BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 15 2011, pages 2104–2111
doi:10.1093/bioinformatics/btr323

Genetics and population analysis

Advance Access publication June 7, 2011

MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects

Harm-Jan Westra[1], Ritsert C. Jansen[2], Rudolf S. N. Fehrmann[1], Gerard J. te Meerman[1], David van Heel[3,†], Cisca Wijmenga[1,†] and Lude Franke[1,3,†,*]

[1]Department of Genetics, University Medical Center Groningen, [2]Groningen Bioinformatics Center, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, 9700AB, Groningen, The Netherlands and [3]Blizard Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Sample mix-ups can arise during sample collection, handling, genotyping or data management. It is unclear how often sample mix-ups occur in genome-wide studies, as there currently are no *post hoc* methods that can identify these mix-ups in unrelated samples. We have therefore developed an algorithm (*MixupMapper*) that can both detect and correct sample mix-ups in genome-wide studies that study gene expression levels.

Results: We applied *MixupMapper* to five publicly available human genetical genomics datasets. On average, 3% of all analyzed samples had been assigned incorrect expression phenotypes: in one of the datasets 23% of the samples had incorrect expression phenotypes. The consequences of sample mix-ups are substantial: when we corrected these sample mix-ups, we identified on average 15% more significant *cis*-expression quantitative trait loci (*cis*-eQTLs). In one dataset, we identified three times as many significant *cis*-eQTLs after correction. Furthermore, we show through simulations that sample mix-ups can lead to an underestimation of the explained heritability of complex traits in genome-wide association datasets.

Availability and implementation: *MixupMapper* is freely available at http://www.genenetwork.nl/mixupmapper/

Contact: lude@ludesign.nl

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on March 15, 2011; revised on May 19, 2011; accepted on May 22, 2011

1 INTRODUCTION

Genome-wide studies have identified many disease-associated variants for a wide plethora of complex human diseases (Hindorff *et al*., 2009) [such as celiac disease (Dubois *et al*., 2010), type 1 diabetes (Barrett *et al*., 2009) and type 2 diabetes (Voight *et al*., 2010), Crohn's disease (Franke *et al*., 2010)], and complex continuous phenotypes [such as lipid levels (Teslovich *et al*., 2010), body mass index (BMI) (Speliotes *et al*., 2010) and height (Lango Allen *et al*., 2010)]. Many of these studies (Dubois *et al*., 2010; Voight *et al*., 2010; Teslovich *et al.,* 2010; Speliotes *et al*., 2010; Lango Allen *et al*., 2010) also assess the effect of the identified genetic variants on gene expression variation [i.e. genetical genomics (Jansen and Nap, 2001)], by mapping expression quantitative trait loci (eQTL). As such, these studies involve many steps before actual analysis of the data, during each of which samples could be accidentally swapped. Since these studies are pushing toward larger sample-sizes in order to be able to identify ever smaller effects, the presence of sample mix-ups becomes almost unavoidable.

It is known from simulations that sample mix-ups can have an effect on the power to detect genetic associations in genome-wide studies (Buyske *et al*., 2009; Gordon and Finch, 2004; Ho and Lange, 2010; Samuels *et al*., 2009; Zheng and Tian, 2005), which may present a problem to detect variants with small effects. However, it is unclear how often such sample mix-ups actually occur in studies investigating gene expression. The common method to detect sample mix-ups in genome-wide association studies (GWAS) is to check for heterozygous genotypes for X-chromosomal markers in males. However, this procedure will not identify sample mix-ups between samples of identical gender. While it is also possible to use multiple phenotypes that can be well predicted based on genetic markers [such as eye color and hair color (Sulem *et al*., 2007) and ABO blood group (Yip, 2002)], we are not aware of any study where this has been applied to identify sample mix-ups in GWAS. It is obvious that if there would be a considerable number of such phenotypes available, identification of nearly all sample mix-ups should become feasible. Another method to prevent sample mix-ups that is commonly used in GWA studies involves the genotyping of a small number of variants prior to hybridization to the chip. *Post hoc* concordance analysis then allows to resolve mixed-up

samples, although this method does not resolve mix-ups that might have been introduced during phenotyping. Although these methods are tailored for GWAS, they are also applicable to genetical genomics datasets. This however does not apply to gene expression data, for which to our knowledge no methods to detect sample mix-ups currently exist.

Our method (*MixupMapper*) uses gene expression levels for genes, which are influenced by genetic variation located near these genes (*cis*-eQTLs). On the basis of such *cis*-eQTL effects, our

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the last three authors should be regarded as joint Last Authors.

method measures the difference between actual gene expression levels and predicted expression levels that are solely based on genotype data of *cis*-SNPs. Using this distance measure, *MixupMapper* is able to detect and correct sample mix-ups with high sensitivity and specificity. In this study, we analyzed five publicly available human genome-wide studies on gene expression and observe that sample mix-ups are frequent. Subsequently, we show that correcting sample mix-ups can yield a substantial increase in the number of significant *cis*-eQTLs. Furthermore, we show through simulations that sample mix-ups have large effects in GWAS as well, when detecting genome-wide significant associations, which may account for a considerable proportion of the missing heritability problem that affects many current GWAS studies (Manolio *et al*., 2009).

