

RECONSTRUCTING DNA COPY NUMBER
BY JOINT SEGMENTATION OF MULTIPLE SEQUENCES

By

Zhongyang Zhang
Kenneth Lange
Chiara Sabatti

Technical Report 261
March 2012

Division of Biostatistics
STANFORD UNIVERSITY
Stanford, California



RECONSTRUCTING DNA COPY NUMBER
BY JOINT SEGMENTATION OF MULTIPLE SEQUENCES

By

Zhongyang Zhang
Department of Statistics, UCLA

Kenneth Lange
Departments of Biomathematics,
Human Genetics and Statistics, UCLA

Chiara Sabatti
Departments of Health Research and Policy
and Statistics, Stanford University

Technical Report 261
March 2012

**This research was supported in part by
National Institutes of Health grants
GM053275 and MH075007.**

**Division of Biostatistics
STANFORD UNIVERSITY
Sequoia Hall, 390 Serra Mall
Stanford, CA 94305-4065**

<http://statistics.stanford.edu>

Reconstructing DNA Copy Number by Joint Segmentation of Multiple Sequences

Zhongyang Zhang, Kenneth Lange, and Chiara Sabatti

March 2012

Abstract

Variations in DNA copy number carries information on the modalities of genome evolution and misregulation of DNA replication in cancer cells; their study can be helpful to localize tumor suppressor genes, distinguish different populations of cancerous cell, as well identify genomic variations responsible for disease phenotypes. A number of different high throughput technologies can be used to identify copy number variable sites, and the literature documents multiple effective algorithms. We focus here on the specific problem of detecting regions where variation in copy number is relatively common in the sample at hand: this encompasses the cases of copy number polymorphisms, related samples, technical replicates, and cancerous sub-populations from the same individual. We present an algorithm based on regularization approaches with significant computational advantages and competitive accuracy. We illustrate its applicability with simulated and real data sets.

1 Introduction

Duplication and deletion of genomic materials are common in cancer cells and known to play a role in the establishment of the tumor status [24]. As our ability to survey the fine scale of the human genome has increased, it has become apparent that normal cells can also harbor a number of variations in copy number [18, 36]. The last few years have witnessed a steady increase in our knowledge of size and frequency of these variants [10, 19, 23, 30] and their implications in complex diseases [29, 40].

At the same time, statistical methods and algorithms have been developed to better harness the information available. At the cost of oversimplification, two different approaches have become particularly popular: one is based on the hidden Markov model (HMM) machinery and explicitly aims to reconstruct the unobservable discrete DNA copy number; the other, which we will generically call “segmentation”, aims at identifying portions of the genome that have constant copy number, without specifically reconstructing it. The HMM approach takes advantage of the implicitly discrete nature of the copy number process (both when a finite number of states is assumed and when, as in some implementations, less parametric approaches are adopted); furthermore, by careful modeling of the emission probabilities, one can fully utilize the information derived from the experimental results. In the case of genotyping arrays, for example, both the quantification of total DNA amount and relative allelic abundance as well as prior information (for example, minor allele frequencies) can be considered. No a-priori knowledge of the number of copy number states is required the segmentation approach—an advantage in the study of cancer where polyploids and contamination with normal tissues result in a wide range of fractional copy numbers. Possibly for the reasons outlined, HMMs are the methods of choice in the analysis of normal samples [9, 34, 41, 47, 49], while segmentation methods are the standard in cancer studies [26, 53]. A limitation of segmentation methods is that they rely on the data in which the variation in copy number is reflected in the differences in means of the segments—which make them applicable di-

rectly to a substantial portion of the data derived from recent technologies, but not to relative allelic abundance (see the modification suggested in [39] and following description for an exception).

While a number of successful approaches have been derived along the lines described above, there is still a paucity of methodology for the joint analysis of multiple sequences. It is clear that if multiple subjects share the same variation in copy number, there exists the potential to increase power by joint analysis. Wang et al. (2009) [45] presented a methodology that extended [24] to reconstruct the location of tumor suppressor genes from the identification of regions lost in a larger number of samples; the initial steps of the Birdsuite algorithm rely on the identification of suspect signals in the context of multiple samples; PennCNV [47] includes an option of joint analysis of trios; methodology to process multiple samples with the context of change point analysis has been developed in [37, 51, 53]; Efron and Zhang (2011) [14] consider FDR analysis of independent samples to identify copy number polymorphisms (CNPs); and Nowak et al. (2011) [25] use a latent feature model to capture, in joint analysis of array-CGH data from multiple tumor samples, shared copy number profiles, on each of which a fused-lasso penalty is enforced for sparsity. In the present work we consider a setting similar to [53] in that we want joint analysis to inform the segmentation of multiple samples. Our main focus is the analysis of genotyping array data, but the methodology we develop is applicable to a variety of platforms. By adopting a flexible framework we are able, for example, to define a segmentation algorithm that uses all information from Illumina genotyping data. As in [37], we are interested in the situation when not all the samples under consideration carry a copy number variant (CNV): we rather want to enforce a certain sparsity in the vector that identifies which samples carry a given variant. We tackle this problem using a penalized estimation approach, originally proposed in this context by [42], on which we have developed an algorithmic implementation before [54]. Appreciable results are achieved in terms of speed, accuracy and flexibility. In concluding this introduction, we would like to make an important qualification: the focus of our contribution is on segmentation methods, knowing that this is only one of the steps necessary for an effective recovery of CNVs. In particular, normalization and transformation of

the signal from experimental sources are crucial and can have a very substantial impact on final results: we refer the reader to [1, 2, 7, 12, 33, 35], for example. Furthermore, calling procedures that further classify results of segmentation while possibly controlling global error measures [14] are also needed. Indeed, in the data analysis included in this paper we need to resort to both these additional steps and we will describe briefly the fairly standard choices we are making.

The rest of the paper is organized as follows: Section 2 motivates the need for joint analysis of multiple signals and presents the penalized estimation framework. Section 3 describes how the model can be used for data analysis by (a) outlining an efficient estimation algorithm, (b) generalizing it to the case of uncoordinated data, and (c) describing the choice of the penalization parameters. Section 4 illustrates our results on two simulated data sets (descriptive of normal and tumor samples) and two real data sets: in one case multiple platforms are used to analyze the same sample and in the other case samples from related individuals benefit from joint analysis.

2 Multiple sequence segmentation

The goal of the present paper is to develop a flexible methodology for joint segmentation of multiple sequences that are presumed to carry related information on CNVs. We start by illustrating a series of contexts where the joint analysis appears to be useful.

2.1 Motivation

2.1.1 Genotyping arrays and CNV detection

Genotyping arrays have been used on hundreds of thousands of subjects and the data collected through them provides an extraordinary resource for CNV detection and the study of their frequencies in multiple populations. Typically, the raw intensity data representing hybridization strength is processed to obtain two signals: a quantification of total DNA amount (from now on $\log R$

Ratio LRR, following Illumina terminology) and a relative abundance of the two queried alleles (from now on B allele frequency, BAF). Both these signals contain information on CNV and one of the strengths of HMM models has been that they can easily process them jointly. Segmentation models like CBS have traditionally relied only on LRR. While this is a reasonable choice, it can lead to substantial loss of information, particularly in tumor cells, where poliploidity and contamination make information in LRR hard to decipher. To exploit BAF in the context of a segmentation method, a signal transformation has been suggested [39]: mirrored BAF (mBAF) relies on exchangeability of the two alleles and the low information content of homozygous SNPs. The resulting mBAF is defined on a coarser grid than the original BAF, but is characterized by changing means in presence of CNV. While [39] shows that its analysis alone can be advantageous and more powerful than segmentation of LRR in some contexts, clearly a joint analysis of LRR and mBAF should be preferable to an arbitrary selection of one or the other signal.

2.1.2 Multiple platforms

LRR and BAF are just one example of the multiple signals that one can have available for the same sample. Often, as research progresses, the samples are assessed with a variety of technologies. For example, a number of subjects that have been genotyped at high resolution are now being resequenced. Whenever the technology adopted generates a signal that contains some information on copy number, there is an incentive to analyze the available signals jointly.

2.1.3 Tumor samples from the same patient obtained at different sites or different progression stages

In an effort to identify mutations that are driving a specific tumor, as well as study its response to treatment, researchers might want to study CNVs in cells obtained at different tumor sites or at different time points [27]. Copy number is highly dynamic in cancer cells, so that it is to be

expected that some differences be detected over time or across sites. In contrast, the presence of the same CNVs across these samples, can be taken as an indication that the tumors share the same origin: therefore a comparative analysis of CNV can be used to distinguish resurgence of the same cancer from insurgence of a new one, or to identify specific cancer cell populations. Given that the tissue extracted always consists of a mixture of normal and cancer cells, which are in turn a mixture of different populations, joint analysis of the signals from the varied materials is much more likely to lead to the identification of common CNVs, when these exist.

2.1.4 Related subjects

Family data is crucial in genetic investigations and hence it is common to analyze related subjects. When studying individuals from the same pedigree, it is reasonable to assume that some CNVs might be segregating in multiple people: joint analysis would reduce Mendelian errors and increase power of detection.

2.2 A model for joint analysis of multiple signals

Assume we have observed M signals, each measured at N locations, corresponding to ordered physical positions along the genome, with y_{ij} being the observed value of sequence i at location j . The copy number process can be modeled as

$$y_{ij} = \beta_{ij} + \epsilon_{ij}, \quad (1)$$

where ϵ_{ij} represent noise, and the mean values β_{ij} are piece-wise constant: there exists a linearly ordered partition $\{R_1^{(i)}, R_2^{(i)}, \dots, R_{K_i}^{(i)}\}$ of the location index $\{1, 2, \dots, N\}$ such that $\beta_{is} = \dots = \beta_{it} = \mu_k^{(i)}$ for $s, \dots, t \in R_k^{(i)}$ and $1 \leq k \leq K_i$. In other words, most of the increments $|\beta_{ij} - \beta_{i,j-1}|$ are assumed to be zero. When two sequences k and l share a CNV with the same boundaries at location j , both $|\beta_{kj} - \beta_{k,j-1}|$ and $|\beta_{lj} - \beta_{l,j-1}|$ will be different from zero in correspondence of the change point. Modulo an appropriate signal normalization, $\beta_{ij} = 0$ can be interpreted

as corresponding to the appropriate normal copy number equal to 2. We propose to reconstruct the mean values β by minimizing the following function, called hereafter generalized fused lasso (GFL):

$$f(\beta) = \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^N (y_{ij} - \beta_{ij})^2 + \lambda_1 \sum_{i=1}^M \sum_{j=1}^N |\beta_{ij}| + \lambda_2 \sum_{i=1}^M \sum_{j=2}^N |\beta_{ij} - \beta_{i,j-1}| + \lambda_3 \sum_{j=2}^N \left[\sum_{i=1}^M (\beta_{ij} - \beta_{i,j-1})^2 \right]^{\frac{1}{2}}, \quad (2)$$

which includes a goodness-of-fit term and three penalties, whose roles we will explain one at the time. The ℓ_1 penalty $\sum_{i=1}^M \sum_{j=1}^N |\beta_{ij}|$ enforces sparsity within β , in favor of values $\beta_{ij} = 0$, corresponding to the normal copy number. The total variation penalty $\sum_{j=2}^N |\beta_{ij} - \beta_{i,j-1}|$ minimizes the number of jumps in the piece-wise constant means of each sequence and was introduced by [42] in the context of CNV reconstruction from array-CGH data. Finally, the Euclidean penalty on the column vector of jumps $\sqrt{\sum_{i=1}^M (\beta_{ij} - \beta_{i,j-1})^2}$ is a form of the group penalty introduced by [50] and favors common jumps across sequences. As clearly explained in [56], “the local penalty around 0 for each member of a group relaxes as soon as the $|\beta_{ij} - \beta_{i,j-1}|$ for one member i of the group moves off 0.” Bleakley and Vert (2011) [4] also suggested the use of this group-fused-lasso penalty to reconstruct CNV. We here consider the use of both the total variation and the Euclidean penalty on the jumps to achieve the equivalent effect of the sparse group lasso, which, as pointed out in [16], favors CNV detection in multiple samples, allowing for sparsity in the vector indicating which subjects are carriers of the variant. This property is important in situations as presented in Section 2.1.3 and 2.1.4, where one does not want to assume that all the M sequences carry the same CNV.

The incorporation of the latter two penalties can also be naturally interpreted in view of image denoising. To restore an image disturbed by random noise while preserving sharp edges of items in the image, a 2-D total variation penalty $\lambda \sum_{i=1}^M \sum_{j=2}^N |\beta_{ij} - \beta_{i,j-1}| + \rho \sum_{j=1}^N \sum_{i=2}^M |\beta_{ij} - \beta_{i-1,j}|$ is proposed in a regularized least-square optimization [32], where β_{ij} is the true underlying intensity

of pixel (i, j) . In CNV detection problems, signals from multiple sequences can be aligned up in shape of an image, except that pixels in each sequence are linearly ordered while sequences as a group have no certain order a priori; thus one of the two total variation penalties is replaced by the group penalty on the column vector of jumps.

Using matrix notation, and allowing the tuning parameter λ_1 , λ_2 and λ_3 to be sequence specific, we can reformulate the objective function as follows. Let $\mathbf{Y} = (y_{ij})_{M \times N}$ and $\boldsymbol{\beta} = (\beta_{ij})_{M \times N}$. Let $\boldsymbol{\beta}_i$ be the i th row of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{(j)}$ the j th column of $\boldsymbol{\beta}$. Also, let $\boldsymbol{\lambda}_3 = (\lambda_{3,i})_{M \times 1}$. Then we have

$$f(\boldsymbol{\beta}) = \frac{1}{2} \|\mathbf{Y} - \boldsymbol{\beta}\|_F^2 + \sum_{i=1}^M \lambda_{1,i} \|\boldsymbol{\beta}_i\|_{\ell_1} + \sum_{i=1}^M \lambda_{2,i} \|\boldsymbol{\beta}_{i,2:N} - \boldsymbol{\beta}_{i,1:(N-1)}\|_{\ell_1} + \sum_{j=2}^N \|\boldsymbol{\lambda}_3 * (\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)})\|_{\ell_2}, \quad (3)$$

where $\|\cdot\|_F$ is the Frobenius norm of matrix, $\|\cdot\|_{\ell_1}$ and $\|\cdot\|_{\ell_2}$ are ℓ_1 and ℓ_2 norm of vector, $\boldsymbol{\beta}_{i,s:t}$ indicates the sub-vector with elements $\beta_{i,s}, \dots, \beta_{i,t}$ in row vector $\boldsymbol{\beta}_i$, and “*” is used as entry-wise multiplication between two vectors. Note that it would be easy to modify the tuning parameters so as to make them location specific: that is, reduce the penalty for a jump in correspondence of genomic regions known to harbor CNVs.

