Detection of Correlated Breakpoints in Cancer

Brendan Hickey

May 18, 2009

1 Background

During oncogenesis the loss of caretaker genes leads to chromosomal instability: chromosomes mispair during anaphase, and may undergo non-homologous recombination, or break during telophase. This instability leads to changes in the copy number of genes, changes in their expression patterns, and can construct novel genes through fusion or truncation. Array comparative genome hybridization (aCGH) [6] is a high-throughput technique used to assess these copy-number variants.

aCGH assesses these deviations from the normal genome through the use of short probing sequences that are represented uniquely in the human genome. These probes are hybridized to fluorescently tagged experimental genomic DNA and fluorescently tagged reference DNA. The score for each probe is given as $\log_2(E/R)$, where E is the luminosity of the experimental sample, and R is the luminosity of the reference. The $\log_2$ ratio of the luminosities of the sample and reference is a noisy estimate of the relative copy number of genomic material in the experimental sample as compared to the reference. Probes hybridized to regions of normal copy number have a $\log_2$ ratio close to zero, while gains are positive and areas of loss are negative.

While there is no computational impediment to attempting to correlate the measurements at individual aCGH probes there is a statistical barrier. The number of possible probe pairings is quadratic in the number of probes, and the number of possible combinations is exponential in the number of probes. Modern arrays contain upwards of several million probes. Due to multiple hypothesis testing, an attempt to naïvely test the correlation between all pairs in a $244,000$ probe array would necessitate correcting each p-value by a factor of $5.95 * 10^{10}$. To reduce the the number of hypotheses tested we rely on the notion of breakpoints.

A breakpoint is a point on a chromosome that is the terminus of some gross structural change. Intervals are a related structure which are

characterized by a region of the genome that is contiguous in the reference, has a uniform copy number throughout, and is flanked by breakpoints, the centromere or chromosome ends. Through aCGH data we can identify those breakpoints and the corresponding intervals, that are caused by events that alter copy number, deletions, duplications, and some transpositions. Given aCGH data, segmentation algorithms identify breakpoints and the corresponding intervals.

Since we expect the number of breakpoints and intervals to increase sublinearly[1] in the size of the array, an increase in array resolution will not exacerbate the problem of multiple hypothesis testing. Instead of searching for correlations between individuals probes, we can find correlated breakpoints and intervals.

Earlier work by Ben-Dor, *et al.* [1] and Diskin, *et al.* [2] focused on finding recurrent deletions and duplications across multiple array experiments. This work was motivated by biological knowledge of oncogenes and tumor suppressors. When oncogenes are present in higher than normal copy number they promote oncogenesis. Likewise the homozygous loss of a tumor suppressor gene also promotes oncogenesis. The identification of recurrent abberations may generate novel prognostic markers, such as BRCA1 and BRCA2 in breast cancer. We have previously developed asymptotic improvements for the algorithm given in [2]. Rather than adopt an interval-centric we attempt to infer the identity and presence of fusion genes in cancer. By means of gross chromosomal rearrangement (via deletion, duplication, inversion, or transposition), two genes, their regulatory elements, or some fraction thereof, are brought together to generate a transcribed gene. The BCR-ABL fusion is an oncogene present in many cases of chronic myelogenous leukemia. Because BCR-ABL and other fusion genes are only present in cancerous cells and they represent potential drug targets; a compound that preferentially kills cells expressing a fusion transcript will selectively target cancerous cells. The drug imatinib (Gleevec) was designed to target BCR-ABL's protein product binding it and inhibiting cellular proliferation. The TMPRSS-ERG fusion is occurs in prostate cancer cases and results in the androgen mediated overexpression of ERG. [9] This fusion is the result of a deletion on chromosome 21 that joins the genes. Fluorescence *in situ* hybridization (FISH), which uses fluorescence microscopy to localize probes to specific regions of the genome,

[1] As the resolution of arrays increases, our ability to detect finer copy number changes and polymorphisms increases. We expect the number of somatic, cancer-related breakpoints to be independent of the resolution of the array.

can be used to detect TMPRSS-ERG fusions. However FISH is labor-intensive relative to array experiments. Using an existing algorithm, we were able to identify the TMPRSS-ERG fusion as a pair of correlated breakpoints in 7 members of a cohort of 36 prostate cancer patients. Our analysis has revealed cases in which this approach fails to detect the TMPRSS-ERG fusion, thereby motivating the development of a new algorithm.

