SUPPLEMENTARY INFORMATION

1 PROPERTIES OF THE PWC REPRESENTATION

The representation presented in section 2.3 was initially inspired by the concept of wavelet footprints (Dragotti and Vetterli, 2002) where the more general case of piece-wise polynomial signals is considered from a wavelet analysis perspective. The *maximally sparse* representation for PWC signals demonstrated in wavelet footprints is reformulated here using standard linear algebra and extended to arbitrary vector lengths. This also allows us to establish a correspondence between sets of breakpoints and a *nested structure* of vector subspaces which we use here to demonstrate the representation properties.

Mathematically, a PWC vector $\boldsymbol{x}$ can be completely characterized by its change locations (i.e., breakpoints) and the constant values of the regions in between (i.e., segment amplitudes):

DEFINITION 1 (PWC vector). *A piece-wise constant vector $\boldsymbol{x} = (x_1, \ldots, x_M)^t$ is characterized by an ordered set of K discontinuity locations $\mathcal{I} = \{i_1 < i_2 \ldots < i_K\} \subset \{1, \ldots, M-1\}$ (i_k denotes the beginning position of segment k) and a vector with the corresponding $K{+}1$ segment amplitudes $\boldsymbol{a} = (a_0, \ldots, a_K)^t$. Thus:*

$$\boldsymbol{x}^t = \left(\underset{\uparrow 1}{a_0}, \ldots, a_0, \underset{\uparrow i_1}{a_1}, \ldots, a_{k-1}, a_{k-1}, \underset{\uparrow i_k}{a_k}, a_k, \ldots, \underset{\uparrow M}{a_K} \right) \tag{22}$$

With this definition it is easy to show that the breakpoint sets $\mathcal{I}$s induce the following vector subspace properties:

LEMMA 1 (PWC Vector Subspaces). *Let $\mathcal{S}_\mathcal{I}$ be the set of all PWC vectors $\mathbf{x}$ that have breakpoint locations contained in $\mathcal{I}$, and segment amplitudes $\boldsymbol{a} \in \mathbb{R}^{K+1}$. Then, we have that:*
i) $\mathcal{S}_\mathcal{I}$ is a subspace of $\mathbb{R}^M$ of dimension $K+1$.
ii) $\mathcal{S}_{\mathcal{I}_1}$ is a subspace of $\mathcal{S}_{\mathcal{I}_2}$ if and only if $\mathcal{I}_1 \subset \mathcal{I}_2$

PROOF. It is clear that i) holds since, first, for any $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ with breakpoints in $\mathcal{I}$, but different amplitudes $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$; we have that $\boldsymbol{x}_3 = \boldsymbol{x}_1 + \boldsymbol{x}_2$ may remove existing breakpoints but never create a breakpoint outside $\mathcal{I}$, thus $\boldsymbol{x}_3 \in \mathcal{S}_\mathcal{I}$ because it will always have breakpoints contained in the same $\mathcal{I}$, and $\boldsymbol{a}_3 = \boldsymbol{a}_1 + \boldsymbol{a}_2$. Second, for any $\boldsymbol{x}_1 \in \mathcal{S}_\mathcal{I}$ and for all α, $\boldsymbol{x}_4 = \alpha \boldsymbol{x}_1$ will also have breakpoints contained in $\mathcal{I}$ with $\boldsymbol{a}_4 = \alpha \boldsymbol{a}_1$ and thus will belong to $\mathcal{S}_\mathcal{I}$. Furthermore, when $\mathcal{I}$ is fixed $\mathbf{x}$ and $\mathbf{a}$ vector spaces are isomorphic and hence $\mathcal{S}_\mathcal{I}$ has dimension $K+1$; thus, ii) readily follows from i). □

Part ii) of the lemma is equivalent to saying that any PWC vector $\mathbf{x} \in \mathcal{S}_\mathcal{I}$ can be represented as a linear combination of step vectors in $\mathcal{S}_{\{k\}}$, $k = i_1, \ldots, i_K$. With this principle in mind, we now introduce a basis for PWC signal representation that has some desirable properties.

THEOREM 1 (PWC Basis). *Define a matrix $\boldsymbol{F} = [\boldsymbol{f}_0, \boldsymbol{f}_1, ..., \boldsymbol{f}_{M-1}]$, with columns $\boldsymbol{f}_i$ where $\boldsymbol{f}_0 = \mathbf{1}_M/\sqrt{M}$ is a constant vector and the remaining columns are step vectors:*

$$\boldsymbol{f}_i(m) = \begin{cases} -\sqrt{\frac{M-i}{iM}} & m \leq i \\ \sqrt{\frac{i}{M(M-i)}} & m > i \end{cases} \tag{23}$$

Then, we have the following properties:

i) (Complete Basis): *The columns of $\mathbf{F}$ are a basis for $\mathbb{R}^M$, i.e., for any $\boldsymbol{x} \in \mathbb{R}^M$ there exists a unique $\mathbf{w}$ such that $\boldsymbol{x} = \boldsymbol{F}\boldsymbol{w}$.*

ii) (Nested Structure): *The columns of $\boldsymbol{F}_\mathcal{I} = [\boldsymbol{f}_0, \boldsymbol{f}_{i_1}, \ldots, \boldsymbol{f}_{i_K}]$ are a basis for the vector subspace $\mathcal{S}_\mathcal{I}$, formed by PWC vectors with breakpoints at $\mathcal{I} = \{i_1 < i_2 \ldots < i_K\}$.*

iii) (Maximal Sparseness): *Any PWC vector $\boldsymbol{x} \in \mathcal{S}_\mathcal{I}$, can be written as $\mathbf{x} = \mathbf{F}\mathbf{w}$, where $\mathbf{w}$ has as many as $|\mathcal{I}| + 1$ non-zero entries, which is the minimal amount possible (maximal sparseness). Moreover, if the non-zero weights are $\boldsymbol{w}_\mathcal{I} = [w_0, w_{i_1}, \ldots, w_{i_K}]$, we can write $\boldsymbol{x} = \boldsymbol{F}_\mathcal{I}\boldsymbol{w}_\mathcal{I}$, where the subscript $\mathcal{I}$ denotes that only the columns of $\mathbf{F}$ (resp. components of $\mathbf{w}$) at the positions corresponding to the indices in $\mathcal{I}$ are included.*

In order to better understand the previous theorem, we will explain how the basis has been constructed. First, if $\mathbf{x}$ is a constant vector, i.e., it has no discontinuities, $\mathcal{I} = \emptyset$, the dimension of $\mathcal{S}_0 = \mathcal{S}_{\mathcal{I}=\emptyset}$ is one and can be spanned by the constant vector $\boldsymbol{f}_0$. Then, for $k = 1, \ldots, M-1$ the vector spaces $\mathcal{S}_k = \mathcal{S}_{\mathcal{I}=\{k\}}$ of PWC vectors with a single discontinuity between k and $k+1$ can be spanned by adding the element $\boldsymbol{f}_k$, a step vector with a breakpoint at that position. Moreover, the set of vectors now forms a complete basis: from Lemma 1.ii) any $\mathbf{x} \in \mathcal{S}_\mathcal{I}$ can be represented by linearly combining $\{\boldsymbol{f}_0, \boldsymbol{f}_{i_1}, \ldots, \boldsymbol{f}_{i_K}\}$. This proves ii) in the theorem, as well as i) when $\mathcal{I} = \{1, \ldots, M\}$.

