Received: 12 April 2018 | Revised: 10 August 2018 | Accepted: 28 August 2018

DOI: 10.1002/gepi.22167

RESEARCH ARTICLE

WILEY Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype

Ian B. Stanaway[1] | Taryn O. Hall[1] | Elisabeth A. Rosenthal[2] | Melody Palmer[2] | Vivek Naranbhai[1,3] | Rachel Knevel[3] | Bahram Namjou-Khales[4] | Robert J. Carroll[5] | Krzysztof Kiryluk[6] | Adam S. Gordon[2] | Jodell Linder[7] | Kayla Marie Howell[7] | Brandy M. Mapes[7] | Frederick T.J. Lin[8] | Yoonjung Yoonie Joo[8] | M. Geoffrey Hayes[8] | Ali G. Gharavi[6] | Sarah A. Pendergrass[9] | Marylyn D. Ritchie[10] | Mariza de Andrade[11] | Damien C. Croteau-Chonka[3] | Soumya Raychaudhuri[3,12] | Scott T. Weiss[3] | Matt Lebo[3] | Sami S. Amr[3] | David Carrell[13] | Eric B. Larson[13] | Christopher G. Chute[14] | Laura Jarmila Rasmussen-Torvik[8] | Megan J. Roy-Puckelwartz[8] | Patrick Sleiman[15] | Hakon Hakonarson[15] | Rongling Li[16] | Elizabeth W. Karlson[10] | Josh F. Peterson[5] | Iftikhar J. Kullo[11] | Rex Chisholm[8] | Joshua Charles Denny[5] | Gail P. Jarvik[2] | The eMERGE Network[16] | David R. Crosslin[1]

[1]Department of Biomedical Informatics Medical Education, School of Medicine, University of Washington, Seattle, Washington

[2]Division of Medical Genetics, School of Medicine, University of Washington, Seattle, Washington

[3]Harvard Medical School, Harvard University, Cambridge, Massachusetts

[4]Department of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio

[5]Departments of Biomedical Informatics and Medicine, Vanderbilt University, Nashville, Tennessee

[6]Department of Medicine, Columbia University, New York City, New York

[7]Vanderbilt Institute for Clinical and Translational Research, School of Medicine, Vanderbilt University, Nashville, Tennessee

[8]Feinberg School of Medicine, Northwestern University, Chicago, Illinois

[9]Geisinger Research, Rockville, Marland

[10]Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania

[11]Mayo Clinic, Rochester, Minnesota

[12]Program in Medical and Population Genetics, Broad Institute of Massachusetts Technical Institute and Harvard University, Cambridge, Massachusetts

[13]Kaiser Permanente Washington Health Research Institute (Formerly Group Health Cooperative-Seattle), Kaiser Permanente, Seattle, Washington

[14]Schools of Medicine, Public Health, and Nursing, Johns Hopkins University, Baltimore, Maryland

[15]Children's Hospital of Philadelphia, Philadelphia, Pennsylvania

[16]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland

Correspondence

Ian B. Stanaway and David R. Crosslin, Department of Biomedical Informatics Medical Education, School of Medicine,

Abstract

The Electronic Medical Records and Genomics (eMERGE) network is a network of medical centers with electronic medical records linked to existing biorepository

University of Washington, Seattle, WA 98109.
Email: bard@uw.edu and davidcr@uw.edu

Funding information
National Human Genome Research Institute, Grant/Award Number: U01HG8657 (Kaiser Washington/University of Washington), U01HG8685 (Brigham and Women's Hospital), U01HG8672 (Vanderbilt University Medical Center), U01HG8666 (Cincinnati Children's Hospital Medical Center), U01HG6379 (Mayo Clinic), U01HG8679 (Geisinger Clinic), U01HG8680 (Columbia University Health Sciences), U01HG8684 (Children's Hospital of Philadelphia), U01HG8673 (Northwestern University), U01HG8701 (Vanderbilt University Medical Center serving as the Coordinating Center), U01HG8676 (Partners Healthcare/Broad Institute), and U01HG8664 (Baylor College of Medicine), U01HG006828 (Cincinnati Children's Hospital Medical Center/Boston Children's Hospital), U01HG006830 (Children's Hospital of Philadelphia), U01HG006389 (Essentia Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University), U01HG006382 (Geisinger Clinic), U01HG006375 (Group Health Cooperative/University of Washington), U01HG006379 (Mayo Clinic), U01HG006380 (Icahn School of Medicine at Mount Sinai), U01HG006388 (Northwestern University), U01HG006378 (Vanderbilt University Medical Center), and U01HG006385 (Vanderbilt University Medical Center serving as the Coordinating Center), U01HG004438 (CIDR), U01HG004424 (the Broad Institute)

samples for genomic discovery and genomic medicine research. The network sought to unify the genetic results from 78 Illumina and Affymetrix genotype array batches from 12 contributing medical centers for joint association analysis of 83,717 human participants. In this report, we describe the imputation of eMERGE results and methods to create the unified imputed merged set of genome-wide variant genotype data. We imputed the data using the Michigan Imputation Server, which provides a missing single-nucleotide variant genotype imputation service using the minimac3 imputation algorithm with the Haplotype Reference Consortium genotype reference set. We describe the quality control and filtering steps used in the generation of this data set and suggest generalizable quality thresholds for imputation and phenotype association studies. To test the merged imputed genotype set, we replicated a previously reported chromosome 6 *HLA-B* herpes zoster (shingles) association and discovered a novel zoster-associated loci in an epigenetic binding site near the terminus of chromosome 3 (3p29).

KEYWORDS
electronic medical records, genotypes, GWAS, herpes zoster, variants

1 | INTRODUCTION

The Electronic Medical Records and Genomics (eMERGE) network has been developing a unified genome-wide single-nucleotide variant (SNV) genotype array-based association platform for analysis of electronic medical record (EMR)-derived phenotypes for approximately 10 years (Chisholm, 2013; Crawford et al., 2014; Gottesman et al., 2013). The eMERGE network has successfully continued efforts to advance the growth and discovery results of this genotype array and clinical phenotype resource. In the first funding phase, eMERGE 1, discovery efforts were based on the Illumina 660k genotype array with ~20,000 participants being enrolled through five medical centers. Using this common array allowed for a unified analysis in eMERGE 1 without the need for imputation to harmonize datasets. This homogeneous quality controlled variant set (Turner et al., 2011; Zuvich et al., 2011) with EMR phenotype algorithms was used to discover novel genomics findings and showed replication of many prior associations, including diabetes (Ng et al., 2014) and blood cell traits (Crosslin et al., 2012, 2013; Kullo et al., 2011). In eMERGE 2 ~30,000 more individuals with high-density genotype data were ascertained resulting in analyses with ~50,000 individuals. The genotype array platforms and variant selections genotyped were not all the same in eMERGE 2, making merged analyses across all variants not possible without the loss and/or addition of information on the ungenotyped variants not shared by all Illumina and Affymetrix arrays. Progress in the development of statistical imputation software (Browning & Browning, 2009; Howie, Donnelly, & Marchini, 2009) for calculating missing genotypes at ungenotyped variants, and the development of population-based reference haplotype panels (1000 Genomes; Howie, Marchini, & Stephens, 2011) by this time (~2011), made imputation-based unification of the eMERGE 2 genotyped samples possible (S. S. Verma et al., 2014). This enabled discovery of genetic associations with EMR-derived disease phenotypes for the eMERGE network, including cataracts (Hall et al., 2015; Ritchie et al., 2014), glaucoma (S. S. Verma et al., 2016), serum thyroid stimulating hormone levels (Malinowski et al., 2014), angiotensin-converting enzyme inhibitor-associated cough (Mosley et al., 2016), liver function tests (Namjou et al., 2015), cardiovascular (Dumitrescu et al., 2017), obesity (Cronin et al., 2014; De et al., 2015), and lipid traits (De et al., 2017; Rasmussen-Torvik et al., 2012) as well as phenotype-wide association study (PheWAS)/ International classification of diseases (ICD-9) code associations (A. Verma et al., 2016). Additionally, detailed population-based analysis of a common genetic condition, hemochromatosis (Gallego et al., 2015), became possible with the increased sample size in eMERGE 2. This catalog of associations demonstrates the potential of discovering novel associations and their genomic risk in the eMERGE network.

In eMERGE 3, genotype and clinical EMR data of ~33,000 additional participants have been added to the resources available for analysis. Statistical methods for imputation have progressed, and improved reference panels used to infer missing variants have become larger and more cosmopolitan with global ancestries represented in the Haplotype Reference Consortium (HRC1.1), a genotype reference set of ~30,000 individuals with 64,976 haplotypes and ~40 million genetic SNV marker alleles (McCarthy et al., 2016). The HRC1.1 contains the 1000 Genomes globally diverse sample set as well as a large balance of European samples. The Michigan Imputation Server (MIS) uses the HRC1.1 reference haplotypes' linkage disequilibrium (LD) to impute missing SNV genotypes with the minimac3 algorithm (Das et al., 2016; Loh et al., 2016). These imputed and molecularly genotyped sets can then be combined across smaller studies with different genetic variant site selection ascertainment than in the original genotyping chip array platforms, making larger genome-wide association studies possible on the sample-variant level. Here, we describe methods to develop a merged imputed data set for the network, including quality control steps and we validate the data by recapitulating the previous herpes zoster (shingles) association to *HLA-B* (Crosslin et al., 2015) using the EMR ICD-9 phenotype at this larger sample size with discovery potential.

2 | SUBJECTS, MATERIALS AND METHODS

2.1 | Participating medical center's source plink genotype batch bfiles

In the creation of the eMERGE 3 merged multisample genotype set, the imputation of participants' missing variants was performed in 78 batches provided as PLINK bfiles from the 12 contributing medical centers in eMERGE 1, 2, and 3. The institutional review board of each contributing institution approved eMERGE study enrollment and informed consent was obtained from participants.

2.2 | Imputation pipeline

To implement the minimac3 missing genotype variant imputation statistical model, we followed the MIS guidelines (Das et al., 2016; Loh et al., 2016; McCarthy et al., 2016) and imputed each genotype array batch independently. The MIS used the HRC1.1 variation reference in genome build 37 (hg19, hs37d5.fa) coordinates to impute the missing variants across samples in genotype array batches of up to 15,000 samples. The starting genotype array batch data files were provided in genome reference build 37 positions and ACGT allele coding by the contributing medical centers. We have included a pipeline flowchart as Figure S1 in the Supporting Information Materials.

2.3 | Preprocessing

In general, Unix automation by bash, Perl, python, and R scripts prepared data for imputation, sample inventory, merging, and analysis to produce results in a parallel computing environment. In batch processing scripts, the software plink (Chang et al., 2015; Purcell et al., 2007), vcftools, bcftools, checkVCF.py, and William Wrayner's HRC1.1 variant pruning tool were used to format data. (See the Section 8 for software uniform resource locators.) These tools use an allele frequency based format conversion method to fix strand problems that exist on and between the many different genotyping platforms to the HRC reference forward-strand format. Briefly, we assessed the participant missingness at all molecularly genotyped variant sites, then genotype variant site missingness of the participants using plink 1.9, both at an exclusion threshold of >2% in a genotype array batch. We then generated a plink frequency file which is the input along with the plink bfile to Wrayner's variant pruning tool (HRC-1000 G-check-bim.pl). This made a shell script named Run-plink.sh, which we ran to prune each batch. We then split the pruned plink bfiles into chromosome 1–22 files and recoded them as variant call format (VCF) files using plink. With these files, we then used the bcftools + fixref function to ensure the reference allele VCF column is fixed to the reference genome. The checkVCF.py script was used to perform a double-check of the variants to assess whether the data are properly formatted. We also ran vcf-sort and bgzipped the final VCF to have the prepared files for imputation.

