# QUALITY CONTROL AND QUALITY ASSURANCE REPORT

FTP (Conger) Genome-Wide Association Study

|  |  |
|---|---|
| **Genotyping Platforms:** | *Illumina HumanOmni1-Quad v1*<br>*Illumina HumanOmniExpressExome v1* |
| **Number of DNA Samples Genotyped:**<br>(including duplicates and lab controls) | N = 2202 |

Table of Contents

Note: All files and scripts used for this report are located on statgen.colorado.edu in:

/home/mcqueenm/CONGER/

Of particular relevance to this report:

/home/mcqueenm/CONGER/mergeall/qc.sh

includes step by step PLINK commands that were used.

Marker Information (**Illumina HumanOmni1-Quad**)
There were a total of 1,140,419 genetic and structural variants [including single nucleotide polymorphisms (SNPs), copy-number variants (CNVs) and Insertion/Deletion variants (INDELs)] generated through the 1.0 (B) version of the Illumina *HumanOmni1-Quad*. Illumina released an updated marker list, map and annotation file [corresponding to the 1.0 (H) version] in April 2011. As of this report, this is the most recent marker information and it includes a total of 1,134,514 genetic and structural variants. This update is based upon GenomeBuild 37.1 that is the new Human Genome Reference Consortium assembly and annotation. In addition, the set of new marker information provides minor allele frequencies from the HapMap sample, which are highly validated markers. Genotypes for this project were originally generated and annotated using a previous version of the map file [version 1.0 (B)] and needed to be updated to reflect the new marker list to use.

*Marker list discrepancy (Starting=1,140,419)*
There were 5,905 markers that were included in the previous marker list [version 1.0 (B)] for which genotypes were generated for, but are not included in the updated marker list [version 1.0 (H)]. The genotypes from these markers were ignored during the initial step of reading in the raw files. The removal of those markers aligns with the total number of 1,134,514 from the new marker list (5,905 markers removed)

*Markers with unreliable map position (Starting=1,134,514)*
There were 9,837 markers that no longer have a reliable chromosome or base-pair location using the new Human Genome Reference Build (37.1). These markers were removed from the dataset leaving 1,124,677 total markers (9,837 markers removed).

*Marker name changes (Starting=1,124,677)*
There are 71,393 marker names that were changed to reflect the most recent information using the Human Genome Reference Consortium assembly (for example, 1000 genomes reference SNPs are updated to include the prefix “kgp”). These markers were *not* removed from the dataset. There remain 1,124,677 markers (0 markers removed).

*Marker position changes (Starting=1,124,677)*
Out of the 1,124,677 markers, 187 changed chromosomes from the previous map file to the current map file. In addition, 1,124,239 markers changed base pair location (as expected with an updated Human Genome Reference Build). No markers were removed at this step, however the map file was updated to reflect this new information. There remain 1,124,677 markers (0 markers removed).

*CNVs and INDELs (Starting=1,124,677)*
The genotyping platform includes structural variation such as CNVs and INDELs. While potentially useful and interesting for future work, we chose here to focus on biallelic SNPs to establish sample quality and other QC parameters. There were 120,107 CNV marker probes that were removed prior to QC steps. In addition, there were 314 INDELs removed. There remain 1,004,256 SNPs (120,421 markers removed).

*Chromosomes 1-22, X and Y (Starting=1,004,256)*
For the purposes of this project, we chose to remove the mitochondria SNPs (MT; 25 markers) as well as the pseudo-autosomal region of X (XY; 904 markers). This step resulted in a total of 1,003,327 SNPs (929 markers removed).

*Markers with HapMap and 1000 genomes allele frequencies (Starting=1,003,327)*

Markers that are included in the HapMap project and the 1000 genomes project are markers that have been extensively validated in independent samples. Again, for the QC procedures, we will focus exclusively on these markers that have been validated in these databases. Out of the 1,003,327 markers, 960,445 markers were also included in the HapMap project and/or 1000 genomes project (42,882 marker removed).

*Final marker set for QA/QC reporting (**N=959,952**)*

The final set of genotypes used to merge with the ExomeExpress (see next section) is from chromosomes 1-22 and the X chromosome (N=959,952). This set of marker information will allow us to perform various genome-wide checks in addition to checks of biological sex. In the table below we provide a breakdown of the number of markers per chromosome for both the full set and the set based on HapMap/1000 genomes overlap.

*Table: Number of SNPs per chromosome (Quad)*

| ***Chromosome*** | ***N*** | ***N (HapMap)**** |
|---|---|---|
| 1 | 83,718 | 79,326 |
| 2 | 77,107 | 74,629 |
| 3 | 62,857 | 60,661 |
| 4 | 58,800 | 56,828 |
| 5 | 56,804 | 55,101 |
| 6 | 75,749 | 72,049 |
| 7 | 52,566 | 50,326 |
| 8 | 51,787 | 50,201 |
| 9 | 46,122 | 44,352 |
| 10 | 52,474 | 50,616 |
| 11 | 49,444 | 47,381 |
| 12 | 48,157 | 46,034 |
| 13 | 34,426 | 33,569 |
| 14 | 29,997 | 29,091 |
| 15 | 29,772 | 28,416 |
| 16 | 32,269 | 30,394 |
| 17 | 28,782 | 26,900 |
| 18 | 27,681 | 26,873 |
| 19 | 23,024 | 21,319 |
| 20 | 28,277 | 26,208 |
| 21 | 14,051 | 13,528 |
| 22 | 14,397 | 13,646 |
| *Total (1-22)* | *978,261* | *937,448* |
| X | 23,845 | 22,504 |
| *Total (1-22, X)* | *1,002,106* | ***959,952*** |
| Y | 1,221 | 493 |
| *Total (1-22, X & Y)* | *1,003,327* | *960,445* |

*SNPs that have a minor allele frequency reported in the HapMap database.

Marker Information (**Illumina HumanOmniExpressExome**)
There were a total of 958,179 genetic and structural variants [including single nucleotide polymorphisms (SNPs) and Insertion/Deletion variants (INDELs)] generated through the 1.0 (B) version of the Illumina *HumanOmniExpressExome*. The genotypes were called using the most recent information available (GenomeBuild 37.1) from Illumina. There were however, in the original annotation file, SNP markers tagged with "exm" to denote that those markers are from the Exome portion of the platform. For example, SNP "rs6685064" was listed as "exm-rs6685064" in the file. In order to merge these data with data from the Omni1-Quad, the "exm" prefixes were removed.

*Chromosomes 1-22, X and Y (Starting=958,179)*
There were 1,423 markers that were either not assigned to a chromosome (chromosome="0"; 1205) or were mitochondria SNPs (MT; 25 markers). These markers were removed from subsequent steps in the QA/QC process (1,423 markers removed).

*Insertion/Deletion (Starting 956,755)*
There were 136 Insertion/Deletion markers (INDEL; 136) genotyped. For the purposes of the QA/QC process, the focus is on SNPs, therefore these markers were removed (136 markers removed).

*Duplicate Markers (Starting=956,619)*
There were 4,115 markers that were duplicated on the OmniExpressExome chip. This appeared to be largely due to markers from the Exome portion of the platform where there was overlap within the same platform. Duplicate markers were removed to create a unique set of markers (4,115 markers removed).

*Final marker set for QA/QC reporting (**N=951,034**)*

The final set of genotypes used to merge with the Omni1-Quad platform is based upon chromosomes 1-22 and the X chromosome (N=951,034). This set of marker information will allow us to perform various genome-wide checks in addition to checks of biological sex. In the table below we provide a breakdown of the number of markers per chromosome.

*Table: Number of SNPs per chromosome (ExomeExpress)*

| ***Chromosome*** | ***N*** |
|---|---|
| 1 | 82,569 |
| 2 | 73,873 |
| 3 | 60,880 |
| 4 | 49,972 |
| 5 | 52,655 |
| 6 | 61,053 |
| 7 | 48,482 |
| 8 | 45,300 |
| 9 | 42,372 |
| 10 | 47,953 |
| 11 | 51,361 |
| 12 | 47,157 |
| 13 | 31,787 |
| 14 | 30,557 |
| 15 | 29,498 |
| 16 | 32,555 |
| 17 | 32,565 |
| 18 | 25,146 |
| 19 | 29,421 |
| 20 | 24,456 |
| 21 | 21,852 |
| 22 | 15,304 |
| *Total (1-22)* | *927,768* |
| X | 23,266 |
| *Total (1-22, X)* | ***951,034*** |
| Y | 1,470 |
| *Total (1-22, X & Y)* | *952,504* |

**Merged Marker Set for QA/QC (N=582,624)**

To create a single set of markers to be used for the entire sample across both genotyping platforms, we merged the data and selected only markers that are common to both platforms, have a MAF of at least 0.01 and a call rate of at least 0.95. As noted above, this only includes markers on chromosomes 1-22 and X. Overall, 639,671 SNP markers are common to both genotyping platforms (chromosomes 1-22 and X). Of those, 582,624 have a MAF > 0.01. Therefore, the final QA/QC set of markers used to generate this document is based upon these 582,624 SNP markers.