2 METHODS

Datasets: we used five genetical genomics studies (Choy *et al*., 2008; Heinzen *et al*., 2008; Stranger *et al*., 2007; Webster *et al*., 2009; Zhang *et al*., 2009) to assess the prevalence of sample mix-ups (Table 1). To our knowledge, this list includes all publicly available datasets that include both genome-wide genotype and gene expression data (as of October 2010). For the various HapMap datasets (Choy *et al.*, 2008; Zhang *et al.*, 2009; Stranger *et al.*, 2007) we confined ourselves to the 309 565 SNPs present on the commonly used Illumina HumanHap300 platform (to limit the number of calculations). The studies that investigated samples from the HapMap project concentrated on the Central European (CEU), Chinese (CHB), Japanese (JPT) and Yoruban (YRI) populations. We combined the CHB and JPT populations since their sample sizes were very small and both reflect Asian samples. As such, we analyzed three sample sets for the studies that used HapMap samples (CEU, CHB+JPT and YRI) for the datasets of Stranger *et al.* (2007) and Choy *et al*. (2008). We analyzed two sets of samples for Zhang *et al.*'s (2009) dataset as they had only investigated the CEU and YRI subpopulations. The dataset from Heinzen *et al.* (2008) consisted of two separate sets of samples from peripheral blood mononuclear cells (PBMCs) and brain tissue that were analyzed separately. Finally, we included a dataset on brain tissue samples from Webster *et al*. (2009). We also assessed a liver dataset from Wolfs,M.G. *et al.* (unpublished data) but did not include eQTL mapping results, as this dataset was not published as a genetical genomics dataset before.

Cis-eQTL mapping: for the sample mix-up analysis, we performed an initial *cis*-eQTL analysis on each dataset. Although we expected the presence of sample mix-ups to have a large effect on the ability to detect *cis*-eQTLs, we assumed this influence was limited for the *cis*-eQTLs with the strongest effects. Gene expression datasets were quantile normalized and $\log_2$ transformed, if appropriate, prior to eQTL mapping. *Cis*-eQTL mapping was performed by using Spearman rank correlations (minor allele frequency (MAF) >5%, Hardy–Weinberg equilibrium (HWE) *P*-value ≥ 0.0001, SNP call-rate $\geq$95%). Only those SNP-probe pairs were tested that were within a vicinity of 250 kilobases (kb). Multiple testing correction was performed by controlling the false discovery rate (FDR) at 0.05 by permuting the phenotype to genotype sample labels [swapping sample phenotype labels, thus preserving the correlation structure within both the genotype and expression data (Breitling *et al*., 2008)] and re-running the eQTL mapping 1000 times. The numbers of *cis*-eQTLs that we have reported here refer to the numbers of unique probes that show a *cis*-association. It is important to note that we did not correct for any potentially false-positive *cis*-eQTLs caused by primer polymorphisms within the expression probe because they actually assist in determining the correct correspondence between genotype and gene expression data. This approach was applied to all *cis*-eQTL mapping procedures in this study, unless specified otherwise.

Identifying sample mix-ups: the concept behind our method to identify sample mix-ups is straightforward: for some genes, the expression level is very strongly determined by a SNP genotype. For a pair of genotype and gene expression arrays, we can determine the concordance between the expected gene expression level conditional on the putative SNP genotype for such genes (Fig. 1C). By systematically assessing all sample pairs with such a set of *cis*-eQTLs, we can determine which pairs are likely to be correct and which pairs are not. In general, there are several scenarios for errors in sample assignment. These include duplicate genotypes, duplicate expression arrays, absent genotypes, absent expression arrays and sample swaps. We considered each of these scenarios indicative of a sample mix-up. We defined the total number of sample mix-ups as the number of genotyped samples that have had an incorrect expression array assigned.

To identify sample mix-ups, *MixupMapper* uses each significantly detected *cis*-eQTL in the dataset. For each of these *cis*-eQTLs we determined the mean (μAA, μAB and μBB) and standard deviation (σAA, σAB and σBB) of the gene expression values for each of the three genotypes (AA, AB and BB). For this purpose we used the genotype and gene expression pairs that were initially assumed to be correct. For each pair of genotype and gene expression array we determined the SNP genotype (g). We calculated the number of standard deviations that the gene expression value (e) differed

Table 1. Analyzed genetical genomics datasets

Study	Population	Sample size	Expression platform	Accession
Choy *et al.* (2008)	HapMap CHB + JPT	87	Affymetrix HG-U133A	GSE11582
	HapMap CEU	84	Affymetrix HG-U133A	
	HapMap YRI	85	Affymetrix HG-U133A	
Stranger *et al.* (2007)	HapMap CHB+JPT	90	Illumina Sentrix Human6 Beadchip	GSE6536
	HapMap CEU	90	Illumina Sentrix Human6 Beadchip	
	HapMap YRI	90	Illumina Sentrix Human6 Beadchip	
Zhang *et al.* (2009)	HapMap CEU	87	Affymetrix Human ST1.0 Exon array	GSE9703
	HapMap YRI	89	Affymetrix Human ST1.0 Exon array	
Webster *et al.* (2009)	Brain	363	Illumina Human Refseq-8	GSE15222
Heinzen *et al.* (2008)	Brain	93	Affymetrix Human ST1.0 Exon array	http://people.genome.duke.edu/~dg48/SNPExpress/
	PBMC	80	Affymetrix Human ST1.0 Exon array	
Wolfs,M.G. *et al.* (unpublished data)	Liver	73	Illumina Human HT12v3	GSE22070[a]

[a]We excluded the Wolfs,M.G. *et al.* (unpublished data) dataset from eQTL mapping, since this dataset was not previously published as a genetical genomics dataset.

