JOHNS HOPKINS
UNIVERSITY

Center for Inherited Disease Research
1812 Ashland Avenue / Suite 200
Baltimore, MD 21205

September 8, 2017

Memo: Update on GWAS data cleaning for CIDR Program projects

For the last several years, CIDR Program genotyping projects which utilize SNP arrays containing a genome-wide backbone have received a package of additional services. These include data cleaning service (Laurie et al., 2010), imputation to the Haplotype Reference Consortium (McCarthy et al., 2016), and assistance with dbGaP posting. There is no extra charge for these services.

The purpose of this letter is to inform you about changes in the way CIDR performs these services. Prior to fall 2017 these services were provided through an agreement with the Genetic Analysis Center at the University of Washington led by Dr. Bruce Weir and Dr. Cathy Laurie. Over the past ~ 10 years, GWAS data cleaning and imputation procedures have become standardized, in large part due to the efforts of Dr. Laurie and her group. Moving forward CIDR will provide these services under the direction of our statistical geneticist, Dr. Hua Ling and her CIDR colleagues Dr. Peng Zhang and Dr. Elizabeth Pugh. We are working closely with Dr. Laurie and her group to facilitate a smooth transition. The transfer of this process to CIDR should increase efficiency, and shorten project timelines as some steps can be done prior to the full completion and documentation of the genotyping dataset. In addition, certain steps previously performed independently by both groups will not need to be repeated.

The GWAS data cleaning process is critical to these projects as it eliminates many sources of error. The cleaning suite starts with resolving any remaining sample identity issues (gender, Mendelian inconsistencies and cryptic relatedness). Samples are also identified that should be removed for some analyses but may be retained as part of the posting to dbGaP, such as unexpected relatives. Batch effects (samples processed together, DNA source or extraction method, substudy/site) are checked and differences in ethnicity are evaluated and controlled for in analysis. Principal component analysis is used to identify ethnic outliers and to calculate eigenvectors to adjust for population stratification in association analyses. SNP filters are developed including missing data filters, duplicate and Mendelian errors, minor allele frequency and Hardy-Weinberg equilibrium. A relatively simple association ("pre-compute") analysis is performed to confirm there is not significant genomic inflation suggestive of false positives. The pre-compute also serves as a valuable quality control step for investigators who receive approval to access the dataset on dbGaP. Repeating the precompute results allows them to verify they were able to download, merge genotype and phenotype datasets and apply the filters correctly.

A QC report is prepared for posting on dbGaP which extensively documents the dataset, results and recommended filters from the data cleaning process. After the data cleaning process is complete, it will be imputed to the Haplotype Reference Consortium using a third party server and software resources (**) from the University of Michigan. As before, we will perform data integrity and quality control checks on the resulting imputed dataset and transfer it both to the Principal Investigator and to dbGaP for posting.

Please feel free to contact me with any concerns, kdoheny@jhmi.edu or 667-208-7283.

Kimberly F. Doheny

Kimberly F Doheny, Ph.D.
Co-Director, JH Genomics
Director, Center for Inherited Disease Research

1) Laurie CC, Doheny KF, Mirel DB, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. Genetic epidemiology. 2010;34(6):591-602. doi:10.1002/gepi.20516.
2) McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nature genetics. 2016;48(10):1279-1283. doi:10.1038/ng.3643.
3) ** Once imputation is complete neither the source data nor the imputed genotypes will be shared outside of the imputation server and the data is automatically destroyed after 7 days. Neither the University of Michigan nor CIDR have any rights to use your data other than to prepare it for dbGaP posting.
(see https://imputationserver.sph.umich.edu/index.html#!pages/faq for more data security details).