

False Discovery Rates and Copy Number Variation

BY BRADLEY EFRON

*Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford
University, Stanford, CA 94305-4065, USA*
siegmund@stanford.edu

AND NANCY R. ZHANG

*Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford
University, Stanford, CA 94305-4065, USA*
nzhang@stanford.edu

SUMMARY

Copy number changes, the gains and losses of chromosome segments, are a common type of genetic variation among healthy individuals as well as an important feature in tumor genomes. Microarray technology enables us to simultaneously measure, with moderate accuracy, copy number variation at more than a million chromosome locations and for hundreds of subjects. This leads to massive data sets and complicated inference problems concerning which locations are more likely to vary. In this paper we consider a relatively simple false discovery rate approach to copy number analysis. More careful parametric change-point methods can then be focused on promising regions of the genome.

Some key words: DNA copy number, False discovery rate, grouped hypotheses, multiple testing

1. INTRODUCTION

Basic genetics says that we have two copies of each bit of chromosomal information. In fact, however, even healthy individuals show occasional variations, displaying stretches of the genome having more or less than two copies. Within the past decade, significant advances in microarray technology have enabled the genome-wide fine scale measurement of DNA copy number in high throughput fashion, see (Pinkel et al., 1998; Pollack et al., 1999; Snijders et al., 2001; Bignell et al., 2004; Peiffer et al., 2006). This has led to large-scale studies investigating the role of DNA copy number changes in human disease and phenotypic variation. The studies fall into two main categories: changes in DNA copy number can occur as a form of inherited genetic polymorphism in normal human DNA. They can also accompany somatic mutation, as often observed in cancerous tumors. Inherited copy number changes in normal samples have been called copy number variants (cnv), while those that occur in tumors have been referred to as copy number aberrations (cna), to distinguish the fact that they are “aberrant” forms which do not occur as population-wide variation. This paper discusses a false discovery rate approach to the analysis of DNA copy number data.

The statistical properties of copy number data are quite different between normal and tumor samples: In normal samples, cnvs are usually short and spaced far apart, whereas in tumor samples, cnas can be quite long, sometimes spanning entire chromosomes. Since false discovery rates inherently cast the problem in a hypothesis testing framework, the methods in this paper

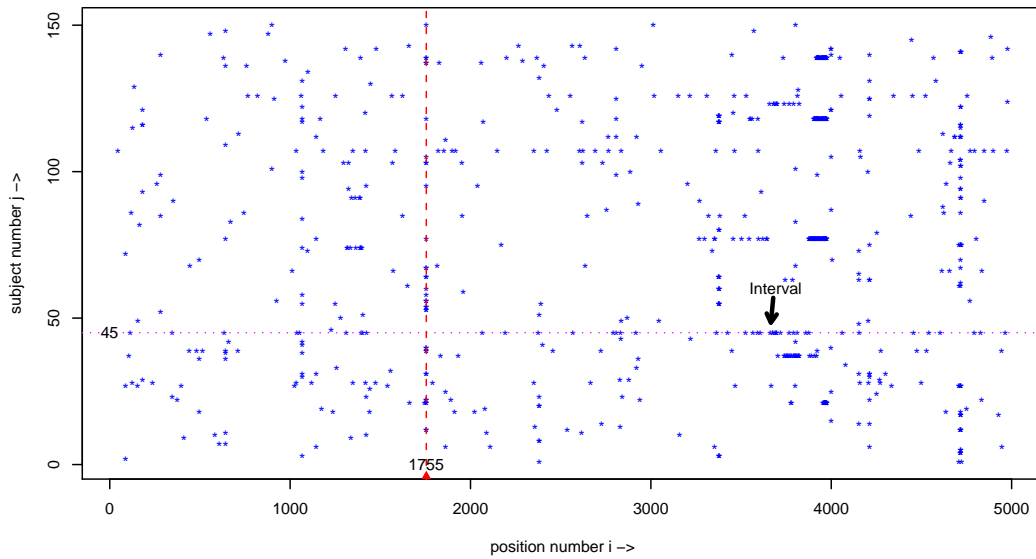


Fig. 1. CNV data for 150 healthy subjects measured at 5000 marker positions. Points indicate positions (i, j) with measurement x_{ij} (1) in the most negative 1/10 of 1%, perhaps indicating copy numbers less than 2. Subject 45 has a long interval of points around position 3800. The marker at position 1755 seems prone to copy numbers < 2 . (Corresponding map for positive x_{ij} values shows less structure.) The matrix has been transposed for convenient display.

perform best when a substantial portion of the data is “null”. This is inherently true for normal samples, and also applies to some tumor samples (see the example in Section 7). In some tumors, chromosomal aberrations cover a significant portion of the genome. Controlling false discovery rates would not be as meaningful in such situations, and the data may be better analyzed from an estimation perspective (e.g. the model selection approach in Zhang & Siegmund (2007), Zhang & Siegmund (2010)).

Figure 1 concerns a cnv data set we will use to illustrate our methodology. Here $n = 150$ healthy subjects have each been assessed at $N = 5000$ marker positions, yielding cnv measurements

$$x_{ij} = \begin{cases} i = 1, 2, \dots, N = 5000 \text{ positions} \\ j = 1, 2, \dots, n = 150 \text{ subjects.} \end{cases} \quad (1)$$

The measurements are normalized probe intensities from microarrays, see Supplementary Materials for details. Roughly speaking, values of x_{ij} much less than zero indicate *less* than two copies, and values much greater than zero *more* than two copies, but there is considerable measurement error. The histogram of all 750,000 x_{ij} is smoothly unimodal, mean and standard deviation -0.018 and 0.188 , showing moderate skewness toward the negative side, with a small percentage of extreme outliers in both directions.

The points in Figure 1 (where the matrix has been transposed) indicate the 750 most negative x_{ij} values, in other words the one-tenth of one percent of (position, subject) combinations giving strongest evidence of copy numbers less than 2. We can see that subject 45, for example, seems to have a long interval of decreased copy numbers around position 3800, while position

1755 might be prone to copy number reductions. A typical question we would like to answer is whether position 1755 is genuinely *cnv*-prone. A key feature of *cnv* problems is the availability of information in both directions. Does subject 45 have less than two copies of the marker at position 1755? The methodology introduced in Section 2 combines the horizontal and vertical features in Figure 1 to answer such questions.

Let \bar{x}_{ij} indicate a moving average of the x_{ij} values for subject j ,

$$\bar{x}_{ij} = \sum_{i'=i-m}^{i+m} x_{i'j} / (2m + 1) \quad (2)$$

for some fixed value of m (with obvious modifications for i near 1 or N). Because *cnv* intervals tend to span a contiguous range of marker positions, \bar{x}_{ij} will be less noisy than x_{ij} ; see Section 5 and the Supplementary Materials for a discussion on the choice of m . It is also helpful to standardize the columns of the $\{\bar{x}_{ij}\}$ matrix, that is, each *subject's* \bar{x}_{ij} values, by defining

$$z_{ij} = (\bar{x}_{ij} - \hat{a}_j) / \hat{b}_j \quad (3)$$

where \hat{a}_j and \hat{b}_j are the median and robust standard deviation (one-half the distance between the 16th and 84th percentiles) of $\{\bar{x}_{ij} : i = 1, 2, \dots, N\}$. Most of our numerical examples will be based on z_{ij} values (2), (3) with $m = 5$. The application of *fdr* methodology to the z -values renders copy number variations far more visible; see Figure 3.

There are by now many different methods for single sample analysis of DNA copy number. These methods process each subject (i.e. each column in the matrix (1.1)) separately, as if the method has never seen a similar sample before, and will never see another sample again. Reviews of single sample methods are given in Lai et al. (2005), Willenbrock & Fridlyand (2005), and Zhang (2010). For both single sample analysis and the simultaneous processing of multiple samples, global change-point tests, scanning over the entire range of positions, have played a central role in the statistical *cnv* literature (Olshen et al., 2004; Zhang et al., 2010; Siegmund et al., 2010). The literature leans heavily on Gaussian process theory, and within that realm produces impressively precise testing algorithms. Wang et al. (2005) propose an *fdr* approach, closer to the methods proposed here.

A large part of this paper is devoted to the question “which positions are more CNV prone than expected under random chance?” The identification of *cnv* prone regions has gained increasing scientific interest, especially in the analysis of tumor samples. Most published methods (Newton et al., 1998; Newton & Lee, 2000; Diskin et al., 2006; Beroukhim et al., 2007; Guttman et al., 2007; Taylor et al., 2008; Rouveirol et al., 2006) take a “post-segmentation” approach: Each sample is first segmented individually, which reduces them to piece-wise constant sequences indicating regions of amplification, deletion, or normal copy number. Then, the samples are aligned, and a statistical model (Newton et al. (1998); Newton & Lee (2000)) or permutation based approach (Diskin et al. (2006)) is used to identify regions of highly recurrent aberration. These post-segmentation approaches rely on the vagaries of the underlying segmentation model. After segmentation, how evidence for gains and losses should be combined across samples is still much debated. Existing strategies range from counting the number of carriers, without weighting by the strength of evidence of each carrier (e.g. Diskin et al. (2006)), to the “G-score” (Beroukhim et al., 2007), defined as the number of carriers times the average amplitude of the signal among carriers. The *fdr*-based approach that we take here arises from a natural likelihood model and is computationally scalable to large data sets.

The paper develops as follows: an iterative algorithm is introduced in Section 2, in which a local false discovery rate estimate (Efron, 2008) is first fit to the combined data, and then modified

145 to take account of differing cnv probabilities at the various positions i . This gives an fdr estimate
 146 for each position and subject, as well as an estimate \hat{k}_i of the number of subjects carrying a cnv at
 147 position i . Sections 3 and 4 develop hypothesis testing and estimation methods based on the \hat{k}_i 's,
 148 aimed at answering the question of which, if any, of the positions are cnv-prone. The iterative
 149 algorithm is examined more closely in Sections 5 and 6, and connected to maximum likelihood
 150 theory. Section 7 examines in more detail the problem of detecting cna prone regions in tumors.
 151 Having located positions prone to copy number changes based on the fdr or \hat{k}_i estimates, one
 152 then needs to conduct a more detailed analysis to more accurately identify the set of carriers and
 153 estimate the CNV breakpoints. A discussion of follow-up estimation procedures is given in the
 154 Supplementary Materials.

