BMC
Genetics

**METHODOLOGY ARTICLE**                                        **Open Access**

# Comparison of multimarker logistic regression models, with application to a genomewide scan of schizophrenia

James MS Wason[1*], Frank Dudbridge[2]

**Abstract**

**Background:** Genome-wide association studies (GWAS) are a widely used study design for detecting genetic causes of complex diseases. Current studies provide good coverage of common causal SNPs, but not rare ones. A popular method to detect rare causal variants is haplotype testing. A disadvantage of this approach is that many parameters are estimated simultaneously, which can mean a loss of power and slower fitting to large datasets. Haplotype testing effectively tests both the allele frequencies and the linkage disequilibrium (LD) structure of the data. LD has previously been shown to be mostly attributable to LD between adjacent SNPs. We propose a generalised linear model (GLM) which models the effects of each SNP in a region as well as the statistical interactions between adjacent pairs. This is compared to two other commonly used multimarker GLMs: one with a main-effect parameter for each SNP; one with a parameter for each haplotype.

**Results:** We show the haplotype model has higher power for rare untyped causal SNPs, the main-effects model has higher power for common untyped causal SNPs, and the proposed model generally has power in between the two others. We show that the relative power of the three methods is dependent on the number of marker haplotypes the causal allele is present on, which depends on the age of the mutation. Except in the case of a common causal variant in high LD with markers, all three multimarker models are superior in power to single-SNP tests.

Including the adjacent statistical interactions results in lower inflation in test statistics when a realistic level of population stratification is present in a dataset.

Using the multimarker models, we analyse data from the Molecular Genetics of Schizophrenia study. The multimarker models find potential associations that are not found by single-SNP tests. However, multimarker models also require stricter control of data quality since biases can have a larger inflationary effect on multimarker test statistics than on single-SNP test statistics.

**Conclusions:** Analysing a GWAS with multimarker models can yield candidate regions which may contain rare untyped causal variants. This is useful for increasing prior odds of association in future whole-genome sequence analyses.

## Background

Genome-wide association studies (GWAS) have provided a wealth of information on associated genetic variants for complex human diseases. To date, GWAS have found over 2,700 new causal variants [1]. Despite the number of variants found, significant amounts of genetic variation remain unexplained. The remaining

variation is unlikely to be due solely to remaining undetected common variants.

Maher [2] proposes several other possible causes for the remaining variation: rare variants, copy number variants, gene-gene interaction, and gene-environment interaction. Rare variants are thought to be important in many diseases, and especially so in Schizophrenia [3].

Current GWAS are generally well-powered to detect common variants with modest effect sizes, but underpowered to detect rare variants, even with larger effect

* Correspondence: james.wason@mrc-bsu.cam.ac.uk
[1]MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 0SR, UK
Full list of author information is available at the end of the article

sizes. Methods such as imputation can add significant power to detect rarer variants [4], but are limited by the size of the reference panel used. The 1000 genomes project [5] will add significant power to detect rare variants, but will still not reliably cover every variant with minor allele frequency (MAF) ≤ 0.01 in the genome.

An exciting recent development is the increasing feasibility of performing a full sequence analysis, which allows the testing of every variant in the genome. Costs have fallen several orders of magnitude in the last few years, meaning the concept of a large scale case-control study using the full sequences of participants will soon be feasible.

New technology generally introduces new challenges, and sequence analysis is no exception. The sheer number of variables in such a study will mean problems in storage and analysis. The prior odds for a specific variant being associated are so low that a highly significant p-value threshold is needed to give confidence that a seemingly statistically significant result is a true positive [6,7]. Having a set of variants with higher prior odds before a sequence analysis begins would allow more relaxed p-value thresholds to be applied without resulting in large numbers of false positives. Prior odds for association can be improved for variants within a region with prior functional or positional evidence for association.

Haplotype analysis has previously been proposed to improve power to detect variants that are not typed on a genotyping chip. Models which consider the effects of marker haplotypes directly have been dismissed as not being as powerful as a model which just considers the main-effects of several nearby SNPs (referred to here as the main-effects model) [8,9]. It has also been claimed to take too long to apply to a genome-wide study [10], although computational power has since increased considerably. In this article, we discuss logistic regression models which include haplotype information with different levels of complexity in order to improve power to detect rare variants and time taken to a large scale GWAS. A region that shows evidence for association, but contains no associated common variants should have a larger prior odds for association with rare variants.

Kim et al. [11] investigated the structure of LD and how it can be partitioned into different orders. The authors showed that, for individuals of European descent, an average of around 65% of total LD in HapMap ENCODE regions is attributable to LD between adjacent SNPs. This finding led us to consider a model, intermediate in complexity between the main-effects model and haplotype testing, that would include LD information parsimoniously. This model, denoted as the main

+adj model, includes the main-effects and statistical interactions between adjacent SNPs. The number of parameters is generally low compared to the haplotype model, but should explain on average a large proportion of the haplotype structure.

The main+adj model was considered by Humphreys and Iles [12] amongst several other codings, and performed fairly well. It was only applied to one dataset (the GAW 14 dataset) so its strengths and weaknesses were not thoroughly examined.

In this paper, we assess the main+adj logistic regression model, and compare it to the main-effects model and haplotype model. We show first that for realistic simulated data all three investigated multimarker models provide higher power to detect untyped variants than single-SNP tests, except in the case of a common variant in high LD with nearby markers. We show that the models which include more haplotype information (main+adj and haplotype models) are more powerful than the main-effects model for rare untyped variants when the causal allele is on a low number of marker haplotypes, which will be the case for more recent mutations [13]. The main-effects model generally performs best when applied to a common causal variant, but the main+adj model does not lose much power in this situation.

We also examine inflation in test statistics under the null when population stratification is present. Using a simple stratification model, we show that including the adjacent interactions results in lower susceptibility of bias due to population stratification. This is a desirable property for real GWAS data with a reduced need for methods such as principal component analysis [14] or use of family data, which is robust to population stratification, but has also been shown to be less cost-effective than case-control data [15].

We then apply the models to the Molecular Genetics of Schizophrenia GWAS of Schizophrenia [16] and show that several regions display evidence of association that is not shown by common variants within the same regions. We discuss some issues with quality control that arise from using multimarker models. Our conclusion is that multimarker models can yield candidate regions not found from imputation which may be fruitful to search with a full sequence analysis.

## Results and Discussion
### Simulated association data
#### *Relative performance of multimarker models*
To assess the relative performance of the three multimarker models, we applied them to realistic simulated data generated from COSI [17] using the supplied parameter values typical of European populations. For this

section, we assume that each observation consists of one haplotype. Later, we consider how to apply these models when the haplotype phase is unknown.

Three different causal SNP MAFs were considered - 0.005, 0.025, and 0.25. Two different methods for picking a causal SNP were considered, one which corresponds to the causal allele being on a moderate number of marker haplotypes, and the other corresponding to the causal allele being on one. More details about the simulation study are given in Methods.

Figures 1, 2 and 3 show power curves for each of the combinations of causal SNP MAFs and causal SNP models. Each line consists of the average power from 200 sets of six marker SNPs, each of which has the power of detecting association estimated from 1000 simulated phenotype vectors (see methods). We grouped sets by similar mean $|r|$ between markers and causal SNP. The median mean $|r|$ value for the group and the mean power over the grouping is plotted.

We first note that the haplotype model performs well when the MAF of the causal SNP is low. It clearly has the highest power for both disease models for MAF 0.005. Additionally it performs best for the causal haplotype case when MAF = 0.025. Previously, Chapman et al. [8] and Clayton et al. [9] indicated that haplotype coding is less powerful than the main-effects coding in most situations; this does not contradict our results, because that work considered causal SNPs with MAF ≥ 0.05.
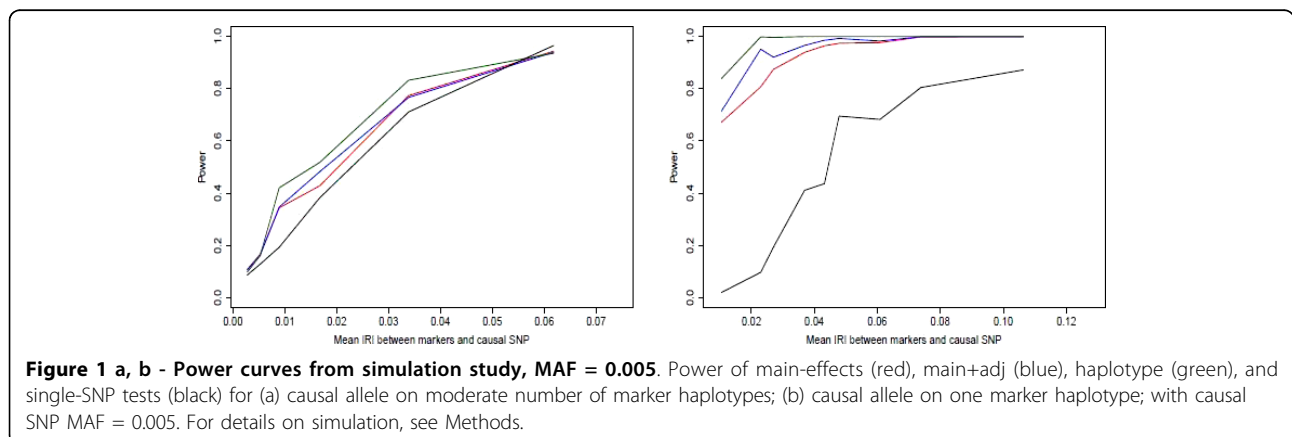
Secondly, for the disease driven by an untyped causal SNP, adding the adjacent interactions to the main-effects generally loses a small amount of power compared to the main-effects only for MAF ≥ 0.025. For the very rare causal SNP case, the main+adj model performs slightly better. When the disease is associated with a SNP on one marker haplotype, adding the adjacent interactions gives a clear power gain for the two rare causal SNP cases, but a small loss for the common SNP case.
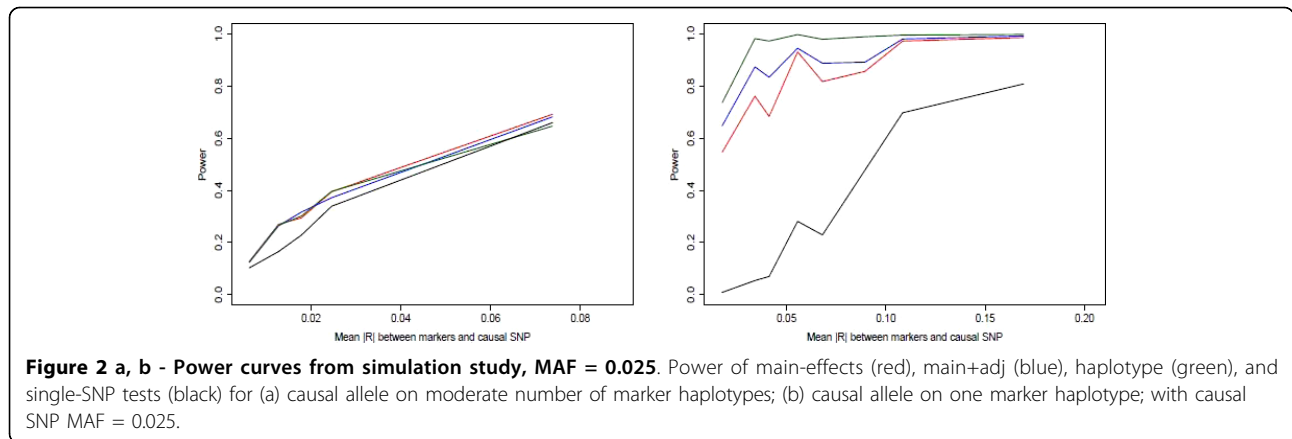
A third interesting result was the big difference between using the multimarker tests and single-SNP tests. The single-SNP test generally performed badly compared to the multimarker tests. The only case in which it provided a power boost was when the disease was generated from a common SNP. Even in this case, the increase in power was present only when the average $|r|$ between markers and causal SNP was high, i.e. almost equivalent to testing the causal SNP directly, which is the ideal situation for single-SNP tests. We had expected that the main+adj model would generally have power in between the main-effects and haplotype model, or in other words it would be a good compromise model to use when the disease model is unknown. This appears to be the case from these simulations, with its power generally close to the power of the best performing model.

