eMERGE Genomics

Marylyn D Ritchie, PhD

January 2014

Scientific projects for genomics workgroup

eMERGE-I

- Genotyping and sequencing
- Quality control of large scale genomic data
- Genome-wide association studies

eMERGE-II

- Imputation
- Genome-wide association studies
- Gene-gene and gene-environment interactions
- Null variants
- PheWAS/EWAS
- PGx

Quality Control

- Developed a QC pipeline for large scale genomic data
- Developed a strategy for merging multiple datasets into one mega-dataset for genomic analysis
- Benefit of eMERGE-I was access to individual level data across the consortium

Curr Protoc Hum Genet. 2011 Jan;Chapter 1:Unit1.19. doi: 10.1002/0471142905.hg0119s68.

Quality control procedures for genome-wide association studies.

Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, de Andrade M, Doheny KF, Haines JL, Hayes G, Jarvik G, Jiang L, Kullo IJ, Li R, Ling H, Manolio TA, Matsumoto M, McCarty CA, McDavid AN, Mirel DB, Paschall JE, Pugh EW, Rasmussen LV, Wilke RA, Zuvich RL, Ritchie MD.

Author information

Abstract

Genome-wide association studies (GWAS) are being conducted at an unprecedented rate in population-based cohorts and have increased our understanding of the pathophysiology of complex disease. Regardless of context, the practical utility of this information will ultimately depend upon the quality of the original data. Quality control (QC) procedures for GWAS are computationally intensive, operationally challenging, and constantly evolving. Here we enumerate some of the challenges in QC of GWAS data and describe the approaches that the electronic MEdical Records and Genom
discuss
chromo
and dis

PMID: 21

Image

Genet Epidemiol. 2011 Dec;35(8):887-98. doi: 10.1002/gepi.20639.

Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality.

Zuvich RL, Armstrong LL, Bielinski SJ, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, de Andrade M, Doheny KF, Haines JL, Hayes MG, Jarvik GP, Jiang L, Kullo IJ, Li R, Ling H, Manolio TA, Matsumoto ME, McCarty CA, McDavid AN, Mirel DB, Olson LM, Paschall JE, Pugh EW, Rasmussen LV, Rasmussen-Torvik LJ, Turner SD, Wilke RA, Ritchie MD.

Author information

Abstract

Genome-wide association studies (GWAS) are a useful approach in the study of the genetic components of complex phenotypes. Aside from large cohorts, GWAS have generally been limited to the study of one or a few diseases or traits. The emergence of biobanks linked to electronic medical records (EMRs) allows the efficient reuse of genetic data to yield meaningful genotype-phenotype associations for multiple phenotypes or traits. Phase I of the electronic MEdical Records and GEnomics (eMERGE-I) Network is a National Human Genome Research Institute-supported consortium composed of five sites to perform various genetic association studies using DNA repositories and EMR systems. Each eMERGE site has developed EMR-based algorithms to comprise a core set of 14 phenotypes for extraction of study samples from each site's DNA repository. Each eMERGE site selected samples for a specific phenotype, and these samples were genotyped at either the Broad Institute or at the Center for Inherited Disease Research using the Illumina Infinium BeadChip technology. In all, approximately 17,000 samples from across the five sites were genotyped. A unified quality control (QC) pipeline was developed by the eMERGE Genomics Working Group and used to ensure thorough cleaning of the data. This process includes examination of sample and marker quality and various batch effects. Upon completion of the genotyping and QC analyses for each site's primary study, eMERGE Coordinating Center merged the datasets from all five sites. This larger merged dataset reentered the established eMERGE QC pipeline. Based on lessons learned during the process, additional analyses and QC checkpoints were added to the pipeline to ensure proper merging. Here, we explore the challenges associated with combining datasets from different genotyping centers and describe the expansion to eMERGE QC pipeline for merged datasets. These additional steps will be useful as the eMERGE project expands to include additional sites in eMERGE-II, and also serve as a starting point for investigators merging multiple genotype datasets accessible through the National Center for Biotechnology Information in the database of Genotypes and Phenotypes. Our experience demonstrates that merging multiple datasets after additional QC can be an efficient use of genotype data despite new challenges that appear in the process.

