Supplementary information to Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion

Ruibin Xi Angela G. Hadjipanayis Lovelace J. Luquette Tae-Min Kim
Eunjung Lee Jianhua Zhang Mark D. Johnson Donna M. Muzny
David Wheeler Richard Gibbs Raju Kucherlapati Peter Park

Contents

1 The statistical model

Suppose that two genomes are sequenced and mapped to a reference genome. In a cancer study, two genomes are just tumor genome and its matched normal genome. For simplicity, we will call the sequencing reads tumor reads and normal reads from the two genomes, respectively. In the following, we will concentrate on the sequencing reads that are mapped on a reference chromosome c. Given a random read R on the reference chromosome c, it consists of two pieces of information, the read's position S on c, and the read's sample information Y. The variable Y is a dichotomous random variable, which can only take values 1 (tumor reads) or 0 (normal reads). In the following, we will treat the random read R and its associated random vector (Y, S) as equivalent, i.e. $R = (Y, S)$. Assume that the joint likelihood of the random vector (Y, S) is $f(y, s)$. Then, by Bayes' formula, we have

$$\begin{aligned} f(y,s) &= Pr(Y=y|S=s)Pr(S=s) \\ &= Pr(Y=y|S=s)f_S(s), \end{aligned}$$

where $f_S(s)$ is the marginal distribution of the read's position S.

The distribution $f_S(s)$ is in fact a mixture of the distribution of tumor read's position $f_1(s)$ and the distribution of normal read's position $f_0(s)$, i.e.

$$\begin{aligned} f(s) &= Pr(Y=1)f_1(s) + Pr(Y=0)f_0(s) \\ &= \pi f_1(s) + (1-\pi)f_0(s), \end{aligned}$$

where $\pi = Pr(S = 1)$. The parameter π can be easily estimated by $\hat{\pi} = N_1/(N_1 + N_0)$, where N_1 and N_0 are the total number of tumor and normal reads that are mapped to the reference genome, respectively. The ratio $r(s) = f_1(s)/f_0(s)$ represents the difference between the two genomes, and $r(s)$ should be 1 if two genomes are identical. If a segment I is duplicated in the tumor genome, a random tumor read would be more likely mapped into the segment I than a random normal read, and hence the ratio $r(s)$ will be greater than 1 for $s \in I$. On the other hand, if the segment I is a deletion for the tumor genome, a random tumor read would be mapped into I with lower probability, so the ratio $r(s)$ will be less than 1 for $s \in I$. Thus, the region with $r(s)$ deviating substantially from 1 will be a CNV region. Chiang et al. (2008) essentially used $r(s)$ as copy ratio and developed a method to estimate $r(s)$. In this paper, we do not estimate $r(s)$ directly but try to estimate a transformation of $r(s)$.

Let $q_s = Pr(Y = 1|S = s)$ be the conditional probability that a random read R is a tumor read given its position s on c. Again by Bayes' formula, we have

$$\begin{aligned} q_s f_S(s) &= Pr(Y=1, S=s) \\ &= Pr(S=s|Y=1)Pr(Y=1) \\ &= f_1(s)\pi. \end{aligned}$$

Hence, we get the following

$$q_s = \frac{\pi f_1(s)}{\pi f_1(s) + (1-\pi)f_0(s)}. \tag{1}$$

Note that $r(s) = q_s(1-\pi)/[\pi(1-q_s)]$. Thus, if we can obtain a good estimate of q_s, we can easily get a good estimate of $r(s)$. If two genomes are identical, we have $f_1(s) = f_0(s)$ and the

probabilities q_s will be constant over the whole reference chromosome c. But if there are some CNVs, the probabilities q_s will be different in different regions of the genome. Thus, to identify the CNV regions, it is enough to identify the genomic regions that have different probabilities q_s.

Let n be the total number of sequencing reads mapped c. Let $R_1 = (y_1, s_1), \cdots, R_n = (y_n, s_n)$ be the n sequencing reads. Assume that the sequencing reads are indexed in increasing order of their mapped positions on the reference chromosome. Then, we have $s_i \leq s_{i+1}$ for $i = 1, \cdots, N-1$. The joint likelihood L_N is

$$L_n = \prod_{i=1}^{n} q_{s_i}^{y_i} (1 - q_{s_i})^{1-y_i} f(s_i). \tag{2}$$

Suppose that there are m breakpoints and let $\tau_1 < \tau_2 < \cdots < \tau_m$ be their chromosome positions. The copy ratios $r(s)$ within the segment (τ_j, τ_{j+1}) are constant. So, the probabilities q_s between any two consecutive breakpoints τ_j and τ_{j+1} would be the same. Denote $\tau_0 = 0$ and $\tau_{m+1} = L_c$, where L_c is the length of the reference chromosome c. Thus, the joint likelihood (2) can be written as

$$L_N = \prod_{j=0}^{m} \prod_{\tau_j < s_i \leq \tau_{j+1}} p_j^{y_i} (1 - p_j)^{1-y_i} f(s_i), \tag{3}$$

where p_j is the common probability q_s between the breakpoints τ_j and τ_{j+1} $(j = 0, \cdots, m)$. If the number of breakpoints is known and one is willing to assume certain form of the distribution function $f(s)$, the breakpoint positions may be estimated by maximizing the likelihood (3). However, the number of breakpoints is unknown in general, and a model with more breakpoints will generally give higher likelihood. Therefore, direct maximization of the likelihood cannot give good estimate of the breakpoint positions. In the next section, we propose to use the BIC as a criterion to estimate positions of the breakpoints and the parameters p_j without assuming a parametric form for the distribution function $f(s)$.

2 CNV identification using the BIC

The BIC was first introduced and discussed in Schwarz (1978) as a general criterion for model selection. The BIC of a model is in general defined as

$$\text{BIC} = -2 \log(L) + k \log(n),$$

where L is the likelihood function evaluated at the maximum likelihood estimate (MLE), k is the number of parameters of the model and n is the total number of observations. For the model (3) with m breakpoints, there are $m + 1$ parameters $p_0, p_1, \cdots, p_m$. We thus define the BIC for the model (3) as

$$\text{BIC}(\lambda) = -2 \sum_{j=0}^{m} \left[k_j \log(\hat{p}_j) + (n_j - k_j) \log(1 - \hat{p}_j) \right] - 2 \sum_{i=1}^{N} f(s_i) + (m+1) \lambda \log(N), \tag{4}$$

where k_j and n_j are the number of tumor reads and the total number of sequencing reads between the breakpoints τ_j and τ_{j+1}, and $\hat{p}_j = k_j / n_j$ is the MLE of the parameter p_j given the breakpoints.

Here, we introduce a tuning parameter $\lambda > 0$ to give more flexibility to our CNV identification method. When $\lambda = 1$, the BIC given in (4) becomes the standard BIC as defined in Schwarz (1978). Given two models, the one with a smaller BIC is preferred. Our goal is then to find a model that can minimize the BIC (4).

Note that the second term in (4) is common for all possible models and it will be canceled out when comparing two different models' BIC. It is therefore unnecessary to know the form of the distribution function $f(s)$ for our model selection procedure based on the BIC (4). For simplicity, we drop the second term in (4) and redefine the BIC as

$$\text{BIC}(\lambda) \quad = \quad -2\sum_{j=0}^{m}\left[k_j \log(\hat{p}_j) + (n_j - k_j)\log(1-\hat{p}_j)\right] + (m+1)\lambda \log(N), \tag{5}$$

The workhorse function of the BIC-SEQ method is BIC-SEQ-LEVEL(B, t, λ). Each *level* of the algorithm attempts to minimize the BIC of the current model by repeatedly choosing the *best* t consecutive bins—the t consecutive bins which minimize the BIC if merged—and merging them.

Given a list of B bins and a window size t, BIC-SEQ first creates $B-t+1$ sliding windows, each of which consisting of t consecutive bins. For each window w, MERGE(w) is simulated allowing the difference in BIC (BIC-DIFF(w): the difference between the model with w's bins merged and the model with w's bins separate) to be computed. Each $\langle$BIC-DIFF$(w), w\rangle$ pair is then inserted into a fast index keyed by BIC-DIFF to facilitate the greedy merging process.