The results are not fully generalisable to real genetic data, because the haplotype phase was assumed to be known and the observations were individual haplotypes rather than haplotype pairs. Assuming that each additional causal allele contributes additively to the log-odds of disease, this should not affect the power of the single-SNP tests or main-effects model at all. It may affect the power of the two higher-order models slightly, since these are affected by haplotype uncertainty. On the whole, we are con dent these results would show a similar pattern if applied to unphased data where the LD is at least moderately strong, as it is for modern GWAS.

### Improvement of haplotype modelling when adjacent interactions are included

The simulation study shows that for rare causal alleles, including the adjacent interaction terms tends to improve power over main effects to detect a disease which is associated with a haplotype effect. To gain greater insight into why this is, we tested whether including adjacent interactions significantly improved the modelling of haplotype frequencies. As further



**Figure 1 a, b - Power curves from simulation study, MAF = 0.005**. Power of main-effects (red), main+adj (blue), haplotype (green), and single-SNP tests (black) for (a) causal allele on moderate number of marker haplotypes; (b) causal allele on one marker haplotype; with causal SNP MAF = 0.005. For details on simulation, see Methods.

**Figure 2 a, b - Power curves from simulation study, MAF = 0.025**. Power of main-effects (red), main+adj (blue), haplotype (green), and single-SNP tests (black) for (a) causal allele on moderate number of marker haplotypes; (b) causal allele on one marker haplotype; with causal SNP MAF = 0.025.

described in Methods, this was done by fitting multinomial logistic models with haplotype carried as the response variable and main-effects or main+adj effects as the explanatory variables.

We randomly chose 1000 sets of 1500 phased haplotype observations at six marker SNPs, with average density one SNP per approximately 5 kb, from the dataset generated by COSI. For comparison purposes, 1000 additional sets with average density one SNP per 20 kb were also chosen. Table 1 shows the number of marker sets where the adjacent interactions significantly improved the fit of the haplotype frequencies. Note that in some sets, no adjacent interactions were included because of high LD between markers. As expected, this occurred less often for the sparser marker sets.

Table 1 shows that, for the vast majority of sets, including the adjacent interactions both improves the modelling of the haplotype frequencies significantly, and reduces the AIC. The proportion of sets for which this is true decreases when the marker SNPs are more dense, but is still over 90%. This indicates that when an untyped causal SNP is associated with a marker haplotype, including the adjacent interactions should improve

modelling the association between the phenotype and genetic information. Whether or not this translates into improving the power to detect association depends on the number of extra parameters used by the adjacent interactions and the level of association between causal SNP and marker haplotype.

**Population stratification**

To compare the effect of population stratification on the multimarker models, we simulated sets of 3-SNP haplotypes from two distinct populations. The frequency of the most common haplotype differed between populations depending on a parameter $\delta$, with larger values of $\delta$ yielding larger differences. The probability of an individual from either population being a case differed depending on a parameter $\mu$, with larger values giving larger differences. More details on these two parameters are given in the Methods section. For each set of parameters, test statistics were calculated from 5000 independent sets of 3000 3-SNP haplotypes. The resulting inflation in test statistics, as measured by the ratio of observed median test statistic to expected median test statistic under the null, $\lambda$, is shown in figure 4.



**Figure 3 a, b - Power curves from simulation study, MAF = 0.25**. Power of main-effects (red), main+adj (blue), haplotype (green), and single-SNP tests (black) for (a) causal allele on moderate number of marker haplotypes; (b) causal allele on one marker haplotype; with causal SNP MAF = 0.25.
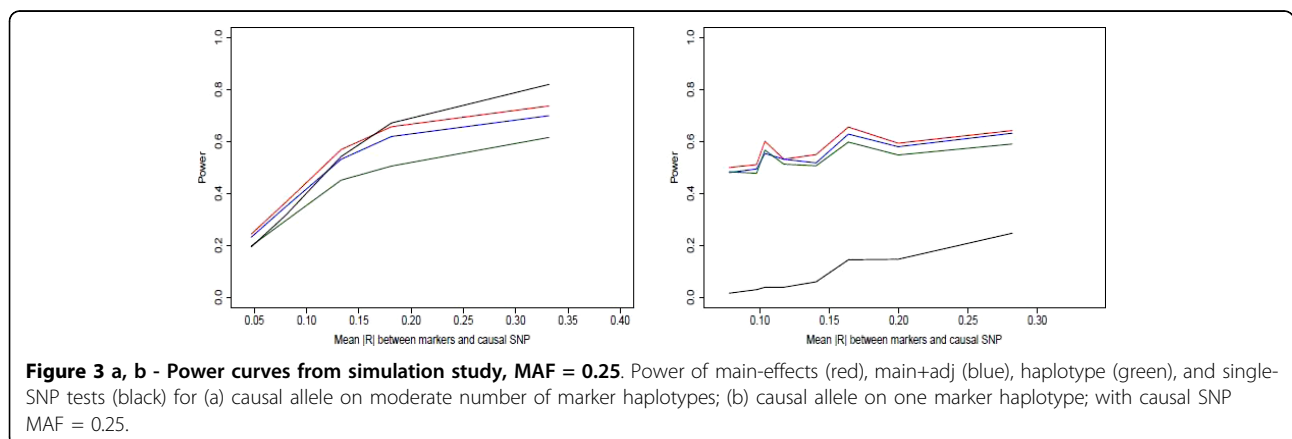
**Table 1 Improvement of haplotype modelling from including adjacent interactions**

| Density | Number of marker SNP sets with adjacent interactions included | Number (proportion) with significant improvement in log-likelihood | Number (proportion) with improvement in AIC |
|---|---|---|---|
| 5 | 727 | 662 (91.1%) | 674 (92.7%) |
| 20 | 983 | 955 (98.3%) | 961 (97.8%) |

Results of whether including adjacent interaction improves modelling of marker haplotype frequencies. Columns refer to 1) the density of marker SNPs (in terms of 1 marker SNP per *x* SNPs in dataset generated by COSI); 2) the number of sets (out of 1000) where adjacent interactions were not fully predictable from main-effects; 3) the number of sets with adjacent interactions included where they resulted in a significant improvement in model fit (at 5% significance level); 4) the number of sets where including adjacent interactions reduced the AIC.

Overall, the results in figure 4 appear to show that inflation in test statistics is low when $\mu \leq 0.04$, but can be high for realistic levels of population stratification when $\mu$ is higher. This shows that care must be taken to ensure that cases and controls are not sampled from different populations unless the analysis allows for this.

Making judgements for individual points on figure 4 is inadvisable, since $\hat{\lambda}$ is estimated from a finite sample. However, it is safer to look at the trend, which is consistent for higher levels of population stratification ($\mu \geq 0.6$). For the multimarker tests, the level of inflation in statistics is highest for the main-effects model, and lowest for the haplotype model. The main+adj model is in-between the two, but is generally slightly closer to the haplotype model. This is encouraging, as it shows the main+adj model controls inflation in statistics better than main-effects for stratified datasets.

Overall, the results show that the multimarker models which consider higher order interaction have lower inflation in statistics in stratified populations. We should be more careful with significant results found using the main-effects model, since they are more likely to be false positives due to population stratification. Only checking the inflation in the median may not give the whole picture, so we also show the QQ-plots of the statistics from the different models. These display the quantiles of the distribution of the observed score statistics against the quantiles of the theoretical distribution under the null. Figure 5 shows the QQ-plots of the different models for $\delta = 0.08$, $\mu = 0.08$, a fairly severe level of stratification.

The dashed horizontal lines in figure 5 represent the medians of the actual test statistics. The plots represent a $\lambda$ of 1.138 for the main-effects model, 1.108 for the main+adj omnibus model, and 1.086 for the haplotype model. The plots in figure 5 show nothing out of the ordinary. The level of inflation (the gradient of the line) is almost constant until deep into the tails of the distribution. Here, the points tend to be more divergent from the line of constant gradient, but this is to be expected due to sampling variance. The higher order multimarker models are less susceptible to population stratification than the main-effects model. This may indicate that it is the main-effect terms which are affected by stratification, but not the statistical interactions between SNPs. If this were the case, then it may result in the average inflation per parameter to be lower for the higher order models, and therefore for the overall effect to be less. Another explanation is that the simulation study assumes just one of the SNPs is affected by stratification. If several were affected in a window, then one might see the statistical interactions being affected too. Thus, if SNPs affected by stratification tended to cluster in groups, then it's possible that the higher order models would be more severely affected than the main-effects model.

Although the higher order multimarker models are less susceptible, they still show a high level of inflation. Thus, methods for correcting for population substructure, such as EIGENSTRAT [14] should be used. EIGENSTRAT yields principal component scores for each individual which can be easily incorporated into the logistic regression model.

**Schizophrenia GWAS analysis**

Simulated data can never fully capture the structure of real data, so we applied the multimarker models to the MGS study of Schizophrenia [16]. This study had genotyped 2,681 cases and 2,653 controls of European descent using the Affymetrix 6.0 chip. Genotypes can be obtained from the dbGaP database (www.ncbi.nlm.nih.gov/gap/). The controls were screened members of the general population. No genome-wide significant results had been found using single-SNP tests, although three SNPs showed association at $p < 1 \times 10^{-6}$. A region on chromosome 6p22.1 was implicated in a subsequent meta-analysis (described in the same paper). We used similar quality control measures as carried out in the actual study, which are described in the methods section. To avoid multicolinearity problems, an LD pruning step was used to filter out SNPs which were almost completely inferrable from nearby SNPs. This step filtered out the most SNPs, leaving 363,579 markers out of 671,422 originally genotyped.

