reGenotyper: Detecting mislabeled samples in genetic data

PLOS ONE

RESEARCH ARTICLE

reGenotyper: Detecting mislabeled samples in genetic data

Konrad Zych[1], Basten L. Snoek[2], Mark Elvin[3], Miriam Rodriguez[2], K. Joeri Van der Velde[4], Danny Arends[1], Harm-Jan Westra[5,6,7], Morris A. Swertz[4], Gino Poulin[3], Jan E. Kammenga[2], Rainer Breitling[8], Ritsert C. Jansen[1], Yang Li[1,9]*

1 Groningen Bioinformatics Centre, University of Groningen, Groningen, The Netherlands, **2** Laboratory of Nematology, Wageningen University, Wageningen, The Netherlands, **3** Faculty of Life Sciences, University of Manchester, Manchester, United Kingdom, **4** Genomics Coordination Center, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands, **5** Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States of America, **6** Partners Center for Personalized Genetic Medicine, Boston, Massachusetts, United States of America, **7** Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America, **8** Manchester Institute of Biotechnology, Faculty of Life Sciences, University of Manchester, Manchester, United Kingdom, **9** University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands

* yang.li@rug.nl

Citation: Zych K, Snoek BL, Elvin M, Rodriguez M, Van der Velde KJ, Arends D, et al. (2017) reGenotyper: Detecting mislabeled samples in genetic data. PLoS ONE 12(2): e0171324. doi:10.1371/journal.pone.0171324

Editor: Suzannah Rutherford, Fred Hutchinson Cancer Research Center, UNITED STATES

Received: July 22, 2016

Accepted: January 19, 2017

Published: February 13, 2017

Data Availability Statement: Gene expression data of C. elegans at two temperatures can be found at NCBI Gene Expression Omnibus with accession numbers: GSE5395 (two temperatures), GSE17071 (three age stages). All files are also available at http://www.wormqtl.org/.

Funding: This work was supported by the 7th Framework Programme of the European Commission under the Research Project PANACEA [Contract No. 222936 to RCJ and JEK]; the Netherlands Organisation for Scientific Research (NWO) VENI grant [n° 863.13.011 to YL]; and the

Abstract

In high-throughput molecular profiling studies, genotype labels can be wrongly assigned at various experimental steps; the resulting mislabeled samples seriously reduce the power to detect the genetic basis of phenotypic variation. We have developed an approach to detect potential mislabeling, recover the "ideal" genotype and identify "best-matched" labels for mislabeled samples. On average, we identified 4% of samples as mislabeled in eight published datasets, highlighting the necessity of applying a "data cleaning" step before standard data analysis.

Introduction

With the development of a wide range of high-throughput molecular profiling methods in recent years, there has been significant growth in the number of genetic studies on molecular profiles from genetically different individuals [1–6], aiming at the identification of the functional consequences of naturally-occurring and induced genetic variation (these studies are often referred to as genetical genomics [7,8] or expression genetics [9] experiments). Statistical significance testing is used to suggest causal relationships between genetic variation (polymorphisms) at the genomic level and phenotypic variation at multiple molecular levels (transcriptomics, proteomics, and/or metabolomics) using methods such as quantitative trait locus (QTL) mapping and genome-wide association studies (GWAS). Obviously, the effectiveness of significance testing depends critically on the accurate labeling of the samples, i.e., the genotype data used for statistical testing needs to correctly represent the true genotypes of the examined individuals. However, samples can be wrongly labeled in the laboratory due to human error, resulting in the assignment of wrong genotypes to individual samples; as we show below, this is a surprisingly common phenomenon [10].

Dutch Carbohydrate Competence Center, which is is co-financed by the European Regional Development Fund, the Dutch Ministry of Economic Affairs (as part of Pieken in de Delta, the government's regional economic agenda), the Municipality of Groningen and the Province of Groningen [CCC WP23 to KZ] The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

A wrong genotype assignment will seriously weaken the significance testing [11] in genetical genomics studies on model organisms, especially for experiments with relatively small sample sizes. Technical genotyping errors (e.g. assigning incorrect SNP nucleotide) may also impact statistical power of genomics studies. However, these can be treated by some of the QTL mapping tools [12,13]. Particularly, mislabeled samples will influence the detection of small genetic effects. This could, on the one hand, explain the lack of consistent results across some experiments reported in meta-analyses of genetical genomics studies [14]. On the other hand, mislabeled samples could also lead to false positives for QTL mapping or GWAS in certain conditions. For example, when performing multiple QTL mapping [15], the mapping of the second (minor) QTL is based on the residuals from the phenotypes after correction for the first (major) QTL effect [16]. When there are mislabeled samples in the data, this correction process is actually turning these samples into outliers, which subsequently results in a (slight) increase in number of significant minor QTLs. Obviously, these extra minor QTLs are all false positives without biological meaning. Unfortunately, permutation methods will not help to alleviate this problem since the correction for the major QTL is done after the permutation. Following a similar reasoning, the mislabeled samples can lead to false results when testing for interactions between QTLs to correct for the deviation between data and model, such as epistasis or QTL-by-environment interactions. Therefore, new methods are necessary to detect and correct mislabeled samples before the data analysis process begins and increase the power to map meaningful QTLs.

