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

Ruibin Xi[a], Angela G. Hadjipanayis[b], Lovelace J. Luquette[a], Tae-Min Kim[a], Eunjung Lee[a,b], Jianhua Zhang[c], Mark D. Johnson[d], Donna M. Muzny[e], David A. Wheeler[e], Richard A. Gibbs[e], Raju Kucherlapati[b,f], and Peter J. Park[a,b,g,1]

[a]Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115; [b]Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115; [c]Belfer Institute for Applied Cancer Science, Dana-Farber Cancer Institute, Boston, MA 02115; [d]Department of Neurological Surgery, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115; [e]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030; [f]Department of Genetics, Harvard Medical School, Boston, MA 02115; and [g]Children's Hospital Informatics Program, Boston, MA 02115

Edited by Robert Tibshirani, Stanford University, Stanford, CA, and accepted by the Editorial Board September 30, 2011 (received for review July 1, 2011)

DNA copy number variations (CNVs) play an important role in the pathogenesis and progression of cancer and confer susceptibility to a variety of human disorders. Array comparative genomic hybridization has been used widely to identify CNVs genome wide, but the next-generation sequencing technology provides an opportunity to characterize CNVs genome wide with unprecedented resolution. In this study, we developed an algorithm to detect CNVs from whole-genome sequencing data and applied it to a newly sequenced glioblastoma genome with a matched control. This read-depth algorithm, called BIC-seq, can accurately and efficiently identify CNVs via minimizing the Bayesian information criterion. Using BIC-seq, we identified hundreds of CNVs as small as 40 bp in the cancer genome sequenced at 10× coverage, whereas we could only detect large CNVs (>15 kb) in the array comparative genomic hybridization profiles for the same genome. Eighty percent (14/16) of the small variants tested (110 bp to 14 kb) were experimentally validated by quantitative PCR, demonstrating high sensitivity and true positive rate of the algorithm. We also extended the algorithm to detect recurrent CNVs in multiple samples as well as deriving error bars for breakpoints using a Gibbs sampling approach. We propose this statistical approach as a principled yet practical and efficient method to estimate CNVs in whole-genome sequencing data.

structural variation | genomic alterations | model selection | semiparametric model

Copy number variations (CNVs), which are gains or deletions of genomic segments, account for a substantial proportion of human genetic variations. CNVs have been shown to be associated with a wide spectrum of human disorders such as autoimmune diseases (1), autism (2), schizophrenia (3, 4), and obesity (5, 6). CNVs can also occur in the form of somatic alterations. For example, cancer genomes often acquire dosage or structural alteration of cancer-related genes, some of which may confer tumor cells a selective growth advantage over normal cells.

Microarray-based platforms have been widely used in genome-wide studies to identify CNVs in cancer genomes (somatic CNVs) (7–10) as well as in the genomes of normal population (germline CNVs) (11–16). These efforts have led to the discovery of previously unrecognized oncogenes and tumor suppressors. However, due to the limited resolution of microarrays, this approach has suffered from poor sensitivity in detecting small CNVs. More recently, the advent of next-generation sequencing and the rapid increase in its throughput have led to the possibility that CNVs can be characterized in much finer detail via whole-genome sequencing. With mate-pair reads, sequencing-based approaches can also be used for the discovery of structure variations (SV) such as insertions, deletions, inversions, and translocations (17–19). Current mate-pair-based algorithms for SV detection cannot estimate the magnitude of copy number change accurately, tend to perform poorly in genomic regions rich in tandem or inverted duplications (20), and have limited power in detecting insertions larger than the insert size (the distance between the two ends of a paired-end read) (21).

