Genome-Wide Association Study on g Methodology

Christopher C. Chang, Stephen D. H. Hsu, James J. Lee,
Laurent C. A. Melchior Tellier, Rui Yang, Bowen Zhao

In their 2007 paper, the Wellcome Trust Case Control Consortium (WTCCC) outlined many methodological guidelines for the conduct of genome-wide association studies (GWAS) that were widely adopted by the human genetics community [6]. This same group has published a new GWAS that introduces or incorporates many methodological advances above and beyond those employed in its earlier paper [2]. This document will outline the precedent GWAS methodological procedures set forth by the WTCCC in a condensed format, and elaborate upon how these will be modified and applied to the CGU GWAS on g.

Contents

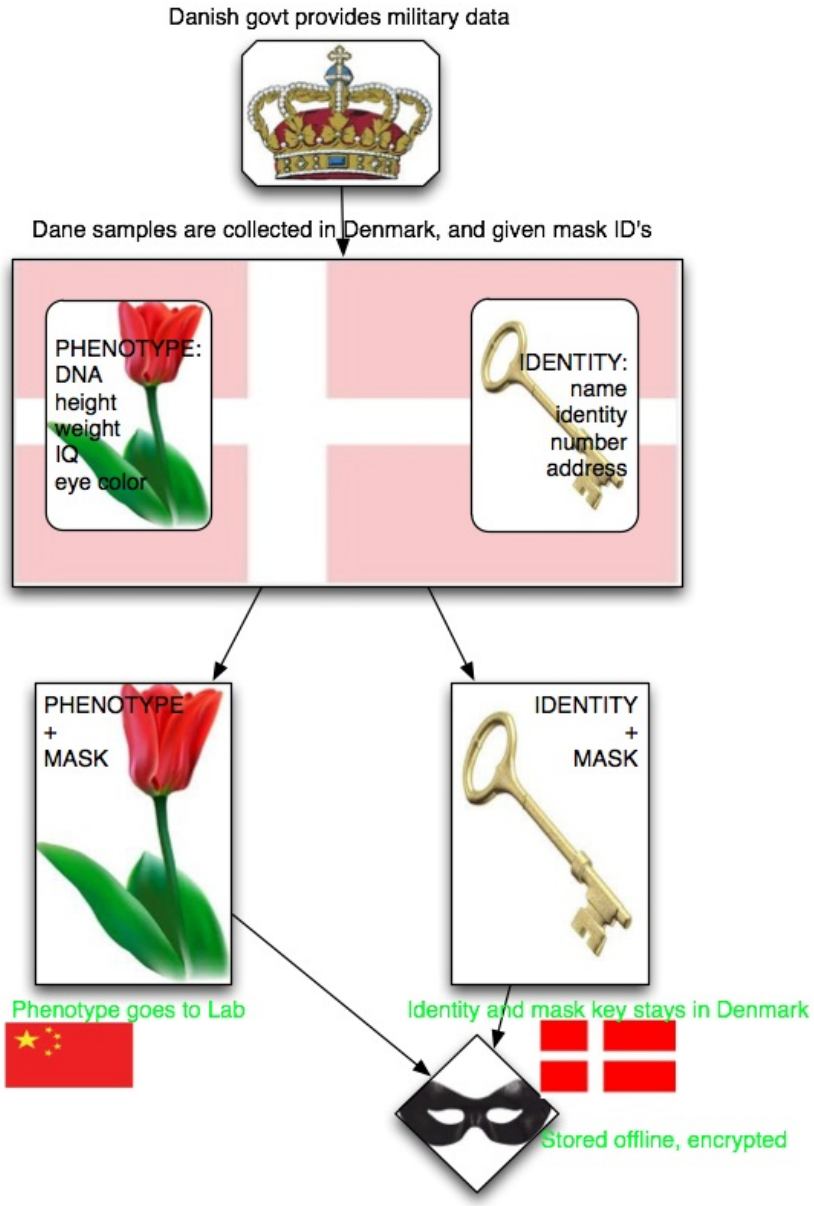

Figure 1: **Schematic of Danish sample set data security via ID masking**. Identity data are divorced from phenotype data, and cached in the country of origin, maintained under multiple layers of encryption.