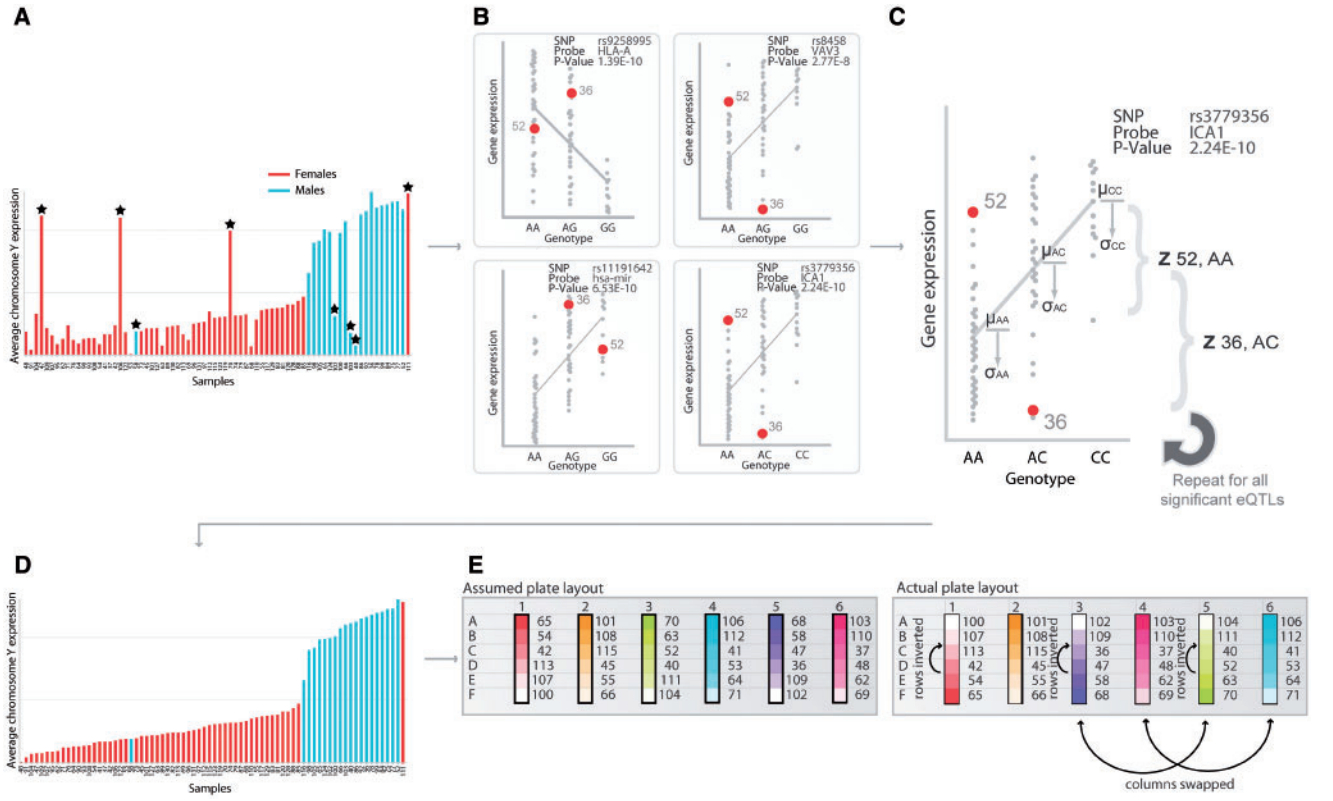

Fig. 1. (**A**) We observed numerous sample mix-ups in a dataset created by our laboratory (Wolfs,M.G. *et al.*, unpublished data), where the chromosome Y expression did not correspond to the genotype-derived sex (indicated with asterisk). (**B**) Four plots of *cis*-eQTLs mapped in the dataset from Wolfs,M.G. *et al.* showed samples 36 and 52 as outliers. These samples generally deviated more from the expected regression line than the other samples in this dataset (samples 36 and 52 highlighted). If this was a general observation over all significant *cis*-eQTLs for this dataset, we gathered evidence that something was wrong with these samples. (**C**) Therefore, for each *cis*-eQTL, we calculated the mean gene expression level (μ) and standard deviation (σ) per genotype (g). This allowed us to determine, per individual (i), to what extent the gene expression level (e_i) was deviating from the regression line using an absolute $Z\text{-score}_i = |e_i - \mu_{gi}|/\sigma_{gi}$. Samples 36 and 52 generally have a higher Z-score compared to other samples. By repeating these calculations for all significant *cis*-eQTLs, and by comparing all pairs of gene expression arrays and genotyping arrays, we could identify those samples that were likely to be mixed-up. (**D**) When we corrected these mix-ups we observed that the chromosome Y expression now corresponded to the sex for most samples: sample mix-ups resolved. (**E**) Inspection of the RNA plate layout indicated that mix-ups had been introduced by pipetting mistakes.

from the expected value associated with the SNP genotype using an absolute Z-score (1).

$$Z = \frac{|e - \mu_g|}{\sigma_g} \quad (1)$$

For each sample pair we summed the absolute Z-scores of all significant *cis*-eQTLs and determined the average Z-score for each sample pair to account for differences in the number of tested eQTLs per sample pair due to missing SNP genotypes.

Expression arrays that have been hybridized to lower quality or degraded RNA tend to result in higher deviations from the individual *cis*-eQTL regression lines, irrespective of what genotyped sample has been tested for such an expression array. As a result, such an expression sample will show higher overall Z-scores on average for each of the genotyped samples to which it is compared. Therefore, in order to standardize the Z-scores for each of the expression arrays, we normalized the Z-scores by subtracting the average of the overall Z-scores for this expression sample and divided it by the standard deviation of the overall Z-scores for this expression sample. Similarly, we normalized the Z-scores by subtracting the average of the overall Z-scores for this genotype sample and divided it by the standard deviation of the overall Z-scores for the genotype sample.

After these normalizations we determined what the expression array was with the lowest overall normalized Z-score for each genotyped sample. We considered this expression sample to reflect this particular genotyped sample.

Once the best matching expression sample had been identified for each genotyped sample, we compared it to what had been initially defined, permitting us to identify which samples were mixed-up.

