|  |  |  |
| --- | --- | --- |
| **XeMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT314 | |
| **Submission Date** | 11/7/2018 | |
| **Project Title** | The Development of an Imputed Structural Variant Genomic Dataset and Association to Neurological and Alcohol Use Disorder Electronic Medical Record Phenotypes with Biobank Scale Subject Ascertainment | |
| **Tentative Lead Investigator** *(first author)* | Ian Byrell Stanaway | |
| **Tentative Senior Author**  *(last author)* | David Russell Crosslin | |
| **All Other Authors** | Gail P. Jarvik, Taryn O. Hall, Melody Palmer, Elisabeth Rosenthal, Adam Gordon, other interested eMERGE co-authors | |
| **Sites Participating** | All eMERGE Sites | |
| **Background / Significance** | We will impute structural variations (SV) for ~100,000 electronic MEdical Record and GEnomics (eMERGE) Network participants using the publicly available 1000 Genomes single nucleotide variant (SNV), polymorphism (SNP) and SV genetic maps [1]. eMERGE has ~84 SNV genotype array batches of samples from 12 medical centers providing the molecularly genotyped marker data. We propose to use the Beagle imputation software [2,3] produced by the Browning Lab at the University of Washington to do this imputation. The Sudmant et al., 2015 [1] paper, "An integrated map of structural variation in 2,504 human genomes" from Dr. Evan Eichler's group at the University of Washington describes in detail the SV genetic map we would use. The variants from this genetic map, "...comprising 42,279 biallelic deletions, 6,025 biallelic duplications, 2,929 mCNVs (multi allelic copy-number variants)...", provides a rich map of SVs to impute. Sudmant et al. also determines that these SVs are statistically correlated with SNV markers: "We found that 73% of SVs with >1% VAF and 68% of rarer SVs (VAF > 0.1%) are in linkage disequilibrium (LD) with nearby SNPs (r2 > 0.6);...", making imputation feasible as we intend to perform using the eMERGE SNP-SNV genotype array marker data.  Genetic disease risk estimations from Genome Wide Association Studies (GWAS) are a continually growing body of evidence we are beginning to use to understand the basis of complex disease. Many of the GWAS works published to date have been based on SNV analyses of common variants with variant allele frequencies >0.01-0.05. In addition to SNVs, SVs are another distinct and diverse class of genomic variation that exists in our populations. These SVs have not been explored in nearly the detail [4] which SNVs have been interrogated for disease association, especially at biobank scale sample sizes. In this proposal, I aim to establish an imputed SV dataset with ~100,000 samples indexed to electronic medical record (EMR) disease phenotypes provided by the electronic Medical Record and Genomics (eMERGE) Network of 12 medical centers. eMERGE has 23 robust clinical phenotype algorithms and ~1,700 EMR phenotypes [5-7] which we could interrogate for association with SVs. I have previously imputed ~40 million SNPs with 83,717 eMERGE human subjects genotyped on 79 different SNVs array batches with disparate marker ascertainments and performed genome wide association to the Herpes Zoster shingles phenotype [8]. SVs, including mCNVs and insertion-deletions are not in the Haplotype Reference Consortium SNV map [9, 10] we previously imputed, making them a logical addition to our genetic maps. We also have the mCNV sub-class of SVs called using log R and B-allele mCNV genotype array based PennCNV [11] calling on ~20,000 eMERGE participants genotyped on the Illumina 660k SNV array as an orthogonal comparison method. We would validate the imputed mCNV-SV calls concordance against these 660k array based SV calls.  Hypothesis 1: Common Structural Variants are readily imputable from single nucleotide polymorphism array data.  Hypothesis 2: Structural Variant alleles will have multiple known and novel disease risk associations to eMERGE defined phenotypes.  Aim 1: Impute Structural Variants onto the eMERGE ~100,000 subjects to make a comprehensive genetic map.  Aim 2: Validate the imputation results by comparing to copy number variant genotype calls in the eMERGE 1 660k single nucleotide polymorphism arrays via the log R and B-allele frequency calling methods for ~20,000 of the samples.  