3 Implementation

3.1 An MM algorithm

While the solution to the optimization problem (3) might have interesting properties, this approach is useful only if an effective algorithm is available. The last few years have witnessed substantial advances in computational methods for ℓ_1 -regularization problems, including the use of coordinate descent [15, 48] and path following methods [4, 17, 43, 55]. The time cost of these methods in the best situation is $O(MNK)$, for K knots along the solution path. It is important to note that these algorithms – some of which are designed for more general applications – may not be the most

efficient for large scale CNV analysis for at least two reasons: on the one hand, reasonable choices of λ might be available, making it unnecessary to solve for the entire path; on the other hand, the number of knots K can be expected to be as large as $O(N)$, making the computational costs of path algorithms prohibitive.

With specific regard to the fused-lasso application to CNV detection, we were successful in developing algorithm with per iteration cost $O(N)$ and empirically fast convergence rate for the analysis of one sequence [54]. We apply the same principles here. We start by modifying the norms in the penalty as follows: rather than the ℓ_1 norm we use $\|x\|_{2,\epsilon} = \sqrt{x^2 + \epsilon}$ for sufficiently small ϵ , and, for computational stability, we also substitute ℓ_2 norm with $\|\mathbf{x}\|_{2,\epsilon} = (\sum_{i=1}^n x_i^2 + \epsilon)^{\frac{1}{2}}$, obtaining a differentiable objective function

$$\begin{aligned} f_\epsilon(\boldsymbol{\beta}) &= \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^N (y_{ij} - \beta_{ij})^2 + \sum_{i=1}^M \lambda_{1,i} \sum_{j=1}^N \|\beta_{ij}\|_{2,\epsilon} \\ &+ \sum_{i=1}^M \lambda_{2,i} \sum_{j=2}^N \|\beta_{ij} - \beta_{i,j-1}\|_{2,\epsilon} + \sum_{j=2}^N \|\boldsymbol{\lambda}_3 * (\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)})\|_{2,\epsilon}. \end{aligned} \quad (4)$$

Adopting an MM framework [21], we want to find a surrogate function $g_\epsilon(\boldsymbol{\beta} | \boldsymbol{\beta}^{(m)})$ for each iteration m such that $g_\epsilon(\boldsymbol{\beta}^{(m)} | \boldsymbol{\beta}^{(m)}) = f_\epsilon(\boldsymbol{\beta}^{(m)})$ and $g_\epsilon(\boldsymbol{\beta} | \boldsymbol{\beta}^{(m)}) \geq f_\epsilon(\boldsymbol{\beta})$ for all $\boldsymbol{\beta}$. At each iteration, then, $\boldsymbol{\beta}^{(m+1)} = \operatorname{argmin} g_\epsilon(\boldsymbol{\beta} | \boldsymbol{\beta}^{(m)})$. A majorizing function with the above properties is readily obtained using the concavity of square-root function $\|x\|_{2,\epsilon} \leq \frac{1}{2\|z\|_{2,\epsilon}}(x^2 - z^2)$, and its vector equivalent $\|\mathbf{x}\|_{2,\epsilon} \leq \frac{1}{2\|\mathbf{z}\|_{2,\epsilon}}(\|\mathbf{x}\|_{\ell_2}^2 - \|\mathbf{z}\|_{\ell_2}^2)$. The resulting

$$\begin{aligned} g_\epsilon(\boldsymbol{\beta} | \boldsymbol{\beta}^{(m)}) &= \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^N (y_{ij} - \beta_{ij})^2 + \sum_{i=1}^M \lambda_{1,i} \sum_{j=1}^N \frac{\beta_{ij}^2}{2\|\beta_{ij}^{(m)}\|_{2,\epsilon}} \\ &+ \sum_{i=1}^M \lambda_{2,i} \sum_{j=2}^N \frac{(\beta_{ij} - \beta_{i,j-1})^2}{2\|\beta_{ij}^{(m)} - \beta_{i,j-1}^{(m)}\|_{2,\epsilon}} + \sum_{j=1}^N \frac{\|\boldsymbol{\lambda}_3 * (\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)})\|_{\ell_2}^2}{2\|\boldsymbol{\lambda}_3 * (\boldsymbol{\beta}_{(j)}^{(m)} - \boldsymbol{\beta}_{(j-1)}^{(m)})\|_{2,\epsilon}} + c^{(m)} \end{aligned}$$

can be decomposed in the sum of similar functions of all the row vectors $\boldsymbol{\beta}_i$

$$g_\epsilon(\boldsymbol{\beta} | \boldsymbol{\beta}^{(m)}) = \sum_{i=1}^M g_i(\boldsymbol{\beta}_i | \boldsymbol{\beta}^{(m)}),$$

where

$$g_i(\boldsymbol{\beta}_i | \boldsymbol{\beta}^{(m)}) = \frac{1}{2}\boldsymbol{\beta}_i\mathbf{A}_i^{(m)}\boldsymbol{\beta}_i^T - [\mathbf{b}_i^{(m)}]^T\boldsymbol{\beta}_i^T + \tilde{c}_i^{(m)}. \quad (5)$$

Here each $\mathbf{A}_i^{(m)}$ is a tridiagonal symmetric matrix, and $\tilde{c}_i^{(m)}$ is irrelevant constant for optimization purpose. In view of the strict convexity of the surrogate function, each $\mathbf{A}_i^{(m)}$ is also positive definite. The nonzero entries of $\mathbf{A}_i^{(m)}$ and $\mathbf{b}_i^{(m)}$ ($i = 1, \dots, M$) are listed in the supplementary material. Each of the surrogate functions in (5) can be minimized solving the linear system $\boldsymbol{\beta}_i = [\boldsymbol{\beta}_i^{(m)}]^T[\mathbf{A}_i^{(m)}]^{-1}$ by the Tri-diagonal Matrix (TDM) algorithm [11]. This results in a per-iteration computational cost of $O(MN)$. This algorithm is empirically observed to achieve an exponential convergence rate [54], although we do not yet have an analytic proof. In practice, this method scales well with joint analysis of tens to hundreds of samples with measurements at millions of locations, with limitations dictated by memory requirements. For analysis of real data, we suggest one or a group of samples to be analyzed chromosome by chromosome, since a CNV region can never extend beyond one chromosome to another. Actual computation times are shown along with different examples in Section 4.

3.2 Stacking observations at different genomic locations

While copy number is continuously defined across the genome, experimental procedures record data at discrete positions, for which we have used the indexes $j = 1, \dots, N$. In reality, repeated evaluations of the same sample (or related samples) will typically result in measurements at only partially overlapping genomic locations: either because different platforms use different sets of probes, or because missing data may occur at different positions across sequences (consider for example, mBAF and LRR from the same experiment on one subject: the mBAF signal will be defined on a subset of the locations where LRR is).

Let S indicate the union of all genomic positions where some measurement is available among the M signals under study. And let S_i be the subset of locations with measurements in sequence

i. We reconstruct β_{ij} for all $j \in S$. When $j \notin S_i$, β_{ij} will be determined simply on the basis of the neighboring datapoints, relying on the regularizations introduced in (3). The goodness-of-fit portion of the objective function is therefore redefined as

$$\frac{1}{2} \sum_{i=1}^M \sum_{j=1}^N (\delta_{ij} y_{ij} - \delta_{ij} \beta_{ij})^2 \quad \text{with} \quad \delta_{ij} = \begin{cases} 1, & \text{if } j \in S_i, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

The MM strategy previously described applies with slight modifications of the matrix $\mathbf{A}_i^{(m)}$ (see the supplementary material).

The attentive reader would have noted that y_{ij} with $j \notin S_i$ can be considered as missing data, and an evaluation of the characteristics of this missingness is appropriate. In general, y_{ij} cannot be considered missing at random. The most important example is the case of mBAF, where homozygous markers result in missing values. Now, homozygosity is more common when copy number is equal to 1 than when copy number is equal to 2 and, therefore, there is potentially more information on β_{ij} to be extracted from the signals than the one we will capture with the proposed methodology. On the other hand, it does appear that the approach outlined does not increase false positive: operationally, then, it can be considered as an improvement over segmentation based on LRR only, even if in theory, it does not completely use the information on BAF. It is also relevant to note that, in reality, most of the information on deletion is obtained through LRR, and BAF is really carrying additional information in case of duplications (where the changes in LRR are limited due to saturation effects).

3.3 Choice of tuning constants and segmentation

One of the limitations of penalization procedures is that a value for the tuning parameters needs to be set and clear guidelines are not always available. Path methods that obtain a solution of the optimization problem (3) for every value of tuning parameters can be attractive, but recent algorithmic advances [4, 43, 55] remain impractical for problems of the size of ours. A number

of recent publications obtain optimal values of penalty parameters under a series of conditions [3, 5, 6, 13]: we rely upon them to propose the following strategy consisting of obtaining a solution of (3) for reasonably liberal values of the tuning parameters, followed by a sequence-by-sequence hard thresholding of the detected jumps with a data-adaptive threshold.

We have found the following guidelines to be useful in choosing penalty parameter values:

$$\begin{aligned}
 \lambda_{1,i} &= c_1 \hat{\sigma}_i, \\
 \lambda_{2,i} &= \rho(p) c_2 \hat{\sigma}_i \sqrt{\log N}, \\
 \lambda_{3,i} &= [1 - \rho(p)] c_3 \hat{\sigma}_i \sqrt{pM} \sqrt{\log N},
 \end{aligned} \tag{7}$$

for $i = 1, \dots, M$, where $\hat{\sigma}_i$ is a robust estimate of standard deviation of y_i , p is roughly the proportion of the M sequences we anticipate to carry CNVs, and c_1 , c_2 and c_3 are positive multipliers adjusted in consideration of different signal-to-noise ratios and CNV sizes.

While a more rigorous justification is provided in the supplementary material, we start by underscoring some of the characteristics of this proposal.

- The sequence-specific penalizing parameters are proportional to an estimate of the standard deviation of the sequence signal: that is, proviso an initial normalization, the same penalties would be used across all signals.
- The tuning parameter for the total variation (fused lasso) and the Euclidean (group fused lasso) penalties on the jumps depend on $\sqrt{\log N}$, where N is the possible number of jumps. This has a “multiple comparison controlling” effect and resembles rates that have been proven optimal under various sparse scenarios [3, 5, 6, 13]. This term does not appear in the expression of λ_1 , as the lasso penalty can be understood as providing a soft thresholding of the solution of (3) when $\lambda_1 = 0$: given the penalization due to λ_2 and λ_3 , this object will have much smaller dimensionality than N .

- The group penalty depends on \sqrt{M} , where M is the number of grouped sequences, as in the original proposal [50].
- The relative weight of the fused-lasso and group-fused-lasso penalties is regulated by ρ , which depends on p , the proportion of the M sequences expected to carry the same CNV. For example, if $M = 2$ and the two sequences are LRR and BAF from the same individual, we anticipate $p = 1$ with $\rho = 0$, enforcing jumps at identical places in the two signals. At the other extreme, for completely unrelated sequences, $p = 0$ and $\rho = 1$.

The standard deviation $\hat{\sigma}_i$ can be estimated robustly as follows. Let $\Delta_{ij} = y_{i,j+1} - y_{i,j}$, for $j = 1, \dots, N - 1$, be the one-order difference of adjacent y_{ij} for sequence i . Then most $\text{Var}(\Delta_{ij}) = 2\sigma_i^2$ except those bridging real change points, so we can take

$$\hat{\sigma}_i = \widehat{SD}(\Delta_i)/\sqrt{2},$$

where $\widehat{SD}(\Delta_i) = \text{Standard Deviation}(\Delta_i)$ or $\widehat{SD}(\Delta_i) = \text{Median Absolute Deviation}(\Delta_i)$ for $\Delta_i = \{\Delta_{i,1}, \dots, \Delta_{i,N-1}\}$.

As mentioned before, the exact values of the penalty parameters should be adjusted depending on the expectations of signal strengths. Following the approach in [31], one can approximate the bias induced by each of the penalties and hence work backwards in terms of acceptable levels. As detailed in the supplementary material,

$$\text{Bias}(\lambda_1) \approx \lambda_1$$

$$\text{Bias}(\lambda_2) \approx \lambda_2/\text{Length of segment}$$

$$\text{Bias}(\lambda_3) \approx \lambda_3/(\text{Length of segment} \times \sqrt{\# \text{ sequences sharing segment}})$$

Following again the approach in [31], one can show that under some relatively strong assumptions, the choices in (7) lead to a consistent behavior as $N \rightarrow \infty$ and M stays bounded (see the supplementary material). Despite the fact that N is indeed large in our studies, it is not clear that

we can assume it to be in the asymptotic regime. As finer scale measurements become available, scientists desire to investigate CNV of decreasing length: the CNVs we are interested in discovering are often covered by a small number of probes. Furthermore we have often little information on the sizes and frequencies of CNV. In this context, we find it advisable to rely on a two-stage strategy:

1. Sequences are jointly segmented minimizing (3) for a relatively lax choice of the penalty parameters.
2. Jumps are further thresholded on the basis of a data-driven cut-off.