Clearly i) holds, since $\mathbf{F}$ is a square invertible matrix, with the rows of its inverse $\boldsymbol{F}^{-1}$ forming the *dual basis*:

$$\begin{aligned} \boldsymbol{F}^{-1} &= \left[\tilde{\boldsymbol{f}}_0, \ldots, \tilde{\boldsymbol{f}}_{M-1}\right]^t \\ \tilde{\boldsymbol{f}}_0 &= \frac{1}{\sqrt{M}} \mathbf{1}_M \\ \tilde{\mathbf{f}}_k(m) &= \begin{cases} -\sqrt{\frac{k(N-k)}{N}} & m = k-1 \\ \sqrt{\frac{k(N-k)}{N}} & m = k \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{24}$$

The most appealing property for our specific application is iii), since copy number vectors will have very few breakpoints ($K << M$) which makes $\mathbf{w}$ a sparse representation. We can prove iii) by the following argument. First, we cannot have less than $K+1$ non-zero elements because this is the minimum required to form a basis for $\mathcal{S}_\mathcal{I}$. Then, for all $m \notin \mathcal{I}$, we have that $x_m - x_{m-1} = 0$, and using the dual basis, $\boldsymbol{w} = \boldsymbol{F}^{-1}\boldsymbol{x}$, we have that for all $m > 0$ $w_m = 0$ if and only if $x_m - x_{m-1} = 0$. Thus, there are exactly $K+1$ non-zero elements, which is indeed the minimum (so the representation is maximally sparse).

2 ALTERNATIVE APPROACHES TO SBL TO EXPLOIT THE PWC REPRESENTATION

Similar problems that involve solving the minimizing problem stated in (8) have already been addressed both in the signal processing and in the statistics literature, leading to similar strategies and methods which are given different names (see Table 4). However, as is discussed in this section, most of these existing methods are severely limited by the high collinearity/coherence (lack of orthogonality) between the columns of $\mathbf{F}$.

The first class of strategies, greedy methods, consists of reducing the search space of all possible predictor subsets 2^M by assuming that the best set of K_1 predictors will often be a subset of the best

Table 4. Relationship between signal processing methods for overcomplete expansions and methods in statistics for variable selection in multiple regression

		Signal Processing	Statistics
Greedy Methods:			
	MP-FS	Matching Pursuit (Mallat and Zhang, 1993)	Forward Selection (Seber and Lee, 2003)
	OMP	Orthogonal Matching Pursuit (Pati *et al.*, 1993)	
Relaxation methods:			
	MoF-Ridge	Method of Frames (Chen *et al.*, 1998)	Ridge regression (Hastie *et al.*, 2001)
	BP-Lasso	Basis Pursuit (Chen *et al.*, 1998)	Lasso (Hastie *et al.*, 2001)

Methods are paired when a similar version of equation (7) is solved (i.e., when the same metrics are chosen). But note that there will be differences in how λ is adjusted, and the size or types of design matrices $\mathbf{F}$ that are used.

set of K_2 predictors, for $K_2 > K_1$. If this assumption is correct, the set of best predictors can be constructed sequentially as in MP-FS; where we start selecting the vector (regressor) with largest projection (largest F-score), and keep adding the vector that most reduces the energy of the residual. This strategy is only optimal when $\mathbf{F}$ is orthogonal, or nearly optimal (Donoho *et al.*, 2006) when the coherence of $\mathbf{F}$ ($C = \max \langle \boldsymbol{f}_k, \boldsymbol{f}_j \rangle \; k \neq j$) is small and the signal is "sufficiently" sparse (i.e., $\|\boldsymbol{w}\|_0$ is small). It is important to note that this result cannot be applied to our case, since the coherence of $\mathbf{F}$ approaches 1, i.e., the set of vectors considered here is highly coherent.

The second class of strategies is based on replacing the l_0 sparseness measure $\|w\|_0$ by some other l_p measure, such that more efficient optimization methods (such as linear programming, projection or gradient methods) can be used. For example, for $p = 2$ (i.e., $\|w\|_2$), we would have a ridge regression in which the two square norms can be easily combined resulting in $\hat{\boldsymbol{w}}_{\text{rigde}} = \left(\boldsymbol{F}^t\boldsymbol{F} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{F}^t\boldsymbol{y}$. However, $\hat{\boldsymbol{w}}_{\text{ridge}}$ is not sparse at all in l_0 sense, and thus we would be interested in using a small as possible p. The l_1 norm is often used, because it is is the minimal one for which the constraints form a convex set and thus convex optimization or linear programming can be used to solve the problem. This is the strategy behind basis pursuit (Chen *et al.*, 1998) and lasso (Hastie *et al.*, 2001), for which there exist a similar result as in MP (Donoho *et al.*, 2006) showing that if the coherence is small then minimizing for l_1 is equivalent to minimizing for l_0. Therefore, when $\mathbf{F}$ is highly coherent, as in our case, these techniques lead to sub-optimal performance and a new approach is needed.

2.1 Comparison to SBL approach

In previous work (Pique-Regi *et al.*, 2007) we demonstrated using the same performance evaluation procedure as in section 3.1 (Willenbrock and Fridlyand, 2005) that the SBL approach presented in this paper had a significantly superior performance than the techniques in Table 4 for the specific application of CNA detection.

The precision-recall operating curves (PROC) were generated for each approach by obtaining the sensitivity and FDR for detected CNA in the simulated CGH dataset at each operating point (Figure 7). SBL had the best performing PROC curve as compared to other approaches for all given values of δ (data shown only for $\delta = 2$).

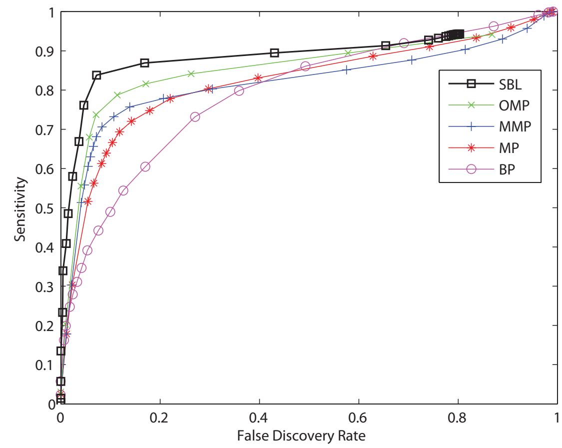

Fig. 7. PROC operational curves for sensitivity vs. false discovery rate in detecting real copy number changes within a $\delta = 2$ sample precision window in the dataset introduced by Willenbrock and Fridlyand (2005).

3 SBL ALGORITHM DETAILS

3.1 The role of the parameter a in SBL

The parameter a controls the shape of the prior distribution over the weights $p(\boldsymbol{w})$ specified by the hierarchical prior defined by $p(\boldsymbol{w}|\boldsymbol{\alpha})$ (12) and $p(\boldsymbol{\alpha})$ (13). Following Tipping (2001), the $\boldsymbol{\alpha}$ hyperparameters can be integrated out to find the marginal "effective" prior $p(\boldsymbol{w})$:

$$\begin{aligned} p(\boldsymbol{w}) &= \int p(\boldsymbol{w}|\boldsymbol{\alpha})\, p(\boldsymbol{\alpha})\, d\boldsymbol{\alpha} \\ &= \prod_{m=1}^{M-1} \int p(w_m|\alpha_m)\, p(\alpha_m)\, d\alpha_m \\ &= \prod_{m=1}^{M-1} p(w_m) \end{aligned} \tag{25}$$

where $p(w_m)$ is:

$$\begin{aligned} p(w_m) &= \int p(w_m|\alpha_m)\, p(\alpha_m)\, d\alpha_m \\ &= \frac{\Gamma(a+1/2)}{\Gamma(a)\sqrt{2\pi a}} \sqrt{\frac{a}{b}} \left(1 + \frac{w^2}{2b}\right)^{-\left(\frac{1}{2}+a\right)} \end{aligned} \tag{26}$$

a t-distribution with $2a$ degrees of freedom and a scale parameter of $\sqrt{a/b}$. When $b \to 0$ and a is small, this distribution peaks very sharply at 0, and has very thick flat tails, as shown in Figure 8.

