Supplement D to Oron Et. Al.: Using Residuals to find DNA Copy Number Irregularities with a verified example from St. Jude's Pediatric ALL Dataset

July 29, 2008

Outlier Detection via Residuals

Supplement B demonstrated several ways to detect outliers in the context of sex assignment errors in the dataset. The problem there was somewhat easier, in that the subject population is naturally divided into two distinct groups, which can be translated into two clear hypotheses regarding chromosome Y expression.

Looking for aneuploidy signals in the dataset is more akin to a general outlier search. It is known that extra DNA copies do lead in general to higher expression levels, however the relationship is weaker than linear and quite noisy (Pollack *et al.*, 2002). Moreover, aneuploidy is intrinsically much more heterogeneous, at least in the following ways:

- In the same dataset there may be various multisomies of a given chromosomes – trisomies, tetrasomies, etc.;
- In a given tissue sample, some of the material may exhibit multisomies and some not;
- Some extra DNA copies may be biochemically "silenced", and hence not detectable via gene expression;
- Some samples may exhibit partial additions or deletions for the same chromosome.

Outlier detection is a *relative* affair: it identifies data points diverging from the main group. As the group of outliers becomes "large" (approaching one quarter of the sample size, if they are all on the same side compared with the main group), the question becomes ill-defined and detection sensitivity will typically be the first to suffer, because weaker outliers will masked by the stronger ones.

Wisnowski *et al.* (2001) numerically examined several common outlier-detection methods in the regression context, and with outlier frequencies of up to 20%. Fortunately, a robust regression residuals approach – which is intuitively simple and directly related to our article – performs quite well in scenarios similar to those encountered in microarray analysis, specifically the DNA copy

number question. Robust regression is an extension of robust location estimation. Most robust estimation methods operate similarly; we have chosen Huber (1981)'s M-estimator which is among the simplest. The algorithm begins from an initial location and scale estimates (usually the median and IQR, respectively). Then, weights are assigned to data points according to their standardized distance from the center, and then location and scale estimates are re-calculated using an weighted averaging formula. The process repeats iteratively until the estimates are stable. We chose the simplest weighting scheme of the M-estimator, known as 'Winsorization': each data point at a standardized distance exceeding some cutoff c is replaced by c in the averaging. This is equivalent to an inverse-weighting scheme for points beyond c scale units away from the center. In robust regression, location estimates are replaced by regression estimates and distances are replaced by residuals.

There exist more aggressive weighting schemes and sophisticated algorithms, such as Yohai (1987)'s MM-estimator which identifies outliers more aggressively – with the almost inevitable side-effect of more false positives. We found the relatively conservative M-estimator to be a reasonably good match for the noisy data environment of gene expression arrays. A downside to using robust methods, is that there is no theoretical reference distribution akin to the t of the classical approach. The solution is to generate a numerical reference distribution, using a large number of repetitions of the same-size dataset and with the same covariate structure – but with no true outliers (i.e., data points consist of the covariate effect plus t-distributed noise). The percentiles on the tail of this numerical null will be used as thresholds to detect outliers at specified significance levels (Wisnowski *et al.*, 2001). Both M and MM regression estimators are available via function `rlm` of the MASS package; for MM, one needs to add the argument `"method=MM"`. If the model is intercept-only, then M estimation can be performed faster using `hubers` from the same package. The default value of c is 1.5. Note that at bottom line, applying robust estimation to the residuals from an intercept-only model (as we are doing here) leads simply to their linear rescaling according to the new location and scale estimates.

The robust method detects outliers relative to the bulk of samples. However, Some chromosomes may exhibit low expression variability across samples – and then outliers may be flagged for relatively small deviations. Just like Hertzberg *et al.* (2007), and given the physical nature of aneuploidy (e.g., 1 or 3 or 4 chromosomes instead of 2, or 2/3/4 instead of 1 for males and the X chromosome), we have found it necessary to impose a second criteria of a minimum fold-change magnitude required for the sample to be considered a candidate for aneuploidy. After consulting literature (Pollack *et al.*, 2002) and some minimal trial-and-error, we set it at a $1:6$ or greater discrepancy, symmetric on the log level, i.e., a sample if flagged only if its average expression level for the chromosome in question is $< 6/7$ or $> 7/6$ of the dataset median.

