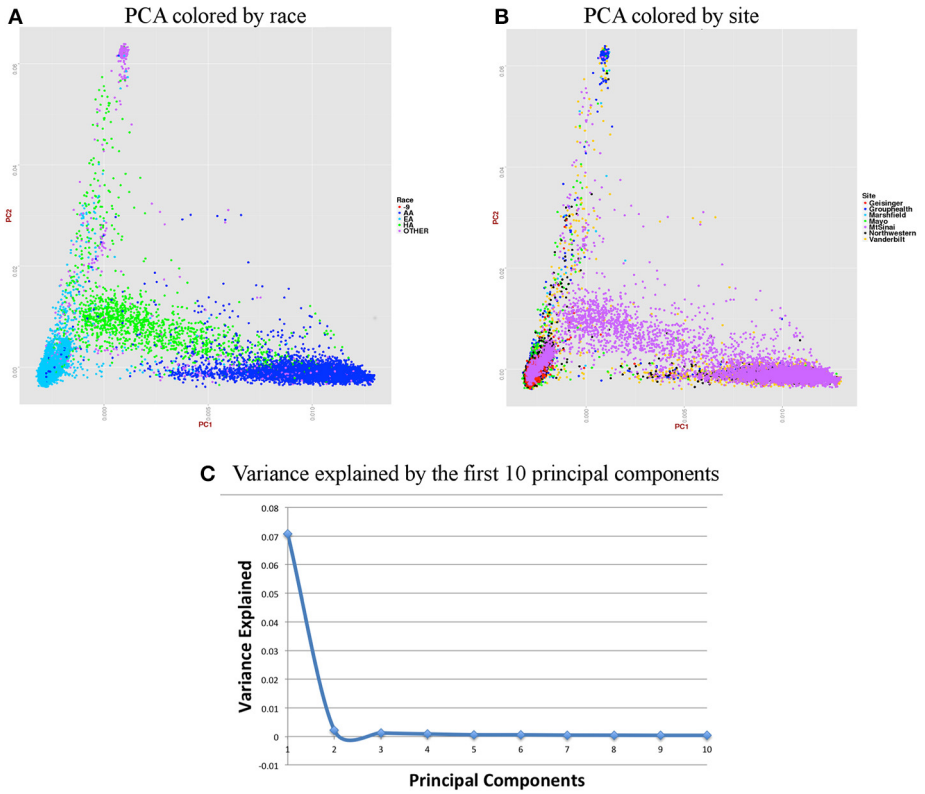


FIGURE 6 | Summary on principal component (PC) analysis for adult DNA samples. (A) PC1 and PC2 colored by self-reported race (AA, African American; EA, European American; HA, Hispanic, Others and -9, missing), **(B)** PC1 and PC2 colored by site, **(C)** Variance explained by first 10 PCs.

at info score = 0.7 and marker call rate = 99%. We did not apply any sample call rate, and MAF filter as that depends on the type of analysis being performed. **Tables 6**, **7** show total counts of SNPs at each threshold we used during quality control for both the adults and pediatric datasets.

POPULATION STRUCTURE

For accurate imputations, it is important that the samples from imputed data cluster closely to the reference panel. We performed Principal Component Analysis (PCA) as it has been shown to reliably detect differences between populations (Novembre and Stephens, 2008). Population stratification can inflate identity-by-descent (IBD) estimates; thus, we used the KING program which is designed to circumvent the inflation of IBD estimates due to stratification (Manichaikul et al., 2010).We used a kinship coefficient threshold of 0.125 (second degree relatives) to identify clusters of close relatives, and we retained only one subject from each relative cluster. We used R package SNPRelate (Zheng et al., 2012) to carry out principal components analysis (PCA), which is a form of projection pursuit capture, because it is computationally efficient, and can be parallelized easily. Principal components (PCs) were constructed to represent axes of genetic variation across all samples in unrelated adult and pediatric datasets that were pruned using the "indep-pairwise" option in PLINK (Purcell et al., 2007) such that all SNPs within a given window size of 100 had pairwise $r^2 < 0.1$ (for adults) and 0.4 (for pediatric) and also only included very common autosomal SNPs (MAF > 10%). We pruned data to reduce the number of markers to approximately 100,000 as previous studies have shown that 100,000 markers not in LD can detect ancestral information correctly (Price et al., 2006). These 100,000 markers included both imputed and genotyped SNPs, as the number of SNPs of overlap across the different genotyping platforms was too small to use only genotyped variants. It has been shown that PCA is most effective when the dataset includes unrelated individuals, low LD, and common variants (Zou et al., 2010; Zhang et al., 2013). We calculated up to 32 PCs, but show only the results for up to the first 10 PCs in scree plots represented in **Figures 6C**, **7C**. It can be noted from these figures that only the first two PCs explain all of the appreciable variance and the other PCs explain very little of the variance.

For the merged imputed adult data, we removed all related individuals (IBD estimation done using KING (Manichaikul et al., 2010) kinship >0.125), performed QC, LD pruned with $r^2 < 0.1$ and MAF > 10% to include only common variants. Thus, PCA included 37,972 samples and 1,948,089 markers. **Figure 6** presents plots for PCs 1 and 2 colored by race and eMERGE site. Population structure is very well evident from these PC plots and it shows the ancestral distributions of the data from all of the eMERGE sites.

For the pediatric data, we removed all related individuals (IBD estimation done using KING[18] kinship >0.125), performed QC, LD pruned with $r^2 < 0.4$ and MAF > 10% to include

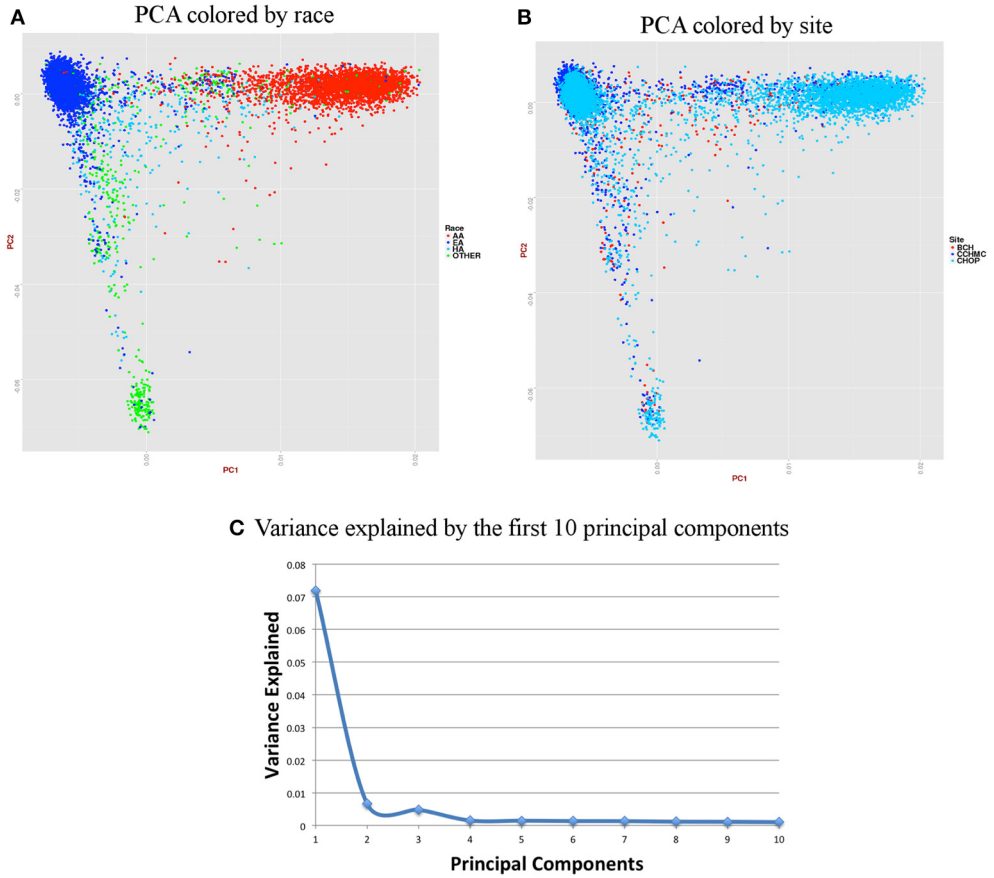


FIGURE 7 | Summary on principal component (PC) analysis for pediatric DNA samples. (A) PC1 and PC2 colored by self-reported race (AA, African American; EA, European American; HA, Hispanic and Others), **(B)** PC1 and PC2 colored by site, **(C)** Variance explained by first 10 PCs.

only common variants. Thus, PCA included 11,798 samples and 162,576 markers.

Figures 6, **7** represent plots for the first two PCs colored by self-reported race or ethnicity and also represent variance explained by the first 10 PCs for both adult and pediatric datasets. Detailed PCA results on the merged eMERGE dataset are described in another publication by the eMERGE Network investigators (Crosslin et al., 2014).

DISCUSSION

We have performed genotype imputation to facilitate the merging of data from all eMERGE datasets. We imputed using the cosmopolitan 1000 Genomes Project reference panel and IMPUTE2 software (after a comparison with BEAGLE software). We also performed initial QC steps after merging the datasets to assess the quality and accuracy of the imputed data. Imputation results appear to be very accurate, based on the high concordance rates in the masked analysis. In addition, there was a clear distinction between the different ancestral groups, as expected, based on the PC analysis. It is very difficult to merge all of the genotype data from different platforms together prior to imputation, as a strategy to perform imputation, due to lack of sufficient overlapping markers between different genotyping platforms. Therefore, our pipeline performs imputations separately on each platform and origin of the genotype data, and then we merged the data together. We obtained very good results using this strategy and therefore consider it is an appropriate approach. It allows for the maximization of the number of genotyped markers available as study SNPs to use as the backbone to initiate imputation. It is suggested to remove all palindromic SNPs from the dataset before running any imputations to future pipelines. We performed a test on two of our datasets, running the imputation both before and after removing palindromic SNPs. Concordance check between the two runs of imputations revealed that the results were 99.8% concordant. More exploration of this issue is important for future work.

This manuscript is meant to serve as an applied, educational resource and to provide guidance for imputation. There are a number of other reviews and comparisons of different imputation packages available (Pei et al., 2008; Ellinghaus et al., 2009; Nothnagel et al., 2009; Hancock et al., 2012; Comparing BEAGLE, IMPUTE2, and Minimac Imputation Methods for Accuracy, Computation Time, and Memory Usage | Our 2 SNPs…®). The imputed genotypes, phenotype information, accompanying marker annotation and quality metrics files for these eMERGE data will be available through the authorized

access portion of the dbGaP (http://www.ncbi.nlm.nih.gov/gap). Numerous references are accessible for users wanting additional information on imputation methods, as well as recommendations for downstream analyses (Marchini et al., 2007; Servin and Stephens, 2007; Browning, 2008; Guan and Stephens, 2008; Li et al., 2009; Aulchenko et al., 2010; International HapMap 3 Consortium et al., 2010; Hancock et al., 2012; Nelson et al., 2012).

ACKNOWLEDGMENTS

The eMERGE Network is funded by NHGRI, with additional funding from NIGMS through the following grants: U01HG004438 to Johns Hopkins University; U01HG004424 to The Broad Institute; U01HG004438 to CIDR; U01HG004610 and U01HG006375 to Group Health Cooperative; U01HG004608 to Marshfield Clinic; U01HG006389 to Essentia Institute of Rural Health; U01HG04599 and U01HG006379 to Mayo Clinic; U01HG004609 and U01HG006388 to Northwestern University; U01HG04603 and U01HG006378 to Vanderbilt University; U01HG006385 to the Coordinating Center; U01HG006382 to Geisinger Clinic; U01HG006380 to Mount Sinai School of Medicine; U01HG006830 to The Children's Hospital of Philadelphia; and U01HG006828 to Cincinnati Children's Hospital and Boston Children's Hospital. We would like to give special thanks to Sarah Nelson and her colleagues at the University of Washington and the GENEVA consortium for guidance and input as we started our imputation process. We would like to thank the members of the eMERGE Genomics Workgroup who participated in weekly phone calls discussing this project and results. We would also like to thank the staffs of the Research Computing Center at the Pennsylvania State University, who were extremely helpful troubleshooting and enabling massive compute cluster usage to complete these imputations.

REFERENCES

(2010). E pluribus unum. *Nat. Methods* 7, 331–331. doi: 10.1038/nmth0510-331

Aulchenko, Y. S., Struchalin, M. V., and van Duijn, C. M. (2010). ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* 11:134. doi: 10.1186/1471-2105-11-134

Browning, B. L., and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84, 210–223. doi: 10.1016/j.ajhg.2009.01.005

Browning, S. R. (2008). Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum. Genet.* 124, 439–450. doi: 10.1007/s00439-008-0568-7

Crosslin, D. R., Tromp, G., Burt, A., Kim, D. S., Verma, S. S., Lucas, A. M., et al. (2014). Controlling for population structure and genotyping platform bias in the eMERGE multi-institutional biobank linked to Electronic Health Records. *Front. Genet.* 5:352. doi: 10.3389/fgene.2014.00352

Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* 10, 5–6. doi: 10.1038/nmeth.2307

Ellinghaus, D., Schreiber, S., Franke, A., and Nothnagel, M. (2009). Current software for genotype imputation. *Hum. Genomics* 3, 371–380.

1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65. doi: 10.1038/nature11632

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., et al. (2013). The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genet. Med.* 15, 761–771. doi: 10.1038/gim.2013.72

Guan, Y., and Stephens, M. (2008). Practical issues in imputation-based association mapping. *PLoS Genet.* 4:e1000279. doi: 10.1371/journal.pgen.1000279

Hancock, D. B., Levy, J. L., Gaddis, N. C., Bierut, L. J., Saccone, N. L., Page, G. P., et al. (2012). Assessment of genotype imputation performance using 1000 Genomes in African American studies. *PLoS ONE* 7:e50610. doi: 10.1371/journal.pone.0050610

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* 44, 955–959. doi: 10.1038/ng.2354

Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3 (Bethesda)* 1, 457–470. doi: 10.1534/g3.111.001198

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5:e1000529. doi: 10.1371/journal.pgen.1000529

International HapMap 3 Consortium, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–58. doi: 10.1038/nature09298

Jostins, L., Morley, K. I., and Barrett, J. C. (2011). Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur. J. Hum. Genet.* 19, 662–666. doi: 10.1038/ejhg.2011.10

Karolchik, D., Hinrichs, A. S., and Kent, W. J. (2011). The UCSC genome browser. *Curr. Protoc. Hum. Genet.* Chapter, Unit18.6. doi: 10.1002/0471142905.hg1806s71

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* 34, 816–834. doi: 10.1002/gepi.20533

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* 10, 387–406. doi: 10.1146/annurev.genom.9.081307.164242

Lin, P., Hartz, S. M., Zhang, Z., Saccone, S. F., Wang, J., Tischfield, J. A., et al. (2010). A new statistic to evaluate imputation reliability. *PLoS ONE* 5:e9697. doi: 10.1371/journal.pone.0009697

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. doi: 10.1093/bioinformatics/btq559

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 39, 906–913. doi: 10.1038/ng2088

McCarty, C. A., Chisholm, R. L., Chute, C. G., Kullo, I. J., Jarvik, G. P., Larson, E. B., et al. (2011). The eMERGE network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics* 4:13. doi: 10.1186/1755-8794-4-13

Nelson, S. C., Doheny, K. F., Laurie, C. C., and Mirel, D. B. (2012). Is "forward" the same as "plus"? and other adventures in SNP allele nomenclature. *Trends Genet.* 28, 361–363. doi: 10.1016/j.tig.2012.05.002

Nelson, S. C., Doheny, K. F., Pugh, E. W., Romm, J. M., Ling, H., Laurie, C. A., et al. (2013). Imputation-based genomic coverage assessments of current human genotyping arrays. *G3 (Bethesda)* 3, 1795–1807. doi: 10.1534/g3.113.007161

Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M., and Franke, A. (2009). A comprehensive evaluation of SNP genotype imputation. *Hum. Genet.* 125, 163–171. doi: 10.1007/s00439-008-0606-5

Novembre, J., and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* 40, 646–649. doi: 10.1038/ng.139

Overby, C. L., Kohane, I., Kannry, J. L., Williams, M. S., Starren, J., Bottinger, E., et al. (2013). Opportunities for genomic clinical decision support interventions. *Genet. Med.* 15, 817–823. doi: 10.1038/gim.2013.128

Pei, Y.-F., Li, J., Zhang, L., Papasian, C. J., and Deng, H.-W. (2008). Analyses and comparison of accuracy of different genotype imputation methods. *PLoS ONE* 3:e3551. doi: 10.1371/journal.pone.0003551

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. doi: 10.1038/ng1847

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and

population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Servin, B., and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* 3:e114. doi: 10.1371/journal.pgen.0030114

Southam, L., Panoutsopoulou, K., Rayner, N. W., Chapman, K., Durrant, C., Ferreira, T., et al. (2011). The effect of genome-wide association scan quality control on imputation outcome for common variants. *Eur. J. Hum. Genet.* 19, 610–614. doi: 10.1038/ejhg.2010.242

Verma, S., Peissig, P., Cross, D., Waudby, C., Brilliant, M. H., McCarty, C. A., et al. (2014). *Benefits of Accurate Imputations in GWAS. LNCS 8602*. Granada, 877–889.

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., et al. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* 8:e1002793. doi: 10.1371/journal.pgen.1002793

Zhang, Y., Guan, W., and Pan, W. (2013). Adjustment for population stratification via principal components in association analysis of rare variants. *Genet. Epidemiol.* 37, 99–109. doi: 10.1002/gepi.21691

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28, 3326–3328. doi: 10.1093/bioinformatics/bts606

Zou, F., Lee, S., Knowles, M. R., and Wright, F. A. (2010). Quantification of population structure using correlated SNPs by shrinkage principal components. *Hum. Hered.* 70, 9–22. doi: 10.1159/000288706

Zuvich, R. L., Armstrong, L. L., Bielinski, S. J., Bradford, Y., Carlson, C. S., Crawford, D. C., et al. (2011). Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality. *Genet. Epidemiol.* 35, 887–898. doi: 10.1002/gepi.20639

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 April 2014; accepted: 03 October 2014; published online: 11 December 2014.

Citation: Verma SS, de Andrade M, Tromp G, Kuivaniemi H, Pugh E, Namjou-Khales B, Mukherjee S, Jarvik GP, Kottyan LC, Burt A, Bradford Y, Armstrong GD, Derr K, Crawford DC, Haines JL, Li R, Crosslin D and Ritchie MD (2014) Imputation and quality control steps for combining multiple genome-wide datasets. Front. Genet. ***5****:370. doi: 10.3389/fgene.2014.00370*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 04 November 2014
doi: 10.3389/fgene.2014.00352

Controlling for population structure and genotyping platform bias in the eMERGE multi-institutional biobank linked to electronic health records

David R. Crosslin[1,2*†], Gerard Tromp[3†], Amber Burt[2], Daniel S. Kim[1,2], Shefali S. Verma[4], Anastasia M. Lucas[4], Yuki Bradford[4], Dana C. Crawford[5,6], Sebastian M. Armasu[7], John A. Heit[8], M. Geoffrey Hayes[9], Helena Kuivaniemi[3], Marylyn D. Ritchie[4], Gail P. Jarvik[1,2], Mariza de Andrade[7] and the electronic Medical Records and Genomics (eMERGE) Network

[1] Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA, USA
[2] Department of Genome Sciences, University of Washington, Seattle, WA, USA
[3] The Sigfried and Janet Weis Center for Research, Geisinger Health System, Danville, PA, USA
[4] Department of Biochemistry and Molecular Biology, Center for Systems Genomics, Pennsylvania State University, University Park, PA, USA
[5] Center for Human Genetics Research, School of Medicine, Vanderbilt University, Nashville, TN, USA
[6] Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN, USA
[7] Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA
[8] Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN, USA
[9] Division of Endocrinology, Metabolism, and Molecular Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

Edited by:
Karen T. Cuenco, Genentech, USA

Reviewed by:
Alexis C. Frazier-Wood, University of Alabama at Birmingham, USA
Tesfaye B. Mersha, Cincinnati Children's Hospital Medical Center, USA

****Correspondence:***
David R. Crosslin, Division of Medical Genetics, Department of Medicine, University of Washington, 1705 NE Pacific Street, HSB, J309A, Box 357720, Seattle, WA 98195, USA
e-mail: david.crosslin@gmail.com

† These authors have contributed equally to this work.

Combining samples across multiple cohorts in large-scale scientific research programs is often required to achieve the necessary power for genome-wide association studies. Controlling for genomic ancestry through principal component analysis (PCA) to address the effect of population stratification is a common practice. In addition to local genomic variation, such as copy number variation and inversions, other factors directly related to combining multiple studies, such as platform and site recruitment bias, can drive the correlation patterns in PCA. In this report, we describe the combination and analysis of multi-ethnic cohort with biobanks linked to electronic health records for large-scale genomic association discovery analyses. First, we outline the observed site and platform bias, in addition to ancestry differences. Second, we outline a general protocol for selecting variants for input into the subject variance-covariance matrix, the conventional PCA approach. Finally, we introduce an alternative approach to PCA by deriving components from subject loadings calculated from a reference sample. This alternative approach of generating principal components controlled for site and platform bias, in addition to ancestry differences, has the advantage of fewer covariates and degrees of freedom.

Keywords: principal component analysis, ancestry, biobank, loadings, genetic association study

1. INTRODUCTION

To reach the statistical power needed for genome-wide association studies, large numbers of participants are needed. This can be achieved through large research networks such as the Electronic Medical Records and Genomics (eMERGE) Network, which comprises a multi-ethnic cohort of ∼57,000 participants linked to electronic health records (EHRs) for phenotype mining from nine participating sites (seven adult; two pediatric) in the United States (U.S.) (Gottesman et al., 2013). When combining genetic data from diverse data sets, understanding the contribution of ancestry, genotyping platform, and site bias are of vital importance.

Through the course of the eMERGE project, multiple genotyping platforms from both Illumina and Affymetrix were utilized (Gottesman et al., 2013; Crawford et al., 2014). Imputation using the BEAGLE software was then carried out to allow merging of the diverse data sets (Verma et al., *Imputation and quality control steps for combining multiple genome-wide data sets*. Manuscript submitted for publication).

There were ancestry or racial/ethnic differences both within and across the eMERGE Network sites in addition to the platform heterogeneity. The majority of eMERGE study sites based race/ethnicity on self-report while Vanderbilt University's BioVU used third-party or administratively assigned race/ethnicity (Dumitrescu et al., 2010). The major group for the entire eMERGE sample set is of European-descent. eMERGE also includes a sizeable African-descent and Hispanic sample (Gottesman et al., 2013). The latter represents a three-way admixture event (Manichaikul et al., 2012) that further contributes to expected ancestral differences within and across eMERGE. There are also both cryptic and known related participants, especially in Marshfield Clinic Research Foundation (Gottesman et al., 2013; Crawford et al., 2014).

We present an example of integrating the diverse genetic data sets from the eMERGE Network in a systematic fashion and provide guidance for other investigators in large research networks. We outline a general approach for selecting variants for input into a sample variance-covariance matrix on the adult participants in eMERGE, the conventional principal component analysis (PCA) approach in human genetics research (Patterson et al., 2006). We also describe how we categorized genetic ancestry based on self-reported race, framed in terms of continental origin, in line with standard protocol in human genetic research (NHGRI, 2005; Ali-Khan et al., 2011).

Given our "sizeable" non-European sample in the presence of platform bias and imputation, the eMERGE Network took great care in not only assessing and adjusting for ancestry, but also exploring alternative methods to do so and increase power. To assess ancestry in related individuals, Zhu et al. (2008) introduced a method of generating principal components (PCs) by deriving SNP loadings from founders, and applying them to the entire sample. We introduce this concept of deriving SNP loadings from the BEAGLE imputation 1000 Genomes reference sample, and apply it to the entire imputed sample set of 57,000 genotyped individuals from the eMERGE Network as an alternative approach to control for site and platform bias in addition to ancestry differences for our large cohort.

2. MATERIALS AND METHODS

The eMERGE Network comprises a multi-ethnic cohort of ∼57,000 participants linked to EHRs for phenotype mining from nine participating sites (seven adult; two pediatric) in the United States (Gottesman et al., 2013) with genotype and imputed data.

2.1. IMPUTATION

The imputation and merging were performed by the eMERGE Coordinating Center (CC) at Pennsylvania State University (PSU). Detailed quality assurance /quality control (QA/QC) measures are outlined in the imputation guide provided on the PSU eMERGE CC web site (see Web Resources). Before imputation, study site data were converted to the same build (Build 37) as the imputation reference data set. Next, strand flipping was employed to account for different strand alignments including Illumina TOP/BOT strand, plus(+) / minus(−), and forward/reverse (Nelson et al., 2012). Finally, phasing and imputation were performed on randomized ancestry sub-samples against a "Cosmopolitan" reference set from the 1000 Genomes containing multiple ancestry groups provided by the BEAGLE software package (Browning and Browning, 2009). While the imputation data presented are derived from using BEAGLE software (Browning and Browning, 2009), it should be noted that IMPUTE2 software (Howie et al., 2012) produced nearly identical results (see Supplementary Figure S1) (Howie et al., 2011; Delaneau et al., 2013).

2.2. PCA

There are multiple software packages for running PCA to estimate genomic ancestry, but we utilized the high-performance computing toolset SNPRelate R package (Zheng et al., 2012) for multiple reasons. First, the increased computational performance allows for PCA analyses of a large number of participants such as eMERGE. Second, this tool allows the extraction of both sample and SNP loadings, which allows the correction of population stratification for related and unrelated participants (Zhu et al., 2008). The two types of matrices are mathematically equivalent and can be derived from one another. Finally, SNPRelate allows for absolute genotype-PC correlation to assess whether a local region of the genome is driving the correlation structure (Zheng et al., 2012).

We derived PCs using three general approaches, each applied to the overall set and to each ancestry group. First, we performed PCA on a combined data set (across sites) after imputation using the BEAGLE software package (Version 3.3.1) (Browning and Browning, 2009). Second, we performed PCA on a pre-imputed merged version (across sites) of the data. Finally, we derived PCs for the entire set using SNP loadings generated from the BEAGLE imputation reference set (Browning and Browning, 2009).

For all genotype data used to generate the variance-covariance matrices and to eliminate redundant SNPs in high linkage disequilibrium (LD), we applied the following thresholds. The autosomal variants were selected after LD pruning at $r > 0.5$ with a 500 kbp (kilo basepairs) sliding window, and a minor allele frequency (MAF) > 0.05. In addition, a variant missingness filter of 0.02 was applied. For both PCA on the combined imputed and the combined preimputed, which is basically the singular value decomposition on the sample covariance matrix as outlined in Patterson et al. (2006).

2.2.1. Deriving PCA using reference sample loadings

We also assessed PCA using the Zhu et al. (2008) method by deriving SNP loadings from the BEAGLE imputation 1000 Genomes reference sample, and applying it to the entire sample set. As such, we utilized their nomenclature with respect to generating the components. This was implemented using the SNPRelate R package (Zheng et al., 2012), specifically the `snpgdsPCASampLoading` and `snpgdsPCASNPLoading` functions (see Web Resources).

We treated the entire eMERGE cohort as one "related" family, and the imputation reference sample as $(a = 1, 2, \ldots, B)$ unrelated. Because of this, the g_{ij} marker genotype value of the jth individual in the ith family as utilized by Zhu et al. (2008), simplified to g_j. The column vector $X_{ij} = (x_{j1}, x_{j2}, \ldots, x_{jM})^T$ of $l = 1, 2, \ldots, M$ biallelic markers, and was coded as an additive model of inheritance.

The variance-covariance matrix for the marker data from the reference sample (unrelated), took on the form $\mathbf{\Sigma} = \Sigma_{a=1}^{B}(X_a - \overline{X})(X_a - \overline{X})^T$, assuming $\overline{X}$ as the overall genotype mean for those samples. Following Zhu et al. (2008), we let e_l be the lth eigenvalue of $\mathbf{\Sigma}$, where $l = 1, 2, \ldots, M$, which is a vector of the SNP loadings. We then derived the lth PC for the individual (j) of the entire cohort or "related" family by $t_{jl} = (X_j - \overline{X})^T e_l$.

2.3. VENOUS THROMBOEMBOLISM ASSOCIATION

The venous thromboembolism (VTE) phenotype was extracted using an EHR-driven algorithm from African ancestry participants (Pathak et al., personal communication), excluding

patients with cancer. A total of 400 VTE cases and 5,065 controls were selected from 4 sites and 4 different genotype platforms (Illumina 660, 1M, and Omni; and Affymetrix 6.0). We performed two logistic regressions for association using the software PLINK v1.07 (Purcell et al., 2007). The first was adjusted for age, sex, stroke, sickle cell genetic variant, site-platform, and conventional PC1 and PC2 and the second was adjusted for age, sex, stroke, sickle cell genetic variant and "loadings" PC1 and PC2.

3. RESULTS

3.1. DEMOGRAPHICS

Table 1 outlines the breakdown of the 38,288 adult participants included in these analyses by eMERGE site, self-reported or administratively assigned ancestry, sex, and genotyping platform. Most sites were predominantly of European ancestry. Compared with most other eMERGE study sites, both Vanderbilt University and Northwestern University had a greater representation of African ancestry (26 and 12%, respectively). Mount Sinai School of Medicine had the greatest proportion of African ancestry (70%), followed by a sizeable proportion of Hispanic participants (19%). Overall, there were more females than males (57% vs. 43%). All sites followed this pattern, except for Geisinger Health System (53% male). Most of the genotyping across all sites was performed using Illumina arrays (610, 660, 1M and Omni), with the exception of Mount Sinai School of Medicine (Affymetrix 6.0).

Eigenvectors 1 and 2 for the 38,288 adult eMERGE participants are illustrated in **Figure 1**, annotated by self-reported race (**Figure 1A**), genotyping platform (**Figure 1B**), and by eMERGE study site (**Figure 1C**). Genetically determined ancestry was assigned by creating subjective boundaries for the African, European and Hispanic groups. These boundaries were set using the respective medians (Q_2) and standard deviations (SD) calculated for each genetic ancestry group, as illustrated in **Figures 2A–C** for the African ($Q_{2_A} \pm 2SD$), European ($Q_{2_E} \pm 4SD$) and Hispanic ($Q_{2_H} \pm 1SD$) groups, respectively.

3.2. EXAMINATION OF THE VARIANCE EXPLAINED PER PC USING SCREE PLOTS

To assess the variance explained from each PCA, we plotted the first ten PCs against the variance explained as illustrated in **Figure 3**. Across the columns of the trellis we show scree plots of joint, African ancestry, European ancestry, and Hispanic groups. Across each row, we have scree plots representing PC analyses on the imputed merged set, pre-imputed merged set, and on the PC analyses using the "loadings" method outlined in Subsection 2.2.1. As expected, eigenvector 1 explains most of the variance for the joint ancestry imputed (∼7%), pre-imputed (∼4%), and "loadings" (∼7%). When we stratified by ancestry (across the trellis), the variance explained by eigenvector 1 for the imputed and pre-imputed data sets was <1%. For the "loadings" approach with the African and European genetic ancestry data sets, the variance explained <1%, and >2% for the Hispanic group. In all scenarios (joint and all ancestry groups) the variance explained approached 0 for eigenvectors 2 through 10 for the imputed and pre-imputed data sets. Interestingly, the "loadings" approach allows for more variance explained for eigenvectors 2 and beyond, especially for the Hispanics. For the joint loadings approach, the variance explained by eigenvector 2 approached ∼4%, while the genetic ancestry groups approached 1%.

3.3. EVALUATION OF THE EFFECT OF ANCESTRY ON PC PLOTS—JOINT AND STRATIFIED ANCESTRY

We evaluated the population structure by plotting eigenvectors 1 and 2 for the joint data set (**Figure 4**) as well as for the African (**Figure 5**), European (**Figure 6**) and Hispanic (**Figure 7**) ancestry

Table 1 | Summary of eMERGE sample by self-reported ancestry, sex, and genotyping platform for the adult participants.

	Geisinger ($N = 3,111$) (%)	Group Health ($N = 3,520$) (%)	Marshfield ($N = 4,193$) (%)	Mayo ($N = 6,836$) (%)	Mt. Sinai ($N = 6,290$) (%)	Northwestern ($N = 4,858$) (%)	Vanderbilt ($N = 9,480$) (%)	Combined ($N = 38,288$)
SELF-REPORTED ANCESTRY								
African	0	4	0	0	70	12	26[†]	20% (7, 651)
European	99	92	99	99	11	88	66[†]	74% (28, 469)
Hispanic	0	0	0	0	19	0	0	3% (1, 258)
Other	0	5	1	0	0	0	7[†]	2% (910)
SEX								
Female	47	57	58	45	59	83	53	57% (21, 802)
Male	53	43	41	55	41	17	47	43% (16, 486)
GENOTYPING PLATFORM								
Affymetrix 6	0	0	0	0	44	0	0	7% (2, 775)
Illumina 1M	0	0	0	0	0	12	21	7% (2, 634)
Illumina 610	0	0	0	45	0	0	0	8% (3, 081)
Illumina 660	0	89	100	55	0	27	42	43% (16, 362)
Illumina Omni	100	11	0	0	56	61	37	35% (13, 436)

[†]*Race/ethnicity is administratively assigned.*

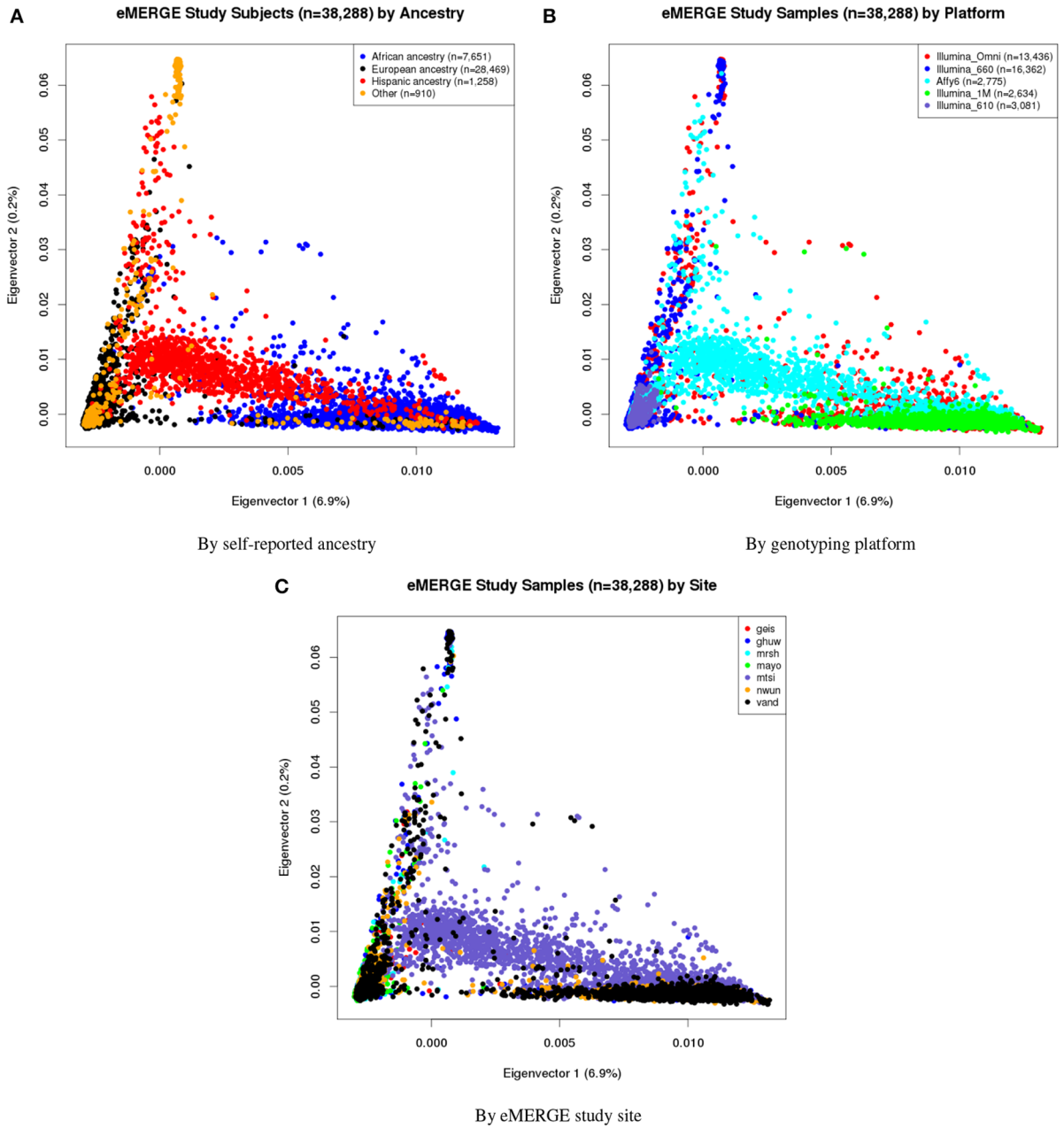


FIGURE 1 | PC plots of PCs 1 and 2 for all adults of eMERGE by self-reported race. (A), genotyping platform (B), and eMERGE study site (C), using BEAGLE imputed data. (1) geis, Geisinger Health System, (2) ghuw, Group Health Research Institute/University of Washington; (3) mrsh, Marshfield Clinic Research Foundation; (4) mayo, Mayo Clinic; (5) mtsi, Mount Sinai School of Medicine; (6) nwun, Northwestern University; and (7) vand, Vanderbilt University.

groups, separately. In each case of ancestry analysis, we plotted the imputed and pre-imputed merged data set, and the data set derived from the "loadings" method. **Figures 4A,B** illustrate the imputation and pre-imputation data sets, respectively, and are generally opposites with respect to eigenvector 1 due to different projections for that component. **Figure 4C** illustrates the "loadings" data set, which offers a different characterization of the joint data set, with the African and European genetic ancestry groups largely represented by two ellipses.

Figures 5A–C illustrate the African ancestry imputation data set, annotated by self-reported race, genotyping platform, and site, respectively. As illustrated in **Figures 5B,C**, there are batch effects by platform and study site. The pre-imputed data set (**Figure 5D**) has two distinct bands for both eigenvectors 1 and 2. The "loadings" approach (**Figure 5E**) produces an ellipse, indicating no effect due to platform or study site. **Figures 6A–C** illustrate the European ancestry imputed and pre-imputed data set, and the "loadings" data set, respectively. Eigenvectors 1 and 2 for the

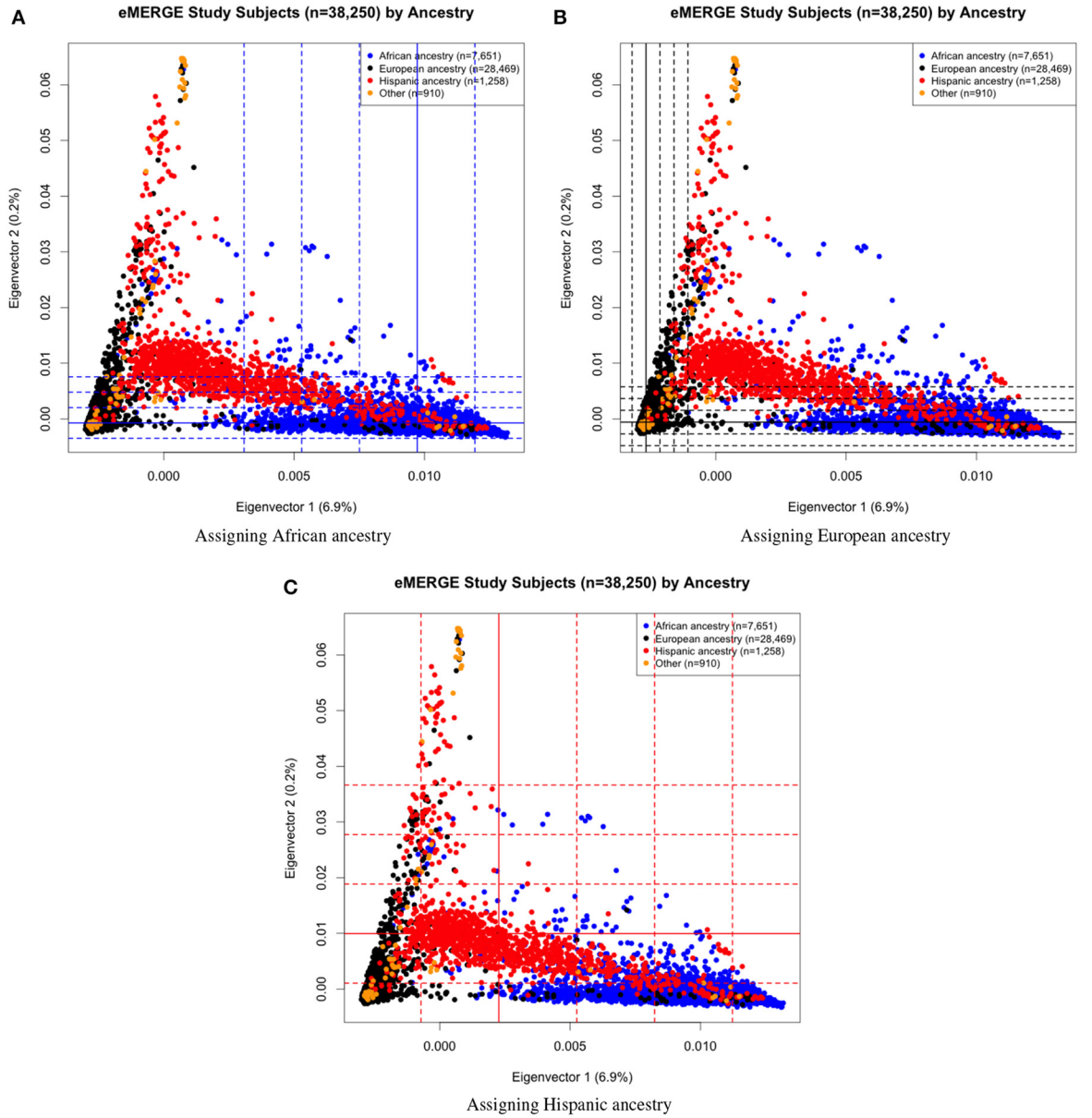


FIGURE 2 | PC plots of PCs 1 and 2 comparing eMERGE genetically determined and self-reported ancestry, using BEAGLE imputed data. (A) African ancestry assigned using ($Q_{2_A} \pm 2SD$) of eigenvectors 1 and 2 for self-reported as African ancestry. **(B)** European ancestry assigned using ($Q_{2_E} \pm 4SD$) of eigenvectors 1 and 2 for self-reported as European ancestry. **(C)** Hispanic assigned using ($Q_{2_H} \pm 1SD$) of eigenvectors 1 and 2 for self-reported as Hispanics.

imputed data set (**Figure 6A**) produce separation much like the joint ancestry analyses, while the pre-imputed data set produces two separate bands (**Figure 6B**). Like the African genetic ancestry "loadings" set, the European set produces an ellipse. Finally, the Hispanic data sets are illustrated in **Figures 7A–C**. With only 994 participants, most of the variance seems to be explained by eigenvector 1 for both the imputed (**Figure 7A**) and pre-imputed (**Figure 7B**). The "loadings" approach (**Figure 7C**) produces the familiar ellipse, with the mixed ancestry in the middle, most likely representing the Hispanic sample.

3.4. EXAMINATION OF SNP-PC CORRELATION

We also illustrate component-genotype absolute correlation plots generated using the SNPRelate R package for the imputed and pre-imputed data sets. Ideally, a component will be driven by genome-wide correlation patterns, as illustrated by eigenvector 3 of the pre-imputed data in **Figure 8A**. However, many times chromosomal artifacts will drive local regions of correlation, resulting in components dominated by that pattern. Examples of this include **Figures 8B,C**. **Figure 8B** illustrates a known chromosome 8 inversion (Feuk et al., 2006) driving the correlation patterns for

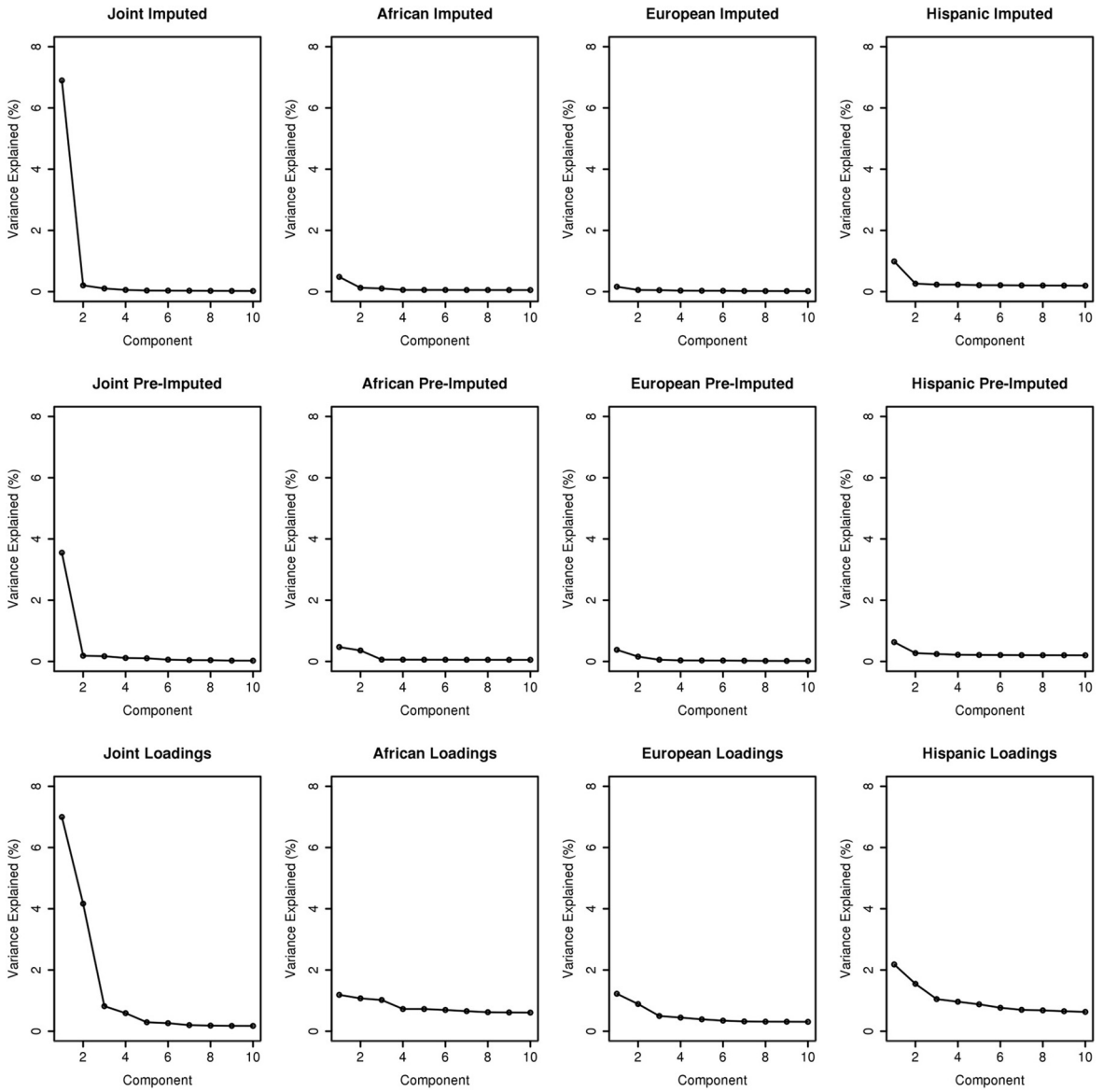


FIGURE 3 | Scree plots illustrating variance explained for PCA outlined in this manuscript.

eigenvector 9 in the imputed data. **Figure 8C** illustrates the correlation pattern driven by the HLA region for eigenvector 10 of the pre-imputed data.

3.5. VENOUS THROMBOEMBOLISM ASSOCIATION

We applied our approach using the eMERGE VTE African ancestry cohort that consists of four adult sites and four genotyping platforms that had previously been analyzed controlling for site, platform and genomic ancestry (Heit et al., 2013). For clarity, the original analysis' first two eigenvectors along with site and platform will be referred to as "PCs." The principal components derived from the imputed data set by the conventional approach will be referred as normal eigenvectors (normal "EIGs"), and derived by the "loadings" approach as "loading" eigenvectors ("loadings EIGs"). We first compared the two first PCs obtained using the eMERGE African ancestry from the original analysis with the two first eigenvectors (PCs) using the "loadings" method (**Figure 9**). We observed that the PCs used in the analysis had similar pattern as the standard eigenvectors (**Figures 9A,B**, first row), but just in a different direction for the projections. **Figure 9C** illustrates a bivariate normal distribution with low variance of the African genetic ancestry when using the "loadings" eigenvectors.

We observed dispersion between the first PC and the first "loading" eigenvector (**Figure 9D**), demonstrating that the "loadings" approach captured a different aspect of variance. The first PC showed an inverse correlation with the first PC and first normal eigenvector (**Figure 9E**). Such an inversion is a consequence of the arbitrary nature of mathematical sign in the computation of PCs resulting in opposite projections. **Figure 9F**

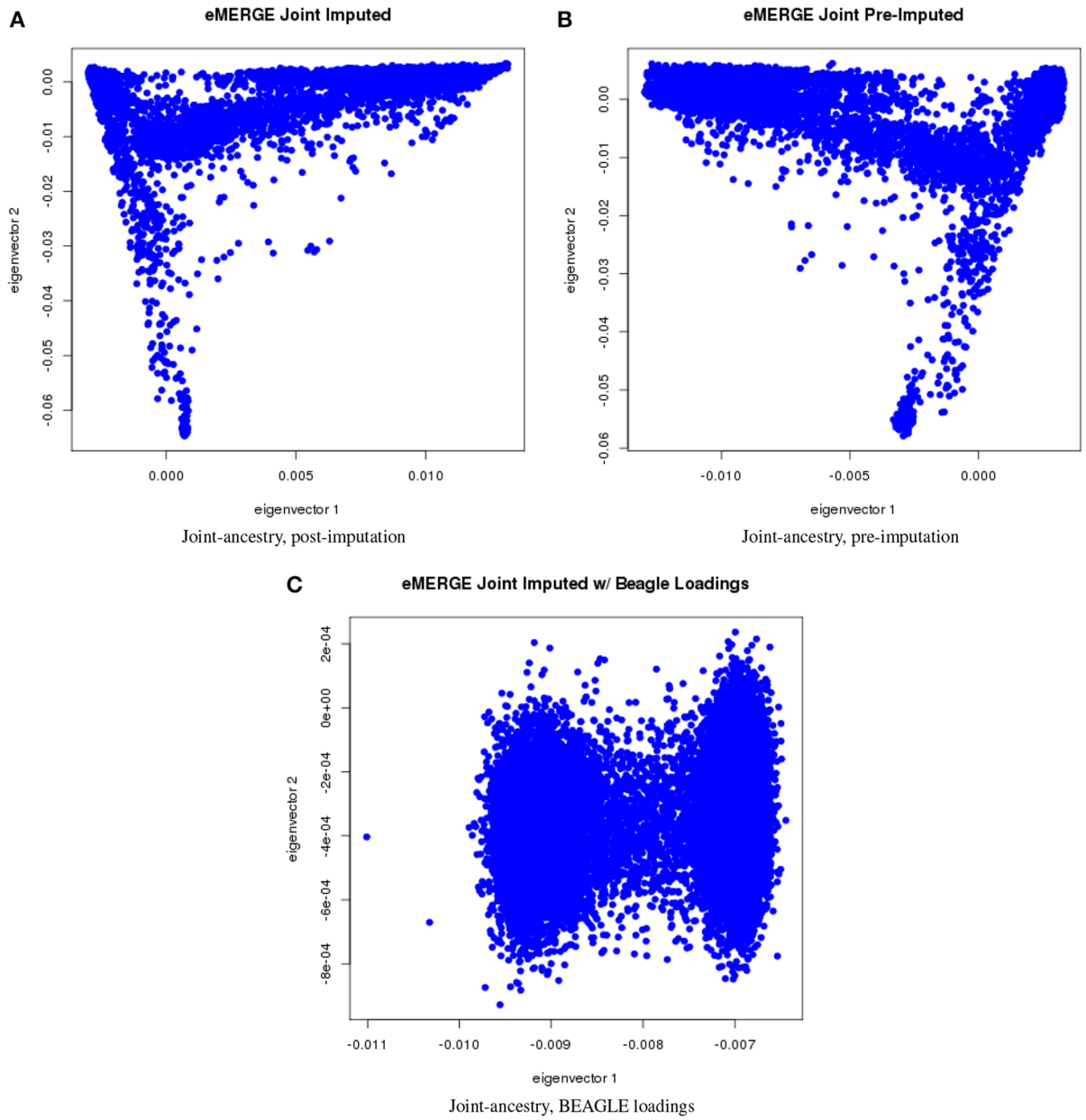


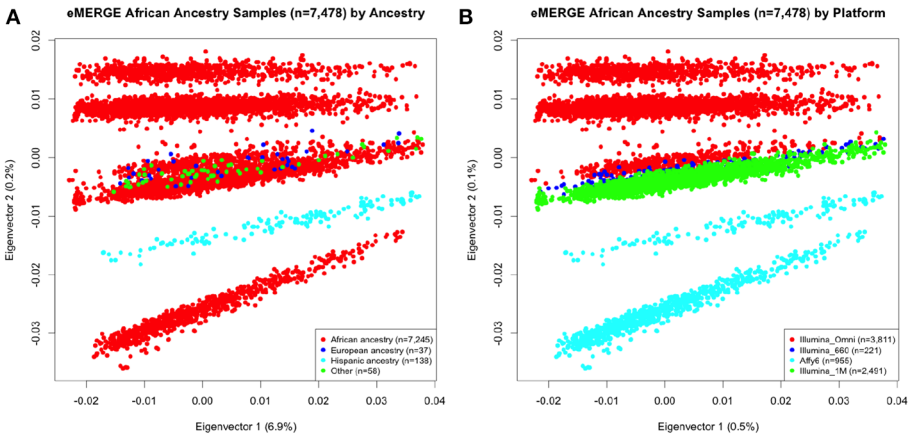
FIGURE 4 | PC plots of eMERGE joint ancestry. (A) Plot of eigenvectors 1 and 2 for the joint imputed data set. **(B)** Plot of eigenvectors 1 and 2 for the joint pre-imputed data set. **(C)** Plot of eigenvectors 1 and 2 for the joint imputed data set using the "loadings" method.

illustrates the second PC compared to the second "loadings" PC, which shows no correlation and some outliers in the PC projection.

Figure 9G depicts the comparison between the second PC with the second normal eigenvector that showed the same outliers observed previously but in a different scale. Thus, by using the BEAGLE loadings we have a more parsimonious model, and the association results in P-values and $-log_{10}(P)$ are tighter for chromosome 22 (**Figures 9H,I**). Finally, **Figures 10A,B** represent the QQ plots for the conventional PC adjusting for site and platform method ($\lambda = 1.01$) and the "loadings" approach ($\lambda = 1.02$), respectively.

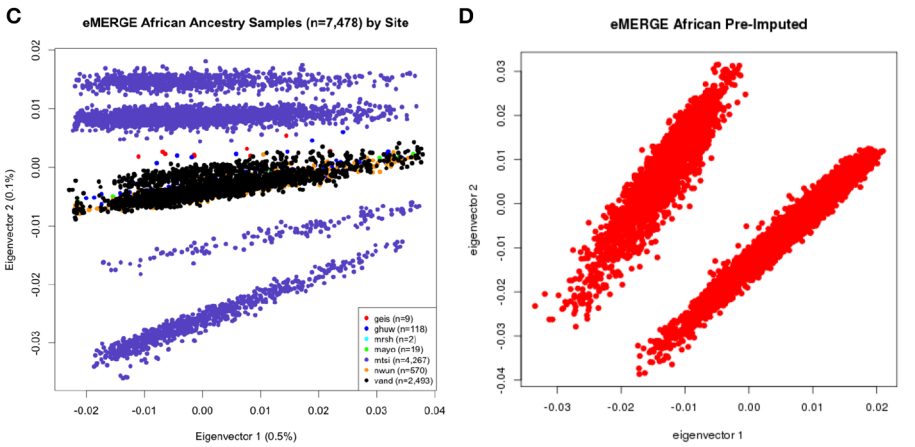
4. DISCUSSION

Imputation depends on how well the genotype data (the observed LD) capture the true underlying LD. The more completely LD is represented, the more accurately the imputation will extend the LD to non-genotyped markers. There is always an inherent risk that the imputed genotypes will not represent the true state of nature accurately; this risk increases as the genotyped density decreases and the genotypes do not capture the underlying LD. We detected effects from the genotyping platform when performing the PCA (here we use platform to indicate the design as well as the method). The effect was most evident when a low-density platform such as the MetaboChip (data not shown)



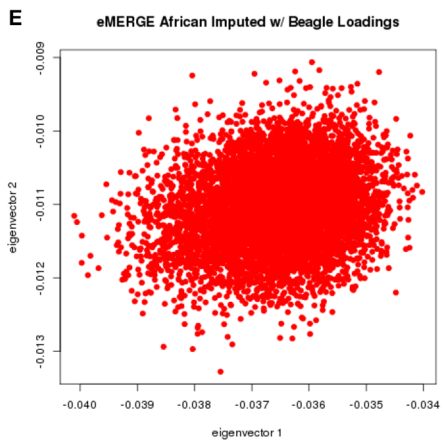
African-ancestry, post-imputation by ancestry

African-ancestry, post-imputation by platform



African-ancestry, post-imputation by study site

African-ancestry, pre-imputation



African-ancestry, BEAGLE loadings

FIGURE 5 | PC plots of eMERGE participants geneticaly determined to be of African ancestry. (A) Plot of eigenvectors 1 and 2 for the imputed data set African ancestry participants, annotated by self-reported ancestry. **(B)** Plot of eigenvectors 1 and 2 for the imputed data set African ancestry participants, annotated by genotyping platform. **(C)** Plot of eigenvectors 1 and 2 for the imputed data set African ancestry participants, annotated by eMERGE site. **(D)** Plot of eigenvectors 1 and 2 for the pre-imputed data set African ancestry participants. **(E)** Plot of eigenvectors 1 and 2 for the imputed data set African ancestry participants using the "loadings" method.

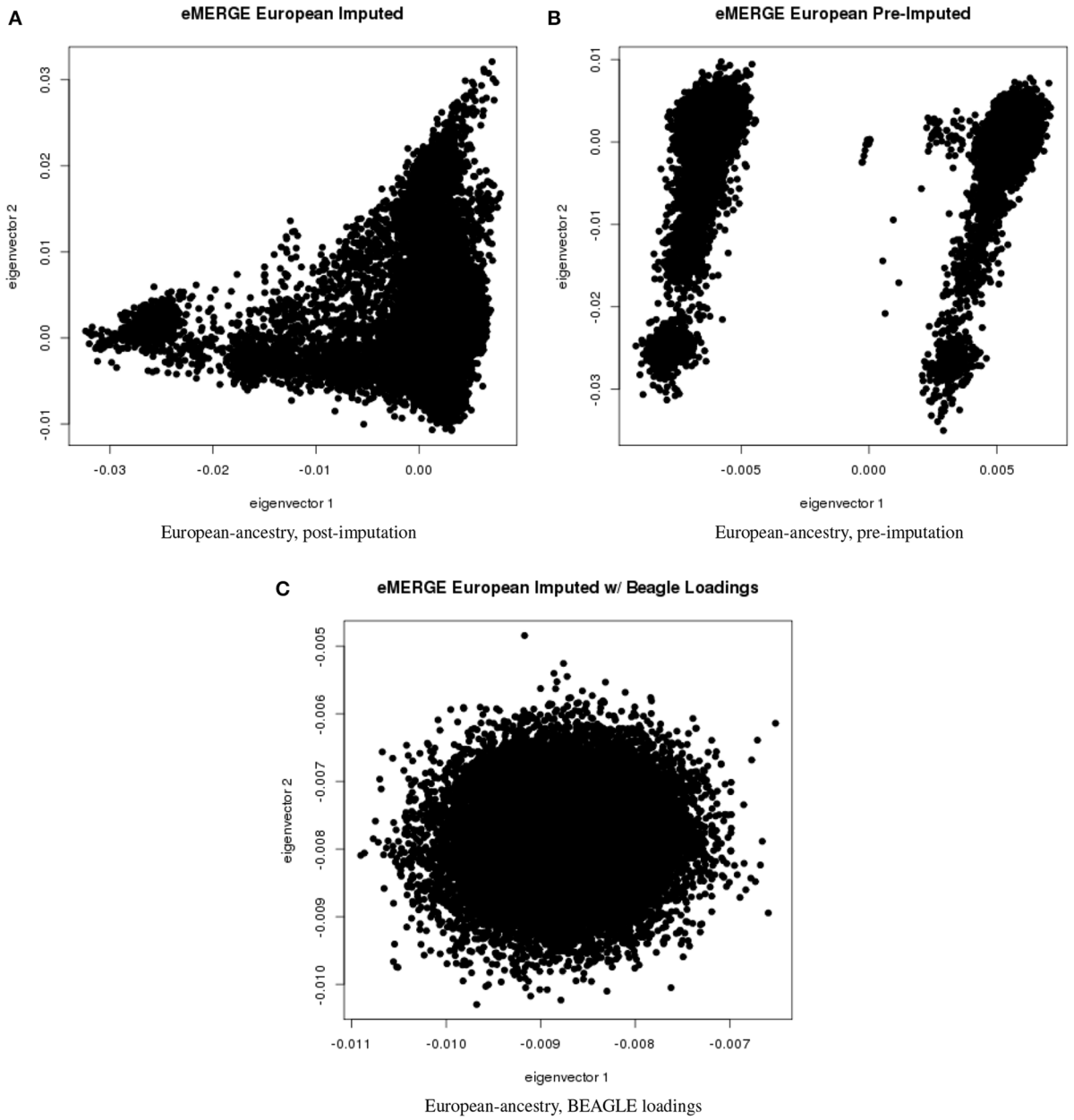


FIGURE 6 | PC plots of eMERGE participants genetically determined to be of European ancestry. (A) Plot of eigenvectors 1 and 2 for the imputed data set Hispanic participants. **(B)** Plot of eigenvectors 1 and 2 for the pre-imputed data set Hispanic participants. **(C)** Plot of eigenvectors 1 and 2 for the imputed data set Hispanic participants using the "loadings" method.

were combined with high-density platforms: the MetaboChip data set was an outlier even at overview scale. Platform differences re-appear when PCA is performed on apparently homogeneous subsets, e.g., African and European genetic ancestry subsets. These platform differences in homogeneous racial groups are amplified as the overall variance in the data set diminishes. Some of the differences might actually reflect subtle differences in LD in the populations due to ethnic stratification correlated with platform, because the populations were not randomly represented in the Biobank and therefore not randomized to platform.

In addition to difference of LD capture by platform, genotype encoding remains problematic when combining large data sets genotyped at different sites and on different platforms. A number of tools, e.g., liftOver (Hinrichs et al., 2006), can be used to standardize the allele states between data sets. Nevertheless, coding remains fraught with problems (Nelson et al., 2012). One data set was initially submitted with non-standard coding resulting in the data set being an outlier even with respect to other data sets on the same platform and chip. Such miscoding results in an extreme form of platform bias, as the LD is misrepresented.

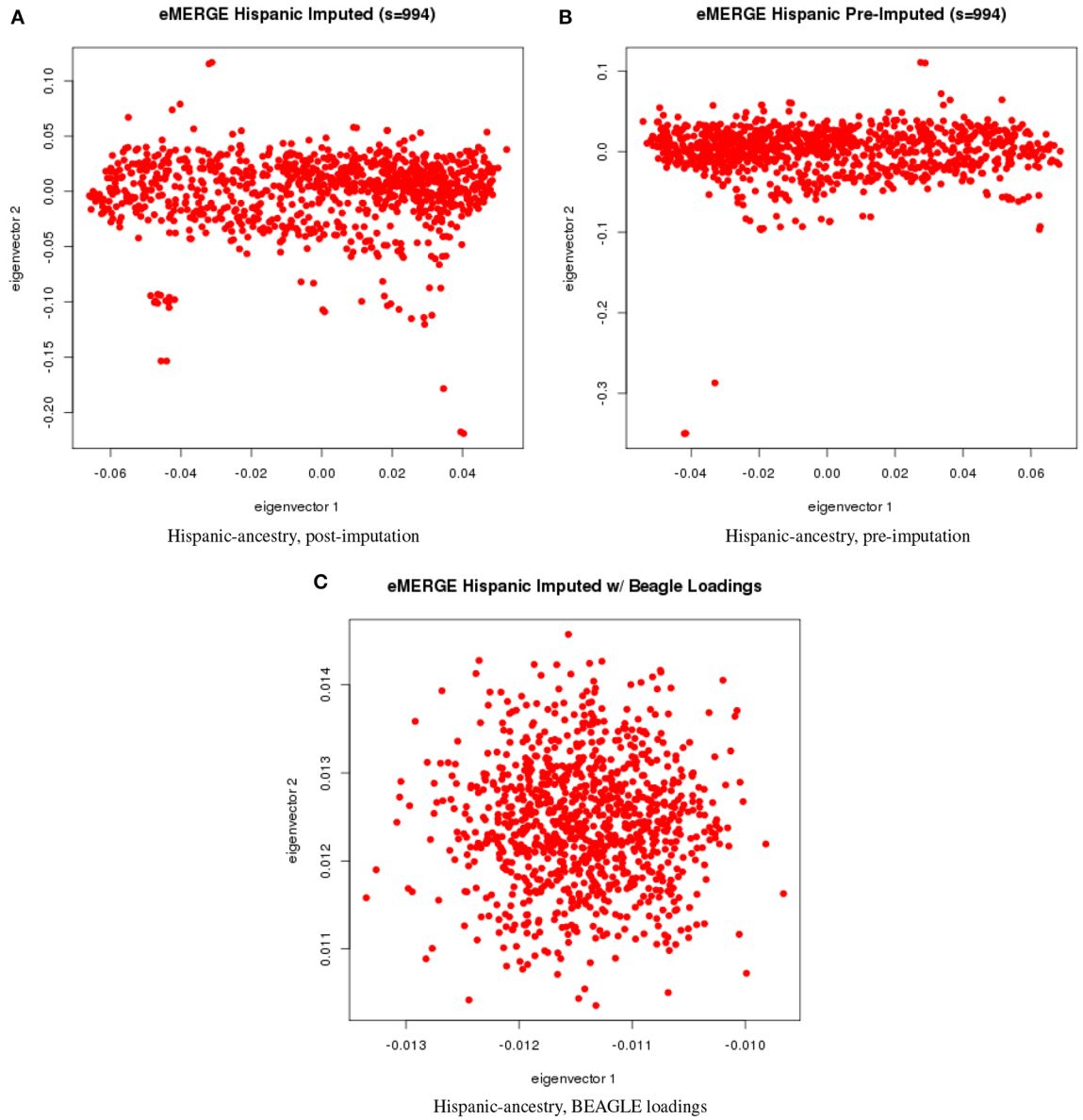


FIGURE 7 | PC plots of eMERGE participants genetically determined to be Hispanic. (A) Plot of eigenvectors 1 and 2 for the imputed data set Hispanic participants. **(B)** Plot of eigenvectors 1 and 2 for the pre-imputed data set Hispanic participants. **(C)** Plot of eigenvectors 1 and 2 for the imputed data set Hispanic participants using the "loadings" method.

Other potential source of bias could be induced by the sites or genotyping center.

It is likely that the imputed data can exaggerate some underlying features. Any chromosomal variation that is poorly represented in the reference set can lead to more uniformity around the variation that causes that chromosome to be selected. Some regions that are promoted (occur prominently in a lower number PC), probably are reflecting rare chromosomes in the reference panel.

We have outlined a general checklist for filtering variants to be utilized with PCA: (1) Ensure uniformity of strand representation among different platforms to avoid the bias induced by site; (2) Select variants on autosomal chromosomes only, no sex chromosomes; (3) Filter variants with LD pruning ($r = 0.50 - 0.84$), in a sliding window of 500 kbp; (4) Filter variants on MAF >0.05, and for missingness <0.02; and (5) Examine plots of absolute correlation between PC and genotype as illustrated in **Figure 10** and remove regions where chromosome artifacts (e.g., HLA, chromosome 8 inversion) are driving the correlation pattern for a given component (Laurie et al., 2010). However, in many cases removing the HLA region will not completely eliminate the correlation pattern in that region (data not shown). Normally the first ten eigenvectors are appropriate, but this depends on the proportion of variance explained and the specific analysis conducted.

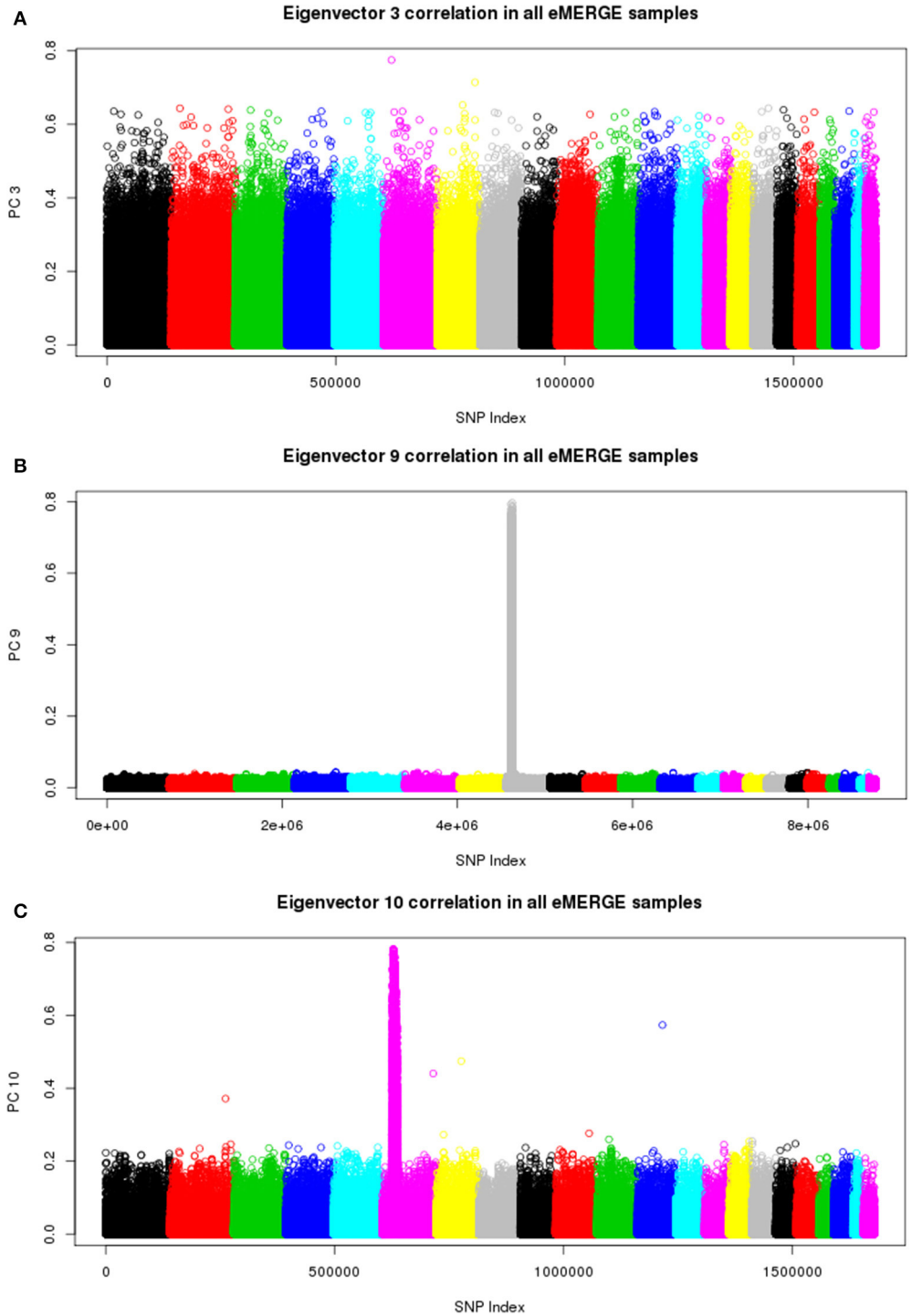


FIGURE 8 | Eigenvector-genotype correlation plots from the joint ancestry PCA analyses representing genome-wide correlation (A), correlation driven the chromosome 8 inversion (B), and correlation driven by the HLA region (C).

As a proof of concept, we repeated a previously presented genome-wide association for VTE in participants of African ancestry (Heit et al., 2013). We compared the performance of the two approaches described above: (a) PCs derived from the "loadings" method and (b) PCs derived from the equivalent of the conventional method. Our results showed that using the "loadings" approach provided similar association results and controlled for inflation while controlling for fewer covariates and

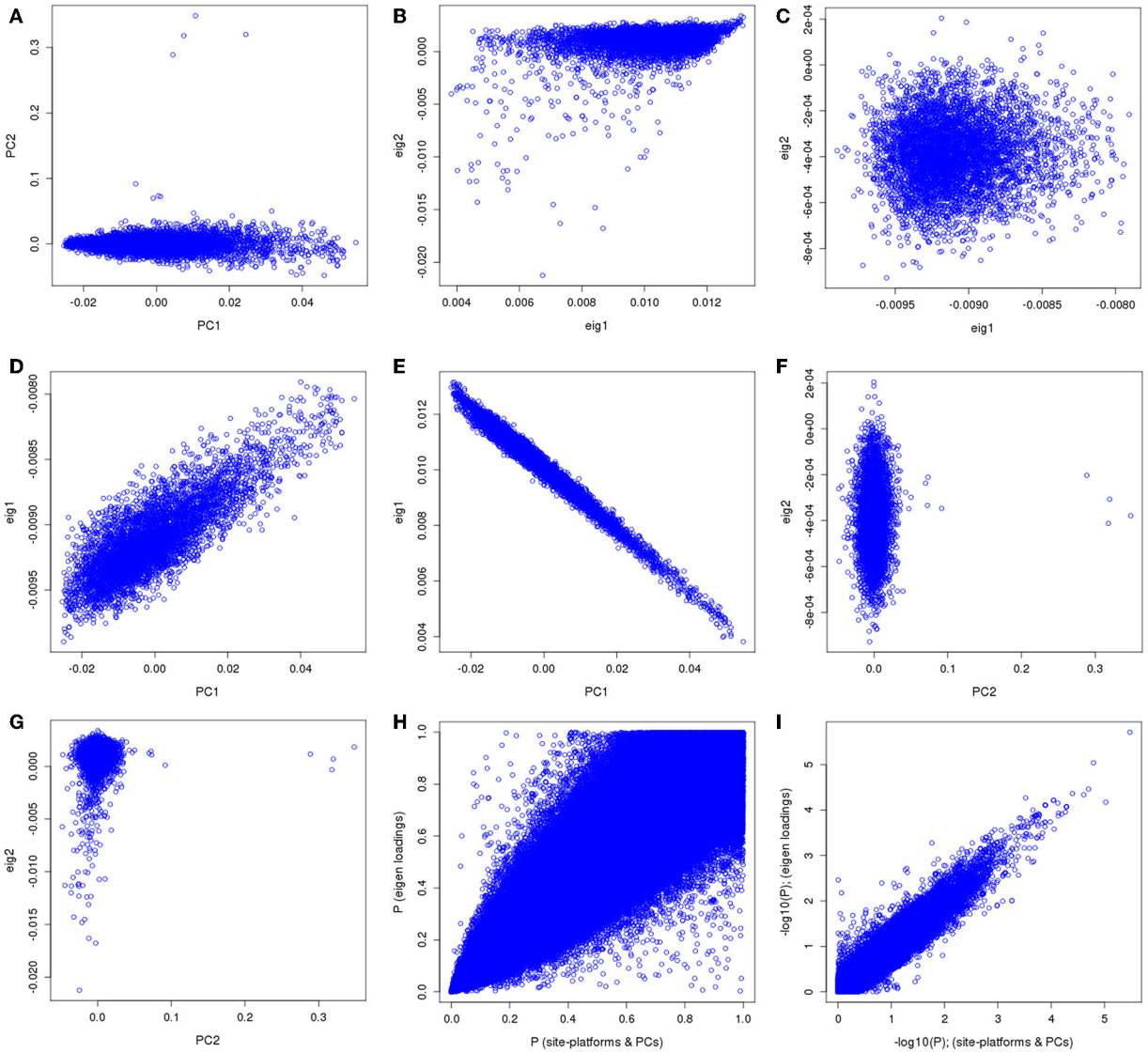


FIGURE 9 | PC comparisons derived from the "loadings" method and PCs derived from the equivalent of the imputation method for venous thromboembolism association in African ancestry participants.

consequently fewer degrees of freedom. This method will need further validation using simulated data, but does seem promising nonetheless.

We have demonstrated that analysis of data across sites in research networks can expose subtle biases and stratification effects. The conventional approach of adjusting for the first number of PCs does not adequately adjust for the bias of platform and site. We recognize that in comparison to most meta analyses which use summary statistics for aggregation, we have both individual subject genotypes as well as information on genotyping platform and site. We hope our research study will serve as a reference for similar projects that attempt to control for confounders and ancestry in large genetic association studies.

5. CONCLUSION

In summary, we outline a general checklist for filtering genetic variants for conventional PCA to avoid the bias induced by platform and site as well as to avoid false-positive results due to the correlation between the PCs and the SNP genotypes. We have also proposed the "loadings" method as an alternative to the conventional method to derive PCs that control for bias due to the site and platform. Furthermore, we demonstrated the applicability of this new approach for the VTE genome-wide association analysis in genetic African ancestry eMERGE participants.

WEB RESOURCES

– eMERGE Coordinating Center genotyping data: http://emerge.mc.vanderbilt.edu/genotyping-data-released

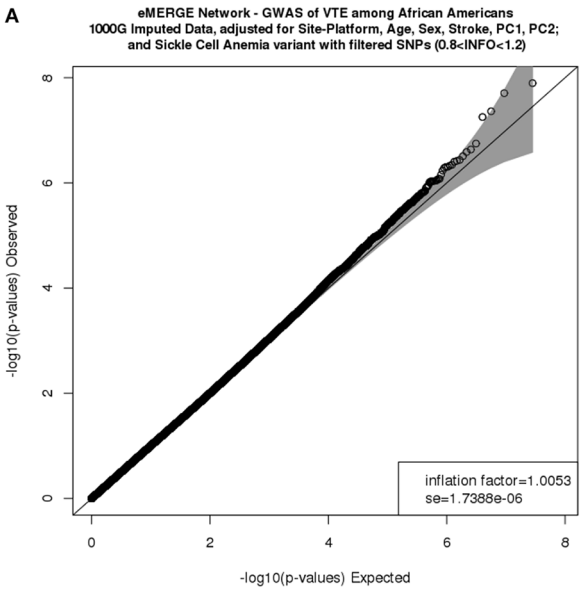


FIGURE 10 | QQ plots of the venous thromboembolism (VTE) association in African ancestry participants. PC comparisons derived from the "loadings" method and PCs derived from the equivalent of the imputation method. **(A)** QQ plots of the VTE association in African ancestry participants using PCs derived from the equivalent of the imputation method. **(B)** QQ plots of the VTE association in African ancestry participants using PCs derived from the equivalent of the "loadings" method.

– R package SNPRelate: https://github.com/zhengxwen/SNPRelate

FUNDING

This study was supported by the following U01 grants from the National Human Genome Research Institute (NHGRI), a component of the National Institutes of Health (NIH), Bethesda, MD, USA: (1) U01HG006375 (Group Health/University of Washington); (2) U01HG006382 (Geisinger Health System); (3) U01HG006379 (Mayo Clinic); (4) U01HG006389 (Essentia Health, Marshfield Clinic Research Foundation, and Pennsylvania State University); (5) U01HG006388 (Northwestern University); (6) HG004438 (Center for Inherited Disease Research, Johns Hopkins University); (7) HG004424 (Broad Institute of Harvard and MIT); (8) U01HG006378, U01HG006385, U01HG006385 (Vanderbilt University and Pennsylvania State University); (9) U01HG006380 (The Mt. Sinai Hospital); (10) U01HG006828 (Cincinnati Children's Hospital Medical Center/Harvard); (11) U01HG006830 (Childrens Hospital of Philadelphia). Additional support was provided by a State of Washington Life Sciences Discovery Fund award to the Northwest Institute of Genetic Medicine (Gail P. Jarvik).

ACKNOWLEDGMENT

We are grateful to all the participants in the eMERGE study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2014.00352/abstract

REFERENCES

Ali-Khan, S. E., Krakowski, T., Tahir, R., and Daar, A. S. (2011). The use of race, ethnicity and ancestry in human genetic research. *HUGO J.* 5, 47–63. doi: 10.1007/s11568-011-9154-5

Browning, B. L., and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84, 210–223. doi: 10.1016/j.ajhg.2009.01.005

Crawford, D. C., Crosslin, D. R., Tromp, G., Kullo, I. J., Kuivaniemi, H., Hayes, M. G., et al. (2014). eMERGEing progress in genomics—the first seven years. *Front. Genet.* 5:184. doi: 10.3389/fgene.2014.00184

Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Meth.* 10, 5–6. doi: 10.1038/nmeth.2307

Dumitrescu, L., Ritchie, M. D., Brown-Gentry, K., Pulley, J. M., Basford, M., Denny, J. C., et al. (2010). Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. *Genet. Med.* 12, 648–650. doi: 10.1097/GIM.0b013e3181efe2df

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nat. Rev. Genet.* 7, 85–97. doi: 10.1038/nrg1767

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., et al. (2013). The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genet. Med.* 15, 761–771. doi: 10.1038/gim.2013.72

Heit, J. A., De Andrade, M., Armasu, M. S., Kullo, I. J., Pathak, J., Chute, C. G., et al. (2013). "Genome-Wide Association Study (GWAS) of Venous Thromboembolism (VTE) in African-Americans from the Electronic Medical Records and Genomics (eMERGE) Network," in *Oral Presentation #458, 55th ASH Annual Meeting and Exposition*. Available online at: http://www.bloodjournal.org/content/122/21/458

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., et al. (2006). The UCSC genome browser database: update 2006. *Nucleic Acids Res.* 34(Suppl. 1), D590–D598. doi: 10.1093/nar/gkj144

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* 44, 955–959. doi: 10.1038/ng.2354

Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3 (Bethesda)* 1, 457–470. doi: 10.1534/g3.111.001198

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., et al. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* 34, 591–602. doi: 10.1002/gepi.20516

Manichaikul, A., Palmas, W., Rodriguez, C. J., Peralta, C. A., Divers, J., Guo, X., et al. (2012). Population structure of hispanics in the united states: the multi-ethnic study of atherosclerosis. *PLoS Genet.* 8:e1002640. doi: 10.1371/journal.pgen.1002640

Nelson, S. C., Doheny, K. F., Laurie, C. C., and Mirel, D. B. (2012). Is 'forward' the same as 'plus'?. . . and other adventures in {SNP} allele nomenclature. *Trends Genet.* 28, 361–363. doi: 10.1016/j.tig.2012.05.002

NHGRI. (2005). The use of racial, ethnic, and ancestral categories in human genetics research. *Am. J. Hum. Genet. Bethesda* 77, 519–532. doi: 10.1086/491747

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2:e190. doi: 10.1371/journal.pgen.0020190

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28, 3326–3328. doi: 10.1093/bioinformatics/bts606

Zhu, X., Li, S., Cooper, R. S., and Elston, R. C. (2008). A unified association analysis approach for family and unrelated samples correcting for stratification. *Am. J. Hum. Genet.* 82, 352–365. doi: 10.1016/j.ajhg.2007.10.009

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 03 July 2014; accepted: 19 September 2014; published online: 04 November 2014.

*Citation: Crosslin DR, Tromp G, Burt A, Kim DS, Verma SS, Lucas AM, Bradford Y, Crawford DC, Armasu SM, Heit JA, Hayes MG, Kuivaniemi H, Ritchie MD, Jarvik GP and de Andrade M (2014) Controlling for population structure and genotyping platform bias in the eMERGE multi-institutional biobank linked to electronic health records. Front. Genet. **5**:352. doi: 10.3389/fgene.2014.00352*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 12 May 2014
doi: 10.3389/fgene.2014.00096

Imputation of *TPMT* defective alleles for the identification of patients with high-risk phenotypes

Berta Almoguera[1]*, Lyam Vazquez[1], John J. Connolly[1], Jonathan Bradfield[1], Patrick Sleiman[1,2], Brendan Keating[1,2] and Hakon Hakonarson[1,2]

[1] *Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA*
[2] *Department of Pediatrics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA*

Edited by:
Helena Kuivaniemi, Geisinger Health System, USA

Reviewed by:
Peter A. Kanetsky, Moffitt Cancer Center, USA
Kelli K. Ryckman, University of Iowa, USA

****Correspondence:***
Berta Almoguera, Center for Applied Genomics, The Children's Hospital of Philadelphia, 3615 Civic Center Boulevard, Abramson Building, Philadelphia, PA 19104, USA
e-mail: castillob@email.chop.edu

Background: The activity of thiopurine methyltransferase (TPMT) is subject to genetic variation. Loss-of-function alleles are associated with various degrees of myelosuppression after treatment with thiopurine drugs, thus genotype-based dosing recommendations currently exist. The aim of this study was to evaluate the potential utility of leveraging genomic data from large biorepositories in the identification of individuals with TPMT defective alleles.

Material and methods: *TPMT* variants were imputed using the 1000 Genomes Project reference panel in 87,979 samples from the biobank at The Children's Hospital of Philadelphia. Population ancestry was determined by principal component analysis using HapMap3 samples as reference. Frequencies of the TPMT imputed alleles, genotypes and the associated phenotype were determined across the different populations. A sample of 630 subjects with genotype data from Sanger sequencing ($N = 59$) and direct genotyping ($N = 583$) (12 samples overlapping in the two groups) was used to check the concordance between the imputed and observed genotypes, as well as the sensitivity, specificity and positive and negative predictive values of the imputation.

Results: Two SNPs (rs1800460 and rs1142345) that represent three *TPMT* alleles (*3A, *3B, and *3C) were imputed with adequate quality. Frequency for the associated enzyme activity varied across populations and 89.36–94.58% were predicted to have normal TPMT activity, 5.3–10.31% intermediate and 0.12–0.34% poor activities. Overall, 98.88% of individuals (623/630) were correctly imputed into carrying no risk alleles (553/553), heterozygous (45/46) and homozygous (25/31). Sensitivity, specificity and predictive values of imputation were over 90% in all cases except for the sensitivity of imputing homozygous subjects that was 80.64%.

Conclusion: Imputation of TPMT alleles from existing genomic data can be used as a first step in the screening of individuals at risk of developing serious adverse events secondary to thiopurine drugs.

Keywords: TPMT, genotype imputation, DNA biobank, pharmacogenetics, Electronic Medical Records

INTRODUCTION

Thiopurine S-methyltransferase (TPMT) is an enzyme involved in the metabolism of purine analogs such as azathioprine, 6-mercaptopurine and thioguanine, drugs that are used as chemotherapeutic and immunosuppressant agents in diseases such as lymphoid malignancies, leukemias, inflammatory bowel disease, and other immune conditions (Relling et al., 2011; Appell et al., 2013). *TPMT* maps to chromosome 6p22.3. It is subject to genetic variation and, to date, 34 alleles have been identified and characterized, most of which are associated with reduced activity *in vitro* (Relling et al., 2011). Alleles *2 (rs1800462), *3A (rs1800460 and rs1142345), *3B (rs1800460), and *3C (rs1142345) account for 95% of all defective alleles and all four involve missense mutations: allele *2 results in the p.Ala80Pro change (chr6:18143955), allele *3A contains two missense changes: p.Ala154Thr (chr6: 6:18139228) and p.Tyr240Cys (chr6:18130918), and alleles *3B and *3C are defined by p.Ala154Thr and p.Tyr240Cys, respectively (reference sequence NP_000358.1). The frequencies of these alleles vary significantly across ethnic populations (Appell et al., 2013): while *3A is the most frequently found in Caucasians (4.5%) (Schaeffeler et al., 2004), *3C is more prevalent in Africans or Asians, with 5.4–7.6% and 0.3–3%, respectively (reviewed in Teml et al., 2007).

TPMT enzymatic activity exhibits a trimodal distribution and approximately 0.3% of the population carry two defective alleles (associated with negligible activity), about 10% are heterozygous (intermediate activity), and 89% have normal activity (Weinshilboum and Sladek, 1980; Schaeffeler et al., 2004). Therefore, both heterozygous and homozygous individuals are at higher risk of developing myelosuppression within a few weeks after starting treatment with conventional doses that can be lethal

if unrecognized, independent of the underlying disease being treated (Sim et al., 2013).

Due to the potential cytotoxicity and narrow therapeutic index of thiopurines, the US Food and Drug Administration (FDA) recommends *TPMT* testing prior to starting treatment with thiopurine drugs, and *TPMT* genotype-guided dosing recommendations are currently in use (Relling et al., 2011, 2013).

Results of genetic tests are potentially relevant over a patient's lifetime and having that information incorporated into patients' medical records may be useful in the improvement and guidance of drug treatments, if ever needed. With electronic medical records (EMR) currently widely implemented at academic hospitals and other treatment institutions, pharmacogenetic actionable variants can be integrated to the already available patient's information, helping optimize clinical decision making and care planning (Gottesman et al., 2013). Moreover, genome-wide data is increasingly accessible due to decreasing costs of genomic technologies and the development of methods that allow for accurate imputation of genotypes not directly probed by specific arrays could influence health care decisions (Marchini and Howie, 2010). Genomic data is frequently stored within large biorepositories where DNA samples are linked with phenotypic data. These biorepositories have been efficient and successful in studies of genotype-phenotype associations and they can be used as a model for the implementation and evaluation of pharmacogenomics in routine clinical practice (Gottesman et al., 2013).

In the present study, we leverage existing genome-wide genotyping data to impute common defective *TPMT* alleles with the aim of identifying individuals carrying high-risk genotypes for thiopurines-induced adverse events.

MATERIAL AND METHODS

SUBJECTS AND GENOTYPING

This study was approved by the institutional review board and the ethics committee of The Children's Hospital of Philadelphia (CHOP). Written informed consent was obtained from each participant in accordance with institutional requirements and the Declaration of Helsinki Principles. Subjects were selected from the biorepository at the Center for Applied Genomics at CHOP. The CHOP biobank has a collection of over 160,000 samples including 60,000 internal pediatric samples and over 100,000 adult and pediatric samples from external collaborators genotyped using standard GWAS arrays from Illumina and Affymetrix (summarized in Gottesman et al., 2013).

Figure 1 illustrates the study process. For *TPMT* imputation, we selected a total of 87,979 samples genotyped with either InfiniumII HumanHap550 (550; $N = 45{,}893$) or Human610-Quad version 1 (Quad; $N = 42{,}086$) arrays (Illumina, San Diego, CA). Genotyping data were used to impute sex using PLINK (http://pngu.mgh.harvard.edu/purcell/plink/) (Purcell et al., 2007); population ancestry was determined by principal component analysis (Eigenstrat 3.0) (Price et al., 2006), and samples were grouped into populations using nearest neighbors analysis and the HapMap3 samples (https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html) as a reference.

IMPUTATION OF TPMT GENOTYPES

Imputation of unobserved genotypes in *TPMT* gene locus (chr6:18,128,545–18,155,374) was carried out with the IMPUTE2 package (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html) (Howie et al., 2009) with the 1,000 Genomes Project reference panel, after prephasing chromosome 6 haplotypes with SHAPEIT version 2 (http://www.shapeit.fr/) (Delaneau et al., 2013). Since rs1800460 is probed on the Illumina HH610 Quad array, prephasing and imputation was performed for each chip type separately. Quality control filters were applied and only SNPs with an info score >0.9 were kept.

VALIDATION OF THE IMPUTED GENOTYPES

To determine the accuracy of imputation, *TPMT* imputed haplotypes were compared to those obtained by other genotyping platforms covering *TPMT* variation. Of the 87,979 samples, 583 also had genotyping data on Illumina Infinium Immunochip (Immunochip), and HumanOmni1-Quad version 1 (Omni), which captured both rs1800460 and rs1142345. Additionally, Sanger sequencing of rs1800460 in exon 7 and rs1142345 in exon 10 was used to validate the imputation results (primers previously described in Schaeffeler et al., 2001). The sample selected for Sanger sequencing consisted of 59 individuals predicted to carry one or two defective alleles by imputation that had been exposed to a TPMT medication based on the EMR.

DETERMINATION OF THE CONCORDANCE AND THE SENSITIVITY, SPECIFICITY, POSITIVE AND NEGATIVE PREDICTIVE VALUES OF THE IMPUTATION FOR THE IDENTIFICATION OF CARRIERS OF TPMT DEFECTIVE ALLELES

We determined the concordance of the imputation as the number of imputed genotypes that correspond with output from direct genotyping or sequencing (expressed as percentage). Sensitivity, specificity and positive and negative predictive values of the imputation in the discrimination of subjects carrying the *1/*1 genotype, and one and two *TPMT* defective alleles were determined as shown in **Figure 2**.

RESULTS

SUBJECTS AND TPMT IMPUTATION

Ancestry, sex estimation and imputation of *TPMT* genotypes for the 89,797 individuals were performed in eleven batches of ∼8,000 samples. An average of 174,911 SNPs with $r^2 < 0.2$ were used for principal components calculation.

There were 50.04% males out of the 89,797 individuals (imputed sex for 0.87% of the individuals was undetermined). Principal component analysis classified 72.74% of individuals as Caucasians, 18.78% with African ancestry, 6.55% Hispanics, and 1.93% Asians (**Table 1**). There were also data on self-reported ethnicity for 24,527 out of the 89,797 individuals, with 50.15% Caucasians ($N = 12{,}304$), 41.81% African Americans ($N = 10{,}263$), 1.6% Asians ($N = 385$), 0.08% American Indians ($N = 21$), 0.03% Native Hawaiians ($N = 7$), 0.02% Indians ($N = 5$), and 6.29% were considered as "Other" ($N = 1542$). Concordance between self-reported and imputed ancestry was >80% for Caucasians, African Americans, and Asians (93.92, 98.34, and 81.30%, respectively). For the remaining groups, 57.58% were

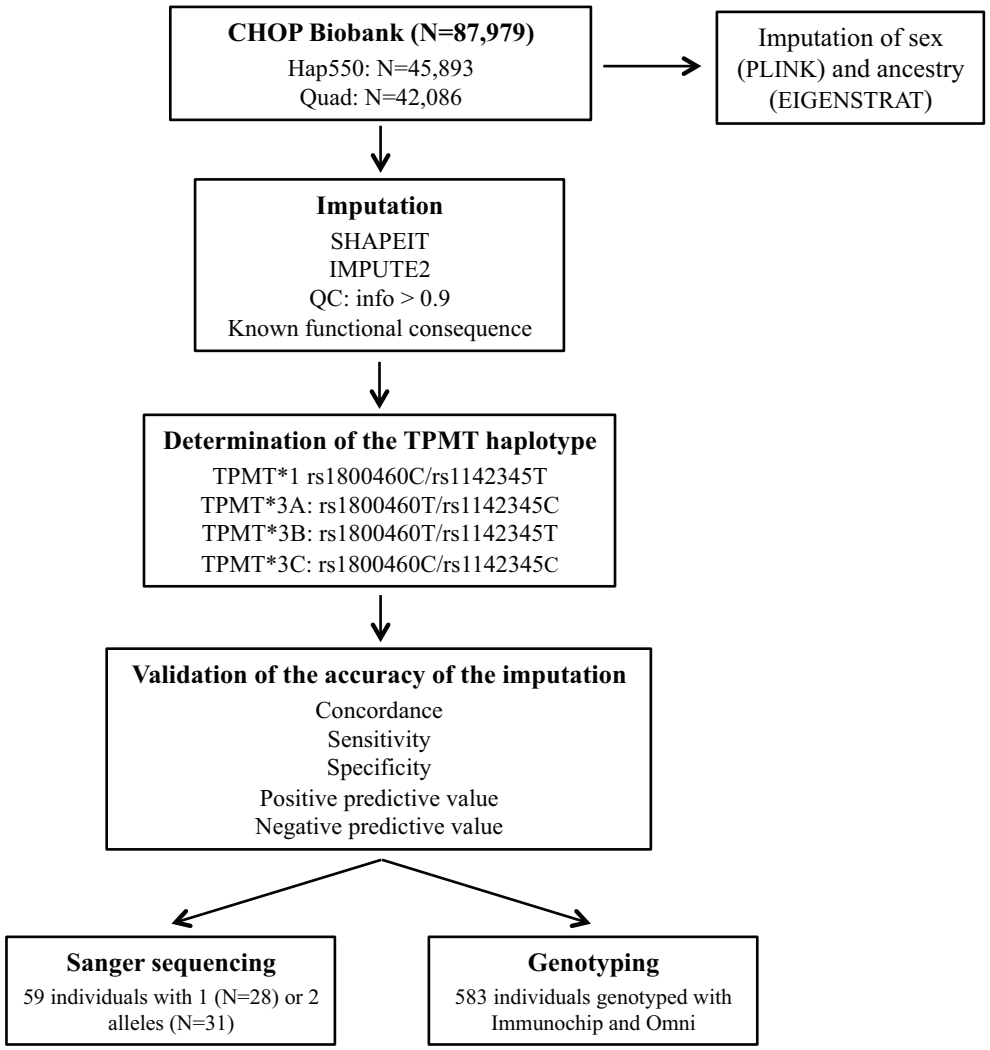


FIGURE 1 | General schema for the study process. Hap550 is the InfiniumII HumanHap550 array, Quad is Human610-Quad version 1 array, Immunochip is the Illumina Infinium Immunochip array and Omni is the HumanOmni1-Quad version 1.

	Observed genotypes	
Imputed genotypes	**Carriers of N alleles**	**Non-carriers of N alleles**
Carriers of N alleles	True positives	False negatives
Non-carriers of N alleles	False positives	True negatives

Sensitivity = True positives/(True positives + False negatives) x 100
Specificity = True negatives/(True negatives + False positives) x 100
Positive predictive value = True positives/(True positives + False positives) x 100
Negative predictive value = True negatives/(True negatives + False negatives) x 100

FIGURE 2 | Definition of true and false positive and negative values and formulae used for the determination of the sensitivity, specificity, and predictive values of the imputation.

classified as Hispanics (19 out of 33), 15.15% as Asians ($N = 5$) and Caucasians ($N = 5$) and 12.12% as African Americans ($N = 4$).

Three hundred and fifty four variants were imputed in the *TPMT* gene, including 322 SNPs and 33 insertion/deletion polymorphisms (*indels*). Out of these, only 117 had an info value $\geq$0.9 (103 SNPs and 14 *indels*). However for the subsequent analysis only those with a known functional significance were considered: rs1800460 and rs1142345, which define alleles *3A, *3B, and *3C. The loss-of-function variant rs1800462 that

Table 1 | Frequencies of the different ethnicities in the sample investigated based on principal component analysis of imputed genotype data.

Population	*N*	(%)
Caucasians	63,998	72.74
CEU, Utah residents with ancestry from northern and western Europe	56,675	64.42
TSI, Toscans in Italy	7,323	8.32
African ancestry	16,519	18.78
ASW, African ancestry in Southwest USA	15,457	17.57
YRI, Yoruba in Ibadan, Nigeria	947	1.08
MKK, Maasai in Kinyawa, Kenya	96	0.11
LWK, Luhya in Webuye, Kenya	19	0.02
Hispanics	5,764	6.55
MEX, Mexican ancestry in Los Angeles, California	4,786	5.44
GIH, Gujarati Indians in Houston, Texas	978	1.10
Asians	1,698	1.93
CHD, Chinese in Metropolitan Denver, Colorado	1,043	1.19
JPT, Japanese in Tokyo, Japan	269	0.91
CHB, Han Chinese in Beijing, China	386	0.43

defines allele *2 was also imputed but did not pass the quality filters, thus it was excluded from the further analyses.

TPMT alleles were assigned as *1 when the rs1800460 C>T/rs1142345 T>C diplotype was CT, *3B when TT, and *3C when CC. For allele *3A, given the high linkage disequilibrium between rs1800460 (*3B) and rs1142345 (*3C) and the low minor allele frequency, whenever an individual carried both variants (rs1800460T and rs1142345C), the allele was assigned as *3A.

TPMT allelic, genotypic and associated phenotypic frequencies for *3A, *3B, and *3C across ethnic groups are illustrated in **Tables 2**, **3** and **4**, respectively. As shown in **Table 2**, the distribution of the three defective alleles varied largely across populations: *3A was more represented among the Caucasians, *3B in the Hispanics and *3C in both Asians and African Americans, being the latter the group with the highest frequency of carriers of *TPMT* defective alleles (5.49 vs. 4.07% in Caucasians, 4.41% in Hispanics and 2.77% in Asians). According to the genotype-associated enzymatic activity, Asians harbored the lowest rates of poor metabolizers with only 0.12% whereas Caucasians, African Americans and Hispanics have a frequency close to 0.33%.

CONCORDANCE, SENSITIVITY, SPECIFICITY, AND POSITIVE AND NEGATIVE PREDICTIVE VALUES OF THE IMPUTATION

Out of the 87,979 samples used for imputation, 583 had genotyping data on both rs1800460 and rs1142345: 94.8% of them carried the genotype *1/*1, 4.5% *1/*3A, and 0.7% the genotype *1/*3C. Concordance of the imputed haplotypes compared to those determined by genotyping was 99.8% (**Table 5**).

Sanger sequencing was performed in a subset of 59 samples predicted to carry 1 ($N = 28$) or 2 ($N = 31$) defective alleles (*3A, or *3C). Twelve of the 59 samples also had genotype data and results were consistent across the two methods and with the imputation. The overall concordance was 84.7% for the total 59 samples. **Table 6** illustrates the concordance of the imputed genotypes after validation with Sanger sequencing.

When taking into account the number of defective alleles, 98.88% of individuals (623 of the 630 individuals—excluding the 12 in the two groups-) were accurately imputed. All of the samples identified as carrying no risk alleles (*1/*1) were confirmed by direct genotyping or sequencing (553/553) and for heterozygous and homozygous individuals, the concordance was lower, with 97.8% (45/46) and 80.64% (25/31), respectively (**Table 7**). Sensitivity, specificity and positive and negative predictive values of imputation of *TPMT* genotypes are summarized in **Table 7**. Since the importance of *TPMT* genotyping lies in the discrimination of individuals carrying defective alleles, these metrics were determined for the discrimination of individuals carrying the *1/*1 genotype, individuals with one defective allele and individuals with two defective alleles. Sensitivity, specificity, and predictive values were close to 100% for all cases except for the positive predictive value of identifying homozygous carriers, which was 80.64%.

DISCUSSION

A growing number of drug-gene interactions, affecting routinely prescribed drugs, are being validated (Relling and Klein, 2011). The FDA has to date recommended the inclusion of pharmacogenetic markers in the labels of more than a 100 drugs (http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm) (Shuldiner et al., 2013) and initiatives such as Pharmacogenomics Knowledge Base (PharmGKB) and the Clinical Pharmacogenetics Implementation Consortium (CPIC) (Relling and Klein, 2011) provide essential pharmacogenetic information and play a major role in establishing recommendations to aid clinicians in guiding therapies. If one considers the report by Schildcrout et al, who demonstrated that up to 65% of patients were exposed to at least one medication with an established drug-gene association within 5 years (Schildcrout et al., 2012), then pharmacogenetics integration into individuals' medical records for clinical use is becoming an urgent need. Availability of pharmacogenetic information prior to patients' treatment has the opportunity to identify individuals potentially benefiting from a given therapy, select adequate medications and doses, in order to ultimately administer the most effective treatment to each patient, with the lower incidence of adverse events.

Using existing genomic data from the CHOP biobank repository we have been able to impute three of the most common defective *TPMT* alleles *3A, *3B, *3C in a cohort of 87,979 individuals. The sensitivity, specificity and positive and negative predictive values of the imputation were sufficiently high to allow discrimination of patients carrying one or two defective alleles from those with a *1/*1 genotype. Concordance between observed and imputed genotypes was 100% for individuals with *1/*1 genotype, and for carriers of *TPMT* alleles, discordant

Table 2 | Distribution of allele frequencies for *TPMT* alleles *3A, *3B, and *3C across the different ethnic groups.

	Caucasian (*N* = 63,998)	AA (*N* = 16,519)	Hispanic (*N* = 5,764)	Asian (*N* = 1,698)	Total (*N* = 87,979)
Allele	***N* (%)**	***N* (%)**	***N* (%)**	***N***	***N* (%)**
*1	122,787 (95.93)	31,225 (94.51)	11,020 (95.59)	3,302 (97.23)	168,333 (95.67)
*3A	4,305 (3.36)	303 (0.92)	334 (2.90)	19 (0.56)	4,961 (2.82)
*3B	86 (0.07)	1 (0.00)	12 (0.10)	0 (0.00)	99 (0.06)
*3C	817 (0.64)	1,509 (4.57)	162 (1.41)	75 (2.21)	2,563 (1.46)

Table 3 | Distribution of genotypes for *TPMT* alleles *3A, *3B, and *3C across the different ethnic groups.