It is, however, also conceivable that our method might incorrectly suggest the presence of sample mix-ups, because of potential overfitting of mean and standard deviations for each of the three genotype groups per *cis*-eQTL. We therefore used a *post hoc* permutation strategy to check if our results were not due to overfitting. We first permuted the phenotype labels and subsequently ran a *cis*-eQTL analysis. As expected, this analysis did not lead to the identification of any significant *cis*-eQTL, although it did permit us to identify the list of top *cis*-eQTLs for this permutation. We chose an equal number of *cis*-eQTLs as identified in the initial *cis*-eQTL analysis (that was based on the non-permuted data). Using this set of *cis*-eQTLs and the permuted phenotype labels, we then performed the sample mix-up identification procedure. This resulted in overall Z-scores for each genotype-expression sample pair and a respective distribution that indicated what could be expected when running the mix-up procedure on randomly permuted data. Based on this distribution, we determined the 5th percentile Z-score threshold

(low Z-values indicating a better agreement). We repeated this permutation strategy 1000 times, resulting in a distribution of 5th percentile Z-score thresholds. We decided to select the 5th percentile Z-score threshold as what was attained in only 5% of the 1000 permutations. Using this strategy we determined a Z-score significance threshold for each of the datasets. We used this threshold for each of the inferred sample mix-ups and only considered the mix-ups significant and real if the mix-up Z-score was lower than the Z-score significance threshold.

Once we had determined the best match for each genotype array, we used the following procedure to decide which genotyped samples to keep, which to correct, and which to remove completely:

(1) For each genotyped sample we first checked if the overall Z-score of the best matching gene expression array was below the permutation Z-score threshold (a lower Z-score corresponds to a better match). We discarded the genotyped sample completely if no well-matched gene expression sample was identified.

(2) For each of the remaining genotyped samples, we checked if the best matching gene expression array corresponded to the one that was originally considered to match it. If they corresponded, this indicated that the sample had not been mixed-up and it would be kept.

(3) For those genotyped samples with an incorrect gene expression array, we applied the following procedure:

- We first determined what other genotyped sample was originally coupled to that particular gene expression array. If that other genotype sample was not considered a mix-up and thus matched the particular gene expression array well, we knew we had to discard the genotype sample that we were assessing.
- Alternatively, if the other genotype sample did not correlate well with the gene expression array, we knew we could safely assign the assessed genotyped sample to this gene expression array, because the other genotype sample was likely to be assigned to another well matching expression sample.

(4) Finally, to ensure that each gene expression array was eventually assigned to a single genotype array, we checked whether there were two genotype samples that were assigned to the same gene expression sample. If this was the case, we discarded the genotype sample that was the worse match to the expression array (i.e. the one with the highest Z-score).

Using this method, we were able to not only correct for sample mix-ups, but also to identify those genotype arrays that clearly did not match any gene expression arrays. Such genotype samples generally had a worse Z-score than those genotype samples that matched a gene expression array well and they were discarded from further analyses.

Effects of sample mix-ups in genetical genomics studies: we assessed the effect of sample mix-ups in each of the assessed datasets by repeating the *cis*-eQTL mapping after we had corrected the sample mix-ups. We also simulated the effect of accidental sample mix-ups in the datasets in which we had not identified any mix-ups. The samples were permuted with increments of ~5%, after which we performed *cis*-eQTL mapping (FDR < 0.05, 100 permutations). We repeated this analysis 100 times to get an accurate estimate of the average number of significantly detected *cis*-eQTLs for the different sets of increasingly mixed-up samples.

Sensitivity and specificity of sample mix-up method: for the datasets where we had deliberately introduced sample mix-ups, we ascertained how many of these mix-ups could be identified by our method to get realistic estimates on its specificity and sensitivity. We defined the true positives (TP) as the number of mixed-up samples that our method had correctly identified. However, we also required that the method had identified the correct alternative expression sample. We defined the false positives (FP) as the number of samples that were either falsely deemed to be a mix-up, or that were mixed-up but not assigned to the correct alternative expression sample. We defined the true negatives (TN) as the number of samples correctly identified as not being mixed-up. This permitted us to determine the true positive rate (TPR, sensitivity) as: TPR = TP/(Number of introduced mix-ups), and the false positive rate (FPR) as: FPR = FP/(TN + FP). We defined the specificity as 1 – FPR.

Effects of sample mix-ups on GWAS that looked at one particular, continuous, phenotype: we simulated the effect of sample mix-ups on GWAS that investigated one particular, continuous, phenotype by first generating genotypes for a population of 500 000 individuals, each with a minor allele frequency (MAF) of 0.5. On the basis of these genotypes, we then generated a random continuous phenotype, based on an error term and the joint effect of 100, 200 or 500 unlinked SNPs (each having an equal effect size), and we assumed that each of these causal variants had been successfully genotyped. Using this method, we created phenotypes that were respectively 90, 80, 70, 60 and 50% heritable. From this population, we then randomly sampled 10 000 samples, and conducted an association analysis. We correlated the phenotype to the genotype using Pearson correlation coefficients. By using a P-value threshold of 5×10^{-8}, which is commonly used to declare genome-wide significance, we were able to determine what proportion of the causative variants were significant. By repeating this procedure 1000 times, we obtained reliable estimates of how many of the variants were declared significant. We then repeated this procedure while swapping the phenotypes of an increasing number of individuals, to ascertain the effect of sample mix-ups.

3 RESULTS

3.1 Identifying sample mix-ups

The issue of sample mix-ups became apparent in a genetical genomics dataset of liver tissue that had been generated in our laboratory, as there were various samples for which the chromosome Y expression did not correspond to the gender as derived from the genotypes (Wolfs,M.G. *et al.*, unpublished data; Fig. 1A).

To identify the exact origin of these sample mix-ups we developed a sample mix-up algorithm (*MixupMapper*) that relies upon gene expression phenotypes that are influenced by genetic variation (*cis*-eQTLs). Figure 1B shows four *cis*-eQTL plots from this dataset in which samples 36 and 52 have been highlighted. These samples generally deviate substantially from the *cis*-eQTL regression line (Fig. 1C), suggesting they have been swapped.

By applying our mix-up identification method to all pairs of genotyped and gene expression samples, we were able to identify 28 sample mix-ups in this dataset (Fig. 1D). These mix-ups were later confirmed by the facility involved in generating the data: when we compared the results to the plate layout of the RNA samples used during pipetting, we observed that some columns had been swapped and some rows had been inverted after hybridization to the gene expression chip (Fig. 1E). We should note that the actual mix-up of samples can have occurred during any of the steps involved in the generation of this dataset, such as during DNA and RNA isolation, aliquoting and hybridization. However, these samples appeared to be mixed-up because of pipetting mistakes prior to hybridization on the RNA-chip, as we did not observe any indications of errors that occurred in the DNA preparation or hybridization process. In total, 30 out of 74 samples (41%) had been assigned wrong expression phenotypes and our method was able to resolve 28 of them.