156 2. FALSE DISCOVERY RATE METHODS

158 Forgetting about cnv structure for a moment, suppose we have M null hypotheses
 159 $H_{01}, H_{02}, \dots, H_{0M}$ to test, based on possibly correlated test statistics z_1, z_2, \dots, z_M . False dis-
 160 covery rate methods can be motivated by the Bayesian *two-groups model* discussed at length
 161 in Efron (2008), in which each case is either null or non-null with prior probability π_0 or
 162 $\pi_1 = 1 - \pi_0$, and with the z values having density either $f_0(z)$ or $f_1(z)$,

$$163 \begin{aligned} 164 \pi_0 &= \Pr\{\text{null}\} & f_0(z) &= \text{density if null} \\ 165 \pi_1 &= \Pr\{\text{non-null}\} & f_1(z) &= \text{density if non-null.} \end{aligned} \quad (4)$$

166 Bayes rule shows that the posterior probability of “null” given z , the *local false discovery rate*,
 167 is

$$169 \text{fdr}(z) = \Pr\{\text{null}|z\} = \pi_0 f_0(z) / f(z) \quad (5)$$

170 where $f(z)$ is the mixture density

$$172 f(z) = \pi_0 f_0(z) + \pi_1 f_1(z). \quad (6)$$

174 An empirical Bayes approach to multiple testing uses the entire vector $\mathbf{z} = (z_1, z_2, \dots, z_M)$
 175 to estimate $\pi_0, f_0(z), f(z)$, and then $\text{fdr}(z)$,

$$177 \widehat{\text{fdr}}(z) = \hat{\pi}_0 \hat{f}_0(z) / \hat{f}(z), \quad (7)$$

179 rejecting the m th null hypothesis H_{0m} if $\widehat{\text{fdr}}(z_m)$ is small, perhaps for $\widehat{\text{fdr}}(z_m) \leq 0.1$ or ≤ 0.01 .
 180 (Replacing densities f_0 and f_1 with their cumulative distribution functions gets us back, almost,
 181 to Benjamini and Hochberg’s 1995 false discovery rate control algorithm, but here it will be more
 182 convenient to deal directly with the densities.)

183 Figure 2 shows $\widehat{\text{fdr}}(z)$ based on the combined data for all $M = 750,000$ z values z_{ij} (3),
 184 computed using the `locfdr` algorithm, Efron (2008); `locfdr` assumes that $f_0(z)$ is normal,
 185 a reasonable assumption here looking at the central portion of the $\{z_{ij}\}$ histogram, which the
 186 averaging in (2) renders quite Gaussian. The numerator estimates in (7) were

$$187 \hat{\pi}_0 = 0.954, \quad \hat{f}_0 \sim \mathcal{N}(0.04, 0.93^2), \quad (8)$$

189 obtained using the “central matching” (geometric) method. Taken literally, this implies 4.6% of
 190 the (i, j) pairs represent cnv locations,

$$192 \hat{\pi}_1 = 0.046. \quad (9)$$

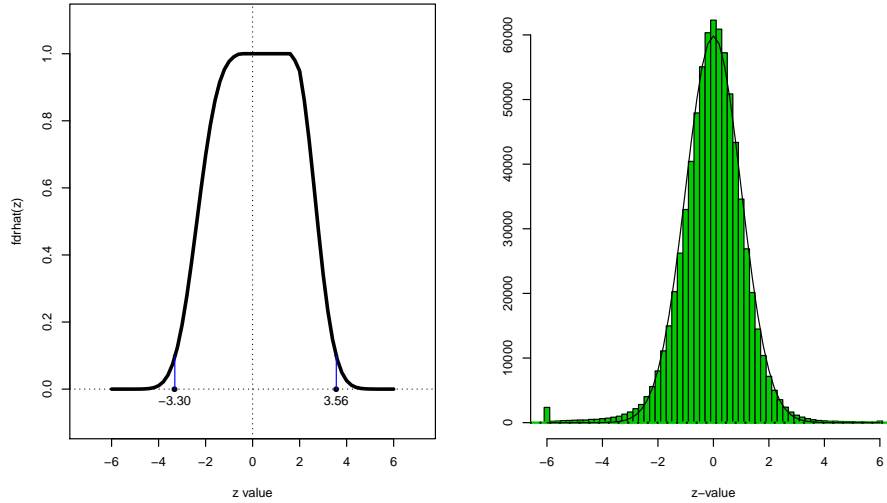


Fig. 2. *Left*: Estimated local false discovery rate $\widehat{\text{fdr}}(z)$ based on all 750,000 values z_{ij} (3) for the *cnv* data; $\widehat{\text{fdr}}(z)$ is ≤ 0.1 for $z \leq -3.30$ and $z \geq 3.56$, respectively 1.2% and 0.3% of the 750,000 cases. Computed from program `locfdr`, Efron (2008). *Right*: A histogram of all z_{ij} values, with the light curve indicating a $N(0, 1)$ density function.

As discussed in Efron (2008), we can expect a majority of such pairs to have disappointingly large values of $\widehat{\text{fdr}}(z_{ij})$. Here only 1.5% of the 750,000 have $\widehat{\text{fdr}}(z_{ij}) \leq 0.1$, those with $z_{ij} \leq -3.30$ or ≥ 3.56 . However, we can improve power by adapting the *fdr* methodology to the two-way structure of *cnv* data.

Let \mathcal{C}_i be the class of pairs (i, j) corresponding to position i ,

$$\mathcal{C}_i = \{(i, j) : j = 1, 2, \dots, n\} \quad (10)$$

so the corresponding z values z_{ij} are all those obtained at position i . We now have N classes $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_N$, one for each position, and can imagine fitting a *separate* two-groups model (4) for each class, yielding separate false discovery rate functions $\widehat{\text{fdr}}_i(z)$. The trouble is that, unless n is very large, a direct approach as in Figure 2 will produce inaccurate estimates $\widehat{\text{fdr}}_i(z)$.

A compromise between using the combined estimate $\widehat{\text{fdr}}(z)$ or completely separate estimates $\widehat{\text{fdr}}_i(z)$ goes as follows: we assume that the null and non-null densities f_0 and f_1 in (4) apply unchanged to each class, but that the null and non-null prior probabilities may differ, say

$$\pi_{i0} = \Pr\{\text{null}|\mathcal{C}_i\} \quad \text{and} \quad \pi_{i1} = 1 - \pi_{i0} = \Pr\{\text{non-null}|\mathcal{C}_i\}. \quad (11)$$

So a *cnv*-prone position would be one having a larger value of π_{i1} than the combined value π_1 .

Using $\widehat{\text{fdr}}_i(z) = \pi_{i0}f_0(z)/f_{(i)}(z)$, with $f_{(i)}(z)$ the mixture density applying to \mathcal{C}_i ,

$$f_{(i)}(z) = \pi_{i0}f_0(z) + \pi_{i1}f_1(z), \quad (12)$$

comparison with (5) gives

$$\widehat{\text{fdr}}_i(z) = \widehat{\text{fdr}}(z)/[1 + \text{tdr}(z) \cdot R_i] \quad (13)$$

193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240

241 where $\text{tdr}(z)$ is the *true discovery rate*

$$242 \quad \text{tdr}(z) = 1 - \text{fdr}(z) = \Pr\{\text{non-null}|z\} \quad (14)$$

243 and

$$244 \quad R_i = \frac{\pi_{i1}/\pi_1}{\pi_{i0}/\pi_0} - 1. \quad (15)$$

245 An equivalent form is

$$246 \quad \text{tdr}_i(z) = \text{tdr}(z)/[1 + \text{fdr}(z) \cdot S_i] \quad (16)$$

247 now where $\text{tdr}_i(z) = 1 - \text{fdr}_i(z)$ is the true discovery rate $\Pr\{\text{non-null}|z, \mathcal{C}_i\}$ applying to \mathcal{C}_i ,
248 and

$$249 \quad S_i = \frac{\pi_{i0}/\pi_0}{\pi_{i1}/\pi_1} - 1. \quad (17)$$

250 Section 6 discusses (13) and (16) in more detail.

251 None of this seems like a step forward since (13) and (16) both require knowledge of π_{i1} ,
252 the non-null proportion in \mathcal{C}_i . There is, however, a simple iterative solution. Given a preliminary
253 estimate $\widehat{\text{tdr}}_i(z)$ of $\text{tdr}_i(z)$, perhaps $\widehat{\text{tdr}}(z)$ from the combined analysis,

$$254 \quad \hat{k}_i = \sum_{j \in \mathcal{C}_i} \widehat{\text{tdr}}_i(z_{ij}) = \sum_{j=1}^n \widehat{\text{tdr}}_i(z_{ij}) \quad (18)$$

255 is the obvious estimate of k_i , the number of non-null cases in \mathcal{C}_i , since $\widehat{\text{tdr}}_i(z_{ij})$ estimates the
256 probability that case (i, j) is non-null.

257 This yields

$$258 \quad \hat{\pi}_{i1} = \hat{k}_i/n = \sum_{j=1}^n \widehat{\text{tdr}}_i(z_{ij}) / n \quad (19)$$

259 as an estimate of π_{i1} , and

$$260 \quad \hat{S}_i = \frac{(1 - \hat{\pi}_{i1})/\hat{\pi}_0}{\hat{\pi}_{i1}/\hat{\pi}_1} - 1 \quad (20)$$

261 for (17) (with $\hat{\pi}_0$ and $\hat{\pi}_1$ obtained from the combined analysis that gave $\widehat{\text{fdr}}$ and $\widehat{\text{tdr}}$, as in (9)).
262 We can now update (16) to

$$263 \quad \widehat{\text{tdr}}_i(z_{ij}) = \widehat{\text{tdr}}(z_{ij}) / \left[1 + \widehat{\text{fdr}}(z_{ij}) \hat{S}_i \right], \quad (21)$$

264 recompute (18), etc. The numerical results that follow stopped after five iterations of (18)–(21),
265 close to the final convergence values; see Section 6. Other examples, involving fewer subjects,
266 required more iterations to reach convergence, though the increase did not noticeably affect sub-
267 sequent inferences.

