RESEARCH ARTICLE

REVISED Whose sample is it anyway? Widespread misannotation of samples in transcriptomics studies [version 2; referees: 2 approved, 1 approved with reservations]

Lilah Toker[1,2], Min Feng[1-3], Paul Pavlidis[1,2]

[1]Department of Psychiatry, University of British Columbia, Vancouver, V6T 2A1, Canada
[2]Michael Smith Laboratories, University of British Columbia, Vancouver, V6T 1Z4, Canada
[3]Graduate Program in Genome Sciences and Technology, University of British Columbia, Vancouver, V5Z 4S6, Canada

v2 **First published:** 30 Aug 2016, **5**:2103 (doi: 10.12688/f1000research.9471.1)
Latest published: 30 Sep 2016, **5**:2103 (doi: 10.12688/f1000research.9471.2)

Abstract

Concern about the reproducibility and reliability of biomedical research has been rising. An understudied issue is the prevalence of sample mislabeling, one impact of which would be invalid comparisons. We studied this issue in a corpus of human transcriptomics studies by comparing the provided annotations of sex to the expression levels of sex-specific genes. We identified apparent mislabeled samples in 46% of the datasets studied, yielding a 99% confidence lower-bound estimate for all studies of 33%. In a separate analysis of a set of datasets concerning a single cohort of subjects, 2/4 had mislabeled samples, indicating laboratory mix-ups rather than data recording errors. While the number of mixed-up samples per study was generally small, because our method can only identify a subset of potential mix-ups, our estimate is conservative for the breadth of the problem. Our findings emphasize the need for more stringent sample tracking, and that re-users of published data must be alert to the possibility of annotation and labelling errors.

This article is included in the Preclinical Reproducibility and Robustness channel.

This article is included in the Neuroinformatics channel.

Open Peer Review

Referee Status: ✓ ✓ ?

	Invited Referees		
	1	**2**	**3**
REVISED **version 2** published 30 Sep 2016			? report
version 1 published 30 Aug 2016	✓ report	✓ report	

1 **Leonard P. Freedman**, Global Biological Standards Institute USA

2 **Hans van Bokhoven**, Radboud University Medical Center Netherlands

3 **Levi Waldron**, City University of New York USA

Discuss this article

Comments (0)

Corresponding authors: Lilah Toker (ltoker@mail.ubc.ca), Paul Pavlidis (paul@chibi.ubc.ca)

How to cite this article: Toker L, Feng M and Pavlidis P. **Whose sample is it anyway? Widespread misannotation of samples in transcriptomics studies [version 2; referees: 2 approved, 1 approved with reservations]** *F1000Research* 2016, **5**:2103 (doi: 10.12688/f1000research.9471.2)

Grant information: The research was supported by grants to PP from NIH (GM076990) and the NeuroDevNet Network of Centres of Excellence, and the UBC Graduate program in Genome Sciences and Technology (MF).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 30 Aug 2016, **5**:2103 (doi: 10.12688/f1000research.9471.1)

REVISED **Amendments from Version 1**

As noted by Hans van Bokhoven, in the original version the the number of studies with mislabelled samples was incorrectly indicated in the text as “31” instead of “32”. This is now corrected.

See referee reports

Introduction

Recent years have seen an increase in concern about the quality of scientific research, along with efforts to improve reliability and reproducibility[1,2]. These issues are highly relevant to genomics studies, which deal with complex and often weak signals measured genome-wide. In transcriptomics studies (our focus here), mRNA is extracted from samples and processed using microarrays or RNA-seq, followed by statistical analysis to identify patterns of interest (*e.g.* differential expression). Much work has been done to raise awareness of technical issues in such studies such as RNA quality[3] and batch effects[4] and many investigators are aware of the need to address them[5]. Alongside, a great effort was put into establishing guidelines for annotation standards of expression data into public repositories[6].

A key step in many scientific experiments, which has received less attention, is the importance of maintaining an accurate correspondence between the experimental conditions or sources of the samples and the eventual data. Simply put, for the analysis to be valid, the samples must not be mixed up. If mix-ups are present but undetected, the conclusions of the analysis might be affected and pollute the literature, as well as create a lurking problem for those who re-use the data.

The obviousness of the need to avoid mix-ups suggests that investigators should be well aware of the risk, and take steps to reduce it, such as careful bookkeeping (*e.g.*, permanent sample tube labels matched to data files). However, we recently became concerned that mix-ups might not be rare. Our concerns came to a head when we reanalyzed four publically available datasets of Parkinson’s disease subjects[7]. As part of our quality checks of the data, we examined expression levels of sex-specific genes (genes expressed only in males or in females), and compared these with the corresponding subject sex meta-data annotations from each of the papers. To our surprise, we found discordance between the sex predicted based on expression levels of sex-specific genes and the manuscript-annotated sex in two out of the four datasets[7] (Supplementary Figure S1). This finding, and other anecdotal observations, led us to examine this issue more broadly.

Sex-specific genes are well-suited for this purpose. In genetics studies, genotypes of the sex chromosome are routinely used to identify mislabeled samples[8,9], moreover, sex check is a built-in option for some of the dedicated software[10]. Given that genetic abnormalities resulting in disagreement between genotypic and phenotypic sex are rare[11], any disagreements are very likely to stem from errors and may also be indicative of other dataset quality issues. Using such genes for quality checks of transcriptome data is not widespread practice, but it is well known that several X- and Y-linked genes show sex-specific patterns of expression. A limitation of this approach is that mix-ups that do not yield conflicting sex labels (*e.g.*, swapping two female samples) cannot be detected. But at the very least the sex-specific-gene-based approach can provide a lower bound for the amount of mix-ups and if any are detected it should trigger a reassessment of the tracking of all samples in the study.