2 Segmentation Algorithms

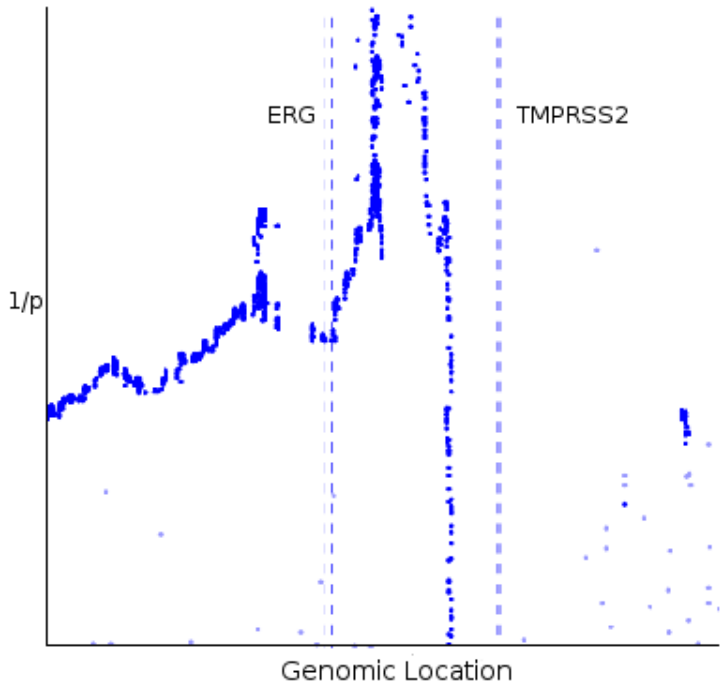

Figure 1: View of CBS segmentation of Sample 22 Chromosome 21. Here CBS conflates the breakpoints associated with TMPRSS2 and ERG into a single breakpoint falling between, rather than incident to, the genes. The y-axis is the reciprocal of the p-value reported by the T-test used in CBS.

Circular Binary Segmentation (CBS) of [5] is a widely used segmentation algorithm for assessing DNA copy number. CBS adopts a greedy approach to segmentation. For a set of probes it uses a single-tailed T-test to identify

the location of the most likely breakpoint. The set is then partitioned at this probe, and CBS recurs on each half, returning the set of partitions as breakpoints. CBS can therefore arrive at a locally optimal segmentation that is a poor approximation of the globally optimal segmentation. We modified the Matlab implementation CBS to report the result of all T-tests for each putative breakpoints tested in its recursive descent rather than just the most likely breakpoints. This modification allows us to observe the choices made by CBS throughout segmentation. In Figure 1, CBS has segmented the 21^{st} chromosome of a patient with prostate cancer and a putative TMPRSS-ERG fusion. CBS greedily places the breakpoint between ERG and TMPRSS2 rather than selecting a locally suboptimal segmentation that places breakpoints near gene. This exemplifies the undesirable consequences of CBS's greedy nature.