The *log* of $p(\boldsymbol{w})$ gives us the sparseness cost measure:

$$\log p(\boldsymbol{w}) = C(a,b) + \left(1 + \frac{a}{2}\right) \sum_{m=0}^{M-1} \log\left(1 + \frac{w_m^2}{2b}\right)$$

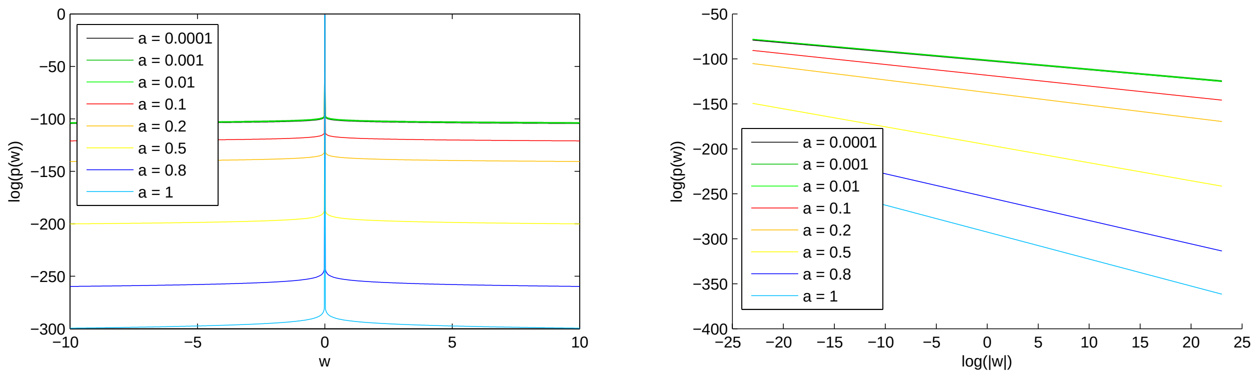

Fig. 8. Plot of the marginal prior distribution on a single weight for different choices of the hyperparameter a. To make the plot we approximated $b \longrightarrow 0$ by $b = 1E - 80$, that is a similar conceptual plot as would be drawing a delta function by a normal distribution with $\sigma = 1E - 80$.

and we are interested in the case when $b \longrightarrow 0$, which gives (18):

$$\log p(\boldsymbol{w}) \underset{b \to 0}{\longrightarrow} C(a) + (1+2a) \sum_{m=0}^{M-1} \log |w_m|$$

This sparseness cost is depicted for a single nonzero weight and several a in Figure 8 and for multiple nonzero weights and a single a in Figure 3. The approximately flat tails makes this sparseness cost a good approximation of the l_0 norm, and much more desirable than the l_1 norm (i.e. Laplacian prior). Considering specifically a in (18) and Figure 8, we can see that the sparseness penalty is proportional to $(1+2a)$. For example, in Figure 8 (left), for $a = 1$ we get a penalty of around 300 for large coefficients as compared to 100 when $a \sim 0$, i.e. $(1+2a)$ times higher. Therefore, we can increase the sparseness by increasing a, this takes mass away from the tails and puts it on the "delta" (point mass at 0) by decreasing the rate on the tail decay. In Figure 8 (right), the tail decay rate is about $(1+2a)$ on the natural logarithmic scale.

The parameter a also has an impact on the convergence rate of the EM algorithm, i.e., the speed of the SBL algorithm. In our experiments, for higher sparseness settings (fewer breakpoints and larger a), the algorithm converges faster than for smaller a. This is also supported with the following argument. The α_m^{-1} parameters, either converge to 0 (breakpoint discarded) or to a finite point (breakpoint accepted). The EM algorithm rate of convergence is governed by the maximum eigenvalue of the Jacobian matrix of the EM mapping defined in (17), (McLachlan and Krishnan, 1997). In that situation, $1/(1+2a)$ would pull out of the derivative of α_m^{-1} in (17); thus speeding up convergence since the maximum eigenvalue is divided by $1/(1+2a)$.

In conclusion, the a parameter controls the sparseness in the SBL algorithm, and the speed of the algorithm. An increase of a leads to a sparser result, fewer breakpoints, and faster convergence.

4 BACKWARD ELIMINATION ALGORITHM DETAILS

The backward elimination (BE) procedure could be used alone for CNA detection. It is based on considering our PWC model (6) as a classical variable regression selection problem, $\boldsymbol{y} \sim \boldsymbol{F}\boldsymbol{w}$; where the regressors w_i with less impact on the residual are sequentially removed one by one. To the best our knowledge, this simple procedure has never been proposed as a standalone technique for CNA detection. This is a greedy approach, which is suboptimal since we may eliminate breakpoints that could be more significant at a later stage. Since errors can be added by each greedy decision, this algorithm tends to be more reliable when the number of regressors (i.e., candidate breakpoints) is smaller. Compared to forward selection (FS), backward elimination (BE) has been seen to perform better in situations where, as in our case, the columns of $\boldsymbol{F}$ have high degree of collinearity (Kohavi and John, 1997). Furthermore, the structure of $\boldsymbol{F}$, the design matrix, can be exploited to efficiently find and remove each breakpoint and produce a ranking list as detailed in Algorithm 2.

Using standard linear regression, for a given fixed breakpoint set $\mathcal{I}$, the least squares estimate for the breakpoint weights $\boldsymbol{w}_{\mathcal{I}}$ is found by solving the normal equations:

$$\begin{aligned} \boldsymbol{F}_{\mathcal{I}}^t \boldsymbol{y} &= \boldsymbol{F}_{\mathcal{I}}^t \boldsymbol{F}_{\mathcal{I}} \hat{\boldsymbol{w}}_{\mathcal{I}} \\ \hat{\boldsymbol{w}}_{\mathcal{I}} &= \left(\boldsymbol{F}_{\mathcal{I}}^t \boldsymbol{F}_{\mathcal{I}}\right)^{-1} \boldsymbol{F}_{\mathcal{I}}^t \boldsymbol{y} \end{aligned} \quad (27)$$

which gives the orthogonal projection $\hat{\boldsymbol{x}}_{\mathcal{I}}$ of the vector $\boldsymbol{y}$ on $\mathcal{S}_{\mathcal{I}}$ as:

$$\hat{\boldsymbol{x}}_{\mathcal{I}} = \boldsymbol{F}_{\mathcal{I}} \hat{\boldsymbol{w}}_{\mathcal{I}} \quad (28)$$

$$\hat{\boldsymbol{x}}_{\mathcal{I}} = \boldsymbol{F}_{\mathcal{I}} \left(\boldsymbol{F}_{\mathcal{I}}^t \boldsymbol{F}_{\mathcal{I}}\right)^{-1} \boldsymbol{F}_{\mathcal{I}}^t \boldsymbol{y} \quad (29)$$

and the residual sum of squares $RSS_{\mathcal{I}}$ or norm of the error is:

$$\begin{aligned} RSS_{\mathcal{I}} &= \|\boldsymbol{y} - \hat{\boldsymbol{x}}_{\mathcal{I}}\|^2 \\ &= \|\boldsymbol{y} - \boldsymbol{F}_{\mathcal{I}} \hat{\boldsymbol{w}}_{\mathcal{I}}\|^2 \end{aligned} \quad (30)$$

All these operations can be solved efficiently by noticing again that $\boldsymbol{H}_{\mathcal{I}} = (\boldsymbol{F}_{\mathcal{I}}' \boldsymbol{F}_{\mathcal{I}})^{-1}$ is a symmetric tridiagonal matrix, with main diagonal $\boldsymbol{h}_0$ (19) and first off-diagonals $\boldsymbol{h}_1$ (20) (see lines 2 and 3 of Algorithm 2).

