Supplement B to Oron et al.: Statistical Evidence for Sex Data-Entry Errors in ALL Dataset

July 29, 2008

Framework

The ALL dataset's Y chromosome expression patterns (see Fig. 3 in main article) compel us to face the question whether there has been a data-entry error with respect to certain subjects' sex. This type of question calls for a more structured decision-making process, using statistical inference. Below are two approaches to test

$$H_0 : \text{All sex entries are correct}$$

vs.

$$H_a : \text{Some sex entries are incorrect.}$$

It is intuitively clear that the tests to use are outlier significance tests of some form. Data patterns indicate that there may be more than one *bona fide* outlier; in fact, the article's Fig. 3 shows five samples ostensibly behaving "against type". This means that outliers might mask (downplay) each other's effect, or vice versa. Therefore, we should proceed via **successive outlier elimination**, stopping when no more outliers pass the significance test. In our case, "elimination" means re-labeling the sex entry and re-calculating the relevant statistics.

We observed that out of 11 non-autosomal Y genes, 10 exhibit a strong sex effect, even with the outliers masking it: per-gene t-statistics for the sex effect (whether or not adjusted for phenotype) are upwards of 4.5 for all ten. The last gene, however (symbol NLGN4Y), hardly shows an effect: log fold-change of 0.03 and t-statistic of 0.4. Whether this is due to an undocumented autosomal copy on the X chromosome, or due to non-expression (raw expression levels are low for this gene), is immaterial for our purposes: since it fails to produce a sex signal, it was excluded from the outlier analysis. We are therefore left with 10 genes.

Another factor masking outlier detection is imperfect normalization (see Fig. 1 in main article). Therefore, we have used the re-normalized version of the data with each sample's median removed.

Outlier-Detection GS Residual Test (Likelihood Ratio Test)

First we note that phenotype is not evenly distributed between males and females. Since the sex effect strongly dominates any other effect for Y-chromosome expression, the GS residual boxplots in the article's Fig. 3 may include some small bias. Therefore we recalculate the GS residuals, this time from an intercept-only model. This is equivalent to removing per-gene means, then standardizing by the per-gene variance. We have 78 relevant, standardized GS residuals (one sample has a missing sex entry), which we can assume are roughly identically distributed. We regress them on sex (as originally labeled), and look at the externally-Studentized residuals from this second regression. As stated in the article's "Methods" section, according to linear-model theory these residuals are t-distributed, and can be interpreted as outlier hypothesis tests with H_a being that the tested point has a different mean than predicted by the model. The degrees of freedom are 74: we lost one d.f. each to the missing-data sample, to the first-stage mean calculations, to the second-stage intercept and sex effect, and to excluding the tested point. Our test will be one-sided, since the sex-assignment null hypothesis can only be rejected for residuals deviating towards the opposite sex's mean; the sign of the rejection region will be negative for males and positive for females.[1]

Care should be taken in setting the critical value. Our search for outliers is not determined *a priori*: any sample behaving like the opposite sex may be suspect. Therefore we must control for multiple testing, or the familywise error rate (FWER). The actual critical value can be determined via

$$\tilde{\alpha} \equiv 1 - (1 - \alpha)^{1/n},$$

which for $\alpha = 0.01$ and $n = 78$ yields $\tilde{\alpha} = 1.3E - 4$ and $|t_{74,\tilde{\alpha}}| = 3.84$. This ensures that the probability of observing at random one or more outliers beyond the critical t value is 0.01. The strongest outlying residuals – belonging to samples `04016, 36002` labeled as males – have t-statistics of -4.21 and -4.01, respectively, passing the 99% significance threshold. The next-largest outlier – sample `22010` labeled as female – has a t-statistic of 2.68 which fails to clear even 95%. However, it is expected to increase after relabeling the first pair and recalculating the regression and its residuals. Additionally, we allow ourselves to label the one missing entry as male based on its Y expression patterns which place it squarely in the male column. The critical-value calculations need to be modified slightly, but the resulting thresholds change very little.

Sample `22010`'s t-statistic is now 3.53 – short of clearing the bar for $\alpha = 0.01$. If we are willing to risk $\alpha = 0.05$, the cutoff becomes 3.34 and this sample clears it. One has to keep in mind, though, that p-values are somewhat compromised by the previous elimination step: they ignore the probability of observing one of the first two outliers randomly (in which case they should still be assigned as males in the current calculations). By the way, the Studentized residual of the missing-data sample (now labeled male) is 0.63.

Suppose we are willing to increase our risk, relabel `22010` and recalculate again. The most egregious outliers now are two females – `24010, 26003` – which are also visible on the article's Fig. 3. Their t-statistics are 3 and 1.83, respectively, failing to clear the threshold even for $\alpha = 0.1$, and even when neglecting the effect of the first three re-assignments. The relabeling process stops.

[1] Note: the residual test is equivalent to a likelihood-ratio test between assigning each sample as male or female – which is, according to theory, the most powerful test using the GS residual data.