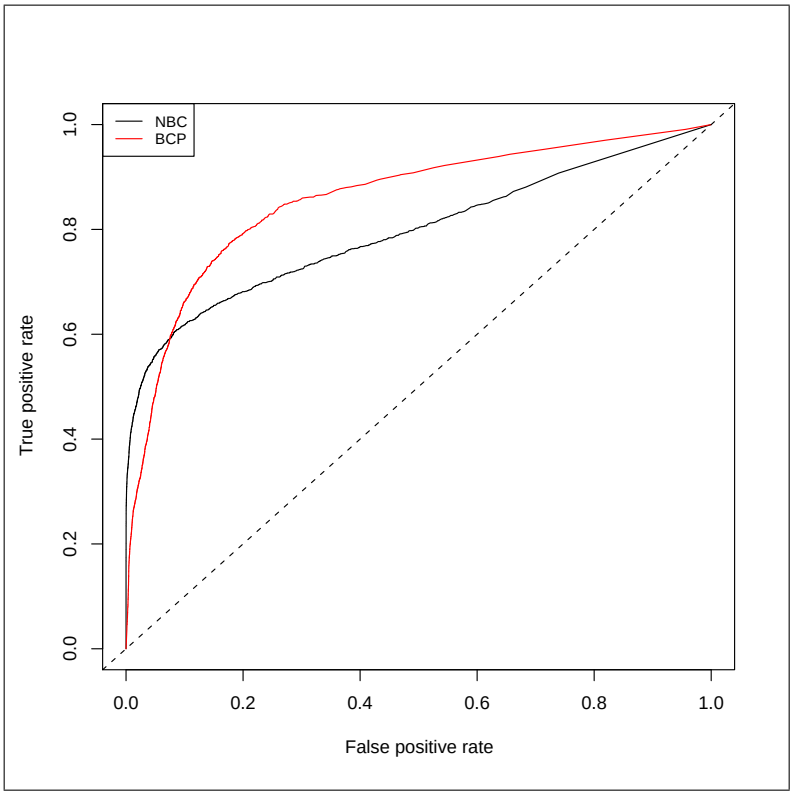

Figure 2: ROC curve for NBC and BCP. Window size = 1.

The Bayesian Change Point (BCP) algorithm of [3] uses an Markov

chain Monte Carlo approach to segment aCGH data. Unlike CBS which returns a single binary segmentation–assigning 1 to the locations that it has identified as breakpoints and 0 to all others–BCP instead computes the probability of a breakpoint occurring at each probe in the array. On real data BCP performs poorly, determining that single-probe aberrations, either the result of experimental error or heritable polymorphisms, are significant breakpoints. NBC, given in [7], attempts to remedy the faults of BCP by extending the work of Liu & Lawrence [4].

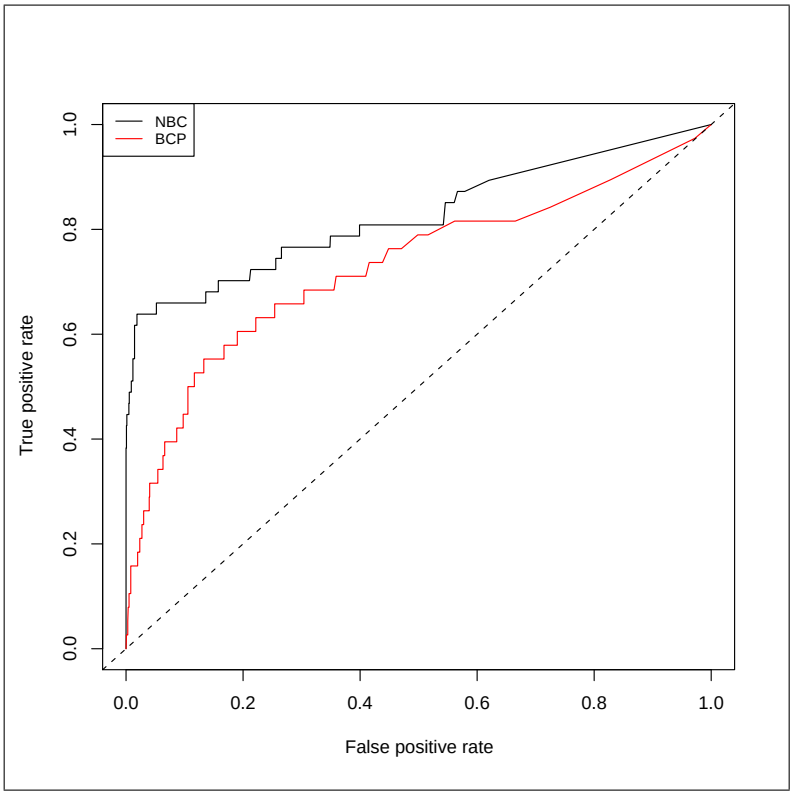

Figure 3: ROC curve for NBC and BCP. Window size = 1. Simulated data with high-amplitude single probe abberations.