In this study, we focused on publically available human expression profiling experiments that included individuals of both sexes. To our surprise, we found strong evidence of mix-ups in nearly half of them. Importantly, for the vast majority of the studies we were able to validate that the disagreement between metadata- and gene-based sex is prevalent in the original manuscript. This indicates that the disagreements are not a result of erroneous sex description during data submission to public repository. An additional 10% of the studies have samples of ambiguous gene-based sex, suggesting the possibility of samples being mistakenly combined or other quality problems. While it is possible that a small number of the cases we identify are due to sex chromosome abnormalities, we regard the most likely explanation for most to be laboratory mix-ups or errors in the meta-data annotations. Our findings suggest a widespread quality control issue in transcriptomics studies.

Methods

Except where mentioned, data analysis was performed using the R/Bioconductor environment[12,13]. Source code for the analysis is available in a Github repository (https://github.com/min110/mislabeled.samples.identification). The archived version of the code at the time of publication can be accessed through Zenodo **mislabeled.samples.identification:** doi:10.5281/zenodo.60313.

We identified datasets containing sex information as experimental factors by searching the Gemma database[14]. Out of an initial 121 datasets we focused on 79 studies run on the Affymetrix HG-U1333Plus_2 and HG-U133A platforms as they have the same sex marker genes (GEO platform identifiers GPL570 and GPL96 respectively). The annotations in Gemma, which originate from GEO sample descriptions augmented with manual annotation, were re-checked against GEO, resulting in the correction of errors for 14 samples. Datasets that contained samples of only one sex, represented data from sex-specific tissues (*e.g*. ovary or testicle) or contained numerous missing values were excluded (nine datasets). A final set of 70 studies (a total of 4160 samples) met the criteria. Table 1 summarizes the data included and full details of each study are in Supplementary Table S1. Whenever possible, data were reanalyzed from .CEL files. The signals were summarized using RMA method from the Affymetrix “power tools” (http://media.affymetrix.com/partners_programs/programs/developer/tools/powertools.affx), $\log_2$ transformed and quantile normalized as part of the general Gemma pre-processing pipeline.

Probeset selection: The male-specific genes *KDM5D* and *RPS4Y1* are represented by a single probeset on both platforms included in our analysis. *XIST* is represented by two probesets on the GPL96 platform and by seven probesets on the GPL570 platform. With the exception of the 221728_x_at probeset, *XIST* probesets were highly correlated with each other, and negatively correlated with the *KDM5D* and *RPS4Y1* expression in all of the datasets analyzed (Supplementary Figure S3). The poor-performing *XIST* probeset (221728_x_at) was excluded from further analysis. The final set was four probesets for GPL96 and eight probesets for GPL570.

Table 1. Summary of discrepancies between the gene expression-based and annotated sex in human microarray datasets. Unclassified samples are samples with disagreement between their classification using k-means clustering and the median expression of the sex specific probesets. Datasets were considered as "correctly annotated" only if they did not contain mismatched or unclassified samples. Eight of the datasets contained both mismatched samples and unclassified samples.

	All Datasets	Non-cancer Datasets	Cancer Datasets	All Samples	Non-cancer Samples	Cancer Samples
Correctly annotated	31 (44%)	29 (53%)	2 (13%)	4043 (97%)	2868 (98%)	1175 (96%)
Mismatched	32 (46%)	24 (44%)	8 (53%)	83 (2%)	58 (1.97%)	25 (2.04%)
Unclassified	15 (21%)	7 (13%)	8 (53%)	34 (0.8%)	11 (0.4%)	23 (1.9%)
Total	70	55	15	4160	2937	1223

Assigning gene-based (biological) sex to samples: The expression data for the selected sex markers were extracted from the normalized data for each dataset. For each of these small expression matrices, we applied standard k-means clustering (using the "kmeans" function from the "stats" package in R[15] to classify the samples into two clusters. We assigned the two clusters as "male" or "female", based on the centroid values of each of the probesets: specifically, the cluster with higher values of the *XIST* probesets centroids and a lower value of *KDM5D* and *RPS4Y1* centroids was assigned as a "female" cluster. To identify samples with ambiguous sex, we calculated the difference between the median expression level of the *XIST* probesets and the median expression level of the *KDM5D* and *RPS4Y1* probesets. We compared this difference with the cluster-based gender, and validated that the difference is positive for samples assigned as females and negative for samples assigned as males. We excluded 34 samples that showed disagreement in this comparison since they could not provide a conclusive result for the gene—expression-based sex. We note that 12 (35%) of these would have been assigned to a cluster contradicting their annotated sex if we had retained them.

Manual validation of the discrepancy between the gene-based sex and the meta-data-based sex: For all the cases where a discrepancy was found between the gene-expression-based sex and the meta-data-based sex, we manually examined the original studies to check if the mismatch was due to incorrect annotation of the sample during the data upload to GEO, or was present in the original paper. Since most of the manuscripts only contain summary statistics of the demographic data (13/32, Supplementary Table S2), direct sample-by-sample validation was not possible for most studies. For these studies we used the highest resolution level of group summary statistics, provided in the publication to validate that the data in the paper corroborate the data in GEO. In addition, for all of the datasets with mismatched samples, we manually evaluated the expression values of the relevant probesets using the GEO2R tool on the GEO website.

Confidence interval estimate for population proportion of studies with misannotated samples: We used the properties of the binomial distribution to compute the confidence interval for the population estimate of affected data sets using the "qbinom" function in R.