Missing Data Rates for Individual Samples (N=2202)

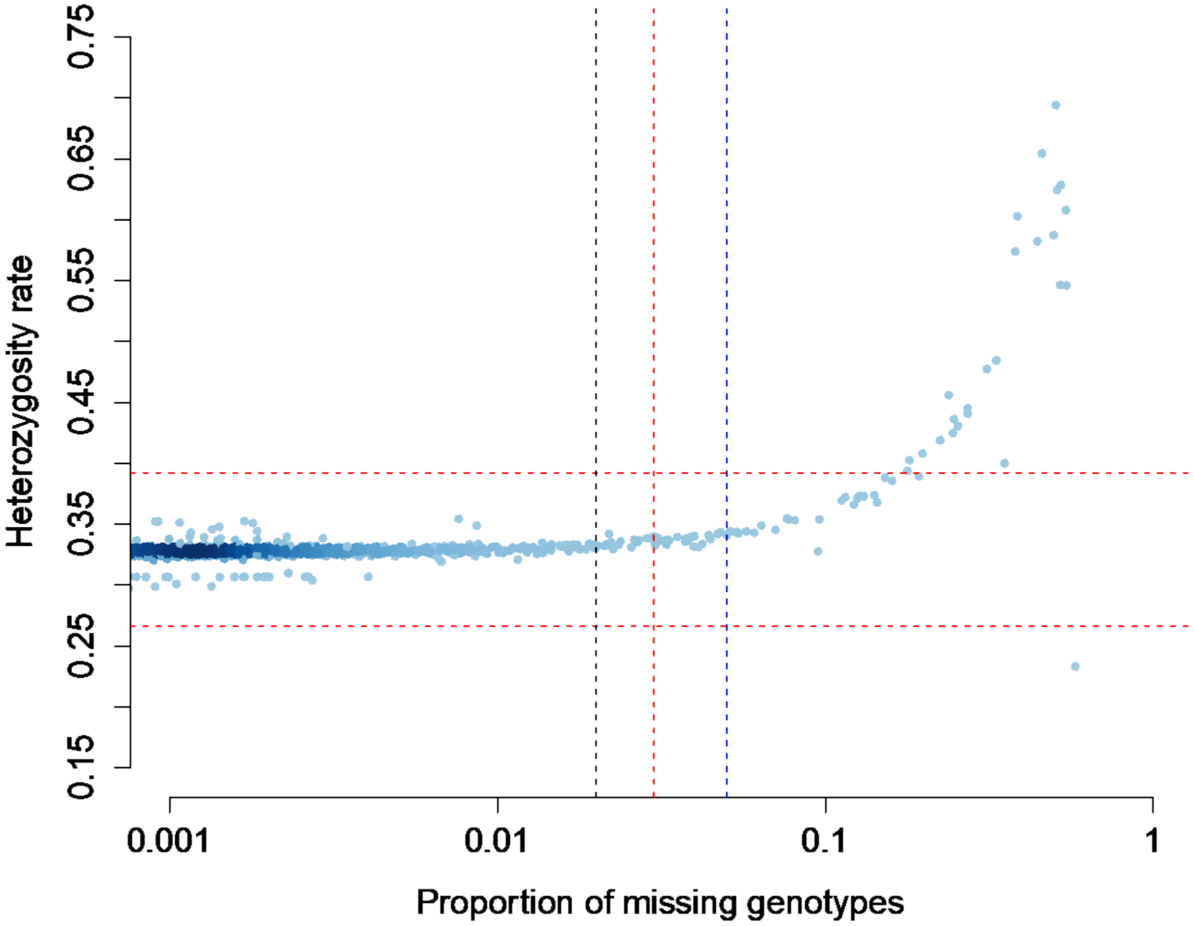
To generate missing data rates for individual samples, we focused on the 582,624 overlapping SNP markers across chromosome 1-22 and X. The average missing data rate for the 2202 samples was 0.0115 (sd=0.0739) with a range of 0.00017 to 1.00. The missing data thresholds that are often used for GWAS range from 3-5%. Further, data is often inspected for the distribution of mean heterozygosity across the autosomes. In general, samples exhibiting excess heterozygosity may be an indicator of sample contamination while less than expected heterozygosity is thought to be an indicator for inbreeding. Mean heterozygosity in this context is defined as (N-O)/O, where N is the number of nonmissing genotypes and O is the observed number of homozygous genotypes.