The methods section describes how the different models were applied to data with unknown haplotypic phase. We used a sliding window with fixed window
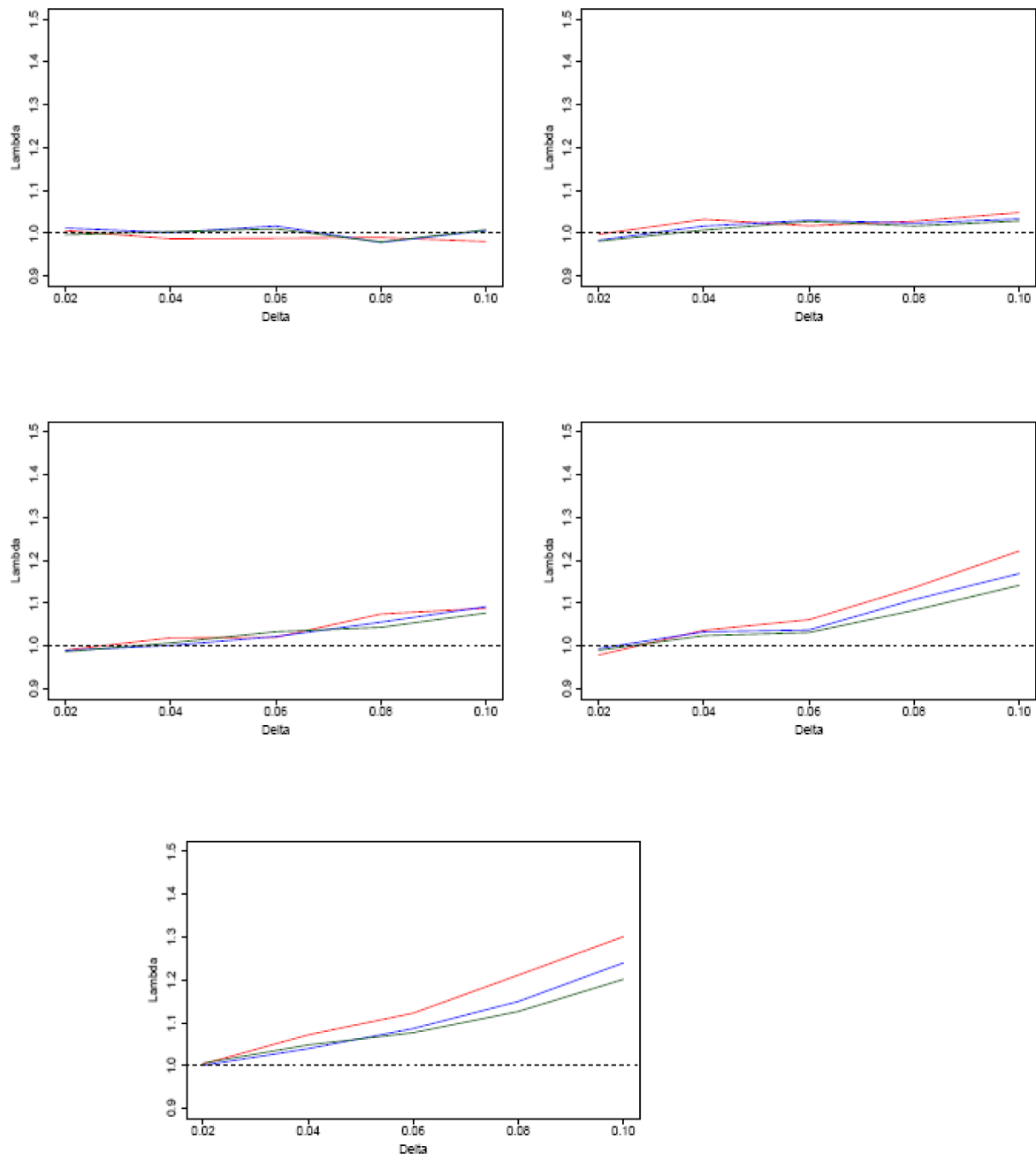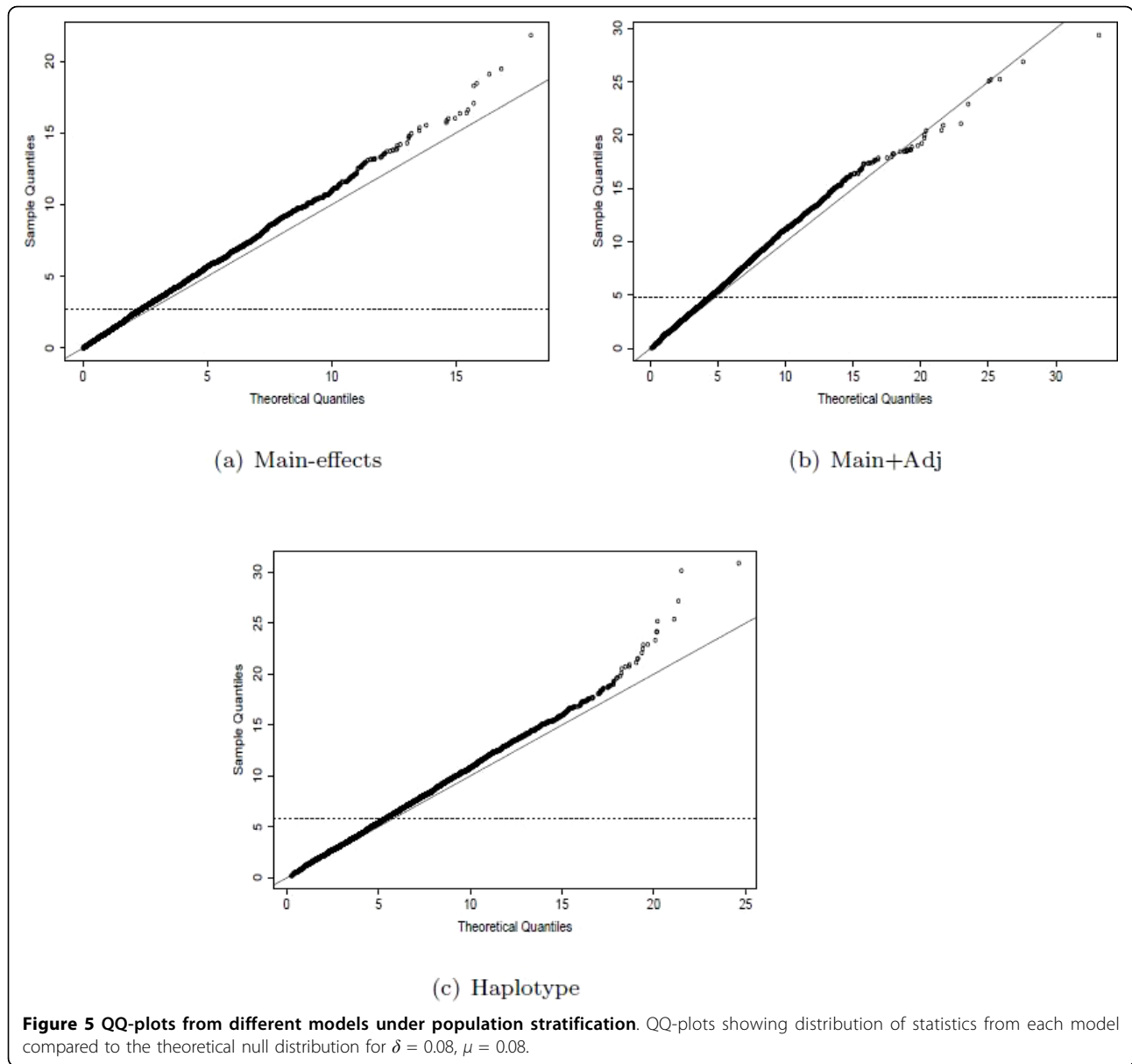
**Figure 4 Inflation in test statistics from different models under population stratification**. Inflation in test statistics for (a) $\delta$ = 0.02, (b) $\delta$ = 0.04, (c) $\delta$ = 0.06, (d) $\delta$ = 0.08, (e) $\delta$ = 0.1 of main-effects (red), main+adj (blue) and haplotype (green) models. $\delta$ represents the difference between populations of the frequency of the most common haplotype, and $\mu$ represents the difference in probabilities of a random individual from each population being a case. Inflation is measured by $\lambda$, the ratio of the median of observed statistics to the median expected under the null.

size of 6 marker SNPs, since window sizes of 5-6 were found to be most powerful for haplotype analysis by Zaykin et al. [18].

### Marginally significant results
To give a high-level comparison of results from each model, we firstly present the number of tests on each chromosome which had p-value $1 \times 10^{-4}$ or less.

Although this threshold is not genome-wide significant, the number of such tests gives an indication as to whether each method is picking up genuine associations. Since each multimarker model is a proxy for testing association with untyped genetic variants, comparing the p-values from the different methods should be valid. One issue that may cause the number of marginally

**Figure 5 QQ-plots from different models under population stratification**. QQ-plots showing distribution of statistics from each model compared to the theoretical null distribution for $\delta = 0.08$, $\mu = 0.08$.
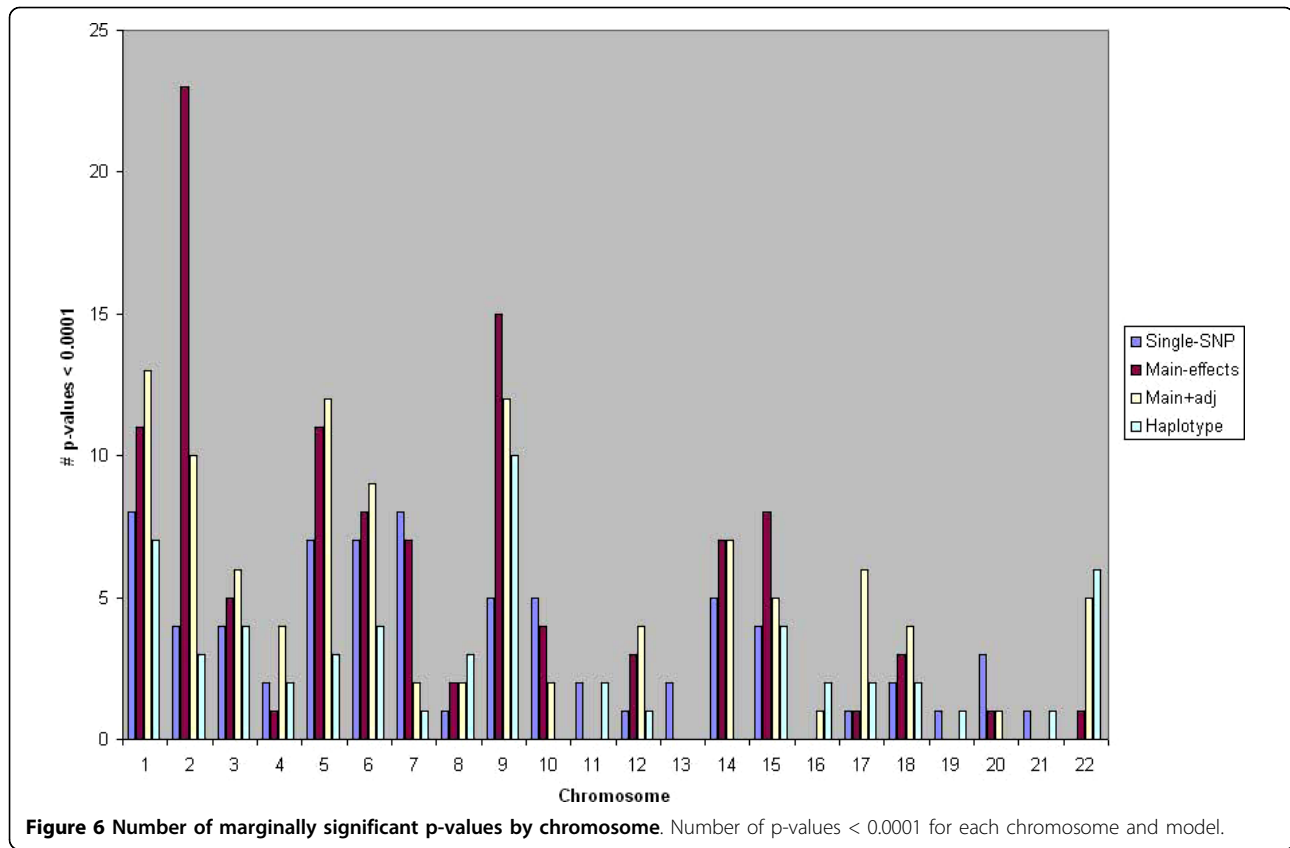
significant p-values to not be comparable is that the models are affected by quality control issues and stratification differently. This is examined later.

Figure 6 shows the number of marginally significant p-values given by each model for each chromosome. The total numbers across the genome were 73, 111, 105, and 58 for the single-SNP tests, main-effects, main+adj, and haplotype models respectively. These numbers show that the main-effects and main+adj models found more significant results than the single-SNP tests. In particular, the main-effects model performs best, with main+adj close behind. On a chromosome level, this ranking is more variable. For some chromosomes, the main+adj model outperforms the main-effects model. This indicates that

the two models find different results, but also may be consistent with sampling variance.

Significant test results do not necessarily correspond to true positives. If we just consider the number of tests performed ($\approx 363, 500$), then we'd expect an average of just over 36 test statistics to be significant at p-value 0.0001 level by chance alone. In addition, factors such as population stratification can cause spurious associations. To get an idea of the inflation in test statistics, we calculated $\hat{\lambda}$ from equation (12) for each model. Figure 7 summarises $\hat{\lambda}$ for each chromosome.

