# Framingham Data Preparation Methods

## Expression Array Preprocessing and Normalization

We utilized the Affymetrix power tools (APT) suite to perform the preprocessing and normalization steps. First the robust multi-array analysis (RMA) protocol was applied which consists of three steps: background correction, quantile normalization, and summarization(Irizarry, 2003). The background correction step utilizes antigenomic probes that do not match known genome sequences to adjust the baseline for detection, and is applied separately to each array. Next, the normalization step utilizes a ‘sketch’ quantile normalization technique instead of a memory-intensive full quantile normalization. The benefit is a much lower memory requirement with little accuracy trade-off for large sample sets such as this one. Finally, the adjusted probe values were summarized (by the median polish method) into log-transformed expression values such that one value is derived per exon or gene. Additionally an analysis of the detection of probes above the background noise (DABG) was carried out. It provides further diagnostic information which can be used to filter out poorly performing probes and weakly expressed genes. The summarized expression values were then annotated more fully using the annotation databases contained in the huex10stprobeset.db (exon-level annotations) and huex10sttranscriptcluster.db (gene-level annotations) R packages available from bioconductor. In both cases gene annotations were provided for each feature.

## Genotype Array Pre-Phasing and Imputation

Plink was used to carry out several data wrangling and cleaning steps. The data wrangling steps included updating probe IDs, unifying data to ‘+’ strand, updating locations to GRCh37. The data cleaning steps included a step to filter for variant and subject missingness and minor alleles, one to filter variants with Hardy-Weinberg exact test, and a step to remove unusual heterozygosity. Additionally we used the HRC-check-bin tool in order to carry out data wrangling steps required to make our data compatible with the Haplotype Reference Consortium (HRC) panel (http://www.well.ox.ac.uk/~wrayner/tools/). Having been prepared thusly, the data were split by chromosome and ‘pre-phased’ with SHAPEIT using the 1000 genomes phase 3 panel and converted to vcf format. These files were then submitted to the Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html) for imputation with the HRC version 1 panel.

## References for tools used

SHAPEIT version 2.r837

O. Delaneau, J. Marchini, JF. Zagury (2012) A linear complexity phasing method for thousands of genomes. Nat Methods. 9(2):179-81. doi: 10.1038/nmeth.1785

O. Delaneau, JF. Zagury, J. Marchini (2013) Improved whole chromosome phasing for disease and population genetic studies. Nat Methods. 10(1):5-6. doi: 10.1038/nmeth.2307

O. Delaneau, B. Howie, A. Cox, J-F. Zagury, J. Marchini (2013) Haplotype estimation using sequence reads. American Journal of Human Genetics 93 (4) 787-696

J. O'Connell, D. Gurdasani, O. Delaneau, et al. (2014) A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genetics

O. Delaneau, J. Marchini, The 1000 Genomes Project Consortium (2014) Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nature Communications 5 3934

Plink version 1.90b.3u

Package : PLINK [version]

Authors : Shaun Purcell, Christopher Chang

URL : <https://www.cog-genomics.org/plink2>

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4.

Haplotype Reference Consortium panel version 1

Shane McCarthy, Sayantan Das, Warren Kretzschmar, Richard Durbin, Goncalo Abecasis, Jonathan Marchini. (2015) A reference panel of 64,976 haplotypes for genotype imputation. bioRxiv. <http://dx.doi.org/10.1101/035170>

Affymetrix Power Tools version 1.16

Irizarry,R.A. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res., 31, 15e–15.

Lockstone,H.E. (2011) Exon array data analysis using Affymetrix power tools and R statistical software. Brief. Bioinform., 12, 634–44.

Annotation DB Bioconductor Packages

MacDonald,J.W. huex10stprobeset.db: Affymetrix huex10 annotation data (chip huex10stprobeset). R package version 8.3.1.

MacDonald,J.W. huex10sttranscriptcluster.db: Affymetrix huex10 annotation data (chip huex10sttranscriptcluster). R package version 8.3.1.