GWAS Discoveries in eMERGE I/II

- Many traits
  - Continuous lab values
  - Binary disease outcomes

frontiers in
GENETICS

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

EMR-linked GWAS study: investigation of variation

INVESTIGATION

Gene
Are
Afric
Reco

Keyue Di
Marylyn D
Abel N. K
*Division of
Informatics,
Research Ins
and §Center
and Molecu
Informatics a
**Departme
†††Departm

ABSTRAC
in African-
into disco
(HGB), her
and mean

Human Molecular Genetics, 2013, Vol. 22, No. 10 2119–2127
doi:10.1093/hmg/ddt010
Advance Access published on January 12, 2013

Genetic variation associated with circulating monocyte count in the eMERGE Network

David R. Crosslin[1,2,*], Andrew McDavid[9], Noah Weston[10], Xiuwen Zheng[3], Eugene Hart[10], Mariza de Andrade[11], Iftikhar J. Kullo[12], Catherine A. McCarty[13,14], Kimberly F. Doheny[15], Elizabeth Pugh[15], Abel Kho[18], M. Geoffrey Hayes[19], Marylyn D. Ritchie[20], Alexander Saip[21], Dana C. Crawford[22,23], Paul K. Crane[4], Katherine Newton[10], David S. Carrell[10], Carlos J. Gallego[1], Michael A. Nalls[24], Rongling Li[26], Daniel B. Mirel[27], Andrew Crenshaw[27], David J. Couper[28], Toshiko Tanaka[29], Frank J.A. van Rooij[30,31], Ming-Huei Chen[32,33], Albert V. Smith[34,35], Neil A. Zakai[36,37], Qiong Yango[32,38], Melissa Garcia[25], Yongmei Liu[39], Thomas Lumley[5], Aaron R. Folsom[40], Alex P. Reiner[6], Janine F. Felix[30,31], Abbas Dehghan[30,31], James G. Wilson[41], Joshua C. Bis[7], Caroline S. Fox[32,42], Nicole L. Glazer[7], L. Adrienne Cupples[32,38], Josef Coresh[16], Gudny Eiriksdottir[34], Vilmundur Gudnason[34,35], Stefania Bandinelli[43], Timothy M. Frayling[44], Aravinda Chakravarti[17], Cornelia M. van Duijn[30,31], David Melzer[45,46], Daniel Levy[32,47], Eric Boerwinkle[48], Andrew B. Singleton[27], Dena G. Hernandez[27,49], Dan L. Longo[50], Jacqueline C.M. Witteman[30,31], Bruce M. Psaty[8,51], Luigi Ferrucci[29], Tamara B. Harris[25], Christopher J. O'Donnell[32,47,52], Santhi K. Ganesh[53], CHARGE Hematology Working Group, Eric B. Larson[10], Chris S. Carlson[9] and Gail P. Jarvik[1,2], The electronic Medical Records and Genomics (eMERGE) Network

Imputation

BEAGLE Imputed Data (Adult Sites only)