3.2 Sample mix-ups in published datasets

As we had identified these mix-ups in our own data, we applied our method to five publicly available human datasets for which both genotype and gene expression data was freely available online

Table 2. *Cis*-eQTL mapping and sample mix-up identification results

Stud	Population	Sample-size	Initial *cis*-eQTLs	Mix-ups detected[a] *n* (%)	Sample-size after correction *n* (%)	*cis*-eQTLs after correction *n* (%)
Choy *et al.* (2008)	CHB+JP	87	138	20 (23)	79 (90)	418 (+203)
	CE	84	558		NA	NA
	YR	85	274	2 (2)	83 (97)	287 (+5)
Stranger *et al.* (2007)	CHB+JP	90	1511		NA	NA
	CE	90	903		NA	NA
	YR	90	663	1 (1)	89 (99)	667 (+1)
Zhang *et al.* (2009)	CE	87	2581		NA	NA
	YR	89	1454	2 (2)	89 (100)	1635 (+12)
Webster *et al.* (2009)	Brai	36	1284	16 (4)	356 (98)	1367 (+6)
Heinzen *et al.* (2008)	Brai	93	349		NA	NA
	PBMC	80	297		NA	NA

[a]In four out of the five studies, sample mix-ups were present in some of the populations investigated by the authors. In a substantial number of cases, these sample mix-ups could be resolved if, for instance, we assumed that two expression samples had been accidentally swapped. Genotyped samples for which no appropriate expression sample could be identified were removed. Numbers of *cis*-eQTLs are number of unique probes with an significant effect (FDR < 0.05). NA, not applicable.

(Table 1) (Choy *et al*., 2008; Heinzen *et al*., 2008; Stranger *et al*., 2007; Webster *et al*., 2009; Zhang *et al*., 2009). Four out of these five datasets contained sample mix-ups (Table 2, Supplementary Fig. S1 and Table S1) and observed that for 3% of all samples, the genotype and expression data did not correspond (41 out of 1238 samples). The number of sample mix-ups was highest in the CHB+JPT subset from Choy *et al.* (2008): out of 87 samples, 20 were incorrect (23%). In total, we were able to correct 21 of the 41 samples that had an incorrect expression phenotype.

We assessed whether the identified mix-ups were not due to extreme expression levels (e.g. because of hybridization problems that lead to poor normalization) or bad-quality genotypes. For this purpose, we assessed the variability of all genotype and gene expression samples within each dataset using principal component analysis (PCA) on the sample correlation matrix: samples that are clear outliers in terms of sample quality, show deviations for the first two principal components (PCs). To test whether the identified sample mix-ups deviated from the remaining samples, we performed a Wilcoxon–Mann–Whitney test on the eigenvalues for these PCs. As we did not observe any significant differences (Bonferroni correction $P < 0.004$), we can conclude that sample quality does not confound the results of our method (Supplementary Fig. S2).

3.3 Sensitivity and specificity

We established that our method is highly specific and sensitive, by conducting simulations in the datasets in which no mix-ups had been identified (Fig. 2A). We observed the best performance in the Stranger *et al.* (2007) CHB+JPT dataset: even when 40% of the samples had been mixed-up, 98% (σ^2: 3%) of these 'errors' could still be successfully identified and successfully corrected. The worst performance was observed in the Heinzen *et al.* (2008) post-mortem brain dataset: when 10% of the samples in this dataset were mixed-up, only 85% (σ^2: 15%) of these could be successfully identified and corrected. Very few samples were wrongly deemed a sample mix-up by our method, indicating our method is highly specific for each of the datasets in which 10% of the samples had been mixed-up, we observed only rarely that non-mixed-up samples were wrongly deemed to be mixed-up (Specificity >99.9%, Fig. 2A).

3.4 Effect of correcting sample mix-ups on the number of detected *cis*-eQTLs

Table 2 shows that the number of detected *cis*-eQTLs increased in each of the datasets after we had corrected for the sample mix-ups (Table 2 and Supplementary Figs S3, S4A–S4F, FDR < 0.05, 1000 permutations). In total, 15% more *cis*-eQTLs were identified for these datasets. For the CHB+JPT population from Choy *et al.* (2008), correcting the identified mix-ups increased the number of significant *cis*-eQTLs by 203% (418 *cis*-eQTLs compared to 138 *cis*-eQTLs before correction). This is a considerable increase, especially since the effective sample size decreased by 9% (79 samples included after correction compared to 87 samples in the original dataset). Furthermore, this indicated that the removed samples effectively amounted to noise and therefore contributed to a decrease in the power to detect *cis*-eQTL effects, especially smaller ones. However, the increase in the number of detected *cis*-eQTLs we describe here could also be explained by an increase of the proportion of false positives that are the result of SNPs being present in the gene expression probe sequences, directly affecting hybridization efficiency (Benovoy *et al.*, 2008). We therefore checked whether the proportion of potential false positives due to probe polymorphisms differed before and after sample mix-up correction for each dataset and found no differences in the assessed datasets (Supplementary Fig. S5).

