Published in final edited form as:

Ann Hum Genet. 2008 May ; 72(Pt 3): 368–374. doi:10.1111/j.1469-1809.2007.00422.x.

Perturbation analysis: a simple method for filtering SNPs with erroneous genotyping in genome-wide association studies

YY Teo^{1,*}, KS Small¹, TG Clark^{1,2}, and DP Kwiatkowski^{1,2}

¹ Wellcome Trust Centre for Human Genetics, University of Oxford, United Kingdom

² Wellcome Trust Sanger Institute, Hinxton, United Kingdom

SUMMARY

We introduce a simple and yet scientifically objective criterion for identifying SNPs with genotyping errors due to poor clustering. This yields a metric for assessing the stability of the assigned genotypes by evaluating the extent of discordance between the calls made with the unperturbed and perturbed intensities. The efficacy of the metric is evaluated by: (1) estimating the extent of over-dispersion of the Hardy-Weinberg equilibrium chi-square test statistics; (2) an interim case-control study, where we investigated the efficacy of the introduced metric and standard quality control filters in reducing the number of SNPs with evidence of phenotypic association which are attributed to genotyping errors; (3) investigating the call and concordance rates of SNPs identified by perturbation analysis which have been genotyped on both Affymetrix and Illumina platforms. Removing SNPs identified by the extent of discordance can reduce the degree of over-dispersion of the HWE test statistic. Sensible use of perturbation analysis in an association study can correctly identify SNPs with problematic genotyping, reducing the number required for visual inspection. SNPs identified by perturbation analysis had lower call and concordance rates, and removal of these SNPs significantly improved the performance for the remaining SNPs.

Keywords

genotyping errors; genome-wide association studies; genotype calling

INTRODUCTION

Large-scale genotyping of hundreds of thousands of genetic variants typically rely on automated calling algorithms for assigning genotypes based on the observed hybridization intensities (Di *et al.* 2005; Rabbee & Speed 2006; Affymetrix Inc 2006; Plagnol *et al.* 2007; The Wellcome Trust Case Control Consortium 2007). Due to the scale involved, the assigned genotypes are usually not manually curated and thus any erroneous genotyping will not be detected. Such erroneous genotype assignments have the potential to introduce artifacts of false positive associations, especially in case-control designs where genotypes for the two phenotype groups may be called independently which introduces differential biases where subtle shifts in the genotype clouds can result in dramatically different genotype distributions (Clayton *et al.* 2005). Even sophisticated genotype calling algorithms which pool information across different groups but procedurally call them separately can produce erroneous calls (Plagnol *et al.* 2007; The Wellcome Trust Case Control Consortium

* Corresponding author: Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, United Kingdom.
teo@well.ox.ac.uk.

2007). In related analyses for identifying the extent of population structure and haplotype phasing, it is also important to use a set of highly accurate genotype calls to avoid introducing biases, especially when comparing against publicly-available datasets with high-quality genotypes, for example from the HapMap project (International HapMap Consortium 2005).

Visual inspection of the assigned calls for SNPs with significant association p-values has been advocated to eliminate the possibility of false positive association due to genotyping errors (The Wellcome Trust Case Control Consortium 2007). In a genome-wide scan involving at least half a million SNPs, this can translate to visually ascertaining the assigned genotypes for hundreds of clusterplots (a clusterplot is a visual representation of the allelic intensities for a collection of individuals at a SNP where colours are typically used in the plot to differentiate between the assigned genotypes). This process of visual inspection is tedious but ultimately manageable. However it should be noted that the process of manually inspecting clusterplots is subject to enormous variations between researchers. As verdicts on the assessment of call accuracy can fluctuate between different researchers who may impose their own criteria for accepting or rejecting the assigned genotypes after checking the clusterplots, there is scope for the development of scientifically rigorous approaches (Plagnol *et al.* 2007) which will at least minimize the number of clusterplots that require visual inspection. Visually verifying the clusterplot for each SNP, however, will not be practical for identifying a panel of SNPs with accurate genotypes for assessing population structure, phasing of haplotypes and comparing the extent of linkage disequilibrium (LD) across populations.