1 Data Security Protocol

The first concern of the study is data security. To this end, the protocol of compartmentalized ID masking is applied. Figure 1 illustrates this protocol in the example of the Danish sample set.

Prior to shipment out of its country of origin, each sample set is subjected to procedural ID masking. Each sample has phenotypic and identifying information segregated into separate databases, reunitable only by demasking via the mask key. The database of masked identities, and the mask key, are respectively stored on two distinct and respectively encrypted disks, in the country of origin. Furthermore, the mask key itself is stored in an offline disk. Both databases are RAIDed in RAID 5 for failsafe redundancy.

2 Data Preprocessing

The application of the QC steps below led the WTCCC to exclude about 20% of their samples and 20% of their SNPs in their study of multiple sclerosis.

Sample QC can be applied to all samples at once regardless of ancestral background. Genotyping and ancestry QC should be applied separately to each of the major racial groups in the overall study.

2.1 Sample Quality Control

1. All samples genotyped specifically for the study are fingerprinted with a panel of at least 30 markers using a middle-plex genotyping platform such as Sequenom, TaqMan, iSelect, or GoldenGate. All samples failing to genotype or yielding data of inadequate quality are excluded. The X-linked middle-plex markers allow the identification of the participant's sex; any samples with a discrepancy between genotyped and self-reported sex are excluded.

2. Each sample's DNA concentration is measured in duplicate with picogreen. A sample is excluded if its concentration is found to be below 50 ngμl^{-1} or if there is more than a 10% difference in the two measurements. All samples passing Pico QC are then normalized to 50 ngμl^{-1}. Running DNA from the normalized samples on an agarose gel may reveal degraded DNA or a weak/absent band. All samples exhibiting such a defect are excluded.

3. In high-density oligonucleotide SNP arrays, hundreds of thousands of probes are arrayed on a small chip, allowing for many SNPs to be interrogated simultaneously.

Because SNP alleles only differ in one nucleotide and because it is difficult to achieve optimal hybridization conditions for all probes on the array, the target DNA has the potential to hybridize to mismatched probes. This is addressed somewhat by using several redundant probes to interrogate each SNP. Probes are designed to have the SNP site in several different locations as well as containing mismatches to the SNP allele. By comparing the hybridization quantities of the target DNA to each of these redundant probes, it is possible to determine specific homozygous and heterozygous alleles.

The samples are genotyped with the chosen SNP array and assigned a call at each SNP using an appropriate algorithm. There are several calling algorithms for each of the available SNP arrays. In its recent study of multiple sclerosis, the WTCCC used the Illuminus package, which is available upon request from its authors [5].

All samples yielding a call rate less than 92.5% are run through the process for a second time. If a sample still fails to meet the 92.5% threshold, it is run through the process for a third time. All samples that do not yield a 92.5% call rate after three attempts are excluded.

4. A signal intensity plot (or cluster plot) is a graphical representation of the results of both the genotyping and calling of a SNP. It is a scatterplot of normalized summary probe intensities; each point represents one individual. Each point is then colored to indicate how the calling algorithm decided to classify that individual—a homozygote for one of the two alleles, a heterozygote, or a null call (NA). Figure 2 shows an example of such a plot. If the genotyping is not performed in-house, then the raw intensity data must be obtained from the collaborator in order to construct this plot. The particular software that was used to call the genotypes is also necessary.

   Each sample's mean intensity difference between the channels for the two alleles (signal intensity Y - signal intensity X) is calculated across a large set of autosomal SNPs. The resulting distribution of mean differences may reveal outliers. Moreover, plotting this mean difference with respect to time of sample processing may show that the outliers tend to have been processed temporally close together. Such a pattern might arise if the samples were in the same problematic well plate. All samples with an outlying mean intensity difference are excluded. All samples from well plates with an outlier rate exceeding 50% are also excluded, regardless of their own mean intensity difference.