To recap, the workflow of our residuals-based DNA suspect-aneuploidy detection method is:

1. Generate per-gene Studentized residuals from the relevant linear model; in the absence of a clear model to adjust for, use a null (intercept-only) model
2. Produce GS residuals for each chromosome using the rescaled-mean formula given in the article
3. Rescale the residuals for each chromosome separately, using the robust M-estimator for location and scale

4. Compare the residuals to a numerically-generated reference distribution of outlier-free data that has undergone a similar robust estimation procedure, and using the appropriate FDR thresholds

5. Further filter out from the suspect-aneuploidy list any sample whose mean expression for the chromosome was between 6/7 and 7/6 of the dataset's median

Verification with St. Jude Ross *et al.* (2003) Dataset for Chromosome 21

Following the referees' suggestion, we applied this workflow, with no further adaptation (and using an intercept-only model), to a dataset for which such information exists at least for chromosome 21 – the Ross *et al.* (2003) pediatric ALL dataset from St. Jude's hospital. A preprocessed and summarized version of the dataset was kindly provided by Claudio Lottaz of the University of Regensburg. Incidentally, this was the same dataset on which Hertzberg *et al.* (2007)'s method was optimized. This dataset has 132 samples, 20 of which were biochemically identified as having over 50 chromosomes, 2 as having 47-50 chromosomes, and 4 as hypodiploid.

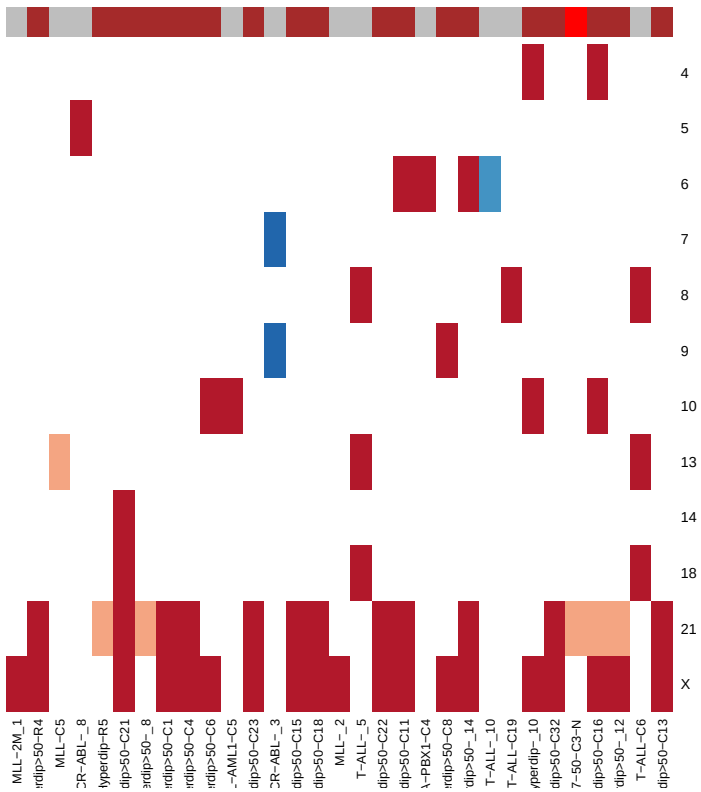

Figure D-1: Suspected aneuploidies among the 132 samples of the (Ross *et al.*, 2003) pediatric ALL dataset. Color keys and statistical methods are identical to those of the main article's Fig. 5.

Fig. D-1 displays a suspect-aneuploidy map analogous to the main article's Fig. 5. The similarities are striking, including but not limited to:

- The most commonly encountered aneuploidies are on chromosomes X and 21, in that order.
- Most hyperdiploid samples (top bar, brown for over 50 chromosomes and red for 47-50) and a much smaller fraction of the rest of the dataset are flagged.
- There are a couple of chromosome-8 suspect multisomies, but interestingly these are indicated only for non-hyperdiploid samples and are usually not observed jointly with chromosome 21 or X extra copies.
- The chromosomes most strongly indicated for missing copies are 7 and 9.

		Hertzberg *et al.* (2007)			
		True Positives	False Positives	True Negatives	False Negatives
Residuals Method (this article)	True Positives	17	0	0	0
	False Positives	0	0	0	0
	True Negatives	0	3	101	0
	False Negatives	7	0	0	4

Table D-1: Performance comparison summary of true chromosome 21 aneuploidy detection in the Ross *et al.* (2003) dataset, by the Hertzberg *et al.* (2007) and residual methods.