268 Figure 3 displays the results. The top panel shows those pairs (i, j) having $\widehat{\text{tdr}}_i(z_{ij}) \geq 0.99$, or
269 equivalently $\widehat{\text{fdr}}_i(z_{ij}) \leq 0.01$. A very long cnv region is centered around $i = 3800$, with shorter
270 but still prominent regions near 1, 1100, 1300, 3000, and 4700. Position $i = 1755$ is less impres-
271 sive, but does show some non-null cases. The bottom panel graphs \hat{k}_i as a function of position i ,
272 with $\hat{k}_{1755} = 39.1$ showing as a small isolated spike. Is this a “significant” result? The next two
273 sections consider testing and estimation questions for the \hat{k}_i values.
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288

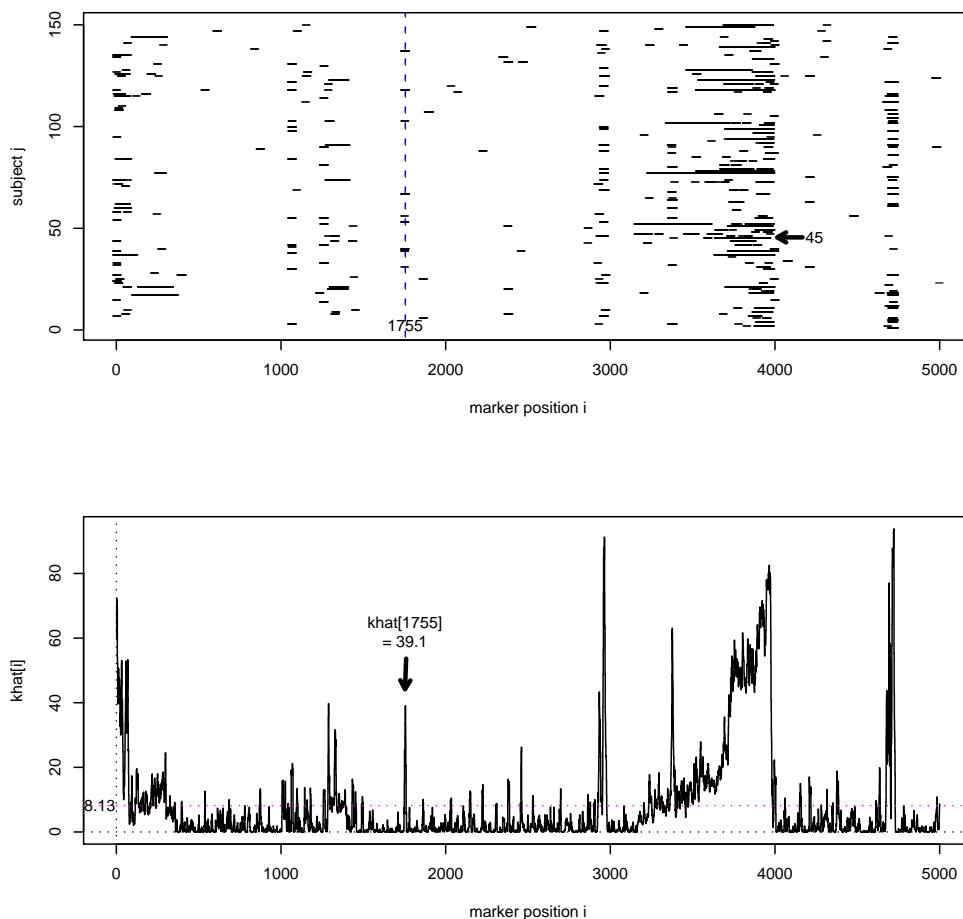


Fig. 3. Algorithm (18)–(21), five iterations, applied to *cnv* data (1). *Top panel*: (position, subject) pairs (i, j) having estimated true discovery rate $\widehat{\text{tdr}}_i(z_{ij}) \geq 0.99$. *Bottom panel*: estimates \hat{k}_i for the number of non-null subjects at marker position i .

3. HYPOTHESIS TESTS FOR POSITION-WISE COPY NUMBER VARIATION

Having obtained estimates \hat{k}_i , $i = 1, 2, \dots, N$, for the number of non-null *cnv* subjects at marker position i , we wish to decide which, if any, of the positions are especially prone to copy number variation. For example, \hat{k}_{1755} equals 39.1, compared to the average $\bar{k} = 8.13$ in Figure 3, which might suggest excess variation at position 1755, a hypothesis we would like to test.

An easy permutation test proceeds as follows: let z_j be the j th column of the \mathbf{Z} matrix $\{z_{ij}\}$ (3), and z_j^* the same vector except shifted left I units (with wraparound),

$$z_j^* = (z_{I+1,j}, z_{I+2,j}, \dots, z_{N,j}, z_{1j}, z_{2j}, \dots, z_{Ij})'. \quad (22)$$

Choosing I as an independent and random integer between 0 and $N - 1$ for each of the n rows yields a permuted matrix,

$$\mathbf{Z}^* = (z_1^*, z_2^*, \dots, z_n^*) \quad (23)$$

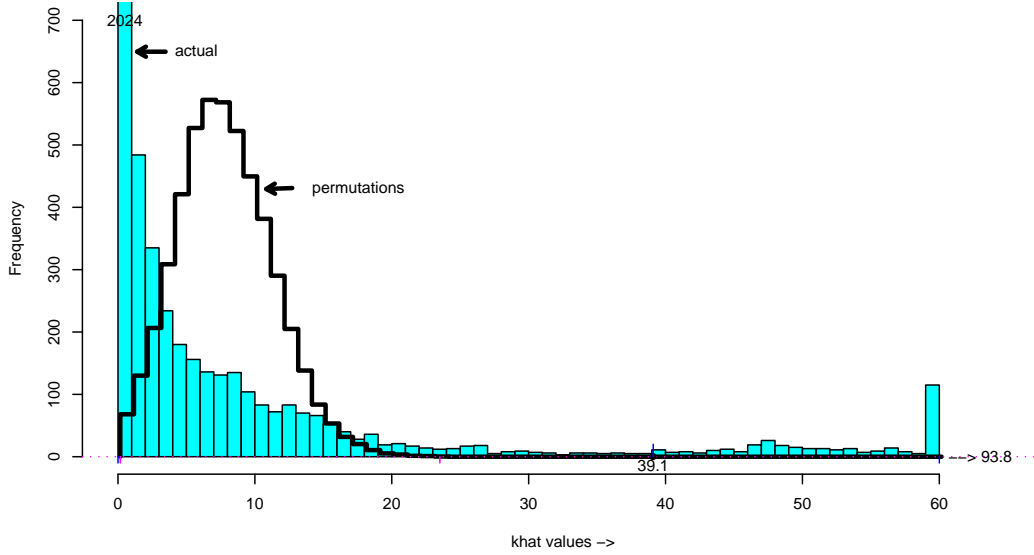


Fig. 4. Histogram of the 5000 \hat{k}_i estimates for cnv data, from 5 iterations of (18)–(21) (solid), compared to 50,000 permutation values \hat{k}_i^* as following (23) (line histogram). Maximum permutation value equals 23.3, far less than $\hat{k}_{1755} = 39.1$. Spike at $\hat{k} = 60$ represents 106 \hat{k}_i values ≥ 60 , maximum 93.8. The 2024 \hat{k}_i values ≤ 1 are significantly too *small* according to the permutation distribution.

in which position-wise structure has been nullified, while any subject-wise structure of cnv intervals is maintained. The permutation test compares \hat{k}_i with the distribution $\{\hat{k}_1^*, \hat{k}_2^*, \dots, \hat{k}_N^*\}$, where the k^* values are obtained by applying algorithm (18)–(21) to \mathbf{Z}^* . (Notice that \mathbf{Z}^* has the same elements as \mathbf{Z} , so that the combined analysis quantities $\hat{\pi}_0, \hat{\pi}_1, \widehat{\text{fdr}}(z)$ and $\widehat{\text{tdr}}(z)$ have the same values as in (18)–(21).)

Ten independent replications of \mathbf{Z}^* were generated for the example of Figure 3, yielding 50,000 \hat{k}_i^* values in total. The line histogram in Figure 4 compares them with the distribution of the 5000 actual \hat{k}_i values: $\max\{\hat{k}_i^*\} = 23.3$ suggesting, for example, that $\hat{k}_{1755} = 39.1$ is strongly significant evidence for excess variation at position 1755. In less-extreme circumstances we could compute permutation p -values,

$$p_i = \text{proportion of permutation values exceeding } \hat{k}_i \quad (24)$$

and use a standard false discovery rate procedure to assess significance among the N p -values.

Basing our significance tests on \hat{k}_i values seems reasonable but perhaps *ad hoc*. It can, however, be motivated in terms of the two-groups model (4), (11). Define

$$r = \pi_{i1}/\pi_1 \quad (25)$$

the ratio of the non-null probability at the i th position to the combined value π_1 ; the null hypothesis H_{0i} that position i is *not* cnv-prone is $H_{0i} : r = 1$.

385 Observation z_{ij} has density $f(z)$ under H_{0i} and density $f_{(i)}(z)$ under (11), (12), giving log
386 likelihood ratio

$$387 \quad l(z_{ij}) = \log \left\{ \frac{f_{(i)}(z_{ij})}{f(z_{ij})} \right\} = \log \left\{ \frac{\pi_{i0}f_0(z_{ij}) + \pi_{i1}f_1(z_{ij})}{\pi_0f_0(z_{ij}) + \pi_1f_1(z_{ij})} \right\} \quad (26)$$

$$390 \quad = \log \{1 + (r - 1)T(z_{ij})\}$$

391 where some calculation yields

$$393 \quad T(z) = \frac{\text{tdr}(z) - \pi_1}{\pi_0}, \quad (27)$$

395 $\text{tdr}(z) = 1 - \text{fdr}(z)$ as in (14). Assuming that subjects were sampled independently, the n ob-
396 servations in $\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{in})$ are independent of each other. The most powerful test of
397 H_{0i} versus a specific alternative choice of $r > 1$ then rejects H_{0i} for large values of

$$399 \quad l_r(\mathbf{z}_i) = \sum_{j=1}^n \log \{1 + (r - 1)T(z_{ij})\}. \quad (28)$$

402 The locally most powerful (lmp) test of $r = 1$ versus $r > 1$ rejects for large values of

$$404 \quad \left. \frac{\partial l_r(\mathbf{z}_i)}{\partial r} \right|_{r=1} = \sum_{j=1}^n T(z_{ij}) = \sum_{j=1}^n \frac{\text{tdr}(z_{ij}) - \pi_1}{\pi_0}, \quad (29)$$

407 an increasing function of $\sum_1^n \text{tdr}(z_{ij})$. In practice we could reject for large values of

$$409 \quad \sum_{j=1}^n \widehat{\text{tdr}}(z_{ij}) = \hat{k}_i^{(1)} \quad (30)$$

412 where $\hat{k}_i^{(1)}$ is from the *first* iteration of k_i in (18), beginning at $\widehat{\text{tdr}}_i(z) = \widehat{\text{tdr}}(z)$. This justifies
413 $\hat{k}_i^{(1)}$ as a preferred test statistic for H_{0i} .