	Caucasian (*N* = 63,998)	AA (*N* = 16,519)	Hispanic (*N* = 5,764)	Asian (*N* = 1,698)	Total (*N* = 87,979)
Genotype	***N* (%)**	***N* (%)**	***N* (%)**	***N* (%)**	***N* (%)**
*1/*1	58,981 (92.16)	14,761 (89.36)	5,275 (91.52)	1,606 (94.58)	80,623 (91.64)
*1/*3A	4,119 (6.44)	286 (1.73)	322 (5.59)	19 (1.12)	4,746 (5.39)
*1/*3B	10 (0.02)	1 (0.01)	1 (0.02)	0 (0.00)	12 (0.01)
*1/*3C	697 (1.09)	1,416 (8.57)	147 (2.55)	71 (4.18)	2,331 (2.65)
*3A/*3A	81 (0.13)	1 (0.01)	5 (0.09)	0 (0.00)	87 (0.10)
*3A/*3B	1 (0.01)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.001)
*3A/*3C	23 (0.04)	15 (0.09)	2 (0.03)	0 (0.00)	40 (0.05)
*3B/*3C	75 (0.12)	0 (0.00)	11 (0.19)	0 (0.00)	86 (0.10)
*3C/*3C	11(0.02)	39 (0.24)	1 (0.02)	2 (0.12)	53 (0.06)

Table 4 | *TPMT* genotype-associated phenotypic frequencies across the different ethnic groups.

	Caucasian (*N* = 63,998)	AA (*N* = 16,519)	Hispanic (*N* = 5,764)	Asian (*N* = 1,698)	Total (*N* = 87,979)
TPMT activity	***N* (%)**	***N* (%)**	***N* (%)**	***N* (%)**	***N* (%)**
Normal	58,981 (92.16)	14,761 (89.36)	5,275 (91.52)	1,606 (94.58)	80,623 (91.64)
Intermediate	4,826 (7.54)	1,703 (10.31)	470 (8.15)	90 (5.30)	7,089 (8.06)
Low	191 (0.30)	55 (0.33)	19 (0.33)	2 (0.12)	267 (0.30)

results were essentially cases of individuals predicted to be heterozygous by imputation that were found to be homozygous by genotyping or sequencing. Additionally, probably because of the rarity of SNPs rs1800460 and rs1142345 and the increase in imputation errors as minor allele frequency decreases, alleles *3A and *3C were frequently switched, and allele *3B was only identified in the subset of samples genotyped with the Quad array. The rarity of allele *2 may also be the explanation for the inability of imputing with adequate quality.

TPMT deficiency exhibits an extensive interethnic variability (Wang et al., 2010; Appell et al., 2013). The population investigated in this study is characterized for being largely admixed with African Americans, Asians and Hispanics accounting for almost 30% of all individuals. As previously described, frequency of alleles *3A, *3B and *3C is population-specific. Whereas *3A and *3B were predominantly found in Caucasians and Hispanics (3.36 and 2.90%, for *3A and 0.07 and 0.1%, for *3B, respectively), the most prevalent defective allele in African Americans and Asians was *3C (4.57 and 2.21%, respectively). These results were similar to frequencies previously reported for those populations (Oliveira et al., 2007; Taja-Chayeb et al., 2008; Appell et al., 2013). Regarding TPMT associated-phenotypes, African

Table 5 | Concordance between the imputed genotypes and genotypes determined by genotyping using Immunochip (Illumina Infinium Immunochip array) and Omni (HumanOmni1-Quad version 1) (*N* = 583).

Imputation	Genotyping			
	***1/*1**	***1/*3A**	***1/*3C**	**Total**
*1/*1	553	0	0	553
*1/*3A	0	26	0	26
*1/*3C	1	0	3	4
Total	554	26	3	583

Americans had the highest proportion of intermediate (10.31%) and low methylators (0.33%), being the ethnic group with the highest risk of developing adverse events derived from TPMT treatment. Conversely, Asians are the lowest risk group, with only 5.4% of individuals carrying one (5.30%) or two alleles (0.12%). Caucasians and Hispanics had a similar percentage of individuals with TPMT intermediate (7.54 and 8.15%, respectively) and low activity (0.30 and 0.33%, respectively). Other population-specific alleles not imputed in the current study, such as *2 that is almost restricted to Caucasians, or *6 (rs75543815) and *8 (rs56161402) that occur at frequencies between 1.5 and 3.5% in some African and Asian populations (Oliveira et al., 2007), are also important contributors of TPMT deficiency. These rare *TPMT* alleles or novel variants will not be detected with this approach and can only be identified by direct genotyping or sequencing. Thus, the frequency of intermediate and low methylators in this study may be slightly underestimated.

It is worth mentioning that approximately 1 in 10 individuals tested from our biobank were found to carry at least one high-risk *TPMT* allele. There are currently over 2.5 million children enrolled in the CHOP healthcare system, and if one extrapolates the results yielded from this study to the entire population at CHOP, then more than 170,000 patients would be expected to be TPMT deficient. Identification of such carriers is especially important in the pediatric population, as thiopurines are commonly prescribed drugs in children. Thiopurines are the backbone drugs for maintenance of acute lymphoblastic leukemia (ALL), which is the most common childhood malignancy (Pui and Evans, 2006), and are also frequently used as chronic immunosuppressive therapy after organ transplantation and in inflammatory bowel disease (Dubinsky, 2004; Relling et al., 2011; Appell et al., 2013). A major limitation of their use is their narrow therapeutic index and the severe myelosuppression they cause, a life threating adverse event highly associated with *TPMT* deficiency (Relling et al., 1999). In a study by Relling and coworkers in 180 children with ALL receiving conventional doses of 6-mercaptopurin, the authors found that the cumulative incidence of toxicity was 100% for homozygous TPMT deficiency, 35% for heterozygous, and 7% for patients homozygous for allele *1 (Relling et al., 1999). This association has been widely replicated, so genotyping of *TPMT* is recommended in US FDA-approved labeling and currently genotype-based dosing recommendations exist (Relling et al., 2011, 2013). The high genotype-phenotype correlation existent for TPMT and the large interethnic variability in the susceptibility to thiopurine hematopoietic toxicity, sustain the need of availing such genetic information to prospectively identify individuals where thiopurine therapy may need to be modified or changed.

Biorepositories where DNA samples are linked to the EMR of patients, such as the CHOP biobank, offer the ideal platform for screening and identification of individuals with high-risk genotypes that may require a modification in the therapy if a given drug is prescribed. CHOP is part of the Electronic Medical Records and Genomics (eMERGE) consortium that is actively working on large-scale testing and integration of information on actionable pharmacogenetic variants, such as *TPMT* alleles, into clinical practice using EMR technologies (Gottesman et al., 2013). One of the goals of eMERGE is the creation of SPHINX (Sequence, Phenotype, and pHarmacogenomics INtegration eXchange http://www.emergesphinx.org/), a web accessible repository of genomic variants derived from a panel of 84 genes involved in the pharmacogenetics of a large number of drugs, designed by the NIH-supported Pharmacogenetics Research Network (PGRN), and linked to clinical information. To date, SPHINX contains data on 2000 of the nearly 9000 subjects that are planned to be enrolled in the project. Interestingly, so far SPHINX lists 174 variants in the *TPMT* gene, including known and novel variants, and the minor allele frequency information. Variant repositories such as SPHINX allow the advance in the knowledge of pharmacogenetics through the exploration of new hypotheses and the further integration of this information into the EMR.

Table 6 | Concordance between imputed genotypes and genotypes determined by Sanger sequencing (*N* = 59).

Imputation	Sanger sequencing						
	*1/*1	*1/*3A	*1/*3C	*3A/*3A	*3A/*3C	*3C/*3C	Total
*1/*1	0	0	0	0	0	0	0
*1/*3A	0	19	0	0	0	0	19
*1/*3C	0	0	9	0	0	0	9
*3A/*3A	1	0	0	6	0	0	7
*3A/*3C	0	0	1	0	3	1	5
*3C/*3C	0	0	4	1	1	13	19
Total	1	19	14	7	4	14	59

Table 7 | Concordance between imputed and observed genotypes according to the number of defective alleles and characteristics of the imputation in terms of sensitivity (S), specificity (SP), positive predictive value (PPV) and negative predictive value (NPV) (*N* = 630).

Imputed genotypes	Observed genotypes			Imputation metrics			
	0 alleles	1 allele	2 alleles	S	SP	PPV	NPV
0 alleles	553	0	0	99.64	100	100	97.40
1 allele	1	45	0	90.00	99.83	97.82	99.14
2 alleles	1	5	25	100	99.01	80.64	100
Total	555	50	25				

The results yielded from this study demonstrate that imputation of *TPMT* alleles from existing genomic data is feasible and may be used as a first step in the screening of high-risk individuals for thiopurine drugs toxicity. Sensitivity, specificity, and predictive values of the imputation were over 90% in all cases, except for the positive predictive value of the imputation of homozygous subjects. Given that around 90% of the population is expected to have two fully functional *TPMT* alleles, being able to accurately identify such individuals based on existing genomic data yields 10% of the population to be screened for high-risk genotypes with direct genotyping methods. The positive and negative predictive values of 100 and 97.40%, respectively, obtained for the discrimination of individuals with the *1/*1 genotype supports the potential utility of imputation in narrowing the target population where *TPMT* genotypes need to be determined. Further integration of such pharmacogenetic information into the EMR, with clinical decision support, may be used to aid clinicians prescribe therapies with the maximum risk-benefit ratio based on each individual's information.

ACKNOWLEDGMENTS

This work was funded by institutional support from the Children's Hospital of Philadelphia and the National Human Genome Research Institute (# U01 HG006830).

REFERENCES

Appell, M. L., Berg, J., Duley, J., Evans, W. E., Kennedy, M. A., Lennard, L., et al. (2013). Nomenclature for alleles of the thiopurine methyltransferase gene. *Pharmacogenet. Genomics* 23, 242–248. doi: 10.1097/FPC.0b013e32835f1cc0

Delaneau, O., Zagury, J. F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* 10, 5–6. doi: 10.1038/nmeth.2307

Dubinsky, M. C. (2004). Azathioprine, 6-mercaptopurine in inflammatory bowel disease: pharmacology, efficacy, and safety. *Clin. Gastroenterol. Hepatol.* 2, 731–743. doi: 10.1016/S1542-3565(04)00344-1

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., et al. (2013). The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* 15, 761–771. doi: 10.1038/gim.2013.72

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5:e1000529. doi: 10.1371/journal.pgen.1000529

Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11, 499–511. doi: 10.1038/nrg2796

Oliveira, E., Quental, S., Alves, S., Amorim, A., and Prata, M. J. (2007). Do the distribution patterns of polymorphisms at the thiopurine S-methyltransferase locus in sub-Saharan populations need revision? Hints from Cabinda and Mozambique. *Eur. J. Clin. Pharmacol.* 63, 703–706. doi: 10.1007/s00228-007-0310-8

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. doi: 10.1038/ng1847

Pui, C. H., and Evans, W. E. (2006). Treatment of acute lymphoblastic leukemia. *N. Engl. J. Med.* 354, 166–178. doi: 10.1056/NEJMra052603

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Relling, M. V., Gardner, E. E., Sandborn, W. J., Schmiegelow, K., Pui, C. H., Yee, S. W., et al. (2013). Clinical pharmacogenetics implementation consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing: 2013 update. *Clin. Pharmacol. Ther.* 93, 324–325. doi: 10.1038/clpt.2013.4

Relling, M. V., Gardner, E. E., Sandborn, W. J., Schmiegelow, K., Pui, C. H., Yee, S. W., et al. (2011). Clinical Pharmacogenetics Implementation Consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing. *Clin. Pharmacol. Ther.* 89, 387–391. doi: 10.1038/clpt.2010.320

Relling, M. V., Hancock, M. L., Rivera, G. K., Sandlund, J. T., Ribeiro, R. C., Krynetski, E. Y., et al. (1999). Mercaptopurine therapy intolerance and heterozygosity at the thiopurine S-methyltransferase gene locus. *J. Natl. Cancer Inst.* 91, 2001–2008.

Relling, M. V., and Klein, T. E. (2011). CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clin. Pharmacol. Ther.* 89, 464–467. doi: 10.1038/clpt.2010.279

Schaeffeler, E., Fischer, C., Brockmeier, D., Wernet, D., Moerike, K., Eichelbaum, M., et al. (2004). Comprehensive analysis of thiopurine S-methyltransferase phenotype-genotype correlation in a large population of German-Caucasians and identification of novel TPMT variants. *Pharmacogenetics* 14, 407–417. doi: 10.1097/01.fpc.0000114745.08559.db

Schaeffeler, E., Lang, T., Zanger, U. M., Eichelbaum, M., and Schwab, M. (2001). High-throughput genotyping of thiopurine S-methyltransferase by denaturing HPLC. *Clin. Chem.* 47, 548–555.

Schildcrout, J. S., Denny, J. C., Bowton, E., Gregg, W., Pulley, J. M., Basford, M. A., et al. (2012). Optimizing drug outcomes through pharmacogenetics: a case for preemptive genotyping. *Clin. Pharmacol. Ther.* 92, 235–242. doi: 10.1038/clpt.2012.66

Shuldiner, A. R., Relling, M. V., Peterson, J. F., Hicks, J. K., Freimuth, R. R., Sadee, W., et al. (2013). The pharmacogenomics research network translational pharmacogenetics program: overcoming challenges of real-world implementation. *Clin. Pharmacol. Ther.* 94, 207–210. doi: 10.1038/clpt.2013.59

Sim, S. C., Kacevska, M., and Ingelman-Sundberg, M. (2013). Pharmacogenomics of drug-metabolizing enzymes: a recent update on clinical implications and endogenous effects. *Pharmacogenomics J.* 13, 1–11. doi: 10.1038/tpj.2012.45

Taja-Chayeb, L., Vidal-Millan, S., Gutierrez, O., Ostrosky-Wegman, P., Duenas-Gonzalez, A., and Candelaria, M. (2008). Thiopurine S-methyltransferase gene (TMPT) polymorphisms in a Mexican population of healthy individuals and leukemic patients. *Med. Oncol.* 25, 56–62. doi: 10.1007/s12032-007-9002-6

Teml, A., Schaeffeler, E., Herrlinger, K. R., Klotz, U., and Schwab, M. (2007). Thiopurine treatment in inflammatory bowel disease: clinical pharmacology and implication of pharmacogenetically guided dosing. *Clin. Pharmacokinet.* 46, 187–208. doi: 10.2165/00003088-200746030-00001

Wang, L., Pelleymounter, L., Weinshilboum, R., Johnson, J. A., Hebert, J. M., Altman, R. B., et al. (2010). Very important pharmacogene summary: thiopurine S-methyltransferase. *Pharmacogenet. Genomics* 20, 401–405. doi: 10.1097/FPC.0b013e3283352860

Weinshilboum, R. M., and Sladek, S. L. (1980). Mercaptopurine pharmacogenetics: monogenic inheritance of erythrocyte thiopurine methyltransferase activity. *Am. J. Hum. Genet.* 32, 651–662.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 January 2014; accepted: 04 April 2014; published online: 12 May 2014.
Citation: Almoguera B, Vazquez L, Connolly JJ, Bradfield J, Sleiman P, Keating B and Hakonarson H (2014) Imputation of TPMT defective alleles for the identification of patients with high-risk phenotypes. Front. Genet. ***5****:96. doi: 10.3389/fgene.2014.00096*
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 03 December 2013
doi: 10.3389/fgene.2013.00268

EMR-linked GWAS study: investigation of variation landscape of loci for body mass index in children

Bahram Namjou[1]*, Mehdi Keddache[2,3], Keith Marsolo[3,4], Michael Wagner[3,4], Todd Lingren[3,4], Beth Cobb[1], Cassandra Perry[5], Stephanie Kennebeck[2,3], Ingrid A. Holm[6], Rongling Li[7], Nancy A. Crimmins[2,3], Lisa Martin[2,3], Imre Solti[3,4], Isaac S. Kohane[8] and John B. Harley[1,2,3,9]

[1] Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[2] Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[3] School of Medicine, University of Cincinnati, Cincinnati, OH, USA
[4] Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[5] Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA
[6] Division of Genetics and Genomics, Department of Pediatrics, The Manton Center for Orphan Disease Research, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
[7] National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
[8] Center for Biomedical Informatics, Harvard Medical School and Children's Hospital Informatics Program, Boston, MA, USA
[9] Department of Veteran Affairs Medical Center, Cincinnati, OH, USA

Edited by:
Mariza de Andrade, Mayo Clinic, USA

Reviewed by:
Ayo Priscille Doumatey, National Institute of Health, USA
Kelli K. Ryckman, University of Iowa, USA

****Correspondence:***
Bahram Namjou, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229, USA
e-mail: Bahram.namjou@cchmc.org

Common variations at the loci harboring the fat mass and obesity gene (FTO), MC4R, and TMEM18 are consistently reported as being associated with obesity and body mass index (BMI) especially in adult population. In order to confirm this effect in pediatric population five European ancestry cohorts from pediatric eMERGE-II network (CCHMC-BCH) were evaluated.

Method: Data on 5049 samples of European ancestry were obtained from the Electronic Medical Records (EMRs) of two large academic centers in five different genotyped cohorts. For all available samples, gender, age, height, and weight were collected and BMI was calculated. To account for age and sex differences in BMI, BMI *z*-scores were generated using 2000 Centers of Disease Control and Prevention (CDC) growth charts. A Genome-wide association study (GWAS) was performed with BMI *z*-score. After removing missing data and outliers based on principal components (PC) analyses, 2860 samples were used for the GWAS study. The association between each single nucleotide polymorphism (SNP) and BMI was tested using linear regression adjusting for age, gender, and PC by cohort. The effects of SNPs were modeled assuming additive, recessive, and dominant effects of the minor allele. Meta-analysis was conducted using a weighted *z*-score approach.

Results: The mean age of subjects was 9.8 years (range 2–19). The proportion of male subjects was 56%. In these cohorts, 14% of samples had a BMI $\geq$95 and 28 $\geq$ 85%. Meta analyses produced a signal at 16q12 genomic region with the best result of $p = 1.43 \times 10^{-7}$ [$p_{(rec)} = 7.34 \times 10^{-8}$) for the SNP rs8050136 at the first intron of FTO gene ($z = 5.26$) and with no heterogeneity between cohorts ($p = 0.77$). Under a recessive model, another published SNP at this locus, rs1421085, generates the best result [$z = 5.782$, $p_{(rec)} = 8.21 \times 10^{-9}$]. Imputation in this region using dense 1000-Genome and Hapmap CEU samples revealed 71 SNPs with $p < 10^{-6}$, all at the first intron of FTO locus. When heterogeneity was permitted between cohorts, signals were also obtained in other previously identified loci, including MC4R (rs12964056, $p = 6.87 \times 10^{-7}$, $z = -4.98$), cholecystokinin CCK (rs8192472, $p = 1.33 \times 10^{-6}$, $z = -4.85$), Interleukin 15 (rs2099884, $p = 1.27 \times 10^{-5}$, $z = 4.34$), low density lipoprotein receptor-related protein 1B [LRP1B (rs7583748, $p = 0.00013$, $z = -3.81$)] and near transmembrane protein 18 (TMEM18) (rs7561317, $p = 0.001$, $z = -3.17$). We also detected a novel locus at chromosome 3 at COL6A5 [best SNP = rs1542829, minor allele frequency (MAF) of 5% $p = 4.35 \times 10^{-9}$, $z = 5.89$].

Conclusion: An EMR linked cohort study demonstrates that the BMI-*Z* measurements can be successfully extracted and linked to genomic data with meaningful confirmatory results. We verified the high prevalence of childhood rate of overweight and obesity in our cohort (28%). In addition, our data indicate that genetic variants in the first intron of FTO, a known adult genetic risk factor for BMI, are also robustly associated with BMI in pediatric population.

Keywords: BMI, obesity, polymorphism, GWAS

INTRODUCTION

The electronic MEdical Records and GEnomics (eMERGE) Network, founded in 2007, is a consortium of multiple adult and pediatric institutions developed to explore the utility of DNA bio repositories linked to electronic medical records (EMR) in advancing genomic medicine (McCarty et al., 2011). For each site, the primary site-specific phenotypes have undergone genome-wide association studies (GWAS) with data and results shared through the network. In the pediatric population, however, genetic studies are challenging due to different developmental phase and growth patterns, different spectrums of disease, and unusual rare genetic or congenital abnormalities.

In both adults and children, obesity is a major risk factor for a number of chronic diseases, a steady and continuous rise in prevalence over the past four decades holds serious and ominous medical and economic burdens (Kopelman, 2000). The phenotype is highly heritable. Family and twin studies have shown that between 40 and 70% of the inter-individual variation in obesity can be attributable to genetic factors (Maes et al., 1997). In recent years, large-scale GWAS have identified many loci associated with Body Mass Index (BMI), the most common measure of obesity, but these loci combined explain only 2–4% of the heritability (Speliotes et al., 2010). Thus far, four waves of GWAS studies for BMI identified 32 loci that reached genome-wide significance and unequivocally were associated with BMI in a large meta-analysis performed by the GIANT (Genetic Investigation of ANtropometric Traits) consortium (Frayling et al., 2007; Scuteri et al., 2007; Loos et al., 2008; Willer et al., 2009; Speliotes et al., 2010; Loos, 2012; Mägi et al., 2013). The firstly identified locus, FTO (fat mass and obesity associated gene), has the largest effect on obesity-susceptibility with obesity at a risk of 1.20 fold. Moreover, the frequency of the BMI-increasing allele is high in white Europeans (i.e., 40%). As a consequence, of all 32 BMI-associated loci, the FTO locus explains the largest proportion of the inter-individual variation in BMI (0.34%; Speliotes et al., 2010; Loos, 2012).

In this study, we investigated the genetic association of pediatric BMI using anthropomorphic measures extracted from medical records in a collection of already genotyped samples from two large pediatric cohort repositories (CCHMC and CHB) in order to confirm and identify additional genetic loci.

MATERIALS AND METHODS

STUDY SUBJECTS

Protocols for this study were approved by the Institutional Review Boards (IRBs) at the institutions where participants were recruited. Only those self-reported to have European ancestry were selected for study. The anthropometric measurements of height and weight, as well as age of measurement and gender, were extracted from the EMR. All enrolled participants with measured weight and height on the same day were included. All inconsistent or out of range values were excluded. Out of range was defined as any height or weight values higher or lower than is considered biologically possible according to Centers of Disease Control and Prevention (CDC) growth charts. Children and teens, aged 2 through 19 years old were included based on CDC growth chart requirements. In addition, three patients with the ICD-9 code for Prader Willi syndrome were excluded from final results. After removing the missing data and outliers, out of a total of 5,049 individuals, 2860 samples were included in the study. The demographic distributions of these samples are shown in **Table 1**.

GENOTYPING

High throughput single nucleotide polymorphism (SNP) genotyping was carried out previously in CCHMC and BCH using different Illumina™ or Affymetrix™ platforms (**Table 1**). Quality control (QC) of the data was performed before imputation. In each genotyped cohort, standard QC criteria were met and SNPs were removed if (a) >10% missing genotyping, (b) out of Hardy–Weinberg equilibrium (HWE, $P < 0.001$), or a minor allele frequency (MAF) <1%. Samples with call rate <98% were excluded. Principle component analysis (PCA) was performed to identify outliers and hidden population structure using EIGENSTRAT (Price et al., 2006). Based on examination of the scree plot, the first two PCs were retained and used as covariates during the association analysis in order to adjust for population stratification.

PHENOTYPING

We obtained height and weight measurements from EMR in order to calculate BMI [wt (kg)/(ht(m)^2]. When multiple measurements were available for a subject, the most recent measurement was selected and all inconsistent measures were excluded. BMI *z*-scores and percentiles were generated using the 2000 CDC growth charts (study[1]). These *z*-scores and percentiles account for the age and sex differences in BMI throughout childhood. All data for BMI-*z* scores (−3 to +3, mean = 0) were scaled to positive value (+4, 1–7, mean = 4) to be used as a quantitative trait for the GWAS study. In order to assess the burden of increased body weight on health and estimate the effect size, standard cut-offs were also used, and the tail BMI distribution was considered as a binary phenotype (≥95% as case and ≤20% as control). For the published FTO locus, Phenome wide association study (PheWas) was also performed in which presence or absence of each ICD-9 codes were considered as binary phenotype. Only ICD-9 codes with 50 or more available samples were included (143 codes) in the analysis.

STATISTICAL ANALYSIS

Genome-wide association studies analysis was performed by cohort in PLINK (Purcell et al., 2007) using regression models and adjusting for age, sex, and the first two principal components (PC). For BMI *z*-score, the primary analysis was performed using an additive effect of the minor alleles. However, as previous BMI associations have reported better model fit with different models, recessive and dominant models were subsequently evaluated. For binary phenotypes (dichotomization of BMI using a tails approach and PheWas analyses), allelic association was assessed between cases and controls by chi-square with 1 degree of freedom (df). Allelic odds ratio (OR) and 95% confidence intervals (95% CIs) were obtained. In PheWas analyses, permutation procedure was performed using sample randomization strategy in which case

[1]http://www.cdc.gov/healthyweight/downloads/BMI_group_calculator_English.xls

Table 1 | Demographic distribution of pediatric cohorts under study.

	# Europeans	# After removing outliers and missing values	M/F	Mean age (95% CI)	Array
CHB	741	613	387/226	13.30 (12.97–13.66)	Affymetrix-Axiom
CCHMC	829	696	338/358	10.80 (10.53–11.12)	Omni-5
	657	405	261/144	7.18 (6.73–7.63)	Omni-1
	1270	942	589/353	7.32 (7.03–7.62)	Illumina-610
	1552	204	28/176	13.70 (13.13–14.23)	Affymetrix-6
Total	5049	2860	1603/1257	9.8 (8.67–10.85)	

and control labels are permuted randomly (×10000) in order to obtain empirical p values and to correct for multiple testing.

For specific target regions, imputation-based analyses were performed using the impute2-Gtool pipeline and the publicly available 1000 Genomes Project as the reference haplotype panel composed of 1092 samples (release version 2 of the 1000 Genomes Project Phase I[2] (Howie et al., 2011). For each batch of imputation runs, the standard Markov Chain Monte Carlo (MCMC) algorithm implemented in impute-2, was used with the following threshold criteria (burnin = 10, iteration = 30, and Ne = 20000, buffer = 250 kb). A threshold of 0.90 for the posterior probability of each genotype was then applied for genotype calling and conversion using Gtool. For each imputation run, the overall genotype concordance rate was more than 95%. Additional post imputation filtering were also implemented to remove poorly imputed variants with low concordance rate according to the impute-2 standard protocol (info >0.4; Howie et al., 2011). To graphically display the results, LocusZoom was used (Pruim et al., 2010).

META-ANALYSES

The results from primary analysis in each cohort were assembled to conduct a fixed effects weighted Z meta-analysis using Metal (Willer et al., 2010). This approach controls the differences in phenotype scaling across the studies and weights the signed Z statistics from each study by its sample size (i.e., weighted sum), from which a probability is calculated. The program also applies the genomic control correction to control type I error rates using summary statistics from each cohort. After QC filtrations in each cohort (as described above), meta-analyses were performed on SNP markers that were overlap among all five cohorts. 92670 SNP markers were in this category. At the next step in order to allow heterogeneity between cohorts, we applied a minimum weight of at least 1000 samples for analyses and identify additional effects. 583824 SNP markers were evaluated in this mode. We considered genome-wide significance thresholds of nominal p-value $<10^{-8}$ for any new findings and report all significant results ($p < 0.001$) of previously known loci that concurred with previous publications in terms of strand direction, MAF, and supporting evidence from nearby region. In addition, to describe the presence or absence of excess variation between cohorts, we evaluated the Q-statistic and I^2 as a measure of heterogeneity (Willer et al., 2010).

[2]ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521

RESULTS

The demographic distribution of the European ancestry population under study (**Table 1**) shows that the overall mean age of participants was 9.8 (95% CI = 8.67–10.85) years old with 56% being male. **Table 2** shows the estimated prevalence of overweight and obesity in these pediatric cohorts with a rate of 28% overweight (≥85th %ile) and 14% of obesity (≥95th %ile). This distribution was consistent across all cohorts (**Table 2**).

Genome-wide analyses were conducted within each cohort. Associations between SNPs and BMI assumed an additive genetic model and summary statistics were subsequently used for meta-analysis using a weighted z-score method. After cleaning the data by applying our QC criteria, the ratio of the observed to expected χ^2 test statistic (lambda) was $\lambda = 1.007$ (**Figure 1B**). The results of the meta-analyses of all studies revealed a significant signal of association at 16q12 (Manhattan plot, **Figure 1A**).

The typed SNP rs8050136 at first intron of FTO gene produced consistent evidence of association in all cohorts with the best overall result of ($P = 1.43 \times 10^{-7}$, $z = 5.26$) and with no heterogeneity between study cohorts ($p = 0.77$; **Figure 1A**; **Table 3**). In addition, the allele frequencies and strand alignment are similar across cohorts and consistent with European ancestry. In consistent with previous publications, when mean of BMI-z was stratified by genotype (AA, AC, CC) for SNP (rs8050136), the additive association with risk allele was observed and it is

Table 2 | Summary of BMI-for-age and prevalence of overweight and obese children from the CCHMC-BCH cohorts.

	Male	Female	Total
Number of children assessed	1603	1257	2860
Underweight (<5th %ile)	10	9	10
Normal BMI (5th–85th %ile)	62	63	62
Overweight or obese (≥85th %ile)*	28	28	28
Obese (≥95th %ile)	15	12	14

*Summary of children's BMI-for-age and prevalence of overweight and obesity. *Terminology based on Barlow and the Expert Committee (2007).*

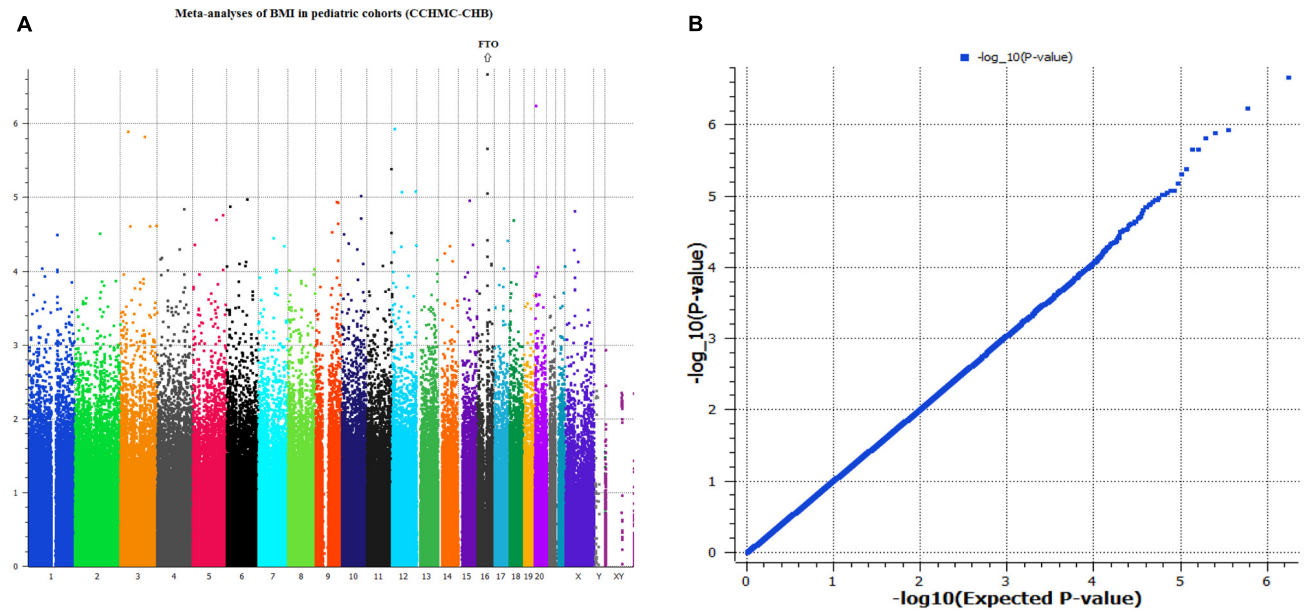


FIGURE 1 | (A,B) Manhattan plot and *Q–Q* plot of SNP markers used for meta-analyses (genomic inflation, $\lambda = 1.007$).

Table 3 | Most associated SNPs with BMI-*z* in CCHMC-BCH pediatric cohorts.

Chr	Position	Gene	SNP	Minor allele	MAF	*Z* score*	$P_{(a_{dd})}$	$P_{(rec)}$	P_{Dom}	*Cochran-q*	I^2 (%)
16	53816275	FTO	rs8050136	A	0.39	5.26	1.43e−07	7.34e−08	0.0008	0.71	0
16	53813367	FTO	rs17817449	G	0.39	5.02	5.56e−07	5.10e−08	0.001	0.49	0
16	53820527	FTO	rs9939609	A	0.39	5.01	6.07e−07	8.53e−08	0.001	0.75	0
16	53800954	FTO	rs1421085	C	0.39	4.65	3.45e−06	8.21e−09	0.02	0.61	0
18	57673799	Near MC4R	rs12964056	A	0.24	−4.98	6.87e−07	0.0004	4.96e−06	0.95	0
3	42299870	CCK	rs8192472	A	0.37	−4.85	1.33e−06	0.0007	0.0002	0.89	0
4	142763570	IL15	rs2099884	T	0.15	4.29	1.27e−05	0.0006	0.0004	0.82	0
2	142855291	LRP1B	rs7583748	G	0.10	−3.81	0.0001	0.03	8.08e−005	0.19	0.41
2	644953	TMEM18	rs7561317	A	0.17	−3.17	0.001	0.09	0.02	0.74	0

**The direction of all effect (weighted z scores) are for the minor alleles. The I^2 inconsistency metric was null to small for all of the markers (I^2 = 0–42%).*

shown in **Figure 2** (risk allele, A). There was 0.4 *z*-score-unit difference in mean of BMI-*z* score between homozygotes with risk and non-risk genotype in our pediatric cohorts (**Figure 2**). We further subdivided all cohorts into two age strata of less than 5 and above 5 years old. In meta-analyses of both strata, the minor allele (A) was associated additively with a higher BMI (**Figure 2**).

Next, we performed imputation-based association followed by conditional analysis to identify independent association in the FTO locus. We identified 71 SNPs at the first intronic region of FTO that were significantly associated with BMI in European ancestry, all of which were of similar magnitude (Z score $4.4 < z < 5.26$, p-values between $10^{-7} < p < 10^{-6}$) and with high linkage disequilibrium (LD) with each other ($r^2 > 0.8$; **Figure 3A**). The haplotype boundaries and recombination rate in this ancestry are also shown in **Figure 3**. Indeed, the SNP rs8050136 was in proxy with other well-known variants in this region including rs9939609 ($r^2 = 0.98$), rs17817449 ($r^2 = 0.99$), rs1421085 ($r^2 = 0.93$) and all resided in the same haplotype and are associated with obesity in adults (4–9). Therefore, no independent effect has been identified. **Table 3** shows the summary of SNP results after regression analyses under additive and recessive models adjusted for age, gender, and PC. No significant difference was observed when the cohort site is included as another covariate (Data not shown). Of note, and only in this intron, more than 10 polymorphic indels were also associated with BMI-*z*, in particular A/AT at chr16:53822169 (MAF = 36%) and TTTC/T at chr16:53829962 (MAF = 35%; $p = 2.41 \times 10^{-5}$, $p = 8.09 \times 10^{-5}$, respectively). We noticed a subtle improvement of overall results using recessive model in our cohorts (**Figure 3B**; **Table 3**). In particular, another published SNP

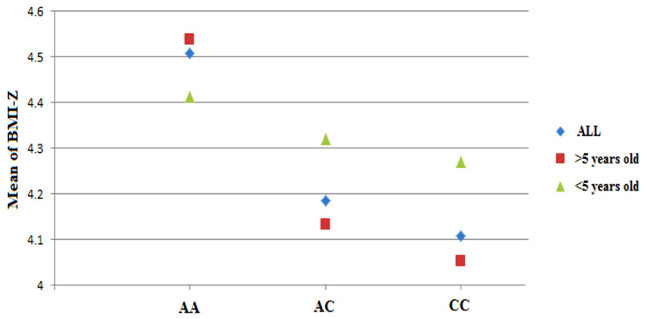


FIGURE 2 | Genotypic correlation of FTO-SNP rs8050136 with mean of BMI-*z* score. The result further divided by age above and less than 5 years old respectively. All BMI-*z* scores (−3 to +3, mean = 0) were transformed to positive value (+4, 1–7, mean = 4).

rs1421085 generated the best result $z = 5.782$, $p_{(\mathrm{Rec})} = 8.21 \times 10^{-9}$ (**Table 2**).

Although BMI is a continuous trait, standard cut-offs were also used to assess the burden of increased body weight on health as a binary phenotype and estimate the effect size. When the tail distribution was considered (>95% as case and <20% as control), an OR of 1.61 (95% CI = 1.31–1.97, $p = 4.04 \times 10^{-6}$) was detected for the best surrogate marker rs8050136, adjusted by age, sex, and PC. The strongest effect size was observed using a recessive model for rs1421085 [(OR = 2.79, 95% CI = 1.89–4.10), $p = 1.83 \times 10^{-7}$].

When heterogeneity was allowed between cohorts, weaker signals were also obtained in loci, such as near the MC4R region (rs12964056, $p = 6.87 \times 10^{-7}$, $z = -4.98$), cholecystokinin CCK (rs8192472, $p = 1.33 \times 10^{-6}$, $z = -4.85$), Interleukin 15 (rs2099884, $p = 1.27 \times 10^{-5}$, $z = 4.34$), low density lipoprotein receptor-related protein 1B [LRP1B (rs7583748, $p = 0.00013$, $z = -3.81$), and near transmembrane protein 18 (TMEM18; rs7561317, $p = 0.001$, $z = -3.17$)], all of which have been previously reported to be associated with obesity or BMI (**Table 3**; 4–10). The imputation result for the MC4R region shows multiple association markers (**Figure 3C**).

We also performed imputation in additional regions of interest. Of note, one new locus that has not been previously reported to be associated with BMI, passed the GWAS significance level in our cohorts ($p < 1.0\mathrm{E}{-10^{-8}}$). The best marker was a relatively infrequent intronic SNP, rs1542829 in the COL6A5 gene in chromosome 3, with overall MAF of 5% that produced a $p = 4.35 \times 10^{-9}$, $z = 5.89$ under an additive model adjusted for age, gender, and PC. The allele frequency of this marker was consistent among cohorts and with CEU-Hapmap data, and it was in HWE. Additional SNP markers in this region, after imputation, produced probabilities at the level of 10^{-5}(**Figure 3D**). Of note, the majority of these SNPs had MAF of less than 10% with less cohesive haplotype boundaries (**Figure 3D**). When we considered the tail distribution (>95% as case and <20% as control) of BMI-*z* as a binary phenotype, an OR of 2.90 (95% CI = 1.93–4.34), $p = 9.03 \times 10^{-8}$ was obtained for this marker with MAF of 9% in cases vs. 3% in controls. Other unreported loci with suggestive associations to pediatric BMI ($10^{-7} < p < 10^{-5}$), include KCNH5 (rs10136789, $p = 4.62 \times 10^{-7}$, $z = 5.05$), APOL5 (rs2016586, $p = 3.26 \times 10^{-}6$, $z = -4.67$), LRRC7 (rs10889850, $p = 1.77 \times 10^{-6}$, $z = -4.78$), and GALNT13 (rs12693973, $p = 1.65 \times 10^{-6}$, $z = -4.80$).

Finally, using ICD-9 diagnostic codes in our collections, we have also performed a Phewas study for the best identified markers in the FTO locus. In this approach, presence or absence of each diagnostic ICD-9 code was included as a binary phenotype, allelic associations were assessed between cases and controls, and the final results were corrected for multiple testing. By this approach, a negative association was detected between cases with hypertrophic cardiomyopathy or valvular structural heart disease and the FTO common risk alleles (ICD-9 codes = 424.0–424.3 and 425) in a subset of 81 cases and 2259 controls. This effect remained significant after 10000 permutations [p(perm) = 0.0009, OR = 0.53 (95% CI 0.37–0.77)]. Suggestive positive associations were also observed between the risk alleles and ICD-9 codes for impaired glucose tolerance test (ICD-9 = 790.2) and myopia (ICD-9 = 367.1; $p < 0.05$), however, this effect did not remain significant after permutation and correcting for multiple testing. It is noteworthy to mention that in our pediatric cohorts, the number of patients with diabetes related diagnostic codes (one of the BMI-related phenotype) was small and not sufficient for independent analysis.

DISCUSSION

In this study, we evaluated BMI in five European ancestry pediatric cohorts with available EMR-linked genotyped data from the CCHMC-BCH eMERGE-II Network site. We successfully utilized anthropomorphic measures to calculate BMI-for-age percentile, derived BMI-*z* scores according to CDC growth charts, performed quantitative trait locus GWAS study and conducted meta analyses. The overall adjusted meta-analysis result of 2860 European samples produced the best signal in the 16q12 genomic region at the first intron of the FTO locus for a cluster of SNPs. The best typed marker rs8050136 produced [$p_{(\mathrm{Rec})} = 7.34 \times 10^{-8}$), $z = 5.26$] and with no heterogeneity between cohorts ($p = 0.77$). When the tail distribution was considered (≥95% as case and ≤20 as control), an OR of 1.61 (95% CI = 1.313–1.965, $p = 4.04 \times 10^{-6}$) was detected. Notably, this OR estimate in our data was relatively higher than previous adult studies (OR = 1.2; Loos, 2012). In any case, considering the effect size (or OR) at the range of 1.2 and the high MAF of the FTO loci (0.40) in Europeans, 700 samples were sufficient for us to achieve an optimum power of 0.8 with a type 1 error level of 0.05. In fact, genetic variants in the first intron of FTO present as the strongest BMI-associated GWAS locus in humans. Over the past few years, the association of the FTO locus has been repeatedly replicated, not only for BMI, but also for obesity risk, body fat percentage, waist circumference, and other obesity-related traits, in particular type II diabetes (Scuteri et al., 2007; Frayling et al., 2007). Recent association with dyslipidemia, hypertension, reduced brain volume, Alzheimer's disease, and dementia has also been reported that could be confounded with obesity and vascular complications (Pausova et al., 2009; Keller et al., 2011).

The functional mechanism of FTO however is still elusive and is currently the subject of intense interest. It is an AlkB-like, Fe(II)-

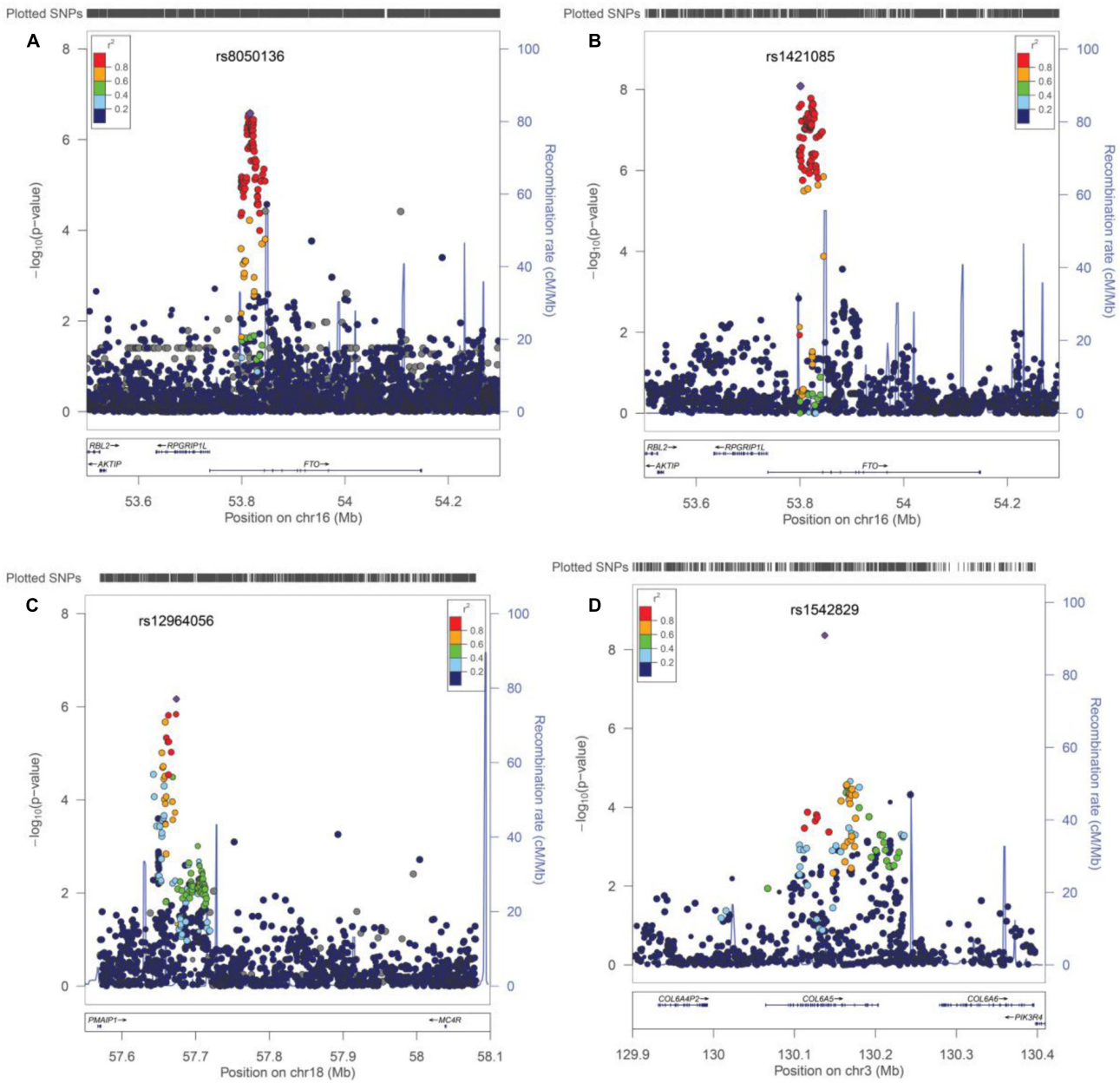


FIGURE 3 | Post imputation results on selected regions. (A) Imputation results and signal at FTO locus contributing to BMI. SNPs are plotted by position in a 0.2 Mb window of chromosome 16 against association with BMI-*z* (−log10 *P*-value). The panel highlights the most significant SNP in a meta-analysis using an additive model. Estimated recombination rates (from HapMap) are plotted in cyan to reflect the local LD structure. The SNPs surrounding the most significant SNP (rs8050135), are color-coded to reflect their LD with this SNP (taken from pairwise r^2 values from the HapMap CEU database, www.hapmap.org). Regional plots were generated using LocusZoom (http://csg.sph.umich.edu/locuszoom). **(B)** Regression results at the FTO locus under recessive model, best marker = rs1421085, p(rec) = 8.21 × 10^{-9}. **(C)** Imputation results near the MC4R locus at chromosome 18, is shown. Best marker rs12964056, p = 6.87 × 10^{-7}, z = −4.98. **(D)** A new effect at the COL6A5 locus in chromosome 3. Best marker rs1542829 p = 4.35 × 10^{-9}, z = 5.889.

and 2-oxoglutarate-dependent nucleic acid demethylase that has been shown to demethylate 3-methylthymine and 3-methyluracil in single-stranded DNA and RNA, respectively (Gerken et al., 2007). A link between FTO demethylase activity and increased fat mass was suggested by recent animal studies. Notably, homozygous mutant *fto*$^{-/-}$ mice show postnatal growth retardation and a significant reduction in adipose tissue and lean body mass, an observation that also was supported by the deleterious mutation Ile367Phe in mouse FTO protein with an impaired activity from a separate study (Church et al., 2009; Gao et al., 2010). In both studies, the leanness of *fto*-deficient mice seems to be the result of increased energy expenditure and systemic sympathetic

activation. Interpreting energy expenditure from these results, however, was challenging because of body composition differences and growth retardation. Overall, these experimental data suggest that inactivation of the FTO gene protects from obesity. On the other hand, mice globally overexpressing FTO are obese, hyperphagic, and exhibit normal energy expenditure when corrected for lean tissue mass (Church et al., 2010). Most human studies also report that obesity-predisposing FTO alleles are associated with increased food intake, but not energy expenditure (Speakman et al., 2008; Haupt et al., 2009). Although the detailed mechanisms are still not clear, this is suggested to be the result of increased expression of FTO in humans, due to cis-regulatory variation in intron 1 of this gene which is consistent with the mouse models, in which decreased levels of FTO cause a lean body habitus (Zabena et al., 2009; Berulava and Horsthemke, 2010). Moreover, a positive correlation of FTO gene expression with other adipocytokine gene expressions, including leptin, perilipin, and visfatin, has also been shown (Haupt et al., 2009).

Consistent with previous studies and because of high LD between FTO intron-1 markers in Europeans (**Figure 2**), conditional analyses didn't produce independent effect; however, we observed a subtle improvement of FTO association using recessive models in comparison to additive models (**Figure 1**; **Table 2**). This effect was unique to the FTO locus, while use of a dominant model was best near the MC4R region (**Table 3**). Indeed, between-study variations in regard to the optimal inheritance model of the FTO polymorphisms have been noticed previously. In one study, eight different meta-analysis results of BMI in women with polycystic ovary syndromes were systematically reviewed and the recessive model was found to fit best in half of them and the additive model worked best in the rest (Wojciechowski et al., 2012). In our pediatrics population in which we had an enrichment of earlier age of obesity with a prevalence rate of 28%, a recessive pattern may be more relevant in comparison to the general population. Under this model, another intron-1 marker, rs1421085, produced the best result $p_{(\mathrm{Rec})} = 8.21 \times 10^{-9}$, $z = 5.78$ (**Table 3**; $r^2 = 0.93$ with rs8050136). Indeed, the variant rs1421085 is particularly interesting. It is located within a highly conserved element and the risk allele C has been predicted to substantially reduce binding affinity for CUX1, a transcription factor implicated in the regulation of FTO (Stratigopoulos et al., 2011; Peters et al., 2013).

Limited evidence suggests that the cross-sectional FTO association with BMI varies by age. Specifically, at early ages (up to 5–7 years), the association between common variation at FTO and BMI appears to be reduced in magnitude (Hardy et al., 2010). Longitudinal twin studies also suggest that with increasing age loci such as FTO may be able to exert a greater effect on BMI which depends on the influence of shared environmental effect and the timing of adiposity rebounds (Haworth et al., 2008; Sovio et al., 2011). In our cohorts when we divide all samples into two strata of less than 5 and above 5 years old, we didn't observe any opposing effect, however a higher magnitude was identified for strata above 5 years old (OR = 1.72) (**Figure 2**). Larger sample size with detailed longitudinal data would seem to be needed to fully elucidate this correlation.

Furthermore, in the context of rare and severe phenotypes, recently, in a large Palestinian Arab consanguineous multiplex family with nine affected, a homozygous R316Q enzyme inactivating mutation in the FTO gene, resulted in a broad spectrum of clinical manifestations including severe intrauterine growth retardation, severe microcephaly, and death from infection before the age of three (Boissel et al., 2009). Of note, six out of eight cases had structural heart defect with cardiomyopathy. This would appear consistent with our unique observation of a negative association of obesity-predisposing FTO alleles with cardiomyopathy [$p_{(\mathrm{corr})} = 0.0009$, OR = 0.53 (95% CI 0.37–0.77)] that could result in lower expression of FTO; although further studies with larger sample sizes are necessary to confirm or refute this finding. In fact, to our knowledge, no SNP in intron 1 of FTO has been previously associated with any trait unrelated to BMI. Recently, an independent effect at the eighth intron of the FTO locus has been reported to be associated with melanoma (GenoMEL Consortium et al., 2013). The best marker, rs16953002, was replicated using 12,313 cases and 55,667 controls of European ancestry in a study conducted by the GenoMEL consortium (combined $P = 3.6 \times 10^{-12}$, OR = 1.16). Notably, in their study, none of the BMI related SNPs in intron 1 were associated with melanoma. Similarly, in our collection, there was no effect observed at the eighth intron of the FTO gene with BMI ($p = 0.54$ for rs16953002, $r^2 < 0.01$ with rs8050136). This suggests independent functions and genetic risks for FTO that broaden the existing paradigm and identify distinct pathogenic effects (GenoMEL Consortium et al., 2013).

In this report, we have supported association in other previously reported BMI loci, in particular loci near the MC4R region (rs12964056, $p = 6.87 \times 10^{-7}$, $z = -4.98$), cholecystokinin CCK (rs8192472, $p = 1.33 \times 10^{-6}$, $z = -4.85$), Interleukin 15 (rs2099884, $p = 1.27 \times 10^{-5}$, $z = 4.34$), low density lipoprotein receptor-related protein 1B [LRP1B (rs7583748, $p = 0.00013$, $z = -3.81$)] and near transmembrane protein 18 (TMEM18; rs7561317, $p = 0.001$, $z = -3.17$; **Table 2**). Because of their lower effect on BMI and obesity risk (OR ∼1.10, in adult meta-data) and lower allele frequency, the identification of these loci requires a quadrupling of the sample size in a random population. Finding all of these loci in our pediatric collections, despite limited sample size, indicate the enrichment of the genetic signal to noise ratio given the shorter amount of time that environment has had an effect.

Additionally, we have detected a new unreported signal at chromosome 3 (Col6A5) (best SNP is rs1542829, MAF of 5% $p = 4.35 \times 10^{-9}$, $z = 5.89$). This marker produced an OR of 2.90 when the tail distributions (>95% as case and <20% as control, 386 cases, and 572 controls) was considered as binary phenotype. Considering this level of OR (2.90), even with MAF of 5%, 500 samples were sufficient for us to achieve the extraordinary power (0.99) with a type 1 error level of 0.05. This could be considered as one of the rare obesity risk loci that we expect to detect in these special cohorts. The α5-containing collagen VI (Col6A5, COL29A1), belongs to the class of collagens containing von Willebrand factor type A domains. These collagens form filaments with globular domains containing vWA motifs, which are involved in protein-ligand interactions for the organization of tissue architecture and

cell adhesion. Collagen VI is a major extracellular matrix (ECM) protein with a critical role in maintaining skeletal muscle functional integrity. It has been suggested that type VI collagen is a fibrotic component that restricts adipose tissue expandability. In humans, Col6A3 gene expression in adipose tissues was found to correlate with visceral adipose tissue mass and pro-inflammatory gene expression (Pasarica et al., 2009). Mutations in different families of this gene have also been associated with myopathy and muscular dystrophy. Recently, it has been shown that COL6A5 is involved in adhesion at myotendinous and dermal–epidermal junctions (Sabatelli et al., 2012). Different polymorphisms in or near this gene have been linked to atopic dermatitis and eczema, but with contradictory reports (Söderhäll et al., 2007; Naumann et al., 2011). In our cohorts, the number of samples with Atopic dermatitis and related conditions (ICD-9 = 691) was only 49 with a trend of association for published SNP rs7629719 ($p = 0.14$); adjusting the results based on presence or absence of atopy, didn't have any effect on overall BMI associations. A larger sample size is necessary to further elucidate this coexistent condition and to determine whether COL6A5 has any role in obesity related conditions.

Four additional novel loci with homogenous but suggestive associations ($10^{-7} < p < 10^{-5}$) to childhood BMI were also identified in this study. These include KCNH5 (rs10136789, $p = 4.62 \times 10^{-7}$, $z = 5.05$), a voltage-gated potassium channel with various function in neurotransmitter regulation, hormone release, cardiac function, and cell volume; APOL5 (rs2016586, $p = 3.26 \times 10^{-6}$, $z = -4.67$), a component of high-density lipoprotein with a potential role in lipid metabolism; LRRC7 or Densin (rs10889850, $p = 1.77 \times 10^{-6}$, $z = -4.78$) a core component of post-synaptic densities and GALNT13 (rs12693973, $p = 1.65\text{x}10^{-6}$, $z = -4.80$) a member of the UDP-*N*-acetyl-alpha-D-galactosamine and a major enzyme responsible for the synthesis of *O*-glycan. Independent cohorts are necessary to confirm these preliminary suggestive findings and their importance in childhood obesity.

Despite the limitations of using an EMR-derived data set for analysis of secondary phenotypes including errors in data extractions, discordant time of sampling, and underlying coexistent conditions of our pediatric cohorts, we demonstrate that a strong signal, larger than seen in adult populations is detectable. We have removed all inconsistent data and outliers to the best of our ability. We have also excluded infants and those less than 2 years old because of the complexity of growth chart pattern and many potential maternal effects on infants from perinatal periods. In addition we assessed the distribution of BMI-*z* in the whole population with a large sample size as a quantitative trait rather than attempting to identify limited cases and controls. From the statistical standpoint, quantitative traits usually are preferred in meta-GWAS studies because they improve power to detect a genetic effect and often have a more interpretable outcome (Bush and Moore, 2012). Furthermore, BMI is a highly heritable trait in humans and, as mentioned above, up to 70% of the inter-individual variation in obesity can be attributable to genetic factors *per se* (Maes et al., 1997); therefore, given the strong genetic confirmation described here, indeed, we managed to repurpose the genotyping data collected for the analyses of another phenotype and successfully find association between the new phenotype and genotypic data.

In summary, using the EMR-linked genotyped data, we have confirmed association of several previously known BMI loci, in particular with the FTO gene [OR of 1.61 (95% CI = 1.31–1.97)]. Our data also support the importance of variants at the FTO locus in childhood obesity and with saturation of an earlier age of onset, these data point to a closer functional variant in this locus.

ACKNOWLEDGMENTS

We are grateful to the individuals who participated in this study. We thank the genotyping core facilities in both academic centers (CCHMC-BCH) and our colleagues who facilitated the genotyping and recruitment of subjects. This work was supported by a grant from the National Human Genomic Research Institute: 1U01HG006828. "TL and IS were partially supported by NIH grants 5R00LM010227-04 and 1R21HD072883-01."

REFERENCES

Boissel, S., Reish, O., Proulx, K., Kawagoe-Takaki, H., Sedgwick, B., Yeo, G. S., et al. (2009). Loss-of-function mutation in the dioxygenase-encoding FTO gene causes severe growth retardation and multiple malformations. *Am. J. Hum. Genet.* 85, 106–111. doi: 10.1016/j.ajhg.2009.06.002

Berulava, T., and Horsthemke B. (2010). The obesity-associated SNPs in intron 1 of the FTO gene affect primary transcript levels. *Eur. J. Hum. Genet.* 18, 1054–1056. doi: 10.1038/ejhg.2010.71

Bush, W. S., and Moore, J. H. (2012). Chapter 11: genome-wide association studies. *PLoS Comput. Biol.* 8:e1002822. doi: 10.1371/journal.pcbi.1002822

Barlow, S. E., and the Expert Committee. (2007). Expert committee recommendations regarding the prevention, assessment, and treatment of child and adolescent overweight and obesity: summary report. *Pediatrics* 120, S164–S192. doi: 10.1542/peds.2007-2329C

Church, C., Moir, L., McMurray, F., Girard, C., Banks, G. T., Teboul, L., et al. (2010). Overexpression of Fto leads to increased food intake and results in obesity. *Nat. Genet.* 42, 1086–1092. doi: 10.1038/ng.713

Church, C., Lee, S., Bagg, E. A., McTaggart, J. S., Deacon, R., Gerken, T., et al. (2009). A mouse model for the metabolic effects of the human fat mass and obesity associated FTO gene. *PLoS Genet.* 5:e1000599. doi: 10.1371/journal.pgen.1000599

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316, 889–894. doi: 10.1126/science.1141634

Gerken, T., Girard, C. A., Tung, Y. C. L., Webby, C. J., Saudek, V., Hewitson, K. S., et al. (2007). The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. *Science* 318, 1469–1472. doi: 10.1126/science.1151710

Gao, X., Shin, Y. H., Li, M., Wang, F., Tong, Q., and Zhang, P. (2010). The fat mass and obesity associated gene FTO functions in the brain to regulate postnatal growth in mice. *PLoS ONE* 5: e14005. doi: 10.1371/journal.pone.0014005

GenoMEL Consortium, Iles, M. M., Law, M. H., Stacey, S. N., Han, J., Fang, S., et al. (2013). A variant in FTO shows association with melanoma risk not due to BMI. *Nat. Genet.* 45, 428–432. doi: 10.1038/ng.2571

Haupt, A., Thamer, C., Staiger, H., Tschritter, O., Kirchhoff, K., Machicao, F., et al. (2009). Variation in the FTO gene influences food intake but not energy expenditure. *Exp. Clin. Endocrinol. Diabetes* 117, 194–197. doi: 10.1055/s-0028-1087176

Hardy, R., Wills, A. K., Wong, A., Elks, C. E., Wareham, N. J., Loos, R. J., et al. (2010). Life course variations in the associations between FTO and MC4R gene variants and body size. *Hum. Mol. Genet.* 19, 545–552. doi: 10.1093/hmg/ddp504

Haworth, C. M., Carnell, S., Meaburn, E. L., Davis, O. S., Plomin, R., and Wardle, J. (2008). Increasing heritability of BMI and stronger associations with the FTO gene over childhood. *Obesity* 16, 2663–2668. doi: 10.1038/oby.2008.434

Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3 (Bethesda)* 1, 457–470. doi: 10.1534/g3.111.001198

Keller, L., Xu, W., Wang, H. X., Winblad, B., Fratiglioni, L., and Graff, C. (2011). The obesity related gene, FTO, interacts with APOE, and is associated with Alzheimer's

disease risk: a prospective cohort study. *J. Alzheimers Dis.* 23, 461–469. doi: 10.3233/JAD-2010-101068.

Kopelman, P. G. (2000). Obesity as a medical problem. *Nature* 404, 635–643.

Loos, R. J. (2012). Genetic determinants of common obesity and their value in prediction. *Best Pract. Res. Clin. Endocrinol. Metab.* 26, 211–26. doi: 10.1016/j.beem.2011.11.003

Loos, R. J., Lindgren, C. M., Li, S., Wheeler, E., Zhao, J. H., Prokopenko, I., et al. (2008). Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat. Genet.* 40, 768–775. doi: 10.1038/ng.140

Maes, H. H., Neale, M. C., and Eaves, L. J. (1997). Genetic and environmental factors in relative body weight and human obesity. *Behav. Genet.* 27, 325–351. doi: 10.1023/A:1025635913927

McCarty, C. A., Chisholm, R. L., Chute, C. G., Kullo, I. J., Jarvik, G. P., Larson, E. B., et al. (2011). The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics* 4:13.doi: 10.1186/1755-8794-4-13.

Mägi, R., Manning, S., Yousseif, A., Pucci, A., Santini, F., Karra, E., et al. (2013). Contribution of 32 GWAS-identified common variants to severe obesity in European adults referred for Bariatric surgery. *PLoS ONE* 7:e70735. doi: 10.1371/journal.pone.0070735

Naumann, A., Söderhäll, C., Fölster-Holst, R., Baurecht, H., Harde, V., Müller-Wehling, K., et al. (2011). A comprehensive analysis of the COL29A1 gene does not support a role in eczema. *J. Allergy Clin. Immunol.* 127, 1187.e7–1194.e7. doi: 10.1016/j.jaci.2010.12.1123

Pasarica, M., Gowronska-Kozak, B., Burk, D., Remedios, I., Hymel, D., Gimble, J., et al. (2009). Adipose tissue collagen VI in obesity. *J. Clin. Endocrinol. Metab.* 94, 5155–5162. doi: 10.1210/jc.2009-0947

Pausova, Z., Syme, C., Abrahamowicz, M., Xiao, Y., Leonard, G. T., Perron, M., et al. (2009). A common variant of the FTO gene is associated with not only increased adiposity but also elevated blood pressure in French Canadians. *Circ. Cardiovasc. Genet.* 2, 260–269. doi: 10.1161/CIRCGENETICS.109.857359

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. doi: 10.1038/ng1847

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., et al. (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26, 2336–2337. doi: 10.1093/bioinformatics/btq419

Peters, U., North, K. E., Sethupathy, P., Buyske, S., Haessler, J., Jiao, S., et al. (2013). A systematic mapping approach of 16q12.2/FTO and BMI in more than 20,000 African Americans narrows in on the underlying functional variation: results from the Population Architecture using Genomics and Epidemiology (PAGE) study. *PLoS Genet.* 9:e1003171. doi: 10.1371/journal.pgen.1003171

Stratigopoulos, G., LeDuc, C. A., Cremona, M. L., Chung, W. K., and Leibel, R. L. (2011). Cut-like homeobox 1 (CUX1) regulates expression of the fat mass and obesityassociated and retinitis pigmentosa GTPase regulator-interacting protein-1-like (RPGRIP1L) genes and coordinates leptin receptor signaling. *J. Biol. Chem.* 286, 2155–2170. doi: 10.1074/jbc.M110.188482

Söderhäll, C., Marenholz, I., Kerscher, T., Rüschendorf, F., Esparza-Gordillo, J., Worm, M., et al. (2007). Variants in a novel epidermal collagen gene (COL29A1) are associated with atopic dermatitis. *PLoS Biol.* 5:e242. doi: 10.1371/journal.pbio.0050242

Sabatelli, P., Gualandi, F., Gara, S. K., Grumati, P., Zamparelli, A., Martoni, E., et al. (2012). Expression of collagen VI a5 and a6 chains in human muscle and in Duchenne muscular dystrophy-related muscle fibrosis. *Matrix Biol.* 31, 187–196. doi: 10.1016/j.matbio.2011.12.003

Sovio, U., Mook-Kanamori, D. O., Warrington, N. M., Lawrence, R., Briollais, L., Palmer, C. N., et al. (2011). Association between common variation at the FTO locus and changes in body mass index from infancy to late childhood: the complex nature of genetic association through growth and development. *PLoS Genet.* 7:e1001307. doi: 10.1371/journal.pgen.1001307

Speakman, J. R., Rance, K. A., and Johnstone, A. M. (2008). Polymorphisms of the FTO gene are associated with variation in energy intake, but not energy expenditure. *Obesity (Silver Spring)* 16, 1961–1965. doi: 10.1038/oby.2008.318

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* 42, 937–948. doi: 10.1038/ng.686

Scuteri, A., Sanna, S., Chen, W. M., Uda, M., Albai, G., Strait, J., et al. (2007). Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet.* 3:e115. doi: 10.1371/journal.pgen.0030115

Wojciechowski, P., Lipowska, A., Rys, P., Ewens, K. G., Franks, S., Tan, S., et al. (2012). Impact of FTO genotypes on BMI and weight in polycystic ovary syndrome: a systematic review and meta-analysis. *Diabetologia.* 55, 2636–2645. doi: 10.1007/s00125-012-2638-6

Willer, C. J., Li, Y., Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190–2191. doi: 10.1093/bioinformatics/btq340

Willer, C. J., Speliotes, E. K., Loos, R. J., Li, S., Lindgren, C. M., Heid, I. M., et al. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* 41, 25–34. doi: 10.1038/ng.287

Zabena, C., González-Sánchez, J. L., Martínez-Larrad, M. T., Torres-García, A., Alvarez-Fernández-Represa, J., Corbatón-Anchuelo, A., et al. (2009). The FTO obesity gene. Genotyping and gene expression analysis in morbidly obese patients. *Obes. Surg.* 19, 87–95. doi: 10.1007/s11695-008-9727-0

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 October 2013; paper pending published: 31 October 2013; accepted: 16 November 2013; published online: 03 December 2013.

Citation: Namjou B, Keddache M, Marsolo K, Wagner M, Lingren T, Cobb B, Perry C, Kennebeck S, Holm IA, Li R, Crimmins NA, Martin L, Solti I, Kohane IS and Harley JB (2013) EMR-linked GWAS study: investigation of variation landscape of loci for body mass index in children. Front. Genet. ***4****:268. doi: 10.3389/fgene.2013.00268*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

METHODS ARTICLE
published: 29 April 2014
doi: 10.3389/fgene.2014.00095

Using previously genotyped controls in genome-wide association studies (GWAS): application to the Stroke Genetics Network (SiGN)

Braxton D. Mitchell[1,2]*, Myriam Fornage[3], Patrick F. McArdle[1], Yu-Ching Cheng[1,2], Sara L. Pulit[4], Quenna Wong[5], Tushar Dave[1], Stephen R. Williams[6,7], Roderick Corriveau[8], Katrina Gwinn[8], Kimberly Doheny[9], Cathy C. Laurie[5], Stephen S. Rich[6] and Paul I. W. de Bakker[4,10], on behalf of the Stroke Genetics Network (SiGN)

[1] *Department of Medicine and Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD, USA*
[2] *Veterans Administration Medical Center, Baltimore, MD, USA*
[3] *Department of Medicine, University of Texas Health Science Center, Houston, TX, USA*
[4] *Department of Medical Genetics, University Medical Center Utrecht, Utrecht, Netherlands*
[5] *Department of Biostatistics, University of Washington, Seattle, WA, USA*
[6] *School of Medicine, Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA*
[7] *School of Medicine, Cardiovascular Research Center, University of Virginia, Charlottesville, VA, USA*
[8] *National Institute of Neurological Disorders and Stroke, Bethesda, MD, USA*
[9] *Center for Inherited Disease Research, Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA*
[10] *Department of Epidemiology, University Medical Center Utrecht, Utrecht, Netherlands*

Edited by:
Marylyn D. Ritchie, The Pennsylvania State University, USA

Reviewed by:
David Fardo, University of Kentucky, USA
Andrew DeWan, Yale School of Public Health, USA

****Correspondence:***
Braxton D. Mitchell, Department of Medicine, University of Maryland School of Medicine, 302 MSTF, Baltimore, MD 21201, USA
e-mail: bmitchel@medicine.umaryland.edu

Genome-wide association studies (GWAS) are widely applied to identify susceptibility loci for a variety of diseases using genotyping arrays that interrogate known polymorphisms throughout the genome. A particular strength of GWAS is that it is unbiased with respect to specific genomic elements (e.g., coding or regulatory regions of genes), and it has revealed important associations that would have never been suspected based on prior knowledge or assumptions. To date, the discovered SNPs associated with complex human traits tend to have small effect sizes, requiring very large sample sizes to achieve robust statistical power. To address these issues, a number of efficient strategies have emerged for conducting GWAS, including combining study results across multiple studies using meta-analysis, collecting cases through electronic health records, and using samples collected from other studies as controls that have already been genotyped and made publicly available (e.g., through deposition of de-identified data into dbGaP or EGA). In certain scenarios, it may be attractive to use already genotyped controls and divert resources to standardized collection, phenotyping, and genotyping of cases only. This strategy, however, requires that careful attention be paid to the choice of "public controls" and to the comparability of genetic data between cases and the public controls to ensure that any allele frequency differences observed between groups is attributable to locus-specific effects rather than to a systematic bias due to poor matching (population stratification) or differential genotype calling (batch effects). The goal of this paper is to describe some of the potential pitfalls in using previously genotyped control data. We focus on considerations related to the choice of control groups, the use of different genotyping platforms, and approaches to deal with population stratification when cases and controls are genotyped across different platforms.

Keywords: genome-wide association study, case-control study, genetic association study, population stratification, power

INTRODUCTION

Genome-wide association studies (GWAS) have been widely used in recent years as a tool for identifying susceptibility loci for a number of complex human traits and, in particular, multifactorial diseases. Indeed, the NHGRI-maintained Catalog of Published Genome-Wide Association Studies includes 1788 publications and 12,329 SNP (single nucleotide polymorphism)-trait associations as of 1/10/2013 (http://www.genome.gov/gwastudies/) (Hindorff et al., 2009). With few exceptions, the associated loci have small effect sizes, and large sample sizes were required to detect them.

One popular approach to increase sample size and power for GWAS has been to combine information (either individual-level data or summary statistics) across multiple studies through meta-analysis (Panagiotou et al., 2013). As more and more genotype data are being made publicly available through various databases such as the database of Genotypes and Phenotypes (dbGaP) or the European Genome-phenome Archive (EGA), an

alternative approach to increase the statistical power of a study at no extra cost is to devote available clinical and genotyping resources almost entirely to cases and use publicly available data from already genotyped samples as controls. Using available controls may be particularly attractive for registry-based studies from which a large number of cases can be rapidly identified. While this strategy has the obvious benefit of allocating scarce resources toward genotyping a larger number of cases, it can also introduce potential bias into the experimental design leading to spurious associations if not applied carefully. For example, control populations must be comparable to cases in terms of ancestry so that any allele frequency differences observed between cases and controls can be attributed to disease susceptibility loci and not just to differences in ancestral background between the two populations (i.e., population stratification). Ascertainment of the population to be used as controls is also critical. For example, a study of low-density lipoprotein (LDL) concentration could recruit participants at high LDL (a risk factor for stroke) as well as low LDL (not a risk factor); thus, inclusion of all participants as controls could introduce a different bias in the comparison with cases. Equally important is the requirement that genotyping quality and allele calls be comparable between the two groups. Even minor differences in genotype calling, possibly attributable to a laboratory or technician bias, may translate into subtle but systematic differences in allele frequencies between cases and controls that can result in false positive associations.

The goal of this paper is to describe some of the potential pitfalls in using previously genotyped control data. To provide context for these discussions we present as an example the Stroke Genetics Network (SiGN), an international collaboration initiated to carry out a GWAS of ischemic stroke and stroke subtypes that utilizes site-collected cases and already genotyped controls for nearly all sites. We focus on considerations related to the choice of control groups, the use of different genotyping platforms, and approaches to deal with population stratification when cases and controls are genotyped on different platforms.

OVERVIEW OF THE STROKE GENETICS NETWORK (SiGN)

The SiGN was initiated in 2009 to carry out a GWAS of ischemic stroke and stroke subtypes using previously collected DNA samples from multiple centers throughout the US and Europe. These centers included 19 sites contributing 9789 cases to be genotyped. Because of the well-recognized heterogeneity within ischemic stroke, a key feature of SiGN was its focus on standardizing the assignment of stroke subtypes (presumed etiology) for the purpose of performing subtype-specific association analyses. In order to increase the sample size, the decision was made to channel resources into genotyping as many cases as possible and use publicly available control genotypic data wherever possible. A detailed description of the design of SiGN has been previously published, including collection of stroke cases at each study site and the standardizing procedures for assigning stroke subtype (Meschia et al., 2013).

Briefly, stroke research centers with carefully phenotyped ischemic stroke cases were invited to join SiGN and have their stroke cases genotyped using an existing GWAS array. The three requirements for joining SiGN were (1) that the stroke research center have at least 100 cases with DNA immediately available for genotyping, (2) that participating sites must have informed consent on the participants to permit genotypes to be deposited into dbGaP, and (3) that sufficient imaging and additional clinical information had been collected to allow assignment of stroke subtype by Causative Classification of Stroke (CCS) methodology (Ay et al., 2007). CCS phenotyping was performed under a standardized protocol using a web-based system (Meschia et al., 2013).

As indicated in **Table 1**, the 19 participating sites contributed a total of 11,033 samples for genotyping. With the exception of two sites (Leuven, Belgium and Krakow, Poland) all sites provided cases only. The decision was made to genotype both cases and controls from Leuven and Krakow because of the difficulty in locating previously genotyped controls from those areas.

Genotyping of SiGN cases was performed at the Center for Inherited Disease Research (CIDR) in Baltimore, Maryland, using the Illumina HumanOmni 5M Exome genotyping array. This array consists of a total of 4,511,703 variants, including 1,084,398 (24%) "rs" (refSNP) SNPs, 3,178,220 (70%) "kgp" (1000 Genomes) SNPs, 231,910 (5%) "exm" (exome) SNPs, and 17,175 (0.4%) other SNPs.

IMPROVEMENT IN POWER IN SiGN BY PREFERENTIALLY GENOTYPING CASES

There may be multiple reasons to consider utilizing already genotyped control groups for a genetic association study. Foremost among these is the increase in sample size of cases for the same genotyping budget to allow detection of variants with smaller effect sizes, assuming a sufficient number of genotyped controls. Within the context of SiGN, we contrasted the power to detect stroke-associated loci using the strategy of genotyping cases only (with already genotyped controls) vs. genotyping a comparable number of cases and controls at each site. The results of these analyses are shown in **Figure 1** for a range of allele frequencies and an alpha level of $p = 5 \times 10^{-8}$. Sample size estimates are guided by the SiGN genotyping budget of ∼11,000 subjects. Power is shown for three sets of results: (1) genotyping 11,000 cases and utilizing 27,000 previously genotyped controls (as per SiGN); (2) genotyping 5500 cases and 5500 controls and utilizing no previously genotyped controls; and (3) genotyping 5500 cases and 5500 controls but also utilizing an additional 21,500 previously genotyped controls for a total of 27,000. Shown in **Figure 1** are the minimal odds ratios detectable at 80% power at a genome-wide significance alpha level of $p = 5 \times 10^{-8}$.

As indicated in **Figure 1**, substantially lower odds ratios can be detected at 80% for sample 1, which includes 11,000 genotyped cases and 27,000 previously genotyped controls, vs. sample 2, which includes only 5500 genotyped cases and 5500 genotyped controls. While much of the gain in power seen in sample 1 comes from the increased number of controls, sample 3 shows that there remains a sizable increase in power in sample 1 that is attributable to genotyping more cases even when the same number of controls is used (e.g., detectable odds ratios of 1.11–1.18 across a range of minor allele frequencies in sample 1 vs. 1.14–1.23 in sample 3). One caveat about applying power calculations to data that includes publicly available controls is that if controls have

Table 1 | Ischemic stroke cases genotyped as part of the SiGN study and previously genotyped control groups, according to study site*.

Site	Location	Genotype platform	Cases (*n*)	Controls (*n*)
CASES GENOTYPED THROUGH SiGN (US SITES)				
GASROS	Boston, USA	Illumina HumanOmni 5M Exome	470	
GCNKSS	Greater Cincinnati region, USA	Illumina HumanOmni 5M Exome	499	
ISGS	Multi-center, USA	Illumina HumanOmni 5M Exome	187	
MCISS	New Jersey, USA	Illumina HumanOmni 5M Exome	630	
MIAMISR	Miami, USA	Illumina HumanOmni 5M Exome	299	
NHS	National sample, USA	Illumina HumanOmni 5M Exome	316	
NOMAS(S)	Manhattan, USA	Illumina HumanOmni 5M Exome	363	
REGARDS	National sample, USA	Illumina HumanOmni 5M Exome	311	
SPS3	Multi-center; USA; Latin America, Spain	Illumina HumanOmni 5M Exome	962	
SWISS	Multi-center, USA	Illumina HumanOmni 5M Exome	271	
WHI	National sample, USA	Illumina HumanOmni 5M Exome	458	
WUSTL	St. Louis, USA	Illumina HumanOmni 5M Exome	455	
CASES GENOTYPED THROUGH SiGN (INTERNATIONAL SITES)				
BASICMAR	Barcelona, Spain	Illumina HumanOmni 5M Exome	930	
BRAINS	London, England	Illumina HumanOmni 5M Exome	114	
GRAZ	Graz, Austria	Illumina HumanOmni 5M Exome	639	
KRAKOW	Krakow, Poland	Illumina HumanOmni 5M Exome	952	776
LEUVEN	Leuven, Belgium	Illumina HumanOmni 5M Exome	482	468
LUND	Lund, Sweden	Illumina HumanOmni 5M Exome	651	
SAHLSIS	Gothenburg, Sweden	Illumina HumanOmni 5M Exome	800	
PREVIOUSLY GENOTYPED CONTROL GROUPS				
HABC	Multi-center, USA	Illumina 1M-Duo		2802
HRS	Multi-center, USA	Illumina HumanOmni 2.5M		12507
OAI	Multi-center, USA	Illumina HumanOmni 2.5M		4011
ADHD	Barcelona, Spain	Illumina HumanOmni 1M		435
GRAZ	Graz, Austria	Illumina 610		829
INMA	Barcelona, Spain	Illumina HumanOmni 1M		1061
KORA	Southern Germany	Illumina Human 550		820
WTCCC	United Kingdom	Illumina 660		5186

**SiGN cases genotyped at the Center for Inherited Diseases (CIDR) on the Illumina HumanOmni 5M Exome array.*

not been screened for the absence of disease, there is the potential for misclassification and a subsequent loss in power. Potential misclassification was not taken into account in the power calculations above. We further note that such misclassification bias will be more pronounced for common diseases. The power calculations provided also assume equivalent type 1 error rates across the three samples—i.e., no inflation of type 1 error rates introduced by use of publicly available controls.

CHOICE OF ALREADY GENOTYPED CONTROL GROUPS

An important consideration in the design of case-control studies is that cases and controls come from comparable underlying populations so that any differences observed between the groups can be attributed to the exposure under study and not to other unmeasured factors that might differ between the groups (i.e., confounding). Compared to other types of epidemiologic studies, however, genetic association studies are well-suited for utilizing already available control groups because of the limited role of confounding in genetic epidemiology studies. When germ-line variation in DNA sequence is the measured exposure of interest, confounding is limited to the presence of population stratification, that is, the ancestral differences between cases and controls. Choosing already genotyped controls from a similar ancestral background as the cases is thus highly important. Fortunately, the high density of coverage of modern SNP platforms makes this a feature that can be empirically investigated from the data themselves without making any assumptions.

The multicenter design of SiGN that included cases with diverse ancestral origins required the inclusion of multiple control groups. In addition to the samples included for GWAS genotyping through CIDR, SiGN also included cases previously genotyped on multiple genotyping arrays and platforms (see **Table 1**). To reduce variability between cases and controls that could be introduced solely from artifacts related to genotyping platform, the availability of potential controls was limited to those genotyped on a platform believed to be compatible with the Illumina HumanOmni 5M Exome genotyping array used to genotype the cases. This decision limited potential controls groups to sets that had previously been genotyped on a compatible Illumina array and were available to the Network upon request.

Three different multicenter studies were identified for SiGN to serve as controls for the cases from the US sites: The Health and Retirement Study (Juster and Suzman, 1995)

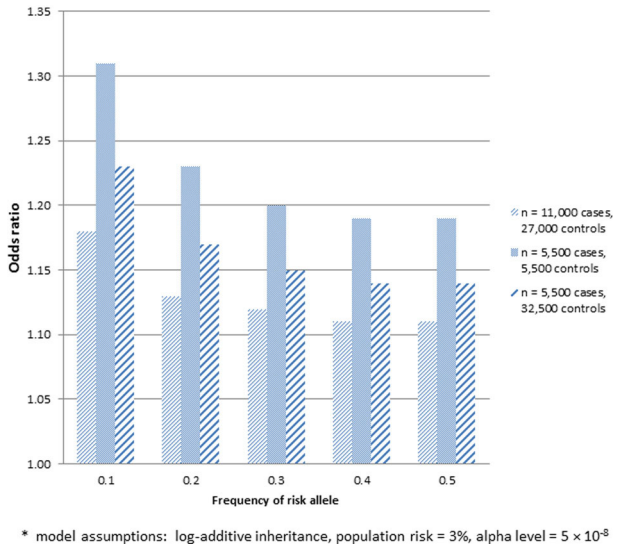


FIGURE 1 | Minimum odds ratio to detect SNP associations at 80% power for three different sample sets. (1) 11,000 cases and 27,000 previously genotyped controls; (2) 5500 cases and 5500 controls; and (3) 5500 cases and 27,000 previously genotyped controls*.

(HRS; phs000428.v1.p1), The Osteoarthritis Initiative (Lester, 2008) (OAI; http://www.oai.ucsf.edu/datarelease/About.asp), and HealthABC (Yaffe et al., 2009) (HABC; phs000169.v1.p1). These studies were selected because of their large sizes (12,500, 4000, and 2800, for HRS, OAI, and HABC, respectively), their geographic diversity within the US, and the dense genotyping available on each (Illumina HumanOmni 2.5M array in HRS and OAI and Illumina Human 1M array for HABC). All three studies included substantial representation of European Caucasians and African Americans, while HRS also included substantial numbers of Hispanics.

For each international (non-US) site, studies with previously genotyped controls were identified from the same ancestral background. For two sites (Leuven and Krakow), previously genotyped ancestry-matched control groups could not be identified, and so controls at these sites were genotyped alongside SiGN cases. For other sites (e.g., Barcelona), multiple control groups were identified to allow cases to be matched to suitable controls at a later stage, where we initially included as many genotyped controls as possible to improve power.

AVAILABILITY OF DISEASE RISK FACTORS AND OTHER COVARIATES

One drawback of using an already available control group is that that the clinical and covariate data may be limited or even absent altogether. This issue is of particular importance when effect decomposition is of interest, for example, whether a SNP acts through a modifiable risk factors such as smoking, or interacts with such a factor (Vanderweele and Hernan, 2012). Additionally, utilizing properly selected publically available controls can produce unbiased estimates of total genetic effects, even in the presence of gene by environment interactions, but these estimates may not be generalizable to populations with drastically different covariate distributions. If covariate information is missing in the controls, extending research findings to other populations may be limited. This limitation is mitigated in the absence of gene by environment interaction or the low prevalence of the genetic variant.

A second potential drawback of using already available controls is that misclassification bias can result if controls are not "disease free." In studying an aging-related disease such as stroke, one may want to choose already genotyped controls that are disease-free and older so that genetically susceptible individuals are under-represented in the control pool. To the extent that phenotypic characterization is limited and disease status unknown, the use of publicly available controls may be better suited for studies of rare/uncommon diseases for which the likelihood that controls are affected is small. The prevalence of stroke in the adult population is approximately 3–4% (Go et al., 2014). We note that misclassification bias only reduces power and does not influence type 1 error (false positives).

COMPARABILITY OF GENOTYPING PLATFORMS BETWEEN CASES AND CONTROLS

In case-control studies it is critical to obtain measurements from cases and controls in comparable fashion to ensure that any measurement differences between groups are not due to artifacts in measurement procedures. In the case of genetic association studies, spurious differences between cases and controls can occur by virtue of systematic differences in sample processing, genotype assays (choice of genotyping platform), and genotype calling procedures. Potential biases due to different genotyping procedures constitute perhaps the biggest challenge for genetic association studies that utilize previously genotyped controls. This potential source of bias can be minimized by choosing control groups that have been previously genotyped on the same, or a highly compatible, platform as the one used for cases.

As additional quality control, it may be useful to genotype a small number of previously genotyped individuals alongside the cases to evaluate genotype discordance across different platforms. SiGN cases are genotyped using the Illumina HumanOmni 5 M Exome genotyping array, but controls had previously been genotyped on different arrays, primarily the Illumina Omni 1 M and the Illumina Omni 2.5 M. To evaluate genotyping quality between these platforms, DNA from 30 previously genotyped subjects were identified from five of the control populations (HRS, OAI, INMA GRAZ, and LUND) and then re-genotyped at CIDR alongside the cases so that genotype calls from the same sample could be compared across the two arrays. Genotype concordance rates were calculated across each set of 30 samples and all SNPs having one or more discordant genotypes ($n = 17,401$ SNPs) were flagged as potentially problematic and excluded from subsequent imputation and case-control analysis. The effectiveness of this filter can be evaluated empirically by assessing type 1 error rates in association analysis among these SNPs.

POST-GENOTYPING QUALITY CONTROL PROCEDURES TO ENHANCE GENOTYPE COMPARABILITY BETWEEN CASES AND CONTROLS

Differential genotyping quality between newly genotyped cases and previously genotyped controls is a primary source of spurious results in GWAS. Thorough quality control analysis of the case

and control data sets is therefore critical. The first step is to identify poor quality samples in cases and controls and remove these from further analyses. In SiGN preference was given to control groups genotyped and cleaned by the same labs that processed case genotype data. Genotyping of cases was performed on the Illumina HumanOmni5Exome-4v1 array at CIDR, which also performed initial quality control (QC), including manual review and, as needed, manual re-clustering of SNPs selected as potentially problematic (such as many low minor allele frequency exome SNPs). Additional QC was performed at the University of Washington using methods described by Laurie et al. (2010) in general and by Meschia et al. (2013) specifically for SiGN. Among ∼11,000 subjects genotyped, 8 were excluded due to unresolved identity issues (e.g., sex mismatches and unexpected relatedness).

Using the KING-robust method (Manichaikul et al., 2010) to analyze cryptic relatedness revealed that 99% of the subjects are mutually unrelated. We defined a related pair of subjects as connected by a kinship coefficient achieving the lower limit of the 95% prediction interval for second-degree relative pairs (KC $>$ 0.088). Among 4.5 million SNPs assayed, 4.2% were either failed by CIDR or flagged as potentially low quality. Starting with 4.3M non-monomorphic and unique SNPs, 110K were failed by CIDR (for various reasons, including manual review of zCall-flagged SNPs), an additional 60K for missing call rate $\geq$2%, an additional 5K for 3 or more discordant calls among 343 duplicate pairs, an additional 1.8K for 2 or more Mendelian errors among 24 HapMap trios, and an additional 2.5K for HWE p-value $<$0.0001 (in controls only), resulting in 4.1M SNPs passing QC. The median call rate was 99.9% and the error rate estimated from 343 pairs of sample duplicates was 2×10^{-5}, indicating very high quality data.

The QC procedures were applied to all SNPs regardless of minor allele frequency, but standard quality metrics have less power to detect problems with rare than with common variants. A post-processing procedure has been proposed for modifying GenomeStudio calls to improve accuracy of genotypes for rare variants (Goldstein et al., 2012). CIDR used a modified version of zCall to flag SNPs with potential problems as those with specific differences in genotype calling between GenomeStudio and zCall. Specifically, the CIDR QC process for low MAF SNPs includes running zCall to identify SNPs where possible heterozygous clusters were missed by GenCall (parameters $T = 21$ and $I = 0.2$). SNPs with 4 or more possible new heterozygotes were manually reviewed and manually re-called (or failed) as needed.

ASSESSMENT OF POPULATION STRATIFICATION, IMPUTATION, AND DATA ANALYSIS STRATEGY

Identifying matching control groups from the same ancestral background as cases is a necessary step to ensure case-control comparability in GWAS studies. Analysis of population substructure was particularly challenging in SiGN because of the desirability in generating population substrata that were based not only on ancestry but also on array content to minimize the pairing of samples genotyped on very dense arrays (e.g., Illumina HumanOmni 5M Exome) with samples genotyped on relatively sparse (e.g., Illumina 610) arrays. The approach we took in SiGN to accommodate these two competing strategies is summarized in **Figure 2**. Our first step was to define four array groups (Illumina 610, Illumina 660, Illumina 1M, and Illumina 2.5/5M). Within each array group, we then defined three different continental groupings (Europe, Africa, and Admixed) using principal components analysis (PCA) (Price et al., 2010), projecting onto the HapMap 3 samples. Only "high-quality" SNPs were used for these analyses, defined as those with extremely low missingness

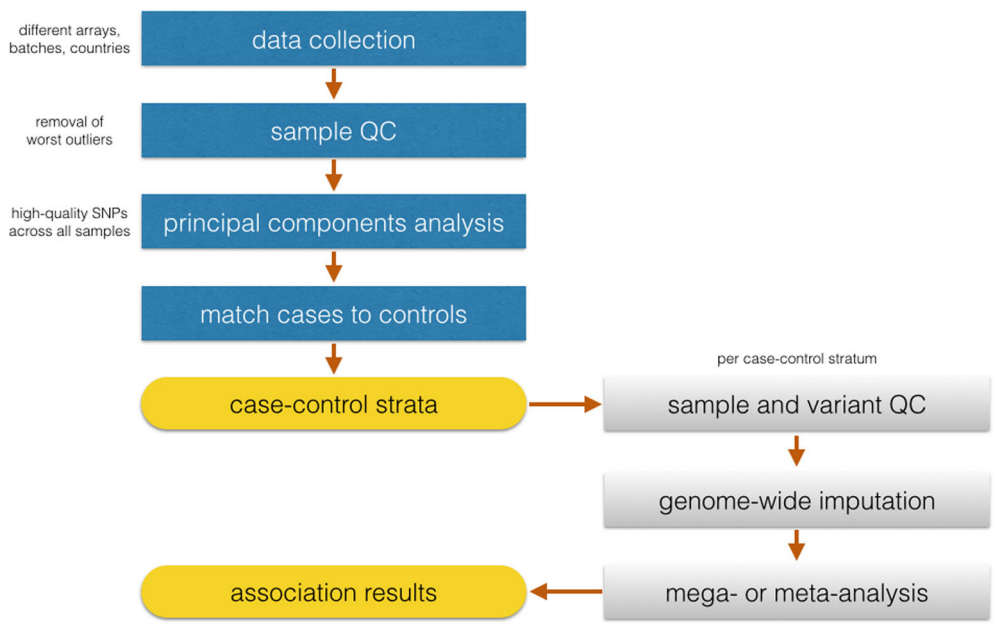


FIGURE 2 | Flowchart of the proposed analysis by SiGN.

Table 2 | Considerations when using already genotyped controls.

SELECTION OF CONTROLS:
• When possible, identify control sets that are from similar ethnic ancestry and were genotyped on the same platform as the cases.
• Consider using multiple control groups, especially when cases and controls are genotyped on different platforms and/or when the size of available control groups is small.
• Cross-study duplicates: if possible, re-genotype a small number of previously genotyped controls to allow evaluation of SNP concordance rates across the two platforms.
POPULATION SUBSTRUCTURE, IMPUTATION, AND ASSOCIATION ANALYSIS:
• Combine cases and previously genotyped controls together for assessment of population substructure, using a subset of non-imputed markers common to all samples (and after excluding SNPs found to be discordant from analysis of cross-study duplicates).
• Impute genotypes of cases and controls within population substrata.
• For confirmation, it is prudent to replicate observed associations after re-genotyping cases and control samples together.

(e.g., <0.1%) on all platforms, high frequency (e.g., >20%, as these are easier to genotype than low-frequency SNPs), outside of regions, such as the MHC or lactase (*LCT*) gene, that tend to be highly diverse even across populations of similar ancestry, and LD-pruned at an r^2 of 0.2.

Once continental groupings were defined within array groups, we then performed a second round of quality control analyses within ancestry by array group strata to remove problematic samples and SNPs, such as those samples or SNPs with high missingness rates or samples with inbreeding coefficients further than 3 SD from the mean of the sample distribution.

With the QCed set of samples, the next task was to combine continental groupings across array groupings to investigate population stratification across the full study sample. To do this, we started with a set of SNPs that were common across all samples and arrays ($n = 206,476$ SNPS) and selected high-quality sites only (as described above). After this SNP selection, the remaining 50–60K SNPs (depending on continental group) were used for PC analysis to check case-control clustering across all groups. Only 10 cases were missing matched controls and were removed from the analysis.

Iterative logistic regression and evaluation of statistical inflation (lambda) (Devlin and Roeder, 1999) will be necessary to recognize the extent of false-positives in the data and remove SNPs showing association to the trait due to systematic genotyping differences. Following identification of discrete case-control strata with well-behaved association statistics, imputation will be performed in continent-specific and array-specific groups. The SiGN analysis plan is for case-control analysis for stroke and its subtypes to be performed separately within each stratum using logistic regression, and then merged across strata using standard meta-analysis procedures.

SUMMARY

GWAS have been undeniably successful in identifying novel disease susceptibility loci (e.g., Billings and Florez, 2010; Teslovich et al., 2010; Chasman et al., 2012). Nonetheless, results from GWAS have also made clear that very large sample sizes are required to detect trait-associated SNPs that have small effect sizes. Large collections of cases suitable for genetic studies can often be obtained by pooling cases from a variety of sources, such as case reports, registries or large epidemiologic studies or, as demonstrated more recently, through the use of electronic health records (Ritchie et al., 2010). As we describe in this manuscript, there can be immense efficiency achieved in power by devoting genotyping resources to cases and using previously genotyped controls.

Availability of large collections of previously genotyped controls has been greatly facilitated by the decision of NIH that all genotypes for GWAS studies funded by federal dollars be made available for further research. In 2007 the tool dbGaP was introduced to facilitate community-wide access to these data (Mailman et al., 2007). It was this decision, the making available of publicly funded genotyping data, that affords researchers the opportunity to expand further scientific discoveries, as outlined here. Genetic researchers are thus favorably positioned to take advantage of this tremendous resource and are not as beholden to the initial study design as other etiologic research. This benefit does not come without a cost. We have outlined here, and summarized in **Table 2**, some considerations researchers may wish to consider as they design case-control studies using publicly available data.

ACKNOWLEDGMENTS

The SiGN study is funded by a cooperative agreement grant from the National Institute of Neurological Disorders and Stroke NINDS U01 NS069208. This project received partial support from The Mid-Atlantic Nutrition Obesity Research Center (P30 DK072488) from the NIH National Institute of Diabetes and Digestive and Kidney Diseases.

REFERENCES

Ay, H., Benner, T., Arsava, E. M., Furie, K. L., Singhal, A. B., Jensen, M. B., et al. (2007). A computerized algorithm for etiologic classification of ischemic stroke: the causative classification of stroke system. *Stroke* 38, 2979–2984. doi: 10.1161/STROKEAHA.107.490896

Billings, L. K., and Florez, J. C. (2010). The genetics of type 2 diabetes: what have we learned from GWAS? *Ann. N.Y. Acad. Sci.* 1212, 59–77. doi: 10.1111/j.1749-6632.2010.05838.x

Chasman, D. I., Fuchsberger, C., Pattaro, C., Teumer, A., Boger, C. A., Endlich, K., et al. (2012). Integration of genome-wide association studies with biological knowledge identifies six novel genes related to kidney function. *Hum. Mol. Genet.* 21, 5329–5343. doi: 10.1093/hmg/dds369

Devlin, B., and Roeder, K. (1999). Genomic control for association studies. *Biometrics* 55, 997–1004. doi: 10.1111/j.0006-341X.1999.00997.x

Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Blaha, M. J., et al. (2014). Heart disease and stroke statistics–2014 Update: a report from the american heart association. *Circulation* 129, e28–e292. doi: 10.1161/01.cir.0000441139.02102.80

Goldstein, J. I., Crenshaw, A., Carey, J., Grant, G. B., Maguire, J., Fromer, M., et al. (2012). zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics* 28, 2543–2545. doi: 10.1093/bioinformatics/bts479

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9362–9367. doi: 10.1073/pnas.0903103106

Juster, F. T., and Suzman, R. (1995). An overview of the health and retirement study. *J. Hum. Res.* 30, S7–S56. doi: 10.2307/146277

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., et al. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* 34, 591–602. doi: 10.1002/gepi.20516

Lester, G. (2008). Clinical research in OA–the NIH Osteoarthritis Initiative. *J. Musculoskelet. Neuronal Interact* 8, 313–314.

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., et al. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* 39, 1181–1186. doi: 10.1038/ng1007-1181

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. doi: 10.1093/bioinformatics/btq559

Meschia, J. F., Arnett, D. K., Ay, H., Brown, R. D. Jr., Benavente, O. R., Cole, J. W., et al. (2013). Stroke Genetics Network (SiGN) study: design and rationale for a genome-wide association study of ischemic stroke subtypes. *Stroke* 44, 2694–2702. doi: 10.1161/STROKEAHA.113.001857

Panagiotou, O. A., Willer, C. J., Hirschhorn, J. N., and Ioannidis, J. P. (2013). The power of meta-analysis in genome-wide association studies. *Annu. Rev. Genomics Hum. Genet.* 14, 441–465. doi: 10.1146/annurev-genom-091212-153520

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. doi: 10.1038/nrg2813

Ritchie, M. D., Denny, J. C., Crawford, D. C., Ramirez, A. H., Weiner, J. B., Pulley, J. M., et al. (2010). Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* 86, 560–572. doi: 10.1016/j.ajhg.2010.03.003

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466, 707–713. doi: 10.1038/nature09270

Vanderweele, T. J., and Hernan, M. A. (2012). Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. *Am. J. Epidemiol.* 175, 1303–1310. doi: 10.1093/aje/kwr458

Yaffe, K., Fiocco, A. J., Lindquist, K., Vittinghoff, E., Simonsick, E. M., Newman, A. B., et al. (2009). Predictors of maintaining cognitive function in older adults: the Health ABC study. *Neurology* 72, 2029–2035. doi: 10.1212/WNL.0b013e3181a92c36

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 January 2014; accepted: 04 April 2014; published online: 29 April 2014.
Citation: Mitchell BD, Fornage M, McArdle PF, Cheng Y-C, Pulit SL, Wong Q, Dave T, Williams SR, Corriveau R, Gwinn K, Doheny K, Laurie CC, Rich SS and de Bakker PIW (2014) Using previously genotyped controls in genome-wide association studies (GWAS): application to the Stroke Genetics Network (SiGN). Front. Genet. ***5****:95. doi: 10.3389/fgene.2014.00095*
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 25 June 2014
doi: 10.3389/fgene.2014.00166

The *ATXN2-SH2B3* locus is associated with peripheral arterial disease: an electronic medical record-based genome-wide association study

Iftikhar J. Kullo[1]*, Khader Shameer[1], Hayan Jouni[1], Timothy G. Lesnick[2], Jyotishman Pathak[2], Christopher G. Chute[2] and Mariza de Andrade[2]

[1] *Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN, USA*
[2] *Biomedical Statistics and Informatics, Health-Related Sciences, Mayo Clinic, Rochester, MN, USA*

Edited by:
Hakon Hakonarson, University of Pennsylvania, USA

Reviewed by:
Sarah Buxbaum, Jackson State University, USA
Linda E. Kelemen, Alberta Health Services-Cancer Care, Canada

****Correspondence:***
Iftikhar J. Kullo, Division of Cardiovascular Diseases, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA
e-mail: kullo.iftikhar@mayo.edu

Objectives: In contrast to coronary heart disease (CHD), genetic variants that influence susceptibility to peripheral arterial disease (PAD) remain largely unknown.

Background: We performed a two-stage genomic association study leveraging an electronic medical record (EMR) linked-biorepository to identify genetic variants that mediate susceptibility to PAD.

Methods: PAD was defined as a resting/post-exercise ankle-brachial index (ABI) $\leq$0.9 or $\geq$1.4 and/or history of lower extremity revascularization. Controls were patients without history of PAD. In Stage I we performed a genome-wide association analysis adjusting for age and sex, of 537, 872 SNPs in 1641 PAD cases (66 $\pm$ 11 years, 64% men) and 1604 control subjects (61 $\pm$ 7 year, 60% men) of European ancestry. In Stage II we genotyped the top 48 SNPs that were associated with PAD in Stage I, in a replication cohort of 740 PAD cases (70 $\pm$ 11 year, 63% men) and 1051 controls (70 $\pm$ 12 year, 61% men).

Results: The SNP rs653178 in the *ATXN2-SH2B3* locus was significantly associated with PAD in the discovery cohort ($OR = 1.23$; $P = 5.59 \times 10^{-5}$), in the replication cohort ($OR = 1.22$; 8.9×10^{-4}) and in the combined cohort ($OR = 1.22$; $P = 6.46 \times 10^{-7}$). In the combined cohort this SNP remained associated with PAD after additional adjustment for cardiovascular risk factors including smoking ($OR = 1.22$; $P = 2.15 \times 10^{-6}$) and after excluding patients with ABI > 1.4 ($OR = 1.24$; $P = 3.98 \times 10^{-7}$). The SNP is in near-complete linkage disequilibrium (LD) ($r^2 = 0.99$) with a missense SNP (rs3184504) in *SH2B3*, a gene encoding an adapter protein that plays a key role in immune and inflammatory response pathways and vascular homeostasis. The SNP has pleiotropic effects and has been previously associated with multiple phenotypes including myocardial infarction.

Conclusions: Our findings suggest that the *ATXN2-SH2B3* locus influences susceptibility to PAD.

Keywords: genome-wide association study, peripheral arterial disease, ankle-brachial index, electronic medical records, biorepository

INTRODUCTION

Peripheral arterial disease (PAD) affects nearly 10 million people in the US and more than 200 million people worldwide (Hirsch et al., 2001; Fowkes et al., 2013). PAD is associated with significant mortality and morbidity, underscoring the need to discover genetic variants that mediate susceptibility to this disease (Leeper et al., 2012). In contrast to coronary heart disease (CHD), genetic variants that influence susceptibility to PAD remain unknown. A genome-wide association study (GWAS) of smoking quantity revealed a variant in *CHRNA3* that was associated with PAD and lung cancer (Thorgeirsson et al., 2008).

Repositories of DNA from patients seen in the clinical setting and linked to the electronic medical record (EMR) systems can be leveraged to conduct genotyping or sequencing studies to identify genetic variants associated with human diseases and related quantitative traits. Extensive clinical data residing in the EMR can be leveraged for high-throughput phenotyping of medically relevant traits (Kullo et al., 2010). Such an approach may reduce the time, effort, and cost involved in conducting genomic studies to identify disease susceptibility loci.

The Electronic Medical Records and Genomics (eMERGE) consortium (McCarty et al., 2011) was created to develop and implement approaches for leveraging biorepositories linked to the EMR for large-scale genomic research, including but not limited to GWAS, sequencing, and structural variation (Kho et al., 2011). We undertook a GWAS of PAD cases and controls identified

from the EMR using a two-stage study design. In Stage I we performed a GWAS of 1641 PAD cases and 1604 controls, and in Stage II we attempted replication of the top significant SNPs in an independent sample of 740 PAD cases and 1051 controls.

MATERIALS AND METHODS

STUDY PARTICIPANTS

All participants gave written informed consent for participation in the study and the use of their data for future research. The Institutional Review Board of the Mayo Clinic approved the study protocol.

ASCERTAINMENT OF PAD CASES AND CONTROLS

The PAD patients were recruited from the non-invasive vascular laboratory at the Mayo Clinic Rochester, MN, based on the following criteria: (1) an ankle brachial index (ABI) of $\leq$0.9 at rest or 1 min after exercise, along with an abnormal continuous wave Doppler signal in one of the lower extremity arteries; (2) history of lower extremity revascularization if the ABI was normal; and (3) ABI $\geq$ 1.4 or ankle systolic BP $>$ 250 mm Hg, representing poorly compressible arteries. Exclusion criteria included PAD secondary to vasculitis, radiation to the abdomen or lower extremities, trauma to a lower extremity artery, thrombophilia, and arterial thrombosis. Controls were identified from patients referred to the Cardiovascular Health Clinic for exercise ECG to screen for cardiovascular disease. We excluded patients who had a positive exercise ECG, were younger than age 50, or had an abnormal ABI or history of PAD. A proportion (60%) of the subjects who underwent exercise ECG also underwent measurement of ABI. The prevalence of an abnormal ABI in patients who had a negative stress ECG was $<$1%.