2.4 | Imputation postprocessing

All samples in the 78 batches of bfiles which passed initial missingness QC were uploaded to the MIS, imputed, and the results downloaded. If samples were duplicated on different array batches, they were analyzed in duplicate or more in each of these batches for the imputation results. In subsequent merging steps, we selected the unique sample from the array batch with the most variants genotyped using a custom Perl sample inventory script that batches vcftools command-line calls into shell scripts run on a computing cluster. The finished MIS imputation data are provided by chromosome in VCF format with an underlying allele-dose model in a diploid 0–2 continuous variable range and the statistically phased hard genotype calls (0|0, 0|1, 1|0, 1|1) with a posterior probability of >50%. The accompanying *.info files contain the *R*-squared quality correlation for each variant between the true genotype and the imputed value at each variant.

2.5 | Variant frequency analysis

The minor allele frequency (MAF) of the merged imputed variant genotypes was calculated using vcftools (Danecek et al., 2011). For summary purposes, we counted a variant as molecularly genotyped if it was present in one or more arrays after imputation preprocessing and imputed only if it was imputed in all genotype array batches. To produce the site frequency spectrum and counts of imputed and molecularly genotyped variants, we used a Perl script that interprets the vcftools frequency output to bin counts and formatted tables for plotting with an R script. If variants had more alleles than a biallelic variant we only considered the more common minor allele.

3 | POPULATION AND ANCESTRY SUBSTRUCTURE

3.1 | Principal component analysis (PCA)

The hard genotype call merged sample set of unique imputed samples was analyzed by PCA using the plink2–pca approx fast pca method for large sample sizes and participant groupings were compared to the self-reported or observed-reported ancestry. We performed PCA on the 83,717 participant multisample with variants MAF > 5%, variant missingness of 0.1, and LD-pruned to an *R*-squared threshold of 0.7. We then used the R *k*-means() function with three groups and PCs 1 and 2 to define the canonical Asian, European, and African ancestry groups (Lee, Abdool, & Huang, 2009; Solovieff et al., 2010). We then excluded the Hispanic, Native American, and other ancestry groups to refine the canonical European, Asian, and African ancestry groups by requiring the observed/self-reported ancestry to match the PCA-based *k*-means group. Within these ancestry groups, we calculated PCAs with the same SNV pruning reapplied.

3.2 | Identity by descent (IBD)

The plink2–genome function for IBD calculation was performed with the same SNV pruning and preprocessing as the PCA to generate ~3.5 billion pairwise comparisons. Z0 indicates the probability that zero alleles are shared IBD and Z1 represents the probability that one allele is shared IBD. We defined families using sample pairs with ($Z0 < 0.83$ and $Z1 > 0.1$) and then included everyone who has a shared pair with another sample to include founders and distantly related branches into the same family.

3.3 | *R*-squared imputation quality

The quality of missing genotype imputation in the 78 genotype array batches was assessed using the *R*-square imputation quality metric in the chr*.info files provided by the MIS (Das et al., 2016; Loh et al., 2016; McCarthy et al., 2016). This value (0–1) is an estimate of the squared correlation between the unobserved genotypes and the imputed genotypes. Since the true genotypes are not available at imputed sites, the minimac3 algorithm makes the assumption that the population frequency affects the quality of imputed genotypes counts by shrinking them to their expectation derived from this frequency.

3.4 | Regression of batch imputation quality

We calculated the mean of the ~40 million variants *R*-square imputation qualities for each of the 78 array batches. We used this array batch mean *R*-square value as the dependent variable in a linear regression to determine the influence of the input batch sample count and batch molecular genotype count on imputation quality. The linear regression took the following form:

Batch mean *R*-square imputation quality ~ log(batch genotype count) + log(batch sample count).

3.5 | Variant genotype average qualities

We calculated the mean, median, maximum, and minimum of each variants imputation quality *R*-square across the 78 batches for the ~40 million variants. A histogram of the variants' *R*-squared means was used to summarize the data for viewing the distribution and assist in deciding the quality threshold for inclusion for subsequent genome-wide association study (GWAS). We also analyzed the imputation *R*-square quality metric by the frequency bins used in the frequency analysis. We plotted boxplots of the distribution of the means of the batches *R*-squared for each variant frequency bin and observed the expected decay in *R*-square imputation quality from common to rare variant bins. Additionally, an empirical *R*-squared measure was calculated for the variants on the various genotype array chips by imputing while leaving the know genotype out, then comparing the imputed values to the known values. We report the percentage of empirical *R*-squared values greater than 0.8.

3.6 | Heterozygosity–homozygosity analysis

We used vcftools (Danecek et al., 2011) to calculate the inbreeding coefficient, F as a measure of homozygosity (1-heterozygosity), for chromosomes 1–22. A method of moments calculation is used to estimate *F* for each individual. Higher values indicate less heterozygosity and more homozygosity. We then compared *F* to the batch mean *R*-square imputation quality and *k*-means ancestry groups determined by the PCA.

4 | GWAS OF THE HERPES ZOSTER ICD-9 PHENOTYPE

Given this eMERGE 3 unified imputed genotype set, we sought to validate the data by recapitulating the previous eMERGE 2 association to the *HLA-B* gene region variation with herpes zoster (shingles; Crosslin et al., 2015) with the larger sample size offered by the eMERGE 3 data. The previous herpes zoster phenotype GWAS participants are contained as a subset of the included samples making this analysis a validation, with *HLA-B* as an expected positive control. The eMERGE 3 data have approximately twice the sample size of the previously reported eMERGE 2 data. The zoster phenotype has been defined here by the ICD-9-derived PheWAS codes (Denny et al., 2010, 2013) from the eMERGE network to identify case and control status of the individual participants. We selected the Phecode 053 from the eMERGE records, requiring at least one observation of the code in cases. We removed participants from the three pediatric sites, sites which had fewer than 50 cases identified and genotype imputation batches where the mean imputation quality *R*-square was <0.3. We removed anomalous IBD clusters and suspected twins which may be duplicated sample errors. We retained one individual from each family, a case if available. Regression was performed with the PLINK 1.9 software (Chang et al., 2015) in additive mode on the hard-called genotypes with a posterior probability of >50%. We applied covariate adjustment for the 12 medical centers and gender as well as PCs 1, 2, and 3 as continuous variables to control for major population stratification (Crosslin et al., 2014).

We first analyzed all samples jointly. Next, we stratified individuals to the major African and European continental ancestral groups as defined by the intersection of the PCA-based *k*-means clusters (Lee et al., 2009; Solovieff et al., 2010) and the observed/self-reported ancestry. We used the PCs 1, 2, and 3 of the secondary (ancestry group specific) PCAs as covariates in the stratified regression models. The PCAs control for both population structure and platform bias (Crosslin et al., 2014). We relaxed the medical center site minimum case count to 15 (from 50 in the joint analysis) for medical center participant exclusions in the ancestry-stratified regressions. We also repicked the cases and controls from families within the ancestry subgroups.

For SNVs in common variant regressions, we required MAF > 0.05 and *R*-squared imputation quality ≥ 0.3 inclusion thresholds to achieve a high confidence set of results. The *R*-squared imputation quality threshold we applied in two ways, initially we excluded imputation batches which had a mean *R*-square of <0.3 across the ~40 million variants. We also excluded regressions with imputed variant genotypes which had a mean *R*-square of <0.3 across the 78 genotype array batches. Using the variants which meet these association and quality parameters, we plotted the Manhattan and Quantile–Quantile plots using the GWASTools (Gogarten et al., 2012) R package and observed the genomic control lambda (λ; Devlin & Roeder, 1999). λ has an expected value close to one in a valid genome-wide association. To declare genome-wide significant variants, we used the generally accepted p-value threshold of <5 × 10e–8. We inspected all associated loci using the University of California, Santa Cruz (UCSC) genome browser (Hinrichs et al., 2006) and the NCBI bioinformatics online tools to summarize the gene and nucleotide elements present.

We also inspected the genomic inflation control of λ by running the regression models with a variant MAF threshold of ≥0.001 and at all variants that converge without a per variant *R*-square quality or MAF inclusion threshold. This allowed us to observe how imputation uncertainty, quality, and variant frequency may affect the association statistics and genomic inflation. We report these lower quality lambdas, low frequency, and low-quality associations in the supplemental materials.

TABLE 1 Number of unique participant eMERGE IDs and reported demographics

		Arrays	Gender	Gender	African/	American			Pacific	Hispanic/	
Medical center	**Participants**	**Batches**	**Male**	**Female**	**Black**	**Indian**	**Asian**	**White**	**Islander**	**Latino**	**Unknown**
Boston Children's	1,019	1	596	423	66	2	21	676	0	125	129
CCHMC	5,717	12	3,262	2,455	601	5	67	4,673	5	143	223
CHOP	10,465	21	5,630	4,835	4,666	7	161	4,890	3	321	417
Columbia	2,065	2	1,058	1,007	179	6	77	619	2	448	734
Geisinger	3,111	1	1,638	1,473	9	2	0	3,085	0	13	2
Harvard	10,095	3	4,626	5,469	509	0	172	8,579	0	474	361
Kaiser/GHC/UW	3,316	3	1,428	1,888	109	12	89	2,922	6	69	109
Marshfield Clinic	4,756	5	1,878	2,878	2	3	12	4,690	0	14	35
Mayo Clinic	10,256	16	5,193	5,063	23	18	21	8,810	0	1,043	341
Mt. Sinai	6,255	4	2,555	3,700	4,046	33	3	679	0	1,297	197
Northwestern	4,848	2	817	4,031	598	0	0	4,207	0	36	7
Vanderbilt	21,814	10	9,868	11,946	3,854	16	102	17,313	0	211	318
Total	83,717		38,549	45,168	14,662	104	725	61,143	16	4,194	2,873

Note. eMERGE: Electronic Medical Records and Genomics.

5 | RESULTS

After removal of low-call-rate samples (>2% missingness) and duplicated samples, the data set resulted in 83,717 unique imputed participants based on the eMERGE subject IDs from 77 imputation batches. One batch was duplicated completely by samples on other arrays. The medical centers and batch counts for genotypic data, gender, and self-reported or observed ancestry is shown in Table 1. The sampled participants were imputed collectively for 39,127,678 variants genome wide from 5,166,481 variants molecularly genotyped on one or more arrays. Only 13 variants were shared by all 78 arrays in the data input to the MIS, while 181,713 variants are shared among any 50 arrays. Supporting Information Tables S1 to S3 show the participant sample counts by batch array bfile name, medical center, and eMERGE 1, 2, and 3 project development cycles.

5.1 | Variant frequency analysis

Among the 5,166,481 molecularly genotyped variants, 404 were monomorphic in the merged imputed genotype set. Similarly, 668,157 monomorphic variants and 33,293,040 polymorphic variants were imputed across all genotype array batches. In the site frequency spectrum (see Supporting Information Figure S2), 2,312,962 molecularly genotyped variants have MAF > 0.05 and comprise ~45% of the genotyped variants with a MAF > 0.05. There are 3,447,308 imputed common variants with MAF > 0.05. Collectively, imputation makes 5,760,270 common variants available for subsequent analysis. There is a large drop-off in the number of variants both molecularly observed and imputed between the 0.0001 and 0.00001 genotype frequency bins. The frequency bins 0.01 to 0.00001 represent the bulk (27,119,930 variants) of the imputed genotype spectrum. See Supporting Information Table S4 for counts of imputed and genotyped variants by MAF bin.

