UKB_WCSGAX: UK Biobank 500K Samples Genotyping Data Generation by the Affymetrix Research Services Laboratory

April, 2015

Contents

Overview

Genotyping analysis of the UK Biobank UKB_WCSGAX and UKBiLEVEAX datasets was generally done according to the Best Practices Workflow as described in the Axiom® Genotyping Solution Data Analysis Guide (see Reference 1). This document provides details regarding advanced approaches used for analysis. To generate the genotypes in these datasets, samples were extracted from blood by UK Biobank personnel. The DNA was delivered to Affymetrix in barcoded 96-well microtiter plates. The samples were then processed in the approximate order received to produce genotype data using the Affymetrix Axiom® platform with a custom-designed array described in the UK Biobank Axiom® Array Content Summary (see Reference 2). Processing was done using a LIMS system to track instrumentation, Axiom consumables arrays and reagents and operators. The process is described in the Affymetrix UKB_WCSGAX Lab Processing document (see Reference 3).

Genotype data was generated in batches of approximately 4800 individuals each. The batches were generally chosen based on data quality and/or samples for which reprocessing had already been attempted. Data for all samples from the same plate are not necessarily included in the same batch.

Rare Variants

Observation

As mentioned above, for this project genotyping cluster batches consisted of approximately 4800 individuals or the equivalent of 50 plates of 96 samples. Additional analytic measures were undertaken to ensure proper calling of SNPs with rare minor alleles for cases when the number of heterozygote individuals is relatively small at the SNP site (less than three in the 50 plate batch or less than approximately 0.06%).

Approach

Affymetrix bioinformatics scientists determined empirically that genotyping samples grouped in a single 96 sample plate (single plate genotyping) would improve rare variant detection. Consequently, the process for calling rare variants employed for data delivery was as follows:

1) Identify all probesets with <6 minor allele calls (up to 5 heterozygotes) in any single batch of ~4,800 samples. This was done using data from batches 1 - 22.

2) Identify all single sample plates with >= 80 samples.

3) Re-call the probesets from (1) on the plates from (2), in single-plate genotyping mode.

4) Adjust calls

   a) For each batch, adjust the calls for each probeset from (1) that originally had MAC < 6 and now has MAC < 11 in the single-plate results summed over the batch in question.

   b) If the 50-plate genotype call for a given loci x sample was no-call, and the corresponding genotype in single-plate genotyping call is heterozygote, the genotype call was changed from no-call to het. Otherwise, the original 50-plate genotype call was kept (i.e., keep the 50-plate call for the vast majority of probeset/sample combinations).

5) Return data

   a) 50-plate genotype calls, confidence values, and posterior model files

   b) Single-plate genotyping adjusted genotype calls: for a given batch, return the mix of 50-plate and single-plate genotype calls for the probesets from (1)

ApoE Markers

The standard process for analyzing ApoE markers is outlined the document "Analysis workflow for UK Biobank Axiom$^{®}$ Array" (Reference 4) Genotypes generated as described have been provided for all UK Biobank genotype batches.

SNP specific priors

Observation

Use of "SNP specific priors" (described below) from a sample set on which the probeset cluster is consistent with established quality thresholds allows for genotyping consistency improvements for a subset of the probesets genotyped across many batches. Axiom genotyping is executed by *AxiomGT1*, a clustering algorithm that adapts pre-positioned genotype cluster locations called *priors* to the sample data in a Bayesian step and computes three posterior cluster locations. Priors can be *generic*, meaning the same pre-positioned location is provided for every SNP, or *SNP specific*, meaning the different pre-positioned locations are provided on a SNP by SNP basis. The default workflow uses generic priors to

determine cluster boundaries during genotyping. The use of SNP specific priors can improve genotype call consistency data for a subset of markers across batches.

Approach

Affymetrix bioinformatics scientists developed a workflow to generate SNP specific priors for a subset of variants analyzed. In some cases, performance was improved by selecting SNP specific priors from one of the first 16 batches that were processed. Genotypes generated using those priors are included in the results; however, if the probeset did not appear to be providing sufficient cluster resolution and passing established QC thresholds, it was added to the exclusion list (discussed below).

Exclusion list

Observation

Since content on this array includes experimental probe designs, it was expected that some portion of them may not pass the strict QC metrics imposed. Results for all probesets not meeting QC requirements were investigated.

Approach

Probesets systematically below QC thresholds across batches were added to the exclusion list. Reasons for exclusion were primarily due to the experimental nature of specific probe designs as well as multiallelic variants which are currently not supported by Axiom analysis software.

References

1. Axiom® Genotyping Solution Data Analysis Guide

    http://media.affymetrix.com/support/downloads/manuals/axiom_genotyping_solution_analysis_guide.pdf

2. UK Biobank Axiom® Array Content Summary (design information document)

    http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UK-Biobank-Axiom-Array-Content-Summary-2014.pdf

3. Affymetrix UKB_WCSGAX Lab Processing document

    http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=155583

4. Analysis workflow for UK Biobank Axiom® Array

    http://media.affymetrix.com/support/downloads/manuals/ukbiobankarray_analysis_note.pdf