

Supplementary Materials for *False Discovery Rates and Copy Number Variation*

Bradley Efron Nancy Zhang

Department of Statistics
Stanford University
390 Serra Mall Stanford, CA 94305 U.S.A.

1 Data Preprocessing.

The matrix $\{x_{ij}\}$ represents normalized intensities obtained from microarray experiments. Different microarray platforms require different, sometimes complicated, preprocessing procedures. See Pinkel et al. (1998) Pollack et al. (1999), Snijders et al. (2001), Bignell et al. (2004), and Peiffer et al. (2006) for details on the data from individual platforms. Generally, for two-color arrays, the $\{x_{ij}\}$'s are the log ratios of the intensities for the two colors on each probe. For genotyping microarrays such as Affymetrix and Illumina, they are the log of the $A + B$ intensities, normalized to a population reference.

For data from any platform, it is essential to remove the local trends described in Olshen et al. (2004), Diskin et al. (2008), and Marioni et al. (2007). These trends are artifacts that, if present in the data, could swamp out the signal. They can be partially removed by a regression model on the GC content (Diskin et al., 2008) or by removing a low-rank component from the data, as described in Siegmund, Yakir and Zhang (2010). The latter method is similar to the approaches of Leek and Storey (2008) and Friguet, Kloareg and Causeur (2009) for treating latent factors in simultaneous hypothesis testing.

The data shown in Figure 1 of the manuscript is part of a larger data set assayed on the Affymetrix Genomewide SNP Array 6.0. The raw copy numbers were first obtained using the software `aroma.affymetrix` (Bengtsson, Irizarry, Carvalho and Speed, 2008). The glioblastoma data in Section 7 of the manuscript are simply the log R (log ratio of total copy number to reference) quantities from the Illumina HumanHap 550k platform. The method from Siegmund et al. (2010) was applied to both data sets to remove local trends.

2 Choice of the moving average window width

Here we explore in more detail the choice of the moving average window width M . The adjusted moving average from Equation (2) of the manuscript,

$$z_{ij} = \sqrt{M}\bar{x}_{ij} \tag{2.1}$$

has distributions

$$\text{null } z_{ij} \sim \mathcal{N}(0, 1) \quad \text{non-null } z_{ij} \sim \mathcal{N}(\mu_M, 1) \tag{2.2}$$

where, if the moving average is taken entirely within a contiguous non-null block, we have

$$\mu_M = \sqrt{M}\mu. \quad (2.3)$$

Averaging increases the null/non-null separation in this case, improving the power of our detection procedure, as made explicit next.

The ratio of true to false discovery rates in position i is

$$\frac{\text{tdr}_i(z)}{\text{fdr}_i(z)} = \frac{\Pr\{\text{non-null}|z, \mathcal{C}_i\}}{\Pr\{\text{null}|z, \mathcal{C}_i\}} = \frac{\pi_{i1}f_1(z)}{\pi_{i0}f_0(z)}, \quad (2.4)$$

in the notation of Section 2 of the manuscript. Then (2.2) yields

$$\frac{\text{tdr}_i(z)}{\text{fdr}_i(z)} = \frac{\pi_{i1}}{\pi_{i0}} e^{\mu_M(z-\mu_M/2)}. \quad (2.5)$$

Under (2.3),

$$\frac{\text{tdr}_i(z)}{\text{fdr}_i(z)} = \frac{\pi_{i1}}{\pi_{i0}} e^{M\mu^2/2} \quad (2.6)$$

at $z = \mu_M$, its non-null expected value, so increasing the window width M raises the ratio exponentially fast. Put simply, large M produces non-null z -values far from 0, at least at positions inside long non-null blocks.

