applied biosystems

UKB_WCSGAX: UK Biobank 500K Samples Genotyping Data Generation by the Affymetrix[1] Research Services Laboratory

June 2017

[1] Now part of Thermo Fisher Scientific

ThermoFisher SCIENTIFIC

Contents

Overview

This document provides details regarding the genotyping data generation by the by the Applied Biosystems™ Microarray Research Services Laboratory (formerly the Affymetrix™ Research Services Laboratory [ARSL]) for the UK Biobank genotyping project. The UK Biobank project is a prospective cohort study of approximately 500,000 samples from across the United Kingdom. The first 50,000 samples were genotyped on the Applied Biosystems™ UK BiLEVE Axiom™ Array, which will be referred to as the UK BiLEVE array in this document. The rest of the samples were genotyped Applied Biosystems™ UK Biobank Axiom™ Array (1), which will be referred to as the UK Biobank array in this document. There is a 95% overlap between the markers on the two arrays.

In order to generate the genotypes, samples were extracted from blood by UK Biobank personnel. The DNA was delivered to the Applied Biosystems Microarray Research Services Laboratory in barcoded 96-well microtiter plates. The samples were then processed in the approximate order received to produce genotype data using the Applied Biosystems™ Axiom™ platform. Processing was done using a Laboratory Information Management System (LIMS) to track instrumentation, Applied Biosystems™ Axiom™ consumables (arrays and reagents), and operators. The process is described in the UKB_WCSGAX Lab Processing document (2).

Due to the size of the study, genotype data generation was performed in batches of approximately 4,700 samples which were genotyped on approximately 50 Axiom 96-format array plates, for a total of 95 UK Biobank array batches and 11 UK BiLEVE array batches. The batches were analyzed with the Axiom biobank genotyping analysis workflow (3), which consists of two rounds of genotyping. The first round of genotyping was done according to the Axiom best practices workflow as described in the Axiom Genotyping Solution Data Analysis Guide (4). After the first round of standard genotyping, all batches were analyzed to select an exemplar batch for each probe set as the source of SNP specific prior (SSP) information. These SSPs are used by the AxiomGT1 algorithm to improve consistency and accuracy in the other batches in the second round of genotyping. Other probe set-specific modifications to the genotyping algorithm, such as changes to algorithmic parameters or advanced normalization to attenuate variation, were applied to selected probe sets as well. A final round of analysis identified probe sets that were found to be systematically problematic or redundant. All genotypes from such probe sets were excluded from the data delivery.

Axiom array terminology

A marker refers to the genetic variation at a specific genomic location in the DNA of a sample that is being assayed by the Axiom Genotyping Solution. Both SNPs and indels can be genotyped on this platform. The Axiom identifier for a marker is referred to as an *affy_snp_id*. An affy_snp_id is comprised of the prefix "Affx-" followed by an integer, for example Affx-19965213, as shown below (Figure 1). A set

of one or more probe sequences whose intensities are combined to interrogate a marker site is referred to as a *probe set*. Most Axiom™ markers are interrogated with one or two probe sets; one derived from the forward strand sequence and/or one derived from the reverse strand sequence. The Axiom identifier for a probe set is referred to as a *probeset_id*. A probeset_id is comprised of the prefix "AX-" followed by an integer, for example AX-33782819.

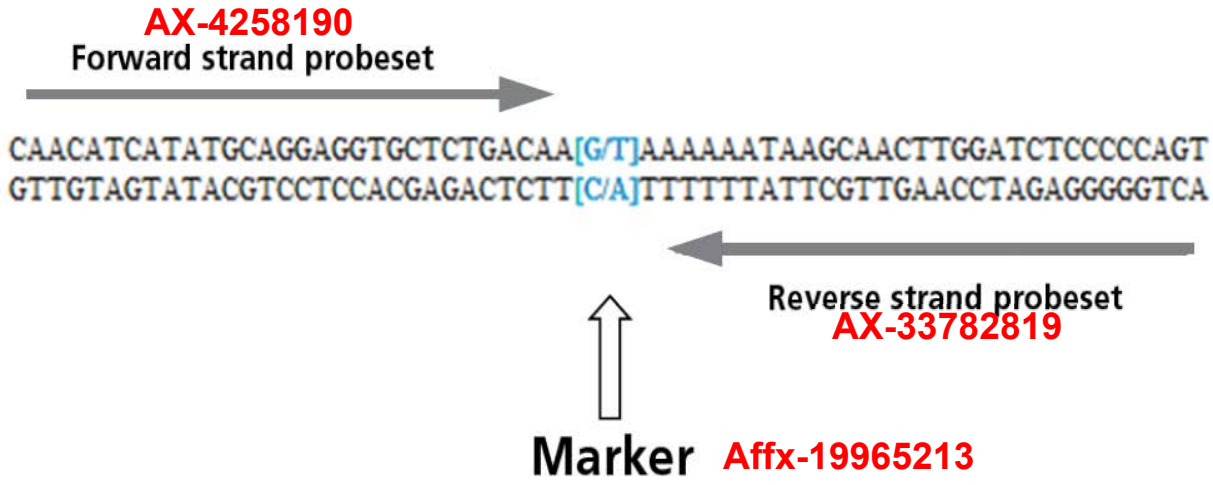

Figure 1

Steps of Axiom biobank genotyping analysis workflow

1. Divide samples into batches of approximately 4,700 samples (50 Axiom 96-format array plates). There were 95 UK Biobank array batches and 11 UK BiLEVE array batches.
2. First round of genotyping with the Axiom best practices genotyping workflow.
3. SNP specific prior (SSP) selection to create SSPs.
4. Second round of genotyping with SSPs and advanced normalization.
5. "Multi-cluster" detection to flag markers with complex genetics.

The biobank genotyping analysis workflow reduces missing information and increases allele frequency consistency across the batches (3).

First round genotyping with the Axiom best practices genotyping workflow

For each batch:
1. Sample QC to exclude poor quality samples.
2. Genotype with generic SNP priors. "SNP priors" refers to pre-positioned cluster locations that are used by the AxiomGT1 genotyping algorithm. Generic SNP priors

use the same locations across all SNPs with weak weights while SNP specific priors use different locations on a marker by marker basis.
3. SNP QC to assign probe sets to one of six quality classes (see Figure 2 below with genotype cluster plot examples for six probe sets). Three classes are recommended (denoted by arrows and shown in the upper row) and three classes are not-recommend (lower row). If a probe set is classified into one of the not-recommended classes, the genotype calls for the batch are set to missing.
See the Axiom Genotyping Solution Data Analysis Guide (4) for more detail.