Here we present *reGenotyper*, which implements a fast and accurate algorithm to identify samples that are likely to have been mislabeled, based on the measured phenotypes (Fig 1). Our approach is based on a data perturbation strategy and exploits the highly parallel nature of molecular profiling in modern genetic studies. reGenotyper aims to detect mislabeled samples in genetical genomics studies based on the heuristic "significance change value". Our tool is able to quickly and precisely detect potential sample mislabeling using combined power of hundreds or thousands of QTLs. This makes it perfectly suited for studies based on high-throughput molecular profiling but unfeasible for traditional QTL studies based on limited number of phenotypes.

We show that the highly parallel nature of genetical genomics studies allows the sensitive and specific detection of the problematic samples. We also show that the frequency of wrongly labeled samples in such studies can be high (on average, 4% of samples are mixed up in the worm studies we analyzed [17,18], and similarly high values have been reported for human [19] and mice [10] studies [19]. As wrong sample labels seriously affect the power of QTL detection in studies that are already operating at the limit of statistical viability [20–22], the application of a mislabeled sample detection (and correction) step would be strongly recommended for any GWAS or QTL study on large-scale molecular profiling data. We implemented *reGenotyper* in the R programming language [23]. The R package and documentation are freely available from the CRAN repository and at http://www.molgenis.org/regenotyper.

Materials and methods

For sake of simplicity, we will showcase our algorithm on transcriptome data from a population of recombinant inbred lines (RILs). RILs are homozygous individuals that result from repeated self-mating or sibling mating, starting from an F_1 of two homozygous parents, carrying alleles of type *A* and type *B*, respectively. The genome of a RIL is therefore a mosaic of the "founder" genomes, making them a perfect population to illustrate the mode of action of our algorithm. Nonetheless, our method can be applied to any type of high-throughput phenotype data (e.g. proteomics, metabolomics) from a variety of other population types.

