eMERGE Network Proposal for Analysis

Project/Manuscript Concept Sheet

Submission Date	
Project Title	A Phenome-wide Survey of the Phenotypic Effects of High-Frequency Human-Derived Alleles
Tentative Lead Investigator (first author)	Corinne Simonti
Tentative Senior Author (last author)	John A. Capra
All other authors	Josh Denny, other eMERGE collaborators TBD as appropriate
Sites Involved	All analyses will be performed at Vanderbilt University. Data from all adult sites will be analyzed.
Background / Significance	Building on our success applying the PheWAS methodology on eMERGE network data to characterize phenotypic effects of Neanderthal DNA in modern human populations, we propose to apply this approach to address another important question in recent human evolution. By comparing the sequences of modern humans to those of closely related great ape populations, we can identify new alleles that are specific to the human lineage and show signatures of significant selection against the ancestral allele (i.e., the one present in the human-chimp ancestor) in humans. We have identified ~100,000 such variants for which the derived (human-specific) allele is at 90% or greater frequency in human populations. By testing for these alleles in ancient DNA samples, we can further predict the timing of their appearance and increase in frequency. PheWAS provides a comprehensive framework in which to explore the phenotypic effects of these variants that will shed light on phenotypes that have experienced strong selective pressure on the human lineage.
Outline of Project	From available genotype data, we will identify nearly fixed human-specific alleles in each individual. Then, using ICD9-derived phenotypes we will perform phenome-wide association tests on the human-specific alleles. We will further test using GCTA whether these variants explain a significant portion of the proportion of phenotypic variance in several common, evolutionarily relevant traits.
Desired Variables (essential for analysis indicated by *)	Phenotypes derived from ICD9 code groups for each subject for phenome-wide scans. Standard demographic information, e.g., sex, age, and PCs, to be used as covariates. We may also test specific hypotheses based on other phenotypic data, e.g., bone mineral density from DEXA scans, when available.

Desired data	Genotype calls and imputed data in eMERGE release 1 and 2 set.
Planned Statistical Analyses	We will identify human-specific variants by combining genome-wide human variation data from diverse human populations from the 1000 Genomes Project, great ape variation data across 100s of great apes from the Great Ape Genome Project, and ancient DNA from Neanderthals, Denisovans, and several early humans. We will follow a similar strategy for identifying human-specific alleles at high frequency and signatures of recent selection used in Prüfer et al. (2013), but we will incorporate the population-level great ape data and more recently sequenced archaic hominin DNA into these analyses. To characterize the phenotypic effects of these human-specific derived alleles, we will follow a similar analysis strategy as in our recent Neanderthal admixture paper (Simonti et al. 2016). In short, we will perform PheWAS discovery and replication meta-analyses (in eMERGE1 and eMERGE2) on the variants of interest using the ICD9-based phenotypes for which there are sufficient cases. All PheWAS analyses will be performed with the PheWAS R package. Since the ancestral alleles will be at low frequency, we will explore recent statistical innovations to the methodology, such as permutation-based p-value calculation strategies, to account for the highly correlated nature of the tests. We will further estimate the proportion of phenotypic variance in common phenotypes that is explained by the human-specific variants in aggregate. We will use the GREML framework implemented in GCTA and consider several different models with multiple GRMs for different sets of human-specific and non-human-specific variants. We will correct for multiple testing in these analyses using q-value-based false discovery rate control.
Ethical considerations	None.
Target Journal	Hope for *Nature Genetics* if major findings. Otherwise, *PLoS Genetics, AJHG,* or *Molecular Biology and Evolution*.
Milestones**	Project duration: 6 months. Initiate analysis: February 2016; Complete main PheWAS and GCTA analyses: May 2016; Complete follow-up analyses and manuscript draft: June 2016; Review of draft and submission: 2016.

** This section should include: Timeline for completion of project, including approval, project duration, first and second draft of the paper and submission.