Suppose though that the non-null block length M_{non} is less than M . The same reasoning as in (2.6) gives

$$\frac{\text{tdr}_i(\mu_M)}{\text{fdr}_i(\mu_M)} = \frac{\pi_{i1}}{\pi_{i0}} e^{(M_{\text{non}}^2/M)\mu^2/2} \quad (2.7)$$

at the block's central position, so that increasing M is now harmful. The ideal choice is $M = M_{\text{non}}$, the well-known *signal matching criterion*, but of course in practice we won't know M_{non} .

Other considerations come into play: larger M improves the normality of z_{ij} , null normality being an important assumption in (2.5); correlation between nearby x_{ij} 's decreases the advantage of averaging; long non-null intervals like those seen near $i = 3800$ in Figure 3 (main manuscript) may include sub-blocks of negative as well as positive cnv effect. For more on this point, see the discussion on one-sided procedures in Section 5 of the manuscript.

3 A closer look at the 8500-9200 region of TCGA Glioblastoma Chromosome 1

Figure 1 shows a heatmap of the original $\{x_{ij}\}$ matrix of the 700 marker region containing the recurrent deletions covering genes FAF1 and CDKN2C. The color scheme of the heatmap has the baseline value of 0 as green, with losses visible as horizontal streaks of blue. Dark blue streaks, such as the one for sample 77 from around 8700 to around 8900, are evidence for loss of both copies of the chromosome (homozygous deletions). Less conspicuous light blue streaks are evidence for loss of one of the two chromosomal copies (hemizygous deletions). The dash-plot below the heatmap shows those signals that fall below an FDR threshold of 0.05. All of the homozygous deletions seem to be captured at this threshold. Some of the hemizygous deletions fall above this threshold and are not shown in the dash plot, but they contribute to \hat{k}_i , which takes on a maximum value of 12 in this region.

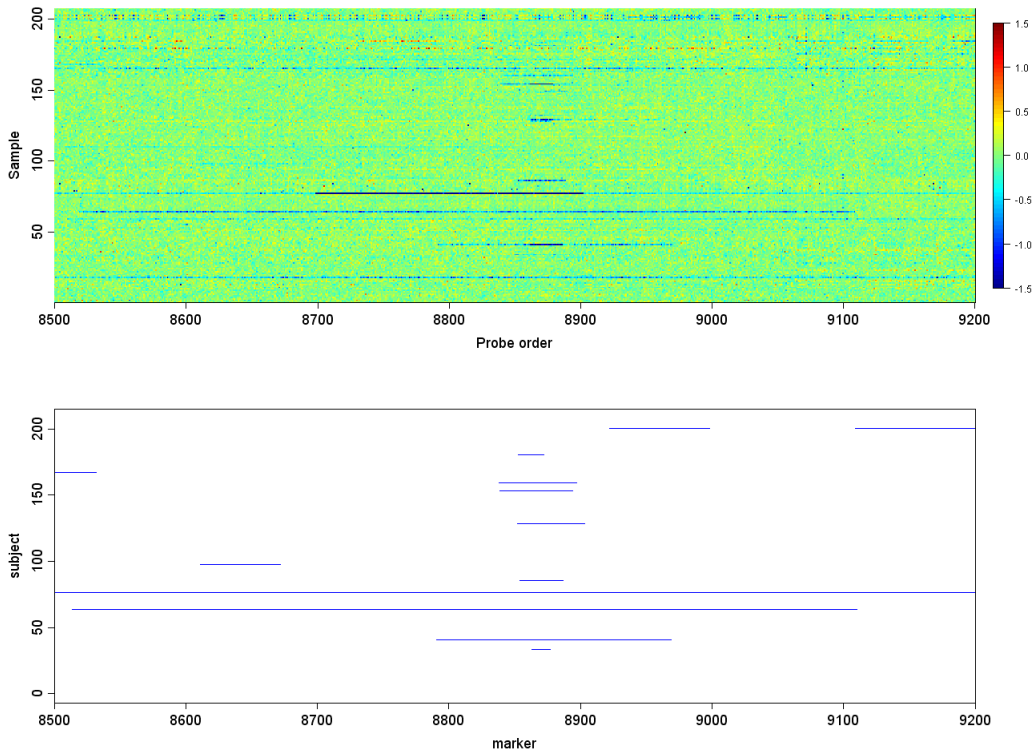


Figure 1: The 8500-9200 region of chromosome 1 of the TCGA glioblastoma samples. At the top is a heatmap of the original $\{x_{ij}\}$ matrix, followed by a dash-plot showing the candidate gains and losses by a 0.05 threshold on the local false discovery rate.