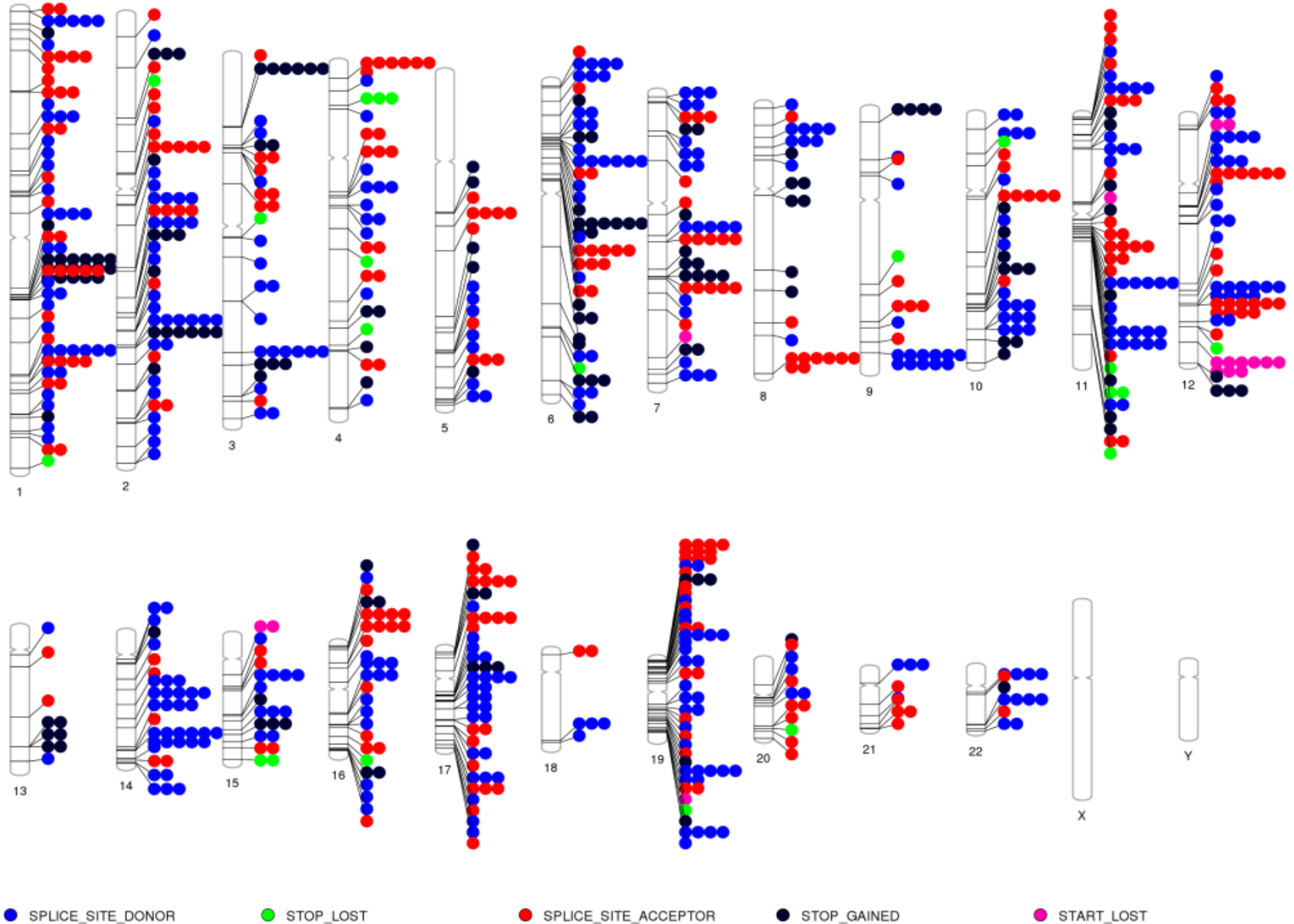
	# Genotyped Samples	# BEAGLE Imputed SNPs
Merged eMERGE-I 1M	2,634	
Merged eMERGE-I 660	16,029	
Adult sites (unmerged)	*19,625*	
Adult Site Total	**38,288**	**15,212,466**

Impute2 Imputed Data (Adult and Pediatric)	
	# Genotyped Samples
Merged eMERGE-I 1M	2,634
Merged eMERGE-I 660	16,029
Geisinger	3,111
Group Health	731
Marshfield	500
Mayo	3121
Mt. Sinai	6,290
NU	2,951
Vanderbilt	3,461
BCH	1,038
CCHMC	4,322
CHOP	6,850
Total - All Impute2 Imputed Samples	**51,038**