Patient-level data elements in the Mayo EMR included demographics, outpatient visits and hospitalizations, providers, diagnosis and procedure codes, and results of non-invasive lower extremity arterial evaluation. Birth date, race, sex, and ethnicity were obtained from the demographic database; the categories for race were "White," "Black or African American," "Hispanic," "Asian/Pacific Islander," "American Indian/Alaskan Native," "Others," and "Unknown."

STAGE-I: HIGH-DENSITY GENOTYPING OF DISCOVERY COHORT

Genotyping was performed using the Illumina 660W-Quad BeadChip at the Center for Genotyping and Analysis at the Broad Institute, Cambridge, MA. This platform consists of 561,490 SNPs and 95,876 intensity-only probes. In addition to 3347 patient DNA samples, 58 blind duplicates, and 37 Coriell controls were genotyped. The Coriell controls include 1 trio (3 unique samples) that was duplicated on each plate. Genotyping calls were made using BeadStudio version 3.3.7 (2010).

Analysis tools used for quality control (QC) procedures included Illumina BeadStudio (2010), PLINK (Purcell et al., 2007), *R* (The R Development Core Team, 2007), STRUCTURE (Pritchard et al., 2000), and Eigenstrat in the Eigensoft package (Price et al., 2006). Data were cleaned using the QC pipeline developed by the eMERGE Genomics Working Group (Turner et al., 2011). This process includes evaluation of sample and marker call rate, gender mismatch and anomalies, duplicate and HapMap concordance, batch effects, Hardy-Weinberg equilibrium, sample relatedness, and population stratification. The data from all the patients, in addition to the HapMap II populations, were evaluated for population structure/substructure using Eigenstrat (Price et al., 2006). Of the 3347 unique samples, 3336 passed genotyping QC (see Supplementary Data and Figures S1–S3).

STAGE-II: GENOTYPING OF LEAD SNPs IN THE REPLICATION COHORT

The replication cohort consisted of 744 (470 males and 274 females) patients who had PAD based on the criteria listed above and 1053 (645 males and 408 females) controls with no prior history of PAD. The top 48 SNPs associated with PAD in the discovery cohort were genotyped using an Illumina custom genotyping panel with primers and probes from Assay-by-Design (Applied Biosystems, Foster City, CA). Custom capture and genotyping was performed at Mayo Clinic's Genotyping Core lab/Genotyping Shared Resource Lab.

Standard QC procedures were applied including evaluation of sample and marker call rate, HapMap concordance, and Hardy-Weinberg in controls only. We excluded six patients with low call rates ($<$95%). Of the 48 SNPs selected for replication, one (rs7900716) had a low call rate. All the 47 remaining SNPs had call rates $>$99% and Hardy-Weinberg P-value $>$ 0.05 in the controls.

STATISTICAL ANALYSES

Statistical analyses were conducted using SAS v. 9.3 {SAS Institute Inc., Cary, NC} and PLINK v1.07 (Purcell et al., 2007), and plots were created using *R* v2.11.0 (The R Development Core Team). Descriptive analyses were performed for the covariates and outcome variables using *t*-tests for continuous variables and chi-square tests for discrete variables. To adjust for population stratification, we used principal components to identify outliers in the study cohort (Price et al., 2006). Quantile-quantile (QQ) plots of observed $-\log_{10}$ P-values for PAD association versus the expected $-\log_{10}$ P-values under the null hypothesis of no association were generated to display the potential significant associations and to calculate the genomic inflation factor λ and to check for over dispersion of the test statistics. For each locus, we determined the set of HapMap SNPs in linkage disequilibrium (LD) ($r^2 > 0.5$) with the most significantly associated SNP. We then bounded the associated interval by the flanking HapMap recombination hotspots. These windows are likely to contain the causal variants explaining the associations. We used logistic regression analyses that adjusted for age and sex to identify the SNPs associated with PAD case/control status in the discovery, replication, and combined sets. All analyses were forced to test the same allele as the original sample. We performed sensitivity analyses by including additional adjustment variables for smoking, CHD, statin use, diastolic and systolic blood pressure, and diabetes. Since the additional adjustment variables did not have a qualitative impact on the final inferences, the results are not shown.

FUNCTIONAL ANNOTATION OF THE LEAD SNP

Data for the SNP rs3184504 (c.784T$>$C), which is in nearly complete LD with the most significant SNP, were obtained from the

Exome Variant Server. The impact of the variant was assessed using SIFT (Ng and Henikoff, 2003), PoplyPhen2 (Adzhubei et al., 2010), and conservation based measures such as PhastCons (Siepel et al., 2005), GRANTHAM (Grantham, 1974), and GERP (Cooper et al., 2005) scores. We performed Gene Ontology (GO) term enrichment analysis of *SH2B3* using first-degree interacting partners that were obtained from the protein-protein interaction database "STRING" (http://www.string-db.org). To understand the impact of SNP rs3184504 on protein structure, we performed a molecular dynamics simulation using GROMACS v4.5.7 (http://www.gromacs.org/) of the pleckstrin homolog (PH) domain of the SH2B3 protein where the SNP is localized.

RESULTS

DISCOVERY

After exclusions based on QC, including removal of related individuals and those of non-European Ancestry, a total of 3245 individuals—1641 PAD subjects and 1604 controls—were included in the analyses. No evidence of population stratification was found and therefore correction for population stratification was not needed in the analyses. Since the estimate of λ was 1.0, the test statistics showed no significant over dispersion. The study population demographic and clinical characteristics by case-control status are presented in **Table 1**. Among PAD cases, 64.3% were men, while among the controls, 60.3% were men. The mean age of the PAD patients was higher than the mean age of the control patients (65.7 years vs. 60.8 years) (**Table 1**). Assuming an additive genetic model and adjusting for age and sex, 60 SNPs were associated with PAD at $P < 1 \times 10^{-4}$. **Figure 1** presents a Manhattan plot of the *P*-values. Of these 60 SNPs, 48 were selected for replication based on Illumina designability score, LD, and minor allele frequency (MAF) in controls (see Supplementary Data for details).

REPLICATION

Characteristics of participants in the discovery and replication cohorts are presented in **Table 1**. The allele C of the intronic SNP rs653178 at the *ATXN2-SH2B3* locus on chromosome 12 was present more frequently in PAD cases (52%) than in controls (47%) with a resulting odds ratio (OR) of 1.23 (95% CI, 1.11–1.36, $P = 5.59 \times 10^{-5}$) in the discovery cohort (**Table 2**). In the replication cohort, the OR was 1.25 (95% CI, 1.10–1.40, $P = 8.94 \times 10^{-4}$) and in the combined sample, the OR was 1.22 (95% CI, 1.13–1.32, $P = 6.46 \times 10^{-7}$) (**Table 2**). The lead SNP rs653178 is in strong LD ($r^2 = 0.99$) with a missense SNP (rs3184504) in *SH2B3*, an adapter protein that plays a key role in immune and inflammatory response pathways and vascular homeostasis (Devalliere and Charreau, 2011; Devalliere et al., 2012). A locus specific visualization of lead variants associated with PAD is provided in the Supplementary Data (Figure S4).

Two additional SNPs rs11726269 (intronic region of *MAPK10*) and rs131408 (intergenic region between *LOC388882* and *IGLL1*) were significant at $P < 0.05$ in the replication cohort with similar direction of effect. However the *P* values exceeded the Bonferroni threshold for testing 48 SNPs (see Supplementary Table S1). The two most significant SNPs in the discovery cohort, rs7795096 in *PRKAG2* on chromosome 7 and rs2587888 in *GNAO1* on chromosome 16, did not replicate.

STRUCTURAL AND FUNCTIONAL IMPLICATIONS OF THE W262R VARIANT IN THE SH2B3 PROTEIN

Our analyses indicate that *SH2B3* encodes a multi-functional protein involved in diverse molecular pathways. Comparative protein sequence analyses using wild type and mutant sequences indicated that that rs3184504 leads to substitution of tryptophan with arginine (W262R) thereby introducing a new cAMP phosphorylation site in the PH domain of SH2B3 (see Supplementary Data and Figure S5). The PH domain in SH2B3 is important for lipid binding, membrane tethering and protein-protein interactions. GO terms (Ashburner et al., 2000) that are enriched among proteins interacting with *SH2B3* include blood coagulation; wound healing, and cell signaling events (see Supplementary Data, Table S2 and Figure S6). Conservation measures like Genomic

Table 1 | Participant characteristics.

	Discovery cohort		Replication cohort	
	Cases ($n = 1641$)	**Controls ($n = 1604$)**	**Cases ($n = 740$)**	**Controls ($n = 1051$)**
Men, *n* (%)	1055 (64.3)	968 (60.3)	468 (63.2)	643 (61.2)
Age, years	65.7 ± 10.68	60.8 ± 7.41‡	70.6 ± 11.60	70.2 ± 12.42
European ancestry, *n* (%)	1547 (94.3)	1512 (94.3)	721 (97.4)	1023 (97.3)
"Ever" smoker, *n* (%)	1322 (80.5)	963 (60.1)	632 (85.4)	641 (61.0)
ABI (pre-exercise)	0.72 ± 0.25	1.1 ± 0.07‡	0.79 ± 0.30	1.07 ± 0.16‡
ABI (post-exercise)	0.54 ± 0.25	1.1 ± 0.12‡	0.56 ± 0.28	1.03 ± 0.19‡
Hypertension, *n* (%)	1358 (82.8)	843 (52.6)‡	583 (78.8)	634 (60.3)‡
Type 2 diabetes, *n* (%)	507 (30.9)	141 (8.8)‡	225 (30.4)	126 (12.0)‡
Statin use, *n* (%)	774 (49.2)	398 (24.8)‡	532 (72.0)	326 (61.1)‡
CHD, *n* (%)	903 (55)	251 (15.6)‡	483 (65.3)	235 (22.4)‡

Continuous traits are depicted as mean ± standard deviation and categorical traits as count (percent); ABI, ankle-brachial index; CHD, coronary heart disease.
‡P < 0.001 for differences between PAD cases and controls.

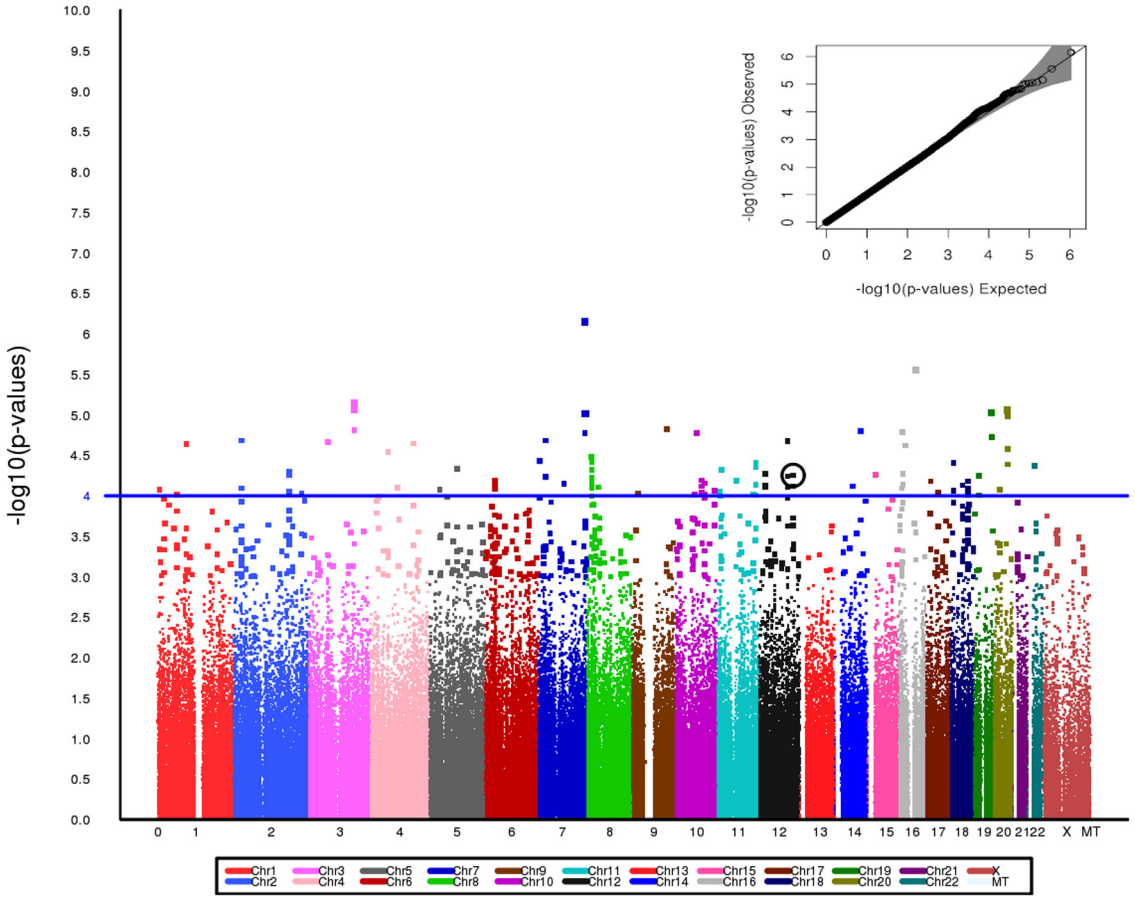


FIGURE 1 | Manhattan plot for the Stage I discovery cohort showing genome wide *P*-value distribution and the corresponding Q-Q plot (inset). SNP rs653178 is circled. A line at $p = 1 \times 10^{-4}$ is included on the Manhattan plot to provide a visual reference for *p*-values across the chromosomes.

Table 2 | Association of rs653178 with PAD in discovery, replication, and combined cohorts, after adjustment for age and sex in logistic regression models.

Cohort	PAD (*n*)	Controls (*n*)	Risk allele (frequency*)	OR (95% CI)	*P*-value
Discovery	1641	1604	C (0.469)	1.23 (1.11,1.36)	5.59×10^{-5}
Replication	740	1051	C (0.475)	1.25 (1.10,1.40)	8.90×10^{-4}
Combined	2381	2655	C (0.469)	1.22 (1.13,1.32)	6.46×10^{-7}

CI, confidence interval; OR, odds ratio; PAD, peripheral disease.
**Controls.*

Evolutionary Rate Profiling (GERP: 2.97) and phastCons (posterior probability: 0.159) suggest the variant is marginally conserved. Effect prediction analysis using Variant Effect Predicator (McLaren et al., 2010) indicate the variant as tolerant (SIFT: score = 1) benign (PolyPhen-2; score = 0.0), and moderately radical (GRANTHAM; score = 101). Molecular dynamic simulation suggested that the mutation in the PH domain of the SH2B3 results in structural perturbations and conformational changes (see Supplementary Data; Figures S7 and S8).

DISCUSSION

A better understanding of the genetic basis of PAD is required to improve risk stratification and identify new pathophysiologic pathways and drug targets. Conventional linkage and association approaches have failed to identify replicable susceptibility loci for PAD (Leeper et al., 2012) and the genome-wide association approach is currently the most promising design to uncover such loci. Heritable factors contribute to the risk of developing PAD. In the large population-based Swedish Twin Registry (Wahlgren and Magnusson, 2011), the odds ratio of having PAD in persons whose twin had PAD compared with persons whose twin did not have PAD was 17.7 (95% CI, 11.7–26.6) for monozygotic twins and 5.7 (95% CI, 4.1–7.9) for dizygotic twins. In a large case control study we found that family history of PAD was associated with doubling the odds of the presence of PAD (Khaleghi et al., 2014). Heritability estimates for ABI have varied from 0.21 (Kullo et al., 2006; Murabito et al., 2006) to 0.48 (Carmelli et al., 2000). In spite of evidence supporting the presence of heritable contribution to PAD, little is known about the genetic determinants of PAD.

In the present study, the SNP most strongly associated with PAD was an intronic SNP rs653178 in *ATXN2* on chromosome 12q24-12q24.1. This SNP is in near-complete LD with a missense

SNP in SH2B adaptor protein 3 gene (*SH2B3)* (rs3184504; $r^2 = 0.99$) that is likely the causal SNP. The SNP in *SH2B3* results in a substitution of tryptophan (large size and aromatic side chain) by arginine (large size and basic side chain) that induces changes in the structure and hydrophilic properties of the pleckstrin homology domain. This may result in altered lipid binding and protein–protein interactions as indicated by our molecular dynamics analyses. The variant also introduces a new phosphorylation site in the pleckstrin homology domain which may influence signaling pathways mediated by SH2B3. The SNP rs3184504 exhibits significant pleiotropic effects and has been implicated in immunological disorders, cardiovascular diseases (Gudbjartsson et al., 2009) and hematologic traits such as platelet count, mean-platelet volume (Gieger et al., 2011) and eosinophil count (Barrett et al., 2009). A summary of disease/trait associations of rs3184504 and rs653178 in the *ATXN2-SH2B3* locus is provided in the Supplement (Table S3).

The pleiotropic nature of *SH2B3* may be due to its role in immune and inflammatory signaling pathways including erythropoietin, cytokine receptor-mediated and integrin signaling (20). The protein also regulates hematopoietic cell lineage and endothelial cells, and influences adhesion and migration of platelets by modulating actin cytoskeleton organization (Takizawa et al., 2010; Gieger et al., 2011; Devalliere et al., 2012; Shameer et al., 2014). *SH2B3* is also involved in platelet production via megakaryocyte development; mice lacking *SH2B3* ($Lnk/SH2B3^{-/-}$) (Kwon et al., 2009) have altered platelet function and thrombus development (Tong et al., 2005). The relatively high frequency of this SNP in the general population is speculated to be due to a protective effect against bacterial infection (Zhernakova et al., 2010). We (Ding and Kullo, 2011) and others (Pickrell et al., 2009) have previously demonstrated that the SNP may have been subject to natural selection.

Two GWAS in European ancestry cohorts have reported variants associated with PAD. Thorgeirsson et al (Thorgeirsson et al., 2008) found a common variant in the nicotinic acetylcholine receptor gene cluster on chromosome 15q24 to affect nicotine dependence, smoking quantity, and the risk of PAD and lung cancer. A synonymous SNP (rs1051730) within the cholinergic receptor nicotinic alpha 3 gene (*CHRNA3*) was significantly associated with PAD ($OR = 1.19$). In a meta-analysis (Murabito et al., 2012) of GWAS for ABI consisting of 21 population-based cohort studies and 41,692 participants of European ancestry among whom 3409 participants had PAD (ABI < 0.90), six SNPs were associated ($P = 1 \times 10^{-6}$) with PAD, but none at a genome-wide significance level. The *ATXN2-SH2B3* locus was not associated with PAD in this study. One possible explanation may be the differences in case ascertainment, the present study including symptomatic PAD patients from the clinical setting whereas in the meta-analyses by Murabito et al, most individuals had undergone ABI measurement as part of prospective cohort studies. Koriyama et al. (2010) found the *OSBPL10* locus to be associated with PAD in a Japanese cohort. We assessed the strength of association of these SNPs in our dataset and found that the 9p21 variant and the *OSBPL10* variants were not associated, whereas the *CHRNA3* variant was weakly ($P = 1 \times 10^{-3}$) associated with PAD case status.

In conclusion, our findings suggest that SNP rs653178 in the *ATXN2-SH2B3* locus is associated with clinically defined PAD. The SNP is in near complete LD with rs3184504, a non-synonymous SNP in *SH2B3*, a gene implicated in immune, inflammatory, and hematopoietic pathways. This SNP is emerging as a key pleiotropic genetic variant influencing multiple cardiovascular traits. Our findings motivate additional investigation of this locus including sequencing, gene expression and drug targeting studies as well as studies in experimental animals.

SOURCES OF FUNDING

This work was supported by grants HG-04599 and HG-06379 from the National Human Genome Research Institute (NHGRI), Bethesda, MD. The eMERGE Network was initiated and funded by NHGRI, with additional funding from National Institute of General Medical Sciences (NIGMS), Bethesda, MD.

ACKNOWLEDGMENT

We acknowledge Genotyping Core Lab/Genotyping Shared Resource Lab at Mayo Clinic for technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2014.00166/abstract

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi: 10.1038/nmeth0410-248

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Gene Ontol. Consort. Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., et al. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* 41, 703–707. doi: 10.1038/ng.381

Carmelli, D., Fabsitz, R. R., Swan, G. E., Reed, T., Miller, B., and Wolf, P. A. (2000). Contribution of genetic and environmental influences to ankle-brachial blood pressure index in the NHLBI Twin Study. National Heart, Lung, and Blood Institute. *Am. J. Epidemiol.* 151, 452–458. doi: 10.1093/oxfordjournals.aje.a010230

Cooper, G. M., Stone, E. A., Asimenos, G., Green, E. D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* 15, 901–913. doi: 10.1101/gr.3577405

Devalliere, J., and Charreau, B. (2011). The adaptor Lnk (SH2B3): an emerging regulator in vascular cells and a link between immune and inflammatory signaling. *Biochem. Pharmacol.* 82, 1391–1402. doi: 10.1016/j.bcp.2011.06.023

Devalliere, J., Chatelais, M., Fitau, J., Gerard, N., Hulin, P., Velazquez, L., et al. (2012). LNK (SH2B3) is a key regulator of integrin signaling in endothelial cells and targets alpha-parvin to control cell adhesion and migration. *FASEB J.* 26, 2592–2606. doi: 10.1096/fj.11-193383

Ding, K., and Kullo, I. J. (2011). Geographic differences in allele frequencies of susceptibility SNPs for cardiovascular disease. *BMC Med. Genet.* 12:55. doi: 10.1186/1471-2350-12-55

Fowkes, F. G., Rudan, D., Rudan, I., Aboyans, V., Denenberg, J. O., McDermott, M. M., et al. (2013). Comparison of global estimates of prevalence and risk factors for peripheral artery disease in 2000 and 2010: a systematic review and analysis. *Lancet* 382, 1329–1340. doi: 10.1016/S0140-6736(13)61249-0

Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., et al. (2011). New gene functions in megakaryopoiesis and platelet formation. *Nature* 480, 201–208. doi: 10.1038/nature10659

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science* 185, 862–864. doi: 10.1126/science.185.4154.862

Gudbjartsson, D. F., Bjornsdottir, U. S., Halapi, E., Helgadottir, A., Sulem, P., Jonsdottir, G. M., et al. (2009). Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. *Nat. Genet.* 41, 342–347. doi: 10.1038/ng.323

Hirsch, A. T., Criqui, M. H., Treat-Jacobson, D., Regensteiner, J. G., Creager, M. A., Olin, J. W., et al. (2001). Peripheral arterial disease detection, awareness, and treatment in primary care. *JAMA* 286, 1317–1324. doi: 10.1001/jama.286.11.1317

Illumina. (2010). *Illumina BeadStudio Data Analysis Software Modules*. Available online at: http://www.illumina.com/Documents/products/datasheets/datasheet_beadstudio.pdf (Illumina, Inc.)

Khaleghi, M., Isseh, I. N., Jouni, H., Bailey, K. R., and Kullo, I. J. (2014). Family history as a risk factor for peripheral arterial disease. *Am. J. Cardiol.* (in press).

Kho, A. N., Pacheco, J. A., Peissig, P. L., Rasmussen, L., Newton, K. M., Weston, N., et al. (2011). Electronic medical records for genetic research: results of the eMERGE consortium. *Sci. Transl. Med.* 3, 79re71. doi: 10.1126/scitranslmed.3001807

Koriyama, H., Nakagami, H., Katsuya, T., Sugimoto, K., Yamashita, H., Takami, Y., et al. (2010). Identification of evidence suggestive of an association with peripheral arterial disease at the OSBPL10 locus by genome-wide investigation in the Japanese population. *J. Atheroscler. Thromb.* 17, 1054–1062. doi: 10.5551/jat.4291

Kullo, I. J., Fan, J., Pathak, J., Savova, G. K., Ali, Z., and Chute, C. G. (2010). Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. *J. Am. Med. Inform. Assoc.* 17, 568–574. doi: 10.1136/jamia.2010.004366

Kullo, I. J., Turner, S. T., Kardia, S. L., Mosley, T. H. Jr., Boerwinkle, E., and de Andrade, M. (2006). A genome-wide linkage scan for ankle-brachial index in African American and non-Hispanic white subjects participating in the GENOA study. *Atherosclerosis* 187, 433–438. doi: 10.1016/j.atherosclerosis.2005.10.003

Kwon, S. M., Suzuki, T., Kawamoto, A., Ii, M., Eguchi, M., Akimaru, H., et al. (2009). Pivotal role of lnk adaptor protein in endothelial progenitor cell biology for vascular regeneration. *Circ. Res.* 104, 969–977. doi: 10.1161/CIRCRESAHA.108.192856

Leeper, N. J., Kullo, I. J., and Cooke, J. P. (2012). Genetics of peripheral artery disease. *Circulation* 125, 3220–3228. doi: 10.1161/CIRCULATIONAHA.111.033878

McCarty, C. A., Chisholm, R. L., Chute, C. G., Kullo, I. J., Jarvik, G., Larson, E. B., et al. (2011). The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics* 4:13. doi: 10.1186/1755-8794-4-13

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. doi: 10.1093/bioinformatics/btq330

Murabito, J. M., Guo, C. Y., Fox, C. S., and D'Agostino, R. B. (2006). Heritability of the ankle-brachial index: the Framingham Offspring study. *Am. J. Epidemiol.* 164, 963–968. doi: 10.1093/aje/kwj295

Murabito, J. M., White, C. C., Kavousi, M., Sun, Y. V., Feitosa, M. F., Nambi, V., et al. (2012). Association between chromosome 9p21 variants and the ankle-brachial index identified by a meta-analysis of 21 genome-wide association studies. *Circ. Cardiovasc. Genet.* 5, 100–112. doi: 10.1161/CIRCGENETICS.111.961292

Ng, P. C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi: 10.1093/nar/gkg509

Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19, 826–837. doi: 10.1101/gr.087577.108

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 doi: 10.1038/ng1847

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Shameer, K., Denny, J. C., Ding, K., Jouni, H., Crosslin, D. R., de Andrade, M., et al. (2014). A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum. Genet.* 133, 95–109. doi: 10.1007/s00439-013-1355-7

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050. doi: 10.1101/gr.3715005

Takizawa, H., Nishimura, S., Takayama, N., Oda, A., Nishikii, H., Morita, Y., et al. (2010). Lnk regulates integrin alphaIIbbeta3 outside-in signaling in mouse platelets, leading to stabilization of thrombus development *in vivo*. *J. Clin. Invest.* 120, 179–190. doi: 10.1172/JCI39503

The R Development Core Team. (2007). *R: A Language and Environment for Statistical Computing*. Available online at: http://www.r-project.org/

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., et al. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 452, 638–642. doi: 10.1038/nature06846

Tong, W., Zhang, J., and Lodish, H. F. (2005). Lnk inhibits erythropoiesis and Epo-dependent JAK2 activation and downstream signaling pathways. *Blood* 105, 4604–4612. doi: 10.1182/blood-2004-10-4093

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., et al. (2011). Quality control procedures for genome wide association studies. *Curr. Protoc. Hum. Genet.* Chapter 1; Unit 1: 19. doi: 10.1002/0471142905.hg0119s68

Wahlgren, C. M., and Magnusson, P. K. (2011). Genetic influences on peripheral arterial disease in a twin population. *Arterioscler. Thromb. Vasc. Biol.* 31, 678–682. doi: 10.1161/ATVBAHA.110.210385

Zhernakova, A., Elbers, C. C., Ferwerda, B., Romanos, J., Trynka, G., Dubois, P. C., et al. (2010). Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. *Am. J. Hum. Genet.* 86, 970–977. doi: 10.1016/j.ajhg.2010.05.004

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 12 February 2014; accepted: 19 May 2014; published online: 25 June 2014.
Citation: Kullo IJ, Shameer K, Jouni H, Lesnick TG, Pathak J, Chute CG and de Andrade M (2014) The ATXN2-SH2B3 locus is associated with peripheral arterial disease: an electronic medical record-based genome-wide association study. Front. Genet. ***5****:166. doi: 10.3389/fgene.2014.00166*
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 05 August 2014
doi: 10.3389/fgene.2014.00222

Extension of GWAS results for lipid-related phenotypes to extreme obesity using electronic health record (EHR) data and the Metabochip

Ankita Parihar[1], G. Craig Wood[2], Xin Chu[2], Qunjan Jin[3], George Argyropoulos[2], Christopher D. Still[2], Alan R. Shuldiner[1,4], Braxton D. Mitchell[1,4] and Glenn S. Gerhard[3]*

[1] *Department of Medicine and Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD, USA*
[2] *Geisinger Clinic, Geisinger Obesity Institute, Danville, PA, USA*
[3] *Department of Pathology and Laboratory Medicine, Department of Biochemistry and Molecular Biology, Institute for Personalized Medicine, Pennsylvania State University College of Medicine, Hershey, PA, USA*
[4] *Geriatric Research and Education Clinical Center, Veterans Administration Medical Center, Baltimore, MD, USA*

Edited by:
Helena Kuivaniemi, Geisinger Health System, USA

Reviewed by:
Gregory T. Jones, University of Otago, New Zealand
Rob Igo, Case Western Reserve University, USA

****Correspondence:***
Glenn S. Gerhard, Department of Pathology and Laboratory Medicine, Department of Biochemistry and Molecular Biology, Institute for Personalized Medicine, Pennsylvania State University College of Medicine, Room C5750500 University Drive, MC H171, Hershey, PA 17033, USA
e-mail: ggerhard@hmc.psu.edu

A variety of health-related data are commonly deposited into electronic health records (EHRs), including laboratory, diagnostic, and medication information. The digital nature of EHR data facilitates efficient extraction of these data for research studies, including genome-wide association studies (GWAS). Previous GWAS have identified numerous SNPs associated with variation in total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and triglycerides (TG). These findings have led to the development of specialized genotyping platforms that can be used for fine-mapping and replication in other populations. We have combined the efficiency of EHR data and the economic advantages of the Illumina Metabochip, a custom designed SNP chip targeted to traits related to coronary artery disease, myocardial infarction, and type 2 diabetes, to conduct an array-wide analysis of lipid traits in a population with extreme obesity. Our analyses identified associations with 12 of 21 previously identified lipid-associated SNPs with effect sizes similar to prior results. Association analysis using several approaches to account for lipid-lowering medication use resulted in fewer and less strongly associated SNPs. The availability of phenotype data from the EHR and the economic efficiency of the specialized Metabochip can be exploited to conduct multi-faceted genetic association analyses.

Keywords: GWAS, lipids, obesity, EHR

INTRODUCTION

Genome-wide association studies (GWAS) have been highly successful at identifying SNPs associated with a wide variety of phenotypes, including lipid disorders, although such studies require very large sample sizes (Willer et al., 2013). This limits their utility because of economic considerations and the need to acquire phenotype data from across diverse sources. These limitations can be minimized using data obtained from electronic health records (EHRs), which can be an efficient means to obtain robust and extant phenotype data (Pathak et al., 2012) from potentially a large number of individuals, including for metabolic traits (Wood et al., 2012) and genetics studies (Wood et al., 2008). Furthermore, this approach provides the opportunity to assess the relevance of genetic associations in real-world patient populations with selected phenotypic characteristics such as extreme obesity. Because lipid screening is part of standard of care testing and body weights are often measured, these data are commonly present in EHRs. The electronic nature of EHR data facilitates efficient extraction for research studies (Prokosch and Ganslandt, 2009). However, the accuracy of EHR-based data depends upon how the data were obtained and entered and how it was extracted. Certain portions of the EHR are more standardized, such as laboratory measures. Other data, such as medications, may not be as straight-forward because of the complexity of coding for medication use.

Large meta-analyses of GWAS have identified numerous genetic loci associated with variation in lipid phenotypes, including 39 loci for total cholesterol (TC), 22 loci for low-density lipoprotein cholesterol (LDL-C), 31 loci for high-density lipoprotein cholesterol (HDL-C), and 16 loci for triglycerides (TG) (Van Dongen et al., 2013) as well as body mass index (BMI) (Sandholt et al., 2012). These loci are estimated to underlie about one-quarter to one-third of the genetic basis for these traits, a result that has motivated the search for additional loci through even larger GWAS studies (Willer et al., 2013). Few GWAS have been conducted in populations with extreme obesity (Sarzynski et al., 2011; Rinella et al., 2013), which may differ significantly from the large GWAS population-based samples in prevalence of co-morbidities such as dyslipidemia and use of corresponding lipid lowering medications.

Results from GWAS data have led to the development of specialized platforms designed to identify additional genetic loci as follow-up to initial analyses and to allow for finer genetic mapping of previously identified loci. An economical genotyping platform, the Illumina Metabochip, was custom designed for analysis of traits related to coronary artery disease and type 2 diabetes (Voight et al., 2012). As a proof-of-principle use of EHR data, we genotyped DNA from a cohort of individuals with extreme obesity, ascertained on the basis of undergoing bariatric surgery and on whom a rich database of EHR-derived phenotype data were available (Still et al., 2013), using the Metabochip to evaluate associations with lipid traits. The presence of SNPs on the Metabochip residing in loci known to be previously associated with blood lipid levels enabled extension of findings to the context of extreme obesity. The use of EHR data and the Metabochip platform thus provided an effective strategy to test the relevance of lipid trait GWAS findings in this patient population.

MATERIALS AND METHODS

STUDY PARTICIPANTS, EHR SOURCE DATA, AND COLLECTION OF BLOOD SAMPLES

Study participants were patients treated in the Geisinger Clinic Center for Nutrition and Weight Management who met clinical inclusion and exclusion criteria for bariatric surgery and were prospectively recruited into a research program on obesity from 2004 to 2012. The Geisinger Health System is an integrated health care delivery system that serves residents in central and northeastern Pennsylvania that includes the Geisinger Clinic, a network of 37 community-based primary care practices that provide care to over 400,000 patients. All sites have used the EpicCare™ EHR since 2001 (Allen-Ramey et al., 2013).

Data used for this study were obtained from an obesity database based on the EHR as previously described (Wood et al., 2012). Source data included patient demographics, clinical measures, problem list based on ICD-9 codes, medical history, medication history, and lab results. Blood drawn for lipid measurements and DNA isolation was obtained as part of a standard of care phlebotomy performed during the pre-surgery period, which consisted of a 6–12 month program during which a comprehensive medical history was obtained, a physical exam conducted, body weight, waist circumference, and height measured, and disease-specific, standard of care laboratory tests obtained, including fasting TC, HDL-C, LDL-C, and TG. Clinical data were recorded in the EHR. Blood for DNA isolation was transported to the research laboratory for processing and storage. Genomic DNA was isolated from patient whole blood samples as previously described (Chu et al., 2008), arrayed into microplates, and transported to the University of Maryland Translational Genomics Laboratory for Metabochip genotyping. The research was approved by the Geisinger Clinic and Penn State Hershey Institutional Review Boards and all participants provided written informed consent.

GENOTYPING AND GENOTYPE CLEANING

A total of 1851 samples was selected for genotyping using the Illumina Metabochip. The Metabochip array consists of ∼200,000 SNPs that include: (1) "replication" SNPs corresponding to validated associations; (2) a set of 63,450 SNPs that were the most significantly associated with over 20 traits related to coronary artery disease or T2D, including lipids, (3) SNPs previously associated with BMI and waist circumference, as well as 122,241 SNPs to fine-map these loci; and (4) 16,992 other SNPs selected for a variety of reasons, including those that reached genome-wide significance in any GWAS (Voight et al., 2012; Shah et al., 2013). Genotyping of the Metabochip was performed as per the manufacturer's protocol. A total of 196,725 were polymorphic. Samples that had call rates across all SNPs of <95% were removed, leaving a total of 1827 samples (Supplementary Table 1). Eight of the 1827 samples were excluded due to missing phenotype data. After excluding samples discordant for reported and genetically determined sex, unresolvable duplicates, and samples related to another sample (Supplementary Methods), the remaining number of subjects available for the analysis set was 1686 (Supplementary Table 1).

A series of analyses were also conducted to identify potentially problematic SNPs. Starting with the 196,725 polymorphic SNPs, we identified a total of 6279 problematic SNPs (Supplementary Methods) that were excluded, with the final cleaned dataset consisting of a total of 190,446 SNPs of which 63,134 SNPs had minor allele frequencies <0.01.

ASSOCIATION ANALYSIS

Statistical association testing between individual SNPs and lipid phenotypes was conducted under an additive model by regressing the genotype score (coded as number of copies of the reference allele) against the outcome lipid variable. Age and sex were included in the model as covariates. Further analyses addressed the issue of use of lipid-lowering agents (see Results). Our initial aim was to assess associations with 21 SNPs present on the Metabochip previously associated with lipid levels in prior meta-analysis of GWAS results (Kathiresan et al., 2008). For these analyses, we regarded a p-value of 0.0024 (0.05/21) to be statistically significant. We additionally assessed associations of 21 SNPs previously associated with body mass index and waist circumference (Willer et al., 2009; Speliotes et al., 2010; Sandholt et al., 2012), regarding a p-value of 0.002 (0.05/21) to be statistically significant.

We performed a secondary analysis to assess associations of all Metabochip SNPs with lipid and body weight traits in which we adjusted for the total number of SNPs tested, defining the significance cut-off as $p < 2.6 \times 10^{-7}$ after Bonferroni's correction ($p = 0.05/190{,}446$). We estimated that our final sample size of 1686 individuals provided 80% power to detect SNPs explaining 2–2.5% of the variation in lipid or BMI levels at this significance level.

RESULTS

COHORT CHARACTERISTICS

The demographic, anthropometric, and lipid profiles of the population (**Table 1**) were characteristic of a bariatric surgery cohort (Wood et al., 2012). Over 99% of the population was Caucasian/European ancestry. Just under 46% of study subjects reported taking one or more lipid-lowering medications, the

Table 1 | Demographic and laboratory data.

Trait	Female ($n = 1365$)		Male ($n = 321$)		Total ($n = 1686$)	
	Mean	Stdev	Mean	Stdev	Mean	Stdev
BMI (kg/m^2)	46.5	8.1	48.6	9.1	46.9	8.3
WaistCir (inches)	50.0	12.5	56.8	15.1	51.3	13.3
TG (mg/dl)	162.7	106.2	190.6	145.5	168.0	115.0
TC (mg/dl)	184.8	46.7	168.4	51.1	181.8	48.0
HDL-C (mg/dl)	46.9	13.7	38.4	12.6	45.3	13.9
LDL (mg/dl)	104.3	38.4	88.6	42.5	101.4	39.7
TC/HDL-C	3.8	2.1	4.1	2.7	3.9	2.2

majority taking statins (Supplementary Table 2). The percent of patients with a diagnosis of hypertension was 48.7%. The diagnosis of type 2 diabetes was 35.2% (Supplementary Table 2), which was reflected in the concomitant use of diabetes medications (Supplementary Table 2).

ASSOCIATION OF SNPS AT KNOWN LIPID LOCI WITH LIPID LEVELS

Results from association testing of previously identified lipid loci (Kathiresan et al., 2008) are shown in **Table 2**. Of the 21 lipid-associated SNPs tested, 12 were nominally associated with one or more lipid traits at a $p < 0.05$, including 3 that remained significant following adjustment for multiple comparisons ($p < 0.002$). The loci marked by these three SNPs included *GCKR* (associated with TG levels at 5.3×10^{-4}), *LPL* (associated with HDL-C levels at 1.4×10^{-5}), and *CETP* (associated with HDL-C levels at 4.1×10^{-11}). The directions of the observed effects for all of the 12 SNPs nominally or significantly associated with lipid levels were directionally consistent with those previously reported.

ADJUSTMENT FOR LIPID-LOWERING MEDICATIONS

A significant proportion (46%) of subjects in this cohort were being treated with lipid-lowering medications. Medication use was not associated with levels of LDL-C (beta $= -2.74$; $p = 0.12$) or TC (beta $= 2.70$; $p = 0.19$), but was significantly associated with levels of HDL-C (beta $= -2.78$; $p = 1 \times 10^{-6}$) and TG (beta $= 0.22$; $p = 1 \times 10^{-17}$). The observed values TC, LDL-C, and TG levels would likely have been higher, and HDL-C levels lower, had they not been taking lipid-lowering medications. We therefore considered three additional analytic approaches to accommodate the effect of the medications. Our first approach was to repeat the association analyses after removing all subjects on lipid-lowering medications (final $n = 945$). Our second approach was to include use of lipid-lowering medications as a covariate (medication user vs. non-user) in the regression model (final $n = 1686$). Our final approach was to restrict analysis to subjects taking lipid-lowering medications (final $n = 741$).Results of the association analyses of SNPs at known lipid loci using all three approaches to address the use of lipid-lowering medications are shown in **Table 2**. Results obtained from analysis restricted to subjects not taking lipid-lowering medications were generally consistent with those obtained from the initial analysis of the entire cohort. With only a few exceptions (e.g., rs6544713 near ABCG8), the effect sizes at most loci tended to be of the same magnitude, although the *p*-values tended to be less significant in the sample with medication-users removed, consistent with a smaller sample size. The same trend, i.e., comparable effect sizes but lower statistical significance, was also observed when analyses were restricted to subjects taking lipid-lowering medications. Inclusion of medication use as a covariate in the model had virtually no effect on the genotype-lipid phenotype association at any of the tested SNPs.

Array-wide association analysis of lipid levels was also carried out using the same three approaches to evaluate the impact of lipid-lowering medication use. Manhattan plots for these results are shown in Supplementary Figures 3–5. In these analyses, we detected the association of HDL-C with the *CETP* locus at array-wide significance thresholds in subjects not taking lipid-lowering medications and with medication use as a covariate, but not in the subgroup taking lipid-lowering medications (Supplementary Table 4). A similar result was obtained for association of LDL-C with the *APOE* locus. The association of TG with the *APOA1-APOA3-APOA4-APOA5* locus was detected only when using medication use as a covariate.

Association of SNPs at known BMI and waist circumference loci For BMI and waist circumference, no SNP achieved a *p*-value of less than 0.002 (Supplementary Table 5).

ARRAY-WIDE ASSOCIATION ANALYSIS

Following analysis of the candidate SNPs, association analysis was undertaken for all SNPs on the array using an additive genetic model for 7 phenotypes; BMI, waist circumference, TC, LDL-C, HDL-C, TG, and TC/HDL-C ratio. For BMI and waist circumference, no SNP achieved a *p*-value of less than 1×10^{-6}(Supplementary Figures 1 and 2).

Results of the array-wide association analyses for 5 lipid phenotypes are summarized in Manhattan plots shown in **Figures 1A–E**. SNPs at three loci achieved *p*-values at less than 1×10^{-7} in association with HDL-C (**Figure 1A**). A cluster of SNPs with *p*-values less than 1×10^{-12} was identified at the *HERPUD1-CETP* locus on chromosome 16 (Keebler et al., 2009). All associated SNPs were in high linkage disequilibrium with rs173539 (**Figure 2**), the peak SNP identified in this region previously associated with HDL-C (Kathiresan et al., 2008). A cluster of SNPs at the *LPL* locus on chromosome 8 also associated with HDL-C levels, as has been previously reported (Heid et al., 2008). As shown in **Figure 3**, the associated SNPs were in high linkage disequilibrium with rs12678919, the peak SNP previously identified in this region associated with HDL-C (Kathiresan et al., 2008). The third locus associated with HDL-C levels was tagged by only a single SNP with a *p*-value of 7.46×10^{-9} was located at the *NPAS3* locus.

For TC (**Figure 1B**), the peak association occurred with multiple SNPs on chromosome 1 at the *CELSR2-PSRC1-SORT1* locus (peak association: $p < 2.6 \times 10^{-7}$), which has previously been associated with TC in multiple studies (Lu et al., 2010; Ma et al., 2010). Associations at this locus were also apparent for LDL-C levels (**Figure 1C**), as has been reported in previous studies (Kathiresan et al., 2008; Nakayama et al., 2009), although the associations did not achieve the Bonferroni-corrected

Table 2 | Associations of SNPs at known lipid-associated loci with lipid traits.

Trait	SNP	CHR	CHR position (HG18)	GENE	Ref. allele	Loci previously associated with lipid levels through GWAS from Kathiresan et al. (2008)			All Subjects ($n = 1686$)			Exclude subjects taking lipid-lowering medications ($n = 945$)		Include medication use as a covariate ($n = 1686$)		Include ONLY subjects taking lipid-lowering medications ($n = 741$)	
						Allele freq	Beta†	*P*-value	Allele freq	Beta†	*P*-value	Beta†	*P*-value	Beta†	*P*-value	Beta†	*P*-value
LDL	rs11206510	1	55,268,627	PCSK9	G	0.19	−0.09	4.00 E-08	0.18	−3.32	0.030	−2.51	0.225	−3.38	0.028	−2.70	0.311
HDL	rs4846914	1	228,000,000	GALNT2	G	0.40	−0.05	4.00 E-08	0.40	−0.04	0.923	0.35	0.617	−0.05	0.905	−0.33	0.564
TG	rs7557067	2	21,061,717	APOB	G	0.22	−0.08	9.00 E-12	0.25	−0.04	0.025	−0.06	0.057	−0.05	0.022	0.00	0.809
LDL	rs515135	2	21,139,562	APOB	A	0.20	−0.16	5.00 E-09	0.19	−3.74	0.013	−5.26	0.013	−3.88	0.010	−1.02	0.702
TG	rs1260326	2	27,584,444	GCKR	A	0.45	0.12	2.00 E-31	0.43	0.06	5.30 E-04	0.07	0.010	0.06	7.60 E-04	0.02	0.101
LDL	rs6544713	2	43,927,385	ABCG8	A	0.32	0.15	2.00 E-20	0.31	3.00	0.019	0.33	0.853	3.12	0.015	3.23	0.145
TG	rs714052	7	72,502,805	MLXIPL	G	0.12	−0.16	3.00 E-15	0.11	0.00	0.888	−0.03	0.517	−0.01	0.791	0.02	0.266
TG	rs7819412	8	11,082,571	XKR6 AMAC1L2	A	0.48	−0.04	3.00 E-08	0.48	0.00	0.931	0.02	0.526	0.01	0.895	0.00	0.805
TG	rs12678919	8	19,888,502	LPL	G	0.10	0.23	2.00 E-41	0.10	−0.08	0.006	−0.10	0.021	−0.08	0.007	−0.02	0.166
HDL	rs12678919	8	19,888,502	LPL	G	0.10	0.23	2.00 E-34	0.10	2.78	1.37 E-05	2.74	0.012	2.75	1.59 E-05	0.36	0.714
TG	rs2954029	8	126,490,972	TRIB1	T	0.44	−0.11	3.00 E-19	0.47	−0.04	0.011	−0.03	0.226	−0.04	0.022	−0.03	0.164
HDL	rs1883025	9	107,000,000	ABCA1	A	0.26	−0.08	1.00 E-09	0.27	−0.45	0.298	0.16	0.837	−0.43	0.311	−0.15	0.826
TG	rs174547	11	61,327,359	FADS1 FADS2 FADS3	G	0.33	−0.09	2.00 E-14	0.32	0.03	0.165	0.02	0.541	0.03	0.110	0.01	0.088
HDL	rs174547	11	61,327,359	FADS1 FADS2 FADS3	G	0.33	−0.09	2.00 E-12	0.32	−0.73	0.069	−0.87	0.202	−0.77	0.053	−1.05	0.543
LDL	rs2650000	12	121388962	HNF1A	A	0.36	0.07	2.00 E-08	0.37	1.84	0.127	0.39	0.811	1.89	0.117	1.73	0.425
HDL	rs10468017	15	56,465,804	LIPC	A	0.30	0.1	0.008	0.28	1.10	0.097	0.75	0.293	1.09	0.010	0.67	0.302
HDL	rs173539	16	55,545,545	CETP	A	0.32	0.25	4.00 E-75	0.32	2.66	4.09 E-11	3.19	3.15 E-06	2.67	2.85 E-11	0.74	0.242

(Continued)

Table 2 | Continued

Trait	SNP	CHR	CHR position (HG18)	GENE	Ref. allele	Loci previously associated with lipid levels through GWAS from Kathiresan et al. (2008)			All Subjects ($n = 1686$)			Exclude subjects taking lipid-lowering medications ($n = 945$)		Include medication use as a covariate ($n = 1686$)		Include ONLY subjects taking lipid-lowering medications ($n = 741$)	
						Allele freq	Beta†	*P*-value	Allele freq	Beta†	*P*-value	Beta†	*P*-value	Beta†	*P*-value	Beta†	*P*-value
HDL	rs4939883	18	45,421,212	LIPG	A	0.17	−0.14	7.00 E-15	0.17	−1.42	0.005	−0.43	0.647	−1.47	0.004	−2.33	0.003
HDL	rs2967605	19	8,375,738	ANG	A	0.16	−0.12	1.00 E-08	0.18	0.34	0.493	0.58	0.479	0.19	0.700	0.87	0.251
LDL	rs6511720	19	11,063,306	LDLR	A	0.10	−0.26	2.00 E-26	0.12	−5.03	0.005	−5.20	0.026	−5.27	0.003	−7.74	0.017
LDL	rs10401969	19	19,268,718	NCAN CILP2	G	0.06	−0.05	2.00 E-08	0.07	−1.64	0.488	0.58	0.856	−1.87	0.431	−9.89	0.030
TG	rs17216525	19	19,523,220	NCAN CILP2 PBX4	A	0.07	−0.11	4.00 E-11	0.07	−0.07	0.031	−0.08	0.123	−0.06	0.086	−0.04	0.198
TG	rs7679	20	44,009,909	PLTP	G	0.19	−0.07	7.00 E-11	0.17	0.05	0.044	0.08	0.026	0.06	0.023	0.00	0.902
HDL	rs7679	20	44,009,909	PLTP	G	0.19	−0.07	4.00 E-09	0.17	−0.93	0.060	−1.25	0.157	−1.00	0.043	−0.06	0.934

**SNPs at known lipid-associated loci from Kathiresan et al. (2008).*

† Effect on lipid levels (expressed in SD units) associated with each copy of the reference allele.

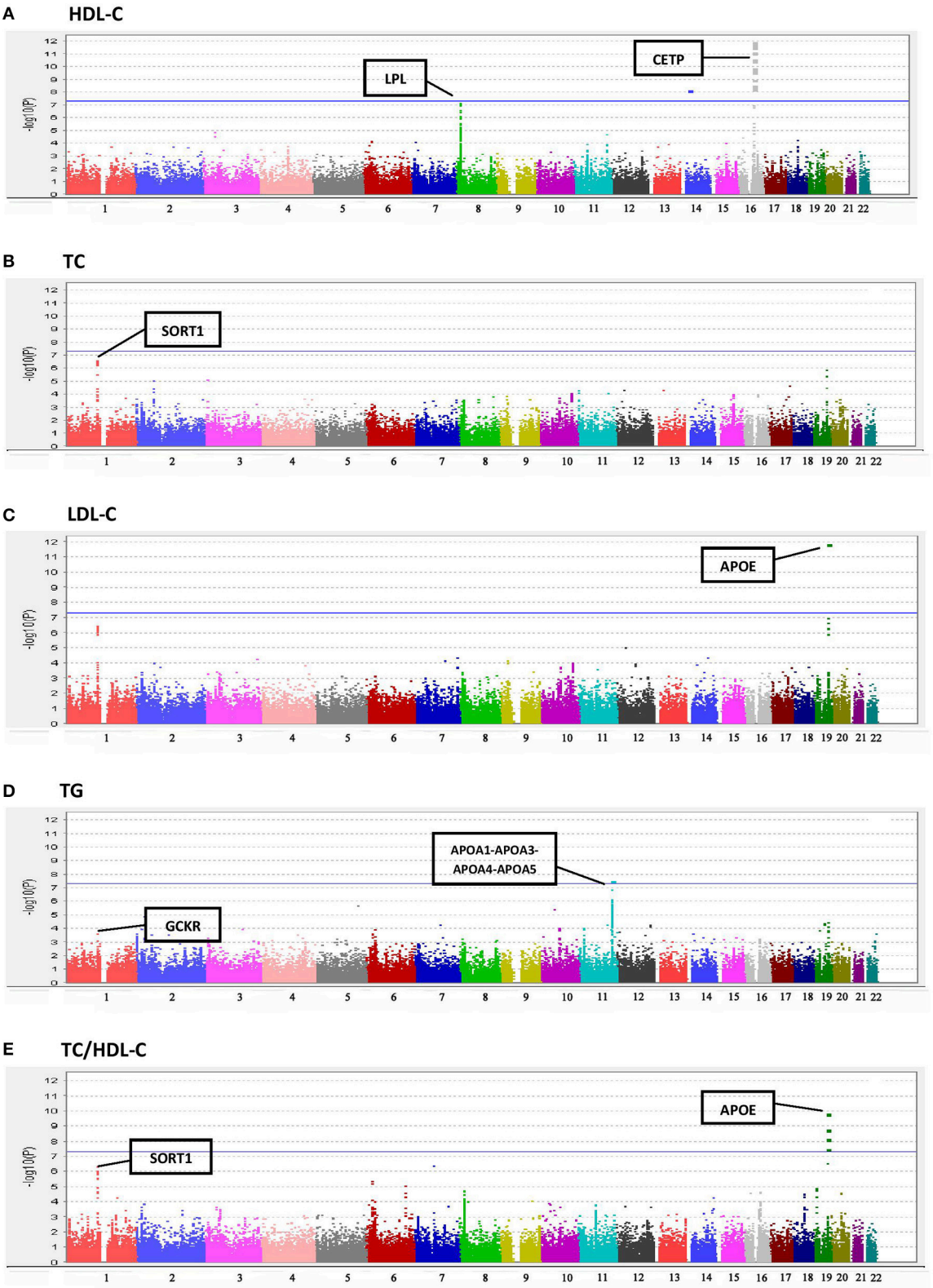


FIGURE 1 | Manhattan plots for associations of TC, HDL-C, LDL-C, TG, and TC/HDL-C with 190,446 SNPS from the Metabochip. (A) HDL-C. **(B)** TC. **(C)** LDL-C. **(D)** TG. **(E)** TC/HDL-C.

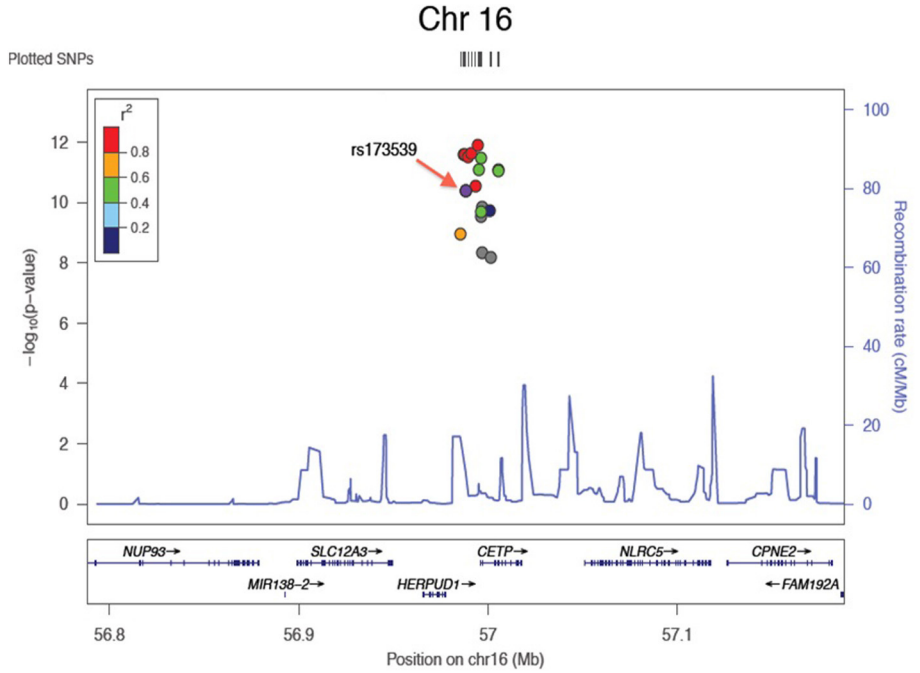


FIGURE 2 | Zoom plot showing association of SNPs in CETP region (Chromosome 16) associated with HDL-C, and linkage disequilibrium between associated SNPs and index SNP.

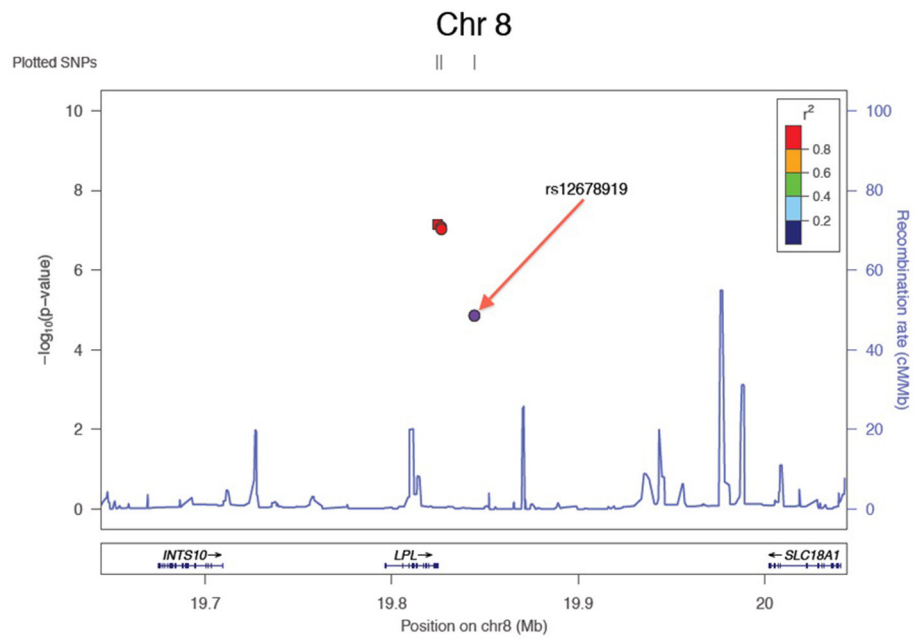


FIGURE 3 | Zoom plot showing association of SNPs in LPL region (Chromosome 8) associated with HDL-C, and linkage disequilibrium between associated SNPs and index SNP.

level of statistical significance. The well-documented association of LDL-C levels with *APOE* on chromosome 19 (Waterworth et al., 2010), was also detected, with a *p*-value of 1.56×10^{-12}.

A significant association was observed between TG and the *APOA1-APOA3-APOA4-APOA5* locus on chromosome 11 (SNP rs number not available; resides at position 116,156,325; see **Figure 1D**) as has been previously reported (Johansen et al.,

2011). Statistically significant associations with TC/HDL-C ratio were found at three loci (**Figure 1E**), two overlapping with TC and LDL-C, the *CELSR2-PSRC1-SORT1* and *APOE* regions.

DISCUSSION

Large-scale genomic studies can be costly due to the large sample sizes required for sufficient power to detect statistically significant differences of alleles with small to moderate effect sizes. Both phenotyping and genotyping can be expensive depending upon what is required for phenotyping and which genotype platform is selected. For this study, we used phenotype data derived from a clinical database constructed using EHR data (Wood et al., 2012). The laboratory data were obtained as part of clinical standard of care saving on the costs of thousands of blood lipid analyses; the anthropomorphic and demographic data were obtained by professionally certified clinicians and providers remunerated as part of clinical care. Importantly, the data were obtained in an electronic format allowing for the efficient construction of a flexible database. Despite the electronic format, careful data quality control and scrubbing were required in order to ensure robust results. In particular, the recording of body weights in a cohort of patients with extreme obesity must be carefully curated to ensure accuracy. Similarly, despite selection for high success rates, generating accurate SNP calls for the Metabochip platform required quite extensive manual curation and data cleaning.

We successfully extended to an extremely obese cohort associations between SNPs identified by large meta-GWAS studies and lipid traits. Failure to replicate associations with many of the SNPs was likely due to insufficient power from the smaller sample size in our cohort vis-à-vis the very large populations analyzed in previous studies. Alternatively, it is possible that some of the non-replicated SNPs have smaller effect sizes in extreme obese populations. However, the direction of effects and the effect sizes were similar to that previously reported by the large meta GWAS studies in all but two of the SNPs analyzed.

The SNPs we associated with TC and LDL-C largely overlapped. This is not surprising since the value for LDL-C is calculated using the Friedewald equation based on adjusting TC for HDL-C and TGs levels (subtract HDL-C and one-fifth of the TGs from TC). They are thus correlated values, so linear regression analyses will identify similar associations. Medication use was also not associated with either TC or LDL-C. Levels of HDL-C appear to be under strong genetic control, yet despite such high heritability (Ober et al., 2006), GWAS loci do not explain a large proportion of HDL-C variation (Willer and Mohlke, 2012). Nevertheless, we replicated several known loci. One locus, *LPL*, has also been robustly associated with risk for cardiovascular disease (Deloukas et al., 2013). Our results indicate that this locus may therefore also be a risk locus for CVD in patients with extreme obesity similar to previous studies of other CVD loci (Wood et al., 2008). Our results for HDL-C are similar to those reported for a bariatric surgery cohort of similar sample size in which a total of 60 SNPs in the ATP-binding cassette, sub-family A member 1 (*ABCA1*), apolipoprotein A1/C3/A4/A5 cluster (*APOA5*), cholesterol ester transfer protein (*CETP*), UDP-GalNAc transferase 2 (*GALNT2*), hepatic lipase (*LIPC*), endothelial lipase (*LIPG*), lipoprotein lipase (*LPL*), and the methylmalonic aciduria cblB type (*MMAB*)/mevalonate kinase (*MVK*) loci were genotyped (Sarzynski et al., 2011). Only SNPs in the *LPL*, *LIPC*, and *CETP* loci were statistically associated with pre-operative HDL-C level, similar to our results, although the multiple test correction factor was far less stringent than ours.

We found only a single locus associated with TG that replicated from the loci reported by the Global Lipids Genetics Consortium meta-analysis of over 100,000 individuals comprised of multi-ethnic and multi-racial populations (Teslovich et al., 2010). This is likely due to the much smaller sample size or alternatively, variation at the *APOA1/C3/A4/A5* gene cluster (Lai et al., 2005) may be the only genetic locus of the previously identified loci that associates with TG in extreme obesity. Which gene or genes in the *APOA1/C3/A4/A5* cluster harbors the TG influencing variant is not known.

A limitation of our study is that many of our subjects were on lipid lowering medications that may have masked our ability to identify genetic associations. A priori, lowering (or raising in the case of HDL-C) lipid levels through medication use may be expected to disproportionately occur in subjects with dyslipidemia due to a genetic predisposition, thus decreasing the ability to identify lipid-genotype associations. To address this issue, we performed three complementary analyses, including removing all subjects on lipid-lowering medications, the strategy employed by large meta-GWAS (Kathiresan et al., 2008). This predictably led to a major loss of statistical power, perhaps acceptable for very large sample sizes but not for our cohort. The approach of adjusting for medication use as a covariate had virtually no effect on the sensitivity of detecting SNP-lipid associations. However, this approach is biased by the indication for medication being high lipid levels, thus any identified associations between genotype and lipid levels are not independent of medication use.

Another potential limitation is that some patients may not have been fasting for a sufficient length of time prior to the blood draw to avoid an artifactual dietary effect on blood lipid measurements. For example, triglyceride levels are particularly sensitive to prandial state and other influences (Yuan et al., 2007). About 25 patients had triglyceride levels greater than 500 mg/dl, considered the highest category of hypertriglyceridemia by the Adult Treatment Panel III of the National Cholesterol Education Program (Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults, 2001). Hypertriglyceridemia may be expected to have a higher prevalence in populations with extreme obesity. In addition, assuming that the probability of non-fasting was independent of genotype, one would expect that measurement error due to non-fasting would obscure gene-lipid associations, not create false positives. The replication of associations at known loci supports the utility of using clinical samples.

No SNPs were found to be significantly associated with either BMI or waist circumference. The SNPs selected for the Metabochip included those from GWAS from the Genetic Investigation of Anthropometric Traits (GIANT) consortium, which studied anthropometric traits BMI and waist circumference. A total of 18,211 SNPs from 24 loci (Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults, 2001; Brahm and Hegele, 2013) found to be associated

with BMI, plus 5055 replication SNPs, were included on the Metabochip, along with 1374 SNPs and 1048 replication SNPs from 2 loci associated with waist circumference (Yuan et al., 2007). These studies involved multiple cohorts each with a mean BMI of about 27–28 kg/m^2 with a standard deviation of less than 5 kg/m^2. The proportion of these cohorts with a BMI of greater 40 kg/m^2 is thus likely less than 5–10%, with a relatively limited range of BMIs of less than 25 kg/m^2. The cohort with extreme obesity studied here had an average BMI of 48 ± 8 kg/m^2, representing a much higher average BMI, as well as a much wider range in BMI. The BMI of individuals at the upper range of the human body weight distribution may represent a distinct phenotype (Still et al., 2011) and harbor rarer variants with higher penetrance and larger effect sizes than the common variants interrogated by the Metabochip platform. Next generation sequencing may be required to identify those variants (Gerhard et al., 2013).

In summary, we conducted a GWAS of major lipid traits using EHR derived data to analyze SNPs that had previously been associated with lipid phenotypes, as well as other SNPs residing on the Metabochip, in an extremely obese cohort. Although several lipid loci replicated, other previously identified lipid and body weight loci did not. Possible differences may be due to the use of EHR data for phenotyping, characteristics of the cohort, and/or decreased statistical power. Nevertheless, the availability of extant EHR phenotype data and the relatively low cost of the specialized Metabochip can be effectively used to conduct a GWAS.

ACKNOWLEDGMENTS

Support for this project was provided by National Institutes of Health grants U01HG006382, P30 DK072488, and R01 DK088231.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2014.00222/abstract

REFERENCES

Allen-Ramey, F. C., Nelsen, L. M., Leader, J. B., Mercer, D., Kirchner, H. L., and Jones, J. B. (2013). Electronic health record-based assessment of oral corticosteroid use in a population of primary care patients with asthma: an observational study. *Allergy Asthma Clin. Immunol.* 9:27. doi: 10.1186/1710-1492-9-27

Brahm, A., and Hegele, R. A. (2013). Hypertriglyceridemia. *Nutrients* 5, 981–1001. doi: 10.3390/nu5030981

Chu, X., Erdman, R., Susek, M., Gerst, H., Derr, K., Al-Agha, M., et al. (2008). Association of morbid obesity with FTO and INSIG2 allelic variants. *Arch. Surg.* 143, 235–240; discussion 241. doi: 10.1001/archsurg.2007.77

Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T. L., Thompson, J. R., et al. (2013). Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.* 45, 25–33. doi: 10.1038/ng.2480

Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. (2001). Executive summary of the third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). *JAMA* 285, 2486–2497. doi: 10.1001/jama.285.19.2486

Gerhard, G. S., Chu, X., Wood, G. C., Gerhard, G. M., Benotti, P., Petrick, A. T., et al. (2013). Next-generation sequence analysis of genes associated with obesity and nonalcoholic fatty liver disease-related cirrhosis in extreme obesity. *Hum. Hered.* 75, 144–151. doi: 10.1159/000351719

Heid, I. M., Boes, E., Muller, M., Kollerits, B., Lamina, C., Coassin, S., et al. (2008). Genome-wide association analysis of high-density lipoprotein cholesterol in the population-based KORA study sheds new light on intergenic regions. *Circ. Cardiovasc. Genet.* 1, 10–20. doi: 10.1161/CIRCGENETICS.108.776708

Johansen, C. T., Kathiresan, S., and Hegele, R. A. (2011). Genetic determinants of plasma triglycerides. *J. Lipid Res.* 52, 189–206. doi: 10.1194/jlr.R009720

Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N. P., Rieder, M. J., et al. (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Genet.* 40, 189–197. doi: 10.1038/ng.75

Keebler, M. E., Sanders, C. L., Surti, A., Guiducci, C., Burtt, N. P., and Kathiresan, S. (2009). Association of blood lipids with common DNA sequence variants at 19 genetic loci in the multiethnic United States National Health and Nutrition Examination Survey III. *Circ. Cardiovasc. Genet.* 2, 238–243. doi: 10.1161/CIRCGENETICS.108.829473

Lai, C. Q., Parnell, L. D., and Ordovas, J. M. (2005). The APOA1/C3/A4/A5 gene cluster, lipid metabolism and cardiovascular disease risk. *Curr. Opin. Lipidol.* 16, 153–166. doi: 10.1097/01.mol.0000162320.54795.68

Lu, Y., Feskens, E. J., Boer, J. M., Imholz, S., Verschuren, W. M., Wijmenga, C., et al. (2010). Exploring genetic determinants of plasma total cholesterol levels and their predictive value in a longitudinal study. *Atherosclerosis* 213, 200–205. doi: 10.1016/j.atherosclerosis.2010.08.053

Ma, L., Yang, J., Runesha, H. B., Tanaka, T., Ferrucci, L., Bandinelli, S., et al. (2010). Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the Framingham heart study data. *BMC Med. Genet.* 11:55. doi: 10.1186/1471-2350-11-55

Nakayama, K., Bayasgalan, T., Yamanaka, K., Kumada, M., Gotoh, T., Utsumi, N., et al. (2009). Large scale replication analysis of loci associated with lipid concentrations in a Japanese population. *J. Med. Genet.* 46, 370–374. doi: 10.1136/jmg.2008.064063

Ober, C., Pan, L., Phillips, N., Parry, R., and Kurina, L. M. (2006). Sex-specific genetic architecture of asthma-associated quantitative trait loci in a founder population. *Curr. Allergy Asthma Rep.* 6, 241–246. doi: 10.1007/s11882-006-0041-4

Pathak, J., Kiefer, R. C., and Chute, C. G. (2012). Using semantic web technologies for cohort identification from electronic health records for clinical research. *AMIA Jt. Summits Transl. Sci. Proc.* 2012, 10–19.

Prokosch, H. U., and Ganslandt, T. (2009). Perspectives for medical informatics. Reusing the electronic medical record for clinical research. *Methods Inf. Med.* 48, 38–44. doi: 10.3414/ME9132

Rinella, E. S., Still, C., Shao, Y., Wood, G. C., Chu, X., Salerno, B., et al. (2013). Genome-wide association of single-nucleotide polymorphisms with weight loss outcomes after Roux-en-Y gastric bypass surgery. *J. Clin. Endocrinol. Metab.* 98, E1131–E1136. doi: 10.1210/jc.2012-3421

Sandholt, C. H., Hansen, T., and Pedersen, O. (2012). Beyond the fourth wave of genome-wide obesity association studies. *Nutr. Diabetes* 2, e37. doi: 10.1038/nutd.2012.9

Sarzynski, M. A., Jacobson, P., Rankinen, T., Carlsson, B., Sjostrom, L., Carlsson, L. M., et al. (2011). Association of GWAS-based candidate genes with HDL-cholesterol levels before and after bariatric surgery in the Swedish obese subjects study. *J. Clin. Endocrinol. Metab.* 96, E953–E957. doi: 10.1210/jc.2010-2227

Shah, T., Engmann, J., Dale, C., Shah, S., White, J., Giambartolomei, C., et al. (2013). Population genomics of cardiometabolic traits: design of the University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol (UCLEB) Consortium. *PLoS ONE* 8:e71345. doi: 10.1371/journal.pone.0071345

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* 42, 937–948. doi: 10.1038/ng.686

Still, C. D., Wood, G. C., Chu, X., Erdman, R., Manney, C. H., Benotti, P. N., et al. (2011). High allelic burden of four obesity SNPs is associated with poorer weight loss outcomes following gastric bypass surgery. *Obesity (Silver Spring)* 19, 1676–1683. doi: 10.1038/oby.2011.3

Still, C. D., Wood, G. C., Chu, X., Manney, C., Strodel, W., Petrick, A., et al. (2013). Clinical factors associated with weight loss outcomes after Roux-en-Y gastric bypass surgery. *Obesity (Silver Spring)* 22, 888–894. doi: 10.1002/oby.20529

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466, 707–713. doi: 10.1038/nature09270

Van Dongen, J., Willemsen, G., Chen, W. M., De Geus, E. J., and Boomsma, D. I. (2013). Heritability of metabolic syndrome traits in a large population-based sample. *J. Lipid Res.* 54, 2914–2923. doi: 10.1194/jlr.P041673

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., et al. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* 8:e1002793. doi: 10.1371/journal.pgen.1002793

Waterworth, D. M., Ricketts, S. L., Song, K., Chen, L., Zhao, J. H., Ripatti, S., et al. (2010). Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arterioscler. Thromb. Vasc. Biol.* 30, 2264–2276. doi: 10.1161/ATVBAHA.109.201020

Willer, C. J., and Mohlke, K. L. (2012). Finding genes and variants for lipid levels after genome-wide association analysis. *Curr. Opin. Lipidol.* 23, 98–103. doi: 10.1097/MOL.0b013e328350fad2

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* 45, 1274–1283. doi: 10.1038/ng.2797

Willer, C. J., Speliotes, E. K., Loos, R. J., Li, S., Lindgren, C. M., Heid, I. M., et al. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* 41, 25–34. doi: 10.1038/ng.287

Wood, G. C., Chu, X., Manney, C., Strodel, W., Petrick, A., Gabrielsen, J., et al. (2012). An electronic health record-enabled obesity database. *BMC Med. Inform. Decis. Mak.* 12:45. doi: 10.1186/1472-6947-12-45

Wood, G. C., Still, C. D., Chu, X., Susek, M., Erdman, R., Hartman, C., et al. (2008). Association of chromosome 9p21 SNPs with cardiovascular phenotypes in morbid obesity using electronic health record data. *Genomic Med.* 2, 33–43. doi: 10.1007/s11568-008-9023-z

Yuan, G., Al-Shali, K. Z., and Hegele, R. A. (2007). Hypertriglyceridemia: its etiology, effects and treatment. *CMAJ* 176, 1113–1120. doi: 10.1503/cmaj.060963

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 13 January 2014; accepted: 26 June 2014; published online: 05 August 2014.
*Citation: Parihar A, Wood GC, Chu X, Jin Q, Argyropoulos G, Still CD, Shuldiner AR, Mitchell BD and Gerhard GS (2014) Extension of GWAS results for lipid-related phenotypes to extreme obesity using electronic health record (EHR) data and the Metabochip. Front. Genet. **5**:222. doi: 10.3389/fgene.2014.00222*
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 09 May 2014
doi: 10.3389/fgene.2014.00125

Genome wide association study of SNP-, gene-, and pathway-based approaches to identify genes influencing susceptibility to *Staphylococcus aureus* infections

Zhan Ye[1], Daniel A. Vasco[2], Tonia C. Carter[2], Murray H. Brilliant[2], Steven J. Schrodi[2] and Sanjay K. Shukla[2]*

[1] *Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI, USA*
[2] *Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, WI, USA*

Edited by:
Helena Kuivaniemi, Geisinger Health System, USA

Reviewed by:
Lili Ding, Cincinnati Children's Hospital Medical Center, USA
David Fardo, University of Kentucky, USA

****Correspondence:***
Sanjay K. Shukla, Center for Human Genetics, Marshfield Clinic Research Foundation, 1000 N Oak Avenue – MLR, Marshfield, WI 54449, USA
e-mail: shukla.sanjay@mcrf.mfldclin.edu

Background: We conducted a genome-wide association study (GWAS) to identify specific genetic variants that underlie susceptibility to diseases caused by *Staphylococcus aureus* in humans.

Methods: Cases ($n = 309$) and controls ($n = 2925$) were genotyped at 508,921 single nucleotide polymorphisms (SNPs). Cases had at least one laboratory and clinician confirmed disease caused by *S. aureus* whereas controls did not. R-package (for SNP association), EIGENSOFT (to estimate and adjust for population stratification) and gene- (VEGAS) and pathway-based (DAVID, PANTHER, and Ingenuity Pathway Analysis) analyses were performed.

Results: No SNP reached genome-wide significance. Four SNPs exceeded the $p < 10^{-5}$ threshold including two (rs2455012 and rs7152530) reaching a p-value $< 10^{-7}$. The nearby genes were *PDE4B* (rs2455012), *TXNRD2* (rs3804047), *VRK1* and *BCL11B* (rs7152530), and *PNPLA5* (rs470093). The top two findings from the gene-based analysis were *NMRK2* ($p_{gene} = 1.20E\text{-}05$), which codes an integrin binding molecule (focal adhesion), and *DAPK3* ($p_{gene} = 5.10E\text{-}05$), a serine/threonine kinase (apoptosis and cytokinesis). The pathway analyses identified epithelial cell responses to mechanical and non-mechanical stress.

Conclusion: We identified potential susceptibility genes for *S. aureus* diseases in this preliminary study but confirmation by other studies is needed. The observed associations could be relevant given the complexity of *S. aureus* as a pathogen and its ability to exploit multiple biological pathways to cause infections in humans.

Keywords: *Staphylococcus aureus*, skin and soft tissue infection, GWAS, disease susceptibility, integrin and keratin disease pathway

INTRODUCTION

Staphylococcus aureus is a complex human pathogen due to its ability to survive both as a carriage organism, and behave as an opportunistic pathogen in a susceptible host. It is a leading cause of invasive bacterial infection that contributes to substantial morbidity and mortality worldwide. This bacterium can cause a variety of diseases ranging from mild to severe skin and soft tissue infections, keratitis, and osteomyelitis to life-threatening bacteremia, pneumonia, endocarditis, and sepsis (Lowy, 1998; Rehm, 2008). Even though *S. aureus* colonizes human anterior nares and other sites on skin in 30–50% of the general population (Graham et al., 2006; Kuehnert et al., 2006; Gordon and Lowy, 2008), not everyone who is colonized gets infected. One reason could be differences in genetic susceptibility to colonization and infections. The established role of host susceptibility in other infectious diseases lends support for a role of host genetics in *S. aureus* infection (de Bakker and Telenti, 2010). Genetic susceptibility to *S. aureus* infections is expected to be complex because this pathogen uses a wide variety of virulence factors that interact with several host pathways to cause disease in humans.

Numerous alleles segregating at a large number of loci contribute to complex disease susceptibility (Yang et al., 2011) with contributions from both common and rare alleles. Because genetic risk factors for *S. aureus* infections have not been previously studied on a genome-wide scale, we utilized the Personalized Medicine Research Project (PMRP) of Marshfield Clinic—a large, population-based biobank of DNA samples—to perform a preliminary genome-wide association study (GWAS) of laboratory-confirmed *S. aureus* infections to discover the underlying host-pathogen interactions.

MATERIALS AND METHODS

SUBJECTS

This study utilized Marshfield Clinic's PMRP biobank, a cohort of ∼20,000 individuals from 14 Zip Codes surrounding Marshfield, Wisconsin, USA. DNA samples from 3234 PMRP participants were genotyped at >500,000 SNPs as part of the NHGRI/NIGMS-funded eMERGE network (McCarty et al., 2011). These genotypic data are linked to longitudinal

electronic medical records (EMR) at the Marshfield Clinic and have served as a powerful resource in previous studies (McCarty et al., 2011; Cross et al., 2012; Hebbring et al., 2013). The population of PMRP is stable and highly homogeneous with a predominant Northern European genetic background and so carries lower risk of confounding by population stratification. All individuals in the study provided informed consent and the study was approved by the Marshfield Clinic IRB.

STUDY DESIGN

This study examined host susceptibility genes for *S. aureus* infection regardless of *S. aureus* colonization status. We performed a case/control GWAS on a subgroup of the PMRP population: 3234 subjects (309 cases/2925 controls). All study subjects were over 49 years of age. Cases were defined as individuals who had at least one laboratory confirmed test in their medical record of a disease caused by *S. aureus.* Controls were subjects who did not have any evidence of infection due to *S. aureus* in their EMR. We reasoned that because cases had to have evidence of laboratory confirmed *S. aureus* infection in the EMR, patients with no evidence of infection in the EMR would form an ideal control group. The results using our control group are likely very similar to those that would be obtained using population-based controls, as is often employed in GWAS studies (Burton et al., 2007). Only subjects with self-identified Northern European ancestry were included; therefore, both cases and controls had the same genetic ancestry. Additionally, a principal components analysis of their genome-wide genotypes in all study subjects (without knowledge of case/control status) was performed, revealing no evidence of population substructure (**Supplemental Figure 1**). With this filtered set of case/control subjects, we performed three levels of statistical analysis: SNP-based, Gene-based, and Pathway-based, to identify novel polymorphisms, genes, and pathways involved in susceptibility to *S. aureus* diseases. This hierarchical investigation of genetic effects allows for the incorporation of a gradation of biological information into the statistical tests—the SNP-based scan is the most comprehensive and agnostic to prior biological knowledge, the gene-based analysis uses positional information and collapses effects from the same protein-coding region, and the pathway-based analysis incorporates information obtained from various molecular biology studies.

PHENOTYPES

We extracted demographic and medical information on all case/control subjects from the Marshfield Clinic EMR. All cases had an active infection that had yielded *S. aureus* as the major or the only bacterium on a culture plate from a clinical sample such as blood, sputum, etc. Hospital surveillance subjects who were positive for *S. aureus* colonization by PCR were excluded from the study. Age, sex, body mass index (BMI), and Type 2 Diabetes (T2D) status as determined from the EMR were tested for association with case/control status. The binary variables were analyzed using a Fisher's exact test, and the continuous variables were analyzed through a two-tailed *T*-test to test mean differences between cases and controls.

GENOTYPIC DATA

The Illumina 660W-QUAD Beadchip array (Illumina, San Diego, California, USA) was used to generate genotype data on over 500 K SNPs from cases and controls. To ease automation of analysis, only data obtained from autosomes were analyzed. After filtering out SNPs with low minor allele frequency (<0.01), missing genotype data (≥ 0.05 of the study population with missing genotypes) or a significant departure from Hardy-Weinberg equilibrium (HWE) ($p < 1E\text{-}5$), there were 508,921 SNPs that passed the quality control screening and were used in subsequent analysis. For a sample to be included in the analysis, we ensured that it had at least 99% of the non-missing SNPs and for each SNP to be included in the analysis, we required that the SNP should have at least 95% of the non-missing subjects.

POPULATION STRATIFICATION ESTIMATION USING EIGENSOFT

We used the EIGENSOFT program (Price et al., 2006) to examine population stratification in our dataset. The program combines population genetics methods and the use of PCA to explicitly capture ancestry differences between cases and controls along continuous axes of variation.

SNP-BASED ANALYSIS USING PLINK AND R–PACKAGE

PLINK (Purcell et al., 2007), a whole genome association analysis tool set, was used to perform the filtering and QC procedures (as described under "Genotypic data" above) on the raw dataset to generate the data set employed in this study. R is a free software programming language and software environment for statistical computing and graphics (R. Core Team, 2013). The glm() function within R was used to establish the logistic model assuming an additive mode of inheritance with adjustment for the following risk factors: age, gender, BMI, diabetes and the top three principal components estimated from EIGENSOFT for population stratification. The *p*-values calculated from the logistic model were used to test the associations between SNPs and case/control phenotype.

GENE-BASED ANALYSIS USING VEGAS

VEGAS is a software program that tests association between a gene and phenotype trait. VEGAS uses SNP-level data to incorporate information from a full set of markers annotated to each gene and accounts for linkage disequilibrium (LD) between markers (Neale and Sham, 2004; Liu et al., 2010). All SNPs annotated to a gene were used in the calculation, and the method adjusted by the linkage disequilibrium structure by using HapMap data. Monte Carlo simulations from multivariate Gaussian random variables and Cholesky decomposition matrices were employed by VEGAS to produce disease association *p*-values per gene, correcting for the correlation structure between nearby SNPs. This type of analysis has several advantages including the collapsing of effects for all genotyped SNPs within each gene, and reducing the multiple testing burdens. The LD structure of each gene region is factored into the analysis through applying decomposition matrices in the analysis, effectively factoring-out the correlational structure between tightly linked SNPs. Additionally, gene-based results enable the subsequent use of many pathway analysis packages designed to use a single measurement from each gene.

PATHWAY-BASED ANALYSIS BY DAVID, PANTHER, AND INGENUITY PATHWAY ANALYSIS (IPA) PROGRAMS

All genes with a p-value = 0.01 from the VEGAS analysis were used as input to perform separate pathway analysis by DAVID, PANTHER, and IPA. The DAVID analyses consisted of three steps: measurement of the functional relationship of gene pairs, a DAVID agglomeration procedure to partition genes into functional gene groups, and visualization of the results (Huang et al., 2007, 2009a,b). PANTHER is a publically-available database having gene ontology, functional annotation, and evolutionary conservation information. PANTHER Pathways are available within the database and these pathways were used in our analysis. We calculated p-values for enrichment of *S. aureus*-associated genes within specific PANTHER pathways using a standard hypergeometric statistical approach. The statistical over-representation test was implemented in the PANTHER program. A binomial test was used to compare our gene list to a reference list (all human genes) to determine over- or under- representation of genes from our list in PANTHER gene function categories using an experiment-wise approach (experiment-wise $\alpha = 0.05$) (Mi and Thomas, 2009). Gene-based analyses using IPA (Ingenuity® Systems, www.ingenuity.com, Redwood City, California, USA, www.ingenuity.com) were used to generate protein-protein interaction networks.

RESULTS

POPULATION STRATIFICATION

Using EIGENSOFT to perform PCA of the 508,921 genotypes from all study subjects (without knowledge of case/control status), we found no evidence of strong population stratification (**Supplemental Figure 1**). Therefore none of the 3234 study subjects were excluded because of population stratification.

STUDY SUBJECT CHARACTERISTICS

It has been previously noted that male sex, age, high BMI and T2D are risk factors for invasive *S. aureus* infection (Graffunder and Venezia, 2002). As expected, the percent of males was significantly greater in the cases than the control group (51 vs. 39%; $p =$ 5.50E-05) (**Table 1**). There was no significant difference in mean age of case and controls. The prevalence of T2D was significantly higher among cases than controls (22.3 vs. 12.7%; $p =$ 1.03E-05) and so was the mean BMI (cases: 32.1 vs. control: 29.5 kg/m^2; $p =$ 3.50E-8). Our findings are consistent with those of previous reports.

Table 1 | Demographic and other phenotypic characteristics.

Characteristic	Cases (N = 309)	Controls (N = 2925)	*p*-value*
Age (years) (mean ± *SD*)	74.5 ± 10.9	73.5 (10.8)	0.13
Males (*N*, %)	159 (51.5)	1153 (39.4)	5.50E-05
Type 2 diabetes (N, %)	69 (22.3)	370 (12.7)	1.03E-05
Body mass index (kg/m^2) (mean ± *SD*)	32.1 ± 8.0	29.5 ± 6.1	3.50E-08

**Two-tailed t-test assuming unequal variances used to compare continuous variables (age and body mass index) and Fisher's exact test used to compare categorical variables (% males and % with type 2 diabetes).*

Q-Q PLOT

We performed Q-Q analysis on the p-values obtained using logistic model assuming additive model of inheritance (**Figure 1**). The plot showed no evidence of population stratification, confounding effects, or systematic bias in the results from the statistical routines employed.

HWE

Eight hundred forty three SNPs were excluded from the analysis due to departure from HWE exceeding $\alpha =$ 1E-05.

SNP-BASED ANALYSES

SNP associations were tested using logistic regression analysis after adjusting for risk factors such as age, gender, BMI, diabetes and three principal components. No single SNP in the GWAS reached the level of genome-wide significance ($p < 5 \times 10^{-8}$) (**Figure 2**). However, four SNPs exceeded the $p < 10^{-5}$ threshold, including two SNPs (rs2455012 and rs7152530) with a $p < 10^{-7}$ (**Table 2**). Out of these four SNPs, two were intronic (*PDE4B* and *TXNRD2*), one was intergenic with respect to *VRK1* and *BCL11B*, and one was in the 3'UTR of *PNPLA5*. Of the four SNPs on chromosome 14 (**Table 2**), rs1892234 was in weak linkage disequilibrium (LD) with the other three SNPs ($r^2 < 0.42$) which were in strong LD with each other ($r^2 = 0.80$). The two SNPs in *XRN1* were in strong LD ($r^2 = 0.92$), two of the SNPs on chromosome 22 (rs470093 and rs9614174) were in moderate LD ($r^2 = 0.49$), and the two SNPs on chromosome 19 exhibited low LD ($r^2 =$ 0.09).

VEGAS-BASED GENE ANALYSES

Table 3 shows the top 15 genes ranked by their p-values from the VEGAS analysis and four additional interesting genes that could potentially have a role in *S. aureus*-caused diseases based on their known involvement in immune and inflammatory processes. The topmost hit was *NMRK2* (or *ITGB1BP3*; $p =$ 1.20E-05) which encodes nicotinamide riboside kinase 2. Two of the 15 genes also featured in the list from the SNP-based analysis: *DAPK3* ($p =$ 5.10E-05) and *XRN1* ($p =$ 1.85E-04).

PATHWAY ANALYSES

The top-ranked 196 genes (p-value = 0.01) resulting from the VEGAS analysis were selected for the DAVID analyses but none of the gene groups were statistically significant. One of the top gene groups (gene group 1; Supplemental Table 1) included *CST8* (cystatin 8), *SERPINA6* (serine peptidase inhibitor, clade A, member 6), *SERPINA10* (serine peptidase inhibitor, clade A, member 10), and *SPINK1* (serine peptidase inhibitor, Kazak type 1). The enriched genes in group 2 (Supplemental Table 1) included four keratin genes: *KRT24* (form intermediate filament), *KRT82* (type II hair keratin), *KRT12* (type I intermediate filament keratin 12), and *KRT75* (form intermediate filament in in the cytoplasm of epithelial cells). As expected, the PANTHER analysis also suggested enrichment for intermediate filament cytoskeleton pathway (Supplemental Table 2). Using the same gene set input as in DAVID and PANTHER, the IPA was also explored with the intent

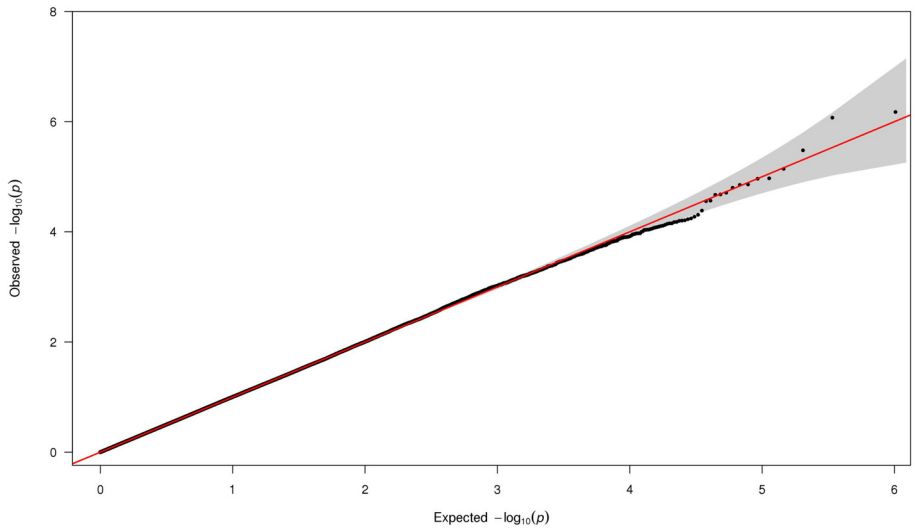


FIGURE 1 | The Q-Q plot of the *p*-values from all 508,921 SNPs. The x-axis shows the expected $-\log_{10}$(*p*-value). The y-axis shows the observed $-\log_{10}$(*p*-value).

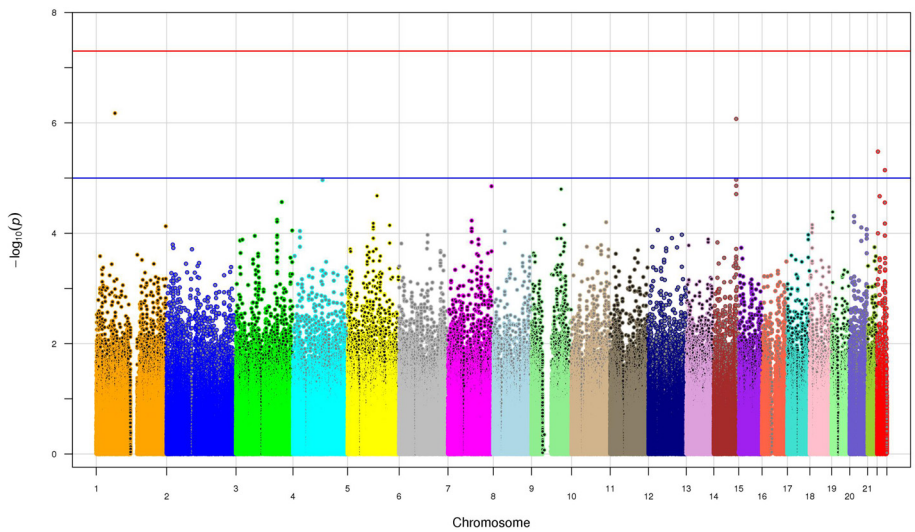


FIGURE 2 | The manhattan plot of the *p*-values from all 508,921 SNPs. The x-axis shows the chromosome numbers. The y-axis is the -$\log_{10}$ (*p*-value). The blue line is *p*-value of 10^{-5} whereas the red line shows the genome wide significance *p*-value of 5×10^{-8}.

of hypothesis-generating protein-protein interaction networks associated our gene data set. The IPA yielded 12 protein networks (data not shown) of which network 7, involving cell death, and survival, appeared interesting (**Supplemental Figure 2**).

DISCUSSION

Multiple diseases (endocarditis, skin infections, etc.) caused by *S. aureus* are mediated by two main classes of virulence factors, adhesion and secretory proteins, which interact with host receptors and inflammatory/anti-inflammatory molecules to produce the disease phenotype (Gordon and Lowy, 2008). Adhesion proteins help the pathogen to attach to the skin and survive on the epidermis and in the sub-epidermal layer through a repertoire of molecules collectively known as MSCRAMMS (microbial surface components recognizing adhesive matrix molecules). The MSCRAMMS can bind to fibronectin, fibrinogen, and platelets among others. Subsequent to attachment, *S. aureus* can secrete tissue and organ-specific virulence proteins (e.g., coagulase, proteases, toxins, superantigens) with a wide range of virulence functions that enable the pathogen to infect its host.

So far, most of the genetic susceptibility data related to *S. aureus* has been limited to *S. aureus* colonization. For example, the glucocorticoid receptor gene polymorphisms are associated with carriage risk (van den Akker et al., 2006), whereas *DEFB1* has

Table 2 | Top 20 SNPs, their chromosomal locations, associated genes, major and minor alleles, minor allele frequency of cases and controls, *p*-value and odds ratio and 95% confidence interval.

db SNP ID	Chr	Position (hg19)	Nearby genes	Variant location	Major/minor alleles	Case MAF*	control MAF	P-value	Odds ratio (95% CI)
rs2455012	1	66520998	*PDE4B*	Intron	C/T	0.090615	0.045190	6.68*E*-07	2.17 (1.59, 2.93)
rs7152530	14	98641215	*VRK1, BCL11B*	Intergenic	G/A	0.270227	0.366746	8.47*E*-07	0.62 (0.51, 0.75)
rs3804047	22	19879637	*TXNRD2*	Intron	T/C	0.368932	0.281143	3.32*E*-06	1.52 (1.27, 1.81)
rs470093	22	44276171	*PNPLA5*	3′ UTR	G/T	0.211974	0.146838	7.19*E*-06	1.63 (1.31, 2.02)
rs1381281	14	98615771	*VRK1, BCL11B*	Intergenic	A/G	0.322006	0.410024	1.07*E*-05	0.67 (0.56, 0.80)
rs13107325	4	103188709	*SLC39A8*	Exon 8	G/A	0.113269	0.065299	1.09*E*-05	1.88 (1.41, 2.47)
rs1892234	14	98720904	*VRK1, BCL11B*	Intergenic	T/C	0.364078	0.455897	1.38*E*-05	0.68 (0.57, 0.81)
rs6948646	7	152945284	*ACTR3B, DPP6*	Intergenic	G/A	0.444984	0.362393	1.41*E*-05	1.46 (1.23, 1.73)
rs1012203	9	104626708	*GRIN3A, CYLC2*	Intergenic	A/G	0.048544	0.104103	1.59*E*-05	0.44 (0.29, 0.62)
rs987514	14	98628943	*VRK1, BCL11B*	Intergenic	C/T	0.330097	0.415356	1.95*E*-05	0.68 (0.56, 0.81)
rs11950651	5	103009768	*NUDT12, EFNA5*	Intergenic	C/T	0.032362	0.079829	2.09*E*-05	0.37 (0.22, 0.57)
rs12160908	22	25928620	*LRP5L, ADRBK2*	Intergenic	C/T	0.067961	0.127401	2.13*E*-05	0.49 (0.35, 0.67)
rs12696090	3	158669688	*MFSD1, IQCJ-SCHIP1*	Intergenic	G/T	0.150485	0.226635	2.72*E*-05	0.61 (0.48, 0.76)
rs9614174	22	44082454	*EFCAB6*	Intron	A/C	0.215210	0.151795	2.78*E*-05	1.56 (1.26, 1.92)
rs7255123	19	3958397	*DAPK3*	3′ near gene	G/A	0.263754	0.198632	4.11*E*-05	1.51 (1.24, 1.84)
rs6135407	20	15397676	*MACROD2*	Intron	C/T	0.220065	0.158858	4.91*E*-05	1.54 (1.25, 1.90)
rs4807532	19	3928369	*ATCAY*	3′ near gene	G/A	0.461165	0.375897	5.32*E*-05	1.42 (1.20, 1.69)
rs7643377	3	142114694	*XRN1*	Intron	C/T	0.483819	0.400342	5.70*E*-05	1.41 (1.19, 1.67)
rs2535368	7	83157207	*SEMA3E*	Intron	G/A	0.202265	0.144494	5.90*E*-05	1.56 (1.25, 1.93)
rs9867210	3	142132749	*XRN1*	Intron	TIC	0.495146	0.411453	6.17*E*-05	1.41 (1.19, 1.67)

*Minor allele frequency.

been shown to promote persistent colonization by modulating beta-defensin expression in keratinocytes (Nurjadi et al., 2013). In contrast, the Danish middle-aged/elderly twin study showed that host genetics had a modest influence only on *S. aureus* carrier state (Andersen et al., 2012). While our manuscript was under review, a study by Nelson et al did not find any SNP that has genome wide significance for association with *S. aureus* bacteremia (SAB) although an intronic SNP in *CDON* was speculated to be associated with complicated SAB (Nelson et al., 2014). In a report by Stappers et al., four SNPs from three toll-like receptor (*TLR*) genes, *TLR1*, *TLR2*, and *TLR6*, increase the susceptibility to complicated skin and soft tissue infections caused by staphylococci, streptococci, and enterococci (Stappers et al., 2014).

In our multi-tiered GWAS-based investigation, we have identified a number of potentially interesting genes that need further investigation. Although the individual SNP results did not pass genome-wide significance, the top-tier SNPs, Gene-based, and Pathway-based results were enriched for genes that have plausible functions in bacterial infections. This includes genes that have roles in intracellular signaling, inflammation, zinc transport, and integrin binding. Thus, these genes have relatively high prior probabilities for involvement in *S. aureus* infection susceptibility—an aspect of the results that we believe supports considerable interest in subsequent studies to follow-up on these findings.

Notably, two genes (*DAPK3* and *XRN1*) were identified in both the SNP-based and gene-based analyses. *DAPK3* is a protein kinase that modulates apoptosis-related signaling pathways (Wu et al., 2010) and, in interaction with *RhoD*, modulates actin filament assembly and focal adhesion reorganization (Nehru et al., 2013). *S. aureus* is known to induce apoptosis of host cells during host invasion, leading to a compromised host immune response (Haslinger-Löffler et al., 2005). *XRN1* encodes a 5′–3′ exonuclease family member involved in cellular mRNA turnover (Nagarajan et al., 2013). The gene is shown to complete host mRNA degradation initiated by viral pathogens (Gaglia et al., 2012). These functions suggest possible roles for *DAPK3* and *XRN1* in susceptibility to diseases caused by *S. aureus*.

Some of the other genes we identified have been implicated in infectious disease processes. *PDE4B* is involved in modulating bacteria-induced inflammation (Komatsu et al., 2013). *PNPLA5* appears to be critical for autophagosome functions, including microbial clearance (Dupont et al., 2014). Mammalian hosts are known to reduce the level of free zinc to thwart pathogen growth (Kehl-Fie and Skaar, 2010) and it is plausible that *SLC39A8*, a zinc transporter, could be involved. *BCL11B* encodes a transcriptional repressor involved in T-cell development (Wakabayashi et al., 2003). *NMRK2*, an integrin beta1 binding protein, could function in host responses to bacterial function based on the finding that host fibronectin forms a bridge between *S. aureus* fibronectin-binding proteins and host cell beta1 integrins during *S. aureus* cellular invasion (Fowler et al., 2000). Keratin intermediate filaments are shown to have a protective role during infection with *Bartonella henselae* in cat scratch disease (Zhu et al., 2013). Interleukin 1 cytokine family members (*IL1A*, *IL1B*, *IL1R1*,

Table 3 | Top 15 and four other potentially interesting gene hits from VEGAS gene-based analysis.

Gene	Number of SNPs	Chr.	Start position (hg19)	End position (hg19)	*p*-value
NMRK2 (ITGB1BP3)	20	19	3933101	3942414	1.20E-05
DAPK3	18	19	3958452	3969827	5.10E-05
NPM3	4	10	103541082	103543170	0.000142
EEF2	17	19	3976054	3985461	0.000167
XRN1	12	3	142025449	142166853	0.000185
CDK7	*11*	5	68530622	68573257	0.000206
FBXL4	17	6	99321601	99395849	0.000257
ATR	20	3	142168077	142297668	0.000269
FGF8	3	10	103529887	103535759	0.000294
MRPS36	8	5	68513573	68525985	0.000423
CCDC125	*10*	5	68576519	68616407	0.000424
ATCAY	33	19	3880618	3928080	0.000571
KCNIP2	6	10	103585731	103603677	0.000637
MGEA5	6	10	103544200	103578222	0.000654
LDOC1L	30	22	44888450	44894178	0.000664
**IL1RL2* (75)	60	2	102803433	102855811	0.0031
**IL1B* (80)	20	2	113587337	113594356	0.00322
**ILIA* (93)	22	2	113531492	113542971	0.00363
**IL1R1* (168)	54	2	102770401	102796334	0.00806

*Ranks of these genes are mentioned in parenthesis.

and *IL1RL2*) are known mediators of immune and inflammatory responses (Garlanda et al., 2013).

It is known that many Mendelian and oligogenetic immunodeficiency disorders confer risk to staphylococcal infection including lymphocyte deficiencies such as severe combined immunodeficiency, chronic granulomatous disease, and hyper-IgE syndrome (Stephan et al., 1993; Grimbacher et al., 1999; Van de Vosse et al., 2009). These disorders are typified by highly disruptive mutations occurring in genes central to lymphoid cell competency including *STAT3*, *JAK3*, *DOCK8*, and *CD18*, among others (Hogg et al., 1999; Kalman et al., 2004; Jiao et al., 2008; Zhang et al., 2009). However, most cases of severe staphylococcal infection are not attributable to these more rare conditions and have unknown genetic etiology.

In summary, this preliminary GWAS applied a SNP-to-gene-to-disease-pathway approach to identify susceptibility genes against a broad umbrella of laboratory confirmed *S. aureus* infections. While no one SNP and gene was found to be highly significant in this study, we suspect that for a versatile pathogen like *S. aureus*, that needs to overcome barriers presented by a variety of tissues and defense systems to infect various sites in the body, there are bound to be several genes involved in host susceptibility. Not everyone exposed to a virulent or a colonizing strain of *S. aureus* has similar severity of infection. It is reasonable to speculate that effects of variants segregating at multiple genes contribute to the severity of *S. aureus* infection. Similarly, there could be protective alleles that may lower the risk of clinically-attended infection as well. Additional studies will be needed to confirm these findings but eventually functional studies will be needed to illuminate the detailed mechanisms of how these variants confer predisposition to infection. Once consensus disease loci and pathways are identified, they can serve as targets for future pharmaceutical development and further elucidation of how aberrant cellular processes/signaling give rise to *Staphylococcus*-induced pathologies.

ACKNOWLEDGMENTS

We would like to thank Crystal Jacobson, Amy Aswani, and Rob Strenn for very helpful informatics expertise and Dr. Matt Hall for valuable discussions and advice regarding the clinical aspects of *S. aureus* infection. We would also like to sincerely thank the PMRP participants. The project described was supported by the Clinical and Translational Science Award (CTSA) program, previously through the National Center for Research Resources (NCRR) grant 1UL1RR025011 and the National Center for Advancing Translational Sciences (NCATS) grant 9U54TR000021, and now by the NCATS grant UL1TR000427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The eMERGE Network is funded by the NHGRI, with additional funding from the National Institute of General Medical Sciences through the following grants: U01HG004438 to the Center for Inherited Disease Research; U01HG004608 to Essentia Institute for Rural Health/Marshfield Clinic Research Foundation. This study was also funded through support from the Marshfield Clinic, the Wisconsin Genomics Initiative, and philanthropic gifts in support of medical research at the Marshfield Clinic.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/2014.00125/abstract

Supplemental Figure 1 | A principal components analysis of genome-wide genotypes in all study subjects.

Supplemental Figure 2 | One of the IPA outputs of the protein-protein interaction network using VEGAS based 196 genes ($p \leq 0.01$) as input. Three genes, *DAPK3, KRT12,* and *TXNRD2* have been shown to be in direct network UBC (ubiquitin C). Boxes in pink color are input genes from the VEGAS analysis. Network shapes Symbols: ◇ = enzyme; ○ = Other; ▽ = Kinase; — = direct interaction; ---- = indirect interaction. Additional details about the symbols can be found at www.ingenuity.com.