Analysis of Stanley Foundation datasets: CEL files and sample metadata were downloaded directly from the Stanley Medical Research Institute genomic database (https://www.stanleygenomics.org/stanley/). CEL files were pre-processed, quantile normalized and $\log_2$ transformed using the rma function from the "affy" package in R Bioconductor[12,13].

Results

We identified a corpus of 70 human gene expression studies that had sample sex annotation (4160 samples in total) run on two platforms. We developed a simple robust method for classifying samples by sex based on three sex specific genes – *XIST*, *RPS4Y1* and *KDM5D*. *XIST* (*X-inactive specific transcript*) is expressed from the inactive X chromosome and acts to silence its expression and thus, is only expressed in female subjects. *KDM5D* (*Lysine (K)-Specific Demethylase 5D*) and *RPS4Y1* (*Ribosomal Protein S4, Y-Linked 1*) are both located on the Y chromosome, and thus are only expressed in male subjects. Although additional sex-specific genes exist, we determined that *KDM5D, RPS4Y1* and *XIST* are the only sex-specific genes consistently showing high expression levels in the associated sex in all tissues. Our method assigns a predicted sex based on gene expression to each sample, which we refer to as "gene-based sex" (see Methods). We also performed a second analysis to identify samples where a gene-based sex could not be confidently assigned. Such samples might reflect technical problems, but could also be due to true biological effects; for example, *XIST* expression is altered in some cancers and in early stages of development[16]. We then compared gene-based sex to the sex according to the provided sample annotations ("meta-data-based sex") for the 70 studies, seeking samples with disagreements. Figure 1 shows examples of studies with no discrepant samples (1A) and with discrepancies (1B). Similar plots for all datasets analyzed are shown in Supplementary Figure S2. All calls of discrepant or ambiguous sex were followed by manual confirmation.

We found samples with a discrepancy between the meta-data sex information and the gene-based sex in 32/70 (46%) of the datasets (ambiguous samples excluded; summarized in Table 1; details in Supplementary Table S2). Although datasets containing mismatched samples were more prevalent among cancer datasets (53% vs 44%, cancer vs. non-cancer, respectively), the proportion of mismatched samples was similar in cancer and non-cancer

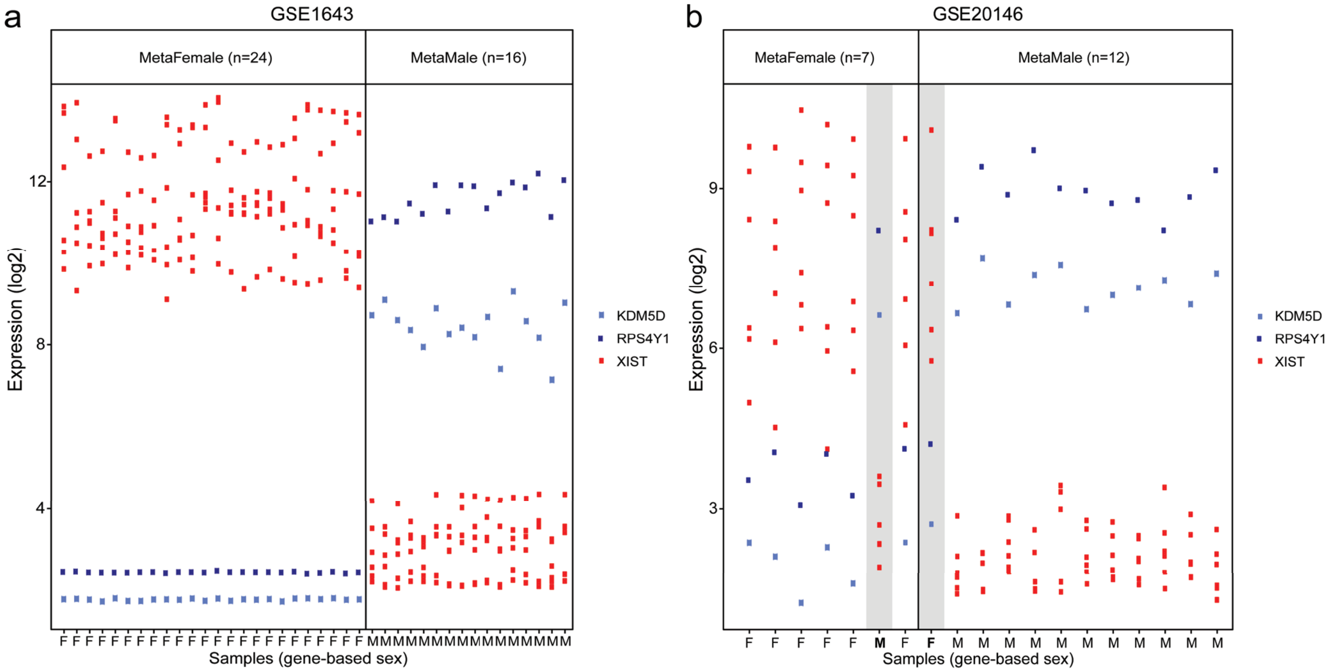

Figure 1. Representative plots showing expression levels of sex-specific probesets. Expression level of probesets representing the *XIST* (red), *KDM5D* (black) and *RPS4Y1* (blue) genes. "MetaFemale" and "MetaMale" indicate the meta-data annotated sex of the samples and their total number in brackets. The "M" and "F" along the X axis indicates the gene-based sex of the samples, as determined by k-means clustering. Log_2-transformed expression levels are plotted. (**a**) Representative dataset with no mismatched samples. (**b**) Representative dataset with two mismatched samples (highlighted with grey bars). Gene-based sex that contradicts the annotated sex of the sample is highlighted in bold at bottom.

samples (2.04% vs 1.97%; Table 1). This discrepancy might be explained by on average higher number of samples in cancer datasets from our corpus (Supplementary Table S1). As expected, the proportion of samples with ambiguous gene-based sex was much higher in cancer as compared to non-cancer samples: 23/1223 (1.97%) in cancer vs. 11/2937 (0.4%; Table 1). In total, 34 samples were flagged as ambiguous, though we note that 12/34 (35%) would have been signed to the discrepant sex by our method. Ambiguous samples were found in 15/70 (21%) of the studies (eight of which also contained mismatched samples).