Step 2 allows us to be adaptive to the signal strength and can be carried on with multiple methods. For example, one can adopt the modified Bayesian Information Criteria (mBIC) [52]. For sequence i , the jumps are sorted as $\{\hat{d}_{i(1)}, \dots, \hat{d}_{i(N-1)}\}$ in the descending order of their absolute values. And then we choose the first \hat{k} change points where \hat{k} is given by

$$\hat{k} = \operatorname{argmax}_k \text{mBIC}(k).$$

In data analysis, we often apply an even simpler procedure where the threshold for jumps is defined as a fraction of the maximal jump size observed for every sequence. Specifically, for sequence i , let $\hat{D}_i = \max_{2 \leq j \leq N} \{|\hat{d}_{ij}|\}$, where $\hat{d}_{ij} = \hat{\beta}_{ij} - \hat{\beta}_{i,j-1}$, be the largest observed jump for sequence i . Then we define

$$\gamma_i = \max\{a\hat{\sigma}_i, \min\{\hat{D}_i, b\hat{\sigma}_i\}\}, \quad \text{for } a < b,$$

as a ‘‘ruler’’ reflecting the scale of a possible real jump size, taking $c\gamma_i$ as the cut-off in removal of most small jumps. In all analyses for this paper, we fix $a = 1$, $b = 5$ and $c = 0.2$. In our experience, this heuristic procedure works well for both tumor and normal tissue CNV data.

3.4 Calling Procedure

Even if this is not the focus of our proposal, in order to compare the performance of our segmentation algorithm with HMM approaches, it becomes necessary to distinguish acquisitions from losses of copy number. While the same segmentation algorithm can be applied to a wide range of data sets, calling procedures depend more closely on the specific technology used to carry out the experiments. Since our data analysis relies on Illumina genotyping arrays, we limit ourselves to this platform, and briefly describe the calling procedure we adopt in Section 4.

Analyzing one subject at the time, each segment with constant mean is assigned to one of five possible copy number states ($c = 0, 1, 2, 3, 4$). Let R collect the indexes of all SNPs comprising one segment and let $(\mathbf{x}_R, \mathbf{y}_R) = \{(x_j, y_j), j \in R\}$ be the vectors of values for BAF and LRR in the segment. On the basis of typical pattern for BAF and LRR in the different copy number states (see [9, 45, 47]), we can write log-likelihood ratio

$$\text{LR}(c) = \log \frac{L_{\text{BAF}}(\mathbf{x}_R; c)}{L_{\text{BAF}}(\mathbf{x}_R; 2)} + \log \frac{L_{\text{LRR}}(\mathbf{y}_R; c)}{L_{\text{LRR}}(\mathbf{y}_R; 2)}, \quad c = 0, 1, 3, 4, \quad (8)$$

explicitly defined in the supplementary material. Segment R is assigned a CNV state \hat{c} that maximize $\text{LR}(c)$, only if $\text{LR}(\hat{c}) > r_1$, where r_1 is a pre-specified cut-off.

As noted in [53], the LRR data for a segment with $c = 2$, ideally normalized to have mean 0, often has a small non-zero mean, due to experimental artifacts. If the number of SNPs in R is sufficiently large, a log-likelihood-ratio criterion as the above would result in the erroneous identification of a copy number different from 2. To avoid this, we also require that the size of the absolute difference of the mean of LRR from zero be larger than a threshold $|\bar{y}_R| > r_2\sigma$.

4 Results

We report the results of the analysis of two simulated and two real data sets, which overall exemplify the variety of situations where joint segmentation of multiple sequences is attractive, as

described in the introduction. In all cases, we compare the performance of the proposed procedure with a set of relevant, often specialized, algorithms. The penalized estimation method we put forward in this manuscript shows competitive performance in all cases and often a substantial computational advantage. Its versatility and speed make it a very convenient tool for initial exploration. To calibrate the run times reported in what follows, it is relevant to know that all our analyses were run on a Mac OS X (10.6.7) machine with 2.93 GHz Intel Core 2 Duo and 4 GB 1067 MHz DDR3 memory.

4.1 Simulated CNV in normal samples

We consider one of the simulated data sets described in [54]: relatively short deletion and duplication (300 comprising 5,10, 20, 30, 40, 50 SNPs each) are inserted in the middle of 13000 SNPs long sequences, using a combination of male and female X chromosome data from Illumina HumanHap550 array, appropriately pre-processed to avoid biases (these steps included a scrambling of SNP positions, so to avoid long-range signal fluctuation). This setting mimics the small rare CNVs possibly occurring in the genome of normal individuals: in our main analysis, therefore, we process one individual at the time, reflecting the typical level of information available to scientists in these contexts. HMM methods, like PennCNV, are expected to be the most effective in this problem; segmentation methods like CBS are closer to our own and therefore also make an interesting comparison. As repeatedly discussed, Illumina platform produces two signals for one subject: LRR and BAF. A segmentation method that can process one signal at the time would give its best results using LRR, which carries most of the information. Given this background, we compare four methods: PennCNV, CBS on LRR, fused lasso on LRR only, and group fused lasso on LRR and mBAF. The implementations we use are those reflected in the software packages: PennCNV (version 2010May01), R package DNACopy for CBS (version 1.24.0) [44] and our own R package Piet (version 0.1.0). Tuning parameters for PennCNV and CBS are set at the

default values; the fused lasso implementation corresponds to $\lambda_1 = 0.1$, $\lambda_2 = 2 \times \sqrt{13000}$, and $\lambda_3 = 0$ and the group fused lasso to $\lambda_1 = 0.1$, $\lambda_2 = 0$, and $\lambda_3 = 2 \times \sqrt{13000}$. To call deletion and duplication with CBS and the two fused-lasso approaches, we use both LRR and BAF data (before transformed to mBAF) with the following cut-off values: $r_1 = 10$ and $r_2 = 1(1.5)$ for duplication (deletion). Performance is evaluated by the same indexes we used in [54]: true positive rate (TPR or sensitivity) and false discovery rate (FDR), all defined on a per SNP basis. Results are summarized in Table 1.

Not surprisingly, all algorithms perform similarly well for larger deletions/duplications and it is mainly for variants that involve ≤ 10 SNPs that differences are visible. Algorithms that rely only on LRR (as CBS and fused lasso) underperform in the detection of small duplications (comparison is particularly easy for duplications of size 10 SNP, where the selected parameter values lead to similar FDRs in the three segmentation methods). The group fused lasso can almost entirely recover the performance of PennCNV and outperforms CBS in this context.

For curiosity, we analyzed all sequences simultaneously. While this represents an unrealistic amount of prior information, it allows us to evaluate the possible gain of joint analysis: FDR practically become 0 ($<0.02\%$) for all CNV sizes, but power increases only for CNV including less than 10 SNPs.

Finally, it is useful to compare running times. Summary statistics of the per sample time are reported in Table 1: while all algorithms are rather fast, the two implementations of the fused lasso are dominating.

4.2 A simulated tumor data set

To explore the challenges presented by tumor data, we rely on a data set created by [39], with the specific goal of studying the effect of contamination between normal and cancer cells. The HapMap sample NA06991, genotyped on Illumina HumanHap550 array, was used to simulate a

cancer cell line, by inserting a total of 10 structure variation regions, including one-copy losses, one-copy gains, and copy neutral loss-of-heterozygosity (CN-LOH) (see Supplementary Table 2). The signal from this artificial “tumor” sample was then contaminated *in silico* with that of the original “normal” sample, resulting in 21 data sets, with a percentage of normal cells ranging from 0% to 100%. Note that most simulated CNV or CN-LOH regions are very large—some spanning an entire chromosome—and the challenge in detection is really due to the contamination levels.

For ease of comparison, we evaluate the accuracy of calling procedures as in the original reference [39]: sensitivity is measured for each variant region as the percentage of heterozygous SNPs that are assigned the correct copy number; and specificity is the percentage of originally heterozygous SNPs in unperturbed regions that are assigned CN=2. We compare the performance of GFL to BAFsegmentation [39] and PSCN [8] representing, respectively, a version of segmentation and HMM approaches specifically developed to deal with contaminated tumor samples (both these algorithms have been tested with success on this simulated data set).

Following other analyses, we do not pre-process the data prior to CNV detection. BAFsegmentation and PSCN were run using recommended parameter values. For each of the diluted data sets, we applied the GFL model on each chromosome at one time using both LRR and mBAF, whose standard deviations are normalized to 1. Tuning constants are set to $\lambda_1 = 0$, $\lambda_2 = 0.5 \times 3 \times \sqrt{\log N}$, and $\lambda_3 = 0.5 \times 3 \times \sqrt{\log N}$, varying specifically for chromosome interrogated by N SNPs. The change points resulting from hard segmentation on LRR and mBAF are combined to make a finer segmentation of the genome. Finally, we adopt the same calling procedure described by [39]. For ease of comparison with PSCN, only analysis of simulated tumor data are reported, even if BAFsegmentation and GFL would gain from using the genotype of normal cell in defining mBAF.

Figure 1 summarizes the sensitivity of each method, as a function of percentage of normal cell in the sample. Sensitivity is calculated for each of the 10 regions separately. All three methods work reasonably well under a wide range of percentages of normal cell contamination (in 5 out of the 10 regions, GFL appears to lead to best results, while in the other 5 PSCN does). The

CNV region that comprises the smallest amount of SNP is the hemizygous loss on Chromosome 13: in this case GFL in our hands behaved in the most stable manner. GLF outperforms the two comparison methods in terms of specificity (Figure 2): while the specificity values might appear very high in any case, this is somewhat of an artifact due to the adopted definition of this index. It is relevant to note that the performance of PSCN in our hands does not correspond to the published one [8]. While we tried our best to set the parameter values, we have not succeeded in replicating the authors' original results, which should be considered in the interest of fairness.

PSCN, like GFL, is implemented in R with some computationally intensive subroutines coded in C. BAFsegmentation relies its segmentation part on the R package DNACopy, whose core algorithms are implemented in C and Fortran, and it is wrapped in Perl. A comparison of run times indicate that GLF and BAFsegmentation are comparable, while PSCN is fifty times slower than GFL (see Supplementary Table 3).

4.3 One sample assayed with multiple replicates and multiple platforms

We use the data from a study [28] assessing the performance of different array platforms and CNV calling methods to illustrate the advantages of joint analysis of multiple measurements on the same subject. DNA from four individuals was analyzed in triplicate on each of 5 platforms: Affymetrix 6.0, Illumina 1M, 660W, Omni1-Quad (O1Q) and Omni2.5-Quad (O2Q) (among others [28]). We use the results on the first three to define “true” copy numbers and try to reconstruct them using data from O1Q and O2Q. The nine “reference” experiments were analyzed with 4 or 5 CNV calling algorithms (see [28]) and a CNV was identified using majority votes: consistent evidence was required from at least 2 analysis tools, on at least 2 platforms, and in at least 2 replicates (see Supplementary Table 4). Here CNVs detected in two replicates/algorithms/platforms are regarded as the same CNV and collapse down to one CNV with the outmost boundaries when they overlap with each other.

The test experiments are based on 1,020,596 and 2,390,395 SNPs on autosomes after some quality control, at a total of 2,657,077 unique loci. Since our focus here is to investigate how to best analyze multiple signals on the same subject, rather than on the specific properties of any CNV calling method, we carry out all the analyses using different settings of GFL in segmentation while keeping the same CNV calling and summarizing procedure. All segmentation is done on LRR only while calling procedure uses both LRR and BAF (with cut-off $r_1 = 10$ and $r_2 = 1$). Here we compare three segmentation settings to analyze these 6 experiments per subject (see Supplementary Table 5 for more details about tuning parameters):

1. The signals from the three technical replicates with one platform are averaged and then segmented and subject to calling procedure separately. The final CNV list is the union of CNV calls from the two platforms.
2. The signals from the three technical replicates with one platform are each segmented and subject to calling procedure separately. A majority vote is used to summarize CNV result for each platform: a CNV needs to be called in at least two replicates out of three. The final CNV list is the union of the two platforms' results.
3. The signals from the three technical replicates of both platforms (6 LRR sequences) are segmented jointly. Calling procedure is still done on each replicate separately, and the same majority vote is used to summarize CNV result for each platform. Again, the final CNV list is the union of the two platforms' results.

To benchmark the result of joint analysis we use MPCBS [51], a segmentation method, specifically designed for multi-platform CNV analysis. The segments output from MPCBS are proceeded to the same calling, majority voting, and summarizing procedure.

Table 2 presents the results: averaging results from different technical replicates leads to loss of power, while joint analysis of all the signals leads to the most effective performance. GFL joint

analysis leads to results comparable to those of MPCBS, but it is at least 30 times faster than the competing method.

4.4 Multiple related samples assayed with the same platform

In the context of a study of the genetic basis of bipolar disorder, the Illumina Omni2.5-Quad chip was used to genotype 455 individuals from 11 Columbian and 13 Costa Rican pedigrees. We use this data set to explore the advantages of a joint segmentation of related individuals. In absence of a reference evaluation of CNV status in these samples, we rely on two indirect methods to assess the quality of the predicted CNVs. We used the collection of CNVs observed in HapMap Phase III [19] to compile a list of 426 copy number polymorphisms (selecting all those CNVs with frequency ≥ 0.05 in pooled samples from 11 populations) and assumed that if we identify in our sample a CNV corresponding to one of these regions, we should consider it a true positive. For the purposes of this analysis we considered a detected CNV to correspond to one identified in HapMap if there was any overlap between the two regions.

Another indirect measure of the quality of CNV calls derives from the amount of Mendelian errors encountered in the pedigrees when we consider the CNV as a segregating site. De novo CNVs are certainly a possibility, and in their case Mendelian errors are to be expected. However, when the CNV in question is a common one (already identified in HapMap), it is reasonable to expect that it segregate in the pedigrees as any regular polymorphism. We selected a very common deletion on Chromosome 8 (HapMap reports overall frequency > 0.4 in 11 populations) and compared different CNV calling procedures on the basis of how many Mendelian errors they generate.

As mentioned before, PennCNV represents a state-of-the-art HMM method for the analysis of normal samples and, therefore, we included it in our comparisons. However, the parameters of the underlying HMM algorithm had not been tuned on the Omni2.5-Quad at the time of writing,

resulting in sub-standard performance. Segmentation methods are less dependent on parameter optimization; hence, GFL analysis of LRR and BAF one subject at a time can provide a better indication of the potential of single-sample methods. We considered two multiple-sample algorithms: GFL and MSSCAN [53], both applied on LRR with group defined by pedigree memberships. (While a trio-mode is available in PennCNV [46], this does not adapt to the structure of our families.) A final qualification is in order. While the authors of MSSCAN kindly shared with us a beta-version of their software, we find it not to be robust. Indeed, we were unable to use it to segment the entire genome. However, we successfully used it to segment Chromosome 8, so that we could include MSSCAN in the comparison based on Mendelian error rates.