6 | POPULATION AND ANCESTRY SUBSTRUCTURE

6.1 | PCA

A total of 1,003,235 autosomal biallelic SNVs with a MAF > 0.05 after LD pruning were used to create the participant by genotype correlation matrix for PCA. The

PCA results were used to quality control the data for genetic versus self-reported ancestry and provide genomic control of *p*-value inflation in regression analysis. The scree of variance explained by eigenvectors and principal components is plotted in Figure 1a,b. PC 1 explained 88.3% of the imputed genotype variance and represented the African to European ancestry cline. PC 2 showed 6.1% of the genotype variance while representing the Asian to European ancestry cline. PCA ancestry analysis showed concordance with self-reported race and displayed the canonical European, African, and Asian ancestry sample groupings without evidence of gross batch effects in the first two principal components (Figure 1). The participants were classified to 15,980 African, 2,604 Asian, and 65,133 European subjects, using *k*-means with three groups on PCs 1 and 2 (see Supporting Information Figure S3). When the observed/self-reported categories not matching the continental *k*-means ancestry groups were removed, we had an intersection of 14,380 African, 646 Asian, and 60,747 European participants.

6.2 | Ancestry-specific PCA

We computed three additional PCAs, one for each ancestry subgroup from the *k*-means of the joint PCA while requiring the intersected observed/self-reported ancestry to match to obtain refined ancestry-specific PCs for stratified analyses of the main ancestral types. After the MAF filtering and LD pruning the African samples had 1,889,435 variants, Asians had 776,084 variants, and Europeans had 741,856 variants for PCA. In Figure 1c, we plot the African subgroup PCA where we see two self-identified African specific clusters separated by PC 2 explaining 7.3% of the variance, while PC 1 explains 56.6% of the variance. In Figure 1d we plot the Asian subgroup PCA where we see four main clusters, three of these are mainly separated by PC 1. The fourth cluster appears as a long cline defined by PC 2 on the right side of PC 1. The Asian PC 1 explains 56.2% of the variance while PC 2 explains 8.3%. In Figure 1e, we plot the European subgroup PCA where we see a triangle shape with no separation in clusters. The European PC 1 explains 43.4% of the variance and PC 2 explains 9.9% of the variance.

6.3 | IBD

The plot of 3,504,226,187 pairwise sample IBD comparisons is presented in Figure 2. Upon inspection of the IBD plot, it was noted that there were 114 pairs (228 individuals) close to the origin that are either twins or sample duplicates with different eMERGE IDs. We also noted a grouping of 4,211 pairs (139 individuals) with noncanonical IBD plot positioning in the region of Z0 between 0.4 and 0.65 with Z1 < 0.4. Inspection of this IBD cluster determined that the majority of these samples (108/139) are from a single contributing center (Marshfield) and represent a family study. We recommend to exclude these related samples in association analyses that do not take into account a relatedness metric. Initially, we defined a family as those sample pairs falling in the IBD metrics Z0 < 0.83 and Z1 > 0.1 with any pair connecting a family network to include the samples with unrelated founders in the same family. We additionally included any sample with Z0 < 0.63 to capture the twins/sample duplicates near the origin into

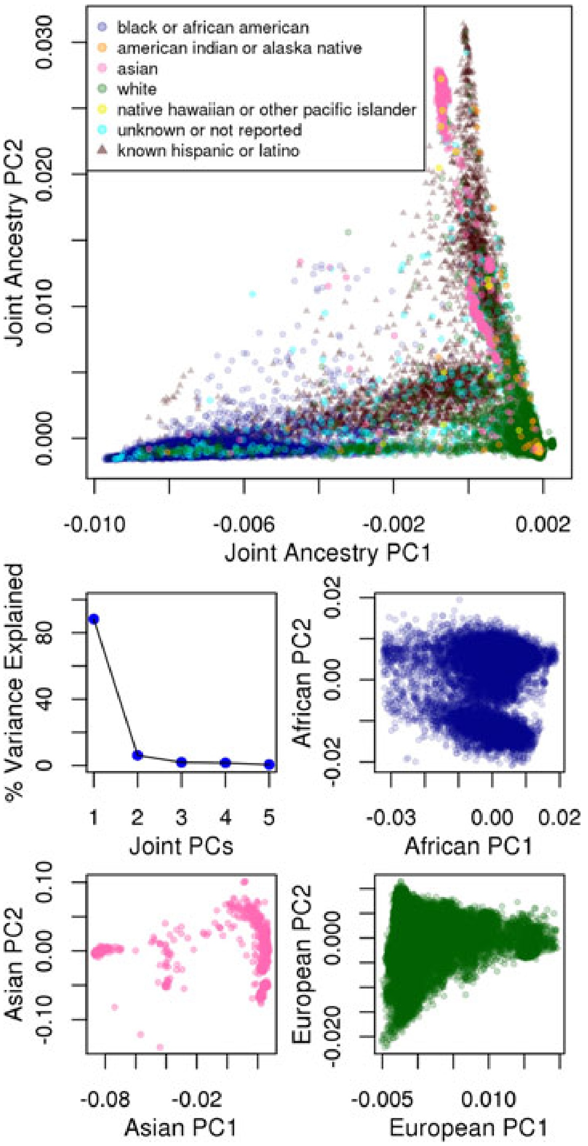

FIGURE 1 PCA and Screen plot using MAF 5%, LD-pruned *R*-square of 0.7 and missingness of 10% by joint ancestry, and stratified by African, Asian, and European ancestry PCAs defined by the intersection of the *k*-means and observed/self-reported race. LD: linkage disequilibrium; MAF: minor allele frequency; PCA: principal component analysis

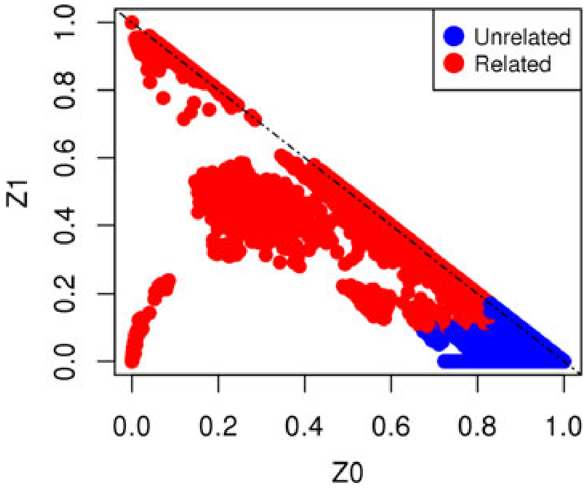

FIGURE 2 Z0 Z1 identity by descent plot of eMERGE 3 imputation ($n = 3{,}504{,}226{,}187$ pairwise comparisons). eMERGE: Electronic Medical Records and Genomics

families. Using this, we identified 13,152 individuals that grouped into 4,529 families.

6.4 | The *R*-squared metric of imputation quality

The mean *R*-square of the array batches has a range of 0.089–0.68 across the entire variant frequency spectrum. A linear model using the variables, the count of directly genotyped variants (See Figure 3a; range 50,911–3,410,557) and batch sample count (see Figure 3b; range 4–9,315) showed the log of the batches sample counts ($p < 2 \times 10e{-}16$; 62% of variance) and the log of the batches directly genotyped variant counts as input to the imputation algorithm (p~$5.6 \times 10e{-}6$; 21% of variance) were both significantly associated with the batch mean *R*-square imputation quality. Imputation quality is very poor (*R*-square < 0.3) with small batch sizes, so we elected to exclude the 27 batches with a mean *R*-square of <0.3, leaving 50 batches for GWAS analysis. Supporting Information Table S5 provides a summary of the batches, sample counts, and genotype counts.

The histogram of the variants' mean *R*-squares in Figure 3c shows a trimodal distribution with two minima and three maxima. Two of the maxima are at the ends of the distribution close to zero and one with a middle maximum at ~0.45 with the minima on either side. The right side lower minimum is at ~0.8 and the left side higher minimum shoulder is at ~0.3. We chose the left minimum for variants with an imputation quality *R*-square of >0.3 to be the decision threshold for inclusion in subsequent GWAS analysis summaries. There are 21,924,838 variants (56%) with a mean *R*-square imputation quality >0.3, and 17,294,872 variants (44%) with a median *R*-square imputation quality of >0.3. Among the variants with genotypes on various arrays where the empirical *R*-square was calculated, 94% had an empirical *R*-square of >0.8.

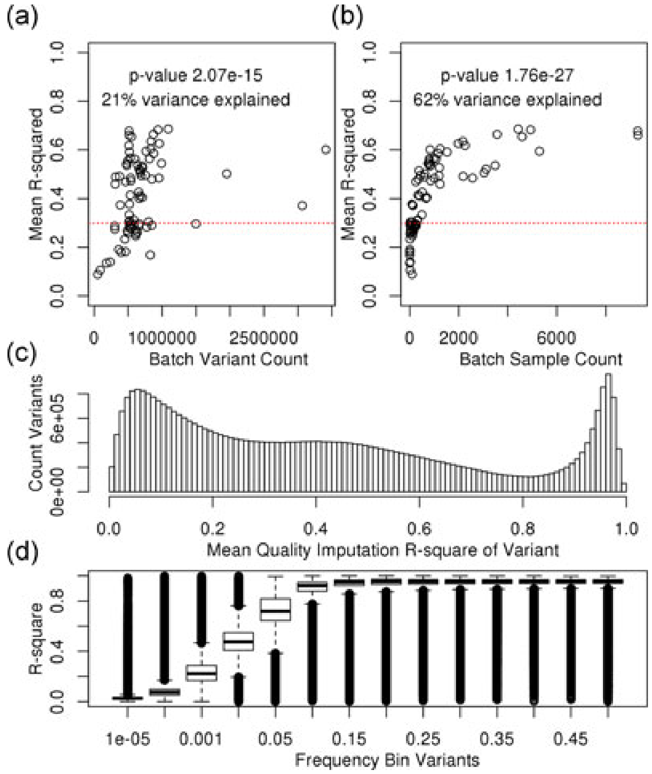

FIGURE 3 Plots of genotype array batch mean *R*-square imputation quality regression variables of samples size (a) and variant count (b). Histogram (c) of each variants mean *R*-square imputation quality across imputation batches. Boxplots (d) of variants mean *R*-square imputation quality by frequency bins

We also analyzed the *R*-squared imputation quality metric by the MAF bins as shown in the site frequency spectrum analysis. The median *R*-squared quality values were all >0.9 down to the 0.05–0.1 MAF bin. See Figure 3d for boxplot distributions of *R*-square values of each MAF bin. The median *R*-square imputation quality value then begins to drop and passed 0.3 in the 0.00001–0.0001 MAF bin. Based on the distribution of variability in the imputation quality *R*-squares in these MAF bins, the common variants generally impute very well while rare variants have less certainty in their imputation.