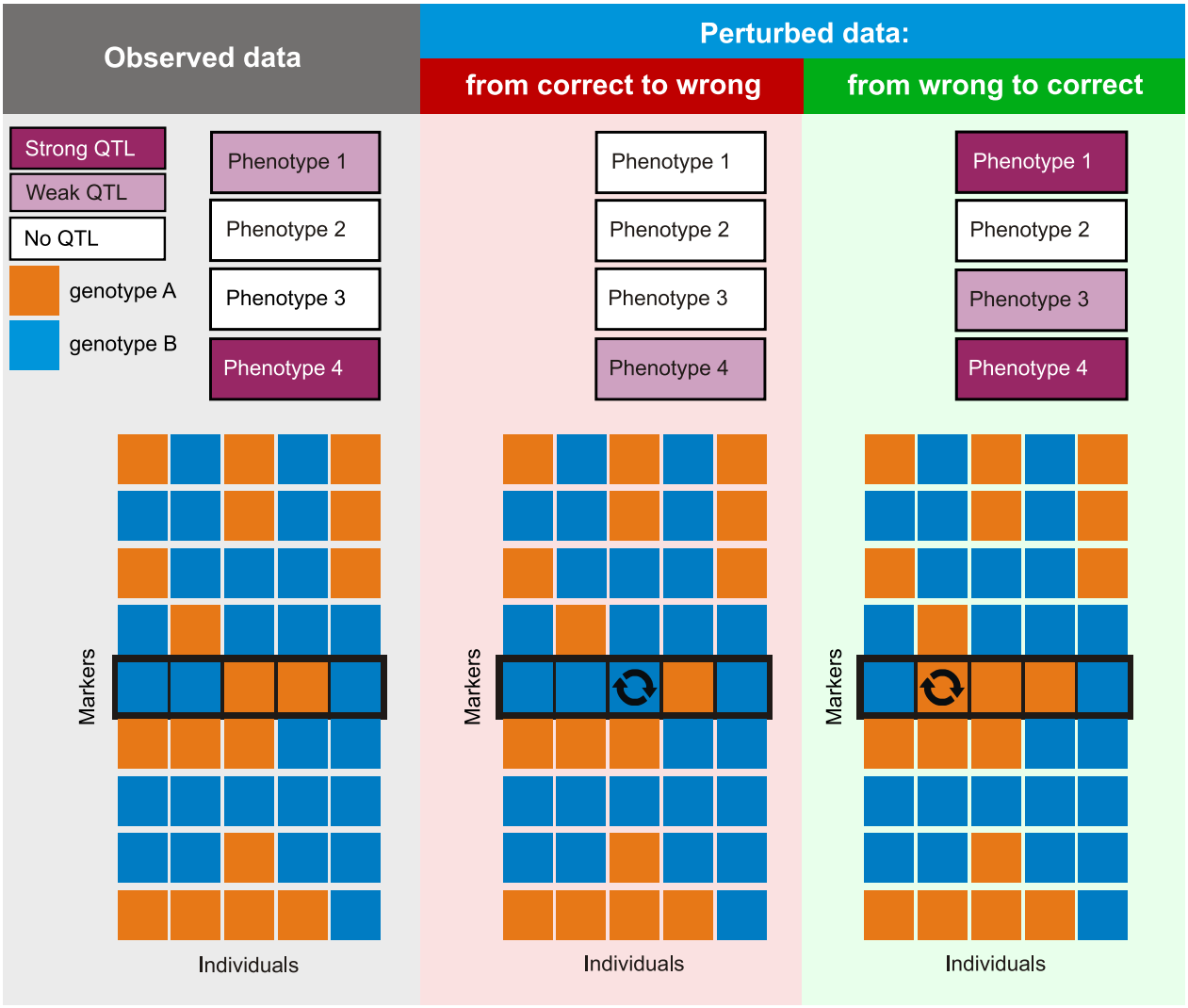

Fig 1. Graphical summary of the reGenotyper algorithm. reGenotyper uses a data perturbation strategy and exploits the highly parallel nature of molecular profilinrg in modern genetic studies. 1) Observed genotype data and QTLs. The data matrix in the middle contains the genotype information at each marker position (row) for each sample (column), where orange and blue represent two different genotypes. The observed QTL significance of Phenotypes 1–4 from a standard QTL mapping technique is shown in graded shades of purple, with a darker color representing a stronger QTL significance. 2) Perturbation of true genotypes. Specifically, perturbing the genotype at a particular marker of a correct sample (correct → wrong) will lead to a decreased QTL significance for all molecular traits mapping to a QTL near that marker. In this panel, the genotype of the 3rd sample at the 5th marker position (indicated by an arrow circle) is randomly perturbed (changed from the orange to the blue allele). Then we re-map the QTLs using the perturbed genotype data and unchanged phenotype matrix and observe that for most of the QTL the significance decreases (dark color changes to light color), i.e. the QTL loses significance if noise is added. 3) Correction of wrong genotypes. The genotype of the 2rd sample at the 5th marker position (indicated by an arrow circle) is randomly perturbed (changed from the blue to orange allele). Then we re-map the QTLs using the perturbed genotype data and unchanged phenotype matrix. In this case, for most of the QTL the significance increases after perturbation (light color changes to dark color) since the original genotype was wrong. When such an increase is observed for a number of phenotype —marker pairs, it suggests that the genotype of this sample was mislabeled.

doi:10.1371/journal.pone.0171324.g001

reGenotyper mode of action

The reGenotyper algorithm works as follows: perturbing the genotype at a particular marker of a correct sample (correct → wrong) will lead to a decreased QTL significance (e.g. $\Delta t = t^{new} - t^{old} < 0$, when using the t statistic) for all molecular traits mapping to a QTL near that marker (Fig 1). This would be true for as many molecular traits—marker combinations as there are QTLs,

when the genotype of this sample was originally correctly labelled. Therefore, the distribution of Δt values for the correct sample n ($\Delta t^n_{k,m}$ values from K traits at M markers) would roughly show a single-component distribution, and the mean of the distribution is expected to be smaller than zero: the reason is that when perturbing the genotype of a correct sample, the significance of QTLs is expected to be reduced by the introduction of wrong marker data. In contrast, perturbing the genotype at a particular marker of a mislabeled sample (wrong → correct) will lead to an improved QTL significance (e.g. $\Delta t = t^{new} - t^{old} < 0$, when using t statistic) for all molecular traits mapping to a QTL near that marker (Fig 1). This would be true for as many molecular traits—marker combinations as there are QTLs, when the genotype of this sample was originally mislabeled. Therefore, when a sample has been mislabeled, the distribution of Δt values for this sample n ($\Delta t^n_{k,m}$ values from K traits at M markers) would show a mixture distribution, with one extra component with a mean larger than 0. Although Δt for a certain trait at a certain marker might show a non-zero value simply by chance, the identification of mislabeled samples is based on the Δt values from a large number of traits measured in parallel (e.g., the top x QTLs at one specific marker) at a number of marker positions (different and largely independent sets of QTLs at those markers), thus making use of the data-richness of genetical genomics experiments. Therefore, the distribution of all $\Delta t^n_{k,m}$ ($k = 1,\ldots,K; m = 1,\ldots,M$) from one sample can actually be used to evaluate the chance that a sample is mislabeled. reGenotyper also provides an estimate of the "ideal" genotype of the potential mislabeled sample and identifies the best-matched labels for each sample, from a collection of potential genotypes. The mathematical details of the algorithm and the significance assessment are provided in the S1 File.

Test datasets

We showcase reGenotyper as a method to detect and correct mislabeled samples using five previously published [17,18] and two unpublished *C. elegans* datasets. The recombinant inbred lines (RILs) used in these studies were derived from *C. elegans* wild-types N2 and CB4856. In the temperature experiment [17], genome-wide gene expression levels of 80 RILs were profiled at two conditions (16˚C and 24˚C). In the aging experiment [18], gene expression profiles of 35 RILs were measured at three stages (40 h, 96 h and 214 h). In a more recent unpublished experiment on lines from the same collection, gene expression levels of 46 and 38 RILs were profiled in two different conditions. In total, 200 different RILs were used, with considerable overlap between the studies, and the lines were SNP-genotyped with an average density of 1 marker per centi-Morgan. All the published datasets can be accessed using WormQTL [www.wormqtl.org] [24] or WormQTLHD [www.wormqtl-hd.org] [25] web portals.