Figure 7 shows that the main-effects model generally has highest $\hat{\lambda}$, followed by the main+adj model. Statistics from single-SNP tests and the haplotype model

**Figure 6 Number of marginally significant p-values by chromosome**. Number of p-values < 0.0001 for each chromosome and model.

generally show lower inflation; on average $\hat{\lambda}$ from the haplotype model is slightly lower. The average chromosome $\hat{\lambda}$ across the genome for each model is 1.070, 1.101, 1.092, and 1.064 for the single-SNP, main-effects, main+adj, and haplotype tests respectively. There are several possible explanations for this inflation:

1. It may be due to genuine associations, since there could be many untyped SNPs with small effect size which are picked up by the main-effects model.
2. Population stratification can cause inflation in multimarker test statistics, as seen earlier.
3. Problems with poor quality data may also yield inflated test statistics.
4. If parameters in a GLM are estimated from few observations, then the distribution of test statistics under the null tends to be inflated.

Problems with asymptotics do not explain the pattern of $\hat{\lambda}$ seen above, since the multimarker model with fewest parameters has the higher inflation in the median. Other quality control factors could be responsible, but the pre-study QC step should have gone some way to reduce false positives due to genotyping error and HWE.

Population stratification is a plausible cause of inflated test statistics here. The ranking of multimarker models in terms of $\hat{\lambda}$ is the same as it was for the stratification simulation. We therefore repeated the analysis with the top five principal components included. Test statistics from the multimarker models became the score test for the effects of the genetic covariates after correcting for the principal components. Table 2 gives the genome-wide inflation in the median test statistic for each of the models before and after correction for stratification.

For the three multimarker models, including the principal components significantly reduces the inflation in the test statistics. It does not, however, completely eliminate the inflation. After adjustment, the main-effects model still has the highest inflation, with the haplotype model having the lowest; on the other hand, the difference in $\hat{\lambda}$ between the models has shrunk noticeably. The remaining inflation may be due to one of the other reasons above, or may be because of residual population stratification. We included the top five principal components, but lower ones may also be informative (though are likely to be significantly less so).

Although we considered PC adjustment for single-SNP tests, the type of statistic used differs. Before PC adjustment, the Cochran-Armitage trend test was used.

**Figure 7 Inflation in test statistics by chromosome**. $\hat{\lambda}$ for each chromosome and model.
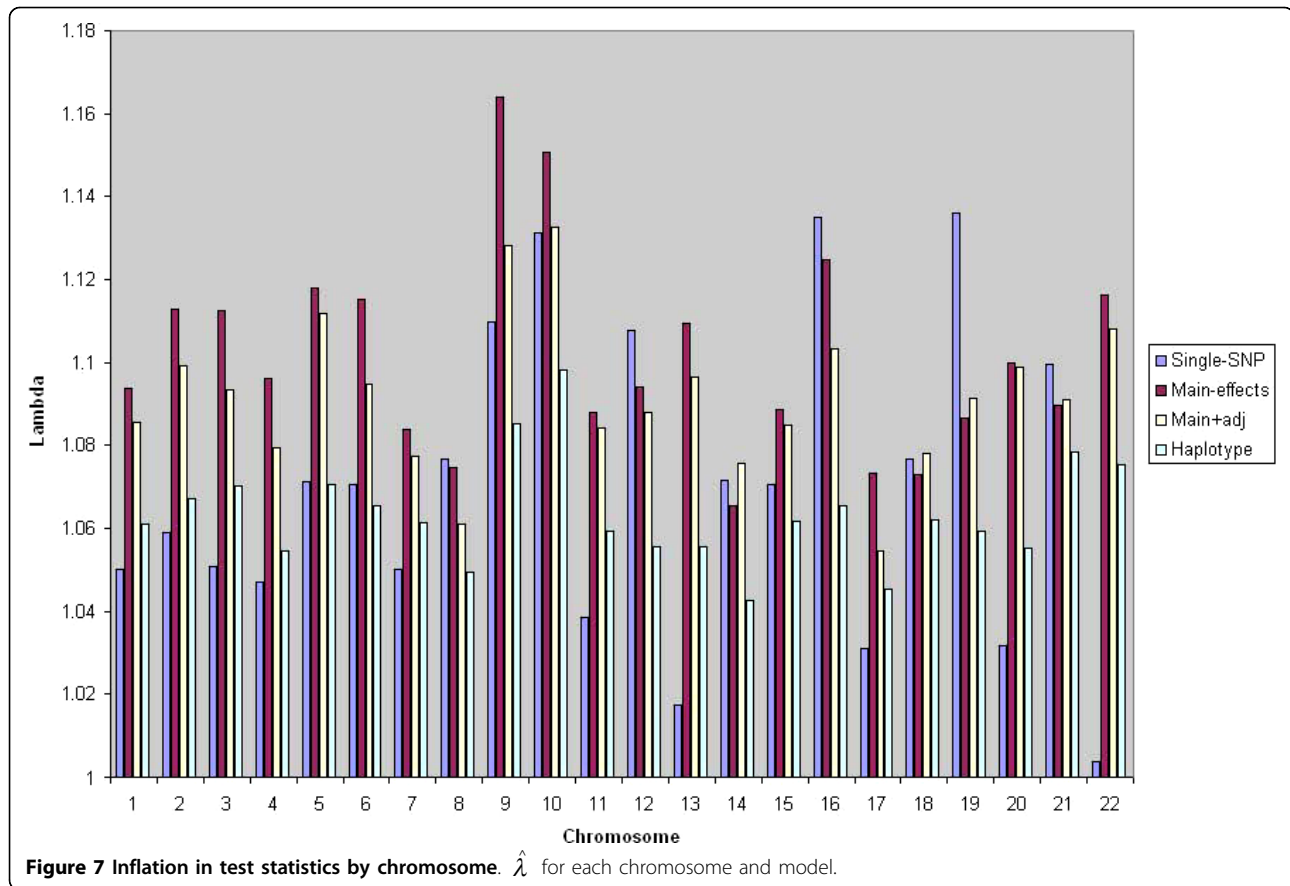
Adjustment of the trend test is possible, as described by Rakovski and Stram [19]. However, currently widely used software packages, such as PLINK [20], do not implement this adjustment. To incorporate PCs using PLINK, a suitable test is a 1 d.f. single-SNP logistic regression test. Asymptotically, under the null, these should give the same results. This does not apply when there are biases or genuine associations present. Thus, we do not consider the comparison here to be valid.

**Table 2 Inflation before and after principal component correction**

| Model | $\hat{\lambda}$ without PC adjustment | $\hat{\lambda}$ with PC adjustment |
|---|---|---|
| Single-SNP tests* | 1.070 | 1.077 |
| Main-effects | 1.101 | 1.059 |
| Main+adj | 1.092 | 1.052 |
| Haplotype | 1.064 | 1.037 |

Comparison of inflation in median test statistics ( $\hat{\lambda}$ ) for each model with and

without principal components. *Single-SNP tests without PC adjustment were Cochran-Armitage trend statistics which are not able to incorporate PC adjustment. The single-SNP statistics with PC adjustment are the 1 d.f. logistic regression.

Since $\hat{\lambda}$ has shrunk significantly, there should also be fewer marginally significant p-values. Table 3 shows that this is indeed the case, with 29 fewer for both main-effects and main+adj models and 10 fewer for the haplotype model. All models still have more significant associations than expected under the null. Although we have explored the marginal associations, a question of interest is whether any highly significant associations are present, and whether correcting for population stratification changes this.

Although principal component approaches provide a convenient method of adjusting test statistics for population stratification within a logistic regression framework, there have been recent publications which propose different methods. Population stratification is closely related to the problem of cryptic relatedness, where the assumption of independent data is flawed due to kinship between individuals (which is generally greater within cases). Both population stratification and cryptic relatedness result in hidden structure which can cause confounding. Rakovski and Stram [19] use Gauss-Markov estimators from a linear model where the covariance structure is imposed to be proportional to a relatedness matrix estimated from the data. This

**Table 3 Marginally significant p-values before and after principal component correction**

| Model | # marginally significant p-values without PC adjustment | # marginally significant p-values with PC adjustment |
|---|---|---|
| Main-effects | 111 | 82 |
| Main+adj | 105 | 76 |
| Haplotype | 58 | 48 |

Number of p-values $< \times 10^{-4}$ for each model with and without principal components.

approach is shown to have advantages over the principal component method of EIGENSTRAT. Astle and Balding [21] review a variety of methods, and conclude that using linear mixed models is a powerful and computationally efficient approach which makes the most explicit use of kinship. None of these techniques have been applied to multimarker tests, and this would be an interesting piece of work in its own right.

### Highly significant results

Tables 4, 5 and 6 give the top 25 most significant results for each multimarker model. Also included is the p-value after correcting for the top five principal components. Note that the SNP listed is the first one in the window of 6.

The most significant p-values are generally found using the main-effects model, followed by the main+adj model, and lastly the haplotype model. There is also a large overlap in which windows were most significant from each model. The most significant test for each model is for the 6 SNP window starting at SNP rs7850685 on Chromosome 9. The minimum of p-values from the single-SNP statistics for the 6 SNPs in this window was 0.002, although 4 of the other 5 SNPs had p-values of 0.1 or less.

There are windows for which the main+adj model is most significant, for example the window starting at rs1554268 on chromosome 2. The difference is not that great before correction, and still within an order of magnitude after correction for population stratification. Most windows have more significant p-values using the main-effects model, which may indicate that the associations are due to common untyped SNPs rather than rare ones.

A few windows which appear in the top 25 results from the haplotype model do not appear in either of the

**Table 4 Highly significant main-effects windows**

| Chromosome | First SNP in window | Kb map location | p-value | PC corrected p-value |
|---|---|---|---|---|
| 9 | rs7850685 | 99803.7 | 1.36E-08 | 2.06E-08 |
| 9 | rs10967991 | 27564.8 | 7.88E-08 | 8.30E-07 |
| 9 | rs13049 | 99862.9 | 1.25E-07 | 1.57E-07 |
| 9 | rs872251 | 99797.2 | 1.30E-07 | 2.89E-07 |
| 9 | rs2783010 | 27571.2 | 4.14E-07 | 3.70E-06 |
| 9 | rs1982915 | 27569.6 | 7.14E-07 | 5.60E-06 |
| 17 | rs9904870 | 72109.7 | 9.59E-07 | 7.79E-06 |
| 9 | rs10967992 | 27572 | 1.53E-06 | 1.49E-07 |
| 5 | rs17176973 | 10864.5 | 1.86E-06 | 5.15E-06 |
| 5 | rs13155209 | 10841.9 | 2.69E-06 | 2.41E-06 |
| 5 | rs2964798 | 10841.8 | 2.96E-06 | 3.14E-06 |
| 18 | rs4542737 | 23750.3 | 3.03E-06 | 2.18E-05 |
| 12 | rs10774841 | 114863 | 3.44E-06 | 3.65E-06 |
| 15 | rs3784405 | 86489 | 3.86E-06 | 9.38E-06 |
| 6 | rs12211360 | 143162 | 5.15E-06 | 1.57E-05 |
| 6 | rs197508 | 143139 | 5.18E-06 | 1.65E-05 |
| 5 | rs17771257 | 10863.6 | 6.08E-06 | 1.42E-05 |
| 6 | rs6928738 | 31039 | 6.35E-06 | 1.42E-05 |
| 2 | rs1554268 | 22284.6 | 8.22E-06 | 1.98E-05 |
| 6 | rs12205474 | 143141 | 8.72E-06 | 1.52E-05 |
| 1 | rs12025436 | 80078.6 | 8.85E-06 | 1.64E-05 |
| 1 | rs722589 | 79946.9 | 8.94E-06 | 6.65E-05 |
| 1 | rs10158916 | 79946.2 | 9.13E-06 | 6.50E-05 |
| 2 | rs12151565 | 22352.1 | 1.04E-05 | 8.31E-05 |
| 1 | rs17422551 | 79964.9 | 1.04E-05 | 8.10E-05 |

Top 25 results by significance for main-effects omnibus tests.