415 In our *cnv* example, $\hat{k}_i^{(5)}$, the fifth iterate, performed a little better than $\hat{k}_i^{(1)}$, almost matching
416 the most powerful test statistic (26) over the range $1 < r \leq 4$. This seems to put the significance
417 of position 1755 as *cnv-prone* on safe footing.

418 Figure 4 is not completely reassuring in this regard: the permutation distribution does not
419 look much like a reasonable null hypothesis, since it makes “significant” a majority of the 5000
420 positions. In particular, the 2024 positions having $\hat{k}_i \leq 1$ are significantly too *small* by the per-
421 mutation criterion. Perhaps we should be estimating the accuracy of the \hat{k}_i values rather than
422 testing them for nullness, a point of view taken up in Section 4.

424 4. THE ACCURACY OF POSITION-WISE ESTIMATES

426 How accurate is \hat{k}_i as an estimate of k_i , the number of non-null cases at position i ? A sim-
427 ple answer is obtained by resampling the n subjects and calculating non-parametric bootstrap
428 estimates of standard deviations.

429 Let $\mathbf{z}_j = (z_{1j}, z_{2j}, \dots, z_{Nj})'$ be the N -vector of data for subject j . A typical bootstrap data
430 matrix is

$$432 \quad \mathbf{Z}^* = (\mathbf{z}_{j1}, \mathbf{z}_{j2}, \dots, \mathbf{z}_{jn}) \quad (31)$$

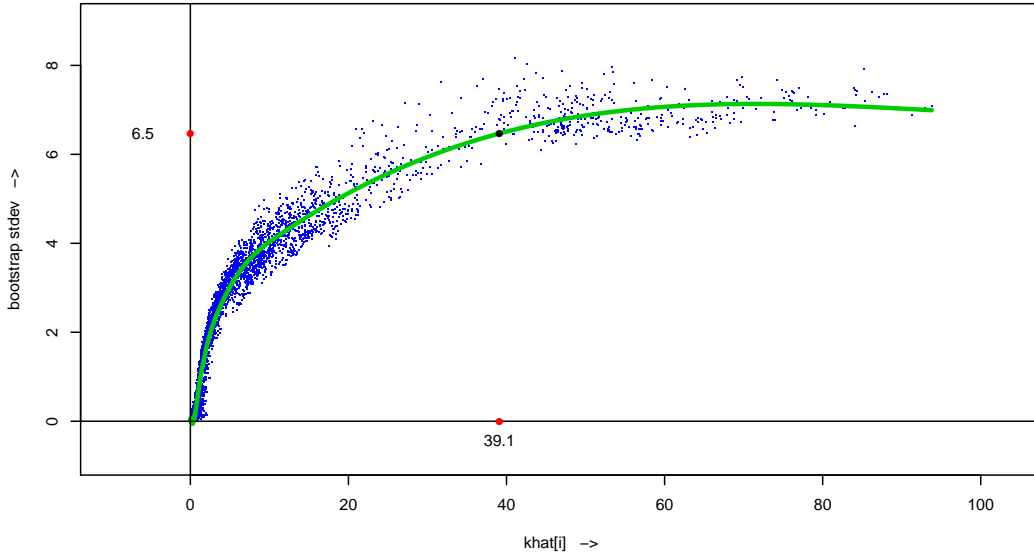


Fig. 5. Bootstrap standard deviations of \hat{k}_i estimates (18)–(21), 5 iterations, for *cnv* data (1), plotted versus \hat{k}_i . Smooth curve is natural spline least squares fit, 4 degrees of freedom. At $\hat{k}_{1755} = 39.1$ it gives $\widehat{sd}_{1755} = 6.5$.

where j_1, j_2, \dots, j_n is a random sample taken with replacement from the integers $(1, 2, \dots, n)$. We calculate k_i^* , $i = 1, 2, \dots, N$, from Z^* according to (18)–(21), including the five iterations. Doing so B times gives bootstrap standard deviation estimates

$$\widehat{sd}_i = \sqrt{\frac{\sum_{b=1}^B (\hat{k}_i^{*b} - \hat{k}_i^*)^2}{(B-1)}} \quad (32)$$

where $\hat{k}_i^* = \sum_1^B \hat{k}_i^{*b} / B$.

Figure 5 plots \widehat{sd}_i versus \hat{k}_i for the *cnv* data (1), $i = 1, 2, \dots, 5000$, based on $B = 200$ bootstrap replications. A smooth curve has been drawn through the 5000 $(\hat{k}_i, \widehat{sd}_i)$ points, giving for example $\widehat{sd}_{1755} = 6.5$ at $\hat{k}_{1755} = 39.1$. This yields approximate 95% confidence intervals $\hat{k}_i \pm 2 \cdot \widehat{sd}_i$, in particular,

$$k_{1755} \in (26.1, 52.1). \quad (33)$$

The lower limit is far above $\bar{k} = 8.3$, providing further evidence that position 1755 is *cnv*-prone. *Note:* The bootstrap calculations did not include recomputation of the combined quantities $\hat{\pi}_0, \hat{\pi}_1, \widehat{tdr}(\cdot)$, and $\widehat{fdr}(\cdot)$ in (18)–(21), which were kept at their original values. This amounts to treating them as fixed ancillary statistics, as is effectively done in the permutation test of Section 3. Recomputing the combined quantities for each bootstrap replication considerably increased the standard deviation estimates ($\widehat{sd}_{1755} = 9.1$ for example), and seemed inappropriately conservative here.

At this point, *selection bias* needs to be considered: positions such as 1755 come to our attention because of their unusual \hat{k}_i values, which can be misleadingly large when selected from thousands of possibilities. Frequentist corrections for bias are difficult here, but a simple empirical Bayes calculation offers some insight.

481 Consider the univariate Bayesian model in which a parameter μ is drawn from prior density
 482 $g(\cdot)$ and then $x \sim \mathcal{N}(\mu, \sigma^2)$ is observed,

$$483 \mu \sim g(\cdot) \quad \text{and} \quad x|\mu \sim \mathcal{N}(\mu, \sigma^2). \quad (34)$$

484 Let $f(x)$ be the marginal density of x ,

$$485 f(x) = \int_{-\infty}^{\infty} \varphi_{\sigma}(x - \mu)g(\mu) d\mu, \quad (35)$$

486 $\varphi_{\sigma}(x) = \exp\{-x^2/2\sigma^2\}/\sqrt{2\pi\sigma^2}$, and define $l(x) = \log\{f(x)\}$.

487 LEMMA 1. *The posterior expectation and standard deviation of μ given x are*

$$488 E\{\mu|x\} = x + \sigma^2 l'(x) \quad (36)$$

489 and

$$490 \text{Sd}\{\mu|x\} = \sigma \cdot [1 + \sigma^2 l''(x)]^{1/2} \quad (37)$$

491 where $l'(x)$ and $l''(x)$ are the first and second derivations of $l(x)$.

492 *Proof.* According to Bayes theorem, the posterior density of μ given x is

$$493 g(\mu|x) = \varphi_{\sigma}(x - \mu)g(\mu)/f(x) \quad (38)$$

$$494 = e^{x\mu/\sigma^2 - \psi(x)} [g(\mu)e^{-\mu^2/2\sigma^2}]$$

495 with

$$496 \psi(x) = \log \{f(x)/\varphi_{\sigma}(x)\}; \quad (39)$$

497 (38) is a one-parameter exponential family with canonical parameter x and sufficient statistic
 498 μ/σ^2 . Differentiating $\psi(x)$ twice yields the mean and variance of μ/σ^2 given x , verifying the
 499 lemma. \square

500 Formula (36) goes back, at least, to Robbins (1956), who credits correspondence with M.
 501 Tweedie, though (37) seems less familiar. They are ideal for empirical Bayes purposes: having
 502 observed x_1, x_2, \dots, x_N from repeated realizations (μ_i, x_i) of (34), we can directly estimate
 503 $f(x)$ and $l(x)$, and differentiate to get $E\{\mu|x\}$ and $\text{sd}\{\mu|x\}$ from (36)–(37). The key point is
 504 that deconvolution for the estimation of the prior $g(\mu)$ is completely avoided.

505 Now let (k_i, \hat{k}_i) play the role of (μ_i, x_i) in (34). The 200 bootstrap replications for Figure 5
 506 showed

$$507 \hat{k}_i \sim \mathcal{N}(k_i, \sigma_i^2) \quad (40)$$

508 to be a reasonable approximation, with σ_i as indicated in Figure 5. A density estimate $\hat{f}(k)$
 509 was obtained by fitting a smooth curve to the histogram heights in Figure 4 (using a Poisson
 510 generalized linear model based on a natural spline with five degrees of freedom, as described
 511 in Remark D of Efron, 2009). This gave $\hat{l}(k) = \log\{\hat{f}(k)\}$, $\hat{l}'(k)$, $\hat{l}''(k)$, and then $\hat{E}\{k_i|\hat{k}_i\}$ and
 512 $\widehat{\text{Sd}}\{k_i|\hat{k}_i\}$, with σ^2 obtained from the fitted curve in Figure 5.

513 1 displays $\hat{E}\{k|\hat{k}\}$ and the posterior bounds $\hat{E}\{k|\hat{k}\} \pm 2 \cdot \widehat{\text{sd}}\{k|\hat{k}\}$. Unlike the examples in
 514 Efron (2009), the results here are not very different from the frequentist estimates $\hat{k}_i \pm 2 \cdot \widehat{\text{sd}}_i$.
 515 In particular, for $\hat{k}_{1755} = 39.1$ we get $\hat{E} = 41.3$ and posterior interval (26.5,56.0), almost the
 516 same as (33). Figure 4 shows a greater concentration of \hat{k}_i values within a couple of standard
 517

\hat{k}	10	20	30	40	50	60	70	80
$\hat{E} - 2 \cdot \widehat{\text{sd}}$	-1.0	5.6	13.4	28.1	41.3	41.7	50.8	68.6
\hat{E}	7.4	16.3	27.7	42.4	49.6	56.1	68.8	83.3
$\hat{E} + 2 \cdot \widehat{\text{sd}}$	15.7	27.1	42.0	56.8	57.8	70.4	86.8	98.1

Table 1. Empirical Bayes estimates $\hat{E}\{k|\hat{k}\}$ and posterior limits $\hat{E}\{k|\hat{k}\} \pm 2 \cdot \widehat{\text{Sd}}\{k|\hat{k}\}$.

deviations to the right of 39.1 than to the left, accounting for the slightly increased Bayesian estimate and interval limits.