Results

reGenotyper has high power to detect mislabeled samples

The ability of reGenotyper to detect wrongly labelled samples will obviously depend on how similar the assumed (wrong) genotype is to the real genotype. We first evaluated this effect through simulation using the area under the (receiver operating characteristic) curve (AUC) approach. Specifically, we studied the performance of the method when the original true genotypes showed different proportions of common marker genotypes with the mistakenly used genotype. The more similar the mistakenly used genotype is to the correct genotype, the more difficult it should be to identify the sample as mislabeled. In S2b Fig we can see that the AUC is around 0.8 when the true genotype shares 90% of its marker genotypes with the mislabeled sample, although in real datasets such a high level of genetic likeness is extremely rare. In the case of mouse recombinant inbred lines, the probability that two randomly selected lines have the same

genotypes in 18 out of 20 chromosomes (i.e. high similarity = 90%) can be roughly estimated to be $0.5^{18} = 3.8\times10^{-6}$. With increasing dissimilarity between the two genotypes, the performance increases to an AUC of almost 100%, i.e. a perfect detection of all wrongly labeled samples.

reGenotyper detects mislabeled samples in seven sets of expression genetics data from worm studies

We applied the reGenotyper on seven *C. elegans* datasets, as described in the Test Datasets subsection of the Materials and Methods (Fig 1). Four samples showed strong evidence (mislabeling score > 0.9) of being mislabeled in at least two different experiments (Fig 2 and S3 Fig). The WN53 line was the top candidate for being wrongly labeled: this line had the first rank amongst potentially mislabeled samples in all four experiments in which it had been used.

If samples have been mistakenly swapped with other samples in the same RIL collection, the "true" RIL genotype can be recovered by comparing the estimated genotype with all the known genotypes. Applying reGenotyper to the four mislabeled samples detected above, all of them show consistent candidate RILs in at least three independent experiments. Most interestingly, WN54 was identified as a candidate for being the true RIL for WN53 in all four experiments. In order to verify our findings, we re-genotyped the WN53 line using an Illumina SNP array. The genotyping results based on 96 equally spaced SNP markers clearly showed that indeed sample labeled as WN53 is actually WN54. A follow-up quantitative re-analysis of the temperature experimental data demonstrated the substantial gain from applying reGenotyper. Correcting the genotype for the 5 mislabeled samples leads to better consistency in the detected *trans*-eQTL, i.e. the number of *trans*-eQTL shared between two temperatures increased by 111% (from 93 to 190, FDR$<$0.05) (S4 Fig). This result indicates that correcting mislabeled samples could resolve, at least in part, the commonly observed lack of consistency in *trans*-QTL across studies [14,26].

reGenotyper detects mislabeled samples in publicly available expression genetics data from mouse, worm and yeast studies

As we had identified a number of mislabeled samples in the worm datasets affecting the power to detect eQTLs, we applied our algorithm to several publicly available datasets for which the genotype and transcriptome data were available online.

1. Among 208 recombinant inbred advanced intercross lines from *C. elegans* [27], four samples (1.9%) showed a mislabeling score of $>$90%, based on 1000 permutations. The corrected mapping can be found in WormQTL [24] and WormQTLHD [25] portals.
2. Among 96 recombinant inbred lines from mouse [28], two samples (2%) showed strong evidence of being mislabeled (mislabeling score $>$90%), based on 1000 permutations.
3. None of the yeast segregants [29] was detected as mislabeled with the same threshold, but one sample showed suggestive evidence of being mislabeled (S5 Fig).

In each case, a number of additional samples were highlighted as suspicious and could be followed up experimentally when resources permit.

reGenotyper can help to correct the mislabeled samples it detects in three user-defined ways

1. The simplest way is removing those unreliable samples from the dataset and performing the further analysis based on the remaining samples. This most conservative approach