**Table 5 Highly significant main+adj windows**

| Chromosome | First SNP in window | Kb map location | p-value | PC corrected p-value |
|---|---|---|---|---|
| 9 | rs7850685 | 99803.7 | 1.84E-07 | 4.14E-07 |
| 9 | rs10967991 | 27564.8 | 4.44E-07 | 7.64E-06 |
| 9 | rs2783010 | 27571.2 | 1.12E-06 | 1.60E-05 |
| 2 | rs1554268 | 22284.6 | 1.24E-06 | 3.98E-06 |
| 9 | rs872251 | 99797.2 | 1.93E-06 | 5.56E-06 |
| 9 | rs13049 | 99862.9 | 2.41E-06 | 2.88E-06 |
| 5 | rs13155209 | 10841.9 | 2.59E-06 | 3.79E-06 |
| 17 | rs9904870 | 72109.7 | 2.59E-06 | 1.96E-05 |
| 9 | rs10967992 | 27572 | 2.96E-06 | 4.26E-05 |
| 9 | rs1982915 | 27569.6 | 3.00E-06 | 3.81E-05 |
| 6 | rs12211360 | 143162 | 7.46E-06 | 1.31E-05 |
| 1 | rs12025436 | 80078.6 | 8.85E-06 | 1.64E-05 |
| 1 | rs722589 | 79946.9 | 8.94E-06 | 6.65E-05 |
| 1 | rs10158916 | 79946.2 | 9.13E-06 | 6.50E-05 |
| 12 | rs10774841 | 114863 | 9.18E-06 | 1.26E-05 |
| 2 | rs340764 | 16321.5 | 1.01E-05 | 1.23E-05 |
| 1 | rs17422551 | 79964.9 | 1.04E-05 | 8.10E-05 |
| 5 | rs12186813 | 10855.7 | 1.07E-05 | 1.85E-05 |
| 1 | rs12144807 | 79945.4 | 1.09E-05 | 5.37E-05 |
| 6 | rs10805983 | 74663.4 | 1.24E-05 | 5.72E-05 |
| 6 | rs6928738 | 31039 | 1.38E-05 | 2.82E-05 |
| 22 | rs178259 | 19654.2 | 1.52E-05 | 2.15E-05 |
| 1 | rs1524183 | 79952.5 | 1.68E-05 | 1.03E-04 |
| 18 | rs4542737 | 23750.3 | 1.75E-05 | 9.70E-05 |
| 22 | rs178268 | 19660.3 | 1.80E-05 | 1.88E-05 |

Top 25 results by significance for main+adj omnibus tests.

other two, for example, the window on chromosome 16 starting at rs17618203. The window is still significant using the other two models, but the p-value is an order of magnitude higher. It does seem that each model has windows for which it is the most significant. This is backed up by the simulation results earlier which showed that in situations where the main-effects model performed best, the other two models did still find significant p-values for datasets which the main-effects model did not.

Test statistics which correct for population stratification are generally less significant than those that do not. The level of correction differs greatly between windows. Some, for example the window on chromosome 16 starting at rs17618203 using the haplotype test, actually become more significant after correction; others are corrected by an order of magnitude or more.

### QC of top ranked windows

We examined quality control statistics for the most significant results found. For each of the 25 most significant main-effects omnibus tests, p-values for SNP-HWE and differential missingness by phenotype were calculated for each SNP. Table 7 lists the minimum SNP-

HWE and differential missingness p-value, together with the haplotype-HWE p-value for each window.

The table shows that most of the top results do not have significant tests for any of the QC criteria. Some results show some potential problems, for example two of the windows on Chromosome 9 show evidence against haplotypes being in HWE. The p-values from testing for differential missingness may seem to be closer to 0 than they should be, but as each one is the minimum of 6 p-values, the distribution is close to expected under the null of no differential missingness. The p-values for the SNP-HWE tests are more severely away from expectation under the null of HWE, perhaps indicating some false positives due to genotyping error.

A possible cause of false positives in haplotype testing is when sporadic but differential genotyping error causes a non-existent haplotype to appear in just cases (or just controls). The significant deviation from haplotypic HWE in the two consecutive windows on chromosome 9 may be due to this type of error. We used UNPHASED [22] to examine the haplotype distribution in cases and controls.

UNPHASED gives the case and control counts of haplotypes inferred from the data, together with a p-value

**Table 6 Highly significant haplotype windows**

| Chromosome | First SNP in window | Kb map location | p-value | PC corrected p-value |
|---|---|---|---|---|
| 9 | rs7850685 | 99803.7 | 2.14E-07 | 8.215372E-07 |
| 9 | rs13049 | 99862.9 | 1.95E-06 | 1.898616E-06 |
| 22 | rs178260 | 19654.3 | 7.51E-06 | 6.082905E-06 |
| 1 | rs10158916 | 79946.2 | 8.43E-06 | 5.806333E-05 |
| 22 | rs178259 | 19654.2 | 1.20E-05 | 1.794159E-05 |
| 15 | rs17264145 | 88904.7 | 1.35E-05 | 5.959549E-05 |
| 5 | rs13155209 | 10841.9 | 1.43E-05 | 2.769336E-05 |
| 1 | rs12144807 | 79945.4 | 1.44E-05 | 6.107985E-05 |
| 9 | rs10967991 | 27564.8 | 1.49E-05 | 1.232903E-04 |
| 9 | rs872251 | 99797.2 | 1.60E-05 | 5.476211E-05 |
| 9 | rs551120 | 134879 | 1.72E-05 | 3.428334E-05 |
| 22 | rs9613019 | 24504.4 | 1.84E-05 | 4.384690E-06 |
| 15 | rs12911150 | 86469.7 | 2.05E-05 | 5.855712E-05 |
| 16 | rs17618203 | 65704.5 | 2.55E-05 | 9.370586E-06 |
| 5 | rs12186813 | 10855.7 | 2.67E-05 | 4.666928E-05 |
| 19 | rs11882629 | 11319.7 | 2.97E-05 | 8.237138E-05 |
| 1 | rs11591084 | 80077.3 | 3.23E-05 | 3.804365E-05 |
| 9 | rs4740559 | 14120.2 | 3.29E-05 | 2.475267E-05 |
| 1 | rs722589 | 79946.9 | 3.49E-05 | 2.592247E-04 |
| 2 | rs340764 | 16321.5 | 3.67E-05 | 5.459454E-05 |
| 9 | rs3132315 | 136567 | 3.68E-05 | 1.894568E-05 |
| 9 | rs10967992 | 27572 | 3.76E-05 | 4.216994E-04 |
| 1 | rs17422551 | 79964.9 | 3.78E-05 | 2.412767E-04 |
| 7 | rs16868596 | 88502.2 | 3.80E-05 | 5.435478E-05 |
| 12 | rs1993498 | 10689.9 | 3.86E-05 | 5.445096E-05 |

Top 25 results by significance for haplotype omnibus tests.

of association between phenotype and each haplotype. Edited output for the most significant window (starting at rs10818472) is given in table 8.

The second and third columns in table 8 show the estimated haplotype counts in cases and controls. There are several haplotypes which have a significant p-value. Of these, the most significant is 2 - 2 - 1 - 4 - 3 - 1 which has estimated count 57.67 in cases and 113 in controls. Since it is present in more controls, its estimated effect is protective. There are several other haplotypes with significant effect, one of which is present only in controls (3 - 4 - 3 - 4 - 3 - 1). This may have resulted from differential genotyping error. In this case, the haplotype may actually not exist, but appear to due to genotyping errors in SNP 2. The haplotype 3 - 2 - 3 - 4 - 3 - 1 is common in both cases and controls, so even a small chance of error per observation may result in 3 - 4 - 3 - 4 - 3 - 1 appearing in controls only and therefore showing a significant association. On the other hand, the significance of the window is not solely due to this haplotype, so this doesn't necessarily mean the result is a false positive.

## Imputation analysis of regions with highly significant results

Several regions appear to be associated with schizophrenia according to our results. Two regions on Chromosome 9 contain many highly significant omnibus tests. The single-SNP tests show some p-values which are significant at 5%, but no highly significant p-values.

One of these regions is between 27564.8 kb and 27572 kb, the other is between 99797.2 kb and 99862.9 kb. Although none of the single-SNP tests of typed SNPs in the region shows a highly significant p-value, it is possible that an imputation analysis would find significant associations with one or more of the untyped SNPs in the region. Even if not, it might show many of the untyped SNPs displaying association at a more marginal level, which could provide evidence of a genuine association. With this in mind, we decided to carry out an imputation analysis of both regions using BEAGLE [23]. BEAGLE imputes both common and rare SNPs well, and is quick to run on large regions. We extended each region by 50 kb in either direction to detect any nearby associations.

**Table 7 Summary of QC statistics for most significant windows**

| Chromosome | SNP | Min differential missingness p-value | Haplotype HWE p-value | Min SNP HWE p-value |
|---|---|---|---|---|
| 2 | rs6759206 | 8.20E-02 | 4.20E-01 | 2.29E-01 |
| 2 | rs2317011 | 9.86E-02 | 3.90E-01 | 2.29E-01 |
| 2 | rs1962550 | 9.86E-02 | 3.90E-01 | 2.29E-01 |
| 5 | rs2964798 | 4.99E-02 | 4.90E-01 | 1.57E-02 |
| 5 | rs13155209 | 4.99E-02 | 7.60E-01 | 1.57E-02 |
| 5 | rs10043680 | 4.86E-02 | 8.20E-01 | 2.47E-01 |
| 5 | rs17771257 | 4.86E-02 | 5.80E-01 | 3.75E-01 |
| 5 | rs17176973 | 4.86E-02 | 8.80E-01 | 7.01E-02 |
| 5 | rs17075700 | 7.73E-02 | 3.90E-01 | 4.29E-02 |
| 7 | rs10227362 | 1.16E-01 | 6.70E-01 | 3.85E-01 |
| 7 | rs12534223 | 1.16E-01 | 9.30E-01 | 1.33E-01 |
| 9 | rs10967991 | 1.82E-01 | < 0.001 | 1.07E-01 |
| 9 | rs1982915 | 1.82E-01 | < 0.001 | 1.07E-01 |
| 9 | rs2783010 | 1.82E-01 | 1.10E-02 | 1.07E-01 |
| 9 | rs10967992 | 7.94E-02 | 1.60E-02 | 1.07E-01 |
| 9 | rs10967993 | 7.94E-02 | 1.60E-01 | 1.07E-01 |
| 9 | rs872251 | 6.24E-03 | 3.20E-01 | 7.78E-03 |
| 9 | rs7850685 | 6.24E-03 | 5.00E-02 | 7.78E-03 |
| 9 | rs13049 | 6.24E-03 | 5.70E-01 | 1.03E-01 |
| 10 | rs682664 | 5.85E-02 | 4.70E-01 | 3.94E-02 |
| 12 | rs10774841 | 3.08E-01 | 3.10E-01 | 2.40E-02 |
| 15 | rs3784405 | 1.99E-01 | 6.30E-01 | 3.83E-01 |
| 17 | rs9904870 | 1.88E-01 | 4.80E-01 | 2.50E-01 |
| 18 | rs4542737 | 4.05E-02 | 7.60E-01 | 6.47E-04 |
| 20 | rs967417 | 2.42E-01 | 4.30E-01 | 1.23E-01 |

QC statistics for 25 most significant main-effects windows.