5. All samples showing a concordance between the SNP-array genotyping and middleplex fingerprinting below 90% are excluded.

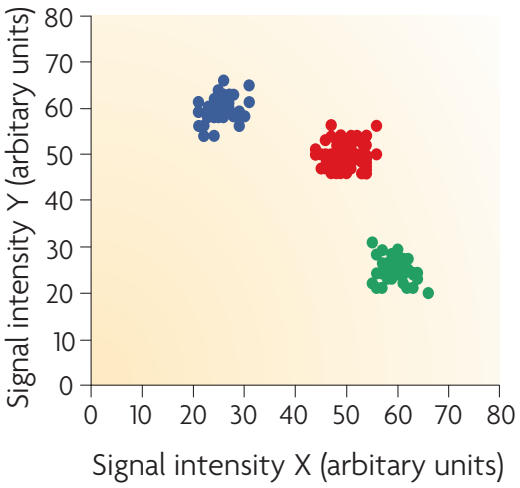

Figure 2: **An idealized signal intensity plot**. Each axis represents the intensity of target DNA hybridization to the probes for one of the two alleles. The well-defined clusters indicate that the genotyping and calling were highly accurate.

The initial fingerprinting of the samples with a middle-plex genotyping platform does not seem to be a universal practice, perhaps because of relatively high capital expenditure. However, if a suitable platform is already available, the low marginal cost of fingerprinting each sample is worthwhile. This procedure provides an inexpensive sex check and a means of eliminating samples who have been accidentally swapped or misplaced within the pipeline.

6. The WTCCC now advocates using "Bayesian cluster analysis" to identify and exclude samples whose genotyping quality is consistently poor across SNPs. Although the WTCCC's descriptions of this approach are not very detailed, it seems to amount to the following: plot the mean heterozygosity (proportion of SNPs where the participant is called as heterozygous) versus the logit of the call rate and exclude those samples that do not belong to the main distribution. Recall that earlier a hard floor on call rate was set at 92.5%.

7. The mean intensity in a single channel from the SNPs in the non-pseudo-autosomal part of the X chromosome should show a substantially higher value in females than in males. Any sample with a mean intensity at a great distance from its nearest sex-specific distribution or inconsistent with the reported sex is excluded.

2.2 Genotyping Quality Control

If the samples are not randomized or balanced with respect to phenotype in their progress through the genotyping pipeline, then differences in how their genotypes are called may be mistaken for differences in allele frequency indicative of an association with the phenotype. Therefore it is important that the SNPs retained for analysis meet a minimum threshold of genotyping accuracy.

At several points below, the arbitrary thresholds employed by the WTCCC are given. They are arbitrary because the proper balance of false positives and negatives is a matter of judgment. When the threshold is a p-value, the "right" threshold will depend on sample size. That is, if the sample size is small, then no SNPs will fail a given significance threshold; if the sample size is very large, then "too many" SNPs will fail. A reasonable procedure in our case may be to adopt the WTCCC thresholds and determine whether the percentage of SNPs excluded is similar. If the percentage is much smaller, then we should carefully consider whether there is reason to believe that our genotyping pipeline was unusually accurate. If it is much larger, then we should consider relaxing the thresholds and manually verifying genotyping quality for all putative associations.

Some of the steps below, such as filtering SNPs based on minor allele frequency (MAF) and deviations from Hardy-Weinberg equilibrium, can be performed with PLINK. Some of the other steps may have been implemented in the packages developed by the WTCCC (available at `http://www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html`).

1. All SNPs must satisfy the following criteria in order to be included in the association analysis: minor allele frequency (MAF) $> .01$, statistical information $> .90$, Hardy-Weinberg equilibrium p-value $> 10^{-50}$, and plate association p-value $> 10^{-50}$.

   I cannot find any publication that gives a clear description of "statistical information," although the WTCCC does state that this filter requires higher call rates for SNPs with lower MAF. Thus this quantity appears to be a function of (at least) call rate and MAF. In any event this quantity is calculated by the SNPTEST package, and inspection of the source code may turn out to be informative.