Because the sex annotations we used to this point were obtained from the sample descriptions in GEO, there was a possibility that the discrepancies we identified were due to mistakes introduced during the communication of the data from the submitter to GEO. If this was the case, the results in the original publication (29/32 of the affected studies had an associated publication) would be unaffected, though users of the GEO data would still be affected. To check this possibility, we went back to the 29 original publications to see if the sex labels provided in the paper matched those in GEO (detailed in Supplementary Table S2). This check was not always possible because many publications did not provide detailed meta-data in the paper or Supplementary materials; GEO provides the only record. In 12/29 cases, sufficient detail was provided for us to confirm that the discrepant sex labels were present in the publication, and in all of them there was agreement between the meta-data in the publication and the meta-data in GEO. In 13 cases only summaries were given in the publication (*e.g.* "10 males and nine females in group X"). In 10 of these 13 studies, the summary counts in the publication agree with GEO. In the other three, both GEO and gene-based totals disagree with the publication-based totals. In other words, there seems to have been miscommunication with GEO in addition to a sex annotation discrepancy in the original study report. Finally, for four datasets meta-data was not provided or ambiguously described in the paper. We failed to find any unambiguous case in which we would infer the only problem was a miscommunication with GEO.

The analysis presented cannot distinguish between actual sample mix-ups (*e.g.*, tube swaps) and errors in the meta-data (incorrect recording of the subject's sex). Fortuitously, we identified data sets where it can be determined that at least in some cases, samples were probably physically mixed up. In addition to the 70 datasets used above, we analyzed four datasets that all used human brain RNA from the same collection of subjects (Stanley Medical Research Institute, Array Collection, https://www.stanleygenomics.org/stanley/). In this case the meta-data is common across the four laboratories since they are all analyzing the same individuals (though not all studies analyzed all the individuals). If the meta-data is incorrect, then all of the studies should show discrepancies for the same samples. If the samples were mixed up in a particular laboratory (or by the sample provider at the time they were sent to the laboratory), each study would have different discrepancies. We found that out of the four available datasets with data corresponding to the same subjects, two datasets contained mismatched samples (a single mismatched sample was identified in the "AltarA" study, and five in the "Dobrin" study; Figure 2). Importantly, the mismatched subjects differed between the datasets and samples from the same subjects appeared as correctly annotated in the other datasets. This suggests that the mismatched cases are likely to represent mislabeled samples rather than mistakes taking place during the recording of the subjects' sex.

We were surprised that nearly 50% of studies had at least one labeling error, and were concerned that this might be an overestimate by chance, due to sampling error. To address this we computed confidence intervals for our estimate of the fraction of affected studies, yielding a 95%-confident lower bound of 36% and a 99% lower bound of 33% (upper confidence bounds were 56% and 60% respectively). We also note that our independent observations of 2/4 datasets containing misannotations described in Santiago *et al.*[17] and in 2/4 of the Stanley data set are in agreement with a relatively high estimate. Thus we project that, with 99% certainty, if