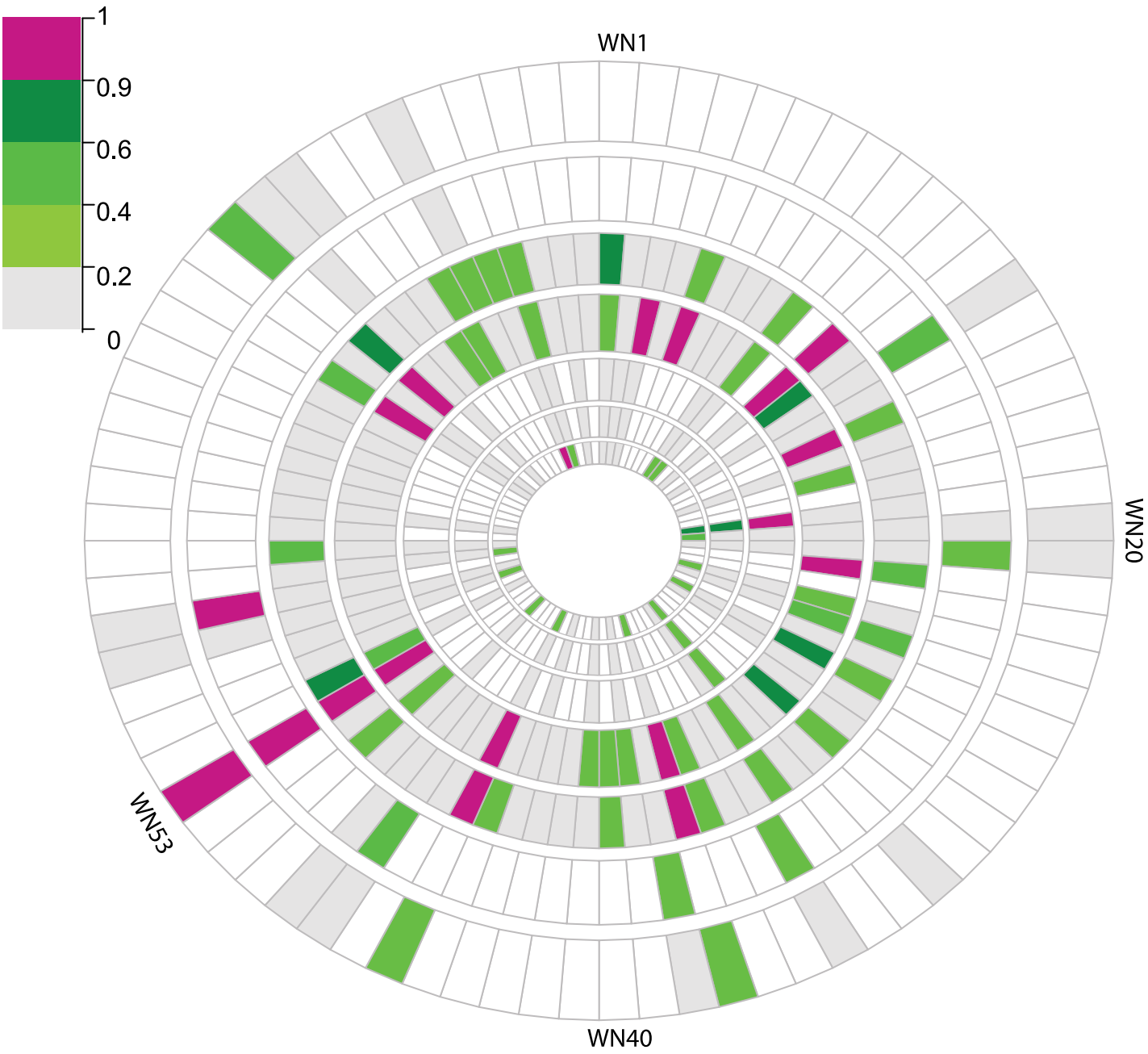

Fig 2. Individual evidence of the samples (arranged around the circle) being detected as potentially mislabeled sample (MS) across seven different experiments (each represented by a circle) from *C. elegans* studies using the reGenotyper method. Different shades of green represent the mislabeling score, with a darker color corresponding to a higher score, and magenta indicating that the sample has been detected as MS with a score larger than 90%. White indicates that the sample was not used in this experiment (as not all samples were used in all experiments). The samples with consistent strong evidence of being potentially mislabeled across experiments (i.e., showing high scores in multiple experiments) are more likely to indeed be mislabeled. Note that sample WN53 shows a mislabeling score larger than 0.9 in four independent experiments, making it very likely that it was indeed mislabeled, as confirmed by subsequent experiments.

doi:10.1371/journal.pone.0171324.g002

would remove dubious information from the original data, but it also leads to a decrease in the sample size and a concomitant reduction of statistical power. We performed a simulation study to show an effect of removing the samples. We used the median of the detected QTL significance distribution as a benchmark score. We removed samples in order of their

chance of being mislabeled (i.e. according to their mislabeling score). This increased the benchmark score, until the point of removing samples with little evidence of being mislabeled (i.e. correct samples). (S6 Fig).

2. A possibly more efficient way of handling the mislabeled data would be to recover the true genotypes for the mislabeled samples. Not discarding molecular profiling information of the mislabeled samples could potentially lead to obtaining even more power for QTL detection. For example, in a simulated genetical genomics experiment (details can be found in the S1 File), the genotypes of six samples (10%) were manually swapped (i.e. exchanged between two samples that were both included in the dataset). The proportion of common marker genotypes between true genotype and mistakenly used one was on average 60% (i.e. the original genotype and the wrong new genotype still had, on average, 60% identical marker genotypes but differed for the remaining 40%). In 200 simulations, on average 23% additional simulated QTLs could be detected using the inferred true genotypes compared to using mislabeled genotypes (S7 Fig).
3. In the case that genotype information is available for a larger collection of many strains, reGenotyper can also identify the best-matched genotype for the detected mislabeled samples. This step helps to correct the swapping of sample labels in the lab (as showcased in results section).