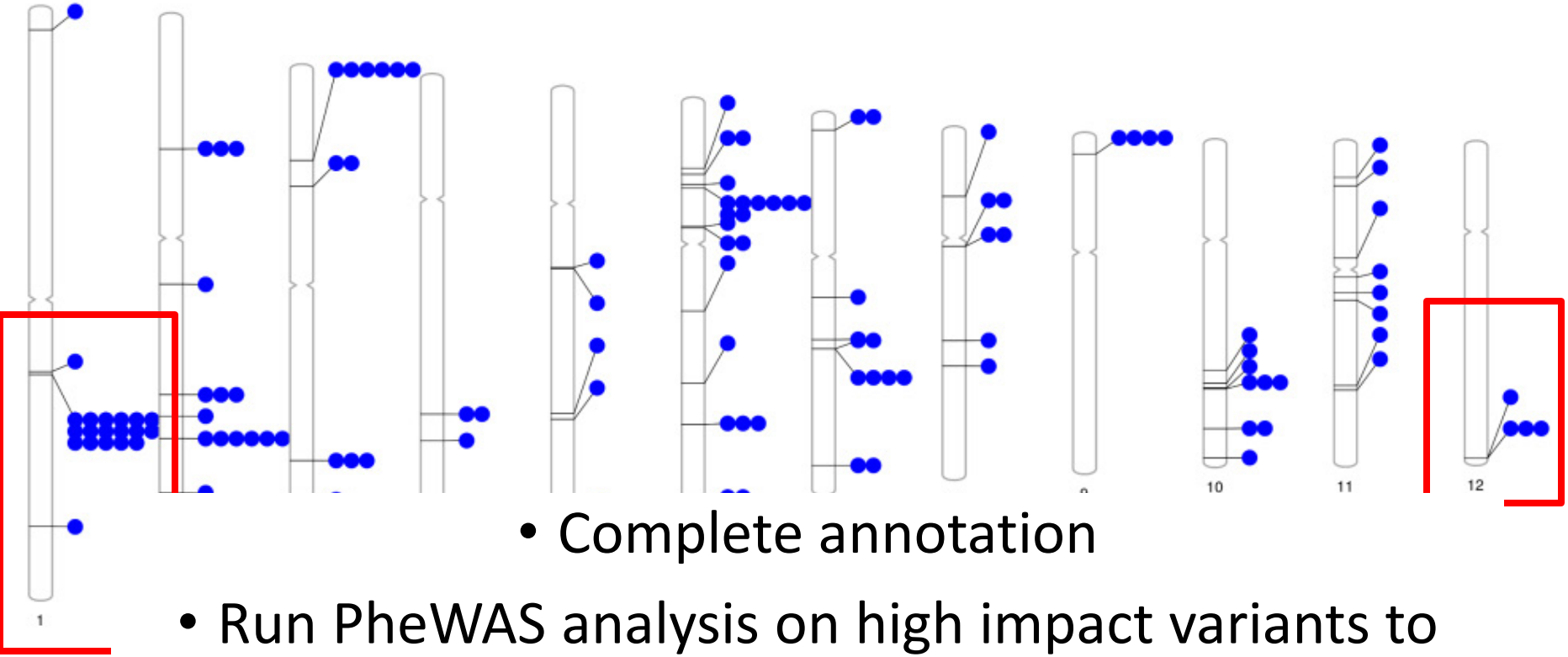
Null variants

- Hypothesis: With our large eMERGE dataset, we should be able to identify multiple rare, null variants and look for correlation with clinical traits.
- Used bioinformatics approach to predict null variants
- Explored genotyped and imputed datasets for the occurrence of null variants

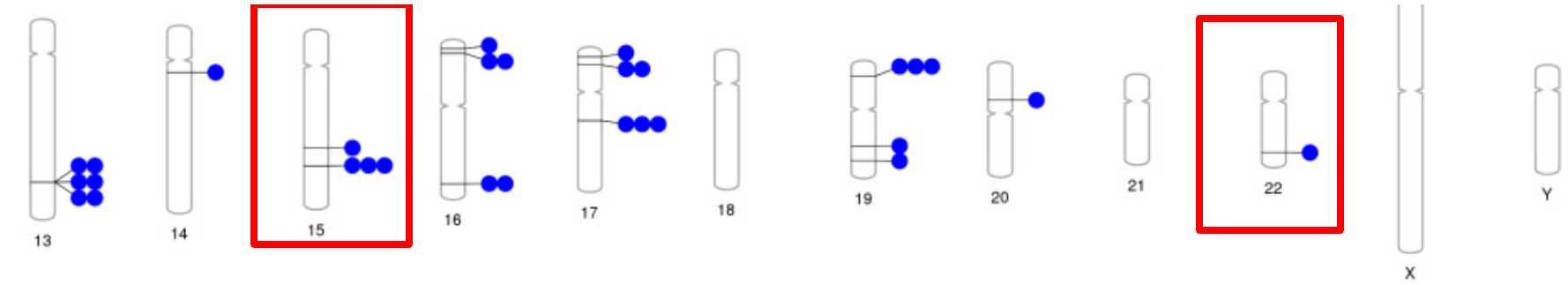
Phenogram of High Impact Variants



Phenogram of Stop Gain Variants



- Complete annotation
- Run PheWAS analysis on high impact variants to look for association with clinical characteristics