A number of filters based on Mendelian inheritance errors (where applicable), conformity to Hardy-Weinberg equilibrium (HWE), the extent of missingness and minor allele frequencies have been implemented (The Wellcome Trust Case Control Consortium 2007) to remove SNPs which may be prone to erroneous genotyping. Typically these filters work by removing SNPs which fall into the extremes of each criterion. For example, the Wellcome Trust Case Control Consortium (WTCCC) removed a SNP if at least one of the following condition was true: (i) HWE test statistic > 25 (or p-value $< 5.7 \times 10^{-7}$); (ii) call rate $< 95\%$ if minor allele frequency $> 5\%$; (iii) call rate $< 99\%$ if minor allele frequency $< 5\%$. Such filter-by-extreme procedures do not actually remove all the SNPs with problematic genotyping, and in fact the WTCCC had to use more stringent filtering criteria to create a 'very clean' set of data for the assessment of over-dispersion caused by population structure.

For most of the genotyping platforms, the hybridization intensity data can be summarized through well-designed normalization schemes to yield a pair of coordinates for each individual, which essentially correspond to a measure of allelic expression for each of the two alleles (see Rabbee & Speed (2006), Affymetrix Inc (2006), Plagnol *et al.* (2007) and the WTCCC (2007) for Affymetrix GeneChip platforms, Kermani (2006) for Illumina BeadArray platforms). For most of the SNPs, assigning genotypes to these normalized intensity data is straightforward as the genotype clouds are clearly distinct (Fig 1a). However, this can be problematic for a small fraction of SNPs due to overlapping genotype clusters (Fig 1b). Most of the available algorithms rely on comparing particular characteristics of the genotype clusters in assigning calls, and the occasional shifts in genotype clouds for particular SNPs can result in high-confidence assignments of genotypes which are clearly erroneous (Teo *et al.* 2007a).

In this report, we propose a simple mathematical procedure to produce a metric for quantifying the quality of the assigned genotypes at each SNP. By perturbing the allelic intensities of each individual slightly, we can assess the stability of the calls for each SNP by comparing the concordance between two runs of the calling algorithm using the perturbed

and unperturbed intensities. This method is based on the premise that SNPs with clearly distinct genotype clouds will result in stable and identical calls between both runs, while SNPs with overlapping clusters may yield higher rates of discordance. This procedure mimics the decision-making process of an individual who is manually assessing the clusterplots, by flagging and rejecting SNPs with signals of putative associations and non-distinct genotype clouds. We evaluate the performance of this metric with genotype data from the Affymetrix GeneChip 500K and the Illumina HumanHap650Y BeadArray platforms.

METHODS

Data Set

DNA samples from 2,370 individuals (972 cases and 1,398 controls) have been genotyped on the Affymetrix GeneChip 500K platform as part of a case-control study on malaria. Of these individuals, 252 cases have also been genotyped on the Illumina HumanHap650Y BeadChip array. The raw intensity data (CEL files) from the Affymetrix experiment have been normalized and processed with the scheme identical to that used by the WTCCC (2007). The raw data from the Illumina experiment have been normalized and processed using the scheme implemented in the BeadStudio suite [8], which yields a pair of allelic signals for each individual at each SNP.

Suppose x and y denote the final form of the allelic intensities of the two pre-designated alleles for an individual at a SNP prior to calling, we can add a perturbation factor by the introduction of white noise to each of the two values

$$\begin{aligned} x' &= x + \varepsilon_1 \\ y' &= y + \varepsilon_2 \end{aligned}$$

where ε_1 and ε_2 are independent and identically distributed as Gaussian distributions $\mathcal{N}(0, s^2)$. While the concept of perturbation is generic across genotyping platforms, the exact choice of s^2 will depend on the calling algorithm (and thus the scale) used. In our analyses for the Affymetrix data, we have chosen s to take values which correspond approximately to 2%, 10% and 20% of the across-platform mean signal intensity of allele A for the BB genotype cloud (note that the definition of alleles A and B are interchangeable).