6.5 | Homozygosity

We compared the chromosome 1–22 inbreeding coefficient (*F*) homozygosity to the batch mean *R*-square imputation quality and *k*-means ancestry groups determined by the PCA (see Figure 4, top and bottom panels

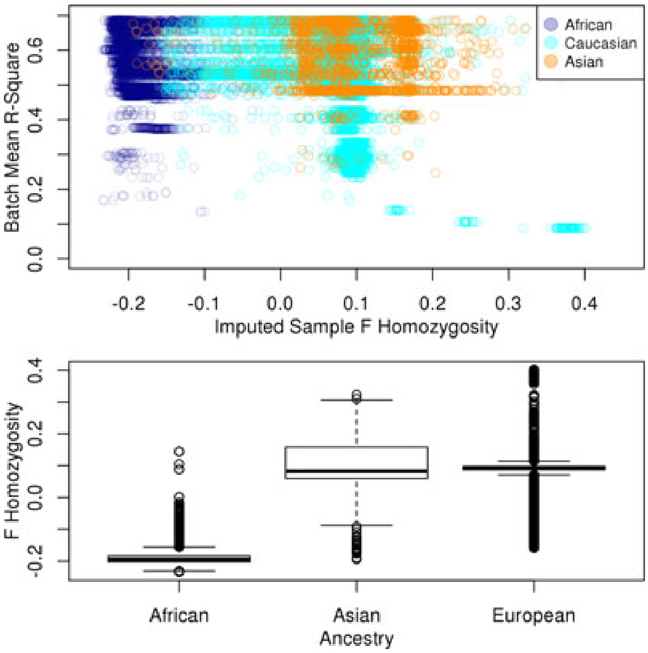

FIGURE 4 The chromosomes 1–22 inbreeding coefficient F, as a measure of homozygosity plotted versus the batch mean R-square imputation quality in the top panel and k-means Principal component analysis ancestries in the top and bottom panels

respectively). African ancestry participants in general have inbreeding coefficients centered around −0.2 with the Asian and European ancestries centered on ~0.1. Batches with mean R-square imputation qualities down to ~0.2 were able to resolve African ancestry individuals by F. Three European only batches with very low imputation quality (R-square < 0.2) began to skew their inbreeding coefficients to higher values (~0.15–0.4; 98th to 100th percentiles of F values) suggesting the imputation is under calling these samples' batches making them appear as more homozygous.

6.6 | Imputed data set validation by GWAS of previous eMERGE association

eMERGE has previously identified *HLA-B* variants on chr6p21 as associated with herpes zoster in a case–control analysis of 22,981 participants (2,280 cases; Crosslin et al., 2015). In the current analysis, 27 genotype batches had imputation mean R-squared values of <0.3 thus, we excluded the 2,234 participants in these batches from GWAS. We removed the anomalous IBD cluster of participants from the Marshfield Clinic, the suspected twins and left one individual (a case if available) from each family. See Table 2 for a count of the included zoster cases and controls by contributing medical center. A total of 46,350 participants (3,763 cases) from nine medical centers were selected based on inclusion criteria and availability of an ICD-9 record.

Logistic regression results are presented in Table 3 for the 28 common variants which reach genome-wide significance ($p < 5 \times 10e{-}8$). We included variants with MAF ≥ 0.05 and mean imputation quality R-squares of ≥0.3 for a total of 5,752,274 common variants that converge in the regressions. The GWAS λ statistic is ~1.02 in this high confidence set. We observe two peaks in the Manhattan plot (Figure 5), one at chr3p29, near the terminus of chromosome 3, and the second on chromosome 6 in the *HLA-B* locus, which replicates the previous GWAS (Crosslin et al., 2015).

The 12 zoster-associated variants on chromosome 3 are in or closely adjacent to an annotated brain tissue histone H3K4me3 peak (Maunakea et al., 2010), which also is in the footprint of transcription factors family of transcription (*FOS*), signal transducer and activator of transcription 3 (*STAT3*), and transcription factor 7-like 2 (*TCF7L2*) Chip-seq DNA precipitation peaks and also is a DNAase I hypersensitivity site that all occupy the same genomic bases near the end of the chromosome (3p29; Hinrichs et al., 2006). The closest

TABLE 2 Counts of herpes zoster cases and controls included in the final zoster regression model

Site	Cases	Controls	Male	Female	African	Asian	European
KPUW	641	1,625	984	1,282	94	78	2,094
MRSH	758	1,244	851	1,151	2	13	1,987
COLU	51	1,511	817	745	368	178	1,016
GEIS	285	1,983	1,232	1,036	6	1	2,261
NWUN	178	3,415	588	3,005	399	8	3,186
HARV	461	7,063	3,539	3,985	405	205	6,914
MTSI	286	3,648	1,507	2,427	2,846	73	1,015
MAYO	348	6,128	3,426	3,050	19	36	6,421
VAND	755	15,970	7,679	9,046	2,692	164	13,869
Total 46,350	3,763	42,587	20,623	25,727	6,831	756	38,763

Note. Ancestry is reported based on the k-means three clusters of PC 1 and PC 2.

TABLE 3 28 common variants that reach genome-wide significance

rsID	Rsq	G/I	M/MA	MAF	chr	position	p	OR	gene
rs9810195	0.92	8/70	G/A	0.13	3	192746326	2.805e-09	1.26	TFBS
rs9848218	0.92	0/78	T/C	0.13	3	192746451	3.438e-09	1.26	TFBS
rs6784731	0.92	0/78	C/G	0.13	3	192746746	3.538e-09	1.26	TFBS
rs6784850	0.92	0/78	A/G	0.13	3	192746850	3.604e-09	1.26	TFBS
rs1039219	0.91	0/78	G/A	0.12	3	192747249	2.966e-08	1.24	TFBS
rs1039220	0.94	0/78	T/C	0.14	3	192747381	1.096e-08	1.23	TFBS
rs11916599	0.92	0/78	G/A	0.13	3	192747851	6.238e-09	1.25	TFBS
rs11924420	0.94	4/74	T/C	0.15	3	192748011	2.09e-08	1.23	TFBS
rs4371461	0.94	0/78	T/C	0.15	3	192748673	2.911e-08	1.23	TFBS
rs7428308	0.94	0/78	G/A	0.14	3	192749047	3.481e-08	1.22	TFBS
rs73071839	0.94	0/78	G/A	0.15	3	192749140	2.387e-08	1.23	TFBS
rs112062423	0.94	0/78	A/C	0.15	3	192749169	2.367e-08	1.23	TFBS
rs2844584	0.86	0/78	G/A	0.09	6	31321524	2.043e-08	0.772	HLA-B intron
rs2769	0.88	3/75	A/G	0.12	6	31321882	3.027e-10	0.772	HLA-B 3 prime UTR
rs1093	0.88	3/75	G/A	0.17	6	31321906	1.541e-09	0.884	HLA-B 3 prime UTR
rs17199328	0.89	0/78	G/A	0.12	6	31322395	2.284e-10	0.771	HLA-B intron, missense
rs2854001	0.88	0/78	A/G	0.17	6	31323012	8.484e-09	0.819	HLA-B intron
rs1050723	0.88	0/78	A/G	0.12	6	31323321	2.528e-10	0.77	HLA-B missense
rs9266266	0.89	12/66	T/C	0.15	6	31326011	6.869e-09	0.807	HLA-B upstream
rs9266269	0.89	0/78	A/G	0.15	6	31326055	4.808e-09	0.805	HLA-B upstream
rs9266270	0.89	0/78	A/G	0.15	6	31326072	6.724e-09	0.806	HLA-B upstream
rs116583816	0.89	0/78	C/G	0.13	6	31326123	2.36e-09	0.785	HLA-B upstream, TFBS
rs2523591	0.90	34/44	A/G	0.42	6	31326960	9.609e-09	0.863	HLA-B upstream
rs2523586	0.90	27/51	T/G	0.23	6	31327435	1.461e-08	0.839	HLA-B upstream
rs2596477	0.89	20/58	A/G	0.13	6	31327723	3.531e-10	0.774	HLA-B upstream
rs2523577	0.89	0/78	G/A	0.13	6	31328739	4.399e-10	0.776	HLA-B upstream
rs9266853	0.86	0/78	C/G	0.08	6	31387725	3.718e-08	0.754	HPC5, HLA-B upstream
rs3893526	0.88	0/78	A/G	0.08	6	31413742	2.102e-08	0.751	HPC5, LINC01149

Note. G/I: genotyped/imputed batch count; M/MA: major/minor allele; MAF: minor allele frequency; OR: odds-ratio; Rsq: Imputation quality *R*-square mean; TFBS: transcription factor binding site.

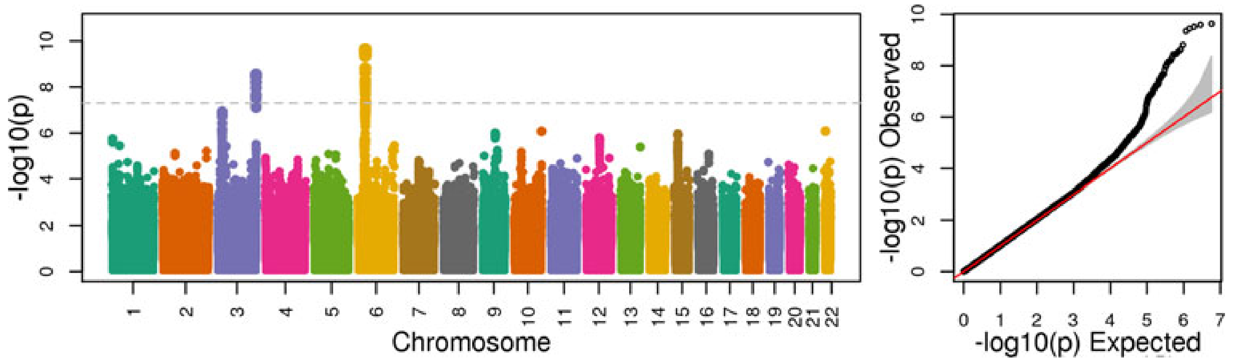

FIGURE 5 Joint ancestry GWAS Manhattan (left panel) and quantile–quantile plots (right panel) of herpes zoster (shingles) with 3,763 cases and 42,587 controls. Variant inclusion stringency is set to *R*-square of ≥0.3 and minor allele frequency of ≥0.05. Covariate adjustments were made for PCs 1, 2, and 3, gender and the nine contributing medical centers which were included. Genomic control is close to one with a λ of ~1.02

genes to these transcription factor binding site variants are *MB21D2* towards the centromere and *HRASLS* towards the terminus, both approximately 100 kb away (see Figure 6). We could find no prior associations with herpes zoster to this locus.

Sixteen significantly zoster-associated variants are annotated to be in the *HLA-B* gene region on chromosome 6 (see Figure 7). *HLA-B* as previously noted (Crosslin et al., 2015) likely plays a role in viral suppression. Variants in this region have also been associated with delayed development of AIDS with HIV-infected individuals (Fellay et al., 2009). Similarly, we also see a protective effect of the associated variants' minor alleles with odds ratios of ~0.77. Table 3 provides summary descriptions of the variants. All these variants display similar association statistics, MAF and are in LD with each other suggesting they share a common haplotype.

6.7 | Common variant African and European stratified ancestry association statistics

In the ancestry-stratified regression models five medical centers had African ancestry participants (398 cases and 5,922 controls) and eight medical centers had European ancestry participants (3,201 cases and 34,960 controls) which are summarized in Supporting Information Table S6. The ancestry-stratified regressions showed no statistically significant common variants in African ancestry ($\lambda = \sim 1.007$; Supporting Information Figure S4) and 46 genome-wide significant common variants across the *HLA-B* locus in Europeans ($\lambda = \sim 0.999$; Supporting Information Figure S4 and Table S7). These 46 variants include 14 variants also identified in the joint ancestry analysis. The chromosome 3p29 locus variants in the ancestry-stratified analysis did not reach genome-wide significance, but did display the same variants with low *p*-values in European (*p*, ~0.001–0.0003; MAF, ~0.08–0.1) and African (*p*, ~0.0006–0.06; MAF, ~0.3) ancestral groups. Similarly the African ancestry-stratified analysis displayed low *p* values (~0.001–0.1) with similar odds ratios (~0.77) for many of the same variants in the *HLA-B* region. The minor allele frequencies of the protective alleles in the African stratified analysis are ~0.06–0.08, slightly less than observed in the joint analysis (~0.13 MAF). We are likely not powered to detect the *HLA-B* association at genome-wide significance in the African stratified analyses due to sample size.