The criteria to decide which breakpoint to remove can be seen in three different but equivalent ways.

First, we might consider removing the breakpoint which increases the least the $RSS_{\mathcal{I}}$. If we denote RSS_j to be the residual sum of the squares after removing i_j from $\mathcal{I}$, then the increase in RSS is:

$$RSS_j - RSS_{\mathcal{I}} = \frac{\hat{\boldsymbol{w}}_{\mathcal{I}}^2(j)}{\boldsymbol{h}_0(j)} \tag{31}$$

Furthermore, when the noise is normal $N(0, \sigma^2 \boldsymbol{I})$,

$$F_j = \frac{RSS_j - RSS_{\mathcal{I}}}{RSS_{\mathcal{I}}/(M-K)} \tag{32}$$

is distributed as $F_{1,M-K}$ distribution (M is the number of candidate breakpoints, and $K = |\mathcal{I}|$ the number of breakpoints in the model). If the σ^2 is known, or $M >> K$,then $RSS_{\mathcal{I}}/(M-K) \rightarrow \sigma^2$ and $F_{1,\infty} \sim \chi_1^2$; thus

$$t_j^2 = \frac{RSS_j - RSS_{\mathcal{I}}}{\sigma^2} = \frac{\hat{\boldsymbol{w}}_{\mathcal{I}}^2(j)}{\sigma^2 \boldsymbol{h}_0(j)} \tag{33}$$

is distributed as a χ_1^2 distribution.

Second, if we assume that the noise is normal $N(0, \sigma^2 \boldsymbol{I})$, and σ^2 is known. Then the least squares estimate for $\hat{\boldsymbol{w}}_{\mathcal{I}}$ is also normally distributed:

$$\hat{\boldsymbol{w}}_{\mathcal{I}} \sim N\left(\boldsymbol{w}_{\mathcal{I}}, \boldsymbol{H}_{\mathcal{I}}/\sigma^2\right) \tag{34}$$

Therefore, under the hypothesis that $w_{\mathcal{I}}(j) = 0$

$$t_j \equiv \frac{\hat{\boldsymbol{w}}_{\mathcal{I}}(j)}{\sqrt{\sigma^2 \boldsymbol{h}_0(j)}} \sim N(0,1) \tag{35}$$

Third, developing what t_j represents in terms of $\boldsymbol{y}$ and σ^2 by performing all the operations in (27), we can see that:

$$t_j = \frac{\left(\frac{1}{i_{j+1}-i_j} \sum_{m=i_j+1}^{i_{j+1}} y_m\right) - \left(\frac{1}{i_j - i_{j-1}} \sum_{m=i_{j-1}+1}^{i_j} y_m\right)}{\sigma \sqrt{\frac{1}{i_{j+1}-i_j} + \frac{1}{i_j - i_{j-1}}}} \tag{36}$$

which can be interpreted as the difference between the sample mean of the right and the left segment of i_j breakpoint divided by the square root of the variance of that difference. Even if the noise is not normal, but has a finite variance σ^2, (36) tells us that as the size of the segments increases, under the null hypothesis of no difference, t_j will converge to $N(0,1)$ because of the central limit theorem.

Recalculation of the weights after each removal, can be done efficiently with very few (a constant amount of) operations using the weights already calculated (see lines 9,12 and 16 on Algorithm 2). Thus the overall order of complexity to rank a breakpoint set $\mathcal{I}$ is linear with the size of the set $O(|\mathcal{I}|)$.

4.1 The role of the T parameter in BE ranking

The ranking provided by the backward elimination procedure, Algorithm 2, can be used to quickly return a breakpoint set with different degrees of sparseness that contains the breakpoints with the strongest evidence. This is done by cutting the ranking r at some specified threshold T, such that all the remaining breakpoints have a $|t_j| \geq T$. Both true positives and false positives will decrease with increasing level of sparseness, i.e. higher T; but if $P(|t_j| \geq T | w_j = 0) < P(|t_j| \geq T | w_j \neq 0)$ the expected proportion of false breakpoints on the returned set, i.e. the false discovery rate FDR, will be monotonically decreasing with T. The previous condition is true for Gaussian noise but will also be true for other symmetrically bell shaped noise distributions.

Additionally, we can associate a p-value for any particular value of t, or a significance cutoff $\alpha = P(|t_j| \geq T | w_j = 0)$ for any T, if we assume the noise is normal, using (35). If the noise is not Gaussian, the p-value will still be a good a approximation for the breakpoints with large flanking segments (i.e. the two neighboring breakpoints are far apart), since t will converge to a normal distribution under the null hypothesis for any iid finite variance noise. Alternatively, or for small segments we could estimate the p-value by a resampling method or replace the t score by a non-parametric ranksum test. In any case, it is important to notice that the computed p-value is associated with a single breakpoint in one of the many possible segmentations; thus, it does not take into account all the possible segmentations that are effectively tested during the algorithm, i.e. multiple hypothesis testing or multiple comparison problem.

Commonly used tools to solve this problem are not recommended here because they do not take into account the special correlation structure that exists between the t scores of overlapping or neighboring segmentations, and the independence between the t scores separated by one breakpoint or more. Solving the problem of the multiple testing in this scenario, in the sense of being able to provide a T that controls for the FDR being bellow some bound is beyond the scope of this paper. However, since the FDR is monotonically decreasing with T, we can adjust it to achieve a particular degree of sparseness, and then estimate the FDR that corresponds to that T either using results from multiple samples, replicates or by a resampling procedure.

5 SEGMENT ALTERATION DETECTION

The SBL and BE procedures are segmentation approaches that make no assumptions on the amplitude of the reconstructed segments. The objective is to provide a nearly optimal set of amplitudes and breakpoint positions that best fits the hybridization intensities observed in the array as described in (8). Once the breakpoints are fixed, in order to achieve the minimal residual error RSS, the amplitude corresponding to each segment is given by the average hybridization level of all the probes that fall inside that segment. The consequence of this model, is that segments that correspond to the same underlying copy number state may be given a different reconstruction amplitude; and, an additional step has to be done to classify these segments into a copy number (0, 1, 2, 3, 4, ...) or alteration status (*Non-Altered*, *Gain* and *Loss*).

There already exists two alternatives to perform this additional step, since it is also required in other segmentation procedures like DNAcopy (Olshen *et al.*, 2004) and CGHseg (Picard *et al.*, 2005). The first alternative, also used in smoothing and thresholding methods (Pollack *et al.*, 1999; Huang *et al.*, 2004), assumes or estimates a baseline *Non-Altered* mean hybridization level and classifies all the segments whose amplitude are significantly above (bellow) that level as *Gain* (*Loss*), or *Non-Altered* otherwise. The second alternative is the MergeLevels algorithm (Willenbrock and Fridlyand, 2005), that reduce the number of different reconstruction amplitudes by recursively merging those that are the least significantly different. The final smaller set of levels may be associated with a copy number state (0, 1, 2, 3, 4, ...).

Other CNA detection approaches, specially those that are based on HMMs automatically incorporate the classification encoded in the different states of a hidden variable associated with each probe. However, as we discussed in the introduction, this may not be a good model when the number of hidden states that has been assumed does not match the true number of underlying mean hybridization levels. This is specially likely to occur when analyzing tumor samples which represent mixtures of cells with different copy number state, because cancer genomes are inherently unstable and heterogeneous (Garraway *et al.*, 2005).