With the fast index in hand, we select the window with the minimum key and merge its bins. This process is repeated until all windows with negative BIC-DIFF values are exhausted. It should be noted that this heuristic is greedy and that there is not, in general, an optimal substructure to guarantee a globally minimized BIC.

BIC-SEQ-LEVEL(B, t, λ)

```
 1  W ← list(overlapping windows of t bins in B)
 2  T ← EMPTY
 3  for w ∈ W
 4        do Simulate MERGE(w)
 5           d ← BIC-DIFF(w)
 6           INSERT(T, ⟨d, w⟩)
 7  while T ≠ EMPTY
 8        do ⟨d, w⟩ ← min T
 9           if d ≤ 0
10              then MERGE(w)
11              else return
```

The MERGE procedure, which we will not detail here, consists of bookkeeping operations to maintain the validity of the fast index T. Namely, there are $t-1$ bin mergers; $t-1$ permanent deletions from T; and t BIC-DIFF computations, each of which necessitates one deletion and reinsertion in T. For a red-black tree T, the complexity of MERGE is $O(t \log |W|)$.

The complexity of MERGE does not affect the complexity of the MERGE simulation. In fact, the simulation and BIC-DIFF computations run in $O(t)$ time. If T is a red-black tree, insertion requires $O(\log |W|)$ and constructing the fast index for a given level t requires $O(|W| \log |W|)$ time.

It follows that the complexity of BIC-SEQ-LEVEL is dominated by the time spent in the MERGE loop, which is $O(t \cdot |W| \cdot \log |W|)$.

BIC-SEQ-LEVEL is applied to the model for each value $t = 2, 3, \ldots$ until no merger of t_0 consecutive bins lowers the BIC of the model. Then, BIC-SEQ will go backwards from $t = t_0 - 1$ to 2 until some merger can be made for some $t \geq 2$ or no merger can be made for all $t = t_0 - 1 \cdots, 2$. In the first case, BIC-SEQ will again go forwards and try to made higher level mergers. In the latter case, BIC-SEQ will stop and return the segments.

3 Copy ratio estimation and significance assignment

Suppose that $\tau_1, \tau_2, \cdots, \tau_m$ are all breakpoints identified by the above algorithm and $\hat{p}_j$'s ($j = 0, \cdots, m$) are their associated estimated probabilities. By (1), the copy ratio r_j in the region $(\tau_j, \tau_{j+1}]$ can be estimated by

$$\hat{r}_j = \frac{\hat{p}_j(1-\hat{\pi})}{\hat{\pi}(1-\hat{p}_j)}, \tag{6}$$

where $\hat{\pi} = \frac{N_1}{N_0}$. However, there are often whole chromosomal or whole chromosomal arm gains or losses in many tumor genomes. The copy ratio estimated by (6) would be inaccurate for these arm-level or chromosomal-level CNVs. To get a more accurate copy ratio estimates of these CNVs, we remove the regions with $|\log_2(\hat{r}_j)| \geq 0.2$ and re-estimate π as $\hat{\pi} = \frac{\tilde{N}_1}{\tilde{N}_0}$, where $\tilde{N}_1$ and $\tilde{N}_0$ are the total numbers of tumor and normal reads in the regions with $|\log_2(\hat{r}_j)| < 0.2$. Then, we plug in the new estimate $\hat{\pi}$ to (6) and re-estimate the copy ratios.

Notice that the test for $r_j = 1$ is equivalent to the test for $p_j = \pi$. Under the null hypothesis that $r_j = 1$ or $p_j = \pi$, we have, approximately,

$$\sqrt{n_j}(\hat{p}_j - \pi) \sim \mathcal{N}(0, \pi(1-\pi)),$$

where n_j is the total number of reads in the region. If we use $\hat{\pi}$ as an estimate of π, the significance or p-value of the estimated copy ratio $\hat{r}_j$ is given by

$$\text{p-value} = 2\Phi\left(-\sqrt{\frac{n_j}{\hat{\pi}(1-\hat{\pi})}} \cdot |\hat{p}_j - \hat{\pi}|\right),$$

where Φ is the cumulative distribution function of the standard normal distribution.

4 Credible intervals of breakpoints

We now develop a Gibbs sampler to assign confidence intervals to the breakpoints given by BIC-seq. Given a genomic window (a, b), assume there is only one breakpoint τ. Suppose that $D = (R_{l_1}, \cdots, R_{l_2})$ are all the reads in the interval (a, b). Let p_1, p_2 be the probabilities of a read being a tumor read before and after the breakpoint τ. Then, conditional on τ, we have the following distribution

$$f(D|\tau, p_1, p_2) = \prod_{k=l_1}^{l_2} \left[p_1^{y_k}(1-p_1)^{1-y_k} I(s_k \leq \tau) + p_2^{y_k}(1-p_2)^{1-y_k} I(s_k > \tau)\right] f(s_k).$$

Assuming uniform priors on τ, p_1 and p_2, i.e. $\pi(\tau) = I(a < \tau < b)/(b-a)$, $\pi(p_1) = \pi(p_2) = 1$, the full conditional distribution of p_1 is

$$\begin{aligned} f(p_1|D,\tau,p_2) &\propto \prod_{a<s_k\le\tau} p_1^{y_k}(1-p_1)^{1-y_k} \\ &= p_1^{\sum_{a<s_k\le\tau} y_k}(1-p_1)^{\sum_{a<s_k\le\tau}(1-y_k)}. \end{aligned}$$

That is, the full conditional distribution of p_1 is the Beta distribution Beta($\sum_{a<s_k\le\tau} y_k+1, \sum_{a<s_k\le\tau}(1-y_k)+1$). Similarly, the full conditional distribution of p_2 is also a Beta distribution Beta($\sum_{\tau<s_k<b} y_k+1, \sum_{\tau<s_k<b}(1-y_k)+1$). The full conditional distribution of τ is

$$\begin{aligned} f(\tau|D,p_1,p_2) &\propto f(D|\tau,p_1,p_2)\pi(\tau) \\ &\propto \prod_{k=l_1}^{l_2} \left[p_1^{y_k}(1-p_1)^{1-y_k} I(s_k \le \tau) + p_2^{y_k}(1-p_2)^{1-y_k} I(s_k > \tau)\right] I(a<\tau<b), \end{aligned}$$

which can be easily sampled using its inverse cumulative distribution. Given these full conditional distributions, we then can use the Gibbs sampler to get a credible interval for the breakpoint τ.

5 FDR estimate and an alternative way of specifying λ

We use a resampling method to estimate the FDR of a set of CNV calls. Assume that there are N_n and N_t normal and tumor reads, respectively. We first pool the tumor and normal reads together. Then, we randomly sample (with replacement) N_1 and N_2 reads as tumor and normal reads from the pooled data, respectively, where N_1 and N_2 are random sampled from the binomial distributions Binom(N_n+N_t, π) and Binom($N_n+N_t, 1-\pi$) with $\pi = N_t/(N_t+N_n)$. Since the read positions of the resampled data are limited by the read positions of the original data set, the FDR estimate based on this direct resampling strategy would under-estimate the true FDR, especially for the low coverage data. To correct for this bias, after a read is resampled, we randomly perturb its position in a neighborhood U of the read. Suppose that n reads are mapped to a chromosome. Let s_1, s_2 be the smallest and largest mapped position of the n reads. The neighborhood U is chosen as $(s-d, s+d)$, where s is the mapped position of the resampled read and $d = (s_2-s_1)/n$ is roughly the mean of the distances between neighboring reads. Candidate CNVs are called using BIC-seq with the same criterion as used in the original data set. The ratio between the numbers of CNVs called from the resampled data and from the original data is then an FDR estimate. We repeat this process many times (e.g. 100 times) and take the mean of the FDRs as the final FDR estimate.