The plot provided (see next page) is the proportion of missing genotypes (log-scale on the x-axis) versus heterozygosity rate (y-axis). The two horizontal lines indicate 2x the standard deviation of the mean heterozygosity in this sample. The black, red and blue vertical lines represent a missing genotype threshold of 0.01, 0.03 and 0.05 respectively. Based upon the distribution of missing data it could be argued that a missing data rate of 0.03 could be reasonably adopted (red vertical line). This threshold would remove 82 samples (~4% of the total sample), which is well within the range of acceptable sample loss among traditional GWA studies.

Removal of Individual Samples Based Upon Genotyping (N=2120)

Two different criteria were used to flag individual samples for potential removal from the analysis data set. First, missing data rate > 0.03. Second, individual mean heterozygosity rate that exceeds ± 2(SD) of the mean heterozygosity rate of the entire sample. For this sample, there were no samples removed because of excess heterozygosity. However, based upon the missing data rate of > 0.03, we removed 82 samples. As noted above, this results in a sample loss of approximately 4%. Sample size used for the following QA/QC steps is N=2120. Note this number includes duplicate samples, as part of the QC/QA is to assess duplicate concordance.

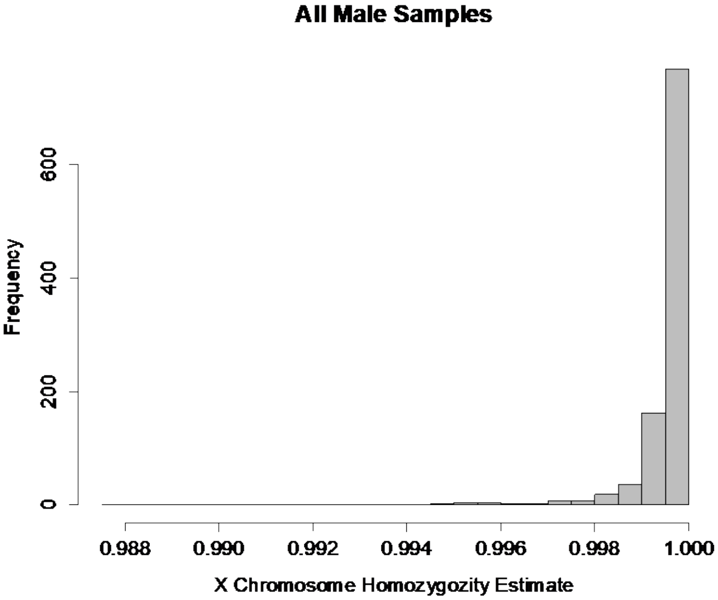
*Figure: Proportion of missing genotypes versus Heterozygosity rates among the 2202 samples*

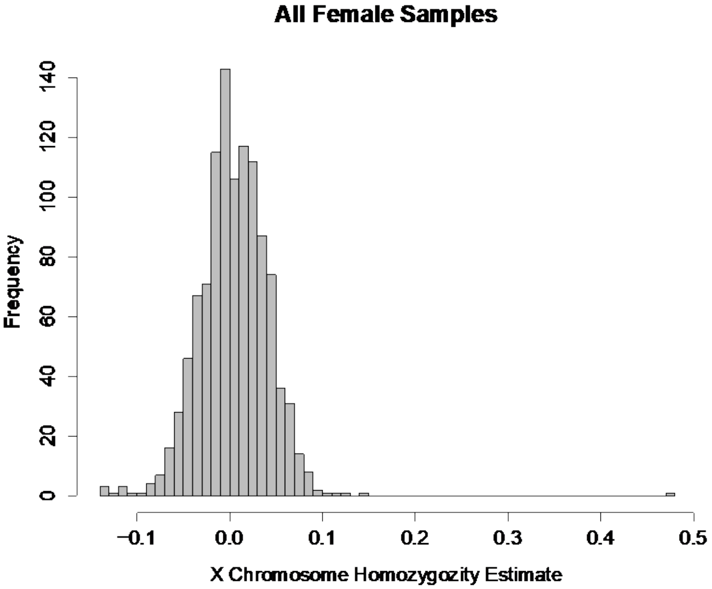


Sex Check (N=2120)

After removal of 82 samples based on missing data rates, 138 samples were “flagged” by PLINK using an inbreeding (homozygosity) estimate for the X chromosome. PLINK expects a homozygosity estimate > 0.80 for males and < 0.20 for females. Any estimates between 0.20 and 0.80 are flagged. This is considered very conservative. The following two graphs display the X chromosome homozygosity for (assumed) males and females respectively.