REFERENCES

Andersen, P. S., Pedersen, J. K., Fode, P., Skov, R. L., Fowler, V. G. Jr., Stegger, M., et al. (2012). Influence of host genetics and environment on nasal carriage of *Staphylococcus aureus* in Danish middle-aged and elderly twins. *J. Infect. Dis.* 206, 1178–1184. doi: 10.1093/infdis/jis491

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678. doi: 10.1038/nature05911

Cross, D. S., McCarty, C. A., Hytopoulos, E., Beggs, M., Nolan, N., Harrington, D., et al. (2012). Coronary risk assessment among intermediate risk patients using

a clinical and biomarker based algorithm developed and validated in two population cohorts. *Curr. Med. Res. Opin.* 28, 1819–1830. doi: 10.1185/03007995

de Bakker, P. I., and Telenti. A. (2010). Infectious diseases not immune to genome-wide association. *Nat. Genet*. 42, 731–732. doi: 10.1038/ng0910-731

Dupont, N., Chauhan, S., Arko-Mensah, J., Castillo, E. F., Masedunskas, A., Weigert, R., et al. (2014). Neutral lipid stores and lipase PNPLA5 contribute to autophagosome biogenesis. *Curr. Biol*. 24, 609–620. doi: 10.1016/j.cub.2014.02.008

Fowler, T., Wann, E. R., Joh, D., Johansson, S., Foster, T. J., Höök, M., et al. (2000). Cellular invasion by *Staphylococcus aureus* involves a fibronectin bridge between the bacterial fibronectin-binding MSCRAMMs and host cell beta1 integrins. *Eur. J. Cell Biol.* 79, 672–679. doi: 10.1078/0171-9335-00104

Gaglia, M. M., Covarrubias, S., Wong, W., and Glaunsinger, B. A. (2012). A common strategy for host RNA degradation by divergent viruses. *J. Virol.* 86, 9527–9530. doi: 10.1128/JVI.01230-12

Garlanda, C., Dinarello, C. A., and Mantovani, A. (2013). The interleukin-1 family: back to the future. *Immunity* 39, 1003–1018. doi: 10.1016/j.immuni.2013.11.010

Gordon, R. J., and Lowy, F. D. (2008). Pathogenesis of methicillin-resistant *Staphylococcus aureus* infection. *Clin. Infect. Dis.* 46(Suppl. 5), S350–S359. doi: 10.1086/533591

Graffunder, E. M., and Venezia, R. A. (2002). Risk factors associated with nosocomial methicillin-resistant *Staphylococcus aureus* (MRSA) infection including previous use of antimicrobials. *J. Antimicrob. Chemother*. 49, 999–1005. doi: 10.1093/jac/dkf009

Graham, P. L. 3rd., Lin, S. X., and Larson, E. L. (2006). A. US population based survey of Staphylococcus aureus colonization. *Ann. Intern. Med*. 144, 318–325. doi: 10.7326/0003-4819-144-5-200603070-00006

Grimbacher, B., Holland, S. M., Gallin, J. I., Greenberg, F., Hill, S. C., Malech, H. L., et al. (1999). Hyper-IgE syndrome with recurrent infections—an autosomal dominant multisystem disorder. *N. Engl. J. Med*. 340, 692–702. doi: 10.1056/NEJM199903043400904

Haslinger-Löffler, B., Kahl, B. C., Grundmeier, M., Strangfeld, K., Wagner, B., Fischer, U., et al. (2005). Multiple virulence factors are required for *Staphylococcus aureus*-induced apoptosis in endothelial cells. *Cell. Microbiol*. 7, 1087–1097. doi: 10.1111/j.1462-5822.2005.00533.x

Hebbring, S., Slager, S., Epperla, N., Mazza, J. J., Ye, Z., Zhou, Z., et al. (2013). Genetic Evidence of *PTPN22* effects on chronic lymphocytic leukemia. *Blood* 121, 237–238. doi: 10.1182/blood-2012-08-450221

Hogg, N., Stewart, M. P., Scarth, S. L., Newton, R., Shaw, J. M., Law, S. K., et al. (1999). A novel leukocyte adhesion deficiency caused by expression but nonfunctional beta-2 integrins Mac-1 and LFA-1. *J. Clin. Invest.*103, 97–106. doi: 10.1172/JCI3312

Huang, D. A. W., Sherman, B. T., and Lempicki, R. A. (2009a). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Huang, D. A. W., Sherman B. T., and Lempicki, R. A. (2009b). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923

Huang, D. A. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., et al. (2007). The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol*. 8:R183. doi: 10.1186/gb-2007-8-9-r183

Jiao, H., Tóth, B., Erdos, M., Fransson, I., Rákóczi, E., Balogh, I., et al. (2008). Novel and recurrent STAT3 mutations in hyper-IgE syndrome patients from different ethnic groups. *Mol. Immunol*. 46, 202–206. doi: 10.1016/j.molimm.2008.07.001

Kalman, L., Lindegren, M. L., Kobrynski, L., Vogt, R., Hannon, H., Howard, J. T., et al. (2004). Mutations in genes required for T-cell development: IL7R, CD45, IL2RG, JAK3, RAG1, RAG2, ARTEMIS, and ADA and severe combined immunodeficiency: HuGE review. *Genet. Med*. 6, 16–26. doi: 10.1097/01.GIM.0000105752.80592.A3

Kehl-Fie, T. E., and Skaar, E. P. (2010). Nutritional immunity beyond iron: a role for manganese and zinc. *Curr. Opin. Chem. Biol.* 14, 218–224. doi: 10.1016/j.cbpa.2009.11.008

Komatsu, K., Lee, J. Y., Miyata, M., Hyang Lim, J., Jono, H., Koga, T., et al. (2013). Inhibition of PDE4B suppresses inflammation by increasing expression of the deubiquitinase CYLD. *Nat. Commun*. 4, 1684. doi: 10.1038/ncomms2674

Kuehnert, M. J., Kruszon-Moran, D., Hill, H. A., McQuillan, G., McAllister, S. K., Fosheim, G., et al. (2006). Prevalence of *Staphylococcus aureus* nasal colonization in the United States, 2001-2002. *J. Infect. Disease.* 193, 169–171. doi: 10.1086/499632

Liu, J. Z., McRae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., et al. (2010). A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet*. 87, 139–145. doi: 10.1016/j.ajhg.2010.06.009

Lowy, F. D. (1998). *Staphylococcus aureus* infections. *N. Engl. J. Med.* 339, 520–532. doi: 10.1056/NEJM199808203390806

McCarty, C. A., Chisholm, R. L., Chute, C. G., Kulo, I. J., Jarvi, G. P., and Larson, E. B. (2011). The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics* 4:13. doi: 10.1186/1755-8794-4-13

Mi, H., and Thomas, P. (2009). PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol. Biol*. 563, 123–140. doi: 10.1007/978-1-60761-175-2_7

Nagarajan, V. K., Jones, C. I., Newbury, S. F., and Green, P. J. (2013). *XRN* 5'?3' exoribonucleases: structure, mechanisms and functions. *Biochim. Biophys. Acta* 1829, 590–603. doi: 10.1016/j.bbegrm.2013.03.005

Neale, B. M., and Sham, P. C. (2004). The future of association studies: gene-based analysis and replication. *Am. J. Hum. Genet*. 75, 353–362. doi: 10.1086/423901

Nehru, V., Almeida, F. N., and Aspenström, P. (2013). Interaction of RhoD and ZIP kinase modulates actin filament assembly and focal adhesion dynamics. *Biochem. Biophys. Res. Commun*. 433, 163–169. doi: 10.1016/j.bbrc.2013.02.046

Nelson, C. L., Pelak, K., Podgoreanu, M. V., Ahn, S. H., Scott, W. K., Allen, A. S., et al. (2014). A genome-wide association study of variants associated with acquisition of *Staphylococcus aureus* bacteremia in a healthcare setting. *BMC Infect. Dis*. 14:83. doi: 10.1186/1471-2334-14-83

Nurjadi, D., Herrmann, E., Hinderberger, I., and Zanger, P. (2013). Impaired β-defensin expression in human skin links DEFB1 promoter polymorphisms with persistent *Staphylococcus aureus* nasal carriage. *J. Infect. Dis.* 207, 666–674. doi: 10.1093/infdis/jis735

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., Reich, D., et al. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet*. 38, 904–909. doi: 10.1038/ng1847

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

R. Core Team., (2013). "R: a language and environment for statistical computing," in *R Foundation for Statistical Computing*, Vienna, Austria. Available online at: http://www.R-project.org/

Rehm, S. J. (2008). *Staphylococcus aureus*: the new adventures of a legendary pathogen. *Cleve. Clin. J. Med.* 75, 177–180, 183–186, 190–192. doi: 10.3949/ccjm.75.3.177

Stappers, M. H., Thys, Y., Oosting, M., Plantinga, T. S., Ioana, M., Reimnitz, P., et al. (2014). TLR1, TLR2, and TLR6 gene polymorphisms are associated with increased susceptibility to complicated skin and skin structure infections. *J. Infect. Dis.* doi: 10.1093/infdis/jiu080. [Epub ahead of print].

Stephan, J. L., Vlekova, V., Le Deist, F., Blanche, S., Donadieu, J., De Saint-Basile, G., et al. (1993). Severe combined immunodeficiency: a retrospective single-center study of clinical presentation and outcome in 117 patients. *J. Pediatr*. 123, 564–572. doi: 10.1016/S0022-3476(05)80951-5

Van de Vosse, E., van Wengen, A., van Geelen, J. A., de Boer, M., Roos, D., and van Dissel, J. T. (2009). A novel mutation in NCF1 in an adult CGD patient with a liver abscess as first presentation. *J. Hum. Genet*. 54, 313–316. doi: 10.1038/jhg.2009.24

van den Akker, E. L., Nouwen, J. L., Melles, D. C., van Rossum, E. F., Koper, J. W., Uitterlinden, A. G., et al. (2006). *Staphylococcus aureus* nasal carriage is associated with glucocorticoid receptor gene polymorphisms. *J. Infect. Dis*. 194, 814–818. doi: 10.1086/506367

Wakabayashi, Y., Watanabe, H., Inoue, J., Takeda, N., Sakata, J., Mishima, Y., et al. (2003). Bcl11b is required for differentiation and survival of alphabeta T lymphocytes. *Nat. Immunol*. 4, 533–539. doi: 10.1038/ni927

Wu, Y., Yan, Q., Zuo, J., Saiyin, H., Jiang, W., Qiao, S., et al. (2010). Link of Dlk/ZIP kinase to cell apoptosis and tumor suppression. *Biochem. Biophys. Res. Commun.* 392, 510–515. doi: 10.1016/j.bbrc.2010.01.054

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M.,et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet*. 43, 519–525. doi: 10.1038/ng.823

Zhang, Q., Davis, J. C., Lamborn, I. T., Freeman, A. F., Jing, H., Favreau, A. J., et al. (2009). Combined immunodeficiency associated with DOCK8 mutations. *N. Engl. J. Med*. 361, 2046–2055. doi: 10.1056/NEJMoa0905506

Zhu, C., Bai, Y., Liu, Q., Li, D., Hong, J., Yang, Z., et al. (2013). Depolymerization of cytokeratin intermediate filaments facilitates intracellular infection of HeLa cells by Bartonella henselae. *J. Infect. Dis*. 207, 1397–1405. doi: 10.1093/infdis/jit040

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 January 2014; accepted: 21 April 2014; published online: 09 May 2014.
*Citation: Ye Z, Vasco DA, Carter TC, Brilliant MH, Schrodi SJ and Shukla SK (2014) Genome wide association study of SNP-, gene-, and pathway-based approaches to identify genes influencing susceptibility to Staphylococcus aureus infections. Front. Genet. **5**:125. doi: 10.3389/fgene.2014.00125*
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 18 November 2014
doi: 10.3389/fgene.2014.00401

Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links *PLCL1* to speech language development and *IL5-IL13* to Eosinophilic Esophagitis

Bahram Namjou[1,2]*, Keith Marsolo[2,3], Robert J. Caroll[4], Joshua C. Denny[4,5], Marylyn D. Ritchie[6], Shefali S. Verma[6], Todd Lingren[2,3], Aleksey Porollo[1,2,3], Beth L. Cobb[1], Cassandra Perry[7], Leah C. Kottyan[1,2,8], Marc E. Rothenberg[8], Susan D. Thompson[1,2], Ingrid A. Holm[9], Isaac S. Kohane[10] and John B. Harley[1,2,11]

[1] *Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA*
[2] *College of Medicine, University of Cincinnati, Cincinnati, OH, USA*
[3] *Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA*
[4] *Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, USA*
[5] *Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN, USA*
[6] *Center for Systems Genomics, The Pennsylvania State University, Philadelphia, PA, USA*
[7] *Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA*
[8] *Division of Allergy and Immunology, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA*
[9] *Division of Genetics and Genomics, Department of Pediatrics, The Manton Center for Orphan Disease Research, Harvard Medical School, Boston Children's Hospital, Boston, MA, USA*
[10] *Children's Hospital Informatics Program, Center for Biomedical Informatics, Harvard Medical School, Boston, MA, USA*
[11] *U.S. Department of Veterans Affairs Medical Center, Cincinnati, OH, USA*

Edited by:
Mariza De Andrade, Mayo Clinic, USA

Reviewed by:
Andrew Skol, University of Chicago, USA
Albert Vernon Smith, Icelandic Heart Association, Iceland
Shelley Cole, Texas Biomedical Research Institute, USA

****Correspondence:***
Bahram Namjou, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229, USA
e-mail: bahram.namjou@cchmc.org

Objective: We report the first pediatric specific Phenome-Wide Association Study (PheWAS) using electronic medical records (EMRs). Given the early success of PheWAS in adult populations, we investigated the feasibility of this approach in pediatric cohorts in which associations between a previously known genetic variant and a wide range of clinical or physiological traits were evaluated. Although computationally intensive, this approach has potential to reveal disease mechanistic relationships between a variant and a network of phenotypes.

Method: Data on 5049 samples of European ancestry were obtained from the EMRs of two large academic centers in five different genotyped cohorts. Recently, these samples have undergone whole genome imputation. After standard quality controls, removing missing data and outliers based on principal components analyses (PCA), 4268 samples were used for the PheWAS study. We scanned for associations between 2476 single-nucleotide polymorphisms (SNP) with available genotyping data from previously published GWAS studies and 539 EMR-derived phenotypes. The false discovery rate was calculated and, for any new PheWAS findings, a permutation approach (with up to 1,000,000 trials) was implemented.

Results: This PheWAS found a variety of common variants (MAF $>$ 10%) with prior GWAS associations in our pediatric cohorts including Juvenile Rheumatoid Arthritis (JRA), Asthma, Autism and Pervasive Developmental Disorder (PDD) and Type 1 Diabetes with a false discovery rate $<$ 0.05 and power of study above 80%. In addition, several new PheWAS findings were identified including a cluster of association near the *NDFIP1* gene for mental retardation (best SNP rs10057309, $p = 4.33 \times 10^{-7}$, $OR = 1.70$, 95%CI $= 1.38 - 2.09$); association near *PLCL1* gene for developmental delays and speech disorder [best SNP rs1595825, $p = 1.13 \times 10^{-8}$, $OR = 0.65(0.57 - 0.76)$]; a cluster of associations in the *IL5-IL13* region with Eosinophilic Esophagitis (EoE) [best at rs12653750, $p = 3.03 \times 10^{-9}$, $OR = 1.73$ 95%CI $= (1.44 - 2.07)$], previously implicated in asthma, allergy, and eosinophilia; and association of variants in *GCKR* and *JAZF1* with allergic rhinitis in our pediatric cohorts [best SNP rs780093, $p = 2.18 \times 10^{-5}$, $OR = 1.39$, 95%CI $= (1.19 - 1.61)$], previously demonstrated in metabolic disease and diabetes in adults.

Conclusion: The PheWAS approach with re-mapping ICD-9 structured codes for our European-origin pediatric cohorts, as with the previous adult studies, finds many previously reported associations as well as presents the discovery of associations with potentially important clinical implications.

Keywords: PheWAS, ICD-9 code, genetic polymorphism

INTRODUCTION

Phenome-wide association study (PheWAS) is a relatively new genomic approach to link clinical conditions with published variants (Denny et al., 2010). The concept, although not new, was originally applied to genomic research by the eMERGE (electronic MEdical Records and GEnomics) network, which is in a unique position to access tens of thousands of Electronic Medical Records (EMR) linked to ICD-9 codes in structured data. Multiple eMERGE PheWAS results have been published that primarily address adult cohorts (Denny et al., 2011, 2013). The phenotypic data used in PheWAS may include ICD-9 codes, epidemiologic data in health surveys, biomarkers, intermediate or quantitative traits (Pendergrass et al., 2011, 2013; Neuraz et al., 2013; Liao et al., 2014). By virtue of this inclusive approach, new hypotheses may be generated that provide insight into genetic architecture of complex traits. Challenges with PheWAS include multiple test corrections across the thousands of phenotypes tested and autocorrelation of some of the phenotypes. Nevertheless, novel robust insights have resulted from PheWAS, for example, genetic association findings with heart rate variability are notable (Ritchie et al., 2013).

PheWAS combines multiple phenotypes from previous GWAS, and identify common SNPs affecting different traits. In this study, we used this approach to evaluate whether known GWAS variants identified in adult diseases can be also identified in children using two EMR-linked pediatric datasets from eMERGE. PheWAS in pediatrics is particularly important because it not only assesses the effect of early age of onset on many established adult-GWAS loci, but also may provide insights into how a primary phenotype during child development develops into one or more diseases in adulthood. A priori, there are several reasons that in principle might make a pediatric PheWAS more challenging. These include the change in heritability with age for several traits (St Pourcain et al., 2014), the flux in the recommendations for pediatric monitoring for traits that are routinely measured in adults (Gidding, 1993; Klein et al., 2010) and the use of cross-sectional standardization rather than longitudinal standardization of developmental traits such as height (Tiisala and Kantero, 1971).

To determine whether robust association signals would be present in the context of these challenges, we conducted the first PheWAS study in pediatrics on our available samples. We successfully translated 93,724 specific ICD-9 diagnostic codes into 1402 distinct PheWAS code groups and 14 major disease concept paths and evaluated 2481 previously published variants. After quality control, only 2476 genetic variants were analyzed in 539 diseases in the two pediatric sites. Finally we replicated 24 genetic variants and identified 14 new possible associations confirming our hypothesis. Our primary results highlight the utility of an EMR-based PheWAS approach as a new line of investigation for discovery of genotype-phenotype associations in pediatrics.

MATERIALS AND METHODS

STUDY SUBJECTS

Protocols for this study were approved by the Institutional Review Boards (IRBs) at the institutions where participants were recruited. All study participants provided written consent prior to study enrolment; consent forms were obtained at each location under IRB guidelines. Children and teens, aged through 19 years old were included. The EMR-linked pediatric emerge cohorts consist of 4560 subjects from Cincinnati Children's Hospital Medical Center (CCHMC) and 1000 subjects from Boston Children's Hospital (BCH). Only those self-reported to have European ancestry were selected for this study (**Table 1**).

SNP PRIORITIZATION

We limit our investigation to particular genetic variants: First, we obtained the list of all previously published SNPs from different public domain databases including The National Human Genome Research Institute (NHGRI) catalog of published Genome-Wide Association Studies (http://www.genome.gov/gwastudies), Genetic Association of Complex Diseases and Disorders (GAD, http://geneticassociationdb.nih.gov), the UCSC Genome Browser database (UCSC, http://genome.ucsc.edu/), Online Mendelian Inheritance in Man (OMIM, http://www.omim.org/), and PharmGKB (pharmgkb, https://www.pharmgkb.org). After linking this collection to PubMed reference numbers, only those with at least one reported of positive associations were selected regardless of the previously observed p values or number of publications. In addition, all downloaded databases were current at the time of this submission. From the filtered variants, 2476 variants were available and assessed in our clean, post-imputation genotyping dataset for analysis.

GENOTYPING AND STATISTICAL ANALYSES

High throughput SNP genotyping was carried out previously in CCHMC and BCH using Illumina™ or Affymetrix™ platforms, as previously described (Namjou et al., 2013). Quality control (QC) of the data was performed before imputation. In each genotyped cohort, standard quality control criteria were met and single nucleotide polymorphisms (SNPs) were removed if (a) >5% of the genotyping data was missing, (b) out of Hardy-Weinberg equilibrium (HWE, $p < 0.001$) in controls, or a minor allele frequency (MAF) <1%. Samples with call rate <98% were excluded.

Recently all eMERGE cohorts have also undergone whole genome imputation. The details of these procedures are available in this issue of Frontiers in Genetics (Setia et al., 2014). Briefly, the imputation pipeline was implemented using IMPUTE2 program and the publicly available 1000-Genomes Project as the reference haplotype panel composed of 1092 samples (release version 2 from March 2012 of the 1000 Genomes Project Phase I, ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521) (Howie et al., 2011). The eMERGE imputed data provided to us were already filtered, i.e., imputed data with a threshold of 0.90 for the genotype posterior probability and with a IMPUTE2 info score > 0.7 (Howie et al., 2011). Principle component analysis (PCA) performed to identify outliers and hidden population structure using EIGENSTRAT (Price et al., 2006). The first two principle components explained most of the variance and were retained and used as covariates during the association analysis in order to adjust for population stratification. In addition, 14 outlier samples were removed. To illustrate the overall inflation rate a phenotype with sufficient number of cases and

Table 1 | The demographic distribution of the European ancestry population (CCHMC-BCH).

	Cohort names	#Europeans	M/F	Mean age (95%CI)	Array
BCH*	The gene partnership	727	449/278	13.30(12.97–13.66)	Affymetrix-Axiom
CCHMC**	Cytogenetics	1228	758/470	7.32(7.03–7.62)	Illumina-610
	Cytogenetics	609	373/236	7.18(6.73–7.63)	Illumina-Omni-1
	EoE†	543	394/149	12.27 (11.70–12.67)	Illumina-Omni-5
	JIA‡	488	101/387	13.70(13.13–14.23)	Affymetrix-6
	Cincinnati- control cohorts	673	329/344	13.50(13.25–13.84)	Illumina-Omni-5
Total		4268	2403/1865	11.52(11.16–11.91)	

*BCH, Boston Children's Hospital; **CCHMC, Cincinnati Children's Hospital Medical center; †, Eosinophilic Esophagitis (EoE) cohorts; ‡, Juvenile Idiopathic Arthritis cohorts (JIA). The details of platforms used have been described elsewhere (Namjou et al., 2013).

controls has been selected (autism) and the inflation of $\lambda = 1.03$ was obtained.

Next, from our prioritized SNP list mentioned above, 2481 variants were available. Five of these SNPs had a site-specific effect with either CCHMC or BCH ($p < 10^{-5}$ for the difference between sites) and were removed from final analyses. For each phenotype, logistic regression was performed between cases and control adjusted for two principal components using PLINK (Purcell et al., 2007). To investigate whether either the phenotype or the genotype has an effect on the outcome variable, we perform phenotypic and genotypic conditional analyses, controlling for the effect of a specific SNP or phenotype. After pruning of highly correlated SNPs ($r^2 > 0.5$), we used false discovery rate (FDR) methods to correct for multiple testing using the Benjamini–Hochberg procedure implemented in PLINK (Purcell et al., 2007). As a result of LD pruning 1828 independent variants were used for the purpose of FDR estimation. *Q* values correspond to the proportion of false positives among the results. Thus, *Q* values less than 0.05 signify less than 5% of false positives and are accepted as a measure of significance (FDR < 0.05) in this study. For any novel PheWAS findings, an adaptive permutation approach was performed using a sample randomization strategy in which case and control labels were permuted randomly (with up to 1,000,000 trials) in order to obtain empirical *p* values [PLINK (Purcell et al., 2007)]. We also report previous known effects that only produce suggestive findings in our study ($0.05 < p < 0.001$). Sample size and power calculations based on the size effect and risk allele frequency were estimated using QUANTO (Gauderman and Morrison, 2006). To graphically display results, LocusZoom was used (Pruim et al., 2010).

PHENOTYPING

A phenome-wide association analysis (PheWAS) was performed in which presence or absence of each PheWAS code [mapped from translated ICD-9 codes as per Carroll et al., 2014)] were considered as a binary phenotype. The per-patient ICD-9 codes were obtained from the i2b2 Research Patient Data Warehouse at CCHMC and BCH. Also, these PheWAS codes were used to define comparison control groups by excluding the PheWAS case- code and those closely related to them in the ICD-9 hierarchy. Control groups for Crohn's Disease (CD), for instance, excluded CD, ulcerative colitis, and several other related gastrointestinal complaints. Similarly, control groups for myocardial infarction excluded patients with myocardial infarctions, as well as angina and other evidence of ischemic heart disease. The current PheWAS map and PheWAS script written in R is available [http://phewascatalog.org, (Carroll et al., 2014)]. In this study, subgroups of European cases with more than 20 samples were selected for PheWAS association study (539 subgroups) and the available published SNPs that passed quality controls were evaluated. The case cohorts for the two phenotypes of Juvenile Idiopathic Arthritis (JIA) and Eosinophilic Esophagitis (EoE) have both been previously published as parts of larger phenotype specific studies (Rothenberg et al., 2010; Thompson et al., 2012; Hinks et al., 2013). The origin of all case records is presented in **Table 1**. In this study, Juvenile Onset Rheumatoid Arthritis (JRA) is identified by ICD-9 codes and designated as JRA; when the criteria for Juvenile Idiopathic Arthritis (JIA) were applied in the studies of others (Thompson et al., 2012), then this phenotype was referred to as JIA.

RESULTS

In this study only European ancestry was included in the analysis to avoid potential bias induced by ancestry. The demographic distribution of the European ancestry population under study (**Table 2**) had 93,724 specific ICD-9 diagnostic codes representing 1402 distinct PheWAS code groups and 14 major disease concept paths. The frequencies of concept path hierarchy of the ontology (**Figure 1**) show the neuropsychiatric concept path as the most frequent and neoplastic and infection paths as the least frequent.

Replication of existing associations using PheWAS

We compared SNPs with previous GWAS-reports and present association findings (FDR-q < 0.05) after corrected for population stratification and standard quality control (**Table 2**).

First, for the two phenotypes of JRA and EoE samples overlap largely with those previously reported phenotype specific GWAS study (Rothenberg et al., 2010; Thompson et al., 2012; Kottyan et al., 2014). We reproduced the major findings of those publications using different methodology. For JRA, association with *PTPN22* is a consistent finding. As expected, we replicated a previous report of association of *PTPN22* at non-synonymous coding SNP rs2476601 with this phenotype and with the same direction

Table 2 | Replication of previous GWAS association results in CCHMC/BCH pediatric cohorts.

Chr	SNP	Position	Gene	Minor allele	Case	Control	*p* value	FDR*q* value	OR	Description	Case/ Control
1	rs2476601	114377568	*PTPN22*	A	0.16	0.09	9.10E-07	8.01E-06	1.87 (1.46–2.41)	JRA	272 /3412
1	rs2476601	114377568	*PTPN22*	A	0.28	0.10	2.78E-05	4.16E-04	3.44 (1.80–6.57)	Thyroiditis	23/3571
1	rs2476601	114377568	*PTPN22*	A	0.18	0.10	0.007	NS	1.96 (1.16–3.31)	T1DM	47/3609
1	rs6679677	114303808	*PTPN22*	A	0.16	0.09	3.63E-07	4.15E-06	1.92 (1.49–2.47)	JRA	272/3412
1	rs6679677	114303808	*PTPN22*	A	0.28	0.10	2.00E-05	4.16E-04	3.52 (1.84–6.74)	Thyroiditis	23/3571
1	rs6679677	114303808	*PTPN22*	A	0.18	0.10	0.005	NS	2.00 (1.18–3.38)	T1DM	47/3609
2	rs3771180	102953617	*IL1RL1*	T	0.19	0.14	5.71E-05	0.0005	1.46 (1.19–1.80)	EoE or Food Allergy	599/2346
2	rs7574865	191964633	STAT4	T	0.32	0.24	0.004	NS	1.46 (1.11–1.92)	Wheezing	125/3372
3	rs78122814	85200034	*CADM2*	A	0.08	0.05	4.34E-05	0.0004	1.72 (1.32–2.24)	Autism	601/1840
5	rs3806932	110405675	*TSLP*	G	0.35	0.44	5.59E-07	8.38E-06	0.69 (0.59–0.80)	EoE	446/2586
5	rs272889	131665378	*SLC22A4*	A	0.46	0.37	1.53E-05	0.0003	1.45 (1.22–1.71)	Atopic Dermatitis	298/3031
5	rs12653750	131971902	*IL5-IL13*	T	0.27	0.20	9.74E-05	0.005	1.50 (1.22–1.84)	Eosinophilia	250/3344
6	rs75732170	101845494	*GRIK2*	A	0.06	0.03	8.49E-06	0.0002	2.00 (1.47–2.73)	Autism	601/1840
6	rs477515	32569691	*HLA-DRB1*	A	0.17	0.33	1.15E-12	8.62E-12	0.41 (0.32–0.53)	JRA	272/3412
6	rs477515	32569691	*HLA-DRB1*	A	0.07	0.33	1.12E-06	2.60E-05	0.16 (0.08–0.38)	Uveitis	51/3089
6	rs622137	32569852	*HLA-DRB1*	A	0.17	0.32	4.98E-13	5.78E-12	0.41 (0.32–0.53)	JRA	272/3412
6	rs2516051	32570184	*HLA-DRB1*	T	0.17	0.32	5.78E-13	5.78E-12	0.41 (0.32–0.53)	JRA	272/3412
6	rs2516049	32570400	*HLA-DRB1*	C	0.14	0.32	1.49E-15	4.48E-14	0.36 (0.27–0.46)	JRA	272/3412
6	rs660895	32577380	*HLA-DRB1*	G	0.42	0.21	7.85E-07	1.65E-05	2.73 (1.80–4.13)	T1DM	47/3609
6	rs9388489	126698719	*CENPW*	G	0.68	0.47	3.07E-05	0.0003	2.46 (1.58–3.80)	T1DM	47/3609
6	rs1490388	126835655	*CENPW*	T	0.68	0.47	4.29E-05	0.0003	2.42 (1.56–3.74)	T1DM	47/3609
9	rs7850258	100549013	*FOXE1*	A	0.15	0.34	0.005	NS	0.35 (0.15–0.78)	Thyroiditis	23/3571
9	rs1443438	100550028	*FOXE1*	T	0.15	0.34	0.009	NS	0.35 (0.15–0.78)	Thyroiditis	23/3571
10	rs12411988	65315397	*REEP3*	C	0.20	0.14	9.50E-05	0.005	1.53 (1.23–1.92)	JRA	272/3412
10	rs7903146	114758349	*TCF7L2*	T	0.44	0.29	0.001	NS	2.00 (1.29–3.08)	Abnormal Glucose Test	42/3609
16	rs12924729	11187783	*CLEC16A*	A	0.26	0.35	3.34E-08	9.08E-06	0.67 (0.58–0.77)	EoE or Food Allergy	599/2346
17	rs8067378	38051348	*GSDMB*	A	0.57	0.49	3.13E-06	0.0001	1.37 (1.19–1.57)	Asthma	499/3175
17	rs2290400	38066240	*GSDMB*	C	0.43	0.50	1.05E-05	0.0002	0.74 (0.64–0.84)	Asthma	499/3175
17	rs8074094	45348021	*ITGB3*	C	0.30	0.25	2.00E-05	0.0002	1.29 (1.15–1.45)	PDD	1141/1840
20	rs716316	14908741	*MACROD2*	T	0.32	0.39	2.01E-05	0.0003	0.74 (0.65–0.85)	Autism	601/1840

False discovery rate (FDR-q < 0.05) was set for the threshold of significance. The calculated odds ratio was based on minor allele frequency and the coded alleles were shown. All positions were based on NCBI build 37. NS (not significant). The p-values and q-values are ordered based on chromosome and position.

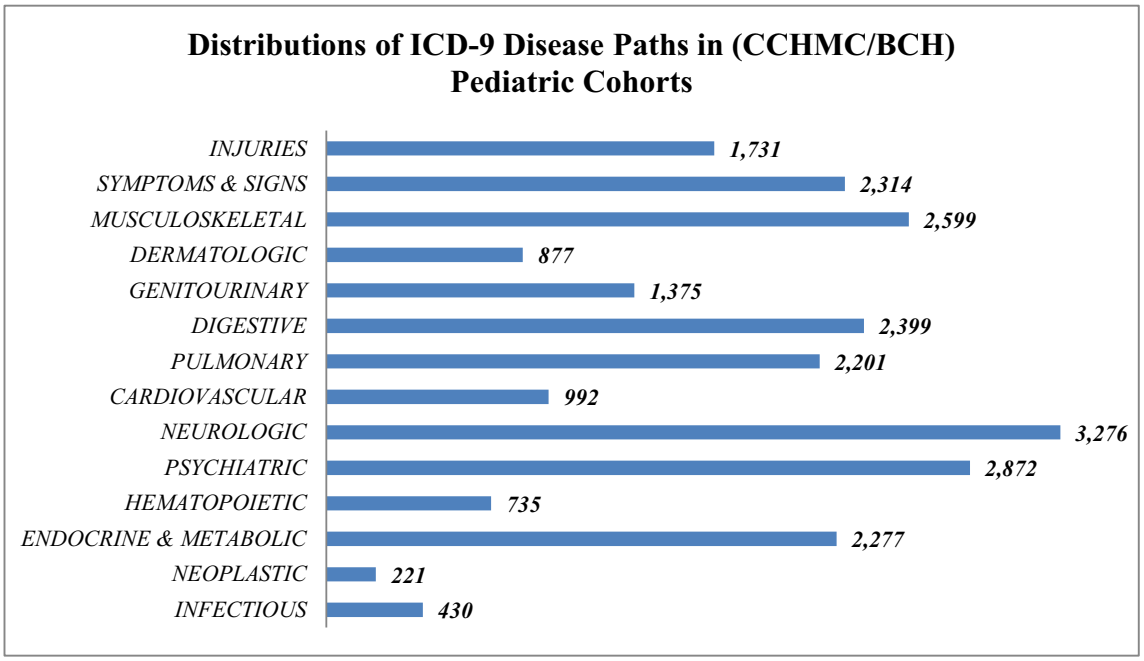


FIGURE 1 | Frequency and distribution of 14 major ontology concept path categories from CCHMC/BCH European pediatric cohorts.

of allele frequency, ($p = 9.10 \times 10^{-7}$, $OR = 1.87$, 95%CI 1.46 − 2.40). The SNP in proxy (rs6679677, $r^2 = 1$) also produced a similar result (**Table 2**). In our cohorts, variants in *PTPN22* are also associated with thyroiditis as well as Type 1 diabetes mellitus (T1DM), consistent with previous reports and despite low sample size (**Table 2**) (Plenge et al., 2007; Todd et al., 2007; Lee et al., 2011). From these three known associations of *PTPN22*, i.e., JRA, T1DM, and thyroiditis, the largest magnitude of the association is with pediatric onset thyroiditis (**Table 2**, $OR = 3.52$ 95%CI 1.84 − 6.75).

For JRA, multiple loci in the HLA region were also associated at the level of $p < 10^{-12}$ including rs477515 and rs2516049 near *HLA-DRB1* (**Table 2**). Of note, the size effect of HLA related SNPs, were highest for those with coexisting uveitis (best SNP rs477515, $OR = 6.5$, 95% CI = 2.73 − 15.68 for the risk allele, **Table 2**). In addition, for JRA, another previously published association (rs12411988 in *REEP3*) was also found and with the same size effect as previously described ($OR = 1.53$) (**Table 2**) (Thompson et al., 2012).

Furthermore, with regard to EoE traits, we also replicated previous major finding of association of SNP rs3806932 located at the vicinity of the *TSLP* gene at 5q22 region [$p = 5.59 \times 10 - 7$, $OR = 0.69$ (95%CI = 0.59 − 0.80)] in these cohorts (**Table 2**) (Rothenberg et al., 2010; Kottyan et al., 2014).

For asthma, the best PheWAS results were detected at 17q21 which includes *GSDMB* and has been previously reported to be associated specifically with childhood onset Asthma (Verlaan et al., 2009). In fact, the best associated SNP rs8067378 in our cohorts [$p = 3.13 \times 10^{-6}$, $OR = 1.37$ (1.19 − 1.57)], tags the asthma associated haplotype in which the allele-specific expression analyses for this haplotype has previously shown strong association with Asthma risk (Verlaan et al., 2009). There is strong support for this association from a cluster of variants in this neighborhood (**Figure 2A**).

The minor allele (T) of the intronic SNP rs7903146 in *TCF7L2* is one of the larger magnitude and more frequently identified associations in Type 2 diabetes mellitus (T2DM) and hyperlipidemia in many adult GWAS studies (Lyssenko et al., 2007; Huertas-Vazquez et al., 2008). In fact, the best PheWAS trait in our cohorts at this variant was also related to T2DM and hyperlipidemia as well, although our sample size was small. In this family of ICD-9 codes the best suggestive result was obtained for an abnormal glucose test with [$p = 0.001$, $OR = 2.00$ (95%CI 1.29 − 3.08)] (**Table 2**).

Specifically, for T1DM, in addition to the positive association with *PTPN22* mentioned above, additional published loci were confirmed and with relatively larger effect sizes ($OR > 2$) including known HLA-SNP rs660895 [$p = 7.85 \times 10^{-7}$, $OR = 2.73$ (95%CI = 1.80 − 4.13)], as well as variants near *CENPW* that previously have been reported for this trait (**Table 2**) (Barrett et al., 2009).

Other effects

Several loci previously associated with autism and pervasive developmental disorders (PDD) (GWAS or copy number variations reports) including those at *MACROD2, ITGB3, CADM2,* and *GRIK2* (Jamain et al., 2002; Weiss et al., 2006; Thomas et al., 2008; Anney et al., 2010) also provided evidence of association in our cohorts for these traits (**Table 2**). Variants in the *FOXE1* gene that have been previously associated with primary hypothyroidism and thyroiditis in adult eMERGE cohorts (Denny et al., 2011), produced a trend of association and consistent in

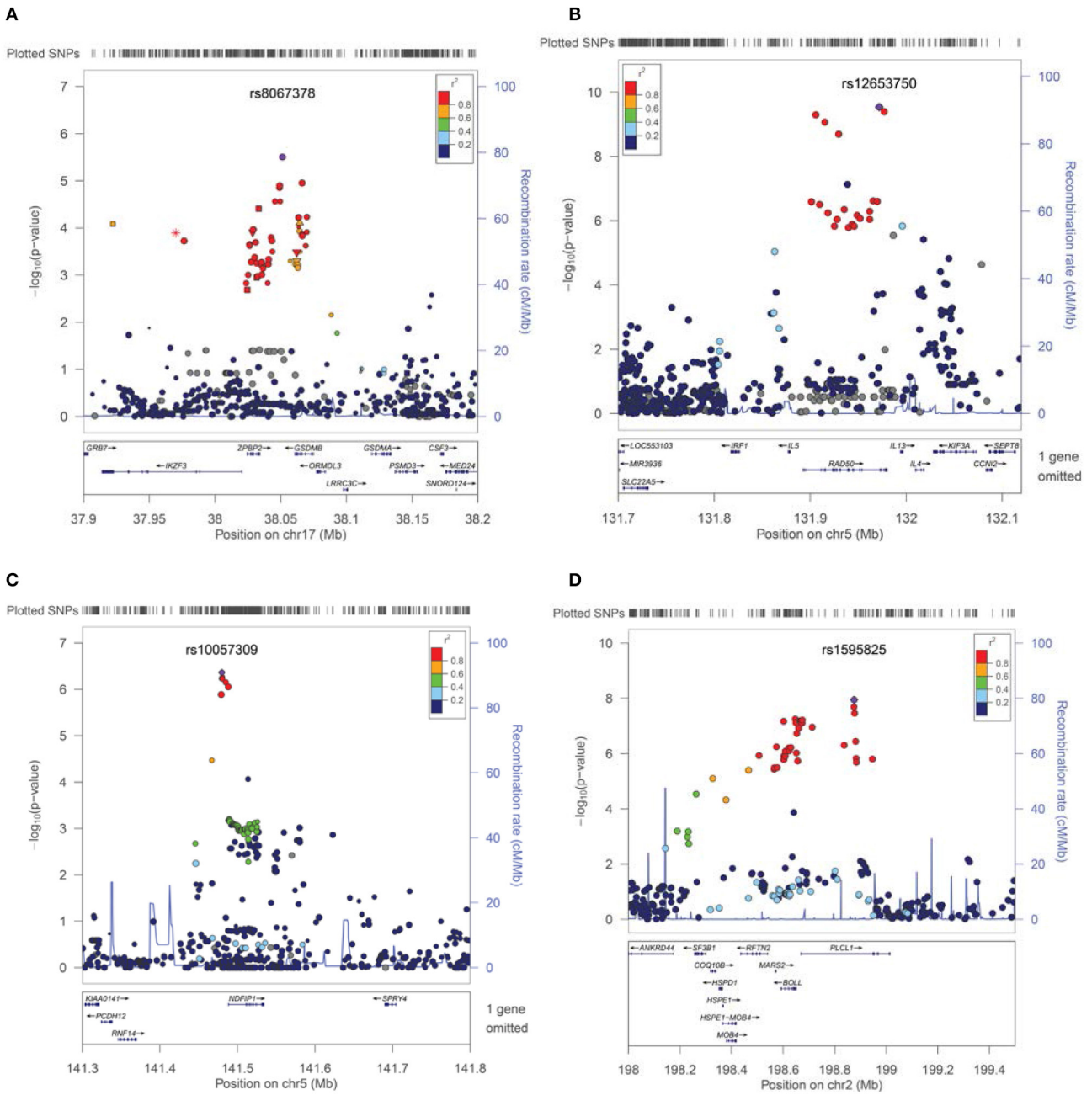


FIGURE 2 | Association results and signals contributing to Asthma, Eosinophilic Esophagitis, Mental Retardation, and Developmental Delays. SNPs are plotted by position in a 0.2 Mb window against association signals (−log10 *P*-value). For each trait, the most significant SNP is highlighted. Estimated recombination rates (from HapMap) are plotted in cyan to reflect the local LD structure. The SNPs surrounding the most significant SNP, are color-coded to reflect their LD with identified SNP (taken from pairwise r^2 values from the HapMap CEU database, www.hapmap.org). Regional plots were generated using LocusZoom (http://csg.sph.umich.edu/locuszoom). **(A)** Cluster of the association effect for asthma at 17q21 near the gasdermin-B (GSDMB) gene. **(B)** Association signal for Eosinophilic Esophagitis at 5q31 (IL5-IL13 cluster region). **(C)** Cluster of association near the NDFIP1 gene for Mental Retardation traits. **(D)** Plot of association effects in the PLCL1 region for Developmental Delays-Speech Disorders.

directionality with thyroiditis in our pediatric cohorts despite low sample size (**Table 2**). No gene-gene interaction was evident between *PTPN22* and *FOXE1* for hypothyroidism in these data. Rs7574865 is a SNP in the third intron of the *STAT4* that has been associated with SLE and related autoimmune diseases (Namjou et al., 2009). In these cohorts, pediatric onset lupus was under-represented (less than 20 cases), however, suggestive associations with wheeze and asthma were detected [$p = 0.004$, OR 1.46 (95%CI = 1.11 − 1.92) (**Table 2**)] with the same direction of the difference in allele frequency previously observed in autoimmune

traits. This possible association has also been reported in another study (Pykäläinen et al., 2005). Of note, in contrast to rheumatoid arthritis, the *STAT4* association effect was weak for JRA in our cohorts (effect size = 1.12, $p = 0.17$). GWAS studies have linked Inflammatory Bowel disease (IBD) to a number of IL-23 pathway genes, in particular *IL23R*. The well-known coding variant in the IL23 receptor (rs11209026) also showed a trend toward association with IBD in our cohorts with the same allelic direction but due to low sample size (31 cases) it did not reach significance (FDR-q > 0.05) (Li et al., 2010) (data not shown).

Novel findings from this PheWAS

A number of potentially novel associations remained significant after the permutation procedure to assess the probability of the observed distribution with beta > 0.8 FDR-q < 0.05 (**Table 3**). Variants in the Glucokinase Regulator gene (*GCKR*) have been previously implicated in metabolic disease, diabetes and hypertriglyceridemia in adults (Bi et al., 2010; Onuma et al., 2010) and were mostly associated with allergic rhinitis in our pediatric cohorts [best SNP rs780093 $p = 2.18 \times 10^{-5}$, $p_{(\text{perm})} = 8.06 \times 10^{-5}$, $OR = 1.39, 95\%\text{CI} = (1.19 - 1.61)$] (**Table 3**), while no significant association was found for diabetes. Indeed, conditional analyses, controlling for diabetes related traits suggest that this is an independent effect (p-conditional $= 6.75 \times 10^{-5}$). Another major regulatory locus for diabetes in adults, *JAZF1*, also was associated with allergic rhinitis in our cohorts (**Table 3**) even after controlling for diabetes (p-conditional $= 8.46 \times 10^{-5}$, for rs1635852). No significant gene-gene interaction was detected between these two loci or with *TCF7L2*.

Variants in a cytokine cluster of the *IL5-IL13* region, which is known to be associated with Asthma, Allergy, Atopic Dermatitis (AD) and Eosinophilia, produced a cluster of association with EoE in our cohorts [best SNP rs12653750, $p = 3.03 \times 10^{-9}$, $p_{(\text{perm})} = 1.00 \times 10^{-6}$, $OR = 1.73\ (1.44 - 2.07)$] (Bottema et al., 2008; Granada et al., 2012). There is a cluster of significant variants in this neighborhood of chromosome 5 (5q31) associated with EoE (**Figure 2B**). In our cohorts, weaker associations can be detected for all allergy-related phenotypes with the association with Eosinophilia being the most impressive [$p = 9.74 \times 10^{-5}$ (**Table 2**)]. However, conditional analyses and controlling for Asthma and Eosinophilia suggest that an independent effect still exists for EoE at this locus using EMR data (conditional $p = 9.74 \times 10$ - 5 for rs20541). Moreover, no long distance linkage disequilibrium between rs3806932 in *TSLP* gene at 5q22 and rs20541 was detected in this population ($r^2 = 0.0002$, $D' = 0.02$).

We also observed association with AD within this cytokine cluster consistent with previous reports (Paternoster et al., 2011). However, the best associated SNP for AD (rs272889) was located at *SLC22A4* in our population (**Table 2**). These two variants, rs272889 and rs12653750, were separated by more than 300kb with low linkage disequilibrium ($r^2 < 0.1$). A residual effect still exists for AD and rs272889 after controlling for EoE status or the rs12653750 variant that suggests a distinct effect (p-conditional = 0.002). Noteworthy, with regard to AD, another reported SNP (rs2897442) downstream of this cluster at *KIF3A* gene produced only a suggestive association ($p = 0.005$) in our cohort (data not shown).

Table 3 | Novel PheWAS findings in CCHMC/BCH pediatric cohorts.

Description	Case/Control	Chr	SNP	Position	Gene	Minor allele	Case	Control	*p* value	p-permute	Case needed*	OR
Allergic rhinitis	408/2754	2	rs1260326	27730940	*GCKR*	T	0.48	0.41	7.02E-05	1.21E-04	250	1.36 (1.17–1.58)
Allergic rhinitis	408/2754	2	rs780094	27741237	*GCKR*	T	0.47	0.40	2.94E-05	9.61E-05	250	1.38 (1.19–1.60)
Allergic rhinitis	408/2754	2	rs780093	27742603	*GCKR*	T	0.47	0.40	2.18E-05	8.06E-05	250	1.39 (1.19–1.61)
Allergic rhinitis	408/2754	7	rs864745	28180556	*JAZF1*	C	0.43	0.50	9.02E-05	1.11E-04	220	0.76 (0.65–0.88)
Allergic rhinitis	408/2754	7	rs1635852	28189411	*JAZF1*	C	0.43	0.50	6.58E-05	5.97E-05	220	0.75 (0.65–0.87)
Eosinophilic Esophagitis	446/2586	5	rs4143832	131862977	*IL5-IL 13*	T	0.24	0.18	4.70E-06	1.70E-05	200	1.55 (1.29–1.87)
Eosinophilic Esophagitis	446/2586	5	rs12653750	131971902	*IL5-IL 13*	T	0.28	0.19	3.03E-09	1.00E-06	100	1.73 (1.44–2.07)
Eosinophilic Esophagitis	446/2586	5	rs20541	131995964	*IL5-IL 13*	A	0.26	0.19	3.72E-07	3.00E-06	150	1.61 (1.34–1.94)
Mental retardation	297/1840	5	rs11167764	141479065	*NDFIP1*	A	0.29	0.20	1.29E-06	4.00E-06	150	1.66 (1.35–2.04)
Mental retardation	297/1840	5	rs77110703	141479833	*NDFIP1*	T	0.29	0.20	5.83E-07	2.00E-06	150	1.69 (1.38–2.08)
Mental retardation	297/1840	5	rs10057309	141479870	*NDFIP1*	T	0.29	0.20	4.33E-07	2.00E-06	150	1.70 (1.39–2.09)
Developmental disorders	975/1840	2	rs1595825	198875464	*PLCL1*	A	0.15	0.21	1.13E-08	2.00E-06	150	0.65 (0.57–0.76)
Supporative otitis media	362/3082	1	rs10801047	191559356	*near RGS1*	A	0.13	0.08	1.61E-06	2.00E-06	250	1.77 (1.40–2.24)
Depression	107/2864	14	rs7141420	79899454	*NRXN3*	C	0.66	0.46	4.76E-05	1.10E-04	100	1.78 (1.34–2.34)

**P (permute): empirical permutation p values after case and control labels are permuted randomly (up to 1,000,000). All results were at the level of FDR-q < 0.05.*

*** "Cases needed" refers to the estimated number of cases needed to achieve 80% power to detect an association at alpha = 0.05 given the identified odds ratio and the MAF in this population.*

Because of the pleotropic effects between EoE and other allergy related traits, in addition to conditional analyses, we also found possible synergistic effects. One of the closely related phenotypes with EoE is the presence of food allergy. When we combined these two as a subgroup, two additional effects were identified. One cluster was in *IL1RL1* that was previously associated with the related phenotype, i.e., allergy and asthma (best SNP rs3771180, $p = 5.71 \times 10^{-5}$, **Table 2**, Torgerson et al., 2011) and another was in *CLEC16A*, previously associated with different autoimmune diseases [best SNP rs12924729, $p = 3.34 \times 10^{-8}$ (**Table 2**), (Mells et al., 2011)] and was reported as a suggestive effect in recent GWAS study for EoE (Kottyan et al., 2014).

Variants near *RGS* cluster of genes on chromosome 1, previously reported to be associated with IBD and other autoimmune diseases (Hunt et al., 2008; Esposito et al., 2010), were associated with susceptibility to infection, in particular suppurative otitis media [best SNP rs10801047, $p = 1.61 \times 10^{-6}$, $p_{(\text{perm})} = 2.00 \times 10^{-6}$, $OR = 1.77$ 95%CI $= 1.398 - 2.24$].

New association signals have been detected near the *NDFIP1* gene for mental retardation related traits. Variants near this gene that is expressed mostly in brain, were previously reported to be associated with IBD through an unknown mechanism and with a risk effect for major allele (SNP = rs11167764) (Franke et al., 2010). Instead, we found a risk effect for the minor allele [best SNP rs10057309, $p = 4.33 \times 10^{-7}$, $p_{(\text{perm})} = 2.00 \times 10^{-6}$, $OR = 1.702$, 95%CI $= 1.38 - 2.09$] (**Table 3**). Similarly, cerebral palsy, which is linked to mental retardation, was also associated with this variant ($p = 9.00 \times 10^{-4}$). However, conditional analyses controlling for cerebral palsy suggest an independent effect for overall mental retardation (conditional $p = 8.00 \times 10^{-4}$). Furthermore, excluding the small number of samples with known chromosomal abnormalities ($N < 40$) did not affect this result. The overall cluster effect in this neighborhood for mental retardation bolsters the suspicion that an association is found here (**Figure 2C**).

Additionally, for developmental delays of speech and language, a novel signal effect was detected in the *PLCL1* gene at chromosome 2 [best SNP rs1595825, $p = 1.13 \times 10^{-8}$, $OR = 0.65$ $(0.57 - 0.76)$] (**Figure 2D**, **Table 3**). Weaker associations ($0.01 > p > 0.00001$) were also detected for related neurologic phenotypes including abnormal movement, lack of coordination and epilepsy at this locus (data not shown).

NRXN3 polymorphisms that have been previously reported to be associated with substance dependence (Docampo et al., 2012), smoking behavior and attention related problems (Stoltenberg et al., 2011), were associated with depression in our pediatric cohorts (**Table 3** Noteworthy, the major allele of our reported SNP (rs7141420) has been linked to obesity in adult cohorts (Berndt et al., 2013), while we found association with the minor allele for depression [$p = 4.76 \times 10^{-5}$, $OR = 1.78$ $(1.34 - 2.34)$, **Table 3**]. Furthermore, rare micro-deletions in this gene were previously reported for Autism case reports but these rare variants are not available to assess in our genotyped cohorts (Vaags et al., 2012).

DISCUSSION

This first pediatric PheWAS finds 38 associations, 24 previously known phenotype-genotype associations in a pediatric population using EMR-linked eMERGE databases and identified 14 new possible associations at beta > 0.8 and FDR-q < 0.05. From analysis performed on EMR-linked data from 4268 European individuals, we successfully confirmed several major effects for phenotypes with moderate to large sample size, in particular for Asthma, Autism, and neurodevelopmental disease as well as several effects for Type 1 and Type 2 Diabetes (T1DM, T2DM) and Thyroiditis. Almost all of the significant phenotype associations were with common variants (MAF $> 10\%$) (**Tables 2**, **3**). In addition, we compared and verified the consistency of allele frequency of reported markers among cohorts, sample collection sites and with CEU-Hapmap data. Considering a desired power of 0.8, for variants at the fixed allele frequency of 10% and size effect of 1.5 or above, 200 cases are sufficient to detect association at an alpha level of 0.05. Indeed, we have surpassed this level for most of our reported traits. In addition, for all reported phenotypes the control sample was at least two or three times larger than cases (**Tables 2**, **3**). Importantly, since our control samples for each trait are an EMR-derived population and not healthy individuals, this large number of control samples provides minor allele frequencies consistent with hapmap-CEU frequencies for all of our reported variants.

The results for JRA and EoE depend upon previously published studies of these phenotypes. While the case samples are mostly identical, the control samples were substantially different. Consequently, we cannot refer to these particular findings as constituting confirmation and yet our results and different methodology support the previous reports.

In addition, we also identified several novel PheWAS findings for pediatric traits in particular for Allergic Rhinitis, Otitis Media, EoE, Mental Retardation, and Developmental Delays all with sufficient power (beta > 0.8) (**Table 3**, **Figures 2B–D**). This study, however, is underpowered to make discoveries for rare variants or uncommon traits. The power to detect a finding in PheWAS is determined by many factors, including sample size, risk allele frequency, effect size, model of inheritance, the effect of environment and the prevalence of a phenotype within the population.

Similar to previous studies, we also observed pleiotropy for a number of loci in particular *PTPN22* for JRA, T1DM, and Thyroiditis, IL5 for Eosinophilia, Asthma, and EoE and *NDFIP1* for Mental Retardation traits and Cerebral Palsy. These pleotropic effects are specifically expected to be due to underlying biologic correlations. On the other hand, we rarely observed simultaneous robust associations with multiple unrelated phenotypes that had sufficient power. Furthermore, one of the advantages of PheWAS studies is the ability to control the granularity of a database with regard to related phenotypes. For example, by combining two related phenotypes such as uveitis with JRA or food allergy with EoE, we were able to evaluate new subgroups and identify new loci responsible for shared underlying pathways that otherwise cannot be detected or require much larger sample sizes. Further studies with larger sample sizes would be useful to test and perhaps corroborate these findings.

Association of Allergic Rhinitis with loci responsible for diabetes in adults (*GCKR-JAZF1*) may highlight a shared underlying mechanism. In fact, the connection between allergy and diabetes

has been previously suggested in humans but cannot be explained by the Th1/Th2 paradigm (Dales et al., 2005). Moreover, in animal experiments, treating mice with mast cell-stabilizing agents reduced diabetes manifestations (Liu et al., 2009). It is also possible that in our pediatric cohorts we have under-diagnosed children who are diagnosed with diabetes which would appear in a later stage of development. In fact, *GCKR* is an inhibitor of glucokinase (*GCK*), a gene responsible for the autosomal dominant form of T2DM that usually develops later in life and in adulthood. Of note, neither of these two loci showed significant association with Body Mass Index (BMI) in our previous report with these data nor has the obesity link been established in adult studies (Namjou et al., 2013).

The novel association of a cytokine cluster in the *IL5-IL13* region for the EoE trait is particularly interesting since anti-IL5 monoclonal antibodies have been recommended as a novel therapeutic agent for EoE and other eosinophilia–related traits (Corren, 2012). In general, both *IL5* and *IL13* play a major role for regulation of maturation, recruitment, and survival of eosinophils and the variant reported here has been previously associated with other allergic-related traits and with the same direction of allele frequency difference (Bottema et al., 2008; Granada et al., 2012). In particular, a non-synonymous polymorphism in the IL13 gene, rs20541 (R130Q) (**Table 3**), has been shown to be associated with increased IL-13 protein activity, altered IL-13 production, and increased binding of nuclear proteins to this region (van der Pouw Kraan et al., 1999). Perhaps, the association is a reflection of linkage disequilibrium with another polymorphism in the 5q31 region. In fact, in our analyses residual effect still exists for the best SNP (rs12653750), shown in **Figure 2B** after controlling for rs20541 (p-conditional = 2.27×10^{-5}) ($r^2 = 0.35$). This possible association did not reach significance in previous GWAS studies for EoE and had only produced a suggestive effect ($0.05 < p < 0.001$). Perhaps, this behavior is explained partly by phenotypic heterogeneity since minor allele frequency of independent set of both control populations were the same. Indeed, we found that those with the subphenotype of EoE with Eosinophilia had the strongest size effect ($OR = 1.83$, 95%CI $= 1.44 - 2.32$) and our cohorts were enriched with this subphenotype [177 of total 446 EoE cases (40%)]. Of note, the SNPs in this region were originally selected because of eosinophilia-related publications (Bottema et al., 2008; Granada et al., 2012).

Moreover, combining subgroups of patients with food allergy and EoE revealed two new loci that may explain shared etiology. Indeed, the connection between allergy and Interleukin 1 receptor-like-1 (*IL1R1*) is already known (Torgerson et al., 2011). The ligand for IL1R1, IL-33, is a potent eosinophil activator (Bouffi et al., 2013). Interestingly, there is also a report of association of *CLEC16A* variants with allergy in large analysis with more than 50,000 subjects from 23andMe Inc. (Hinds et al., 2013). C-type lectin domain family 16, also known as *CLEC16A*, is mostly associated with autoimmune related traits and is highly expressed in B lymphocytes and natural killer cells. The molecular and cellular functions of CLEC16A are currently under investigation.

Our conditional analyses suggest an independent effect at the *SLC22A4* gene for Atopic Dermatitis. This solute carrier family gene is predominantly expressed in CD14 cells and has an important role for elimination of many endogenous small organic cations as well as a wide array of drugs and environmental toxins. The associated SNP, rs272889, has been previously shown to be correlated with blood metabolite concentration (Suhre et al., 2011). Other variants in this gene were associated with IBD and Crohns disease as well (Feng et al., 2009). Of note, a key substrate of this transporter is ergothioneine, a natural antioxidant, which Mammalia acquire exclusively from their food. Ergothionine is a powerful antioxidant though its precise physiological purpose remains unclear.

Asthma is associated at the 17q21 in our cohorts (**Figure 1**). The best associated SNP, rs8067378, is known to function as a cis-regulatory variant that correlates with expression of the *GSDMB* gene (Verlaan et al., 2009). Variants in *GSDMB* have been shown to determine multiple asthma related phenotypes specifically in childhood asthma including associations with lung function and disease severity (Tulah et al., 2013). These gasdermin-family genes are implicated in the regulation of apoptosis mostly in epithelial cells and have also been linked to cancer; however, their actual function with respect to disease association remains unknown. The associated variants in this cluster are suspected to be regulatory SNPs that govern the transcriptional activity of at least three nearby genes (*ZPBP2, GSDMB, and ORMDL3*) (Verlaan et al., 2009).

We confirmed several loci responsible for Autism and Pervasive Developmental Disease including *MACROD2, ITGB3, CADM2, and GRIK2. ITGB3* has been known as a quantitative trait locus (QTL) for whole blood serotonin levels (Weiss et al., 2004, 2006). Serotonin is a monoamine neurotransmitter that has long been implicated in the etiology of Autism. In fact, about 30 percent of patients with autism have abnormal blood serotonin levels (Weiss et al., 2004). Similarly, *GRIK2* is an ionotropic glutamate receptor associated with autism (Cook, 1990; Cook et al., 1997). *CADM2* is a member of the synaptic cell adhesion molecule with roles in early postnatal development of the central nervous system (Thomas et al., 2008). The function of *MACROD2* (previously c20orf133) is still largely unknown. For Autism that is more commonly seen in males, we found no significant gender effect for these loci.

Association of variants in the neighborhood of RGS cluster genes with suppurative otitis media is another novel finding. SNPs in this region have been previously linked to celiac disease, multiple sclerosis and other autoimmune diseases (Hunt et al., 2008; Esposito et al., 2010). The link between susceptibility to infection and autoimmunity has been long suggested given the fact that the level and regulation of RGS proteins in lymphocytes also significantly impact lymphocyte migration and function. In our pediatric cohort the number of patients with celiac disease was small ($n = 23$) and the association was not detected. Interestingly, one of the major risk variant for celiac disease, rs13151961 (KIAA1109), as well as known HLA variants, produced a tread toward association for celiac disease but did not pass the FDR threshold (data not shown).

Finally we also detected a novel association between mental retardation and the *NDFIP1* gene (**Figure 2C**, **Table 3**). Of note, no effect was detected with Autism at this locus. Indeed, the

only other effect observed in this region was related to Cerebral Palsy ($p = 9.00 \times 10^{-4}$) and, as mentioned above, an independent effect exists for Mental Retardation. The PheWAS code for mental retardation includes ICD-9 codes for mild, moderate and profound degrees of retardation as well as not-otherwise-specified (MR-NOS). Indeed, an additive correlation can also be detected when we score these subgroups according to severity excluding the MR-NOS subgroup ($p = 3.00 \times 10^{-4}$). Larger sample size is necessary to fully elucidate this interesting effect. The Nedd4 family-interacting protein 1 (Ndfip1) is an adaptor protein for the Nedd4 family of E3 ubiquitin ligases important for axon and dendrite development. In fact, cerebral atrophy is one of the main findings in Ndfip1 KO mice (Hammond et al., 2014). Another neurodevelopmental association effect was observed in the vicinity of the Phospholipase C-Like 1 (PLCL1, PRIP-1) gene for overall Developmental Delays-Speech and Language Disorder (**Table 3**, **Figure 2D**). This gene which is expressed predominantly in brain, regulates the turnover of GABA-receptors, contributes to the maintenance of GABA-mediated synaptic inhibition, and has been implicated in several pathologies in animal models and human including epilepsy, bone density and cancer (Liu et al., 2008; Zhu et al., 2012). Finally, we also detected a link between Neuroxin-3 and early onset depression in this study (**Table 3**). In fact, this gene has a major role in synaptic plasticity and function in the nervous system as a receptor and cell adhesion molecule.

In summary, by using the PheWAS approach and re-mapping the ICD-9 codes on our European ancestry pediatric cohorts we have been able to verify and confirm a variety of previously reported associations as well as discover new effects that potentially have clinical implications. Similar to adult PheWAS studies, our data also support the importance of this approach in pediatrics. We replicated known phenotype-genotype associations in a pediatric population using these EMR-linked eMERGE databases, and also noted a number of new possible associations that warrant additional study, especially including the relationship of *PLCL1* to speech and language development and *IL5-IL13* to EoE. Some of the limitations to the current PheWAS map include the fact that current map does not take into account of the correlation between some phenotypes and treat them as independent. Future pediatric PheWAS directions will include enhancements of a PheWAS map for more precise modeling of trait associations as well as improvements for richer querying and filtering.

ACKNOWLEDGMENTS

We are grateful to the individuals who participated in this study. We thank the genotyping core facilities in both academic centers (CCHMC, BCH) and our colleagues who facilitated the genotyping and recruitment of subjects.

This work was supported by a grant from the National Human Genomic Research Institute: 1U01HG006828 with other NIH support (R37 AI024717, P01 AI083194, U19 AI066738, and P01 AR049084), the US Department of Veterans Affairs, the Campaign Urging Research For Eosinophilic Diseases (CURED) Foundation, as well as the Food Allergy Research Education (FARE) Foundation.

REFERENCES

Anney, R., Klei, L., Pinto, D., Regan, R., Conroy, J., Magalhaes, T. R., Correia, C., et al. (2010). A genome-wide scan for common alleles affecting risk for autism. *Hum. Mol. Genet*. 19, 4072–4082. doi: 10.1093/hmg/ddq307

Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., et al. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet*. 41, 703–707. doi: 10.1038/ng.381

Berndt, S. I., Gustafsson, S., Mägi, R., Ganna, A., Wheeler, E., Feitosa, M. F., et al. (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat. Genet*. 45, 501–512. doi: 10.1038/ng.2606

Bi, M., Kao, W. H., Boerwinkle, E., Hoogeveen, R. C., Rasmussen-Torvik, L. J., Astor, B. C., et al. (2010). Association of rs780094 in GCKR with metabolic traits and incident diabetes and cardiovascular disease: the ARIC Study. *PLoS ONE* 5:e11690. doi: 10.1371/journal.pone.0011690

Bottema, R. W., Reijmerink, N. E., Kerkhof, M., Koppelman, G. H., Stelma, F. F., Gerritsen, J., et al. (2008). Interleukin 13, CD14, pet and tobacco smoke influence atopy in three Dutch cohorts: the allergenic study. *Eur. Respir. J.* 32, 593–602. doi: 10.1183/09031936.00162407

Bouffi C. 1st., Rochman, M., Zust, C. B., Stucke, E. M., Kartashov, A., Fulkerson, P. C., et al. (2013). IL-33 markedly activates murine eosinophils by an NF-?B-dependent mechanism differentially dependent upon an IL-4-driven autoinflammatory loop. *J. Immunol*. 91, 4317–4325. doi: 10.4049/jimmunol.1301465

Carroll, R. J., Bastarache, L., and Denny, J. C. (2014). R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* 30, 2375–2376. doi: 10.1093/bioinformatics/btu197

Cook, E. H. Jr., Courchesne, R., Lord, C., Cox, N. J., Yan, S., Lincoln, A., et al. (1997). Evidence of linkage between the serotonin transporter and autistic disorder. *Mol. Psychiatry* 2, 247–250.

Cook, E. H. (1990). Autism: review of neurochemical investigation. *Synapse* 6, 292–308. doi: 10.1002/syn.890060309

Corren, J. (2012). Inhibition of interleukin-5 for the treatment of eosinophilic diseases. *Discov. Med*. 13, 305–312.

Dales, R., Chen, Y., Lin, M., and Karsh, J. (2005). The association between allergy and diabetes in the Canadian population: implications for the Th1-Th2 hypothesis. *Eur. J. Epidemiol*. 20, 713–717. doi: 10.1007/s10654-005-7920-1

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* 31, 1102–1110. doi: 10.1038/nbt.2749

Denny, J. C., Crawford, D. C., Ritchie, M. D., Bielinski, S. J., Basford, M. A., Bradford, Y., et al. (2011). Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am. J. Hum. Genet.* 89, 529–542. doi: 10.1016/j.ajhg.2011.09.008

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., et al. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 26, 1205–1210. doi: 10.1093/bioinformatics/btq126

Docampo, E., Ribasés, M., Gratacòs, M., Bruguera, E., Cabezas, C., Sánchez-Mora, C., et al. (2012). Association of Neurexin 3 polymorphisms with smoking behavior. *Genes Brain Behav*. 11, 704–711. doi: 10.1111/j.1601-183X.2012.00815.x

Esposito, F., Patsopoulos, N. A., Cepok, S., Kockum, I., Leppä, V., Booth, D. R., et al. (2010). IL12A, MPHOSPH9/CDK2AP1 and RGS1 are novel multiple sclerosis susceptibility loci. *Genes Immun.* 11, 397–405. doi: 10.1038/gene.2010.28

Feng, Y., Zheng, P., Zhao, H., and Wu, K. (2009). SLC22A4 and SLC22A5 gene polymorphisms and Crohn's disease in the Chinese Han population. *J. Dig. Dis.* 10, 181–187. doi: 10.1111/j.1751-2980.2009.00383.x

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* 12, 1118–1125. doi: 10.1038/ng.717

Gauderman, W. J., and Morrison, J. M. (2006). *QUANTO 1.1: A Computer Program for Power and Sample Size Calculations for Genetic-epidemiology Studies*. Available online at: http://hydra.usc.edu/gxe

Gidding, S. S. (1993). The rationale for lowering serum cholesterol levels in American children. *Am. J. Dis. Child*. 147, 386–392.

Granada, M., Wilk, J. B., Tuzova, M., Strachan, D. P., Weidinger, S., Albrecht, E., et al. (2012). A genome-wide association study of plasma total IgE concentrations in the Framingham Heart Study. *J. Allergy Clin. Immunol*. 129, 840–845.e21. doi: 10.1016/j.jaci.2011.09.029

Hammond, V. E., Gunnersen, J. M., Goh, C. P., Low, L. H., Hyakumura, T., Tang, M. M., et al. (2014). Ndfip1 is required for the development of pyramidal neuron dendrites and spines in the neocortex. *Cereb. Cortex* 24, 3289–3300. doi: 10.1093/cercor/bht191

Hinds, D. A., McMahon, G., Kiefer, A. K., Do, C. B., Eriksson, N., Evans, D. M., et al. (2013). A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. *Nat. Genet*. 45, 907–911. doi: 10.1038/ng.2686

Hinks, A., Cobb, J., Marion, M. C., Prahalad, S., Sudman, M., Bowes, J., et al. (2013). Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nat. Genet*. 45, 664–669. doi: 10.1038/ng.2614

Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3 (Bethesda)*. 1, 457–470. doi: 10.1534/g3.111.001198

Huertas-Vazquez, A., Plaisier, C., Weissglas-Volkov, D., Sinsheimer, J., Canizales-Quinteros, S., Cruz-Bautista, I., et al. (2008). TCF7L2 is associated with high serum triacylglycerol and differentially expressed in adipose tissue in families with familial combined hyperlipidaemia. *Diabetologia* 51, 62–69. doi: 10.1007/s00125-007-0850-6

Hunt, K. A., Zhernakova, A., Turner, G., Heap, G. A., Franke, L., Bruinenberg, M., et al. (2008). Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.* 40, 395–402. doi: 10.1038/ng.102

Jamain, S., Betancur, C., Quach, H., Philippe, A., Fellous, M., Giros, B., et al. (2002). Linkage and association of the glutamate receptor 6 gene with autism. *Mol. Psychiatry* 7, 302–310. doi: 10.1038/sj.mp.4000979

Klein, J. D., Sesselberg, T. S., Johnson, M. S., O'Connor, K. G., Cook, S., Coon, M., et al. (2010). Adoption of body mass index guidelines for screening and counseling in pediatric practice. *Pediatrics* 125, 265–272. doi: 10.1542/peds.2008-2985

Kottyan, L. C., Davis, B., Sherrill, J. D., Liu, K., Rochman, M., Kaufman, K., et al. (2014). Identification of genome-wide susceptibility loci for eosinophilic esophagitis elucidates tissue-specificity of this allergic disease. *Nat. Genet.* 46, 895–900. doi: 10.1038/ng.3033

Lee, H. S., Kang, J., Yang, S., Kim, D., and Park, Y. (2011). Susceptibility influence of a PTPN22 haplotype with thyroid autoimmunity in Koreans. *Diabetes Metab. Res. Rev*. 27, 878–882. doi: 10.1002/dmrr.1265

Li, Y., Mao, Q., Shen, L., Tian, Y., Yu, C., Zhu, W. M., et al. (2010). Interleukin-23 receptor genetic polymorphisms and Crohn's disease susceptibility: a meta-analysis. *Inflamm. Res*. 59, 607–614. doi: 10.1007/s00011-010-0171-y

Liao, K. P., Diogo, D., Cui, J., Cai, T., Okada, Y., Gainer, V. S., et al. (2014). Association between low density lipoprotein and rheumatoid arthritis genetic factors with low density lipoprotein levels in rheumatoid arthritis and non-rheumatoid arthritis controls. *Ann. Rheum. Dis*. 73, 1170–1175. doi: 10.1136/annrheumdis-2012-203202

Liu, J., Divoux, A., Sun, J., Zhang, J., Clément, K., Glickman, J. N., Sukhova, G. K., et al. (2009). Genetic deficiency and pharmacological stabilization of mast cells reduce diet-induced obesity and diabetes in mice. *Nat. Med.* 15, 940–945. doi: 10.1038/nm.1994

Liu, Y. Z., Wilson, S. G., Wang, L., Liu, X. G., Guo, Y. F., Li, J., et al. (2008). Identification of PLCL1 gene for hip bone size variation in females in a genome-wide association study. *PLoS ONE* 3:e3160. doi: 10.1371/journal.pone.0003160

Lyssenko, V., Lupi, R., Marchetti, P., Del Guerra, S., Orho-Melander, M., Almgren, P., et al. (2007). Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. *J. Clin. Invest.* 117, 2155–2163. doi: 10.1172/JCI30706

Mells, G. F., Floyd, J. A., Morley, K. I., Cordell, H. J., Franklin, C. S., Shin, S. Y., et al. (2011). Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nat. Genet.* 43, 329–332. doi: 10.1038/ng.789

Namjou, B., Keddache, M., Marsolo, K., Wagner, M., Lingren, T., Cobb, B., et al. (2013). EMR-linked GWAS study: investigation of variation landscape of loci for body mass index in children. *Front. Genet*. 4:268. doi: 10.3389/fgene.2013.00268

Namjou, B., Sestak, A. L., Armstrong, D. L., Zidovetzki, R., Kelly, J. A., Jacob, N., et al. (2009). High-density genotyping of STAT4 reveals multiple haplotypic associations with systemic lupus erythematosus in different racial groups. *Arthritis Rheum*. 60, 1085–1095. doi: 10.1002/art.24387

Neuraz, A., Chouchana, L., Malamut, G., Le Beller, C., Roche, D., Beaune, P., et al. (2013). Phenome-wide association studies on a quantitative trait: application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics. *PLoS Comput. Biol*. 9:e1003405. doi: 10.1371/journal.pcbi.1003405

Onuma, H., Tabara, Y., Kawamoto, R., Shimizu, I., Kawamura, R., Takata, Y., et al. (2010). The GCKR rs780094 polymorphism is associated with susceptibility of type 2 diabetes, reduced fasting plasma glucose levels, increased triglycerides levels and lower HOMA-IR in Japanese population. *J. Hum. Genet*. 55, 600–604. doi: 10.1038/jhg.2010.75

Paternoster, L., Standl, M., Chen, C. M., Ramasamy, A., Bønnelykke, K., Duijts, L., et al. (2011). Meta-analysis of genome-wide association studies identifies three new risk loci for atopic dermatitis. *Nat. Genet*. 44, 187–192. doi: 10.1038/ng.1017

Pendergrass, S. A., Brown-Gentry, K., Dudek, S., Frase, A., Torstenson, E. S., Goodloe, R., et al. (2013). Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet.* 9:e1003087. doi: 10.1371/journal.pgen.1003087

Pendergrass, S. A., Brown-Gentry, K., Dudek, S. M., Torstenson, E. S., Ambite, J. L., Avery, C. L., et al. (2011). The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet Epidemiol.* 35, 410–422. doi: 10.1002/gepi.20589

Plenge, R. M., Seielstad, M., Padyukov, L., Lee, A. T., Remmers, E. F., Ding, B., et al. (2007). TRAF1-C5 as a risk locus for rheumatoid arthritis—a genomewide study. *N. Engl. J. Med.* 357, 1199–1209. doi: 10.1056/NEJMoa073491

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. doi: 10.1038/ng1847

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., et al. (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26, 2336–2337. doi: 10.1093/bioinformatics/btq419

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*. 81, 559–575. doi: 10.1086/519795

Pykäläinen, M., Kinos, R., Valkonen, S., Rydman, P., Kilpeläinen, M., Laitinen, L. A., et al. (2005). Association analysis of common variants of STAT6, GATA3, and STAT4 to asthma and high serum IgE phenotypes. *J. Allergy Clin. Immunol.* 115, 80–87. doi: 10.1016/j.jaci.2004.10.006

Ritchie, M. D., Denny, J. C., Zuvich, R. L., Crawford, D. C., Schildcrout, J. S., Bastarache, L., et al. (2013). Genome-and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* 127, 1377–1385. doi: 10.1161/CIRCULATIONAHA.112.000604

Rothenberg, M. E., Spergel, J. M., Sherrill, J. D., Annaiah, K., Martin, L. J., Cianferoni, A., et al. (2010). Common variants at 5q22 associate with pediatric eosinophilic esophagitis. *Nat. Genet*. 42, 289–291. doi: 10.1038/ng.547

Setia, S., Andrade, M., Tromp, G., Kuivaniemi, H., Pugh, E., Namjou, B., et al. (2014). Imputation and quality control steps for combining multiple genome-wide datasets. *Front. Genet.* 5:370. doi: 10.3389/fgene.2014.00370

Stoltenberg, S. F., Lehmann, M. K., Christ, C. C., Hersrud, S. L., and Davies, G. E. (2011). Associations among types of impulsivity, substance use problems and neurexin-3 polymorphisms. *Drug Alcohol Depend*. 119, e31–e38. doi: 10.1016/j.drugalcdep.2011.05.025

St Pourcain, B., Skuse, D. H., Mandy, W. P., Wang, K., Hakonarson, H., Timpson, N. J., et al. (2014). Variability in the common genetic architecture of social-communication spectrum phenotypes during childhood and adolescence. *Mol. Autism* 5:18. doi: 10.1186/2040-2392-5-18

Suhre, K., Shin, S. Y., Petersen, A. K., Mohney, R. P., Meredith, D., Wägele, B., et al. (2011). Human metabolic individuality in biomedical and pharmaceutical research. *Nature* 477, 54–60. doi: 10.1038/nature10354

Thomas, L. A., Akins, M. R., and Biederer, T. (2008). Expression and adhesion profiles of SynCAM molecules indicate distinct neuronal functions. *J. Comp. Neurol.* 510, 47–67. doi: 10.1002/cne.21773

Thompson, S. D., Marion, M. C., Sudman, M., Ryan, M., Tsoras, M., Howard, T. D., et al. (2012). Genome-wide association analysis of juvenile idiopathic arthritis identifies a new susceptibility locus at chromosomal region 3q13. *Arthritis Rheum*. 64, 2781–2791. doi: 10.1002/art.34429

Tiisala, R., and Kantero, R. L. (1971). Studies on growth of Finnish children from birth to 10 years. 3. Comparison of height and weight distance curves based

on longitudinal and cross-sectional series from birth to 10 years. *Acta Paediatr Scand. Suppl.* 220, 13–7.

Todd, J. A., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V., et al. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.* 39, 857–864. doi: 10.1038/ng2068

Torgerson, D. G., Ampleford, E. J., Chiu, G. Y., Gauderman, W. J., Gignoux, C. R., Graves, P. E., et al. (2011). Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat. Genet.* 43, 887–892. doi: 10.1038/ng.888

Tulah, A. S., Holloway, J. W., and Sayers, I. (2013). Defining the contribution of SNPs identified in asthma GWAS to clinical variables in asthmatic children. *BMC Med. Genet.* 14:100. doi: 10.1186/1471-2350-14-100

Vaags, A. K., Lionel, A. C., Sato, D., Goodenberger, M., Stein, Q. P., Curran, S., et al. (2012). Rare deletions at the neurexin 3 locus in autism spectrum disorder. *Am. J. Hum. Genet.* 90, 133–141. doi: 10.1016/j.ajhg.2011.11.025

van der Pouw Kraan, T. C., van Veen, A., Boeije, L. C., van Tuyl, S. A., de Groot, E. R., Stapel, S. O., et al. (1999). An IL-13 promoter polymorphism associated with increased risk of allergic asthma. *Genes Immun.* 1, 61–65. doi: 10.1038/sj.gene.6363630

Verlaan, D. J., Berlivet, S., Hunninghake, G. M., Madore, A. M., Larivière, M., Moussette, S., et al. (2009). Allele-specific chromatin remodeling in the ZPBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. *Am. J. Hum. Genet.* 85, 377–393. doi: 10.1016/j.ajhg.2009.08.007

Weiss, L. A., Kosova, G., Delahanty, R. J., Jiang, L., Cook, E. H., Ober, C., et al. (2006). Variation in ITGB3 is associated with whole-blood serotonin level and autism susceptibility. *Eur. J. Hum. Genet.* 14:923–931. doi: 10.1038/sj.ejhg.5201644

Weiss, L. A., Veenstra-Vanderweele, J., Newman, D. L., et al. (2004). Genomewide association study identifies ITGB3 as a QTL for whole blood serotonin. *Eur. J. Hum. Genet.* 12, 949–954. doi: 10.1038/sj.ejhg.5201239

Zhu, G., Yoshida, S., Migita, K., Yamada, J., Mori, F., Tomiyama, M., et al. (2012). Dysfunction of extrasynaptic GABAergic transmission in phospholipase C-related, but catalytically inactive protein 1 knockout mice is associated with an epilepsy phenotype. *J. Pharmacol. Exp. Ther.* 340, 520–528. doi: 10.1124/jpet.111.182386

Conflict of Interest Statement: The Guest Associate Editor Mariza De Andrade declares that, despite having collaborated with authors Bahram Namjou, Joshua C. Denny, Leah C. Kottyan, Marylyn D. Ritchie, and Shefali S. Verma, the review process was handled objectively and no conflict of interest exists. The Review Editor Andrew Skol declares that, despite having collaborated with author John B. Harley, the review process was handled objectively and no conflict of interest exists. Marc E. Rothenberg is a consultant for Immune Pharmaceuticals and has an equity interest. Marc E. Rothenberg has a royalty interest in reslizumab being developed by Teva Pharmaceuticals. Marc E. Rothenberg, John B. Harley, and Leah C. Kottyan are co-inventors of a patent application, being submitted by CCHMC, concerning the genetics of EoE. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 May 2014; accepted: 31 October 2014; published online: 18 November 2014.

Citation: Namjou B, Marsolo K, Caroll RJ, Denny JC, Ritchie MD, Verma SS, Lingren T, Porollo A, Cobb BL, Perry C, Kottyan LC, Rothenberg ME, Thompson SD, Holm IA, Kohane IS and Harley JB (2014) Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis. Front. Genet. ***5****:401. doi: 10.3389/fgene.2014.00401*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 05 August 2014
doi: 10.3389/fgene.2014.00250

Phenome-wide association studies demonstrating pleiotropy of genetic variants within *FTO* with and without adjustment for body mass index

Robert M. Cronin[1,2]*, Julie R. Field[3], Yuki Bradford[4], Christian M. Shaffer[4], Robert J. Carroll[1], Jonathan D. Mosley[1,5], Lisa Bastarache[2], Todd L. Edwards[6], Scott J. Hebbring[7], Simon Lin[8], Lucia A. Hindorff[9], Paul K. Crane[10], Sarah A. Pendergrass[11], Marylyn D. Ritchie[11], Dana C. Crawford[4], Jyotishman Pathak[12], Suzette J. Bielinski[13], David S. Carrell[14], David R. Crosslin[15], David H. Ledbetter[16], David J. Carey[17], Gerard Tromp[17], Marc S. Williams[16], Eric B. Larson[14], Gail P. Jarvik[10,15], Peggy L. Peissig[8], Murray H. Brilliant[7], Catherine A. McCarty[18], Christopher G. Chute[12], Iftikhar J. Kullo[19], Erwin Bottinger[20], Rex Chisholm[21], Maureen E. Smith[21], Dan M. Roden[1,5] and Joshua C. Denny[1,2]*

[1] *Department of Medicine, Vanderbilt University, Nashville, TN, USA*
[2] *Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA*
[3] *Office of Research, Vanderbilt University, Nashville, TN, USA*
[4] *Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University, Nashville, TN, USA*
[5] *Department of Pharmacology, Vanderbilt University, Nashville, TN, USA*
[6] *Vanderbilt Epidemiology Center, Vanderbilt University, Nashville, TN, USA*
[7] *Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, WI, USA*
[8] *Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI, USA*
[9] *Division of Genomic Medicine, National Human Genome Research Institute, Bethesda, MD, USA*
[10] *Department of Medicine, University of Washington, Seattle, WA, USA*
[11] *Department of Biochemistry and Molecular Biology, Center for Systems Genomics, The Pennsylvania State University, University Park, PA, USA*
[12] *Divisions of Biomedical Informatics and Statistics, Mayo Clinic, Rochester, MN, USA*
[13] *Division of Epidemiology, Mayo Clinic, Rochester, MN, USA*
[14] *Group Health Research Institute, Seattle, WA, USA*
[15] *Department of Genome Sciences, University of Washington, Seattle, WA, USA*
[16] *Genomic Medicine Institute, Geisinger Health System, Danville, PA, USA*
[17] *Weis Center for Research, Geisinger Health System, Danville, PA, USA*
[18] *Essentia Institute of Rural Health, Duluth, MN, USA*
[19] *Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN, USA*
[20] *The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA*
[21] *Department of Cell and Molecular Biology, Feinberg School of Medicine, Northwestern University, Evanston, IL, USA*

Edited by:
Helena Kuivaniemi, Geisinger Health System, USA

Reviewed by:
Qing Lu, Michigan State University, USA
Sarah Buxbaum, Jackson State University, USA

****Correspondence:***
Robert M. Cronin, Department of Biomedical Informatics, Vanderbilt University Medical Center, 220 Garland 440 EBL, Nashville, TN 37232, USA
e-mail: robert.cronin@vanderbilt.edu;
Joshua C. Denny, Department of Biomedical Informatics and Department of Medicine, Vanderbilt University Medical Center, 2525 West End Avenue, Suite 600, Nashville, TN 37203-8820, USA
e-mail: josh.denny@vanderbilt.edu

Phenome-wide association studies (PheWAS) have demonstrated utility in validating genetic associations derived from traditional genetic studies as well as identifying novel genetic associations. Here we used an electronic health record (EHR)-based PheWAS to explore pleiotropy of genetic variants in the fat mass and obesity associated gene (*FTO*), some of which have been previously associated with obesity and type 2 diabetes (T2D). We used a population of 10,487 individuals of European ancestry with genome-wide genotyping from the Electronic Medical Records and Genomics (eMERGE) Network and another population of 13,711 individuals of European ancestry from the BioVU DNA biobank at Vanderbilt genotyped using Illumina HumanExome BeadChip. A meta-analysis of the two study populations replicated the well-described associations between *FTO* variants and obesity (odds ratio [OR] = 1.25, 95% Confidence Interval = 1.11–1.24, $p = 2.10 \times 10^{-9}$) and *FTO* variants and T2D ($OR = 1.14$, 95% $CI = 1.08$–1.21, $p = 2.34 \times 10^{-6}$). The meta-analysis also demonstrated that *FTO* variant rs8050136 was significantly associated with sleep apnea ($OR = 1.14$, 95% $CI = 1.07$–1.22, $p = 3.33 \times 10^{-5}$); however, the association was attenuated after adjustment for body mass index (BMI). Novel phenotype associations with obesity-associated *FTO* variants included fibrocystic breast disease (rs9941349, $OR = 0.81$, 95% $CI = 0.74$–0.91, $p = 5.41 \times 10^{-5}$) and trends toward associations with non-alcoholic liver disease and gram-positive bacterial infections. *FTO* variants not associated with obesity demonstrated other potential disease associations including non-inflammatory disorders of the cervix and chronic periodontitis. These results suggest that genetic variants in *FTO* may have pleiotropic associations, some of which are not mediated by obesity.

Keywords: PheWAS, genetic association, pleiotropy, Exome chip, *FTO*, BMI

INTRODUCTION

Pleiotropy, the phenomenon in which a single gene or genetic variant is associated with multiple phenotypes, is essential to the functionality of the human genome (Crespi, 2010; Wagner and Zhang, 2011; Pavlicev and Wagner, 2012). Through comparing multiple genome-wide association studies (GWAS) and candidate gene studies, pleiotropy has been noted in many single nucleotide polymorphisms (SNPs) and genes, potentially providing greater insight into putative biological mechanisms (Sivakumaran et al., 2011; Stranger et al., 2011; Solovieff et al., 2013). The increasing prevalence of DNA biobanks linked to rich phenotype resources and large epidemiological databases have enabled the development of phenome-wide association study (PheWAS) method as an additional tool to investigate pleiotropy (Denny et al., 2010a; Pendergrass et al., 2011). As a complement to GWAS, PheWAS enables both the validation of genotype-phenotype associations identified through traditional GWAS and the generation of new hypotheses, identifying potentially novel associations as well as putative instances of genetic pleiotropy (Denny et al., 2011; Pendergrass et al., 2013). A recent application of PheWAS to 3144 GWAS-identified variants, replicated 210 known associations and noted 63 new, pleiotropic associations (Denny et al., 2013).

The Electronic Medical Records and Genomics (eMERGE) Network was formed in 2007 to use phenotypes derived from electronic health record (EHR) data to perform GWAS and other genomic investigations (Kullo et al., 2011; Pathak et al., 2011; Crosslin et al., 2013; Ding et al., 2013). eMERGE investigators have also used EHR-based PheWAS methods to evaluate multiple phenotypes associated with specific genetic variants (Denny et al., 2010a; Pathak et al., 2012; Hebbring et al., 2013). PheWAS has been used to enhance our understanding of the genetic determinants of complex traits discovered through GWAS. For example, a PheWAS of variants associated with longer cardiac conduction (Ritchie et al., 2013) demonstrated an association with atrial fibrillation, and a PheWAS of variants affecting platelet count and size identified associations with autoimmune diseases (Shameer et al., 2013).

Variants in the fat mass and obesity associated gene (*FTO)* have been studied since 2007, when it was discovered that some were associated with body mass index (BMI) and obesity (Frayling et al., 2007). Multiple GWAS have demonstrated further associations between variants in *FTO* and obesity (Jacobsson et al., 2012). Some of these variants have also been noted to be associated with both obesity and type 2 diabetes (T2D) (Hertel et al., 2011; Rees et al., 2011; Li et al., 2012) including SNPs rs9939609 and rs8050136, which are in high linkage disequilibrium (LD) with each other in people of European ancestry ($r^2 = 1.00$; using 1000 Genomes Pilot 1 reference in the CEU population). The SNP rs8050136 is located in an intronic region where the transcription factor cut-like homeobox (*CUTL1*) protein (Li et al., 2000) is predicted to bind (Stratigopoulos et al., 2008). This variant has been associated with T2D and obesity in Han Chinese and European populations (Hubacek et al., 2008; Liu et al., 2010; Hotta et al., 2011) but other studies found no association between this variant and T2D or obesity in the Chinese Han population (Li et al., 2008; Xi and Mi, 2009). These differences in associations of SNPs with phenotypes have been further analyzed through fine mapping of BMI loci (Gong et al., 2013). This study reported that GWAS studies primarily performed in European populations of numerous loci associated with BMI are not generalizable to other ethnic groups, for example African Americans. Another study demonstrated that rs8050136 was associated with increased energy intake from fat with similar total energy intake (Park et al., 2013). A more recent study noted that the mechanism of action for common variants in *FTO* may be through regulation of *IRX3* expression, which is highly expressed in the brain (Smemo et al., 2014).

There is also evidence of other putative disease associations with *FTO* variants that have not achieved genome-wide significance, such as pancreatic cancer, Alzheimer's disease, attention deficit hyperactivity disorder, alcoholism, and osteoarthritis (Keller et al., 2011; Lurie et al., 2011; Sobczyk-Kopciol et al., 2011; arcOGEN Consortium et al., 2012; Corella et al., 2012; Reitz et al., 2012; Velders et al., 2012). These varied disease-SNP associations suggest that SNPs in *FTO* may have pleiotropic effects. Utilizing the population and diagnostic diversity contained within the real-world clinical environment for variants within *FTO*, our goal was to determine whether an EHR-based PheWAS could identify genetic pleiotropy that might otherwise remain undetected in traditional cohort study designs. In the present study, we utilized PheWAS method and data sets from the eMERGE network (McCarty et al., 2011; Gottesman et al., 2013) to evaluate pleiotropy of variants in *FTO*.

MATERIALS AND METHODS

PARTICIPATION OF eMERGE SITES

The eMERGE Network data used in this study consists of seven institutions (Group Health Cooperative and University of Washington, Marshfield Clinic, Mayo Clinic, Northwestern University, Mount Sinai, Geisinger Health System, and Vanderbilt University Medical Center), each with DNA biorepositories linked to their EHRs. Each site pulled demographic, vital sign, and billing data from their EHR research data repositories for this study. All projects were either approved by local IRBs or classified as IRB exempt as non-human subjects research.

GENOTYPING OF eMERGE SUBJECTS

Variants for eMERGE subjects were selected from extant genome-wide genotypes with either the Human660W-Quadv1_A or Illumina OmniExpress chips. The Human660W-Quadv1_A BeadChip was completed at the Center for Genotyping and Analysis at the Broad Institute, and the Center for Inherited Disease Research at Johns Hopkins University. Genotyping for Illumina OmniExpress BeadChips was performed at the University of Pittsburgh Genomics and Proteomics Core Laboratories. These genotyping data comprised 10,487 individuals of European ancestry, as designated in the EHRs.

Quality-control (QC) of the genotype data was performed using a pipeline developed by the eMERGE Genomics Working Group (Turner et al., 2011). This process included call rate restrictions listed below, identification of sex mismatch and anomalies, checking duplicate and HapMap concordance, as well as identifying batch effects, sample relatedness, and minor allele frequency (MAF). Population stratification was evaluated using

STRUCTURE (Pritchard et al., 2000) and EIGENSTRAT (Price et al., 2006). Only SNPs with call rates >99% and MAF >0.01 in unrelated samples were included for further study. Relatedness was determined on the basis of identity by descent (IBD) estimates generated from the genome-wide genotype data in PLINK (Purcell et al., 2007). All study sites had pairs of individuals with an IBD estimate greater than 0.25; only one of the individuals in each related pair was randomly selected and used in the analysis. Additional genotypes were imputed using IMPUTE2 (Marchini et al., 2007) and 1000 Genomes Project as the reference (1000 Genomes Project Consortium et al., 2010). We used imputed SNPs with a minimum info score of 0.7 and called genotypes based on the maximum posterior probability.

We selected 54 SNPs, of which 51 were imputed in at least one site, located in *FTO* that met the QC criteria above and were previously associated with obesity (Jacobsson et al., 2012). QC and subsequent association tests were performed using PLINK (Purcell et al., 2007) and the R statistical package (R Core Team, 2013).

GENOTYPING OF VANDERBILT SUBJECTS USING HumanExome BeadChips

We selected 13,711 individuals of European ancestry from the BioVU DNA databank with BMI data who were genotyped using the Illumina Infinium HumanExome BeadChip, which includes >240,000 markers, mostly within exonic regions, as well as SNPs from the GWAS catalog (Welter et al., 2014) including rs8050136 in *FTO*. Genotyping was performed at the Vanderbilt Technologies for Advanced Genomics (VANTAGE) Core, and genomic data were processed by the Vanderbilt Technologies for Advanced Genomics Analysis and Research Design (VANGARD) Core. Clustering was performed using GenomeStudio's GenTrain clustering algorithm followed by manual review and reclustering; genotype calling was performed using GenomeStudio's GenCall algorithm. Genotyping quality was evaluated using SNP call rates and concordance rates with HapMap controls; SNPs with <99.8% call rate or <98% concordance were excluded. In the first analysis, we focused on rs8050136, which had a call rate of >99.9%. In the subsequent analyses, we further analyzed eight *FTO* SNPs on the Exome chip, which had call rates greater than 99.8% and MAFs >0.01. Similar to the eMERGE set, for individuals with an IBD estimate greater than 0.25; only one of the individuals in each related group was selected randomly and used in our analyses.

PheWAS ANALYSES

We first tested the 54 eMERGE SNPs for association with BMI using linear regression. We calculated LD with our reference SNP rs8050136, chosen as the reference because of its GWAS associations with BMI and T2D in the literature and since it was directly genotyped on all of the platforms. To evaluate phenotype associations and potential pleiotropy among different FTO SNPs, we grouped SNPs into three groups for convenience based on their LD with rs805136: high LD ($r^2 > 0.80$), moderate LD ($0.80 \geq r^2 > 0.60$) and low LD ($r^2 \leq 0.60$) with rs8050136. Our hypothesis was that SNPs in high LD would show similar patterns of phenotype associations with rs8050136, and that different patterns may be observed in SNPs with lower LD.

Analyses for the eMERGE and the BioVU datasets were conducted separately and then meta-analyzed. The eMERGE population had 54 SNPs and the BioVU population had nine SNPs for analysis, which were also present in the eMERGE dataset. We used logistic regression adjusted for age, sex, eMERGE site, and the first three principal components as calculated for each dataset by EIGENSTRAT, using an additive genetic model. We performed PheWAS using each SNP using methods and code groupings described previously (Denny et al., 2013) using the R PheWAS package (Carroll et al., 2014), briefly, calculating comprehensive associations between SNPs and a total of 1645 clinical phenotypes derived from the International Classification of Disease, 9th CM (ICD-9) edition codes from each site's EHR. The ICD-9 codes that are associated with each phenotype can be found at the PheWAS catalog located online at http://phewas.mc.vanderbilt.edu/. Cases for a given disease were defined as having at least two relevant ICD-9 codes on different days. The PheWAS method also defines control groups for each disease, which ensures that related diseases do not serve as controls for the current disease being analyzed. We performed association testing for all PheWAS phenotypes occurring in at least 20 individuals (effectively 20 "cases").

We then compared our results to performing PheWAS for each *FTO* SNP adjusting for BMI. The BMI, obtained from each site's EHR, was estimated using the average BMI from individuals within our dataset. To minimize erroneous data, we only used BMI measurements between 15 and 70, a range that we have used in prior studies and has good precision (Denny et al., 2010b). Plotting was performed in R using the PheWAS and ggplot2 packages.

Meta-analysis was performed using the inverse-variance method (Hunter et al., 1982) for the nine shared SNPs. There were 1010 phenotypes that were in common across both datasets and met our minimum case criteria of at least 20 cases. This yields a Bonferroni corrected p-value of 4.95×10^{-5}, ($p = 0.05/1010 = 4.95 \times 10^{-5}$), for a single SNP. We chose a single SNP, phenome-wide correction threshold since most of the SNPs in this analysis were in high LD with each other and thus do not represent truly independent tests. A false discovery rate (FDR) of $q = 0.05$, calculated with the Benjamin and Hochberg method using the R p.adjust method, yields a p-value of 2.48×10^{-4} (Benjamini and Hochberg, 1995). For our latter analyses, we considered a total of 54 SNPs. Since many phenotypes are correlated with each other and many of the SNPs are in LD, we also used simpleM (Gao et al., 2010) to estimate the number of unique tests performed, leading to an adjustment of $p = 2.36 \times 10^{-6}$. All analyses assumed a two-tailed distribution.

RESULTS

OVERVIEW

A total of 24,198 individuals were used in our analyses (**Table 1**). Both the eMERGE and BioVU datasets were similar in median age, sex, and BMI. Our analysis of the association of the *FTO* SNPs with BMI (**Table 2**) showed that most SNPs in high linkage disequilibrium with rs8050136 ($r^2 > 0.80$) have highly significant p-values ($< 3 \times 10^{-9}$) and betas for BMI (**Table 2**). SNPs with

Table 1 | Characteristics of the study sets.

	eMERGE ***n* = 10,487**	**BioVU** ***n* = 13,711**	**Combined** ***n* = 24,198**
Genotyping Platform	Illumina Human660W-Quadv1_A	Illumina HumanExome	
Number of SNPs	54	9	9
Total number of phenotypes	1094	1254	1010
Median age (IQR)	58 (48–68)	60 (47–72)	59 (48–70)
Female (%)	52.24	54.31	53.35
BMI (average ± *SD*)	30.86 ± 7.48	28.43 ± 6.44	29.54 ± 7.04
Most frequent diagnoses	Hypertension (66%) Hyperlipidemia (61%) Pain in limb (47%) Malaise and fatigue (39%) Abdominal/pelvic symptoms (36%)	Hypertension (63%) Malaise and fatigue (51%) Eye infection, viral (50%) Hyperlipidemia (40%) Pain in limb (39%)	Hypertension (64%) Hyperlipidemia (49%) Malaise and fatigue (46%) Pain in limb (43%) GERD (34%)

This table shows the main characteristics of the study populations of European ancestry, including age, sex, BMI and the five most significant PheWAS phenotypes observed in the datasets. The sample size included 10,487 from the eMERGE population and 13,711 from the BioVU population for a total of 24,198 people. For a given phenotype, in the combined dataset our maximum number of cases was 14,592 in hypertension and the minimum number of cases was 44.

lower correlations with rs8050136 have highly variable associations with BMI.

PheWAS OF *FTO* rs8050136 UNADJUSTED FOR BMI

In the BioVU population, we observed that obesity ($OR = 1.22$, $p = 1.4 \times 10^{-6}$, 95% $CI = 1.13$–1.33) was significantly associated with rs8050136. Three obesity-related diseases also trended toward significance; T2D ($OR = 1.14$, $p = 5.3 \times 10^{-5}$, 95% $CI = 1.07$–1.21), obstructive sleep apnea (OSA; $OR = 1.15$, $p = 4.6 \times 10^{-3}$, 95% $CI = 1.04$–1.26) and chronic non-alcoholic liver disease (NAFLD; $OR = 1.20$, $p = 6.06 \times 10^{-3}$, 95% $CI = 1.05$–1.38) (Supplementary Table 1). We observed similar odds ratios for obesity and T2D in eMERGE (obesity: $OR = 1.37$, $p = 1.88 \times 10^{-4}$, 95% $CI = 1.16$–1.61; T2D: $OR = 1.16$, $p = 0.014$, 95% $CI = 1.03$–1.32). eMERGE results also demonstrated similar trends toward significant associations with OSA ($OR = 1.14$, $p = 2.4 \times 10^{-3}$, 95% $CI = 1.05$–1.24) (Supplementary Table 1). After meta-analysis, obesity ($OR = 1.25$, $p = 2.1 \times 10^{-9}$, 95% $CI = 1.16$–1.35), morbid obesity ($OR = 1.34$, $p = 1.07 \times 10^{-7}$, 95% $CI = 1.20$–1.48), and two obesity-related diseases, T2D ($OR = 1.14$, $p = 2.3 \times 10^{-6}$, 95% $CI = 1.08$–1.21) and OSA ($OR = 1.15$, $p = 3.3 \times 10^{-5}$, 95% $CI = 1.07$–1.22), were associated with rs8050136 (**Table 3**). Additionally, the associations with NAFLD and fibrocystic breast disease were also $q < 0.05$.

PheWAS OF *FTO* rs8050136 ADJUSTED FOR BMI

After adjusting for average BMI, some of the associations were greatly attenuated, while others remained relatively unchanged (**Table 3**, **Figure 1**). The associations with obesity and OSA were largely attenuated by adjustment for BMI (obesity: $OR = 1.11$, $p = 0.017$, 95% $CI = 1.02$–1.22; morbid obesity: $OR = 1.17$, $p = 0.016$, 95% $CI = 1.03$–1.33; OSA: $OR = 1.07$, $p = 0.040$, 95% $CI = 1.00$–1.15). Chronic non-alcoholic liver disease demonstrated a possible association with rs8050136, which was only slightly attenuated between unadjusted and BMI-adjusted analyses (*OR*: 1.23 vs. 1.19; *p*: 2.2×10^{-4} vs. 1.9×10^{-3}, 95% $CI = 1.10$–1.37 vs. 1.07–1.33). Additional phenotypes trended toward association with rs8050136, including fibrocystic breast disease ($OR = 0.84$, $p = 4.8 \times 10^{-4}$, 95% $CI = 0.75$–0.92), staphylococcal infections ($OR = 1.16$, $p = 5.8 \times 10^{-3}$, 95% $CI = 1.04$–1.29), streptococcal infections ($OR = 1.21$,$p = 6.6 \times 10^{-3}$, 95% $CI = 1.05$–1.39), osteomyelitis ($OR = 1.21$, $p = 0.011$, 95% $CI = 1.04$–1.41), and joint effusions ($OR = 1.22$, $p = 6.9 \times 10^{-3}$, 95% $CI = 1.06$–1.41). These were not notably changed by BMI adjustment. Due to the number of gram-positive bacterial infections, we tested *post hoc* for the association between the SNP and a composite phenotype of all gram-positive infections, which were defined as staphylococcal infections, streptococcal infections, pneumococcal pneumonia, and gram positive septicemia. When combining all gram-positive phenotypes, the result was similar to the individual phenotypes ($n = 1095$, $OR = 1.15$ 95% confidence interval [95% CI] = 1.06–1.26).