Aim 3: Perform genome wide structural variant and single nucleotide variant associations to the eMERGE defined neurological and alcohol use disorder phenotypes to establish disease association genomic risk evidence.  Research Strategy  (a) Significance  This project has the potential to increase eMERGE's catalog of risk variants types available to be implicated in complex disease pathogenesis. The joint SV and SNV [8] imputed eMERGE genome wide variant data on ~100,000 human subjects with ~1,700 EMR and 23 clinical algorithmic phenotypes [5-7] make for a deep systematic genotype-phenotype catalog as starting material to build robust genomic risk estimates from. These genomic risk variants have potential utility for prediction, intervention and prevention of disease in genomic risk prone individuals.  (b) Innovation  Adding imputed SVs to the genetic mapping resources available for analysis will be novel for the eMERGE Network. eMERGE has not published a large genome wide interrogation of the ~100,000 subject data with SVs. This imputed SV dataset will give us the potential to show the contribution of SVs to clinically defined disease etiology on a sample size scale that has not been possible until now.  (c) Dissemination  We will write one or more manuscripts on the methods we employed, associations and results we generate for this SV imputation projects. This imputed SV data set will also be available to the eMERGE Network for others to interrogate for their phenotypes of interest. Dr. Browning has also expressed interest in posting the merged SNV-SV imputation reference map I have phased on his Beagle software website for the genomics community to use in similar projects.  (d) Approach of Research Design  Aim 1: Impute Structural Variants onto the eMERGE ~100,000 subjects genotype array data.  Rational: Structural variants have been an understudied source of genomic variation in the eMERGE Network. Providing this SV resource for the network to utilize in many analyses will increase the utility of the data for genomic disease risk estimation.  Methods: The eMERGE Network of 12 medical centers has shared 99,211 consented patients' genotype array information in the making of a biobank scale national research effort. I intend to use Dr. Browning's and Beagle Imputation Software (http://faculty.washington.edu/browning/beagle/beagle.html) [2, 3] to perform chromosomal phasing and imputation of SV calls provided by the 1000 Genomes SV map [1].  Aim 2: Validate the imputation results by comparing to copy number variant genotype calls in the eMERGE 1 660k single nucleotide polymorphism arrays via the log R and B-allele frequency PennCNV calling method [11] for ~20,000 of the samples.  Rational: The eMERGE 1 660k array CNV calls will allow for an orthogonal validation of the imputed SV results in a large subset (~20%) of the samples which we statistically imputed.  Methods: We will bioinformatically catalog the common and rare variants detected in the 660k CNV data and the 1000 Genomes reference set. Among these variants shared by the reference and the 660k method, we will then see which impute well or are missed. Conversely, we will also be able to catalog which variants are imputing at high confidence and are not being detected by PennCNV [11] at the marker density in the array data.  Aim 3: Perform genome wide structural variant and single nucleotide variant associations to the eMERGE electronic medical record defined neurological and alcohol use disorder phenotypes [5-7] to establish structural variant disease association risk evidence.  Rational: The genomic risk variants we identify will be able to inform precision medicine to the best practices aimed at prediction, intervention and prevention of disease in risk prone individuals. Additionally SV have not been interrogated at this biobank scale (~100,000 individuals) for genomic risk.  