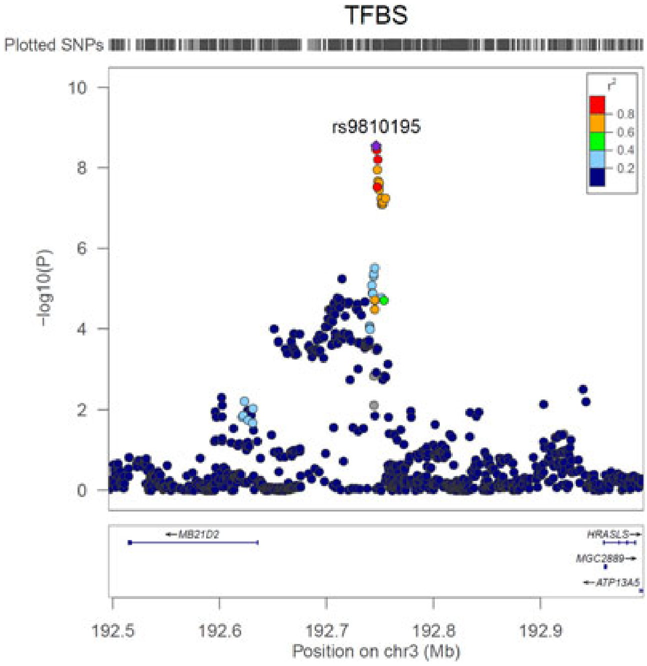

FIGURE 6 Chromosome 3p29 site LocusZoom plot of zoster association in the joint ancestry regressions. SNP: single-nucleotide polymorphism; TFBS: transcription factor binding site

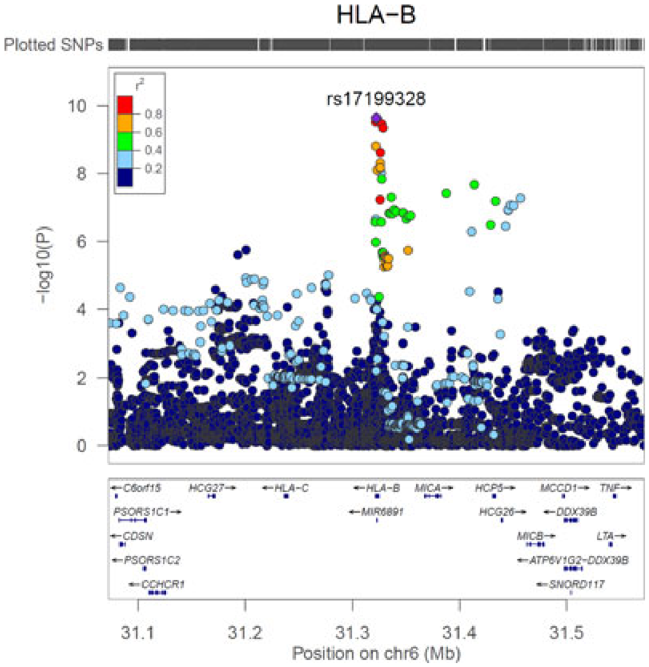

FIGURE 7 Chromosome 6 *HLA-B* LocusZoom plot of zoster association in the joint ancestry regressions. SNP: single-nucleotide polymorphism

7 | DISCUSSION

The metrics of variant *R*-square imputation quality (>0.3), batch mean of variants *R*-square imputation quality (>0.3), frequency thresholds, PCAs, and IBD will assist in defining best practices for performing association analyses with the rich eMERGE phenotype data derived from electronic medical records. The diverse genomic array datasets unified here show that sharing resources with careful data management can work for genomic association replication and discovery with EMR phenotypes as cohorts grow in size.

Imputation with the HRC1.1 is able to discriminate European, Asian, and African ancestry samples as seen in the PCA and homo/heterozygosity results, even with the large contingent of European HRC1.1 imputation reference samples. Additionally one can see the diversity of the Hispanic/Latinos in the space between and adjacent the African and Asian arms and in the V adjacent to the Europeans (see Figure 1), hinting at the admixture between Native American, European, and African ancestry many Latinos can have. Both the PCA-based genetic ancestry grouping methods of *k*-means and observed/self-reported compare well with the reported ancestry for African and White/European but the Asian observed/self-reported ($n = 725$; see Table 1) is overestimated by the PCA-based ancestry grouping methods (n, ~2,604; see Supporting Information Figure S3) making the intersection of *k*-means and self-reported ancestry important for stratified analyses. This is due to the inclusion of many Hispanic/Latino reported samples in the *k*-means groups. We attempted to capture the Hispanics in their own group by including a $k = 4$ with PC 1–3 analysis, but found this did not capture the Hispanic substructure and centered mainly on the European samples with arms extending into the African and Asian groups. Hispanics/Latinos have a complex admixture with different proportions based on local and urban versus rural influences, making clustering a difficult endeavor in a single PCA analysis.

We interpreted the performance of array batches by mean imputation quality by batch with respect to genetic ancestry defined by the PCAs as the proportion of African, Asian, and European samples (see Supporting Information Table S5). The three large African American batches performed well with imputation quality mean *R*-squares of ~0.5. Additionally, we have ~6 larger chip batches with ~50% African Americans that have good mean quality *R*-squares of ~0.5–0.6 near the maximum batch quality values we observe. This suggests that mixed and ancestry unbalanced imputation array batches can perform well with the HRC1.1 imputation reference set given enough genotyped input variants and samples.

Collectively batch sample and variant count sizes had a large effect (~80% of variance) on the mean imputation quality of the variants. With a few exceptions in general, batch sizes above ~500 samples are usable, while mean *R*-square is maximized above ~2,000 samples. In general, the smaller less dense genotyping arrays also did not perform as well as the higher density arrays. Many of the arrays with fewer variants after pruning and strand fixing (<~500 k variants) had lower mean variant *R*-square values of <0.3. These batch size analysis results give estimates of sample and genotype array size for future projects to plan imputation batches. Among the 14 Affymetrix arrays, many performed as well as the Illumina arrays (see Supporting Information Table S5). Homozygosity analysis by the inbreeding coefficient *F* by batch (see Figure 4) showed that there was no inflation in the *F* inbreeding homozygosity statistic until the batch mean *R*-square imputation quality became <~0.2 in the sample batches plotted. This supports our selection of the 0.3 threshold for the mean batch imputation quality. We also explored the lower quality associations and show that including lower frequency variants ($\text{MAF} > 0.001$) obtains similar acceptable lambda values (see Supporting Information Materials Table S8 and Figure S6). The QC and variant selection methods suggested here show a validation of the data using standard GWAS tools. Many parallel projects in eMERGE will utilize this imputed genotype resource and include different subsets of the data appropriate for their respective phenotypes. These projects may choose to recalculate the imputation quality *R*-square means and inclusions based on the subset of samples and arrays they select for their study. This may include assessing the common variants only in the calculation of chip imputation performance to ascertain cases among the samples on the 27 low performing arrays.

The IBD relatedness calculation and relatedness thresholds we used are likely to be over conservative in the exclusions for relatedness within families. This is due to the admixed and multiancestry composition of this cohort. Other methods (i.e., Pcrelate [Conomos, Reiner, Weir, & Thornton, 2016] and KING [Manichaikul et al., 2010]) are reported to be able to robustly distinguish relatedness among these cosmopolitan assemblages of diverse individuals. These methods may need to be explored for studies where the phenotype of interest has more related individuals within eMERGE. Regardless, for the zoster phenotype, eMERGE has accrued the sample size to power this association while avoiding complications due to related individuals and cryptic relatedness using the conservative method used here. The PCA methods we used to control for *p*-value inflation were not fitted with an exterior reference genotype set such as the 1000 Genomes (Howie et al., 2011). Of note, in initial association runs we used all 10 PCs with regressions and obtained an acceptable lambda value (~1.016) with the same *HLA-B* and chr3p29 hits, but we noticed that there were three nonconvergent regions of the genome with three false-positive variants at the borders of these regions. We decided to try regressions with just PCs 1, 2, and 3 since these PCs showed very similar genomic control. We obtained the same major *HLA-B* and chr3p29 hits with a similar acceptable lambda (~1.02). The nonconvergent regions and false positives at their margins disappeared with just PCs 1, 2, and 3 used as covariates. Due to this we chose to report results from just the regressions with PCs 1, 2, and 3, while we view the exploration of the PCs 1–10 in preliminary trial regressions to show that genomic control is sufficient in comparison with just the first three PCs.

The positive control recapitulation of the previous *HLA-B* locus association provides a validation of the genotype data management, cleaning, and imputation process. This *HLA-B* association has also recently been replicated by another large cohort of ~200,000 participants from 23andMe (Tian et al., 2017). The novel epigenetic binding site we discover on chromosome 3p29 has three transcription factors (*FOS*, *STAT3*, and *TCF7L2*) known to bind there and several DNAase hypersensitivity sites flanking this transcription factor binding site and in the span of associated variants. Among these various epigenetic sites in the associated span, ~36+ diverse tissue types are represented including brain, astrocytes, and lymphocytes. Many of these tissues could hypothetically have interaction with zoster and other herpes viruses. These associated variants may be affecting the epigenetic and noncoding genome function in this region. These same transcription factors have also previously been shown to have interactions with viral response and transcription in molecular biology based analyses. HIV infection has been shown to induce activated protein-1 which contains the *FOS* transcription factor subunit (Chirmule et al., 1995). *STAT3* protein products are phosphorylated by receptor-associated kinases stimulated by cytokines and growth factors. *STAT3* proteins induce a cellular antiviral state when they form dimers which translocate to the nucleus and function as transcriptional activators (Danziger, Pupko, Bacharach, & Ehrlich, 2018; Roca Suarez, Van Renne, Baumert, & Lupberger, 2018). The Kaposi sarcoma-associated herpes virus (KSHV) kaposin A and human immunodeficiency virus (HIV-1) Tat protein complex activates the *MEK/ERK*, *STAT3*, and *PI3K/Akt* signals (Chen et al., 2009; Zeng et al., 2007). Evidence shows KSHV also affects chromatin looping through transcription factor binding sites and is positively correlated with viral production (Campbell et al., 2018). There are also *STAT3* interactions with Epstein–Barr virus in B-lymphocytes (Martorelli et al., 2015). *TCF7L2* expression in human B cells is downregulated by the HIV-1 gp120 protein (Jelicic et al., 2013). T-cell Factor 4 (*TCF4*) is a common synonym for *TCF7L2*. *TCF4* can inhibit HIV-1 Tat docking in the HIV-1 LTR (Henderson, Sharma, Monaco, Major, & Al-Harthi, 2012; Wortman, Darbinian, Sawaya, Khalili, & Amini, 2002). Viral gene transcription is thought to be regulated by *TCF4*, Sp1, and Tat interaction (Rossi et al., 2006). The closest genes to these transcription factor binding site variants are *MB21D2* towards the centromere and *HRASLS* towards the terminus, both approximately 100 kb away (see Figure 5). The nearby gene, *HRASLS*, is most highly expressed in EBV transformed lymphocytes and transformed fibroblast cells (Hinrichs et al., 2006). Interestingly, one of the annotated UCSC genome browser DNAase hypersensitivity peaks (chr3:192748121–192748415) is shown in the Huh-7.5 hepatocellular carcinoma cell line selected for high levels of hepatitis C replication. Collectively, this GWAS result and these prior references implicate this transcription factor binding and flanking epigenetic sites in the response to the herpes zoster virus and suggest this epigenetic locus has function in a general viral response program. This novel herpes zoster susceptibility locus shows the continued potential of discovering novel genome associations with EMR phenotypes contained in the eMERGE network resources.