Prior to analysis, the data was normalized using the GC-content correction implemented in PennCNV [12]. For individual analysis, the GFL parameters were $\lambda_1 = 0.1$, $\lambda_2 = 0$, and $\lambda_3 = 2 \times \sqrt{\log N}$, where N is the number of SNPs deployed on each chromosome; for pedigree analysis, the GFL parameters were $\lambda_1 = 0.1$, $\lambda_2 = 0.5 \times 2 \times \sqrt{\log N}$, and $\lambda_3 = 0.5 \times 2 \times \sqrt{0.3M} \times \sqrt{\log N}$, where M is the number of individuals in each pedigree. For MSSCAN, CNV size is constraint to be less than 200 SNPs and the maximum number of change points is set as 50. The calling procedure with $r_1 = 10$ and $r_2 = 1$ was applied to both the GFL and MSSCAN results.

Table 3 summarized the total number of copy number polymorphisms (CNPs) identified in our sample by different approaches and their overlap with known CNPs from HapMap. For the purpose of this comparison we considered as a CNP a variant with frequency at least 10% in our sample. All analysis modes of GFL agree more with HapMap list than PennCNV in the sense of percentage of overlap. It is also clear that GFL-pedigree analysis achieves larger overlap with HapMap data than GFL-individual analysis. The time cost per sample for pedigree is reasonable and scales well with the increment of sample size.

Table 4 summarizes the results of our investigation of a 154kb CNP region on Chromosome 8p (from 39,351,896 to 39,506,122 on NCBI Build 36 coordinate). All methods but PennCNV show detected deletions only; this coincides with the observation from HapMap data. We used option

Mistyping of Mendel (version 11.0) [22, 38] to detect Mendelian errors. Joint segmentation methods discover more hemizygous deletions than individual analysis, resulting in fewer Mendelian errors. MSSCAN discovers the largest number of hemizygous deletions. Figure 3 shows an example of large pedigree, where 3 out of 4 Mendelian errors are removed by joint analysis.

Table 1: Detection accuracy (as percentage of SNPs) and computation times for PennCNV, CBS, Fused Lasso and Group Fused Lasso on a simulated set of CNVs in normal samples. Overall accuracy are calculated pooling all sequences with a given type of CNVs. The average (and standard deviation) of the number of seconds required for the analysis of one sequence is reported.

CNV Size	CNV Type	PennCNV		CBS		Fused Lasso		Group Fused Lasso	
		TPR	FDR	TPR	FDR	TPR	FDR	TPR	FDR
5	Deletion	83.80	4.92	78.20	0.68	63.93	1.74	64.27	1.83
	Duplication	58.53	4.67	11.67	10.26	20.00	37.76	39.87	14.33
10	Deletion	95.03	1.45	88.37	0.56	88.50	0.60	88.87	0.56
	Duplication	93.43	0.78	56.50	4.40	83.90	12.60	91.60	3.85
20	Deletion	94.63	0.58	90.50	0.39	90.80	0.47	90.83	0.47
	Duplication	96.13	0.92	86.22	3.58	92.77	4.95	94.98	2.13
30	Deletion	94.57	0.28	93.30	0.29	89.38	0.52	89.77	0.53
	Duplication	96.09	0.05	90.77	1.61	94.32	1.78	94.98	1.29
40	Deletion	97.83	0.59	97.58	0.09	97.28	0.19	97.28	0.19
	Duplication	94.61	0.46	92.77	0.98	93.94	1.15	94.63	0.75
50	Deletion	94.33	0.07	92.76	0.04	90.47	0.11	90.48	0.11
	Duplication	94.50	0.09	93.81	0.74	93.11	0.79	93.64	0.49
Overall Deletion		95.02	0.55	93.06	0.19	91.08	0.33	91.19	0.34
Overall Duplication		93.82	0.44	86.92	1.55	90.56	2.85	92.46	1.38
Overall		94.42	0.49	89.99	0.85	90.82	1.60	91.83	0.87
Time (sec.)		0.48 (0.01)		0.78 (0.69)		0.22 (0.13)		0.28 (0.05)	

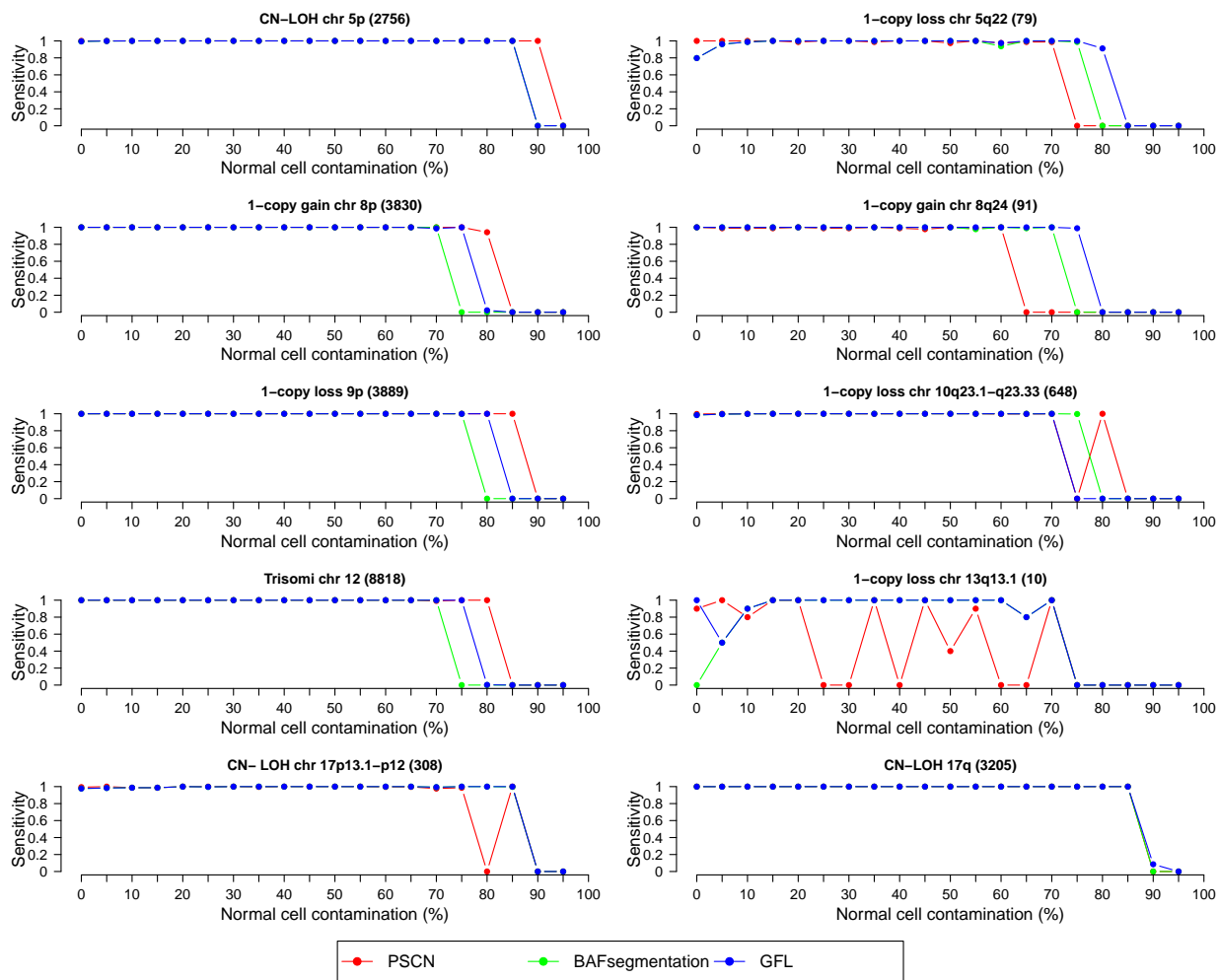


Figure 1: Sensitivity as function of percentage contamination by normal cells in the 10 different simulated CNV regions. Sensitivity is not defined at 100% contamination.

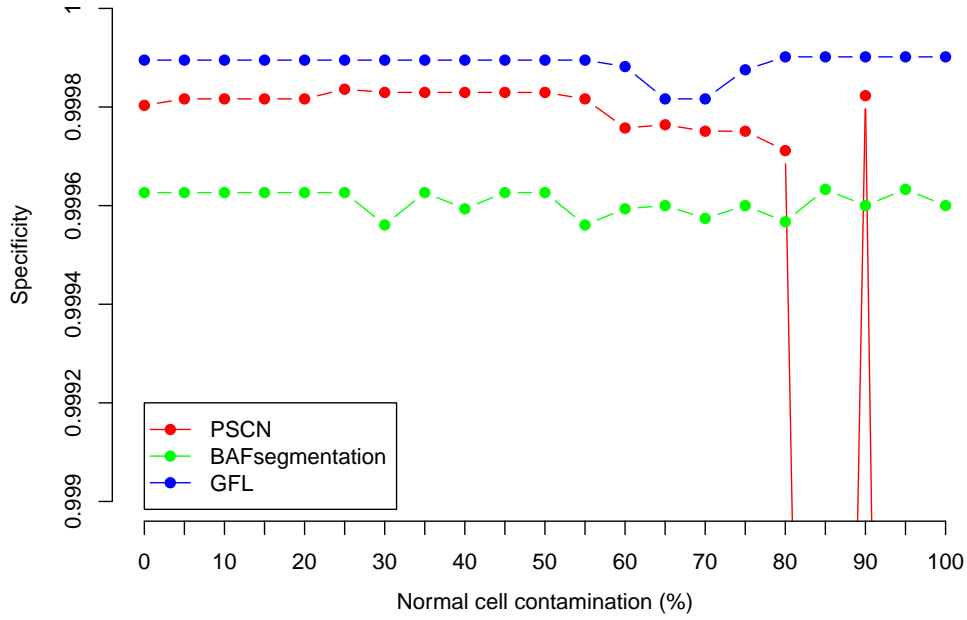


Figure 2: Specificity as function of percentage contamination by normal cells. Note that [8] reports better performance of PSCN in correspondence of contamination levels 85% , 95% and 100%.

Table 2: Number of CNVs detected (Det.) and overlapping (Ovlp.) with reference results as well as average computation time for four samples under different analyses.

	NA15510		NA18517		NA18576		NA18980		
Analysis	# Det.	# Ovlp.	# Det.	# Ovlp.	# Det.	# Ovlp	# Det.	# Ovlp	Time (min.)
Analysis 1	170	38	144	34	160	25	145	22	1.2
Analysis 2	102	36	109	33	93	25	91	20	3.7
Analysis 3	80	38	82	32	69	25	56	15	8.5
MPCBS	98	34	88	28	59	18	68	21	313.9

Table 3: The number of detected CNP regions with frequency ≥ 0.1 in our sample by different methods and their overlap with a list of CNP regions compiled from HapMap data. Computation time (in minute) is per sample.

Method	# detected CNVR	# Overlap	% Overlap	Time (min.)
PennCNV	189	63	33.33%	3.44
GFL-Individual (LRR+BAF)	95	50	52.63%	3.90
GFL-Pedigree (LRR)	106	62	58.49%	1.57

Table 4: Detected copy numbers in a common deletion on Chromosome 8. Across the various algorithms, subjects are assigned to one of 4 types of copy number: for each algorithm, we report the total numbers of $CN \neq 2$ identified; the total number of “core” families with Mendelian errors; and the average computation time (in minute) per sample for the analysis of Chromosome 8.

Method	# CN=0	# CN=1	# CN=3	# families with Mendelian errors	Time (min.)
PennCNV	125	39	102	35	0.19
GFL-Individual	123	97	0	20	0.21
GFL-Pedigree	123	137	0	15	0.09
MSSCAN-Pedigree	123	154	0	15	0.11

5 Discussion

We have presented a segmentation method based on penalized estimation and capable of processing multiple signals jointly. We have shown how this leads to improvements in the analysis of normal samples (where segmentation can be applied to both total intensity and allelic proportion), tumor sample (where we are able to deal with contamination effectively), measurements from multiple platforms, and related individuals. Given that copy number detection is such an active area of research, it is impossible to compare one method to all the others available. However, for each of the situations we analyzed, we tried to select approaches that represented the most successful state-of-the-art. In comparison to these, the algorithm we presented performs well: its accuracy is always comparable to that of the most effective competitor and its computation time often more contained. We believe that for its versatility and speed, GFL is particularly useful for initial screening.

There are of course many aspects of CNV detection that we have not analyzed in this paper: from normalization and signal transformation to FDR control of detected CNVs. There are also a number of improvements to our approach that appear promising, but at this stage are left for further work: for example, it is easy to modify algorithms so that the penalization parameters are location dependent to incorporate prior information on known copy number polymorphisms; more challenging is developing theory and method to select the values of these regularization parameters in a data-adaptive fashion.

Finally, while our scientific motivation has been the study of copy number variations, the joint segmentation algorithm we present is not restricted to specific characteristics of these data types, and we expect it will be applied in other contexts.

Software implementation

All the code used to run the analysis presented in this paper is available at the web-page of the authors. We have implemented the segmentation routine, which is our core contribution, in an R

package (`Piet`) to be submitted to R-forge (<http://r-forge.r-project.org>). To demonstrate a visualization of the CNV results on Chromosome 8 in the bipolar disorder study (see Section 4.4), we refer the interested audience to Supplementary Figure 2 in the supplementary material.

Acknowledgments

The authors thank Nelson Freimer and all the collaborators in the Bipolar Endophenotype Mapping project for authorizing use of the genotype data. We also thank Susan Service and Joseph DeYoung for assistance in data management and interpretation, and Pierre Neuvial and Henrik Bengtsson for helpful discussion. C. S. gratefully acknowledges support from NIH/NIGMS GM053275 and MH075007 and K. L. from NIH/NIGMS GM053275.