**Table 8 UNPHASED output**

| Haplotype | Case | Control | P-value |
|---|---|---|---|
| 2-2-1-2-1-1 | 6.115 | 6.69 | 0.2675 |
| 2-2-1-2-1-2 | 0.9372 | 0 | 0.308 |
| 2-2-1-2-3-1 | 0.6905 | 0 | 0.9992 |
| 2-2-1-4-1-1 | 828.9 | 834 | 0.8783 |
| 2-2-1-4-1-2 | 9.433 | 10.21 | 0.8322 |
| 2-2-1-4-3-1 | 57.67 | 113 | 3.747E-05 |
| 2-2-1-4-3-2 | 1.007 | 0 | 0.3188 |
| 2-2-3-2-1-1 | 12.98 | 6.27 | 0.3267 |
| 2-2-3-2-1-2 | 0 | 0.6707 | 1 |
| 2-2-3-2-3-1 | 1.204 | 0 | 1 |
| 2-2-3-4-1-1 | 0.867 | 1.918 | 0.3793 |
| 2-2-3-4-1-2 | 15.44 | 19.16 | 0.7131 |
| 2-2-3-4-3-1 | 30.39 | 43.74 | 0.2237 |
| 2-2-3-4-3-2 | 2.114 | 0.9249 | 0.609 |
| 2-4-1-4-1-1 | 1.627 | 3.416 | 0.365 |
| 2-4-3-2-1-2 | 0.8596 | 2.21 | 0.6195 |
| 2-4-3-4-1-1 | 1.591 | 13.85 | 0.002506 |
| 2-4-3-4-1-2 | 137.2 | 122.5 | 0.4528 |
| 2-4-3-4-3-1 | 0 | 2.456 | 0.01213 |
| 3-2-1-2-1-1 | 1.148 | 0.6285 | 0.7064 |
| 3-2-1-4-1-1 | 85.07 | 64.9 | 0.07011 |
| 3-2-1-4-1-2 | 1.305 | 0.2028 | 0.3483 |
| 3-2-1-4-3-1 | 3.732 | 0 | 0.3045 |
| 3-2-3-2-1-1 | 295.5 | 334.6 | 0.1975 |
| 3-2-3-2-1-2 | 7.775 | 4.195 | 0.2888 |
| 3-2-3-2-3-1 | 4.929 | 8.097 | 0.1492 |
| 3-2-3-4-1-1 | 5.902 | 0.8962 | 0.06286 |
| 3-2-3-4-1-2 | 48.69 | 19.48 | 0.0004034 |
| 3-2-3-4-3-1 | 1114 | 1153 | 0.3141 |
| 3-2-3-4-3-2 | 5.975 | 2.85 | 0.2993 |
| 3-4-1-4-1-1 | 1.055 | 4.961 | 0.1278 |
| 3-4-1-4-1-2 | 1.301 | 2.3 | 0.6549 |
| 3-4-1-4-3-1 | 1.013 | 0 | 0.3212 |
| 3-4-1-4-3-2 | 0 | 0.6906 | 0.09517 |
| 3-4-3-2-1-1 | 7.049 | 1.218 | 0.3176 |
| 3-4-3-2-1-2 | 12.86 | 31.44 | 0.008042 |
| 3-4-3-4-1-1 | 17.34 | 20.11 | 0.555 |
| 3-4-3-4-1-2 | 2422 | 2309 | 0.02225 |
| 3-4-3-4-3-1 | 0 | 12.7 | 0.004522 |

UNPHASED output for window beginning with SNP rs17046067.

As a reference panel, we used filtered HapMap phase II release 23 data. The filtering removed SNPs with MAF less than 0.01, and ones with genotyping rate less than 0.95. Since HapMap phase II has a small sample size, SNPs with a very low MAF are prone to error. We used just the individuals of European descent, since the MGS samples were of European descent. The reference panel had 60 individuals genotyped at 2.3 million SNPs.

BEAGLE provides an estimate of the squared Pearson correlation between the imputed and actual allelic dosage. This gives an assessment of the quality of imputation, with values of 0.95 or higher indicating excellent quality imputation. We used BEAGLE to impute the two regions of chromosome 9, using the whole of chromosome 9 HapMap data as a reference panel. This should incorporate long range LD information and improve the quality of the imputation. Imputed SNPs with allelic dosage squared correlations of 0.95 or greater were tested for association with case-control status using the 1 d.f. allelic test.

Tables S1 and S2 in additional files 1 and 2 show the results from applying this imputation procedure to the two aforementioned regions of chromosome 9.

The results in table S1 show that of the 41 good quality SNPs, nine have significant p-values at the 5% significance level. Six of these were in the original dataset. Although the minimum p-value is not near genome-wide significance, the fact that there are a number of marginal associations provides evidence that there may be a genuine association within the region as a whole,

although the number itself means little because several of the SNPs which show association could be in LD with one another.

Results from table S2 are very similar. Out of 39 good quality SNPs, 9 have p-values ≤ 0.05. Two of these were from the original dataset, so the imputed SNPs give more new evidence of association than the previous region.

The results from imputation do indicate that associations within the region could be due to untyped variants. If the imputation analysis had been performed on the entire genome, it is unlikely that the regions would have been picked as promising for further analysis - none of the p-values are below even $1 \times 10^{-4}$. The multimarker models on the other hand show several highly significant windows within the region. This provides an example of how using multimarker models may complement an imputation analysis in finding regions to investigate further.

### Summary of performance of multimarker models on GWAS and potential for sequence analysis

Overall, the analysis of the schizophrenia GWAS shows that multimarker tests highlight potential associations that single-SNP tests do not. However, they also result in more spurious associations, especially the main-effects model.

Including the adjacent interactions reduces the inflation in statistics, as was seen in simulated data previously. It also results in fewer highly significant p-values for this dataset. This is to be expected, since the haplotype model performed worse than the main-effects model and simulated data showed that the power of the main+adj model is generally in between the power of the other two models.

Of the most significant test results found, most appear to be based on good quality data, with little deviation from HWE and no evidence for differential missingness found. This is encouraging, as it appears that the results found uniquely through multimarker tests are not easily dismissed as false positives. However, detailed examination of the raw genotyping data may be needed to rule out differential genotyping error.

A next stage would be to look at regions where the multimarker models gave results which were far more significant than the single-SNP tests. It should be pointed out that a lot of these significant multimarker tests may be explained by SNPs that were filtered out by the LD pruning step. As an example, the 6 SNP window starting at rs9904870 on chromosome 17 actually had 8 SNPs before LD pruning. However, the two additional SNPs in this case do not show any association with the phenotype.

For two of the regions which multimarker tests had indicated might be associated with schizophrenia, we carried out an imputation analysis using BEAGLE. Both regions contained several marginal associations, but nothing near genome-wide significance. The imputation was not optimal for detecting associations with rare variants. For example, the reference panel used consisted of the individuals of European descent from HapMap phase II. This gave just 60 individuals in the reference panel - not enough to accurately model rare variants. The analysis would be far more accurate when a larger reference panel, such as that currently being developed by the 1000 genomes project, becomes available.

All of the models used have the highest power to detect associations when each causal allele contributes additively to log-odds of disease. This may not be the case, with some confirmed causal SNPs showing dominance effects. In this case, the multimarker models will lose power (as will the 1 d.f. single-SNP tests). Chapman and Clayton [24] discuss this issue, and suggest that if associations with such SNPs are anticipated, then an extra parameter can be added to the logistic regression model in order to improve the power to detect association. This slightly reduces the power for additive causal SNPs, but considerably improves the power for non-additive causal SNPs.

### Conclusions

In this paper we have shown that straightforward multimarker models have higher power than single-SNP tests to detect rare untyped variants. The relative ranking of models with different levels of complexity depends on the frequency of the causal variant, and the number of haplotypes the causal allele is present on. The best case scenario for the haplotype model is a recent mutation which is still rare and appears on just one marker haplotype. The best case scenario for the main-effects model is a common causal allele which appears on many marker haplotypes. This would be the case when the mutant allele appeared in the population long in the past. If, in addition, the causal SNP is in high LD with the marker SNPs, single-SNP tests have the highest power. The main+adj model performs well in all situations, and is thus a good 'compromise model'.

We have also shown that including the adjacent interactions results in lower inflation in test statistics due to population stratification. A possible reason is that the stratification manifests as a common-effect, and thus is better modelled by the main-effects model.

When applied to the schizophrenia GWAS, several regions were found which had highly significant multimarker test statistics, but did not contain any SNPs

which had significant test statistics when considered alone. This should not be taken as saying these regions are definitely associated, but instead as illustrations of the kind of regions which could later be analysed with a full sequence analysis with a higher prior odds for association.

A good candidate for a sequence analysis would be a region that shows: 1) highly significant associations with multimarker methods; 2) No obvious quality control problems; 3) No single-SNP tests with marginally significant p-value; and 4) No highly significant imputation results.

Although the main-effects model performed best on the schizophrenia GWAS, it may be that more samples are needed in order to have power to detect the rare untyped variants which the main+adj model detects with higher power. Perhaps comparing the three multimarker models on a larger study may show different results.

Overall, we have shown that multimarker logistic regression provides a straightforward and useful complement to single-SNP analysis. Applying it to existing GWAS data may provide new candidate regions with highly penetrant rare causal alleles which explain more of the heritability in complex human diseases.

## Methods
### Multimarker models
Logistic regression provides a flexible method of modelling the probability of an individual being a case as a function of genetic and non genetic factors. For n phased haplotypes at m SNPs, let $X_{ij}$ ($i \in \{1,...,n\}$ $j \in \{1,...,m\}$) be defined as:

$$X_{ij} = 1 \text{ if observation } i \text{ carries the minor allele at SNP } j$$
$$= 0 \text{ if observation } i \text{ carries the major allele at SNP } j. \quad (1)$$

and $Z_{ij}$ ($i \in \{1,...,n\}$ $j \in \{1,...,h\}$) be defined as:

$$Z_{ij} = 1 \text{ if observation } i \text{ carries haplotype } j$$
$$= 0 \text{ if observation } i \text{ does not carry haplotype } j, \quad (2)$$

where h is the total number of haplotypes.

Using the notation given in (1) and (2), with $Y_i$ the phenotype of the ith observation, the main-effects model is:

$$Y_i \sim \text{Bin}\left(1, \pi_i\right)$$
$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mu + \sum_{j=1}^{m} X_{ij}\beta_j. \quad (3)$$

The haplotype model is:

$$Y_i \sim \text{Bin}\left(1, \pi_i\right)$$
$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \sum_{j=1}^{h} Z_{ij}\beta_j, \quad (4)$$

and the main+adj model is:

$$Y_i \sim \text{Bin}\left(1, \pi_i\right)$$
$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mu + \sum_{j=1}^{m} X_{ij}\beta_j + \sum_{j-1}^{m-1} X_{ij}X_{ij+1}\beta_{j+m}. \quad (5)$$

In practice, we include an adjacent interaction term only if the LD between the two SNPs is below a threshold value. Let $\pi_{11}$, $\pi_{12}$, $\pi_{21}$, $\pi_{22}$ represent the frequencies for the four possible phased haplotypes at two adjacent markers. Since these frequencies sum to 1, there are three identifiable parameters modelling the effect of the two-SNP haplotype on disease risk: the main-effects $\beta_j$ and $\beta_{j+1}$ model two of them, and the adjacent interaction $\beta_{j+m}$ models the third. If the two SNPs are in complete LD ($D' = 1$), then at least one of the cell frequencies is 0, and there are at most two identifiable parameters. If the two SNPs are close to $D' = 1$, then the third parameter is identifiable, but is estimated from few observations. According to Peduzzi et al. [25], logistic regression experiences problems with asymptotics if fewer than ten observations are used for estimating a parameter. Thus, we include an adjacent interaction in the model if the observed count of the rarest 2-SNP haplotype is ≥ 10.