Other next steps would be to more robustly extend the analysis of *R*-square quality variant inclusion in the rare frequency spectrum of variation, where the majority of genetic variation resides (Coventry et al., 2010; Keinan & Clark, 2012). This would require gene or genomic segment burden-based collapsing tests like sequence kernel association test (SKAT) analysis (Wu et al., 2011) of imputed variants. Additionally, we did not yet explore the power gains possible with the use of the continuous 0–2 variant dose imputation calls and these imputed variant doses' relations to the imputation quality. We simply used the hard-called genotype data with a posterior probability of >50% in this first pass validation of the imputed data.

This eMERGE 3 data set includes EMR and genomic data from 12 medical centers and can power the investigation of novel associations and known replications in both GWAS and PheWAS contexts. The EMR phenotyping algorithm components of the eMERGE network provides access to ICD-9 phenotype codes, numerous clinically generated phenotypes and laboratory medicine assay values unified across the sampled human participants (Bielinski et al., 2015; Carroll, Eyler, & Denny, 2015; McCarty et al., 2011; McDavid et al., 2013; Newton et al., 2013; Pathak et al., 2011). eMERGE currently has ~1,800 ICD-9-derived PheWAS codes which could be investigated by similar analyses to this herpes zoster shingles example using the same imputed genotype data. We expect the models derived from this Big Genomic Data genotype set will inform genetic risk assessments in a clinically defined context, providing practical and pragmatic solutions to assess genomic disease risk in an *a priori* manner. eMERGE also aims to improve and inform transfer of genetic information to the usable clinical record. A forward thinking next step would be to compare concordance of association direction between the PheWAS-based case-control status with the NHGRI GWAS Catalog phenotypes and other similar biobank scale efforts, as was done on a smaller scale previously within eMERGE (Denny et al., 2013).

The ascertainment methods of eMERGE make for a progressive use of existing resources. The pooling of biobank resources, direct querying of ICD-9/10 records, standardized intake and consent procedures during clinic visits and follow-up by the medical practitioners make for efficient study enrollment. Other biobank recruitment efforts have found that clinic visit ascertainment of participants to be the most cost-effective at reference population biobank building (Salowe et al., 2017). Salowe et al. identified a saturating effect of their ascertainment potential in a short enrollment period and that community outreach-oriented ascertainment programs have low rates of enrollment in comparison to the clinical ascertainments (Salowe et al., 2017). This suggests that after initial identification in the existing clinic population, the temporal intake rate will be defined by the new diagnosis rate of in situ residents. Big Genomic Data with many predictors homogenized across the ascertainment by imputation and/or molecular methods, large sample size and EMR phenotype information powers genetic risk discovery. ICD-9 is translatable to a direct economic indicator of the burden of cost for each variant we discover.

There are several biobank projects that have amassed tens of thousands to hundreds of thousands of samples linked to medical records and are aimed at population genetic risk assessment. The United Kingdom Biobank is a particularly notable resource (Thompson & Willeit, 2015) with many phenotypes and genetic associations for ~500,000 individuals (Clarke et al., 2017; Ge, Chen, Neale, Sabuncu, & Smoller, 2017; Howard, Adams et al., 2017; Howard, Hall et al., 2017; Luciano et al., 2018; Pilling et al., 2017; Taylor, Davey Smith, & Munafo, 2018; Ward et al., 2017) of public health importance. The Chinese Kadoorie (Dong et al., 2017; Sun et al., 2018; L. X. Wang et al., 2017; M. Wang et al., 2017; L. Yang et al., 2017; Yu et al., 2017) and Guangzhou Biobanks (Au Yeung et al., 2017, 2018; Pan et al., 2017; Xu, Jiang, Lam et al., 2017; Xu, Jiang, Schooling et al., 2017; Xu, Lam et al., 2017; S. Yang et al., 2017) have also grown extensively to be leading investigative bodies in phenotype and genotype risk estimation. Many consortia have also joined resources to power and replicate their discoveries (Hagenaars et al., 2016; Hoffmann et al., 2016; Hoffmann et al., 2017; Kraja et al., 2017; Liu et al., 2017; Scott et al., 2017). eMERGE exists as a cost-effective early effort among this ecosphere of medical informatics and genomics biobanks. Looking for concordance between ascertainment strategies, imputation, association results, and estimation of genetic risks will be the weight of evidence informing our present and future genomic medicine risk assessments. We are just at the beginning of personalized genomics evidence-based medicine. The United States has also recently started two new programs, the Million Veteran's Program (https://www.research.va.gov/mvp/) and the 'All of Us' Research Program (https://allofus.nih.gov/), which seeks to enroll 1 million participants. Health Maintenance Organizations (Kaiser Permanente, Geisinger, etc.) are also integrating routine discovery consent procedures of regular bio-specimen samples into the clinical care practice. These efforts will make the discovery and predictive statistics describing personalized genomic healthcare more tractable and reproducible by multiple institutions. This will also robustly inform clinical practice to the interpretation of underlying variables of medical risk and diagnosis in genomics and multivariate predictive medicine.

8 | WEB RESOURCES

Michigan Imputation Server (MIS) (https://imputationserver.sph.umich.edu/)
Minimac3 (http://genome.sph.umich.edu/wiki/Minimac3)
plink (http://pngu.mgh.harvard.edu/~purcell/plink/index.shtml)
plink1.9 and plink2 (https://www.cog-genomics.org/plink2)
vcftools (http://vcftools.sourceforge.net/man_latest.html)
bcftools (https://samtools.github.io/bcftools/bcftools.html)
William Wrayner's Haplotype Reference Consortium Variant Pruning Tool (http://www.well.ox.ac.uk/~wrayner/tools/HRC-1000 G-check-bim.v4.2.5.zip)
checkVCF.py (https://github.com/zhanxw/checkVCF)
Factorbook (http://www.factorbook.org/human/chipseq/tf/)
NCBI (https://www.ncbi.nlm.nih.gov/)

ACKNOWLEDGMENTS

In eMERGE network (Phase 3 ascertainment), this phase of the eMERGE Network was initiated and funded by the NHGRI through the following grants: U01HG8657 (Kaiser Washington/University of Washington); U01HG8685 (Brigham and Women's Hospital); U01HG8672 (Vanderbilt University Medical Center); U01HG8666 (Cincinnati Children's Hospital Medical Center); U01HG6379 (Mayo Clinic); U01HG8679 (Geisinger Clinic); U01HG8680 (Columbia University Health Sciences); U01HG8684 (Children's Hospital of Philadelphia); U01HG8673 (Northwestern University); U01HG8701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG8676 (Partners Healthcare/Broad Institute); and U01HG8664 (Baylor College of Medicine). In eMERGE network (Phase 1 and 2 ascertainment), the eMERGE Network was initiated and funded by NHGRI through the following grants: U01HG006828 (Cincinnati Children's Hospital Medical Center/Boston Children's Hospital); U01HG006830 (Children's Hospital of Philadelphia); U01HG006389 (Essentia Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University); U01HG006382 (Geisinger Clinic); U01HG006375 (Group Health Cooperative/University of Washington); U01HG006379 (Mayo Clinic); U01HG006380 (Icahn School of Medicine at Mount Sinai); U01HG006388 (Northwestern University); U01HG006378 (Vanderbilt University Medical Center); and U01HG006385 (Vanderbilt University Medical Center serving as the Coordinating Center) with U01HG004438 (CIDR) and U01HG004424 (the Broad Institute) serving as Genotyping Centers.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

AUTHORS' CONTRIBUTIONS

I. B. S. imputed, merged, and QC/QAed the genotype information, performed associations, and wrote the manuscript. R. J. C. provided the phenotype information. D. R. C. oversaw the conception and development of the project. All authors read and made suggestions for the drafted manuscript versions and/or contributed to the collective biobanks' phenotype and genotype information and development of the eMERGE network collaborations.

ORCID

Ian B. Stanaway http://orcid.org/0000-0002-0783-0918

REFERENCES

Au Yeung, S. L., Jiang, C., Cheng, K. K., Xu, L., Zhang, W., Lam, T. H., … Schooling, C. M. (2017). Age at menarche and cardiovascular risk factors using Mendelian randomization in the Guangzhou Biobank Cohort Study. *Preventive Medicine*, *101*, 142–148. https://doi.org/10.1016/j.ypmed.2017.06.006

Au Yeung, S. L., Jiang, C., Cheng, K. K., Xu, L., Zhang, W., Lam, T. H., … Schooling, C. M. (2018). Age at menarche and depressive symptoms in older Southern Chinese women: A Mendelian randomization study in the Guangzhou Biobank Cohort Study. *Psychiatry Research*, *259*, 32–35. https://doi.org/10.1016/j.psychres.2017.09.040

Bielinski, S. J., Pathak, J., Carrell, D. S., Takahashi, P. Y., Olson, J. E., Larson, N. B., … Roger, V. L. (2015). A robust e-epidemiology tool in phenotyping heart failure with differentiation for preserved and reduced ejection fraction: The Electronic Medical Records and Genomics (eMERGE) network. *Journal of Cardiovascular Translational Research*, *8*(8), 475–483. https://doi.org/10.1007/s12265-015-9644-2

Browning, B. L., & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics*, *84*(2), 210–223. https://doi.org/10.1016/j.ajhg.2009.01.005

Campbell, M., Watanabe, T., Nakano, K., Davis, R. R., Lyu, Y., Tepper, C. G., … Izumiya, Y. (2018). KSHV episomes reveal dynamic chromatin loop formation with domain-specific gene regulation. *Nature Communications*, *9*(1), 49. https://doi.org/10.1038/s41467-017-02089-9

Carroll, R. J., Eyler, A. E., & Denny, J. C. (2015). Intelligent use and clinical benefits of electronic health records in rheumatoid arthritis. *Expert Review of Clinical Immunology*, *11*(3), 329–337. https://doi.org/10.1586/1744666X.2015.1009895

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 7. https://doi.org/10.1186/s13742-015-0047-8

Chen, X., Cheng, L., Jia, X., Zeng, Y., Yao, S., Lv, Z., … Lu, C. (2009). Human immunodeficiency virus type 1 Tat accelerates Kaposi sarcoma-associated herpesvirus Kaposin A-mediated tumorigenesis of transformed fibroblasts in vitro as well as in nude and immunocompetent mice. *Neoplasia*, *11*(12), 1272–1284.

Chirmule, N., Goonewardena, H., Pahwa, S., Pasieka, R., Kalyanaraman, V. S., & Pahwa, S. (1995). HIV-1 envelope glycoproteins induce activation of activated protein-1 in CD4+ T cells. *Journal of Biological Chemistry*, *270*(33), 19364–19369.