   Plate association indicates that there are significant differences in called allele frequency across well plates.

   The WTCCC imposed more stringent QC thresholds (statistical information $> .975$, SNP call rate $> .98$, Hardy-Weinberg equilibrium p-value $> 10^{-20}$) on those samples that it did not genotype itself. This is probably because it was not able

to carry out all of the QC measures described in this document on these external samples.

2. The samples will typically go through the genotyping and calling pipeline in distinct batches. For instance, the cases may be in one batch and the controls in another. The following algorithm ensures that each retained SNP shows a consistent placement of called genotypes across batches.

   - For each of the B batches, fit a bivariate normal distribution to each of the genotype clusters using the calls provided. We now have B sets of fitted clusters.
   - Recall the genotypes in each batch B times; each time we use the set of clusters fitted to one of the batches. The WTCCC does not state what "recalling" precisely entails; it is perhaps natural to assume that the Illuminus software can accept the mean vector and covariance matrix of a bivariate normal distribution as a fixed input. In any event it should be a simple matter to recall the genotypes by maximum likelihood or, perhaps preferably, Bayesian estimation using the appropriate HapMap or 1000 Genomes dataset to provide the prior.
   - For each batch there are now $B + 1$ sets of calls: the original calls and the calls obtained by fitting clusters to one of the B batches.
   - For each batch perform a chi-squared test for a difference in allele frequency between the original calls and the B sets of fitted clusters.
   - Exclude any SNP where any of the $B \times B$ p-values is $< 10^{-10}$.

   This algorithm excludes any SNP where the clusters are not well described by bivariate normal distributions. For example, a cluster that is split across two disjoint distributions will lead to the exclusion of the SNP. The algorithm also excludes any SNP where the difference in the position or size of the clusters between batches is sufficient to substantially alter the estimated allele frequency.

   Note that this algorithm does not seem to have been implemented in freely available software.

3. It is possible that the batches will come from different ethnic groups or countries. In an attempt to identify SNPs with misspecified genotypes in one or more of these batches, we compare estimated allele frequencies across batches with that expected under a beta-binomial evolutionary model. Although the WTCCC's description of

this procedure is unclear and possibly contains some typos, the procedure appears to be as follows.

Suppose that we have genotyped L SNPs across B control batches. Let p_{ij} be the frequency of the counted allele at SNP j in the population contributing batch i. We assume that p_{ij} follows a beta distribution with mean (μ_j) representing the allele frequency in the ancestral population common to all batches and variance (σ_i^2) proportional to the amount of genetic drift accumulated since the breakup of the ancestral population. Given the relations $\alpha_{ij} = (\mu_j^2 - \mu_j^3 - \mu_j\sigma_i^2)/\sigma_i^2$ and $\beta_{ij} = [(\mu_j - 1)(\sigma_i^2 + \mu_j^2 - \mu_j)]/\sigma_i^2$, we can write the probability density function of p_{ij} as

$$f(p_{ij} \,|\, \alpha_{ij}, \beta_{ij}) = \frac{\Gamma(\alpha_{ij} + \beta_{ij})}{\Gamma(\alpha_{ij})\Gamma(\beta_{ij})} p_{ij}^{\alpha_{ij}-1}(1 - p_{ij})^{\beta_{ij}-1}. \tag{1}$$

Let x_{ij} be the allele count at SNP j in batch i and n_{ij} the total number of chromosomes. For autosomal loci n_{ij} will be twice the sample size after previous QC steps. We assume that the count follows the mixture distribution

$$\begin{aligned} P(x_{ij} \,|\, p_{ij}, r_{ij}, e_{ij}) = (1 - e_{ij})\binom{n_{ij}}{x_{ij}} p_{ij}^{x_{ij}}(1 - p_{ij})^{n_{ij}-x_{ij}} \\ + e_{ij}\binom{n_{ij}}{x_{ij}} r_{ij}^{x_{ij}}(1 - r_{ij})^{n_{ij}-x_{ij}}. \end{aligned} \tag{2}$$

One distribution is binomial with probability of success equal to a draw from the batch's beta distribution; the other is binomial with probability of success equal to the parameter r_{ij}. The second component of the mixture distribution is meant to represent genotyping errors or an unusual evolutionary trajectory in the population from which the batch was taken.