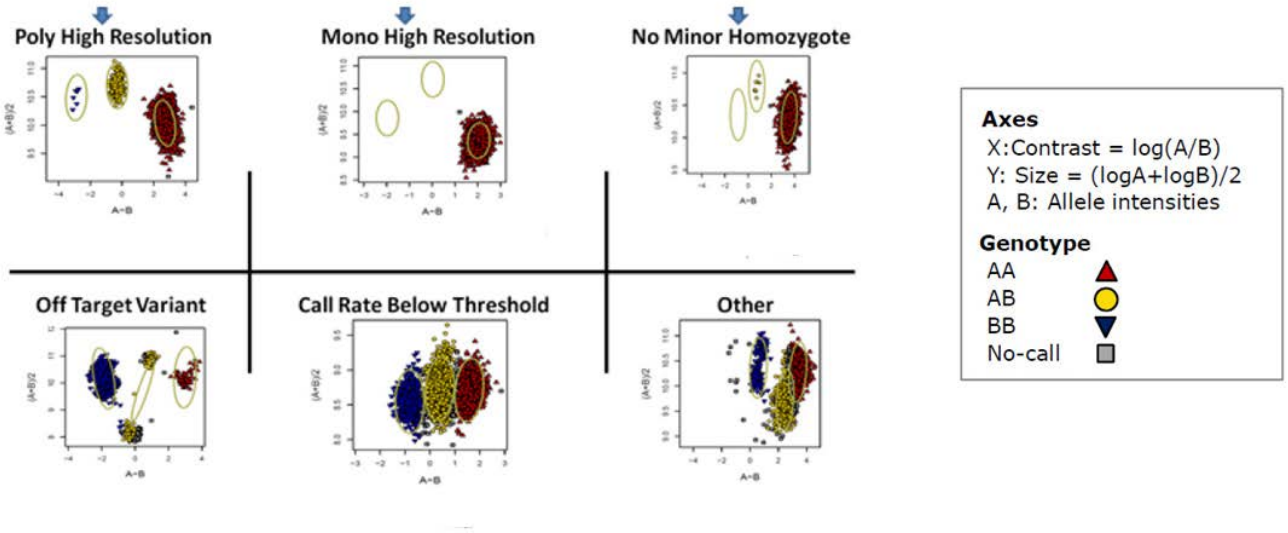

Figure 2

Characteristics of quality classes

- Poly High Resolution: Good cluster resolution and examples of both the major and minor homozygous clusters.
- Mono High Resolution: All genotyped samples are monomorphic, and the monomorphic cluster is offset from zero in Contrast (X) space.
- No Minor Homozygote: Good cluster resolution, at least one heterozygous example, and no examples of the minor homozygous cluster.
- Off Target Variant: Examples of non-hybridizing targets, potentially due to double deletions, or alternative sequences. Cluster is low in size space (Y) and near zero in Contrast space (X).
- Call Rate Below Threshold: Normal cluster properties, but call rate is below threshold.
- Other: One or more cluster properties are below threshold values.

SNP specific prior selector

The SNP specific prior selector calculates informative metrics for cluster shape and position for each probe set in each batch, and selects an exemplar batch for each probe set. If the probe set in the exemplar batch meets quality standards, then it is used as the source of SNP specific priors for that probe set. SNP specific priors are generated from the *posterior* cluster positions (cluster centers and variances that are the output of AxiomGT1 genotyping (4)) in the exemplar batch, and used to genotype the probe set in all remaining batches. This greatly reduces mis-clustering events that can cause a small percentage of probe sets to be classified as not-recommended.
An example is shown below (Figure 3). In this case, the genotype calls produced for Batch 4 with generic priors in Round 1 of analysis would be classified as "Other" (and therefore not recommended) because the homozygote cluster is incorrectly called AA (shown in red) due to the influence of an outlying sample (near the arrow). After genotyping with SNP specific priors selected from one of the correctly genotyped batches, the Batch 4 homozygote cluster is correctly called BB (shown in blue) and the results fall into a recommended class.