4 Local Tests for CNV

Once a genomic region has been identified to contain a candidate cnv by the *fdr*-based method, other methods can be used to extract more detailed information. At this stage, many questions remain: When multiple nearby locations within a sample fall below the *fdr* threshold, do they belong to a single contiguous stretch of cnv? If they do, can we accurately estimate the locations of the break-points? For a cnv-prone location, identified by a high \hat{k}_i value, can the carriers be identified more accurately than by thresholding the local *fdr*? Also, the set of candidates reported by the *fdr*-based method would no doubt contain false positives. Could we achieve better accuracy by a more detailed follow-up analysis, that examines each candidate cnv and tosses out those that look like imposters? In this section, we seek solutions to these remaining problems. Since our analysis is limited only to those genomic regions that are labeled as “interesting” by the *fdr* method, we will call the ensuing analysis “local”, as opposed to a “global” analysis covering the entire data matrix.

In the local analysis, we return to the original $\{x_{ij}\}$ matrix of normalized intensity values. The *fdr* method described in the previous sections works off the matrix of *z*-values, which were obtained from the normalized data through a smoothing step that averages adjacent probes. The original *x* matrix, if properly normalized, contains entries that are approximately i.i.d. standard normal under the null hypothesis. An effective normalization procedure based on a



Figure 2: Schematic of a hypothetical region containing 30 positions, with the positions that fall below the fdr threshold marked in black. If we define “nearby” to be ≤ 5 markers, then this region would contain two index sets, I_1 and I_2 , as shown.

low-rank factorization followed by probe-specific standardization is given in Siegmund et al. (2010).

We consider the situation where a set of nearby positions

$$\mathcal{I} = \{i_1, i_2, \dots, i_l\}$$

fall below the fdr threshold in a given sample. By “nearby”, we mean that they are close enough for us to suspect that they may belong to the same contiguous cnv. Since it is common for cnvs in normal samples to cover 10 kilobases or more, and very uncommon for two *different* cnvs to be separated by less than 10 kilobases, we might define “nearby” to be within 10 kilobases of genomic distance, which equates to 1-10 probes depending on the microarray platform. If a position falls below the fdr threshold, and no nearby positions are significant, then we would have $l = 1$, in which case \mathcal{I} would contain only the single position. We assume that the index set \mathcal{I} can not be expanded further, that is, there is no fdr-significant position in the given sample that is nearby, but that does not belong to \mathcal{I} . It is easy to see that the set of all called positions for a given sample can be uniquely partitioned in this way into non-overlapping index sets. An example is shown in Figure 2

For each index set \mathcal{I} (say, corresponding to a sample j), a change-point model can be used to estimate the location(s) of one or more possible change-points in the genomic region containing \mathcal{I} . Let the indices in \mathcal{I} be ordered by genome position, and let $s = i_1 - L$, $t = i_l + L$, where L is a value that is large but much smaller N . We then extract the values $\{x_{s,j}, x_{s+1,j}, \dots, x_{t,j}\}$ from the x matrix, which we re-name element-wise as y_1, \dots, y_T , $T = t - s + 1$, for convenience. If we were to take a hypothesis testing approach this step, the null hypothesis that there is actually nothing going on in this region can be formulated as

$$H_0 : y_i \sim N(0, 1), \quad i = 1, \dots, T,$$

with the alternative hypothesis that there is a cnv interval at $[\tau_1, \tau_2)$ formulated as

$$H_A : y_i \sim N(\mu_i, 1), \quad \mu_i = \begin{cases} \mu, & i = \tau_1, \dots, \tau_2 - 1; \\ 0, & \text{otherwise.} \end{cases}$$