3.5 Replication of detected *cis*-eQTLs

We reasoned that we could gain further evidence that the identified sample mix-ups were indeed mix-ups by replicating the *cis*-eQTLs that were identified after the correction. It has been shown that substantial overlap exists in *cis*-eQTL effects between different populations and also between different tissues (Bullaughey *et al*., 2009; Heap *et al*., 2009; Stranger *et al*., 2007). We therefore compared the datasets that had been run on the same expression platform, which amounts to comparing the different HapMap

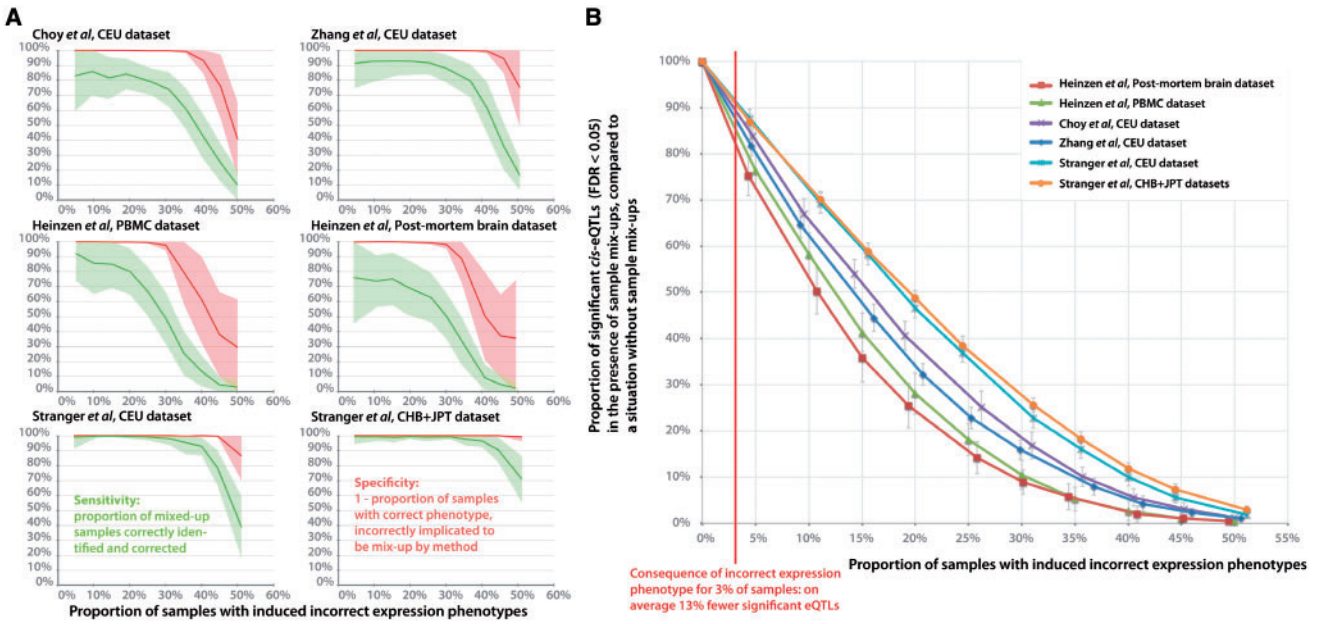

Fig. 2. (**A**) Robustness analyses on the sample mix-up identification method shows that if the fraction of mixed-up samples is <25%, most of the sample mix-ups are detected and corrected by our method with high specificity and sensitivity. Variability is generally low (shaded areas around graphs), especially for the specificity. (**B**) Sample mix-ups were introduced into the six datasets that did not initially contain any mix-ups. Deliberately introducing sample mix-ups resulted in a substantial decrease in the number of significantly detectable *cis*-eQTLs (FDR < 0.05). If 3% of the samples had an incorrect phenotype assigned, the average number of detectable *cis*-eQTLs decreased by 13%.

populations. For the studies of Choy *et al.* (2008), Zhang *et al.* and Stranger *et al.* (2007), we assessed number of the significantly detected *cis*-eQTLs in one HapMap population that had also been detected in another HapMap population (Supplementary Fig. S6). After correcting the sample mix-ups, the number of shared *cis*-eQTLs increased in each of the population comparisons. A total of 99.7% of the eQTLs that were shared between at least two populations showed identical allelic directions. We observed comparable increases in shared *cis*-eQTLs after mix-up correction, when comparing identical HapMap populations that had been run on different expression platforms (Supplementary Fig. S7).

3.6 Effect of sample mix-ups on GWAS studies on continuous traits

We systematically explored the effect of different proportions of sample mix-ups on the power to detect *cis*-eQTLs. We investigated the datasets in which we had not identified any sample mix-ups and deliberately introduced sample mix-ups by swapping the expression array measurements for an increasing number of samples. We observed a considerable decrease of, on average, 13% when only 3% of the samples were deliberately mixed-up (Fig. 2B).

Furthermore, we decided to model the influence of sample mix-ups on GWAS studies that investigate traits caused by hundreds of variants, such as height or body mass index. We simulated a trait with several degrees of heritability and different numbers of causative variants in a population of 500 000 individuals. From this population we randomly sampled 10 000 individuals on which we conducted a linear regression analysis and determined the percentage of SNPs that were genome-wide significant ($P < 510^{-8}$, Fig. 3A). To measure the effect of sample mix-ups, we then deliberately randomized the trait measurements of an increasing number of randomly selected individuals (Fig. 3B and 3C).

A very limited number of sample mix-ups had severe consequences for traits that have a heritability of 50%, of which the heritable fraction is due to 500 causal variants. If 3% of all samples had an incorrect phenotype (as observed in the genetical genomics datasets), we could only significantly detect 77% of the causal variants that we would have identified if no mix-ups had been present (at $P < 5 \times 10^{-8}$, Fig. 3B). Likewise, we observed a decrease in the proportion of the heritability that could be explained by the genome-wide significant associations (Fig. 3C): the explained heritability of genome-wide significant variants decreased with 1.24-fold (from 3.7% to 2.8%) when 3% of the samples were mixed-up.

If these sample mix-ups could be detected and corrected, it would be possible to explain 1.24-fold more of the heritability for such traits. Methods to detect such sample mix-ups therefore have the potential to substantially increase the explained heritability and power to detect genetic effects in GWAS on complex phenotypes such as height, BMI or some diseases.