References

- [1] H. Bengtsson, P. Neuvial, and T. Speed. Tumorboost: Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. *BMC bioinformatics*, 11(1):245, 2010.
- [2] H. Bengtsson, P. Wirapati, and T. P. Speed. A single-array preprocessing method for estimating full-resolution raw copy numbers from all affymetrix genotyping arrays including genomewidesnp 5 & 6. *Bioinformatics*, 25(17):2149–2156, 2009.
- [3] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.
- [4] K. Bleakley and J. P. Vert. The group fused lasso for multiple change-point detection. *Arxiv preprint arXiv:1106.4199*, 2011.

- [5] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.
- [6] E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n . *The Annals of Statistics*, 35(6):2313–2351, 2007.
- [7] B. Carvalho, H. Bengtsson, T. P. Speed, and R. A. Irizarry. Exploration, normalization, and genotype calls of high-density oligonucleotide snp array data. *Biostatistics*, 8(2):485–499, 2007.
- [8] H. Chen, H. Xing, and N. R. Zhang. Estimation of parent specific dna copy number in tumors using high-density genotyping arrays. *PLoS Computational Biology*, 7:e1001060, 2011.
- [9] S. Colella, C. Yau, J. M. Taylor, G. Mirza, H. Butler, P. Clouston, A. S. Bassett, A. Seller, C. C. Holmes, and J. Ragoussis. Quantisnp: An objective bayes hidden-markov model to detect and accurately map copy number variation using snp genotyping data. *Nucleic Acids Research*, 35:2013–2025, 2007.
- [10] D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. MacArthur, J. R. MacDonald, I. Onyiah, A. W. C. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, The Wellcome Trust Case Control Consortium, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712, 2009.
- [11] S. D. Conte and C. deBoor. *Elementary Numerical Analysis*. McGraw-Hill, New York, 1972.
- [12] S. J. Diskin, M. Li, C. Hou, S. Yang, J. Glessner, H. Hakonarson, M. Bucan, J. M. Maris, and K. Wang. Adjustment of genomic waves in signal intensities from whole-genome snp genotyping platforms. *Nucleic Acids Research*, 36:e126, 2008.

- [13] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.
- [14] B. Efron and N. R. Zhang. False discovery rates and copy number variation. *Biometrika*, 98(2):251–271, 2011.
- [15] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1:302–332, 2007.
- [16] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *Arxiv preprint arXiv:1001.0736*, 2010.
- [17] H. Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010.
- [18] A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee. Detection of large-scale variation in the human genome. *Nature Genetics*, 36:949–951, 2004.
- [19] M. Jakobsson, S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, H. C. Fung, Z. A. Szpiech, J. H. Degnan, K. Wang, R. Guerreiro, J. M. Bras, J. C. Schymick, D. G. Hernandez, B. J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H. M. Cann, J. A. Hardy, N. A. Rosenberg, and A. B. Singleton. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451:998–1003, 2008.
- [20] J. O. Korbel, A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, D. Palejev, N. J. Carriero, L. Du, B. E. Taillon, Z. Chen, A. Tanzer, A. C. Eugenia Saunders, J. Chi, F. Yang, N. P. Carter, M. E. Hurles, S. M. Weissman, T. T. Harkins, M. B. Gerstein, M. Egholm, and M. Snyder. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318:420–426, 2007.

- [21] K. Lange. *Optimization*. Springer, New York, 2004.
- [22] K. Lange, R. Cantor, S. Horvath, M. Perola, C. Sabatti, J. Sinsheimer, and E. Sobel. Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. *Am J Hum Genet*, 69(Suppl 1):504, 2001.
- [23] R. E. Mills, K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, A. Chinwalla, D. F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L. M. Iakoucheva, Z. Iqbal, S. Kang, J. M. Kidd, M. K. Konkel, J. Korn, E. Khurana, D. Kural, H. Y. K. Lam, J. Leng, R. Li, Y. Li, C.-Y. Lin, R. Luo, X. J. Mu, J. Nemesh, H. E. Peckham, T. Rausch, A. Scally, X. Shi, M. P. Stromberg, A. M. Stutz, A. E. Urban, J. A. Walker, J. Wu, Y. Zhang, Z. D. Zhang, M. A. Batzer, L. Ding, G. T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, K. Ye, E. E. Eichler, M. B. Gerstein, M. E. Hurles, C. Lee, S. A. McCarroll, J. O. Korbel, and 1000 Genomes Project. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, 2011.
- [24] M. A. Newton and Y. Lee. Inferring the location and effect of tumor suppressor genes by instability-selection modeling of allelic-loss data. *Biometrics*, 56(4):1088–1097, 2000.
- [25] G. Nowak, T. Hastie, J. R. Pollack, and R. Tibshirani. A fused lasso latent feature model for analyzing multi-sample acgh data. *Biostatistics*, 12(4):776–791, 2011.
- [26] 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*, 5:557–572, 2004.
- [27] I. Ostrovnyaya, A. B. Olshen, V. E. Seshan, I. Orlov, D. G. Albertson, and C. B. Begg. A metastasis or a second independent cancer? evaluating the clonal origin of tumors using array copy number data. *Statistics in medicine*, 29(15):1608–1621, 2010.

- [28] D. Pinto, K. Darvishi, X. Shi, D. Rajan, D. Rigler, T. Fitzgerald, A. C. Lionel, B. Thiruvahindrapuram, J. R. MacDonald, R. Mills, A. Prasad, K. Noonan, S. Gribble, E. Prigmore, P. K. Donahoe, R. S. Smith, J. H. Park, M. E. Hurler, N. P. Carter, C. Lee, S. W. Scherer, and L. Feuk. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotechnology*, 29:512–520, 2011.
- [29] D. Pinto, A. T. Pagnamenta, L. Klei, R. Anney, D. Merico, R. Regan, J. Conroy, T. R. Magalhaes, C. Correia, B. S. Abrahams, J. Almeida, E. Bacchelli, G. D. Bader, A. J. Bailey, G. Baird, A. Battaglia, T. Berney, N. Bolshakova, S. Bölte, P. F. Bolton, T. Bourgeron, S. Brennan, J. Brian, S. E. Bryson, A. R. Carson, G. Casallo, J. Casey, B. H. Y. Chung, L. Cochrane, C. Corsello, E. L. Crawford, A. Crossett, C. Cytrynbaum, G. Dawson, M. de Jonge, R. Delorme, I. Drmic, E. Duketis, F. Duque, A. Estes, P. Farrar, B. A. Fernandez, S. E. Folstein, E. Fombonne, C. M. Freitag, J. Gilbert, C. Gillberg, J. T. Glessner, J. Goldberg, A. Green, J. Green, S. J. Guter, H. Hakonarson, E. A. Heron, M. Hill, R. Holt, J. L. Howe, G. Hughes, V. Hus, R. Iglizzi, C. Kim, S. M. Klauck, A. Klevzon, O. Korvatska, V. Kustanovich, C. M. Lajonchere, J. A. Lamb, M. Laskawiec, M. Leboyer, A. Le Couteur, B. L. Leventhal, A. C. Lionel, X.-Q. Liu, C. Lord, L. Lotspeich, S. C. Lund, E. Maestrini, W. Mahoney, C. Mantoulan, C. R. Marshall, H. McConachie, C. J. McDougale, J. McGrath, W. M. McMahon, A. Merikangas, O. Migita, N. J. Minshew, G. K. Mirza, J. Munson, S. F. Nelson, C. Noakes, A. Noor, G. Nygren, G. Oliveira, K. Papanikolaou, J. R. Parr, B. Parrini, T. Paton, A. Pickles, M. Pilorge, J. Piven, C. P. Ponting, D. J. Posey, A. Poustka, F. Poustka, A. Prasad, J. Ragoussis, K. Renshaw, J. Rickaby, W. Roberts, K. Roeder, B. Roge, M. L. Rutter, L. J. Bierut, J. P. Rice, J. Salt, K. Sansom, D. Sato, R. Segurado, A. F. Sequeira, L. Senman, N. Shah, V. C. Sheffield, L. Soorya, I. Sousa, O Stein, N. Sykes, V. Stoppioni, C. Strawbridge, R. Tancredi, K. Tansey, B. Thiruvahindrapduram, A. P. Thompson, S. Thomson, A. Tryfon, J. Tsiantis, H. Van Engeland, J. B. Vincent, F. Volkmar, S. Wallace, K. Wang,

- Z. Wang, T. H. Wassink, C. Webber, R. Weksberg, K. Wing, K. Wittmeyer, S. Wood, J. Wu, B. L. Yaspan, D. Zurawiecki, L. Zwaigenbaum, J. D. Buxbaum, R. M. Cantor, E. H. Cook, H. Coon, M. L. Cuccaro, B. Devlin, S. Ennis, L. Gallagher, D. H. Geschwind, M. Gill, J. L. Haines, J. Hallmayer, J. Miller, A. P. Monaco, J. I. Nurnberger Jr, A. D. Paterson, M. A. Pericak-Vance, G. D. Schellenberg, P. Szatmari, A. M. Vicente, V. J. Vieland, E. M. Wijsman, S. W. Scherer, Sutcliffe J. S., and C. Betancur. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466:368–372, 2010.
- [30] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles. Global variation in copy number in the human genome. *Nature*, 444:444–454, 2006.
- [31] A. Rinaldo. Properties and refinements of the fused lasso. *The Annals of Statistics*, 37:2922–2952, 2009.
- [32] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259–268, 1992.
- [33] R. B. Scharpf, M. E. Irizarry R. A., Ritchie, B. Carvalho, and I. Ruczinski. Using the r package crlmm for genotyping and copy number estimation. *Journal of Statistical Software*, 40(i12):1–32, 2011.

- [34] R. B. Scharpf, G. Parmigiani, J. Pevsner, and I. Ruczinski. Hidden markov models for the assessment of chromosomal alterations using high throughput snp arrays. *The Annals of Applied Statistics*, 2:687–713, 2008.
- [35] R. B. Scharpf, I. Ruczinski, B. Carvalho, B. Doan, A. Chakravarti, and R. A. Irizarry. A multilevel model to address batch effects in copy number estimation using snp arrays. *Biostatistics*, 12(1):33–50, 2011.
- [36] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T. C. Gilliam, B. Trask, N. Patterson, A. Zetterberg, and M. Wigler. Large-scale copy number polymorphism in the human genome. *Science*, 305:525–528, 2004.
- [37] D. O. Siegmund, B. Yakir, and N. R. Zhang. Detecting simultaneous intervals in aligned sequences. *The Annals of Applied Statistics*, 5:645–668, 2011.
- [38] E. Sobel, J. C. Papp, and K. Lange. Detection and integration of genotyping errors in statistical genetics. *The American Journal of Human Genetics*, 70(2):496–508, 2002.
- [39] J. Staaf, D. Lindgren, J. Vallon-Christersson, A. Isaksson, H. Göransson, G. Juliusson, R. Rosenquist, M. Höglund, A. Borg, and M. Ringnér. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome snp arrays. *Genome Biology*, 9:R136, 2008.
- [40] H. Stefansson, D. Rujescu, S. Cichon, O. P. H. Pietiläinen, A. Ingason, S. Steinberg, R. Fossdal, E. Sigurdsson, T. Sigmundsson, J. E. Buizer-Voskamp, T. Hansen, K. D. Jakobsen, P. Muglia, C. Francks, P. M. Matthews, A. Gylfason, B. V. Halldorsson, D. Gudbjartsson, T. E. Thorgeirsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, A. Bjornsson, S. Mattiasdottir, T. Blondal, M. Haraldsson, B. B. Magnúsdóttir, I. Giegling, H.-J. Möller, A. Hartmann,

- K. V. Shianna, D. Ge, A. C. Need, C. Crombie, G. Fraser, N. Walker, J. Lonnqvist, J. Suvisaari, A. Tuulio-Henriksson, T. Paunio, T. Touloupoulou, E. Bramon, M. Di Forti, R. Murray, M. Ruggeri, E. Vassos, S. Tosato, M. Walshe, T. Li, C. Vasilescu, T. W. Mühleisen, A. G. Wang, H. Ullum, S. Djurovic, I. Melle, J. Olesen, L. A. Kiemeny, B. Franke, Genetic Risk and Outcome in Psychosis (GROUP), C. Sabatti, N. B. Freimer, J. R. Gulcher, U. Thorsteinsdottir, A. Kong, O. A. Andreassen, R. A. Ophoff, A. Georgi, M. Rietschel, T. Werge, H. Petursson, D. B. Goldstein, M. M. Nöthen, L. Peltonen, D. A. Collier, D. St Clair, and K. Stefansson. Large recurrent microdeletions associated with schizophrenia. *Nature*, 455:232–236, 2008.
- [41] W. Sun, F. A. Wright, Z. Tang, S. H. Nordgard, P. Van Loo, T. Yu, V. N. Kristensen, and C. M. Perou. Integrated study of copy number states and genotype calls using high-density snp arrays. *Nucleic Acids Research*, 37(16):5365–5377, 2009.
- [42] R. Tibshirani and P. Wang. Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics*, 9:18–29, 2008.
- [43] R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011.
- [44] E. S. Venkatraman and A. B. Olshen. A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, 23(6):657–663, 2007.
- [45] H. Wang, J. H. Veldink, H. Blauw, L. H. van den Berg, R. A. Ophoff, and C. Sabatti. Markov models for inferring copy number variations from genotype data on illumina platforms. *Human Heredity*, 68:1–22, 2009.
- [46] K. Wang, Z. Chen, M. G. Tadesse, J. Glessner, S. F. A. Grant, H. Hakonarson, M. Bucan, and M. Li. Modeling genetic inheritance of copy number variations. *Nucleic Acids Research*, 36:e138, 2008.