6 ADJUSTMENT OF THE SBL AND BE PARAMETERS IN GADA

Both the SBL or the BE procedure could be used independently to estimate copy number changes. However, the best results and flexibility are obtained with the combination of these two algorithms as was discussed in section 2.8.

The objectives of this section are: 1) show that SBL and BE elimination produce breakpoint sets that are subsets of those obtained from higher sparseness settings, higher T or a, and can produce equivalent breakpoint sets; 2) propose a strategy for efficient parameter adjustment in the most general case; 3) evaluate the effectiveness of this strategy in the Willenbrock and Fridlyand simulated dataset.

The experiments consist of drawing simulated chromosomes of different lengths M (M=100, 200, 500, 1000 and 2000 probes per chromosome), in the following conditions:

i. Simulation of null hypothesis, (no breakpoints) using normal noise with different levels of variance.

ii. Simulation of normal copy number variations (few breakpoints and short segments) with real noise obtained by randomly sampling segments of data of size M from a pool of a normal (diploid genome) CEPH cell line samples analyzed by Affymetrix 250K Nsp array platform.

iii. Simulation on cancer copy number variations, by sampling random chunks of data of size M from cancer samples analyzed in section 3.3.

iv. Evaluation on the simulated dataset analyzed in section 3.1, (only M=100).

For i. to iii. we simulated $L = 10000$ chromosomes, for the last case iv. all the $L = 500 \times 20 = 2000$ chromosomes of size $M = 100$ were used. Each sample, i.e. chromosome, was analyzed with different options of a and T, and the returned breakpoint sets were evaluated using different metrics. The sparseness of each set was computed as the number of returned breakpoints divided by the size of the chromosome, i.e. $|\mathcal{I}|/M$, and λ denotes the average sparseness across all samples. When comparing two breakpoint sets $\mathcal{A}$ and $\mathcal{B}$ obtained for the same sample but with different parameter settings, we denote $\mathcal{A} \cap \mathcal{B}$ the set of common breakpoints, which in our case includes all breakpoints in $\mathcal{A}$ such that there exists a breakpoint in $\mathcal{B}$ less than δ probes away (if there are two breakpoints in $\mathcal{A}$ closer than δ to a breakpoint on $\mathcal{B}$ then only the closest one is assigned to the intersection). We then computed the averages of the following metrics along the L simulated samples:

$$P(\mathcal{A} = \mathcal{B}) = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|} \quad (37) \qquad P(\mathcal{A} \subset \mathcal{B}) = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A}|} \quad (38)$$

which represent respectively the proportion of breakpoints that are the same on both sets, i.e. concordance, and the proportion of breakpoints on $\mathcal{A}$ that are also in $\mathcal{B}$, i.e., inclusiveness of $\mathcal{A}$ in $\mathcal{B}$.

6.1 Experiments adjusting a and T in GADA

In Figure 9, we can see that the initial breakpoint sets provided by SBL (at $T = 0$), a higher a setting increases sparseness (bottom plots $T = 0$); but at the same time the breakpoints remain the same since the $P(\mathcal{A} \subset \mathcal{B}) > 99\%$ in all the cases. That means that breakpoint sets obtained with higher a tend to be subsets of those obtained with lower a.

As T increases we can see on the bottom plot that we are monotonically obtaining sparser sets. The breakpoints that we are removing with BE might be different depending on the initial conditions; for example, $a = 0.8$ already has a high degree of sparseness so it will not start removing anything until $T > 2.88$, where the sparseness will start to curb down and eventually will converge to the curves obtained with lower a. On the top plot, we can see that this convergence is not only on the degree of sparseness alone but also on the breakpoint sets themselves too, since as T increases concordance goes to 1. That means that as we increase T we remove the extra part that it was in the breakpoint set obtained with lower a and we end up with the same breakpoints. Following the example with $a = 0.8$, we can see in Figure 9 A, that for $T > 4.15$, the concordance to starting with a lower a is higher than 80%; and for $T > 4.25$ and $T > 4.35$ we obtain concordances that are respectively higher than 90% and 95%.

This results indicate that we can adjust the sparseness of the result equivalently with a and T in a wide margin of settings to give the same breakpoint set. This behavior has been observed in all the experiment settings (i.–iv.). If there is something to be detected, true copy number alterations or outliers ii. then the probability of detection is higher and the high concordance is reached for smaller values of T than in the i. case (compare A and C , and B and C, on Figure 9). For iii. case (data not shown) the concordance is even higher since cancer samples contain more CNA. The size of the chromosome M have also some impact on the convergence; on chromosomes with larger M high concordance is reached at a higher T, but for $M > 2000$ it does not move further more to the right. Additionally, our results on case i. are exactly the same for different noise power σ^2 because both a and T have already been corrected by $\hat{\sigma}^2$.

6.2 Strategy to adjust a and T in GADA

Adjusting sparseness with T can be done at no additional computational cost, while adjusting a requires to run the EM algorithm again. Thus, a good strategy is to select a small a for SBL that provides an initial breakpoint set that reduces most of the unlikely breakpoints but still ensures a high sensitivity; i.e. we do not want to miss anything on the first step that would require us to switch back to a lower a. Then, the final degree of sparseness will be adjusted with T.

From the previous experiment in concordance between sets, we can see that a good sensitivity means that we do not remove anything that would not be removed with a lower a at the same T. The worst case, i.e. requiring a higher T for the same concordance, is where there is nothing to be detected. In other words, the true copy number alterations that produce an observation that resembles more pure Gaussian noise are the hardest to be detected.

Moreover, dense arrays (higher M) will be more sensitive because CNA will be sampled with more probes and will produce statistically larger t. Thus, small arrays will be those requiring the smallest T to be highly sensitive. Even smallest arrays with 100 probes per chromosome, $T = 4$ provides enough initial sensitivity. Thus, we find that $a = 0.2$ should be small enough in general, and is the value that we have always used in all the results on section 3.

It is always possible to determine if $a = 0.2$ (or any other choice) is small enough for a particular T of interest by rerunning the algorithm with a lower a, e.g. $a = a/2$, and checking if the set of breakpoints returned for that particular T and different a's are essentially the same (e.g., $> 95\%$ concordant).

6.3 Sensitivity to the adjustments of a and T

We will use the simulation case iv., to evaluate the impact of the parameter setting strategy described in the previous section in terms of accuracy. This is the same dataset as the one used in section 3.1 and by Willenbrock and Fridlyand (2005), where the underlying breakpoints are known, so we can exactly evaluate the FDR and the sensitivity for different choices of T and a.

In Figure 10, curves corresponding to different a have different starting point in terms of sensitivity and FDR, but as T increases we decrease the FDR and similar operational points in terms of sensitivity and FDR are reached compared to those obtained from different a. The proposed $a = 0.2$ in the previous section offers and initial sensitivity and FDR such that all the remaining points in the curve are reached adjusting only T, providing all the levels of sensitivity or FDR that we might be interested in using without having to switch to another a.

Compared to CBS, we are able to obtain a wider margin of operating points of the PROC curve. Moreover, independently of the initial a we always have a point with similar or better average performance either in terms of FDR or sensitivity.