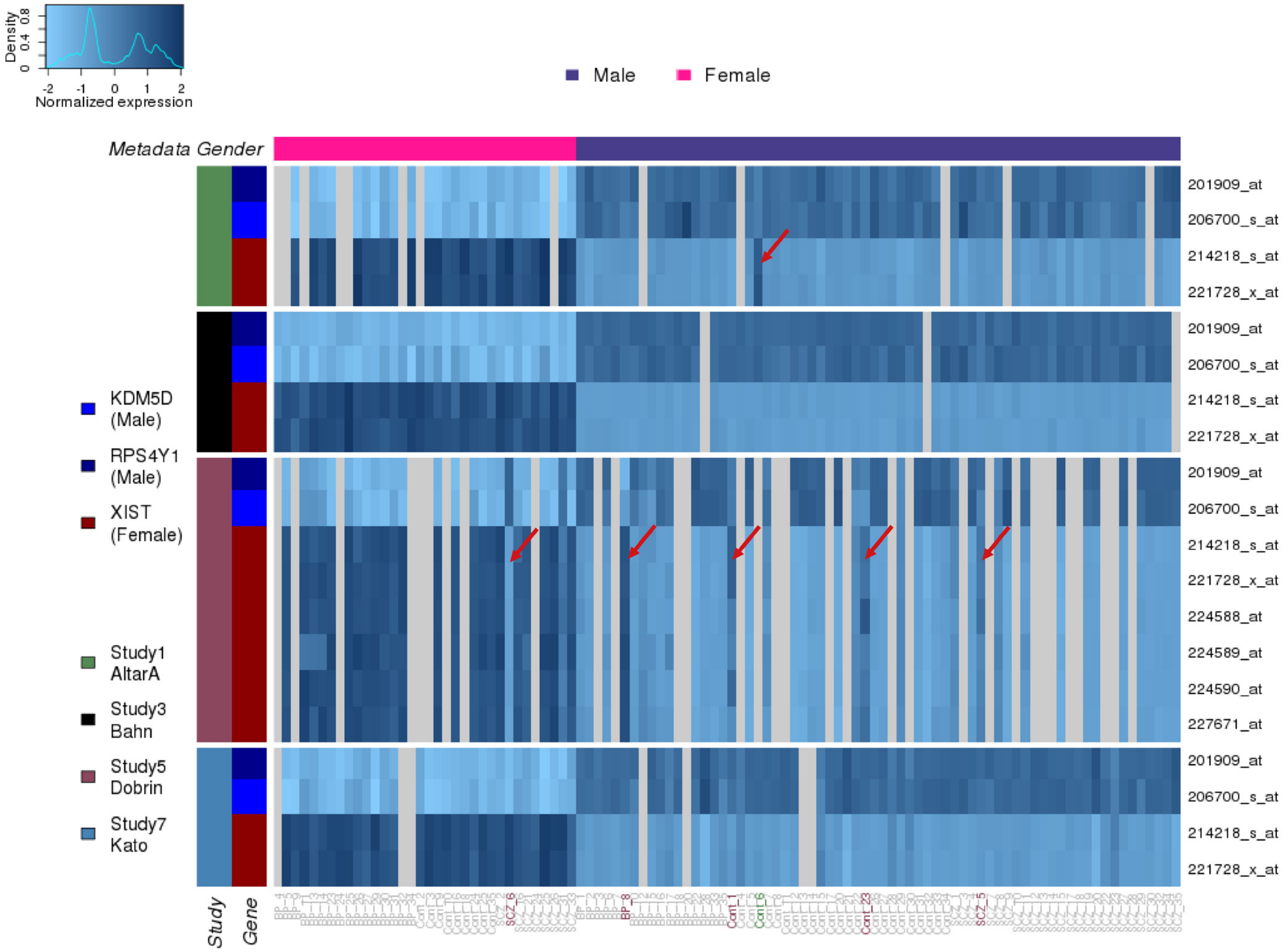

Figure 2. Gene-based and metadata-based sex in four datasets of similar subjects from Stanley Array collection. The heatmap represents z-transformed expression values of *KDM5D*, *RPS4Y1* and *XIST* probeset in four datasets of microarray data from Stanley Array Collection cohort of subjects. The datasets are designated - Study1 AltarA, Study3 Bahn, Study5 Dobrin, Study7 Kato, in correspondence to their names on the Stanley collection site. Each column represents a subject and each raw represents a probeset. The four studies are represented on the left color bar on the side of the heatmap. The gene names corresponding to each probeset are shown by the right color bar on the side of the heatmap. Three of the studies – AltarA, Bahn and Kato were performed on the GPL96 platform on which *XIST* is represented by two probesets. The Dobrin dataset is on the GPL570 platform containing additional 5 *XIST* probesets, one of which was removed from the analysis. The annotated sex of each subject (metadata gender) is represented by the top color bar (females – pink, males – purple). Missing samples (samples that were excluded from the original studies) are shown in grey. Arrows point to the mismatched samples.

all expression studies in GEO could be checked for mix-ups based on sex-specific genes, the fraction affected would be at least 33%.

Discussion

Using a simple approach to compare sample annotations for sex to expression patterns, we found that nearly 50% of datasets we checked contain at least one discrepancy. Our findings are also in general agreement with another study that examined this issue in a cohort of predominantly cancer datasets[18], although in cancer there is an expectation of more ambiguity of sex marker expression[16]. In the case of the Stanley brain datasets, we could determine that the problem is likely to stem from laboratory mix-ups rather than an error in recording the subject's sex. While our analysis is limited to a corpus of studies where sex information was available along with the presence of good markers on the microarray platform, our data suggest a widespread problem.

What is the impact of this issue? Viewed optimistically, a single mixed-up sample is not likely to dramatically affect the conclusions of a well-powered study. In addition, our analysis suggests a lower (99% confident) estimate of "only" 33% of studies with a sex mislabeling, which might provide a small amount of comfort to optimists – it could be worse. However, the sample mislabeling we identified might be the tip of the iceberg, because sex-specific genes can only reveal mixed-up samples with differing sex. We also suggest that sample mix-ups might correlate with other quality problems. Indeed, many of the misannotated datasets we found have additional issues such as undocumented batch effects, outlier samples, other apparent sample misannotations (not sex-related), and discordance in sample descriptions reported in different parts of the relevant publication (Supplementary Table S2). The presence of samples with ambiguous gene-based sex in non-cancer samples is suggestive of even more quality problems. This is because expression patterns of sex-specific genes could be treated as a positive quality control for the expression data as a whole, serving as indicators for the reliability of other gene signals. Deviations from the expected pattern might indicate samples were mixed together, or suggest problems with RNA quality.

Our conclusions are two-fold. First, there is an alarming degree of apparent mislabeling of samples in the transcriptomics literature. In at least the specific cases we identified, the trust in the reliability of the findings reported is certainly not improved. Second, because it is simple to check the expression patterns of sex markers, the tests we performed should become a routine part of all omics studies where sex can be inferred from the data.

Data and software availability

All data analyzed in this manuscript were previously published and can be accessed through the GEO repository (http://www.ncbi.nlm.nih.gov/gds) using accession numbers indicated in Table S1.

Stanley Medical Research Institute data can be accessed through https://www.stanleygenomics.org/stanley/.

Source code for the analysis is available in a Github repository (https://github.com/min110/mislabeled.samples.identification).

The archived version of the code at the time of publication can be accessed through Zenodo **mislabeled.samples.identification:** doi:10.5281/zenodo.60313[19]

Author contributions

LT and PP conceived the study idea. MF and LT performed the analyses, LT and PP prepared the manuscript. All the authors agreed to the final content of the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

The research was supported by grants to PP from NIH (GM076990) and the NeuroDevNet Network of Centres of Excellence, and the UBC Graduate program in Genome Sciences and Technology (MF).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We thank Patrick Tan for assistance with Gemma query and Ben Callaghan for assistance with the Stanley data sets.