Genotypes for the Affymetrix data have been called using Xtyping (Plagnol *et al.* 2007) and Chiamo (The Wellcome Trust Case Control Consortium 2007). Both algorithms produced posterior probabilities for the three valid genotypes. We employed the fuzzy calling procedure described by Plagnol and colleagues for Xtyping to obtain the genotype calls (Plagnol *et al.* 2007). Chiamo calls are obtained by thresholding the posterior probabilities using a threshold of 0.90, where a missing genotype is assigned if the maximum posterior probability across the three valid genotypes is less than 0.90. Genotypes for the Illumina data have been assigned using GenCall – a proprietary calling algorithm distributed by Illumina as part of their BeadStudio software (Peiffer *et al.* 2006). A filter of 0.2 is used to threshold the confidence score associated with each genotype call, and a missing genotype is assigned if the confidence score is less than 0.2.

RESULTS

We performed a series of analyses for the SNPs on five chromosomes in order to assess the performance of perturbation analysis. Specifically for the Affymetrix data, we calculated the HWE chi-square test statistic at each SNP for the 1,398 controls and measured the extent of

over-dispersion in a manner similar to genomic control (Devlin & Roeder 1999). This statistic was first calculated across all 47,012 SNPs (unfiltered), and then calculated for SNPs that remained after applying standard filters (35,315 and 40,188 for Chiamo and Xtyping respectively). The filters involved removing SNPs with: (i) call rate < 95%; (ii) minor allele frequency < 1%; (iii) χ^2 test statistic for HWE > 25. We also measured the extent of over-dispersion for SNPs that remained after applying only a perturbation analysis discordance filter. SNPs with greater than 1%, 2%, 5% and 10% discordant calls using the perturbed and unperturbed intensity data are identified and removed. We used the extent of over-dispersion of the HWE chi-square test statistic as a surrogate for the quality of SNPs which remained (see Table 1). The inflation factors for the unfiltered and filtered (using standard filters) SNPs called using Chiamo were 1.44 and 1.40 respectively, and the corresponding figures for genotypes called using Xtyping were 2.22 and 1.80. We caution against using these figures as a comparison of the two genotype calling algorithms. There is a clear trend that as the threshold for discordance decreases, the inflation factors converge towards the expected value of 1. For example, for SNPs with 10% perturbation, the inflation factor was 1.23 and 1.62 for Chiamo and Xtyping respectively after removing SNPs with more than 2% discordant genotypes (21,756 and 10,731 SNPs removed respectively). This relationship is observed in both situations where the genotypes have been called using Chiamo and Xtyping. When the degree of the perturbation is large (i.e. 20% of the mean signal intensity of allele *A* for the *BB* genotype), more than 75% of the SNPs are removed and this may be excessive and undesirable (see Discussion). We are also interested to find out what proportion of SNPs identified by the standard filters has been identified by perturbation analysis (Table 2). By measuring the degree of over-dispersion for SNPs identified by perturbation analysis but not by standard QC filters, we see that the additional SNPs identified by perturbation analysis in general have inflated test statistics, suggesting that most of these SNPs experience systematic biases which may be attributed to genotyping errors. The SNPs identified by the standard QC filters but not by perturbation analysis instead showed under-dispersion, suggesting that most of these SNPs conformed to HWE, although a small fraction of these SNPs display gross departures to HWE (for example, at a discordance threshold of 2%, slightly more than 10% of these SNPs have χ^2 test statistic > 30). This suggests that the use of perturbation analysis should optimally be complemented by the use of sensible QC filters.

In addition to the identification of a set of SNPs with accurate and stable genotypes for use in downstream analyses, we also assessed the use of perturbation analysis in reducing the number of SNPs in an association study with significant p-values which are attributed to genotyping errors. In an interim analysis across 47,012 SNPs to identify the genetic variants associated with increased susceptibility to and protection from severe malaria, we removed any SNP with more than 10% of the genotypes missing, minor allele frequency less than 1% or has a HWE test statistic of greater than 25 for the control cohort. We have tolerated a higher rate of missingness (10% rather than the conventional 5%) due to the use of whole genome amplified DNA which has been found to result in greater extent of missingness (Teo *et al.* – in review). We assessed the stability of the assigned genotypes from Chiamo by perturbation analysis with an *s* of 0.05, which corresponded to approximately 10% of the mean signal intensity of allele *A* for the *BB* genotype cloud. If we consider 10^{-4} as the significance threshold, this yielded 105 SNPs as putatively associated with the phenotype (Fig 2a). These SNPs thus require subsequent manual verification of the genotype calls to eliminate the possibility that the association is due to erroneous genotyping. The clusterplots for these 105 SNPs have been assessed independently and 10 SNPs out of the 105 were found to have unambiguous genotypes. The remaining 95 SNPs with significant evidence of phenotypic association had questionable genotyping. By removing any SNP with a discordance of at least 5% between the perturbed and unperturbed calls for either the case or control cohorts, this number decreased to 31 SNPs. This set contained all 10 SNPs with