Chisholm, R. L. (2013). At the interface between medical informatics and personalized medicine: The eMERGE network experience. *Healthcare Informatics Research*, *19*(2), 67–68. https://doi.org/10.4258/hir.2013.19.2.67

Clarke, T. K., Adams, M. J., Davies, G., Howard, D. M., Hall, L. S., Padmanabhan, S., … McIntosh, A. M. (2017). Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank (N = 112 117). *Molecular Psychiatry*, *22*(10), 1376–1384. https://doi.org/10.1038/mp.2017.153

Conomos, M. P., Reiner, A. P., Weir, B. S., & Thornton, T. A. (2016). Model-free estimation of recent genetic relatedness. *American Journal of Human Genetics*, *98*(1), 127–148. https://doi.org/10.1016/j.ajhg.2015.11.022

Coventry, A., Bull-Otterson, L. M., Liu, X., Clark, A. G., Maxwell, T. J., Crosby, J., … Sing, C. F. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications*, *1*, 131. https://doi.org/10.1038/ncomms1130

Crawford, D. C., Crosslin, D. R., Tromp, G., Kullo, I. J., Kuivaniemi, H., Hayes, M. G., … Ritchie, M. D. (2014). eMERGEing progress in genomics-the first seven years. *Frontiers in Genetics*, *5*, 184. https://doi.org/10.3389/fgene.2014.00184

Cronin, R. M., Field, J. R., Bradford, Y., Shaffer, C. M., Carroll, R. J., Mosley, J. D., … Denny, J. C. (2014). Phenome-wide association studies demonstrating pleiotropy of genetic variants within FTO with and without adjustment for body mass index. *Frontiers in Genetics*, *5*, 250. https://doi.org/10.3389/fgene.2014.00250

Crosslin, D. R., Carrell, D. S., Burt, A., Kim, D. S., Underwood, J. G., Hanna, D. S., … Jarvik, G. P. (2015). Genetic variation in the HLA region is associated with susceptibility to herpes zoster. *Genes and Immunity*, *16*(1), 1–7. https://doi.org/10.1038/gene.2014.51

Crosslin, D. R., McDavid, A., Weston, N., Nelson, S. C., Zheng, X., Hart, E., … Jarvik, G. P. (2012). Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE network. *Human Genetics*, *131*(4), 639–652. https://doi.org/10.1007/s00439-011-1103-9

Crosslin, D. R., McDavid, A., Weston, N., Zheng, X., Hart, E., de Andrade, M., … Jarvik, G. P. (2013). Genetic variation associated with circulating monocyte count in the eMERGE network. *Human Molecular Genetics*, *22*(10), 2119–2127. https://doi.org/10.1093/hmg/ddt010

Crosslin, D. R., Tromp, G., Burt, A., Kim, D. S., Verma, S. S., Lucas, A. M., … de Andrade, M. (2014). Controlling for population structure and genotyping platform bias in the eMERGE multi-institutional biobank linked to electronic health records. *Frontiers in Genetics*, *5*, 352. https://doi.org/10.3389/fgene.2014.00352

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danziger, O., Pupko, T., Bacharach, E., & Ehrlich, M. (2018). Interleukin-6 and interferon-alpha signaling via JAK1-STAT differentially regulate oncolytic versus cytoprotective antiviral states. *Frontiers in Immunology*, *9*, 94. https://doi.org/10.3389/fimmu.2018.00094

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/ng.3656

De, R., Verma, S. S., Drenos, F., Holzinger, E. R., Holmes, M. V., Hall, M. A., … Gilbert-Diamond, D. (2015). Identifying gene-gene interactions that are highly associated with body mass index using quantitative multifactor dimensionality reduction (QMDR). *BioData Mining*, *8*, 41. https://doi.org/10.1186/s13040-015-0074-0

De, R., Verma, S. S., Holzinger, E., Hall, M., Burt, A., Carrell, D. S., … Gilbert-Diamond, D. (2017). Identifying gene-gene interactions that are highly associated with four quantitative lipid traits across multiple cohorts. *Human Genetics*, *136*(2), 165–178. https://doi.org/10.1007/s00439-016-1738-7

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., … Roden, D. M. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, *31*(12), 1102–1110. https://doi.org/10.1038/nbt.2749

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., … Crawford, D. C. (2010). PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, *26*(9), 1205–1210. https://doi.org/10.1093/bioinformatics/btq126

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*(4), 997–1004.

Dong, W., Pan, X. F., Yu, C., Lv, J., Guo, Y., Bian, Z., … Li, L. (2017). Self-rated health status and risk of ischemic heart disease in the China Kadoorie Biobank Study: A population-based cohort study. *Journal of the American Heart Association*, *6*(9), e006595. https://doi.org/10.1161/JAHA.117.006595

Dumitrescu, L., Ritchie, M. D., Denny, J. C., El Rouby, N. M., McDonough, C. W., Bradford, Y., … Crawford, D. C. (2017). Genome-wide study of resistant hypertension identified from electronic health records. *PLOS One*, *12*(2), e0171745. https://doi.org/10.1371/journal.pone.0171745

Fellay, J., Ge, D., Shianna, K. V., Colombo, S., Ledergerber, B., Cirulli, E. T., … Goldstein, D. B. (2009). Common genetic variation and the control of HIV-1 in humans. *PLOS Genetics*, *5*(12), e1000791. https://doi.org/10.1371/journal.pgen.1000791

Gallego, C. J., Burt, A., Sundaresan, A. S., Ye, Z., Shaw, C., Crosslin, D. R., … Jarvik, G. P. (2015). Penetrance of hemochromatosis in hfe genotypes resulting in p.Cys282Tyr and p.[Cys282Tyr];[His63Asp] in the eMERGE network. *American Journal of Human Genetics*, *97*(4), 512–520. https://doi.org/10.1016/j.ajhg.2015.08.008

Ge, T., Chen, C. Y., Neale, B. M., Sabuncu, M. R., & Smoller, J. W. (2017). Phenome-wide heritability analysis of the UK Biobank. *PLoS Genetics*, *13*(4), e1006711. https://doi.org/10.1371/journal.pgen.1006711

Gogarten, S. M., Bhangale, T., Conomos, M. P., Laurie, C. A., McHugh, C. P., Painter, I., … Laurie, C. C. (2012). GWASTools: An R/bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics*, *28*(24), 3329–3331. https://doi.org/10.1093/bioinformatics/bts610

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., … Williams, M. S. (2013). The Electronic Medical Records and Genomics (eMERGE) network: Past, present, and future. *Genetics in Medicine*, *15*(10), 761–771. https://doi.org/10.1038/gim.2013.72

Hagenaars, S. P., Harris, S. E., Davies, G., Hill, W. D., Liewald, D. C. M., Ritchie, S. J., … Deary, I. J. (2016). Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N = 112 151) and 24 GWAS consortia. *Molecular Psychiatry*, *21*(11), 1624–1632. https://doi.org/10.1038/mp.2015.225

Hall, M. A., Verma, S. S., Wallace, J., Lucas, A., Berg, R. L., Connolly, J., … Ritchie, M. D. (2015). Biology-driven gene-gene interaction analysis of age-related cataract in the eMERGE network. *Genetic Epidemiology*, *39*(5), 376–384. https://doi.org/10.1002/gepi.21902

Henderson, L. J., Sharma, A., Monaco, M. C. G., Major, E. O., & Al-Harthi, L. (2012). Human immunodeficiency virus type 1 (HIV-1) transactivator of transcription through its intact core and cysteine-rich domains inhibits Wnt/beta-catenin signaling in astrocytes: Relevance to HIV neuropathogenesis. *Journal of Neuroscience*, *32*(46), 16306–16313. https://doi.org/10.1523/JNEUROSCI.3145-12.2012

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., & Kent, W. J. (2006). The UCSC Genome Browser Database: Update 2006. *Nucleic Acids Research*, *34* (Database issue), D590–D598. https://doi.org/10.1093/nar/gkj144

Hoffmann, T. J., Ehret, G. B., Nandakumar, P., Ranatunga, D., Schaefer, C., Kwok, P. Y., … Risch, N. (2017). Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nature Genetics*, *49*(1), 54–64. https://doi.org/10.1038/ng.3715

Hoffmann, T. J., Keats, B. J., Yoshikawa, N., Schaefer, C., Risch, N., & Lustig, L. R. (2016). A large genome-wide association study of age-related hearing impairment using electronic health records. *PLOS Genetics*, *12*(10), e1006371. https://doi.org/10.1371/journal.pgen.1006371

Howard, D. M., Adams, M. J., Clarke, T. K., Wigmore, E. M., Zeng, Y., Hagenaars, S. P., ... McIntosh, A. M. (2017). Haplotype-based association analysis of general cognitive ability in Generation Scotland, the English Longitudinal Study of Ageing, and UK Biobank. *Wellcome Open Research*, *2*, 61. https://doi.org/10.12688/wellcomeopenres.12171.1

Howard, D. M., Hall, L. S., Hafferty, J. D., Zeng, Y., Adams, M. J., Clarke, T. K., ... McIntosh, A. M. (2017). Genome-wide haplotype-based association analysis of major depressive disorder in Generation Scotland and UK Biobank. *Translational Psychiatry*, *7*(11), 1263. https://doi.org/10.1038/s41398-017-0010-9

Howie, B., Marchini, J., & Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3*, *1*(6), 457–470. https://doi.org/10.1534/g3.111.001198

Howie, B., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLOS Genetics*, *5*(6), e1000529. https://doi.org/10.1371/journal.pgen.1000529

Jelicic, K., Cimbro, R., Nawaz, F., Huang, D. W., Zheng, X., Yang, J., ... Fauci, A. S. (2013). The HIV-1 envelope protein gp120 impairs B cell proliferation by inducing TGF-beta1 production and FcRL4 expression. *Nature Immunology*, *14*(12), 1256–1265. https://doi.org/10.1038/ni.2746

Keinan, A., & Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, *336*(6082), 740–743. https://doi.org/10.1126/science.1217283

Kraja, A. T., Cook, J. P., Warren, H. R., Surendran, P., Liu, C., Evangelou, E., ... Howson, J. M. M. (2017). New blood pressure-associated loci identified in meta-analyses of 475 000 individuals. *Circulation: Cardiovascular Genetics*, *10*(5), e001778. https://doi.org/10.1161/CIRCGENETICS.117.001778

Kullo, I. J., Ding, K., Shameer, K., McCarty, C. A., Jarvik, G. P., Denny, J. C., ... Chute, C. G. (2011). Complement receptor 1 gene variants are associated with erythrocyte sedimentation rate. *American Journal of Human Genetics*, *89*(1), 131–138. https://doi.org/10.1016/j.ajhg.2011.05.019

Lee, C., Abdool, A., & Huang, C. H. (2009). PCA-based population structure inference with generic clustering algorithms. *BMC Bioinformatics*, *10*(Suppl 1), S73. https://doi.org/10.1186/1471-2105-10-S1-S73

Liu, D. J., Peloso, G. M., Yu, H., Butterworth, A. S., Wang, X., Mahajan, A., ... Kathiresan, S. (2017). Exome-wide association study of plasma lipids in >300,000 individuals. *Nature Genetics*, *49*(12), 1758–1766. https://doi.org/10.1038/ng.3977

Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Y, Y., H, H., ... L Price, A. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, *48*(11), 1443–1448. https://doi.org/10.1038/ng.3679

Luciano, M., Hagenaars, S. P., Davies, G., Hill, W. D., Clarke, T. K., Shirali, M., ... Deary, I. J. (2018). Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nature Genetics*, *50*(1), 6–11. https://doi.org/10.1038/s41588-017-0013-8

Malinowski, J. R., Denny, J. C., Bielinski, S. J., Basford, M. A., Bradford, Y., Peissig, P. L., ... Crawford, D. C. (2014). Genetic variants associated with serum thyroid stimulating hormone (TSH) levels in European Americans and African Americans from the eMERGE Network. *PLOS One*, *9*(12), e111301. https://doi.org/10.1371/journal.pone.0111301

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Martorelli, D., Muraro, E., Mastorci, K., Dal Col, J., Faè, D. A., Furlan, C., ... Dolcetti, R. (2015). A natural HIV p17 protein variant up-regulates the LMP-1 EBV oncoprotein and promotes the growth of EBV-infected B-lymphocytes: Implications for EBV-driven lymphomagenesis in the HIV setting. *International Journal of Cancer*, *137*(6), 1374–1385. https://doi.org/10.1002/ijc.29494

Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., D'souza, C., Fouse, S. D., ... Costello, J. F. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, *466*(7303), 253–257. https://doi.org/10.1038/nature09165

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., & Haplotype Reference, C. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. https://doi.org/10.1038/ng.3643