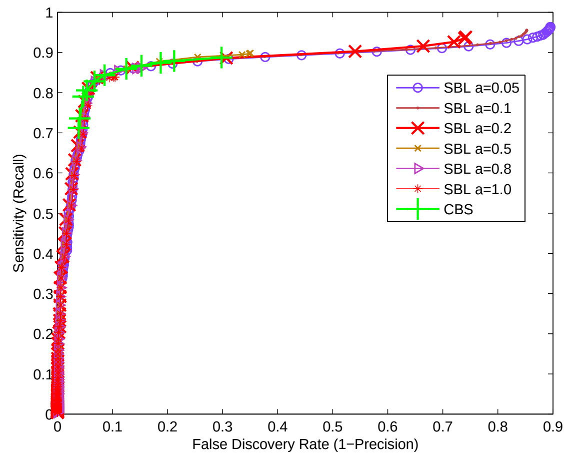

Fig. 10. PR operational curves for sensitivity vs. false discovery rate in detecting copy number changes within $\delta = 2$ probe window. Each line corresponds to SBL+BE with different starting breakpoint sets ($a = 0.05, 0.1, 0.2, 0.5, 0.8, 1.0, 1.5$) and varying T (T increases as we traverse the curve from right to left, i.e. FDR decreases). The light green curve represents the operating points obtained by CBS with different $\alpha = 1E-4, 0.001, 0.002, 0.005, 0.01, 0.05$

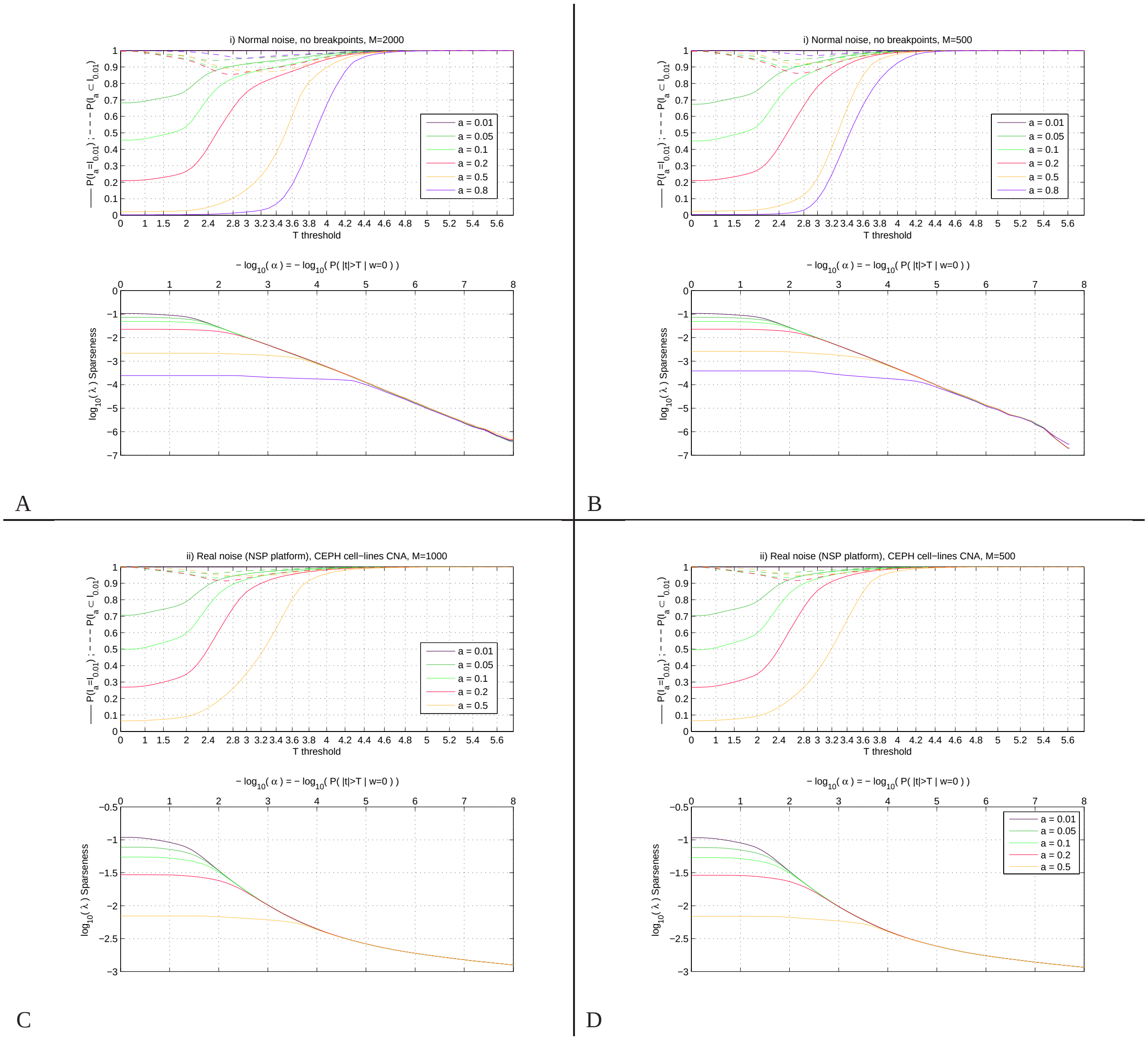

Fig. 9. The four panels (A,B,C and D) represent a different experimental dataset, with the results of applying different settings of a and T parameters. Each color corresponds to different setting of $a = 0.01, 0.05, 0.1, 0.2, 0.5$, and the x-axis increasing values of T or its associated significance level $log_0(\alpha)$. On the top plot we have represented the inclusiveness $P(\mathcal{I}_a \subset \mathcal{I}_{a=0.01})$ (dashed line); and the concordance $P(\mathcal{I}_a = \mathcal{I}_{a=0.01})$. The concordant breakpoints are defined within a window of $\delta = 2$ probes. The bottom plot represents the sparseness which on A and B also represent specificity because there are no underlying breakpoints. A and B use the normal noise simulation described in i. with chromosome lengths of $M = 500$ and $M = 2000$ (different noise levels σ^2 generate exactly the same curves); and C and D use the simulation described in ii. with chromosome lengths of $M = 500$ and $M = 1000$.

7 SUPPLEMENTARY RESULTS TABLES AND FIGURES

Table 5. Significant copy number alterations found in four neuroblastoma cell lines

Chr.	SK-N-BE2	SMS-KAN	LAN-6	CHLA-20
1:		–(pEnd–p13.3)	–(pEnd–p36.12)	
	+(p21.3–qEnd)		+(p21.1–qEnd)	+(p21.1–qEnd)
2:	+(pEnd–p21)	+(pEnd–p24.1)	+(pEnd–p22.1)	++(pEnd–p16.1)
	++p24.3 **MYCN**	++p24.3 **MYCN**		+(p16.1–q31.1)
		++p24.1	+q35	
		–q22.1		
		–q23.3		+(q32.2–q37.2)
3:	–(pEnd–p14.2)		–(pEnd–p14.3)	
		+(p12.1–p11.1)		
4:				–(p16.1–p15.33)
		–(q12–q22.1)		–p24 **KLHL5**
		–q22.3	+(q35.1–q35.2)	+(q34.1–qEnd)
5:			–(q35.3–qEnd)	+q11.2
6:			–(q12–p16.3)	
			–(q22.31–qEnd)	
			– – q26 **PARK**	
7:			+(pEnd–p15.1)	
			–q21.1 **AHR**	+7
			–(p14.3–q11.21)	
			–(p11.21–q11.22)	–q33 **SEC8L1**
8:			–(pEnd–p12)	
			–(q22.1–q23.3)	+q21.3
	– –q24.23			+(q22.2–q24.1)
9:			–p24.2 **GLIS3**	
			–(p23–p21.2)	
			– –p21.3**MTAP**	
			–p13.3**RECK**	
10:		+(pEnd–p11.23)		
		–q22.3 **PTPRE**		
11:	+(q13.1–qEnd)	–(q14.2–q23.3)	+(q13.4–q25)	–q14.1
			– –(q25–qEnd)	+q22.1 **CNTN5**
12:			+(q23.3–q24.33)	+12
			++ q24.33	
13:			–q31.1	
14:			–(q23.2–qEnd)	
			+0(q31.3) **TTC8**	
15:				
16:		+16q	+(pEnd–p13.3) **LEP**	
17:			– –p11.2 **EPN2**	
	-(pEnd–q11.2)		-(pEnd–q11.2)	+17
			+(q21.2-qEnd)	
18:	-18			+(p11.23-qEnd)
19:		–(q13.2–q13.33)		+19p
20:	-p13			
21:				
22:				
X:	X	XX	X	XX

Table listing the most significant copy number alterations $T = 5$, that have been found on at least two of the platforms (Xba,Nsp,Sty,Nsp+Sty) being analyzed.