Synthetic data comprising forty synthetic chromosomes each containing 5000 probes was prepared using the simulator described in this report. The simulated data was processed with BCP, NBC and CBS. For a given p-value threshhold the probability of a breakpoint, as computed by BCP and NBC, was compared with the ground truth from the simulator. Each non-

overlapping window [2] is classified according to its agreement with the ground truth. Both BCP and NBC proved to be insensitive to variations in window size (*data not shown*). We used ROCR [8] to plot the sensitivity versus specificity of BCP and NBC (Shown in Figure 2), varying the probability threshhold for the identification of a breakpoint. Sensitivity is:

$$\frac{TruePositives}{TruePositives + FalseNegatives}$$

Specificity is:

$$\frac{TrueNegatives}{TrueNegatives + FalsePositives}$$

The simulator generates $G = (g_1, g_2, \ldots, g_n)$, where $g_i = 1$ if there is a breakpoint at the i^{th} probe in the ground truth, and $g_i = 0$ otherwise. A segmentation algorithm, like NBC or BCP produces $D = (d_1, d_2, \ldots, d_n)$, where $d_i =$ probability(There is a breakpoint at the i^{th} probe)[3] p is a probability threshhold that is varied to generate the plot.

$$TruePositives = \Sigma_i TP_i$$

$$TP_i = \begin{cases} g_i & \text{if } d_i \geq p, \\ 0 & \text{if } d_i < p. \end{cases}$$

$$FalseNegatives = \Sigma_i FN_i$$

$$FN_i = \begin{cases} 0 & \text{if } d_i \geq p, \\ g_i & \text{if } d_i < p. \end{cases}$$

$$TrueNegatives = \Sigma_i TN_i$$

$$TN_i = \begin{cases} 0 & \text{if } d_i \geq p, \\ 1 - g_i & \text{if } d_i < p. \end{cases}$$

$$FalsePositives = \Sigma_i FP_i$$

[2] Note that for $n = 1$, each probe is classified in isolation.

[3] In the case of CBS and other segmentation algorithms that produce a single segmentation, these probabilities are binary.

$$FP_i = \begin{cases} 1 - g_i & \text{if } d_i \geq p, \\ 0 & \text{if } d_i < p. \end{cases}$$

Unlike real data the simulated corpus in Figure 2 did not contain high-amplitude single probe noise. This led to a reduction in BCP's false positive rate versus its performance on real data. When this type of noise was introduced, BCP's specificity fell while NBC is demonstrably robust to this type of noise. (Shown in Figure 3). In practice NBC's robustness is desirable because the number of hypotheses tested when searching for correlated breakpoints is quadratic in the number of breakpoints–the admission of false positives can cause otherwise significantly correlated breakpoints to be rendered insignificant.

3 Simulated Data

We built simulators in R and Mathematica to provide data for the assessment of our techniques.

The aCGH data simulator of [10] generates synthetic data by sampling a corpus of segmented data. It then introduces gaussian noise with a fixed per sample variance ($\mu = 1, \sigma = 0.1$) to simulate systematic per array errors. We extended this simulator to support the introduction of correlated breakpoints and modified the underlying noise model.