Bayes estimates are immune to selection bias. If in fact the posterior expectation of k_{1755} equals 41.3, then it does not matter why position 1755 came to our attention. The reassuring message of 1 is that selection bias is not a serious problem here.

There are reasons for skepticism:

- Model (34) has σ constant, whereas it varies in our application. More careful calculations show that the effect is small for this situation (only slightly raising the estimates for position 1755).
- At best, the calculations are approximating $E\{k_i|\hat{k}_i\}$, not $E\{k_i|\hat{\mathbf{k}}\}$, the posterior expectation given *all* the k values.
- The \hat{k}_i estimates are correlated with each other. This does not invalidate the use of Lemma 1, but degrades the accuracy of the empirical Bayes estimates; see Efron (2010a).

These last two points emphasize the fact that empirical Bayes is not actual Bayes, and provides no strict theoretical basis for ignoring selection bias. Nevertheless, the results in 1 offer a useful guide for interpreting the estimates \hat{k}_i .

Various numerical experiments were carried out investigating the accuracy of \hat{k}_i calculations. The observations x_{ij} in (1) actually were each the average of two independent replications x_{ij1} and x_{ij2} . Applying algorithm (18)–(21) separately to the two sets gave nearly the same results, both being slightly degraded versions of the analysis based on (1).

Another test involved “spiking in” artificial cnv signals at non-active positions of data (1); for example, adding a square wave signal to 40 of the subjects at positions 2233 through 2239. The corresponding \hat{k}_i values edged up to 40 as the size of the square wave increased, topping out at about 50 for enormous signals. (The window width $2m + 1$ in (2) was kept at 11 as before.) Large numbers of low values of $\widehat{\text{tdr}}_i(z_i)$ in (18) were responsible for the upward bias, which perhaps suggests imposing a cut-off threshold. Section 5 briefly discusses the relation of window width to power and bias.

At this point, we reveal the fact that probe number 1755, which is located at genome base pair position 17,952,757 (NCBI human genome build 36), indeed falls into a region containing previously identified deletions. The deletions in this region have been detected by Conrad et al. (2006) using SNP genotyping arrays and by Mills et al. (2006) and McKernan et al. (2009) using short read sequencing. These studies differ in their estimated boundaries, but all agree that there is a deletion covering probe 1755 in at least one subject in their study. We swear that these facts were discovered only after position 1755 was identified by our analysis.

5. ESTIMATION OF FALSE DISCOVERY RATES

The procedures used for the estimation of $\widehat{\text{fdr}}_i(z)$ (13), the local false discovery rate applying to position i , raise some questions discussed in this section.

A preliminary question concerns the choice of moving average window width $M = 2m + 1$ involved in the construction of the z -values z_{ij} (2)–(3). Some insight is gained from a simple model in which the observations x_{ij} in (1) are independent normal variates with expectation either 0 or μ ,

$$\text{null } x_{ij} \sim \mathcal{N}(0, 1) \quad \text{non-null } x_{ij} \sim \mathcal{N}(\mu, 1) \quad (41)$$

and where the non-null cases for subject j occur in contiguous blocks. The well-known signal matching criterion says that M should match the width of the contiguous blocks, but of course that will usually be unknown. A brief discussion of the trade-off between too wide or too narrow window width appears in the Supplementary Material. In our case, changing M from 11 to 21 produced small increases in most of the larger \hat{k}_i values seen in Figure 3, a notable exception being at $i = 1755$ — inside a very short block — where \hat{k}_i was halved. The value $M = 11$ performed satisfactorily on several other data sets, though the specific choice never seemed crucial.

Our data set (1) includes copy number variations in both negative and positive directions, that is, having less or more than two copies. This can be seen in Figure 2, where the combined local false discovery rate $\widehat{\text{fdr}}(z)$ decreases to zero at both ends of the z scale. As a consequence, the estimates \hat{k}_i produced by algorithm (18)–(21) are *two-sided*: if we begin the iteration with $\widehat{\text{tdr}}_i(z) = \widehat{\text{tdr}}(z) = 1 - \widehat{\text{fdr}}(z)$ then \hat{k}_i is increased for z_{ij} values that are extreme in either direction.

As we will show in the following, two-sidedness can have undesirable effects. It is simple, and probably preferable, to calculate instead both *one-sided* \hat{k}_i estimates. Beginning the iteration at (18) with

$$\widehat{\text{tdr}}_i(z) = \begin{cases} \widehat{\text{tdr}}(z) & \text{if } z \leq 0 \\ 0 & \text{if } z > 0 \end{cases} \quad (42)$$

instead of $\widehat{\text{tdr}}(z)$ produces “left-sided” \hat{k}_i estimates, sensitive only to negative z_{ij} values. A similar tactic gives right-sided \hat{k}_i estimates. The sum of the left- and right-sided estimates is similar to the two-sided estimates of Section 2, but there is an interpretive advantage in observing both sides.

Some of the positions for data set (1) (though not 1755) displayed large z_{ij} in both directions. These can be genuine, but we might worry that an uncontrolled effect, perhaps a microarray reading difficulty at position i , has artificially broadened the distribution of the n z_{ij} values. A drastic cure is to standardize positions as well as subjects, that is, to perform a second standardization (3) with the roles of i and j reversed. Doing so seemed to remove more signal than noise for data (1), and is not recommended. Nevertheless, a plot of robust standard deviations as a function of position i may help reveal systematic reading problems.

Formula (13), which is the basis of our iterative algorithm (18)–(21), depends on the strong assumption that $f_0(z)$ and $f_1(z)$, the null and non-null densities in the two-groups model (4), apply unchanged to each position \mathcal{C}_i . A more general result that allows the non-null density to depend on \mathcal{C}_i (while $f_0(z)$ is still assumed fixed) is developed in Efron (2009) and in Section 10 of Efron (2010b). Define

$$w_i(z) = \Pr\{\mathcal{C}_i|z\}. \quad (43)$$

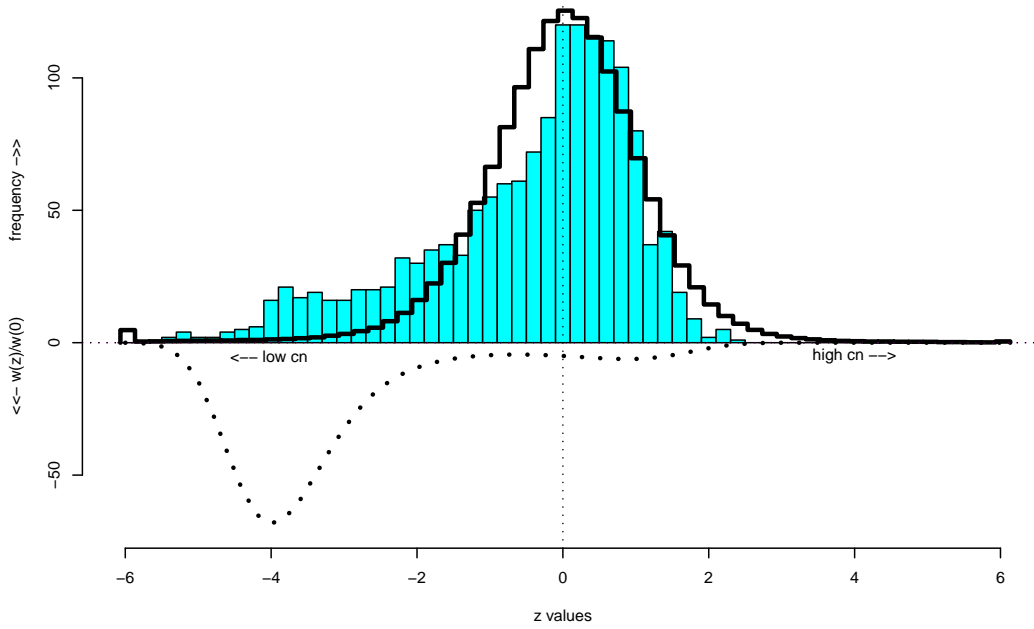


Fig. 6. Solid histogram the 1500 z_{ij} values for positions i in 1750–1759, cnv data (1); line histogram for the 748,500 other z_{ij} ; dotted curve cubic logistic regression estimate (45) for $w_i(z)/w_i(0)$ (43) (multiplied by -5 for display).

Then, to a good approximation,

$$\widehat{\text{fdr}}_i(z) \doteq \widehat{\text{fdr}}(z) \frac{w_i(0)}{w_i(z)}. \quad (44)$$

Figure 6 and Figure 7 concern the application of (44) to the amalgamated set of positions 1750 through 1759. The solid histogram in Figure 6 shows the 1500 z_{ij} at these 10 positions having an excess of negative values, compared to the distribution of all the rest. A logistic regression of the indicator

$$I_{ij} = \begin{cases} 1 & \text{if } i \in 1750 : 1759 \\ 0 & \text{otherwise} \end{cases} \quad (45)$$

as a cubic function of z_{ij} gave estimate $\hat{w}_i(z)$; the ratio $\hat{w}_i(z)/\hat{w}_i(0)$ is plotted below the horizontal axis.

Three estimates $\widehat{\text{fdr}}_i(z)$ for the local false discovery rate applying to positions 1750–1759 appear in Figure 7: the combined estimate of Figure 2, obtained from all 750,000 z_{ij} values; *Method 1*, from five iterations of algorithm (18)–(21), applied in the two-sided fashion of Section 2; and *Method 2*, from formula (44), with $w_i(z)/w_i(0)$ as shown in Figure 6.

On the left side, both Method 1 and Method 2 yield much smaller estimates than the combined curve, for instance $\widehat{\text{fdr}}_i(-3) = 0.019$ from Method 2, compared to the combined estimate 0.125. Both Methods are adjusting the combined estimates $\widehat{\text{fdr}}(z_{ij})$ downward to account for the excess cnv activity observed at positions 1750–1759.

The story is different on the right side. Method 2 has $\widehat{\text{fdr}}_i(z) = 1$ for $z > 0$, which is intuitively correct since Figure 6 shows no tendency toward large positive z -values in positions 1750–1759.