Table 8. Differences in copy number breakpoint placing between chips

		MAD [BP]		K-S p-value	
Chips compared	# cases	GADA	CBS	GADA larger	CBS larger
$\min(\lvert Xba - Sty\rvert, \lvert Xba - Nsp\rvert)$:					
	59	95670	93132	0.54	0.76
$\lvert Nsp - Sty\rvert$:					
	61	88024	71265	0.35	1.0
$\lvert (Nsp\&Sty) - Illumina\rvert$:					
	91	22784	21388	0.58	0.95

For the confirmed breakpoints and excluding those near the centromere, we computed the median absolute difference in breakpoint location between chips (units in base pairs [BP]), and the p-value associated with the Kolmogorov-Smirnoff test for the hypothesis that differences are stochastically larger (i.e. less accurate) in one algorithm vs. the other. No significant changes in accuracy have been found.

Table 6. Copy number breakpoints found on all platforms

Cell-line name	Chr & Cytoband	GADA Position [BP]					CBS Position [BP]				
		Xba	Nsp	Sty	Nsp+Sty	Illumina	Xba	Nsp	Sty	Nsp+Sty	Illumina
SK-N-BE2	1p21.3	97045920	96895983	97183094	96895983	96808701	96602215	96564172	95918556	96486927	96808701
SK-N-BE2	2p24.3	15977810	15978001	15977810	15978001	15979864	15977810	15978001	15977810	15978001	15979864
SK-N-BE2	2p24.3	16419609	16453092	16462002	16462002	16463522	16419609	16453092	16462002	16462002	16465097
SK-N-BE2	2p21	48447814	47722197	48071628	47629563	47840828	48447814	47728339	48071628	47738916	47840828
SK-N-BE2	3p14.2	61381385	61301823	61423021	61159361	61312730	61227509	61301823	61137444	61241447	61312730
SK-N-BE2	8q24.23	137748993	137757306	137735555	137747078	137747933	137748993	137746403	137735555	137746403	137747933
SK-N-BE2	8q24.23	137892295	137955330	137924208	137924208	137919630	137892295	137931617	137924208	137931617	137919630
SK-N-BE2	11q13.1	64310154	64977325	65339642	65010150	65335248	64310154	64977325	65339642	65339642	65193464
SK-N-BE2	17q11.2	28109086	28263828	28164827	28283675	28247634	28109086	28263828	28164827	28266739	28214976
SK-N-BE2	20p13	2406160	2773972	3036010	3036010	2987115	2406160	2773972	3036010	3036010	2987115
SMS-KAN	1p13.3	108157301	107888773	108203626	108218567	108177825	108157301	108208349	108186644	108178227	108177825
SMS-KAN	2p24.3	15721907	15868241	15853157	15868241	15869663	15721907	15868241	15853157	15868241	15869663
SMS-KAN	2p24.3	16419609	16576876	16551640	16576876	16578889	16419609	16576876	16551640	16576876	16578889
SMS-KAN	2p24.1	21887032	21974989	22013932	22013932	21992435	21887032	21974989	22013932	22013932	21992435
SMS-KAN	2p24.1	22466771	22475341	22470539	22475341	22475673	22466771	22475341	22470539	22475341	22475673
SMS-KAN	3p12.1	83843045	84210511	84386931	83873910	84165323	85157384	84137907	84386931	83960187	84165323
SMS-KAN	3p11.1	88163152	90346746	97369003	90346746	90472437	88163152	90346746	96620438	90346746	90472437
SMS-KAN	4q12	55018889	55049094	55058835	55049094	55040244	55152302	55049094	55056941	55049094	55040244
SMS-KAN	4q22.2	94894653	94792613	94948871	94957181	94937901	94894653	94833508	94919018	94872717	94940338
SMS-KAN	10p11.23	30297356	30435753	30721148	30685964	30559838	30297356	30552253	30721148	30721148	30551022
SMS-KAN	10q26.2	129582283	129682011	129741611	129741611	129689191	129582283	129686948		129693110	129689191
SMS-KAN	10q26.2	130102560	130079710	130148252	130148252	130172807	130095140	130168048		130148252	130172807
SMS-KAN	11q14.2	85287454	85435237	85383124	85383124	85381622	85287454	85367689	85383124	85383124	85381622
SMS-KAN	11q23.3	117735474	117791205	118235879	118235879	117802601	117735474	117791205	117823148	117823148	117802601
SMS-KAN	19q13.2	45796700	45571967	45879279	45879279	45910451	45796700	45571967	45879279	45879279	45910451
SMS-KAN	19q13.33	57641156	57395071	57400799	57395071	57374324	57641156	57395071	57322747	57395071	57374324
LAN-6	1p36.33	21940846	22438012	22476248	22476248	22480144	22445846	22438012	22476248	22476248	22480144
LAN-6	1q21.1	142661525	143607676	143798352	143887291	144106312	142661525	143607676	143798352	143798352	144106312
LAN-6	2p22.1	42656869	41405461	33909326	41340562	41358977	42656869	41405461	41335143	41340562	41358977
LAN-6	2q36	218185198	218577787	218754682	218756027	218628755	218395997	218577787	218754682	218754682	218628755
LAN-6	2q36	220406369	220398375	220629809	220629809	220449867	220406369	220534849	220475305	220480390	220449867
LAN-6	3p14.3	57324905	57032227	57238434	57238434	57215650	57324905	57032227	57238434	57238434	57215650
LAN-6	6q12	68022441	68353881	68154792	68095015	68197504	68022441	68095015	67263660	68095015	68197504
LAN-6	6p16.3	106344198	105090334	105177437	105110313	105135994	105275051	105090334	105110313	105110313	105112419
LAN-6	6q22.31	123295546	123710384	123731764	123710384	123710826	123295546	123710384	123605395	123710384	123710826
LAN-6	7p15.1	27784099	24892359	24256367	24892359	24919337	27784099	31129325	25064458	24892359	24919337
LAN-6	7p14.3	31126902	31129325	31135758	31135758	31145274	31126902	31129325	31135758	31135758	31145274
LAN-6	7q11.21	62171000	62343621	63050316	63050316	62336389	62171000	62910986	62283233	62934268	62336389
LAN-6	7q11.21	63993279	63940919	63137057	63940919	63963370	63993279	63940919	65748163	63940919	63963370
LAN-6	7q11.22	69847573	69835520	68907809	69835520	69831960	69738283	69835520	69807385	69835520	69831960
LAN-6	8p12	39331612	37835460	37192319	38093651	37869447	37458281	37835460	38093651	38093651	37869447
LAN-6	8q22.1	93948610	95445117	95274323	95445117	95414304	96630283	95445117	95372632	93495648	95414304
LAN-6	8q23.3	116469762	116142633	116478611	116140376	116480735	116519714	116472717	116471437	116472717	116480735
LAN-6	9p24.2	3394434	3579281	3316679	3579281	3585674	3394434	3579281	3316679	3579281	3585674
LAN-6	9p24.2	4947650	4635935	4685068	4648449	4647040	4947650	4630212	4624827	4624827	4647040
LAN-6	9p23	12741741	12643846	13524659	12649691	12706172	12741741	12643846	12164190	12754534	12716962
LAN-6	9p21.3	21451790	21460464	21460997	21460997	21468318	21451790	21460464	21460997	21460997	21484643
LAN-6	9p21.3	22185820	22158464	22197037	22197037	22197037	22404973	22158464	22197037	22197037	22197037
LAN-6	9p21.2	28417657	28860162	28929272	28820009	28857478	28765609	28853202	28742971	28820009	28844830
LAN-6	11q13.4	71592372	71549242	71591974	71591974	71607855	71592372	71549242	71591974	71591974	71634231
LAN-6	12q23.3	105993065	106086197	106095939	106095939	106074551	105993065	106086197	106232150	106095939	106074551
LAN-6	14q23.1	60468351	60341891	60393691	60343301	60386017	60468351	60436455	60393691	60393691	60386017
LAN-6	17q11.2	25755541	24749129	24833230	24836351	24865310	25755541	24836351	24833230	24836351	24865310
LAN-6	17q21.2	36391251	35264341	36690164	35852756	35294289	36095487	35264341	36556209	36393487	35294289
LAN-6	17q22	50618719	51862430	51475956	51862430	51856911	50618719	51862430	51475956	51862430	51855630
CHLA-20	1q21.1	120089986	143607676	120928505	143887291	143328536	142661525	142756696	143798352	143657867	14802010
CHLA-20	2p16.1	68885099	57662314	55736176	57577846	57662175	57629406	57662314	58097954	57625311	57583694
CHLA-20	2q31.1	174726584	174705263	174793287	174717018	174730921	174726584	174705263	174717018	174717018	174730921
CHLA-20	2q32.2	178651035	178581699	179377095	178574520	178576047	178214924	178574520	179028748	178574520	178576047
CHLA-20	4p16.1	5722378	5842107	5913372	5842107	5844271	5722378	5842107	5913372	5842107	5842107
CHLA-20	4p15.33	12551237	12300938	12392275	12321237	12316278	12551237	12300938	12391799	12321237	12300938
CHLA-20	4q34.1	175185953	174895669	175097390	174895669	174897540	175185953	174895669	174890618	174895669	174895669
CHLA-20	8q21.3	87234127	87640754	87583206	87583206	87594384	87858742	87640368	87583206	87583206	87594384
CHLA-20	8q21.3	90473708	90386811	90407394	90407394	90366715	90138115	90372039	90407394	90376886	90366715
CHLA-20	8q22.2	100574413	99817691	99940387	99940387	99638911	99803560	99817691	99940387	99940387	99638911
CHLA-20	8q24.1	127917518	128903451	128922998	128922998	128913903	127917518	128903451	128922998	128922998	128913903
CHLA-20	18p11.23	8617957	8510927	8200848	8510927	8392640	8617957	8393289	8309751	8448484	8392640
CHLA-20	19q12	33171613	32761177	23876259	32690406	24095263	33171613	32761177	24165666	32690406	24053526