PheWAS OF OTHER *FTO* SNPs ASSOCIATED WITH OBESITY

The results of SNPs in high LD with rs8050136 ($r^2 > 0.8$) showed a similar pattern of phenotypes to rs8050136 (**Figures 2A,B**). Rs9941349, which is in LD with rs8050136 ($r^2 = 0.92$) trended toward association with cystic mastopathy prior to BMI adjustment ($p = 5.4 \times 10^{-5}$, $OR = 0.81$, 95% $CI = 0.73$–0.90). SNPs with moderate to low correlation with rs8050136 had much different patterns of associations. Some of these SNPs demonstrated associations with obesity (e.g., rs9939609, rs9941349), and some did not (e.g., rs6499640, rs7199182; see **Table 2**). Of these SNPs, we only had eMERGE and BioVU data for rs6499640 (**Figure 3A**). All other SNPs were only available in the eMERGE data. "Non-inflammatory disorders of the cervix" was associated with some *FTO* SNPs (rs16952520: $n = 21$, $p = 1.92 \times 10^{-6}$, $OR = 6.76$, 95% $CI = 3.08$–14.84), and was unaffected by adjustment for BMI ($OR = 6.66$, 95% $CI = 3.03$–14.64, $p = 2.36 \times 10^{-6}$) (**Figure 3B**, MAF = 0.087). One less

Table 2 | Association between *FTO* variants and average BMI.

SNP	Minor allele	Major allele	MAF	Beta (95%CI)	p†	r^2
rs8050136	**A**	**C**	**0.41**	**0.535 (0.363, 0.707)**	**1.28E-09**	**1.00**
rs9935401	A	G	0.41	0.535 (0.363, 0.707)	1.26E-09	1.00
rs11075990	G	A	0.42	0.534 (0.362, 0.706)	1.29E-09	1.00
rs9923233	C	G	0.42	0.534 (0.362, 0.706)	1.29E-09	1.00
rs9926289	A	G	0.42	0.534 (0.362, 0.706)	1.29E-09	1.00
rs9936385	C	T	0.42	0.534 (0.362, 0.706)	1.29E-09	1.00
rs9939609	A	T	0.42	0.534 (0.362, 0.706)	1.29E-09	1.00
rs8043757	T	A	0.41	0.539 (0.367, 0.711)	9.71E-10	1.00
rs7185735	G	A	0.42	0.536 (0.364, 0.708)	1.17E-09	1.00
rs17817449	G	T	0.40	0.548 (0.376, 0.720)	5.07E-10	1.00
rs7193144	C	T	0.41	0.529 (0.357, 0.701)	1.96E-09	1.00
rs3751812	T	G	0.34	0.572 (0.400, 0.744)	9.34E-11	0.99
rs55872725	T	C	0.35	0.561 (0.389, 0.733)	1.67E-10	0.94
rs1558902	A	T	0.35	0.560 (0.388, 0.732)	1.84E-10	0.94
rs62048402	A	G	0.35	0.560 (0.388, 0.732)	1.84E-10	0.94
rs11642015	T	C	0.35	0.561 (0.389, 0.733)	1.70E-10	0.94
rs1421085	C	T	0.35	0.561 (0.389, 0.733)	1.70E-10	0.94
rs9941349	T	C	0.37	0.564 (0.392, 0.736)	1.42E-10	0.92
rs9931494	G	C	0.37	0.561 (0.389, 0.733)	1.72E-10	0.92
rs12149832	A	G	0.35	0.560 (0.388, 0.732)	1.71E-10	0.90
rs1121980	A	G	0.44	0.522 (0.351, 0.693)	2.40E-09	0.88
rs9939973	A	G	0.43	0.528 (0.357, 0.699)	1.48E-09	0.88
rs9940646	G	C	0.43	0.528 (0.357, 0.699)	1.48E-09	0.88
rs9940128	A	G	0.43	0.527 (0.356, 0.698)	1.61E-09	0.88
rs9937053	A	G	0.43	0.530 (0.359, 0.701)	1.35E-09	0.88
rs9930333	G	T	0.44	0.534 (0.363, 0.705)	9.67E-10	0.88
rs9932754	C	T	0.39	0.544 (0.373, 0.715)	4.63E-10	0.85
rs9930506	G	A	0.39	0.544 (0.373, 0.715)	4.63E-10	0.85
rs9922619	T	G	0.39	0.553 (0.382, 0.724)	2.37E-10	0.85
rs8057044	G	A	0.47	0.530 (0.359, 0.701)	1.25E-09	0.72
rs17817288	G	A	0.48	0.528 (0.357, 0.699)	1.19E-09	0.68
rs9922047	C	G	0.44	0.502 (0.331, 0.673)	7.21E-09	0.64
rs1861866	C	T	0.44	0.498 (0.327, 0.669)	9.63E-09	0.64
rs8055197	G	A	0.44	0.498 (0.327, 0.669)	9.63E-09	0.64
rs10852521	T	C	0.44	0.497 (0.326, 0.668)	1.02E-08	0.64
rs8047395	G	A	0.43	0.496 (0.325, 0.667)	1.10E-08	0.64
rs8044769	T	C	0.42	0.504 (0.333, 0.675)	6.64E-09	0.62
rs3751813	G	T	0.45	0.419 (0.247, 0.591)	2.06E-06	0.57
rs4783819	G	C	0.33	0.414 (0.236, 0.592)	5.43E-06	0.41
rs1477196	A	G	0.32	0.410 (0.232, 0.588)	6.74E-06	0.40
rs7190492	A	G	0.33	0.426 (0.248, 0.604)	2.83E-06	0.40
rs7186521	G	A	0.45	0.251 (0.080, 0.422)	3.79E-03	0.09
rs1861869	G	C	0.47	0.274 (0.103, 0.445)	1.62E-03	0.08
rs1861868	T	C	0.44	0.256 (0.087, 0.425)	3.04E-03	0.08
rs6499640	G	A	0.39	0.264 (0.090, 0.438)	3.15E-03	0.06
rs11075986	G	C	0.12	0.065 (− 0.251, 0.381)	0.69	0.06
rs16945088	G	A	0.12	0.001 (− 0.317, 0.319)	0.99	0.06
rs8063946	T	C	0.12	0.101 (− 0.260, 0.462)	0.58	0.04
rs1075440	G	A	0.28	0.173 (− 0.011, 0.357)	0.06	0.04
rs16952520	G	A	0.09	0.205 (− 0.238, 0.648)	0.36	0.03
rs12447107	C	G	0.08	0.246 (− 0.379, 0.871)	0.44	0.01
rs7204609	C	T	0.10	0.469 (− 0.111, 1.049)	0.11	0.01
rs7199182	G	A	0.06	2.346 (0.472, 4.220)	0.01	0.00
rs1108102	A	T	0.03	1.045 (− 1.732, 3.822)	0.46	0.00

Analysis used an additive genetic model and linear regression adjusted for age, sex, and first three principal components using the imputed eMERGE samples. The SNPs below are sorted by p-value. The beta represents the kg/m^2 increase in BMI per minor allele. Linkage disequilibrium (r^2) was calculated between rs8050136 (bolded) and other FTO SNPs using the eMERGE imputed set. The Bonferroni correction alpha = 0.05 for 54 SNPs is 9.26×10^{-3}.

†Values are not corrected for multiple testing.

Table 3 | Meta-analysis PheWAS results for rs8050136 with and without adjustment for average BMI.

Phenotype	Cases	Not adjusted for BMI		Adjusted for BMI	
		$p^{\dagger}$	*OR* (95% CI)	$p^{\dagger}$	*OR* (95% *CI*)
Overweight	3943	1.38×10^{-8}	1.17 (1.11–1.24)	0.185	1.05 (0.98–1.12)
Obesity	1662	2.10×10^{-9}	1.25 (1.16–1.35)	0.017	1.11 (1.02–1.22)
Morbid obesity	756	1.07×10^{-7}	1.34 (1.20–1.48)	0.016	1.17 (1.03–1.33)
Type 2 diabetes	3936	2.34×10^{-6}	1.14 (1.08–1.21)	4.56×10^{-4}	1.09 (1.03–1.15)
Sleep apnea	2335	3.33×10^{-5}	1.14 (1.07–1.22)	0.040	1.07 (1.00–1.15)
Cystic mastopathy	967	2.00×10^{-4}	0.82 (0.74–0.91)	4.75×10^{-4}	0.84 (0.75–0.92)
Chronic Nonalcoholic Liver disease	684	2.22×10^{-4}	1.23 (1.10–1.37)	1.86×10^{-3}	1.19 (1.07–1.33)
Chronic Ulcer of Leg or Foot	768	8.31×10^{-4}	1.19 (1.08–1.32)	2.55×10^{-3}	1.17 (1.06–1.30)
Acute Renal Failure	2047	1.12×10^{-3}	1.12 (1.05–1.20)	3.74×10^{-3}	1.11 (1.03–1.19)
Staphylococcus infections	723	2.44×10^{-3}	1.18 (1.06–1.31)	5.76×10^{-3}	1.16 (1.04–1.29)
Superficial cellulitis and abscess	2861	5.65×10^{-3}	1.09 (1.02–1.15)	0.039	1.06 (1.00–1.13)
Streptococcus infection	428	4.26×10^{-3}	1.21 (1.05–1.39)	6.56×10^{-3}	1.21 (1.05–1.39)
Osteomyelitis	352	6.15×10^{-3}	1.23 (1.06–1.43)	0.011	1.21 (1.04–1.41)
All gram positive infections	1095	6.21×10^{-4}	1.16 (1.07–1.27)	1.3×10^{-3}	1.15 (1.06–1.26)
Joint effusions	387	2.35×10^{-3}	1.25 (1.08–1.44)	6.90×10^{-3}	1.22 (1.06–1.41)

This table includes all phenotypes with p-value less than 1.00×10^{-4} prior to BMI adjustment. The Bonferroni alpha = 0.05 equates to a p-value of 4.95×10^{-5}, and an FDR of q = 0.05 gives a p-value of 2.48×10^{-4}. OR, Odds ratio; CI, confidence interval. The ICD-9 codes that are associated with each phenotype can be found at the PheWAS catalog located online at http://phewas.mc.vanderbilt.edu/.

†*Values are not corrected for multiple testing.*

common genetic variant rs7199182 (**Figure 3C**, MAF = 0.064) was associated with chronic periodontitis (202 cases, $OR = 14.58$, 95% $CI = 3.97–53.57$, $p = 5.40 \times 10^{-5}$), and was not changed with adjustment for BMI with the signal being slightly stronger ($OR = 14.66$, 95% $CI = 3.99–53.84$, $p = 5.20 \times 10^{-5}$). Neither rs16952520 nor rs7199182 were associated with obesity or T2D. Detailed results for selected SNPs are shown in Supplementary Tables 3, 4.

DISCUSSION

We studied the pleiotropic patterns for *FTO* variants with and without adjustment for BMI using phenome-wide associations in two large EHR cohorts. Consistent with other studies, we identified statistically significant associations with obesity, morbid obesity, and T2D among SNPs known to be associated with BMI; these associations were attenuated by adjustment for BMI. We also identified an association with OSA and trends toward association with NAFLD, fibrocystic breast disease, and infections, primarily gram-positive, with obesity-related SNPs. Some of these potential associations seem independent of BMI adjustment. Fibrocystic breast changes are a common benign breast disease and traditionally not thought related to obesity, including several epidemiological studies (Friedenreich et al., 2000; Baer et al., 2005; Li et al., 2005). Gram-positive infections could be explained in part by higher incidence of T2D in genetic variants of *FTO*. By analyzing other SNPs not significantly associated with BMI in our analysis, we also identified a few other potential associations with less common traits not associated with obesity (periodontitis, non-inflammatory diseases of the cervix); neither of these SNPs is in high LD with obesity-related SNPs. The most common ICD-9 code for "non-inflammatory disorders of the cervix" is cervical stenosis or stricture not related to congenital abnormalities or labor, which can result from surgical procedures, radiation, trauma, repeated vaginal infections, or menopause-related atrophy. These results, along with the recent association of *FTO* variants with *IRX3* regulation (Smemo et al., 2014), suggest a broader role for *FTO* beyond that of regulating fat mass.

The question of whether the association of *FTO* variants and T2D is influenced by obesity or both obesity and *FTO* has been studied previously. A UK study of 9103 individuals demonstrated the loss of association after adjustment for BMI, as the T2D-*FTO* association prior to adjustment for BMI showed an $OR = 1.15$, $p = 9 \times 10^{-6}$ and after adjustment showed an $OR = 1.03$, $p = 0.44$ (Frayling et al., 2007). However, other studies suggest that T2D's association with *FTO* remains after adjustment for BMI (Hertel et al., 2011; Li et al., 2012). Li et al. studied 96,551 East and South Asians and demonstrated an association with T2D ($OR = 1.15$, $p = 5.5 \times 10^{-8}$) that was only partially attenuated after adjustment for BMI ($OR = 1.10$, $p = 6.6 \times 10^{-5}$) (Li et al., 2012). Similarly, Hertel et al. observed a significant T2D-*FTO* association even after adjustment for BMI in 41,504 Scandinavians, with the *OR* prior to adjustment of 1.13, $p = 4.5 \times 10^{-8}$ and after adjustment, $OR = 1.09$, $p = 1.2 \times 10^{-4}$ (Hertel et al., 2011). Finally a meta-analysis of 24,198 individuals demonstrated *FTO* rs9939609 (in high LD with rs8050136 with $r^2 > 0.8$) was highly significantly associated with T2D before and after adjustment for BMI (before adjustment $OR = 1.14$, 95% $CI = 1.12–1.16$, $p = 1.00 \times 10^{-41}$; after adjustment $OR = 1.07$, 95% $CI = 1.05–1.09$, $p = 6.42 \times 10^{-41}$) (Xi et al., 2014). However, among individuals of European ancestry, the association was markedly attenuated after adjustment for BMI (before

adjustment $OR = 1.14$, 95% $CI = 1.11$–1.16, $p = 1.36 \times 10^{-36}$; after adjustment $OR = 1.06$, 95% $CI = 1.04$–1.09, $p = 3.51 \times 10^{-8}$). In our study, the association between *FTO* and T2D did not decrease after adjustment for BMI as markedly as phenotypes such as obesity or sleep apnea. The effect sizes of these associations with T2D in our study closely parallels these larger studies (before BMI adjustment: $OR = 1.14$, 95% $CI = 1.08$–1.21, $p = 2.11 \times 10^{-6}$; after adjustment: $OR = 1.09$, 95% $CI = 1.03$–1.15, $p = 2.62 \times 10^{-3}$). Although these results show an association of *FTO* with T2D, a mediation analysis first demonstrating the associations of *FTO* SNPs with BMI and pre-diagnostic BMI with T2D, and subsequently modeling both *FTO* SNPs and pre-diagnostic BMI on T2D would help determine the direct and indirect effects of *FTO* on T2D.

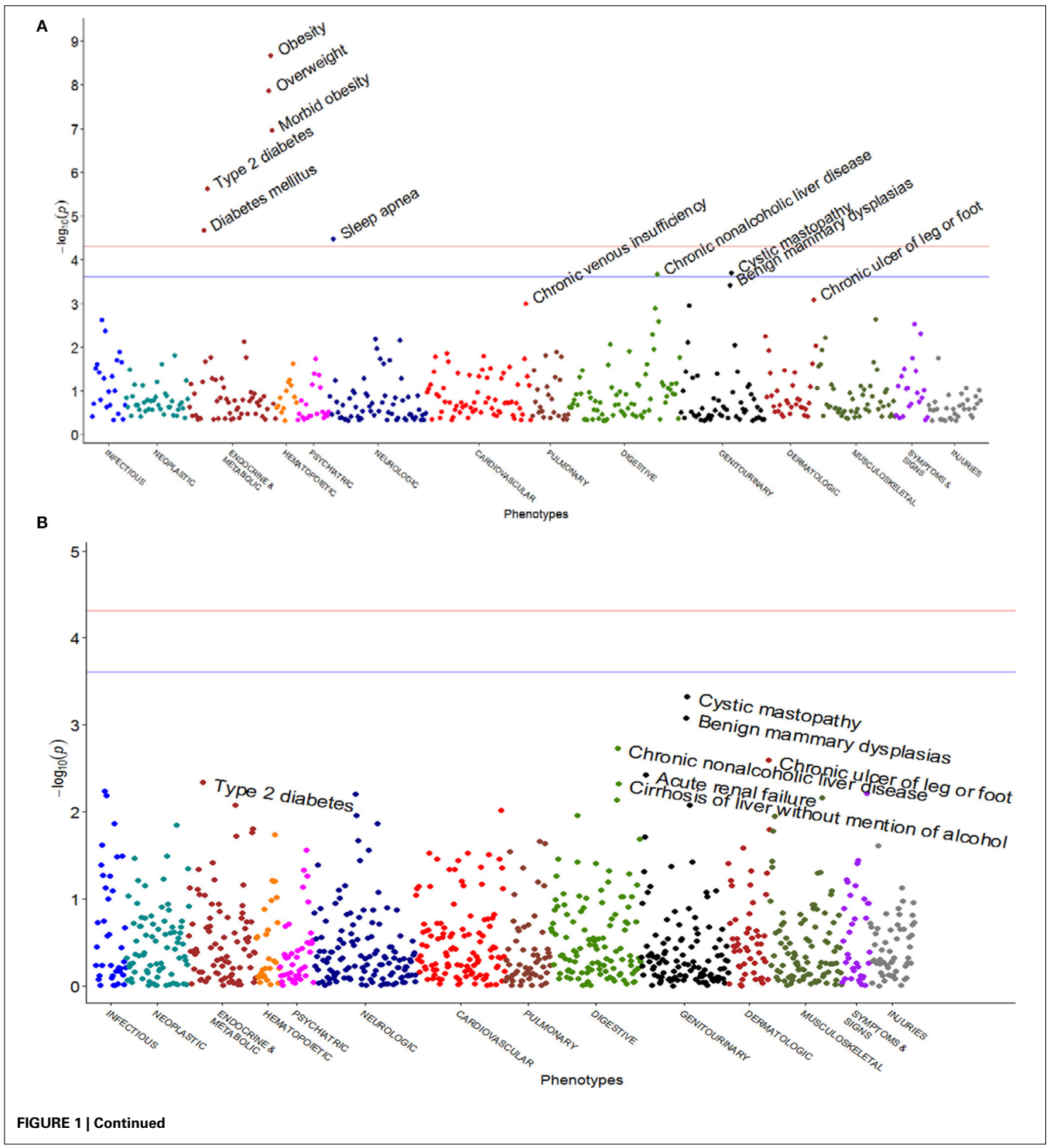


FIGURE 1 | Continued

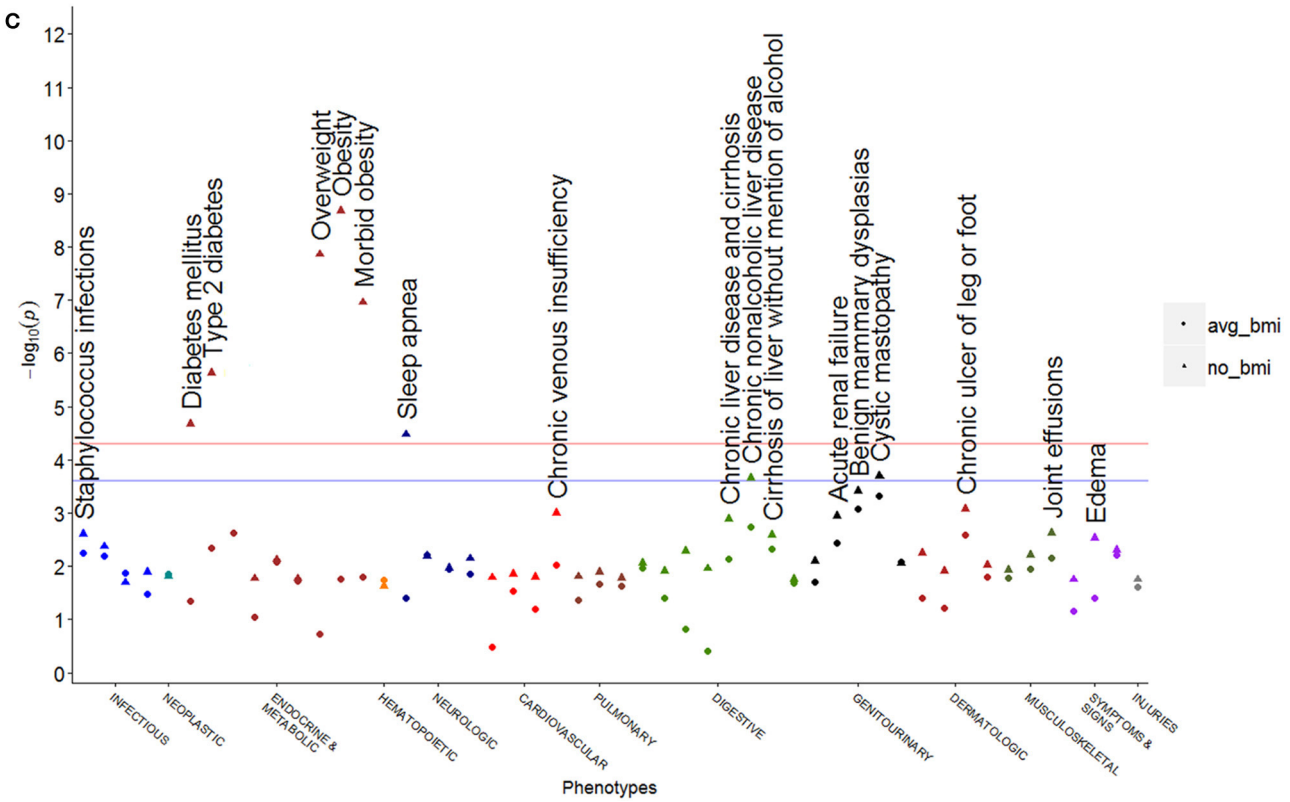


FIGURE 1 | PheWAS plots for *FTO* rs8050136 with and without BMI adjustment. The pink horizontal line represents $p = 4.95 \times 10^{-5}$, which is the Bonferroni correction, and the blue horizontal line represents an FDR $q = 0.05$ ($p = 2.48 \times 10^{-4}$). **(A)** without BMI adjustment, **(B)** with BMI adjustment, and **(C)** most significant phenotypic associations before and after BMI adjustment (BMI-unadjusted values are shown as triangles and average BMI values are shown as dots) plotted on the same axis. The colors of points indicate the membership according to the phenotype classes identified on the X axis.

Many of our findings, while having strong signals, were not significant after Bonferroni correction. The significant associations using Bonferroni correction included obesity, T2D, and OSA prior to BMI adjustment. After adjustment for average BMI, no associations retained statistical significance, but multiple phenotypes approached significance including T2D, NAFLD, and the protective effect on fibrocystic breast disease.

There is still much debate and uncertainty about both phenotypic association and protein functionality of *FTO*. Human *FTO* protein expression studies fail to replicate *FTO*'s association with obesity observed in mouse models (Klöting et al., 2008; Wåhlén et al., 2008; Grunnet et al., 2009). Recent studies have shown that the SNPs in *FTO* that are associated with obesity regulate *IRX3* expression, which is highly expressed in the brain (Smemo et al., 2014). Studies have described the association between *FTO* and obesity, while the association between T2D and *FTO* is debated (Hubacek et al., 2008; Li et al., 2008; Xi and Mi, 2009; Liu et al., 2010; Hotta et al., 2011). More studies with larger populations are required to assess the validity of many of these associations. The results of these associations show the power of the PheWAS method to efficiently detect known and novel pleiotropic associations of genetic variants.

BMI is an inexact surrogate for adiposity. Indeed, individuals with a high BMI do not necessarily have a high body fat percentage, thus BMI may not be the optimal definition of the phenotype (Müller et al., 2010). However, BMI has been shown to be as good a surrogate for obesity and diabetes as other central obesity indicators in multiple studies and meta-analyses (Vazquez et al., 2007; Nyamdorj et al., 2008, 2009).

Prior studies have suggested several other phenotypes that may be associated with *FTO* variants, including pancreatic cancer, Alzheimer's disease, attention deficit hyperactivity disorder, and alcoholism (Keller et al., 2011; Lurie et al., 2011; Sobczyk-Kopciol et al., 2011; arcOGEN Consortium et al., 2012; Corella et al., 2012; Reitz et al., 2012; Velders et al., 2012). We did not find evidence for these associations in our data set ($p > 0.05$) (**Table 4**), but in these cases we may be underpowered to find an association, with case sizes of 76 (attention deficit hyperactivity disorder), 183 (pancreatic cancer), 192 (Alzheimer's disease), and 267 (alcoholism) in our population. A trend toward association between *FTO* rs8044769 and osteoarthritis was observed in a previous GWAS study (rs8044769, $r^2 = 0.647$ with rs8050136, $p = 4 \times 10^{-6}$) (arcOGEN Consortium et al., 2012). Our observation of a trend toward associations with joint effusions, which may be caused by osteoarthritis, lends some support to this inflammatory association.

Further analysis of multiple SNPs associated with obesity in *FTO* yielded some interesting results. First, the SNPs that are

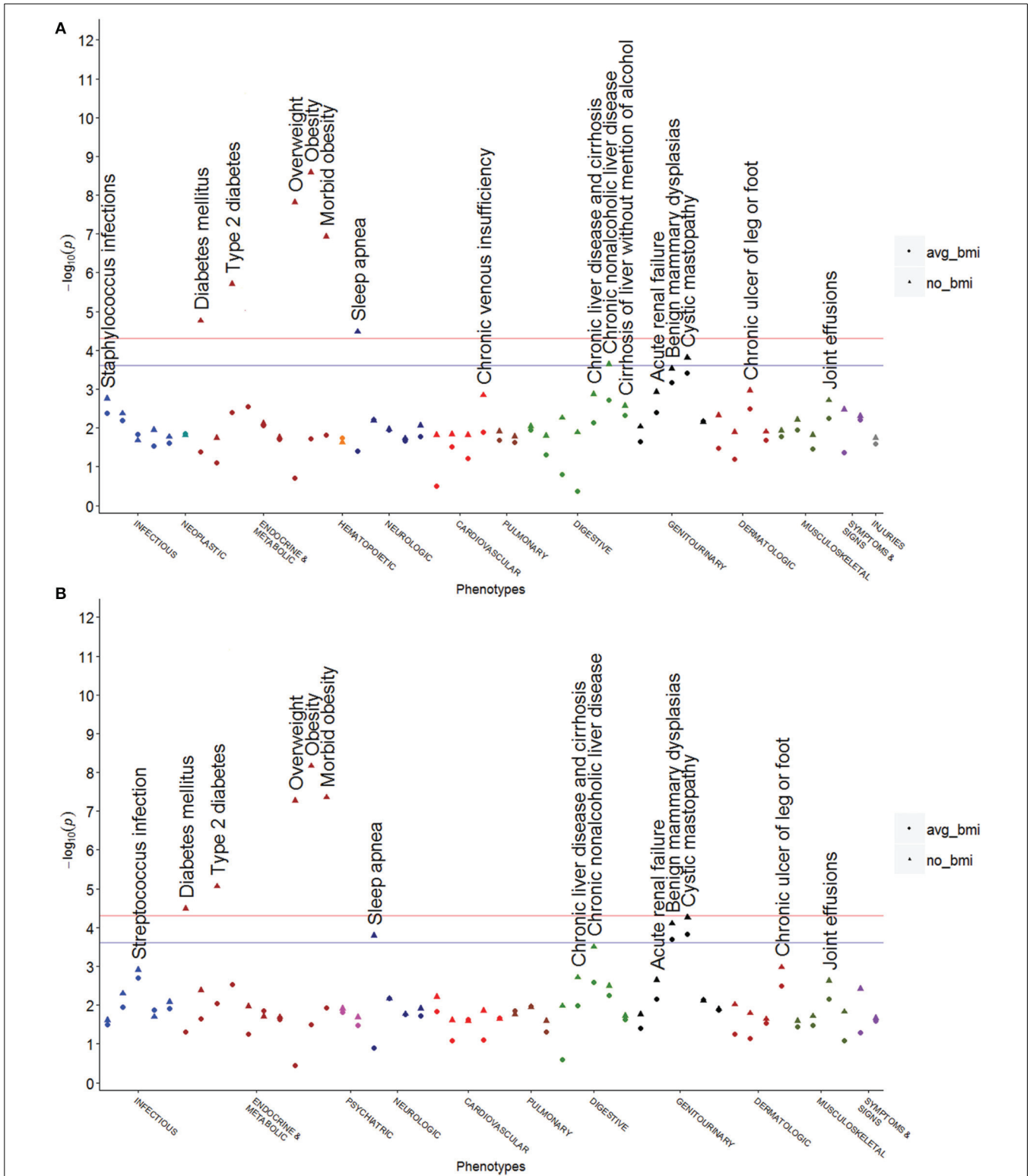


FIGURE 2 | PheWAS plots for other obesity associated SNPs in high LD with rs8050136. These plots show unadjusted values and the average BMI adjusted values on the same axis. These SNPs are associated with BMI and have different correlations with rs8050136. These SNPs are present in both datasets and are presented as meta-analyses below. The pink horizontal line represents $p = 4.95 \times 10^{-5}$, which is the Bonferroni correction, and the blue horizontal line represents an FDR $q = 0.05$ ($p = 2.48 \times 10^{-4}$). **(A)** rs9939609 is reported widely in the literature and has a nearly identical pattern of associations to rs8050136 ($r^2 = 0.96$). **(B)** rs9941349 also has a similar pattern to rs8050136 but cystic mastopathy is marginally more associated ($p = 5.41 \times 10^{-5}$, $OR = 0.81$ before BMI adjustment) than in rs8050136 ($r^2 = 0.88$).

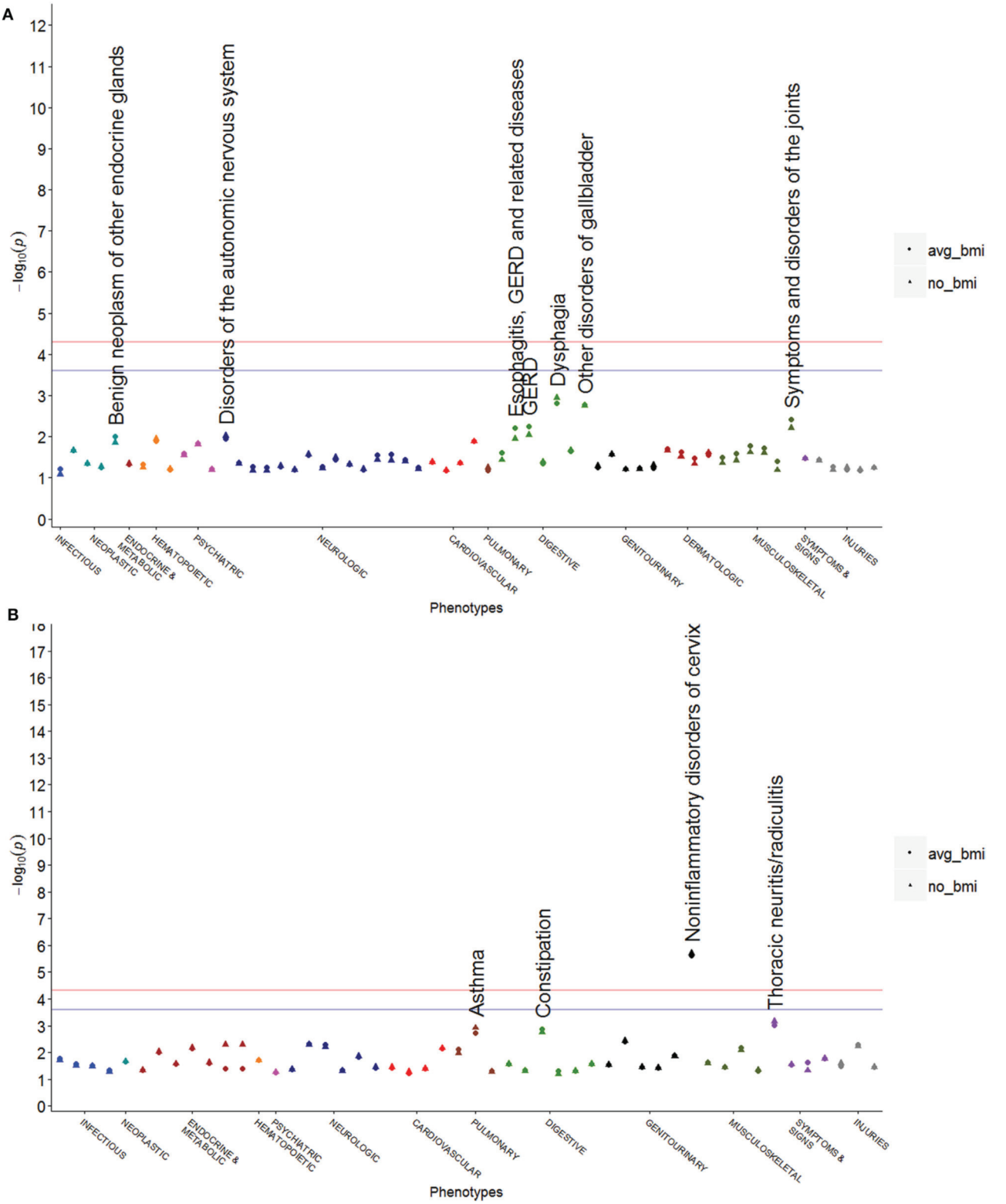


FIGURE 3 | Continued

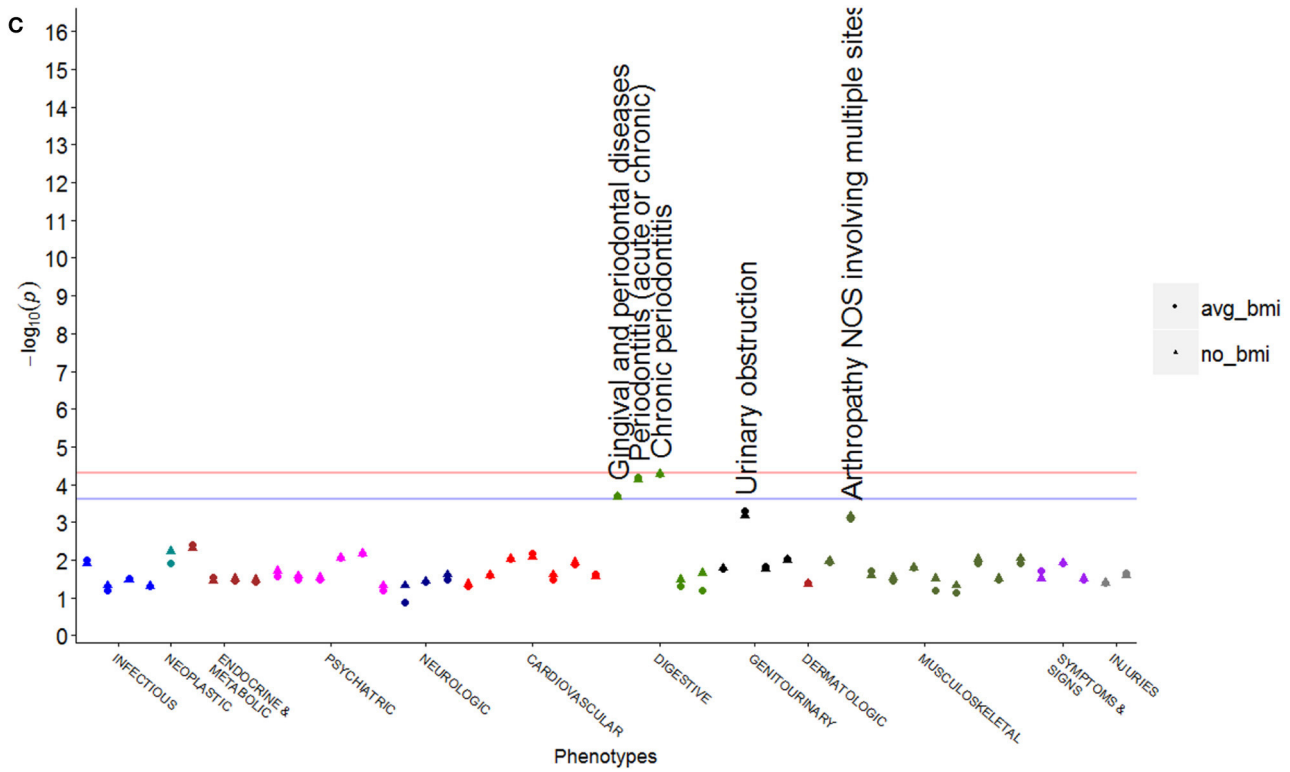


FIGURE 3 | PheWAS plots for other obesity associated SNPs in low LD with rs8050136. These plots show values without adjustment for BMI (shown as triangles) and with adjustment for average BMI (shown as dots) plotted on the same axis. **(A)** rs6499640 is in both datasets with a lower LD with rs8050136 ($r^2 = 0.06$) and has a different phenotype pattern than rs8050136 **(B)** rs16952520 is only present in the eMERGE population and has low LD with rs8050136 ($r^2 = 0.03$) and while not strongly associated with obesity or diabetes does show significant association with non-inflammatory disorders of the cervix ($OR = 6.76$, $p = 1.92 \times 10^{-6}$), unaffected by adjustment for BMI ($OR = 6.66$, $p = 2.36 \times 10^{-6}$) **(C)** rs7199182 is only present in the eMERGE population and has a low LD with rs8050136 ($r^2 = 0.04$) and is associated with chronic periodontitis before and after BMI adjustment (no adjustment: $p = 5.40 \times 10^{-5}$; BMI adjustment: $p = 5.20 \times 10^{-5}$).

Table 4 | Meta-analysis PheWAS results of rs8050136 for previously reported phenotypes associated with genetic variants.

Phenotype	Cases	Not adjusted for BMI		Adjusted for BMI	
		$p^{\dagger}$	*OR* (95% *CI*)	$p^{\dagger}$	*OR* (95% *CI*)
Attention deficit hyperactivity disorder	76	0.085	0.74 (0.52–1.04)	0.11	0.75 (0.53–1.06)
Pancreatic cancer	183	0.23	1.14 (0.92–1.40)	0.19	1.15 (0.93–1.42)
Alcoholism	267	0.37	1.08 (0.91–1.29)	0.32	1.09 (0.92–1.30)
Senile dementia	192	0.90	0.99 (0.80–1.22)	0.90	0.99 (0.80–1.22)
Osteoarthritis	6328	0.20	1.03 (0.98–1.08)	0.88	1.00 (0.95–1.06)

This table includes select phenotypes that have been previously reported in the literature. The Bonferroni alpha = 0.05 equates to a p-value of 4.95×10^{-5}, and an FDR of q = 0.05 gives a p-value of 2.48×10^{-4}. OR, Odds ratio; CI, confidence interval.

†Values are not corrected for multiple testing.

in high correlation with rs8050136 ($r^2 > 0.8$) have very similar results to rs8050136, which is what we would expect. There are also SNPs that were associated with fibrocystic breast disease prior to adjustment for BMI. rs7199182, is in low LD with rs8050136 ($r^2 < 0.01$), showed significant associations with chronic periodontitis before and after adjustment for BMI. Further analysis of this SNP and its association with chronic periodontitis will need to be investigated to validate this finding.

One important consideration of this analysis is the small overlap of genotyped SNPs between the BioVU and eMERGE population. There are multiple SNPs that are present in both datasets and are highly correlated with rs8050136, but only rs6499640, which is in weak LD with rs8050136 ($r^2 = 0.06$), was genotyped in both datasets. We are unable to impute the BioVU. The lack of overlapping SNPs limits our sample size to evaluate more of the potentially novel findings. Limitations

caution interpretation of this study. Some of the case sizes were small and will require larger populations to validate. PheWAS analyses require robust EHR systems that can query patient cohorts efficiently. We used ICD-9 codes for the determination of phenotypes, codes which can be unreliable, inaccurate, and incomplete (Kern et al., 2006; Campbell et al., 2011); however, this could tend to result in missed, rather than false, associations. In addition to the caveats of ICD-9 codes, there are limitations of multiple hypothesis testing that come with comparisons of over 1000 phenotypes. Significance corrections like Bonferroni may be too strict; some of the near-significant pleiotropic associations may, in fact, represent genuine associations. Further testing with larger populations and more carefully defined phenotypes are needed to determine whether these associations are real.

Here we demonstrate the use of the PheWAS method to illustrate pleiotropic effects of variation in the gene *FTO*. When examining this gene with known pleiotropy, we were able to reproduce previously-discovered associations and identify potential new associations, some of which appear independent of obesity.

ACKNOWLEDGMENTS

This work was supported by the eMERGE Network. The eMERGE Network was initiated and funded by NHGRI through the following grants: U01HG006389 (Essentia Institute of Rural Health); U01HG006382 (Geisinger Clinic); U01HG006375, UO1AG06781 (Group Health Cooperative and University of Washington); U01HG06379 (Mayo Clinic).); U01HG006380 (Mount Sinai School of Medicine); U01HG006388 (Northwestern University); U01HG006378 (Vanderbilt University); and U01HG006385 (Vanderbilt University serving as the Coordinating Center); and by National Center for Advancing Translational Sciences (NCATS) grant UL1TR000427 (Marshfield Clinic). Development of the PheWAS method is also supported by R01-LM010685 from the National Library of Medicine. BioVU received and continues to receive support through the NCATS grant 2 UL1 TR000445.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2014.00250/abstract

REFERENCES

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073. doi: 10.1038/nature09534

arcOGEN Consortium, arcOGEN Collaborators, Zeggini, E., Panoutsopoulou, K., Southam, L., Rayner, N. W., et al. (2012). Identification of new susceptibility loci for osteoarthritis (arcOGEN): a genome-wide association study. *Lancet* 380, 815–823. doi: 10.1016/S0140-6736(12)60681-3

Baer, H. J., Schnitt, S. J., Connolly, J. L., Byrne, C., Willett, W. C., Rosner, B., et al. (2005). Early life factors and incidence of proliferative benign breast disease. *Cancer Epidemiol. Biomark. Prev.* 14, 2889–2897. doi: 10.1158/1055-9965.EPI-05-0525

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B Methodol.* 57, 289–300.

Campbell, P. G., Malone, J., Yadla, S., Chitale, R., Nasser, R., Maltenfort, M. G., et al. (2011). Comparison of ICD-9-based, retrospective, and prospective assessments of perioperative complications: assessment of accuracy in reporting. *J. Neurosurg. Spine* 14, 16–22. doi: 10.3171/2010.9.SPINE10151

Carroll, R. J., Bastarache, L., and Denny, J. C. (2014). R PheWAS: data analysis and plotting tools for phenome wide association studies in the R environment. *Bioinformatics*. doi: 10.1093/bioinformatics/btu197. [Epub ahead of print].

Corella, D., Ortega-Azorín, C., Sorlí, J. V., Covas, M. I., Carrasco, P., Salas-Salvadó, J., et al. (2012). Statistical and biological gene-lifestyle interactions of MC4R and FTO with diet and physical activity on obesity: new effects on alcohol consumption. *PLoS ONE* 7:e52344. doi: 10.1371/journal.pone.0052344

Crespi, B. J. (2010). The origins and evolution of genetic disease risk in modern humans. *Ann. N.Y. Acad. Sci.* 1206, 80–109. doi: 10.1111/j.1749-6632.2010.05707.x

Crosslin, D. R., McDavid, A., Weston, N., Zheng, X., Hart, E., de Andrade, M., et al. (2013). Genetic variation associated with circulating monocyte count in the eMERGE Network. *Hum. Mol. Genet.* 22, 2119–2127. doi: 10.1093/hmg/ddt010

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* 31, 1102–1110. doi: 10.1038/nbt.2749

Denny, J. C., Crawford, D. C., Ritchie, M. D., Bielinski, S. J., Basford, M. A., Bradford, Y., et al. (2011). Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am. J. Hum. Genet.* 89, 529–542. doi: 10.1016/j.ajhg.2011.09.008

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., et al. (2010a). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 26, 1205–1210. doi: 10.1093/bioinformatics/btq126

Denny, J. C., Ritchie, M. D., Crawford, D. C., Schildcrout, J. S., Ramirez, A. H., Pulley, J. M., et al. (2010b). Identification of genomic predictors of atrioventricular conduction: using electronic medical records as a tool for genome science. *Circulation* 122, 2016–2021. doi: 10.1161/CIRCULATIONAHA.110.948828

Ding, K., de Andrade, M., Manolio, T. A., Crawford, D. C., Rasmussen-Torvik, L. J., Ritchie, M. D., et al. (2013). Genetic variants that confer resistance to malaria are associated with red blood cell traits in African-Americans: an electronic medical record-based genome-wide association study. *G3 (Bethesda)* 3, 1061–1068. doi: 10.1534/g3.113.006452

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316, 889–894. doi: 10.1126/science.1141634

Friedenreich, C., Bryant, H., Alexander, F., Hugh, J., Danyluk, J., and Page, D. (2000). Risk factors for benign proliferative breast disease. *Int. J. Epidemiol.* 29, 637–644. doi: 10.1093/ije/29.4.637

Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D., and Province, M. A. (2010). Avoiding the high Bonferroni penalty in genome-wide association studies. *Genet. Epidemiol.* 34, 100–105. doi: 10.1002/gepi.20430

Gong, J., Schumacher, F., Lim, U., Hindorff, L. A., Haessler, J., Buyske, S., et al. (2013). Fine Mapping and Identification of BMI Loci in African Americans. *Am. J. Hum. Genet.* 93, 661–671. doi: 10.1016/j.ajhg.2013.08.012

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., et al. (2013). The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* 15, 761–771. doi: 10.1038/gim.2013.72

Grunnet, L. G., Nilsson, E., Ling, C., Hansen, T., Pedersen, O., Groop, L., et al. (2009). Regulation and function of FTO mRNA expression in human skeletal muscle and subcutaneous adipose tissue. *Diabetes* 58, 2402–2408. doi: 10.2337/db09-0205

Hebbring, S. J., Schrodi, S. J., Ye, Z., Zhou, Z., Page, D., and Brilliant, M. H. (2013). A PheWAS approach in studying HLA-DRB1*1501. *Genes Immun.* 14, 187–191. doi: 10.1038/gene.2013.2

Hertel, J. K., Johansson, S., Sonestedt, E., Jonsson, A., Lie, R. T., Platou, C. G. P., et al. (2011). FTO, type 2 diabetes, and weight gain throughout adult life: a meta-analysis of 41,504 subjects from the Scandinavian HUNT, MDC, and MPP studies. *Diabetes* 60, 1637–1644. doi: 10.2337/db10-1340

Hotta, K., Kitamoto, T., Kitamoto, A., Mizusawa, S., Matsuo, T., Nakata, Y., et al. (2011). Association of variations in the FTO, SCG3 and MTMR9 genes with

metabolic syndrome in a Japanese population. *J. Hum. Genet.* 56, 647–651. doi: 10.1038/jhg.2011.74

Hubacek, J. A., Bohuslavova, R., Kuthanova, L., Kubinova, R., Peasey, A., Pikhart, H., et al. (2008). The FTO gene and obesity in a large Eastern European population sample: the HAPIEE study. *Obesity (Silver Spring)* 16, 2764–2766. doi: 10.1038/oby.2008.421

Hunter, J. E., Schmidt, F. L., and Jackson, G. B. (1982). *Meta-Analysis*. Beverly Hills, CA: Sage Publ.

Jacobsson, J. A., Schiöth, H. B., and Fredriksson, R. (2012). The impact of intronic single nucleotide polymorphisms and ethnic diversity for studies on the obesity gene FTO. *Obes. Rev.* 13, 1096–1109. doi: 10.1111/j.1467-789X.2012.01025.x

Keller, L., Xu, W., Wang, H.-X., Winblad, B., Fratiglioni, L., and Graff, C. (2011). The obesity related gene, FTO, interacts with APOE, and is associated with Alzheimer's disease risk: a prospective cohort study. *J. Alzheimers Dis.* 23, 461–469. doi: 10.3233/JAD-2010-101068

Kern, E. F. O., Maney, M., Miller, D. R., Tseng, C.-L., Tiwari, A., Rajan, M., et al. (2006). Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. *Health Serv. Res.* 41, 564–580. doi: 10.1111/j.1475-6773.2005.00482.x

Klöting, N., Schleinitz, D., Ruschke, K., Berndt, J., Fasshauer, M., Tönjes, A., et al. (2008). Inverse relationship between obesity and FTO gene expression in visceral adipose tissue in humans. *Diabetologia* 51, 641–647. doi: 10.1007/s00125-008-0928-9

Kullo, I. J., Ding, K., Shameer, K., McCarty, C. A., Jarvik, G. P., Denny, J. C., et al. (2011). Complement receptor 1 gene variants are associated with erythrocyte sedimentation rate. *Am. J. Hum. Genet.* 89, 131–138. doi: 10.1016/j.ajhg.2011.05.019

Li, H., Kilpeläinen, T. O., Liu, C., Zhu, J., Liu, Y., Hu, C., et al. (2012). Association of genetic variation in FTO with risk of obesity and type 2 diabetes with data from 96,551 East and South Asians. *Diabetologia* 55, 981–995. doi: 10.1007/s00125-011-2370-7

Li, H., Wu, Y., Loos, R. J. F., Hu, F. B., Liu, Y., Wang, J., et al. (2008). Variants in the fat mass- and obesity-associated (FTO) gene are not associated with obesity in a Chinese Han population. *Diabetes* 57, 264–268. doi: 10.2337/db07-1130

Li, S., Aufiero, B., Schiltz, R. L., and Walsh, M. J. (2000). Regulation of the homeodomain CCAAT displacement/cut protein function by histone acetyltransferases p300/CREB-binding protein (CBP)-associated factor and CBP. *Proc. Natl. Acad. Sci. U.S.A.* 97, 7166–7171. doi: 10.1073/pnas.130028697

Li, W., Ray, R. M., Lampe, J. W., Lin, M.-G., Gao, D. L., Wu, C., et al. (2005). Dietary and other risk factors in women having fibrocystic breast conditions with and without concurrent breast cancer: a nested case-control study in Shanghai, China. *Int. J. Cancer* 115, 981–993. doi: 10.1002/ijc.20964

Liu, Y., Liu, Z., Song, Y., Zhou, D., Zhang, D., Zhao, T., et al. (2010). Meta-analysis added power to identify variants in FTO associated with type 2 diabetes and obesity in the Asian population. *Obesity (Silver Spring)* 18, 1619–1624. doi: 10.1038/oby.2009.469

Lurie, G., Gaudet, M. M., Spurdle, A. B., Carney, M. E., Wilkens, L. R., Yang, H. P., et al. (2011). The obesity-associated polymorphisms FTO rs9939609 and MC4R rs17782313 and endometrial cancer risk in non-Hispanic white women. *PLoS ONE* 6:e16756. doi: 10.1371/journal.pone.0016756

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 39, 906–913. doi: 10.1038/ng2088

McCarty, C. A., Chisholm, R. L., Chute, C. G., Kullo, I. J., Jarvik, G. P., Larson, E. B., et al. (2011). The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics* 4:13. doi: 10.1186/1755-8794-4-13

Müller, M. J., Bosy-Westphal, A., and Krawczak, M. (2010). Genetic studies of common types of obesity: a critique of the current use of phenotypes. *Obes. Rev.* 11, 612–618. doi: 10.1111/j.1467-789X.2010.00734.x

Nyamdorj, R., Qiao, Q., Söderberg, S., Pitkäniemi, J., Zimmet, P., Shaw, J., et al. (2008). Comparison of body mass index with waist circumference, waist-to-hip ratio, and waist-to-stature ratio as a predictor of hypertension incidence in Mauritius. *J. Hypertens.* 26, 866–870. doi: 10.1097/HJH.0b013e3282f624b7

Nyamdorj, R., Qiao, Q., Söderberg, S., Pitkäniemi, J. M., Zimmet, P. Z., Shaw, J. E., et al. (2009). BMI compared with central obesity indicators as a predictor of diabetes incidence in Mauritius. *Obesity (Silver Spring)* 17, 342–348. doi: 10.1038/oby.2008.503

Park, S. L., Cheng, I., Pendergrass, S. A., Kucharska-Newton, A. M., Lim, U., Ambite, J. L., et al. (2013). Association of the FTO obesity risk variant rs8050136 with percentage of energy intake from fat in multiple racial/ethnic populations: the PAGE study. *Am. J. Epidemiol.* 178, 780–790. doi: 10.1093/aje/kwt028

Pathak, J., Kiefer, R. C., Bielinski, S. J., and Chute, C. G. (2012). Mining the human phenome using semantic web technologies: a case study for Type 2 Diabetes. *AMIA Annu. Symp. Proc.* 2012, 699–708.

Pathak, J., Wang, J., Kashyap, S., Basford, M., Li, R., Masys, D. R., et al. (2011). Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. *J. Am. Med. Inform. Assoc.* 18, 376–386. doi: 10.1136/amiajnl-2010-000061

Pavlicev, M., and Wagner, G. P. (2012). A model of developmental evolution: selection, pleiotropy and compensation. *Trends Ecol. Evol.* 27, 316–322. doi: 10.1016/j.tree.2012.01.016

Pendergrass, S. A., Brown-Gentry, K., Dudek, S., Frase, A., Torstenson, E. S., Goodloe, R., et al. (2013). Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet.* 9:e1003087. doi: 10.1371/journal.pgen.1003087

Pendergrass, S. A., Brown-Gentry, K., Dudek, S. M., Torstenson, E. S., Ambite, J. L., Avery, C. L., et al. (2011). The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet. Epidemiol.* 35, 410–422. doi: 10.1002/gepi.20589

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. doi: 10.1038/ng1847

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/.

Rees, S. D., Islam, M., Hydrie, M. Z. I., Chaudhary, B., Bellary, S., Hashmi, S., et al. (2011). An FTO variant is associated with Type 2 diabetes in South Asian populations after accounting for body mass index and waist circumference. *Diabet. Med.* 28, 673–680. doi: 10.1111/j.1464-5491.2011.03257.x

Reitz, C., Tosto, G., Mayeux, R., Luchsinger, J. A., NIA-LOAD/NCRA Family Study Group, and Alzheimer's Disease Neuroimaging Initiative. (2012). Genetic variants in the Fat and Obesity Associated (FTO) gene and risk of Alzheimer's disease. *PLoS ONE* 7:e50354. doi: 10.1371/journal.pone.0050354

Ritchie, M. D., Denny, J. C., Zuvich, R. L., Crawford, D. C., Schildcrout, J. S., Bastarache, L., et al. (2013). Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* 127, 1377–1385. doi: 10.1161/CIRCULATIONAHA.112.000604

Shameer, K., Denny, J. C., Ding, K., Jouni, H., Crosslin, D. R., de Andrade, M., et al. (2013). A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum. Genet.* 133, 95–109. doi: 10.1007/s00439-013-1355-7

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., et al. (2011). Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* 89, 607–618. doi: 10.1016/j.ajhg.2011.10.004

Smemo, S., Tena, J. J., Kim, K.-H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 507, 371–375. doi: 10.1038/nature13138

Sobczyk-Kopciol, A., Broda, G., Wojnar, M., Kurjata, P., Jakubczyk, A., Klimkiewicz, A., et al. (2011). Inverse association of the obesity predisposing FTO rs9939609 genotype with alcohol consumption and risk for alcohol dependence. *Addiction* 106, 739–748. doi: 10.1111/j.1360-0443.2010.03248.x

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* 14, 483–495. doi: 10.1038/nrg3461

Stranger, B. E., Stahl, E. A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* 187, 367–383. doi: 10.1534/genetics.110.120907

Stratigopoulos, G., Padilla, S. L., LeDuc, C. A., Watson, E., Hattersley, A. T., McCarthy, M. I., et al. (2008). Regulation of Fto/Ftm gene expression in mice and humans. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 294, R1185–R1196. doi: 10.1152/ajpregu.00839.2007

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., et al. (2011). Quality control procedures for genome-wide association studies. *Curr. Protoc. Hum. Genet.* Chapter 1, Unit 1.19. doi: 10.1002/0471142905.hg0119s68

Vazquez, G., Duval, S., Jacobs, D. R. Jr., and Silventoinen, K. (2007). Comparison of body mass index, waist circumference, and waist/hip ratio in predicting incident diabetes: a meta-analysis. *Epidemiol. Rev.* 29, 115–128. doi: 10.1093/epirev/mxm008

Velders, F. P., De Wit, J. E., Jansen, P. W., Jaddoe, V. W. V., Hofman, A., Verhulst, F. C., et al. (2012). FTO at rs9939609, food responsiveness, emotional control and symptoms of ADHD in preschool children. *PLoS ONE* 7:e49131. doi: 10.1371/journal.pone.0049131

Wagner, G. P., and Zhang, J. (2011). The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nat. Rev. Genet.* 12, 204–213. doi: 10.1038/nrg2949

Wåhlén, K., Sjölin, E., and Hoffstedt, J. (2008). The common rs9939609 gene variant of the fat mass- and obesity-associated gene FTO is related to fat cell lipolysis. *J. Lipid Res.* 49, 607–611. doi: 10.1194/jlr.M700448-JLR200

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–D1006. doi: 10.1093/nar/gkt1229

Xi, B., and Mi, J. (2009). FTO polymorphisms are associated with obesity but not with diabetes in East Asian populations: a meta-analysis. *Biomed. Environ. Sci.* 22, 449–457. doi: 10.1016/S0895-3988(10)60001-3

Xi, B., Takeuchi, F., Meirhaeghe, A., Kato, N., Chambers, J. C., Morris, A. P., et al. (2014). Associations of genetic variants in/near body mass index-associated genes with type 2 diabetes: a systematic meta-analysis. *Clin. Endocrinol. (Oxf.)*. doi: 10.1111/cen.12428. [Epub ahead of print].

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 May 2014; accepted: 10 July 2014; published online: 05 August 2014.
Citation: Cronin RM, Field JR, Bradford Y, Shaffer CM, Carroll RJ, Mosley JD, Bastarache L, Edwards TL, Hebbring SJ, Lin S, Hindorff LA, Crane PK, Pendergrass SA, Ritchie MD, Crawford DC, Pathak J, Bielinski SJ, Carrell DS, Crosslin DR, Ledbetter DH, Carey DJ, Tromp G, Williams MS, Larson EB, Jarvik GP, Peissig PL, Brilliant MH, McCarty CA, Chute CG, Kullo IJ, Bottinger E, Chisholm R, Smith ME, Roden DM and Denny JC (2014) Phenome-wide association studies demonstrating pleiotropy of genetic variants within FTO with and without adjustment for body mass index. Front. Genet. ***5****:250. doi: 10.3389/fgene.2014.00250*
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in
GENETICS

MINI REVIEW ARTICLE
published: 30 April 2014
doi: 10.3389/fgene.2014.00106

Analysis pipeline for the epistasis search – statistical versus biological filtering

Xiangqing Sun[1], Qing Lu[2], Shubhabrata Mukherjee[3], Paul K. Crane[3], Robert Elston[1] and Marylyn D. Ritchie[4]*

[1] *Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA*
[2] *Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, USA*
[3] *Department of Medicine, University of Washington, Seattle, WA, USA*
[4] *Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA, USA*

Edited by:
Mariza De Andrade, Mayo Clinic, USA

Reviewed by:
Frida Renstrom, Lund University, Sweden
Tao Wang, Albert Einstein College of Medicine, USA

Correspondence:
Marylyn D. Ritchie, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512A Wartik Lab, University Park, PA 16802, USA
e-mail: marylyn.ritchie@psu.edu

Gene–gene interactions may contribute to the genetic variation underlying complex traits but have not always been taken fully into account. Statistical analyses that consider gene–gene interaction may increase the power of detecting associations, especially for low-marginal-effect markers, and may explain in part the "missing heritability." Detecting pair-wise and higher-order interactions genome-wide requires enormous computational power. Filtering pipelines increase the computational speed by limiting the number of tests performed. We summarize existing filtering approaches to detect epistasis, after distinguishing the purposes that lead us to search for epistasis. Statistical filtering includes quality control on the basis of single marker statistics to avoid the analysis of bad and least informative data, and limits the search space for finding interactions. Biological filtering includes targeting specific pathways, integrating various databases based on known biological and metabolic pathways, gene function ontology and protein–protein interactions. It is increasingly possible to target single-nucleotide polymorphisms that have defined functions on gene expression, though not belonging to protein-coding genes. Filtering can improve the power of an interaction association study, but also increases the chance of missing important findings.

Keywords: epistasis, genetic interaction, biological interaction, filtering pipeline, optimal search

INTRODUCTION

Genome-wide association studies (GWAS) and next generation sequencing association studies based on single marker tests can identify many associated genetic variants, but typically explain only a small portion of the total estimated heritability. Gene–gene interactions may play an important role in the genetic etiology underlying complex phenotypes and statistical analyses that consider interaction may increase the power to detect epistatic genetic associations, especially among low-marginal-effect markers.

Bateson (1909) defined epistasis as distortions from Mendelian segregation ratios due to one gene masking the effects of another. Fisher (1918) introduced the term "epistacy," considering it to be any departure from a linear model in which the phenotypic effects of genotypes at two or more loci are assumed to be additive. Ever since, the terms "epistasis" and "gene–gene interaction" have often been used interchangeably and we make no distinction between these two terms here. However, the purpose of including such terms in any genetic model must be considered. If, for example, we know that segregation at each of two loci affects a particular phenotype, whether quantitative or binary, we already know there must be *biological* interaction. So, unless our purpose is to describe that interaction, no further analysis is necessary to *detect* its presence. In the case of a quantitative trait, whether or not there are interactions can depend on the scale of measurement, so the scale of the outcome is relevant. Factors that are additive with respective to the outcome measured on one scale may not be additive on another (Elston, 1961; Frankel and Schork, 1996; Greenland et al., 1998; Wang et al., 2010; Steen, 2012). Similarly, in the analysis of a binary trait, the link function used in a generalized linear model may determine whether or not interaction terms are necessary (Satagopan and Elston, 2012). If no transformation or change in link function can remove the interaction, it is called essential; in that case the best way to describe the interaction depends on how much of it is removable by a transformation or change of link function, and how much is essential. Simply describing the interaction by an appropriate statistical model may be useful for prediction in the same population as that sampled, but a prediction model may not be generalizable to other populations unless it is based on biological function.

Detecting pair-wise or higher-order statistical interactions can require enormous computational time. In a genome-wide analysis, the increased computational cost makes it impractical to examine whether interactions are non-essential or can be better described by removing non-additivity. Advances in computational methods, such as using a GPU framework (Yung et al., 2011; Zhu et al., 2013) and parallel computing strategies may overcome this limitation. However, the multiple hypothesis testing issue needs to be considered: this is the major reason why most existing epistasis studies are limited to searching for pair-wise interactions among a moderate number of genetic markers.

STATISTICAL METHODS FOR DETECTING STATISTICAL INTERACTIONS

Regression-based approaches are mostly used to model and test interactions. The regression approach has been implemented in the epistasis module of PLINK (Purcell et al., 2007) to test pair-wise diallelic by diallelic epistasis for both quantitative and binary traits. An extension of the PLINK epistasis module, FastEpistasis, uses an efficient parallel computation algorithm to test pair-wise interactions. FastEpistasis is 15 times faster than PLINK using a single core computer (Schüpbach et al., 2010). Marchini et al. (2005) proposed an approach for joint association analyses allowing for pair-wise interactions based on logistic models; their approach uses an exhaustive search among single-nucleotide polymorphisms (SNPs) meeting some low marginal significance threshold. The software package PLATO can perform linear or logistic regression interaction analysis, calculating the full model, the reduced model, and the likelihood ratio test comparing the two (Grady et al., 2010).

The advantages of regression-based approaches are the clear interpretation of the model and the parameters that relate genotypes to phenotype. However, regression-based approaches have many technical and computational disadvantages for testing higher-order interactions and require many more tests: the number of parameters to be tested increases exponentially with the number of SNPs in the model.

Model-free approaches, such as machine learning and pattern recognition, afford an alternative strategy, and are capable of detecting high-dimensional non-linear interactions. This approach generally does not estimate parameters. It finds combinations of SNPs that can best separate cases and controls associated with the disease by epistatic interactions or joint effects. Some model-free approaches collapse high dimensional data into two dimensions, such as the combinatorial partitioning method (CPM; Nelson et al., 2001), restricted partition method (RPM; Culverhouse et al., 2004), set association (Wille et al., 2003), and multifactor dimensionality reduction (MDR; Ritchie et al., 2001, 2003; Hahn et al., 2003).

Unsupervised pattern recognition has also been used to detect interactions. Li et al. (2011) proposed a method for family based studies to detect differentially inherited SNP modules by hierarchically clustering SNPs that could be interactively associated with a disease. They first construct a genomic context-based SNP network based on adjacency on the chromosome. The association between each SNP and disease is evaluated on the basis of mutual information between SNP identity by descent sharing and affection status sharing of pairs of siblings. Then they use a hierarchical clustering algorithm to find risk SNP modules (clusters) for which discriminative scores are locally maximal. In each module, the SNPs are within a certain network distance (defined as the number of edges separating connected SNPs), and the discriminative score of a module is the maximum mutual information of the SNPs in the module, reflecting the risk associated with the module.

A likelihood ratio-based Mann–Whitney approach (Lu et al., 2012) and its extension (Wei et al., 2013) are other non-parametric methods for detecting interaction. They use a multi-locus Mann–Whitney statistic to evaluate the joint association of a SNP combination. Using a computationally efficient forward selection algorithm makes these methods feasible for genome-wide gene–gene interaction analyses. Nevertheless, they require at least one SNP in the combination to have a significant marginal association. The non-parametric approaches do not suffer from the issue of an increasing number of parameters when modeling high-order interactions, but it is difficult to determine how the detected SNP combinations affect the disease, either via the single marker associations or via their interactions.

Some studies test marker–marker interactions by testing linkage disequilibrium (LD) in the diseased population (Zhao et al., 2006), or test the contrast of LD or Pearson correlation in cases and controls (Kam-Thong et al., 2010; Prabhu and Pe'er, 2012). These methods are based on the idea that, if two unlinked markers are interactively associated with a disease, the two markers will have LD patterns in the disease population. If controls are not studied, these methods assume that the controls do not exhibit similar patterns.

FILTERING PIPELINES FOR EPISTATIC INTERACTIONS PRIOR TO ANALYSIS

In GWAS, an exhaustive search among millions of SNPs for higher-order statistical interactions, or even just pair-wise interactions, could be computationally and statistically challenging. Filtering pipelines limit the number of tests performed between selected SNPs, whereas the use of computational technology and optimal algorithms increases the computational speed, and accelerates convergence if maximization is involved. While data driven filtering such as statistical filtering cleans the data to avoid the analysis of bad and least informative data, other types of filtering can be used purely to improve the power of interaction association analyses. In particular, filtering using biological knowledge limits the analysis to find the biologically most likely interactions.

Knowledge-driven filtering

Interaction models that are constructed based on specific biological knowledge are more likely to make sense. Research over the last several decades has accumulated vast amounts of biological information that is stored in public databases. These include gene ontology annotation, gene–gene interaction databases, pathways, disease related gene networks and systems, as shown in **Table 1**. This information can greatly assist GWAS to find epistatic interactions. Many recent studies have used such biological knowledge and databases for filtering in their interaction studies. The databases have helped identify biological pair-wise interactions among SNPs in pathways, and hence new associations and potential drug targets. For example, Liu et al. (2012) generated genome-wide SNP pairs based on multiple biological pathways such as KEGG, STRING, T2DGADB, etc.

Biofilter is an analysis pipeline that catalogs biological information by integrating data from the Reactome, KEGG, GO, DIP, Pfam, Ensembl, and NetPath (Bush et al., 2009; Pendergrass et al., 2013b). It can build SNP–SNP models based on known interactions between genes and proteins in curated pathways and networks. Grady et al. (2011) utilized the Biofilter

Table 1 | Biological information databases on gene ontology annotation, gene–gene interactions, pathways, disease related gene networks and systems.

Database	URL	Description	Reference
KEGG	http://www.genome.jp/kegg/pathway.html	KEGG is a collection of manually drawn pathway maps representing knowledge on the molecular interaction and reaction networks for metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases, and drug development.	Kanehisa and Goto (2000)
GO	http://www.geneontology.org/	GO provides an ontology of defined terms representing gene product properties. The ontology covers three domains: cellular component, molecular function, and biological processes.	Ashburner et al. (2000)
DIP	http://dip.doe-mbi.ucla.edu/dip/	Databases of experimentally determined interactions between proteins.	Xenarios et al. (2000)
BioGRID	http://thebiogrid.org/	A comprehensive resource of protein–protein and genetic interactions for all major model organism species.	Stark et al. (2006)
NetPath	http://www.netpath.org/	Resource of signal transduction pathways in humans.	Kandasamy et al. (2010)
IntAct	http://www.ebi.ac.uk/intact/	Database of molecular interactions that are derived from literature curation or direct user submissions.	Orchard et al. (2014)
MINT	http://mint.bio.uniroma2.it/mint/	MINT focuses on experimentally verified protein–protein interactions mined from the scientific literature by expert curators. MINT now uses the IntAct database infrastructure to limit the duplication of efforts and to optimize future software development.	Chatr-aryamontri et al. (2007)
MIPS	http://mips.helmholtz-muenchen.de/proj/yeast/CYGD/interaction/	The MIPS mammalian protein–protein interaction Database is a collection of manually curated high-quality interactions.	Pagel et al. (2005)
Pfam	http://pfam.sanger.ac.uk/	The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models. There are two kinds of entries in Pfam: Pfam-A entries are high quality, manually curated families; Pfam-B entries have lower quality.	Punta et al. (2012)
STRING	http://string-db.org	A database of known and predicted protein interactions, including direct (physical) and indirect (functional) associations.	Szklarczyk et al. (2011)
MSigDB	http://www.broadinstitute.org/gsea/msigdb/	Molecular signatures database, a collection of annotated gene sets integrating canonical pathways representing biological processes.	Subramanian et al. (2005)
BioCarta	http://www.biocarta.com/genes/	Includes classical pathways as well as current suggestions for new pathways.	Nishimura (2001)
Reactome	http://www.reactome.org/PathwayBrowser/	The Reactome pathway database aims to provide intuitive bioinformatics tools for visualization, interpretation and analysis of pathway knowledge.	Croft et al. (2011)
T2DGADB	http://t2db.khu.ac.kr:8080/	A disease gene network database for type 2 diabetes.	Lim et al. (2010)

software to look for epistasis contributing to the risk of virologic failure. Approximately two million SNP–SNP interaction models were produced by Biofilter, and Grady et al. (2010) tested these models by using logistic regression via the software package PLATO. They identified interactions between SNPs in the TAP1 and ABCC9 genes. Pendergrass et al. (2013a) identified five significant GxG interactions associated with cataract using Biofilter. Bush et al. (2011) studied multiple sclerosis susceptibility with Biofilter, identifying gene–gene interactions of susceptibility loci involved in the central nervous system and neuron function. Turner et al. (2011) used Biofilter to detect associations with low density lipoprotein cholesterol level, identifying 11 significant GxG interactions, eight of which were replicated in a second cohort. In each of these examples, Biofilter generated biologically plausible gene–gene and SNP–SNP interaction models that were replicated in an independent study.

Some studies reduce the number of tests by performing a gene-based, as opposed to a SNP-based, interaction test. Baranzini et al. (2009) combined the SNP-wise *P*-values to form a gene-wise *P*-value for each gene (such as using the minimum *P*-value for the gene), and superimposed the gene-wise *P*-values on a human protein interaction network to identify sub-networks containing a higher proportion of genes associated with multiple sclerosis than expected by chance. Ma et al. (2013) tested interactions of SNP pairs that are separately located in two different genes as marker-based tests. To test the interaction between each pair of genes, they combined these marker-based interactions and the LD between markers into a gene-based statistic.

Knowledge-driven filtering approaches can test models of genes that participate in the same biological pathway or network, and the interpretation of the interactions is then more straightforward. But their precision and power are hard to validate by simulation. Because such approaches depend on prior knowledge, which may not be accurate or may not be applicable to a particular dataset, they may miss what could be important findings among the genes for which we have little knowledge.

Data-driven filtering

Filtering based on statistical tests is data-driven. Statistical data-driven filtering includes, apart from SNP quality control, single marker associations, feature selection to keep only the most informative markers, and statistical tests to screen for potential interactions. Using data-driven filtering in GWAS can dramatically decrease the search space used to find interactions, so that subsequent statistical tests and machine learning methods can be applied as an exhaustive search among a smaller number of SNPs. The performance of data-driven filtering depends on the assumptions that the statistical tests or filtering algorithms make. Single marker association filtering can only screen interactions among SNPs showing at least a moderate effect on the trait of interest, while feature selection filtering and variance heterogeneity filtering can be used to detect SNP interactions with very weak marginal SNP effects.

Filtering according to single marker association. Filtering SNPs based on their marginal effects is frequently used for a high-dimensional gene–gene interaction search. It is often combined with biological filtering to identify interactions among SNPs that are marginally associated with a phenotype (Baranzini et al., 2009; Grady et al., 2011; Turner et al., 2011; Ma et al., 2012; Pendergrass et al., 2013a). This approach follows the principles of hierarchical model building in the general linear model, where the interaction terms are tested only after all main-effect terms are deemed statistically significant. Typically the significance threshold used is less stringent than the usual genome-wide threshold of 5×10^{-8}. The advantage of this filtering is that it is easy to implement; its disadvantage is that it has low power for detecting interactions among low-marginal-effect SNPs.

Filtering by feature selection algorithms. Feature selection algorithms such as Relief (Kira and Rendell, 1992), ReliefF (Kononenko, 1994), Tuned ReliefF (TuRF; Moore and White, 2007), and Spatially Uniform ReliefF (SURF; Greene et al., 2009) can also be used. They screen pairs of diallelic SNPs that can cluster individuals with similar phenotypes, on the basis of the nine two-SNP genotypes, into two distinct classes (e.g., cases versus controls). For each individual only a small subset of neighboring individuals, i.e., individuals most similar to that individual over all the SNPs, is examined. Iterating over each individual and its chosen subset of neighboring individuals, SNPs are up-weighted for selection on the basis of belonging to the SNP pairs most frequently found in all such sets. Simulation results have indicated this is able to identify SNP pairs with purely non-additive effects in genome-wide datasets. Evaporative cooling (McKinney et al., 2007) is another feature selection approach which couples mutual information and thermodynamics theory. It filters SNPs by removing those with least information for epistatic interactions. Such feature selection filtering is able to retain pure epistatic (i.e., essential) interaction between markers with low-marginal effects, offering a powerful alternative to single-marker filtering.

Filtering by testing variance heterogeneity of phenotype among SNP genotypes. For a quantitative trait, the presence of gene–gene interactions will result in heterogeneity of the phenotype variances among the genotypes of a single SNP, and this heterogeneity of phenotype variance has been proposed as a screen to prioritize SNPs for interaction testing (Paré et al., 2010; Struchalin et al., 2010). SNPs selected on the basis of variance heterogeneity would then be used for later gene–gene or gene–environment interaction analyses. However, unless the phenotypic means are the same for all the SNP genotypes, a transformation corresponding to a non-linear change in the scale of measurement may equalize the variances (Sun et al., 2013). This transformation, if it can be found, would eliminate any interactions detected this way.

USING OPTIMAL SEARCH ALGORITHMS AND COMPUTATIONAL TECHNOLOGY TO SPEED A SCAN FOR INTERACTIONS

Exhaustive search of interactions among millions of SNPs in GWAS data is computationally time-consuming. However, heuristic stochastic searching algorithms and efficient computational

technology, such as parallel computing and bit operation, can boost the computational speed and, if maximization is involved, speed the convergence required to calculate test statistics. Some interaction studies use optimal searching and computational technology to search the whole space for potential interactions. An ultrafast genome-wide scan approach for SNP–SNP interactions, SIXPAC, employs a randomization searching algorithm – probability approximate complete (PAC) testing – to drastically trim the universe of SNP combinations. The approach samples small groups of cases and highlights combinations of alleles carried by all individuals in the group. By further incorporating bit operation technology, SIXPAC can scan genome-wide pair-wise interactions in a few hours, compared to PLINK in weeks (Prabhu and Pe'er, 2012).

Lu et al. (2012) developed a likelihood ratio-based Mann–Whitney approach that can test high-order interactions. It is computationally efficient and only conducts one test for all the identified interaction, so that no adjustment is necessary for multiple testing. A further extension of the approach introduces a randomizing algorithm into the scan, using ensemble tree models (Wei et al., 2013), to increase the computational efficiency and convergence precision.

Schüpbach et al. (2010) developed an efficient extension of the PLINK epistasis module by using a parallel computing algorithm running on multiple processors to increase the speed of an exhaustive scan of all SNP pairs.

Heuristic or randomized search is much more efficient than exhaustive search, so it can perform a genome-wide scan of interactions among millions of SNPs without any filtering in reasonable time. However, it cannot guarantee reaching the optimal solution, which means it may not find all the biologically relevant interactions.

CONCLUSION

Numerous approaches have been proposed for the analysis of epistatic interactions, each of which has advantages and disadvantages. Regression models are easy for model interpretation, but they are less suitable for modeling high-order interaction on a large number of markers. Model-free approaches do not give an explicit explanation of interaction findings, but they are good at detecting high dimensional non-linear interactions. Tests for interactions by contrasting LD between cases and controls or by studying phenotype variance heterogeneity among the different genotypes of a SNP, are two special tests for detecting epistasis in the absence of any main-effect.

With the emergence of massive amounts of genome sequencing data, developing efficient searching algorithms and filter pipelines are especially important. Heuristic searching is much faster than exhaustive searching, at the cost of missing some true positive results and finding more false positive results. Filtering pipelines based on biological knowledge have the advantage of providing a clearer biological explanation for the detected interactions, but the assumed knowledge may be limited and not error-free, in which case such filtering may also lead to testing some irrelevant interaction models and may miss novel and important signals. Data-driven filtering cleans the data by removing low quality and the least informative SNPs, but its performance depends on the underlying assumptions of the filter. Because statistical and biological filtering each has unique features, they should be viewed as complementary to, rather than as competing with, each other. Through novel approaches for filtering and modeling GxG interactions, we may identify more of the missing heritability for common, complex traits.

ACKNOWLEDGMENTS

This work was supported by U.S. Public Health Service grant 1U01HG006382 and U01 HG006389 from the National Human Genome Research Institute, HL065962 which funds the PGRN Statistical Analysis Resource (P-STAR), and by the Korean Government grant NRF-2011-220-C00004 from the National Research Foundation of Korea.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Baranzini, S. E., Galwey, N. W., Wang, J., Khankhanian, P., Lindberg, R., Pelletier, D., et al. (2009). Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Genet.* 18, 2078–2090. doi: 10.1093/hmg/ddp120

Bateson, W. (1909). *Mendel's Principles of Heredity*. Cambridge: Cambridge University Press. doi: 10.5962/bhl.title.44575

Bush, W. S., Dudek, S. M., and Ritchie, M. D. (2009). Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac. Symp. Biocomput.* 2009, 368–379.

Bush, W. S., McCauley, J. L., DeJager, P. L., Dudek, S. M., Hafler, D. A., Gibson, R. A., et al. (2011). A knowledge-driven interaction analysis reveals potential neurodegenerative mechanism of multiple sclerosis susceptibility. *Genes Immun.* 12, 335–340. doi: 10.1038/gene.2011.3

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., et al. (2007). MINT, the molecular interaction database. *Nucleic Acids Res.* 35, D572–D574. doi: 10.1093/nar/gkl950

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., et al. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* 39, D691–D697. doi: 10.1093/nar/gkq1018

Culverhouse, R., Klein, T., and Shannon, W. (2004). Detecting epistatic interactions contributing to quantitative traits. *Genet. Epidemiol.* 27, 141–152. doi: 10.1002/gepi.20006

Elston, R. C. (1961). On additivity in the analysis of variance. *Biometrics* 17, 209–219. doi: 10.2307/2527987

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edin.* 52, 399–433. doi: 10.1017/S0080456800012163

Frankel, W. N., and Schork, N. J. (1996). Who's afraid of epistasis? *Nat. Genet.* 14, 371–373. doi: 10.1038/ng1296-371

Grady, B. J., Torstenson, E. S., Dudek, S. M., Giles, P., and Ritchie, M. D. (2010). Finding unique filter sets in PLATO: a precursor to efficient interaction analysis in GWAS data. *Pac. Symp. Biocomput.* 2010, 315–326.

Grady, B. J., Torstenson, E. S., McLaren, P. J., De Bakker, P. I., Haas, D. W., Robbins, G. K., et al. (2011). Use of biological knowledge to inform the analysis of gene-gene interactions involved in modulating virologic failure with efavirenz-containing treatment regimens in art-naïve ACTG clinical trials participants. *Pac. Symp. Biocomput.* 2011, 253–264.

Greene, C. S., Penrod, N. M., Kiralis, J., and Moore, J. H. (2009). Spatially uniform relieff (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData Min.* 2:5. doi: 10.1186/1756-0381-2-5

Greenland, S., Lash, T. L., and Rothman, K. J. (1998). "Concepts of interaction," in *Modern Epidemiology*, 2nd Edn, eds K. J. Rothman and S. Greenland (Philadelphia, PA: Lippincott-Raven), 71–86.

Hahn, L. W., Ritchie, M. D., and Moore, J. H. (2003). Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics* 19, 376–382. doi: 10.1093/bioinformatics/btf869

Kam-Thong, T., Czamara, D., Tsuda, K., Borgwardt, K., Lewis, C. M., Erhardt-Lehmann, A., et al. (2010). EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. *Eur. J. Hum. Genet.* 19, 465–471. doi: 10.1038/ejhg.2010.196

Kandasamy, K., Mohan, S. S., Raju, R., Keerthikumar, S., Kumar, G. S., Venugopal, A. K., et al. (2010). NetPath: a public resource of curated signal transduction pathways. *Genome Biol.* 11:R3. doi: 10.1186/gb-2010-11-1-r3

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kira, K., and Rendell, L. A. (1992). "The feature selection problem: traditional methods and a new algorithm," in *AAAI-92: Proceedings of the Tenth National Conference on Artificial Intelligence*, ed. D. W. R. Swartout (San Jose, CA: AAAI Press/The MIT Press), 129–134.

Kononenko, I. (1994). Estimating attributes: analysis and extensions of RELIEF. *Lect. Notes Comput. Sci.* 784, 171–182. doi: 10.1007/3-540-57868-4_57

Li, C., Li, Y., Xu, J., Lv, J., Ma, Y., Shao, T., et al. (2011). Disease-driven detection of differential inherited SNP modules from SNP network. *Gene* 489, 119–129. doi: 10.1016/j.gene.2011.08.026

Lim, J. E., Hong, K. W., Jin, H. S., Kim, Y. S., Park, H. K., and Oh, B. (2010). Type 2 diabetes genetic association database manually curated for the study design and odds ratio. *BMC. Med. Inform. Decis. Mak.* 10:76. doi: 10.1186/1472-6947-10-76

Liu, Y., Maxwell, S., Feng, T., Zhu, X., Elston, R. C., Koyutürk, M., et al. (2012). Gene, pathway and network frameworks to identify epistatic interactions of single nucleotide polymorphisms derived from GWAS data. *BMC Syst. Biol.* 6(Suppl. 3), S15. doi: 10.1186/1752-0509-6-S3-S15

Lu, Q., Wei, C., Ye, C., Li, M., and Elston, R. C. (2012). A likelihood ratio-based Mann-Whitney approach finds novel replicable joint gene action for type 2 diabetes. *Genet. Epidemiol.* 36, 583–593. doi: 10.1002/gepi.21651

Ma, L., Brautbar, A., Boerwinkle, E., Sing, C. F., Clark, A. G., and Keinan, A. (2012). Knowledge-driven analysis identifies a gene-gene interaction affecting high-density lipoprotein cholesterol levels in multi-ethnic populations. *PLoS Genet.* 8:e1002714. doi: 10.1371/journal.pgen.1002714

Ma, L., Clark, A. G., and Keinan, A. (2013). Gene-based testing of interactions in association studies of quantitative traits. *PLoS Genet.* 9:e1003321. doi: 10.1371/journal.pgen.1003321

Marchini, J., Donnelly, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* 37, 413–417. doi: 10.1038/ng1537

McKinney, B. A., Reif, D. M., White, B. C., Crowe, J. E. Jr., and Moore, J. H. (2007). Evaporative cooling feature selection for genotypic data involving interactions. *Bioinformatics* 23, 2113–2120. doi: 10.1093/bioinformatics/btm317

Moore, J. H., and White, B. C. (2007). Tuning reliefF for genome-wide genetic analysis. *Lect. Notes Comput. Sci.* 4447, 166–175. doi: 10.1007/978-3-540-71783-6_16

Nelson, M., Kardia, S., Ferrell, R., and Sing, C. (2001). A combinatorial partitioning approach to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res.* 11, 458–470. doi: 10.1101/gr.172901

Nishimura, D. (2001). BioCarta. *Biotech Software Internet Rep.* 2, 117–120. doi: 10.1089/152791601750294344

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., et al. (2014). The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42, D358–D363. doi: 10.1093/nar/gkt1115

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., et al. (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics* 21, 832–834. doi: 10.1093/bioinformatics/bti115

Paré, G., Cook, N. R., Ridker, P. M., and Chasman, D. I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. *PLoS Genet.* 6:e1000981. doi: 10.1371/journal.pgen.1000981

Pendergrass, S. A., Verma, S. S., Holzinger, E. R., Moore, C. B., Wallace, J., Dudek, S. M., et al. (2013a). Next-generation analysis of cataracts: determining knowledge driven gene-gene interactions using Biofilter, and gene-environment interactions using the PhenX Toolkit. *Pac. Symp. Biocomput.* 2013, 147–158. doi: 10.1142/9789814447973_0015

Pendergrass, S. A., Frase, A., Wallace, J., Wolfe, D., Katiyar, N., Moore, C., et al. (2013b). Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development. *BioData Min.* 6:25. doi: 10.1186/1756-0381-6-25

Prabhu, S., and Pe'er, I. (2012). Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. *Genome Res.* 22, 2230–2240. doi: 10.1101/gr.137885.112

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2012). The Pfam protein families database. *Nucleic Acids Res.* 40, D290–D301. doi: 10.1093/nar/gkr1065

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Ritchie, M. D., Hahn, L. W., and Moore, J. H. (2003). Power of multifactor dimensionality reduction for detecting gene–gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet. Epidemiol.* 24, 150–157. doi: 10.1002/gepi.10218

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., et al. (2001). Multifactor dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* 69, 138–147. doi: 10.1086/321276

Satagopan, J. M., and Elston R. C. (2012). Evaluation of removable statistical interaction for binary traits. *Stat. Med.* 32, 1164–1190. doi: 10.1002/sim.5628

Schüpbach, T., Xenarios, I., Bergmann, S., and Kapur, K. (2010). FastEpistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics* 26, 1468–1469. doi: 10.1093/bioinformatics/btq147

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic Acids Res.* 34, D535–D539. doi: 10.1093/nar/gkj109

Steen, K. V. (2012). Travelling the world of gene-gene interactions. *Brief. Bioinform.* 13, 1–19. doi: 10.1093/bib/bbr012

Struchalin, M. V., Dehghan, A., Witteman, J. C., van Duijn, C., and Aulchenko, Y. S. (2010). Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations. *BMC Genet.* 11:92. doi: 10.1186/1471-2156-11-92

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Sun, X., Elston, R., Morris, N., and Zhu, X. (2013). What is the significance of difference in phenotypic variability across SNP genotypes? *Am. J. Hum. Genet.* 93, 390–397. doi: 10.1016/j.ajhg.2013.06.017

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., et al. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* 39, D561–D568. doi: 10.1093/nar/gkq973

Turner, S. D., Berg, R. L., Linneman, J. G., Peissig, P. L., Crawford, D. C., Denny, J. C., et al. (2011). Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks. *PLoS ONE* 6:e19586. doi: 10.1371/journal.pone.0019586

Wang, X., Elston, R. C., and Zhu, X. (2010). Statistical interaction in human genetics: how should we model it if we are looking for biological interaction? *Nat. Rev. Genet.* 12, 74. doi: 10.1038/nrg2579-c2

Wei, C., Schaid, D. J., and Lu, Q. (2013). Trees Assembling Mann-Whitney approach for detecting genome-wide joint association among low-marginal-effect loci. *Genet. Epidemiol.* 37, 84–91. doi: 10.1002/gepi.21693

Wille, A., Hoh, J., and Ott, J. (2003). Sum statistics for the joint detection of multiple disease loci in case-control association studies with SNP markers. *Genet. Epidemiol.* 25, 350–359. doi: 10.1002/gepi.10263

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Res.* 28, 289–291. doi: 10.1093/nar/28.1.289

Yung, L. S., Yang, C., Wan, X., and Yu, W. (2011). GBOOST: a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. *Bioinformatics* 27, 1309–1310. doi: 10.1093/bioinformatics/btr114

Zhao, J., Jin, L., and Xiong, M. (2006). Test for interaction between two unlinked loci. *Am. J. Hum. Genet.* 79, 831–845. doi: 10.1086/508571

Zhu, Z., Tong, X., Zhu, Z., Liang, M., Cui, W., Su, K., et al. (2013). Development of GMDR-GPU for gene-gene interaction analysis and its application to WTCCC GWAS data for type 2 diabetes. *PLoS ONE* 8:e61943. doi: 10.1371/journal.pone.0061943

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 January 2014; accepted: 10 April 2014; published online: 30 April 2014.

Citation: Sun X, Lu Q, Mukherjee S, Crane PK, Elston R and Ritchie MD (2014) Analysis pipeline for the epistasis search – statistical versus biological filtering. Front. Genet. ***5****:106. doi: 10.3389/fgene.2014.00106*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

GENERAL COMMENTARY
published: 09 October 2014
doi: 10.3389/fgene.2014.00350

Corrigendum: Analysis pipeline for the epistasis search – statistical versus biological filtering

Shubhabrata Mukherjee*

Department of Medicine, University of Washington, Seattle, WA, USA
**Correspondence: smukherj@u.washington.edu*

Edited and reviewed by:
David B. Allison, University of Alabama at Birmingham, USA

Keywords: epistasis, genetic interaction, biological interaction, filtering pipeline, optimal search

A corrigendum on

Analysis pipeline for the epistasis search - statistical versus biological filtering
by Sun, X., Lu, Q., Mukheerjee, S., Crane, P. K., Elston, R., and Ritchie, M. D. (2014) Front Genet. 5:106. doi: 10.3389/fgene.2014.00106

My name was misspelled as Shubhabrata Mukheerjee. The correct spelling is Shubhabrata Mukherjee. The original article has been updated.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 09 July 2014; accepted: 18 September 2014; published online: 09 October 2014.
*Citation: Mukherjee S (2014) Corrigendum: Analysis pipeline for the epistasis search – statistical versus biological filtering. Front. Genet. **5**:350. doi: 10.3389/fgene.2014.00350*
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in
GENETICS

REVIEW ARTICLE
published: 18 March 2014
doi: 10.3389/fgene.2014.00051

Copy number variation analysis in the context of electronic medical records and large-scale genomics consortium efforts

John J. Connolly[1], Joseph T. Glessner[1,2], Berta Almoguera[1], David R. Crosslin[3], Gail P. Jarvik[3], Patrick M. Sleiman[1,2] and Hakon Hakonarson[1,2]*

[1] The Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, USA
[2] Department of Pediatrics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
[3] Departments of Medicine (Medical Genetics) and Genome Sciences, University of Washington Medical Center, Seattle, WA, USA

Edited by:
Marylyn D. Ritchie, Pennsylvania State University, USA

Reviewed by:
Lifeng Tian, University of Pennsylvania, USA
Scott Matthew Williams, Dartmouth College, USA

****Correspondence:***
Hakon Hakonarson, The Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
e-mail: hakonarson@chop.edu

The goal of this paper is to review recent research on copy number variations (CNVs) and their association with complex and rare diseases. In the latter part of this paper, we focus on how large biorepositories such as the electronic medical record and genomics (eMERGE) consortium may be best leveraged to systematically mine for potentially pathogenic CNVs, and we end with a discussion of how such variants might be reported back for inclusion in electronic medical records as part of medical history.

Keywords: CNV, copy number, structural variation, eMERGE, review

WHAT ARE COPY NUMBER VARIATIONS?

Copy number variations (CNVs) are deletions and duplications in the genome that vary in length from ∼50 base pairs to many megabases (50 base pair to 1 kilobase CNVs are typically considered indels). Events that cause CNVs include non-allelic homologous recombination, non-homologous end-joining, transposition of transposable elements, transposition of pseudogenes, variable numbers of tandem repeats, and replication errors following template-switching or fork stalling. CNVs are the primary mode by which an individual acquires a mutation, and occur at a rate of approximately 1.7×10^{-6} per locus as opposed to 1.8×10^{-8} for sequence variation (Lupski, 2007). Estimates of CNV frequency vary depending on the size of the structural variation classed as CNV – some estimates suggest that up to 12% of the genome may be variable in copy number, and that the cumulative result of CNV inheritance may constitute more than 10% of the human genome (Carter, 2007; Lupski et al., 2010). Recent studies suggest that the average human genome contains >1000 CNVs, covering approximately four million base pairs (Conrad et al., 2010; Mills et al., 2011), and occur at a rate of 0.07–0.12 per generation (Cordaux and Batzer, 2009; Itsara et al., 2010; Beck et al., 2011; Malhotra and Sebat, 2012). The Database of Genomic Variation (DGV)[1] currently lists over 100,000 published, unique, CNVs across the genome. While the majority continues to be benign, an increasing number of CNVs have been associated with disease susceptibility. Common functional consequences of CNVs typically demonstrate gene dose effect and include truncated protein sequences, eliminated/reduced protein expression (typically the result of deletions), or increased protein expression (typically caused by duplications).

[1]http://dgv.tcag.ca/dgv/app/home

HOW ARE COPY NUMBER VARIATIONS IDENTIFIED?

ARRAY-BASED APPROACHES

A range of approaches are available for detecting CNVs (**Figure 1**). The most common methods rely on computational methods, which leverage signals from genotyping and sequencing to infer CNVs. For example, large chromosomal anomalies can be detected through log R ratio (LRR) and B-allele frequency (BAF), data routinely generated and provided with single nucleotide polymorphism (SNP) and exome microarrays (e.g., **Figure 2**). For replication and validation, quantitative PCR – which compares the threshold cycles of a target versus reference sequence – is still widely deployed. In a similar vein, paralogs-ratio testing and molecular copy number counting are also used for validation.

For high-throughput CNV detection, the most common platforms are genome hybridization (CGH) arrays, genome-wide association (GWA) arrays, and second-generation sequencing (SGS). CGH arrays use artificial bacterial chromosomes or long synthetic oligonucleotides to probe either specific regions of interest or the entire genome (Greshock et al., 2007; Haraksingh et al., 2011). While this method has relatively low spatial resolution (typically >5–10 Mb; Kallioniemi et al., 1993) and requires a relatively large volume of DNA, CGH does offer high sensitivity and specificity (Greshock et al., 2007; Haraksingh et al., 2011), which is critical in a diagnostic context.

Single nucleotide polymorphism (SNP) arrays are more commonly used for CNV analysis, and CNVs can be identified from

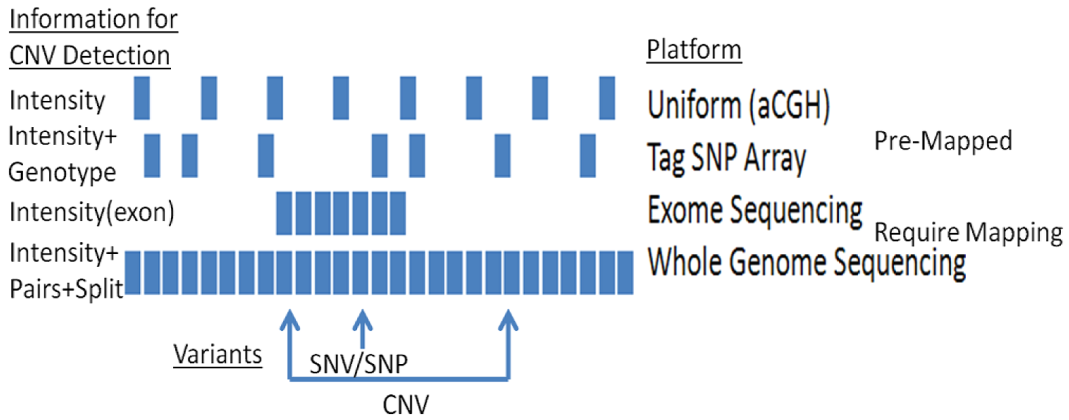


FIGURE 1 | CNV detection using different platforms: platforms vary in their capacities to detect CNVs.

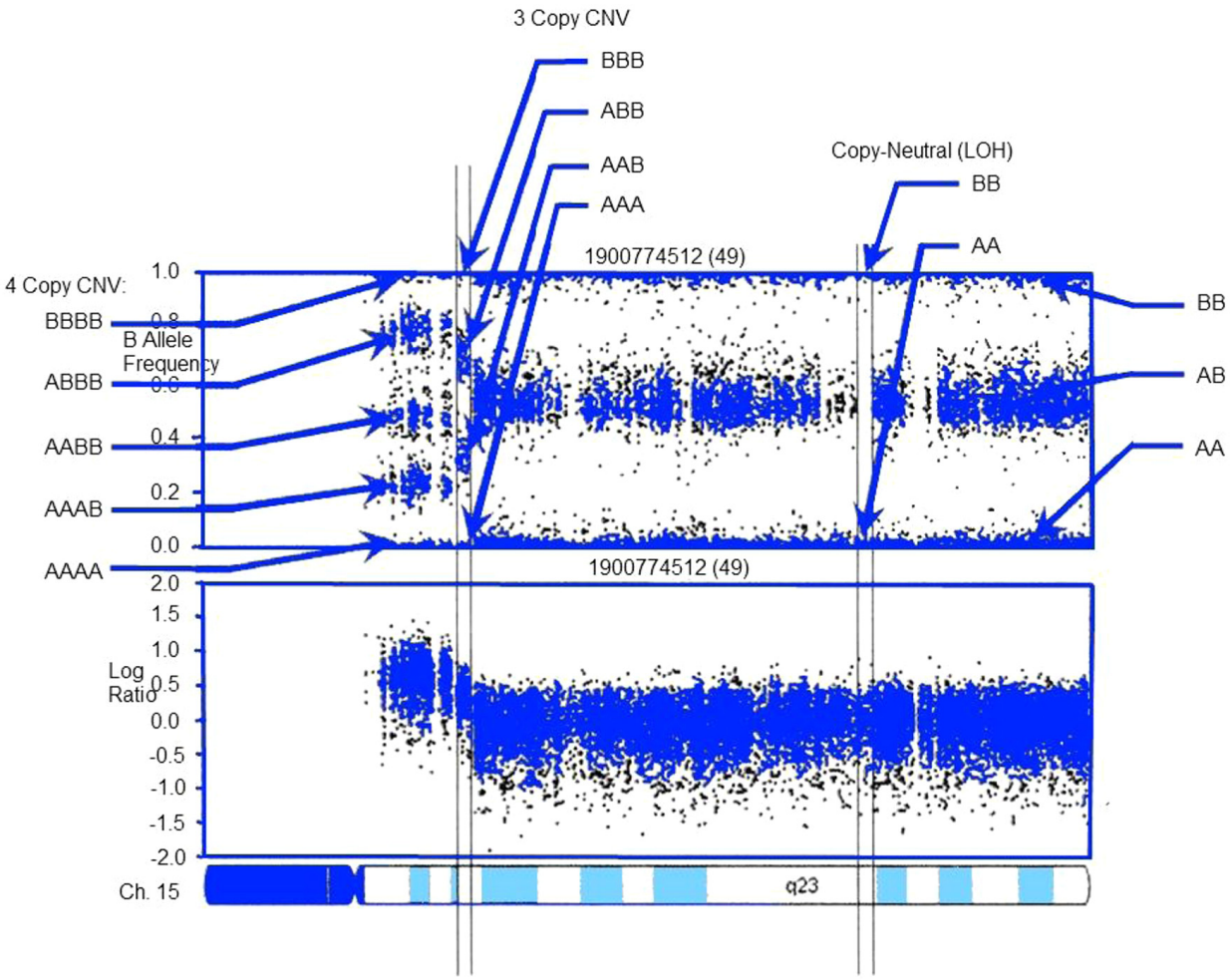


FIGURE 2 | CNV detection in SNP-array data using PennCNV: example log R ratio (LRR) and B Allele Freq (BAF) values for the chromosome 15 q-arm of an individual. Three normal chromosomal BAF genotype clusters (AA, AB, and BB genotypes) have LRR values around zero. The copy-neutral loss-of-heterozygosity (LOH) region has normal LRR values, but no AB cluster. Increased copy number can be observed in the increased number of peaks in the BAF distribution and increased LRR values. LRR and BAF patterns are different for different CNV regions, and can be used to generate CNV calls. Adapted from Wang et al. (2007).

standard GWA array signals, or from arrays that utilize custom probes. Custom probes offer greater coverage of non-SNP sites, and can offer high sensitivity, particularly with regard to breakpoint resolution (Haraksingh et al., 2011). While conventional (i.e., non-custom) SNP arrays offer less specificity, they nevertheless represent a cost-effective option for characterizing CNVs and have been successfully applied to a wide range of phenotypes to date (Connolly and Hakonarson, 2012).

Importantly, it is possible to retroactively characterize CNVs from existing genome-wide association study (GWAS) data. In this context, the observed SNP signal of an allele relative to the normalized intensity of the allele can be used to deduce a deletion (decreased intensity) or duplication (increased intensity; Glessner et al., 2012). This possibility constitutes a major opportunity for custodians of large biorepositories such as electronic medical record and genomics (eMERGE), where a large volume of GWAS data has already been generated. Since its founding in 2007, the eMERGE consortium has produced dozens of GWASs on a range of phenotypes including lipids (Rasmussen-Torvik et al., 2012), arrhythmia (Ritchie et al., 2013), and white blood cell count (Crosslin et al., 2012) to name a few. For many of these phenotypes, no CNV studies have been published to date. This, we believe, represents an opportunity to identify new disease-associated loci without the generation of new genotype data, and will be addressed by the consortium in the immediate future. Similarly, we note that a large number of studies listed in the NHGRI GWAS catalog[2] do not have complementary CNV data, suggesting a largely under-utilized resource.

For array-based analyses, a range of packages are available. Both Affymetrix and Illumina – the two primary purveyors of SNP arrays – offer free software packages for CNV analysis. Independently developed toolsets are also available. These include circular binding segmentation (Olshen et al., 2004) MixHMM (Liu et al., 2010), GADA (Pique-Regi et al., 2008), PennCNV (**Figure 2**; Wang et al., 2007), and ParseCNV (Glessner et al., 2013a; the latter two were developed by eMERGE researchers and are widely used).

SEQUENCING-BASED APPROACHES

Common CNVs are well-covered by SNPs in existing arrays (Conrad et al., 2010; Wellcome Trust Consortium et al., 2010). However, a resequencing study by Pang et al. (2010) suggests that coverage of rare CNVs may be less comprehensive. The authors identified over 12,000 structural variants in 4,867 genes across 40 + mb of sequence (the Venter genome), which had been initially unreported. More than 24% of these CNVs would not have been imputed by SNP-association. Given that rare alleles can have large effect sizes and a high penetrance, these results underline the limitations of SNP arrays to identify certain pathogenic CNVs. SGS, which is far more proficient at identifying rare CNVs, offers an attractive solution in this regard – particularly in identifying novel insertions absent in the reference genome. This has obvious clinical utility. SGS also confers a number of other critical advantages in terms of ability to identify smaller CNVs (<50 bp), and an enhanced capability for detecting breakpoints (Li and Olivier, 2013). Indeed, because SGS allows us to probe breakpoints at the level of base pairs, it facilitates capture of the signature of potential mutational mechanisms (Li and Olivier, 2013).

With SGS data, the most common methods for CNV identification from short-read analysis (Medvedev et al., 2010) are read-depth analysis (Xie and Tammi, 2009; Yoon et al., 2009; Abyzov et al., 2011), split-read mapping (Mills et al., 2006), paired-end read mapping (Korbel et al., 2009), and clone-based sequencing (Kidd et al., 2008). For all approaches, the most important determinants of accuracy are alignment and read-length. The average length of (reliable) reads is ∼ from 100 to 150 bp, which is insufficient to eliminate erroneous mapping. As this metric improves, CNV-calling algorithms will become more accurate.

A large number of algorithms have been developed for indentifying CNVs from sequencing data, including CNVnator (Abyzov et al., 2011), PennCNV-Seq (in press), GenomeStrip (Handsaker et al., 2011), cnvHiTSeq (Bellos et al., 2012), and XHMM (Fromer et al., 2012). Different CNV algorithms have different strengths and weaknesses (see Li and Olivier, 2013 for review), and the most effective strategy in terms of minimizing erroneous CNV calls is to incorporate multiple toolsets, which can be validated computationally via local *de novo* assembly (e.g., see SVMerge, Wong et al., 2010).

DISEASE-ASSOCIATED COPY NUMBER VARIATIONS

As discussed elsewhere in this issue, GWASs have been successful in identifying common risk variants, particularly where the frequency of such variants is >5%. In addition to common variants, certain disorders have been shown to be enriched for rare CNVs (Conrad et al., 2010; Pang et al., 2010). In terms of functional impact, CNVs have been shown to be enriched in genes involved in immune responses, cell–cell signaling, and retrovirus- and transposition-related protein coding (Li and Olivier, 2013). A large number of phenotypes have now been associated with CNVs, including several rare diseases (Matsuura et al., 1997) and a range of neurodevelopmental disorders (Glessner et al., 2012), including depression (Glessner et al., 2010c), schizophrenia (Glessner et al., 2010b), and autism (Glessner et al., 2009). Autism provides a particularly good example of how our understanding of genetic risk factors and etiology is enhanced by CNV research, as demonstrated by a recent exome sequencing study (Iossifov et al., 2012) involving 343 families from the Simons Simplex Collection.