The parameters μ, τ_1, τ_2 are not known. For some platforms, it has been noted that the noise variance increases for CNV regions, which may motivate the addition of an extra variance term σ^2 to the observations within $[\tau_1, \tau_2)$ under the alternative. However, we have found, as does Olshen et al. (2004) and Wen et al. (2006), that the heterogeneous variance model does not significantly improve detection accuracy.

Under the above model, the generalized likelihood ratio assuming known τ_1, τ_2 and maximized over μ has the form

$$L(\tau_1, \tau_2) = \sum_{i=\tau_1}^{\tau_2-1} y_i / (\tau_2 - \tau_1).$$

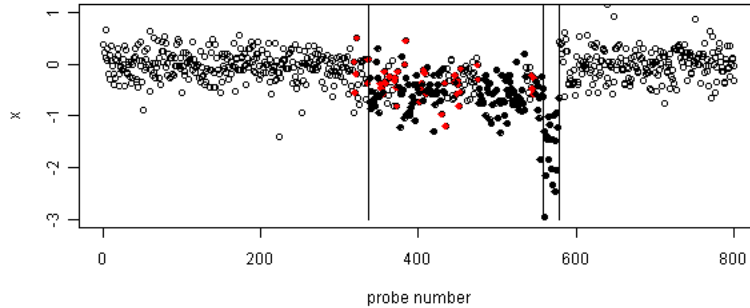


Figure 3: Example of a local cnv analysis. This region contains probes 3400-4200 of sample 41 of the data shown in Figure 1 of the manuscript. The vertical axis is the normalized (but unsmoothed) intensity values (the x_{ij} 's). The red points are locations with $\text{fdr} < 0.05$, and the black points are locations with $\text{fdr} < 0.005$. The vertical lines show the change-points determined by the modified BIC method Zhang and Siegmund (2007).

Maximizing over τ_1 and τ_2 , the generalized likelihood ratio test of H_0 versus H_A is $L(\hat{\tau}_1, \hat{\tau}_2)$, where

$$(\hat{\tau}_1, \hat{\tau}_2) = \operatorname{argmax}_{1 \leq \tau_1 < \tau_2 \leq M} L(\tau_1, \tau_2).$$

The CBS algorithm of Olshen et al. (2004) and the MBIC method of Zhang and Siegmund (2007) use a similar statistic, but adjusted for an unknown baseline mean. Significance values for tests using $L(\hat{\tau}_1, \hat{\tau}_2)$ are given in Siegmund (2007). Since this region has already passed a global filter based on $\widehat{\text{fdr}}$, a less conservative test is more appropriate for the local analysis. In our experience, thresholds of 0.05 or 0.1, without adjusting for the multiple testing across regions, work well. Since the set of regions reported by the fdr procedure should be heavily enriched for true cnvs, it seems more fitting to treat the analysis of each region as an estimation problem rather than as a testing problem. Instead of asking the question, “is there a CNV in this region?” we instead ask, “how many break-points does this region contain, and what are their locations?” This framework is especially fitting for index sets that contain multiple cnvs, or complex variants with nested changes. In this sense, the BIC approach described in Yao (1988) and Zhang and Siegmund (2007) seems to be appropriate. The models in Yao (1988) and Zhang and Siegmund (2007) assume that there are m change-points τ_1, \dots, τ_m (m is unknown). The data is assumed Gaussian, with the mean shifting at each change-point, but with the variance remaining constant. While Yao (1988) showed that the traditional BIC (Schwarz, 1978) is consistent for the estimation of m , Zhang and Siegmund (2007) showed that it is not consistent in estimating the Bayes factor, the quantity that underlies the classic BIC. Zhang and Siegmund (2007) gave a modified form of the BIC, which improves the small sample accuracy for estimating m .