- [47] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. F. A. Grant, H. Hakonarson, and M. Bucan. Penncnv: An integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. *Genome Research*, 17:1665–1674, 2007.
- [48] T. T. Wu and K. Lange. Coordinate descent algorithm for lasso penalized regression. *The Annals of Applied Statistics*, 2:224–244, 2008.
- [49] C. Yau, O. Papaspiliopoulos, G. O. Roberts, and C. Holmes. Bayesian non-parametric hidden markov models with applications in genomics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:37–57, 2011.
- [50] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68:49–67, 2006.
- [51] N. R. Zhang, Y. Senbabaoglu, and J. Z. Li. Joint estimation of dna copy number from multiple platforms. *Bioinformatics*, 26:153–160, 2010.
- [52] N. R. Zhang and D. O. Siegmund. A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63:22–32, 2007.
- [53] N. R. Zhang, D. O. Siegmund, H. Ji, and J. Z. Li. Detecting simultaneous changepoints in multiple sequences. *Biometrika*, 97:631–645, 2010.
- [54] Z. Zhang, K. Lange, R. Ophoff, and C. Sabatti. Reconstructing dna copy number by penalized estimation and imputation. *The Annals of Applied Statistics*, 4:1749–1773, 2010.
- [55] H. Zhou and K. Lange. A path algorithm for constrained estimation. *Arxiv preprint arXiv:1103.3738*, 2011.
- [56] H. Zhou, M. E. Sehl, J. S. Sinsheimer, and K. Lange. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, 26(19):2375–2382, 2010.

**Reconstructing DNA Copy Number by Joint Segmentation of Multiple
Sequences**

SUPPLEMENTARY MATERIAL

Zhongyang Zhang, Kenneth Lange, and Chiara Sabatti

March 2012

TDM algorithm

The non-zero entries in \mathbf{A}_i and \mathbf{b}_i in the re-shaped surrogate function (5) are listed as follows:

$$\begin{aligned}
a_i^{(m)}(1, 1) &= 1 + \frac{\lambda_{1,i}}{\|\beta_{i1}^{(m)}\|_{2,\epsilon}} + \frac{\lambda_{2,i}}{\|\beta_{i2}^{(m)} - \beta_{i1}^{(m)}\|_{2,\epsilon}} + \frac{\lambda_{3,i}^2}{\|\boldsymbol{\lambda}_3 * (\boldsymbol{\beta}_{(2)}^{(m)} - \boldsymbol{\beta}_{(1)}^{(m)})\|_{2,\epsilon}}; \\
a_i^{(m)}(j, j) &= 1 + \frac{\lambda_{1,i}}{\|\beta_{ij}^{(m)}\|_{2,\epsilon}} + \frac{\lambda_{2,i}}{\|\beta_{ij}^{(m)} - \beta_{i,j-1}^{(m)}\|_{2,\epsilon}} + \frac{\lambda_{2,i}}{\|\beta_{i,j+1}^{(m)} - \beta_{ij}^{(m)}\|_{2,\epsilon}} \\
&\quad + \frac{\lambda_{3,i}^2}{\|\boldsymbol{\lambda}_3 * (\boldsymbol{\beta}_{(j)}^{(m)} - \boldsymbol{\beta}_{(j-1)}^{(m)})\|_{2,\epsilon}} + \frac{\lambda_{3,i}^2}{\|\boldsymbol{\lambda}_3 * (\boldsymbol{\beta}_{(j+1)}^{(m)} - \boldsymbol{\beta}_{(j)}^{(m)})\|_{2,\epsilon}}, \\
&\qquad\qquad\qquad j = 2, \dots, n-1; \\
a_i^{(m)}(n, n) &= 1 + \frac{\lambda_{1,i}}{\|\beta_{in}^{(m)}\|_{2,\epsilon}} + \frac{\lambda_{2,i}}{\|\beta_{in}^{(m)} - \beta_{i,n-1}^{(m)}\|_{2,\epsilon}} + \frac{\lambda_{3,i}^2}{\|\boldsymbol{\lambda}_3 * (\boldsymbol{\beta}_{(n)}^{(m)} - \boldsymbol{\beta}_{(n-1)}^{(m)})\|_{2,\epsilon}}; \\
a_i^{(m)}(j, j-1) &= -\frac{\lambda_{2,i}}{\|\beta_{ij}^{(m)} - \beta_{i,j-1}^{(m)}\|_{2,\epsilon}} - \frac{\lambda_{3,i}^2}{\|\boldsymbol{\lambda}_3 * (\boldsymbol{\beta}_{(j)}^{(m)} - \boldsymbol{\beta}_{(j-1)}^{(m)})\|_{2,\epsilon}}, \quad j = 2, \dots, n; \\
a_i^{(m)}(j, j+1) &= -\frac{\lambda_{2,i}}{\|\beta_{i,j+1}^{(m)} - \beta_{ij}^{(m)}\|_{2,\epsilon}} - \frac{\lambda_{3,i}^2}{\|\boldsymbol{\lambda}_3 * (\boldsymbol{\beta}_{(j+1)}^{(m)} - \boldsymbol{\beta}_{(j)}^{(m)})\|_{2,\epsilon}}, \quad j = 1, \dots, n-1; \\
b_i^{(m)}(j) &= y_{ij}, \quad j = 1, \dots, n.
\end{aligned}$$

When staking measurements at different positions, the item 1 in $a_i^{(m)}(j, j)$ is replaced by δ_{ij} and $b_i^{(m)} = y_{ij}$ is replaced by $b_i^{(m)} = \delta_{ij}y_{ij}$.

Bias estimation

Let x_{ij} be the data for sequence i at locus j after σ_i of each sequence is normalized to 1. With such normalization, the model (3) is reduced to a simpler form with global tuning parameters to each sequence for easier interpretation:

$$f(\boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^N (x_{ij} - \beta_{ij})^2 + \lambda_1 \sum_{i=1}^M \sum_{j=1}^N |\beta_{ij}| + \lambda_2 \sum_{i=1}^M \sum_{j=2}^N |\beta_{ij} - \beta_{i,j-1}| + \lambda_3 \sum_{j=2}^N \left[\sum_{i=1}^M (\beta_{ij} - \beta_{i,j-1})^2 \right]^{\frac{1}{2}}. \quad (\text{S.1})$$

The solution to minimize $f(\boldsymbol{\beta})$ is unique for $f(\boldsymbol{\beta})$ is strictly convex. Denote the solution as $\hat{\boldsymbol{\beta}} = (\hat{\beta}_{ij})_{M \times N}$. Suppose sequence i is partitioned into \hat{K}_i consecutive segments $\{\hat{R}_1^{(i)}, \dots, \hat{R}_{\hat{K}_i}^{(i)}\}$,

delimited with change points $\hat{\mathcal{J}}_i = \{\hat{j}_1^{(i)}, \dots, \hat{j}_{\hat{K}_i-1}^{(i)}\} \subset \{2, \dots, N\}$ (left end of segment $2, \dots, \hat{K}_i$). The fitted means of each segment is denoted as $\hat{\boldsymbol{\mu}}^{(i)} = (\hat{\mu}_1^{(i)}, \dots, \hat{\mu}_{\hat{K}_i}^{(i)})$, i.e., $\hat{\beta}_{ij} = \hat{\mu}_k^{(i)}$, if $j \in \hat{R}_k^{(i)}$. The length (number of SNPs) of each segment is $\hat{L}_k^{(i)} = |\hat{R}_k^{(i)}|, k = 1, \dots, \hat{K}_i$. Thus, the estimated mean vector for sequence i can be written as

$$\hat{\boldsymbol{\beta}}_i = \sum_{k=1}^{\hat{K}_i} \hat{\mu}_k^{(i)} I_{\hat{R}_k^{(i)}}.$$

$\hat{\boldsymbol{\beta}}$ is the optimal solution if and only if it satisfies the subgradient condition $\partial f(\hat{\boldsymbol{\beta}}) = 0$; that is,

$$\hat{\beta}_{ij} = y_{ij} - \lambda_1 s_{ij}^{(1)} - \lambda_2 s_{ij}^{(2)} - \lambda_3 s_{ij}^{(3)}, \quad (\text{S.2})$$

where $s_{ij}^{(1)}, s_{ij}^{(2)}$ and $s_{ij}^{(3)}$ are coordinates of subgradient corresponding to β_{ij} 's appearing in each of the three penalty terms. Both bias estimation and asymptotic analysis rely on the analytic form of subgradient. Now we discussed the bias induced by each penalty separately.

Bias induced by lasso penalty

It is easy to verify that the subgradient for the lasso penalty can be written as

$$s_{ij}^{(1)} = \text{sign}(\beta_{ij}),$$

where, with a bit abuse of notation,

$$\text{sign}(x) = \begin{cases} 1, & \text{if } x > 0, \\ -1, & \text{if } x < 0, \\ z \in [-1, 1], & \text{if } x = 0. \end{cases} \quad (\text{S.3})$$

Hence, the lasso penalty term merely plays as a soft-thresholding on the fitted values resulted from the model (S.1) with $\lambda_1 = 0$, denoted as $\hat{\beta}_{ij}(0, \lambda_2, \lambda_3)$; that is, for any $\lambda_1 > 0$,

$$\hat{\beta}_{ij}(\lambda_1, \lambda_2, \lambda_3) = \text{sign} \left[\hat{\beta}_{ij}(0, \lambda_2, \lambda_3) \right] \left[\hat{\beta}_{ij}(0, \lambda_2, \lambda_3) - \lambda_1 \right]_+,$$

where $(x)_+ = \max\{x, 0\}$. This is also highlighted in Lemma A.1 of [15] for model (S.1) with $\lambda_3 = 0$.

Bias induced by fused-lasso penalty

In model (S.1) with $\lambda_1 = 0$ and $\lambda_3 = 0$ (only fused-lasso penalty involved), Lemma 2.1 in [31] gives an insightful characterization of $\hat{\mu}^{(i)}$:

$$\hat{\mu}_k^{(i)} = \frac{1}{\hat{L}_k^{(i)}} \sum_{j \in \hat{R}_k^{(i)}} x_{ij} + \hat{c}_k^{(i)}, \quad k = 1, \dots, \hat{K}_i,$$

where

$$\hat{c}_1^{(i)} = \begin{cases} -\frac{\lambda_2}{\hat{L}_1^{(i)}}, & \text{if } \hat{\mu}_2^{(i)} - \hat{\mu}_1^{(i)} > 0, \\ \frac{\lambda_2}{\hat{L}_1^{(i)}}, & \text{if } \hat{\mu}_2^{(i)} - \hat{\mu}_1^{(i)} < 0, \end{cases}$$

$$\hat{c}_{\hat{K}_i}^{(i)} = \begin{cases} \frac{\lambda_2}{\hat{L}_{\hat{K}_i}^{(i)}}, & \text{if } \hat{\mu}_{\hat{K}_i}^{(i)} - \hat{\mu}_{\hat{K}_i-1}^{(i)} > 0, \\ -\frac{\lambda_2}{\hat{L}_{\hat{K}_i}^{(i)}}, & \text{if } \hat{\mu}_{\hat{K}_i}^{(i)} - \hat{\mu}_{\hat{K}_i-1}^{(i)} < 0, \end{cases}$$

and, for $k = 2, \dots, \hat{K}_i - 1$,

$$\hat{c}_k^{(i)} = \begin{cases} \frac{2\lambda_2}{\hat{L}_k^{(i)}}, & \text{if } \hat{\mu}_k^{(i)} - \hat{\mu}_{k-1}^{(i)} < 0, \hat{\mu}_{k+1}^{(i)} - \hat{\mu}_k^{(i)} > 0, \\ -\frac{2\lambda_2}{\hat{L}_k^{(i)}}, & \text{if } \hat{\mu}_k^{(i)} - \hat{\mu}_{k-1}^{(i)} > 0, \hat{\mu}_{k+1}^{(i)} - \hat{\mu}_k^{(i)} < 0, \\ 0, & \text{if } (\hat{\mu}_k^{(i)} - \hat{\mu}_{k-1}^{(i)})(\hat{\mu}_{k+1}^{(i)} - \hat{\mu}_k^{(i)}) > 0. \end{cases}$$

The result implies that the sample mean (as an unbiased estimate of true mean) of a local minimum/maximum segment (except it is located at either end) is shifted towards 0 due to fused-lasso penalty. The bias is positively proportional to λ_2 and negatively proportional to the length of the segment. It is more important to notice that there exists no configuration where a local minimum/maximum segment has a jump size (relative to neighboring segments) less than the amount of bias. It means that a CNV with small jump size or small length could possibly be merged into neighboring segments, if λ_2 is set too large.

Bias induced by group-fused-lasso penalty

The subgradient for group-fused-lasso penalty is given in the following Proposition 1.

Proposition 1: The β_{ij} 's involved in group-fused-lasso penalty have subgradient given by

$$s_{ij}^{(3)} = \begin{cases} -e_{i2}, & \text{if } j = 1, \\ e_{ij} - e_{i,j+1}, & \text{if } 1 < j < N, \\ e_{iN}, & \text{if } j = N, \end{cases} \quad (\text{S.4})$$

for $i = 1, \dots, M$, where $\mathbf{e}_j = (e_{1j}, \dots, e_{Mj})^T$ for $j = 2, \dots, M$ are given by

$$\mathbf{e}_j = \begin{cases} \left(\frac{\beta_{1j} - \beta_{1,j-1}}{\|\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)}\|_{\ell_2}}, \dots, \frac{\beta_{Mj} - \beta_{M,j-1}}{\|\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)}\|_{\ell_2}} \right)^T, & \text{if } \|\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)}\|_{\ell_2} > 0, \\ \text{any } (e_{1j}, \dots, e_{Mj})^T \text{ s.t. } \|\mathbf{e}_j\|_{\ell_2} \leq 1, & \text{if } \|\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)}\|_{\ell_2} = 0. \end{cases} \quad (\text{S.5})$$

Proof: The proof follows a similar technique used in the proof of Lamma A.1 in [31]. Let $\mathbf{T} = [-\mathbf{I}_M, \mathbf{I}_M]$, where \mathbf{I}_M is $M \times M$ identity matrix. Then, for any $2 \leq j \leq N$,

$$h(\boldsymbol{\beta}_{(j-1)}, \boldsymbol{\beta}_{(j)}) \triangleq \|\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)}\|_{\ell_2} = \|\mathbf{T}[\boldsymbol{\beta}_{(j-1)}^T, \boldsymbol{\beta}_{(j)}^T]^T\|_{\ell_2}.$$

For the j such that $\|\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)}\|_{\ell_2} > 0$, the sub-gradient is reduced to regular gradient, and thus can be derived in a usual way. We now focus on the j such that $\|\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)}\|_{\ell_2} = 0$, i.e., the subgradient of β_{ij} at 0. By Cauchy-Schwartz inequality, we have

$$\begin{aligned} h(\boldsymbol{\beta}_{(j-1)}, \boldsymbol{\beta}_{(j)}) &\geq \|\mathbf{T}[\boldsymbol{\beta}_{(j-1)}^T, \boldsymbol{\beta}_{(j)}^T]^T\|_{\ell_2} \|\mathbf{e}_j\|_{\ell_2} \\ &\geq \langle \mathbf{T}[\boldsymbol{\beta}_{(j-1)}^T, \boldsymbol{\beta}_{(j)}^T]^T, \mathbf{e}_j \rangle \\ &= h(\mathbf{0}) + \langle [\boldsymbol{\beta}_{(j-1)}^T, \boldsymbol{\beta}_{(j)}^T]^T - \mathbf{0}, \mathbf{T}^T \mathbf{e}_j \rangle \end{aligned}$$

where \mathbf{e}_j is any vector such that $\|\mathbf{e}_j\|_{\ell_2} \leq 1$. It follows by the definition of subgradient that $\mathbf{T}^T \mathbf{e}_j = [-\mathbf{e}_j^T, \mathbf{e}_j^T]^T$ is the subgradient for $[\boldsymbol{\beta}_{(j-1)}^T, \boldsymbol{\beta}_{(j)}^T]^T$. \square

The bias induced by the group-fused-lasso penalty can be derived from the analytic form of subgradient accordingly and is given in the following Proposition 2.

