GWAS Quality Control Methodology

(Approved and posted February 2013)

Genotyping quality control was performed following previously described guidelines[1, 2]. 10,503 DNA samples from COPDGene subjects, including duplicates and controls, were genotyped by Illumina (San Diego, CA) on the HumanOmniExpress array. GenomeStudio quality control, including manual review of cluster plots, was performed by Illumina. Genotype calls and intensities were exported for further quality control. Subjects were assessed for SNP missingness (>1.5%), SNP heterozygosity (6 standard deviations above the mean), chromosomal aberrations (analysis of B allele frequency), gender (X and Y chromosome intensity), and cryptic relatedness by estimated IBD (>0.125). After genotype cleaning, an additional set of subjects were excluded based for ineligibility in the primary study, including the presence of other lung disease and inadequate smoking history. An additional ten subjects were dropped because of discordance in alpha-1 genotyping and plasma protein phenotyping results.

Principal component analysis was performed to identify racial mismatches and population outliers. Autosomal SNPs present in the HapMap3 dataset with a minor allele frequency of >5% and Hardy-Weinberg P value >0.01 were pruned in plink[3] using an initial r^2 of 0.12 across a 1500 SNP window, and further pruned using an r^2 of 0.05 within a 50 SNP window. Principal components were generated using EIGENSOFT 3.0[4] and were assessed using both COPDGene and HapMap3 subjects. After removal of racial mismatches, 42 non-Hispanic whites were found to fall beyond six standard deviations on the first three principal components and were removed. No outliers were found in the African American subjects. A summary of the subject exclusions is shown in Table 1.

Markers were cleaned based on SNP concordance (<99%), missingness (>2% for allele frequency <5%, otherwise >5%), and Hardy-Weinberg equilibrium in controls ($<10^{-8}$). Markers with low minor allele frequency (<1%) were additionally excluded for the primary analysis. A summary of the excluded markers is shown in Table 2.

1. Laurie CC, Doheny KF, Mirel DB, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol;34:591-602.
2. Weale ME. Quality control for genome-wide association studies. Methods Mol Biol 2010;628:341-72.
3. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-75.
4. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904-9.

Table 1: Summary of Subject Exclusions

COPDGene samples with genotypes	10,503
Duplicates	-207
Controls	-4
>1.5% missing genotypes	-6
Gender Mismatch	-9
Unintended Duplicates & Gender Mismatch	-4
Unintended Duplicates	-37
Estimated IBD > 0.125	-104
Racial mismatches	-11
Ineligible for primary study	-101
Alpha-1 genotype mismatch	-2
Alpha-1 deficiency	-8
Outliers identified among NHW subjects	-40
COPDGene subjects passing QC	9,970

Table 2: Summary of SNP Marker Exclusions

	Non-Hispanic White	**African American**
Missingness	5,282	5,034
Concordance <99%	254	254
HWE P < 10^{-8}	1,597	2322
MAF < 0.01	76,290	19,822
Remaining SNPs	**646,125**	**701,709**