Supplementary material

Table S1. Description of datasets included in the study.

Click here to access the data.

Table S2. Detailed description of all datasets with mislabeled samples.

Click here to access the data.

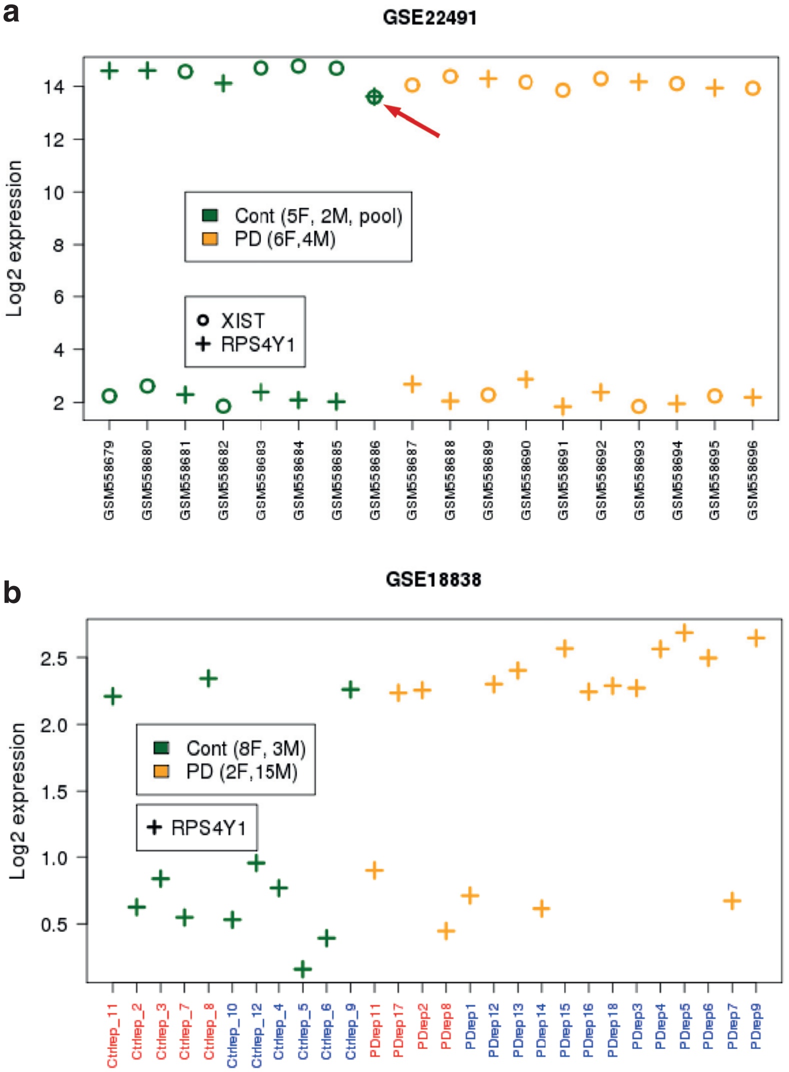

Figure S1. Disagreement between gene-based and annotated sex in three datasets participating in the metaanalysis of Parkinson's disease[1,2]. Santiago and Potashkin included four datasets in their metaanalysis. When available (three out of the four datasets) we used sample characteristics provided in the associated manuscripts to identify existence of mislabelled samples. Gene-based males are defined by high *RPS4Y1* and low *XIST* expression. Cont – control subjects (green), PD – Parkinson's disease (orange). In brackets, the corresponding number of females (F) and males (M) reported in the original manuscript. *XIST* and *RPS4Y1* genes were present in datasets GSE22491, but only *RPS4Y1* was present in GSE18838. (**a**) Based on the sex-genes expression, dataset GSE22491 contains at two 2 mislabelled samples. Of notice, in the pooled sample (indicated by an arrow) containing equal amount of males and females, the two genes are expressed at similar levels. (**b**) This is the only dataset for which sex of individual samples was available on GEO. Red – GEO annotated females, blue – GEO annotated males. Based on the manuscript's sample characteristics there should be 8F, 3M controls, and 2F, 15M PD. However, metadata provided on GEO, describes 5F, 6M controls, and 4F, 13M PD. Both of these annotations disagree with the gene-based sex of the samples (Cont – 8F, 3M, PD – 5F, 12M).

Figure S2. **Expression of probesets corresponding to the sex-specific genes *XIST, KDM5D* and *RPS4Y1* in datasets analyzed in the current study.** Each plot represents a separate dataset. The mismatched samples are highlighted in grey. For presentation interests, samples with undetermined gene-based sex were excluded. Each point represents a value of a single probeset in one sample. *XIST* – filled red circles, *KDM5D* filled black circles, *RPS4Y1* – open black circles. X axis shows the predicted gene-based sex of each sample.

Click here to access the data.

