Overview: JH/CIDR procedures for data QC prior to release of a GWAS Raw Genotype Dataset
Matise_PAGEII_WGHumMEGA_040413_1
Initial Raw Dataset Release

Samples

We generate a SNP barcode (~96 SNPs) on every sample upon sample receipt and verify that this barcode matches the downstream released genotyping data. We verify gender and parent-offspring family relationships (if present). We identify unexpected duplicate samples and confirm the expected duplicate samples.

We attempt to predict poor performance in the GWA assay based on the performance of the sample in the SNP barcode assay and give the investigator the opportunity to replace the sample before going to the GWA array. Any other problems identified by this process are communicated directly to the study PIs.

Prior to redos, we estimated the contamination level for all samples using the VerifyBamID[1] software. The level of contamination, α, is estimated via linear regression using the BAF value and called genotypes:

$$BAF = \beta_0 + \alpha f + \tau \mathit{II}\,(g = \mathrm{AA}) + \varepsilon$$

where α is the estimated proportion of contamination (`verifyID'), β_0 is the intercept, g is the called genotype, f is the minor allele frequency, τ is expected difference in BAF between AA and BB genotypes, and II is the indicator function.

[1] G. Jun, et. al. (2012) Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. American journal of human genetics 91(5):839-848

Samples are monitored for unusual X and Y chromosome patterns. X marker heterozygosity rates and Y chromosome call rates are compared to stated gender and outliers reviewed. X and Y chromosome intensities are also reviewed.

We repeated any genomic DNA sample which resulted in less than 98% call rate one time in the lab (we may choose to repeat samples with call rates higher than this if we know or suspect a lab issue has caused the decreased data quality).

SNPs

SNP Clustering definitions:

We use Illumina's clustering algorithm (GenTrain 1.0) as implemented in GenomeStudio to cluster the data. We do not define clusters a plate at a time. For the released data set, we use a majority of the project's sample data to define the SNP cluster boundaries. The only samples excluded from this process would be those on the lower end of the quality spectrum and/or remaining redo or replacement samples which are finishing in the lab while we call the project data.

Technical Filter Specifics for your Project

We are not releasing genotypes for 'technical failure' SNPs defined as:

1. Any attempted non-Y SNP which has a call rate <85%
2. Any attempted SNP with more than two replicate errors (we include positive control replicates with our studies).
3. Any attempted SNP whose cluster means is too close together (Cluster separation value less than 0.2)
4. Any attempted SNP with a heterozygote rate greater than or equal to 80%
5. SNPs which were manually reviewed and deemed uncallable after manual review
6. Any additional criteria outlined in the SNP ReadMe document.

'Technical failure' for a SNP means that we will not pass on genotypes to the PI; we WILL release all intensity data and all raw data files for ALL attempted SNPs.

A subset of SNPs were manually reviewed and edited or dropped as appropriate for this project:

- All Y chromosome and mitochondrial SNPs were manually reviewed.
- Autosomal and X chromosome SNPs: We first filtered SNPs with less than 85% call rate, cluster separation of less than 0.2, or a heterozygote rate greater than 80%.
  The following autosomal and X chromosome SNPs were then targeted for manual review:
  - Flag possible SNPs where rare heterozygous cluster was missed by Gencall algorithm: The resulting subset of SNPs was run through the zCall pipeline to look for SNPs with possible heterozygous clusters which were not called by Illumina's clustering algorithm. The parameters used are as follows:
    - T=21 and I=0.2
  - Manual review and edited or dropped as appropriate:
    - We manually reviewed SNPs with 10 or more possible new heterozygous points as defined by zCall and a minor allele frequency less than 0.017%

    zCall: a rare variant caller for array-based genotyping: genetics and population analysis.
    Goldstein JI, Crenshaw A, Carey J, Grant GB, Maguire J, Fromer M, O'Dushlaine C, Moran JL, Chambert K, Stevens C; Swedish Schizophrenia Consortium; ARRA Autism Sequencing Consortium, Sklar P, Hultman CM, Purcell S, McCarroll SA, Sullivan PF, Daly MJ, Neale BM.

Quality Score cutoff for Genotypes

Illumina's GenomeStudio software also evaluates all genotypes using a quantitative genotype quality score called GenCall (GC) score. A GC score ranges from 0 to 1 and reflects the proximity within a cluster plot of the intensities of that genotype to the centroid of the nearest cluster. The scores range from 0 – 1 with 1 being the best. Genotypes below a threshold are set to missing (-,-) in all released data and are counted as missing data in all reports and only intensity values released. We used a 0.15 quality score cutoff for this project.

Quality Control and Monitoring

1. A small set of "barcode" SNPs (see above) are genotyped for all samples as they are received. These "barcode" genotypes are used to confirm the correct genotyping dataset is associated with the correct sample at the time of data release.
2. Study case and control samples are mixed within the plate of samples being processed together in the lab to avoid false associations due to laboratory artifacts. Samples that are within a family are maintained on the same plate if possible.

3. Positive control samples are included on each 96-well plate of study samples. At least 2 positive controls are included per plate. Different individuals are used, replicates and trios are created. These controls are used to generate replicate and Mendelian inheritance error rates error rates on a per SNP level. Concordance with HapMap and 1000Genome genotypes is also calculated.
4. Investigators supply blind duplicate samples which are placed across the study plates. This 'blind' is used to generate a study specific error rate after all final data filtering/cleaning decisions are made.

Copy Number Variant Analysis

Although CIDR does not provide copy number variant polymorphism calls as part of data release, Log R Ratio, B allele frequency and raw and normalized intensity values are reported for each sample. In addition, CIDR assesses the suitability of a sample for CNV analysis by calculating the Log R Ratio from GenomeStudio. We suggest that a sample that may have a standard deviation of Log R Ratio greater than 0.20 not be used for CNV analysis. That information can be found in the Sample Table in the Sample_Information directory.