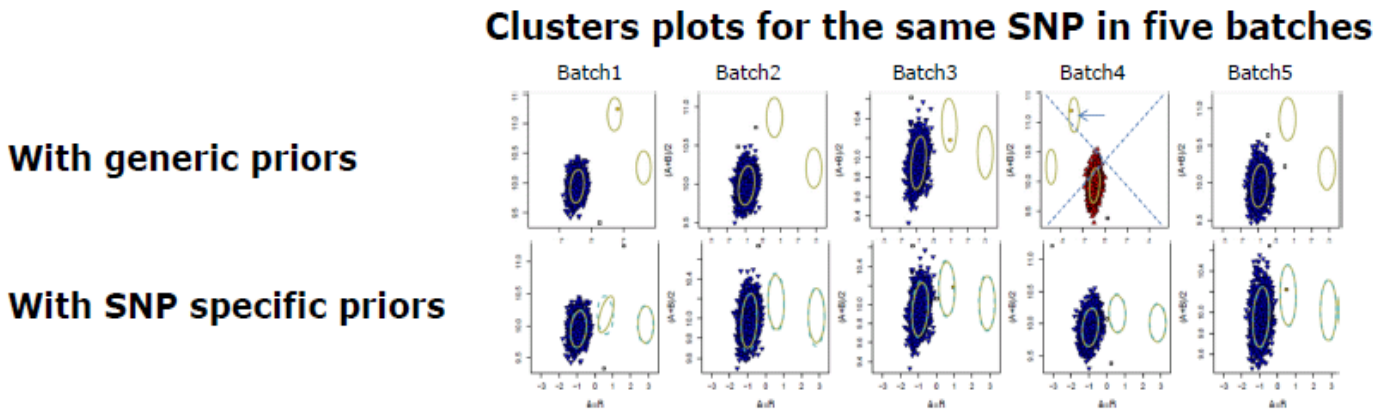

Figure 3

Advanced normalization

The AxiomGT1 genotyping algorithm adapts to signal variation that is assumed to arise from the differences in signal produced by the AA, AB, and BB genotypes. Sometimes signal variations occur due to systematic non-genetic sources. This can cause mis-clustering events resulting in classifying a small percentage of probe sets as not-recommended. One such non-genetic source is variability across the Axiom™ plates in which 96 samples are processed together. Under the assumptions of homogeneity of samples across the plates we applied a log-linear regression of signal to plate to remove irrelevant plate biases through normalization of the signals. The AxiomGT1 genotyping algorithm is executed again using the normalized signals.

Advanced normalization was not applied to all probe sets,only to those for which there was some potential for a plate bias. In particular, advanced normalization was applied to probe sets classified as not-recommended and also to probe sets flagged as having one or more outlier plates. Outlier plates are defined as having an allele frequency value that is out of line with allele frequency values produced by the rest of the plates.

An example is shown below for three SNPs that were classified as not-recommended (upper row). Advanced normalization (lower row) decreased signal variance within genotype clusters resulting in accurate assignments of genotypes by the genotyping algorithm.