*Figure: X Chromosome Homozygosity*





Sex Check (cont)

The following tables provide a breakdown on the samples that passed X chromosome homozygosity checks and those that were flagged. There was only 1 subject flagged and this subject was presumed to be female. This female subject exhibits homozygosity estimates consistent with chromosomal abnormalities such as Turner's Syndrome. No samples were removed based on this step.

*Table: Samples where pedigree file sex matched X chromosome homozygosity (N=1785)*

| ***Assumed Sex*** | ***N*** | ***Average F (SD)*** |
|---|---|---|
| Male | 1022 | 0.999 (0.0009) |
| Female | 1097 | 0.004 (0.0354) |

*Table: Flagged samples defined as males with F < 0.80 and females with F > 0.20 (N=138)*

| ***Assumed Sex*** | ***N*** | ***Average F (SD)*** | ***Range*** |
|---|---|---|---|
| Male | 0 | - | - |
| Female | 1 | 0.4772 | NA |

Estimation of IBS, IBD and Kinship

Using information from chromosome 1-22, we estimated Identity by State (IBS), Identity by Descent (IBD) and Kinship Coefficient using PLINK and **K**inship-based **In**ference for **G**WAS (KING). This information is used to test duplicate concordance, confirm expected biological relationships, identify unknown or cryptic relatedness in the sample and assess genetic ancestry. The relationship measures are calculated pairwise for all individuals in the dataset. Before estimation, as generally recommended, we pruned autosomal SNPs to establish an approximately independent set of markers to be used for IBS, IBD and Kinship Coefficient. We used a linkage disequilibrium threshold ($r^2$) of 0.20 with a SNP window size of 50 and number of SNPs to shift window at each step of 5 *(PLINK command of --indep-pairwise 50 5 0.20*). After pruning, a final set of 114,201 autosomal SNP markers in approximate linkage equilibrium was used to estimate the relationship measures. Pairwise average identity-by-descent (IBD), which is a measure of how “related” two samples are, is estimated using PLINK (“PI_HAT”). However, PLINK’s estimates of IBD may be biased in stratified samples. Therefore, we also used the KING package that provides estimates of relationship that are robust to stratification (“Kinship”).

Duplicates

There were 236 known duplicate pairs that passed initial QC (see *Removal in Individual Samples Based on Genotyping* above). Pairwise mean IBD values that exceed 0.90 are thought to be either duplicate samples or MZ twins. PI_HAT values have a maximum of 1.0 (perfect concordance). The following is the table of IBD for the 236 known duplicate samples. Of the 236 duplicate pairs, 18 duplicate pairs are samples from the study while 218 pairs are laboratory controls. As can be seen in the table, concordance is high for duplicate samples.

*Table: Mean IBD for 236 Known Duplicate Samples*

| ***Sample*** | ***N (Pairs)*** | ***PLINK Average Mean IBD (Range)*** |
|---|---|---|
| *Controls* | 218 | 0.9999 (0.9995, 1.000) |
| *Conger* | 18 | 0.9999 (0.9994, 1.000) |
| *Total* | 236 | 0.9999 (0.9995, 1.000) |

Unexpected Duplicates and/or MZ twin pairs

There were 41 pairs of unexpected duplicates or MZ twin pairs according to unique identifiers provided in the data. These may be additional “blind” duplicates built into the experimental procedure, samples genotyped twice (on different chip platforms), samples mixed up during processing, MZ twins, or some combination of those possibilities. This list can be found in the “/home/mcqueenm/CONGER/mergeall/unex_dups.txt” file.

Relationship Checks

As part of the estimation of IBD, PLINK also provides probabilities of the pair of individuals being test as sharing 0, 1 or 2 alleles IBD. These probabilities are referred to as Z0, Z1 and Z2 respectively and they sum to 1. PLINK's PI_HAT is defined as [(Z1/2) + Z2]. Using the distribution of Z0, Z1 and Z2, we can examine expected relationships in terms of IBD. In addition, the Kinship Coefficient may be used to infer relationships as well (details can be found in Manichaikul et al., 2010). The following table provides the expected Z0 and Z1.