4 DISCUSSION

We have identified sample mix-ups in four out of five genetical genomics studies by applying a novel method (*MixupMapper*). On average, 3% of all samples were mixed-up. After correction for these sample mix-ups by our method, we detected on average 15% more *cis*-eQTLs. Correcting mix-ups in one dataset in which 23% of the samples were incorrect led to three times as many significant *cis*-eQTLs being detected. The consequences of only

A

Number of causal variants	Herita-bility of trait	Proportion of causal variants with association $P < 5 \times 10^{-8}$	Prop. heritability explained by causal variants with $P < 5 \times 10^{-8}$
100	0.9	100%	100%
	0.8	100%	100%
	0.7	100%	100%
	0.6	99%	99%
	0.5	95%	97%
200	0.9	90%	94%
	0.8	81%	89%
	0.7	68%	80%
	0.6	51%	65%
	0.5	33%	48%
500	0.9	11%	22%
	0.7	4%	10%
	0.5	1%	4%

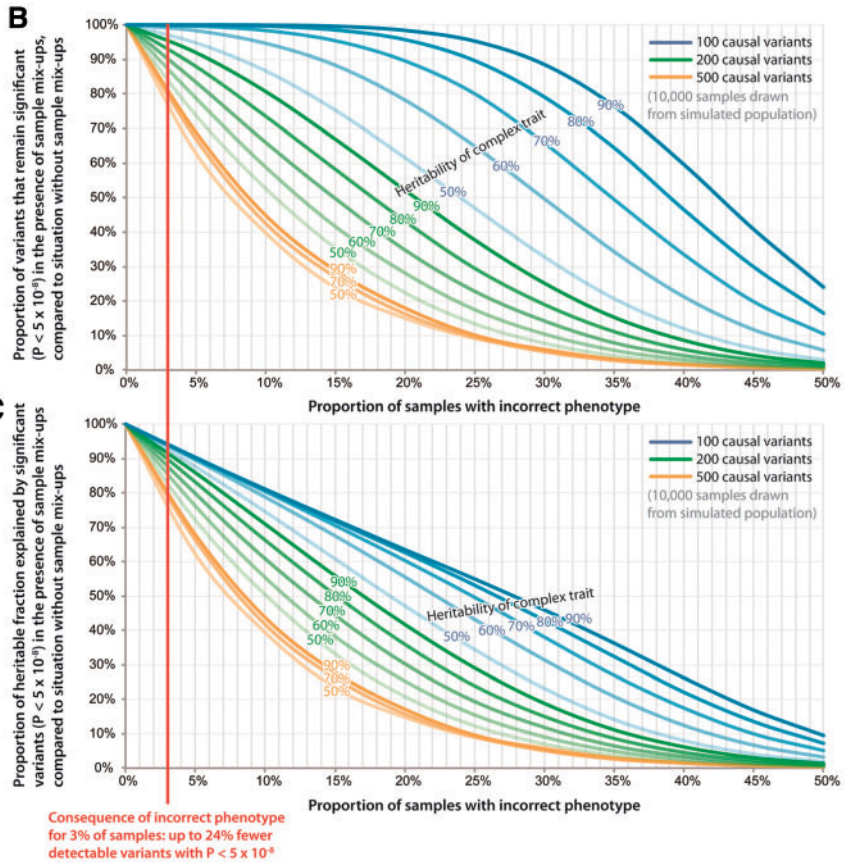

Fig. 3. Continuous traits with different levels of heritability (50, 60, 70, 80 and 90%) were modeled, assuming 100, 200 or 500 causative variants. (**A**) Results from linear regression analysis in 10 000 samples, drawn from simulated population if no sample mix-ups exist. By sampling 10 000 individuals and conducting linear regression, it was established what proportion of causal variants showed genome-wide association and what the explained heritable proportion was. (**B**) Loss in proportion of genome-wide significant casual variants in the presence of sample mix-ups. If sample mix-ups were present, we observed substantial loss in power to detect association. If 3% of samples had incorrect phenotypes, 24% fewer genome-wide significant loci were observed for a 50% heritable trait caused by 500 unlinked variants. (**C**) Likewise we observed a substantial loss in the explained heritable proportion.

3% sample mix-ups on the heritable fraction that can be explained by significantly associated variants can also be substantial. For some simulated complex traits with a moderate to high heritability, the explained heritability of the genome-wide significant variants increased 1.24-fold, when these sample mix-ups could be detected and corrected.

A considerable proportion of the heritability of complex diseases and traits is currently 'missing'. There is debate on whether the missing heritability problem is caused by rare variants with a large effect, by many more common variants, each with a very small effect size, by overestimation of the heritability estimates or through other means (Manolio *et al.*, 2009; McCarthy and Hirschhorn, 2008). As current genome-wide studies are pushing toward associating ever smaller effect sizes, sample sizes have to increase substantially to discover loci with smaller effect sizes (Park *et al.*, 2010). Our results indicate that especially for these small-effect size loci, sample mix-ups could have consequences on the power to detect such loci for both genetical genomics studies as well as GWAS. As such, a proportion of the missing heritability could possibly be explained by the presence of sample mix-ups in genome-wide datasets.

However, it remains a question whether the frequency of sample mix-ups we observed in genetical genomics samples is a realistic estimate for other types of genome-wide datasets. Different types of genome-wide datasets require many different handling steps, and therefore their frequencies for the presence of sample mix-ups may differ. For example for case–control GWAS the cases and controls are often collected and processed separately from each other. Therefore, the frequency and hence the consequence of sample mix-ups in such case–control studies might be lower compared to the studies presented here. As such, GWAS in general might contain fewer sample mix-ups.

In the case of genetical genomics datasets, more *cis*-eQTLs could be detected in each of the datasets after correction, although the number of included samples had actually decreased for three of these datasets. This effectively demonstrates that increasing the sample size is not the only way of increasing statistical power for determining complex traits; increasing the phenotypic accuracy can be equally helpful. In addition to the method described here, phenotypic accuracy may be increased by, for example, including relevant co-variates in GWAS, or by using principal components

analysis in *cis*-eQTL studies (Dubois *et al*., 2010; Leek and Storey, 2007). Although these methods are helpful in increasing the phenotypic accuracy, they do not help to identify or overcome sample mix-ups.