As discussed in the main text of the paper, the merging process of BIC-seq is equivalent to performing a series of likelihood ratio tests, allowing us to specify the number of type I errors in the merging process as an alternative way of specifying λ. Now, suppose that we are now trying to merge a pair of bins. Let $\mathrm{BIC}(\lambda) = -2\log(L) + (m+1)\lambda\log(N)$ and $\mathrm{BIC}_a(\lambda) = -2\log(L_a) + m\lambda\log(N)$ be the BIC before and after merging the pair, where L and L_a are the corresponding likelihood. Then, we would merge the pair if the BIC difference is less than 0, i.e. if $2\log(L) - 2\log(L_a) < \lambda\log(N)$. We have that the log likelihood difference $2(\log(L)-\log(L_a))$ is asymptotically χ^2_1 distributed under the null that two bins contain no breakpoint, or in other words, the two bins are homogeneous (see, for example, Bickel and Doksum (2001)). Suppose that X is χ^2_1-distributed, $\alpha(\lambda, N) = Pr(X \ge$

$\lambda \log(N)$), and that there are K initial bins. The merging process need perform $K-1$ bin-pair merging attempts, and the expected number of type I errors in the merging process is thus about $(K-1)\alpha(\lambda, N)$. On the other hand, if the expected number of type I errors in the merging process is given, we can easily obtain the corresponding λ using the inverse cumulative distribution function of the χ^2_1 distribution. Similarly, the likelihood ratio test corresponding to the tuples of k bins ($k > 2$) is asymptotically χ^2_{k-1} distributed under the null that the k bins contain no breakpoint. However, the critical value of χ^2_{k-1}-distribution for a given significance level is different for different k. And, given a set of initial bins, it is hard to give a prior estimate of the number of k-tuple bin merging attempts ($k \geq 3$). Therefore, we modified BIC-seq to only merge pairs of bins, if the number of expected type I errors in the merging process is provided.

6 Distribution of short reads on the reference genome

In this section, we show that the distribution of read counts cannot be well described with the Poisson distribution and negative binomial distribution of constant parameter (Figure S1). We aligned the short reads from three normal genomes considered in the paper, GBM, AML1 and AML2, onto the reference genome (hg18) allowing for 3 mismatches. Then, we binned the uniquely aligned short reads into 1000 bp bins. Figure S1 shows the plots of the empirical quantile against the theoretical quantile (quantile-quantile plot or QQ-plot). The theoretical quantiles were calculated based on the Poisson and the negative binomial model. The parameters for the Poisson and negative binomial distribution were estimated based on the read counts in the bins. For example, the parameter in the Poisson distribution is estimated as the mean of read counts in all bins. For each genome, we re-scaled both theoretical and empirical quantile by the total number of reads (in millions) to make the QQ-plot of the three genomes comparable. Figure S1 illustrates that the semi-parametric model proposed in this paper is more appropriate for CNV detection using sequencing data.

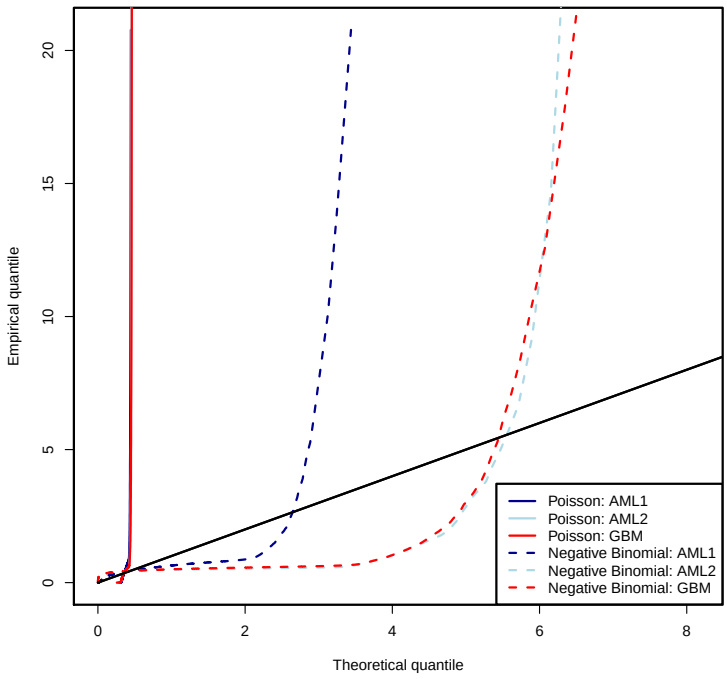


Fig. S1: QQ-plot of read counts in 1 Kb bins. The black line corresponds to the line $y = x$.

7 Examples of outliers

Some genomic positions can have several magnitude more reads than their neighboring positions. In the paper, we described methods to remove these outliers before further analysis. Figure S2 shows some examples of these genomic positions.

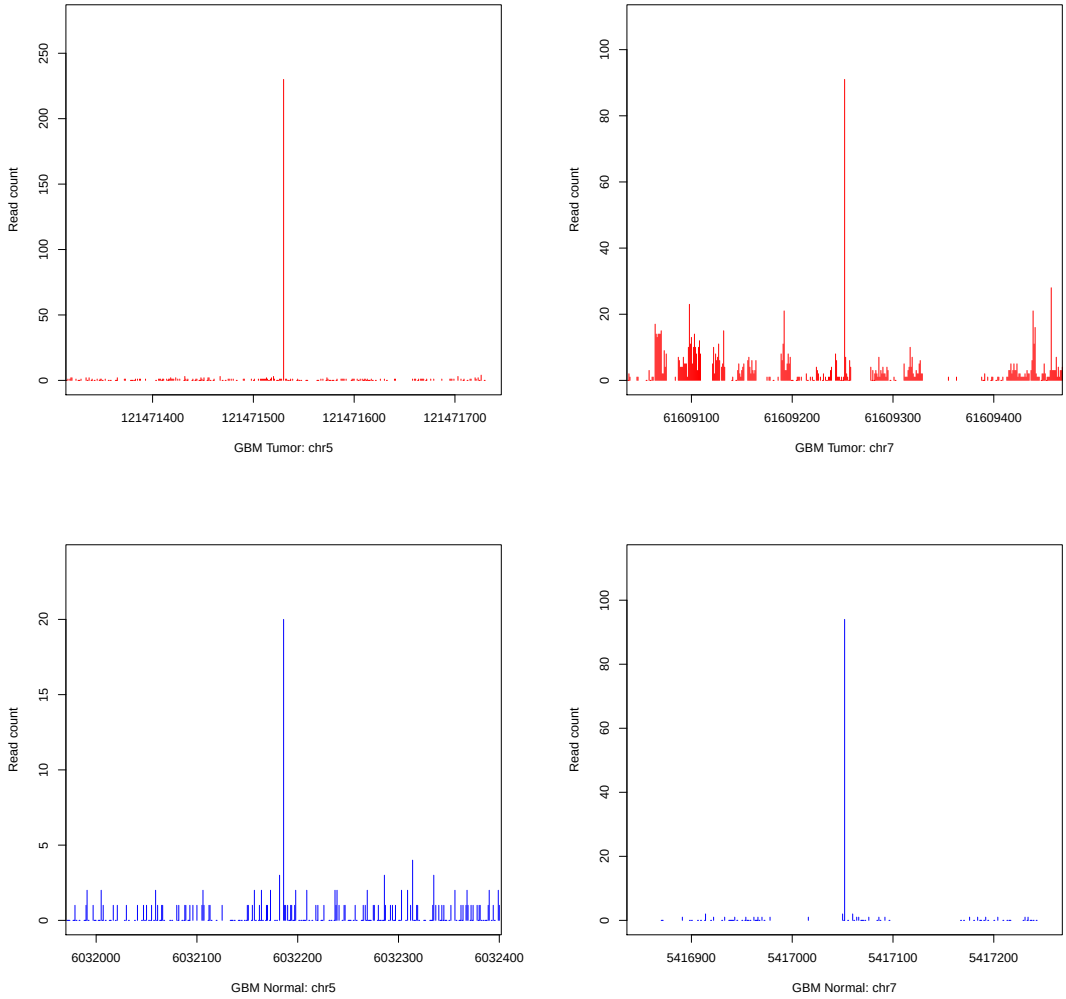


Fig. S2: Examples of genomic positions that have considerably more reads than their neighboring positions.

8 Effect of the tuning parameter λ

In this section, we use simulation to evaluate the effect of the tuning parameter λ on the algorithm's sensitivity and specificity.

To make the simulation data sets more realistic, all simulated data sets were generated based on the sequencing reads on chromosome 5 from the NCI-H2347 cell line (Chiang et al., 2008). Assume that N_n sequence reads from normal sample and N_t sequence reads from tumor sample are aligned

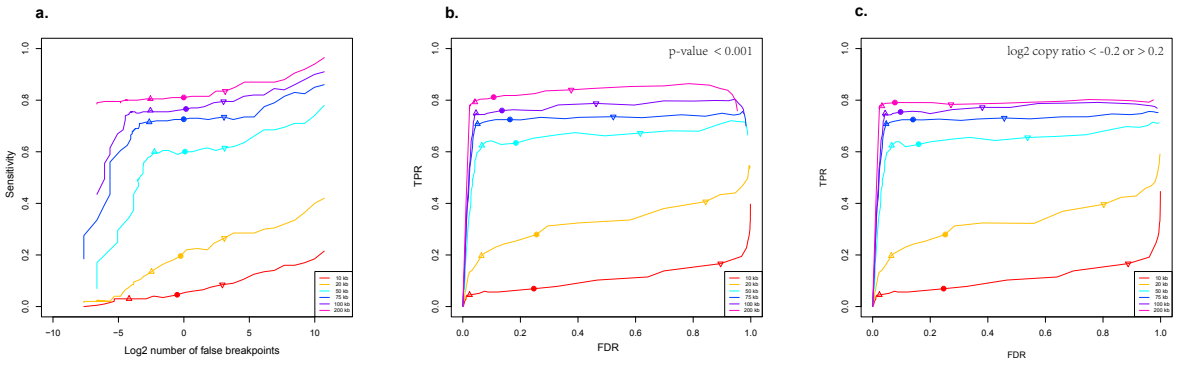


Fig. S3: (**a**) Sensitivity against the log2 number of false breakpoints: the point-up triangles, the bullets and the point-down triangles correspond to $\lambda = 1.2$, 1 and 0.8, respectively. (**b**) The TPR against the FDR: the point-up triangles, the bullets and the point-down triangles correspond to $\lambda = 1.2$, 1 and 0.8, respectively. The candidate CNVs were chosen as regions with p-value less than 0.001. (**c**) Similar to (**b**). The candidate CNVs were chosen as regions with log2 copy ratio less than -0.2 or greater than 0.2.

to a reference chromosome c . Let us call the set of normal sequence reads D_n and the set of tumor sequence reads D_t. We created a larger data set D_v by adding to D_n virtual reads between all pairs of adjacent normal reads. Then, we randomly partitioned the reads in D_v into "normal" reads and "tumor" reads. To create an aberrant region of length L, we randomly selected a chromosome position m and replaced the sequence reads between chromosome positions m and $m + L$ with the corresponding tumor reads in D_t. Since the real tumor/normal ratio is $\sim$1.4, this procedure could generate a one copy gain CNV region in the simulated tumor genome. The above method of generating simulated sequencing data set was the "spike-in" method used in Chiang et al. (2008) for optimizing the tuning parameters in the algorithm SegSeq. The length L of the aberrant region was set as 10 Kb, 20 Kb, 50 Kb, 75 Kb, 100 Kb and 200 Kb. We generated 200 sequencing data sets for each choice of L. The coverage of the real data set is about 0.3X.

We used a variety of values for λ and ran BIC-seq on these 200 data sets for each λ. The range of the tuning parameter λ was from 0.4 to 16. The initial bins we used were 1000 bp equally spaced bins. We recorded a true discovery if both predicted positions of the breakpoints m and $m + L$ were less than $0.3L$ base pairs away from the breakpoints m and $m + L$, respectively. All other breakpoints were recorded as false positives.

Figure S3 (**a**) plots the sensitivity of BIC-seq against the logarithm of the number of false breakpoints. The sensitivity here is defined as the mean of the number of true positives over the 200 data sets. The number of false breakpoints on the x-axis is defined as the mean of the number of false positives for each data set. The point-up triangles, the bullets and the point-down triangles represent $\lambda = 1.2$, 1 and 0.8, respectively. As the tuning parameter λ gets smaller, both sensitivity and number of false breakpoints are increasing. At this sequencing depth, BIC-seq achieves good sensitivity while keeping low FDR rates for $\lambda \sim 1$.

We calculated true positive rates (TPR) and false discovery rates (FDR) for each choice of λ and each size L of the aberrant region and plotted the TPR against FDR by varying λ. The TPR is defined as the length of the significant regions inside the aberration divided by the size of the aberration. The FDR is defined as the length of the significant regions outside the aberration divided by the total length of the significant regions. Figure S3 (**b**,**c**) shows the relationship between

TPR and FDR as λ varies for different aberration sizes. CNVs in Figure S3 (**b**) and (**c**) were chosen as regions with P-value less than 0.001 and regions with log2 ratio < -0.2 or > 0.2, respectively. Here, we are most interested in the effect of λ. A too stringent P-value cutoff would filter out many false positives predicted with smaller λ (e.g., $\lambda < 1$), but a too lenient p-value cutoff would call the regions with copy ratios close to one as CNV candidates. The significant level used in Figure S3 (**b**) can remove the regions whose copy ratios are close to one while leaving others as candidates. This allows us to examine the effect of the tuning parameter λ clearly, with minimal effect from the choice of the p-value cutoff. This p-value cutoff is only used in this section. Generally speaking, both TPR and FDR tend to increase as the tuning parameter λ gets smaller. However, for aberration size 50 Kb, 75 Kb, 100 Kb and 200 Kb, the TPR starts to decrease as λ becomes smaller in the neighborhood of $\lambda = 0.4$ (Figure S3 (**b**)). The reason for this is that when λ was small, the whole aberrant region was divided into smaller, sometimes not significant regions by false breakpoints inside the aberration; while for larger λ, these smaller regions were merged into larger statistically significant regions. So a λ slightly larger than 0.4 can give higher TPR than $\lambda = 0.4$.

The above simulation study shows the effect of λ for low coverage data set. We now study the effect of λ for the data sets of 0.3X, 3X and 30X sequencing coverage. The data sets used here are the simulation data sets in the main text (Materials and Methods). We run BIC-seq on these data sets with $\lambda = 1$, 1.2, 1.5, 2, and 4, and with 1 expected type I error in the merging process. To clearly show the effect of λ, we select candidate CNVs as regions with log2 copy ratios greater than 0.2 or less than -0.2. Figure S4 illustrates the power and the boxplots of the observed FDR and the estimated FDR (based on 100 resamplings) of the 100 simulations. This demonstrates that while the choice of λ is important for small and low copy change CNVs, the power for detecting large and high copy change CNVs are roughly the same for all λ used here. For instance, at 30X coverage, the power is almost the same for $\lambda = 1$ and $\lambda = 2$.