Proposition 2: In model (S.1) with $\lambda_1 = 0$ and $\lambda_2 = 0$, the fitted means of segments for sequence i can be expressed as

$$\hat{\mu}_k^{(i)} = \frac{1}{\hat{L}_k} \sum_{j \in \hat{R}_k^{(i)}} x_{ij} + \hat{c}_k^{(i)}, \quad k = 1, \dots, \hat{K}_i,$$

where

$$\hat{c}_k^{(i)} = \begin{cases} \frac{\lambda_3}{\hat{L}_1^{(i)}} \cdot r_i(\hat{j}_1^{(i)}), & \text{if } k = 1, \\ -\frac{\lambda_3}{\hat{L}_k^{(i)}} \cdot [r_i(\hat{j}_{k-1}^{(i)}) - r_i(\hat{j}_k^{(i)})], & \text{if } 2 \leq k \leq \hat{K}_i - 1, \\ -\frac{\lambda_3}{\hat{L}_{\hat{K}_i}^{(i)}} \cdot r_i(\hat{j}_{\hat{K}_i-1}^{(i)}), & \text{if } k = \hat{K}_i, \end{cases}$$

and

$$r_i(j) \triangleq \frac{\hat{\beta}_{ij} - \hat{\beta}_{i,j-1}}{\|\hat{\beta}_{(j)} - \hat{\beta}_{(j-1)}\|_{\ell_2}}.$$

Proof: The proof follows a similar technique used in the proof of Lemma 2.1 in [31]. Following the subgradient condition (S.2) in case $\lambda_1 = 0$ and $\lambda_2 = 0$, we have

$$\hat{\mu}_k^{(i)} = \frac{1}{\hat{L}_k} \sum_{j \in \hat{R}_k^{(i)}} \hat{\beta}_{ij} = \frac{1}{\hat{L}_k} \sum_{j \in \hat{R}_k^{(i)}} x_{ij} - \frac{\lambda_3}{\hat{L}_k} \sum_{j \in \hat{R}_k^{(i)}} s_{ij}^{(3)}.$$

By Proposition 1 and simple algebra, we have

$$\sum_{j \in \hat{R}_k^{(i)}} s_{ij}^{(3)} = \begin{cases} -e_{i,\hat{j}_1^{(i)}}, & \text{if } k = 1, \\ e_{i,\hat{j}_{k-1}^{(i)}} - e_{i,\hat{j}_k^{(i)}}, & \text{if } 2 \leq k \leq \hat{K}_i - 1, \\ e_{i,\hat{j}_{\hat{K}_i-1}^{(i)}}, & \text{if } k = \hat{K}_i. \end{cases}$$

Note that at jump points, subgradient has explicit form as shown in Proposition 1. It follows that $e_{i,\hat{j}_k^{(i)}} = r_i(\hat{j}_k^{(i)})$, for $k = 1, \dots, \hat{K}_i - 1$, where $r_i(\cdot)$ is defined in Proposition 2. \square

Some interesting implications follow immediately. For sequence i , consider one of its fitted segment k with end points $[\hat{j}_{k-1}^{(i)}, \hat{j}_k^{(i)} - 1]$. If no other sequences share change points at these two

ends, then the bias term $\hat{c}_k^{(i)}$ reduces to what it appears in model (S.1) with fused-lasso term only ($\lambda_1 = 0$ and $\lambda_3 = 0$). If m out of M sequences share change points at these two ends and also assume the jump size at these two locations for all the m sequences are roughly the same, then the absolute value of the bias term can be approximately written as $\frac{2\lambda_3}{\hat{L}_k^{(i)}} \cdot \frac{1}{\sqrt{m}}$. It means that if more than one sequences share change points at the same coordinate, then they can benefit from each other to reduce their individual bias, relative to the bias induced by fused-lasso penalty specific to each individual sequence.

Asymptotic behavior

Now we try to give a justification of the order of the magnitude of λ_2 and λ_3 in compatible with their large sample behavior, say, as $N \rightarrow \infty$. When the number of sequences M in segmentation task is relatively large, extra caution is needed for λ_3 . Again, we discuss asymptotic behavior of the solution influenced by fused-lasso and group-fused-lasso separately for easier exhibition.

Asymptotic behavior for fused-lasso penalty

In fused-lasso model ($\lambda_1 = 0$ and $\lambda_3 = 0$), the justification is directly inspired by the proof of Theorem 2.3 in [31]. Denote the event

$$\mathcal{E}_i = \{\hat{\mathcal{J}}_i = \mathcal{J}_i\} \cap \{\text{sign}(\hat{\beta}_{ij} - \hat{\beta}_{i,j-1}) = \text{sign}(\beta_{ij} - \beta_{i,j-1}), \forall j \in \mathcal{J}_i\},$$

for $i = 1, \dots, M$ respectively. This event means that all jump points and the direction of jumps are correctly identified for each sequence i . A necessary condition required for λ_2 is summarized in Proposition 3.

Proposition 3: It is required that $\lambda_2 = O(\sqrt{\log N})$ to ensure $\lim_{N \rightarrow \infty} \mathbb{P}(\mathcal{E}_i) = 1$ for $i = 1, \dots, M$, at the linear rate.

This asymptotic behavior follows directly the proof of Theorem 2.3 in [31]. We have some quick remarks:

- 1) If the signal of each sequence is not normalized, then $\lambda_{2,i} = c_2\sigma_i\sqrt{\log N}$, specific to sequence i .
- 2) In order to ascertain a CNV segment with length L and jump size δ , the bias needs to satisfy $\frac{2\lambda_{2,i}}{L} = \frac{2c_2\sigma_i\sqrt{\log N}}{L} < \delta$, i.e., $c_2 < \frac{1}{2\sqrt{\log N}} \cdot \frac{\delta}{\sigma_i}L$. Here, $\frac{\delta}{\sigma_i}$ can be interpreted as signal-to-noise ratio (SNR). For a specific platform, one may get a sense of the magnitude of SNR and L from prior knowledge. In practice, it is desired to take as large value of c_2 as possible to ensure the sparsity of the segmentation, but not too large in order to compensate for the constraint of signal strength ($\frac{\delta}{\sigma_i}L$). Based on our experiences of analysis of Illumina data [54], the results are not sensitive to the choice of c_2 , provided that it falls into a reasonable range.

Asymptotic behavior for group-fused-lasso penalty

In group-fused-lasso model ($\lambda_1 = 0$ and $\lambda_2 = 0$), we have similar requirement of λ_3 as for λ_2 , which is given in Proposition 4.

Proposition 4: It is required that $\lambda_3 = O(\sqrt{M}\sqrt{\log N})$ to ensure $\lim_{N \rightarrow \infty} \mathbb{P}(\cap_{i=1}^M \mathcal{E}_i) = 1$, at the linear rate.

Proof: For simplicity, we prove under the condition that ϵ_{ij} are i.i.d. $\mathcal{N}(0, 1)$ (after σ_i is normalized to 1), while this condition can be relaxed [31]. We also follow the same technique used in the proof of Theorem 2.3 in [31]. Let $d_{ij} = \beta_{ij} - \beta_{i,j-1}$, $\hat{d}_{ij} = \hat{\beta}_{ij} - \hat{\beta}_{i,j-1}$, and $d_{ij}^\epsilon = \epsilon_{ij} - \epsilon_{i,j-1}$. Also denote $\mathbf{d}_j^\epsilon = (d_{1j}^\epsilon, \dots, d_{Mj}^\epsilon)^T$ and $\mathcal{J} = \cup_{i=1}^M \mathcal{J}_i$. By the subgradient condition (S.2), for each i , \mathcal{E}_i holds if and only if

$$d_{ij}^\epsilon = \lambda_3(2e_{ij} - e_{i,j-1} - e_{i,j+1}), \quad \text{for } j \in \mathcal{J}_i^c, \quad (\text{S.6})$$

and

$$|\hat{d}_{ij}| > 0, \quad \text{for } j \in \mathcal{J}_i. \quad (\text{S.7})$$

Condition (S.7) has direct relevance to the bias issue, as discussed above. Now we focus on condition (S.6), which implies that

$$\max_{j \in \mathcal{J}^c} \|\mathbf{d}_j^\epsilon\|_{\ell_2} = \max_{j \in \mathcal{J}^c} \lambda_3 \|2\mathbf{e}_j - \mathbf{e}_{j-1} - \mathbf{e}_{j+1}\|_{\ell_2} < 4\lambda_3.$$

It is left to show that $\mathbb{P}(\max_{j \in \mathcal{J}^c} \|\mathbf{d}_j^\epsilon\|_{\ell_2} \geq 4\lambda_3) = \mathbb{P}(\max_{j \in \mathcal{J}^c} \|\mathbf{d}_j^\epsilon/\sqrt{2}\|_{\ell_2}^2 \geq 8\lambda_3^2) \rightarrow 0$ as $N \rightarrow \infty$ for $i = 1, \dots, M$. Note that for each j , $d_{1j}^\epsilon, \dots, d_{Mj}^\epsilon$ are i.i.d. $\mathcal{N}(0, 2)$, so $\|\mathbf{d}_j^\epsilon/\sqrt{2}\|_{\ell_2}^2 \sim \chi_M^2$.

Then we have

$$\begin{aligned} & \mathbb{P}(\max_{j \in \mathcal{J}^c} \|\mathbf{d}_j^\epsilon/\sqrt{2}\|_{\ell_2}^2 \geq 8\lambda_3^2) \\ &= \mathbb{P}(\cup_{j \in \mathcal{J}^c} \|\mathbf{d}_j^\epsilon/\sqrt{2}\|_{\ell_2}^2 \geq 8\lambda_3^2) \\ &\leq \sum_{j \in \mathcal{J}^c} \mathbb{P}(\|\mathbf{d}_j^\epsilon/\sqrt{2}\|_{\ell_2}^2 \geq 8\lambda_3^2) \\ &= |\mathcal{J}^c| \mathbb{P}(\|\mathbf{d}_j^\epsilon/\sqrt{2}\|_{\ell_2}^2 \geq 8\lambda_3^2) \\ &\leq \exp \left[-\frac{1}{2}(8\lambda_3^2 - M) + \log |\mathcal{J}^c| - \frac{M}{2} \log \frac{M}{8\lambda_3^2} \right]. \end{aligned}$$

Here the first inequality is due to union bound and the second inequality is due to Chernoff's bound for χ_M^2 distribution. Under the assumption on sparsity of the change points, we have $|\mathcal{J}^c| = O(N)$ for fixed M . In our settings, M is fixed (which may rise up to thousands) while $N \rightarrow \infty$, yet in practice, M is not negligible with respect to $\sqrt{\log N}$. For example, $\sqrt{\log(10^6)} \approx 3.72$, and it is not uncommon to have more than 4 sequences for joint segmentation. Therefore, it is necessary to have $\lambda_3 = O(\sqrt{M}\sqrt{\log N})$. \square

We also have some remarks on how to determine λ_3 :

- 1) If the signal of each sequence is not normalized, then $\lambda_{3,i} = c_3 \sigma_i \sqrt{pM} \sqrt{\log N}$. The choice of p is decided case by case and discussed in the main text.
- 2) Following the above discussion about bias induced by group-fused-lasso penalty, if m out of M sequences carry CNVs with exactly the same boundary, the bias can be approximately

written as $\frac{2c_3\sigma_i\sqrt{\log N}}{\hat{L}_k^{(i)}} \cdot \frac{\sqrt{pM}}{\sqrt{m}}$. On one hand, if p is over estimated so that pM is much larger than m , the model would be over penalized and introduce more bias than that is attributed to individual fused-lasso penalty, and thus does not benefit from joint analysis; On the other hand, if pM is set too small, we have insufficient control on the sparsity of each sequence, so that it has to be compensated by the fused-lasso penalty. This is the reason why we need to incooperate $\rho(p)$ to re-weight the relative influence of the two penalties.

Details in calling procedure

We specify the likelihood functions of LRR and BAF signals in the log-likelihood ratio (8) as follows. For BAF signal, the likelihood is usually modeled for different copy number states as a mixture of densities surrounding a few possible BAF values corresponding to different genotypes [9, 47]. When population frequencies for allele A and B, p_A and p_B , are available or can be estimated from data, we have

$$L_{\text{BAF}}(x; c) = \sum_{s=0}^c \binom{c}{s} p_A^{c-s} p_B^s \phi_s(x; \mu_s, \sigma_s^2), \quad \text{for } c = 0, 1, 2, 3, 4,$$

where $\phi_s(\cdot; \mu_s, \sigma_s^2)$ is normal density for state s . The details in model and parameter specification are listed in Supplementary Table 1.

