GWAS Data Generation
Details as of 26/11/2012

Mothers

<table>
<thead>
<tr>
<th>GWAS platform</th>
<th>Illumina human660W-quad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calling algorithm</td>
<td>Illumina GenomeStudio</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>QC/Before imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software used for QC filtering</td>
</tr>
<tr>
<td>SNP call rate filter</td>
</tr>
<tr>
<td>SNP HWE filter</td>
</tr>
<tr>
<td>SNP MAF filter</td>
</tr>
<tr>
<td>Individual call rate filter</td>
</tr>
<tr>
<td>Cryptic relatedness filter</td>
</tr>
<tr>
<td>Autosomal heterozygosity filter</td>
</tr>
<tr>
<td>X chromosome heterozygosity filter</td>
</tr>
<tr>
<td>Population Stratification</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Before QC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of samples</td>
</tr>
<tr>
<td>Number of SNPs</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>After QC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of samples</td>
</tr>
<tr>
<td>Number of SNPs</td>
</tr>
<tr>
<td>Imputation software and version</td>
</tr>
<tr>
<td>NCBI; HapMap CEU version for imputation</td>
</tr>
</tbody>
</table>

Centre National de Génomapping (CNG) carried out DNA genotyping on the Illumina human660W-quad array and genotypes were called with Illumina GenomeStudio. PLINK (v1.07) (Purcell et al 2007) was used to carry out quality control measures on an initial set of 10,015 subjects and 557,124 directly genotyped SNPs. SNPs were removed if they displayed more than 5% missingness or a Hardy-Weinberg equilibrium P value of less than 1.0e-06. Additionally SNPs with a minor allele frequency of less than 1% were removed. Samples were excluded if they displayed more than 5% missingness, had indeterminate X chromosome heterozygosity or extreme autosomal heterozygosity. Samples showing evidence of population stratification were identified by multidimensional scaling of genome-wide identity by state pairwise distances using the four HapMap populations as a reference, and then excluded. Cryptic relatedness was assessed using a Pi hat of more than 0.125, which is expected to correspond to roughly 12.5% alleles shared identical by descent (IBD) or relatedness at the first cousin level.

A total of 8,340 subjects and 526,688 SNPs passed these quality control filters. We imputed autosomal SNPs against the HapMap CEU population (release 22) using MaCH (v1.0.16) and NCBI build 36, HapMap 3 release 2 (Feb 2009) for the X chromosome using Minimac (v4.4.3).
A total of 9912 subjects were genotyped using the Illumina HumanHap550 quad genome-wide SNP genotyping platform by 23andMe subcontracting the Wellcome Trust Sanger Institute, Cambridge, UK and the Laboratory Corporation of America, Burlington, NC, USA. Individuals were excluded from further analysis on the basis of having incorrect gender assignments; minimal or excessive heterozygosity (<0.320 and >0.345 for the Sanger data and <0.310 and >0.330 for the LabCorp data); disproportionate levels of individual missingness (>3%); evidence of cryptic relatedness (>10% IBD) and being of non-European ancestry (as detected by a multidimensional scaling analysis seeded with HapMap 2 individuals, EIGENSTRAT analysis revealed no additional obvious population stratification and genome-wide analyses with other phenotypes indicate a low lambda). The resulting data set consisted of 8365 individuals. SNPs with a minor allele frequency of <1% and call rate of <95% were removed. Furthermore, only SNPs which passed an exact test of Hardy–Weinberg equilibrium (P>5×10^{-7}) were considered for analysis. After quality control, 8365 unrelated individuals who were genotyped at 500527 SNPs, were available for analysis. EIGENSTRAT principal components analysis was used to generate the top 100 principal components after the removal of known regions of long linkage disequilibrium in the data (Price et al. 2006; Price et al. 2008). Known autosomal variants were imputed with MACH 1.0.16 Markov Chain Haplotyping software (Li, Willer et al. 2009; Li, Willer et al. 2010), using CEPH individuals from phase 2 of the HapMap project (HG18) as a reference set (release 22). For the X chromosomal variants, imputation was performed using MiniMac (v 4.43) (Howie et al. 2012) and CEPH individuals from phase 3 of the HapMap project (HG18) were used as the reference set.
References:


