Supplementary Material to "Detecting Common Copy Number Variants in High-Throughput Sequencing Data by using JointSLM algorithm"

Alberto Magi, Matteo Benelli, Seungtai Yoon, Franco Roviello, Francesca Torricelli.

Supplementary Methods

SLM is an HMM

The joint probability distribution of equation 5 (Methods) has the following form:

$$\begin{aligned} p(x,m,z|\theta) &= p(x|m,\Sigma_\epsilon)\cdot p(m|z,\mu,\Sigma_\mu)\cdot p(z|\eta) = \\ &= \textstyle\prod_{i=1}^{N} p(x_i|m_i,\Sigma_\epsilon)\cdot p(m_0)\times \prod_{i=0}^{N} p(m_{i+1}|m_i,z_i,\mu,\Sigma_\mu)\cdot p(z_i|\eta), \end{aligned} \quad \text{(SM1)}$$

where:

- $p(x_i|m_i,\Sigma_\epsilon) = N(x_i|m_i,\Sigma_\epsilon)$ is the probability distribution of x_i given m_i and the parameters.
- $p(m_{i+1}|m_i,z_i,\mu,\Sigma_\mu) = (1-z_i)\cdot\delta(m_{i+1}-m_i) + z_i\cdot N(m_{i+1}|\mu,\Sigma_\mu)$ is the probability distribution of the latent variable m_{i+1} given m_i, z_i and the parameters; δ is the Dirac delta function.
- $p(z_i|\eta) = \eta\cdot\delta(z_i-1) + (1-\eta)\cdot\delta(z_i)$ is the probability density function of z_i.

Equation SM1 defines a Hidden Markov Model (HMM) of order one, in which a single state variable, $q_i = (m_i, z_i)$, summarizes all the relevant past information of the underlying process.

In the model defined by SM1 the elements of the HMM are the following:

- the state transition probability distribution is:
  $p(q_{i+1}|q_i,\theta) = p(m_{i+1}|m_i,z_i,\mu,\Sigma_m u)\cdot p(z_i|\eta)$
- the emission probability distribution is:
  $p(x_i|q_i,\theta) = p(x_i|m_i,\Sigma_\epsilon)$
- the initial state probability distribution is:
  $p(q_0|\theta) = p(m_0|\mu,\Sigma_\mu)$

The JointSLM algorithm

The fact that the multivariate SLM is an HMM of order one with state variable $q_i = (m_i, z_i)$ and multivariate emission probability, allows us to make use of the several algorithms developed for these kinds of models. To maximize the likelihood of the multivariate extension of shifting level model we use a procedure similar to that used in (1).
We introduce a markovian stochastic process $s_1, ..., s_k$ taking values in $S = \{1, 2, ..., K\}$. We assume that the conditional probability of x_i, given $s_i = k$, is a multivariate normal with mean $\mu_k = (\mu_{1k}, ..., \mu_{Nk})$ and variance σ_ϵ, and the parameter μ_k is associated to each state of the markovian stochastic process and represents an approximation of the m_i latent variables of the SLM.
Remembering equations (SM1) and (8) the emission probability distribution has the following form:

$$f_k(x_i) = \prod_{t=1}^{M} \frac{1}{\sqrt{2\pi}\sigma_{\epsilon,j}} \exp\left[-\frac{1}{2}\left(\frac{x_{it} - \mu_{kt}}{\sigma_{\epsilon t}}\right)^2\right]. \tag{SM2}$$

In the same way we assume that the conditional probability of m_i, given m_{i-1} and the parameters, is Normal with mean μ and variance σ_μ. By using the state transition probability of Supplementary Equation 1 we can write the state transition matrix in the following form:

$$P_{jk} = \begin{cases} (1-\eta) + \eta \cdot g_{jk} & j = k \\ \eta \cdot g_{jk} & j \neq k \end{cases} \tag{SM3}$$

where:

$$\begin{aligned} g_{jk} &= \textstyle\prod_{t=1}^{M} c_j \cdot e^{-\frac{(\mu_{kt}-\mu_t)^2}{2\sigma_{\mu t}^2}}, \\ c_j &= \textstyle\sum_{k=1}^{K} \left(\prod_{t=1}^{M} e^{-\frac{(\mu_{kt}-\mu_t)^2}{2\sigma_{\mu t}^2}}\right)^{-1}. \end{aligned} \tag{SM4}$$

To estimate the parameters of the Multivariate Shifting Level Model we develop a two step algorithm that follows the idea of (1), based on dynamic programming. In the first step we estimate the parameters μ_k by means of the Baum and Welch re-estimation strategy, while in the second step we estimate the best state sequence s (the z_i variables) by means of the Viterbi algorithm. The inputs to the algorithm are the sequences $x = \{x_1, ..., x_M\}$ to be jointly segmented, the initial estimate of the number of states K_0 and the parameters ω and η. The initialization step consists of the estimation of the means μ_t and the variances σ_t^2 , $\sigma_{\epsilon,t}^2$, and $\sigma_{\mu,t}^2$ with the following formulas:

$$\begin{aligned} \mu_t &= \frac{\sum_{i=1}^{N} x_{it}}{N}, \\ \sigma_t &= \sqrt{\frac{\sum_{i=1}^{N}(x_{it}-\mu_t)^2}{(N-1)}}, \\ \sigma_{\mu t}^2 &= \omega \cdot \sigma_t^2, \\ \sigma_{\epsilon t}^2 &= (1-\omega) \cdot \sigma_t^2. \end{aligned} \tag{SM5}$$

In the first step we estimate the parameters μ_{kt} by means of the Baum and Welch re-estimation strategy:

$$\mu_{kt}^{(j)} = \frac{\sum_{i=1}^{N} \gamma_k(i) x_{it}}{\sum_{i=1}^{N} \gamma_k(i)}, \tag{SM6}$$

where $\gamma_k(i)$ is the probability of being in state k at step i. Equation (SM6) is the Baum-Welch re-estimation formula for the means of an HMM. The probability of state occupation $\gamma_k(i)$ is calculated using the Forward-Backward algorithm:

$$\gamma_k(i) = \frac{\alpha_k(i) \cdot \beta_k(i)}{P(x)} \quad \text{with} \quad P(x) = \sum_{k=1}^{K} \alpha_k(1) \cdot \beta_k(1), \tag{SM7}$$

where α and β are the forward and backward probabilities respectively (2).
The re-estimation of $\mu_{kt}(j)$, of the emission and of the transition probabilities is repeated until the asymptotic value of $P(x)$ is reached.
In the second step we estimate the points of mean shift by means of the Viterbi algorithm. Finally, we convert the data from log space to copy number space and we calculate the median of the data that belong to each segment.

Parameter Study

To understand the effect of parameters η, ω and K_0 on the power of JointSLM algorithm to segment multiple sequential processes, we made an extensive experimentation on synthetic data and evaluated its performance by means of the Area Under the Receiver Operating Characteristic Curve (AUC). In particular, we studied the ability of the algorithm to detect common shifts of various sizes, shared among different fraction of sequential processes with various Signal to Noise Ratio (SNR).
To this end, we performed two distinct simulations on multiple synthetic chromosomes: the first one for the analysis of η and ω and the second one for K_0. Each synthetic chromosome was generated as a series of 1000 points drawn from a normal distribution and alterations added at fixed positions of the sequence as square-wave signal profiles as in (3).
In the first simulation (for η and ω, type-A simulation) we used multiple synthetic sequences made of 10 chromosomes with a common alteration of width w (with $w = (10, 20, 30, 50)$) in a fraction of samples f (with $f = (0.3, 0.5, 0.7, 1)$) and with SNR $= (1, 2, 3, 4)$. SNR is defined as the mean magnitude of the alteration (i.e. signal) divided by the standard deviation of the superimposed Gaussian noise.
In the second simulation (for K_0, type-B simulation) we used 10 synthetic chromosomes with 20 common alterations of width $w = (10, 20)$ and different copy number inserted at fixed position in a fraction of the samples $f = (0.3, 0.5, 0.7, 1)$.
For each simulation we obtained ROC curve by calculating TPR and FPR at different threshold values. TPR is the number of probes inside the aberration whose fitted values are above the threshold level divided by the number of probes in the aberration, and FPR is the number of probes outside the aberration whose fitted values are above the threshold level divided by the total number of probes outside the aberration. The results of all the simulations are summarized in Supplemental Figure 11 and in Supplemental Figures 12-14. Each point of the levelplot is obtained by averaging the AUC over 100

multiple chromosomes.

The scaling parameter ω modulates the relative weight between the experimental variance (σ_ϵ^2) and the biological variance (σ_μ^2) of each sequential process. When ω is close to one, the biological variance is much larger than the experimental one and SLM takes tiny variations of the sequential process as real biological level shifts, while for values of ω close to zero the experimental noise gives the leading contribution to the total variance.

The parameter η corresponds to the probability that a transition to a new mean level vector occurs at any position i of the multiple sequential process. Figure 11.a shows two distinct regions of low efficiency of our algorithm: the first one corresponds to values of ω smaller than 0.1 (left side of the plot), while the second one corresponds to large values of η and ω (right upper corner). When ω is smaller than 0.1, JointSLM loses the ability to identify shifts in sequential processes: in this case we have high performance in terms of specificity and low performance in terms of sensitivity. On the other hands, when the values of ω and η are large, JointSLM becomes more sensitive at the expense of specificity (right upper corner). This effect becomes more evident for small values of SNR, w and f (Supplemental Figures 12-14).

The results of these simulations suggest that the use of small values of ω and η allows the algorithm to control type-I error at the expense of type-II error, while large values of ω and η are able to control type-II error at the expense of type-I error. Since we are interested in identifying a robust set of genomic variants, we will use values of the parameters ω and η that lies in the left side of the plot and allow us to obtain a small number of False Positives (FP).

For the analysis of parameter K_0 (the number of states of the SLM/HMM), we generated multiple chromosomes made of 20 recurrent alterations with randomly assigned copy number. The DNA copy number of each segment were randomly selected in the range [0,30] and each alteration is separated by a region of width $w = 20$ in normal status: in this manner a chromosome made of 1000 points is almost completely full of alterations. The principal aim of this simulation is to understand if the choice of K_0 affect the ability of the JointSLM algorithm in identifying a large number of common alterations with different DNA copy number.

The results of Figure 11.b, Figure 11.c and Supplemental Figures 15-18 clearly show that K_0 has a weak effect on the global performances of our algorithm. This is due to the fact that JointSLM is able to correctly identify all the mean shifts (z_i) also when the number of states K_0 is much smaller than the true number of mean shifts. In fact, when an altered state m_i does not have its own μ_k, the viterbi algorithm associates the alteration to the most lilkely altered m_i, allowing the identification of all the mean shifts z_i. This algorithmic behaviour does not affect the estimation of the DNA copy number of each segment since after the application of the Viterbi algorithm, the median of the data that belong to each segment is recalculated. In this way our algorithm is able to locate the correct position of all the mean shifts and the correct DNA copy number of each segment also for small values of K_0.

However, when we analyze processes made of hundreds of thousands of data we can wait to have a large number of alterations and consequently the choice of K_0 becomes hard. To overcome this problem we choose to break up the data into subsequences made of 1000 points and sequentially apply our estimation method to each subset. After the initialization step, in which we estimate the global parameters of all the sequential processes, we apply the two step algorithm (baum and welch and viterbi) to each of the 1000-points processes. Hence we can use values of K_0 that ranges between 10 and 20 without loosing detection power.

Calling

As final step, after the DOC data have been segmented, the copy number of each segment needs to be estimated in order to identify altered regions. To this end, we first filtered out all the segments that have two copies and then we estimated copy number by rounding the median of each segment to the nearest integer. A simple filtering method based on thresholds was used to remove two copy number events and to select threshold we make use of ROC curve analysis.

We generated synthetic data from chromosome 1 and X of the male individual NA18507 to simulate normal and deleted segments of known size and location. We applied JointSLM to simulated multiple chromosomes with deletions of size $w = (10, 20, 30, 50)$ in a fraction of samples $f = (0.3, 0.5, 0.7, 1)$. Once the data have been segmented, we build ROC curve by calculating True Positive Rate and False Positive Rate for different filtering thresholds and selected the best threshold as the point in the curve with the shortest distance to the point with both maximum sensitivity and specificity, (i.e. the point (0.0, 1.0)). The results of all these simulations are reported in Supplemental Figure 19, and clearly show that the optimal threshold value ranges between 1.1 and 1.3. For this reason we decided to use a threshold values of 1.2 (and a symmetrical threshold of 2.8 for the amplifications) for further analyses.

Copy Number Estimation

Once the median of each segment has been calculated, we estimate DNA copy number by rounding the median of each segment to the nearest integer. The approach of estimating DNA copy number by rounding the median of each segment to the nearest integer was introduced by Yoon et al. (4).