Similarly for the haplotype model, we only include a haplotype parameter if it has at least ten observations. This is to avoid the problem of haplotype parameters being estimated from too few observations, and assumes that very rare haplotypes contribute negligible power to detect a causal variant. The remaining haplotypes are merged into one 'rare haplotype' category, with a parameter estimated for the effect of an individual carrying one of these. This is a straightforward, and indeed, simplistic method of dealing with rare haplotypes. Other methods have been proposed [26-29] which use inferred population histories to merge haplotypes which are likely to have similar effects. For window sizes of six or less dense markers, there will typically be few rare haplotypes, so merging them together should not be significantly less powerful than a more complicated method. For large GWAS, a haplotype with ten observations or less is extremely rare, and there is very low power to detect a significant association with it.

Generally, each observation is an individual with two haplotype strands. If the haplotypic phase is known, then each row of the design matrices in models 3 - 5 consists of the sum of the codings from the individual haplotype strands. For instance, if individual $i$ carries two copies of haplotype $k$, then $Z_{ik} = 2$ in the haplotype model. If the haplotypic phase is unknown, different approaches may be used and are discussed later on.

For the purposes of this paper the omnibus test, of all non-intercept parameters being 0 against the alternative of at least one non-intercept parameter being 0, is of interest. When the observed data is in phased format, the likelihood ratio test can be used to assess significance.

### Simulation study

We used COSI [17] to simulate 20,000 phased haplotype observations of a 1 Mb chromosome segment. COSI allows the user to simulate realistic data using specified population genetics parameters. We used the default set of parameters which produces data with allele frequencies and LD consistent with individuals of European descent. The resulting dataset had 4,164 mutation sites, or one every 240 bp. Many of these mutations were perfectly correlated with nearby mutations. After removing these, we were left with 2,474 mutations, 955 of which had MAF $\geq 0.05$.

Parameters that were varied in the simulation study include the MAF of the causal SNP, the density of the marker SNPs, and the relative risk of the risk allele at the causal locus. We considered two approaches to picking a causal SNP. These are:

1. The causal SNP is picked at random from the set of all COSI-generated SNPs such that it has the required MAF. Marker SNPs are sampled from the set of nearby SNPs with MAF $\geq 0.05$ such that they have the required density. The causal SNP is not included as a marker SNP, but is located in the middle of them.

2. A set of marker SNPs is picked from the COSI-generated dataset with the required density. The observed frequencies of all the distinct marker haplotypes are calculated, and the marker haplotype with frequency closest to the required causal SNP MAF is picked. Observations that carry this marker haplotype are assigned the causal allele, with other observations being assigned the non-causal allele.

In the first case, the causal allele is generally on a moderate number of marker haplotypes, whereas in the second case it is on just one. The two situations correspond to different ages of the causal mutation, with the latter case resulting in the founder haplotype being present only in its original form, without recombination or mutation, i.e. the causal mutation arising recently.

Testing a large number of SNPs together is impractical in terms of time taken to fit the model, asymptotics of the model (i.e. not enough observations per parameter), and power (degrees of freedom of the test statistic being too large). For the simulation study we looked at sets of six marker SNPs, which has been suggested previously as being close to optimal for haplotype testing [18].

For each replicate, a disease is generated for all observations in the COSI dataset from the causal SNP. Observations with the non-causal allele have a baseline risk of disease, and observations with the causal allele have an increased risk. For each causal SNP MAF, the relative risk was chosen to give 80% power to detect the causal SNP directly using a 2 d.f. genotype test. A sample of 1500 case haplotypes and 1500 control haplotypes are taken as the dataset for that replicate. This process simulates a case-control study design.

For each dataset, four models were fitted:

1. Main-effects model from equation (3).
2. Main+adj model from equation (5).
3. Haplotype model from equation (4).
4. Logistic regression of phenotype on each SNP individually (asymptotically equivalent to a 1 d.f. allelic trend test). The minimum p-value from these tests was used

The aim of the simulation study was to compare the power of the four models for different causal SNP frequencies; different LD levels between causal SNP and marker SNPs; and different relations between the causal allele and marker haplotypes. We considered very rare, semi-rare, and common causal alleles with MAFs 0.005, 0.025, and 0.25 respectively. The LD between causal SNP and marker SNP was varied by taking different densities for the sets of marker SNPs. The possible densities were one marker per 5, 10, or 20 SNPs in the COSI dataset. These densities correspond to an average of roughly one SNP every 5 kb, 10 kb and 20 kb respectively.

To estimate the power we generated 1000 case-control studies using the causal SNP and specified relative risk. To control the type I error at 5%, we also generated 1000 independent sets of phenotype observations (the vector $Y$ in (3)-(5)) under the null hypothesis of no association for each dataset. A p-value for association for each model was found using the likelihood ratio test. The p-values from the null datasets gave an empirical null distribution, from which the 5% quantile could be used to control the type I error at 5%. The proportion of p-values from the 1000 non-null case-control studies

below the 5% empirical null p-value threshold gives an estimate of power for each method.

We repeated this process a number of times for each combination of causal SNP MAF, density, and disease model. We grouped together results for simulated datasets with the same causal SNP MAF and similar mean $|r|$ between marker SNPs and causal SNP values, calculating the mean of the power estimates for each grouping.

## Haplotype modelling

To test whether including the adjacent interaction terms improved the modelling of which haplotypes were carried by different individuals, we fitted multinomial logistic models. For a specific set of marker SNPs, let haplotypes 1,..., $h$ be the unique haplotypes observed, with $p_1,...,p_h$ the associated haplotype frequencies, and $n_1,...,n_h$ the number of each observed in the dataset. The observed haplotypes can be regarded as a categorical random variable, with the multinomial log-likelihood for the haplotype frequencies being (up to an additive constant):

$$\sum_{i=1}^{h} n_i \log(p_i). \qquad (6)$$

The multinomial logistic model specifies the first haplotype as the reference haplotype, and models the frequencies of the remaining haplotypes in relation to the first:

$$\log\left(\frac{p_i}{p_1}\right) = x_i^T \beta \; i = 2,...,h, \qquad (7)$$

where $x_i^T$ is the $i$th row of the design matrix corresponding to haplotype $i$, and $\beta$ is a set of parameters to be estimated.

$p_1$ is set to be $\frac{1}{1+\sum_{j=2}^{h} \exp\left(x_j^T \beta\right)}$, with $p_i$ being

$\frac{\exp(x_i^T \beta)}{1+\sum_{j=2}^{h} \exp\left(x_j^T \beta\right)}$.

This constrains $(p_1...,p_h)$ to be a probability distribution. After maximizing the log-likelihood under the main-effects model and main+adj model, we tested the improvement of fit using the likelihood ratio test. Under the null hypothesis that the main-effects alone determine the haplotype frequencies, the likelihood ratio statistic is distributed asymptotically as a $\chi_p^2$ random variable, where $p$ is the number of adjacent interactions included. If the haplotype frequencies are not solely determined by the main-effects, the likelihood ratio statistic will tend to be higher. As an alternative measure

of parsimony, we also calculate the reduction in the Akaike information criterion (AIC).

## Applying to unphased data

Applying the multimarker models to unphased data requires us to use an adjusted model which fits directly, or to infer the underlying phased data and apply the model to that, adjusting for uncertainty in the haplotype distribution. Using the latter approach, adjustment is not needed for the main-effects model, since uncertainty in the haplotype distribution makes no difference to the main-effects. For the main+adj and haplotype models, we used expectation-substitution (ES). For each individual, estimated probabilities of different haplotype pairs are found (using a haplotype phasing method, e.g. the EM algorithm), and the resulting design matrix row for each possible haplotype pair found. The design matrix row used in the GLM is the expectation over the distribution of possible haplotype pairs.

ES makes several assumptions, as described in Stram et al. [30]. All the assumptions are satisfied if the genotype information alone determines the haplotype distribution (as opposed to the genotype and phenotype). This is valid under the null hypothesis of no association. ES has been assessed for correct type I error and bias and found to perform as well as more computationally intensive methods such as multiple imputation or a full likelihood approach [31,32].

We have previously compared the ES approach using estimates of phased haplotype frequencies from a straightforward EM approach [33] with estimates from the more sophisticated Beagle package [23]. Using phased haplotype estimates from Beagle made little or no difference to either the type I error rate or power of any of the multimarker models for the MGS dataset (unpublished data). This is despite the improvement in phasing accuracy.

As mentioned above, we used the likelihood-ratio test to assess significance when the data was in phased format. Kent [34] considers the distribution of the LR statistic under model mis-specification. In that case, the distribution of the LR statistic under the null is equivalent to:

$$\sum_{i=1}^{p} \mu_i Z_i^2 \qquad (8)$$

where $Z_i$ are independent N(0,1) random variables, and $\mu_i \; i \in \{1,...,p\}$ are the eigenvalues of the matrix $H^{-1}J$. Here, J is the covariance of the score vector, and H is the expected score derivative matrix:

$$J(\theta) = \int U(\gamma;\theta)U(\gamma;\theta)^T f(\gamma;\theta)d\gamma \qquad (9)$$

$$H = -\int \frac{\partial U(\gamma;\theta)}{\partial \theta^T} f(\gamma;\theta)d\gamma, \qquad (10)$$

Where $\theta$ are the parameters of the model being assessed, $f(\gamma;\theta)$ is the likelihood function, and $U(\gamma;\theta)$ is the score vector, the vector of partial derivatives of f with respect to each entry of $\theta$.

Assuming the haplotype frequencies are correct, and the model is correctly specified and satisfies certain regularity conditions, $(H^{-1}J) = I$, and so the LR statistic is asymptotically distributed as $\chi_p^2$ under the null. If the haplotype frequencies are incorrect, it is likely that the distribution will deviate from $\chi_p^2$.

Kent [34] proposes alternative forms of the score and Wald statistics which have the correct asymptotic distribution under the null (i.e. $\chi_p^2$ instead of that given in (8)), even if the model is misspecified. There is no robust version of the likelihood-ratio statistic, but a robust version of the score statistic is:

$$S = U^T(\theta_0)J^{-1}(\theta_0)U(\theta_0) \qquad (11)$$

Under the null, phenotype information gives no information about the haplotype distribution, so the expected score over the inferred haplotype distribution can be used instead of U above. In addition, since we are considering a binomial distribution for the phenotype, which is an exponential family, the covariance of the score is equivalent to the Fisher information. This implies that the score test as proposed by Rao [35] is robust to model mis-specification (e.g. errors in inferred haplotype distribution), whereas the LR test is not. Thus, for analysing unphased data using the ES approach, we used the score test instead of the LR test.

If one or more of the SNPs are completely determined by a linear function of the other SNPs, the generalised inverse must be used in the score statistic. Calculating the generalised inverse of a matrix requires more computation than the inverse. An alternative is to remove SNPs from the dataset which are determined from nearby SNPs. This does not lose much information, since the dropped SNPs are determined from the retained SNPs, but allows quicker analysis.

All of the results in this paper assume a case-control study design. Although this design is by far the most common for modern GWAS, family-based designs are still used. Cordell and Clayton [36] provide an excellent discussion of applying logistic regression models to family-based designs.

## Population stratification simulation

For the population stratification simulation, we considered data consisting of seven possible 3-SNP haplotypes for individuals from one of two populations. In the first population, the most common haplotype had underlying frequency $0.3 + \delta$, and the next most common had frequency $0.3 - \delta$. In the second population, both had frequency 0.3. The other five haplotypes had the same frequency in both populations, which ranged between 0.05 and 0.15. We took $\delta \in \{0.02, 0.04, 0.06, 0.08, 0.1\}$. The haplotypes were chosen so that the first SNP displayed population stratification, but not the other two.

Wright's statistic, $F_{ST}$, is used as a measure of the genetic distance between populations. We estimated $F_{ST}$ by assuming that the allele frequencies of SNP 1 in the two subpopulations followed the Balding-Nichols model [37]. In that case, the variance of the subpopulation's allele frequencies would be $p(1 - p)F_{ST}$, where $p$ is the background allele frequency; we calculated the actual variance, and then estimated $F_{ST}$ as the ratio of actual variance to $p(1 - p)$. According to Cavalli-Sforza et al. [38], a $F_{ST}$ of 0.01 is equivalent to divergence between different European populations. Table 9 gives estimates of $F_{ST}$ for SNP 1, for each value of $\delta$.