An example, shown in Figure 3, is the well-known cnv region on the p-arm of chromosome 22. The figure shows the break-points estimated by maximum likelihood under this Gaussian model, with the number of break-points estimated using the modified BIC criterion of Zhang and Siegmund (2007). As indicated by the coloring of points, while most of the locations in the nested homozygous deletion (between 559 and 578) pass the 0.005 fdr threshold, some of

the locations in the hemizygous deletion (between 337 and 559) do not even pass the 0.05 threshold. Thus, a local change-point analysis is useful for refining the fdr result and building a more complete picture for each cnv region.

The BIC may report that the candidate region contains no change-points. There would be two possible reasons for this: The region is a false positive, or the signal is so weak that it is missed by the local analysis. Local analysis with a well formulated change-point model should be more powerful than global analyses, because the multiplicity has been much reduced. Thus, if the BIC reports 0 change-points, we conclude that the region is a false positive.

References

- Bengtsson, H., Irizarry, R., Carvalho, B. and Speed, T. (2008). Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* 24: 759–767.
- Bignell, G. R., Huang, J., Greshock, J., Watt, S., Butler, A., West, S., Grigorova, M., Jones, K. W., Wei, W., Stratton, M. R., Futreal, P. A., Weber, B., Shaperro, M. H. and Wooster, R. (2004). High-resolution analysis of dna copy number using oligonucleotide microarrays. *Genome Research* 14: 287–295.
- Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J. M. and Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome snp genotyping platforms. *Nucl. Acids Res.* 36: e126+.
- Friguet, C., Kloareg, M. and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association* 104: 1406–1415.
- Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* 105: 18718–18723.
- Marioni, J. C., Thorne, N. P., Valsesia, A., Fitzgerald, T., Redon, R., Fiegler, H., Andrews, D. T., Stranger, B. E., Lynch, A. G., Dermitzakis, E. T., Carter, N. P., Tavare, S. and Hurles, M. E. (2007). Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biology* 8: R228+.
- Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics* 5: 557–572.
- Peiffer, D. A., Le, J. M., Steemers, F. J., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C. A., Belmont, J., Cheung, S. W., Shen, R. M., Barker, D. L. and Gunderson, K. L. (2006). High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. *Genome Research* 16: 1136–1148.
- Pinkel, D., Seagraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W. L., Chen, C., Zhai, Y., Dairkee, S. H., Ljung, B. M., Gray, J. W. and Albertson, D. G. (1998). High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* 20: 207–11.
- Pollack, J., Perou, C., Alizadeh, A., Eisen, M., Pergamenschikov, A., Williams, C., Jeffrey, S., Botstein, D. and Brown, P. (1999). Genome-wide analysis of dna copy-number changes using cdna microarrays. *Nature Genetics* 23: 41–46.

- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.
- Siegmund, D., Yakir, B. and Zhang, N. (2010). Detecting simultaneous variant intervals in aligned sequences. *Submitted* .
- Siegmund, D. O. (2007). Approximate tail probabilities for the maxima of some random fields. *Annals of Probability* 16: 487–501.
- Snijders, A. M., Nowak, N., Segaves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J. P., Gray, J. W., Jain, A. N., Pinkel, D. and Albertson, D. G. (2001). Assembly of microarrays for genome-wide measurement of dna copy number. *Nature genetics*. 29: 263–264.
- Wen, C., Wu, Y., Huang, Y., Chen, W., Liu, S., Jiang, S., Juang, J., Lin, C., Fang, W., Hsiung, C. and Chang, I. (2006). A bayes regression approach to array-cgh data. *Statistical Applications in Molecular Biology* 5.
- Yao, Y.-C. (1988). Estimating the number of change-point via schwarz' criterion. *Statistics and Probability Letters* 6: 181–189.
- Zhang, N. and Siegmund, D. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* .