The study identified 59 "likely gene disruptions (LGDs)" in autism cases. Interestingly, the 59-strong LGD shared overlapped strongly with a set of 842 proteins that interact with the fragile X protein, FMRP. In total, 14 of the 59 LGDs encoded FMRP-interacting proteins ($P = 0.006$), as did 13 of 72 CNV candidates from the group's previous CNV paper ($P = 0.0004$). Thus, 26 of 129 candidates were FMRP-related ($P < 1 \times 10^{-13}$). These results mark the fragile X mental retardation 1 (*FMR1*) gene as a high-profile autism candidate. Screening upstream targets of *FMR1*, the same group identified a deletion in *GRM5* that removes a single amino acid, causing an additional substitution at the same site. *GRM5* encodes the glutamate receptor mGluR5 (Bear et al., 2004),

[2]http://www.genome.gov/gwastudies/

which has been proposed as translational target in both ASD and ADHD (Elia et al., 2012; Silverman et al., 2012).

Several other CNV studies of autism have uncovered rare recurrent CNVs that have been informative. Our laboratory recently identified a range of CNVs in two major gene networks, ubiquitins and neuronal cell adhesion molecules that predispose to autism (Glessner et al., 2009). The ubiquitin–proteasome system is known to operate at pre- and post-synapses, and mediate neurotransmitter release, recycling of synaptic vesicles in pre-synaptic terminals, and modulating changes in dendritic spines and post-synaptic density (Yi and Ehlers, 2005). Neuronal cell adhesion molecules contribute to neurodevelopment by facilitating axon guidance, synapse formation and plasticity, and neuron–glial interactions.

Results from these and several other CNV studies suggest that genomic hotspots may be particularly vulnerable, which for autism include loci on chromosomes 1q21, 3p26, 15q11–q13, 16p11, and 22q11 (Bucan et al., 2009; Glessner et al., 2009; Pinto et al., 2010). Interestingly, these hotspots are part of large gene networks that are important to neural signaling and neurodevelopment, and have additionally been associated with other neuropsychiatric disorders. For example, studies of schizophrenia have highlighted structural mutations incorporating chromosomes 1q21, 15q13, and 22q11 (Glessner et al., 2010b). From an etiological perspective, autism and schizophrenia seem extremely different and it would seem counter-intuitive that associated loci should overlap. Some authors have addressed this peculiarity by proposing that the two disorders may in fact be opposite poles of the same spectrum (Crespi and Badcock, 2008). While such propositions await confirmation, they do highlight the potential of CNV studies to generate new hypotheses about the nature of complex diseases. Although individual structural variants explain relatively little by way of genetic variance, their cumulative is likely to be considerable. For autism, Marshall et al. (2008) suggested that CNVs play a causal role in 7% cases.

Beyond neuropsychiatric diseases, CNV studies have been published across a range of disease types, including heart disease (Goldmuntz et al., 2011), obesity (Glessner et al., 2010a), and cancer (Kuusisto et al., 2013). They have also recently been implicated in altered lifespan through alternative splicing mechanism (Glessner et al., 2013b).

COPY NUMBER VARIATIONS IN THE CONTEXT OF THE EMERGE CONSORTIUM

As illustrated in **Table 1**, the eMERGE consortium biorepository includes ∼60,000 individuals that have been genotyped on high-density GWA arrays[3], all of which have been linked with electronic medical records (EMRs). The size and diversity of the repository is such that it invokes the possibility for deep mining of disease-associated variants across multiple phenotypes. It is inevitable that a reasonable proportion of these individuals have disease-associated CNVs, and a larger proportion may be carriers of structural variants in recessive disease genes. By systematically characterizing CNVs across the biorepository, we have a very obvious opportunity to catalog CNVs and their disease-burden status. We have now run PennCNV on eMERGE Phase I data (2007–2011), and will soon have circular binary segmentation analyses complete for the same set (50-kb to whole-chromosome). Relevant analyses will play a major role in the consortium's Phase II genomics program (2012–2015).

Similarly, the eMERGE consortium recently embarked upon a large-scale pharmacogenomics project [$n = \sim$9000, review at Rasmussen-Torvik et al. (2012) in this issue], featuring a targeted sequencing platform developed by the Pharmacogenomics Research Network (PGRN), and covering 84 genes considered important for drug–gene interactions[4]. While the primary purpose of this project is to screen for existing pathogenic variants, this does offer an important opportunity to probe for novel variants in existing candidate genes, and to return results to patients' medical records. This clearly cannot be accomplished without paying heed to extensive medical, psychological, and ethical considerations, which are addressed elsewhere in this issue and in previous literature (Green et al., 2013). Assuming, however, that such considerations are adequately addressed, the section below considers how this might be accomplished and the potential to impact clinical care.

INTEGRATING CNVs WITH MEDICAL RECORDS – WHAT ARE THE OBSTACLES?

As discussed at length in this issue, the possibility of linking genomics data with EMRs represents a potentially major healthcare opportunity. What variants/results and how to report them remains open to debate, and indeed part of the remit of the eMERGE consortium is to think through these hurdles.

An obvious first step is determining the pathogenicity of relevant CNVs. Traditionally (e.g., cytogenetics), interpretation of CNVs has concentrated on diseases where the mode of inheritance was dominant, and relied on simple case–control comparisons to discriminate pathogenic from non-pathogenic variations. Where the CNV was common (i.e., frequency >1–5%), it was typically classed as non-pathogenic. Thus, by process, "rare" implied "pathogenic." With SGS and the increased capacity to detect smaller CNVs, this assumption falls down to a certain extent. We have started to see numerous studies where control and case *de novo* rate of small CNVs is as high as 5–10%. For rare CNVs in complex diseases, there is often insufficient power on which to base a judgment. Public databases that catalog pathogenic and non-pathogenic CNVs are therefore critical to determining frequencies of CNVs in disease cases and healthy controls.

Perhaps the most widely used catalog is the DGV, which aims to provide a "comprehensive summary of structural variation in the human genome" based on peer-review of relevant studies. While the DGV has obvious clinical and research relevance, several recent commentaries (Duclos et al., 2011; Hehir-Kwa et al., 2013) have urged caution in relying too heavily on its frequency and mapping statistics. As highlighted by Lee et al. (2007), many CNVs in the DGV are derived from single platforms/technologies, which may not necessarily translate to alternate approaches. Several recent studies (Perry et al., 2008; Conrad et al., 2010) suggest

[3]http://www.genome.gov/27540473

[4]www.pgrn.org

Table 1 | Summary of biorepositories and electronic medical records (EMRs) at 10 eMERGE-Institutions. Adapted from Gottesman et al. (2013).

Institution	Biorepository	Recruitment model	Biorepository size	Race/ethnicity and age of donors
Boston Children's Hospital	Gene Partnership	Outpatient and hospital-based	3,372	83% European 9% African 6% Asian 11% Hispanic/Latino Mean age: 23 years
Children's Hospital of Philadelphia	A Study of the Genetic Causes of Complex Pediatric Disorders	Population-based and disease-specific	60,000 internal (plus 100,000 external)	47.0% European 43.3% African 7.0% Admixed 1.7% Asian 0.8% Hispanic 0.2% Native Amer. Mean age: 11 years
Cincinnati Children's Hospital	Better Outcomes for Children	Outpatient and hospital-based	8,472	73% European 10% African Mean age: 9 years
Geisinger Clinic	MyCode®	Population-based and disease-specific	35,000	98% European Age: < 89 years
Group Health Seattle	ACT Study; Alzheimer's Disease Patient Registry (ADPR); Northwest Institute of Genetic Medicine (NWIGM)	Disease-specific and HMO-based	5,859	92% European Age: > 50 years
Marshfield Clinic Research Foundation	Personalized Medicine Research Project	Population-based	20,000	98% European Mean age: 48 years
Mayo Clinic	Vascular disease biorepository (VDB); Mayo Clinic Biobank; other disease-specific	Outpatient-based	36,000	97% European Mean age: 63 years
Mount Sinai School of Medicine	Bio*Me*™, The Charles Bronfman Institute for Personalized Medicine Biobank Program	Outpatient and hospital-based	25,000	40% Hispanic/Latino 25% African 25% European
Northwestern University	NUgene	Outpatient and hospital-based	12,000	9% Hispanic/Latino 12% African 78% European Mean age: 48 years
Vanderbilt University	BioVU	Outpatient and hospital-based	155,000	2% Hispanic/Latino 15% African 80% European Mean age: 49 years

that because of relatively low resolution in some studies, the size of relevant CNVs may be smaller than outlined in the DGV. Duclos et al. (2011) drew similar conclusions, stressing the "urgent need to validate the frequencies and boundaries of the CNVs recorded in the DGV." This conclusion is based on the groups finding that some of the recorded CNVs are erroneously listed as polymorphic, which, if implemented in a medical setting may led to a deleterious CNV being called benign. Alternate CNV databases (e.g., dbVar; Lappalainen et al., 2013) have been established, but all are restrained by the quality of data on which they are based.

Other obstacles that have hampered development of CNV databases are inconsistent annotation of genomic data across studies, ill-defined curation protocols (e.g., QC-reporting, CNV-calling parameters), and incomplete phenotypic data. In each case, there is potential for consortium-led efforts to delineate best practices. To address the challenge of incomplete phenotypes, there is a particular opportunity for the eMERGE network. The majority of individuals enrolled in the eMERGE repository have their longitudinal EMRs linked to their genotype. This affords far greater potential for determining pathogenicity than traditional case–control studies, where controls may be categorized as lacking a specific disease state, with no other phenotype data. Completeness-of-EMR is critical in this regard. For patients enrolled in the biorepository at The Children's Hospital of Philadelphia, the mean duration of EMRs is ∼5.5 years, and is similar across other eMERGE sites. Relevant data include all ICD-9 diagnoses, lab values, procedures, and medications. Data of this length and depth should be considered minimal requirements for addressing pathogenicity on a large scale, while supplementation with disease-specific measures is also highly desirable.

Another major challenge in returning CNV data to patients' EMR concerns the nature of inheritance. An interesting study by Boone et al. (2013) recently sought to determine the rate of CNVs in recessive disease genes. The group used CGH to characterize deletion CNVs in 21,470 individual, identifying 3,212 heterozygous potential carrier deletions in 419 unique disease-associated genes. While many of these CNVs are likely benign polymorphisms, the group identified 206 heterozygous CNVs in multiple recessive genes, spanning 2–6 genes in each deletion. These CNVs, therefore, confer carrier status for multiple recessive conditions. Similarly, 307 individuals had multiple deletions in recessive disease genes. While many of these gene pairs

have unrelated function, a non-trivial proportion belongs to a shared pathway. Indeed, one participant had a CNV spanning three recessive immune genes *PSMB8*, *TAP1*, and *TAP2*, which are associated with autoinflammation, lipodystrophy, dermatosis syndrome (*PSMB8*), and type I bare lymphocyte syndrome (*TAP1* and *TAP2*). He also had a CNV in *CD19*, mutations of which are associated with common variable immunodeficiency. The authors were unable to determine whether the individual had a compromised immune system or presented with a history of immune disease (samples were anonymized). Nevertheless, he was clearly a multiple-deletion carrier, as were ∼1.5% of the cohort: such information may be of direct clinical relevance to individuals' offspring – whether this should be shared remains open to debate.

Inherited CNVs pose a similar set of problems. While the majority of inherited CNVs may be in loci that lead to recessive disorders, this is not always the case. Indeed, one of the best-known CNVs is duplication at 15q11–q13, which accounts for up to 3% of autism cases (Sebat et al., 2007; Marshall et al., 2008). A complex scenario was recently described by Knijnenburg et al. (2009), where a child with a homozygous deletion in 15q13.3 (inherited from non-consanguineous, hemizygous carrier parents), resulted in hearing loss. Critically, if the CNV is a gain, three copies may have no phenotypic effect but four copies may have clinical consequences (Giorda et al., 2011). Conversely, when one parent carries a CNV loss in a recessive disease gene and the other parent carries a mutation in the same gene, this can result in compound heterozygosity in offspring (Hehir-Kwa et al., 2013; Paciorkowski et al., 2013). These findings stress the point that not only is the size, location, and direction of the CNV important, but so too is the number of copies. A range of other inheritance scenarios are reviewed by Hehir-Kwa et al. (2013), including X-linked CNVs (wide vary widely across individuals), and mosaic imbalances (Kousoulidou et al., 2013; may vary across an individual's cell types; Biesecker and Spinner, 2013; Forsberg et al., 2013).

Another point concerning CNV interpretation is the phenomenon of pleiotropy. As discussed above, a large proportion of reported recurrent CNVs have replicated *across* diseases (Cooper et al., 2011; Girirajan et al., 2011; Sahoo et al., 2011; Williams et al., 2011). Thus, the same microduplications at 1q21.1 have been associated with both autism and schizophrenia (Weiss et al., 2008; McCarthy et al., 2009). Relevant factors influencing the expressivity of this microduplication are a combination of environmental, epigenetic, and oligogenic (other modifier genes; Girirajan et al., 2010) factors. The precise mechanisms of causality that lead to a particular etiology are thus likely to be extremely complex, which calls into question what, if anything, might be reported in patients' EMRs. Such questions are the subject of ongoing debate (Fabsitz et al., 2010; Cassa et al., 2012), and are beyond the scope of this review. However, it is obvious that as genomic data becomes increasingly ubiquitous, we will require extensive guidelines in determining how CNV results should be interpreted and shared. For the same reason, it is critical that healthcare professionals receive adequate training and resources to understand and communicate test results.

Additionally, due to large numbers of cell divisions, CNVs, particularly deletions, can be acquired in the hematogenic progenitor cells. We have previously shown that acquired mosaicism increases with age and can be associated with hematological disorders (Laurie et al., 2012; Schick et al., 2013). However, when analyzing CNVs associated with neurological disorders, such acquired CNVs must be distinguished from germline mutations that are represented in non-hematological tissues, such as brain.

CONCLUSION

To date, a large number of diseases, across a large range of fields, have been associated with CNVs. We are still in our relative infancy in terms of deciding-upon the pathogenicity of such structural variants. We have stressed the need for a large, publicly accessible, and curated repository where CNVs that have been validated across platforms and technologies are stored. Whether this repository stems from improving existing catalogs or is developed *ab initio* remains to be determined, but the necessity of such a resource is compelling. Several eMERGE-led projects could funnel directly into such a repository, which would have real potential to impact healthcare.

A number of obstacles have stymied result-sharing – difficulties identifying CNVs (particularly in regions enriched for repetitive content), a shortage of standards, and the nature of CNV disease burden. These problems have attracted much attention in the past several years, and are well-characterized. While there is general agreement that such obstacles are substantial, there is a similar degree of optimism that benefits to be derived from solving these problems far outweigh the costs required. Again, consortium-led initiatives will likely be the most effective platforms for standardizing CNV-calling algorithms and developing guidelines for clinical care. The time is ripe for such initiatives, and we expect to see CNV-driven research make a major impact in clinical care in the next decade.

ACKNOWLEDGMENTS

This work was funded by institutional support from the Children's Hospital of Philadelphia and the National Human Genome Research Institute (# U01HG006830 and U01HG006375).

REFERENCES

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 21, 974–984. doi: 10.1101/gr.114876.110

Bear, M. F., Huber, K. M., and Warren, S. T. (2004). The mGluR theory of fragile X mental retardation. *Trends Neurosci.* 27, 370–377. doi: 10.1016/j.tins.2004.04.009

Beck, C. R., Garcia-Perez, J. L., Badge, R. M., and Moran, J. V. (2011). LINE-1 elements in structural variation and disease. *Annu. Rev. Genomics Hum. Genet.* 12, 187–215. doi: 10.1146/annurev-genom-082509-141802

Bellos, E., Johnson, M. R., and Coin, L. J. M. (2012). cnvHiTSeq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data. *Genome Biol.* 13, R120. doi: 10.1186/gb-2012-13-12-r120

Biesecker, L. G., and Spinner, N. B. (2013). A genomic view of mosaicism and human disease. *Nat. Rev. Genet.* 14, 307–320. doi: 10.1038/nrg3424

Boone, P. M., Campbell, I. M., Baggett, B. C., Soens, Z. T., Rao, M. M., Hixson, P. M., et al. (2013). Deletions of recessive disease genes: CNV contribution to carrier states and disease-causing alleles. *Genome Res.* 23, 1383–1394. doi: 10.1101/gr.156075.113

Bucan, M., Abrahams, B. S., Wang, K., Glessner, J. T., Herman, E. I., Sonnenblick, L. I., et al. (2009). Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. *PLoS Genet.* 5:e1000536. doi: 10.1371/journal.pgen.1000536

Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.* 39, S16–S21. doi: 10.1038/ng2028

Cassa, C. A., Savage, S. K., Taylor, P. L., Green, R. C., McGuire, A. L., Mandl, K. D., et al. (2012). Disclosing pathogenic genetic variants to research participants: quantifying an emerging ethical responsibility. *Genome Res.* 22, 421–428. doi: 10.1101/gr.127845.111

Connolly, J. J., and Hakonarson, H. (2012). The impact of genomics on pediatric research and medicine. *Pediatrics* 129, 1150–1160. doi: 10.1542/peds.2011-3636

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., et al. (2010). Origins and functional impact of copy number variation in the human genome. *Nature* 464, 704–712. doi: 10.1038/nature08516

Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., et al. (2011). A copy number variation morbidity map of developmental delay. *Nat. Genet.* 43, 838–846. doi: 10.1038/ng.909

Cordaux, R., and Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* 10, 691–703. doi: 10.1038/nrg2640

Crespi, B., and Badcock, C. (2008). Psychosis and autism as diametrical disorders of the social brain. *Behav. Brain Sci.* 31, 241–261; discussion 261–320. doi: 10.1017/S0140525X08004214

Crosslin, D. R., McDavid, A., Weston, N., Nelson, S. C., Zheng, X., Hart, E., et al. (2012). Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE network. *Hum. Genet.* 131, 639–652. doi: 10.1007/s00439-011-1103-9

Duclos, A., Charbonnier, F., Chambon, P., Latouche, J. B., Blavier, A., Redon, R., et al. (2011). Pitfalls in the use of DGV for CNV interpretation. *Am. J. Med. Genet. A* 155, 2593–2596. doi: 10.1002/ajmg.a.34195

Elia, J., Glessner, J. T., Wang, K., Takahashi, N., Shtir, C. J., Hadley, D., et al. (2012). Genome-wide copy number variation study associates metabotropic glutamate receptor gene networks with attention deficit hyperactivity disorder. *Nat. Genet.* 44, 78–84. doi: 10.1038/ng.1013

Fabsitz, R. R., Fabsitz, R. R., McGuire, A., Sharp, R. R., Puggal, M., Beskow, L. M., et al. (2010). Ethical and practical guidelines for reporting genetic research results to study participants: updated guidelines from a National Heart, Lung, and Blood Institute working group. *Circ. Cardiovasc. Genet.* 3, 574–580. doi: 10.1161/CIRCGENETICS.110.958827

Forsberg, L. A., Absher, D., and Dumanski, J. P. (2013). Republished: non-heritable genetics of human disease: spotlight on post-zygotic genetic variation acquired during lifetime. *Postgrad. Med. J.* 89, 417–426. doi: 10.1136/postgradmedj-2012-101322rep

Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., et al. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* 91, 597–607. doi: 10.1016/j.ajhg.2012.08.005

Giorda, R., Beri, S., Bonaglia, M. C., Spaccini, L., Scelsa, B., Manolakos, E., et al. (2011). Common structural features characterize interstitial intrachromosomal Xp and 18q triplications. *Am. J. Med. Genet. A* 155, 2681–2687. doi: 10.1002/ajmg.a.34248

Girirajan, S., Brkanac, Z., Coe, B. P., Baker, C., Vives, L., Vu, T. H., et al. (2011). Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genet.* 7:e1002334. doi: 10.1371/journal.pgen.1002334

Girirajan, S., Rosenfeld, J. A., Cooper, G. M., Antonacci, F., Siswara, P., Itsara, A., et al. (2010). A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat. Genet.* 42, 203–209. doi: 10.1038/ng.534

Glessner, J. T., Bradfield, J. P., Wang, K., Takahashi, N., Zhang, H., Sleiman, P. M., et al. (2010a). A genome-wide study reveals copy number variants exclusive to childhood obesity cases. *Am. J. Hum. Genet.* 87, 661–666. doi: 10.1016/j.ajhg.2010.09.014

Glessner, J. T., Reilly, M. P., and Hakonarson, H. (2010b). Strong synaptic transmission impact by copy number variations in schizophrenia. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10584–10859. doi: 10.1073/pnas.1000274107

Glessner, J. T., Wang, K., Sleiman, P. M., Zhang, H., Kim, C. E., Flory, J. H., et al. (2010c). Duplication of the SLIT3 locus on 5q35.1 predisposes to major depressive disorder. *PLoS ONE* 5:e15463. doi: 10.1371/journal.pone.0015463

Glessner, J. T., Connolly, J. J., and Hakonarson, H. (2012). Rare genomic deletions and duplications and their role in neurodevelopmental disorders. *Curr. Top. Behav. Neurosci.* 12, 345–360. doi: 10.1007/7854_2011_179

Glessner, J. T., Li, J., and Hakonarson, H. (2013a). ParseCNV integrative copy number variation association software with quality tracking. *Nucleic Acids Res.* 41, e64. doi: 10.1093/nar/gks1346

Glessner, J. T., Smith, A. V., Panossian, S., Kim, C. E., Takahashi, N., Thomas, K. A., et al. (2013b). Copy number variations in alternative splicing gene networks impact lifespan. *PLoS ONE* 8:e53846. doi: 10.1371/journal.pone.0053846

Glessner, J. T., Wang, K., Cai, G., Korvatska, O., Kim, C. E., Wood, S., et al. (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 459, 569–573. doi: 10.1038/nature07953

Goldmuntz, E., Paluru, P., Glessner, J., Hakonarson, H., Biegel, J. A., White, P. S., et al. (2011). Microdeletions and microduplications in patients with congenital heart disease and multiple congenital anomalies. *Congenit. Heart Dis.* 6, 592–602. doi: 10.1111/j.1747-0803.2011.00582.x

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., et al. (2013). The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genet. Med.* 15, 761–771. doi: 10.1038/gim.2013.72

Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., et al. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* 15, 565–574. doi: 10.1038/gim.2013.73

Greshock, J., Feng, B., Nogueira, C., Ivanova, E., Perna, I., Nathanson, K., et al. (2007). A comparison of DNA copy number profiling platforms. *Cancer Res.* 67, 10173–10180. doi: 10.1158/0008-5472.CAN-07-2102

Handsaker, R. E., Korn, J. M., Nemesh, J., and McCarroll, S. A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* 43, 269–276. doi: 10.1038/ng.768

Haraksingh, R. R., Abyzov, A., Gerstein, M., Urban, A. E., and Snyder, M. (2011). Genome-wide mapping of copy number variation in humans: comparative analysis of high resolution array platforms. *PLoS ONE* 6:e27859. doi: 10.1371/journal.pone.0027859

Hehir-Kwa, J., Pfundt, R., Veltman, J., and de Leeuw, N. (2013). Pathogenic or not? Assessing the clinical relevance of copy number variants. *Clin. Genet.* 84, 415–421. doi: 10.1111/cge.12242

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., et al. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron* 74, 285–299. doi: 10.1016/j.neuron.2012.04.009

Itsara, A., Wu, H., Smith, J. D., Nickerson, D. A., Romieu, I., London, S. J., et al. (2010). De novo rates and selection of large copy number variation. *Genome Res.* 20, 1469–1481. doi: 10.1101/gr.107680.110

Kallioniemi, O. P., Kallioniemi, A., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., et al. (1993). Comparative genomic hybridization: a rapid new method for detecting and mapping DNA amplification in tumors. *Semin. Cancer Biol.* 4, 41–46.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature* 453, 56–64. doi: 10.1038/nature06862

Knijnenburg, J., Oberstein, S. A., Frei, K., Lucas, T., Gijsbers, A. C., Ruivenkamp, C. A., et al. (2009). A homozygous deletion of a normal variation locus in a patient with hearing loss from non-consanguineous parents. *J. Med. Genet.* 46, 412–417. doi: 10.1136/jmg.2008.063685

Korbel, J. O., Abyzov, A., Mu, X. J., Carriero, N., Cayting, P., Zhang, Z., et al. (2009). PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.* 10, R23. doi: 10.1186/gb-2009-10-2-r23

Kousoulidou, L., Tanteles, G., Moutafi, M., Sismani, C., Patsalis, P. C., Anastasiadou, V., et al. (2013). 263.4 kb deletion within the TCF4 gene consistent with Pitt–Hopkins syndrome, inherited from a mosaic parent with normal phenotype. *Eur. J. Med. Genet.* 56, 314–318. doi: 10.1016/j.ejmg.2013.03.005

Kuusisto, K. M., Akinrinade, O., and Schleutker, J. (2013). Copy number variation analysis in familial BRCA1/2-negative Finnish breast and ovarian cancer. *PLoS ONE* 8:e71802. doi: 10.1371/journal.pone.0071802

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., et al. (2013). DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res.* 41, D936–D941. doi: 10.1093/nar/gks1213

Laurie, C. C., Laurie, C. A., Rice, K., Doheny, K. F., Zelnick, L. R., McHugh, C. P., et al. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.* 44, 642–650. doi: 10.1038/ng.2271

Lee, C., Iafrate, A. J., and Brothman, A. R. (2007). Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat. Genet.* 39, S48–S54. doi: 10.1038/ng2092

Li, W., and Olivier, M. (2013). Current analysis platforms and methods for detecting copy number variation. *Physiol. Genomics* 45, 1–16. doi: 10.1152/physiolgenomics.00082.2012

Liu, Z., Li, A., Schulz, V., Chen, M., and Tuck, D. (2010). MixHMM: inferring copy number variation and allelic imbalance using SNP arrays and tumor samples mixed with stromal cells. *PLoS ONE* 5:e10909. doi: 10.1371/journal.pone.0010909

Lupski, J. R. (2007). Genomic rearrangements and sporadic disease. *Nat. Genet.* 39, S43–S47. doi: 10.1038/ng2084

Lupski, J. R., Reid, J. G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D. C., Nazareth, L., et al. (2010). Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. *N. Engl. J. Med.* 362, 1181–1191. doi: 10.1056/NEJMoa0908094

Malhotra, D., and Sebat, J. (2012). CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* 148, 1223–1241. doi: 10.1016/j.cell.2012.02.039

Marshall, C. R., Noor, A., and Scherer, S. W. (2008). Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* 82, 477–488. doi: 10.1016/j.ajhg.2007.12.009

Matsuura, T., Sutcliffe, J. S., Fang, P., Galjaard, R. J., Jiang, Y. H., Benton, C. S., et al. (1997). De novo truncating mutations in E6-AP ubiquitin-protein ligase gene (UBE3A) in Angelman syndrome. *Nat. Genet.* 15, 74–77. doi: 10.1038/ng0197-74

McCarthy, S. E., Makarov, V., Kirov, G., Addington, A. M., McClellan, J., Yoon, S., et al. (2009). Microduplications of 16p11.2 are associated with schizophrenia. *Nat. Genet.* 41, 1223–1227. doi: 10.1038/ng.474

Medvedev, P., Fiume, M., Dzamba, M., Smith, T., and Brudno, M. (2010). Detecting copy number variation with mated short reads. *Genome Res.* 20, 1613–1622. doi: 10.1101/gr.106344.110

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., et al. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* 16, 1182–1190. doi: 10.1101/gr.4565806

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., et al. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature* 470, 59–65. doi: 10.1038/nature09708

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5, 557–572. doi: 10.1093/biostatistics/kxh008

Paciorkowski, A. R., Keppler-Noreuil, K., Robinson, L., Sullivan, C., Sajan, S., Christian, S. L., et al. (2013). Deletion 16p13.11 uncovers NDE1 mutations on the non-deleted homolog and extends the spectrum of severe microcephaly to include fetal brain disruption. *Am. J. Med. Genet. A* 161, 1523–1530. doi: 10.1002/ajmg.a.35969

Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., et al. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* 11, R52. doi: 10.1186/gb-2010-11-5-r52

Perry, G. H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C. W., et al. (2008). The fine-scale and complex architecture of human copy-number variation. *Am. J. Hum. Genet.* 82, 685–695. doi: 10.1016/j.ajhg.2007.12.010

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., et al. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 466, 368–372. doi: 10.1038/nature09146

Pique-Regi, R., Monso-Varona, J., Ortega, A., Seeger, R. C., Triche, T. J., Asgharzadeh, S., et al. (2008). Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics* 24, 309–318. doi: 10.1093/bioinformatics/btm601

Rasmussen-Torvik, L. J., Pacheco, J. A., Wilke, R. A., Thompson, W. K., Ritchie, M. D., Kho, A. N., et al. (2012). High density GWAS for LDL cholesterol in African Americans using electronic medical records reveals a strong protective variant in APOE. *Clin. Transl. Sci.* 5, 394–399. doi: 10.1111/j.1752-8062.2012.00446.x

Ritchie, M. D., Denny, J. C., Zuvich, R. L., Crawford, D. C., Schildcrout, J. S., Bastarache, L., et al. (2013). Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* 127, 1377–1385. doi: 10.1161/CIRCULATIONAHA.112.000604

Sahoo, T., Theisen, A., Rosenfeld, J. A., Lamb, A. N., Ravnan, J. B., Schultz, R. A., et al. (2011). Copy number variants of schizophrenia susceptibility loci are associated with a spectrum of speech and developmental delays and behavior problems. *Genet. Med.* 13, 868–880. doi: 10.1097/GIM.0b013e3182217a06

Schick, U. M., McDavid, A., Crane, P. K., Weston, N., Ehrlich, K., Newton, K. M., et al. (2013). Confirmation of the reported association of clonal chromosomal mosaicism with an increased risk of incident hematologic cancer. *PLoS ONE* 8:e59823. doi: 10.1371/journal.pone.0059823

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., et al. (2007). Strong association of de novo copy number mutations with autism. *Science* 316, 445–449. doi: 10.1126/science.1138659

Silverman, J. L., Smith, D. G., Rizzo, S. J., Karras, M. N., Turner, S. M., Tolu, S. S., et al. (2012). Negative allosteric modulation of the mGluR5 receptor reduces repetitive behaviors and rescues social deficits in mouse models of autism. *Sci. Transl. Med.* 4, 131ra51. doi: 10.1126/scitranslmed.3003501

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., et al. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* 17, 1665–1674. doi: 10.1101/gr.6861907

Weiss, L. A., Shen, Y., Korn, J. M., Arking, D. E., Miller, D. T., Fossdal, R., et al. (2008). Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* 358, 667–675. doi: 10.1056/NEJMoa075974

Wellcome Trust Consortium, Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., et al. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 464, 713–720. doi: 10.1038/nature08979

Williams, N. M., Franke, B., Mick, E., Anney, R. J., Freitag, C. M., Gill, M., et al. (2011). Genome-wide analysis of copy number variants in attention deficit hyperactivity disorder: the role of rare variants and duplications at 15q13.3. *Am. J. Psychiatry* 169, 195–204.

Wong K., Keane T. M., Stalker J., and Adams D. J. (2010). Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol.* 11:R128. doi: 10.1186/gb-2010-11-12-r128

Xie, C., and Tammi, M. T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* 10:80. doi: 10.1186/1471-2105-10-80

Yi, J. J., and Ehlers, M. D. (2005). Ubiquitin and protein turnover in synapse function. *Neuron* 47, 629–632. doi: 10.1016/j.neuron.2005.07.008

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* 19, 1586–1592. doi: 10.1101/gr.092981.109

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 13 December 2013; accepted: 18 February 2014; published online: 18 March 2014.

Citation: Connolly JJ, Glessner JT, Almoguera B, Crosslin DR, Jarvik GP, Sleiman PM and Hakonarson H (2014) Copy number variation analysis in the context of electronic medical records and large-scale genomics consortium efforts. Front. Genet. ***5****:51. doi: 10.3389/fgene.2014.00051*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in
GENETICS

METHODS ARTICLE
published: 12 February 2014
doi: 10.3389/fgene.2014.00016

The struggle to find reliable results in exome sequencing data: filtering out Mendelian errors

Zubin H. Patel[1,2†], Leah C. Kottyan[1,3†], Sara Lazaro[1,3], Marc S. Williams[4], David H. Ledbetter[4], Gerard Tromp[4], Andrew Rupert[5], Mojtaba Kohram[5], Michael Wagner[5], Ammar Husami[6], Yaping Qian[6], C. Alexander Valencia[6], Kejian Zhang[6], Margaret K. Hostetter[7], John B. Harley[1,3] and Kenneth M. Kaufman[1,3]*

[1] Division of Rheumatology, Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[2] Medical Scientist Training Program, University of Cincinnati College of Medicine, Cincinnati, OH, USA
[3] Department of Veterans Affairs, Veterans Affairs Medical Center – Cincinnati, Cincinnati, OH, USA
[4] Genomic Medicine Institute, Geisinger Health System, Danville, PA, USA
[5] Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[6] Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[7] Division of Infectious Disease, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Edited by:
Helena Kuivaniemi, Geisinger Health System, USA

Reviewed by:
Tatiana Foroud, Indiana University School of Medicine, USA
Goo Jun, University of Michigan, USA

****Correspondence:***
Kenneth M. Kaufman, Division of Rheumatology, Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229, USA
e-mail: kenneth.kaufman@cchmc.org

† Zubin H. Patel and Leah C. Kottyan have contributed equally to this work.

Next Generation Sequencing studies generate a large quantity of genetic data in a relatively cost and time efficient manner and provide an unprecedented opportunity to identify candidate causative variants that lead to disease phenotypes. A challenge to these studies is the generation of sequencing artifacts by current technologies. To identify and characterize the properties that distinguish false positive variants from true variants, we sequenced a child and both parents (one trio) using DNA isolated from three sources (blood, buccal cells, and saliva). The trio strategy allowed us to identify variants in the proband that could not have been inherited from the parents (Mendelian errors) and would most likely indicate sequencing artifacts. Quality control measurements were examined and three measurements were found to identify the greatest number of Mendelian errors. These included read depth, genotype quality score, and alternate allele ratio. Filtering the variants on these measurements removed ∼95% of the Mendelian errors while retaining 80% of the called variants. These filters were applied independently. After filtering, the concordance between identical samples isolated from different sources was 99.99% as compared to 87% before filtering. This high concordance suggests that different sources of DNA can be used in trio studies without affecting the ability to identify causative polymorphisms. To facilitate analysis of next generation sequencing data, we developed the Cincinnati Analytical Suite for Sequencing Informatics (CASSI) to store sequencing files, metadata (eg. relatedness information), file versioning, data filtering, variant annotation, and identify candidate causative polymorphisms that follow either *de novo*, rare recessive homozygous or compound heterozygous inheritance models. We conclude the data cleaning process improves the signal to noise ratio in terms of variants and facilitates the identification of candidate disease causative polymorphisms.

Keywords: whole exome sequencing, variant filtering, next-generation sequencing, disease causative polymorphisms, Mendelian errors, Mendel errors, CASSI

INTRODUCTION

Next-generation sequencing (NGS) has emerged as a powerful tool to investigate the genetic etiology of diseases. The use of NGS data has revolutionized clinical treatment and bench research. In general, the data generated in a NGS study are massive by comparison to that generated by a genome-wide genotyping array. In NGS, a fastq file of millions of short DNA sequences is generated for each sample. These fastq files are aligned to the reference genome using one of many different alignment tools. The alignment programs create a sequence alignment/map (SAM file) or a binary alignment/map (BAM file; Yu and Sun, 2013). It is widely appreciated that NGS generates a large number of sequencing errors. The extraordinary quantity of data generated even with a low error rate generates a large number of sequencing artifacts which will likely be called variants. This gives the appearance that NGS does not compare well with Sanger sequencing or genotyping arrays (4), but we show herein that the error rate of NGS of the called variants can be substantially reduced with the relative preservation of the vast majority of the data. To address the limitations imposed upon NGS studies by sequencing artifacts, we find refuge in redundancy. Typically, researchers obtain 40–200 reads of each base. Therefore, SAM and BAM files are large files and contain hundreds of millions of short sequences aligned to the reference genome. Variant callers such as the Genome Analysis Tool Kit (GATK) are used to generate a list of the variants in the variant call format (VCF; McKenna et al., 2010). VCF files contain meta-information for each variant

relative to a known reference genome sequence, as well as quality measurements for each subject's individual genotypes. These individual quality metrics include the overall number of reads at each position as well as the number and depth of alleles detected. In addition to predicting the nucleotide base or generating a base call calculated from a statistical algorithm, GATK also calculates a confidence score for the predicted nucleotide, the genotype quality score (Nielsen et al., 2011).

A multi-sample VCF file includes all of the genotypes for which at least one subject has a variant. Due to the flexibility of the format, the information contained in these files can vary widely. Furthermore, different variant callers are known to produce different calls (Rosenfeld et al., 2012; Liu et al., 2013). To further complicate matters, there is no currently agreed upon consensus to guide the analytical choices that are made when deciding which variant calls to include in a VCF file [as reviewed in Nekrutenko and Taylor (2012)]. One approach is to exclude (or filter) variants from a VCF file based on various criteria. These filters are based on a meta- or individual-sequencing parameter used to remove a particular variant. For example, variants can be filtered based upon the read depth (the number of the times the variant was detected), ratio of reads that contained the reference and alternate genotype calls (alt read), or by genotype quality scores. A recent comparison of the most common next-generation sequencing platforms and methodologies demonstrated that only 57% of the variants are common amongst five different pipelines using the same initial data (O'Rawe et al., 2013).

Typically a whole exome NGS experiment will generate ∼50–70 million bases of sequence. Greater than 99.99% of the bases match the reference genome. The remaining 0.01% of bases that differ from the reference genome is identified as variants. Importantly, most sequencing artifacts do not match the reference genome and are mis-identified as variants. Thus, identifying variants also has the effect of concentrating the sequencing artifacts. These sequence artifacts can be detected by identifying non-concordance of sequence from multiple assays of the same samples or as Mendelian errors if family data are available. Mendelian errors are genotypes that are found in the child that could not have been inherited from either parent.

After obtaining the data from a whole exome NGS experiment, analytical strategies range from identifying novel variants, to performing genetic association studies, to identifying variants that are candidates for potentially causing disease. Within the last 5 years, exome sequencing methods have been employed to successfully identify mutations in novel genes for a number of genetic conditions, including Sensenbrenner syndrome, Kabuki syndrome, and Miller syndrome (Gilissen et al., 2010; Ng et al., 2010a,b). One highly successful strategy uses the healthy parents of a patient with a severe disease to identify genetic variants in the patient that were not inherited, termed *de novo* variants. In fact, disruptive *de novo* variants appear to cause a substantial proportion of intellectual disability and many rare genetic disorders (Hoischen et al., 2010, 2011; Vissers et al., 2010; Bartnik et al., 2011; Filges et al., 2011; Gilman et al., 2011; Girard et al., 2011; Gonzalez-del Pozo et al., 2011; Paulussen et al., 2011; Xu et al., 2011; Bujakowska et al., 2012; Dauber et al., 2012; Harakalova et al., 2012; Iossifov et al., 2012; Lederer et al., 2012; Lin et al., 2012; Neale et al., 2012; Need et al., 2012; Neveling et al., 2012; O'Roak et al., 2012b; Riviere et al., 2012; Sanders et al., 2012; Santen et al., 2012; Schrier et al., 2012; Tsurusaki et al., 2012; Van Houdt et al., 2012; Whalen et al., 2012).

Using a trio study design (father, mother, and child) we can identify non-inherited variants in a child. These variants are sequencing errors, somatic mutations, or *de novo* mutations. We have used this analysis of trios as an opportunity to identify methodologies to filter the data to remove sequencing artifacts while retaining true mutations. In this study, we systematically assessed quality metrics to minimize Mendelian errors and identified a set of filters that remove these erroneous variants. These critical metrics are the depth of read (DP), the genotype quality score (GQ), and the alternate allele ratio. Filters based on these metrics were applied to a trio in which each family member was sequenced using three different sources of DNA (blood, saliva, and buccal cells). We tested the efficiency and specificity of our filters to remove sequencing artifacts by measuring the number of Mendelian errors and total variants removed by each filter singly and in combination. After testing the efficiency of our variant calling filters, we evaluated the filters on the concordance rate between identical samples from different DNA tissue sources. In order to make these analyses accessible to clinician researchers with limited command line programming experience, we have developed the Cincinnati Analytical Suite for Sequencing Informatics (CASSI) to seamlessly integrate the data storage, versioning, filtering, and annotation of NGS data through a web-based interface.

METHODS

DATA DESCRIPTION

We performed whole exome sequencing on a family trio. For this trio, three sources of DNA were obtained: blood, buccal, and saliva. Blood samples were collected from the three individuals using EDTA Vacutainer^{TM} Tubes (BD Franklin Lakes, New Jersey, USA). The buccal cells were collected by taking a cheek swab of each individual using the OGR-575 tubes from DNA-Genotek (Kanata, ON, Canada) and the saliva samples were collected by having each individual directly spit into the OGR-500 tube from DNA-Genotek. DNA was extracted using the DNeasy Blood and Tissue kit from Qiagen (Valencia, CA, USA). Each subject gave informed consent or assent approved by the institutional review board at Cincinnati Children's Hospital Medical Center. We studied all samples by exome capture using the Illumina HiSeq 2000 100-base pair-end platform with the IlluminaTruSeq kit. (San Diego, CA, USA; In our experience, exome data generated with the AgilentSureSelect capture kit behaves similar to the data presented in this paper). Samples were sequenced at Perkin Elmer (Branford, CT, USA). These filters have been applied to data generated with IlluminaTrueSeq and AgilentSureSelect capture technologies.

Reads were aligned to the UCSC reference human genome assembly 37.68[1] using BWA with the following commands: aln-o 1-e 10-i 5-k 2-l 32-t 4 (Li and Durbin, 2010). The mapping files in SAM format were converted to the BAM format using SAM tools version 0.1.19 (Li et al., 2009). The variants were called with the Broad Institute's Genome Analysis Tool Kit (McKenna et al.,

[1] http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/

2010; DePristo et al., 2011) using the following commands: -T Unified Genotyper-dcov 1000-stand_call_conf 30.0-stand_emit_conf 30.0 – min_base_quality_score 20 -A Depth Of Coverage -A IndelType -A QualByDepth -A ReadPosRankSumTest -A FisherStrand -A MappingQualityRankSumTest -l INFO -glm.

We obtained an average of 94.5 million reads (range 80–115 million reads per subject, with 106-fold mean depth in the target regions). On average, approximately 98% of these reads were mapped to the human reference genome.

DATA ANALYSIS

The VCF file generated by GATK was analyzed using Golden Helix Software (ver. 7.7.8) (Bozeman, MT, USA) and the newly developed CASSI. Variants located on the X and Y chromosomes were excluded from this analysis due to limitations in the Golden helix software. Only informative genotypes for each family were considered (genotypes where all three members of the trio were homozygous and identical to the reference sequence were removed). When sequencing data from multiple DNA sources were compared, only informative SNPs within the trio from one particular DNA source were included in the analysis. Variants were only removed based on individual quality measurements. When assessing the number of variants present in the child, all variants that remained in the child's dataset after the filters were applied were counted (i.e., if the genotypes for both parents were removed with a filter, but the child's genotype remained, this variant was still counted for the child), Mendelian errors were calculated for each variant by determining genotypes in the child which could not be inherited from the parents based on the parent genotypes. Mendelian errors were inferred for variants with a missing parental genotype if one parent and the child had opposite homozygous genotypes. The Mendelian error calculation did not include cases in which the child was heterozygous and only one parent was called. People interested in using CASSI should contact the corresponding author.

THE CINCINNATI ANALYTICAL SUITE FOR SEQUENCING INFORMATICS

Cincinnati Analytical Suite for Sequencing Informatics was developed to address the data management requirements of next-generation sequencing data and to facilitate access to state-of-the-art open source analysis packages through a centralized web-based interface. CASSI analysis pipelines are run on the CCHMC 700 core Linux-based computational cluster and can also run on a local Linux-based machine. It leverages existing open source VCF file parsers and annotation tools including VCF tools, ANNOVAR, UCSC Genome Browser, Exome Variant Server, and dbGAP (Mailman et al., 2007; Wang et al., 2010; Danecek et al., 2011; Meyer et al., 2013).

Cincinnati Analytical Suite for Sequencing Informatics consists of a web-based front end driven by a MySQL backend. Users are able to upload their NGS data in the form of VCF files along with files that contain the family relatedness information for each sample (fam files). CASSI performs basic quality control checks on the uploaded files before they are accepted into the database. These checks include looking for an abundance of Mendelian errors and verifying the sex of the uploaded samples.

Fields that are commonly queried, such as sample name, family ID, and variant position are parsed out of the VCF file and indexed in the MySQL database. Storing only commonly queried fields in the database while keeping the genotype information in the original VCF file keeps the database size to a minimum while allowing quick access to the original VCF file and sample information. Analysis begins by sample selection and analysis type selection from the CASSI web interface. Using this information, CASSI then dynamically generates a custom pipeline for the specific type of analysis, which is launched using the LONI pipeline software (Rex et al., 2003; Dinov et al., 2010). Pipeline parameters can be changed through the LONI pipeline's point-and-click interface. This allows for a seamless transition between search and analysis interfaces without requiring the user to have programmatic experience.

The LONI Pipeline simplifies computational cluster workflow creation using a drag and drop interface. CASSI users can launch and modify existing processing workflows directly from their web browser by using Java Web Start technology. The selected pipeline is preloaded into the LONI Pipeline client along with any sample data retrieved from the web interface. This is achieved by injecting the file locations of the sample data into a template .pipe LONI Pipeline file. Users are then free to modify the workflow. Input parameters are easily modified via the modules within the LONI Pipeline. The LONI Pipeline server interfaces with existing high performance computing environments in order to handle task dependencies and parallelization. In our case it communicates with the LSF job scheduler, but can also be used to communicate with other scheduling systems such as Oracle Grid Engine. Each modified LONI workflow can then be saved as XML and versioned using existing source control solutions (Subversion, Git, CVS, etc.). These XML files can be saved, shared, and submitted directly to a Linux-based machine.

For trio analysis, each member of the trio is initially extracted into a separate VCF file using VCFtools and then filtered on parameters selected by the user. After filtering for high quality variants, the samples are then scanned for amino acid altering variants (non-synonymous, splicing, insertions, deletions, and variations that alter initiation codons or stop codons) using the UCSC genome browser build 37 human Reference Sequence Gene table. Rare and novel variants are identified by filtering against the 1000 genomes project phase 1 v3 database[2] and the NHLBI exome sequence project ESP6500 variant frequency data[3]. We also generated and use an internal allele frequency table of 312 whole exomes analyzed at CCHMC.

Individual and summary reports are generated for all candidate causative variants. These variants are annotated with chromosome, position, minor allele frequency, Gene name (hyperlinked to www.Genecards.org), transcript, and protein ID, amino-acid position and functional predictions based on dbSNP functional predictions Version 2 table.

IDENTIFICATION OF POTENTIALLY CAUSATIVE MUTATIONS

Three different models of inheritance were used to identify candidate causative variants. We defined *de novo* variants as

[2] ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/

[3] http://evs.gs.washington.edu/EVS/

non-synonymous polymorphisms in which both parents are homozygous for the reference allele and the proband contained a heterozygous genotype. For homozygous recessive variants we required both parents to be heterozygous for the variant and the proband to be homozygous for the non-synonymous rare allele. For compound heterozygous polymorphism, we required the proband to contain at least two heterozygous non-synonymous polymorphisms in the same gene and neither parent could contain both variants. One variant could have a minor allele frequency in the general population up to 5% (based on the 1000 genome and exome sequence project); however, the other polymorphism had to have a minor allele frequency below 1%.

RESULTS

We collected biological samples from the blood, saliva, and buccal cells of a child and the two biological parents. By extracting DNA from these nine samples and sequencing the exome, we used Mendelian errors to identify those variants that were most likely to be sequencing artifacts. In developing informatics filters for the NGS exome data, we aimed to retain the largest number of total variants while removing the largest possible number of Mendelian errors in the child.

The vast majority of Mendelian errors in the unfiltered NGS data is due to sequencing error rather than *de novo* mutations based on the high fidelity of DNA replication in humans (Schmitt et al., 2009; Korona et al., 2011) and provide a method of tracking the effect of filters on sequencing artifacts. The initial analysis of the VCF file from the DNA obtained from blood revealed 2519 Mendelian errors compared to 79,911 called variants (3.15%). (These sequencing reads were mapped to 50 million bases, and for more than 99% of these calls, each of the subjects were homozygous for the reference base.) The mapped sequencing reads from different DNA samples of the same trio showed similar sequencing quality parameters (**Table 1**), similar proportion of Mendelian errors(3.05–3.15%), and total number of variants (79,234–79,911) called. We systematically applied filters to the VCF files until we identified the most efficient way to remove erroneous genotype calls while retaining the greatest number of true genotype calls. The first filter was based on the read depth (DP-number of sequencing reads that contain the variant) called within the trio (**Figure 1**). The read depth histogram of all variants in the proband using DNA isolated from blood shows a left-skewed distribution with two peaks located approximately at 5 reads and at 80 reads (near the mean read depth for this sample). A histogram for the same sample for the Mendelian errors shows the majority have a read depth below 12 reads and a sharp drop in the number of Mendelian errors as the read depth increases. Based on these results, we created filters with increasing stringency with a goal of removing the largest portion of the Mendelian errors while retaining the most variant calls. When applied to the unfiltered data a Read Depth < 10 removed 55% of the Mendelian errors, while retaining 92% of the called variants. With a Read Depth < 15 we were able to remove 59.2% of the Mendelian errors while retaining 90% of the called variants. Increasing the Read Depth filter above 15 had little effect on the number of Mendelian Errors removed (**Figure 1C**). Similar results were obtained with DNA isolated from buccal cells (61.6%) and saliva (56.1%).

There were a number of variants called with a read depth > 2000. It is possible that the sequences for these variants are the result of a PCR artifact during library construction or corresponds to repetitive regions of the genome. We assessed filters that excluded variants with Read-Depth > 800, >500, and >300. After applying these filters, we removed 5, 8, and 12% of Mendelian errors and 3, 6, and 15% of the total variants, respectively. These data suggested that by filtering out variants with a large relative mean number of reads we were not specifically filtering out Mendelian errors, rather we were randomly removing Mendelian errors by decreasing the number of variants. Thus, we did not exclude variants with a relatively large read depth.

Our second filter was based on the genotype quality score (GQ) of each of the variants called within the trio (**Figure 2**). The genotype quality score assesses the quality of sequencing information at each of the bases and ranges from 0 to 99 (see also discussion). A genotype quality score histogram for all variants found in the child blood DNA showed a right-skewed distribution with nearly all variants having GQ > 95. A similar histogram for the Mendelian errors shows a bi-modal distribution with a large portion of the data with a GQ < 20. Based on these results, we developed filters using increasingly stringent criteria and determined the effects of

Table 1 | Sequencing quality parameters for all three individuals in blood, buccal, and saliva trio.

Sample	Percentage of reads with GQS > 30 (%)	Mean GQS	Percentage of targeted sequence covered (%)	Mean read depth
Blood – proband	84.19	33.46	97.07	167
Blood – father	83.58	33.25	96.46	146
Blood – mother	84.57	33.57	94.69	150
Buccal– proband	84.12	33.4	95.89	90
Buccal – father	84.82	33.63	96.73	155
Buccal – mother	85.04	33.71	97.35	136
Saliva – proband	83.38	33.17	95.95	106
Saliva – father	84.21	33.46	95.93	154
Saliva – mother	84.31	33.48	96.3	132

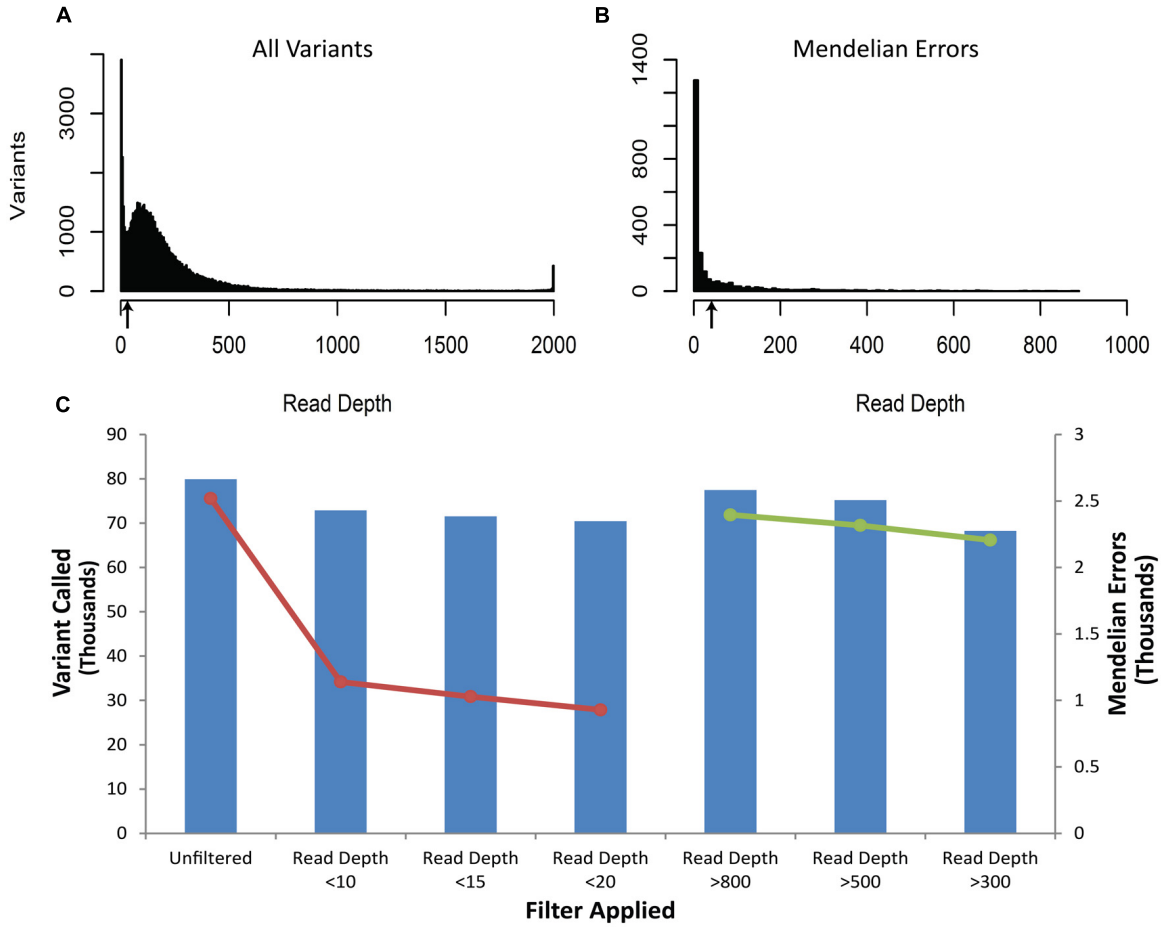


FIGURE 1 | Depth of Coverage: The histograms depict the read depth by all called variants **(A)** and for the Mendelian errors **(B)** in the child. Similar histograms were obtained for the other samples regardless of the DNA source. The arrows depict the coverage depth cutoff (Depth < 15 reads) used to remove sequencing artifacts from the data. The bar graph depicts the number of variants remaining after applying an increasingly stringent read depth filter **(C)**. The line graph (-•-) depicts the number of Mendelian errors remaining after applying an increasingly stringent read depth filter **(C)**. The sequencing data of the DNA extracted from blood are shown and are representative of the other two DNA sources.

those filters on the number of Mendelian errors and the number of variants. The GQ < 20 filter removed 72.4% of the Mendelian errors while retaining 94% of the variants. These data suggest that the GQ filter is very selective and effective at removing Mendelian errors (**Figure 2**). This filter removed 70.8% of the Mendelian errors in the DNA isolated from buccal cells and 70.1% from saliva.

The third filter was based upon the expected alternate allele ratio (alt ratio) for a particular genotype (**Figure 3**). Variants are determined to be homozygous reference, heterozygous, or homozygous non-reference based upon algorithms in the caller. The alternate allele ratio is the proportion of the number of reads with the alternate allele at a position relative to the total number of reads at that same position. We use this metric to identify genotypes that are unlikely to be accurate given the available allele read depth. The histogram for variants with a heterozygous genotype displayed a distribution centered on 0.5. Interestingly, the heterozygous genotypes generated a peak in the histogram near 0.2 and often a smaller peak near 0.8. One possible explanation for these peaks is that the misalignment of two or more regions of the genome that are nearly identical but unevenly sequenced generate these ratios (**Figure 3**). As expected, the histogram for variants with a homozygous genotype for the reference allele showed a left-skewed distribution and the histogram for variants with a homozygous genotype for the alternate allele had a right-skewed distribution (**Figure 3**). Unlike the previous two filters, which used the same criteria for all the variants, the alt ratio filter has different selection criteria based upon the genotype of the sample for each variant. For this particular filter, all homozygous reference variants with alt ratio >0.15 were removed, all homozygous alternate variants with alt ratio <0.85 were removed, and all heterozygous variants with alt ratio <0.3 or alt ratio >0.7 were removed. With this filter, we were able to remove 61.8% of the Mendelian errors while retaining 88% of the total variants. In buccal cell DNA 57.7% of the Mendelian errors were removed and 61.5% in saliva.

Our goal was to use multiple low stringency filters to selectively remove Mendelian errors while maintaining as much data as possible. Each of our filters based on the mean number of

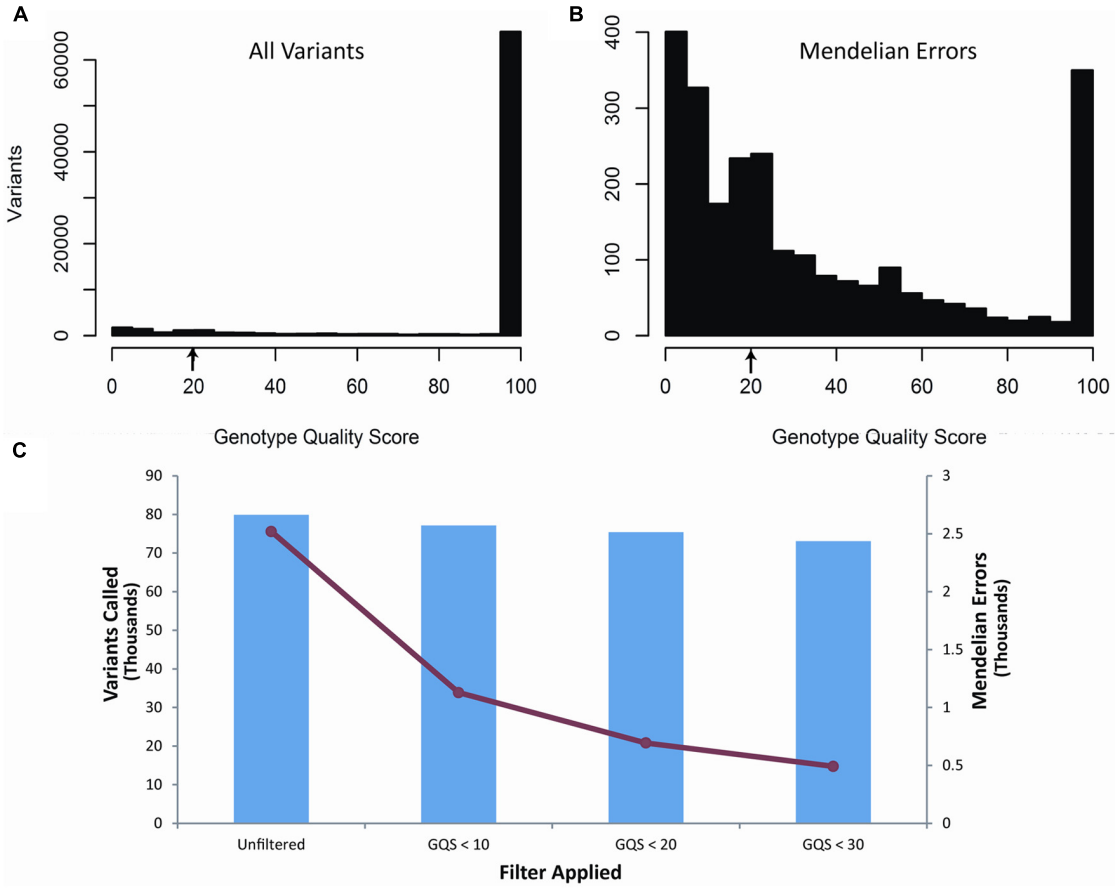


FIGURE 2 | Genotype Quality Score: The histograms depict the distribution of genotype quality scores by all called variants **(A)** and by all the Mendelian errors **(B)** in the child. The arrows depict the genotype quality score cutoff (GQS < 20) used to remove sequencing artifacts from the data. The bar graph depicts the number of variants remaining after applying an increasingly stringent genotype quality score filter **(C)**. The line graph (-●-) depicts the number of Mendelian errors remaining after increasingly stringent filters are applied **(C)**. The sequencing data of the DNA extracted from blood are shown and are representative of the other two DNA sources.

reads, genotype quality score, or alternate allele ratio was able to remove over half of the Mendelian errors in all of the DNA sources tested (Total Mendelian Errors: 2430–2519) while retaining a majority of the called variants (∼90%). We determined the cut-off for each filter based on the variant and Mendelian Error histograms for each parameter (**Figures 1–3**) and a cost-benefit analysis setting the filter at the point at which increasing the filter stringency removed the same proportion of total variants as Mendelian errors. To improve the filtering, we sequentially applied these filters to our trio data (**Figure 4**). As mentioned previously, we were able to exclude 61.8% of the Mendelian errors while retaining 88% of the data by excluding variants with alternate allele ratios differing by 0.2 or greater from the expected alternate allele ratio (**Figure 3**). By adding a filter that also excluded variants with GQ < 20, we were able to exclude 92.7% of the Mendelian errors while retaining 85% of the original sequencing data in the blood sample (**Figure 5**). By excluding variants with read depth less than 15, we were able to further remove 50 Mendelian Errors. Although this may not seem to be a large decrease in the number of Mendelian Errors, these 50 Mendelian Errors comprise approximately 30% of the Mendelian Errors remaining after the Genotype Quality Score and the Alternate Allele Ratio filter are applied. By combining the three filters, we were able to remove 95% of the Mendelian errors, while retaining nearly 80% of the called variants. As shown in **Figure 6**, nearly 60% of the excluded variants are removed by only one filter, supporting our strategy of using multiple low-stringency filters to remove sequencing artifacts.

The Mendelian Error rate in unfiltered data is 3.7%. Based on the observation that the true error rate is three to four times the Mendelian Error rate detected by SNPs (Gordon et al., 1999) we estimate the actual error rate to be 9–10% in unfiltered NGS data. This estimate is in agreement with concordance rates seen when DNA from three different sources: blood, buccal-cells, and saliva for the same sample were compared. We assessed the concordance rates of non-filtered variants that were found in all of the DNA sources and found ∼96% concordance for variants which were present in all three DNA sources (**Table 2**). The concordance dropped to ∼84% if we also considered genotypes which were non-reference in one DNA source, but not

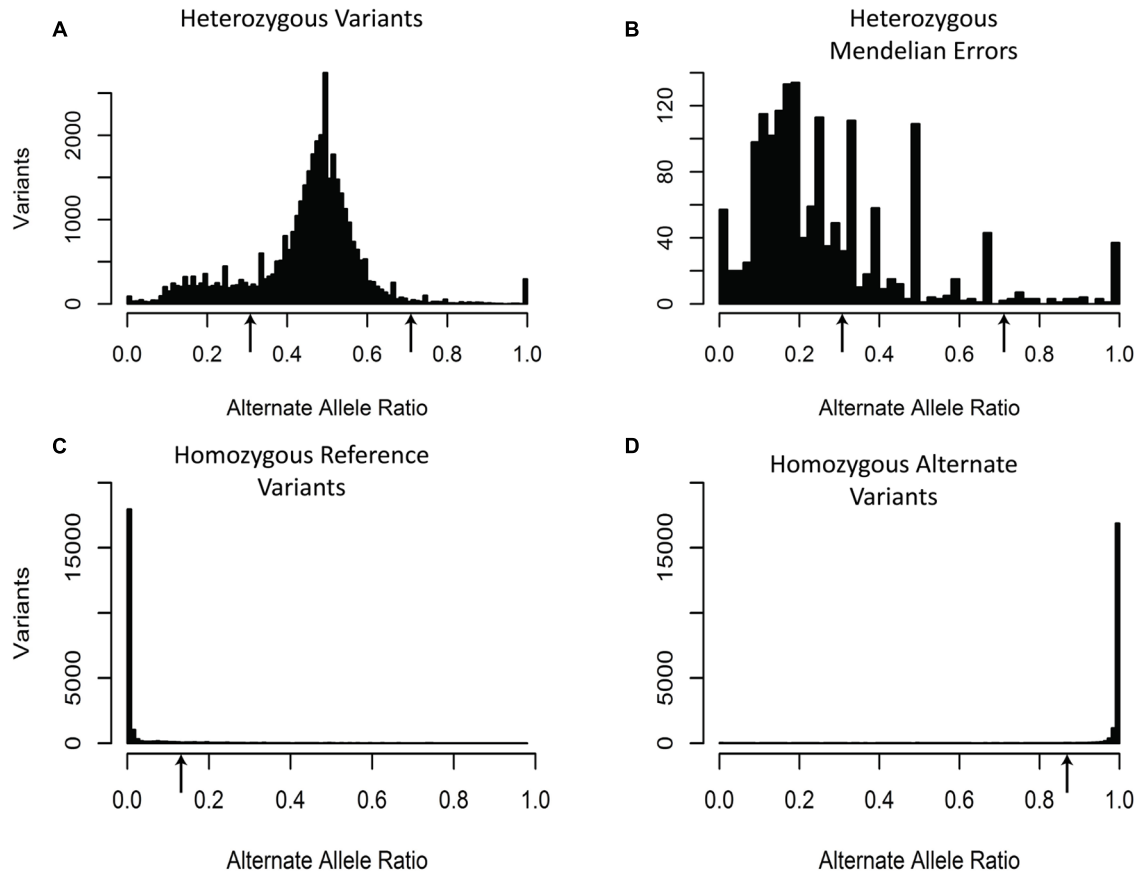


FIGURE 3 | Alternate Allele Ratio: The histograms depict the distribution of Alternate Allele Ratio by all called variants with a heterozygous genotype **(A)**, by all the Mendelian errors with a heterozygous genotype **(B)**, by all the called variants with a homozygous reference genotype **(C)**, and by all the called variants with a homozygous alternate genotype **(D)**. The arrows depict the alternate allele ratios used to remove sequencing artifacts for heterozygous genotype calls (Alt-Allele Ratio $>$ 0.7 or Alt-Allele Ratio $<$ 0.3), homozygous reference genotype calls (Alt-Allele Ratio $>$ 0.15), and homozygous alternate genotype calls (Alt-Allele Ratio $<$ 0.85). The sequencing data of the DNA extracted from blood are shown and are representative of the other two DNA sources.

called in one of the other two as being discordant. These unique genotypes were probably enriched for sequencing errors, as the vast majority were removed after applying the three filters described above (**Table 3**). After applying the filters, we were able to increase this concordance to greater than 99.999% amongst the variants that were common between the DNA sources (**Table 2**).

A trio study design is often used to identify candidate causative rare variants. In order to identify those amino-acid changing variants most likely to contribute functionally to a phenotype, we performed analyses to identify *de novo*, recessive homozygous (with less than 1% allele frequency in public sequencing databases), and compound heterozygous mutations. Filtering the sequencing data before this functional analysis reduced the apparent *de novo* mutations from 321 to 1. Similarly, potentially causal recessive homozygous variants were reduced from 32 to 3 and potentially causal compound heterozygous variants were reduced from 242 to 47. When these analyses were applied to each of the three DNA sources, we further reduced the number of potential causal variants to 0 apparent *de novo*, 3 rare homozygous, and 17 compound heterozygous variants which are identified in all three samples from different DNA sources (**Table 4**).

We developed CASSI to meet the need to store, version, filter, and annotate NGS data. CASSI is an application that seamlessly integrates file storage, metadata storage (e.g., family structure), and downstream processing with a web-based front-end that contains a user-friendly query interface (**Figure 7**). The web interface of CASSI enables biologists and clinicians without any computer science background to launch sophisticated analytical workflows to analyze next-generation sequencing data in an automated procedure. For example, the interface allows users to directly interface with annotation and filtering packages (such as vcftools, variant tools, and ANNOVAR), which are executed on a high-performance cluster at CCHMC.

The key technical component in CASSI is the LONI pipeline engine from UCLA, which is a graphical user interface for executing complex workflows on a cluster that can be launched directly from a web browser. Query results obtained through the CASSI web interface are made available as a data source in the LONI pipeline, and users can choose from a large number of filtering and

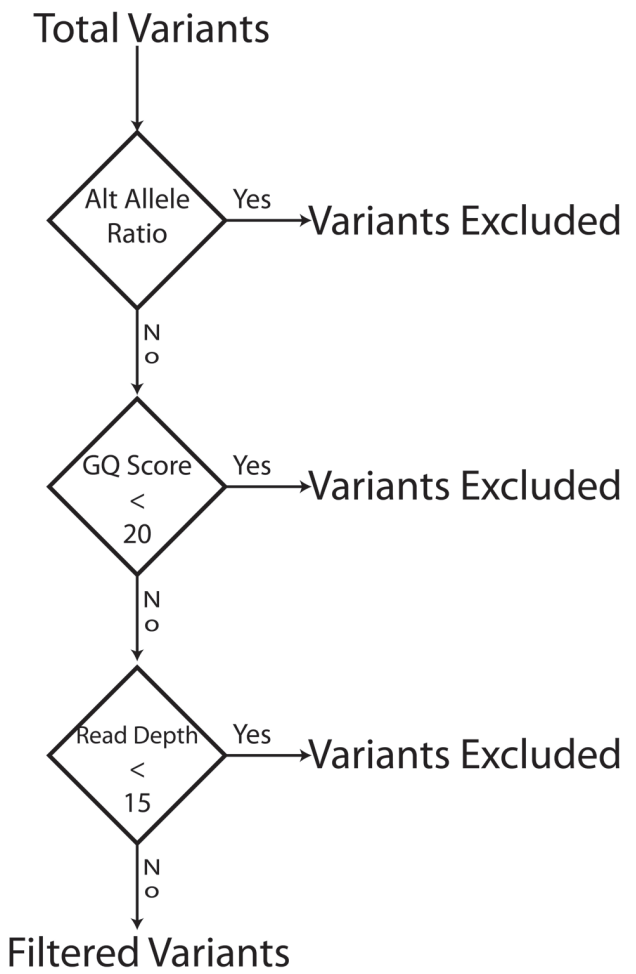


FIGURE 4 | Filter Schema: The bioinformatics filters are sequentially applied. Variants that fail a filter are excluded and the next filter in the sequence is applied to the variants that pass a filter. To be excluded a variant has to fail only one filter. The order of filter is not important.

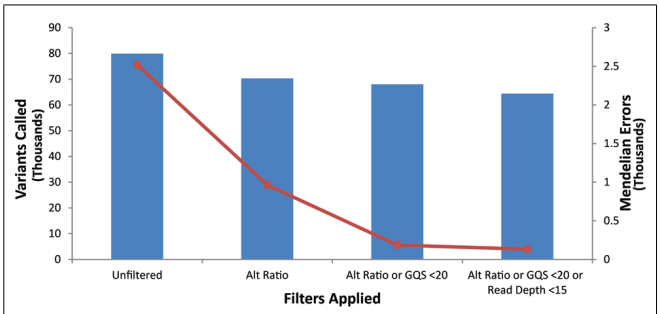


FIGURE 5 | Effect of applying multiple filters: The bar-graph depicts the number of variants remaining after the application of filters on the data. The line-graph depicts the number of Mendelian Errors remaining after the application of additional filters (Top Panel). The sequencing data of the DNA extracted from blood is shown.

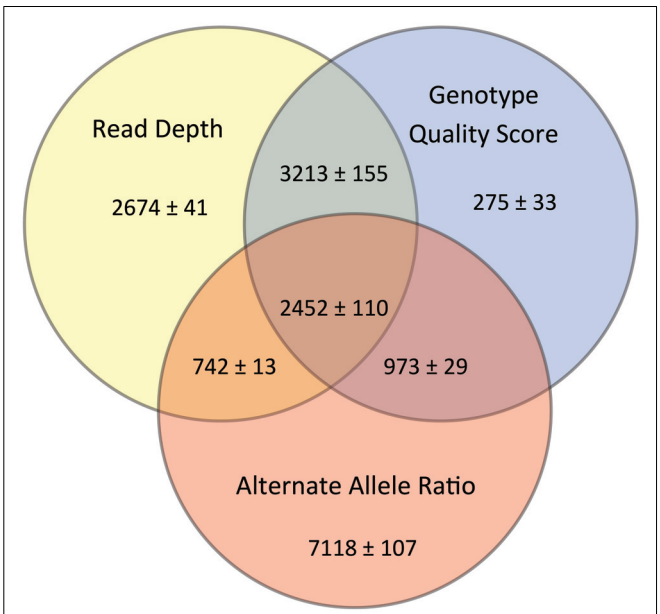


FIGURE 6 | Impact of Filters upon Data Quality: The Venn diagram shows the number of variants excluded by each of the filters. The numbers represent a mean ± range/2 from all three DNA sources.

annotation workflows to analyze variant data. CASSI allows the user to efficiently compare various filtering strategies; for example, it can easily record the number of variants and Mendelian errors remaining after individual filters are implemented. Most importantly, CASSI can be used to assess concordance between samples and to identify *de novo*, rare recessive, and compound heterozygous variants. The flexibility of the pipeline facilitates the implementation of new analytical strategies directly from the interface. Other groups have independently developed a genomic pipeline using LONI, supporting the utility of this resource for sequencing data (Dinov et al., 2011; Torri et al., 2012; **Figure 6**).

DISCUSSION

Next-generation sequencing provides investigators with the ability to quickly and economically generate human sequencing data including the presence of SNPs and insertions/deletions (O'Rawe et al., 2013). This ability to generate large volumes of data also presents the challenge of determining which variants to validate and study biologically (Nielsen et al., 2011). As reviewed in *Nekrutenko and Taylor*, there is no generally accepted method for filtering variants in clinical studies (Nekrutenko and Taylor, 2012). The usual approach for shortening the list of top variants relies on filtering on two parameters, read-depth, and PHRED quality score (Girard et al., 2011; Xu et al., 2011; O'Roak et al., 2012a; Sanders et al., 2012). Although these particular methods successfully remove many of the variants, due to the stringency of filters, they are also excluding real variants present within the sequencing data. By using a combination of three filters based on the intrinsic characteristics of NGS, we removed a large proportion of the Mendelian errors, while retaining the highest portion reasonable of variants called. We estimate that these filters removed 90% of the sequencing artifacts at a cost of 20% of the data.

Table 2 | Concordance analysis of DNA from three individuals was collected from three biological sources and sequenced.

DNA source	Sample	Concordance rate no filter applied (%)	Concordance rate all filters applied, not including variants that are unique to a single DNA source (%)	Concordance rate all filters applied, including variants that are unique to a single DNA source (%)
Blood vs. buccal	Individual 1	96.23	99.99	84.06
	Individual 2	96.61	99.99	84.71
	Individual 3	96.62	99.98	84.48
Blood vs. saliva	Individual 1	96.30	99.99	84.22
	Individual 2	96.53	99.99	84.42
	Individual 3	96.50	99.99	83.79
Buccal vs. saliva	Individual 1	96.27	99.99	84.14
	Individual 2	96.86	99.98	85.05
	Individual 3	96.73	99.99	84.22

Using variants that were common amongst all of the DNA sources, we assessed the genotype concordance. We observed a significant improvement in the concordance after applying the three bioinformatics filters.

One limitation of using Mendelian errors to identify sequencing artifacts is that they under represent the true sequencing error rate as they have low power to detect errors in bi-allelic polymorphisms. In cases where both parents are heterozygous for a polymorphism the child could have any one of the potential genotypes and it would follow Mendelian inheritance. In genotyping experiments it has been estimated that Mendelian errors only predict one-third to one-fourth the number of actual errors (Gordon et al., 1999). Furthermore, the identification of Mendelian errors does not indicate which sample's genotype is erroneous. Even with the limitations, Mendelian errors provide a useful method to determine the quality of the data.

The number of variants with a particular depth of coverage demonstrated a clear peak around 120 (**Figure 1**), which was close to the target coverage depth of 100 reads. On the other hand, the histogram for the depth of coverage amongst Mendelian errors (**Figure 1**) confirmed that the majority of Mendelian errors had a low depth of coverage. This low depth of coverage for the Mendelian errors indicated that many of them may be occurring due to selective sequencing of one chromosome rather than equal sequencing of both chromosomes. This would be particularly relevant for heterozygous SNPs. If only one of the chromosomes was sequenced, the individual would be called either a homozygous reference or homozygous alternate at a particular variant. As the number of reads increases, the probability of sequencing the same chromosome for each read decreases exponentially. By sheer chance at a read depth of 10 with 50-million total reads, there will be 50,000 instances of only one chromosome being read. If the read-depth is increased to 15 reads, this number decreases to approximately 1,500 instances. Based on the difference in the distribution of Mendelian errors and total variants for depth of coverage, a filter which excludes variants with low depth of coverage (15 reads or <20% of average reads) removed a small portion of the total variants while removing a large portion of Mendelian errors (**Figure 2**).

Table 3 | DNA from the proband (child) was collected from three biological sources and sequenced.

DNA source		Unique compared to blood	Unique compared to buccal	Unique compared to saliva	Unique compared to the other two sources
Blood	Unfiltered		2636	1997	1095
	Filtered		10	4	2
Buccal	Unfiltered	1267		1437	535
	Filtered	0		2	0
Saliva	Unfiltered	1268	1438		669
	Filtered	1	0		0

The data set from each DNA source had unique variants that were not found in one or both of the other sources. After three bioinformatics filters were applied to the genotyping data, the number of unique genotypes was considerably reduced. The number of variants in one DNA source that are unique compared to another DNA source (e.g., blood compared to buccal) is different than the inverse comparison (e.g., buccal compared to blood).

Table 4 | Candidate causative sequence variants were identified in unfiltered and filtered data from the same trio that was sequenced three times using different DNA sources.

DNA source	De novo variants			Recessive homozygous variants			Compound heterozygous		
	Called	Unique to a single DNA source	Common to all DNA sources	Called	Unique to a single DNA source	Common to all DNA Sources	Called	Unique to a single DNA source	Common to all DNA sources
Unfiltered									
Blood	321	228	12	32	3	23	242	47	153
Buccal	306	219	12	36	12	23	285	79	153
Saliva	304	230	12	28	4	23	284	80	153
Filtered									
Blood	1	0	0	3	0	3	47	21	17
Buccal	0	0	0	3	0	3	39	11	17
Saliva	1	0	0	3	0	3	45	16	17

Only novel or rare amino acid altering variants were considered. The numbers of unique variants (not found in any other DNA source) are indicated. Common variants were found in all three DNA sources. Variants found in two of the DNA sources, but not in the third are not included in this table. Filters applied to the variants included read depth >15, genotype quality score >20 and alternate allele ratio less than 0.15 for all homozygous reference, greater than 0.85 for homozygous alternate allele and between 0.3 and 0.7 for heterozygous genotypes.

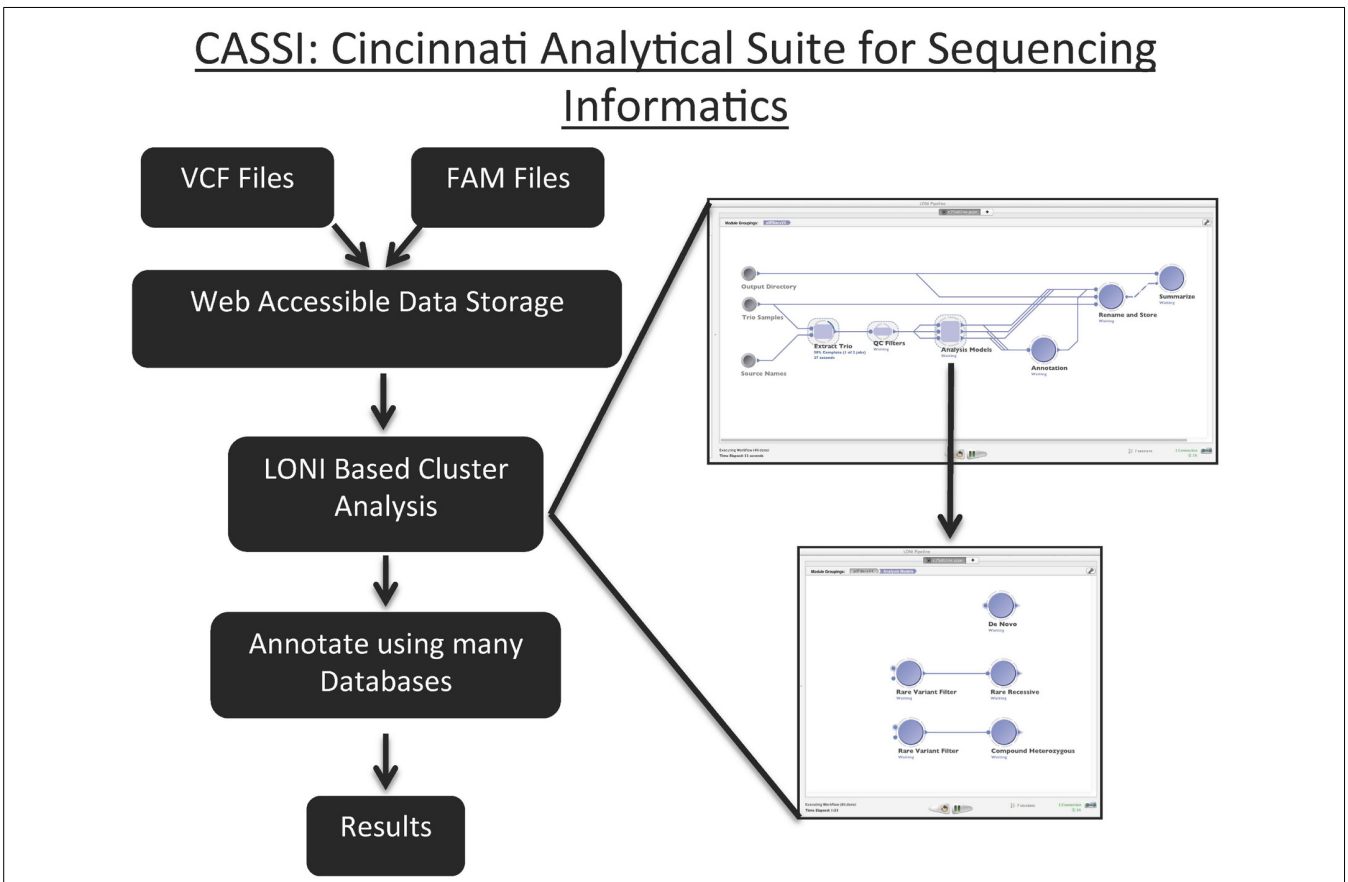


FIGURE 7 | Graphical Representation of the Cincinnati Analytical Suite for Sequencing Informatics. Files with VCF and FAM are uploaded to CASSI's web interface. These data are parsed and stored in web accessible data storage. The LONI based analysis allows users to analyze data through the established pipeline for identifying *de novo*, rare recessive and compound heterozygous variants. These pipe lines can also be changed to accommodate specific analyses. After variants are annotated using many databases, the results are versioned, saved in the database, and available for downloading.

The GQ is computed based on the likelihood of a particular genotype being called in comparison to the likelihood of the other two genotypes being called: L(0/1) versus L(0,0) and L(1,1). Our data based on the GQ demonstrated that the majority of variants have high GQ, whereas the Mendelian errors have a much lower score (**Figure 2**). The difference in the GQ represents the likelihood in the variant calls. For many of the Mendelian errors, the low GQ suggests a low confidence in those calls, which may be due to inconsistent individual sequencing reads at that particular location due to low coverage depth, difficult alignment, or poor sequencing reads. This is expected for sequencing artifacts, since it is unlikely that a sequencing artifact will consistently produce the same sequencing read at a particular location. On the other hand, a true Mendelian Error such as a *de novo* mutation would produce a consistent sequencing read since it is a true difference in the sequence. We exploited the differences in the genotype quality scores by generating a filter that excludes variants with a genotype quality score less than 20. This allowed us to exclude the Mendelian errors present on the left peak of the histogram without excluding a large portion of the called variants, which are located within the peak on the right side of the GQ histogram (**Figure 2**).

We added an additional filter based on the alternate allele ratio (alt ratio; DePristo et al., 2011; Girard et al., 2011; Xu et al., 2011; O'Roak et al., 2012a; Riviere et al., 2012; Sanders et al., 2012). Due to the high depth of coverage for most variants, we expected our variants to have alternate allele ratios close to the theoretical values: 0, 0.5, and 1 representing homozygous reference, heterozygous, and homozygous alternate, respectively. In effect, this filter assesses the consistency of the variant call based on all the sequencing reads. The majority of the variants that the alt ratio filter removes were heterozygous Mendelian errors which were enriched in the peak at 0.2 (p-value $< 10^{-50}$; **Figure 3**) suggesting that homozygous reference and homozygous alternate variant calls were more reliable than heterozygous variant calls.

We combined these three individual filters and observed the increased efficiency of the combined filters in removing the sequencing artifacts. As is evident from the bar graph representing the total number of variants (**Figure 5**), and the line graph representing the number of Mendelian errors called, there was a 93% decrease, in the number of Mendelian errors by the addition of the GQ filter to the Alt Ratio without a large reduction in the number of variants removed (14.9%). This trend continued as we added the depth of read filter to the other filters. By excluding variants that fail the depth of read filter, the Genotype Quality Score Filter, or the Alt Ratio filter, we were able to exclude over 95% of the Mendelian errors. In our test trio, this lowered the number of Mendelian errors called to approximately 130 variants. We attributed the drastic decrease in the number of Mendelian errors to the low likelihood of a sequencing artifact passing all the filters. Approximately 80% of the variants passed all three filters.

After identifying a combination of filters that removed the vast majority of the Mendelian errors, while retaining a large portion of the variants called, we assessed the concordance between identical samples isolated from different DNA sources. The unfiltered sequencing data-set of samples from blood, buccal cells, and saliva had a concordance of ∼84% (including unique calls as discordant). After applying our filters the concordance rate increased to >99.9% between all three samples from the three different DNA sources. It is necessary for the filtering method to generate concordant data, since clinical DNA samples can be collected from any one of various different sources including blood, saliva, and buccal cells.

Next-generation sequencing experiments are often used to find rare or novel variants that lead to disease. Unfortunately, sequencing artifacts can often mimic and confound the identification of these variants. Sequencing artifacts contribute to a large number of false positive disease-causing candidates. The vast majority of apparent *de novo* variants identified in the unfiltered data are sequencing artifacts (**Table 2**). However in previous experiments, after filtering the data on read depth, genotype quality score and alt ratio our confirmation rate by Sanger sequencing is greater than 95% for *de novo* variants.

Cincinnati Analytical Suite for Sequencing Informatics is a suite that allows users with varying degrees of programming sophistication to perform documented, reproducible studies with NGS data to gain insight into the etiology of disease. In the case of the current study, CASSI allowed us to quickly and reproducibly assess different filtering strategies through the calculation of Mendelian errors and total variants remaining after specific filters were applied. With the incorporation of a LONI pipeline we have created a fully automated system that can filter, annotate and apply various genetic models to identify candidate causative variants. The LONI pipeline provides investigators the ability to apply predefined values for filtering or customize the pipeline to fit the type and quality of the data being analyzed.

The filters and methods presented reproducibly generate robust and accurate data sets with low levels of sequencing artifacts. Both the genotype quality score and alt allele ratio filters can be applied to data sets regardless of read depth. In data sets with >75× average read depth we recommend using a hard filter of 15× for read depth. In data sets with less than 75× coverage we suggest using a filter of 20% of the average read depth. While these data sets will contain higher amounts of false variant calls, a hard filter would remove too many true variants from the data set. These filters also have the potential to have bias towards removing variants caused by mosaicism. The alt allele ratio would be particularly sensitive to variants if the cell population with different genotypes is not close to 50%.

As the NGS technology progresses further and the per-base sequencing cost decrease, researchers will be able to generate NGS data-sets with increased depth of coverage and longer read lengths. Both of these improvements will yield better calling of variants. Additionally, the longer read lengths will allow researchers to more accurately predict insertions and deletions. A recent review further identifies ways to improve the fidelity of NGS data, including the use of filtering strategies such as the one presented herein (Robasky et al., 2014).

In summary, our three filters of NGS data selectively exclude the sequencing artifacts, measured as Mendelian errors, while limiting the removal of the true variation amongst the samples. In addition, we show that DNA isolated from different sources (blood, buccal cells, and saliva) have greater than 99.9% concordance and thus mixed DNA sources can be used for causative variant

identification. Our work flow is based on obtaining the most accurate data set possible and results in an extremely small number of candidate causative variants for consideration and interpretation (usually fewer than 10 genes per trio). These methods have been automated through CASSI and greatly increase the ability of investigators and clinicians to understand and discover genetic causes of disease by quickly identifying potential causative variations.

ACKNOWLEDGMENTS

As part of the University of Cincinnati Medical Scientist Training program, Zubin H. Patel is partially supported by the NIGMS Medical Scientist Training Program T-32 GM063483. We are grateful for support from the US Department of Veterans Affairs, the Department of Defense (PR094002), and the National Institutes of Health (HG006828, HG006382, AI024717, AI083194, AR048929, AR049084 and HL10533).

REFERENCES

Bartnik, M., Chun-Hui Tsai, A., Xia, Z., Cheung, S. W., and Stankiewicz, P. (2011). Disruption of the scn2a and scn3a genes in a patient with mental retardation, neurobehavioral and psychiatric abnormalities, and a history of infantile seizures. *Clin. Genet.* 80, 191–195. doi: 10.1111/j.1399-0004.2010.01526.x

Bujakowska, K., Audo, I., Mohand-Said, S., Lancelot, M. E., Antonio, A., Germain, A., et al. (2012). Crb1 mutations in inherited retinal dystrophies. *Hum. Mutat.* 33, 306–315. doi: 10.1002/humu.21653

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and vcftools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Dauber, A., Nguyen, T. T., Sochett, E., Cole, D. E., Horst, R., Abrams, S. A., et al. (2012). Genetic defect in cyp24a1, the vitamin d 24-hydroxylase gene, in a patient with severe infantile hypercalcemia. *J. Clin. Endocrinol. Metab.* 97, E268–E274. doi: 10.1210/jc.2011–1972

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Dinov, I., Lozev, K., Petrosyan, P., Liu, Z., Eggert, P., Pierce, J., et al. (2010). Neuroimaging study designs, computational analyses and data provenance using the loni pipeline. *PLoS ONE* 5:e13070. doi: 10.1371/journal.pone.0013070

Dinov, I. D., Torri, F., Macciardi, F., Petrosyan, P., Liu, Z., Zamanyan, A., et al. (2011). Applications of the pipeline environment for visual informatics and genomics computations. *BMC Bioinformatics* 12:304. doi: 10.1186/1471-2105-12–304

Filges, I., Shimojima, K., Okamoto, N., Rothlisberger, B., Weber, P., Huber, A. R., et al. (2011). Reduced expression by setbp1 haploinsufficiency causes developmental and expressive language delay indicating a phenotype distinct from schinzel-giedion syndrome. *J. Med. Genet.* 48, 117–122. doi: 10.1136/jmg.2010.084582

Gilissen, C., Arts, H. H., Hoischen, A., Spruijt, L., Mans, D. A., Arts, P., et al. (2010). Exome sequencing identifies wdr35 variants involved in sensenbrenner syndrome. *Am. J. Hum. Genet.* 87, 418–423. doi: 10.1016/j.ajhg.2010.08.004

Gilman, S. R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M., and Vitkup, D. (2011). Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* 70, 898–907. doi: 10.1016/j.neuron.2011.05.021

Girard, S. L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., et al. (2011). Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat. Genet.* 43, 860–863. doi: 10.1038/ng.886

Gonzalez-del Pozo, M., Borrego, S., Barragan, I., Pieras, J. I., Santoyo, J., Matamala, N., et al. (2011). Mutation screening of multiple genes in spanish patients with autosomal recessive retinitis pigmentosa by targeted resequencing. *PLoS ONE* 6:e27894. doi: 10.1371/journal.pone.0027894

Gordon, D., Heath, S. C., and Ott, J. (1999). True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. *Hum. Hered.* 49, 65–70. doi: 10.1159/000022846

Harakalova, M., van Harssel, J. J., Terhal, P. A., van Lieshout, S., Duran, K., Renkens, I., et al. (2012). Dominant missense mutations in abcc9 cause cantu syndrome. *Nat. Genet.* 44, 793–796. doi: 10.1038/ng.2324

Hoischen, A., van Bon, B. W., Gilissen, C., Arts, P., van Lier, B., Steehouwer, M., et al. (2010). De novo mutations of setbp1 cause schinzel-giedion syndrome. *Nat. Genet.* 42, 483–485. doi: 10.1038/ng.581

Hoischen, A., van Bon, B. W., Rodriguez-Santiago, B., Gilissen, C., Vissers, L. E., de Vries, P., et al. (2011). De novo nonsense mutations in asxl1 cause bohring-opitz syndrome. *Nat. Genet.* 43, 729–731. doi: 10.1038/ng.868

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., et al. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron* 74, 285–299. doi: 10.1016/j.neuron.2012.04.009

Korona, D. A., Lecompte, K. G., and Pursell, Z. F. (2011). The high fidelity and unique error signature of human DNA polymerase epsilon. *Nucleic Acids Res.* 39, 1763–1773. doi: 10.1093/nar/gkq1034

Lederer, D., Grisart, B., Digilio, M. C., Benoit, V., Crespin, M., Ghariani, S. C., et al. (2012). Deletion of kdm6a, a histone demethylase interacting with mll2, in three patients with kabuki syndrome. *Am. J. Hum. Genet.* 90, 119–124. doi: 10.1016/j.ajhg.2011.11.021

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and samtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lin, Z., Chen, Q., Lee, M., Cao, X., Zhang, J., Ma, D., et al. (2012). Exome sequencing reveals mutations in trpv3 as a cause of olmsted syndrome. *Am. J. Hum. Genet.* 90, 558–564. doi: 10.1016/j.ajhg.2012.02.006

Liu, X., Han, S., Wang, Z., Gelernter, J., and Yang, B. Z. (2013). Variant callers for next-generation sequencing data: a comparison study. *PLoS ONE* 8:e75619. doi: 10.1371/journal.pone.0075619

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., et al. (2007). The ncbi dbgap database of genotypes and phenotypes. *Nat. Genet.* 39, 1181–1186. doi: 10.1038/ng1007-1181

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Meyer, L. R., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Kuhn, R. M., Wong, M., et al. (2013). The ucsc genome browser database: extensions and updates 2013. *Nucleic Acids Res.* 41, D64–D69. doi: 10.1093/nar/gks1048

Neale, B. M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K. E., Sabo, A., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 485, 242–245. doi: 10.1038/nature11011

Need, A. C., Shashi, V., Hitomi, Y., Schoch, K., Shianna, K. V., McDonald, M. T., et al. (2012). Clinical application of exome sequencing in undiagnosed genetic conditions. *J. Med. Genet.* 49, 353–361. doi: 10.1136/jmedgenet-2012–100819

Nekrutenko, A., and Taylor, J. (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat. Rev. Genet.* 13, 667–672. doi: 10.1038/nrg3305

Neveling, K., Collin, R. W., Gilissen, C., van Huet, R. A., Visser, L., Kwint, M. P., et al. (2012). Next-generation genetic testing for retinitis pigmentosa. *Hum. Mutat.* 33, 963–972. doi: 10.1002/humu.22045

Ng, S. B., Bigham, A. W., Buckingham, K. J., Hannibal, M. C., McMillin, M. J., Gildersleeve, H. I., et al. (2010a). Exome sequencing identifies mll2 mutations as a cause of kabuki syndrome. *Nat. Genet.* 42, 790–793. doi: 10.1038/ng.646

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., et al. (2010b). Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* 42, 30–35. doi: 10.1038/ng.499

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and snp calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–451. doi: 10.1038/nrg2986

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., et al. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* 5, 28. doi: 10.1186/gm432

O'Roak, B. J., Vives, L., Fu, W., Egertson, J. D., Stanaway, I. B., Phelps, I. G., et al. (2012a). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* 338, 1619–1622. doi: 10.1126/science.1227764

O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., et al. (2012b). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 485, 246–250. doi: 10.1038/nature10989

Paulussen, A. D., Stegmann, A. P., Blok, M. J., Tserpelis, D., Posma-Velter, C., Detisch, Y., et al. (2011). Mll2 mutation spectrum in 45 patients with kabuki syndrome. *Hum. Mutat.* 32, E2018–E2025. doi: 10.1002/humu.21416

Rex, D. E., Ma, J. Q., and Toga, A. W. (2003). The loni pipeline processing environment. *Neuroimage* 19, 1033–1048. doi: 10.1016/S1053-8119(03)00185-X

Riviere, J. B., Mirzaa, G. M., O'Roak, B. J., Beddaoui, M., Alcantara, D., Conway, R. L., et al. (2012). De novo germline and postzygotic mutations in akt3, pik3r2 and pik3ca cause a spectrum of related megalencephaly syndromes. *Nat. Genet.* 44, 934–940. doi: 10.1038/ng.2331

Robasky, K., Lewis, N. E., and Church, G. M. (2014). The role of replicates for error mitigation in next-generation sequencing. *Nat. Rev. Genet.* 15, 56–62. doi: 10.1038/nrg3655

Rosenfeld, J. A., Mason, C. E., and Smith, T. M. (2012). Limitations of the human reference genome for personalized genomics. *PLoS ONE* 7:e40294. doi: 10.1371/journal.pone.0040294

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485, 237–241. doi: 10.1038/nature10945

Santen, G. W., Aten, E., Sun, Y., Almomani, R., Gilissen, C., Nielsen, M., et al. (2012). Mutations in swi/snf chromatin remodeling complex gene arid1b cause coffin-siris syndrome. *Nat. Genet.* 44, 379–380. doi: 10.1038/ng.2217

Schmitt, M. W., Matsumoto, Y., and Loeb, L. A. (2009). High fidelity and lesion bypass capability of human DNA polymerase delta. *Biochimie* 91, 1163–1172. doi: 10.1016/j.biochi.2009.06.007

Schrier, S. A., Bodurtha, J. N., Burton, B., Chudley, A. E., Chiong, M. A., D'Avanzo M, G., et al. (2012). The coffin-siris syndrome: a proposed diagnostic approach and assessment of 15 overlapping cases. *Am. J. Med. Genet. A* 158A, 1865–1876. doi: 10.1002/ajmg.a.35415

Torri, F., Dinov, I. D., Zamanyan, A., Hobel, S., Genco, A., Petrosyan, P., et al. (2012). Next generation sequence analysis and computational genomics using graphical pipeline workflows. *Genes (Basel)* 3, 545–575. doi: 10.3390/genes3030545

Tsurusaki, Y., Okamoto, N., Ohashi, H., Kosho, T., Imai, Y., Hibi-Ko, Y., et al. (2012). Mutations affecting components of the swi/snf complex cause coffin-siris syndrome. *Nat. Genet.* 44, 376–378. doi: 10.1038/ng.2219

Van Houdt, J. K., Nowakowska, B. A., Sousa, S. B., van Schaik, B. D., Seuntjens, E., Avonce, N., et al. (2012). Heterozygous missense mutations in smarca2 cause nicolaides-baraitser syndrome. *Nat. Genet.* 44, 445–449, S441. doi: 10.1038/ng.1105

Vissers, L. E., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P., et al. (2010). A de novo paradigm for mental retardation. *Nat. Genet.* 42, 1109–1112. doi: 10.1038/ng.712

Wang, K., Li, M., and Hakonarson, H. (2010). Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164. doi: 10.1093/nar/gkq603

Whalen, S., Heron, D., Gaillon, T., Moldovan, O., Rossi, M., Devillard, F., et al. (2012). Novel comprehensive diagnostic strategy in pitt-hopkins syndrome: clinical score and further delineation of the tcf4 mutational spectrum. *Hum. Mutat.* 33, 64–72. doi: 10.1002/humu.21639

Xu, B., Roos, J. L., Dexheimer, P., Boone, B., Plummer, B., Levy, S., et al. (2011). Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat. Genet.* 43, 864–868. doi: 10.1038/ng.902

Yu, X., and Sun, S. (2013). Comparing a few snp calling algorithms using low-coverage sequencing data. *BMC Bioinformatics* 14:274. doi: 10.1186/1471-2105-14–274

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 22 November 2013; accepted: 16 January 2014; published online: 12 February 2014.

*Citation: Patel ZH, Kottyan LC, Lazaro S, Williams MS, Ledbetter DH, Tromp G, Rupert A, Kohram M, Wagner M, Husami A, Qian Y, Valencia CA, Zhang K, Hostetter MK, Harley JB and Kaufman KM (2014) The struggle to find reliable results in exome sequencing data: filtering out Mendelian errors. Front. Genet. **5**:16. doi: 10.3389/fgene.2014.00016*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 29 April 2014
doi: 10.3389/fgene.2014.00105

Assessing the functional consequence of loss of function variants using electronic medical record and large-scale genomics consortium efforts

Patrick Sleiman[1,2] *, Jonathan Bradfield[1], Frank Mentch[1], Berta Almoguera[1], John Connolly[1] and Hakon Hakonarson[1,2] *

[1] *Center for Applied Genomics, Abramson Research Center, The Children's Hospital of Philadelphia, Philadelphia, PA, USA*
[2] *Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*

Edited by:
Mariza De Andrade, Mayo Clinic, USA

Reviewed by:
Lijun Ma, Wake Forest University Health Sciences, USA
Hui-Qi Qu, The University of Texas School of Public Health, USA

****Correspondence:***
Patrick Sleiman and Hakon Hakonarson, Center for Applied Genomics, Abramson Research Center, The Children's Hospital of Philadelphia, 3615 Civic Center Boulevard, Philadelphia, PA 19104-4318, USA
e-mail: sleimanp@email.chop.edu; hakonarson@email.chop.edu

Estimates from large scale genome sequencing studies indicate that each human carries up to 20 genetic variants that are predicted to results in loss of function (LOF) of protein-coding genes. While some are known disease-causing variants or common, tolerated, LOFs in non-essential genes, the majority remain of unknown consequence. We explore the possibility of using imputed GWAS data from large biorepositories such as the electronic medical record and genomics (eMERGE) consortium to determine the effects of rare LOFs. Here, we show that two hypocholesterolemia-associated LOF mutations in the *PCSK9* gene can be accurately imputed into large-scale GWAS datasets which raises the possibility of assessing LOFs through genomics-linked medical records.

Keywords: loss of function (LOF), imputation, PCSK9, eMERGE, biorepository

INTRODUCTION

Complete loss of function (LOF) variants are defined as variants expected to correlate with complete LOF of affected transcripts; i.e., nonsense mutations, splice site mutations, and insertion/deletion (indel) variants that result in downstream premature stop codons, or larger deletions removing either the first exon or more than 50% of the protein-coding sequence of the affected transcript (MacArthur et al., 2012). Partial LOF variants reduce gene activity but do not ablate it completely.

Data from the 1000 genomes project (1KGP), a large scale human genome sequencing study of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing data indicates that on average individuals carry ∼150 LOFs (Genomes Project Consortium et al., 2012). However, as detailed in **Table 1**, the majority of LOFs are common (>5%) and are distributed across a very small number (100–200) of genes. Genes containing common LOFs are strongly enriched for functional categories related to olfactory reception that are apparently unessential and do not result in any severe medical consequence. LOF enriched genes are typically depleted for genes implicated in protein-binding, transcriptional regulation, and anatomical development. Common LOFs are also enriched at the 3′ ends of genes as these mutations escape nonsense-mediated decay and are less subject to purifying natural selection. Finally, at the most highly conserved coding sites, more than 90% of stop-gain and splice-disrupting variants have a frequency below 0.5%. The population frequency of individual LOFs would therefore appear to correlate with their potential to adversely affect human health.

The 1KGP data indicates that each individual carries 10–20 LOF variants with a minor allele frequency (MAF) below 0.5% (**Table 1**). As these LOFs are under purifying selection they are less likely to be present in non-essential genes and at low conservation sites and therefore are likely to present pathological candidates.

The population frequency of rare variants differs considerably compared with common variation. Variants with frequencies above 10% were found in all of the populations studied in the 1KGP (Genomes Project Consortium et al., 2012), albeit with differences in MAF. Low-frequency variants in the 0.5–5% range were also largely shared between ancestral groups with only 17% of variants observed in a single ancestry group. For rare frequency variants with MAFs <0.5%, the majority (53%) were observed in a single population. Population stratification therefore represents a major confounder for rare variant analyzes which would ideally be controlled using principal component analysis from high-density GWAS arrays to select ancestrally matched cases and controls.

As a consequence of their rarity, LOFs will have largely been overlooked in GWAS studies which are best suited to the study of variants with minor alleles >3–5%. However, due to their rarity, very large, GWAS-type sample sets will be necessary to determine phenotypic association.

PCSK9 is expressed primarily in the liver, it is a secreted protein that acts by reducing the amount of low density lipoprotein receptor (LDLR) at the cell surface. Structurally, the PCSK9 protein product is composed a signal peptide, a prodomain, a catalytic domain, and a C-terminal domain. Cleavage of the prodomain is required for PCSK9 maturation and secretion. Cleaved PCSK9 is transported along the secretory pathway, which ultimately

Table 1 | Loss of function allele counts in 1,092 human genomes across three allele frequency bins.