Under the assumption that the sequencing process is uniform, the number of read that maps to a genomic region is expected to be proportional to the number of times the regions appears in the DNA sample. To study the relationship between DNA copy number and read counts, we examined several broad genomic regions that have previously reported to have DNA copy numbers equal to 0, 1, 2, 3 and 4 by McCarrol et al. (5).

We compared the DNA copy number estimated by McCarrol et al. (5) with the median of the normalized read counts that belong to those regions: we found that normalized read counts increase linearly with DNA copy numbers. The results of these analyses are reported in the boxplot of Supplemental Figure 20.

Supplemental Validation

When we compared the set of 3000 calls inferred by our joint model with the known CNVs of DGV version 10, we found that 1722 regions overlap with known CNVs obtaining a global validation rate of 57%.

For regions that range between 1-5 Kb the validation rate is around $70 - 80\%$, and goes up to $95 - 100\%$ for genomic events greater than 5 Kb. On the other hands, when we take into consideration CNVs smaller than 1 Kb, the validation rate ranges between 40% and 60%. The total number of regions that do not overlap with DGV is 1278: 381 (29%) are smaller than 500 bp, 660 (51%) ranges between 500 and 1000 bp, 228 (19%) ranges between 1 and 5 Kb and 9 (1%) are larger than 5 Kb (Supplemental Table 1): 99% of these CNVs regions are smaller than 5 kb in size.

In order to test the information content of these CNV regions, we applied to them the Ward's hierarchical clustering with the aim to group individuals. The results of these analyses are reported in Supplemental Figure 21. The 1278 CNVs are able to segregate the ancestry of the eight individuals in two main clusters: the first cluster include the european ancestry family and the chinese individual, while the second cluster include the nigerian ancestry family and the Yoruban individual NA18507. This result suggests that the set of 1278 CNV regions detected by JointSLM is highly informative.

The fact that almost all the CNVs regions that do not overlap with DGV are smaller than 5 kb in size can be mostly accounted for to the detection limits of the technologies used in the DGV. The database of genomic variants version 10 contains 7666 structural variants originally described in healthy controls. Of all these variants, 4707 (62%) are discovered by using microarray tecniques (CGH array and SNP array), 1862 (25%) by using PEM methods with HTS technologies and the remaining 13% with other approaches. As stated in the introduction of this paper, the resolution of currently available array platforms have a lower limit of detection of $\sim 5 - 10$. Moreover, as explained in the results section of our paper, PEM- and DOC-based approaches allow to detect different classes of SVs. For these reasons, a large part of the small regions inferred by our algorithm can not be found in the database of genomic variants.

Computational performance

In order to test the computational performance of JointSLM in analysing multiple DOC profiles simultaneously, we applied our algorithm to the synthetic data generated from the GC-adjusted DOC data of chromosomes 1 and X of the male individual NA18507, as described in section Synthetic Data Analysis section. In these simulations, we fixed the values of the JointSLM parameters η and ω ($\eta = 10^{-6}$ and $\omega = 0.1$) and the fraction f of the altered samples ($f = 70\%$). The results for different values of K_0 ($K_0 = 5, 10, 15, 20$) and for 10, 20, 30, 40 and 50 multiple samples are reported in Supplemental Figure 22. The time taken by JointSLM in segmenting a 1 Mb synthetic chromosome grows while K_0 and the number of multiple samples increase.

With regards to the real data analysis, JointSLM (with $K_0 = 20$) need about 2.3 hours to segment chromosome 1 of the eight individuals simultaneously on a 2.6 GHz Intel Core 2 Duo with 4 GB RAM.

At present we are working on the parallelization of the algorithm in order to improve its computational performance.

Single experiment analysis

In order to test the performance of our algorithm in analysing single experiment, we made an intensive simulation on single profiles generated from the GC-adjusted DOC data of chromosomes 1 and X of the male individual NA18507.

As in Synthetic Data analysis section, to estimate specificity (FPR) we generated synthetic chromosomes by sampling 10000 100-bp windows from chromosome 1 to simulate normal copy number. To estimate sensitivity (TPR) we added to the normal copy number chromosomes nine deletions of size 200 bp, 300 bp, 400 bp, 500 bp, 700 bp, 1 kbp, 2.5 kbp, 5 kbp, and 10 Kb sampled from chromosome X.

We applied our algorithm with different parameter settings on 100 synthetic chromosomes with normal copy number and 100 chromosomes with deletions and we calculated TPR and FPR as in synthetic data analysis section. The results of all the simulations performed are reported in Supplemental Figure 23 and , as expected, are very similar to the results obtained in the multiple samples analysis: the specificity of the algorithm can be controlled by the parameters η and ω, while the sensitivity, is markedly affected by only the parameter ω. The algorithm is able to correctly detect alterations as short as 1 Kb in size with a TPR greater than 80%. However, we suggest to use conservative values of the parameters (η and ω) to contain the type-II error.

References

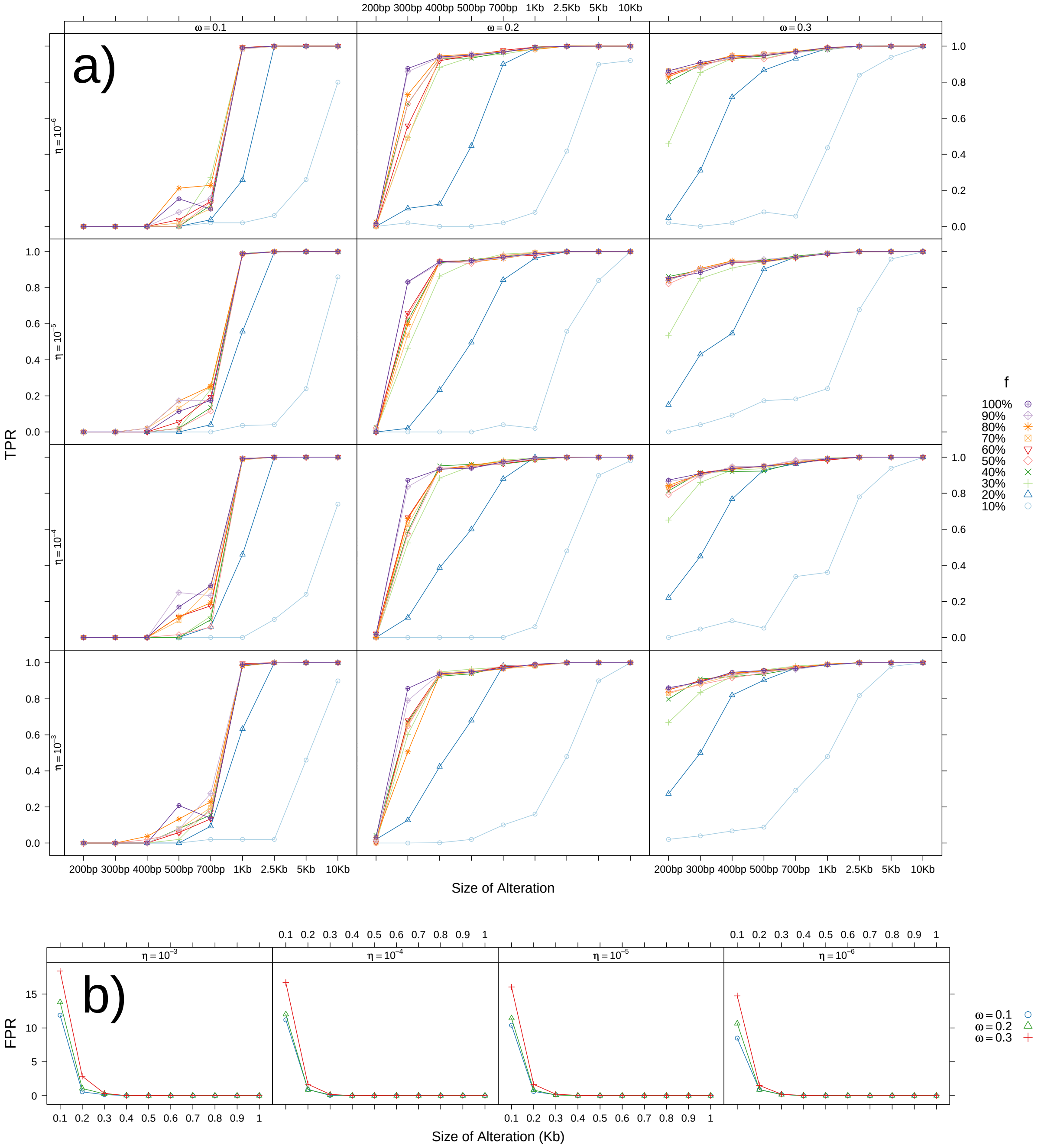
[1] Magi,A., Benelli,M., Marseglia,G., Nannetti,G., Scordo,M.R., Torricelli,F. (2010) A shifting level model algorithm that identifies aberrations in array-CGH data. *Biostatistics*, **11**, 265-280.

[2] Rabiner,L.R. (1988) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**, 257-286.

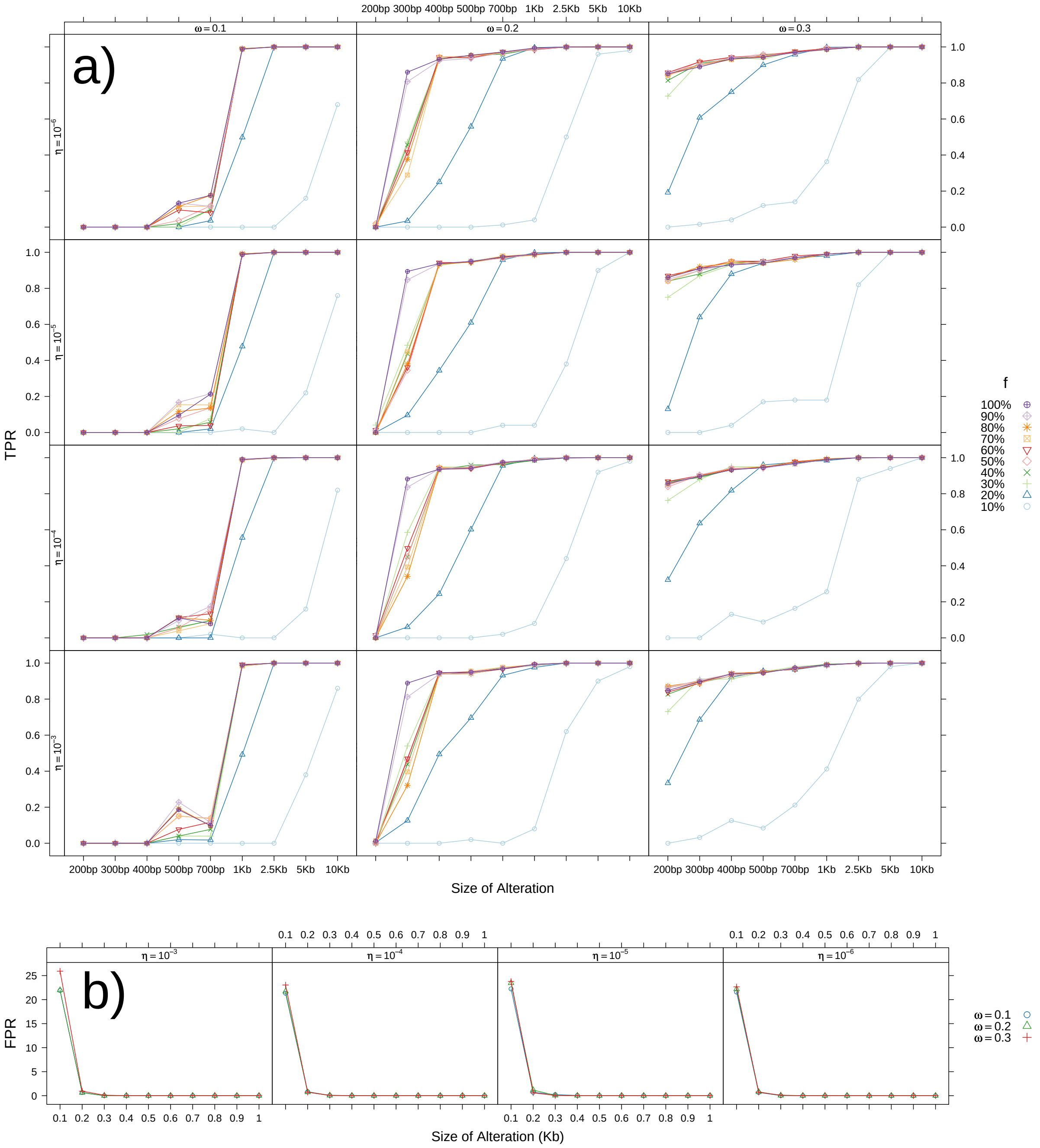
[3] Lai,W.R.R., Johnson,M.D.D., Kucherlapati,R. and Park,P.J.J. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array-CGH data. *Bioinformatics*, **21**, 3763-3770.

[4] Yoon, S., Xuan, Z., Makarov, V., Ye, K and Sebat, J. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res*, **19**, 1586-1592.

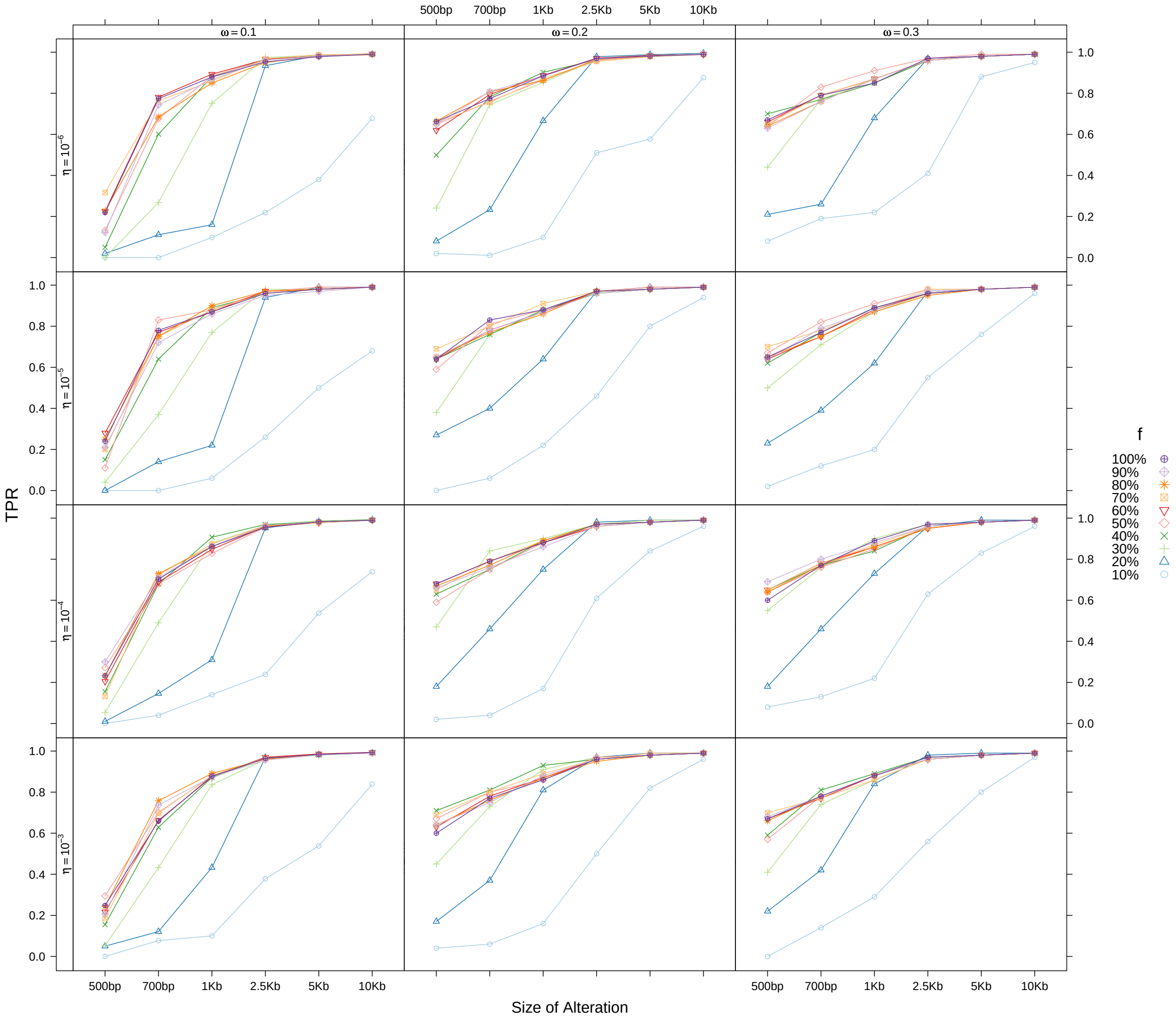
[5] McCarroll,S., Kuruvilla,F., Korn,J., Cawley,S., Nemesh,J., Wysoker,A., Shapero,M., de Bakker,P., Maller,J., Kirby,A. *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet*, **40**, 1166-1174.



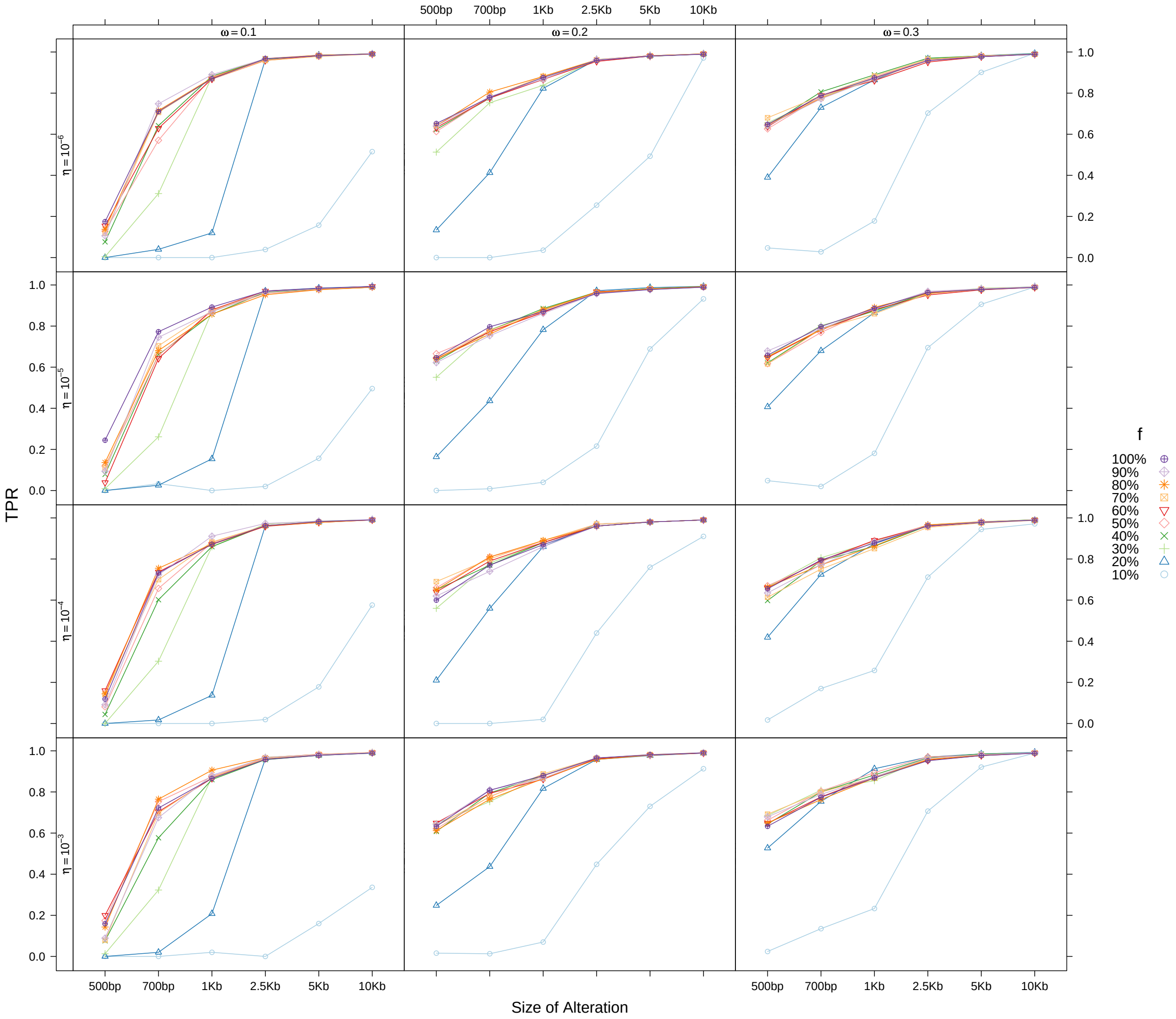
Supplemental Figure 1: TPR and FPR estimate for different values of η and ω on synthetic data made of 30 chromosomes. Each point of the plot is obtained by averaging the JoinSLM results over 100 repeated simulations. (a) Each curve represents the TPR estimate against deletion events of different size. In each plot are reported the curves for different values of fraction of altered samples f (with f that ranges between 0.1 and 1). (b) Each curve represent the FPR estimate against the size of false detected events.



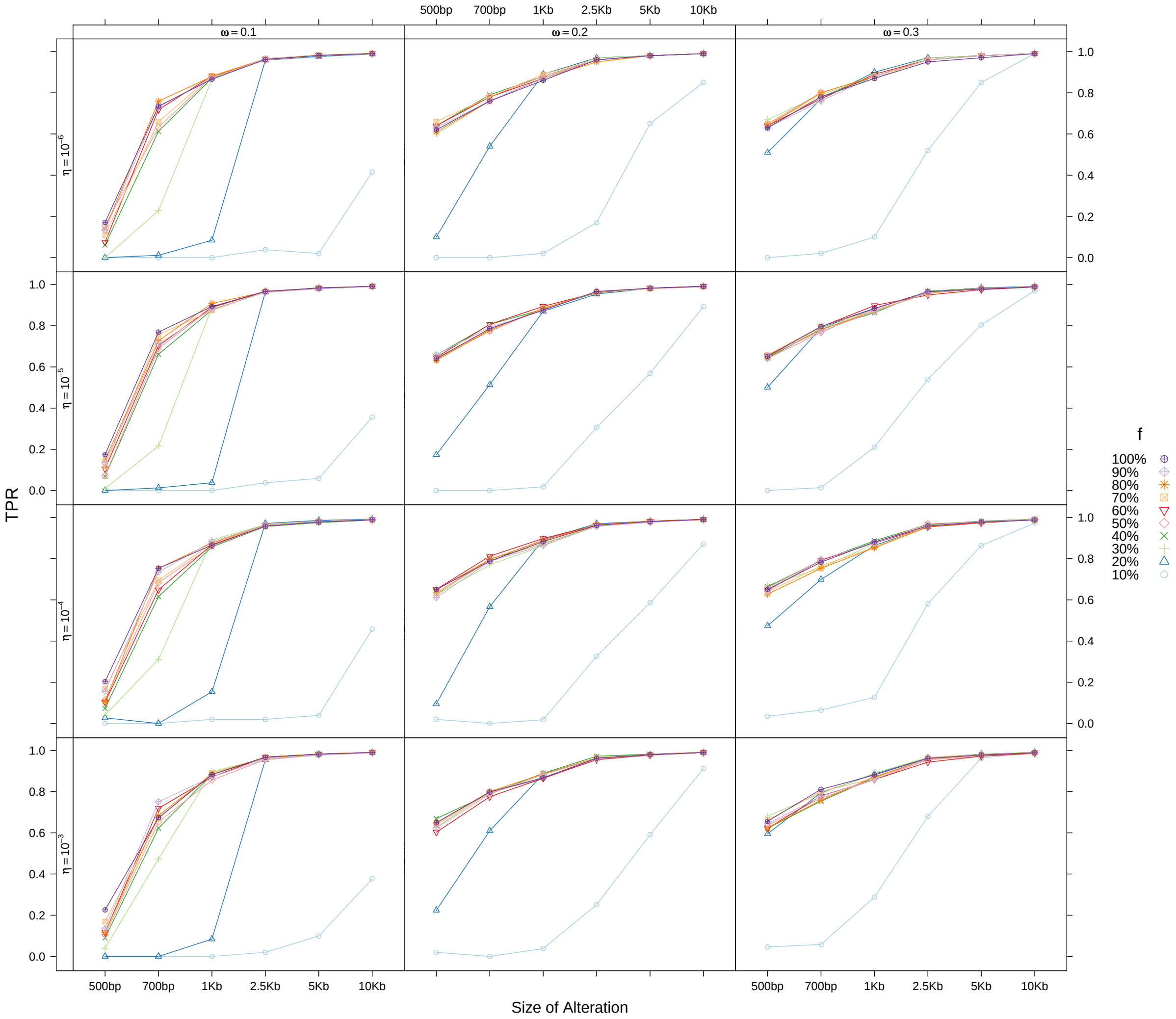
Supplemental Figure 2: TPR and FPR estimate for different values of η and ω on synthetic data made of 50 chromosomes. Each point of the plot is obtained by averaging the JoinSLM results over 100 repeated simulations. (a) Each curve represents the TPR estimate against deletion events of different size. In each plot are reported the curves for different values of fraction of altered samples f (with f that ranges between 0.1 and 1). (b) Each curve represent the FPR estimate against the size of false detected events.



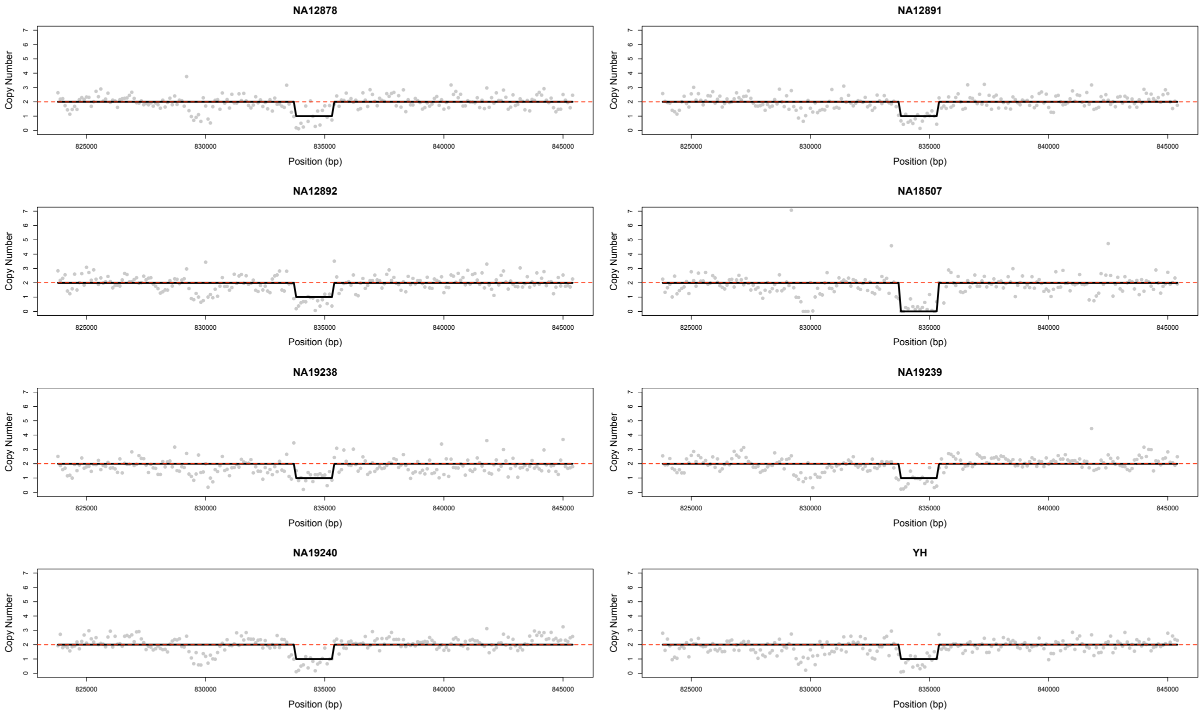
Supplemental Figure 3: TPR estimate for different values of η and ω on synthetic data made of 10 chromosomes with randomly shifted deletions. Each point of the plot is obtained by averaging the JoinSLM results over 100 repeated simulations. (a) Each curve represents the TPR estimate against deletion events of different size. In each plot are reported the curves for different values of fraction of altered samples f (with f that ranges between 0.1 and 1).



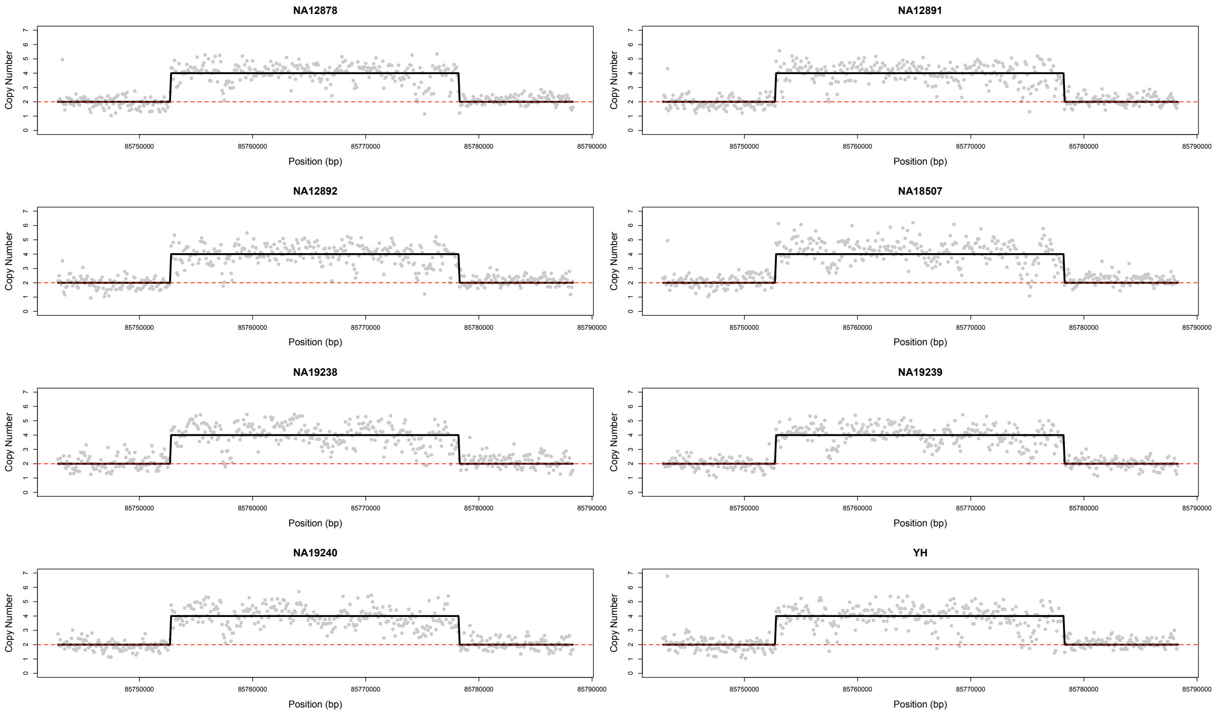
Supplemental Figure 4: TPR estimate for different values of η and ω on synthetic data made of 30 chromosomes with randomly shifted deletions. Each point of the plot is obtained by averaging the JoinSLM results over 100 repeated simulations. (a) Each curve represents the TPR estimate against deletion events of different size. In each plot are reported the curves for different values of fraction of altered samples f (with f that ranges between 0.1 and 1).



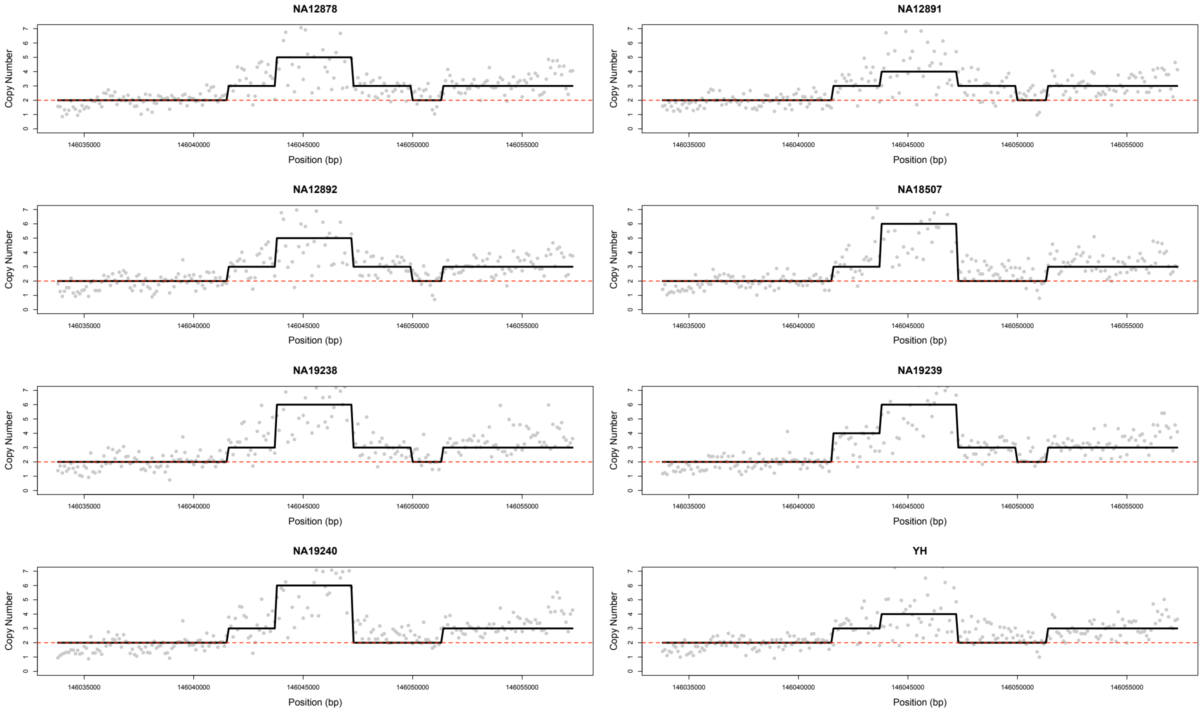
Supplemental Figure 5: TPR estimate for different values of η and ω on synthetic data made of 50 chromosomes with randomly shifted deletions. Each point of the plot is obtained by averaging the JoinSLM results over 100 repeated simulations. (a) Each curve represents the TPR estimate against deletion events of different size. In each plot are reported the curves for different values of fraction of altered samples f (with f that ranges between 0.1 and 1).



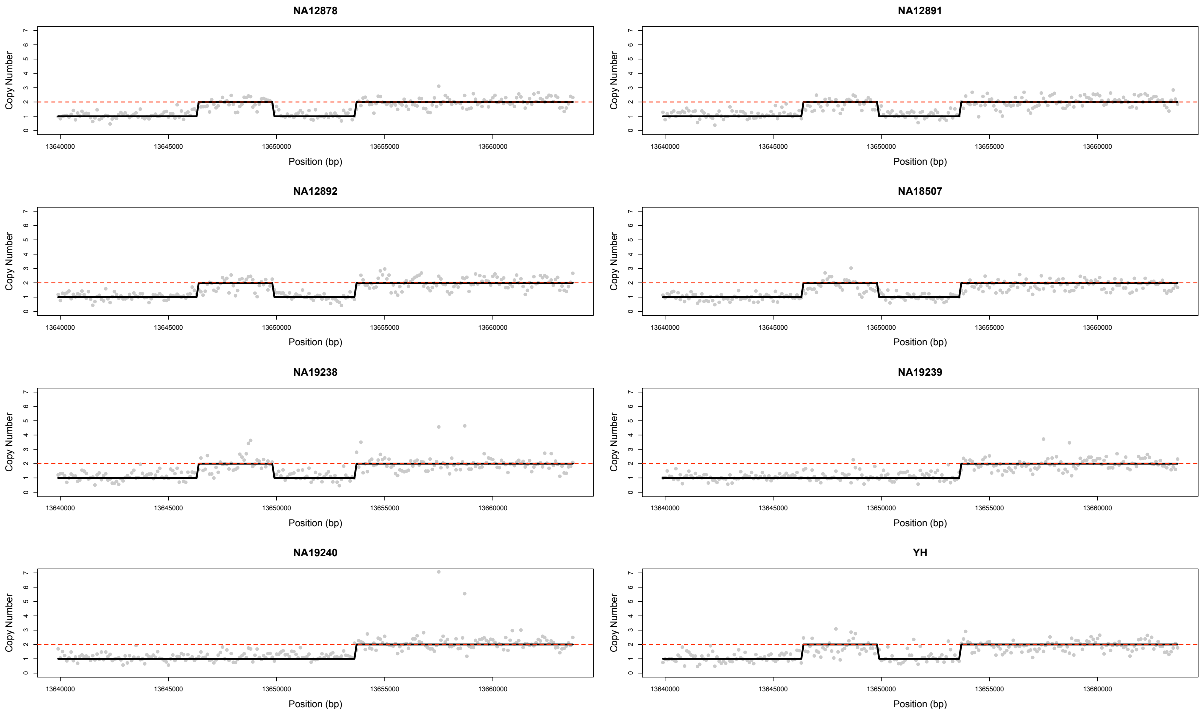
Supplemental Figure 6: Example of recurrent CNVs detected by JointSLM. We report an example of recurrent losses detected by JointSLM in the eight individuals. The x-axis represents the genomic coordinates (in Mb) and the y-axis represents read depth of coverage median-normalized to copy number 2. In each panel, plots are for NA12878, NA12891, NA12892, NA19238, NA19239, NA19240, NA18507, and YH from top to bottom and left to right. The genomic coordinates are chr1: 823,801-845,401.



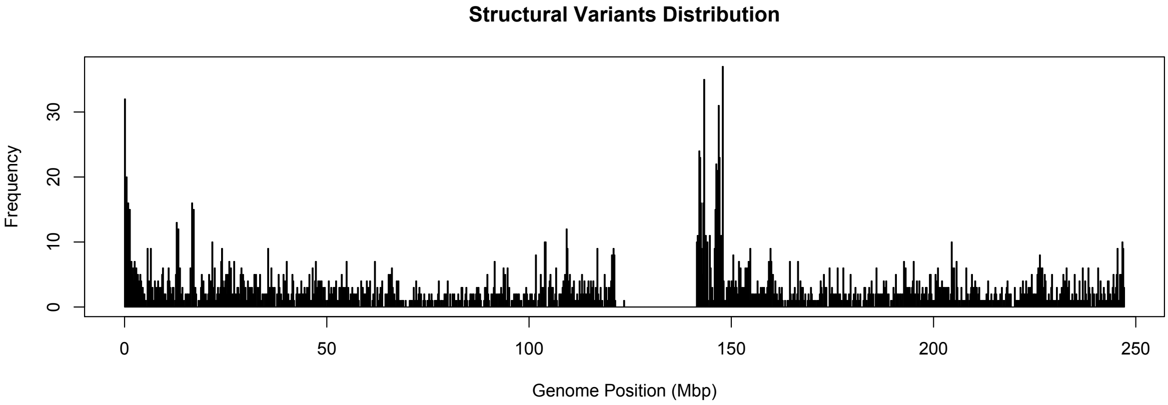
Supplemental Figure 7: Example of recurrent CNVs detected by JointSLM. We report an example of recurrent amplifications detected by JointSLM in the eight individuals. The x-axis represents the genomic coordinates (in Mb) and the y-axis represents read depth of coverage median-normalized to copy number 2. In each panel, plots are for NA12878, NA12891, NA12892, NA19238,NA19239, NA19240, NA18507, and YH from top to bottom and left to right. The genomic coordinates are chr1: 85,742,801-85,788,301.



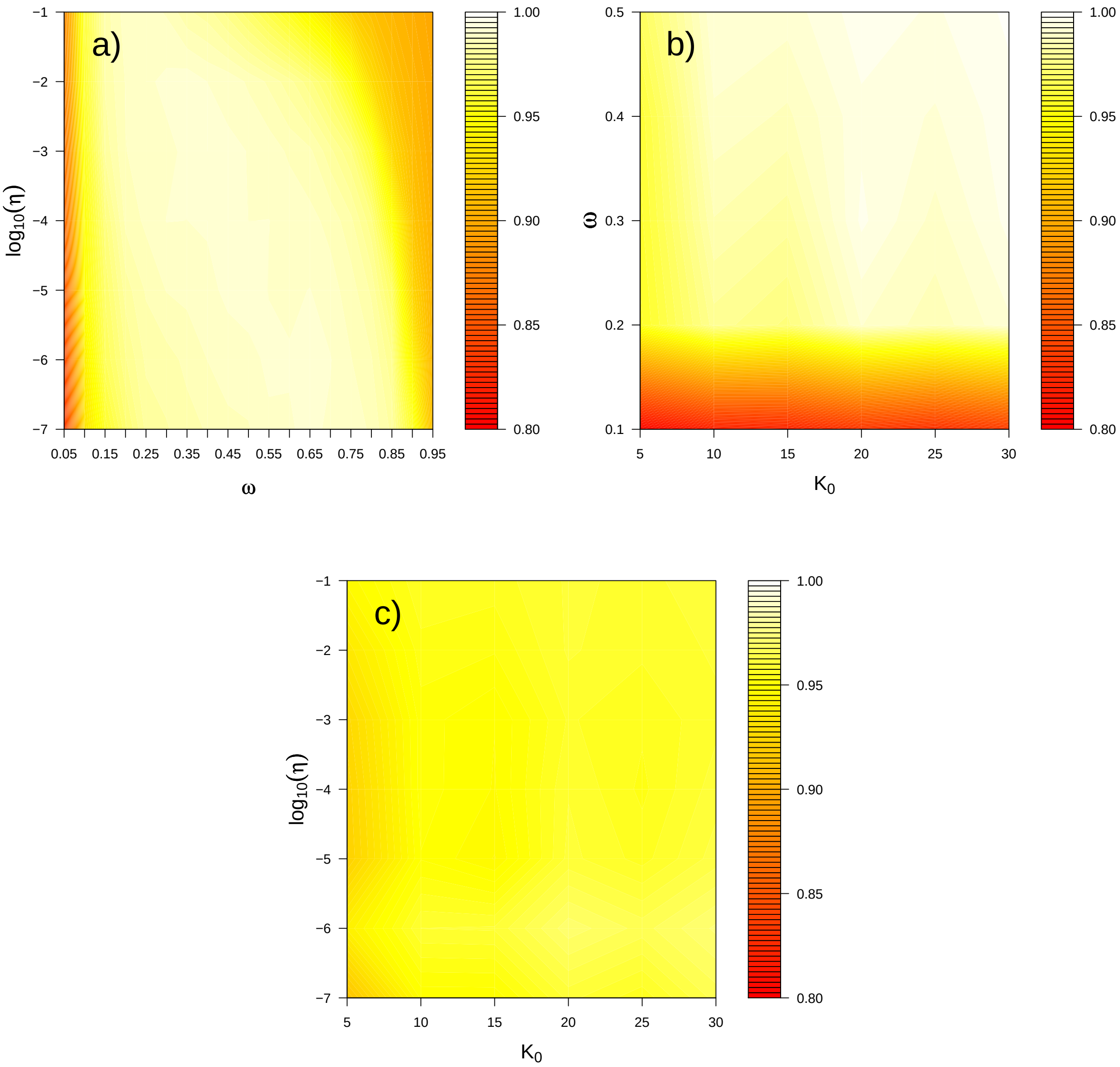
Supplemental Figure 8: Example of recurrent CNVs detected by JointSLM. We report an example of complex structural variation detected by JointSLM in the eight individuals. The x-axis represents the genomic coordinates (in Mb) and the y-axis represents read depth of coverage median-normalized to copy number 2. In each panel, plots are for NA12878, NA12891, NA12892, NA19238,NA19239, NA19240, NA18507, and YH from top to bottom and left to right. The genomic coordinates are chr1: 146,033,801-146,057,301.



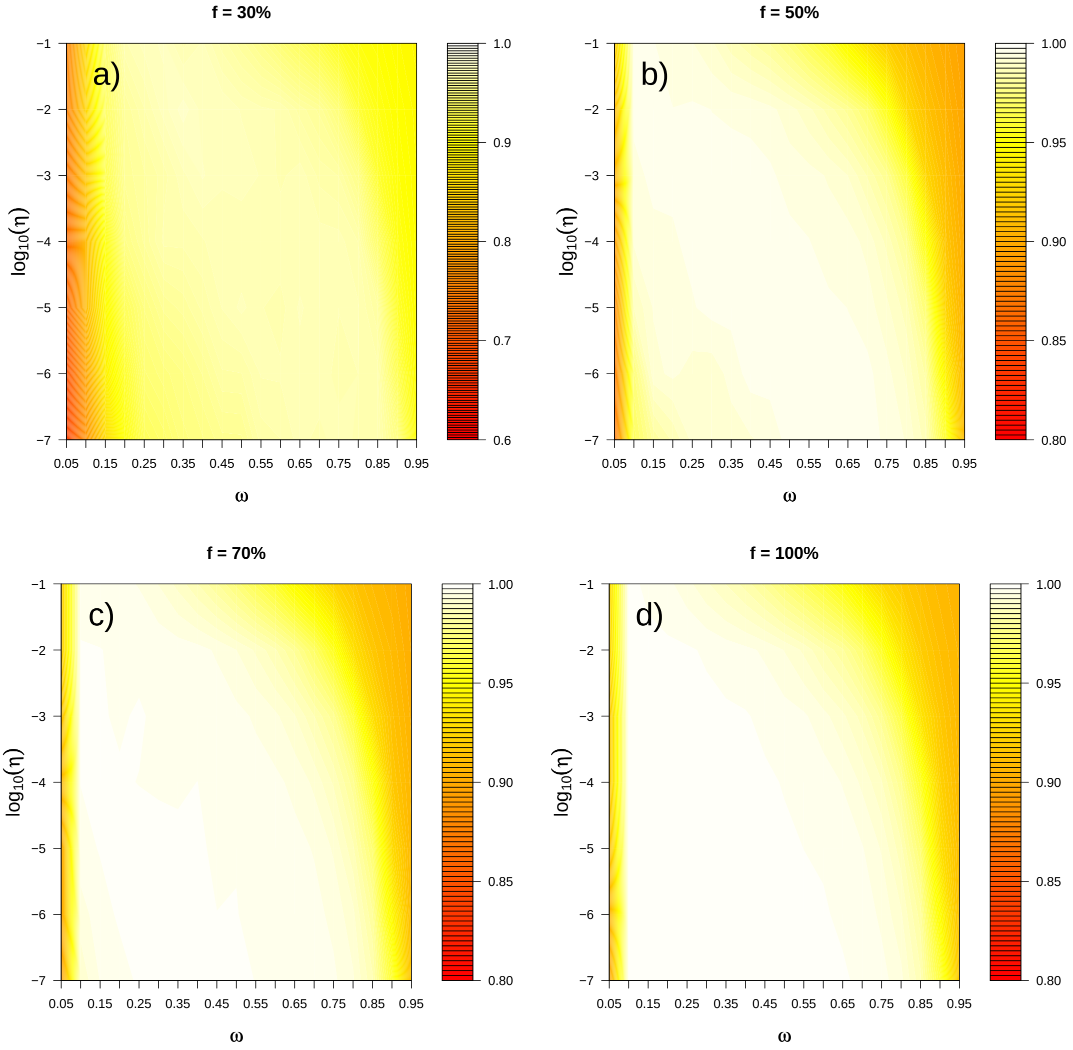
Supplemental Figure 9: Example of recurrent CNVs detected by JointSLM. We report an example of complex structural variation detected by JointSLM in the eight individuals. The x-axis represents the genomic coordinates (in Mb) and the y-axis represents read depth of coverage median-normalized to copy number 2. In each panel, plots are for NA12878, NA12891, NA12892, NA19238,NA19239, NA19240, NA18507, and YH from top to bottom and left to right. The genomic coordinates are chr1: 1,263,001-1,284,401.



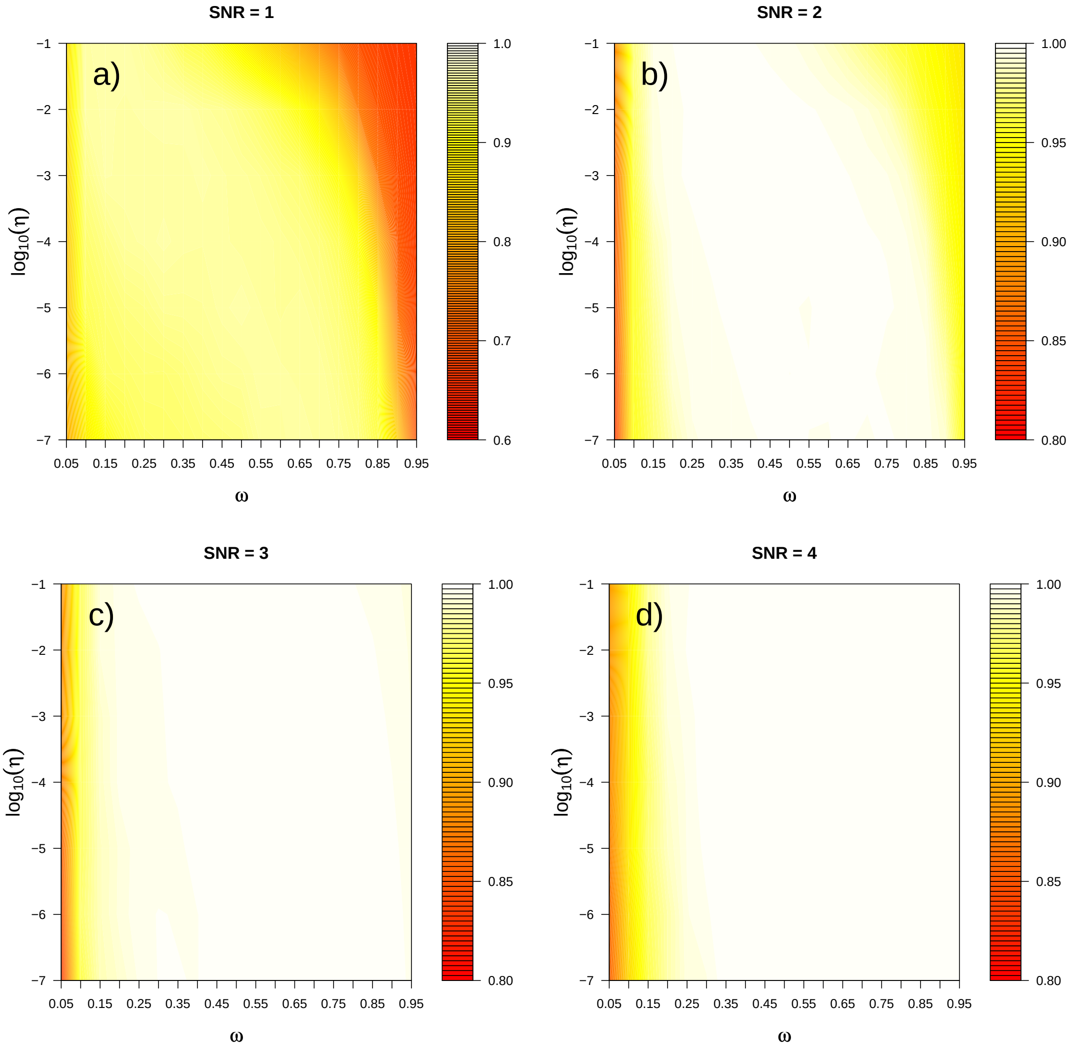
Supplemental Figure 10: Histogram of the CNV positions along chromosome 1. We observed an overrepresentation of the CNVs identified by JointSLM in the genomic regions close to telomeres and centromeres.



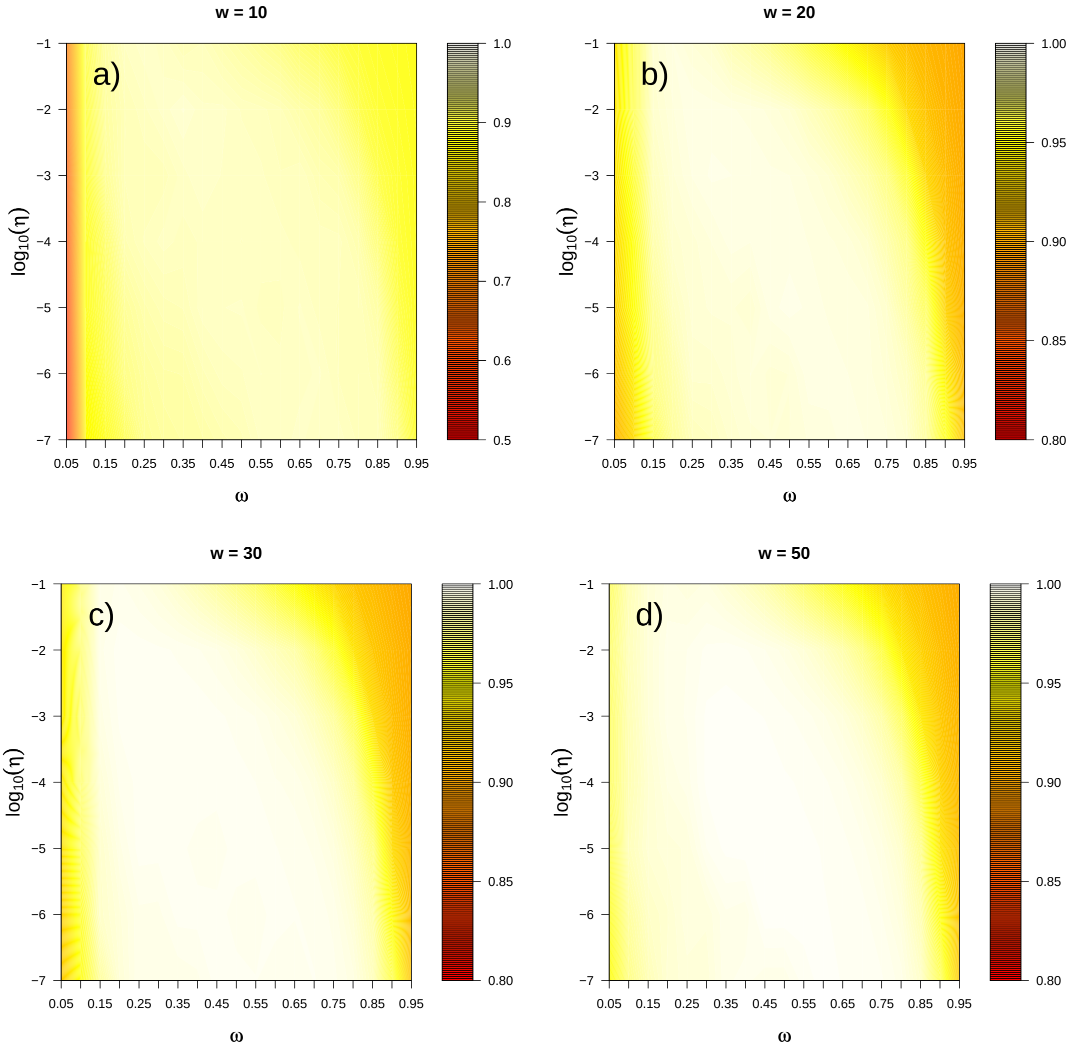
Supplemental Figure 11: Effect of parameters η, ω and K_0 on the global performance of the JointSLM algorithm. In each plot are reported the values of Area Under the ROC curve for different values of the parameters. Each point of the heatplot is obtained by averaging AUC over 100 repeated simulations. (a) The plot represents the performance of our algorithm while varying η and ω on the type-A simulation (see text for more details). (b) and (c) The plot reports the global performance of the algorithm while varying eta ω and K_0 on the type-B simulations.



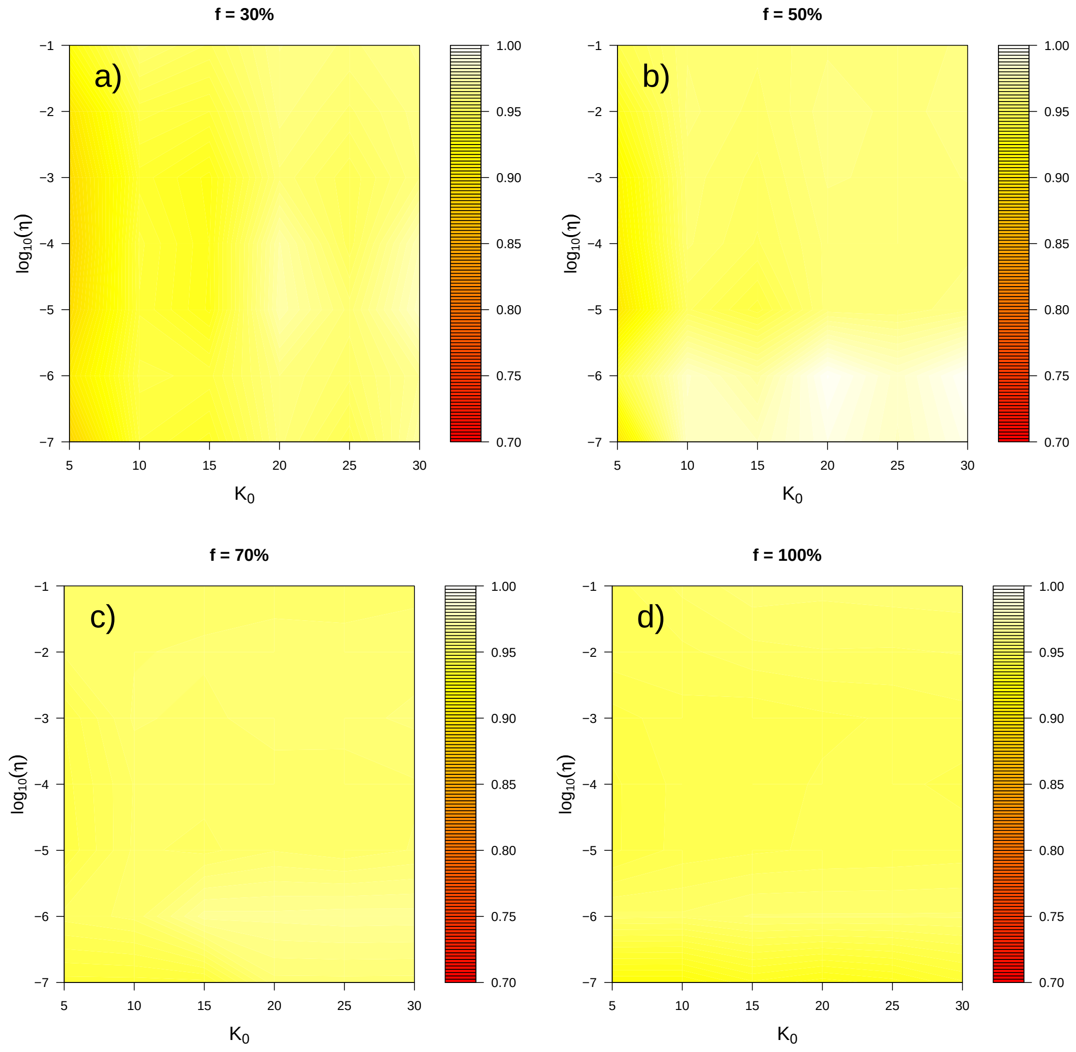
Supplemental Figure 12: Effect of parameters η and ω on the performance of the JointSLM algorithm on multiple synthetic chromosomes with different fraction f of altered samples. Each point of the heatplot is obtained by averaging AUC over 100 repeated simulations of type-A (see text for more details). (a) $f = 30\%$, (b) $f = 50\%$, (c) $f = 70\%$ and (d) $f = 100\%$.



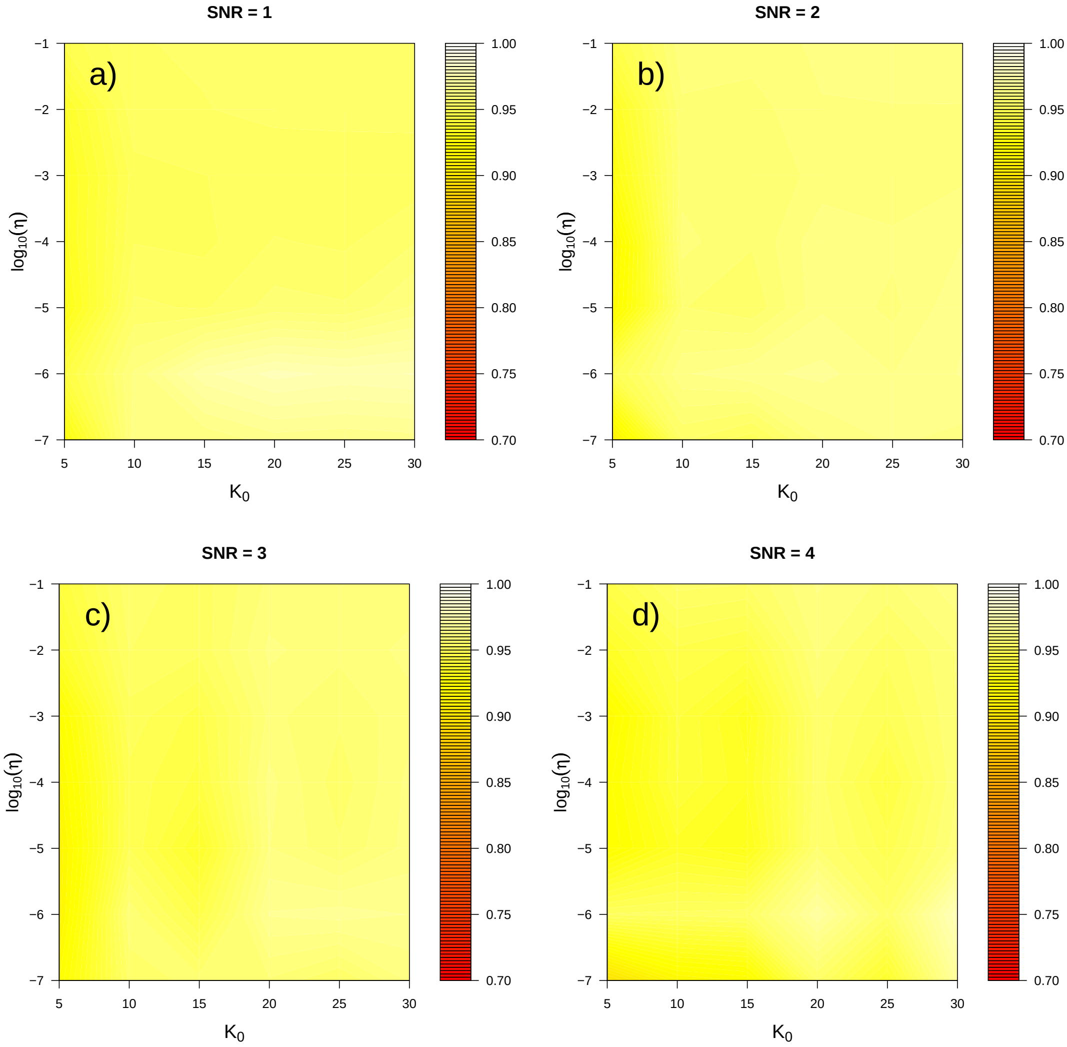
Supplemental Figure 13: Effect of parameters η and ω on the performance of the JointSLM algorithm on multiple synthetic chromosomes with different levels of Signal to Noise Ratios (SNR). Each point of the heatplot is obtained by averaging AUC over 100 repeated simulations of type-A (see text for more details). (a) SNR=1, (b) SNR=2, (c) SNR=3 and (d) SNR=4.



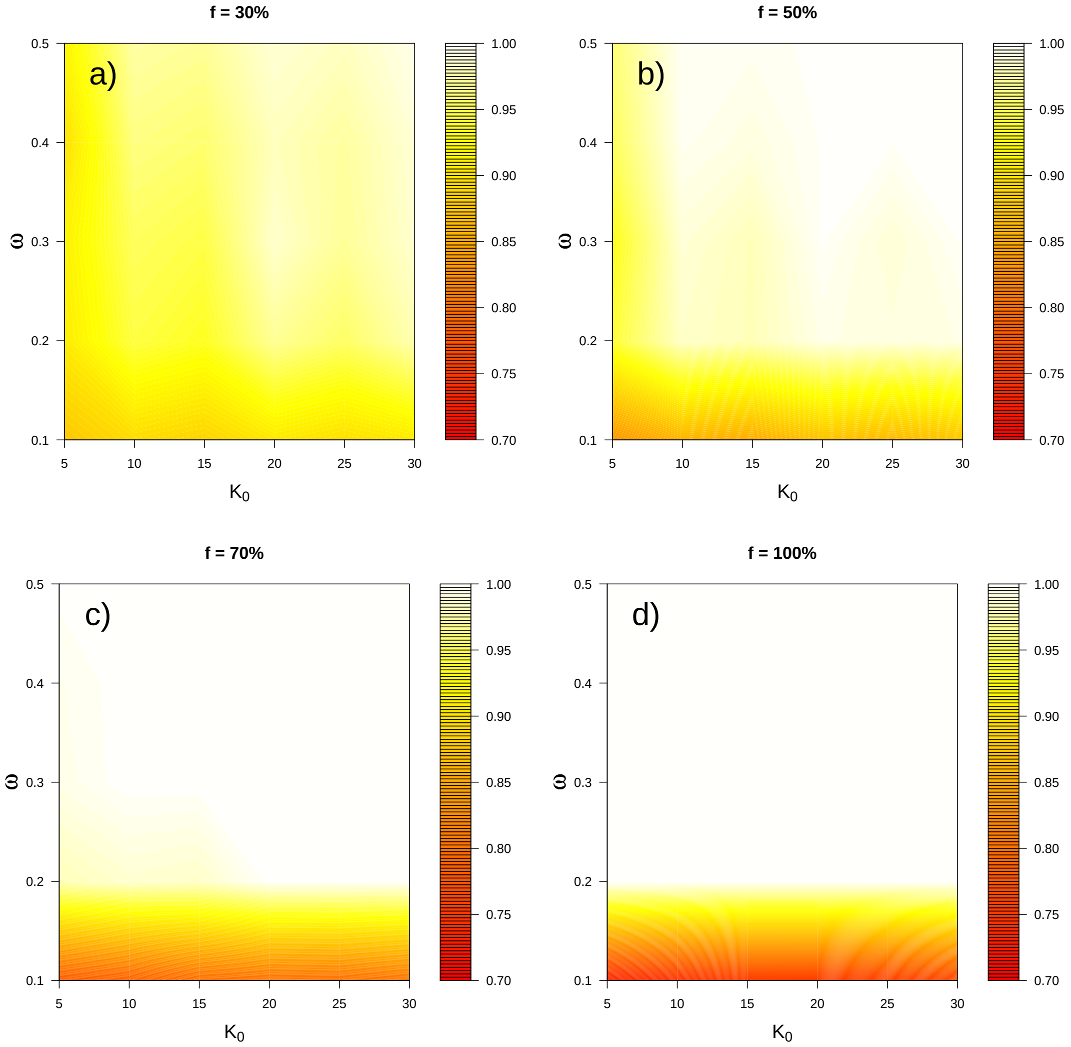
Supplemental Figure 14: Effect of parameters η and ω on the performance of the JointSLM algorithm on multiple synthetic chromosomes with alterations of different width (w). Each point of the heatplot is obtained by averaging AUC over 100 repeated simulations of type-A (see text for more details). (a) $w = 10$, (b) $w = 20$, (c) $w = 30$ and (d) $w = 50$.



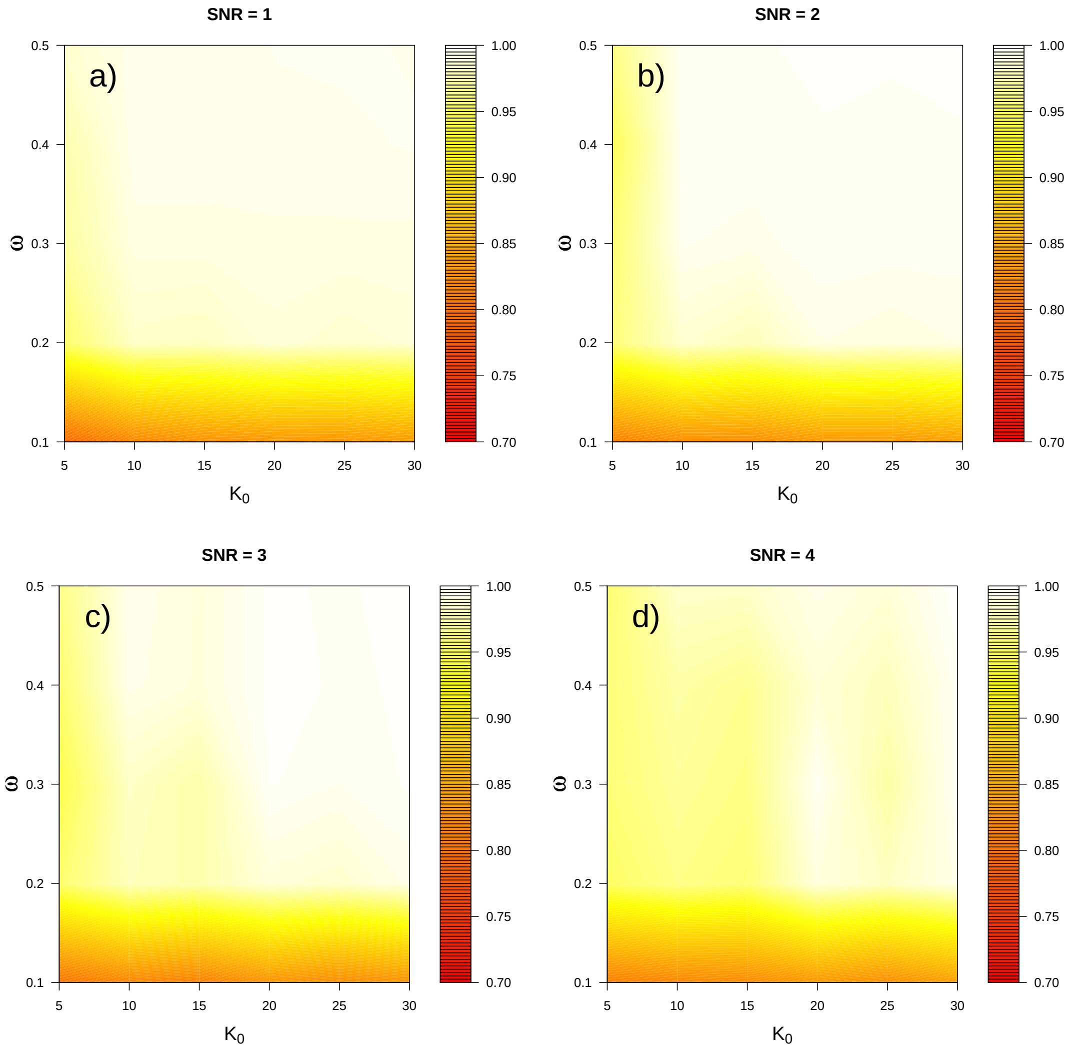
Supplemental Figure 15: Effect of parameters η and K_0 on the performance of the JointSLM algorithm on multiple synthetic chromosomes with different fraction f of altered samples. Each point of the heatplot is obtained by averaging AUC over 100 repeated simulations of type-B (see text for more details). (a) $f = 30\%$, (b) $f = 50\%$, (c) $f = 70\%$ and (d) $f = 100\%$.



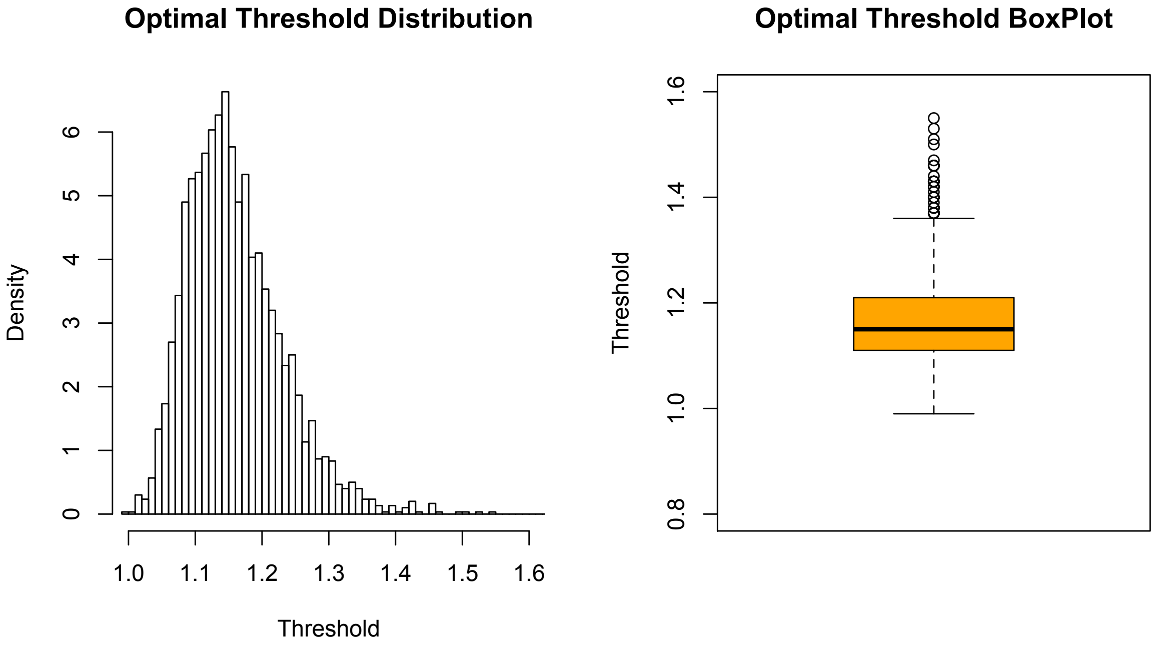
Supplemental Figure 16: Effect of parameters η and K_0 on the performance of the JointSLM algorithm on multiple synthetic chromosomes with different levels of Signal to Noise Ratios (SNR). Each point of the heatplot is obtained by averaging AUC over 100 repeated simulations of type-B (see text for more details). (a) SNR=1, (b) SNR=2, (c) SNR=3 and (d) SNR=4.



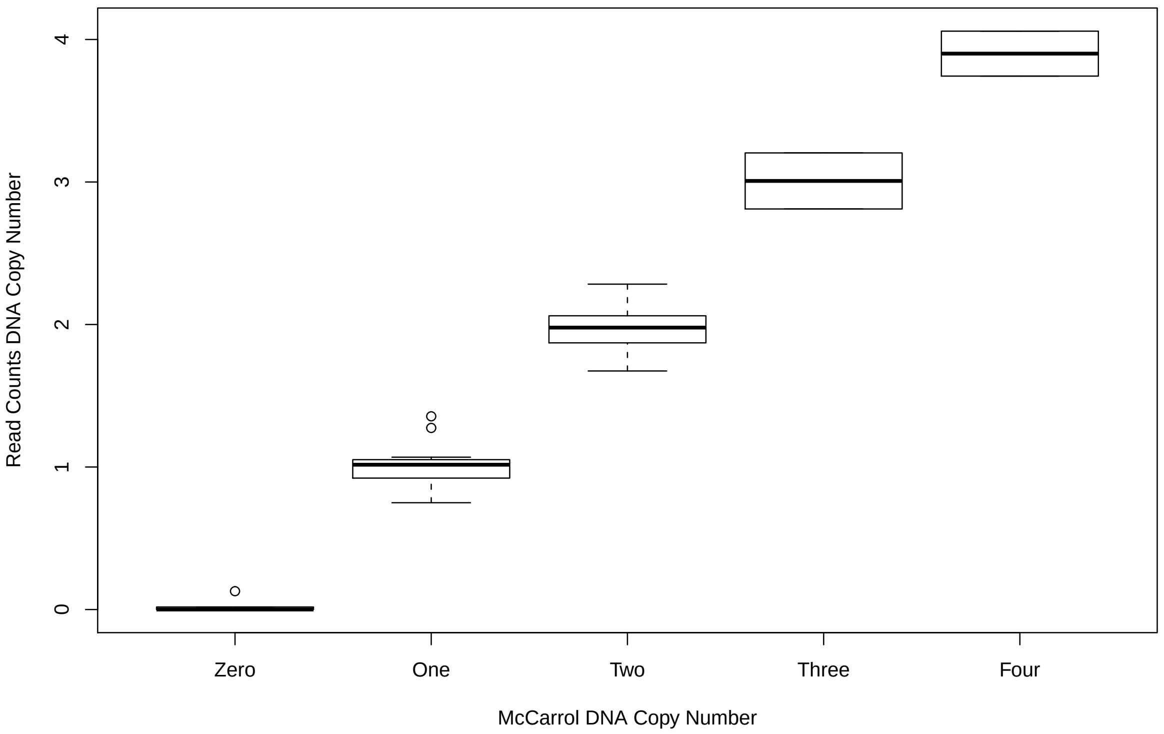
Supplemental Figure 17: Effect of parameters ω and K_0 on the performance of the JointSLM algorithm on multiple synthetic chromosomes with different fraction f of altered samples. Each point of the heatplot is obtained by averaging AUC over 100 repeated simulations of type-B (see text for more details). (a) $f = 30\%$, (b) $f = 50\%$, (c) $f = 70\%$ and (d) $f = 100\%$.



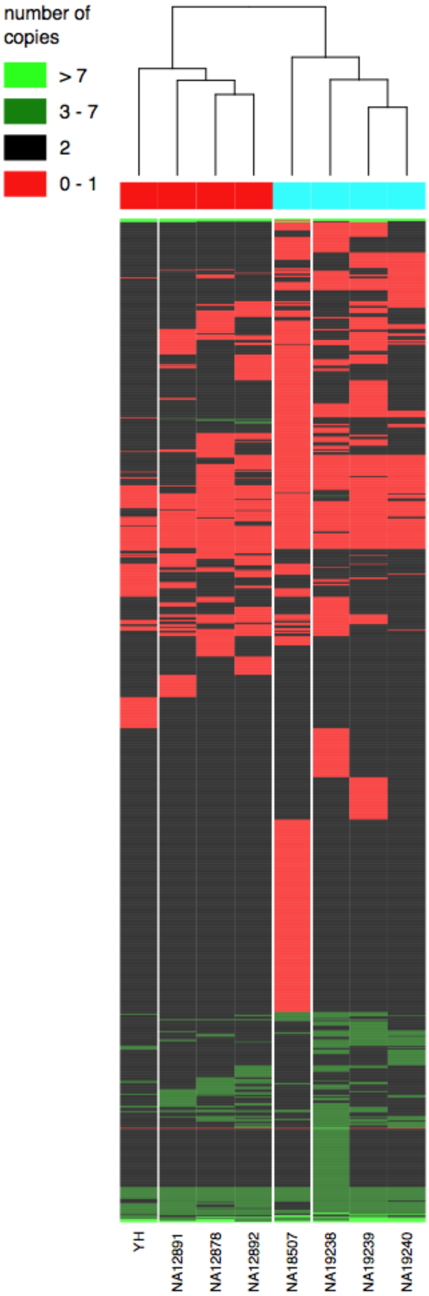
Supplemental Figure 18: Effect of parameters ω and K_0 on the performance of the JointSLM algorithm on multiple synthetic chromosomes with different levels of Signal to Noise Ratios (SNR). Each point of the heatplot is obtained by averaging AUC over 100 repeated simulations of type-B (see text for more details). (a) SNR=1, (b) SNR=2, (c) SNR=3 and (d) SNR=4.



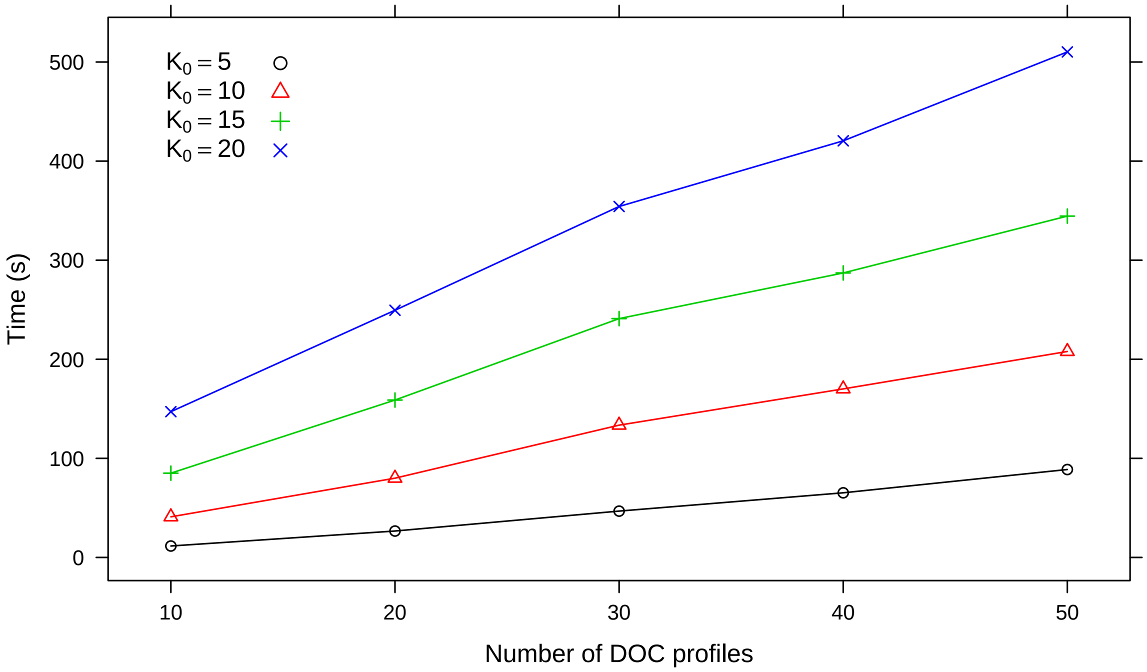
Supplemental Figure 19: Histogram and Boxplot of the optimal threshold values.



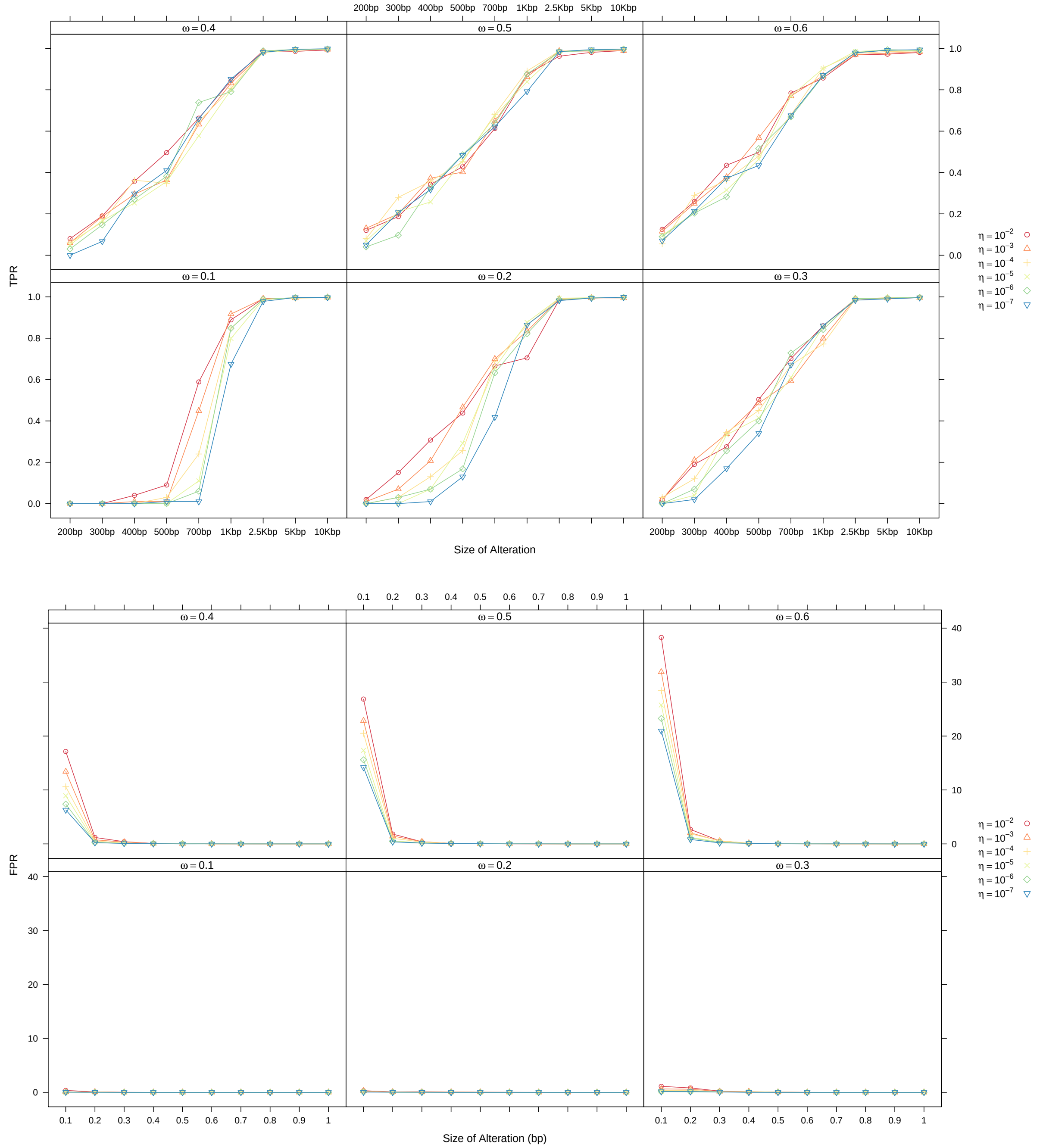
Supplemental Figure 20: Comparison between DNA copy number estimated by McCarrol et al. and the DNA copy number estimated by normalized read counts.



Supplemental Figure 21: Hierarchical clustering on the estimated copy number of the 1278 CNV regions detected by JointSLM on chromosome 1 that do not overlap with known CNVs of Database of Genomic Variants. Each row represents a separate CNVs region and each column a separate individual.



Supplemental Figure 22: Time taken by JointSLM in segmenting a 1 Mb synthetic chromosome while varying the number of multiple samples analysed and for different values of the parameter K_0 . Each value of the plot is obtained by averaging over 100 repeated simulations (see text for details).



Supplemental Figure 23: TPR and FPR estimate for different values of η and ω on single chromosome analysis. Each point of the plot is obtained averaging the JoinSLM results over 100 repeated simulations. (top) Each curve represents the TPR estimate against deletion events of different size. In each plot are reported the curves for different values eta. (bottom) Each curve represent the FPR estimate against the size of false detected events.

Supplemental Table 1: Summary statistics for the CNVs detected by JointSLM on chromosome 1 that do not overlap with the regions of Database of Genomic Variants. The number of CNVs are listed separately for different sizes and number of samples that share the alteration.

# Samples that share the alterations	$100 - 500$ bp	$500 - 1000$ bp	$1 - 5$ Kb	$5 - 10$ Kb	> 10 Kb
1	67	308	142	0	0
2	39	125	29	0	0
3	56	64	18	0	0
4	32	51	10	0	0
5	19	24	6	0	0
6	23	23	7	0	0
7	34	22	4	0	2
8	111	33	12	2	5
Total	381	660	228	2	7