BCP [3] and NBC [7] assume a gaussian noise model. Any comparison between these three methods using synthetic data generated with gaussian noise is prejudiced against CBS. We therefore introduced a random walk into the sample variance to alleviate this bias. Each interval is generated with a fixed sample variance and a per interval variance that is subject to a uniformly distributed random walk (interval $[-0.01, 0.01]$). The simulator can also introduce high-amplitude single probe aberrations at a tunable frequency (Single probe aberrations were added to the data used to generate Figure 3 at a rate of 1%.) Introducing uniform random noise on top of the gaussian noise was also attempted, but did not yield robust results.

The option to generate correlated breakpoints was implemented to facilitate an assessment of the accuracy of detecting this sort of aberration across multiple samples. These aberrations are parameterized over absolute frequency, relative frequency of each member of the pair, amplitude, and position, subject to random variation. The variation in position allows for one end of an interval to remain relatively fixed, while the other end varies freely within the chromosome. This models our intuition about the

formation of fusion transcripts: while there is a narrow physical region in which breaks will yield a fusion, the corresponding breaks produced by these duplications or deletions are not similarly constrained.

The corpus used for generating synthetic data was taken from a CBS segmentation of a set of 257 aCGH experiments carried out on glioblastoma tissue as part of the TCGA project. Short intervals ($n \leq 10$) were discarded from the corpus, as these events are not indicative of gross chromosomal instability and instead seem to be artifacts of CBS. Long intervals ($n \leq 1000$) were discarded for computational efficiency.

4 Correlated Breakpoints

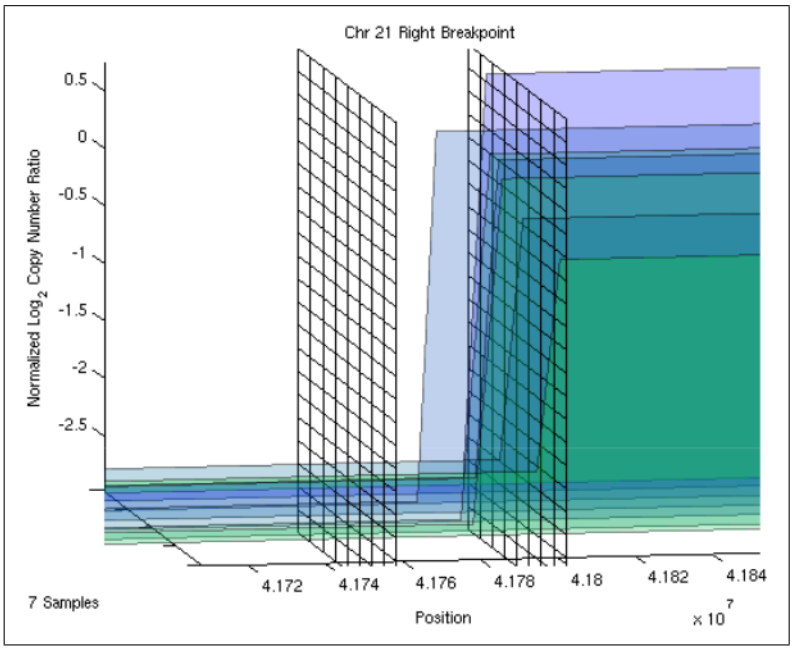

Figure 4: The change in $\log_2$ ratios for those patients determined to have a break in TMPRSS2. The grid indicates the gene boundaries. [7]

We have developed two methods for identifying correlated breakpoints. The first approach utilizes a sliding window. For a fixed genomic width, we assign a score to each window equal to the number of samples with at least one breakpoint in each window across the genome. Next we remove

all windows that are overlapped by a window with greater score. For the remaining windows we apply Fisher's exact test to all pairs, using the Bonferroni correction for multiple hypothesis testing. Using this method, a window of size $350,000$ base pairs and a segmentation produced by CBS, we identified TMPRSS-ERG fusions in 7 out of 36 patients ($p \leq 5.799^{-6}$) in our prostate cancer cohort.