unequivocal genotyping while removing 74 SNPs with questionable genotyping (Fig 2b). Increasing the stringency of the discordance threshold to 2% further reduced the number of SNPs to be inspected to 11, of which the 10 SNPs with accurate genotyping were a subset of (Fig 2c).

We also evaluated the use of perturbation analysis on data from the Illumina HumanHap650Y genotyping array. Given that there are only 252 samples which have been genotyped on the Illumina HumanHap650Y, we chose instead to investigate the effects of perturbation analysis on the concordance and call rates of the assigned genotypes. As these 252 samples have also been genotyped on the Affymetrix array, we identified 9,723 SNPs in five chromosomes which are common across both arrays and compared the concordance and call rates of the assigned genotypes across both platforms. The Chiamo call rate for all the common SNPs on the Affymetrix platform was 96.165%. The valid genotype calls have a concordance rate of 95.626% with the calls made by GenCall on the Illumina data. Running perturbation analysis with an s of 0.05 on the Affymetrix intensities, we identified SNPs with discordances greater than 2% between the perturbed and unperturbed calls. The call and concordance rates of the identified SNPs were significantly lower at 92.326% and 94.689% respectively. Removing these SNPs increased the call and concordance rates of the remaining SNPs to 99.139% and 96.352%.

DISCUSSION

Automated algorithms for assigning genotypes based on hybridization intensities can produce erroneous genotype calls. We have introduced a simple procedure for assessing the sensitivity of the assigned genotypes to small perturbations in the normalized intensities. The extent of discordance between the genotypes called using the unperturbed and perturbed intensities has been shown to be a useful metric for assessing the quality of the assigned genotypes. This is expected to be useful for identifying a set of SNPs with accurate genotypes which can be used for downstream analyses of population structure, haplotype phasing and linkage disequilibrium mapping. The metric has also been shown to be effective at reducing the number of SNPs which require visual inspection of the clusterplots in an association study. We have also shown using data from two independent genotyping platforms that the effectiveness of perturbation analysis can be attributed to the ability to identify SNPs with lower call accuracy.

The use of perturbation analysis is particularly powerful in identifying SNPs with overlapping genotype clouds (as seen in Fig 1). However, as the underlying concept of perturbation analysis relies on the design of the calling algorithms, we do not expect the procedure to identify SNPs in regions of copy number variations, tri-allelic (or quad-allelic) SNPs, or SNPs which map to multiple polymorphisms in the genome unless the calling algorithm can specifically identify these features. Calling genotypes for different phenotype groups separately can potentially produce significant association when one of the groups has been called erroneously, and perturbation analysis may not be as effective for such errors. As perturbation analysis fundamentally depends on the relative positions of the genotype clusters, the efficacy of the approach may also be limited in situations where genotyping errors are attributed to sparse clusters (for SNPs with rare alleles) or extreme positional shifts in the clusters.

The use of overly-stringent filtering procedures can affect the power of an association study since this increases the likelihood of removing SNPs with accurate genotyping. In our application of perturbation analysis, we have seen that if the degree of perturbation is high and the discordance threshold is strict, more than 75% of the data may be removed. The choice of the discordance threshold and the degree of perturbation are thus not independent.

From our observations with data from the Affymetrix and Illumina platforms, we recommend the use of 10% perturbation of the normalized intensity data. The choice of the discordance threshold depends on the context of the analysis. In assessing SNPs for phenotypic association, a stringent discordance threshold (say 2%) may be employed to minimize the number of SNPs with significant evidence of association that are attributed solely to genotyping errors. However, in scenarios where we are interested to identify a set of SNPs with minimum genotyping errors for use in downstream analysis related to population structure and LD patterns, a less stringent discordance threshold (say 5% or 10%) may be employed.