	Allele frequency (%)		
Variant type	<0.5	0.5–5	>5
Stop-gain	3.9–10	5.3–19	24–28
Stop-loss	1.0–1.2	1.0–1.9	2.1–2.8
Indel frameshift	1.0–1.3	11–24	60–66
Splice site donor	1.7–3.6	2.4–7.2	2.6–5.2
Splice site acceptor	1.5–2.9	1.5–4.0	2.1–4.6

promotes LDLR degradation [for review see (Marais et al., 2012)]. One LOF missense mutation in PCSK9, Q152H, has been shown to impair cleavage and hence inhibit PCSK9 secretion (Mayne et al., 2011). The Q152H LOF mutation was shown to result in a 79% decrease in circulating PCSK9 and a 48% decrease in LDL-C in carriers compared with non-carriers (Mayne et al., 2011). The C679*X* mutation results in a processed, partially-folded protein that remains in the ER and is not secreted. As LDLR is degraded at the cell surface and endosomes, the C679*X* mutant has no activity toward the LDLR because of its inability to leave the ER and traffic to LDLR (Benjannet et al., 2006). R46L is also a LOF PCSK9 mutation, the R46L-PCSK9 undergoes near normal autocatalytic cleavage and is secreted, yet cells expressing the mutant displayed a 16% increase in of cell surface LDLR and a 35% increase in internalized LDL compared with WT-PCSK9, suggesting that R46L causes hypocholesterolemia through a decreased ability to degrade LDLR (Cameron et al., 2006).

Mutations in *PCSK9* were first identified in two French families with hypercholesterolemia that screened negative for mutations in both the LDLR and the apolipoprotein B (apoB) genes (Abifadel et al., 2003). The hypercholesterolemia *PCSK9* mutations were all missense variants that are thought to confer a gain of function as overexpression of *pcsk9* in the liver of mice produces hypercholesterolemia by reducing LDLR numbers (Lambert et al., 2006).

In 2005, causative LOF mutations in *PCSK9* were identified in individuals with low plasma LDL-C levels, the LOF variants were shown to be present in ∼2% of the African–American population but rare in European Americans (<0.1%; Cohen et al., 2005). LOF mutation carriers displayed reduced or no PCSK9 activity, and their plasma LDL-C levels were reduced by 40% compared with non-carriers. Further, coronary heart disease risk in those individuals was reduced by 88% compared to non-carriers (Cohen et al., 2006). This observation sparked interest in the biology of PCSK9 and led to the development of several LDL-reducing drugs (Stein et al., 2012).

While the cost of whole genome and exome sequencing experiments has dropped dramatically with improvements in yield from second generation sequencing technologies, very large scale studies remain prohibitively expensive. For sample sets with existing genotypes from dense whole-genome arrays, genotype imputation presents a viable alternative to direct sequencing. Data generated from large sequencing projects such as the 1KGP (Genomes Project Consortium et al., 2012) and the NHLBI exome sequencing project (ESP; Tennessen et al., 2012) is phased (Delaneau et al., 2012) and the haplotypes can be used as reference panel to impute missing variation into the sample genotype data (Howie et al., 2009). Recent improvements in imputation algorithms and the expansion of reference datasets have improved accuracy of imputation for even low MAF variants. Imputed data can then be annotated using tools developed for the annotation of sequencing data such as SnpEff (Cingolani et al., 2012) which determine the genomic location (i.e., exonic, intronic or intergenic, and the effects of variants, missense, nonsense etc. on known genes). Imputed LOF variants can then be assessed against binary phenotypes or quantitative laboratory values derived from patients electronic medical records (EMR).

We sought to determine if two PCSK9 LOF mutations that are present in the 1KGP data, the C679X nonsense mutation and the R46L missense mutation, could be imputed into our dataset and the previously reported association of the LOFs with decreased serum LDL-C replicated.

MATERIALS AND METHODS

The Center for Applied Genomics (CAG) at The Children's Hospital of Philadelphia (CHOP) maintains a biorepository of over 160,000 genotyped samples, 60,000 of which are pediatric samples randomly recruited from CHOP with complete EMRs. As a proof of principle, we imputed the proprotein convertase subtilisin kexin type 9 (*PCSK9*; NM_174936) LOFs C679X (dbSNP:rs28362286) and R46L (dbSNP:rs11591147) into a random selection of 8,028 unrelated samples of Northern European ancestry genotyped on the Illumina HumanHap 550 array from the CAG biorepository. The study was approved by the Institutional Review Board at the CHOP, and written informed consent for sample collection and DNA genotyping/sequencing was provided by the parents of all participating children.

Genetic ancestry was determined by computing principal components on the dataset using smartpca, a part of the EIGENSTRAT package, on 100,000 random autosomal SNPs in linkage equilibrium. Samples were clustered into 4 Continental ancestry groups (Caucasian, African including admixed African–American, Asian, and native American/admixed Hispanic) by K-means clustering using the kmeans package in R. The European ancestry grouping in our dataset mapped most closely to the HapMap CEU population of Utah residents with Northern and Western European ancestry from the CEPH collection[1].

Duplicate samples and cryptic relatedness were assessed by pairwise IBD. IBD values were generated for all 8,028 samples of Northern European ancestry using the plink genome command. A random sample from any pair with a PI_HAT value exceeding 0.3 was excluded from further analysis.

Imputation of untyped markers (∼39 M) was carried out using IMPUTE2 after prephasing with SHAPEIT. Each chromosome was prephased separately. Reference phased cosmopolitan haplotypes and recombination rates were obtained from the 1000 genomes project (1000 Genomes Phase I integrated variant set b37 March 2012 release). Imputation was carried out in 5Mb intervals using an effective population size of 20000 as recommended. As

[1] http://hapmap.ncbi.nlm.nih.gov

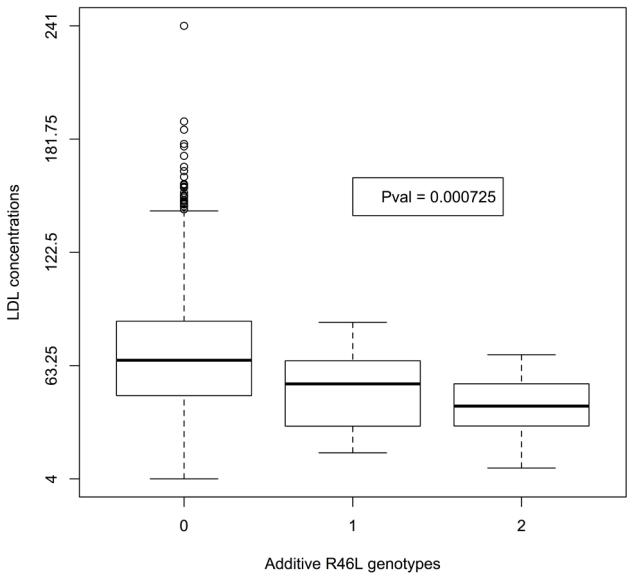


FIGURE 1 | Box and whisker plot of additively encoded *PCSK9* R46L genotypes on the X-axis versus age corrected serum LDL concentrations on the Y-axis ($n = 1432$; min LDL concentration 4; max LDL concentration 241; mean LDL concentration 85). Genotype 0 indicates no copies of the mutant allele, genotype 1 indicates one copy of the mutant allele and genotype 2 indicates 2 copies of the mutant allele.

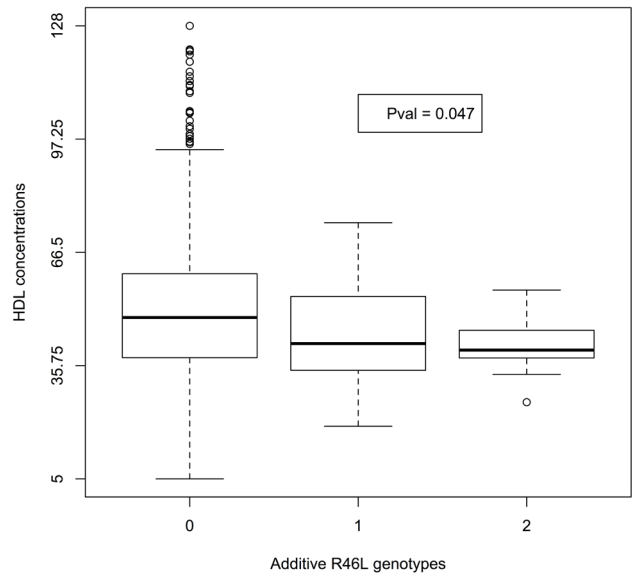


FIGURE 2 | Box and whisker plot of additively encoded *PCSK9* R46L genotypes on the X-axis versus age corrected serum HDL concentrations on the Y-axis ($n = 1495$; min HDL concentration 5; max HDL concentration 128; mean HDL concentration 47). Genotype 0 indicates no copies of the mutant allele, genotype 1 indicates one copy of the mutant allele and genotype 2 indicates 2 copies of the mutant allele.

a measure of the overall imputation accuracy we compared the concordance between the imputed and known genotypes in the subset of SNPs for which genotyping data was available. At a call threshold of 0.9, over 99% of the imputed genotypes were called and over 96% of those were concordant with the known genotypes.

RESULTS

Following imputation using SHAPEIT[2] and IMPUTE2[3] and annotation using SnpEff[4], we extracted and additively re-encoded genotypes for C679X and R46L from the 8,028 European American samples from the CAG biorepository. Both variants were imputed with high confidence, info scores C679X = 0.9 and R46L = 1. The C679X mutation was previously reported to be present in 0.1% of European Americans (Cohen et al., 2005). We identified nine C679X carriers out of 8,028 samples for a frequency of 0.11%, consistent with previous reports. As the samples were randomly selected from the biorepository, not all contained serum lipid data in their EMR. Three of the nine C679X carriers had serum LDL data. The frequency of the R46L was also consistent with the NHLBI ESP data, homozygous wild-type R46L 0.98 (1432 unique individuals with lab values mean age 12.1 years); heterozygous R46L 0.02 (10 unique individuals with lab values mean age 13.5) and homozygous derived allele R46L 0.001 (12 unique individuals with lab values mean age 11.5). A total of twenty-two R46L carriers had LDL data in the EMR.

There was insufficient data to assess the statistical significance of C679X genotypes. Linear regression of EMR-derived age-corrected serum LDL concentrations against R46L genotypes was statistically significant (P-value 7×10^{-4}) and directions of effect consistent with the LOF allele reducing LDL cholesterol (**Figure 1**). Serum HDL concentrations also showed a trend toward association (P-value 0.04; **Figure 2**). By contrast, serum triglyceride levels showed no association with R46L genotype (P-value 0.58; **Figure 3**) as previously described (Kotowski et al., 2006). The mean age-adjusted LDL concentration for R46 wild-type homozygotes was 85.7, mean age-adjusted LDL concentration for R46L heterozygotes was 63 and 62.6 for R46L homozygotes which corresponds approximately to a 26% decrease of serum LDL consistent with the 23.5 mean LDL-C difference previously reported in European American R46L carriers (Kotowski et al., 2006).

DISCUSSION

Recent genome sequencing studies have shown that each individual carries a significant number of variants that are predicted to result in a loss of protein function. The phenotypic effect of the majority of these LOFs remains to be determined. Here, we have shown a successful proof of concept that rare LOFs can be imputed into high density genotyping array data using data from large scale sequencing projects such as the 1KGP as a reference. While second generation sequencing remains prohibitively expensive in large numbers, high density genotyping data has been generated on hundreds of thousands of individuals. The eMERGE consortium biorepository includes ~60,000 individuals

[2]http://www.shapeit.fr

[3]http://mathgen.stats.ox.ac.uk/impute/impute_v2.html

[4]http://snpeff.sourceforge.net

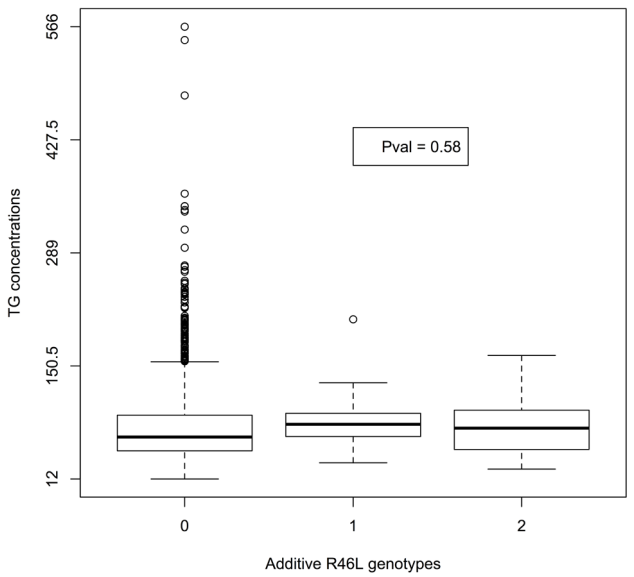


FIGURE 3 | Box and whisker plot of additively encoded *PCSK9* R46L genotypes on the X-axis versus age corrected serum triglyceride concentrations on the Y-axis ($n = 1856$; min TG concentration 12; max TG concentration 566; mean TG concentration 82). Genotype 0 indicates no copies of the mutant allele, genotype 1 indicates one copy of the mutant allele and genotype 2 indicates 2 copies of the mutant allele.

that have been genotyped on high-density GWA arrays (review at http://www.genome.gov/27540473), all of which has been linked with EMRs. As such eMERGE would be ideally suited for the assessment of rare LOF variants across multiple phenotypes either by direct assessment through single variant tests or through burden tests. For future analyses, in order to identify all possible association signals, the data would be analyzed using more than one statistical approach as detailed below.

Annotated, imputed variants, in vcf format[5], would be analyzed for association using both single point and agglomerative tests. Single variant tests for association against the EMR traits would be implemented in EMMAX (Kang et al., 2010), a mixed model algorithm that controls for both population substructure and relatedness between individuals in the test. In addition to the principal components for population stratification applicable covariates such as age could be included. For the agglomerative gene-based association tests, three complementary algorithms, the sequence kernel association test (SKAT; Ionita-Laza et al., 2013), the variable threshold test (Price et al., 2010) and the combined multivariate and collapsing (CMC) test which assess the burden of variation within the gene (Li and Leal, 2008) would be implemented. Gene-based association tests can achieve substantial increases in power to detect associations with rare variation compared with single variant tests (Ionita-Laza et al., 2013).

[5]http://www.1000genomes.org/wiki/analysis/variant-call-format/vcf-variant-call-format-version-42

We anticipate that for the single variant tests greatest power would be achieved against quantitative phenotypes such as lab values, however, gene burden scores could equally be applied using a pheWAS approach (Denny et al., 2010), i.e., EMR derived ICD9-based pseudo-case control analyzes for binary traits. These approaches will be validated on multiple LOF variants across the eMERGE networks in the near-future.

ACKNOWLEDGMENTS

We are grateful to the study volunteers for participating in the research studies and to the clinicians and support staff for enabling patient recruitment and blood sample collection. Informed consent was obtained from all participants or their parents or guardians. Sample genotyping was funded by an Institutional Development Award to the CAG from CHOP; imputation and association analyzes were funded by an eMERGE consortium 1U01HG006830-01 award from the NHGRI.

REFERENCES

Abifadel, M., Varret, M., Rabès, J. P., Allard, D., Ouguerram, K., Devillers, M., et al. (2003). Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat. Genet.* 34, 154–156. doi: 10.1038/ng1161

Benjannet, S., Rhainds, D., Hamelin, J., Nassoury, N., and Seidah, N. G. (2006). The proprotein convertase (PC) PCSK9 is inactivated by furin and/or PC5/6A: functional consequences of natural mutations and post-translational modifications. *J Biol. Chem.* 281, 30561–30572. doi: 10.1074/jbc.M606495200

Cameron, J., Holla, Ø. L., Ranheim, T., Kulseth, M. A., Berge, K. E., and Leren, T. P. (2006). Effect of mutations in the PCSK9 gene on the cell surface LDL receptors. *Hum. Mol. Genet.* 15, 1551–1558. doi: 10.1093/hmg/ddl077

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., et al. (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. doi: 10.4161/fly.19695

Cohen, J. C., Boerwinkle, E., Mosley, T. H. Jr., and Hobbs, H. H. (2006). Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* 354, 1264–1272. doi: 10.1056/NEJMoa054013

Cohen, J., Pertsemlidis, A., Kotowski, I. K., Graham, R., Garcia, C. K., Hobbs, H. H. (2005). Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.* 37, 161–165. doi: 10.1038/ng1509

Delaneau, O., Marchini, J., and Zagury, J. F. (2012). A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9, 179–181. doi: 10.1038/nmeth.1785

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., et al. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 26, 1205–1210. doi: 10.1093/bioinformatics/btq126

Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65. doi: 10.1038/nature11632

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5:e1000529. doi: 10.1371/journal.pgen.1000529

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., and Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* 92, 841–853. doi: 10.1016/j.ajhg.2013.04.015

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. doi: 10.1038/ng.548

Kotowski, I. K., Pertsemlidis, A., Luke, A., Cooper, R. S., Vega, G. L., Cohen, J. C., et al. (2006). A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. *Am. J. Hum. Genet.* 78, 410–422. doi: 10.1086/500615

Lambert, G., Jarnoux, A. L., Pineau, T., Pape, O., Chetiveaux, M., Laboisse, C., et al. (2006). Fasting induces hyperlipidemia in mice overexpressing proprotein convertase subtilisin kexin type 9: lack of modulation of very-low-density

lipoprotein hepatic output by the low-density lipoprotein receptor. *Endocrinology* 147, 4985–4995. doi: 10.1210/en.2006-0098

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. doi: 10.1016/j.ajhg.2008.06.024

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335, 823–828. doi: 10.1126/science.1215040

Marais, D. A., Blom, D. J., Petrides, F., Goueffic, Y., and Lambert, G. (2012). Proprotein convertase subtilisin/kexin type 9 inhibition. *Curr. Opin. Lipidol.* 23, 511–517. doi: 10.1097/MOL.0b013e3283587563

Mayne, J., Dewpura, T., Raymond, A., Bernier, L., Cousins, M., Ooi, T. C., et al. (2011). Novel loss-of-function PCSK9 variant is associated with low plasma LDL cholesterol in a French-Canadian family and with impaired processing and secretion in cell culture. *Clin. Chem.* 57, 1415–1423. doi: 10.1373/clinchem.2011.165191

Price, A. L., Kryukov, G. V., de Bakker, P. I. W., Purcell, S. M., Staples, J., Wei, L.-J., et al. (2010). Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* 86, 832–838. doi: 10.1016/j.ajhg.2010.04.005 +

Stein, E. A., Olsson, A. G., Scott, R., Kim, J. B., Xue, A., Gebski, V., et al. (2012). Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *N. Engl. J. Med.* 366, 1108–1118. doi: 10.1056/NEJMoa1105803

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337, 64–69. doi: 10.1126/science.1219240

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 January 2014; accepted: 10 April 2014; published online: 29 April 2014.
*Citation: Sleiman P, Bradfield J, Mentch F, Almoguera B, Connolly J and Hakonarson H (2014) Assessing the functional consequence of loss of function variants using electronic medical record and large-scale genomics consortium efforts. Front. Genet. **5**:105. doi: 10.3389/fgene.2014.00105*
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

REVIEW ARTICLE
published: 02 June 2014
doi: 10.3389/fgene.2014.00162

Genetic-based prediction of disease traits: prediction is very difficult, especially about the future†

Steven J. Schrodi[1]*, Shubhabrata Mukherjee[2], Ying Shan[3], Gerard Tromp[4], John J. Sninsky[5], Amy P. Callear[1,6], Tonia C. Carter[1], Zhan Ye[7], Jonathan L. Haines[8], Murray H. Brilliant[1], Paul K. Crane[2], Diane T. Smelser[4], Robert C. Elston[8] and Daniel E. Weeks[3]

[1] Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, WI, USA
[2] Department of Medicine, School of Medicine, University of Washington, Seattle, WA, USA
[3] Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh, PA, USA
[4] Sigfried and Janet Weis Center for Research, Geisinger Health System, Danville, PA, USA
[5] Subsidiary of Quest Diagnostics, Discovery Research, Celera Corporation, Alameda, CA, USA
[6] Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA, USA
[7] Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI, USA
[8] Department of Epidemiology and Biostatistics, Case Western Reserve School of Medicine, Cleveland, OH, USA

Edited by:
Marylyn D. Ritchie, The Pennsylvania State University, USA

Reviewed by:
Andrew Skol, University of Chicago, USA
Hui-Qi Qu, The University of Texas School of Public Health, USA

****Correspondence:***
Steven J. Schrodi, Center for Human Genetics, Marshfield Clinic Research Foundation, 1000 N Oak Ave. Marshfield, WI 54449, USA
e-mail: schrodi.steven@mcrf.mfldclin.edu

Translation of results from genetic findings to inform medical practice is a highly anticipated goal of human genetics. The aim of this paper is to review and discuss the role of genetics in medically-relevant prediction. Germline genetics presages disease onset and therefore can contribute prognostic signals that augment laboratory tests and clinical features. As such, the impact of genetic-based predictive models on clinical decisions and therapy choice could be profound. However, given that (i) medical traits result from a complex interplay between genetic and environmental factors, (ii) the underlying genetic architectures for susceptibility to common diseases are not well-understood, and (iii) replicable susceptibility alleles, in combination, account for only a moderate amount of disease heritability, there are substantial challenges to constructing and implementing genetic risk prediction models with high utility. In spite of these challenges, concerted progress has continued in this area with an ongoing accumulation of studies that identify disease predisposing genotypes. Several statistical approaches with the aim of predicting disease have been published. Here we summarize the current state of disease susceptibility mapping and pharmacogenetics efforts for risk prediction, describe methods used to construct and evaluate genetic-based predictive models, and discuss applications.

Keywords: predictive model, genetic risk, human genetics, prognosis, clinical utility

INTRODUCTION AND BACKGROUND

Multiple lines of evidence strongly support the notion that the large majority of common, chronic diseases have complex causes. Environmental components such as infection, caloric flux, and chemical exposure, along with heritable elements such as DNA variants, methylation patterns, and epigenetic RNA effects, are interacting co-conspirators resulting in common diseases. In this background of convoluted and entangled etiology, discovery and use of disease predisposing alleles present a considerable challenge to the human genetics community (Clerget-Darpoux and Elston, 2013). Recent technological advances in high-throughput genotyping, RNA expression, and massively parallel sequencing have accelerated interrogation of genetic variation for the purpose of understanding human disease and drug response. Among the more important uses of these discoveries is providing detailed, mechanistic insight into the molecular pathogenesis of disease states. The two primary avenues of utilizing this explosion in genetic information for the purpose of improving clinical practice are in (1) drug development stemming from the identification of molecular targets and (2) the prediction of disease susceptibility, pharmacogenetic response, and disease severity/trajectory (Khoury et al., 1985; Holtzman and Marteau, 2000; Evans and Relling, 2004). Although only a small minority of current pharmaceuticals originated directly from genetic findings serving as drug targets, the list is expanding and includes inflammatory cytokine-based monoclonal antibodies and targeted cancer therapeutics, among others. These therapeutics often target specific biochemical pathways to improve clinical treatment, often with a reduction in adverse reactions. Disease prediction and diagnosis with genetic testing is a broad field with diverse applications, ranging from karyotyping for chromosomal abnormalities to enhancement of disease risk profiles using single nucleotide polymorphisms (SNPs) previously found to be disease-susceptibility markers, such as *HFE* missense polymorphisms which can lead to hemochromatosis, or the variants in the tumor suppressors *BRCA1* and *BRCA2* that increase risk to breast and ovarian cancers. Clinical genetics testing can provide physicians with an additional tool for better diagnosis and improved medical care.

Much of the variation in disease course, severity, and response to medication is reflective of the underlying allelic repertoire

†This is inspired by a humorous quote that is variably attributed to Mark Twain, Niels Bohr, the Danish Parliament, Samuel Goldwyn, and Yogi Berra.

existing in each individual, offering the opportunity for genetics to facilitate early treatment, preventative medicine, preemptive selection of efficacious drugs, and more accurate estimation of risk for those thought to be at intermediate risk using traditional factors. As the cost and complexity of medical care escalates, the promise of human genetics to provide directly actionable, individualized information to address impediments to optimal and cost-effective medical practice carries increasing weight and urgency (Chen and Snyder, 2013). This review has multiple aims: (1) provide a brief overview of the current state of human disease mapping as this provides the foundational knowledge for genetic-based disease prediction, (2) describe the process of disease prediction in a simple probabilistic framework detailing the general qualities of clinically useful predictive models and also detailed examples, (3) provide an overview of the basic classes of genetic-based prediction models and measures of prognostic utility, and (4) illustrate the application of genetic-based predictive models to data from biobanks and prospective cohorts.

Identification of replicated susceptibility variants provides considerable material for understanding biochemical pathways that govern diseases, particularly when the variants reside within the coding or regulatory regions of well-understood genes and are validated by functional studies (Manolio, 2010). Unfortunately, many disease-associated variants are located in regions of the genome that have not yet been functionally characterized. Indeed, 39% of the National Human Genome Research Institute (NHGRI) Genome-Wide Association Study (GWAS) catalog SNPs are annotated as intergenic and more than 36% are reported as intronic (Welter et al., 2014). The genes and pathways discovered can become targets for pharmaceutical intervention, especially when integrated with corroborating studies from disease models, signal transduction experiments, bioinformatics, and protein biochemistry. Examples of using specific genes or their products as pharmaceutical targets have rapidly accumulated over the past decade and include mipomersen, an antisense therapeutic targeting *APOB* RNA for the treatment of hypercholesterolemia (Raal et al., 2010), ivacaftor which targets the G551D mutation in *CFTR* found in approximately 4% of individuals with cystic fibrosis (Ramsey et al., 2011), inflammatory cytokines and their receptors (e.g., IL-1β, IL-12/23p40, IL-17A, IL-6R) (e.g., Krueger et al., 2007), other immune cell signaling proteins (e.g., CTLA-4, CD30), a variety of tumorigenesis genes that harbor somatic variants useful for individualized cancer treatment (e.g., *BRAF*, *KRAS*, and *EGFR*), and lipid transfer proteins (e.g., CETP, PCSK9), among many others. As more human genetics studies are conducted, the number of these druggable targets will expand. While the use of genetic results in pharmaceutical development is impressive, some of the most highly touted uses of genetic susceptibility data have been the accurate prognosis of diseases (e.g., Mendelian and oligogenetic disorders, such as Tay-Sachs disease, phenylketonuria, Charcot-Marie-Tooth, or rare ciliopathies including polycystic kidney disease and Bardet-Biedl syndrome), or other areas that impact medical decisions, such as choice of drug, selection of dose, avoidance of side-effects, or determining the optimal intensity of clinical monitoring. Unlike identifying potential drug targets, genetic-based prediction models may serve a clinical purpose in advance of precise identification of the functional motifs and molecular mechanisms that drive genetic association/linkage signals. Instead, the utility of predictive models is derived primarily from the correlation patterns—provided that these are robust across intended populations. However, the strength and robustness of the correlation are critical for a genetic prediction model to be clinically useful.

As clinical decisions are specific to individuals, physicians aim to assess the probability of medical traits for each patient. This is a dynamic process where physicians update assessments as additional relevant information, such as laboratory tests (both genetic and non-genetic), or changes in physiology become known. In this way, clinical decisions are informed as variation in an individual's risk to disease, severity of disease and response to medication are progressively revealed. Thus, results from clinical tests, including genetic-based predictive models, are useful when they more accurately discern the likelihood of the medical trait (e.g., disease occurrence or response to medication), compared to the pre-test assessment. For example, if a physician had estimated that a patient had a 40% chance of having a particular disease prior to the results of a clinical test, and the 40% prediction remains unaltered following the results of the test, then the clinical test and new prediction may be of little value. Further, whether the magnitude of this posterior-prior probability departure carries clinical utility depends on the specific application. As an illustration of this process, suppose a patient is referred to a rheumatologist. Prior to the visit, the rheumatologist may not have sufficient information to modify the assessment of the probability that the individual has, for example, rheumatoid arthritis (RA). Upon learning that the patient self-reported symptoms of symmetric sore joints that are partially remediated by non-steroidal anti-inflammatory medication, the rheumatologist proceeds to update the probability of RA and of other conditions. Some diseases would increase in their likelihood, while others would decrease from their initial values. Following a standard evaluation of the classification criteria for rheumatoid arthritis assessing joint involvement, serology, acute-phase reactants, and symptom duration (Aletaha et al., 2010), the rheumatologist proceeds to further update the probability that the patient suffers from RA. Subsequent testing of a genetic panel of known RA-susceptibility markers, including polymorphisms within *HLA-DRB1*, *PTPN22*, *STAT4*, *CTLA4*, *TRAF1*, *CD40*, etc., may further modify the posterior probability. This additional updated posterior probability may be particularly useful in situations where a definitive diagnosis was not available with non-genetic approaches alone. This is, of course, neither a new nor complete account of the diagnostic process, but it underscores the general nature of many medical decisions, where accumulation of information typically results in increasing accuracy in the appraisal of a medical trait probability for an individual. The process of serial refinement based on accumulating data is a hallmark of the diagnostic process and, statistically, can be codified as Bayesian updating of the posterior probability of the trait. The aim of genetic-based predictive models is to augment existing laboratory, imaging, and other clinical data to improve the posterior probabilities (i.e., drive the posterior probabilities toward 0 or 1) of medical traits in a cost-effective manner.

In the context discussed here, predictive models are methods designed to use clinical, analyte, genetic, or other types of data for

the purpose of forecasting a medical trait. Predictive models—including those based on genetic markers—are most beneficial when they yield actionable and individualized results. However, they are of reduced value if they only substantially modify medical decisions for an exquisitely small fraction of the patient population. Hence, the ideal genetic-based predictive model for clinical applications (1) substantially modifies the posterior probability of medical traits compared to that obtained from existing clinical assessment and tests—enough to enable changes in medical decisions and patient management, and (2) impacts a substantial fraction of individuals to whom it is applied and provides improved outcomes. While other considerations are essential, such as more cost-effective care and the ease of adoption and implementation of diagnostic tests, it is this concurrent maximization of (1) modifying the posterior probability of the trait in the context of the benefits and risks of the specific medical decisions and (2) broad applicability that defines an archetypal genetic-based predictive model. For example, genetic testing of *CFTR* mutations for cystic fibrosis is successful in that the recessive disease alleles have very high penetrance and the large majority of pathogenic mutations are covered with contemporary panels. Similarly, multi-gene tests for related rare diseases, each with high penetrance, can also serve as useful clinical tests (Rehm, 2013).

In an attempt to develop such predictive models, many have used genome-wide association study (GWAS) results as they are a ubiquitous source of genetic information (Manolio, 2010). Attempts to use genetic information alone have not been as successful as previously hoped, with posterior probabilities that do not approach 1 or 0 and the vast majority of individuals having decidedly intermediate posterior probabilities. A seminal question is the extent to which genetic information can further modify posterior probabilities for those individuals thought to be of intermediate risk using traditional factors. Wray et al. offer an excellent review of the challenges involved in complex trait prediction with GWAS results (Wray et al., 2013).

The discovery of genetic markers for the prediction of medical traits is entirely dependent on the underlying genetic model that gives rise to the trait. That is, the number of loci and the number, frequency and penetrance of predisposing alleles determine both the likelihood of identifying causal markers and the clinical utility of using those markers in a patient population. For example, monogenic disorders such as phenylketonuria, Tay-Sachs, or sickle cell anemia are likely fully penetrant with allele frequencies that are not exceedingly rare; and therefore genetic tests for such diseases have clinical applications, provided that disease avoidance or disease-modifying treatments exist. However, traits like Alzheimer's disease, diabetes, or response to statins have etiologies that remain enigmatic. Whether or not these complex traits follow extremely polygenic modes of inheritance (i.e., weakly penetrant alleles, and several hundreds to thousands of loci), high locus/allelic heterogeneity (having highly penetrant but unique loci and alleles involved across individuals), high levels of epistasis (e.g., genotypic effects that vary based on genetic background or other specific genotypes), ubiquitous epigenetic effects (e.g., methylation patterns, histone acetylation patterns, or transgenerational RNA artifacts affecting the trait), gene-environment interactions, or some combination thereof, directly impacts the identification of predictive markers as well as their utility. GWAS interrogate the common allelic architecture for disease predisposing markers exhibiting low degrees of allelic and locus heterogeneity, whereas sequencing-based studies in families can facilitate the discovery of rare disease-associated variants, but are not optimal for identifying ancestral disease-predisposing alleles. Therefore, it is reasonable to expect that genetic markers from GWAS may modify posterior probabilities across a large segment of the population, but with a muted impact on those probabilities. On the other hand, rare sequence variants, on the other hand, may have substantial impact on the posterior probabilities for specific individuals, but with little widespread effect.

A review of the potential of genetic-based predictive models to change medical practice in the short-term indicates that three areas have shown promise for improving clinical care: cancer genomics, population screening for Mendelian diseases, and pharmacogenetics. These three areas profit from high penetrances of the genetic variants identified to date, though only a fraction of patients benefit from these tests. As these areas emerge from their infancy and additional genetic results accumulate, the proportion of individuals benefiting will likewise increase.

PREDICTION USING TUMOR GENOMICS

The advent of genetic testing in tumor cells, through harnessing the throughput and read depth of next-generation sequencing platforms, has enabled detailed and clinically actionable molecular pathology genetic tests for numerous cancers. Multiplex sequencing-based assays for biopsies compared to normal tissue are now available and have demonstrated usefulness in augmenting many clinical decisions. The utility of these tests relies on the clear relationship that has been delineated over the past two decades between specific driver mutations, treatment variants and cancer progression, and drug selection (Liaw et al., 1997; Paez et al., 2004; Agrawal et al., 2011; Walter et al., 2012; Kandoth et al., 2013; Vogelstein et al., 2013). Intratumor (Gerlinger et al., 2012) and single-cell sequencing methods (Navin et al., 2011) offer the possibility of inferring the evolutionary history and driver mutations in clonal expansions of cancer cells. These techniques have been successfully applied to several cancers with excellent prognostic utility, for example, kidney cancer (Xu et al., 2012). For well-defined activating mutations such as those within *BRAF* (Loupakis et al., 2009; Borras et al., 2011), *KRAS* (Linardou et al., 2008) and *EGFR* (Lynch et al., 2004), the posterior probability of efficacious treatment selection is also high. Indeed, there seems to be a clear path to incorporating panels of well-defined oncogenesis, metastasis, and drug response variants through next-generation sequencing of tumors. Baylor College of Medicine, one institution among several offering a number of clinical genetics tests, has developed a Cancer Gene Mutation Panel through next-generation sequencing that investigates 2855 known mutations within 50 cancer-associated genes for clinical testing (http://www.bcm.edu/cancergeneticslab/test_detail.cfm?testcode=9705). Other efforts include the UCLA Clinical Genomics Center (http://pathology.ucla.edu/body.cfm?id=105), the Emory Genetics Laboratory

(http://genetics.emory.edu/egl/), and the Washington University School of Medicine (http://gps.wustl.edu/). Identification of a small number of specific mutations enables selective treatment courses to be taken with higher expected efficacy, albeit often with limited duration of effect due to the development of drug resistance, an expected consequence of monotherapy. For example, in this Baylor panel *BRAF* mutations are targeted, where treatment with vemurafenib and dabrafenib has demonstrated *BRAF* Val600-specific metastatic melanoma antitumor activity (Jang and Atkins, 2013). Over the past five years, somatic cell and tumor genomics has provided remarkable insights into the molecular pathobiology of cancers. This rapidly progressing field continues to accumulate examples of improved treatment resulting from these genetic discoveries.

PREDICTION IN SCREENING FOR MENDELIAN DISORDERS

Equally impressive has been the progress in interrogating very highly penetrant alleles in population-based screens, particularly in newborns. Next-generation sequencing has enabled rapid, cost-effective multiplex assays that require little DNA. Given the high positive predictive value of these variants and the ability to modify clinical treatment in many of these Mendelian disorders, genetic-based prediction in this area is an efficacious addition to medical practice. For example, Saunders et al. recently showed the feasibility of screening for monogenic diseases across the genome within 50 h in a neonatal clinical setting (Saunders et al., 2012). Importantly, infants identified as having pathogenic genotypes (e.g., Kwan et al., 2013; Stefanutti et al., 2013) can receive appropriate treatment while still hospitalized, often avoiding life-threatening complications. Comprehensive genetic testing may preclude emotionally and financially costly pediatric odysseys (Kingsmore et al., 2011). In addition, the application of high-throughput sequencing approaches to clinically important, expansive gene panels can reliably identify known inherited pathogenic variants and new germline mutations that are potentially pathogenic, thereby driving effective early screening (Kurian et al., 2014).

PREDICTION IN PHARMACOGENETICS

Pharmacogenetics is the third area in which genetic variants can enable physicians to differentially prescribe certain medications to individuals to avoid adverse events or to modify dosing. The importance of these genetic variants in avoiding adverse drug reactions is underscored by FDA black-box warnings (http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm), as well as by recommendations of other groups (https://www.pharmgkb.org/). For example, individuals carrying the HLA-B*5701 allele are warned against taking abacavir (Mallal et al., 2008), dapsone-treated patients with certain *G6PD* variants are at higher risk for hemolysis as are patients receiving primaquine, many sulfonamides, nitrofurantoin, acetanilide, niridazole, and naphthalene (Cappellini and Fiorelli, 2008), and the beta blocker propranolol can cause adverse reactions in those with variants conferring compromised CYP2D6 function (Cascorbi, 2003; Samer et al., 2013). In all, the FDA currently lists 155 pharmacogenetic warnings across numerous therapeutic areas. Again, the validated, high positive predictive value of these pharmacogenetic variants makes immediate clinical utility possible, if not immediately actionable. Clinically useful genetic variants underlying other pharmacogenetic traits, such as differential response to many lipid-modifying medications, metformin, and anti-TNF therapies, still remain largely abstruse.

Importantly, the setting of clinical application of genetic tests is critical to the usefulness of genetic-based predictive models. Traits, including drug response and adverse reactions, that are (1) otherwise easily diagnosed, or (2) for which disease management would not change with the results of the predictive model, are poor candidates for these predictive models. So, results from the use of genetic-based predictive models must serve as a cog in the health management machinery and clearly satisfy an unmet medical need. For example, genetic-based predictive models are unlikely to play a useful role in diagnosing a bone fracture. Similarly, Kimmel et al. recently showed that even though genetics can fairly accurately predict warfarin sensitivity, this information offers no benefit over clinical management of warfarin dosing to achieve therapeutic range (Kimmel et al., 2013). The setting of medical care also plays an important role: nearly half of all patients are not treated in coagulation centers, leaving the question of how diagnostic genetic testing would fare in those environments.

Why is it that these three above areas have enjoyed more success in applying genetic information to clinical practice than other applications, such as prognosis of complex diseases? In large part, the answer lies in the relatively low complexity of the genetic architecture behind these medical traits. The propagation of cancer cells, tumor survival and metastasis are promoted by specific mutations that wield strong effects on promoting clonal expansions: driver mutations. Different driver mutations accomplish this task in different ways, but each driver mutation has profound effects on cellular metabolism, mitosis, and proliferation. Because the effects of these driver mutations are profound and characteristic of specific molecular pathophysiologies, it is not surprising that they are reasonably predictive of disease trajectory and chemotherapy response. Similarly, provided that the false positive rate of prognostic tests is low, population-based screens for Mendelian disorders have been a useful addition to modern medical practice because the penetrance of such traits is typically complete or nearly complete. That is, aside from the measurement error rates, the prediction of disease given a positive genetic test is accurate and reliable. Finally, although not as definitive as Mendelian disorders, pharmacogenetic effects identified to date testify to reduced complexity of these traits. This is particularly true of extreme adverse events (e.g., FDA black box warnings) and response with those drugs having highly targeted substrates. In contrast, therapeutics with multifold actions, such as statins or metformin, have exhibited much more recalcitrance to genetic dissection.

In contrast to the above-mentioned areas, currently the prediction of common diseases presents a considerable challenge. Most common diseases have been relatively reluctant to reveal a large fraction of the genetic component of their etiologies (Manolio et al., 2009). Several studies of complex diseases have shown little improvement to disease prediction when adding genetic data

to already established disease risk factors (e.g., Thanassoulis and Vasan, 2010; Bao et al., 2013; Muhlenbruch et al., 2013); and, even if statistically significant, the models incorporating genetic information may not be clinically useful (Husing et al., 2012). While there are several instances of important, influential markers that have been discovered in some common diseases, such as *APOE* in Alzheimer's, *ARMS2* and *CFH* in age-related macular degeneration (AMD), and numerous alleles in the MHC region for autoimmune and inflammatory diseases, many genetic linkage results are the result of multiple infrequent alleles and most replicated markers from GWAS have modest effect sizes. In combination, the replicated disease susceptibility alleles discovered thus far have yet to demonstrate substantial prognostic utility. That said, there are encouraging exceptions: the combined effect of the multiple identified loci for AMD or Crohn's disease may offer some clinical utility in selected circumstances. AMD is a leading cause of compromised vision and blindness, and individuals at heightened risk for AMD can benefit from more frequent eye exams and early treatment to curb the likelihood of permanent ophthalmic damage. Administration of anti-VEGF monoclonal antibodies have shown efficacy in exudative AMD treatment (Fung et al., 2007; Heier et al., 2012). Recent GWAS studies in AMD have demonstrated that the 19 top AMD risk loci are estimated to explain between 15 and 65% of the genetic portion of the variance in the phenotype (the proportion depends on the assumption of AMD prevalence being between 0.01 and 0.10). This set of SNPs also generates an area under the ROC curve (AUC) of 0.74 (Fritsche et al., 2013), meaning that if you choose pairs of people at random, one with and one without AMD, and used their SNP data, the one with a higher probability of AMD would in fact be the one with AMD 74% of the time (Berrar and Flach, 2012). Incorporation of other known risk factors, such as age and smoking, further improves this prediction. It is possible that other measures, including positive and negative predictive values or those based on posterior probability distributions, could provide better insight into clinical utility. Another promising area is the use of all genetic variants genotyped in a genome-wide array to construct predictive models, rather than restricting the markers to those that are strongly associated with the trait. Purcell et al. investigated the use of thousands of common alleles in predictive models for schizophrenia and bipolar disorder, demonstrating an increase in the proportion of the maximum variance in these traits explained as the trait-association significance level was relaxed (Purcell et al., 2009). In addition, analysis of the Wellcome Trust Case Control Consortium data for Crohn's disease appeared to indicate that expansion of the number of SNPs in a predictive model over just those reaching genome-wide significance improves the model performance (Kooperberg et al., 2010). These are interesting observations and consistent with results from Yang et al. (2010) and Lee et al. (2011) describing the rather dramatic increase in the proportion of heritability explained with all GWAS SNPs compared to top SNP findings. Wei et al. offer a recent example of harnessing this effect for Crohn's disease where expanding the number of variants and using advanced machine learning techniques increased predictive accuracy (Wei et al., 2013). With greater resolution, Yang et al. (2011) showed that the length of chromosomes is linearly correlated with the percentage of the variance attributable to a variety of phenotypes, including von Willebrand factor, height, BMI and QT interval. However, both theoretical and applied work appears to show limited utility of including more than a few hundred SNPs in commonly-used predictive models (Wu et al., 2013; Warren et al., 2014). Nonetheless, methods that exploit the whole genome for disease prediction, such as extensions to Best Linear Unbiased Prediction (BLUP), continue to develop and may improve accuracy metrics for both binary disease and quantitative traits (Zhang et al., 2014).

UTILITY AND METHODS

Given that for many diseases with effective treatments accurate prediction of potential disease can play a critical role in determining robust clinical care that may avert severe disease or even disease onset, it is essential to characterize the important aspects that produce useful predictive models. For traits with polygenic etiologies, methods must be used to combine signals from multiple genetic markers together into a cohesive metric for prediction (Wimmer et al., 2013). Seven main considerations when doing so are: (1) which genetic markers are to be included in the predictive model—i.e., feature selection, (2) the frequencies of the susceptible/protective genotypes at each selected marker, (3) the strength of the correlation between the genotypes at each marker and the predicted trait, (4) the interactions between the effects sizes of different genetic markers, (5) the prevalence of the trait being predicted, (6) how the genetic data are envisioned to integrate into clinical practice in combination with non-genetic tests, and (7) a determination of the robustness of the prognostic signal across multiple populations, including those with varied ancestries. Over the past decade, several methods have been proposed to accomplish these tasks, including genetic risk scores, various types of regression-based approaches, Bayesian networks, and other machine learning methods. Importantly, polygenic disease-prediction models may serve as instrumental variables for Mendelian randomization analyses in the investigation of the causal role of genetic-based predictors in disease (Burgess and Thompson, 2013).

FEATURE SELECTION

Feature selection refers to the decision about which genetic variants are most effective in determining the medical trait and should therefore be included in a predictive model. For example, it would seem reasonable to include SNPs in the *CFH*, *ARMS2*, *C3*, and *C2/CFB/SKIV2L* regions in a model predicting AMD because the evidence for correlation between AMD and these variants is both substantial and well-established. Further, selection of these variants for inclusion in a predictive model would be prioritized over other variants with little or no evidence of utility in AMD prediction. Jakobsdottir et al. have investigated the properties of individual disease-susceptibility SNPs, showing that SNPs with highly significant odds ratios may be insufficient to classify individuals (Jakobsdottir et al., 2009). There are several different methods that can be employed. For a general review see Guyon and Elisseff (2003). Care must be taken when internal validation techniques are applied to datasets, as the feature selection must be incorporated in the internal validation

routine. Ideally, feature selection should be replicated in an independent sample set. Approaches based on stepwise selection of features are popular. The performance of models constructed based on a stepwise selection can be evaluated based on model fit, accounting for the complexity of the model—the Akaike and Bayesian information criteria are examples of measures to do this (Akaike, 1974; Schwartz, 1978). Aside from purely statistical and computational approaches, use of biological information can improve the selection of genetic markers. By integrating information from numerous decades of biochemistry, molecular biology and cellular physiology—the direct phenotypes of genetic variants—one can construct predictive models weighted toward those variants segregating in functionally relevant regions in an effort to improve the robustness of the model and ease of application to related phenotypes. For example, if one is generating a genetic-based predictive model for Crohn's disease response to IL-17 monoclonal antibody therapy, higher prioritization of variants within IL-17-related genes or those polymorphisms that are known to modify T-helper cells expressing IL-17 (Th17) activity may provide complementary information and yield a higher likelihood of the test having utility when applied to other populations or related phenotypes.

GENETIC RISK SCORES

Genetic Risk Scores (GRS), determined simply on the basis of published GWAS results, are among the simplest methods employed for genetic prediction. The majority of these approaches construct the predictive model based on the sum of predisposing genotypes that each individual carries, either unweighted or weighted by the effect size of the specific predisposing genotypes. The essential approach is to take a weighted sum of risk alleles, choosing the risk alleles based on those found to be genome-wide significant in a recent meta-analysis (e.g., for BMI, see Speliotes et al., 2010). Weights are determined for each risk allele by the β estimates from the meta-analyzed GWAS. Unweighted GRS treat each risk locus equally. To illustrate the weighted GRS approach, assume that k SNPs are known to be genome-wide significant and further assume that the corresponding β weights from the GWAS are denoted as w_i for the ith SNP. Then the be calculated as: $GRS = \sum_i^k w_i R_i$; where R_i is the number of risk alleles at the ith SNP. Speliotes et al., using 32 confirmed obesity-associated loci, showed the distribution of the weighted number of risk alleles across the population used in the Atherosclerosis Risk in Communities (ARIC) study, and presented a corresponding AUC for the GRS (Speliotes et al., 2010). Although significantly different from that expected under the null, the AUC for this example was exceedingly modest (0.515), where flipping an unbiased coin would be expected to have an AUC of 0.500. In another example, Ripatti et al. developed a genetic risk score based on 13 SNPs discovered to be associated with coronary heart disease, myocardial infarction or both, in seven reports (Ripatti et al., 2010). For each individual, the effects of these SNPs were combined by summing the number of risk alleles and the resulting risk score was partitioned into quintiles for the purpose of creating a categorical variable. Comparing extreme quintiles, the authors found roughly a 1.7-fold increased risk for coronary heart disease in the top risk quintile compared to the lowest risk. The genetic risk score did not show a significant effect of the net reclassification of individuals over traditional risk factors and family history. The combined genetic effect was able to slightly improve the risk classification of those individuals who were previously thought to have intermediate risk as determined by traditional risk factors, but may not have strong clinical utility. Increasing the number of informative genotypes and/or the traditional risk factors may improve the prognostic performance of GRS. Other applications, including age-related macular degeneration, exhibit more promising performance (Grassmann et al., 2012; Seddon et al., 2014).

REGRESSION METHODS

Regression methods, familiar tools for constructing prediction models for both dichotomous and quantitative traits, can lead to more general predictive models than simple GRSs. One of the first reports of a cohesive method using multiple replicated markers under a general logistic regression model was developed by Yang et al. (2003). Yang and coworkers proposed using a general logistic regression model to estimate the ratio of the probability of the genotype information given disease to the probability of the genotype information within the non-diseased population. Incorporation of covariates and interaction effects are possible with this generalized form. Currently, regression is still commonly used for disease prediction. For example, a search of PubMed revealed 10 articles published in 2013 which applied regression methods for the prediction of a variety of diseases, including cerebrovascular disease, age-related macular degeneration, and hypertrophic cardiomyopathy (Abraham et al., 2013; Borque et al., 2013; Gruner et al., 2013; Harada et al., 2013; Mondul et al., 2013; Romano et al., 2013; Schellekens et al., 2013; Sharma et al., 2013; Tsai et al., 2013; Uddin et al., 2013). In addition, extensions including regression of the whole genome using a Best Linear Unbiased Prediction method (G-BLUP) can produce more highly predictive models (de Los Campos et al., 2013). Importantly, Yang et al. (2009) pointed out that one should not rely on point estimates alone, but also provide a measure of the uncertainty of the risk estimates. Risk estimates depend on a variety of parameters, each of which may be estimated with some uncertainty. Cumulative uncertainty across all estimated parameters leads to uncertainty of the risk estimates.

There are several modeling assumptions made when applying either linear or logistic regression but, in the specific application area of genetics, the following concerns should be emphasized. First, multicollinearity between nearby markers is usually a serious concern. For markers in high linkage disequilibrium with each other, it is common to select the variant with the lowest p-value for inclusion in the model. Principal component regression is another useful way to address concerns arising from multicollinearity. For example, Gauderman et al. found that this approach performs well when applied to a single candidate gene (Gauderman et al., 2007). Another concern is marker-marker interactions. For parsimony, it is common practice to ignore interactions. Interaction analysis is not easy to conduct and can be computationally intensive. Furthermore, substantially larger sample sizes are typically needed to detect interaction effects than

are needed to detect main effects. However, ignoring interactions may underestimate genetic effects, and improvements in the understanding of interactions would be expected to improve genetic risk prediction models (Thanassoulis and Vasan, 2010). Missing data are commonly problematic since genotype success rates are never perfect (Kim and Misra, 2007). One strategy is to drop samples with missing data (Schwender and Ickstadt, 2008). Otherwise, when possible, imputation can be a useful solution for "filling in" missing data (Yuan, 2000).

Usually, for presence vs. absence of disease phenotypes, a predictive model is first developed by analyzing a case-control dataset, and then applied to a particular population. To provide risk estimates that are calibrated to that particular population, an adjustment which depends on the case to control ratio must be made to the intercept term of the case-control regression model (Yang et al., 2003).

Many studies, but not all (Warren et al., 2014), indicate that risk prediction would be more accurate if more predictors could be added in the risk model (De Jager et al., 2009; van Dieren et al., 2012). But the confidence interval (CI) of the risk estimate is often not considered in the evaluation of the risk model. When the model is built using regression in a meta-analysis of many case-control datasets, confidence intervals are often not even estimated.

Provided it is unbiased, a more precise risk estimate with a smaller CI from a model with fewer predictors is better than a less precise risk estimate with a larger CI from a model with more predictors (Shan et al., 2013). To compute the CI for the risk estimates from a meta-analysis, each individual study in the meta-analysis should do a joint analysis and return coefficient estimates and the variance-covariance matrix for the coefficients. Then, these can be combined to estimate the overall variance-covariance matrix and a precise CI for the risk estimates. Goddard et al. developed a method that derives an empirical CI combining all relevant sources of variation in disease risk (Goddard and Lewis, 2010; Crouch et al., 2013).

BAYESIAN NETWORKS

Bayesian Networks have resulted from the application of advances in graph theory to applied probability and carry a high degree of interpretability, along with providing an intuitive framework for obtaining posterior probabilities and the treatment of classification problems (Pearl, 1988; Jordan, 2004). If the features (genetic markers) within the Bayesian Network can be reasonably modeled as being conditionally independent (conditional on the disease trait in our application), then the network is reduced to a highly tractable Naïve Bayes model. Given a set of n genetic markers, using Bayes' rule one can write the posterior probability of the disease trait (PPD), as:

$$PPD_n = P\Big(D \,\Big|\, \bigcap_{i=1}^{n} G_i\Big) = \frac{P\big(\bigcap_{i=1}^{n} G_i \mid D\big)\,P(D)}{P\big(\bigcap_{i=1}^{n} G_i\big)},$$

where D denotes a random variable for the disease trait and n genetic markers are used in the prediction. Under the conditional independence assumption of Naïve Bayes, we can completely factorize the product and, for a binary trait ($D = 1$ to denote disease and $D = 0$ for non-disease), one can re-write the PPD as:

$$PPD_n = \frac{P(D=1)\prod_{i=1}^{n} P(G_i \mid D=1)}{P(D=1)\prod_{i=1}^{n} P(G_i \mid D=1) + P(D=0)\prod_{i=1}^{n} P(G_i \mid D=0)}.$$

To illustrate this type of calculation, **Figure 1** shows scaled PPD values for a rheumatoid arthritis study. In this study (Chang et al., 2008), the PPD for every possible three-locus genotype combination at the risk loci (*HLA-DRB1*, the R620W polymorphism at *PTPN22*, and diplotypes at *TRAF1*) was calculated, and scaled such that the smallest value was set to 1; SRR denotes this scaled ratio (**Figure 1**). While there is substantial variability across the values for different genotype combinations: over a 41-fold difference in predicted rheumatoid arthritis-risk, it is important to keep in mind how these bins are populated with individuals with and without the disease trait (the case-control frequencies given for each combination), for a prognostic loses general utility as intermediate combinations become frequent. In concrete terms, while a 41-fold difference is impressive, only 0.1% of the general population is calculated to carry genotypes producing this level of effect. 3.2% have multi-locus genotypes that generate at least a 21-fold increase in RA risk, and 13.7% carry a multi-locus genotype with >5-fold increase in RA risk (all compared to the lowest category).

Figure 2 displays the results from a simplified model. Five hundred disease susceptibility SNPs, all having equal effect sizes and genotype frequencies, were modeled. A prior probability of disease was set to 0.20 and the predisposing genotype frequency

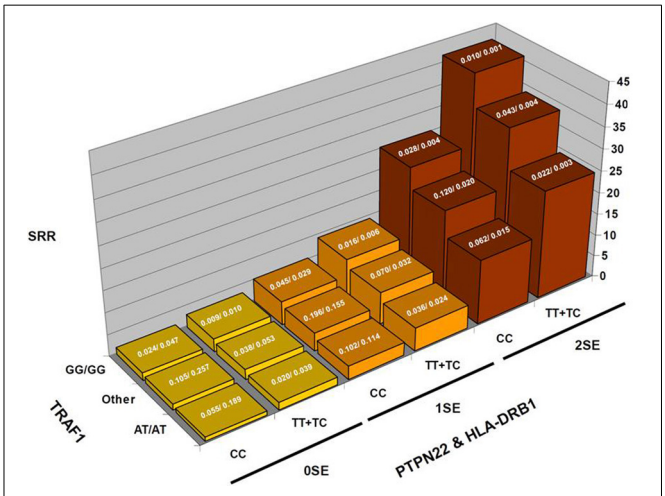


FIGURE 1 | Rheumatoid arthritis scaled posterior probabilities (SRR). Genotype data at three strongly predisposing loci, *HLA-DRB1*, *TRAF1*, and *PTPN22* are combined and the posterior probabilities calculated for every possible multilocus genotype combination. The prior probability was set to the approximate population prevalence of rheumatoid arthritis, 0.01. The posterior probabilities are scaled such that the lowest RA-risk multilocus genotype was set to a value of 1. The results show a 41-fold variation in posterior probabilities. The expected frequencies of the various multilocus genotype combinations in RA patients/controls are shown at the top of each bar.

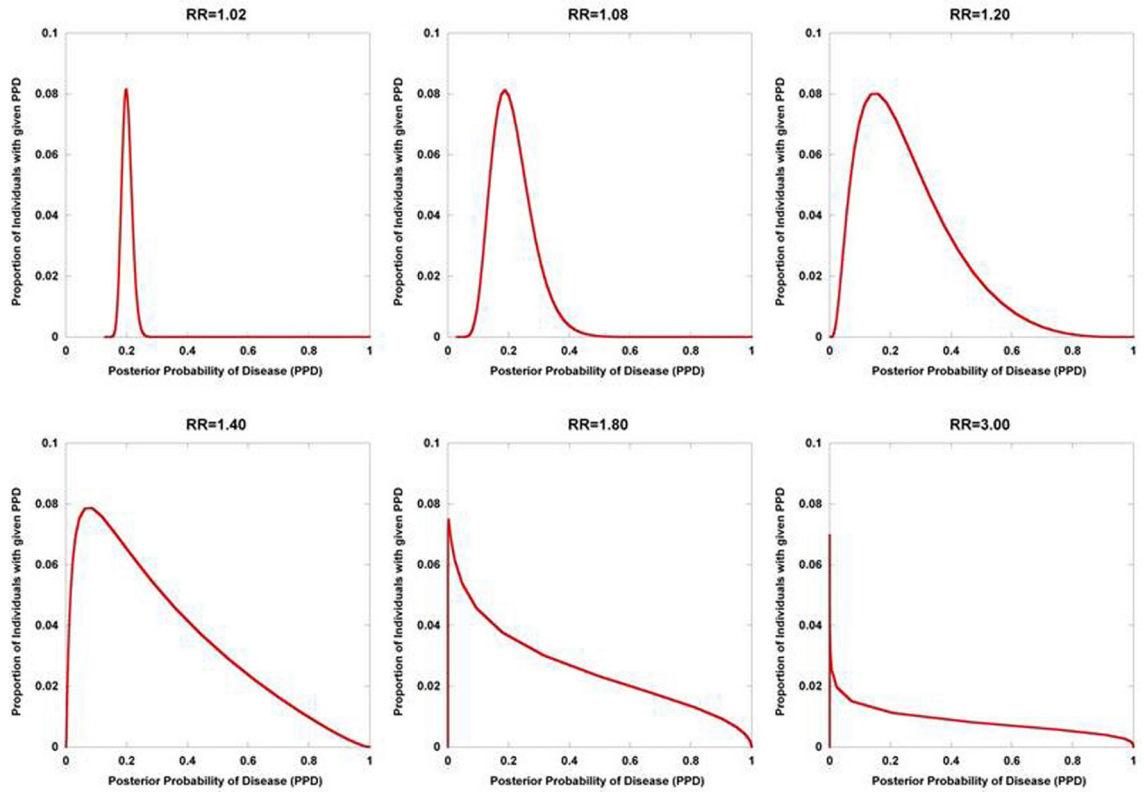


FIGURE 2 | Posterior probability variation with relative risk. The density of posterior probabilities of disease (PPD) are shown under a simplified multilocus disease model. The number of independent, disease-predisposing SNPs was set at 500. Relative risk was modeled as being identical for each predisposing SNP. Frequency of the predisposing genotype in controls was set to 0.05 at each SNP. Prior probability of disease was set at 0.20. Naïve Bayes was used to calculate posterior probabilities. The data points only take on discrete values (The densities are composed of discrete values which are connected by lines to produce the curves. While the sum of the discrete values all equal one in each of the curves, the areas under the curves do not), but are presented with interconnecting lines.

in the general population was set to 0.05 for all 100 SNPs. As expected, for very small effect sizes the number of individuals calculated to have posterior probabilities close to 0.20 is high and rapidly tails off. However, for larger effect sizes, there is an accumulation of individuals with posterior probabilities closer to 1 and 0. Interestingly, even with quite considerable effect sizes (for high frequency alleles) much of the density still resides in the intermediate region between 0.10 and 0.90. If we explore the dynamics as the number of loci is increased, so does the variance in the posterior probability of the disease trait (**Figure 3**).

Diagnosis or prognosis of disease traits with genetic information are classical problems of classification and clustering within machine learning. Hence, numerous machine learning methods, such as neural networks, support vector machines, and random forests can be applied to these types of data sets. Currently, the use of these methods to address problems using gene expression is arguably more advanced than the analogous methods applied to DNA variation data.

QUANTIFYING PROGNOSTIC UTILITY

Within a population studied, once each individual is (1) assigned a score for a risk metric, (2) assigned a posterior probability, (3) clustered or (4) classified, a method for assessing prognostic utility is required to quantify the usefulness in clinical practice. The most common method used is the area under the ROC curve, or AUC. However, although this metric is useful to assess discrimination, it is not the appropriate measure to assess a predicted probability (Cook, 2007). Graphically, the ROC curve is a plot of the performance of the predictor in a space defined by the sensitivity (true positive rate) and 1—specificity (false positive rate). Varying the threshold of calling a result positive or negative, a curve can be produced for the predictive model. The AUC is the integral of the curve. For a completely non-informative predictor, the AUC is 0.50, with larger values (up to unity) indicating improved prognostic utility (**Figure 4**). While useful, sensitivity and specificity are probabilities conditional on the state of the phenotype trait. One may want to consider metrics that have differential performance with the prevalence of the disease trait. Indeed, all other diagnostic factors being equal, a physician should be more prone to diagnose an individual with a more common phenotype than an exceedingly rare one, because the *a priori* likelihood of the disease being the common phenotype is higher than the likelihood for the rare phenotype. Therefore, use of positive and negative predictive values (PPV and NPV) may be more useful in the clinical setting. PPV is defined as the proportion of true positives out of all positive results as

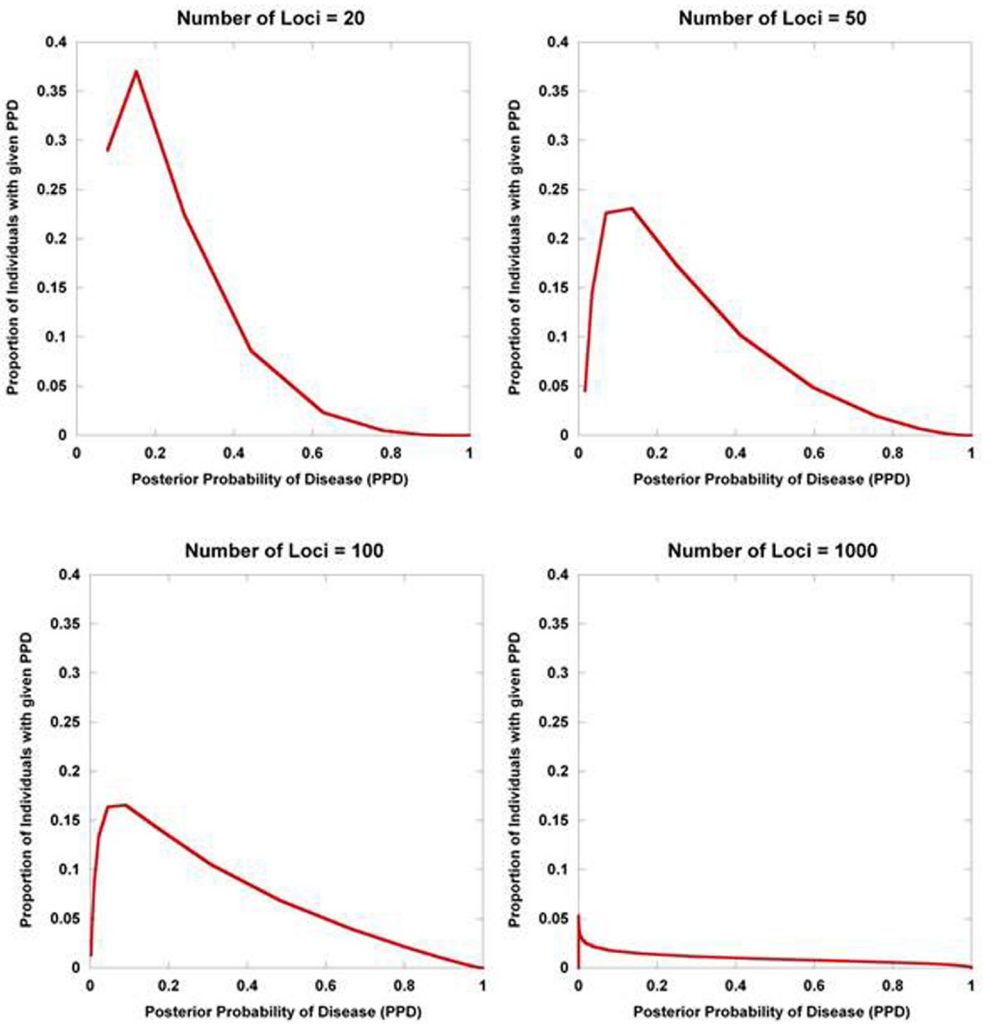


FIGURE 3 | Posterior probability variation with number of predisposing loci. The density of posterior probabilities of disease (PPD) is shown under a simplified multilocus disease model. The relative risk of each independent, disease-predisposing SNP was set to 2.0. Prior probability of disease was set at 0.20. Frequency of the predisposing genotype in controls was set to 0.05 at each SNP. The number of predisposing loci was increased from 20 to 1000. Naïve Bayes was used to calculate posterior probabilities. The data points only take on discrete values (the larger number of loci have many more data points reflecting the larger number of possible multilocus genotype combinations), but are presented with interconnecting lines.

determined from applying a diagnostic test. Conversely, NPV is defined as the proportion of negative results that are indeed truly negative. However, a direct ROC analog of characterizing the tradeoff between PPV and NPV offers challenges. Motivated by this, Pencina et al. suggest that averaging over PPV and NPV may provide an improved metric for characterizing prognostic/diagnostic utility (Pencina et al., 2008). In 2006, a new method for characterizing disease predictions based on proportions of individuals accurately reclassified was presented (Cook et al., 2006). This approach was further developed in subsequent publications, describing the employment of the Hosmer-Lemeshow goodness-of-fit statistic and the net reclassification improvement statistic applied to reclassification categories as predictive measures (Cook, 2007; Cook and Ridker, 2009). The authors applied these approaches to better specify results from cardiovascular risk models.

Another approach would be to characterize the improvement in the distribution of posterior probabilities as compared to the distribution of prior probabilities, where the distribution is taken across all individuals evaluated. The more informative the genetic information becomes, the larger the departure between posterior and prior probability densities. A natural measure for this is the Kullback-Leibler Divergence, which quantifies the departure between two densities (Kullback and Leibler, 1951). Applied to characterizing the improvement in disease prediction following the interrogation of a suite of features such as genetic markers, the Kullback-Leibler Divergence is defined as:

$$D_{KL} = \sum P(Disease \mid G_1, \ldots, G_n) \, log\left[\frac{P(Disease \mid G_1, \ldots, G_n)}{P(Disease)}\right]$$

where G_i are the random variables describing the states of each genetic marker involved in disease susceptibility, and the sum is

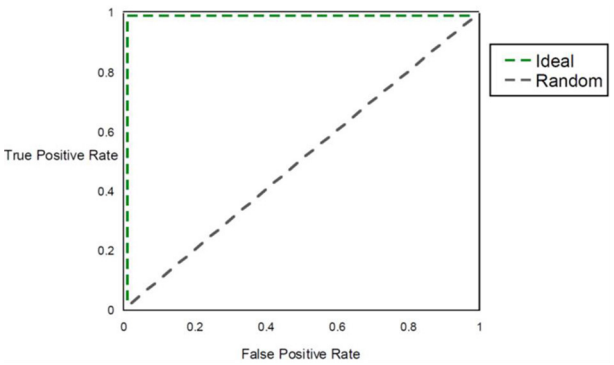


FIGURE 4 | AUC. The figure shows the ROC curve and corresponding area under the ROC curve (AUC). The expected patterns under two extreme scenarios are shown: an ideal diagnostic scenario and the pattern expected using random predictions.

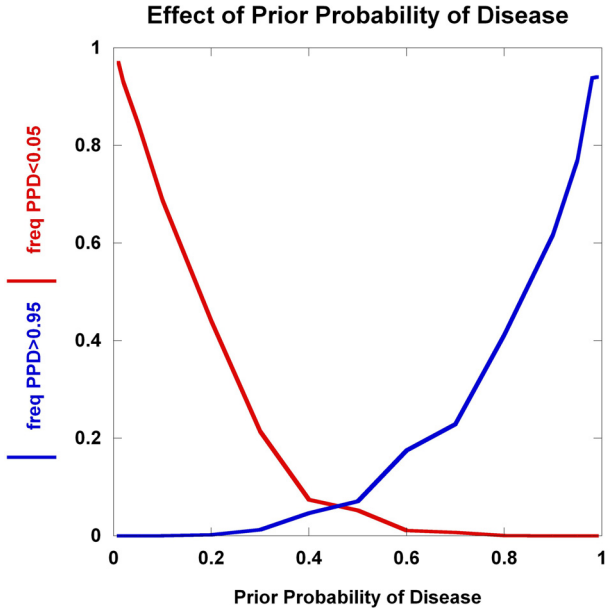


FIGURE 5 | Effect of prior probability. The frequency of multilocus genotype combinations exceeding the C_1 and C_2 thresholds for posterior probabilities of disease (PPD) (set at 0.05 and 0.95, respectively) are presented as a function of the prior probability of disease. 100 predisposing SNPs were used in the model, each having a predisposing genotype frequency of 5% in controls and relative risk of 2.0.

over all possible multilocus genotype combinations. D_{KL} is calculated across the entire population to whom the predictive model is applied. Larger values of D_{KL} indicate enhanced differences between the posterior and prior probabilities across the population, reflecting the greater utility of the genetic information. Hence, the Kullback-Leibler Divergence concurrently captures both the magnitude of the effect the genetic data have on the posterior probabilities for each individual (compared to the prior) and the proportion of tested individuals exhibiting each magnitude of the effect. Empirical-based calibration of this or other measures of prognostic utility can often be accomplished through using well-studied data sets having standard prognostic tests such as the Framingham population and cardiovascular disease risk score (Wilson et al., 1998; Schrodi et al., 2009).

Another possible method for characterizing would be to define some level of probability that is clinically meaningful for the specific application. That is, define critical levels τ_{pos} and τ_{neg} such that exceeding these values with the posterior probability of disease provides actionable information for a clinician. Define the conditions:

C_1: $P\left(Disease \mid Genotype\ Data, other\ features\right) > \tau_{pos}$

C_2: $P\left(Disease \mid Genotype\ Data, other\ features\right) < \tau_{neg}$.

We explore the dynamics of C_1 and C_2 as a function of the prior probability of disease in **Figure 5**. The collective effect of 100 disease-predisposing SNPs, each with relative risk 2.0 and genotype frequency 5%, is clearly not sufficient to concurrently generate high proportions of individuals who are well-classified as either being likely (C_1) or unlikely (C_2) to have disease. At prior probabilities close to 0.50, the majority of individuals do not satisfy either condition C_1 or C_2. It is only in the situations where the prior probability is either close to 0 or 1 that large percentages of the population interrogated achieve very high or very low posterior probabilities. Hence, with current results from disease genetics, it seems reasonable to assume that a clinician should already have a strong suspicion either of a disease diagnosis or the exclusion of a disease to warrant the use of SNPs.

To further explore these prognostic utility patterns, we considered two simplified disease models: a highly polygenic model consisting of 1000 predisposing SNPs, each of appreciable frequency (10% in controls) (**Figure 6**). We set the prior probability of disease to 0.20. As the relative risk of each SNP is increased from 1.02 to 1.80, the C_2 condition exhibits a sigmoidal pattern, climbing to over 80% roughly when the relative risk hits 1.45. In contrast, the C_1 condition peaks at roughly the same relative risk and declines thereafter, but never exceeding 0.02. A typical large GWAS experiment would be well-designed to identify the SNPs with relative risks in excess of roughly 1.1. The collective effect from the 1000 SNPs is not sufficient to overcome the prior probability of 0.20 to promote frequent individual multilocus genotype combinations to exceed the 0.95 threshold of C_1. That said, the proportion of individuals with posterior probabilities exceeding the $C_2 < 0.05$ threshold was much higher. We explored a highly penetrant, rare allele model (**Figure 7**). We constructed this model with 100 predisposing single nucleotide variants (SNVs) with predisposing genotypes being rare (0.1%), offset by large effect sizes ranging from relative risks of 10 to 400. Sequencing studies generate numerous SNVs. In each graph a single effect size was assumed for all SNVs. Again, the prior probability was set to 0.20. Here, the C_1/C_2 dynamics are more complex, with the C_1 and C_2 conditions being very sensitive to individual multilocus genotype combinations. Modeling a distribution of SNV frequencies would smooth this type of graph. These are overly simplified cases examined here and the parameter space is vast—additional work in this area would provide useful insights into the properties of prognostics that result from different genetic-based disease models. That said, the proportion

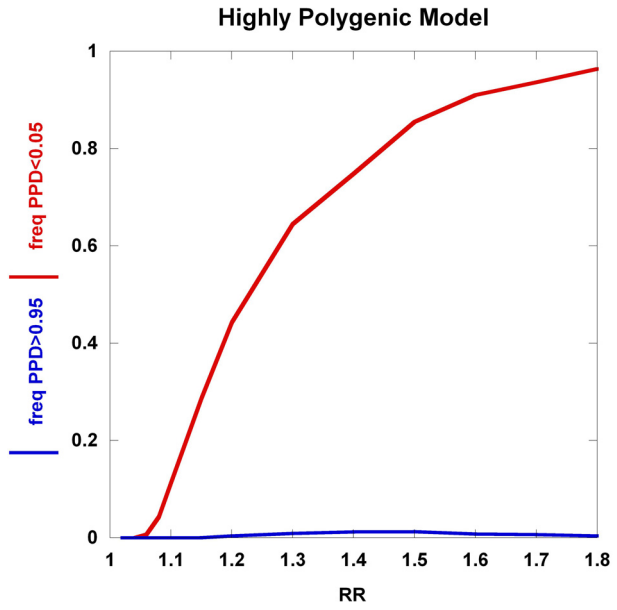


FIGURE 6 | Highly polygenic model. The dynamics of the C_1/C_2 threshold values under a simplified model is shown as the relative risk of the SNPs varies. The *highly polygenic model* has 1000 predisposing SNPs each having predisposing genotype frequencies in controls equal to 10% and a prior probability equal to 0.20. The relative risk was varied from 1.02 to 1.80.

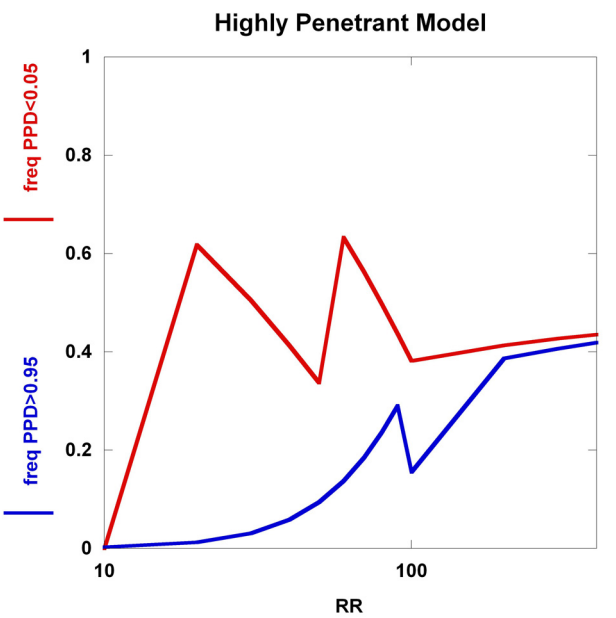


FIGURE 7 | Highly penetrant model. The *highly penetrant model* uses 100 SNPs each having a predisposing genotype frequency of 0.1% and also a prior probability of 0.20. The relative risk takes on values from 10 to 400. Although the *highly polygenic model* yields a large proportion of individuals with posterior probabilities below 0.05, the increasing relative risks have little impact on the proportion of individuals with posterior probabilities above 0.95. The *highly penetrant model* shows an overall increase in the proportions of individuals with posterior probabilities below 0.05 and above 0.95, but the patterns are somewhat unexpected (not smooth, nor monotone). These patterns are generated from all predisposing SNVs having identical genotype frequencies and relative risks, coupled with having specific PPD thresholds.

of individuals satisfying C_1 is dramatically higher than under the highly polygenic model. Further, high values of C_1 and C_2 occur concurrently. Although much more work is needed to fully explore these dynamics, this observation may give some hope to the usefulness of rare, highly penetrant sequence variants in the context of disease prediction. However, one might have expected that the most common results of GWAS analyses—identification of large numbers of common variants each with small impact on disease risk for common diseases—would be more useful, unless there are other effects operating, such as considerable locus heterogeneity for common diseases.

EXAMPLES

GWAS DATA APPLIED TO A TYPE 2 DIABETES PROSPECTIVE COHORT

Type 2 diabetes (T2D) is a common medical condition with rapidly increasing incidence worldwide (Zimmet et al., 2001). This disease is characterized by a multitude of abnormal pathophysiological states involving muted beta cell response, chronic inflammation, and aberrant levels of metabolic markers that ultimately lead to vascular damage, infection, heightened cardiovascular disease risk, and neuropathy (Zimmet et al., 2001). Numerous T2D GWAS have been conducted and have reliably identified new genes and genetic regions involved in T2D susceptibility, albeit with modest effect sizes (McCarthy, 2010). Early prediction of T2D onset and trajectory can be leveraged into improving important medical decisions, including treatment with therapeutics, exercise programs, and diet restriction. It is possible that genetic variants may play a role in improving the prediction of T2D. To test this idea, Shigemizu et al. very recently performed a two-stage study (training and test sets) that resulted in combining nine SNPs with three clinical risk factors (age, gender, and BMI) to develop a predictive model for T2D in a prospective cohort having Japanese ancestry (Shigemizu et al., 2014). The features used in a regression model for T2D-prediction were selected using a Bayes Factor and lasso method. From both genetic and clinical risk factors, the resulting predictive model showed reasonable AUC values in the independent test set (AUC = 0.808). Further, where the sensitivity and sensitivity were concurrently maximized, the model yielded a PPV and NPV of 77.8 and 73.8%, respectively. Although the selected SNPs did add to the diagnostic performance of the prediction model, they only did so in an incremental fashion. The model using SNPs, interactions, and clinical risk factors exhibited a 1.5% increase in the AUC over the clinical risk factors alone. Perhaps the discovery of additional T2D risk variants from sequencing efforts, rarer exome variants, extensive epistatic models, and/or undiscovered epigenetic factors will drive further work in this area to markedly improve the performance of T2D predictive models using heritable information. Until then, there may be greater gains through the use of dynamic markers like metabolite profiling and proteomics. Alternatively, exploration of prediction within T2D subgroups may offer a more fruitful avenue of inquiry.

STROKE PREDICTION USING GENETIC RISK SCORES

Stroke events are major contributors to mortality and morbidity, constituting the fourth leading cause of death in the United States. Accurate prediction of ischemic stroke risk would enable medical interventions to at least partially remediate stroke occurrence and the resulting brain damage. Very recently, two large studies (Ibrahim-Verbaas et al., 2014; Malik et al., 2014) have been published evaluating risk models constructed from a number of stroke and related phenotype-associated GWAS SNPs. The results were consistent with GRSs achieving statistical significance, but adding little in diagnostic utility to clinical features, as measured by AUC. Ibrahim-Verbaas et al. evaluated the performance of a 324-SNP GRS in four population-based cohorts totaling over 22,000 individuals in an effort to improve the discrimination of ischemic stroke over that generated from the Framingham Stroke Risk Score Model, age and sex. The SNPs were selected based on association with stroke-related phenotypes and a GRS constructed using weights from the regression model used to test the disease association. ROC curves from the results for the study show that the weighted GRSs do not substantially add to the AUC over that achieved by the Framingham Stroke Risk Score and sex—the improvement in AUC from the GRS was approximately 0.02 for all stroke as well as for ischemic stroke alone—although the AUC improvement was statistically significant. The authors also examined the impact of the GRS on the net reclassification index, showing statistically significant, but incremental improvement. Concurrently published in the same issue of *Stroke*, Malik et al. presented similar results for their stroke GRS performance in comparison to clinical features using overlapping samples (Malik et al., 2014). The study showed increased stroke risk across quintiles of their GRS, obtained from an analysis of independent samples from the Wellcome Trust Case Control Consortium 2 and the METASTROKE consortium. Slightly under a 1.5-fold increase in risk was found comparing the top quintile to the middle quintile, and a >2-fold increase comparing the lowest quintile to the top quintile. No significant improvement in the net reclassification was observed and the ROC curves with and without the GRS are virtually superimposable for a sample set composed of a clinical trial-based derivation sample set and the replication sample set.

PREDICTION USING BIOBANK DATA

Current efforts to discover and employ genetic risk predictors across multiple health care systems include those of the Electronic Medical Records and Genomics (eMERGE) Network (Gottesman et al., 2013). The eMERGE Network has supported large-scale genotyping efforts in biobanked DNA samples linked to electronic medical records. As such, a repository of genome-wide genetic data can be interrogated with respect to a vast amount of clinical information. One use of these data is to investigate how sets of genetic markers can stratify sample sets for the purpose of performing historical prospective studies. By analyzing longitudinal data, one can specify the sets of individuals to "follow" from a point in time to test for association with various medical traits. In doing so, one can perform a prospective study relating genotypes to the accumulation of various medical outcomes and laboratory values. This is an excellent venue for evaluating genetic-based predictive models. For example, suppose one constructed a predictive model for myocardial infarction (MI) with existing literature findings and then assigned a predicted MI risk for each individual. One could then evaluate how the predicted risk was correlated with the actual conversion rate of non-MI individuals to MI disease states. One can also simultaneously perform association testing between any combination of sequence variants and/or GWAS SNPs and prospectively occurring disease, for the purpose of discovering novel genotype-phenotype correlations. Notably, this type of experimental design is less subject to confounding effects when compared to retrospective case-control designs because a cohort-based design is less likely to impart bias from sample selection being correlated with genetic factors. As noted in a 2010 Institute of Medicine "Rapid Learning" document, the hope is that electronic medical records, biobanks and bioregistries will provide evidentiary support for intervention decisions (National Research Council, 2010). Interesting, Lauer and D'Agostino recently suggested that the next disruptive technology in clinical research would be the randomized registry trial (Lauer and D'Agostino, 2013).

Deeply phenotyped biobanked datasets can also be used to redefine disease states. GWAS have highlighted SNPs that are undoubtedly correlated with susceptibility to common diseases but, as we have discussed, the alleles discovered thus far explain only a marginal amount of disease heritability. The reasons for this are the subject of much debate. Resolution of this perplexing problem will likely involve a multitude of discoveries, not the least of which stem from addressing the opaque correspondence between clinical phenotypes and underlying molecular pathologies. Due to reliance on observations of complex, gross physiology in the clinic, it is reasonable to assume that there may be multiple molecular etiologies that map to a single clinical disease state (e.g., estrogen receptor status now meaningfully partitions previously indistinguishable breast cancers and leads to profound changes in the use of Tamoxifen) (Fisher et al., 1988, 1989; Paik et al., 2004). Conversely, single molecular perturbations may have pleiotropic effects (e.g., the rs2476601 SNP in *PTPN22* is strongly associated with several, clinically distinct autoimmune diseases) (Begovich et al., 2004; Bottini et al., 2004; Kyogoku et al., 2004; Velaga et al., 2004; Canton et al., 2005; Criswell et al., 2005). The medical field is accustomed to defining diseases with regard to visual inspection and gross anatomical measurements, and therefore may (1) aggregate disparate molecular pathophysiologies and (2) partition the same molecular processes into different disease classes. Indeed, there is not a one-to-one mapping between clinical assessments of disease and molecular processes. Thus, it seems reasonable to adopt the reductionist stance that redefining disease states and processes in terms of the underlying genetic and molecular variation may significantly aid investigation of disease etiologies. In this way, one can construct phenotype-based predictive models for sets of genetic/molecular information—a reverse genetics approach. Several groups have recently taken this approach to mapping disease genes: Pendergrass et al. used this method to interrogate data from the PAGE network (Pendergrass et al., 2013), Hebbring et al. (2013), have performed similar types of studies in the Marshfield Personalized Medicine Research Project samples, and Denny et al. utilized data from the eMERGE

Network (Denny et al., 2013). In these studies, clinical phenotypes are screened in electronic medical record (EMR) systems for association with specific genetic variants with known function (or highly likely to have specific impact on biological pathways)—a method pioneered by Ritchie et al. in a large-scale effort to replicate numerous associations using DNA databanks linked to EMRs (Ritchie et al., 2010). Novel disease associations can be discovered through these "PheWAS" studies. In addition, this "bottom-up" (specific genetic variants-to-phenotype) approach can also be viewed as a starting point for using genetic information to redefine disease states in a classification system that more closely mirrors the underlying molecular pathophysiology. For example, screening diseases within a biobank for association with *IL23R* missense variants uncovers sets of disease phenotypes where aberrant Th17 signaling plays a pathogenic role. Autoinflammatory diseases, including ankylosing spondylitis, psoriasis, and Crohn's disease, would all show a common, core aspect to their molecular pathophysiology. Additionally, partitioning by these same variants allows elucidation of disease subgroups. This reclassification can further enable disease prediction, for the phenotypes predicted would exhibit clearer correspondence with the underlying molecular mechanisms.

INFLAMMATORY ARTHRITIS PREDICTION IN THE MARSHFIELD POPULATION

To illustrate how to apply machine learning methods to empirical datasets for the purpose of disease prediction and show some of the difficulties with attaining strong predictive signals from GWAS findings, we present an example of using genetic data and samples from an EMR-linked biorepository for the purpose of distinguishing between inflammatory arthritis conditions.

Worldwide and within the US, inflammatory arthritides are common conditions representing a substantial portion of disabling disease. Early treatment of these conditions can provide substantial benefit in averting disabling articular damage and systemic complications. In general, autoimmune and autoinflammatory diseases such as rheumatoid arthritis and spondyloarthritides have significant heritabilities—a substantial portion of which has been explained by identified polymorphisms, thereby motivating the incorporation of genotype information into prognostic approaches. This study aimed to investigate and characterize the ability of a panel of genetic markers, identified from genome-wide association study results, to classify individuals into the three inflammatory arthritis categories: rheumatoid arthritis, axial spondyloarthritis and psoriatic arthritis (**Table 1**). Using genotyped samples from an independent sample set from Central Wisconsin (Marshfield Clinic), several machine learning methods were applied to a filtered set of these polymorphisms to classify individuals into the three inflammatory arthritis diseases, blinded to the known disease status. The WEKA software package was used to implement the machine learning algorithms (Holmes et al., 1994; Hall et al., 2009). Accuracy was defined as the proportion of positive and negative classification results that were in fact true. The Naïve Bayes classifier attained the highest average accuracy from 10-fold cross validation on the training set (**Table 2**). However, when applied to the Marshfield test set, there was a substantial decline in the performance with an average area under the ROC curve of 0.635 (**Figure 8**). Although the difference between this observed AUC of 0.635 and that expected under the null (AUC = 0.500) is statistically significant, we conclude that additional, orthogonal predictive variables, such as clinical features, circulating cytokine profiles or additional genetic variants, are necessary

Table 1 | Illustration of machine learning methods applied to genetic data: feature selection.

Ankylosing spondylitis	Psoriatic arthritis	Rheumatoid arthritis
Rs13203464 (HLA-B27)	Rs10484554 (HLA-C)	Rs660895 (HLA-DRB1)
Rs30187 (ERAP1)	Rs20541 (IL13)	Rs2476601 (PTPN22)
Rs11209026 (IL23R)	Rs13017599 (REL)	Rs3761847 (TRAF1/C5)
Rs10865531 (2p15)	Rs2066808 (IL23A)	Rs3890745 (MMEL1)
Rs2310173 (IL1R2)	Rs12924903 (RUNX3)	Rs13031237 (REL)
Rs4333130 (ANTXR2)	Rs4795067 (NOS2)	Rs7574865 (STAT4)
Rs378108 (21q22)	Rs4379175 (IL12B)	Rs548234 (PRDM1)
Rs2297909 (KIF2B)	Rs4982254 (PSMA6)	Rs2327832 (TNFAIP3)
Rs10045431 (IL12B)	Rs13151961 (IL2/21)	Rs1569723 (CD40)
Rs10903118 (RUNX3)	Rs11209026 (IL23R)	Rs11574914 (CCL21)
Rs7720838 (PTGER4)	Rs7720838 (PTGER4)	Rs11172254 (KIF5A)
Rs2058276 (Y-marker)		Rs231804 (CTLA4)
		Rs1160542 (AFF3)
		Rs13151961 (IL2/21)

SNPs were selected for the Naïve Bayes Algorithm to determine inflammatory arthritis categories. Feature selection for the SNPs was performed through a literature search on the three diseases of interest from existing GWAS studies: Ankylosing spondylitis (AS), Psoriatic arthritis (PsA), and Rheumatoid arthritis (RA). The most significant and replicable SNPs were used to construct the entire set. Subsequent evaluation of features used Wrapper Subset Evaluation, ChiSq Attribute Evaluation, Classifier Subset Evaluation, and Information Gain from Attributes methods within the program WEKA. These methods were coupled with a variety of search methods including genetic algorithm-based, Greedy Stepwise Selection, and Linear Forward Search methods within WEKA.

Table 2 | Relative performance across machine learning methods.

Algorithm	Number of SNPs	Accuracy (%)	10-fold CV accuracy (%)
Decision tree	33	87	74.1
Neural network	5	76	77.1
Logistic regression	24	78	77.7
Support vector machine	1	73	72.6
K-nearest neighbor	7	77	77.2
Naïve bayes	29	78	77.9

Six Machine Learning algorithms using feature selection and classification were evaluated for accuracy using complied data from the literature in a synthetic training data set. Effect sizes and genotype frequencies were estimated from the literature and incorporated into the synthetic data set to train the algorithms. Accuracy was defined as the proportion of individuals that were correctly predicted (either true positives or true negatives). 10-fold Cross Validation (CV) was performed within WEKA and used as the criterion for selection of an algorithm to use in the test set from Marshfield. Naïve Bayes exhibited slightly higher CV accuracy when compared to other algorithms, with a low amount of overfitting.

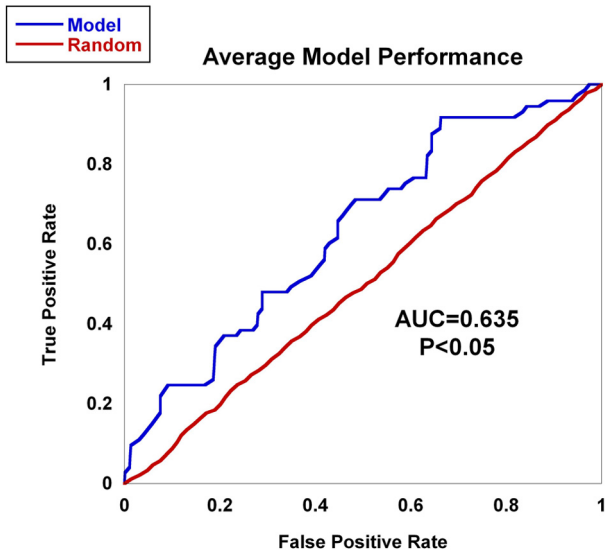


FIGURE 8 | AUC for inflammatory arthritis prediction study for the Marshfield population. The Naïve Bayes classifier developed using data from the literature was applied to the Marshfield population of inflammatory arthritis individuals: Rheumatoid arthritis (RA), Psoriatic arthritis (PsA), and Ankylosing spondylitis (AS). The model generated an AUC of 0.635, which was statistically significant via permutation. In addition, performance on randomized sample sets is shown in red, showing an expected null performance.

to build a clinically useful prognostic test for classifying these diseases.

SUMMARY

We have summarized some of the seminal issues in utilizing genetic information in predictive models for disease traits. Currently, the application of genetic-based predictive models to common diseases is, generally speaking, disappointing from both theoretical and empirical lines of evidence. There are some bright spots, including AMD, Crohn's disease, and special applications to selected populations with increased posterior probabilities due to non-genetic factors. Additionally, if the current wave of sequence-based disease gene mapping uncovers sufficient numbers of highly penetrant alleles, then these may provide clinically relevant prognostic utility. Outside of common disease prognostics, tumor genetics, screening for inherited Mendelian disorders, and some pharmacogenetic applications have exhibited the most progress over the past five years. The reasons for this stem from the reduced complexity of the genetic architecture of these traits, yielding extremely high or extremely low posterior probabilities. Certainly, many questions in the field remain. As our understanding of the nature of elements that resolve the missing heritability problem matures, the path to applying predictive modeling methods will become clearer. What needs to fall in place for clinically useful prediction of complex diseases? We speculate that six critical steps will aid this process:

(1) Through next-generation sequencing platforms applied to both linkage and association designs, identification of additional susceptibility variants will fully cover the allele frequency spectrum and capture disease-predictive alleles. However, the discovery of rare, highly penetrant risk alleles will be most useful as clinical sequencing becomes widespread and applied earlier in life.
(2) As other elements besides DNA sequences are inherited and contribute to phenotypic variance, the interrogation of additional possible contributors to heritability, including DNA methylation patterns, histone modifications, transgenerational effects, and other factors correlated with disease traits, will capture more of the molecularly-defined heritability.
(3) Redefining disease phenotypes to more accurately mirror the underlying molecular pathophysiology will be critical in reducing disease complexity and better enable genetic susceptibility mapping. For example, partitioning diseases by molecular subtypes will identify physiological subgroups with clearer correspondence with the underlying genetics. Within the context of research using biobanks linked to medical records, relevant laboratory tests or imaging information, or both, would also be valuable.
(4) Considerable progress has been made in the field of machine learning, where robust methods have been developed to select features and use them in predictive models. Applying these approaches to genetic data in combination with existing laboratory tests, imaging data, and other established medical tests will offer the best chance of creating viable prognostics.
(5) Metrics that capture prognostic utility in a way that accurately reflects what a clinician requires to inform medical decisions will be developed.
(6) The application of disease predictive models to diverse clinical populations will clarify the performance and limitations of proposed predictive models and improve medical practice.

In summary, while prediction will continue to be challenging, future investigations promise to provide a wealth of information, some of which will be clinically useful if considered in the appropriate context.

ACKNOWLEDGMENTS

We would like to thank Dr. Judy Smith for insightful comments on the manuscript and Dr. Sam Broder for thoughtful discussions. We would also like to thank Drs. Bruce Krawisz, Kajal Sitwala, Tim Uphoff, Ariel Brautbar, and Scott Hebbring for many useful discussions on the use of genetics in clinical applications. We would like to thank Dr. Ray White for sharing his insights into genetic mapping and thoughts on the genetic architecture of diseases. The eMERGE Network is funded by the NHGRI, with additional funding from the National Institute of General Medical Sciences through the following grants: U01HG004438 to Center for Inherited Disease Research; U01HG004608 to Essentia Institute for Rural Health/Marshfield Clinic Research Foundation; U01HG04603 and U01HG006378 to Vanderbilt University; U01HG006385 to the Coordinating Center; U01HG006382 to Geisinger Clinic; and U01 HG006375 to Group Health Cooperative and the University of Washington. The project described was also supported by the Clinical and Translational Science Award (CTSA) program, previously

through the National Center for Research Resources (NCRR) grant 1UL1RR025011 and the National Center for Advancing Translational Sciences (NCATS) grant 9U54TR000021, and now by the NCATS grant UL1TR000427. Additional funding included "Utility of genomic data in population screening for abdominal aortic aneurysm" from The Commonwealth Universal Research Enhancement (CURE) program of the Commonwealth of Pennsylvania (Geisinger and University of Pittsburgh). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

REFERENCES

Abraham, G., Kowalczyk, A., Zobel, J., and Inouye, M. (2013). Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genet. Epidemiol.* 37, 184–195. doi: 10.1002/gepi.21698

Agrawal, N., Frederick, M. J., Pickering, C. R., Bettegowda, C., Chang, K., Li, R. J., et al. (2011). Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* 333, 1154–1157. doi: 10.1126/science.1206923

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control IEEE Trans.* 19, 716–723. doi: 10.1109/tac.1974.1100705

Aletaha, D., Neogi, T., Silman, A. J., Funovits, J., Felson, D. T., Bingham, C. O. 3rd, et al. (2010). 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Ann. Rheum. Dis.* 69, 1580–1588. doi: 10.1136/ard.2010.138461

Bao, W., Hu, F. B., Rong, S., Rong, Y., Bowers, K., Schisterman, E. F., et al. (2013). Predicting risk of type 2 diabetes mellitus with genetic risk models on the basis of established genome-wide association markers: a systematic review. *Am. J. Epidemiol.* 178, 1197–1207. doi: 10.1093/aje/kwt123

Begovich, A. B., Carlton, V. E., Honigberg, L. A., Schrodi, S. J., Chokkalingam, A. P., Alexander, H. C., et al. (2004). A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am. J. Hum. Genet.* 75, 330–337. doi: 10.1086/422827

Berrar, D., and Flach, P. (2012). Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). *Brief Bioinform.* 13, 83–97. doi: 10.1093/bib/bbr008

Borque, A., del Amo, J., Esteban, L. M., Ars, E., Hernandez, C., Planas, J., et al. (2013). Genetic predisposition to early recurrence in clinically localized prostate cancer. *BJU Int.* 111, 549–558. doi: 10.1111/j.1464-410X.2012.11333.x

Borras, E., Jurado, I., Hernan, I., Gamundi, M. J., Dias, M., Marti, I., et al. (2011). Clinical pharmacogenomic testing of KRAS, BRAF and EGFR mutations by high resolution melting analysis and ultra-deep pyrosequencing. *BMC Cancer* 11:406. doi: 10.1186/1471-2407-11-406

Bottini, N., Musumeci, L., Alonso, A., Rahmouni, S., Nika, K., Rostamkhani, M., et al. (2004). A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nat. Genet.* 36, 337–338. doi: 10.1038/ng1323

Burgess, S., and Thompson, S. G. (2013). Use of allele scores as instrumental variables for Mendelian randomization. *Int. J. Epidemiol.* 42, 1134–1144. doi: 10.1093/ije/dyt093

Canton, I., Akhtar, S., Gavalas, N. G., Gawkrodger, D. J., Blomhoff, A., Watson, P. F., et al. (2005). A single-nucleotide polymorphism in the gene encoding lymphoid protein tyrosine phosphatase (PTPN22) confers susceptibility to generalised vitiligo. *Genes Immun.* 6, 584–587. doi: 10.1038/sj.gene.6364243

Cappellini, M. D., and Fiorelli, G. (2008). Glucose-6-phosphate dehydrogenase deficiency. *Lancet* 371, 64–74. doi: 10.1016/S0140-6736(08)60073-2

Cascorbi, I. (2003). Pharmacogenetics of cytochrome p4502D6: genetic background and clinical application. *Eur. J. Clin. Invest.* 33(Suppl. 2), 17–22. doi: 10.1046/j.1365-2362.33.s2.3.x

Chang, M., Rowland, C. M., Garcia, V. E., Schrodi, S. J., Catanese, J. J., van der Helm-van Mil, A. H., et al. (2008). A large-scale rheumatoid arthritis genetic study identifies association at chromosome 9q33.2. *PLoS Genet.* 4:e1000107. doi: 10.1371/journal.pgen.1000107

Chen, R., and Snyder, M. (2013). Promise of personalized omics to precision medicine. *Wiley Interdiscipl. Rev. Syst. Biol. Med.* 5, 73–82. doi: 10.1002/wsbm.1198

Clerget-Darpoux, F., and Elston, R. C. (2013). Will formal genetics become dispensable? *Hum. Hered.* 76, 47–52. doi: 10.1159/000354571

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 115, 928–935. doi: 10.1161/CIRCULATIONAHA.106.672402

Cook, N. R., Buring, J. E., and Ridker, P. M. (2006). The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Ann. Intern. Med.* 145, 21–29. doi: 10.7326/0003-4819-145-1-200607040-00128

Cook, N. R., and Ridker, P. M. (2009). Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Ann. Intern. Med.* 150, 795–802. doi: 10.7326/0003-4819-150-11-200906020-00007

Criswell, L. A., Pfeiffer, K. A., Lum, R. F., Gonzales, B., Novitzke, J., Kern, M., et al. (2005). Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) collection: the PTPN22 620W allele associates with multiple autoimmune phenotypes. *Am. J. Hum. Genet.* 76, 561–571. doi: 10.1086/429096

Crouch, D. J., Goddard, G. H., and Lewis, C. M. (2013). REGENT: a risk assessment and classification algorithm for genetic and environmental factors. *Eur. J. Hum. Genet.* 21, 109–111. doi: 10.1038/ejhg.2012.107

De Jager, P. L., Chibnik, L. B., Cui, J., Reischl, J., Lehr, S., Simon, K. C., et al. (2009). Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score. *Lancet Neurol.* 8, 1111–1119. doi: 10.1016/S1474-4422(09)70275-3

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. doi: 10.1371/journal.pgen.1003608

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* 31, 1102–1110. doi: 10.1038/nbt.2749

Evans, W. E., and Relling, M. V. (2004). Moving towards individualized medicine with pharmacogenomics. *Nature* 429, 464–468. doi: 10.1038/nature02626

Fisher, B., Redmond, C., Fisher, E. R., and Caplan, R. (1988). Relative worth of estrogen or progesterone receptor and pathologic characteristics of differentiation as indicators of prognosis in node negative breast cancer patients: findings from National Surgical Adjuvant Breast and Bowel Project Protocol B-06. *J. Clin. Oncol.* 6, 1076–1087.

Fisher, B., Costantino, J., Redmond, C., Poisson, R., Bowman, D., Couture, J., et al. (1989). A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors. *N. Engl. J. Med.* 320, 479–484. doi: 10.1056/NEJM198902233200802

Fritsche, L. G., Chen, W., Schu, M., Yaspan, B. L., Yu, Y., Thorleifsson, G., et al. (2013). Seven new loci associated with age-related macular degeneration. *Nat. Genet.* 45, 433–439, 9e1–9e2. doi: 10.1038/ng.2578

Fung, A. E., Lalwani, G. A., Rosenfeld, P. J., Dubovy, S. R., Michels, S., Feuer, W. J., et al. (2007). An optical coherence tomography-guided, variable dosing regimen with intravitreal ranibizumab (Lucentis) for neovascular age-related macular degeneration. *Am. J. Ophthalmol.* 143, 566–583. doi: 10.1016/j.ajo.2007.01.028

Gauderman, W. J., Murcray, C., Gilliland, F., and Conti, D. V. (2007). Testing association between disease and multiple SNPs in a candidate gene. *Genet. Epidemiol.* 31, 383–395. doi: 10.1002/gepi.20219

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* 366, 883–892. doi: 10.1056/NEJMoa1113205

Goddard, G. H., and Lewis, C. M. (2010). Risk categorization for complex disorders according to genotype relative risk and precision in parameter estimates. *Genet. Epidemiol.* 34, 624–632. doi: 10.1002/gepi.20519

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., et al. (2013). The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* 15, 761–771. doi: 10.1038/gim.2013.72

Grassmann, F., Fritsche, L. G., Keilhauer, C. N., Heid, I. M., and Weber, B. H. (2012). Modelling the genetic risk in age-related macular degeneration. *PLoS ONE* 7:e37979. doi: 10.1371/journal.pone.0037979

Gruner, C., Ivanov, J., Care, M., Williams, L., Moravsky, G., Yang, H., et al. (2013). Toronto hypertrophic cardiomyopathy genotype score for prediction of a positive genotype in hypertrophic cardiomyopathy. *Circ. Cardiovasc. Genet.* 6, 19–26. doi: 10.1161/CIRCGENETICS.112.963363

Guyon, I., and Elisseff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn.* 3, 1157–1182. Available online at: http://jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explor.* 11, 10–18. doi: 10.1145/1656274.1656278

Harada, H., Miyamoto, K., Yamashita, Y., Nakano, K., Taniyama, K., Miyata, Y., et al. (2013). Methylation of breast cancer susceptibility gene 1 (BRCA1) predicts recurrence in patients with curatively resected stage I non-small cell lung cancer. *Cancer* 119, 792–799. doi: 10.1002/cncr.27754

Hebbring, S. J., Schrodi, S. J., Ye, Z., Zhou, Z., Page, D., and Brilliant, M. H. (2013). A PheWAS approach in studying HLA-DRB1*1501. *Genes Immun.* 14, 187–191. doi: 10.1038/gene.2013.2

Heier, J. S., Brown, D. M., Chong, V., Korobelnik, J. F., Kaiser, P. K., Nguyen, Q. D., et al. (2012). Intravitreal aflibercept (VEGF trap-eye) in wet age-related macular degeneration. *Ophthalmology* 119, 2537–2548. doi: 10.1016/j.optha.2012.09.006

Holmes, G., Donkin, A., and Witten, I. H. (1994). "WEKA: a machine learning workbench," *Intelligent Information Systems, 1994 Proceedings of the 1994 Second Australian and New Zealand Conference.* (Brisbane, QLD), 357–361.