Discussion

reGenotyper can be used to systematically examine the molecular profiling phenotype data to directly identify potentially mislabeled samples

To our best knowledge, there is no other tool capable of such a task. Even though there has been numerous attempts to address sample mix-ups detection [19,30–33]. A method called MixupMapper, was proposed to correct sample mix-ups in gene expression data for human genome-wide association studies [19]. However, it is based solely on the expression of genes which are influenced by genetic variation located near these genes (*cis*-eQTL). MixupMapper aims to detect pairs of samples for which the genotype information has been mistakenly swapped, but both genotypes have been measured and both samples have been included in the experiment. Thus, this method cannot find mislabeled samples that result from slightly more complex (but still very likely) experimental mix-ups, e.g. the use of wrong samples from a larger collection, or accidental duplicate measurements. Additionally, a Bayesian approach was reported to predict SNP genotypes based on RNA expression data, and then to match the predicted genotype to the observed genotype of individuals in large populations [30]. It makes use of the consistency of *cis*-eQTL across different tissues and therefore requires training data from one or more independent tissues; unfortunately this kind of replication is not available for most studies. Moreover, all of the above methods depend on eQTL, which limits their usage to gene expression studies, whereas our reGenotyper method could in principle be applied to any type of high-throughput phenotype data, including the increasingly popular and powerful genetic analysis of protein and metabolite profiles [4,5]. The tool is able to perform an analysis within minutes (*C. elegans* data) to hours (human data) on a personal computer. However, we recommend use of parallel computing (e.g. cloud services) so that more permutations can be performed increasing accuracy of reGenotyper.

Limitations of reGenotyper

The package was built so that detection of sample mix-ups is performed quickly and accurately for large numbers of phenotypes. However, our heuristic "significance change value" approach is only valid if phenotype data is of massively parallel nature. Moreover, the package needs numerous phenotypes with considerable number of highly significant QTLs in order to perform meaningful permutation analysis. This limits reGenotyper usage to the high-throughput molecular studies where measuring thousands of phenotypes is feasible.

On the other side of the spectrum are studies generating truly big data. For example, genomic studies in humans utilize tens of millions of SNP markers [34]. Permuting a dataset with thousands of phenotypes mapped to tens of millions of markers might prove infeasible. It is possible to use our package on a High Performance Computing (HPC) cluster with an aid of specialized platform like xQTL workbench [35], but this requires computing expertise and access to an HPC cluster.

The accuracy of sample mix-up detection may also be affected by gene—environment interactions, as environmental variation induces major changes in phenotype that might not be linked to genetics. This can result in discovering false mix-ups. On the other hand, measuring phenotypes in multiple environments allows untangling environmental and genetic components of observed phenotypical variation [36]. Analyzing phenotypes from each of the environment separately with reGenotyper and then comparing the results increases the accuracy of mislabeling detection even further.

We aimed at as much automation of the workflow as possible. However, there is still a crucial step that requires user intervention. Samples are marked as mislabeled based on a user-defined threshold. Even though the threshold is statistically well-defined (see S1 File), the choice of a more or less stringent cut-off will depend on a user's preferences (e.g, the perceived cost of missing a mislabeled sample) and expectations (e.g., the predicted frequency of mislabeled samples in a dataset).

Supporting information

S1 Fig. Results of a simulated genetical genomics experiment. Experiment without a mislabeled sample (a) and with mislabeled samples (b). Grey and red correspond to values of 0 and 1 for the S element value, respectively. The rows represent markers along the genome and the columns represent different samples. (a) S values of 1 (red spots) are scattered in the figure, indicating that no sample has been mislabeled. (b) Vertical red bars indicate that the corresponding samples (columns) are very likely to have been mislabeled. In this simulated experiment, six (~10%) mislabeled samples sharing about 50% common marker genotypes with the correct sample were included, and these were all clearly identified.
(PNG)

S2 Fig. Detection of mislabelling between similar samples. a) Comparison of the distribution of Δt for a correctly labeled sample (green) and a wrongly labeled sample (magenta) from a simulated dataset. The wrongly labeled sample and the true genotype are 70% identical. b) Box plot of the area under the receiver operating characteristic curve (AUC) for the reGenotyper algorithm for different scenarios. The x-axis represents ten different situations, i.e. different proportions of common marker genotypes between the mislabeled sample and the true genotype of this sample; in the most challenging scenario, the mislabeled sample shares 95% of its marker genotypes with the true sample. Under each scenario, 50 simulations were carried out. In each simulation, a gene expression dataset of 100 samples with 10% mislabeled samples

was simulated.
(PNG)

S3 Fig. Individual evidence of the 200 RIL samples (column) being detected as potentially mislabeled. Sample across seven different experiments (rows) from our lab using the reGenotyper algoritm. Different shades of green gradients represent the mislabeling score. Darker green corresponds to higher scores, while magenta indicates that the sample was detected as MS with a score larger than 90%. White indicates that the sample was not selected for use in this experiment.
(PNG)

S4 Fig. Impact of removal of MS on QTL mapping. Figure compares results of QTL mapping on original data (left) and after removal of 5 MS (right). Removal of MS detected by reGenotyper results in sharp increase in number of QTLs shared between two temperatures used in the study (from 93 to 190–111%).
(PNG)

S5 Fig. Mislabeling scores in different experiments from a) worm, b) mouse and c) yeast. Respectively, 4 out of 208, 2 out of 96 and 0 out of 109 samples show a mislabeling score >0.9 (dashed line).
(PNG)