We can estimate the parameters of this model using Markov chain Monte Carlo (MCMC). The WTCCC does not provide a software implementation of this MCMC procedure or any low-level details of its own analysis. In order to create such an implementation, prior distributions need to be assigned to each parameter. The following priors seems reasonable:

$$\begin{aligned} \mu_j &\sim \text{Unif}(0, 1), \\ \sigma_i^2 &\sim \text{Exp}(100), \\ r_{ij} &\sim \text{Unif}(0, 1), \\ e_{ij} &\sim \text{Exp}(\lambda_i), \\ \lambda_i &\sim \text{Normal}(20, 2). \end{aligned} \tag{3}$$

If the SNPs were a random sample of all polymorphic sites in the human genome, then a natural prior for the μ_j would be proportional to $\mu_j^{-1}(1-\mu_j)^{-1}$. However, because of ascertainment bias in SNP selection, a uniform distribution may be more appropriate. We can thus write the posterior probability of the model parameters as

$$f(\boldsymbol{\mu}, \boldsymbol{\sigma^2}, \mathbf{R}, \mathbf{E}, \boldsymbol{\lambda} \,|\, \mathbf{X}) \propto \prod_{i=1}^{B} f(\lambda_i) f(\sigma_i^2) \times \prod_{j=1}^{L} f(\mu_j) f(r_{ij}) f(e_{ij}) f(p_{ij} \,|\, \mu_j, \sigma_i^2) P(x_{ij} \,|\, p_{ij}, r_{ij}, e_{ij}). \quad (4)$$

The parameters of primary interest are the entries of the matrix $\mathbf{E}$ (the e_{ij}). Each such entry can be interpreted as the probability that the given batch-SNP combination is an "error." The WTCCC treated a posterior expectation exceeding .10 as reason to discard the SNP. Whether this is an appropriate threshold, however, may be sensitive to the prior specification. It will be necessary to ensure that most SNPs flagged by the chosen threshold show some anomaly upon further scrutiny.

Given the vast number of parameters that must be estimated, it will be important to optimize the MCMC algorithm. The WTCCC states that their implementation is analogous to that used by Falush, Stephens, and Pritchard [1], which may be worth studying in detail.

2.3 Ancestry Quality Control

1. It is customary to remove samples in such a way that no remaining sample is closely related to any other. Within a set of close relatives, the sample with the higher genotyping call rate is usually preferred.

   The WTCCC uses a hidden Markov model (HMM) to infer relatedness, but its description of this approach is rather sparse. I suggest the following alternative. Bin the samples into groups of relatively homogenous self-reported ancestry (e.g., all Northwest Europeans, all half-Jews). Let p_i is the frequency in one group of the designated allele at locus i, and let X_{ij} be the number of these alleles carried by individual j. Then $Z_{ij} = X_{ij} - 2p_i/\sqrt{2p_i(1-p_i)}$ is the standardized allele count of individual j. A genome-wide estimate of relatedness between individuals j and

k is then given by

$$R_{jk} = \frac{1}{L} \sum_{i=1}^{L} z_{ij} z_{ik}. \tag{5}$$

Within a set of individuals characterized by pairwise values of $R_{jk} > .05$, retain the individual with the highest genotyping call rate. If the set of related individuals are all cases and their call rates are comparable, then retain the individual judged to be the most phenotypically extreme.

After applying this analysis to each group separately, apply it to all samples in order to detect relatedness among individuals self-reporting different ancestries. In this global analysis, employ a threshold of $R_{jk} > .25$.

2. Principal components analysis (PCA) is applied to remove any samples with substantially deviant ancestry. The input SNPs for PCA will be those surviving all earlier QC steps. These SNPs are further filtered to retain only those with non-complementary alleles (e.g., G/T) and MAF > .05, minimize the correlation between markers due to linkage disequilibrium (LD), exclude the *MHC* region, and exclude SNPs in regions with unusually high loadings. The filtering by MAF and LD can be performed conveniently with PLINK. The final subset of SNPs for PCA should contain at least 100,000 markers.

