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.