In case where population frequencies p_A and p_B are not available, we might use an alternative likelihood function for BAF [54], defined by

$$L_{\text{BAF}}(x; c) = \max_{s \in \{0, \dots, c\}} \phi_s(x; \mu_s, \sigma_s^2), \quad \text{for } c = 0, 1, 2, 3, 4,$$

where all parameters are defined in the same way (see Supplementary Table 1).

For LRR signal, the likelihood function is simply defined by normal density:

$$L_{\text{LRR}}(y; c) = \phi(y; \mu_c, \sigma_c^2).$$

c	s	Genotype	$\phi_s(\cdot)$	μ_s	σ_s
0	0	Null	normal	1/2	$10\hat{\sigma}_x$
1	0, 1	A, B	half normal	0, 1	$\hat{\sigma}_x$
2	0, 2	AA, BB	half normal	0, 1	$\hat{\sigma}_x$
	1	AB	normal	1/2	$\hat{\sigma}_x$
3	0, 3	AAA, BBB	half normal	0, 1	$\hat{\sigma}_x$
	1, 2	AAB, ABB	normal	1/3, 2/3	$\hat{\sigma}_x$
4	0, 4	AAAA, BBBB	half normal	0, 1	$\hat{\sigma}_x$
	1, 2, 3	AAAB, AABB, ABBB	normal	1/4, 1/2, 3/4	$\hat{\sigma}_x$

Supplementary Table 1: Model and parameter specification in BAF signal for each copy number state. $\hat{\sigma}_x$ is empirically estimated from BAF values in (0.4, 0.6) for each individual.

For $c = 0, 1, 3, 4$, μ_c and σ_c^2 are estimated based on the data y_R in segment R being considered, while μ_2 and σ_2^2 are estimated from the data of the whole chromosome on which segment R locates or, locally, from the data of a few hundred markers flanking the segment.

Additional Results

Region	Aberration Type	Chr	bp Start	bp End	#SNP	#hetSNP
1	CN-LOH	5	1	47700000	9397	2756
2	Loss	5	111789971	112521346	156	79
3	Gain	8	1	45200000	12564	3830
4	Gain	8	128432670	129207869	218	91
5	Loss	9	1	50600000	11201	3889
6	Loss	10	84504379	94825178	1988	648
7	Gain	12	1	132449811	27131	8818
8	Loss	13	31766569	31892852	37	10
9	CN-LOH	17	7431864	11747138	1150	308
10	CN-LOH	17	22300000	78774742	9713	3205
Total number of modified heterozygous SNPs						23634
Total number of heterozygous SNPs on autosome						176207
Total number of SNPs on autosome						547359

Supplementary Table 2: Regions of allelic imbalance imposed to the HapMap sample NA06991 [39].

Method	Time per sample in sec. (mean (std dev))
GFL	21.97 (1.31)
BAFsegmentation	41.73 (-)
PSCN	1154.18 (74.73)

Supplementary Table 3: Speed comparison of three methods: GFL, BAFsegmentation and PSCN.

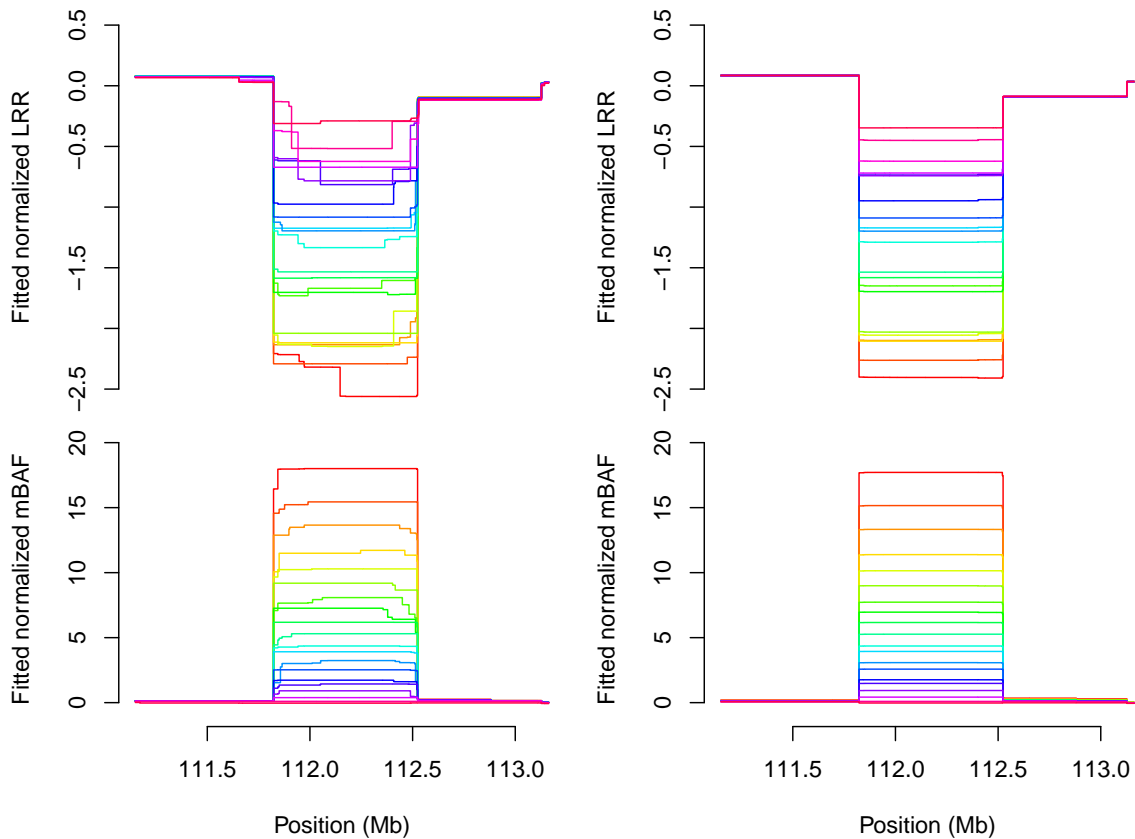
Sample	Gender	Ancestry	Resource	Type	<10k	10–50k	50–100k	>100k	Total
NA15510	Female	European	PDR	loss	12	25	3	7	47
				gain	0	0	1	4	5
				total	12	25	4	11	52
NA18517	Female	YRI	HapMap	loss	10	22	4	4	40
				gain	1	3	1	8	13
				total	11	25	5	12	53
NA18576	Female	CHB	HapMap	loss	13	16	4	5	38
				gain	0	2	2	4	8
				total	13	18	6	9	46
NA18980	Female	JPT	HapMap	loss	8	16	1	4	29
				gain	0	0	1	3	4
				total	8	16	2	7	33

Supplementary Table 4: Sample information and reference CNV regions summarized for each sample by their types and sizes. The ancestry of NA15510 was not recorded but inferred in [20].

Abbreviation: PDR = Polymorphism Discovery Resource.

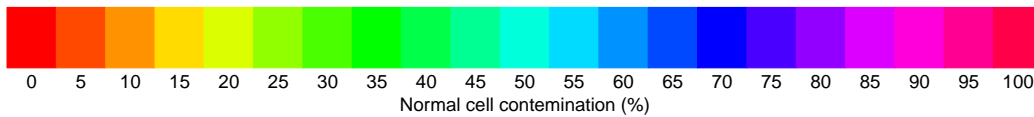
			NA15510		NA18517		NA18576		NA18980		
Analysis	ρ	M	# Det.	# Ovlp.	# Det.	# Ovlp.	# Det.	# Ovlp.	# Det.	# Ovlp.	Time (min.)
Analysis A: GFL done on averaged signal for each platform											
O1Q	1	1	92	34	73	22	71	21	69	20	0.3
O2Q	1	1	114	22	92	24	111	15	95	11	0.9
Union	-	-	170	38	144	34	160	25	145	22	1.2
Analysis B: GFL done on averaged signal of both platforms jointly											
	0	2	128	40	108	33	96	21	104	23	4.2
Analysis C: GFL done on three replicates separately for each platform											
O1Q	1	1	66	31	65	22	43	19	48	15	0.9
O2Q	1	1	68	23	65	22	65	12	59	13	2.8
Union	-	-	102	36	109	33	93	25	91	20	3.7
Analysis D: GFL done on three replicates jointly for each platform											
O1Q	0	3	64	32	66	22	54	21	53	18	1.1
O2Q	0	3	75	22	70	24	65	11	49	12	3.1
Union	-	-	106	36	115	33	96	22	83	21	4.2
Analysis E: GFL done on three replicates of both platforms jointly											
	0	6	80	38	82	32	69	25	56	15	8.5
MPCBS: Segmentation done on three replicates of both platforms jointly											
	-	-	98	34	88	28	59	18	68	21	313.9

Supplementary Table 5: Number of CNVs detected (Det.) and overlapping (Ovlp.) with reference results as well as average computation time for four samples under different analyses. Tuning parameters used in segmentation: $c_1 = 0.1$, $c_2 = 2$, $c_3 = 2$ and $p = 1$; ρ and M are specified for each analysis. Analysis A, C and E correspond to Analysis 1, 2 and 3 respectively in Table 2 of main text.

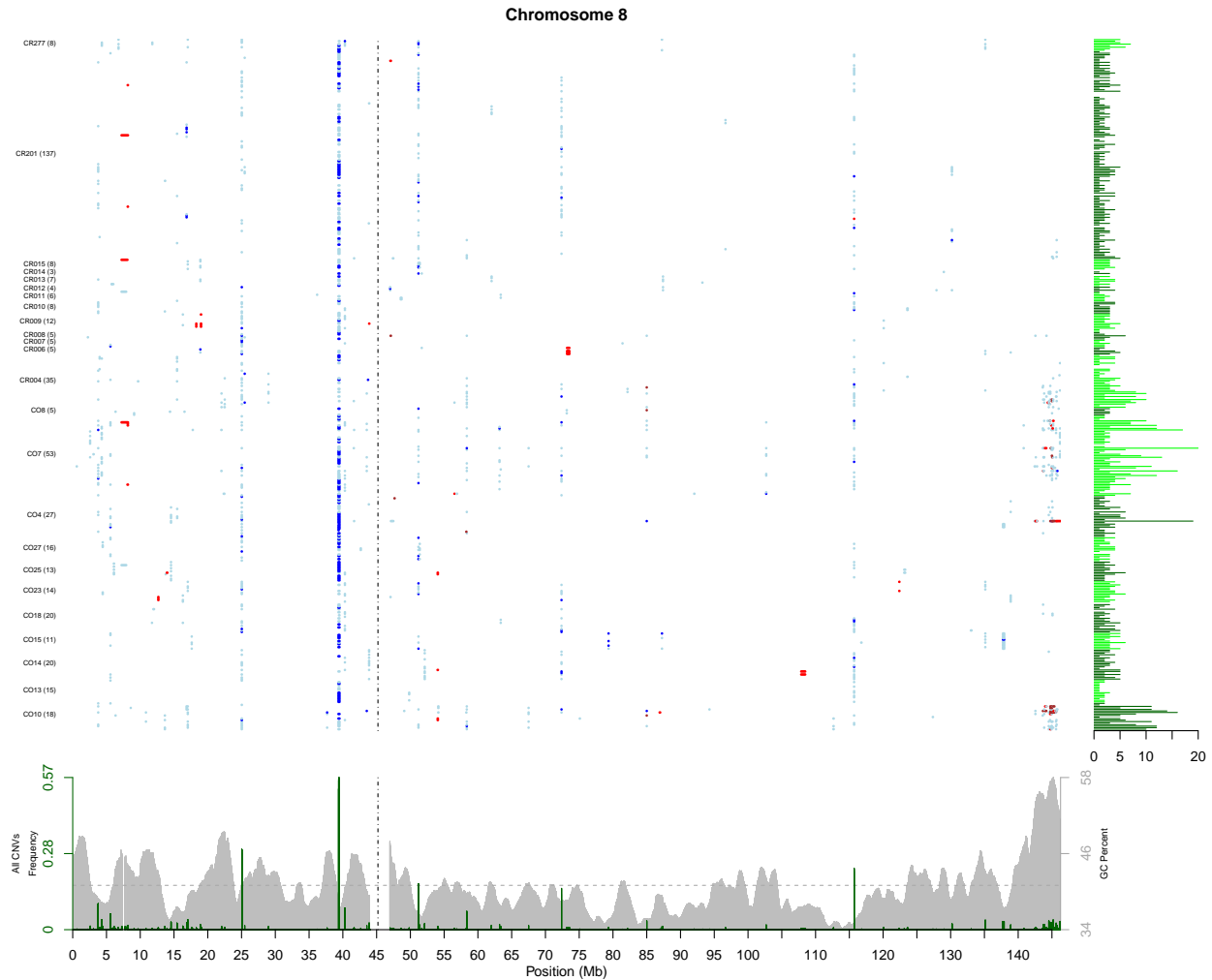


(a) Individual analysis

(b) Joint analysis



Supplementary Figure 1: Comparison of fitted profiles between analysis for each tumor sample with different normal cell contamination levels and joint analysis for all 21 tumor samples. Shown is a hemizygous loss on Chromosome 5q22. In each of the subplots, the upper panel shows the fitted profiles on LRR for each sample distinctly marked by a spectrum of colors, while the lower panel shows their corresponding fitted profiles on mBAF. Shown are data points for heterozygous makers. (a) Individual analysis; (b) Joint analysis.



Supplementary Figure 2: Visualization of pedigree-wise CNV analysis results of Chromosome 8 data in bipolar disorder study. In the main body of the plot, CNVs estimated for each individual are marked by small segments with color code: CN=0 in blue, CN=1 in light blue, CN=3 in red and CN=4 in brown. Each subject is a row, each SNP a column. Subjects belonging to the same pedigree are stacked together. The pedigree names are indicated on the left hand side with the number of pedigree members included in parentheses. On the right hand side, the barplot represents the number of CNV detected per subject. Two shades of green are switched alternately to indicate the pedigree to which the subject belongs. At the bottom, the gray histogram shows the GC content along the chromosome; coordinated with the representation of CNVs in the main body, the green histogram counts the frequency of CNV among the subjects represented. Vertical dotted line marks the centromere.