S6 Fig. Changes in QTL significance after removing samples. The genotype and phenotype data were simulated for 60 RILs in the same way as described in the S1 File. Three (5%) mislabeled genotypes were added to the data set. The samples were removed in decreasing order of their mislabeling score. First, three mislabeled samples (MS) and later samples with little chance of being mislabeled (i.e. low mislabeling score), correct samples, (CS) were removed. The median of detected QTLs (-log10P) significance distribution (orange line) increases after removing MS and drops again after removing CS. (x axis, rm = removing).
(PNG)

S7 Fig. Changes in QTL significance after recovering the true genotypes for the mislabeled samples in simulated data. 200 simulations were performed and QTL mapping was performed a) before swapping samples to simulate mislabeling, b) after swapping samples and c) after swapping samples and recovering true genotypes using reGenotyper. Plots shows rate of: red 1 –phenotypes with simulated QTL that was mapped; blue 2 –phenotypes without simulated QTL, where no QTL was mapped; green 3 –phenotypes without simulated QTL, where QTL was mapped. After recovery with reGenotyper almost all of the QTLs could be mapped correctly, compared to 77% in case with MS.
(PNG)

S1 File. Detailed description of the reGenotyper algorithm.
(DOC)

Acknowledgments

We thank Frank Johannes, Groningen Bioinformatics Centre, University of Groningen, for helpful discussions and Jackie Senior for editing the final text.

This work was supported by the 7th Framework Programme of the European Commission under the Research Project PANACEA [Contract No. 222936 to R.C.J. and J.E.K.]; the Netherlands Organisation for Scientific Research (NWO) VENI grant [n° 863.13.011 to Y.L.]; and the Dutch Carbohydrate Competence Center, which is is co-financed by the European Regional

Development Fund, the Dutch Ministry of Economic Affairs (as part of Pieken in de Delta, the government's regional economic agenda), the Municipality of Groningen and the Province of Groningen [CCC WP23 to K.Z.];

Author Contributions

Conceptualization: JEK RB RCJ YL.

Data curation: KZ KJV DA YL.

Formal analysis: KZ YL.

Funding acquisition: JEK RCJ YL.

Resources: BLS ME MR KJV DA MAS GP JEK.

Software: KZ YL.

Supervision: RB JEK RCJ.

Visualization: KZ DA YL.

Writing – original draft: KZ BLS KJV DA HJW JEK RB RCJ YL.

Writing – review & editing: KZ RB YL.

References

1. Baute J, Herman D, Coppens F, De Block J, Slabbinck B, Dell'Acqua M, et al. Correlation analysis of the transcriptome of growing leaves with mature leaf parameters in a maize RIL population. Genome Biol. 2015; 16: 168. doi: 10.1186/s13059-015-0735-9 PMID: 26357925
2. Ongen H, Andersen CL, Bramsen JB, Oster B, Rasmussen MH, Ferreira PG, et al. Putative cis-regulatory drivers in colorectal cancer. Nature. 2014; 512: 87–90. doi: 10.1038/nature13602 PMID: 25079323
3. Zhang X, Joehanes R, Chen BH, Huan T, Ying S, Munson PJ, et al. Identification of common genetic variants controlling transcript isoform variation in human whole blood. Nat Genet. 2015; 47: 345–352. doi: 10.1038/ng.3220 PMID: 25685889
4. Albert FW, Treusch S, Shockley AH, Bloom JS, Kruglyak L. Genetics of single-cell protein abundance variation in large yeast populations. Nature. 2014; 506: 494–497. doi: 10.1038/nature12904 PMID: 24402228
5. Raffler J, Friedrich N, Arnold M, Kacprowski T, Rueedi R, Altmaier E, et al. Genome-Wide Association Study with Targeted and Non-targeted NMR Metabolomics Identifies 15 Novel Loci of Urinary Human Metabolic Individuality. PLoS Genet. 2015; 11: e1005487. doi: 10.1371/journal.pgen.1005487 PMID: 26352407
6. Draisma HHM, Pool R, Kobl M, Jansen R, Petersen A-K, Vaarhorst AAM, et al. Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. Nat Commun. 2015; 6: 7208. doi: 10.1038/ncomms8208 PMID: 26068415
7. Jansen RC. Studying complex biological systems using multifactorial perturbation. Nat Rev Genet. 2003; 4: 145–51. doi: 10.1038/nrg996 PMID: 12560811
8. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. Trends Genet. 2001; 17: 388–91. PMID: 11418218
9. Sieberts SK, Schadt EE. Moving toward a system genetics view of disease. Mamm Genome. 2007; 18: 389–401. doi: 10.1007/s00335-007-9040-6 PMID: 17653589
10. Broman KW, Keller MP, Broman AT, Kendziorski C, Yandell BS, Sen Ś, et al. Identification and Correction of Sample Mix-Ups in Expression Genetic Data: A Case Study. G3 Bethesda Md. 2015; 5: 2177–2186.
11. Buyske S, Yang G, Matise TC, Gordon D. 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. 2009; 67: 287–292. doi: 10.1159/000194981 PMID: 19172087
12. Broman KW, Wu H, Sen S, Churchill GA. R/qtl: QTL mapping in experimental crosses. Bioinforma Oxf Engl. 2003; 19: 889–890.