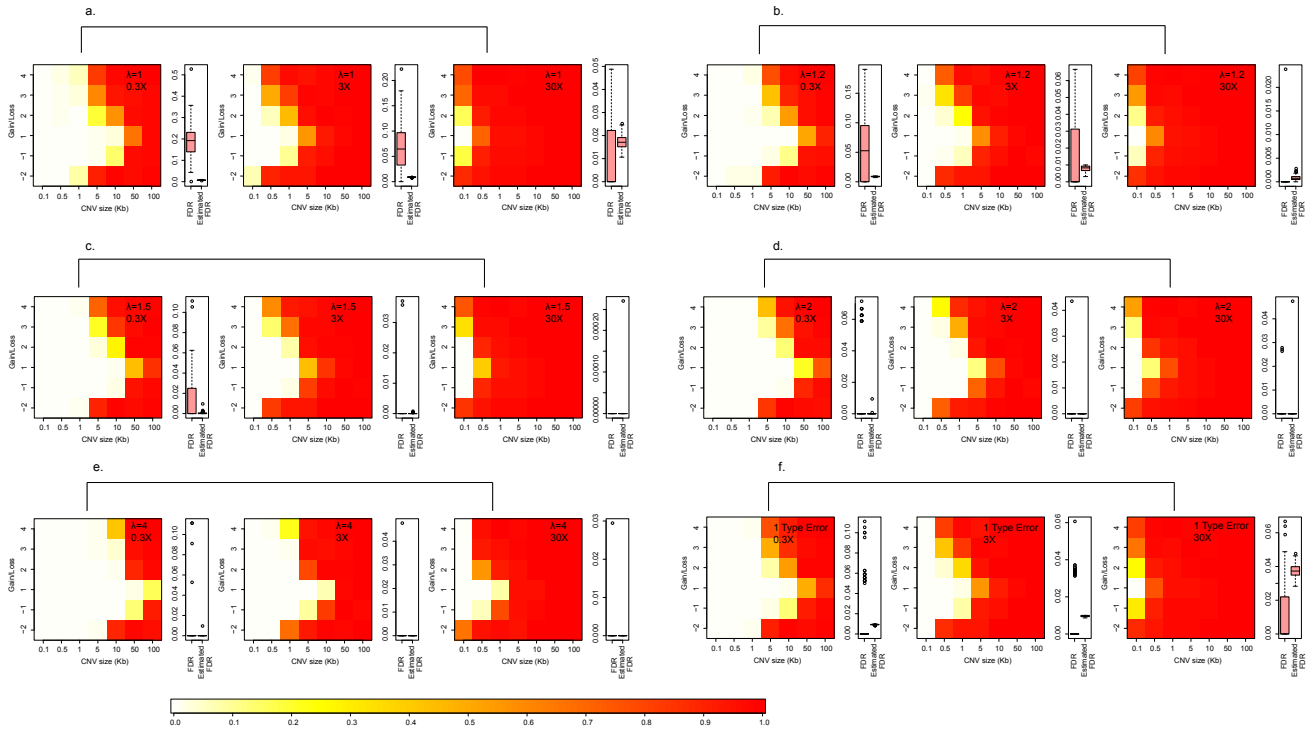


Fig. S4: The CNV detection power of BIC-seq for (**a**) $\lambda = 1$, (**b**) $\lambda = 1.2$, (**c**) $\lambda = 1.5$, (**d**) $\lambda = 2$, (**e**) $\lambda = 4$, (**f**) 1 type I error. The CNVs were chosen as regions with log2 copy ratio greater than 0.2 or less than -0.2.

9 Comparison with SegSeq

We used the "spike-in" simulation data in the last section to compare our algorithm with the algorithm SegSeq(Chiang et al., 2008) (SegSeq 1.0.1). The tuning parameter λ was fixed to be 1. Chiang et al. (2008) showed that 400 was the best choice of SegSeq's parameter w for aberration size 50 Kb, 75 Kb, 100 Kb and 200 Kb, so we used $w = 400$ for these aberration sizes. For aberration sizes 10 Kb and 20 Kb, $w = 400$ was not the best and we used $w = 50$ and $w = 400$. We varied the significant level α, identified the significant regions whose p-values were less than or equal to α, and calculated the corresponding TPR and FDR for SegSeq and BIC-seq.

Figure S5 plots the TPR against the FDR as the significance level α varies. The red solid lines are for BIC-seq and the dashed lines are for SegSeq (blue for $w = 50$ and green for $w = 400$). It is clear that BIC-seq outperforms SegSeq, especially when the aberration size is relatively small, e.g. 50 Kb and 75 Kb. When the aberration size is 200 Kb, BIC-seq and SegSeq perform quite similar.

10 Comparison between sequencing and array platforms

We compared the copy ratios of 291 CNV regions given by BIC-seq with the copy ratios of the same regions given by two array platforms (Affymetrix SNP 6.0 performed at Broad Institute and Agilent 244A performed at Harvard Medical School). For comparison, we used the Level III data

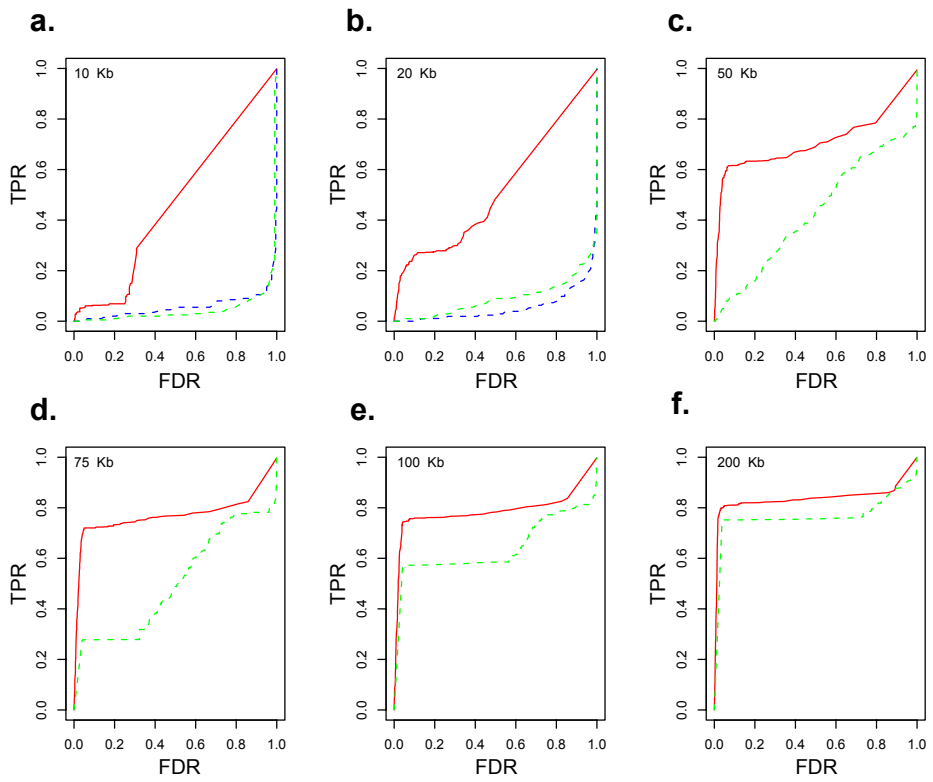


Fig. S5: Comparison of BIC-seq and SegSeq. BIC-seq: red solid lines; SegSeq: blue dashed lines ($w = 50$) and green dashed lines ($w = 400$).

downloaded from the Cancer Genome Atlas (TCGA)(http://tcga-data.nci.nih.gov/tcga/). The Level III data are segmented profiles generated by TCGA consortium. For each platform, there are two profiles, one for the tumor genome and the other for the normal genome. For each CNV detected by BIC-seq, we first identified the segments of the array profiles that overlapped with the CNV. If there was only one such segment, we used the difference between the tumor and normal *seg.mean* as the log2 copy ratio given by the array platform of the CNV region. If there were more than one such segments, we took average of *seg.mean*'s over the overlapped segments, and used the difference between the averages of tumor and normal as the log2 copy ratio of the CNV given by BIC-seq.

For the 291 CNVs discovered by BIC-seq, we regressed the log2 copy ratios given by BIC-seq over the log2 copy ratios given by each of the two array platforms, respectively, using the linear median regression. We found that, in general, the magnitude of log2 copy ratios given by array platforms are smaller than the log2 copy ratios given by BIC-seq (Figure S6).

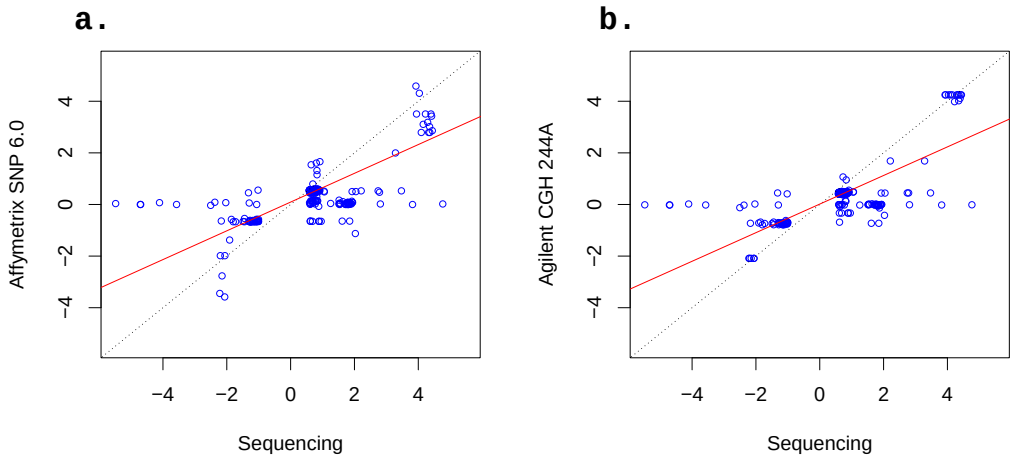


Fig. S6: Comparison of the copy ratio estimates given by BIC-seq and two array platforms, (**a**) Affymetrix SNP 6.0 Array and (**b**) Agilent 244A CGH array. The y-axes are the log2 copy ratio given by the array platforms and the x-axes are the log2 copy ratio given by BIC-seq.

To further study the relationship between the log2 copy ratio given by BIC-seq and array platforms, we classified the CNVs given by BIC-seq into two sets, one set consisting of CNVs less than 15 Kb and the other set consisting of CNVs larger than 15 Kb. Figure S7 shows log2 copy ratios from the array platforms versus those from the sequencing platform. The red lines in the plots are the fitted linear median regression models under each scenario. We clearly see that for the large CNVs (larger than 15 Kb), the log2 copy ratios given by sequencing and array platforms are very similar, especially for the Agilent platform. However, for the small CNVs, the copy ratios given by sequencing and array platforms are quite different. Generally, except for a few outliers, the magnitude of log2 copy ratio of the small CNV given by array platforms is much smaller than that given by sequencing data.

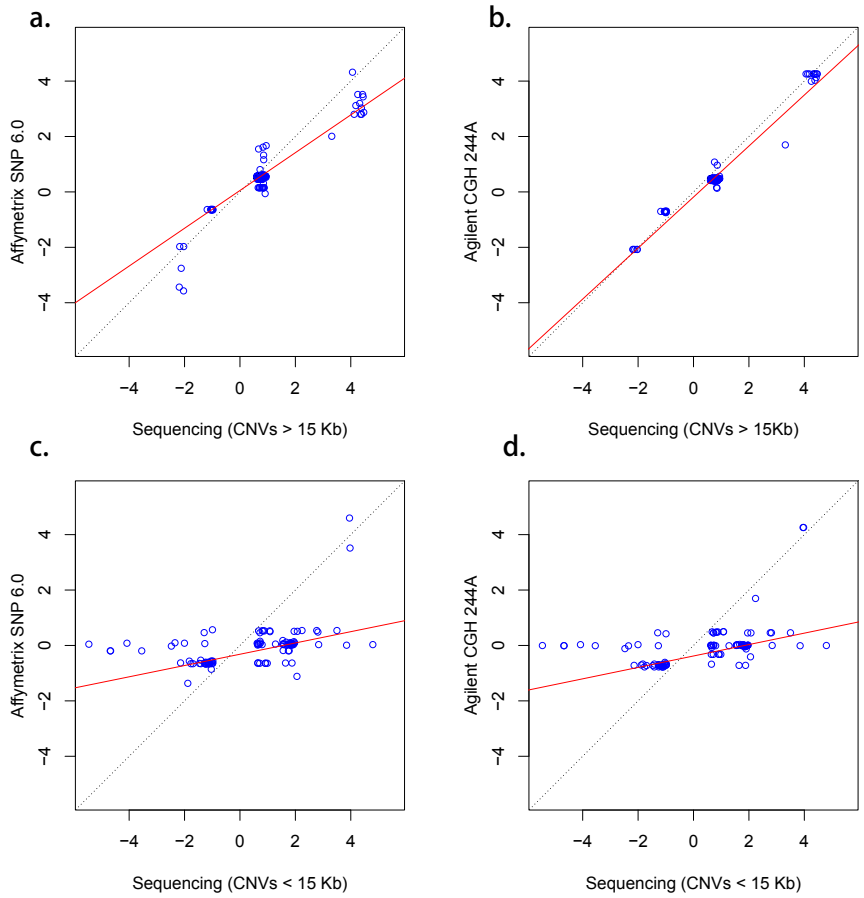


Fig. S7: Comparison of the copy ratio estimates given by BIC-seq and two array platforms; the CNVs given by BIC-seq were classified into two groups, small CNV group (less than 15 Kb) and large CNV group (larger than 15 Kb). The y-axes are the log2 copy ratio given by the array platforms and the x-axes are the log2 copy ratio given by BIC-seq. (**a,c**) Affymetrix SNP Array 6.0. (**a**) CNVs larger than 15 Kb; the slope of the fitted linear model is 0.682 with s.d. 0.03. (**c**) CNVs less than 15 Kb; the slope of the fitted linear model is 0.204 with s.d. 0.053. (**b,d**) Agilent CGH Microarray 244A; (**b**) CNVs larger than 15 Kb; the slope of the fitted linear model is 0.921 with s.d. 0.078. (**d**) CNVs less than 15 Kb; the slope of the fitted linear model is 0.206 with s.d. 0.066.

A large CNV region detected by all three platforms is shown in Figure S8. Here, we only show the tumor profile for Affymetrix and Agilent platforms. This CNV region overlaps with the gene *EGFR*, which was shown to be associated with the development of GBM.

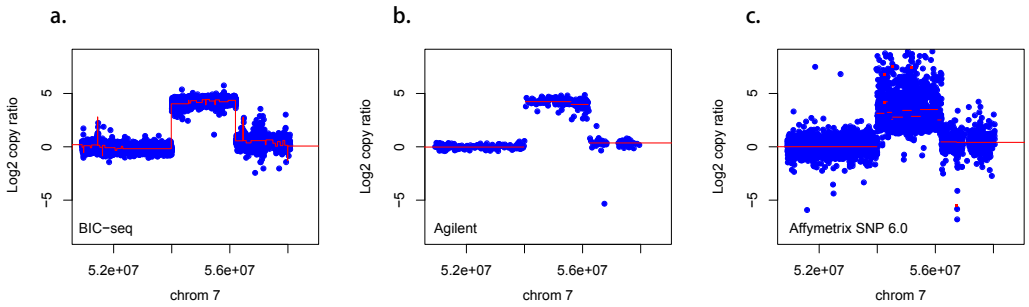
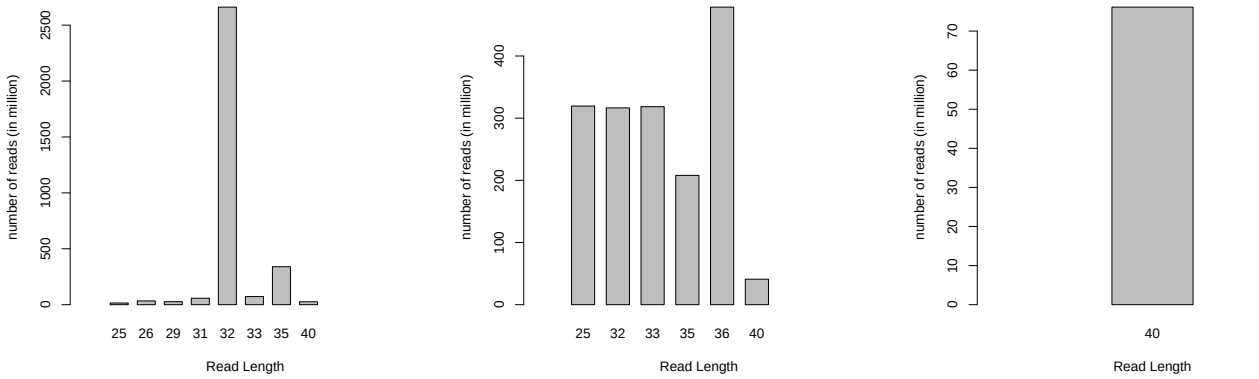


Fig. S8: A large CNV region that was identified by all three platforms.

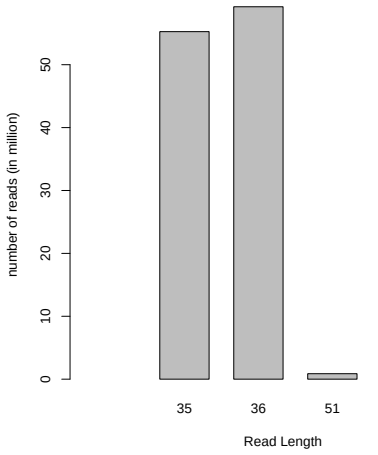
11 Variable read lengths in the AML1 and AML2



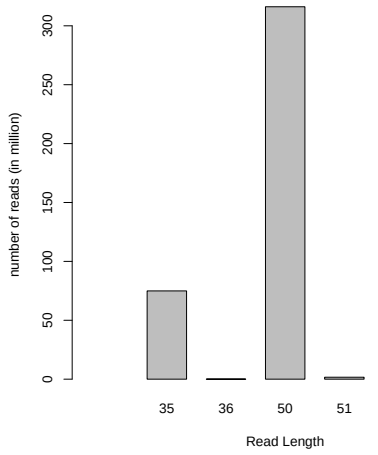
(a) Tumor: single-end (b) Normal: single-end (c) Tumor: paired-end

Fig. S9: Distribution of single-end reads and paired-end reads (AML1)

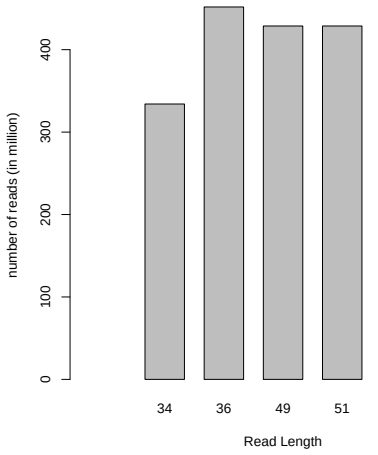
Figure S9 shows the distribution of single-end reads and pair-end reads of AML1. The AML1 normal genome has no paired-end reads. As there are many more single-end reads than the paired-end reads, we only used the single-end reads for CNV detection. Since most of the AML1 tumor single-end reads are 32 bp, we only used the short reads of length larger than or equal to 32 bp. Reads larger than 32 bp were trimmed to 32 bp before alignment. We aligned all obtained reads using Bowtie allowing for 3 mismatches.



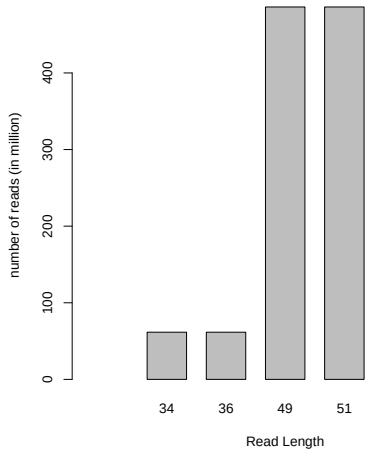
(a) Tumor: single-end



(b) Normal: single-end



(c) Tumor: paired-end



(d) Normal: paired-end

Fig. S10: Distribution of single-end reads and paired-end reads (AML2)

Figure S10 shows the distribution of single-end and paired-end reads of AML2. Simple merging of single-end reads and paired-end reads may lead to false positives and the AML2 tumor genome has many more paired-end reads than single-end reads, we only used the paired-end reads. Furthermore, since most of the AML2 normal paired-end reads are 49 bp and 51 bp, we decided to only use the 49 bp and 51 bp reads. Then, we trimmed all paired-end reads (49 bp and 51 bp) to 49 bp and mapped them back to human hg18 reference genome using Bowtie allowing for 3 mismatches.

Here we show the statistics of the CNVs detected in AML1 and AML2.

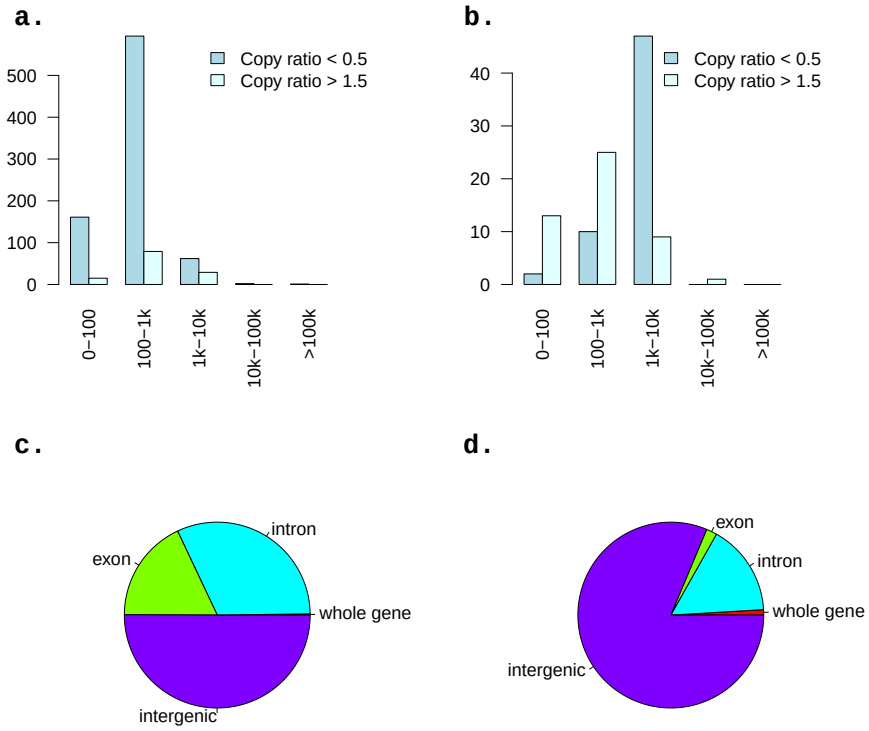


Fig. S11: CNVs detected by BIC-seq in two AML genomes. (**a**,**b**) The distribution of putative CNVs detected in AML1 and AML2 ($\lambda = 4$). Here, the labels on the x-axes demonstrated the size ranges of the CNVs in the corresponding groups. (**c**, **d**) Overlaps of AML1 and AML2 CNVs with Refseq Genes.

12 Comparison with a PEM-based method

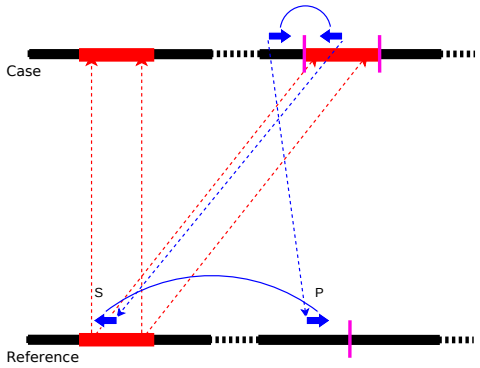


Fig. S12: An insertion that is predicted as a large translocation by BreakDancer

Given an insertion, a simple PEM-based method such as BreakDancer may predict a translocation that is much larger than the actual insertion. Figure S12 shows an example. Here, the red segment on the left of the case genome is duplicated and inserted to the right of the same chromosome. A paired-end read that spans the left breakpoint of the duplicated segment is sequenced from the case genome. The end located within the duplicated segment will be mapped to the original location of the segment (S in the plot), while the other end will be mapped next to the breakpoint (P in the plot). If this happens, BreakDancer will occasionally report the insertion as a large translocation from S to P. In the paper, we used a lenient criterion and counted these predictions as true positive discoveries. However, strictly speaking, these predictions are not correct predictions, since the predicted sizes are in general much larger than the original sizes of the insertion and this makes the subsequent analysis difficult. Figure S13 shows the SV detection power of BreakDancer if we do not count these predictions as true positives. We see that its power for detecting deletions and small insertions is basically the same as Figure 5 in the paper, but the power for detecting large insertion drops significantly.

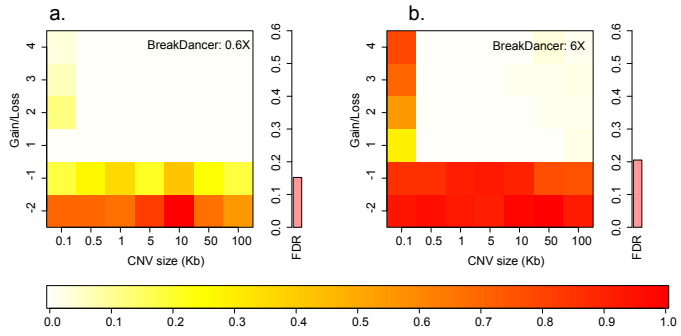


Fig. S13: The SV detection power of BreakDancer. (**a**) 0.6x. (**b**) 6x.

In the main text of this paper, we used single-end data for BIC-seq and paired-end data for

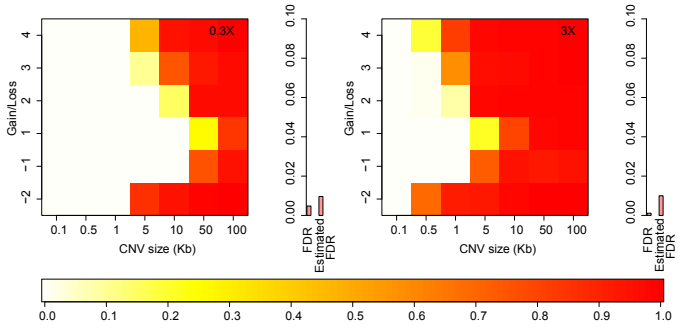


Fig. S14: The CNV detection power of BIC-seq on paired-end data. (**a**) 0.3X. (**b**) 3X.

BreakDancer for comparison of BIC-seq and BreakDancer. Here, we applied BIC-seq to simulated paired-end data to show that BIC-seq is also more powerful than BreakDancer at detecting large insertions using paired-end data. The "tumor" paired-end reads for BIC-seq were randomly sampled 50% from the paired-end reads generated for BreakDancer. The normal paired-end reads were generated using Metasim with insert size 220 bp and s.d. 20 bp. In this simulation, we only consider 0.3X and 3X sequencing coverage (0.3X coverage for both tumor and normal sequencing reads, similarly for 3X coverage). Figure S15 shows the statistical power of BIC-seq at different scenarios. We see that the power of BIC-seq based on paired-end reads drops a slightly compared with that based on single-end reads, but they are similar in general.

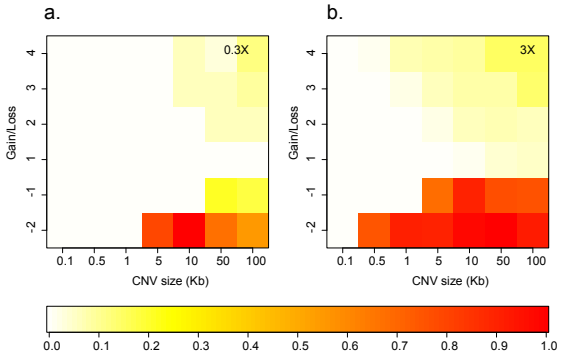


Fig. S15: Common CNVs detected by BIC-seq and BreakDancer. The color bar bar indicates the percentage of the true CNVs detected by both algorithms. Clearly, the deletions are highly overlapping but not amplifications for both 0.3X and 3X coverages.

13 Primers for PCR validation

The primer designs for BIC-seq are listed in Table S1, and the results of the validation are shown in Fig 3 of the main text.

Table S1: Primers Design for BIC-Seq Candidate Copy Number Variations

Copy Number Variation	Forward Primer	Reverse Primer	CNV Size
Chr7: 641189671-64204220	5′-GTGGTGTCATGCAGGTTGTC-3′	5′-AAGCCCTGCTCACTTTCTCA-3′	14,460
Chr1: 554311-560770	5′-CGCTCCCTGAAGAAACACTC-3′	5′-CATGGGATGGAGAGAAAGGA-3′	6460
Chr10: 128040731-128047770	5′-GCCCACAGATTCAAAAGGAA-3′	5′-TTGCTTTCCCCACTTGTACC-3′	7040
Chr10: 131434591-13144100	5′-GGGTATTGTGGCTCACGTCT-3′	5′-TCTACATGGGAGGTGCATCA-3′	6810
Chr10: 31286641-31292530	5′-CTTGGGAGCAGAATCTGAGG-3′	5′-GCAGTGTTCCACATCACCAC-3′	5890
Chr9: 26445131-26450590	5′-CCCATGACATGTGGGGTTAT-3′	5′-AGAAAGCCACCCATGAACAG-3′	5460
Chr5: 134286961-134292120	5′-GCATTCCAACCTGGGAGATA-3′	5′-CTGAAGCGGGTAATTTGCAT-3′	5160
Chr9: 132070481-132074360	5′-ACATGCTCTGCCCTCTGTCT-3′	5′-CACCTGGCCTGTGATAAGGT-3′	3880
Chr4: 66216681-66219490	5′-TGCCATTGGGCTATCACTTT-3′	5′-TCAGATCACCAGCGTTTCAG-3′	2810
Chr13: 108874471-108874700	5′-TGGCCCAATGTACACTTTTT-3′	5′-TTGCCGTAAGGTTAGTAGGAGAA-3′	230
Chr12: 8263201-8263960	5′-GTCTCCCAGGTTCACACCAT-3′	5′-AGGCAGTGGATCACAAGGTC-3′	760
Chr1: 103965011-103965360	5′-GGGAAGCAGAGGTGTTACGA-3′	5′-CTGTCCCCTTGGTGATGAGT-3′	350
Chr7: 151695131-151695480	5′-TGTGTTGGTGCACACCTGTA-3′	5′-TGGAGTGGAGTGGTGCTATT-3′	350
Chr2: 190939261-190939440	5′-CCTCTAGTCCCAGCTGCTCA-3′	5′-CAGCTCACTGCAACTTCCAA-3′	180
Chr7: 39876271-39876430	5′-TTTTCAGCCCAGAATTTAGAAAG-3′	5′-CCAGCAAGCATCAGTCAAGTT-3′	160
Chr12: 40043691-40043800	5′-CTTGGTGTCCTGGGTGTTTT-3′	5′-GGTAAGCTAACCATTTTCTCCAA-3′	110

References

Bickel, P. and Doksum (2001), *Mathematical Statistics: Basic ideas and selected topics*, Pearson.

Chiang, D. Y., Getz, G., Jaffe, D. B., OKelly, M. J. T., Zhao, X., Carter, S. L., Russ, C., Nusbaum, C., Meyerson, M., and Lander, E. S. (2008), "High-resolution mapping of copy-number alterations with massively parallel sequencing," *Nat. Meth.*, 6, 99–103.

Schwarz, G. E. (1978), "Estimating the dimension of a model," *Ann. Stat.*, 6(2), 461–464.