Table listing the locations for the copy number breakpoints detected by GADA and CBS on the four neuroblastoma cell-lines (SK-N-BE2, SMS-KAN, LAN-6, CHLA-20) that have also been found on all array platforms (Xba, Nsp, Sty, Nsp+Sty, Illumina).

Table 7. Copy number breakpoints found by at least two platforms

Cell-line name	Chr & Cytoband	GADA Position [BP]					CBS Position [BP]				
		Xba	Nsp	Sty	Nsp+Sty	Illumina	Xba	Nsp	Sty	Nsp+Sty	Illumina
SMS-KAN	2q22.1	141962960	142006622		142006622	141996840	141962960	141991258		141991258	141996840
SMS-KAN	2q22.1	142284086	142419855		142419855	142418322	142284086	142419855		142386408	142418322
SMS-KAN	2q23.3	152928225	152959722		152945591	152947104	152928225	152959722			
SMS-KAN	2q23.3	153172826	153233532		153356905	153182040	153172826	153231102			
SMS-KAN	4q22.3	98081595	97971534		97971534	97946136	98081595	97971534		97971534	97946136
SMS-KAN	4q22.3	98389453	98794422		98492192	98515047	98389453	98489669		98489669	98515047
LAN-6	4q35.1		187001741	187043679	187043679	187037031		187001741	187248120	186997226	187037031
LAN-6	4q35.2		189693227	189400592	189537964	189715209		189693227	189400592	189693227	189715209
LAN-6	5q35.3		178389593	178435944	178439675	178388353		178389593	178435944	178439675	178388353
LAN-6	6q26		162768919	162672040	162783990	162769931		162768919	162770133	162770133	162769931
LAN-6	6q26		163042210	162863051	163042210	163069487		163073363	162948280	163042210	163069487
LAN-6	7p21.1	17042942	17048506		17048506	17079506					17079506
LAN-6	7p21.1	17194089	17293268		17293268	17187461					17194647
LAN-6	9p13.3		35932406		35934224	35923323					35923323
LAN-6	9p13.3		36095264		36117196	36036596					36036596
LAN-6	11q25		134393784		134408260	134410991					134410991
LAN-6	12q24.32		125117158		125475975	125319087		125257058	124609076	125131448	125319087
LAN-6	12q24.33		127723245	127722879	127723245	127723245		127723245	127835197	127723245	127723245
LAN-6	13q31.1				82988642	82996585					82996585
LAN-6	13q31.1				83045936	83063672					83055928
LAN-6	14q31.3	88300117	88381272		88443831	88310402	88300117	88381272		88443831	
LAN-6	14q31.3	88499809	88623396		88625132	88647502	88499809	88623396		88634883	
LAN-6	16p13.3		5755300	5775884	5775884	5679682		5669239	5775884	5802165	6023611
LAN-6	16q23.3			82340991	82342624	82374996		86364648	82340991	82342624	82502789
LAN-6	17		19109505		19117656	19120783				19117656	19120783
LAN-6	17		19175068		19175068	19145456				19175068	19145456
CHLA-20	2q37.2		236702079	236768287	236768287	236738515		236702079	236844697	236768287	236765691
CHLA-20	4p14		38741486		38741486	38758076		38741486		38741486	38752396
CHLA-20	4p14		38996761		38996761	39068335		38996761		38996761	39006763
CHLA-20	5q11.2		50870182		50824363	50850389		50824363		50824363	50850389
CHLA-20	5q11.2		51529692		51532772	51532772		51529692		51532772	51532772
CHLA-20	7q33		132875721		132875721	132884795		132875721		132875721	132875949
CHLA-20	7q33		133004505		133004505	132996066		133004505		133004505	132996066
CHLA-20	11q14.1		78683236	78694008	78694008	78691521			78694008	78694008	78691521
CHLA-20	11q14.1		78805883	78801531	78801531	78814667			78801531	78801531	78818346
CHLA-20	11q22.1	99371373	99366936		99366936	99378927	99371373	99240362		99366485	99378927
CHLA-20	11q22.1	100243669	100322922		100322922	100299086	100243669	100319819		100322922	100299086

Table listing the locations for the copy number breakpoints detected by GADA and CBS on the four neuroblastoma cell-lines (SK-N-BE2, SMS-KAN, LAN-6, CHLA-20) that have also been found on at two of the array platforms analyzed (Xba, Nsp, Sty, Nsp+Sty, Illumina).