We have measured the over-dispersion of the HWE test statistic as a criterion for assessing the efficacy of perturbation analysis compared to the standard QC filters. This may appear as circular, since departure from HWE is one of the standard QC filter used. However, typical usage of HWE as a QC criterion only filters SNPs with gross departures from HWE, and this does not contribute significantly in correcting the extent of over-dispersion, other than to remove SNPs at the upper extreme tail of the distribution of the HWE values and to shift the median of this distribution downwards very marginally. This means that the HWE QC criterion does not actually remove SNPs with subtle or moderate genotyping errors. As some of the recently introduced model-based calling algorithms (Plagnol *et al.* 2007; The Wellcome Trust Case Control Consortium 2007) incorporate some form of weak adherence to HWE in the models, SNPs with genotyping errors may not necessarily display gross departures from HWE.

The option for performing perturbation analysis has been included as part of the procedure for a novel genotype calling algorithm designed for the Illumina BeadArray platforms (Teo *et al.* 2007b), and the Xtyping algorithm for Affymetrix GeneChip arrays (Plagnol *et al.* 2007).

Perturbation analysis is a simple addition to the arsenal of tools for identifying and removing SNPs with erroneous genotyping. This method will be helpful in situations where there is a need to identify a set of SNPs with accurate genotypes. In association studies, perturbation analysis can reduce the number of SNPs which require manual inspection of the assigned genotypes. We recommend that, in addition to the standard filters using conformity to Hardy-Weinberg equilibrium, extent of missing genotypes and minor allele frequencies, perturbation analysis be used to identify SNPs with problematic genotypes. A comprehensive set of analytical tools for identifying SNPs with erroneous genotyping is especially relevant as the science moves into the next phase involving genotyping arrays with the capability to assay millions of genetic variants.

Acknowledgments

We are grateful to two anonymous reviewers for their constructive comments which have helped to improve this article. This work has been made possible with fundings from the Grand Challenges in Global Health initiative (Gates Foundation, Wellcome Trust and the FNIH). In addition, DPK also acknowledges fundings from the UK Medical Research Council.

REFERENCES

- Affymetrix Inc. BRLMM: an improved genotype calling method for the GeneChip Human Mapping 500K Array Set. 2006 http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf
- Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JM, Faham M, Moorhead M, Jones HB, Falkowski M,

- Hardenbol P, Willis TD, Todd JA. Population structure, differential bias and genomic control in a large-scale case-control association study. *Nat Genet.* 2005; 37:1243–1246. [PubMed: 16228001]
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics.* 1999; 55:997–1004. [PubMed: 11315092]
- Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, Shen MM, Kulp D, Kennedy GC, Mei R, Jones KW, Cawley S. Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics.* 2005; 21:1958–1963. [PubMed: 15657097]
- International HapMap Consortium. A haplotype map of the human genome. *Nature.* 2005; 437:1299–1320. [PubMed: 16255080]
- Kermani, BG. Artificial intelligence and global normalization methods for genotyping. US Patent 20060224529. 2006.
- Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.* 2006; 16:1136–1148. [PubMed: 16899659]
- Plagnol V, Cooper JD, Todd JA, Clayton DG. A method to address differential bias in genotyping in large scale association studies. *PLoS Genet.* 2007; 3:e74. [PubMed: 17511519]
- Rabbee N, Speed TP. A genotype calling algorithm for Affymetrix SNP arrays. *Bioinformatics.* 2006; 1:7–12. [PubMed: 16267090]
- Teo YY, Fry AE, Clark TG, Tai ES, Seielstad M. On the usage of HWE for identifying genotyping errors. *Ann Hum Genet.* 2007a; 71:701–703. [PubMed: 17388941]
- Teo YY, Inouye M, Small KS, Gwilliam R, Deloukas P, Kwiatkowski DP, Clark TG. A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics.* 2007b in print.
- Teo YY, Inouye M, Small KS, Fry AE, Potter SC, Dunstan SJ, Seielstad M, Barroso I, Wareham NJ, Rockett KA, Kwiatkowski DP, Deloukas P. Performance of whole genome amplified DNA in genome-wide association studies. In review.
- The Wellcome Trust Case Control Consortium. Genomewide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007; 447:661–678. [PubMed: 17554300]

Figure 1.