   Using this subset, project the samples onto the top principal components from the PCA of the HapMap data. The WTCCC uses the program SHELLFISH for this purpose. Most genetic applications of PCA, however, rely on the EIGENSOFT package. The relative merits of these two programs are unclear. I recommend trying EIGENSOFT first since it is being continuously maintained.

   Exclude all samples who are outliers from the relevant ancestral cluster. Further consideration should also be given to individuals whose inferred ancestry appears inconsistent with self-report.

3. ADMIXTURE is now the preferred program for estimating the proportion of an individual's ancestry originating from a given source population. Figure 3 shows the output of a typical application.

   Perform an ADMIXTURE analysis of the samples together with the representatives of the ten HapMap Phase3 populations. Those SNPs surviving earlier QC steps and overlapping with those genotyped in HapMap3 are used in this analysis. Some experimentation may be necessary to decide on a reasonable number of source populations; it may help to perform the analysis with and without the

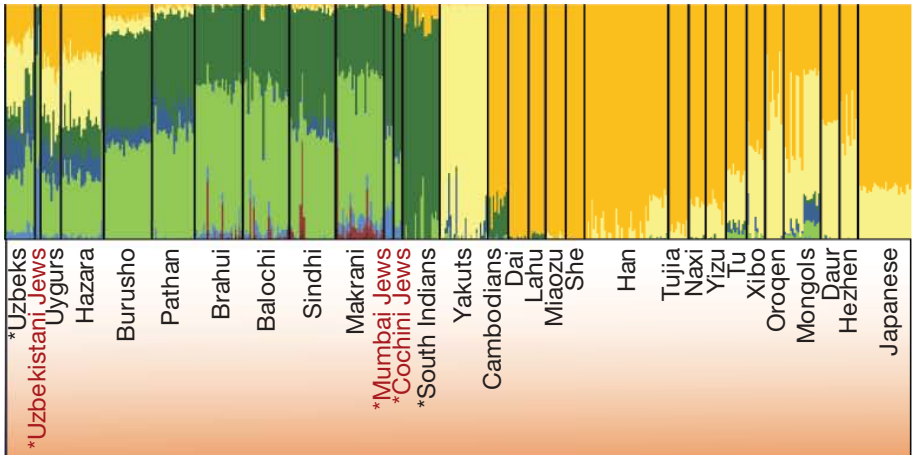

Figure 3: **An admixture plot**. Each individual is represented by a vertical (100%) stacked column of ancestry components shown in color. Note that Han Chinese and their neighbors share a gold component that is mostly absent in the Middle Eastern and South Asian population. Similarly, the Middle Eastern and South Asian populations share dark-green and light-green components that are mostly absent in East Asians.

samples reporting Jewish ancestry. Exclude any sample with African (African American, Luhya, Masai, Yoruba) ancestry in excess of .05.

3 Association Analysis

Here we do not address substantive followup analyses: searching for secondary signals, testing for the enrichment of particular biological pathways, and the like. Our sole focus is a first-pass analysis of primary signals.

3.1 Analysis of Complete SNP Set

1. Use the IMPUTE2 program to impute genotypes at sites that have not been directly assayed or have been excluded as part of QC. The 1000 Genomes Project probably provides the most extensive reference panel, and the WTCCC currently recommends using the interim haplotype release from Phase 1.

   Neither the 1000 Genomes Project nor HapMap 3 seems to contain an Ashkenazi Jewish population. Imputation quality for Ashkenazi Jews should be carefully

monitored.

It is worth pointing out that there exist several programs for genotype imputation, many of which are in common use. Marchini and Howie [4] provide a detailed comparison of these programs.

IMPUTE2 and other imputations programs provide for each imputed SNP an information measure ranging between 0 and 1 indicating the quality of imputation. Marchini and Howie do not recommend excluding imputed SNPs with information measures falling below some threshold; they advocate using a statistical model for association testing that accounts for genotype uncertainty. Such a model is conveniently built into SNPTEST, but at the present time both PLINK and EMMAX lack this feature. It therefore seems prudent to exclude all imputed SNPs with an information measure below .80.