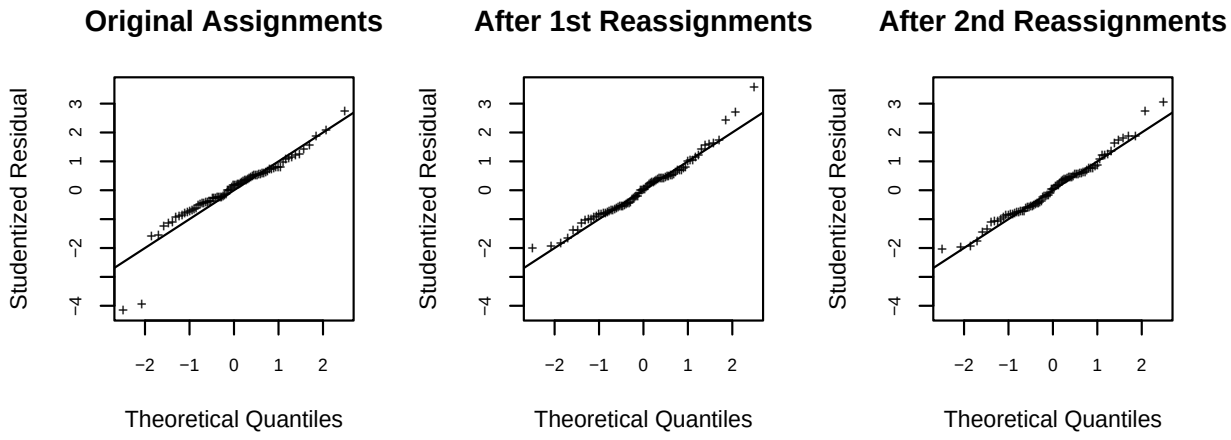

Figure B-1: Cumulative normal plots of Studentized second-stage residuals from all samples, after modeling the effect of sex. The model was run on intercept-only GS residuals aggregated over 10 non-autosomal Y-chromosome genes. From left to right: original sex assignments, after relabeling a pair of males as females, after relabeling two pairs (one male, one female). The diagonal lines represent the standard Normal distribution.

NOTES: With single-stage analysis using robust regression (similar to the DNA extra copy analysis in the main article), the results are nearly identical to those described here. Using the successive-outlier-deletion procedure shown here, but without renormalizing each sample's expression by its median, finds four samples (the above three, plus 26003) to be misassigned at $\alpha = 0.01$, when testing against the occurrence of outliers in closely-matched pairs (as single outliers none of them clears the bar on its own). The fourth sample has a relatively high expression baseline (its median residual from the null model is 9th largest). We also explored the problem using the false-discovery-rate (FDR) threshold approach, with similar results to those reported above.

Per-Gene Proximity Beta-Binomial Test

The previous test uses two successive regressions, and does not directly rely upon individual-gene outcomes. A more direct approach might be to calculate sex-group means for each gene, and determine – for each gene and sample – whether its expression is closer to its own group mean, or to the opposite-sex group mean. We use the same 10 Y chromosome genes as above. We expect true males and females to be closer to their group means for the lion's share of the 10. Incorrectly labeled samples would exhibit mirror-image behavior.

Table B-1 presents the 10-gene summaries for each sex separately. The same three outliers flagged by the regression procedure are evident, with only 0 to 3 "correct" genes. The fourth and

Number of Genes Closer to "Correct" Group Center	0	1	2	3	4	5	6	7	8	9	10
Female	0	0	1	0	0	2	0	3	2	6	14
Male	1	1	0	0	0	0	4	6	5	21	12

Table B-1: Table of the number of non-autosomal Y genes, for which a given sample is closer in its expression level to its sex-group mean than to the opposite-sex mean. Calculated over the 10 genes used above.

fifth samples are again much less egregious. Under inter-gene independence, the counts for each sample should be binomially distributed, with the probability of "success" (i.e., correct-sex mean closer) estimated from the data. A quick comparison between Table 1's counts and binomial probabilities reveals that the data are significantly more disperse – even when ignoring the five strongest outliers. A beta-binomial model with a sex covariate (since males appear to have lower counts, on the average), fits the data reasonably well when outliers are excluded from the calculations.

The three outliers identified in the previous section easily pass the adjusted $\alpha = 0.01$ p-value threshold using beta-binomial probabilities. After relabeling these three and adding the missing-entry sample, we recalculate distances, producing Table B-2.

Number of Genes Closer to "Correct" Group Center	5	6	7	8	9	10
Female	2	0	3	2	6	16
Male	0	4	6	6	21	13

Table B-2: Table of the number of non-autosomal Y genes, for which a given sample is closer in its expression level to its sex-group mean than to the opposite-sex mean, following reassignment of 3 samples' sex and assignment of a missing-entry sample as male.

Samples `24010, 26003` still show a 5-5 split. According to the beta-binomial model, this split has an unadjusted p-value of 0.0035. This again fails to clear the threshold for familywise-adjusted inference, even for $\alpha = 0.1$ and even after removing the first 3 samples from consideration in the multiple-testing adjustment. Repeating the beta-binomial procedure using FDR yielded the same results.

We therefore decided to change the sex labels of three samples only: `04016, 36002` and `22010`.