http://visualization.ritchielab.psu.edu/phenograms/plot

eMERGE-PGx

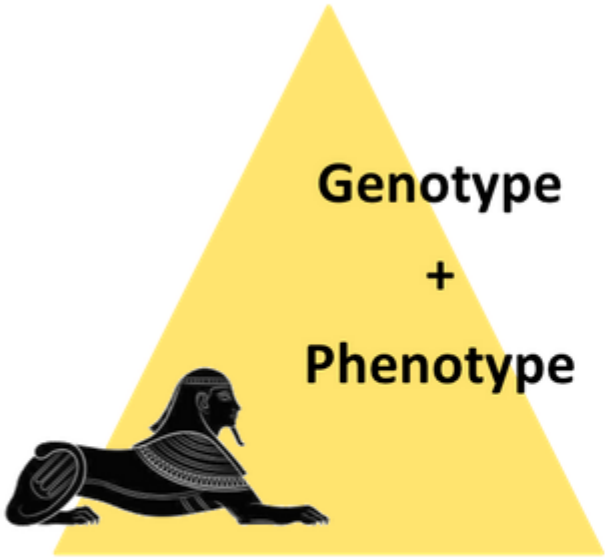
- PGRN-seq platform being used across the network
- 84 pharmacogenes
- Next-generation sequencing of these 84 genes
- eMERGE-CC performing multiple-sample variant calling on the full eMERGE dataset
- Different quality control analyses underway
- Association analyses of both common variants and rare variants from these data

SPHINX

A resource of the *e*MERGE Network

Network Login

*S*equence, *P*henotype, and p*H*armacogenomics *IN*tegration e*X*change



SPHINX is a web-based tool for exploring drug response implications of genetic variation across the eMERGE PGx project cohort.

The eMERGE-PGx project is a multi-center pilot of pharmacogenetic sequencing in clinical practice. Once fully enrolled, SPHINX will contain data on nearly 9000 subjects from participating electronic medical record (EMR) systems. PGRN-Seq sequencing will identify common variants, some with known clinical implications, and also variants of unknown significance.

The eMERGE-PGx project is a multi-center pilot of pharmacogenetic sequencing in clinical practice. Once fully enrolled, SPHINX will contain data on nearly 9000 subjects from participating electronic medical record (EMR) systems. PGRN-Seq sequencing will identify common variants, some with known clinical implications, and also variants of unknown significance. SPHINX has a public-facing gene variant repository that provides variant summary data.

The sites participating in eMERGE and the eMERGE-PGx project include:

- Children's Hospital of Pennsylvania
- Cincinnati Children's Medical Center with Boston Children's Hospital
- Geisinger Health System
- Group Health Cooperative with University of Washington
- Essential Rural Health with Marshfield Clinic and The Pennsylvania State University

eMERGE-III Future Plans

- Enormous potential for future discoveries
  - eMERGE has a set of over 50,000 with shared individual level data
    - Most consortia only have summary statistics
  - EHR enables vast phenotyping potential for GWAS
    - Treatment outcomes
    - Disease subsets based on clinical characteristics
    - Subphenotyping/endophenotypes
    - Extreme phenotypes
    - Direction of causality
    - Longitudinal GWAS

eMERGE-III Future Plans

- Other types of analyses
  - Racial/ancestry disparities
  - Structural variants (CNVs)
  - Low frequency variant analysis
  - GxG and GxE – more traits, alternative methods
  - Pathway analysis – integrating functional data from ENCODE
  - Integrating other epidemiological data
  - Integrating GIS data
- Other molecular data
  - RNAseq
  - Methylation
  - Sequencing data
  - Targeting high throughput genotyping

Gene-Gene and Gene-Environment Interactions

Pac Symp Biocomput. 2013:147-58.

Next-generation analysis of cataracts: determining knowledge driven gene-gene interactions using Biofilter, and gene-environment interactions using the PhenX Toolkit.

Pendergrass SA, Verma SS, Holzinger ER, Moore CB, Wallace J, Dudek SM, Huggins W, Kitchner T, Waudby C, Berg R, McCarty CA, Ritchie MD.

Author information

Abstract

Investigatin
area of rese
phenotypin
Personalize
Genomics
nucleotide
investigatio
called Biofil
Using the E
that require
well as an
SNPs and
previously a
associated
beyond GW
etiology of

PMID: 234241

PLoS One. 2011 May 11;6(5):e19586. doi: 10.1371/journal.pone.0019586.

Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks.

Turner SD, Berg RL, Linneman JG, Peissig PL, Crawford DC, Denny JC, Roden DM, McCarty CA, Ritchie MD, Wilke RA.