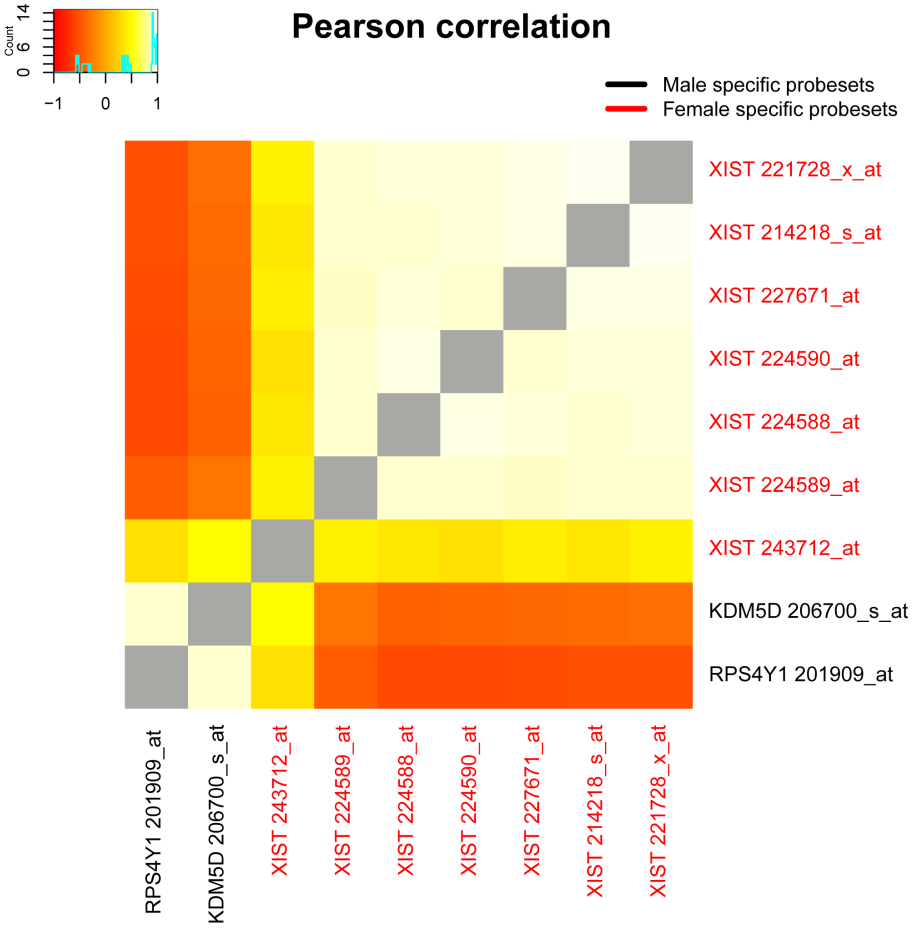

Figure S3. Correlation of probesets corresponding to sex-specific genes. Probeset-probeset Pearson correlation of probesets corresponding to *XIST*, *KDM5D* and *RPS4Y1* genes from all the datasets on in the current study. High correlation was observed between the two probesets corresponding to male specific genes. Six out of the seven *XIST* probesets showed very high positive correlation with each other, and high negative correlation with probesets corresponding to the male genes. Probeset 243712_at showed low positive correlation with other *XIST* probesets and relatively low negative correlation with male specific probesets and thus was excluded from the analysis.

Supplementary references

1. Santiago JA, Potashkin JA: Network-based metaanalysis identifies HNF4A and PTBP1 as longitudinally dynamic biomarkers for Parkinson's disease. *Proc Natl Acad Sci U S A.* **112**, 2257–2262 (2015).

2. Toker L, Pavlidis P: Metaanalysis of flawed expression profiling data leading to erroneous Parkinson's biomarker identification. *Proc Natl Acad Sci U S A.* **112**, E3637 (2015).

References

1. Allison DB, Brown AW, George BJ, *et al.*: **Reproducibility: A tragedy of errors.** *Nature.* 2016; **530**(7588): 27–29.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Begley CG, Ioannidis JP: **Reproducibility in science: improving the standard for basic and preclinical research.** *Circ Res.* 2015; **116**(1): 116–126.
PubMed Abstract | Publisher Full Text

3. Kauffmann A, Gentleman R, Huber W: **arrayQualityMetrics--a bioconductor package for quality assessment of microarray data.** *Bioinformatics.* 2009; **25**(3): 415–416.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Leek JT, Scharpf RB, Bravo HC, *et al.*: **Tackling the widespread and critical impact of batch effects in high-throughput data.** *Nat Rev Genet.* 2010; **11**(10): 733–739.
PubMed Abstract | Publisher Full Text | Free Full Text

5. SEQC/MAQC-III Consortium: **A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium.** *Nat Biotechnol.* 2014; **32**(9): 903–14.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Edgar R, Barrett T: **NCBI GEO standards and services for microarray data.** *Nat Biotechnol.* 2006; **24**(12): 1471–1472.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Toker L, Pavlidis P: **Metaanalysis of flawed expression profiling data leading to erroneous Parkinson's biomarker identification.** *Proc Natl Acad Sci U S A.* 2015; **112**(28): E3637.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Qu C, Schuetz JM, Min JE, *et al.*: **Cost-effective prediction of gender-labeling errors and estimation of gender-labeling error rates in candidate-gene association studies.** *Front Genet.* 2011; **2**: 31.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Tzvetkov MV, Meineke I, Sehrt D, *et al.*: **Amelogenin-based sex identification as a strategy to control the identity of DNA samples in genetic association studies.** *Pharmacogenomics.* 2010; **11**(3): 449–457.
PubMed Abstract | Publisher Full Text

10. Purcell S, Chang C: **PLINK 1.9.**
Reference Source

11. Sax L: **How common is intersex? a response to Anne Fausto-Sterling.** *J Sex Res.* 2002; **39**(3): 174–178.
PubMed Abstract | Publisher Full Text

12. Gentleman RC, Carey VJ, Bates DM, *et al.*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol.* 2004; **5**(10): R80.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Huber W, Carey VJ, Gentleman R, *et al.*: **Orchestrating high-throughput genomic analysis with Bioconductor.** *Nat Methods.* 2015; **12**(2): 115–121.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Zoubarev A, Hamer KM, Keshav KD, *et al.*: **Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data.** *Bioinformatics.* 2012; **28**(17): 2272–2273.
PubMed Abstract | Publisher Full Text | Free Full Text