Given that the datasets are related to the same disease, these similarities increase our confidence that this method captures quite a few true signals.

Table D-1 presents a performance comparison summary between the two data-mining methods on chromosome 21, for which there was laboratory verification of aneuploidies using two methods (cytogenetics and FISH). Table D-2 displays a sample-by-sample list of all verified, complete extra copies of chromosome 21 in the dataset, and whether they were detected by either data-mining method. As Table D-1 indicates, the performance differences between the methods boil down mostly to **threshold placement**: Hertzberg *et al.* (2007) detects all 17 true positives found by the residuals method, and adds 10 more positives – 7 true and 3 false. Moreover, a detailed look (Table D-2) shows that the two methods are perfectly matched on the multiple-extra-copy samples.[1] Another secondary observation is that the two methods markedly differ on 4 `TEL-AML1` samples, a genetic-material rearrangement involving chromosome 21: Hertzberg *et al.* (2007) detect 3 of 4 while the residuals method detects none (we have no information regarding whether any of the former's 3 false-positives also belong to this group). We repeated the residuals approach on this dataset using MM-estimation, with identical results.

[1]Recalling that both methods are expression-driven, we may venture to guess that the 3 multiple-extra-copy samples missed by both may have the extra copies largely silenced. Moreover, these are also the only multiple-extra-copy samples classified by Ross *et al.* (2003) to groups other than the hyperdiploid group, using expression-based clustering.

As stated previously, the residuals methods has not been as thoroughly optimized, and offering a competitor to Hertzberg *et al.* (2007) was not our goal here. However, the similarity in performance is encouraging, given that the two methods probably diverge on data-analysis details at all stages from preprocessing onwards. The residuals method can be automated as a convenient starting point, to get a rough idea regarding potential aneuploidies in a dataset – as long as the bulk of the samples are known to be diploid.

References

Hertzberg, L. *et al.* (2007). Prediction of chromosomal aneuploidy from gene expression data. *Genes Chromosome Cancer*, **46**(1), 75–86.

Huber, P. J. (1981). *Robust statistics*. John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics.

Pollack, J. *et al.* (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Nat. Acad. Sci.*, **99**(20), 12963–12968.

Ross, M. *et al.* (2003). Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, **102**, 2951–2959.

Wisnowski, J. *et al.* (2001). A comparative analysis of multiple outlier detection procedures in the linear regression model. *Comp. Stat. Data Analys.*, **36**(3), 351–382.

Yohai, V. (1987). High breakdown point and high-efficiency robust estimates for regression. **15**(2), 642–656.

Sample ID	Extra Copies	Partial?	Hertzberg et al.	Residuals
HDover50-C13	3		Y	Y
HDover50-12	2		Y	Y
HDover50-14	2		Y	Y
HDover50-8	2		Y	Y
HDover50-C11	2		Y	Y
HDover50-C16	2	x	Y	Y
HDover50-C18	2		Y	Y
HDover50-C21	2		Y	Y
HDover50-C23	2		Y	Y
HDover50-C27-N	2		N	N
HDover50-C32	2		Y	Y
HDover50-R4	2		Y	Y
HDover50-C1	2		Y	Y
HDover50-C4	2	x	Y	Y
Hyperdip47-50-C14-N	2		N	N
TEL-AML1-C38	2		N	N
BCR-ABL-Hyperdip-10	1		Y	N
BCR-ABL-Hyperdip-R5	1		Y	Y
HDover50-C15	1		Y	Y
HDover50-C22	1		Y	Y
HDover50-C6	1		Y	N
HDover50-C8	1		Y	N
Hyperdip47-50-C3-N	1		Y	Y
Pseudodip-6	1		Y	N
TEL-AML1-C5	1		Y	N
T-ALL-C7	1	x	N	N
TEL-AML1-9	1	x	Y	N
TEL-AML1-2M2	1	x	Y	N

Table D-2: Detailed summary of true chromosome 21 aneuploidies by sample in the Ross *et al.* (2003) dataset, and their identification by the Hertzberg *et al.* (2007) and residual methods (two rightmost columns). The 'Partial?' column is marked with an 'x' for those samples for which the aneuploidy signal was not found by both experimental methods, or was reported to exist for less then 60% of the tissue sample.