Each individual was a case with probability $0.5 + \mu$ if from the first population, $0.5 - \mu$ if from the second. We took $\mu \in \{0.02, 0.04, 0.06, 0.08, 0.1\}$. This process produced data where the disease is not actually associated with any of the 3 SNPs, but may appear to be due to population stratification.

For each combination of $\delta$ and $\mu$, we simulated 5,000 datasets. Each dataset contained 1,000 haplotypes from population 1, and 1,000 from population 2. The haplotypes were paired randomly within populations to form unphased genotypes, and the individuals were assigned case/control status randomly according to their population. We then calculated the score statistic for association for the main-effects, main+adj, and haplotype omnibus models.

To assess the inflation in the distribution of test statistics, we calculated $\hat{\lambda}$, a generalised version of the genome inflation factor from Devlin and Roeder [39]:

$$\hat{\lambda} = \frac{\text{Median of observed test statistics}}{\text{Median of expected test statistics under null of no association}} \qquad (12)$$

$\hat{\lambda}$ gives an idea of the inflation in the median of a distribution, but for interpretations beyond this, may not be comparable between models with different degrees of freedom.

## Quality control used in MGS GWAS

Quality control (QC) is extremely important for reducing biases in an analysis of a GWAS [10,40]. The

**Table 9 Estimated $F_{ST}$ values**

| $\delta$ | $F_{ST}$ |
|---|---|
| 0.02 | $4.40 \times 10^{-4}$ |
| 0.04 | $1.76 \times 10^{-3}$ |
| 0.06 | $3.96 \times 10^{-3}$ |
| 0.08 | $7.03 \times 10^{-3}$ |
| 0.1 | $1.10 \times 10^{-2}$ |

Table of the estimated $F_{ST}$ parameters for different values of $\delta$.

paper summarising the results from the WTCCC study provides a useful summary of quality control measures that were used in the study [41]. Some biases are potentially more severe when analysing markers together such as in a multimarker model approach. One example is differential genotyping error, discussed by Clayton et al. [42], in which genotyping errors for a particular SNP occur at different rates in case and controls. Although this can impact a single-SNP analysis, it has the potential to be more severe in a multimarker analysis. If a particular SNP is affected by genotyping error, it can lead to a new haplotype being inferred from the data which actually does not exist. If the genotyping error differs between cases and controls, the spurious haplotype may appear just in cases, or just in controls. This may lead to the new haplotype appearing to be highly associated with phenotype, and thus a false positive result.

Generally, as in a single-SNP analysis, the best approach to use is to attempt to filter out as much low quality data before the analysis whilst keeping data which may be the result of a true association. Most QC checks use straightforward statistics or statistical tests, such as missing data per individual, missing data per SNP, whether a SNP deviates significantly from HWE, whether missingness at a SNP depends on the phenotype, etc. There is a limit as to what can be done before the study, however. Some tests are too computationally intensive to be performed on large numbers of SNPs. An example of this would be testing whether haplotypes are in HWE.

A second problem with pre-study QC for a GWAS is the significance threshold to use for p-values. Many tests will have p-values of < 0.05 by chance, even if they are from high quality data. Setting a very strict p-value threshold (e.g. 0.05) will mean many SNPs, and potentially some causal variants, will be filtered out; setting a relaxed threshold may cause more false positives to appear in the analysis.

One way to overcome this is to examine the distribution of QC test statistics in the most significant tests for disease association. If the distribution in the top results is significantly different to that expected under the null of no association, we should be concerned. If a

particular multimarker association statistic is highly significant, but contains several SNPs which were close to failing the pre-study QC check, it may be appropriate to dismiss the apparent significant result as being a false positive.

In addition to the QC filters used in the MGS study, we applied further steps:

• Exclude SNPs which have more than 3% missing data. This is a stricter threshold than the 5% level used in the MGS study, and reflects that multimarker analysis may be more severely affected by missing data than single-SNP analysis.
• Exclude SNPs for which missing data rates are significantly different in cases and controls (p-value < 0.0001).
• Filter out SNPs which are almost completely determined by nearby SNPs. This is equivalent to taking a tag set of SNPs, and can be done in PLINK [20] using the -indep command.

In addition, the windows with most significant association statistics were tested for haplotype-HWE by using the approach given for multi-allele SNPs in Weir [43].

An alternative to pruning SNPs in high LD with surroufinding SNPs is to t the multimarker models to every window, but use the generalised inverse instead of the standard inverse. Although this approach is valid, it takes longer because: 1) the generalised inverse takes longer to calculate than the standard inverse; and 2) many more windows must be tested. Because we are interested in regions that show association with multimarker models, analysing a sparser dataset loses little of the information, and takes a significantly lower amount of computation to do.

## Additional material

**Additional file 1: Imputation results (1)**. Imputation results from first chromosome 9 region.
**Additional file 2: Imputation results (2)**. Imputation results from second chromosome 9 region.

### Author details
[1]MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 0SR, UK. [2]Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, UK.

## Authors' contributions
JMSW conceived and performed the analysis, and wrote the manuscript. FD supervised the analysis, and wrote the manuscript. Both authors read and approved the final manuscript.

## References
1. National institute of health GWAS website. [http://www.genome.gov/GWAStudies/].
2. Maher B: **The case of the missing heritability.** *Nature* 2008, **456**:18-21.
3. McClellan J, Sussier E, King M: **Schizophrenia: a common disease caused by multiple rare alleles.** *British Journal of Psychiatry* 2007, **190**:194-199.
4. Browning S: **Missing data imputation and haplotype phase inference for genome-wide association studies.** *Hum Genet* 2008, **124**:439-450.
5. 1000 Genomes project website. [http://www.1000genomes.org].
6. Dudbridge F, Gusnanto A: **Estimation of significance thresholds for genomewide association scans.** *Genetic Epidemiology* 2008, **32**:227-234.
7. Hoggart C, Clark T, De Iorio M, Whittaker J, Balding D: **Genome-wide significance for dense SNP and resequencing data.** *Genetic Epidemiology* 2008, **32**:179-185.
8. Chapman J, Cooper JD, Todd TA, Clayton D: **Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power.** *Hum Hered* 2003, **56**:18-31.
9. Clayton D, Chapman J, Cooper J: **Use of unphased multilocus genotype data in indirect association studies.** *Genetic Epidemiology* 2004, **27**:415-428.
10. Balding D: **A tutorial on statistical methods for population association studies.** *Nature Reviews Genetics* 2006, **7**:781-791.
11. Kim Y, Feng S, Zeng Z: **Measuring and partitioning the high-order linkage disequilibrium by multiple order Markov chains.** *Genetic Epidemiology* 2008, **32**:301-312.
12. Humphreys K, Iles M: **Fine-scale mapping in case-control samples using locus scoring and haplotype-sharing methods.** *BMC Genet* 2005, **6(Suppl 1)**:1-5.
13. Clayton D, Jones H: **Transmission/Disequilibrium tests for extended marker haplotypes.** *Am J Hum Gen* 1999, **65**:1161-1169.
14. Price A, Patterson N, *et al*: **Principal componenets analysis corrects for stratification in genome-wide association studies.** *Nature genetics* 2006, **38**:904-909.
15. Bacanu S, Devlin B, Roeder K: **The power of genomic control.** *Am J Hum Gen* 2000, **66**:1933-1944.
16. Shi J, Levinson D, Duan J, *et al*: **Common variants on chromosome 6p22.1 are associated with schizophrenia.** *Nature* 2009, **460**:753-757.
17. Schaffner S, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D: **Calibrating a coalescent simulation of human genome sequence variation.** *Genome Research* 2005, **15**:1576-1583.
18. Zaykin D, Westfall P, Young S, Karnoub M, Wagner M, Ehm M: **Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals.** *Hum Hered* 2001, **53**:79-91.
19. Rakovski C, Stram D: **A kinship-based modification of the Armitage trend test to address hidden population structure and small differential genotyping errors.** *PLoS one* 2009, **4**:e5825.
20. Purcell S, *et al*: **PLINK: a toolset for whole-genome association and population-based linkage analysis.** *Am J Hum Gen* 2007, **81**:559-575.
21. Astle W, Balding D: **Population structure and cryptic relatedness in genetic association studies.** *Statistical Science* 2009, **24**:451-471.
22. Dudbridge F: **Likelihood-based association analysis for nuclear and unrelated subjects with missing genotype data.** *Human Heredity* 2008, **66**:87-98.
23. Browning S, Browning B: **Rapid and accurate haplotype phasing and missing data inference for whole-genome association studies by use of localized haplotype clustering.** *Am J Hum Gen* 2007, **81**:1084-1097.
24. Chapman J, Clayton D: **One degree of freedom for dominance in indirect association studies.** *Genetic Epidemiology* 2007, **31**:261-271.
25. Peduzzi P, Concato J, Kemper E, Holford T, Feinstein A: **A simulation study of the number of events per variable in logistic regression analysis.** *J Clin Epidemiol* 1996, **12**:1373-1379.
26. Durrant C, Zondervan K, Cardon L, Hunt S, Deloukas P, Morris A: **Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes.** *Am J Hum Gen* 2004, **75**:35-43.
27. Waldron E, Whittaker J, Balding D: **Fine mapping of disease genes via haplotype clustering.** *Genetic Epidemiology* 2005, **30**:170-179.
28. Morris A: **A flexible Bayesian framework for modelling haplotype association with disease, allowing for dominance effects of the underlying causative variants.** *Am J Hum Genet* 2006, **79**:679-694.
29. Tachmazidou I, Verzilli C, De Iorio M: **Genetic association mapping via evolution-based clustering of haplotypes.** *PLoS Genetics* 2004, **3**:1163-1177.
30. Stram D, *et al*: **Modeling and E-M estimation of haplotype specific relative risks from genotype data for a case-control study of unrelated individuals.** *Hum Hered* 2003, **55**:179-190.
31. Kraft P, Cox D, Paynter R, Hunter D, De Vivo I: **Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques.** *Genetic Epidemiology* 2005, **28**:261-272.
32. Cordell H: **Estimation and testing of genotype and haplotype effects in case-control studies: comparison of weighted regression and multiple imputation procedures.** *Genetic Epidemiology* 2006, **30**:259-275.
33. Excoffier L, Slatkin M: **Maximum likelihood estimation of molecular haplotype frequencies in a diploid population.** *Mol Biol Evol* 1995, **12**:921-927.
34. Kent J: **Robust properties of likelihood ratio tests.** *Biometrika* 1982, **69**:19-27.
35. Rao CR: **Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation.** *Proc Camb Phil* 1948, **44**:50.
36. Cordell H, Clayton D: **A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes.** *Am J Hum Genet* 2002, **70**:124-141.
37. Balding D, Nichols R: **A method for quantifying differentiation between populations a multi-allelicloci and its implications for investigating identity and paternity.** *Genetica* 1995, **96**:3-12.
38. Cavalli-Sforza L, Menozzi P, Piazza A: *The history and geography of human genes* Princeton university press 1994.
39. Devlin B, Roeder K: **Genomic control for association studies.** *Biometrics* 1999, **55**:997-1004.
40. McCarthy M, Abecasis G, *et al*: **Genome-wide association studies for complex traits: consensus, uncertainty and challenges.** *Nature Reviews Genetics* 2008, **9**:356-369.
41. **Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls.** *Nature* 2007, **447**:661-679.
42. Clayon D, Walker N, *et al*: **Population structure, differential bias and genomic control in a large-scale, case-control association study.** *Nature Genetics* 2005, **37**:1243-1246.
43. Weir B: *Genetic data analysis II: methods for discrete population genetic data* Sinaur associates 1996.