Methods: I will use logistic regression techniques with the case-control EMR phenotype codes and covariates including gender and principal components. We will initially focus on neurological phenotype codes due to the large body of literature describing associations between neurological disorders and various classes of SVs [12-14]. These neurological phenotypes will include alcohol addiction, as previous studies have been able to identify common CNVs with alcohol dependence risk associations [15, 16]. In the initial release of the 83,717 imputed samples, alcohol related disorders phenotype EMR code (317) identifies 3,460 cases and 61,608 controls using the Denny et al. phenotyping algorithm [17, 18]. Given these sample counts, a genotype relative risk of 1.5, an additive model, a 1-year prevalence of 13% [19], a 5x10-8 p-value significance level, and a minor allele frequency of 0.025 there is 68.3% power to detect associated alleles. At a 0.05 minor allele frequency there is 99.7% power to detect [20]. We will also investigate other often comorbid phenotypes which are seen at larger sample sizes, the neurological disorders phenotype code (292) identifies 12,927 cases and 61,938 controls, the depression phenotype code (296.2) identifies 17,135 cases and 48,412 controls, and the attention deficit hyperactive disorder phenotype code (313.1) identifies 3,248 cases and 71,106 controls making it similarly powered to the alcohol use disorder phenotype.  Phenotypic Data: The ~1,700 EMR provides an unprecedented scope of phenotyping ascertainment to explore at this biobank scale of subject recruitment of ~100,000 and growing. Additionally there are 23 validated clinical phenotype algorithms that could benefit from having the SV genomic risks queried by others in the eMERGE Network in other manuscripts. The collective efforts of the eMERGE network have merged these clinical EMRs into a coherent set available for interrogation.  Genomic Data: The publicly available reference 1000 Genomes SV call set data used in the Sudmant et al. paper (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated\_sv\_map/) was produced by the Eichler Lab:  (https://eichlerlab.gs.washington.edu/evan.html) and the 1000 Genomes Consortium [1]. This resource should be used with efforts like eMERGE where we have amassed a large genotype biobank. As proof I can accomplish this endeavor, I have previously imputed SNVs with the 83,717 consented eMERGE Network cohort participants to the Haplotype Reference Consortium reference panel [8]. The use of genotype array data with imputation based on deeply ascertained reference panels is a mature methodology to increase the interrogatable genetic variation to scales of ~40 million variants [9, 10]. My role this last two years has been as the primary curating computational biologist of this delivered and published eMERGE network imputed genomic dataset. Currently this imputed SNV genotype set is growing and now numbers ~99,211 participating subjects and is being utilized to write 54+ additional articles with registered manuscripts concept sheets at the eMERGE data coordinating center. I hope that this imputed SV dataset serves to be similarly useful to the eMERGE Network's manuscript writing goals.  By combining this genomic data with electronic medical records using well developed biostatistics, we will produce biobank scale clinical disease phenotype genomic association variation discovery, novel variant disease risk estimations and replications. This will increase the weight of evidence useful to precision genomic medicine. We are poised with the data and ready for analysis and use. I now will use this resource to explore association with SVs and catalog clinical EMR phenotypes genomic risk. I hope this will be useful to the clinician, patient and the public in understanding our disease risks. These genomic risk estimates are the primary tool which will drive the research and investigation of the best practices for clinical decision and return of results in the arriving times of personalized genomic medicine.  References  1. Sudmant, et al., 2015, "An integrated map of structural variation in 2,504 human genomes.", Nature. 2015 Oct 1;526(7571):75-81.  2. Browning and Browning, 2009, "A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals.", The American Journal of Human Genetics, 84(2):210 { 223, 2009.  3. Browning and Browning, 2016, "Genotype imputation with millions of reference samples.", The American Journal of Human Genetics, 98(1):116-126  4. Macé, et al., 2018, "Copy Number Variation.", Methods Mol Biol. 2018; 1793:231-258  5. Newton, et al., 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, 2013.  6. McCarty, et al., 2011, The emerge network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Medical Genomics, 4:13{13, 2011.  