In the second approach we identify gene windows that are incident to a statistically surprising ($p < 0.05$ after multiple hypothesis correction) number of breakpoints across all patients, discounting windows proportionally to their size. We expect the probability of a breakpoint appearing in a window by chance is proportional to the size of the window. We therefore correct the p-value for each window by dividing it by the length of the window. Given these gene windows, we then assess the joint probability of each pair of windows. This method in conjunction with the posterior breakpoint probability generated by NBC identified an eighth TMPRSS-ERG fusion.

5 Exon Expression

We attempted to corroborate our correlated breakpoints by analyzing gene expression data. Ideally if a breakpoint falls within a gene, we should be able to observe a change in expression level across the gene. To do this, we developed a U-statistic based on the rank order of each exon probeset. For each sample i in G, we assign a rank $r_i[e]$ to each probeset, e, $G_i[e]$ equal to its rank order amongst all $G_j[e]$. Where S is the set of samples determined to contain a breakpoint in the gene of interest. We find $R = (r_{e_1}, r_{e_2}, \ldots, r_{e_n})$ where $r_{e_j} = \Sigma_{i \in S} r_i[e_j]$ and perform a T-test on an appropriately normalized transformation of R.

We developed simulated data to validate this approach. Given a set of real data, we relabeled each data point such that the rank of every member of S was strictly less than every member not in S for a contiguous run on one side of the transcript. This technique identified these simulated changes in expression. It did not corroborate the correlated breakpoints identified by our earlier method. Since the U-statistic is agnostic to the amplitude of expression, modest fluctuations in expression level may correspond to large changes in rank and significance. To validate the behavior of the statistic, we normalized each exon probe on a per patient basis ($\mu = 0, \sigma = 1$) and plotted the mean and variance for all members of S. In the null hypothesis, S is a random subset of patients, we expect to find $\mu = 0$ and $\sigma = 1$. Figure

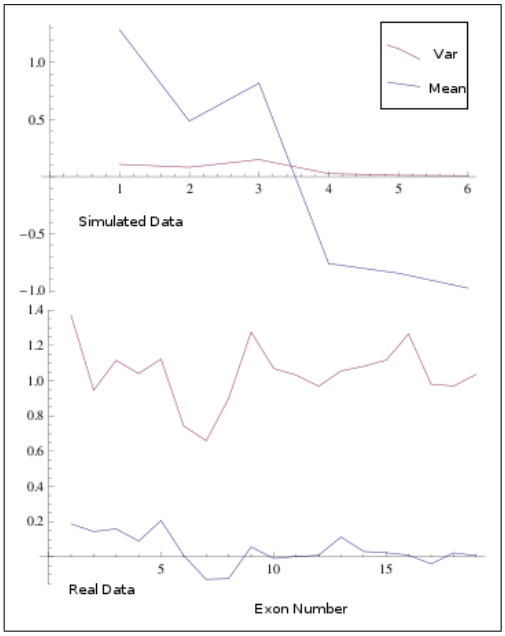

Figure 5: Mean/variance plots for U-statistic performance on real and synthetic data.

5 provides a representative example of the U-statistics behavior on simulated and real data. On simulated data, σ remains close to 0, while μ trends away from 0. Given real data the U-statistic consistently finds $\mu \approx 0$ and $\sigma \approx 1$, and we are therefore unable to reject the null hypothesis.

6 Future Work

With the density of arrays growing faster than computational gains, the adoption of algorithms used to process this data will soon be limited by algorithmic complexity. The dynamic programming step of NBC is quadratic in both space and time. Despite the advantages provided by a Bayesian approach, the complexity of NBC may impede its adoption. We observe that number of somatic breakpoints is independent of the size of an array. This suggests that sampling around putative breakpoints may be useful in reducing the algorithm's computational complexity, while capturing the somatic breakpoints that are relevant to the study of cancer.

7 Acknowledgements

We thank Anna Ritz, Ben Raphael and Hsin-Ta Wu for their collaboration, Chip Lawrence, Daniel Klein and Suzanne Sindi for technical discussions, Collin Collins for providing invaluable data, and Joe Gray for posing motivating questions.