Clusterplots of the allele hybridization intensities at a SNP: the x-axis represents the signal for the designated *A* allele and the y-axis represents the signal for the designated *B* allele. Each point on the plot represents the hybridization signal for an individual at this SNP. The colour of the point represents the assigned genotype: *AA* (red), *AB* (green), *BB* (blue), NULL (grey). Typically, a NULL genotype is assigned when there is insufficient evidence for a valid genotype call. (a) Clusterplot of a SNP with distinct genotype clouds which is expected to yield stable genotype calls upon perturbation; (b) clusterplot of a SNP with overlapping or indistinct genotype clouds which is expected to produce discordant genotypes when the intensities are perturbed marginally.

Figure 2.

Plot of the significance obtained for a test of association with the phenotype for SNPs across five chromosomes based on the unperturbed data. The y-axis for all three plots represents the negative log-significance of the test. Any point which attains a value above 4 corresponds to a SNP that is putatively associated with the phenotype (points in black), and thus require manual verification of the accuracy of the assigned genotypes. The plots correspond to: **(a)** prior to the implementation of perturbation analysis, where 105 SNPs have been flagged as putative hits; **(b)** after removing any SNP with more than 5% discordant data for the genotypes called using the perturbed and unperturbed intensities; **(c)** after removing any SNP with more than 2% discordance.

Table 1

Genomic control inflation factors for HWE at the controls.

Discordance Threshold	Chiamo			Xtyping		
	2%	10%	20%	2%	10%	20%
Unfiltered	1.44	1.44	1.44	2.22	2.22	2.22
Standard filters	1.40	1.40	1.40	1.80	1.80	1.80
> 0.10	1.40	1.32	1.27	2.17	2.00	1.81
> 0.05	1.34	1.28	1.24	2.10	1.84	1.66
> 0.02	1.27	1.23	1.13	1.94	1.62	1.51
> 0.01	1.21	1.11	-	1.78	1.49	1.49

A total of 47,012 SNPs across chromosomes 18, 19, 20, 21 and 22 on the Affymetrix 500K GeneChip genotyping platform are analysed. The inflation factors are calculated as the median χ^2 test statistic for assessing Hardy-Weinberg equilibrium (HWE) at 1,398 control samples. Standard filters refer to those used by the Wellcome Trust Case Control Consortium [5] which consist of removing SNPs on the basis of call rates, allele frequencies and test statistic for HWE. Specifically, we remove any SNP which belongs in any of the following criteria: (i) call rate < 95% (only relevant for Chiamo); (ii) minor allele frequency < 1%; (iii) χ^2 test statistic for HWE > 25. The columns showed the choices of the perturbation factor (*s*) used in the analysis for each genotyping algorithm. For each algorithm, we calculated the over-dispersion factor after removing SNPs where the discordances are greater than the specified threshold. If more than 75% of the SNPs are removed, the inflation factor is not calculated and a dash is shown instead.

Table 2

Extent of overlaps between SNPs identified by the standard filters (SF) and by perturbation analysis (PA).

Discordance Threshold (Chiamo 10%)	PA proportion common with SF (%)	SF proportion common with PA (%)	Over-dispersion PA not common with SF	Over-dispersion SF not common with PA
> 0.10	92.21	45.37	1.484	0.615
> 0.05	75.12	74.91	1.417	0.102
> 0.02	44.89	83.52	1.456	0.040
> 0.01	30.93	89.41	1.443	0.025

Comparisons of SNPs identified by the standard filters and by perturbation analysis. The second column shows the percentage of SNPs identified by perturbation analysis that has been similarly identified by the standard filters; the third column shows the percentage of SNPs identified by the standard filters which has been similarly identified by perturbation analysis; the fourth column shows the inflation factor for SNPs identified by perturbation analysis but not by the standard filters; the last column shows the inflation factor for SNPs identified by the standard filters but not by perturbation analysis.