13. Arends D, Prins P, Jansen RC, Broman KW. R/qtl: high-throughput multiple QTL mapping. Bioinforma Oxf Engl. 2010; 26: 2990–2992.
14. Peirce JL, Li H, Wang J, Manly KF, Hitzemann RJ, Belknap JK, et al. How replicable are mRNA expression QTL? Mamm Genome. 2006; 17: 643–56. doi: 10.1007/s00335-005-0187-8 PMID: 16783644
15. Jansen RC. Controlling the type I and type II errors in mapping quantitative trait loci. Genetics. 1994; 138: 871–881. PMID: 7851782
16. Jansen RC. Quantitative trait loci in inbred lines. In: Balding D, Bishop M, Cannings C, editors. Handbook of Statistical Genetics. 3rd ed. New York: John Wiley & Sons; 2007. p. pp 589-618.
17. Li Y, Alvarez OA, Gutteling EW, Tijsterman M, Fu J, Riksen JA, et al. Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. PLoS Genet. 2006; 2: e222. doi: 10.1371/journal.pgen.0020222 PMID: 17196041
18. Viñuela A, Snoek LB, Riksen JAG, Kammenga JE. Genome-wide gene expression regulation as a function of genotype and age in C. elegans. Genome Res. 2010; 20: 929–937. doi: 10.1101/gr.102160.109 PMID: 20488933
19. Westra H-J, Jansen RC, Fehrmann RSN, te Meerman GJ, van Heel D, Wijmenga C, et al. MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. Bioinforma Oxf Engl. 2011; 27: 2104–2111.
20. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet. 2005; 37: 710–7. doi: 10.1038/ng1589 PMID: 15965475
21. Chaibub Neto E, Ferrara CT, Attie AD, Yandell BS. Inferring causal phenotype networks from segregating populations. Genetics. 2008; 179: 1089–100. doi: 10.1534/genetics.107.085167 PMID: 18505877
22. Li Y, Tesson BM, Churchill GA, Jansen RC. Critical reasoning on causal inference in genome-wide linkage and association studies. Trends Genet TIG. 2010; 26: 493–498. doi: 10.1016/j.tig.2010.09.002 PMID: 20951462
23. R Programming Language for Statistical Computing [Internet]. http://cran.r-project.org/index.html
24. Snoek LB, Van der Velde KJ, Arends D, Li Y, Beyer A, Elvin M, et al. WormQTL—public archive and analysis web portal for natural variation data in Caenorhabditis spp. Nucleic Acids Res. 2013; 41: D738–743. doi: 10.1093/nar/gks1124 PMID: 23180786
25. van der Velde KJ, de Haan M, Zych K, Arends D, Snoek LB, Kammenga JE, et al. WormQTLHD—a web database for linking human disease to natural variation data in C. elegans. Nucleic Acids Res. 2014; 42: D794–801. doi: 10.1093/nar/gkt1044 PMID: 24217915
26. Gutteling EW, Riksen JA, Bakker J, Kammenga JE. Mapping phenotypic plasticity and genotype-environment interactions affecting life-history traits in Caenorhabditis elegans. Heredity. 2006; advance online publication.
27. Rockman MV, Skrovanek SS, Kruglyak L. Selection at linked sites shapes heritable phenotypic variation in C. elegans. Science. 2010; 330: 372–376. doi: 10.1126/science.1194208 PMID: 20947766
28. Gerrits A, Li Y, Tesson BM, Bystrykh L V., Weersing E, Ausema A, et al. Expression quantitative trait loci are highly sensitive to cellular differentiation state. PLoS Genet. 2009; 5: e1000692. doi: 10.1371/journal.pgen.1000692 PMID: 19834560
29. Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc Natl Acad Sci U A. 2005; 102: 1572–7.
30. Schadt EE, Woo S, Hao K. Bayesian method to predict individual SNP genotypes from gene expression data. Nat Genet. 2012; 44: 603–608. doi: 10.1038/ng.2248 PMID: 22484626
31. Lynch AG, Chin S-F, Dunning MJ, Caldas C, Tavaré S, Curtis C. Calling Sample Mix-Ups in Cancer Population Studies. PLoS ONE. 2012; 7: e41815. doi: 10.1371/journal.pone.0041815 PMID: 22912679
32. Ekstrøm CT, Feenstra B. Detecting sample misidentifications in genetic association studies. Stat Appl Genet Mol Biol. 2012; 11: Article 13.
33. Baggerly KA, Coombes KR. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. Ann Appl Stat. 2009; 3: 1309–1334.
34. Danjou F, Zoledziewska M, Sidore C, Steri M, Busonero F, Maschio A, et al. Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. Nat Genet. 2015; 47: 1264–1271. doi: 10.1038/ng.3307 PMID: 26366553
35. Arends D, van der Velde KJ, Prins P, Broman KW, Möller S, Jansen RC, et al. xQTL workbench: a scalable web environment for multi-level QTL analysis. Bioinforma Oxf Engl. 2012; 28: 1042–1044.
36. Li Y, Breitling R, Jansen RC. Generalizing genetical genomics: getting added value from environmental perturbation. Trends Genet TIG. 2008; 24: 518–524. doi: 10.1016/j.tig.2008.08.001 PMID: 18774198