One possible problem with *MixupMapper* is that it depends on an initial set of *cis*-eQTLs that can be detected in the data. Although our method generally shows high sensitivity and specificity when large proportions of samples are mixed-up, in an extreme scenario, where all genotyped samples are randomly assigned to the gene expression samples, the mix-ups cannot be resolved since no *cis*-eQTLs will be initially detectable. However, if a set of *cis*-eQTLs has been independently identified in another set of samples for the particular expression platform used, it is also be possible to resolve these problems (data not shown).

We feel it is important to emphasize that the sample mix-ups that we detected in the five public datasets do not in any way discredit these studies. To our knowledge, we are the first to describe a method to identify mix-ups for these kinds of datasets. If the authors had been aware of the existence of these mix-ups, they would have certainly corrected them, as their goal was to find as many eQTLs as possible. Since we observed a substantial overlap of detected *cis*-eQTLs in the three HapMap populations before correction of sample mix-ups, we assume the detected *cis*-eQTLs in these datasets are still valid. We are convinced that the results and conclusions drawn from these datasets (Choy *et al*., 2008; Heinzen *et al*., 2008; Stranger *et al*., 2007; Webster *et al*., 2009; Zhang *et al*., 2009) remain appropriate.

Although our method is intended for genetical genomics datasets, it can also be applied to other types of genome-wide datasets, as long as there are sufficient numbers of different phenotypes available per individual that are each (strongly) determined by genetic variants or combinations of variants. This requirement will likely be met with the growing interest in population-based cohort studies in which hundreds of phenotypes are collected from the participants. As a consequence, identifying sample mix-ups will then become possible for these datasets as well.

Our results clearly indicate that sample mix-ups occur in many laboratories (including ours). Although a great deal of quality control is conducted in GWAS, it is very difficult to prevent the accidental mislabeling of some samples. This is particularly problematic in studies of unrelated individuals where inheritance patterns cannot be investigated. Nevertheless, these accidental human mistakes or experimental problems can sometimes have far-reaching consequences. We therefore recommend rigorous quality control of the laboratory and administrative processes in order to prevent sample mix-ups from happening. We conclude that fewer sample mix-ups will increase the power to detect significant genetic associations substantially and might resolve a part of the missing heritability.

ACKNOWLEDGEMENTS

We thank Jackie Senior for her critical reading of the manuscript.

Funding: Horizon Breakthrough grant from the Netherlands Genomics Initiative (93519031); VENI grant from the Netherlands Organization for Scientific Research (NWO, ZonMW grant 916.10.135); European Community's Health Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 259867 (to L.F.).

Conflict of Interest: none declared.

REFERENCES

Barrett,J.C. *et al.* (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.,* **41**, 703–707.

Breitling,R. *et al.* (2008) Genetical genomics: spotlight on QTL hotspots *PLoS Genet.*, **4**, e1000232.

Benovoy,D. *et al*. (2008) Effect of polymorphisms within probe-target sequences on olignonucleotide microarray experiments. *Nucleic Acids Res.*, **36**, 4417–4423.

Bullaughey, K. *et al.* (2009) Expression quantitative trait loci detected in cell lines are often present in primary tissues, *Hum Mol Genet*, **18**, 4296–4303.

Buyske,S. *et al.* (2009) When a case is not a case: effects of phenotype misclassification on power and sample size requirements for the transmission disequilibrium test with affected child trios. *Hum. Hered.*, **67**, 287–292.

Choy,E. *et al.* (2008) Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.*, **4**, e1000287.

Dubois,P.C. *et al.* (2010) Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.*, **42**, 295–302.

Franke,A. *et al.* (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.*, **42**, 1118–1125.

Gordon,D. and Finch,S.J. (2004) Consequences of error. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. John Wiley & Sons, Ltd.

Heap,G.A. *et al.* (2009) Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med Genomics*, **2**, 1.

Heinzen,E.L. *et al.* (2008) Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol*, **6**, e1.

Hindorff,L.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.

Ho,L. and Lange,E. (2010) Using public control genotype data to increase power and decrease cost of case–control genetic association studies. *Hum. Genet.*, **128**, 597–608.

Jansen,R.C. and Nap,J.P. (2001) Genetical genomics: the added value from segregation. *Trends Genet.*, **17**, 388–391.

Lango Allen,H. *et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**, 832–838.

Leek,J.T. and Storey,J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, 1724–1735.

Manolio,T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

McCarthy,M.I. and Hirschhorn,J.N. (2008) Genome-wide association studies: potential next steps on a genetic journey. *Hum. Mol. Genet.*, **17**, R156–R165.

Park,J.H. *et al.* (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.*, **42**, 570–575.

Samuels,D.C. *et al*. (2009) Detecting new neurodegenerative disease genes: does phenotype accuracy limit the horizon? *Trends Genet.*, **25**, 486–488.

Speliotes,E.K. *et al.* (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.*, **42**, 937–948.

Stranger,B.E. *et al.* (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848–853.

Sulem,P. *et al.* (2007) Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.*, **39**, 1443–1452.

Teslovich,T.M. *et al.* (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707–713.

Voight,B.F. *et al.* (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.*, **42**, 579–589.

Webster,J.A. *et al.* (2009) Genetic control of human brain transcript expression in Alzheimer disease. *Am. J. Hum. Genet.*, **84**, 445–458.

Yip,S.P. (2002) Sequence variation at the human ABO locus. *Ann. Hum. Genet.*, **66**, 1–27.

Zhang,W. *et al.* (2009) Identification of common genetic variants that account for transcript isoform variation between human populations. *Hum. Genet.*, **125**, 81–93.

Zheng,G. and Tian,X. (2005) The impact of diagnostic error on testing genetic association in case-control studies. *Stat. Med.*, **24**, 869–882.