*Table: Expected Values of Z0, Z1 and the Kinship Coefficient for Different Relationships*

| **Relationship** | **E(Z0)** | **E(Z1)** | **E(Kinship)** |
|---|---|---|---|
| Monozygotic Twin (MZ) | 0.00 | 0.00 | 0.5000 |
| Parent-Offspring | 0.00 | 1.00 | 0.2500 |
| Dizygotic Twin (DZ) | 0.25 | 0.50 | 0.2500 |
| Full Sibling (FS) | 0.25 | 0.50 | 0.2500 |
| Half Sibling (HS) | 0.50 | 0.50 | 0.1250 |
| Avuncular (AV) | 0.50 | 0.50 | 0.1250 |
| Half-Avuncular | 0.75 | 0.25 | 0.0625 |
| First Cousin (CO) | 0.75 | 0.25 | 0.0625 |
| Not Related (NR) | 1.00 | 0.00 | 0.0000 |

The table below shows the breakdown of pairwise relationships for all pairs of individuals. At this point, we have made no assumption about the biological relationship. As can be seen in the table, we have detected 301 duplicate (or MZ twin) pairs, 1,805 1st degree relationships (parent-offspring or sibling pair), 452 2nd degree relationships and 116 3rd degree relationships. The vast majority of the pairwise relationships tested are not biologically related.

*Table: Observed Relationship Status*

**Observed Relationship (Kinship Coefficient)**

| Duplicate/MZ Pair | 1st Degree | 2nd Degree | 3rd Degree | Not Related | *TOTAL* |
|---|---|---|---|---|---|
| 301 | 1,805 | 452 | 116 | 2,243,466 | **1839421** |

It is often helpful to plot observed values of Z0 vs Z1 from PLINK to visualize the pattern of anticipated relationships. Assumed biological relationships can be specified for each given pair of subjects to compare and contrast what is expected and what is observed from the GWAS data.

Genetic Ancestry via Multidimensional Scaling

Using the genome-wide IBS pairwise information, we will be able to use PLINK to identify clusters of individuals based upon genetic similarity for the sample. This method, in combination with self-reported race, is often used to identify clusters of individuals based upon genetic ancestry.

## SNP Marker QA/QC

Using the final set of ‘clean’ samples from unique (excluding duplicates), we will examine the rate of missing data for the markers themselves. At the start of this step, we examined the original 582,624 SNPs from chromosomes 1-22 and X. This step is typically done after poor quality samples are removed and is often the final step prior to association analysis.

There are generally four steps that can be taken for the QC/QA of SNP markers (note that step 4 requires phenotypic information and will need to be conducted in a study-specific manner):
1) Rate of missing data for each SNP
2) Test for Hardy-Weinberg Equilibrium (HWE) for each SNP
3) Removal of SNPs with low minor allele frequency (MAF)
4) Test for different rates of missing data across a phenotypic endpoint (i.e. cases vs controls)

### *Rate of missing data for each SNP*

As is often the case, there are a small number of SNP markers for which no genotypes are available. This is typically a result of the assay failing during the experimental procedure. A 95% genotyping call rate threshold is commonly used to identify potentially problematic markers, however other thresholds such as requiring a 99% genotyping call rate may also be used.

### *Hardy-Weinberg Equilibrium (HWE)*

HWE is highly sensitive to allele frequency differences between subpopulations within a sample. Therefore, it is advisable to test deviations from HWE using a homogenous sample and furthermore, using only one subject per family. For the HWE checks we will use unrelated subjects who self-identify as Caucasian. Markers that exceed a 95% threshold for genotyping call rate will be flagged for potential deviation from HWE ($p < 0.001$). Note, that the threshold of $p < 0.001$ is relatively conservative and thresholds of 0.001 to $1 \times 10^{-7}$ have been used in other published studies.

### *Low minor allele frequency (including monomorphic markers)*

On the basis of minor allele frequency (MAF), some SNP markers may be dropped prior to analysis. However, in this situation, where we have subsets of samples being processed, as long as there is confidence in the reliability of the marker based upon other metrics, it is advisable to retain all markers, allowing independent investigators to perform this on their own.

## Final Marker Set for Analysis

If one were to adopt steps 1-3 and exclude SNP markers on the basis of a > 5% missing data rate, $p < 0.001$ from the HWE test and MAF > 1%, we will construct a reasonable set of markers we are confident in for association analyses.