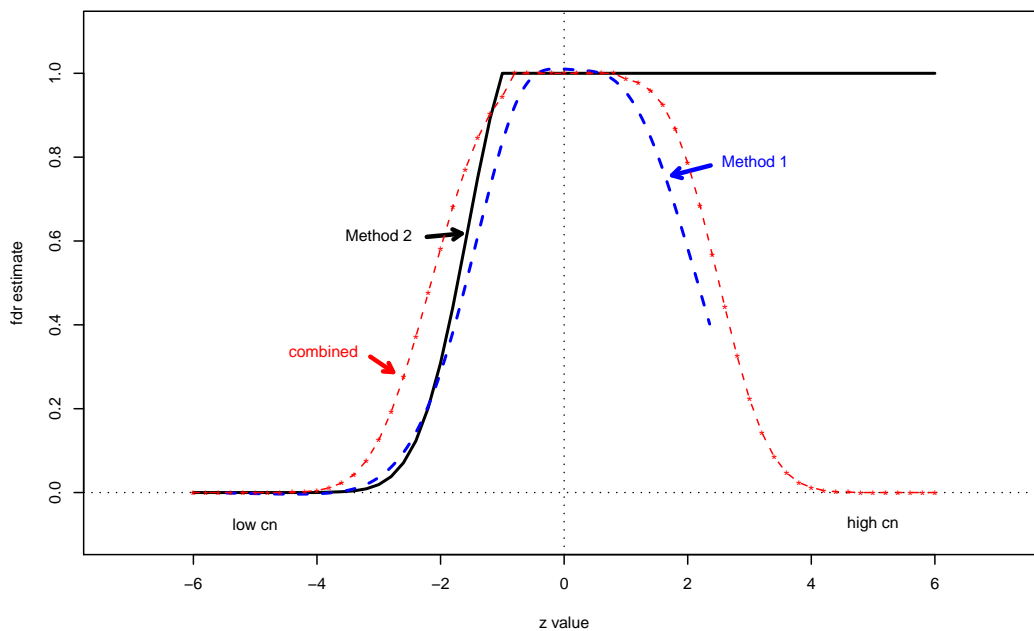


Fig. 7. Three estimates of the local false discovery rate for positions 1750–1759 of cnv data (1): *combined* from all 750,000 z_{ij} , as in Figure 2; *Method 1* from 5 iterations of (18)–(21), two-sided; *Method 2* from (44), using $\hat{w}_i(z)$ as shown in Figure 6.

Many of the other positions *do* show unusually large positive z -values, causing the combined estimate $\widehat{\text{fdr}}(z)$ to decline for large positive z . Method 1’s estimate declines even more sharply. It is using non-null density $f_1(z)$ (4) as estimated from the combined data, and does not “know” that the non-null values in positions 1750–1759 are only left-sided.

Applying Method 1 separately on the left and right, as in (42), resolves the discrepancy with Method 2. Method 2 itself tends to be noisy when applied to individual positions, and the two one-sided versions of Method 1 seem preferable in general.

6. CONVERGENCE PROPERTIES OF THE ITERATIVE ALGORITHM

Algorithm (18)–(21) was stopped after five iterations since numerical convergence of the \hat{k}_i values had nearly been achieved, producing the results pictured in Figure 3, Figure 4, and Figure 5. This section discusses the theoretical convergence point of the algorithm, leading to a formula for its standard error, and a connection with maximum likelihood estimation in model (28). The development will be in terms of $\hat{\pi}_{i1} = \hat{k}_i/n$ (19), rather than \hat{k}_i itself.

Returning to the two-groups notation of Section 2, let p_1 and $p_0 = 1 - p_1$ take values between 0 and 1, and define

$$f(z, p_1) = p_1 f_1(z) + p_0 f_0(z) \quad (46)$$

and

$$\text{tdr}(z, p_1) = \frac{p_1 f_1(z)}{f(z, p_1)} = \frac{1}{1 + \frac{p_0}{p_1} L(z)} \quad (47)$$

where $L(z)$ is the likelihood ratio

$$L(z) = f_0(z)/f_1(z). \quad (48)$$

The actual true discovery rate in class \mathcal{C}_i , $\Pr\{\text{non-null}|z, \mathcal{C}_i\}$, is

$$\text{tdr}_i(z) = \text{tdr}(z, \pi_{i1}) = 1/[1 + (\pi_{i0}/\pi_{i1})L(z)]. \quad (49)$$

(A little algebra shows that (49) equals (16).)

Finally, define

$$h_i(p_1) = p_1 - \int_{\mathcal{Z}} \text{tdr}(z, p_1) f_{(i)}(z) dz \quad (50)$$

where $f_{(i)}(z)$ equals $f(z, \pi_{i1})$, the mixture distribution (12) of z in \mathcal{C}_i , and the integral is taken over \mathcal{Z} , the sample space of z . Since

$$\int_{\mathcal{Z}} \text{tdr}(z, \pi_{i1}) f_{(i)}(z) dz = \int_{\mathcal{Z}} \frac{\pi_{i1} f_1(z)}{f_{(i)}(z)} f_{(i)}(z) dz = \pi_{i1}, \quad (51)$$

the function $h_i(p_1)$ satisfies

$$h_i(\pi_{i1}) = 0; \quad (52)$$

$h_i(\cdot)$ will turn out to determine the convergence point of algorithm (18)–(21), and also the delta-method standard error of the converged estimate.

LEMMA 2. *The derivative of $h_i(p_1)$ is*

$$h'_i(p_1) = 1 - \int_{\mathcal{Z}} \text{tdr}(z, p_1) \text{fdr}(z, p_i) f_{(i)}(z) dz / p_1 p_0 \quad (53)$$

where $\text{fdr}(z, p_1) = 1 - \text{tdr}(z, p_1)$ (47).

Proof. From (47), we calculate

$$\begin{aligned} \frac{\partial \text{tdr}(z, p_1)}{\partial p_1} &= \frac{1}{\left[1 + \left(\frac{1}{p_1} - 1\right) L(z)\right]^2} \frac{L(z)}{p_1^2} = \frac{\text{tdr}(z, p_1)^2 f_0(z)}{p_1^2 f_1(z)} \\ &= \frac{\text{tdr}(z, p_1) p_1 f_1(z) f_0(z)}{p_1^2 f(z, p_1) f_1(z)} = \frac{\text{tdr}(z, p_1) \text{fdr}(z, p_1)}{p_1 p_0} \end{aligned} \quad (54)$$

which gives (53) from (47). \square

The derivative $h'(p_1)$ takes on a convenient form at $p_1 = \pi_{i1}$ (the actual non-null probability in class \mathcal{C}_i), where $h_i(\pi_{i1}) = 0$.

LEMMA 3.

$$h'_i(\pi_{i1}) = \xi_i / \pi_{i1} \pi_{i0} \quad (55)$$

with

$$\xi_i = \int_{\mathcal{Z}} [\text{tdr}(z, \pi_{i1}) - \pi_{i1}]^2 f_{(i)}(z) dz, \quad (56)$$

the variance of $\text{tdr}(z, \pi_{i1}) = \text{tdr}_i(z)$ in \mathcal{C}_i (which is also the variance of $\text{fdr}_i(z)$ in \mathcal{C}_i).

769 *Proof.* Define I to be the null indicator for a random case in \mathcal{C}_i ,

$$770 \quad I = \begin{cases} 1 & \text{if null} \\ 0 & \text{if non-null,} \end{cases} \quad (57)$$

773 so

$$774 \quad \pi_{i1} = \Pr\{I = 0|\mathcal{C}_i\} \quad \text{and} \quad \text{tdr}_i(z) = \Pr\{I = 0|z_i, \mathcal{C}_i\}. \quad (58)$$

775 At $p_1 = \pi_{i1}$, (53) becomes

$$776 \quad h'_i(\pi_{i1}) = 1 - \int_{\mathcal{Z}} \text{tdr}_i(z) \text{fdr}_i(z) f_{(i)}(z) dz / \pi_{i1} \pi_{i0}. \quad (59)$$

777 But $\text{tdr}_i(z) \text{fdr}_i(z)$ equals $\text{var}_i\{I|z\}$, the conditional variance of the Bernoulli random variable I , so

$$778 \quad h'_i(\pi_{i1}) = 1 - E_i \{ \text{var}_i\{I|z\} \} / \pi_{i1} \pi_{i0}, \quad (60)$$

779 E_i indicating expectation with respect to $f_{(i)}(z)$.

780 A standard relationship between conditional and unconditional variances is

$$781 \quad \text{var}_i\{I\} = E_i \{ \text{var}_i\{I|z\} \} + \text{var}_i \{ E_i\{I|z\} \}, \quad (61)$$

782 var_i indicating variance with respect to $f_{(i)}(z)$. Since $E_i\{I|z\} = \text{fdr}_i(z)$, (60)–(61) imply $h'_i(\pi_{i1}) = \text{var}_i\{\text{fdr}_i(z)\} / \pi_{i1} \pi_{i0}$, which is the same as (55). \square

783 *Note.* Since $\pi_{i1} \pi_{i0} = \text{var}_i\{I\}$, we can also write (53) as

$$784 \quad h'_i(\pi_{i1}) = \text{var}_i \{ E_i\{I|z\} \} / \text{var}_i\{I\} \leq 1. \quad (62)$$

785 An empirical version of these theoretical results brings us back to algorithm (18)–(21). Let $z_i = (z_{i1}, z_{i2}, \dots, z_{in})$ be the vector of n observations for position i , and define

$$786 \quad \hat{h}_i(p_1) = p_1 - \frac{1}{n} \sum_{j=1}^n \text{tdr}(z_{ij}, p_1), \quad (63)$$

787 with $\text{tdr}(z, p_1)$ as in (47). The value $\hat{\pi}_{i1}$ having $\hat{h}_i(\hat{\pi}_{i1}) = 0$ satisfies

$$788 \quad \hat{\pi}_{i1} = \frac{1}{n} \sum_{j=1}^n \text{tdr}(z_{ij}, \hat{\pi}_{i1}) = \frac{1}{n} \sum_{j=1}^n \widehat{\text{tdr}}_i(z_{ij}), \quad (64)$$

789 showing that $\hat{\pi}_{i1}$ is the stable point of (19). A familiar estimating-equation argument provides an approximate standard error for $\hat{\pi}_{i1}$ (or equivalently for the convergent value of $\hat{k}_i = n\hat{\pi}_{i1}$).

790 **THEOREM 1.** *The standard deviation of $\hat{\pi}_{i1}$ is approximated by*

$$791 \quad \text{sd}(\hat{\pi}_{i1}) \doteq \pi_{i1} \pi_{i0} / (n \xi_i)^{1/2} \quad (65)$$

792 with ξ_i as in (56).