2. In each distinct racial group, test each genotyped and imputed SNP for association with case-control status using the program EMMAX. The authors of this program suggest that the linear regression model applied to the binary case-control outcome should produce acceptable results [3].

3. Use the meta-analysis program META to combine the association results across racial groups.

3.2 Analysis of Individual SNPs

At this point we have a p-value associated with each SNP. All SNPs with p-values below 5×10^{-8} are taken forward for the post-association QC measures described below.

1. A regional association plot depicts the genomic region surrounding a specified variant. The x-axis is centered around the target variant and extends an equal number of base pairs in both directions. The y-axis indicates the $-\log_{10}(p\text{-value})$ of the genotyped and imputed SNPs in the region. See Figure 4 for an example.

   For each putatively associated SNP, use the web application LocusZoom to create regional association plots extending 200 kb on each side of the SNP. Create separate plots for each racial group. As in Figure 4, the surrounding SNPs in moderate to strong LD with the focal SNP should also show low p-values. If this is not observed, check the local recombination rate and the r^2 estimates for evidence of unusually rapid LD decay in this region. In the absence of such evidence, a failure of nearby SNPs to show signal in one or more of the racial groups is a sign of a genotyping artifact and should lead to the exclusion of the focal SNP.

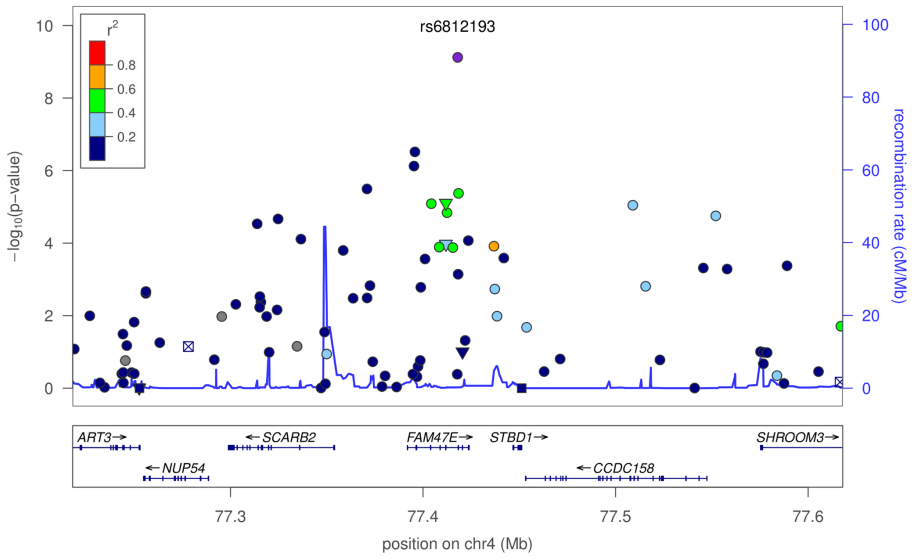

Figure 4: **A regional association plot**. An example LocusZoom plot showing the HDL cholesterol-associated region near the *MMAB* gene.

2. The output of META includes a p-value for Cochran's test of effect heterogeneity across datasets. Form a Q-Q plot of these p-values and look for a significant deviation from a straight line. Exclude any outlier SNP that shows $I^2 > .50$ and a reversal of the effect sign across the separately analyzed groups.

References

[1] D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587, 2003.

[2] International Multiple Sclerosis Genetics Consortium and Wellcome Trust Case Control Consortium 2. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476:214–219, 2011.

[3] H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S.-Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42:348–354, 2010.

[4] J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11:499–511, 2010.

[5] Y. Y. Teo, M. Inouye, K. S. Small, R. Gwilliam, P. Deloukas, D. P. Kwiatkowski, and T. G. Clark. A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics*, 23:2741–2746, 2007.

[6] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–683, 2007.