15. R Core Team: **R: The R Project for Statistical Computing.** In: *The R Project for Statistical Computing.* 2015; [cited 5 Feb 2016].
Reference Source

16. Weakley SM, Wang H, Yao Q, *et al.*: **Expression and function of a large non-coding RNA gene XIST in human cancer.** *World J Surg.* 2011; **35**(8): 1751–1756.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Santiago JA, Potashkin JA: **Network-based metaanalysis identifies HNF4A and PTBP1 as longitudinally dynamic biomarkers for Parkinson's disease.** *Proc Natl Acad Sci U S A.* 2015; **112**(7): 2257–2262.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Lohr M, Hellwig B, Edlund K, *et al.*: **Identification of sample annotation errors in gene expression datasets.** *Arch Toxicol.* 2015; **89**(12): 2265–72.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Toker L: **mislabeled.samples.identification.** *Zenodo.* 2016.
Data Source

Open Peer Review

Current Referee Status: ✓ ✓ ?

Version 2

Referee Report 03 October 2016

doi:10.5256/f1000research.10437.r16396

?

Levi Waldron
School of Public Health, City University of New York, New York, NY, USA

The article is clear and to the point, and the abstract provides an adequate summary of the article, the methods are adequately described, conclusions are balanced and justified, and data are available in a usable format. I do have a few comments though:

1. I have a concern that the use of kmeans clustering assigns individuals to one of the two clusters without any estimate of posterior probability of the assignment. Although the examples shown in Figure 1 imply there is little uncertainty in these assignments, it is unclear whether that is always the case, or whether there are some more ambiguous cases. It would be much more reassuring to use a clustering algorithm that allows estimation of posterior probability of cluster assignment (such as K-nearest neighbors in the "classify" package). Then I would immediately ask, are the biological / annotation mismatches all called with high confidence (say, >99%), or are some ambiguous (e.g. 49%)?

2. The article could have greater impact if the authors provided a ready-to-use tool for others to do the same check. It doesn't have to be fancy; it could be provided even just by providing .RData objects for the fitted classifer object used for the clustering.

3. The "How to replicate my analysis" section on https://github.com/min110/mislabeled.samples.identification README.md is blank, and should be filled in.

4. PURELY OPTIONAL: a paper of mine also points out the frequency of sample duplication, including an instance of several dozen RNA aliquot mix-ups in TCGA [1], should you find it relevant to cite.

References

1. Waldron L, Riester M, Ramos M, Parmigiani G, Birrer M: The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles.*J Natl Cancer Inst*. 2016; **108** (11). PubMed Abstract | Publisher Full Text

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 22 September 2016

doi:10.5256/f1000research.10200.r16144

☑ **Hans van Bokhoven**
Department of Human Genetics & Department of Cognitive Neuroscinders Institute for Brain, Cognition and Behaviour ences, Radboud University Medical Center , Nijmegen, Netherlands

This article by Toker *et al.* reports a retrospective biostatistical analysis of the data reported in 70 RNA expression studies to identify the possible misannotation of samples used in these studies. Their analysis is based on the expression of gender-specific genes (XIST for females; KDM5D and RPS4Y1 for males). Their analysis revealed apparent mismatches between the expression data and the annotated gender for 83 of 4160 samples (2%), encompassing 32 of the 70 studies (46%). This percentage is consistent with a those of a previous analysis in cancer datasets (PUBMED: 26608184). While these figures are already alarming, the actual number of mismatches is likely to be higher, because the gender-analysis can only identify discrepancies based on a gender-mismatch and will not detect mislabelling of samples of the same gender and case-control samples.

For most cases, the reason for the mismatches is not clear, but comparison of identical sample-data presented in different publications revealed that sample mix-ups are likely to be involved.

The mislabelling of samples in transcriptomics studies have an immediate impact on the involved studies, which often only have a modest sample size. In addition, also follow-up studies based on the results reported in such studies can suffer from it. Therefore, the use of controls to check the identity of samples is warranted. The gender-test presented in this work is a simple test that should become routine in expression studies, another option is to use nucleic acid-based bar codes that can be added to the sample early in the processing.

Some minor comments:

- On page 4, second-last line: here 29/31 of the affected studies is mentioned. Shouldn't that be 32? I hope I missed it, but certainly in a report as this one, the figures should be absolutely correct.
- Some typo's in the legend: Figure 2, Stady 5 is Study 5; Supplementary Fig S1: Based on the sex-genes expression, dataset GSE22491 contains at two 2 mislabelled samples.

References

1. Lohr M, Hellwig B, Edlund K, Mattsson JS, Botling J, Schmidt M, Hengstler JG, Micke P, Rahnenführer J: Identification of sample annotation errors in gene expression datasets.*Arch Toxicol*. 2015; **89** (12): 2265-72 PubMed Abstract | Publisher Full Text

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 13 September 2016

doi:10.5256/f1000research.10200.r15949

☑ **Leonard P. Freedman**
Global Biological Standards Institute, Washington, DC, USA

This is an excellent paper highlighting the importance of sample annotation as a critical contributor to reproducible research. Using transcriptomics of sex-specific gene expression levels as an example, the authors do a careful analysis to illustrate the issue of mislabeling. My one concern, which they candidly acknowledge, is that nearly half of the samples appear to have only one error, which may lead many readers to conclude that this actually a quite respectable error rate and that thus this is not really a big problem. It was not entirely clear to me why, as the authors argue, that this is in fact 'the tip of the iceberg".

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.