Holtzman, N. A., and Marteau, T. M. (2000). Will genetics revolutionize medicine? *N. Engl. J. Med.* 343, 141–144. doi: 10.1056/NEJM200007133430213

Husing, A., Canzian, F., Beckmann, L., Garcia-Closas, M., Diver, W. R., Thun, M. J., et al. (2012). Prediction of breast cancer risk by genetic risk factors, overall and by hormone receptor status. *J. Med. Genet.* 49, 601–608. doi: 10.1136/jmedgenet-2011-100716

Ibrahim-Verbaas, C. A., Fornage, M., Bis, J. C., Choi, S. H., Psaty, B. M., Meigs, J. B., et al. (2014). Predicting stroke through genetic risk functions: the CHARGE Risk Score Project. *Stroke* 45, 403–412. doi: 10.1161/STROKEAHA.113.003044

Jakobsdottir, J., Gorin, M. B., Conley, Y. P., Ferrell, R. E., and Weeks, D. E. (2009). Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. *PLoS Genet.* 5:e1000337. doi: 10.1371/journal.pgen.1000337

Jang, S., and Atkins, M. B. (2013). Which drug, and when, for patients with BRAF-mutant melanoma? *The lancet oncology* 14, e60–e69. doi: 10.1016/S1470-2045(12)70539-9

Jordan, M. I. (2004). Graphical models. *Stat. Sci.* 19, 140–155. doi: 10.1214/088342304000000026

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., et al. (2013). Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333–339. doi: 10.1038/nature12634

Khoury, M. J., Newill, C. A., and Chase, G. A. (1985). Epidemiologic evaluation of screening for risk factors: application to genetic screening. *Am. J. Public Health* 75, 1204–1208. doi: 10.2105/AJPH.75.10.1204

Kim, S., and Misra, A. (2007). SNP genotyping: technologies and biomedical applications. *Annu. Rev. Biomed. Eng.* 9, 289–320. doi: 10.1146/annurev.bioeng.9.060906.152037

Kimmel, S. E., French, B., Kasner, S. E., Johnson, J. A., Anderson, J. L., Gage, B. F., et al. (2013). A pharmacogenetic versus a clinical algorithm for warfarin dosing. *N. Engl. J. Med.* 369, 2283–2293. doi: 10.1056/NEJMoa1310669

Kingsmore, S. F., Dinwiddie, D. L., Miller, N. A., Soden, S. E., and Saunders, C. J. (2011). Adopting orphans: comprehensive genetic testing of Mendelian diseases of childhood by next-generation sequencing. *Expert Rev. Mol. Diagn.* 11, 855–868. doi: 10.1586/erm.11.70

Kooperberg, C., LeBlanc, M., and Obenchain, V. (2010). Risk prediction using genome-wide association studies. *Genet. Epidemiol.* 34, 643–652. doi: 10.1002/gepi.20509

Krueger, G. G., Langley, R. G., Leonardi, C., Yeilding, N., Guzzo, C., Wang, Y., et al. (2007). A human interleukin-12/23 monoclonal antibody for the treatment of psoriasis. *N. Engl. J. Med.* 356, 580–592. doi: 10.1056/NEJMoa062382

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* 22, 79–86. doi: 10.1214/aoms/1177729694

Kurian, A. W., Hare, E. E., Mills, M. A., Kingham, K. E., McPherson, L., Whittemore, A. S., et al. (2014). Clinical evaluation of a multiple-gene sequencing panel for hereditary cancer risk assessment. *J. Clin. Oncol.* doi: 10.1200/JCO.2013.53.6607. [Epub ahead of print].

Kwan, A., Church, J. A., Cowan, M. J., Agarwal, R., Kapoor, N., Kohn, D. B., et al. (2013). Newborn screening for severe combined immunodeficiency and T-cell lymphopenia in California: results of the first 2 years. *J. Allergy Clin. Immunol.* 132, 140–150. doi: 10.1016/j.jaci.2013.04.024

Kyogoku, C., Langefeld, C. D., Ortmann, W. A., Lee, A., Selby, S., Carlton, V. E., et al. (2004). Genetic association of the R620W polymorphism of protein tyrosine phosphatase PTPN22 with human SLE. *Am. J. Hum. Genet.* 75, 504–507. doi: 10.1086/423790

Lauer, M. S., and D'Agostino, R. B. Sr. (2013). The randomized registry trial–the next disruptive technology in clinical research? *N. Engl. J. Med.* 369, 1579–1581. doi: 10.1056/NEJMp1310102

Lee, S. H., Wray, N. R., Goddard, M. E., and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* 88, 294–305. doi: 10.1016/j.ajhg.2011.02.002

Liaw, D., Marsh, D. J., Li, J., Dahia, P. L., Wang, S. I., Zheng, Z., et al. (1997). Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. *Nat. Genet.* 16, 64–67. doi: 10.1038/ng0597-64

Linardou, H., Dahabreh, I. J., Kanaloupiti, D., Siannis, F., Bafaloukos, D., Kosmidis, P., et al. (2008). Assessment of somatic k-RAS mutations as a mechanism associated with resistance to EGFR-targeted agents: a systematic review and meta-analysis of studies in advanced non-small-cell lung cancer and metastatic colorectal cancer. *Lancet Oncol.* 9, 962–972. doi: 10.1016/S1470-2045(08)70206-7

Loupakis, F., Ruzzo, A., Cremolini, C., Vincenzi, B., Salvatore, L., Santini, D., et al. (2009). KRAS codon 61, 146 and BRAF mutations predict resistance to cetuximab plus irinotecan in KRAS codon 12 and 13 wild-type metastatic colorectal cancer. *Br. J. Cancer* 101, 715–721. doi: 10.1038/sj.bjc.6605177

Lynch, T. J., Bell, D. W., Sordella, R., Gurubhagavatula, S., Okimoto, R. A., Brannigan, B. W., et al. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* 350, 2129–2139. doi: 10.1056/NEJMoa040938

Malik, R., Bevan, S., Nalls, M. A., Holliday, E. G., Devan, W. J., Cheng, Y. C., et al. (2014). Multilocus genetic risk score associates with ischemic stroke in case-control and prospective cohort studies. *Stroke* 45, 394–402. doi: 10.1161/STROKEAHA.113.002938

Mallal, S., Phillips, E., Carosi, G., Molina, J. M., Workman, C., Tomazic, J., et al. (2008). HLA-B*5701 screening for hypersensitivity to abacavir. *N. Engl. J. Med.* 358, 568–579. doi: 10.1056/NEJMoa0706135

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* 363, 166–176. doi: 10.1056/NEJMra0905980

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494

McCarthy, M. I. (2010). Genomics, Type 2 Diabetes, and Obesity. *N. Engl. J. Med.* 363, 2339–2350. doi: 10.1056/NEJMra0906948

Mondul, A. M., Shui, I. M., Yu, K., Travis, R. C., Stevens, V. L., Campa, D., et al. (2013). Genetic variation in the vitamin d pathway in relation to risk of prostate cancer–results from the breast and prostate cancer cohort consortium. *Cancer Epidemiol. Biomarkers Prev.* 22, 688–696. doi: 10.1158/1055-9965.EPI-13-0007-T

Muhlenbruch, K., Jeppesen, C., Joost, H. G., Boeing, H., and Schulze, M. B. (2013). The value of genetic information for diabetes risk prediction - differences according to sex, age, family history and obesity. *PLoS ONE* 8:e64307. doi: 10.1371/journal.pone.0064307

National Research Council. (2010). *A Foundation for Evidence-Driven Practice: A Rapid Learning System for Cancer Care: Workshop Summary*. Washington, DC: The National Academies Press.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., et al. (2011). Tumour evolution inferred by single-cell sequencing. *Nature* 472, 90–94. doi: 10.1038/nature09807

Paez, J. G., Janne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., et al. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 304, 1497–1500. doi: 10.1126/science.1099314

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* 351, 2817–2826. doi: 10.1056/NEJMoa041588

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann Publishers.

Pencina, M. J., D'Agostino, R. B., Sr., D'Agostino, R. B. Jr., and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat. Med.* 27, 157–172. discussion: 207–212. doi: 10.1002/sim.2929

Pendergrass, S. A., Brown-Gentry, K., Dudek, S., Frase, A., Torstenson, E. S., Goodloe, R., et al. (2013). Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet.* 9:e1003087. doi: 10.1371/journal.pgen.1003087

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 748–752. doi: 10.1038/nature08185

Raal, F. J., Santos, R. D., Bloom, D. J., Marais, A. D., Charng, M. J., Cromwell, W. C., et al. (2010). Mipomersen, an apolipoprotein B synthesis inhibitor, for lowering of LDL cholesterol concentrations in patients with homozygous familial hypercholesterolaemia: a randomized, double-blind, placebo-controlled trial. *Lancet* 375, 998–1006. doi: 10.1016/S0140-6736(10)60284-X

Ramsey, B. W., Davies, J., McElvaney, N. G., Tullis, E., Bell, S. C., Drevinek, P., et al. (2011). A CFTR potentiator in patients with cystic fibrosis and the G551D mutation. *N. Engl. J. Med.* 365, 1663–1672. doi: 10.1056/NEJMoa1105185

Rehm, H. L. (2013). Disease-targeted sequencing: a cornerstone in the clinic. *Nat Rev Genet* 14, 295–300. doi: 10.1038/nrg3463

Ripatti, S., Tikkanen, E., Orhu-Melander, M., Havulinna, A. S., Silander, K., Sharma, A., et al. (2010). A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet* 376, 1393–1400. doi: 10.1016/S0140-6736(10)61267-6

Ritchie, M. D., Denny, J. C., Crawford, D. C., Ramirez, A. H., Weiner, J. B., Pulley, J. M., et al. (2010). Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* 86, 560–572. doi: 10.1016/j.ajhg.2010.03.003

Romano, A., Calabria, L. F., Tavanti, F., Minniti, G., Rossi-Espagnet, M. C., Coppola, V., et al. (2013). Apparent diffusion coefficient obtained by magnetic resonance imaging as a prognostic marker in glioblastomas: correlation with MGMT promoter methylation status. *Eur. Radiol.* 23, 513–520. doi: 10.1007/s00330-012-2601-4

Samer, C. F., Lorenzini, K. I., Rollason, V., Daali, Y., and Desmeules, J. A. (2013). Applications of CYP450 testing in the clinical setting. *Mol. Diagn. Ther.* 17, 165–184. doi: 10.1007/s40291-013-0028-5

Saunders, C. J., Miller, N. A., Soden, S. E., Dinwiddie, D. L., Noll, A., Alnadi, N. A., et al. (2012). Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci. Trans. Med.* 4, 154ra35. doi: 10.1126/scitranslmed.3004041

Schellekens, A. F., Franke, B., Ellenbroek, B., Cools, A., de Jong, C. A., Buitelaar, J. K., et al. (2013). COMT Val158Met modulates the effect of childhood adverse experiences on the risk of alcohol dependence. *Addict. Biol.* 18, 344–356. doi: 10.1111/j.1369-1600.2012.00438.x

Schrodi, S. J., Li, Y., Chang, M., Garcia, V. E., Callis Duffin, K., Nair, R. P., et al. (2009). Trait prediction using multi-locus information: psoriasis as a model for complex disease prognostics. Abstracts from the Eighteenth Annual Meeting of the International Genetic Society. *Genet. Epidemiol.* 33, 752–835. doi: 10.1002/gepi.204

Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Schwender, H., and Ickstadt, K. (2008). Identification of SNP interactions using logic regression. *Biostatistics* 9, 187–198. doi: 10.1093/biostatistics/kxm024

Seddon, J. M., Reynolds, R., Yu, Y., and Rosner, B. (2014). Three new genetic loci (R1210C in CFH, variants in COL8A1 and RAD51B) are independently related to progression to advanced macular degeneration. *PLoS ONE* 9:e87047. doi: 10.1371/journal.pone.0087047

Shan, Y., Smelser, D. T., Tromp, G., Kuivaniemi, H., and Weeks, D. E. (2013). "Genetic risk models: model size and confidence intervals of the risk estimates," in *63rd Annual Meeting of The American Society of Human Genetics* (Boston, MA).

Sharma, N. K., Sharma, S. K., Gupta, A., Prabhakar, S., Singh, R., and Anand, A. (2013). Predictive model for earlier diagnosis of suspected age-related macular degeneration patients. *DNA Cell Biol.* 32, 549–555. doi: 10.1089/dna.2013.2072

Shigemizu, D., Abe, T., Morizono, T., Johnson, T. A., Boroevich, K. A., Hirakawa, Y., et al. (2014). The construction of risk prediction models using GWAS data and its application to a type 2 diabetes prospective cohort. *PLoS ONE* 9:e92549. doi: 10.1371/journal.pone.0092549

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* 42, 937–948. doi: 10.1038/ng.686

Stefanutti, C., Gozzer, M., Pisciotta, L., D'Eufemia, P., Bosco, G., Morozzi, C., et al. (2013). A three month-old infant with severe hyperchylomicronemia: molecular diagnosis and extracorporeal treatment. *Atheroscler. Suppl.* 14, 73–76. doi: 10.1016/j.atherosclerosissup.2012.10.020

Thanassoulis, G., and Vasan, R. S. (2010). Genetic cardiovascular risk prediction: will we get there? *Circulation* 122, 2323–2334. doi: 10.1161/CIRCULATIONAHA.109.909309

Tsai, P. C., Liao, Y. C., Wang, Y. S., Lin, H. F., Lin, R. T., and Juo, S. H. (2013). Serum microRNA-21 and microRNA-221 as potential biomarkers for cerebrovascular disease. *J. Vasc. Res.* 50, 346–354. doi: 10.1159/000351767

Uddin, M., Chang, S. C., Zhang, C., Ressler, K., Mercer, K. B., Galea, S., et al. (2013). Adcyap1r1 genotype, posttraumatic stress disorder, and depression among women exposed to childhood maltreatment. *Depress. Anxiety* 30, 251–258. doi: 10.1002/da.22037

van Dieren, S., Beulens, J. W., Kengne, A. P., Peelen, L. M., Rutten, G. E., Woodward, M., et al. (2012). Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. *Heart* 98, 360–369. doi: 10.1136/heartjnl-2011-300734

Velaga, M. R., Wilson, V., Jennings, C. E., Owen, C. J., Herington, S., Donaldson, P. T., et al. (2004). The codon 620 tryptophan allele of the lymphoid tyrosine phosphatase (LYP) gene is a major determinant of Graves' disease. *J. Clin. Endocrinol. Metab.* 89, 5862–5865. doi: 10.1210/jc.2004-1108

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A. Jr., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science* 339, 1546–1558. doi: 10.1126/science.1235122

Walter, M. J., Shen, D., Ding, L., Shao, J., Koboldt, D. C., Chen, K., et al. (2012). Clonal architecture of secondary acute myeloid leukemia. *N. Engl. J. Med.* 366, 1090–1098. doi: 10.1056/NEJMoa1106968

Warren, H., Casas, J. P., Hingorani, A., Dudbridge, F., and Whittaker, J. (2014). Genetic prediction of quantitative lipid traits: comparing shrinkage models to gene scores. *Genet. Epidemiol.* 38, 72–83. doi: 10.1002/gepi.21777

Wei, Z., Wang, W., Bradfield, J., Li, J., Cardinale, C., Frackelton, E., et al. (2013). Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am. J. Hum. Genet.* 92, 1008–1012. doi: 10.1016/j.ajhg.2013.05.002

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–D1006. doi: 10.1093/nar/gkt1229

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation* 97, 1837–1847. doi: 10.1161/01.CIR.97.18.1837

Wimmer, V., Lehermeier, C., Albrecht, T., Auinger, H. J., Wang, Y., and Schon, C. C. (2013). Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics* 195, 573–587. doi: 10.1534/genetics.113.150078

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. doi: 10.1038/nrg3457

Wu, J., Pfeiffer, R. M., and Gail, M. H. (2013). Strategies for developing prediction models from genome-wide association studies. *Genet. Epidemiol.* 37, 768–777. doi: 10.1002/gepi.21762

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., et al. (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* 148, 886–895. doi: 10.1016/j.cell.2012.02.025

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–569. doi: 10.1038/ng.608

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* 43, 519–525. doi: 10.1038/ng.823

Yang, Q., Khoury, M. J., Botto, L., Friedman, J. M., and Flanders, W. D. (2003). Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. *Am. J. Hum. Genet.* 72, 636–649. doi: 10.1086/367923

Yang, Q., Flanders, W. D., Moonesinghe, R., Ioannidis, J. P., Guessous, I., and Khoury, M. J. (2009). Using lifetime risk estimates in personal genomic profiles: estimation of uncertainty. *Am. J. Hum. Genet.* 85, 786–800. doi: 10.1016/j.ajhg.2009.10.017

Yuan, Y. C. (2000). "Multiple imputation for missing data: concepts and new development," in *SAS Users Group International (SUGI) Proceedings 25*. (Indianapolis, IN), P267–25. Available online at: http://www2.sas.com/proceedings/sugi25/PROCCED.pdf; http://www.ats.ucla.edu/stat/sas/library/multipleimputation.pdf

Zhang, Z., Ober, U., Erbe, M., Zhang, H., Gao, N., He, J., et al. (2014). Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PLoS ONE* 9:e93017. doi: 10.1371/journal.pone.0093017

Zimmet, P., Alberti, K. G. M. M., and Shaw, J. (2001). Global and societal implications of the diabetes epidemic. *Nature* 414, 782–787. doi: 10.1038/414782a

Conflict of Interest Statement: Dr. Steven Schrodi is an inventor on US patents and patent applications, without receiving royalties or any compensation. Dr. John Sninsky an employee of Celera which was recently acquired by Quest Diagnostics. Dr. Sninsky does not receive royalties from patents or separate compensation for patent applications. Daniel E. Weeks holds licensed patents regarding risk prediction for age-related macular degeneration using markers on chromosome 10q26. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 13 February 2014; accepted: 15 May 2014; published online: 02 June 2014.
*Citation: Schrodi SJ, Mukherjee S, Shan Y, Tromp G, Sninsky JJ, Callear AP, Carter TC, Ye Z, Haines JL, Brilliant MH, Crane PK, Smelser DT, Elston RC and Weeks DE (2014) Genetic-based prediction of disease traits: prediction is very difficult, especially about the future. Front. Genet. **5**:162. doi: 10.3389/fgene.2014.00162*
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

METHODS ARTICLE
published: 01 August 2014
doi: 10.3389/fgene.2014.00254

Simple, standardized incorporation of genetic risk into non-genetic risk prediction tools for complex traits: coronary heart disease as an example

Benjamin A. Goldstein[1], Joshua W. Knowles[1], Elias Salfati[1], John P. A. Ioannidis[1,2,3] and Themistocles L. Assimes[1]*

[1] *Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA*
[2] *Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA, USA*
[3] *Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA, USA*

Edited by:
Helena Kuivaniemi, Geisinger Health System, USA

Reviewed by:
Qing Lu, Michigan State University, USA
Braxton D. Mitchell, University of Maryland School of Medicine, USA

****Correspondence:***
Themistocles L. Assimes, Stanford University School of Medicine, Population Health Sciences Building, Suite 300, 1070 Arastradero Road, Palo Alto, CA 94304-1334 USA
e-mail: tassimes@stanford.edu

Purpose: Genetic risk assessment is becoming an important component of clinical decision-making. Genetic Risk Scores (GRSs) allow the composite assessment of genetic risk in complex traits. A technically and clinically pertinent question is how to most easily and effectively combine a GRS with an assessment of clinical risk derived from established non-genetic risk factors as well as to clearly present this information to patient and health care providers.

Materials and Methods: We illustrate a means to combine a GRS with an independent assessment of clinical risk using a log-link function. We apply the method to the prediction of coronary heart disease (CHD) in the Atherosclerosis Risk in Communities (ARIC) cohort. We evaluate different constructions based on metrics of effect change, discrimination, and calibration.

Results: The addition of a GRS to a clinical risk score (CRS) improves both discrimination and calibration for CHD in ARIC. Results are similar regardless of whether external vs. internal coefficients are used for the CRS, risk factor single nucleotide polymorphisms (SNPs) are included in the GRS, or subjects with diabetes at baseline are excluded. We outline how to report the construction and the performance of a GRS using our method and illustrate a means to present genetic risk information to subjects and/or their health care provider.

Conclusion: The proposed method facilitates the standardized incorporation of a GRS in risk assessment.

Keywords: genetic risk scores, personalized medicine, coronary heart disease, electronic health records

INTRODUCTION

As genotyping technologies become more common, the interpretation of genetic risk is becoming a bigger component of clinical decision-making. A particular challenge is the interpretation of such genetic information in the context of other clinical health information. Recently, the electronic MEdical Records and GEnomics (eMERGE) network outlined challenges and opportunities for integrating genetic data into an electronic health records (De Jager et al., 2009) system. One issue identified was the automated interpretation of genetic data (Gottesman et al., 2013; Kho et al., 2013; Marsolo and Spooner, 2013; Ury, 2013). The sheer size of genomic data provides many interpretative challenges, particularly in the age of whole genome sequencing with billions of variant base pairs, many of which are *de novo*.

Genetic Risk Scores (GRSs) are one tool for automating the rendition of one's genetic risk. They provide a means to aggregate the health related risk of a collection of genetic alleles into a single number, which can then be used for risk assessment. Using results from genome-wide association studies, one typically combines the observed (or meta-analyzed) log odds-ratio of the risk associated single nucleotide polymorphisms (SNPs). Such scores have been formulated for a variety of complex traits including coronary heart disease (CHD), diabetes, multiple sclerosis and schizophrenia (De Jager et al., 2009; Purcell et al., 2009; Thanassoulis et al., 2012). Overall, GRSs have been shown to modestly improve risk assessment using both traditional and more recently developed model performance metrics (Cook, 2007; Steyerberg et al., 2012).

We anticipate individuals will increasingly approach their physicians with questions regarding their genetic risk of common diseases as high density genetic profiling becomes progressively more routinely available. In this paper, we consider the emerging scenario where a hospital system decides to incorporate genetic data into their EHR for the purposes of clinical risk assessment. One obstacle hampering the effective incorporation of GRSs into clinical practice is the lack of clarity in how to most readily

combine a GRS with a clinical risk assessment. Here, we describe a relatively straightforward method to combine genetic information at established susceptibility loci with a non-genetic risk prediction tool. We illustrate this approach in the context of CHD using a GRS constructed from the most promising association signals reported to date for this disease. We emphasize that the goal of this study is neither to validate the utility of a GRS in risk prediction nor to assess the best way to construct a GRS but rather to demonstrate how one might interpret a GRS and easily incorporate it into a clinical risk assessment. A GRS can be constructed in a variety of ways (Schrodi et al., 2014). One may select SNPs and define their respective high-risk allele either through the investigation of SNP effects within the cohort itself or within external studies that are typically much larger but not necessarily prospective in nature. One may also weigh the high-risk allele by its effect size observed internally or externally. In this study, we used the weighted approach deriving both the SNPs and weights from external sources. Lastly, we illustrate one way to present risk prediction analyses incorporating GRSs to patients and health care providers.

METHODS

SNP SELECTION AND WEIGHTING

We selected SNPs from the most recent and largest multi–stage meta-analysis of GWAS for coronary artery disease conducted by the CARDIoGRAMplusC4D consortium to construct the GRS (CARDIoGRAMplusC4D Consortium et al., 2013). The study included 63,746 cases and 130,681 controls. The vast majority of the subjects included in this meta-analysis reported white/European ancestry. The meta-analysis added 15 new CHD susceptibility loci and confirmed nearly all loci that had previously reached genome-wide significance. The investigators also identified secondary signals at four established loci. Supplementary Table 9 of the CARDIoGRAMplusC4D manuscript lists all uncorrelated SNPs ($r^2 < 0.2$) with an estimated FDR < 5% (CARDIoGRAMplusC4D Consortium et al., 2013). From this list, we selected the 50 SNPs identified by the consortium as validated SNPs because they had reached a genome-wide level of statistical significance in either the CARDIOGRAMplusC4D meta-analysis or in any previous GWAS.

We expect a subset of SNPs to be influencing the risk of CHD through traditional risk factors as the CARDIOGRAMplusC4D meta-analysis adjusted only for age and sex. Indeed, the CARDIoGRAMplusC4D investigators determined that 12 and 5 of these 50 SNPs likely influence CHD risk through effects on lipids and blood pressure based on their strong association with these traits in the Global Lipids Genetics Consortium and the International Consortium of Blood Pressure meta-analyses of GWAS, respectively (CARDIoGRAMplusC4D Consortium et al., 2013). For the purposes of this study, we classified these 17 SNPs as "risk factor SNPs." The remaining 33 SNPs were classified as "non-risk factor SNPs."

PROSPECTIVE COHORT FOR TESTING GENETIC RISK SCORES

We selected the AtherosclerosisRisk in Communities Study (ARIC) study to develop and test a GRS constructed with the 50 SNPs of interest. The ARIC Study is an ongoing prospective investigation of atherosclerosis and its clinical sequelae involving 15,792 white and black persons aged 45–64 years at recruitment (1987–1989). Detailed descriptions of the study designs, IRB consent process, sampling procedures, methods, definitions of cardiovascular outcomes, and approach to statistical analyses is published elsewhere (White et al., 1996; Volcik et al., 2006).

We selected ARIC for several reasons including the availability of individual level genome-wide data for all participants through the National Institutes of Health (National Human Genome Research Institute) controlled access database of Genotypes and Phenotypes (dbGaP), a prolonged follow up with > 1000 incident cases, and no overlap of incident cases with prevalent cases that were included in the CARDIoGRAMplusC4D consortium study (CARDIoGRAMplusC4D Consortium et al., 2013). The Affymetrix 6.0 array was used to genotype all participants of the ARIC study.

All white/Europeans without a history of CHD, myocardial infarction, or heart failure at baseline among the ARIC cohort subjects in dbGAP were eligible for study inclusion. Incident CHD was defined by the recording for the first time of either non-fatal or fatal myocardial infarction ("mi04," "fatchd04"), CHD related revascularization procedure ("in_by04p"), or silent MI detected by ECG ("in_04s").

The outcome of interest was incident CHD within 10 years. Those without a positive event who died or were lost to follow up prior to their 10th year anniversary of follow up were removed from analysis. All others were deemed event free at 10-years regardless of whether they developed incident CHD sometime after their 10 year anniversary of follow up.

CLINICAL RISK SCORE ASSESSMENT

We calculated two clinical risk scores (CRSs) to assess clinical risk at 10 years. The first was the well-known "external" Framingham Risk Score (FRS) for 10-year risk of CHD. The score is based on one's gender, age, total cholesterol, HDL cholesterol, blood pressure, and diabetes and smoking status. Ten-year risk of CHD was calculated using the published regression coefficients (Wilson et al., 1998). The second score was developed "internally" within the ARIC and tested and incorporated the same FRS risk factor variables using cross-validation (see below). Subjects with one or more missing FRS risk factors were excluded from the analysis.

IMPUTATION OF ARIC RAW GENOTYPE DATA TO 1000 GENOMES

We imputed individual level genotype data from ARIC to the latest build of the 1000 genomes project (1 kGP) used a hidden Markov model to minimize the need to use proxy SNPs in the construction of the GRS (Abecasis et al., 2012; Howie et al., 2012). We first phased each chromosome using MaCH (v1.0.16) by running 20 rounds of the Markov sampler and considering 200 haplotypes (states) when updating each individual. We then used phased haplotypes in each chromosome and the latest release of the 1 kGPcosmopolitan panel (version 3 March 2012 release, 246 AFR + 181 AMR + 286 ASN + 379 EUR) to impute all SNPs in the cosmopolitan panel using the OpenMP protocol based multithreaded version of Minimac (v4.6) with 20 rounds and 300 states for each chromosome. Genotyped SNPs used for imputation were

restricted to those with the following features: MAF > 0.1%, missing data per SNP < 2%, and Hardy-Weinberg equilibrium (HWE) $p > 10^{-6}$. Of the 841,820 autosomal genotyped markers, 543,653 passed the initial quality filters and were used for the imputation of over 37 million SNPs in ARIC. We used GTOOL (Genetics Software Suite, (c) 2007, The University of Oxford) to convert Minimac dosage files to best guess genotype calls.

GRS CONSTRUCTION

We calculated the GRS for an individual in the typical approach as a weighted sum of the number of high risk alleles [1].

$$GRS = \sum_{i \in GRS}^{50} \omega_i \sum_{j=1}^{2} RA_{ij} \quad (1)$$

where the inside summation, RA_{ij}, is the count of high risk alleles and the weight, w_i, is the meta-analyzed log odds-ratio for SNP i. We used the corresponding "combined beta" (i.e., the beta across the stage 1 and 2 CARDIOGRAMplusC4D meta-analysis) to weigh the SNP when constructing the GRS. We carefully identified the high-risk allele for each SNP. We used the GTOOL genotype calls to count high-risk alleles for all SNPs in each individual after first dropping SNPs with a low imputation quality ($r^2 < 0.3$).

There are two primary assumptions in such a construction. Since this summation is over marginal effects, each effect is assumed to be independent. The second is that the effects are linearly additive, i.e., there are no interactions. For the first assumption, care was taken to select SNPs that are not in linkage disequilibrium (i.e., correlated) with one another in white/European descent participants ($r^2 < 0.2$). While the second assumption is likely violated, it is also reasonable to assume that marginal effects capture a majority of genetic risk for CHD (Zdravkovic et al., 2002; Speed et al., 2012). When using the GRS we standardize it to have a mean of 0 and standard deviation of 1.

COMBINING CLINICAL AND GENETIC RISK

We present a simple and easy way to combine one's CRS and GRS by using the following model [2]:

$$\log\left(P(CHD\,|\text{Clinical \& Genetic Factors})\right) = \alpha + \beta_1 CRS + \beta_2 GRS \quad (2)$$

This is a standard generalized linear model, where the outcome is a binary (0–1) indicator for incident CHD within 10 years and the predictor variables are the CRS and GRS, respectively. The CRS represents either a calculated risk due to non-genetic clinical factors (as in FRS) or a summation over multiple clinical risk factors (when using internal coefficients). We emphasize the use of a log link function instead of the more frequently used logistic link function (as in logistic regression). This allows the two coefficients of interest (β_1 and β_2) to represent log relative risks (RR), making the following transformation more straightforward. However, we note that using the logistic link one could perform a similar transformation. After exponentiating equation [2], we obtain:

$$P(CHD\,|\text{Clinical \& Genetic}) = e^{\alpha+\beta_1 CRS} \times e^{\beta_2 GRS} = P(CHD\,|Clinical) \times RR^{GRS}_{(GRS)} \quad (3)$$

In the second line, we have combined the intercept (α) with the effect due to clinical factors. This is generally well captured by a CRS (like FRS) that incorporates the prevalence of disease in the general population. Since we are multiplying the estimated effects for the GRS and CRS, the primary assumption is that the GRS is linearly independent of the CRS. This assumption would potentially be violated if the GRS consisted of SNPs that were thought to act entirely or largely through effects on non-genetic clinical risk factors measured at baseline. However, the impact is mitigated by controlling for the CRS while estimating the RR for the GRS in equation [2].

Therefore, to calculate a probability of CHD based on clinical and genetic factors, we must:

(1) Estimate the RR for a one-unit change in GRS on the probability of CHD within 10 years controlled for CRS.
(2) For a given individual:
   (a) Calculate the probability of CHD based on clinical factors via a FRS or Internal Score
   (b) Calculate the GRS (based on equation 1) and standardize it using population mean and standard deviation (SD)
   (c) Multiply the probability from (a) by the RR from (1) raised to the value of standardized GRS from (b) (based on second line of Equation 3)

EVALUATION OF PERFORMANCE OF RISK SCORES

We used 10-fold cross-validation to test both the CRS and GRS, dividing the cohort into a series of independent training and test sets. We created a series of updated risk scores:

(1) A CRS based solely on the FRS (no genetic information considered)
(2) A CRS based solely on the internal coefficients (no genetic information considered)
(3) A CRS updated with a GRS constructed using all SNPs of interest that were either well genotyped or well imputed in ARIC.
(4) A CRS updated with a GRS constructed using only "non–risk factor" SNPs among the SNPs in (3)
(5) A CRS updated with a GRS constructed using only "risk factor" SNPs among the SNPs in (3)

The overall relative risk for a standardized one-unit change in GRS was estimated while incorporating the CRS (either FRS or internal). Within each of the 10-folds, the training (9/10) and test (1/10), we created a standardized score based on the mean and standard deviation from the training set. The models were estimated on the training split and applied to the test split. We used three forms of assessment. First, we calculated the c-statistic to assess discrimination of the various risk scores. Discrimination refers to a model's ability to separate subjects into distinct groups, in this case, those with CHD from those without. Secondly, we calculated the RR for a one standard deviation change in GRS.

Finally, we calculated the calibration slope to assess each models overall calibration (Kramer and Zimmerman, 2007). The calibration of a model is the extent to which the predicted probability reflects the true underlying probability. The calibration slope is a more interpretable statistic than the more typical Hosmer-Lemeshow statistic, representing the degree of miscalibration (Crowson et al., 2014). A calibration slope of 1.0 indicates perfect calibration while values less than 1.0 suggest over-fitting and above 1.0 poorer calibration. For example a calibration slope of 2.0 indicates a two-fold increase in miscalibration. We chose not to assess our models using the Net Reclassification Index (NRI) or the clinical NRI due to recent concerns about the utility and validity of this metric combined with changing clinical guidelines for cardiovascular disease risk assessment (Paynter and Cook, 2012; Ridker and Cook, 2013; Goff et al., 2014; Kerr et al., 2014; Muntner et al., 2014).

In a sensitivity analysis, we repeated the above comparisons but restricted the cohort to those without prevalent diabetes. We also considered a risk prediction model using only a GRS adjusted for age and gender and no other clinical risk factors to provide a perspective on the overall impact of clinical risk factors compared to the genetic risk score. Finally, we assessed the potential for population stratification by performing a principal components analysis (PCA) with 741 ancestry informative markers (AIMs) using EIGENTRAT (Price et al., 2006) followed by a regression of CHD status onto all significant components, adjusted for the clinical factors.

All analyses were performed in R 3.0.1 (R Core Team, 2014).

RISK REPORTS

Using the generated information, we illustrate one means to provide a risk report about an individual's clinical and genetic risk of disease. Three key pieces of information are included:

(1) The number of risk alleles
(2) How the individual's GRS compares to the distribution of GRSs in a comparative population.
(3) The change in one's overall risk after accounting for genetic risk

The number of risk alleles represents a simple count of the number of alleles that have been associated with an increased risk of CHD. The GRS comparison to the general population is based on the individual's standardized GRS. Finally the updated risk is calculated from equation (3). A fourth piece of information that can be included in the risk report is a statement of how the individual's change in overall risk after accounting for genetic risk influences clinical management. This may be based on some well-accepted guidelines whose recommendations can be easily and reliably automated.

RESULTS

ARIC COHORT EXCLUSIONS

Of the 12,771 from the ARIC cohort with phenotypic and genotypic data, 9633 (75%) were white/European (see **Figure 1**). Among the remaining subjects, 721 (7.5%) had a history of CHD or CHF at baseline and were excluded from further analysis.

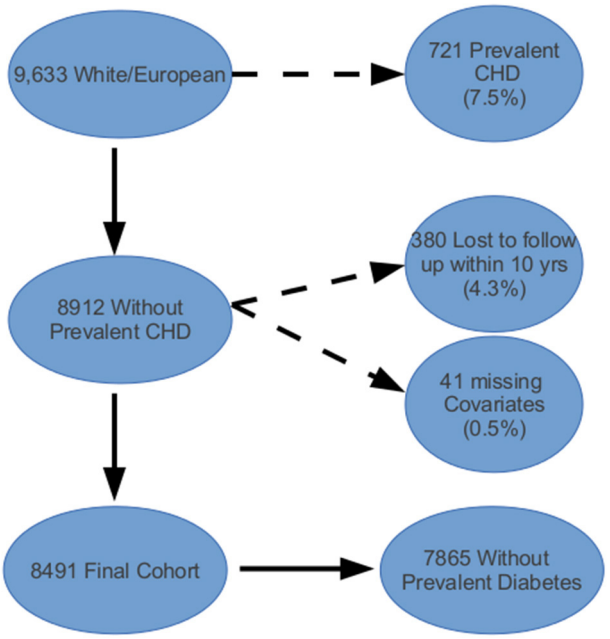


FIGURE 1 | Atherosclerosis Risk in Communities (ARIC) cohort inclusion and exclusion criteria applied to data obtained from the NCBI's database of genotypes and phenotypes (dbGAP).

Lastly, we excluded 380 people who were lost to follow-up or died of non-CHD related factors within 10 years and 41 people with missing covariate information, comprising a final cohort of 8491. **Table 1** shows the baseline characteristics for the ARIC subcohort used in our analyses. The predicted 10-year risk of developing CHD based on the FRS in this subcohort is 7.4% (interquartile range 4.3–12.3%). This predicted risk coincided very well with the observed proportion that developed CHD (7.3%).

RISK SCORES

The 50 SNPs of interest for construction of the GRS are listed in supplemental **Table 1** along with their relationship to risk factors, weights, high risk allele based on the 1000 G reference + strand, imputation quality metrics, and genotype quality control metrics. Of the 50 SNPs, five had an estimated imputation accuracy $r^2 < 0.3$. These five SNPs, which included two SNPs in the *APOE* locus, were dropped from the GRS. The average r^2 of the remaining 45 SNPs was 0.857 (range: 0.361–0.999). The unstandardized mean value of the GRS was 3.17 (SD: 0.347) for all SNPs, 1.95 (0.307) for non-risk factor SNPs alone, and 1.22 (0.160) for risk factor SNPs alone. Interestingly, there was no difference in the unstandardized scores and standard deviations derived from the entire cohort compared to the scores derived from the subset of subjects without diabetes at baseline when considering up to three significant figures. After standardization, the mean and SD of all GRS was 0 and 1 as expected.

PERFORMANCE OF RISK SCORES AND SENSITIVITY ANALYSES

Table 2 summarizes the c-statistics for the 8 risk scores (as well as the age and sex only scores) and the associated RR for a 1-unit

change in the risk score. Adding a GRS improves overall risk discrimination. As expected, the risk score using internal weights demonstrates the best discrimination and calibration. The calibration slope statistics improved (i.e., they become smaller) with the addition of the GRS. A GRS restricted to SNPs that were not related to traditional risk factors performed essentially equally well to a GRS constructed from all SNPs combined, adding about 1 point to the c-statistic. This result suggests that the addition of CHD SNPs that are associated with CHD as well as risk factors will neither aid nor hurt risk assessment. Finally, creating a risk score only with age and sex performed worse than the risk scores with additional clinical factors. However, the improvement in both discrimination and calibration after adding the GRS is comparable to the scores with the full clinical factors.

Table 3 summarizes the same risk score comparisons presented in **Table 2** after removing 626 ARIC participants (7.4%) who reported having diabetes at baseline. We found the general trend of results to be similar to the full cohort despite a smaller sample size. There was a modest improvement in discrimination by about 1 point in the c-statistic as well as improvement in calibration.

PCA revealed eight significant principal components. Only component 3 had a nominal association with CHD ($p = 0.023$, not corrected for number of components tested) suggesting that the addition of PCs into our model for this sample of self reported white/Europeans would not materially influence our results (Supplemental Table 2).

Table 1 | Characteristics of the ARIC subcohort used in analyses ($n = 8491$).

	mean (IQR)
Age (years)	54 (49,59)
SBP (mm/Hg)	116 (106, 128)
DBP (mm/Hg)	71 (65, 78)
HDL (mg/dL)	48 (39, 61)
TC (mg/dL)	211 (187, 238)
	Count (%)
white/European	8491 (100)
Male	3848 (45)
Diabetes	626 (7.4)
SMOKING STATUS	
Current	2010 (24)
Former	2914 (34)
Never	3567 (42)

IQR, inter-quartile range; SBP, Systolic Blood Pressure; DBP, Diastolic Blood Pressure; HDL, High-Density Lipoprotein Cholesterol; TC, Total Cholesterol.

RISK REPORTS

In **Figure 2**, we illustrate a sample report for an individual to show how the addition of a GRS to the model can change the risk assessment that may be used for clinical decision-making. The goal of this report would be to facilitate a conversation around the risk of CHD due to genetics above beyond the known clinical risk factors. At baseline, the participant's estimated risk of CHD at 10 years is 5.5% based on traditional Framingham risk factors. The participant carries 49 of 90 potential risk alleles resulting in a weighted standardized GRS of 1.26 which places the individual in the 89th percentile of genetic risk (i.e., only 11% of the population has a higher risk based on alleles inherited at these 45 SNPs). Combining the participant's genetic risk with their clinical risk results in a final predicted risk of CHD of 7.6% given each SD increase in one's GRS leads to a 38% increase in risk of CHD (**Table 2**). This magnitude of increased risk may

Table 2 | Relative Risks and discrimination metrics for a genetic risk score derived from 50 genome wide significant susceptibility alleles for CHD in the full ARIC sample ($n = 8491$) of white/Europeans subjects.

	Relative Risk (95% CI)	C-statistic*	Calibration Slope
USING FRS FOR CLINICAL RISK SCORE			
FRS alone	–	75.8	7.32
+ full GRS	1.29 (1.20, 1.40)	76.8	6.26
+ GRS restricted to non-risk factor SNPs	1.29 (1.20, 1.40)	76.8	6.29
+ GRS restricted to risk factor SNPs	1.06 (0.98, 1.14)	75.8	7.22
USING INTERNAL COEFFICIENTS FOR CLINICAL RISK SCORE			
Internal coefficients alone	–	77.3	4.34
+ full GRS	1.28 (1.19,1.38)	78.3	4.17
+ GRS restricted to non-risk factor SNPs	1.29 (1.20, 1.39)	78.3	4.18
+ GRS restricted to risk factor SNPs	1.05 (0.97, 1.13)	77.4	4.31
USING ONLY AGE AND SEX			
Internal coefficients alone	–	68.9	11.22
+ full GRS	1.31 (1.22,1.41)	70.4	9.26
+ GRS restricted to non-risk factor SNPs	1.29 (1.20,1.39)	70.1	9.69
+ GRS restricted to risk factor SNPs	1.11 (1.03, 1.20)	69.2	10.79

*CHD, Coronary Heart Disease; ARIC, Atherosclerosis Risk in Communities; FRS, Framingham Risk score; SNPs, Single Nucleotide Polymorphism; GRS, genetic risk score; *performance of second model listed to first model listed.*

Table 3 | Relative Risks and discrimination metrics for a genetic risk score derived from 50 genome wide significant susceptibility alleles for CHD in the ARIC subset of white/Europeans with no diabetes at baseline ($n = 7865$).

	Relative Risk (95% CI)	C-statistic*	Calibration Slope
USING FRS FOR CLINICAL RISK SCORE			
FRS alone	–	75.2	8.84
+ full GRS	1.28 (1.17, 1.39)	76.2	7.02
+ GRS restricted to non-risk factor SNPs	1.30 (1.20, 1.41)	76.3	7.22
+ GRS restricted to risk factor SNPs	1.02 (0.94, 1.11)	75.1	8.67
USING INTERNAL COEFFICIENTS FOR CLINICAL RISK SCORE			
Internal coefficients alone	–	76.7	6.11
+ full GRS	1.28 (1.18, 1.39)	77.6	5.39
+ GRS restricted to non-risk factor SNPs	1.30 (1.20, 1.42)	77.7	5.40
+ GRS restricted to risk factor SNPs	1.03 (0.95, 1.12)	76.6	6.00
USING ONLY AGE AND GENDER			
Internal coefficients alone	–	70.5	12.86
+ full GRS	1.30 (1.20,1.41)	71.8	10.49
+ GRS restricted to non-risk factor SNPs	1.28 (1.18,1.39)	71.6	10.92
+ GRS restricted to risk factor SNPs	1.10 (1.01, 1.19)	70.7	12.44

*CHD, Coronary Heart Disease; ARIC, Atherosclerosis Risk in Communities; FRS, Framingham Risk score; SNPs, Single Nucleotide Polymorphism; GRS, genetic risk score; *performance of second model listed to first model listed.*

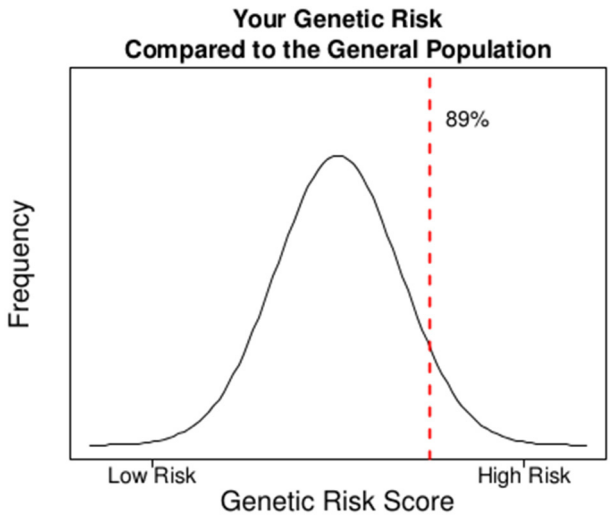


FIGURE 2 | A sample report on CHD risk for an individual in the ARIC study where the incorporation of genetic risk into the model of clinical risk potentially influences clinical management based on current guidelines.

affect the decision to treat this patient with statins (Stone et al., 2014). Ultimately, this person did develop CHD suggesting that the upward adjustment of risk was appropriate.

DISCUSSION

Genetic risk assessment will become an increasingly important component of overall clinical risk assessment. In this context, we ask the question: how can one most easily and effectively incorporate a GRS into an existing clinical risk assessment of a complex trait without compromising effectiveness? We present a straightforward means to combine genetic risk with clinical risk for a given disease where large-scale cohorts with prolonged follow up exist and can be used to evaluate novel biomarkers. Our approach requires knowing only three pieces of information: (1) an individual's GRS, (2) an individual's CRS, and (3) the RR associated with a 1-unit change in standardized GRS within the cohort. Recent studies demonstrate an increasing clinical utility of GRSs for CHD (Brautbar et al., 2012; Hughes et al., 2012; Thanassoulis et al., 2012, 2013; Ganna et al., 2013; Tikkanen et al., 2013). Using our method, we were able to confirm this trend and demonstrate comparable or slightly improved discrimination even when comparing our results to the subset of studies that used a GRS constructed with a similar set of SNPs (Brautbar et al., 2012; Hughes et al., 2012; Thanassoulis et al., 2012; Ganna et al., 2013; Thanassoulis et al., 2013; Tikkanen et al., 2013). We should stress that evidence in the form of a well-executed clinical trial that clearly demonstrates the value of a GRS in improving CHD outcomes does not yet exist (Ioannidis and Tzoulaki, 2010). Thus, we are not endorsing or negating the use of any specific GRS in the primary prevention of CHD on the basis of our results. Ongoing trials are examining the ability of information from GRS to improve outcomes (Knowles et al., 2012; Grant et al., 2013).

Our approach makes the simplifying assumption that the GRS is largely independent of the CRS. This assumption appears reasonable when one reliably restricts SNPs included in the GRS to those influencing risk independent of variables included in the CRS. We tested this assumption by creating two subset GRSs, one restricted to SNPs associated with risk factors and one restricted to SNPs that appear to influence risk of CHD independent of all established risk factors. The non-risk factor GRS performed noticeably better than the risk factor GRS confirming the consequence of grossly violating this assumption. However, we detected no notable difference between the non–risk factor GRS compared to the full GRS. Thus, our approach appears robust to small

violations of this assumption. This confirms others' and our experiences with GRSs that they are fairly robust to alternative constructions (Purcell et al., 2009; Simonson et al., 2011).

An important consideration is the construction of the CRS. We suspect that the ability to derive and make use of such internal coefficients will be facilitated by the increasing availability of EHR with prolonged follow up of individuals receiving care as members of a large-scale health maintenance organization (Ollier et al., 2005; Palmer, 2007; Hoffmann et al., 2011a,b; Kaufman et al., 2012). As expected, the use of internal coefficients led to a slightly more effective CRS compared to the FRS that was developed in a different cohort than ARIC. Despite this observation, we observed a negligible difference in the RR suggesting that perhaps under some circumstances one can develop a GRS using an internal CRS and apply it successfully in other cohorts (or vice-versa). We also note that while the GRS improves calibration, the risk scores overall are still poorly calibrated (> 1), particularly the one using the FRS. This reflects other work that has shown that the external coefficients applied to new populations can often lead to poorly calibrated models (Ridker and Cook, 2013). Finally, the risk score using only age and sex, not surprisingly, performed the worst. Moreover, the improvement in both discrimination (68.9 vs. 77.3) and calibration (11.22 vs. 4.34) after adding additional clinical factors is much greater than after the addition of a GRS highlighting the relative importance of clinical factors collectively at this point in time over the GRS in risk assessment for CHD. However, one should not automatically assume that the current GRS is not clinically useful given its ΔAUC as it is in the same range as that seen for the addition of any single modifiable traditional risk factor to a model that includes all other traditional risk factors.

Several steps need to be followed in reporting of a GRS for a trait using our method to facilitate its testing in additional populations or to easily disseminate its use. First, the cohort in whom the GRS was derived including the age range, sex distribution, risk factor profile, and the ethnicity of its members must be clearly described. The GRS we present here is most relevant to white/Europeans in the age range of 45 to 64 and free of CHD at the time of clinical risk assessment given the eligibility criteria of the ARIC study and the fact that the SNPs used in the GRS were derived from large-scale case-control studies that included subjects in the same race/ethnic group and age range (The ARIC Investigators, 1989; CARDIoGRAMplusC4D Consortium et al., 2013). A different sets of SNPs with different weights will likely be necessary for different race/ethnic groups and possibly different age ranges although we expect substantial overlap across race/ethnic groups in the genomic regions contributing at least one SNP to the GRS (Knowles et al., 2012; Ntzani et al., 2012). Second, one must reliably identify and report which allele was coded as the high-risk allele as this allele is not necessarily the minor allele. Errors in this context due to inadvertent strand flipping either in the original study reporting the susceptibility variant or in the construction of the GRS may have a profound negative impact on the performance of the GRS. Third, the effect estimate for each SNP (generally a log odds ratio) used in the weighting of the GRS should be clearly presented. Lastly, the relative risk for a one-unit change in GRS should be calculated and clearly presented along with the mean and SD of the GRS to facilitate standardization of the score.

We suggest a means to communicate the effect on risk of someone's genetic data when combined with his or her clinical data. Our presentation includes both a contextualization relative to the general population and a statement on how one's inherited variants update one's clinical risk that is based strictly on traditional non-genetic risk factor data. In ongoing clinical investigation, we have applied a similar reporting system within a cardiology clinic (Knowles et al., 2012). Such a report can easily be automated and incorporated into an EHR. Moreover, it can also easily be updated as new susceptibility SNPs are discovered and/or weights refined. Given genome wide genotyping or sequencing is likely to become routine in the near future, more research is needed to identify the optimal way to communicate this information to subjects at risk and health care providers.

Risk scores are likely to evolve over time and practice guidelines may adopt different risk scores. For example, the FRS that we used here forms the basis of the Adult Treatment Panel III (ATPIII) guidelines (2002). Recently, ACC/AHA released new cardiovascular prevention guidelines, with new categories of risk, with a change in the relevant endpoints and in the risk calculation formulas (Goff et al., 2014; Stone et al., 2014). As of this writing, there is still large controversy about the accuracy of the new calculations and the validity of the guidelines (Cook and Ridker, 2013; Ridker and Cook, 2013; Ioannidis, 2014; Muntner et al., 2014). Regardless, our proposed methods can be used to incorporate GRS in any sets of non-genetic predictive models.

In conclusion, we present a simple but effective means to combine a CRS with a GRS and illustrate one way to present such information to an individual interested in understanding how this genetic information influences their risk assessment and thus potentially their clinical management. Furthermore, we highlight information that should be included in all reports of GRSs to facilitate the timely assessment of a new GRS by other investigators in additional populations or, alternatively, to easily incorporate it into clinical practice if its efficacy is no longer in question. We expect the importance of such research to grow over time and hope that future studies will more clearly delineate the optimal way to implement a GRS and how to most effectively disseminate a well-established GRS to patients and their health care providers.

FUNDING SOURCES

Benjamin A. Goldstein is supported by an NIH career development award K25DK097279. Joshua W. Knowles is supported by an American Heart Association, National Fellow to Faculty Award, 10FTF3360005. Themistocles L. Assimes is supported by an NIH career development award K23DK088942.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2014.00254/abstract

REFERENCES

(2002). Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* 106, 3143–3421.

Abecasis, G. R., Auton, A., Brooks, L. D., Depristo, M. A., Durbin, R. M., Handsaker, R. E., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65. doi: 10.1038/nature11632

Brautbar, A., Pompeii, L. A., Dehghan, A., Ngwa, J. S., Nambi, V., Virani, S. S., et al. (2012). A genetic risk score based on direct associations with coronary heart disease improves coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC), but not in the Rotterdam and Framingham Offspring, Studies. *Atherosclerosis* 223, 421–426. doi: 10.1016/j.atherosclerosis.2012.05.035

Consortium, C. A. D., Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T. L., Thompson, J. R., et al. (2013). Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.* 45, 25–33. doi: 10.1038/ng.2480

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 115, 928–935. doi: 10.1161/circulationaha.106.672402

Cook, N. R., and Ridker, P. M. (2013). Response to comment on the reports of overestimation of ASCVD risk using the 2013 AHA/ACC risk equation. *Circulation* 129, 268–269. doi: 10.1161/circulationaha.113.007680

Crowson, C. S., Atkinson, E. J., and Therneau, T. M. (2014). Assessing calibration of prognostic risk scores. *Stat. Methods Med. Res.* doi: 10.1177/0962280213497434. [Epub ahead of print].

De Jager, P. L., Chibnik, L. B., Cui, J., Reischl, J., Lehr, S., Simon, K. C., et al. (2009). Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score. *Lancet Neurol.* 8, 1111–1119. doi: 10.1016/S1474-4422(09)70275-3

Ganna, A., Magnusson, P. K., Pedersen, N. L., De Faire, U., Reilly, M., Arnlov, J., et al. (2013). Multilocus genetic risk scores for coronary heart disease prediction. *Arterioscler Thromb Vasc Biol.* 33, 2267–2272. doi: 10.1161/ATVBAHA.113.301218

Goff, D. C. Jr., Lloyd-Jones, D. M., Bennett, G., Coady, S., D'agostino, R. B., Gibbons, R. Sr., et al. (2014). 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 129, S49–S73. doi: 10.1161/01.cir.0000437741.48606.98

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., et al. (2013). The Electronic Medical Records and Genomics (eMERGE) network: past, present, and future. *Genet. Med.* 15, 761–771. doi: 10.1038/gim.2013.72

Grant, R. W., O'brien, K. E., Waxler, J. L., Vassy, J. L., Delahanty, L. M., Bissett, L. G., et al. (2013). Personalized genetic risk counseling to motivate diabetes prevention: a randomized trial. *Diabetes Care* 36, 13–19. doi: 10.2337/dc12-0884

Hoffmann, T. J., Kvale, M. N., Hesselson, S. E., Zhan, Y., Aquino, C., Cao, Y., et al. (2011a). Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics* 98, 79–89. doi: 10.1016/j.ygeno.2011.04.005

Hoffmann, T. J., Zhan, Y., Kvale, M. N., Hesselson, S. E., Gollub, J., Iribarren, C., et al. (2011b). Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics* 98, 422–430. doi: 10.1016/j.ygeno.2011.08.007

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* 44, 955–959. doi: 10.1038/ng.2354

Hughes, M. F., Saarela, O., Stritzke, J., Kee, F., Silander, K., Klopp, N., et al. (2012). Genetic markers enhance coronary risk prediction in men: the MORGAM prospective cohorts. *PLoS ONE* 7:e40922. doi: 10.1371/journal.pone.0040922

Ioannidis, J. P. (2014). More than a billion people taking statins?: Potential implications of the new cardiovascular guidelines. *JAMA* 311, 463–464. doi: 10.1001/jama.2013.284657

Ioannidis, J. P., and Tzoulaki, I. (2010). What makes a good predictor?: the evidence applied to coronary artery calcium score. *JAMA* 303, 1646–1647. doi: 10.1001/jama.2010.503

Kaufman, D., Bollinger, J., Dvoskin, R., and Scott, J. (2012). Preferences for opt-in and opt-out enrollment and consent models in biobank research: a national survey of Veterans Administration patients. *Genet. Med.* 14, 787–794. doi: 10.1038/gim.2012.45

Kerr, K. F., Wang, Z., Janes, H., McClelland, R. L., Psaty, B. M., and Pepe, M. S. (2014). Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology* 25, 114–121. doi: 10.1097/EDE.0000000000000018

Kho, A. N., Rasmussen, L. V., Connolly, J. J., Peissig, P. L., Starren, J., Hakonarson, H., et al. (2013). Practical challenges in integrating genomic data into the electronic health record. *Genet. Med.* 15, 772–778. doi: 10.1038/gim.2013.131

Knowles, J. W., Assimes, T. L., Kiernan, M., Pavlovic, A., Goldstein, B. A., Yank, V., et al. (2012). Randomized trial of personal genomics for preventive cardiology: design and challenges. *Circ. Cardiovasc Genet.* 5, 368–376. doi: 10.1161/CIRCGENETICS.112.962746

Kramer, A. A., and Zimmerman, J. E. (2007). Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit. Care Med.* 35, 2052–2056. doi: 10.1097/01.CCM.0000275267.64078.B0

Marsolo, K., and Spooner, S. A. (2013). Clinical genomics in the world of the electronic health record. *Genet. Med.* 15, 786–791. doi: 10.1038/gim.2013.88

Muntner, P., Safford, M. M., Cushman, M., and Howard, G. (2014). Comment on the reports of over-estimation of ASCVD risk using the 2013 AHA/ACC risk equation. *Circulation* 129, 266–267. doi: 10.1161/CIRCULATIONAHA.113.007648

National Human Genome Research Institute. *N. DNA Sequencing Costs—Data from the NHGRI Large-Scale Genome Sequencing Program [Online]*. Available online: http://www.genome.gov/sequencingcosts/ [Accessed 2012].

Ntzani, E. E., Liberopoulos, G., Manolio, T. A., and Ioannidis, J. P. (2012). Consistency of genome-wide associations across major ancestral groups. *Hum. Genet.* 131, 1057–1071. doi: 10.1007/s00439-011-1124-4

Ollier, W., Sprosen, T., and Peakman, T. (2005). UK Biobank: from concept to reality. *Pharmacogenomics* 6, 639–646. doi: 10.2217/14622416.6.6.639

Palmer, L. J. (2007). UK Biobank: bank on it. *Lancet* 369, 1980–1982. doi: 10.1016/S0140-6736(07)60924-6

Paynter, N. P., and Cook, N. R. (2012). A bias-corrected net reclassification improvement for clinical subgroups. *Med. Decis. Making* 33, 154–162. doi: 10.1177/0272989x12461856

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. doi: 10.1038/ng1847

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'donovan, M. C., Sullivan, P. F., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 748–752. doi: 10.1038/nature08185

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ridker, P. M., and Cook, N. R. (2013). Statins: new American guidelines for prevention of cardiovascular disease. *Lancet* 382, 1762–1765. doi: 10.1016/S0140-6736(13)62388-0.

Schrodi, S. J., Mukherjee, S., Shan, Y., Tromp, G., Sninsky, J. J., Callear, A. P., et al. (2014). Genetic-based prediction of disease traits: prediction is very difficult, especially about the future(dagger). *Front. Genet.* 5:162. doi: 10.3389/fgene.2014.00162

Simonson, M. A., Wills, A. G., Keller, M. C., and Mcqueen, M. B. (2011). Recent methods for polygenic analysis of genome-wide data implicate an important effect of common variants on cardiovascular disease risk. *BMC Med. Genet.* 12:146. doi: 10.1186/1471-2350-12-146

Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* 91, 1011–1021. doi: 10.1016/j.ajhg.2012.10.010

Steyerberg, E. W., Pencina, M. J., Lingsma, H. F., Kattan, M. W., Vickers, A. J., and Van Calster, B. (2012). Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *Eur. J. Clin. Invest.* 42, 216–228. doi: 10.1111/j.1365-2362.2011.02562.x

Stone, N. J., Robinson, J. G., Lichtenstein, A. H., Bairey Merz, C. N., Blum, C. B., Eckel, R. H., et al. (2014). 2013 ACC/aha guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the american college of cardiology/american heart association task force on practice guidelines. *J. Am. Coll. Cardiol.* 63, 2889–2934. doi: 10.1016/j.jacc.2013.11.002

Thanassoulis, G., Peloso, G. M., and O'donnell, C. J. (2013). Genomic medicine for improved prediction and primordial prevention of cardiovascular disease. *Arterioscler Thromb Vasc Biol.* 33, 2049–2050. doi: 10.1161/ATVBAHA.113.301814

Thanassoulis, G., Peloso, G. M., Pencina, M. J., Hoffmann, U., Fox, C. S., Cupples, L. A., et al. (2012). A genetic risk score is associated with incident cardiovascular disease and coronary artery calcium: the framingham heart study. *Circ. Cardiovasc Genet.* 5, 113–121. doi: 10.1161/CIRCGENETICS.111.961342

The ARIC Investigators. (1989). The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Am. J. Epidemiol.* 129, 687–702.

Tikkanen, E., Havulinna, A. S., Palotie, A., Salomaa, V., and Ripatti, S. (2013). Genetic risk prediction and a 2-stage risk screening strategy for coronary heart disease. *Arterioscler Thromb Vasc Biol.* 33, 2261–2266. doi: 10.1161/ATVBAHA.112.301120

Ury, A. G. (2013). Storing and interpreting genomic information in widely deployed electronic health record systems. *Genet. Med.* 15, 779–785. doi: 10.1038/gim.2013.111

Volcik, K. A., Ballantyne, C. M., Coresh, J., Folsom, A. R., Wu, K. K., and Boerwinkle, E. (2006). P-selectin Thr715Pro polymorphism predicts P-selectin levels but not risk of incident coronary heart disease or ischemic stroke in a cohort of 14595 participants: the Atherosclerosis risk in communities study. *Atherosclerosis* 186, 74–79. doi: 10.1016/j.atherosclerosis.2005.07.010

White, A. D., Folsom, A. R., Chambless, L. E., Sharret, A. R., Yang, K., Conwill, D., et al. (1996). Community surveillance of coronary heart disease in the Atherosclerosis Risk in Communities (ARIC) Study: methods and initial two years' experience. *J. Clin. Epidemiol.* 49, 223–233.

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation* 97, 1837–1847.

Zdravkovic, S., Wienke, A., Pedersen, N. L., Marenberg, M. E., Yashin, A. I., and De Faire, U. (2002). Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. *J. Intern. Med.* 252, 247–254. doi: 10.1046/j.1365-2796.2002.01029.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 18 March 2014; accepted: 10 July 2014; published online: 01 August 2014.
*Citation: Goldstein BA, Knowles JW, Salfati E, Ioannidis JPA and Assimes TL (2014) Simple, standardized incorporation of genetic risk into non-genetic risk prediction tools for complex traits: coronary heart disease as an example. Front. Genet. **5**:254. doi: 10.3389/fgene.2014.00254*
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in Genetics

GENERAL COMMENTARY
published: 07 July 2015
doi: 10.3389/fgene.2015.00231

Corrigendum: Simple, standardized incorporation of genetic risk into non-genetic risk prediction tools for complex traits: coronary heart disease as an example

Benjamin A. Goldstein[1]*, *Joshua W. Knowles*[2], *Elias Salfati*[2], *John P. A. Ioannidis*[3] and *Themistocles L. Assimes*[2]

[1] Department of Biostatistics and Bioinformatics, Center for Predictive Medicine, Duke Clinical Research Institute, Duke University, Durham, NC, USA, [2] Division of Cardiovascular Medicine, Stanford University, Stanford, CA, USA, [3] Department of Medicine, Stanford Prevention Research Center, Stanford University, Stanford, CA, USA

Keywords: risk prediction, genetic risk score (GRS), electronic health records, cardiovascular diseases, coronary disease, biomarkers

OPEN ACCESS

Edited and reviewed by:
Helena Kuivaniemi, Geisinger Health System, USA

***Correspondence:**
Benjamin A. Goldstein, ben.goldstein@duke.edu

Specialty section:
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics

Received: *27 May 2015*
Accepted: *16 June 2015*
Published: *07 July 2015*

Citation:
Goldstein BA, Knowles JW, Salfati E, Ioannidis JPA and Assimes TL (2015) Corrigendum: Simple, standardized incorporation of genetic risk into non-genetic risk prediction tools for complex traits: coronary heart disease as an example. Front. Genet. 6:231. doi: 10.3389/fgene.2015.00231

A corrigendum on

Simple, standardized incorporation of genetic risk into non-genetic risk prediction tools for complex traits: coronary heart disease as an example
by Goldstein, B. A., Knowles, J. W., Salfati, E., Ioannidis, J. P. A. and Assimes TL (2014). Front. Genet. 5:254. doi: 10.3389/fgene.2014.00254

The original Figure 2 did not display the full sample risk report as described in the paper. Here we illustrate how one can convey personalized genetic risk to a patient and how the inclusion of the Genetic Risk Score changes the clinical interpretation of the individual's risk.

Funding

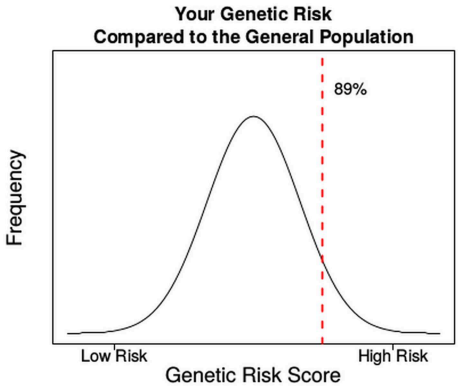
BG is supported by an NIH career development award K25DK097279. JK is supported by an American Heart Association, National Fellow to Faculty Award, 10FTF3360005. TA is supported by an NIH career development award K23DK088942.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

YOUR RISK SCORE

Based on the traditional Framingham risk score, your risk of coronary heart disease over the next 10 years is approximately 5.5%.

We tested for a total of 90 possible risks variants or alleles. Out of these 90, you carry 49 variants that are associated with higher risk. Your genetic profile puts you in the 89 percentile for risk. This means 89% of the general population have a genetic risk score more favorable than you and 11% have a genetic risk score less favorable than you.



Based on the traditional Framingham risk score plus the genetic risk score, your risk of coronary heart disease over the next 10 years is approximately 7.6%.

Your 10 year risk of coronary heart disease risk is ≥7.5% when considering your genetic risk. This information may be discussed with your physician in terms of what would be recommended as most appropriate management given your estimated risk.

frontiers in GENETICS

REVIEW ARTICLE
published: 26 March 2014
doi: 10.3389/fgene.2014.00050

Return of results in the genomic medicine projects of the eMERGE network

Iftikhar J. Kullo[1], Ra'ad Haddad[1], Cynthia A. Prows[2], Ingrid Holm[3], Saskia C. Sanderson[4], Nanibaa' A. Garrison[5], Richard R. Sharp[6], Maureen E. Smith[7], Helena Kuivaniemi[8], Erwin P. Bottinger[4], John J. Connolly[9], Brendan J. Keating[9], Catherine A. McCarty[10], Marc S. Williams[11] and Gail P. Jarvik[12]**

[1] Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN, USA
[2] Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[3] Boston Children's Hospital, Boston, MA, USA
[4] Department of Genetics and Genomic Sciences, Charles R. Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[5] Center for Biomedical Ethics and Society, Department of Pediatrics, Vanderbilt University School of Medicine, Nashville, TN, USA
[6] Biomedical Ethics Program, Mayo Clinic, Rochester, MN, USA
[7] Feinberg School of Medicine, Northwestern University, Chicago, IL, USA
[8] The Sigfried and Janet Weis Center for Research, Geisinger Health System, Danville, PA, USA
[9] Center for Applied Genomics, Abramson Research Center, The Children's Hospital of Philadelphia, Philadelphia, PA, USA
[10] Research Division, Essentia Institute of Rural Health, Duluth, MN, USA
[11] Genomic Medicine Institute, Geisinger Health System, Danville, PA, USA
[12] Departments of Medicine (Medical Genetics) and Genome Sciences, University of Washington, Seattle, WA, USA

Edited by:
Marylyn D. Ritchie, The Pennsylvania State University, USA

Reviewed by:
Yiran Guo, Children's Hospital of Philadelphia, USA
Juli Bollinger, Johns Hopkins University, USA

****Correspondence:***
Iftikhar J. Kullo, Division of Cardiovascular Diseases, Mayo Clinic, 200 First Street Southwest, Rochester, MN 55905, USA
e-mail: kullo.iftikhar@mayo.edu;
Gail P. Jarvik, Medical Genetics, University of Washington, Health Sciences Building, K-253B, Box 357720, Seattle, WA 98195-7720, USA
e-mail: pair@u.washington.edu

The electronic Medical Records and Genomics (eMERGE) (Phase I) network was established in 2007 to further genomic discovery using biorepositories linked to the electronic health record (EHR). In Phase II, which began in 2011, genomic discovery efforts continue and in addition the network is investigating best practices for implementing genomic medicine, in particular, the return of genomic results in the EHR for use by physicians at point-of-care. To develop strategies for addressing the challenges of implementing genomic medicine in the clinical setting, the eMERGE network is conducting studies that return clinically-relevant genomic results to research participants and their health care providers. These genomic medicine pilot studies include returning individual genetic variants associated with disease susceptibility or drug response, as well as genetic risk scores for common "complex" disorders. Additionally, as part of a network-wide pharmacogenomics-related project, targeted resequencing of 84 pharmacogenes is being performed and select genotypes of pharmacogenetic relevance are being placed in the EHR to guide individualized drug therapy. Individual sites within the eMERGE network are exploring mechanisms to address incidental findings generated by resequencing of the 84 pharmacogenes. In this paper, we describe studies being conducted within the eMERGE network to develop best practices for integrating genomic findings into the EHR, and the challenges associated with such work.

Keywords: genomics, electronic health records, incidental findings, implementation, genetic counseling, next generation sequencing, pharmacogenetics

INTRODUCTION

The availability and reduced costs of high-density genotyping and genome sequencing technologies has accelerated genomic discovery (Green et al., 2011). Genome-wide association studies (GWAS) have revealed numerous common genetic variants that influence susceptibility to disease and adverse drug reactions, as well as inter-individual variation in quantitative traits and drug response (Manolio, 2013). Next-generation sequencing has also enabled discovery of variants associated with rare heritable diseases (Yang et al., 2013). Assessing the utility and identifying best practices for integration of this new genomic knowledge into clinical practice to improve patient care is now a major focus in the area of translational genomics (Kullo et al., 2013; Manolio et al., 2013).

The electronic Medical Records and Genomics (eMERGE) network (see Supplementary Figure) was established in 2007 with support from the National Human Genome Research Institute (NHGRI) to further genomic discovery using biorepositories linked to the electronic health record (EHR). The initial phase (Phase I) of the eMERGE network included five sites: Group Health Cooperative/University of Washington, Marshfield Clinic, Mayo Clinic, Northwestern University, and Vanderbilt University. The network was successful in leveraging the EHR to discover new genetic associations including variants that influence traits such as hematologic traits, lipid levels, Alzheimer disease, and

electrocardiographic intervals among many others (Kullo et al., 2010; Naj et al., 2011; Ding et al., 2012, 2013; Rasmussen-Torvik et al., 2012; Crosslin et al., 2013). Phase II of eMERGE began in August 2011 with the addition of Geisinger Health System and Mount Sinai Medical Center to the network. In August 2012, three pediatric sites joined the network: Children's Hospital of Philadelphia, and a joint membership of Cincinnati Children's Hospital Medical Center and Boston Children's Hospital. As was the case for Phase I sites, each new site has its own biorepository linked to clinical phenotypes obtained from the EHR. While continuing efforts at genomic discovery, eMERGE Phase II is also investigating ways of incorporating genomics into the clinical setting (Gottesman et al., 2013), in particular, the return of genomic results in the EHR for use by healthcare providers at the point of care.

Although the focus in Phase I of eMERGE was primarily on genomic discovery efforts, eMERGE investigators also reviewed what research findings should be considered for return to participants (Fullerton et al., 2012). A return of results working group was set up to address the types of genomic findings that might be returned to patients or participants in the future. The genetic data were generated from high-density genotyping arrays and the return of results working group considered two main categories of genomic findings for return to participants. The first related to common variants on these arrays that may have clinical utility, e.g., Factor V Leiden, and hemochromatosis *HFE* variants. The second related to sex chromosome abnormalities such as Klinefelter and Turner syndromes incidentally discovered on signal intensity analysis of fluorescent data from the genotyping arrays. The investigators agreed that the potential to change immediate medical care was an important criterion for considering return of a result. Based on this criterion, homozygosity of single nucleotide polymorphism (SNP) rs6025 (R506Q, also known as the Factor V Leiden mutation) and rs1800562 (*HFE* C282Y, associated with hereditary hemochromatosis) were judged to have the highest clinical relevance (Fullerton et al., 2012). Sex chromosome abnormalities were also considered for return to participants, although context, such as the age of the participant, was felt to be an important factor. The working group summarized these recommendations in a position paper (Fullerton et al., 2012).

In Phase II of eMERGE, genomic discovery efforts continue, but the focus has expanded to include clinical implementation of genomic data. In addition to high-density genotype data in a large number of patients (see Supplementary Table) (Gottesman et al., 2013), targeted next-generation sequencing data of 84 'very important pharmacogenes' (VIPs) will be available for nearly 10,000 patients across the network as part of the eMERGE PGx project. Genetic data being considered for return to participants include SNPs of medical or pharmacogenetic relevance, genetic risk scores for common diseases as well as pharmacogenomic variants, and clinically actionable incidental findings related to next-generation sequencing of 84 VIPs (e.g., *RYR1* and *CACNA1S* genes associated with malignant hyperthermia). Candidate variants for possible return to participants and incorporation into EHR, will be validated in a CLIA-certified laboratory. Clinical integration of genomic data is being explored with active examination of the ethical, legal, and social implications of such integration for both patients and clinicians.

To start addressing the opportunities and challenges in returning genomic results in the clinical setting, each site is conducting genomic medicine pilot projects to return clinically relevant genomic results to participating patients and their health care providers. Additionally, pharmacogenomic information is being placed preemptively in the EHR as part of the network-wide eMERGE PGx project. In this article, we review the current status of the genomic medicine pilot projects at each eMERGE site and the challenges associated with such work. The aim of each project and the genetic variants being considered for return are summarized in **Table 1**. The categories of genomic results being considered for return to patients in eMERGE II include individual or multiplexed SNPs that influence disease susceptibility or drug responses. The following sections provide an overview of the activities in eMERGE II related to return of results in these pilot genomic medicine implementation projects. A separate manuscript is planned to address return of results in the setting of genomic discovery.

INDIVIDUAL SNPs

Individual SNPs associated with disease risk or of pharmacogenomic relevance are being considered for return of results studies within the eMERGE network, and results of this type are being returned in two eMERGE II pilot studies. A variant in the apolipoprotein L1 (*APOL1*) gene that is associated with non-diabetic chronic kidney disease (CKD) in patients of African American ancestry (Tzur et al., 2010; Parsa et al., 2013) is being returned in a pilot study at the Mount Sinai Medical Center. To understand processes and the impact of implementing a screening and decision support system for risk of non-diabetic CKD in African American patients with hypertension and/or family history of renal failure, ∼40 participants will be genotyped for three risk variants in exon 6 of *APOL1*. A catalog of current evidence-based guidelines for the management of hypertension and chronic kidney disease will group the participants into three renal care advice message categories including (i) evaluation of CKD, (ii) identification of CKD progression, and (iii) prevention of CKD progression. In-depth, audio-recorded qualitative interviews are being conducted with patients before and after they receive their *APOL1* results. Analysis of the transcribed interviews will inform the design of the clinical decision support and patient education materials, and the design of quantitative questionnaires for use in a planned larger study of returning *APOL1* results to patients in clinical practice.

At Northwestern University, investigators are following up on the recommendations from the Return of Results workgroup in eMERGE I to include return of potentially actionable findings by reconsenting 150 biobank participants who were genotyped during Phase I of eMERGE. Participants will undergo CLIA-certified genotyping of the Factor V Leiden mutation and the hereditary hemochromatosis *HFE* mutations (C282Y and H63D). Results will be deposited in the EHR and available to physicians. Study participants will be informed of their results via a letter that will

Table 1 | Summary of the eMERGE genomic medicine pilot projects.

Site	Study aim	Variants	Results returned by	EHR integration
Essentia Health	Evaluate genetic markers for increased risk of age-related macular degeneration (AMD) in the clinical setting	Five variants in *CFH*, one variant in *ARMS-2*, one variant in *C3*, and one variant in *ND2*, will be genotyped and results returned	Optometrist	Provider is notified electronically that genetic results and risk score are available in patients' EHR
Geisinger Health System	(1) Incorporate *IL28B* genotyping and genotype-guided therapy into the standard treatment protocol for chronic hepatitis C virus infection (2) Develop a clinical WGS sequencing program. (3) Genetic risk score to identify patients for AAA screening	(1) Two variants in *IL28B* associated with treatment response (2) Identify causal variants based on indication for testing as well as clinically actionable incidental findings (3) Multi-SNP panel	(1) Hepatologist (2) Clinical geneticist and genetic counselor (3) Patient's provider, preventive care team	(1) Electronic Order Set (2) Genomic test report in laboratory section; certain variants (e.g. pharmacogenetic) will have CDS tied to computerized order entry (3) CDS tied to preventive health care reminder system
Group Health Cooperative and University of Washington	Six genes with highly penetrant variants will be sequenced and confirmed pathogenic variants will be returned to participants	Pathogenic variants in *CACNA1C* and *RYR1* (malignant hyperthermia), *KCNH2* and *SCN5A* (long QT syndrome), *RYR2* (catecholaminergic polymorphic ventricular tachycardia), and *LDLR* (hyperlipidemia)	Genetic counselor or medical geneticist	Results placed in the EHR
Mayo Clinic	Disclosure of genomic risk of myocardial infarction using a genetic risk score integrated into the Framingham risk score	28 SNPs associated with coronary heart disease in prior GWAS are genotyped in a CLIA laboratory and returned	Genetic counselor using an EHR-based tool followed by visit with a preventive cardiologist/internist	An EHR-based tool is used to communicate genomic risk. Genotyping results are placed in the EHR
Icahn School of Medicine at Mount Sinai	Evaluate implementation of a screening and decision support system for risk of non-diabetic kidney disease in African Ancestry patients with hypertension and/or family history of renal failure	*APOL1* risk allele status	Genetic counselor or primary care provider	Placed in the EHR and linked to CDS
Northwestern University	Assess the impact and use of genomic results on clinical care by both physicians and patients, to evaluate the use and impact of physician support documents and best practice alerts in the EHR for genomic results, and to evaluate the non-clinical care impact of genomic results on patients	Variants in FV, FII, and *HFE* mutations (*C282Y* and *H63D*)	Physician with referral to study genetic counselor available	Genomic test results available in laboratory section of EHR; Physicians alerted when results received; CDS developed to alert if risk is present; Contextual links to patient and physician information resources
Vanderbilt University	Three major outcomes are being assessed including the efficacy of pharmacogenomics testing in reducing adverse drug events, physician uptake, and patients' knowledge and reaction	Fourteen actionable genetic variants that include: *CYP2C19 *2-*8* (clopidogrel), *CYP2C9 *2-*3, VKORC1 rs9923231* (warfarin), *SLCO1B1 *5* (simvastatin), and *TMPT *1-*3* (thiopurines)	Ordering physician	Genomic test results available in laboratory section of EHR, Patient Summary, and in Patient Portal
Children's Hospital of Philadelphia	The main focus is on prevention of potentially life-threatening drug adverse events	*HLA-B*1502* (carbamazepine induced Stevens-Johnson syndrome) and *TPMT* (thiopurines) variants	Ordering physician	Results shared with providers and placed in the EMR

(Continued)

Table 1 | Continued

Site	Study aim	Variants	Results returned by	EHR integration
Cincinnati Children's Hospital and Boston Children's Hospital	Explore parents' responses to their children's research results	Cincinnati Children's Hospital will return 21 variants of *CYP2D6* test results while Boston Children's Hospital will return hypothetical *CYP2D6* results	Genetics clinical nurse specialist	Results shared with primary care providers but not placed in EHR

AAA, abdominal aortic aneurysm; CDS, clinical decision support; GWAS, genome-wide annotation studies.

be sent by post or accessible via an online patient web-portal called MyChart). Participants will also complete a baseline survey and follow-up surveys at 1 and 6 months after receiving results. Physicians and participants will also have the opportunity to discuss the results during an appointment or be referred to a study genetic counselor. Semi-structured interviews will be conducted with physicians after results are returned to patients and clinical decision support has been triggered in the EHR. The impact and use of genomic results on both physicians and patients will be assessed, as will the use and impact of physician support documents and best practice alerts in the EHR for genomic results.

PHARMACOGENETIC VARIANTS

At the Vanderbilt University site, 14 pharmacogenetic variants including: *CYP2C19* *2–*8 (clopidogrel), *CYP2C9* *2–*3, *VKORC1* rs9923231 (warfarin), *SLCO1B1* *5 (simvastatin), and *TMPT* *1–*3 (thiopurines) will be genotyped in their study population. Vanderbilt University has conducted a number of research studies to assess the efficacy of the Pharmacogenomic Resource for Enhanced Decision in Care and Treatment (PREDICT) program in reducing adverse drug events, physician uptake, and patients' knowledge and attitudes toward such testing. Physician surveys and structured interviews have been performed. Interviews have been conducted in the following groups of patients who have been through the interventional cardiology clinic: (i) no medication change following PREDICT, (ii) a change in clopidogrel medication dosing, (iii) change in statin medication dosing, and (iv) not enrolled in PREDICT. Results of these surveys and interviews are awaited.

At the University of Washington-Group Health site, pathogenic variants in six highly penetrant pharmacogenes will be genotyped in ∼450 participants. These variants include risk alleles in *CACNA1C* (malignant hyperthermia), *RYR1* (malignant hyperthermia), *KCNH2* (long QT syndrome), *SCN5A* (long QT syndrome), *RYR2* (catecholaminergic polymorphic ventricular tachycardia), and *LDLR* (hyperlipidemia). Results are expected to be returned through the Department of Clinical Genetics with appropriate counseling and subsequent documentation in the EHR. Patients will be surveyed regarding their experience. Healthcare providers will be surveyed to assess their use of genetic information in EHR, ease of use, completeness of information in the EHR, and whether any other resources are needed. Additionally, feasibility of implementing genomic clinical decision support will be assessed by interviews with patients and healthcare providers to guide development and testing of prototype interfaces for the EHR.

In a clinical implementation project at Geisinger Health System, all newly diagnosed patients with chronic HCV are genotyped in a CLIA laboratory for two variants in the interleukin 28B gene *(IL28B)* that influence treatment response and can influence medication choice. A genotype-guided treatment decision tree was created in consultation with expert clinicians and an electronic order set was implemented in the EHR to insure that all eligible patients underwent testing. Genotype results for any participant who is prescribed interferon alpha and ribavirin for chronic HCV infection are placed in the EHR and are available to hepatologists initiating treatment. The economic impact of use of these variants is being studied.

Investigators at Children's Hospital of Philadelphia are working on developing tools to engage local practitioners in targeted intervention projects trialing in-house web-based software that integrates with the EHR (Fiks et al., 2013). This new tool can be used to query the institutional biobank as well. So far, 515 patients have been genotyped for *HLA-B*1502* which is associated with risk of developing Stevens-Johnson syndrome following use of carbamazepine (Chen et al., 2011), in addition to 318 patients genotyped for *TPMT* in patients that may be treated with thiopurines.

Investigators at Cincinnati Children's Hospital Medical Center are studying the return of *CYP2D6* variants in pediatric patients in the content of codeine response, and returning these results to the parents. Variants of *CYP2D6* are genotyped using Taqman and long polymerase chain reaction for full gene deletion and duplication. At the time results are returned to parents, they complete a telephone survey about their reactions and plans to share the actual results, and their anticipated preferences regarding receiving hypothetical incidental findings. Follow-up telephone calls are being conducted at 3 and 12 months post-result disclosure to learn how results were used. A subset of parents are also participating in qualitative interviews to further explore their reactions and perceptions to receiving their children's *CYP2D6* results. Although results will not be placed in the EHR, the researchers are asking parents for permission to share the results with their child's primary care providers and will evaluate primary care providers' reactions to receiving the results. At the Boston Children's Hospital site, the same study, including the same survey and qualitative interviews, is being carried out but using hypothetical, not actual *CYP2D6* results, and without the follow-up telephone calls. Providers will be surveyed

about the perceived utility of pharmacogenomics research results in their practice. Between the two sites in Cincinnati and Boston, parents of 400 children will be enrolled. To date, Cincinnati Children's has returned 48 children's *CYP2D6* results to parents and has recontacted 21 parents, all of whom have given permission to share the results with their children's primary care providers.

GENETIC RISK SCORES

Genetic findings of clinical relevance from high-density genotyping arrays include numerous common alleles that influence disease susceptibility. Because most individual variants have modest effect sizes, investigators are exploring the utility of combining the risk variants into genetic risk scores. At the Marshfield/Essentia Health site, a genetic risk score for age-related macular degeneration (AMD) is calculated based on five variants in *CFH*, one variant in *ARMS-2*, one variant in *C3* and one variant in *ND2*. Individuals ($n = 100$) attending optometry clinics are genotyped for these variants (Haines et al., 2005) and a genetic risk score is calculated and incorporated into the EHR by optometrists. Overall risk will be calculated from the genotypes and the patient's smoking status. The results will be shared with the study participants and a telephone survey will be conducted. The participating clinicians will be interviewed after reviewing these results. Recruitment for this study began in June 2013.

At Mayo Clinic, investigators are conducting the Myocardial Infarction Genes (MI-GENES) clinical trial, a pilot study of communicating genetic risk for coronary heart disease (CHD), the leading cause of death in the US. Patients at intermediate risk for CHD based on conventional risk factors will undergo genotyping of ∼28 SNPs that are associated with CHD independent of lipid or blood pressure levels, in a CLIA laboratory (Deloukas et al., 2013). Study participants will be randomized to receive a Framingham risk score or a modified Framingham risk score that incorporates a genetic risk score based on the genotyping results. The genetic risk scores will be placed in the EHR and an EHR-based pictogram will be used to assist in discussing genomic risk of CHD. Patients in the study will be followed for at least 6 months to assess the extent to which there are differences in diet, physical activity, and other lifestyle modifications associated with the communication of CHD-related genomic risk information. Participants will also be evaluated for psychosocial and behavioral changes. Recruitment for MI-GENES started in October 2013.

At Geisinger Health System, investigators are combining clinical risk factors and genomic information to develop a risk score for abdominal aortic aneurysm (AAA), a leading cause of death in older men (Kent et al., 2010; Kuivaniemi et al., 2012). The goal is to develop strategies for population screening that combine genetic susceptibility variants with clinical risk factor data mined from EHR. Patients in the Geisinger MyCode biobank (Gerhard et al., 2013; Gottesman et al., 2013) will be genotyped for variants known to be associated with AAA (Kuivaniemi et al., 2013) and this information used to create a risk score that will then be evaluated in the Geisinger patient population after performing abdominal ultrasonography examination.

TARGETED NEXT-GENERATION SEQUENCING

The eMERGE pharmacogenomics (PGx) project is using targeted sequencing of 84 pharmacogenes to initiate a multi-site test of the concept that genomic sequence information can be coupled to EHRs for use in the clinical setting (Gottesman et al., 2013). The PGRNseq uses a reagent to "capture" exonic sequences of 84 genes important in pharmacokinetic or pharmacodynamics processes for sequencing on next-generation platforms. Genotypes of established pharmacogenomics utility that influence use of simvastatin, clopidogrel, and warfarin will be confirmed in a CLIA environment and be placed pre-emptively in the EHRs of patients who are "at risk" of receiving these drugs.

Sequence information will also be generated for six genes for which the American College of Medical Genetics and Genomics (ACMG) guidelines suggest returning known or expected pathogenic variants given the association with highly penetrant actionable disorders (Green et al., 2013) (**Table 2**). These include genes associated with long QT syndrome genes *(KCNH2* and *SCN5A)*, malignant hyperthermia (*RYR1* and *CACNA1S*), hypercholesterolemia (*LDLR*), and catecholaminergic polymorphic ventricular tachycardia (*RYR2*).

ACMG guidelines emphasize return of mutations in these genes that are known to be pathogenic but also suggest returning novel mutations that are "expected" to be pathogenic. The recommendations have stimulated considerable debate (Green et al., 2013), and in particular, whether patient preferences can or should be incorporated into the return of results pipeline has been highlighted (Allyse and Michie, 2013). While the ACMG recommendations do not apply to research participants, members of the eMERGE network's Consent, Education, Regulation and Consultation (CERC) workgroup have weighed in with concerns about the scope of the ACMG position (Burke and Grefenstette, 2013; Ross et al., 2013). The mechanisms for

Table 2 | Pharmacogenes being sequenced in the eMERGE PGx project and on the ACMG incidental finding list (Green et al., 2013).

Gene	Phenotype	Inheritance*	Variants to report**
RYR2	Catecholaminergic polymorphic ventricular tachycardia	AD	KP
KCNQ1 *KCNH2* *SCN5A*	Romano-Ward long QT syndrome types 1, 2, and 3, Brugada syndrome	AD	KP and EP
LDLR	Familial hypercholesterolemia	SD	KP and EP
RYR1 *CACNA1S*	Malignant hyperthermia susceptibility	AD	KP

SD, semi-dominant inheritance; AD, autosomal dominant.

***EP, expected pathogenic, sequence variation is previously unreported and is of the type that is expected to cause the disorder. KP, known pathogenic variants. Adapted from the ACMG Policy Statement (Green et al., 2013).*

whether and how to return incidental findings are being evaluated at each site. Several sites, in consultation with respective IRBs, are considering confirming pathogenic incidentally found variants for example in *RYR1* and *CACNA1S*, with an orthogonal genotyping method. This would be followed by inviting the study patient to a genetic counseling session during which the option of knowing the research results with confirmation in a CLIA lab will be discussed. **Table 3** summarizes the approach of the various sites toward returning incidental findings generated as part of the eMERGE PGx project.

WHOLE GENOME/EXOME SEQUENCING

Whole genome and exome sequencing are being increasingly utilized in the clinical setting (Yang et al., 2013). Although less than 2% of adults appear to have relevant actionable incidental findings from whole exome sequencing (Dorschner et al., 2013) whether to return such findings is an important topic of debate (Green et al., 2013), particularly in the context of preserving patient autonomy and confidentiality (Klitzman et al., 2013). Institutional committees with diverse expertise will need to adapt guidelines proposed by the ACMG to determine whether and how these can be applied locally, which results will be reported through local EHRs and what format these reports will take. The Clinical Sequencing Exploratory Research (CSER) consortium is also actively investigating return of results in studies utilizing whole genome or whole exome sequencing (https://cser-consortium.org). Several of the eMERGE sites have conducted pilot studies of whole genome/exome sequencing in the clinical setting. For example, at Geisinger Health System, investigators are developing a laboratory report that summarizes results of whole genome sequencing in individuals with intellectual disability and normal chromosomal microarray that have been recruited along with their parents to undergo whole genome sequencing to identify an underlying genetic etiology. Causal variants and incidental findings (from the ACMG list) will be validated using Sanger sequencing. All patients will be informed about the results and undergo counseling. Qualitative

Table 3 | Return of results related to the eMERGE pharmacogenomics (PGx) projects at each of the eMERGE network sites.

Site	Genetic variants to be returned and placed in the EHR	Return of incidental findings of known clinical significance
Geisinger Health System	Pharmacogenetic variants relevant to clopidogrel, warfarin, and simvastatin	IFs will be returned only if they have clear clinical significance.
Group Health Cooperative and University of Washington	PGx variants for carbamazepine sensitivity are approved for inclusion in the EHR	IFs will be returned by a clinical geneticist and this information would be placed in the EHR at the time of the encounter.
Marshfield Clinic	Pharmacogenetic variants relevant to clopidogrel, warfarin, and simvastatin	IFs will be returned only if they are clinically relevant based on input from a physician, clinical geneticist and/or medical specialist in that area of expertise.
Mayo Clinic	Pharmacogenetic variants relevant to clopidogrel, warfarin, and simvastatin	IFs will be reviewed by a multidisciplinary group prior to return.
Icahn School of Medicine at Mount Sinai	4 NYS/CLIA-approved genetic variants relevant to clopidogrel, warfarin, and simvastatin	IFs will not be returned.
Northwestern University	Pharmacogenetic variants relevant to clopidogrel, warfarin, and simvastatin	IFs will not be returned.
Vanderbilt University	11 PGx variants relevant to clopidogrel, warfarin, and simvastatin were already reported	IFs will not be returned.
Children's Hospital of Philadelphia	Variants in several pharmacogenes will be returned: *CYP2D6* and *UGT2B7* (codeine), *CRHR1* (fluticasone propionate), *UGT1A4* (lamotrigine), *KCNH2* (loratadine), *SLCO2B1* (montelukast, *CYP2D6*, *ABCB1*, *OPRM1*, *COMT*, and *UGT2B7* (morphine), *CYP2C19* and *AHR* (omeprazole), *ABCB1* (ranitidine), *ADRB2* (salbutamol), *BDNF* (sertraline), and *IL28B* and *HLA-DR/DQ* (interferon response variants)	IFs will be returned only if they are clinically relevant based on input from a physician, clinical geneticist and/or medical specialist in that area of expertise.
Cincinnati Children's Hospital (CCHMC) and Boston Children's Hospital (BCH)	CCHMC: Genetic variants in *CYP2D6* pre-emptively placed in EMR for children at risk for having surgery. PGx variants placed in EMR at point of care include those relevant for warfarin, thiopurines, tricyclic antidepressants and some SSRIs. BCH: Genetic variants relevant to warfarin	CCHMC: IFs need to be reviewed and approved by IRB before return and placement in EHR. BCH: IFs will be reviewed by Informed Cohort Oversight Board (ICOB) and appropriate action will be determined.

IFs, incidental findings.

data from interviews will be collated and used to improve this process.

INTEGRATION OF GENOMIC FINDINGS INTO THE EHR

A recent theme issue of *Genetics in Medicine* addressed the issues related to integrating genomic findings into the EHR with linkage to clinical decision support at point of care (Kannry and Williams, 2013). Lack of standardized nomenclature for genetic variants is a major hurdle to creating automated decision support. Currently, several groups, including the Health Level 7 Genomics Work Group, are attempting to address this challenge. Additional challenges relate to a number of ethical, legal, and social implications that have been reviewed elsewhere (Hartzler et al., 2013; Hazin et al., 2013). These include uniform provision of genomic CDS to prevent worsening of disparities in healthcare, education of patients and providers, determining which genomic information to include in the EHR, managing incidental findings, privacy and documentation, storage and reinterpretation of genomic data and the results of stakeholder engagement in making these determinations. Education of patients and care providers will be necessary to facilitate the process of return of genomic results. The CERC Work Group in eMERGE has begun to address some of the education issues through jointly developing education materials and a website for patient information (www.myresults.org) and contributing to evaluation of online pharmacogenomics information for physicians. The CERC Working Group also has collaborative interactions with the CSER and Return of Results consortia.

SUMMARY

Genomic technology has advanced greatly through the last decade. As we attempt to implement genomic medicine, many challenges arise, including the need to develop suitable approaches to return results and to deal with incidental findings. We have summarized activities within the eMERGE network that are related to returning genomic results in the EHR setting and also with the return of incidental findings. Initial experiences within the eMERGE network highlight the need for additional studies that address the issues related to disclosure of results from genomic implementation studies. Ongoing work in the eMERGE network will provide important insights into best practices for returning genomic results and dealing with incidental findings using the EHR.

ACKNOWLEDGMENTS

The eMERGE Network is funded by NHGRI, with additional funding from NIGMS through the following grants: U01HG04599 and U01HG006379 to Mayo Clinic; U01HG004610 and U01HG006375 to Group Health Cooperative; U01HG004608 to Marshfield Clinic; U01HG006389 to Essentia Institute of Rural Health; U01HG004609 and U01HG006388 to Northwestern University; U01HG04603 and U01HG006378 to Vanderbilt University; U01HG006385 to the Coordinating Center; U01HG006382 to Geisinger Clinic; U01HG006380 to Icahn School of Medicine at Mount Sinai; U01HG006830 to The Children's Hospital of Philadelphia; and U01HG006828 to Cincinnati Children's Hospital and Boston Children's Hospital. Additional funding included "Utility of genomic data in population screening for abdominal aortic aneurysm" from The Commonwealth Universal Research Enhancement (CURE) program of the Commonwealth of Pennsylvania (Geisinger).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2014.00050/abstract

REFERENCES

Allyse, M., and Michie, M. (2013). Not-so-incidental findings: the ACMG recommendations on the reporting of incidental findings in clinical whole genome and whole exome sequencing. *Trends Biotechnol.* 31, 439–441. doi: 10.1016/j.tibtech.2013.04.006

Burke, D. S., and Grefenstette, J. J. (2013). Toward an integrated meta-model of public health dynamics for preparedness decision support. *J. Public Health Manag. Pract.* 19 (Suppl. 2), S12–S15. doi: 10.1097/PHH.0b013e31828a842f

Chen, P., Lin, J. J., Lu, C. S., Ong, C. T., Hsieh, P. F., Yang, C. C., et al. (2011). Carbamazepine-induced toxic effects and HLA-B*1502 screening in Taiwan. *N. Engl. J. Med.* 364, 1126–1133. doi: 10.1056/NEJMoa1009717

Crosslin, D. R., McDavid, A., Weston, N., Zheng, X., Hart, E., de Andrade, M., et al. (2013). Genetic variation associated with circulating monocyte count in the eMERGE Network. *Hum. Mol. Genet.* 22, 2119–2127. doi: 10.1093/hmg/ddt010

Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T. L., Thompson, J. R., et al. (2013). Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.* 45, 25–33. doi: 10.1038/ng.2480

Ding, K., de Andrade, M., Manolio, T. A., Crawford, D. C., Rasmussen-Torvik, L. J., Ritchie, M. D., et al. (2013). Genetic variants that confer resistance to malaria are associated with red blood cell traits in African-Americans: an electronic medical record-based genome-wide association study. *G3* 3, 1061–1068. doi: 10.1534/g3.113.006452

Ding, K., Shameer, K., Jouni, H., Masys, D. R., Jarvik, G. P., Kho, A. N., et al. (2012). Genetic Loci implicated in erythroid differentiation and cell cycle regulation are associated with red blood cell traits. *Mayo Clin. Proc.* 87, 461–474. doi: 10.1016/j.mayocp.2012.01.016

Dorschner, M. O., Amendola, L. M., Turner, E. H., Robertson, P. D., Shirts, B. H., Gallego, C. J., et al. (2013). Actionable, pathogenic incidental findings in 1,000 participants' exomes. *Am. J. Hum. Genet.* 93, 631–640. doi: 10.1016/j.ajhg.2013.08.006

Fiks, A. G., Grundmeier, R. W., Mayne, S., Song, L., Feemster, K., Karavite, D., et al. (2013). Effectiveness of decision support for families, clinicians, or both on HPV vaccine receipt. *Pediatrics* 131, 1114–1124. doi: 10.1542/peds.2012-3122

Fullerton, S. M., Wolf, W. A., Brothers, K. B., Clayton, E. W., Crawford, D. C., Denny, J. C., et al. (2012). Return of individual research results from genome-wide association studies: experience of the Electronic Medical Records and Genomics (eMERGE) Network. *Genet. Med.* 14, 424–431. doi: 10.1038/gim.2012.15

Gerhard, G. S., Carey, D. J., and Steele, G. D., Jr. (2013). "Electronic health records in genomic medicine," in *Genomic and Personalized Medicine*, eds. G. S. Ginsburg and H. F. Willard (London: Academic Press), 287–294.

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., et al. (2013). The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* 15, 761–771. doi: 10.1038/gim.2013.72

Green, E. D., Guyer, M. S., Manolio, T. A., and Peterson, J. L. (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature* 470, 204–213. doi: 10.1038/nature09764

Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., et al. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* 15, 565–574. doi: 10.1038/gim.2013.73

Haines, J. L., Hauser, M. A., Schmidt, S., Scott, W. K., Olson, L. M., Gallins, P., et al. (2005). Complement factor H variant increases the risk of age-related macular degeneration. *Science* 308, 419–421. doi: 10.1126/science.1110359

Hartzler, A., McCarty, C. A., Rasmussen, L. V., Williams, M. S., Brilliant, M., Bowton, E. A., et al. (2013). Stakeholder engagement: a key component of integrating genomic information into electronic health records. *Genet. Med.* 15, 792–801. doi: 10.1038/gim.2013.127

Hazin, R., Brothers, K. B., Malin, B. A., Koenig, B. A., Sanderson, S. C., Rothstein, M. A., et al. (2013). Ethical, legal, and social implications of incorporating genomic information into electronic health records. *Genet. Med.* 15, 810–816. doi: 10.1038/gim.2013.117

Kannry, J. M., and Williams, M. S. (2013). Integration of genomics into the electronic health record: mapping terra incognita. *Genet. Med.* 15, 757–760. doi: 10.1038/gim.2013.102

Kent, K. C., Zwolak, R. M., Egorova, N. N., Riles, T. S., Manganaro, A., Moskowitz, A. J., et al. (2010). Analysis of risk factors for abdominal aortic aneurysm in a cohort of more than 3 million individuals. *J. Vasc. Surg.* 52, 539–548. doi: 10.1016/j.jvs.2010.05.090

Klitzman, R., Appelbaum, P. S., and Chung, W. (2013). Return of secondary genomic findings vs patient autonomy: implications for medical care. *JAMA* 310, 369–370. doi: 10.1001/jama.2013.41709

Kuivaniemi, H., Ryer, E. J., Yoon, H. R., and Elmore, J. R. (2013). "Genetic risk factors for abdominal aortic aneurysms," in *Aortic Aneurysms: Risk Factors, Diagnosis, Surgery & Repair*, eds D. Fischof and F. Hatig (Hauppauge, NY: Nova Science Publishers, Inc.), 1–30. Available online at: https://www.novapublishers.com/catalog/product_info.php?products_id=37952

Kuivaniemi, H., Tromp, G., Carey, D. J., and Elmore, J. R. (2012). "Molecular biology and genetics of aortic aneurysms," in *Molecular and Translational Vascular Medicine*, eds M. S. Willis, and J. W. Homeister (NewYork, NY: Springer Science+Business Media, Part 1), 3–33. doi: 10.1007/978-1-61779-906-8_1

Kullo, I. J., Ding, K., Jouni, H., Smith, C. Y., and Chute, C. G. (2010). A genome-wide association study of red blood cell taits using the electronic medical record. *PLoS ONE* 5:e13011. doi: 10.1371/journal.pone.0013011

Kullo, I. J., Jarvik, G. P., Manolio, T. A., Williams, M. S., and Roden, D. M. (2013). Leveraging the electronic health record to implement genomic medicine. *Genet. Med.* 15, 270–271. doi: 10.1038/gim.2012.131

Manolio, T. A. (2013). Bringing genome-wide association findings into clinical use. *Nat. Rev. Genet.* 14, 549–558. doi: 10.1038/nrg3523

Manolio, T. A., Chisholm, R. L., Ozenberger, B., Roden, D. M., Williams, M. S., Wilson, R., et al. (2013). Implementing genomic medicine in the clinic: the future is here. *Genet. Med.* 15, 258–267. doi: 10.1038/gim.2012.157

Naj, A. C., Jun, G., Beecham, G. W., Wang, L. S., Vardarajan, B. N., Buros, J., et al. (2011). Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat. Genet.* 43, 436–441. doi: 10.1038/ng.801

Parsa, A., Kao, W. H. L., Xie, D., Astor, B. C., Li, M., Hsu, C.-Y., et al. (2013). APOL1 risk variants, race, and progression of chronic kidney disease. *N. Engl. J. Med.* 369, 2183–2196. doi: 10.1056/NEJMoa1310345

Rasmussen-Torvik, L. J., Pacheco, J. A., Wilke, R. A., Thompson, W. K., Ritchie, M. D., Kho, A. N., et al. (2012). High density GWAS for LDL cholesterol in African Americans using electronic medical records reveals a strong protective variant in APOE. *Clin. Transl. Sci.* 5, 394–399. doi: 10.1111/j.1752-8062.2012.00446.x

Ross, L. F., Rothstein, M. A., and Clayton, E. W. (2013). Mandatory extended searches in all genome sequencing: "incidental findings," patient autonomy, and shared decision making. *JAMA* 310, 367–368. doi: 10.1001/jama.2013.41700

Tzur, S., Rosset, S., Shemer, R., Yudkovsky, G., Selig, S., Tarekegn, A., et al. (2010). Missense mutations in the APOL1 gene are highly associated with end stage kidney disease risk previously attributed to the MYH9 gene. *Hum. Genet.* 128, 345–350. doi: 10.1007/s00439-010-0861-0

Yang, Y., Muzny, D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. A., et al. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* 369, 1502–1511. doi: 10.1056/NEJMoa1306555

Conflict of Interest Statement: The Review Editor Yiran Guo declares that, despite being affiliated to the same institution as authors John J. Connolly and Brendan J. Keating, the review process was handled objectively and no conflict of interest exists.

Received: 12 December 2013; paper pending published: 11 January 2014; accepted: 18 February 2014; published online: 26 March 2014.

*Citation: Kullo IJ, Haddad R, Prows CA, Holm I, Sanderson SC, Garrison NA, Sharp RR, Smith ME, Kuivaniemi H, Bottinger EP, Connolly JJ, Keating BJ, McCarty CA, Williams MS and Jarvik GP (2014) Return of results in the genomic medicine projects of the eMERGE network. Front. Genet. **5**:50. doi: 10.3389/fgene.2014.00050*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on