References

[1] Amir Ben-Dor, Doron Lipson, Anya Tsalenko, Mark Reimers, Lars O. Baumbusch, Michael T. Barrett, John N. Weinstein, Anne-Lise Borresen-Dale, and Zohar Yakhini. Framework for identifying common aberrations in dna copy number data. *RECOMB 2007*, LNBI(4453):122–136, 2007.

[2] S. J. Diskin, T. Eck, J. Greshock, Y. P. Mosse, T. Naylor, C. J. Stoeckert, B. L. Weber, J. M. Maris, and G. R. Grant. STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res.*, 16:1149–1158, Sep 2006.

[3] C. Erdman and J. W. Emerson. A fast Bayesian change point analysis for the segmentation of microarray data. *Bioinformatics*, 24:2143–2148, Oct 2008.

[4] J. S. Liu and C. E. Lawrence. Bayesian inference on biopolymer models. *Bioinformatics*, 15:38–52, Jan 1999.

[5] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004.

[6] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*, 20(2):207–11, 1998.

[7] A. Ritz, H. Wu, B. Hickey, and B. Raphael. Detection of recurrent rearrangments in cancer genomes from array copy number and expression data. 2009.

[8] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):3940–3941, October 2005.

[9] S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X. W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J. E. Montie, R. B. Shah, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan. Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. *Science*, 310(5748):644–8, 2005.

[10] Hanni Willenbrock and Jane Fridlyand. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, 21(22):4084–4091, 2005.

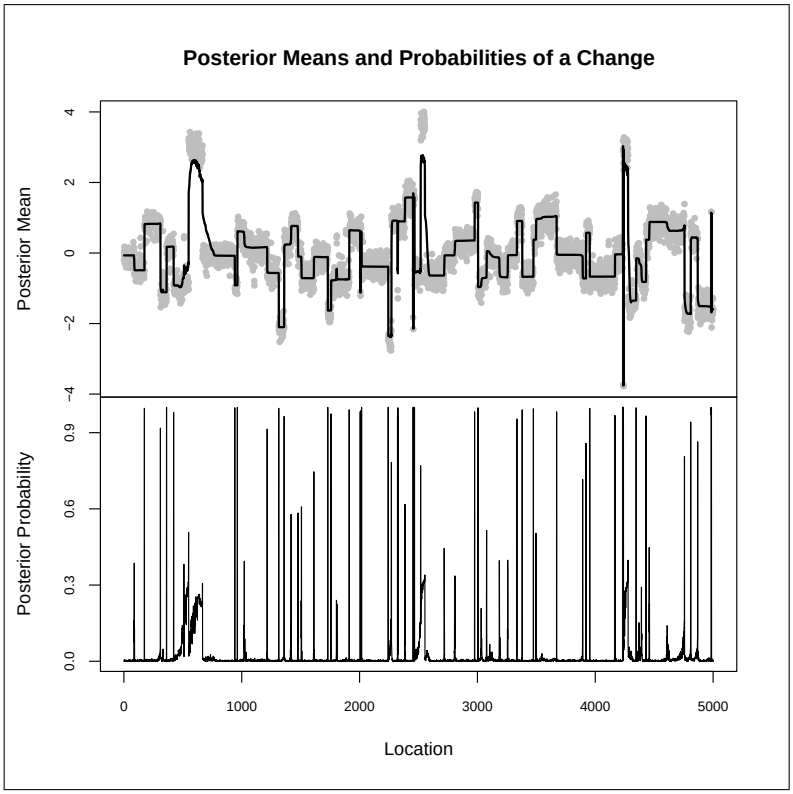

Figure 6: Representative simulated data with high-amplitude single probe abberations added. The top figure shows the raw aCGH data in gray with the interval means, as determined by BCP, in black. The bottom gives BCP's determination of the probability of a breakpoint for each probe in the input data.