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)

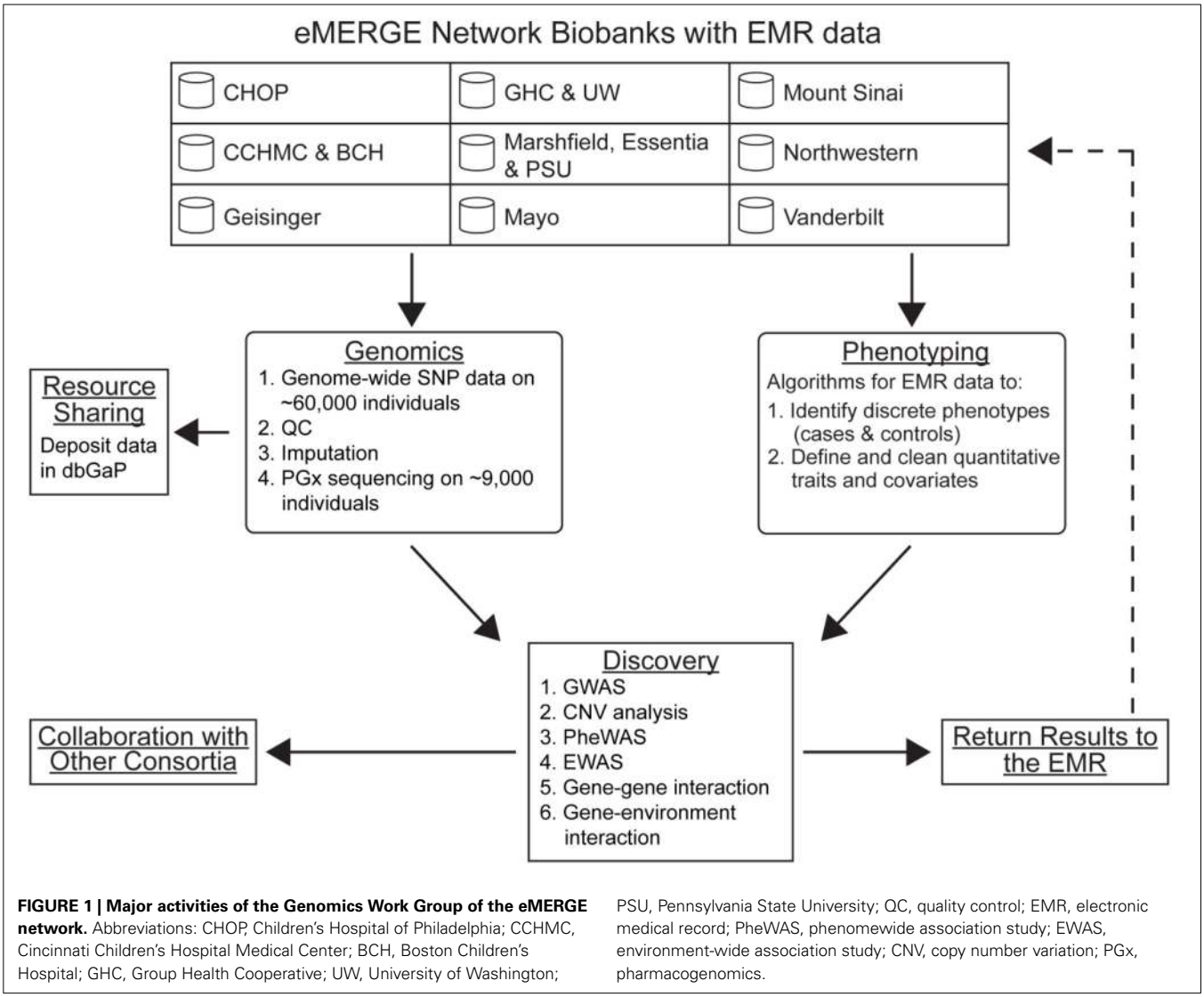

FIGURE 1 | Major activities of the Genomics Work Group of the eMERGE network. Abbreviations: CHOP, Children's Hospital of Philadelphia; CCHMC, Cincinnati Children's Hospital Medical Center; BCH, Boston Children's Hospital; GHC, Group Health Cooperative; UW, University of Washington; PSU, Pennsylvania 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.