Author information

Abstract

Genome-wide association studies (GWAS) are routinely being used to examine the genetic contribution to complex human traits, such as high-density lipoprotein cholesterol (HDL-C). Although HDL-C levels are highly heritable (h(2)~0.7), the genetic determinants identified through GWAS contribute to a small fraction of the variance in this trait. Reasons for this discrepancy may include rare variants, structural variants, gene-environment

www.ashg.org/2013meeting/abstracts/fulltext/f130123027.htm

Apps | Ritchie Lab Blogs | Ritchie Lab Wiki | PASS Explorer | PSU-Websites | Interesting-Websites | Pegula | Other bookmarks

Replication of gene-gene interaction models associated with cataracts in the eMERGE Network. *M. A. Hall[1], S. S. Verma[1], E. R. Holzinger[1], R. Berg[2], J. Connolly[3], D. C. Crawford[4], D. R. Crosslin[5], M. de Andrade[6], K. F. Doheny[7], J. L. Haines[4], J. B. Harley[8], G. P. Jarvik[5], T. Kitchner[2], H. Kuivaniemi[9], E. B. Larson[5,10], G. Tromp[9], S. A. Pendergrass[1], C. A. McCarty[11], M. D. Ritchie[1]* 1) Center for Systems Genomics, The Pennsylvania State University, University Park, PA; 2) Marshfield Clinic, Marshfield, WI; 3) Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA; 4) Center for Human Genetics Research, Vanderbilt University, Nashville, TN; 5) Department of Genome Sciences, University of Washington, Seattle, WA; 6) Mayo Clinic, Rochester, MN; 7) Center for Inherited Disease Research, IGM, Johns Hopkins University SOM, Baltimore, MD; 8) Cincinnati Children's Hospital, University of Cincinnati, Department of Pediatrics, Cincinnati, OH; 9) Geisinger Health System, Danville, PA; 10) Group Health Research Institute, Seattle, WA; 11) Essentia Rural Health, Duluth, MN.

Bioinformatics approaches to examine epistasis provide the means to discover the interactions between multiple genes and pathways that are likely the basis of complex disease. Despite its importance, extensive computational demands and adjusting for multiple testing make uncovering these interactions a challenge when explored with an exhaustive combinatorial search. Here, we address this issue using Biofilter 2.0 to identify putative SNP-SNP models for cataract susceptibility, reducing the number of models for analysis. With Biofilter 2.0, we created biologically relevant SNP-SNP models from genes with published associations, including genes belonging to the same pathway or having known biological interactions. Using PLATO software, we evaluated these models using logistic regression, adjusting for sex and principal components in 3,907 samples (1,354 controls, 2,553 cases) of European (3872), African (1), Asian (14), and other (13) descent from the Marshfield Clinic Personalized Medicine Research Project, part of the Electronic Medical Records & Genomics (eMERGE) Network. All highly significant models from the Marshfield Clinic (likelihood ratio test (LRT) $p < 0.0001$) were then tested in a replication dataset of 3,483 individuals (537 controls, 2,946 cases) of European (3251), African (113), Asian (66), and other (53) descent, using independent samples from additional sites in the eMERGE Network: Mayo Clinic, Group Health Cooperative, and Vanderbilt University Medical Center. Over 100 SNP-SNP models were found in the replicating sample at LRT $p < 0.01$, and 8 models replicated with high significance (LRT $p < 10\text{-}4$). The most significant replicating SNP-SNP models and their nearest genes included rs7749147 (FYN) - rs11017910 (DOCK1), rs9790292 (TGFBR2) - rs8110090 (TGFB1), rs10176426 (UGT1A10) - rs17863787 (UGT1A6), and rs11723463 (UGT2B4) - rs1112310 (UGT1A10). Notably, the genes UGT1A10 and UGT1A6, members of the UDP glucuronosyltransferase 1 family, and UGT2B4, of the UDP glucuronosyltransferase 2 family are involved in the porphyrin and chlorophyll metabolism pathway. This pathway has demonstrated association with cataracts, and therefore, bears further inquiry. These findings indicate the role of epistasis in susceptibility to cataracts and demonstrate the utility of Biofilter 2.0 as a biology-driven method, which can be applied to any GWAS dataset for investigation of the complex genetic architecture of common diseases.

Summary

- There is a lot more discovery that can be done in eMERGE
- These discoveries can be made:
  - By performing more comprehensive analyses with the data that we already have
  - By performing more data generation via sequencing, high-throughout genotyping, other techniques (tissues) to capture genomic variation that we do not yet have genome-wide in eMERGE
- We should try to incorporate more types of data (environmental, etc.)
- Methodological approaches to analyzing these data are needed and critical