793 *Proof.* Only a heuristic derivation of (65) will be given here. The random variable $\text{tdr}(z, \pi_{i1})$ has mean (51) and standard deviation (56),

$$794 \quad \text{tdr}(z, \pi_{i1}) \sim (\pi_{i1}, \xi_i) \quad (66)$$

under the distribution $F_{(i)}$ corresponding to density $f_{(i)}(z)$. Therefore

$$\hat{h}_i(\pi_{i1}) = \pi_{i1} - \frac{1}{n} \sum_{j=1}^n \text{tdr}(z_{ij}, \pi_{i1}) \sim (0, \xi_i/n). \quad (67)$$

The first Newton–Raphson step to find $\hat{\pi}_{i1}$ gives

$$\hat{\pi}_{i1} - \pi_{i1} = -\hat{h}_i(\pi_{i1})/\hat{h}'_i(\pi_{i1}) \doteq -\hat{h}_i(\hat{\pi}_{i1})/h'_i(\pi_{i1}) = \frac{\pi_{i1}\pi_{i0}}{\xi_i} \hat{h}_i(\hat{\pi}_{i1}); \quad (68)$$

(67) and (68) yield

$$\hat{\pi}_{i1} - \pi_{i1} \sim (0, (\pi_{i1}\pi_{i0})^2/(n\xi_i)), \quad (69)$$

which gives (65).

The second step in (68) substitutes $h'_i(\pi_{i1})$ for $\hat{h}'_i(\pi_{i1})$. Using (54),

$$\hat{h}'_i(\pi_{i1}) - h'_i(\pi_{i1}) = \int_{\mathcal{Z}} \frac{\text{tdr}(z, \pi_{i1}) \text{fdr}(z, \pi_{i1})}{\pi_{i1}\pi_{i0}} d\left(F_{(i)} - \hat{F}_{(i)}\right)(z), \quad (70)$$

$F_{(i)} - \hat{F}_{(i)}$ being the difference between the true and empirical distributions in \mathcal{C}_i . Under standard conditions, this will append a factor of only $1 + O(n^{-1/2})$ to the approximation (68). \square

Several relevant points are raised by the previous discussion:

- The key assumption for algorithm (18)–(21) is that the same likelihood ratio $L(z) = f_0(z)/f_1(z)$ applies to all classes \mathcal{C}_i , which is a weaker assumption than $f_0(z)$ and $f_1(z)$ both staying the same. This follows for (47) and (64).
- The stable point $\hat{\pi}_{i1}$ (64) can be found by Newton–Raphson updating, $dp_1 = -\hat{h}_i(p_1)/\hat{h}'_i(p_1)$, (63) and (53). Theoretically, this should converge faster than the EM-type steps in (18)–(21).
- The convergence estimates $\hat{k}_i = n\hat{\pi}_i$ were nearly the same as those shown in Figure 3, for example 39.3 compared to 39.1 at position 1755.
- The standard deviation estimates for \hat{k}_i based on the empirical version of (65) were a good match to those in Figure 5 for positions having $\hat{k}_i \geq 15$. However, (65) gave quite erratic results for $\hat{k}_i < 15$, and is not recommended in general.
- The standard deviation estimate (65) equals the Cramér–Rao lower bound at $r = 1$ in parametric family (28), but not for $r \neq 1$.
- A possible competitor to $\hat{\pi}_{i1}$ would be

$$\tilde{\pi}_{i1} = \hat{\pi}_1 \hat{r}_i \quad (71)$$

where \hat{r}_i is the maximum likelihood estimate of r in (25), (28). Example 7 of Efron (1975) implies that $\tilde{\pi}_{i1}$ would be fully efficient at $r = 1$ but far more variable than Fisher information calculations suggest when r much exceeds 1.

- For our cnv example (1), the ML estimates $\tilde{\pi}_{i1}$ were a nearly perfect linear function of the converged iterative estimates $\hat{\pi}_{i1}$,

$$\tilde{\pi}_{i1} \doteq 1.06 \cdot \hat{\pi}_{i1}. \quad (72)$$

In other words, the \hat{k}_i estimates of Figure 3 nearly equal MLEs from the class-wise two-groups model (4), (11).

7. IDENTIFYING CNV-PRONE REGIONS IN TUMORS

Analysis of chromosome copy number aberrations in tumor samples is now a staple of cancer studies. A central question in this paper has been “which locations are more prone to gain or loss.” The meaning and motivation of this question in the analysis of tumor samples differs from that in the analysis of normal samples. Since tumorigenesis involves the breakdown of DNA repair and maintenance system, many chromosomal gains and losses in tumors are hypothesized to be random events that occur as a due effect of the development of the tumor. In this sense, many of the copy number changes we detect are “passenger” mutations, that, unlike “driver” mutations, do not play a functional role in driving tumor progression. For a recent review, see Stratton et al. (2009). An important goal in the analysis of tumor samples is to find the driver mutations. Since passenger mutations tend to occur more or less randomly throughout the genome, and driver mutations tend to favor certain genome positions containing functionally relevant genes, driver mutations can be identified by finding positions that are more cna-prone than “random” in a cross-sample analysis. This is the scientific problem that motivates our analysis of tumor samples.

As an example, we analyze chromosome 1 of 207 glioblastoma subjects from the Cancer Genome Atlas project (The Cancer Genome Atlas, 2008). This data is a 42,075 by 207 matrix, derived from the 42075 probes that map to chromosome 1 on the Illumina HumanHap 550k array. We applied the hypothesis testing framework of Section 3 to identify locations prone to gains and losses. The estimates \hat{k}_i are computed at each location by equation (3.9). Since gains and losses have completely different biological ramifications in tumors, we used the one-sided statistics described in Section 5. That is, in computing \hat{k}_i , we set the true discovery rate of subjects with $z_{ij} < 0$ to 0 for gains, and vice versa for losses as in (42).

The top plot of Figure 8 shows, at the chromosome level, the locations where the one-sided false discovery rates are lower than 0.05; blue for gains and red for losses. The second plot from top shows the corresponding \hat{k}_i estimates, plotted in the positive direction \hat{k}_i for gains and plotted inverted in the negative direction for losses. The horizontal dashed lines in Figure 8 are the 95% quantiles of the distribution of

$$k^{\max} = \max_{1 \leq i \leq 42075} k_i, \quad (73)$$

separately for gains and losses, estimated from 100 permutations.

The top two plots of Figure 8 show that, for both gains and losses, there are several significant spikes in the $\{\hat{k}_i\}$ profile. There is one very prominent peak for gains at around 32,000. The signals that contribute to this spike are noticeable in the dash-plot as a column of overlapping dashes. A more interesting case is presented by the less conspicuous spike at around position 9,000 for losses (highlighted in gray). In the bottom two plots of Figure 8, we zoom in to the region [8500, 9200] containing this spike. The dash-plot clearly shows that this region contains recurrent deletions in quite a few of the subjects, with \hat{k}_i peaking at the cross subject intersection of the deleted areas. The Supplementary Materials contain a closer look of this 700 marker region via a heatmap of the original $\{x_{ij}\}$ matrix.

The peak (or “valley” in the inverted plot) of the \hat{k}_i profile between markers 8800 and 8900 covers the coding regions for two genes: Fas-associated factor 1 (FAF1) and Cyclin-dependent kinase 4 inhibitor C (CDKN2C), the locations of which are marked in the bottom plot of Figure 8. Notice that the width of FAF1 coincides very well with the width of the peak. FAF1 codes for a protein that enhances FAS-induced programmed cell death (apoptosis). It is well known that cells attain uncontrolled growth in tumors by disrupting apoptosis. CDKN2C also encodes a cell growth regulator protein that prevents tumorigenesis. Thus, it is plausible that deletion of

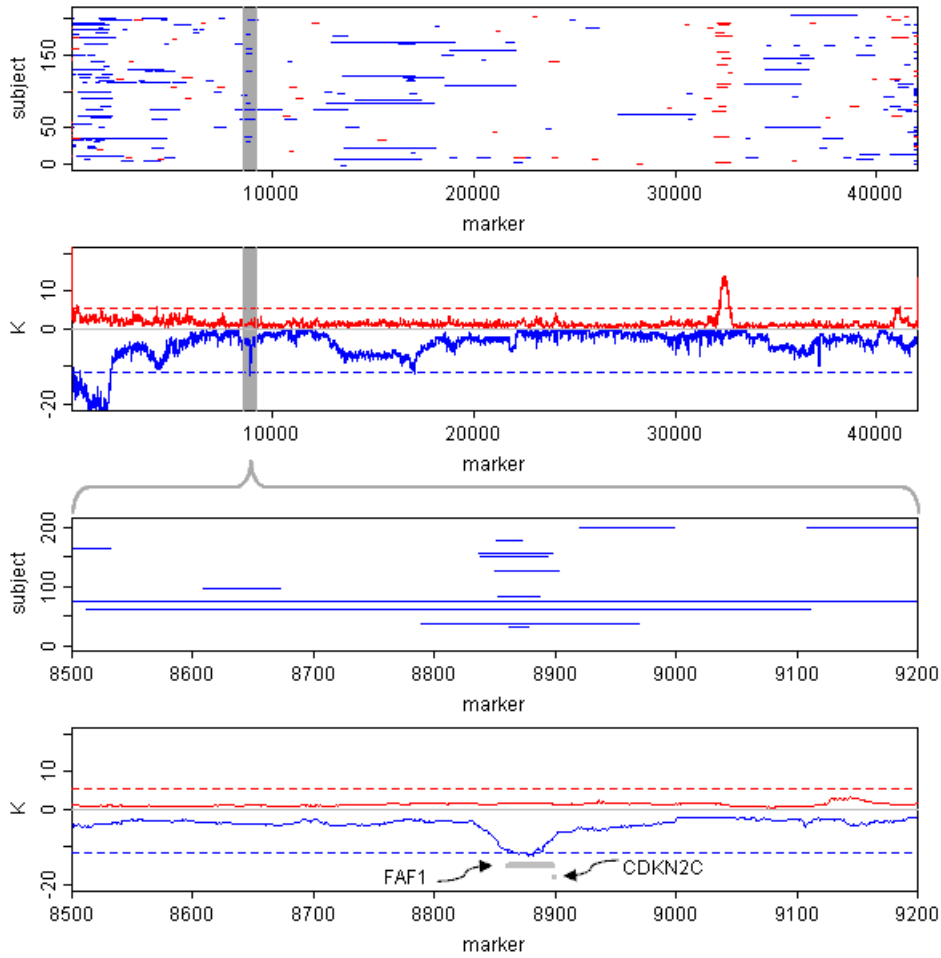


Fig. 8. Analysis of Chromosome 1 of TCGA glioblastoma data. The top two plots show the results at the chromosome level, while the bottom two plots zoom in on the region between markers 8500 and 9200. The dash-plots (first and third from the top) show the candidate gains and losses by a 0.05 threshold on the local false discovery rate. The second and fourth plots from the top show the \hat{k}_i profile for gains (blue, positive axis) and losses (red, plotted inverted on the negative axis), with the horizontal dashed line showing the 95% quantile of $\max_i \hat{k}_i$ computed by permutation. The locations of the genes FAF1 and CDKN2C are shown in the bottom plot.

the region within probes 8800-8900 (mapping to 50 Mb - 51 Mb on the p-arm of chromosome 1) disrupts apoptosis and plays a driving role in the tumorigenesis of its carriers in this cohort.