7. Pathak, et al., 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, 2011.  8. Stanaway, et al., 2018 "The eMERGE Genotype Set of 83,717 Subjects Imputed to ~40 Million Variants Genome Wide and Association with the Herpes Zoster Medical Record Phenotype.", Genetic Epidemiology, 2018 Oct 8. doi: 10.1002/gepi.22167. [Epub ahead of print]  9. McCarthy, et al., 2016, "A reference panel of 64,976 haplotypes for genotype imputation." Nat Genet, 48(10):1279{1283, 10 2016.  10. Sayantan, et al. 2016, "Next-generation genotype imputation service and methods.", Nat Genet, 48(10):1284{1287, 10 2016.  11. Wang, et al., 2007, "PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data Genome Research.", 17:1665-1674, 2007  12. Truty, et al, 2018, "Prevalence and properties of intragenic copy-number variation in Mendelian disease genes.", Genet Med. 2018 Jun 12. doi: 10.1038/s41436-018-0033-5. [Epub ahead of print]  13. Cappuccio, et al., 2016, "New insights in the interpretation of array-CGH: autism spectrum disorder and positive family history for intellectual disability predict the detection of pathogenic variants.", Ital J Pediatr. 2016 Apr 12;42:39.  14. Howell, et al., 2013, "High resolution chromosomal microarray in undiagnosed neurological disorders.", J Paediatr Child Health. 2013 Sep;49(9):716-24.  15. Lin, et al., 2012, "Copy number variations in 6q14.1 and 5q13.2 are associated with alcohol dependence.", Alcohol Clin Exp Res. 2012 Sep;36(9):1512-8.  16. Bae, et al., 2012, "The genetic effect of copy number variations on the risk of alcoholism in a Korean population.", Alcohol Clin Exp Res. 2012 Jan;36(1):35-42. | |
| **Outline of Project** | 1. Make merged 1000 Genomes structural variant and single nucleotide variant imputation reference.  2. Impute the structural variants onto the single nucleotide variant genotype marker data.  3. Merge and QC imputed data.  4. Perform phenotype classification refinements for neurological and alcohol use disorders.  5. Perform GWAS  6. Write the paper. | |
| **Desired Data - Common Variables\***  *(Available from the CC)* | Demographics  ICD9/10 codes  CPT codes  Phecodes  BMI | Common Variable Labs  Common Variable Meds  Other: Case/Control status on Phase I and Phase II phenotypes |
| **Other Desired Data *(Available from participating sites)*** | *Please specifically list out any data elements that participating sites would collect or extract from clinical or other sources for this project (i.e. not common variables above)* | |
| **Desired Genetic Data** | eMERGE I-III Merged set (HRC imputed, GWAS)  eMERGE PGx/PGRNseq data set  eMERGEseq data set (Phase III)  eMERGE Whole Genome sequencing data set  eMERGE Exome chip data set  eMERGE Whole Exome sequencing data set  Other (not listed above): eMERGE 1 660k CNV calls and the other genotype chips CNV calls if completed before the finish of this project. | |
| **Does project pertain to an existing eMERGE Phenotype?** | Yes, if so please list  No | |
| **Planned Statistical Analyses** | 1. Phasing of merge SV-SNV imputation reference.  2. SV imputation of genotype array batches.  3. Merging and PCA based QC.  4. Phenotype refining assessments, assessment of latent phenotype codes  5. GWAS | |
| **Ethical Considerations** | None | |
| **Target Journal** | Nature | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | 12/15/2018 Complete SV Imputation  2/15/2019 Complete Merging and QC of SV data  3/15/2019 Phenotype refining and assessment  4/15/2019 GWAS of 3 phenotypes  5/15/2019 Release of SV imputation data to the eMERGE Network for others projects.  ~7/15/2019 Complete Manuscript | |

**\*Common Variables available across all datasets:**

* Demographics: sex, year of birth, decade of birth, race, ethnicity
* Codes: (repeated values & age at event): ICD, CPT, Phecodes
* BMI: (repeated value & age at event) height, weight, BMI
* Labs: (lab name, repeated lab value & age at event) Serum total cholesterol, LDL, HDL, Triglycerides, Glucose fasting/non-fasting/unknown, & White Blood Cell count
* Medications: (medication name, repeated, & age at event) Cerivastatin sodium, Rosuvastatin, Simvastatin, Fluvastatin, Pravastatin, Lovastatin, Atorvastatin, & Pitavastatin
* Other: Case/Control status on Phase I and Phase II phenotype: only on GWAS dataset participants