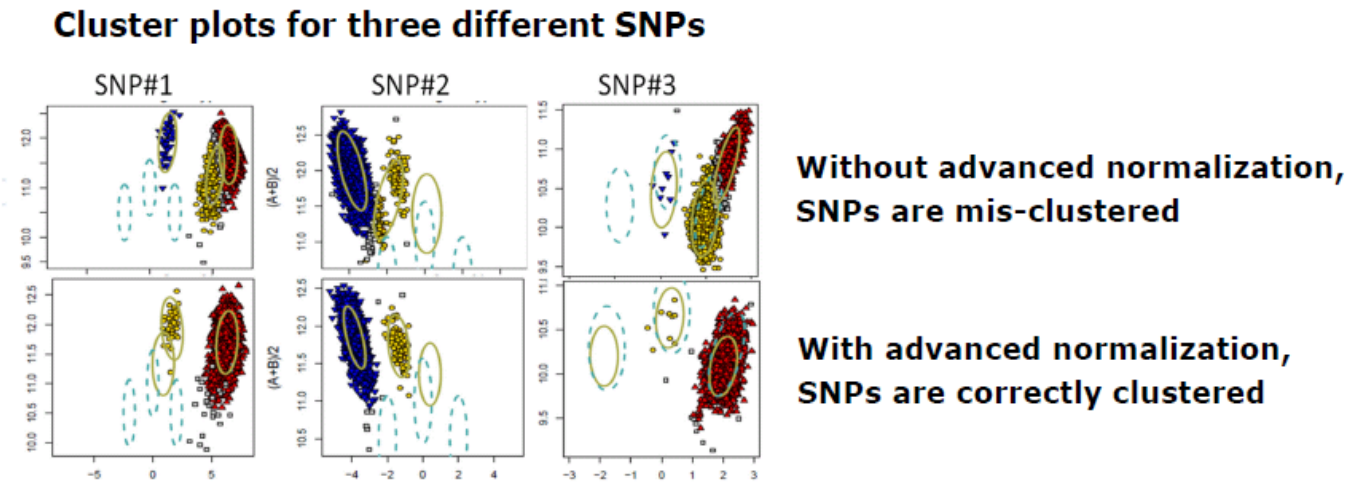

Figure 4

Multi-cluster detection

Some markers are in regions of complex genetic variation such as copy number variation regions, regions with high diversity and interfering mutations, etc. Such variation can produce more than the expected three genotyping states (AA, AB, and BB). We refer to the cluster pattern produced by probe sets interrogating such markers as *multi-cluster*. When samples are randomized across the batches, the same multi-cluster pattern is produced by the probe set across all batches as shown below (Figure 5). A multi-cluster detection program was deployed to flag probe sets producing the multi-cluster pattern. Because no genotyping algorithm exists to call such complex states, genotypes from these probe sets were excluded from the standard biobank data delivery.

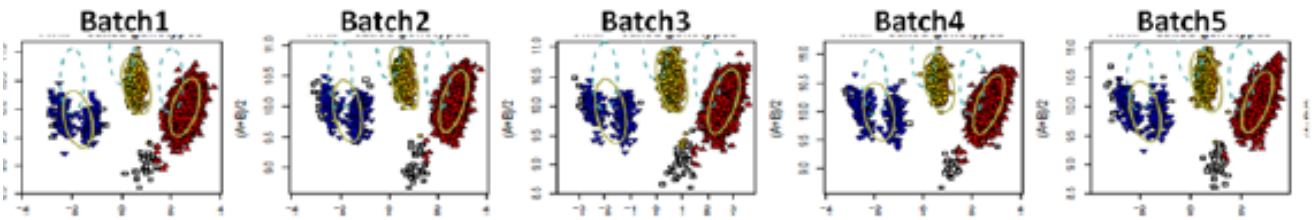

Figure 5

Probe set exclusion

After executing the Axiom biobank genotyping analysis workflow (3), a final round of analysis was executed to flag certain probe sets for exclusion. Exclusion means that all genotypes from flagged probe sets were excluded from all batches in the data delivery to the UK Biobank. Different assessment criteria were created for each of three groups of probe sets: those represented only on the UK Biobank array, those represented only on the UK BiLEVE array and those represented on both arrays.

Exclusion reasons

There are three reasons a probe set could be placed on the exclusion list:

1. The probe set is "not-working", meaning that it does not reliably resolve the genotype clusters. Such probe sets were expected because the array design includes some novel and experimental content.
2. A probe set interrogates a complex locus with more than three possible genotypes, and so cannot be genotyped with the current AxiomGT1 three-cluster genotyping algorithm. These included multi-allelic markers, meaning there were more than two alleles at the variant site, and multi-cluster markers as defined above.
3. A probe set is of one of several probe sets interrogating the same marker. In order to help ensure consistent genotyping across many batches, a single "best" probe set was selected for each marker. Genotypes produced by all other "not-best" probe sets for a marker were excluded from the data delivery. Note that the not-best probe set exclusion rule was executed so that all markers in common between UK Biobank and UK BiLEVE arrays are interrogated by the same "best" probe set for all batches for both arrays.

After executing probe set exclusion, each marker was genotyped with just one common probe set across all batches. Genotypes were delivered for 96.2% (794,409/825,928) of the UK Biobank array markers and 96.3% (778,115/807,411) of UK BiLEVE array markers.

References

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

2. Affymetrix Genotype Sample Processing. [Online] http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=155583.

3. Analysis workflow for UK Biobank AxiomTM Array. [Online] http://tools.thermofisher.com/content/sfs/manuals/ukbiobankarray_analysis_note.pdf.

4. Axiom Genotyping Solution Data Analysis Guide. [Online] http://tools.thermofisher.com/content/sfs/manuals/axiom_genotyping_solution_analysis_guide.pdf.