REFERENCES

BEROUKHIM, R., GETZ, G., NGHIEMPHU, L., BARRETINA, J., HSUEH, T., LINHART, D., VIVANCO, I., LEE, J. C., HUANG, J. H., ALEXANDER, S., DU, J., KAU, T., THOMAS, R. K., SHAH, K., SOTO, H., PERNER, S., PRENSNER, J., DEBIASI, R. M., DEMICHELIS, F., HATTON, C., RUBIN, M. A., GARRAWAY, L. A., NELSON, S. F.,

- 961 LIAU, L., MISCHEL, CLOUGHESY, T. F., MEYERSON, M., GOLUB, T. A., LANDER, E. S., MELLINGHOFF,
962 I. K. & SELLERS, W. R. (2007). Assessing the significance of chromosomal aberrations in cancer: Methodology
963 and application to glioma. *Proceedings of the National Academy of Sciences*, 0710052104+.
- 964 BIGNELL, G. R., HUANG, J., GRESHOCK, J., WATT, S., BUTLER, A., WEST, S., GRIGOROVA, M., JONES, K. W.,
965 WEI, W., STRATTON, M. R., FUTREAL, P. A., WEBER, B., SHAPERO, M. H. & WOOSTER, R. (2004). High-
966 resolution analysis of dna copy number using oligonucleotide microarrays. *Genome Research* **14**, 287–295.
- 967 CONRAD, D., ANDREWS, T., CARTER, N., HURLES, M., & PRITCHARD, J. (2006). A high-resolution survey of
968 deletion polymorphism in the human genome. *Nature Genetics* **38**, 75–81.
- 969 DISKIN, S. J., ECK, T., GRESHOCK, J., MOSSE, Y. P., NAYLOR, T., STOECKERT JR., C. J., WEBER, B. L., MARIS,
970 J. M. & GRANT, G. R. (2006). Stac: A method for testing the significance of dna copy number aberrations across
971 multiple array-cgh experiments. *Genome Research* **16**, 1149–1158.
- 972 EFRON, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann.*
973 *Statist.* **3**, 1189–1242. With a discussion by C. R. Rao, Don A. Pierce, D. R. Cox, D. V. Lindley, Lucien LeCam,
974 J. K. Ghosh, J. Pfanzagl, Niels Keiding, A. P. Dawid, Jim Reeds and with a reply by the author.
- 975 EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23**, 1–22.
- 976 EFRON, B. (2009). Empirical bayes estimates for large-scale prediction problems. *J. Amer. Statist. Assoc.* **104**,
977 1015–1028.
- 978 EFRON, B. (2010a). Correlated z -values and the accuracy of large-scale statistical estimates. *J. Amer. Statist. Assoc.*
979 To appear.
- 980 EFRON, B. (2010b). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, vol. I
981 of *Institute of Mathematical Statistics Monographs*. Cambridge: Cambridge University Press. To be published
982 2011.
- 983 GUTTMAN, M., MIES, C., DUDYCYZ-SULICZ, K., DISKIN, S. J., BALDWIN, D. A., STOECKERT, C. J. & GRANT,
984 G. R. (2007). Assessing the significance of conserved genomic aberrations using high resolution genomic mi-
985 croarrays. *PLoS Genetics* **3**, e143+.
- 986 LAI, W. R., JOHNSON, M. D., KUCHERLAPATI, R. & PARK, P. J. (2005). Comparative analysis of algorithms for
987 identifying amplifications and deletions in array cgh data. *Bioinformatics* **21**, 3763–3770.
- 988 MCKERNAN, K. J., PECKHAM, H. E., COSTA, G. L., MCLAUGHLIN, S. F., FU, Y., TSUNG, E. F., CLOUSER,
989 C. R., DUNCAN, C., ICHIKAWA, J. K., LEE, C. C., ZHANG, Z., RANADE, S. S., DIMALANTA, E. T., HYLAND,
990 F. C., SOKOLSKY, T. D., ZHANG, L., SHERIDAN, A., FU, H., HENDRICKSON, C. L., LI, B., KOTLER, L., STU-
991 ART, J. R., MALEK, J. A., MANNING, J. M., ANTIPOVA, A. A., PEREZ, D. S., MOORE, M. P., HAYASHIBARA,
992 K. C., LYONS, M. R., BEAUDOIN, R. E., COLEMAN, B. E., LAPTEWICZ, M. W., SANNICANDRO, A. E.,
993 RHODES, M. D., GOTTIMUKKALA, R. K., YANG, S., BAFNA, V., BASHIR, A., MACBRIDE, A., ALKAN, C.,
994 KIDD, J. M., EICHLER, E. E., REESE, M. G., DE LA VEGA, F. M. & BLANCHARD, A. P. (2009). Sequence
995 and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using
996 two-base encoding. *Genome Research* **19**, 1527–1541.
- 997 MILLS, R. E. E., LUTTIG, C. T. T., LARKINS, C. E. E., BEAUCHAMP, A., TSUI, C., PITTARD, W. S. S. &
998 DEVINE, S. E. E. (2006). An initial map of insertion and deletion (indel) variation in the human genome. *Genome*
999 *Res* **16**, 1182–1190.
- 1000 NEWTON, M., GOULD, M., REZNIKOFF, C. & HAAG, J. (1998). On the statistical analysis of allelic-loss data.
1001 *Statistics in Medicine* **17**, 1425–1445.
- 1002 NEWTON, M. & LEE, Y. (2000). Inferring the location and effect of tumor suppressor genes by instability-selection
1003 modeling of allelic-loss data. *Biometrics* **56**, 1088–1097.
- 1004 OLSHEN, A. B., VENKATRAMAN, E. S., LUCITO, R. & WIGLER, M. (2004). Circular binary segmentation for the
1005 analysis of array-based dna copy number data. *Biostatistics* **5**, 557–572.
- 1006 PEIFFER, D. A., LE, J. M., STEEMERS, F. J., CHANG, W., JENNIGES, T., GARCIA, F., HADEN, K., LI, J., SHAW,
1007 C. A., BELMONT, J., CHEUNG, S. W., SHEN, R. M., BARKER, D. L. & GUNDERSON, K. L. (2006). High-
1008 resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. *Genome*
Research **16**, 1136–1148.
- PINKEL, D., SEGRAVES, 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. & 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. & BROWN, P. (1999). Genome-wide analysis of dna copy-number changes using cdna microarrays.
Nature Genetics **23**, 41–46.
- ROBBINS, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on*
Mathematical Statistics and Probability, 1954–1955, vol. I. Berkeley and Los Angeles: University of California
Press.
- ROUVEIROL, C., STRANSKY, N., HUPÉ, P., LA ROSA, P., VIARA, E., BARILLOT, E. & RADVANYI, F. (2006).
Computation of recurrent minimal genomic alterations from array-cgh data. *Bioinformatics* **22**, 849–856.

- 1009 SIEGMUND, D., YAKIR, B. & ZHANG, N. (2010). Detecting simultaneous variant intervals in aligned sequences.
1010 *Submitted*.
- 1011 SNIJDERS, A. M., NOWAK, N., SEGRAVES, R., BLACKWOOD, S., BROWN, N., CONROY, J., HAMILTON, G.,
1012 HINDLE, A. K., HUEY, B., KIMURA, K., LAW, S., MYAMBO, K., PALMER, J., YLSTRA, B., YUE, J. P., GRAY,
1013 J. W., JAIN, A. N., PINKEL, D. & ALBERTSON, D. G. (2001). Assembly of microarrays for genome-wide
1014 measurement of dna copy number. *Nature genetics*. **29**, 263–264.
- 1015 STRATTON, M. R., CAMPBELL, P. J. & FUTREAL, P. A. (2009). The cancer genome. *Nature* **458**, 719–724.
- 1016 TAYLOR, B. S., BARRETINA, J., SOCCI, N. D., DECAROLIS, P., LADANYI, M., MEYERSON, M., SINGER, S. &
1017 SANDER, C. (2008). Functional copy-number alterations in cancer. *PLoS ONE* **3**, e3179+.
- 1018 THE CANCER GENOME ATLAS (2008). Comprehensive genomic characterization defines human glioblastoma genes
1019 and core pathways. *Nature* **455**, 1061–1068.
- 1020 WANG, P., KIM, Y., POLLACK, J., NARASIMHAN, B. & TIBSHIRANI, R. (2005). A method for calling gains and
1021 losses in array-cgh data. *Biostatistics* **6**, 45–58.
- 1022 WILLENBROCK, H. & FRIDLAND, J. (2005). A comparison study: applying segmentation to arraycgh data for
1023 downstream analyses. *Bioinformatics* **21**, 4084–4091.
- 1024 ZHANG, N. (2010). Dna copy number profiling in normal and tumor genomes. In *Probability and Statistics and
1025 Their Applications to Biology*, W. Fu, ed. Springer-Verlag.
- 1026 ZHANG, N. & SIEGMUND, D. (2007). A modified bayes information criterion with applications to the analysis of
1027 comparative genomic hybridization data. *Biometrics*.
- 1028 ZHANG, N. & SIEGMUND, D. (2010). The bic criterion for detection of simultaneous change-points in aligned
1029 sequences when the numbers of change-points and sequences are large. *manuscript in preparation*.
- 1030 ZHANG, N., SIEGMUND, D., JI, H. & LI, J. Z. (2010). Detecting simultaneous change- points in multiple sequences.
1031 *Biometrika in press*.
- 1032
- 1033
- 1034
- 1035
- 1036
- 1037
- 1038
- 1039
- 1040
- 1041
- 1042
- 1043
- 1044
- 1045
- 1046
- 1047
- 1048
- 1049
- 1050
- 1051
- 1052
- 1053
- 1054
- 1055
- 1056