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

McDavid, A., Crane, P. K., Newton, K. M., Crosslin, D. R., McCormick, W., Weston, N., ... Carlson, C. S. (2013). Enhancing the power of genetic association studies through the use of silver standard cases derived from electronic medical records. *PLOS One*, *8*(6), e63481. https://doi.org/10.1371/journal.pone.0063481

Mosley, J. D., Shaffer, C. M., Van Driest, S. L., Weeke, P. E., Wells, Q. S., Karnes, J. H., ... Roden, D. M. (2016). A genome-wide association study identifies variants in KCNIP4 associated with ACE inhibitor-induced cough. *Pharmacogenomics Journal*, *16*(3), 231–237. https://doi.org/10.1038/tpj.2015.51

Namjou, B., Marsolo, K., Lingren, T., Ritchie, M. D., Verma, S. S., Cobb, B. L., ... Harley, J. B. (2015). A GWAS study on liver function test using eMERGE network participants. *PLOS One*, *10*(9), e0138677. https://doi.org/10.1371/journal.pone.0138677

Newton, K. M., Peissig, P. L., Kho, A. N., Bielinski, S. J., Berg, R. L., Choudhary, V., ... Denny, J. C. (2013). Validation of electronic medical record-based phenotyping algorithms: Results and lessons learned from the eMERGE network. *Journal of the American Medical Informatics Association*, *20*(e1), e147–e154. https://doi.org/10.1136/amiajnl-2012-000896

Ng, M. C. Y., Shriner, D., Chen, B. H., Li, J., Chen, W. M., Guo, X., ... Bowden, D. W. (2014). Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. *PLOS Genetics*, *10*(8), e1004517. https://doi.org/10.1371/journal.pgen.1004517

Pan, J., Xu, L., Lam, T. H., Jiang, C. Q., Zhang, W. S., Jin, Y. L., … Adab, P. (2017). Association of adiposity with pulmonary function in older Chinese: Guangzhou Biobank Cohort Study. *Respiratory Medicine*, *132*, 102–108. https://doi.org/10.1016/j.rmed.2017.10.003

Pathak, J., Wang, J., Kashyap, S., Basford, M., Li, R., Masys, D. R., & Chute, C. G. (2011). Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: The eMERGE network experience. *Journal of the American Medical Informatics Association*, *18*(4), 376–386. https://doi.org/10.1136/amiajnl-2010-000061

Pilling, L. C., Kuo, C. L., Sicinski, K., Tamosauskaite, J., Kuchel, G. A., Harries, L. W., … Melzer, D. (2017). Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging*, *9*(12), 2504–2520. https://doi.org/10.18632/aging.101334

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

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

Ritchie, M. D., Verma, S. S., Hall, M. A., Goodloe, R. J., Berg, R. L., Carrell, D. S., … McCarty, C. A. (2014). Electronic medical records and genomics (eMERGE) network exploration in cataract: Several new potential susceptibility loci. *Molecular Vision*, *20*, 1281–1295.

Roca Suarez, A. A., Van Renne, N., Baumert, T. F., & Lupberger, J. (2018). Viral manipulation of STAT3: Evade, exploit, and injure. *PLOS Pathogens*, *14*(3), e1006839. https://doi.org/10.1371/journal.ppat.1006839

Rossi, A., Mukerjee, R., Ferrante, P., Khalili, K., Amini, S., & Sawaya, B. E. (2006). Human immunodeficiency virus type 1 Tat prevents dephosphorylation of Sp1 by TCF-4 in astrocytes. *Journal of General Virology*, *87*(Pt 6), 1613–1623. https://doi.org/10.1099/vir.0.81691-0

Salowe, R., O'keefe, L., Merriam, S., Lee, R., Khachatryan, N., Sankar, P., … O'Brien, J. (2017). Cost and yield considerations when expanding recruitment for genetic studies: The primary open-angle African American glaucoma genetics study. *BMC Medical Research Methodology*, *17*(1), 101. https://doi.org/10.1186/s12874-017-0374-9

Scott, R. A., Scott, L. J., Mägi, R., Marullo, L., Gaulton, K. J., Kaakinen, M., … Prokopenko, I. (2017). An expanded genome-wide association study of type 2 diabetes in europeans. *Diabetes*, *66*(11), 2888–2902. https://doi.org/10.2337/db16-1253

Solovieff, N., Hartley, S. W., Baldwin, C. T., Perls, T. T., Steinberg, M. H., & Sebastiani, P. (2010). Clustering by genetic ancestry using genome-wide SNP data. *BMC Genetics*, *11*, 108. https://doi.org/10.1186/1471-2156-11-108

Sun, X., Zheng, B., Lv, J., Guo, Y., Bian, Z., Yang, L., … YU, C. (2018). Sleep behavior and depression: Findings from the China Kadoorie Biobank of 0.5 million Chinese adults. *Journal of Affective Disorders*, *229*, 120–124. https://doi.org/10.1016/j.jad.2017.12.058

Taylor, A. E., Davey Smith, G., & Munafò, M. R. (2018). Associations of coffee genetic risk scores with consumption of coffee, tea and other beverages in the UK Biobank. *Addiction*, *113*(1), 148–157. https://doi.org/10.1111/add.13975

Thompson, S. G., & Willeit, P. (2015). UK Biobank comes of age. *Lancet*, *386*(9993), 509–510. https://doi.org/10.1016/S0140-6736(15)60578-5

Tian, C., Hromatka, B. S., Kiefer, A. K., Eriksson, N., Noble, S. M., Tung, J. Y., & Hinds, D. A. (2017). Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nature Communications*, *8*(1), 599. https://doi.org/10.1038/s41467-017-00257-5

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., … Ritchie, M. D. (2011). Quality control procedures for genome-wide association studies. *Current Protocols in Human Genetics/Editorial Board, Jonathan L. Haines… [et al.]*, *68*, 1.19.1–1.19.18. https://doi.org/10.1002/0471142905.hg0119s68

Verma, A., Verma, S. S., Pendergrass, S. A., Crawford, D. C., Crosslin, D. R., Kuivaniemi, H., … Tromp, G. (2016). eMERGE phenome-wide association study (PheWAS) identifies clinical associations and pleiotropy for stop-gain variants. *BMC Medical Genomics*, *9* (Suppl 1), 32. https://doi.org/10.1186/s12920-016-0191-8

Verma, S. S., de Andrade, M., Tromp, G., Kuivaniemi, H., Pugh, E., Namjou-Khales, B., … Ritchie, M. D. (2014). Imputation and quality control steps for combining multiple genome-wide datasets. *Frontiers in Genetics*, *5*, 370. https://doi.org/10.3389/fgene.2014.00370

Verma, S. S., Cooke Bailey, J. N., Lucas, A., Bradford, Y., Linneman, J. G., Hauser, M. A., … Ritchie, M. D. (2016). Epistatic gene-based interaction analyses for glaucoma in eMERGE and NEIGHBOR consortium. *PLOS Genetics*, *12*(9), e1006186. https://doi.org/10.1371/journal.pgen.1006186

Wang, L. X., Fan, M. Y., Yu, C. Q., Guo, Y., Bian, Z., Tan, Y. L., … China Kadoorie Biobank (CKB) Collaborative Group (2017). Association between body mass index and both total and cause-specific mortality in China: Findings from data through the China Kadoorie Biobank. *Zhonghua Liu Xing Bing Xue Za Zhi*, *38*(2), 205–211. https://doi.org/10.3760/cma.j.issn.0254-6450.2017.02.014

Wang, M., Hu, R. Y., Wang, H., Gong, W. W., Wang, C. M., Xie, K. X., … Li, L. M. (2017). Age at natural menopause and risk of diabetes in adult women: Findings from the China Kadoorie Biobank study in the Zhejiang area. *Journal of Diabetes Investigation*, *9*, 762–768. https://doi.org/10.1111/jdi.12775

Ward, J., Strawbridge, R. J., Bailey, M. E. S., Graham, N., Ferguson, A., Lyall, D. M., … Smith, D. J. (2017). Genome-wide analysis in UK Biobank identifies four loci associated with mood instability and genetic correlation with major depressive disorder, anxiety disorder and schizophrenia. *Translational Psychiatry*, *7*(11), 1264. https://doi.org/10.1038/s41398-017-0012-7

Wortman, B., Darbinian, N., Sawaya, B. E., Khalili, K., & Amini, S. (2002). Evidence for regulation of long terminal repeat transcription by Wnt transcription factor TCF-4 in human astrocytic cells. *Journal of Virology*, *76*(21), 11159–11165.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, *89*(1), 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029

Xu, L., Jiang, C. Q., Schooling, C. M., Zhang, W. S., Cheng, K. K., & Lam, T. H. (2017). Liver enzymes as mediators of association between obesity and diabetes: The Guangzhou Biobank Cohort Study. *Annals of Epidemiology*, *27*(3), 204–207. https://doi.org/10.1016/j.annepidem.2016.11.002

Xu, L., Lam, T. H., Jiang, C. Q., Zhang, W. S., Jin, Y. L., Zhu, T., … Cheng, K. K. (2017). Adiposity and incident diabetes within 4 years of follow-up: The Guangzhou Biobank Cohort Study. *Diabetic Medicine*, *34*(10), 1400–1406. https://doi.org/10.1111/dme.13378

Xu, L., Qiang Jiang, C., Hing Lam, T., Sen Zhang, W., Zhu, F., Li Jin, Y., … Mary Schooling, C. (2017). Mendelian randomization estimates of alanine aminotransferase with cardiovascular disease: Guangzhou Biobank Cohort study. *Human Molecular Genetics*, *26*(2), 430–437. https://doi.org/10.1093/hmg/ddw396

Yang, L., Li, L., Millwood, I. Y., Lewington, S., Guo, Y., Sherliker, P., … Chen, Z. (2017). Adiposity in relation to age at menarche and other reproductive factors among 300 000 Chinese women: Findings from China Kadoorie Biobank study. *International Journal of Epidemiology*, *46*(2), 502–512. https://doi.org/10.1093/ije/dyw165

Yang, S., Xu, L., He, Y., Jiang, C., Jin, Y., Cheng, K. K., … Lam, T. H. (2017). Childhood secondhand smoke exposure and pregnancy loss in never smokers: The Guangzhou Biobank Cohort Study. *Tobacco Control*, *26*(6), 697–702. https://doi.org/10.1136/tobaccocontrol-2016-053239

Yu, C., Shi, Z., Lv, J., Guo, Y., Bian, Z., Du, H., … Li, L. (2017). Dietary patterns and insomnia symptoms in Chinese adults: The China Kadoorie Biobank. *Nutrients*, *9*(3), 232. https://doi.org/10.3390/nu9030232

Zeng, Y., Zhang, X., Huang, Z., Cheng, L., Yao, S., Qin, D., … Lu, C. (2007). Intracellular Tat of human immunodeficiency virus type 1 activates lytic cycle replication of Kaposi's sarcoma-associated herpesvirus: Role of JAK/STAT signaling. *Journal of Virology*, *81*(5), 2401–2417. https://doi.org/10.1128/JVI.02024-06

Zuvich, R. L., Armstrong, L. L., Bielinski, S. J., Bradford, Y., Carlson, C. S., Crawford, D. C., … Ritchie, M. D. (2011). Pitfalls of merging GWAS data: Lessons learned in the eMERGE network and quality control procedures to maintain high data quality. *Genetic Epidemiology*, *35*(8), 887–898. https://doi.org/10.1002/gepi.20639

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Stanaway IB, Hall TO, Rosenthal EA, et al. The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype. *Genet. Epidemiol.* 2019;43:63–81. https://doi.org/10.1002/gepi.22167