Although the density of aligned reads along the genome generally corresponds to the DNA copy number, it is also affected by many sources of bias. These include the higher likelihood of sequencing GC-rich fragments, the uneven representation of genomic regions in library preparation due to variability in DNA fragmentation, and the regional variation in the fraction of short reads that can be aligned to a given position. Thus, somatic CNVs in the tumor genome can be best identified by comparing the read distribution in the tumor genome with that of the matched normal genome. The genomic regions with disproportionate read counts indicate potential CNVs—e.g., the prevalence of tumor reads over normal ones suggests a somatic copy gain in the tumor genome. This approach was adopted by SegSeq (22) and CNV-seq (23). However, both of these methods perform statistical tests based on a Poisson model, in which reads are assumed to be distributed uniformly across the genome. This assumption hardly holds even for normal genomes (*SI Appendix*, Fig. S1) due to the biases mentioned above. In addition, the results are sensitive to the window size that must be specified a priori in these methods, and approaches with initial local detection of breakpoints followed by copy number estimation tend to perform poorly compared to more global approaches (24).

In this paper, we describe a statistically rigorous and computationally efficient algorithm called BIC-seq for detecting CNVs from whole-genome sequencing data. This algorithm does not assume a Poisson or other parametric models on the read distribution as is done in currently available methods, and it is thus more robust to outliers and datasets that cannot be well approximated with a parametric model. It is also fast and able to handle high-coverage genomes effectively. Furthermore, the statistical framework behind BIC-seq can be extended to the problem of identifying recurrent CNVs in multiple cancer genomes. To test

Author contributions: R.X. and P.J.P. designed research; R.X., L.J.L., J.Z., D.M.M., D.A.W., and R.A.G. performed research; R.X., A.G.H., and E.L. analyzed data; and R.X., A.G.H., T.-M.K., M.D.J., R.K., and P.J.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. R.T. is a guest editor invited by the Editorial Board.

Data deposition: The sequences reported in this paper have been deposited in the Genotypes and Phenotypes (dbGaP) database, http://www.ncbi.nlm.nih.gov/gap (accession nos. phs000178.v5.p5 and phs000159.v2.p1).

[1]To whom correspondence should be addressed. E-mail: peter_park@harvard.edu.

See Author Summary on page 18583.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1110574108/-/DCSupplemental.

TEMPUS EX1086

the performance of the algorithm, we applied BIC-seq on the newly sequenced cancer genome of a patient with glioblastoma multiforme (GBM) and its matched control. Large CNVs (e.g., >15 kb) identified by BIC-seq were largely concordant with those identified from the two microarray platforms on which the same genomes were profiled. However, many small CNVs identified by BIC-seq were not detected on the arrays. Subsequent validation on a subset of small CNVs by quantitative PCR (qPCR) showed >80% true positive rates and confirmed the accuracy of the magnitude of copy number ratios.

Results

CNV Detection Algorithm Based on the Bayesian Information Criterion. Current high-throughput sequencing platforms generate short (36–100 bp) sequenced tags from one or both ends of hundreds of millions of DNA fragments. These short reads can then be aligned to the reference genome using alignment tools such as MAQ (Mapping and Assembly with Quality) (25), Bowtie (26), and BWA (Burrows-Wheeler Aligner) (27). The basic idea behind CNV detection is that larger or smaller than the expected number of tags in a genomic region corresponds to gain or loss of DNA, respectively. Although it is tempting to design an algorithm to detect such a variation in a single sample, this would result in a large number of false positives because the tag distribution along the genome fluctuates considerably due to biases in sequencing and genome content as described above. Thus, the "read-depth" methods including ours use the disease-vs.-matched control approach.

In the BIC-seq algorithm, the short reads are aligned onto the reference genome and the uniquely aligned reads are sorted according to their genomic coordinates. Single genomic positions having many orders of magnitude more mapped reads than their neighboring positions (*SI Appendix*, Fig. S2) are filtered out because they are likely to be a result of amplification bias (*Materials and Methods*). Then, tumor and normal reads are mixed and binned (1-bp bins are allowed). The idea of BIC-seq is to iteratively identify and merge the most similar pair of bins (Fig. 1). Here, we choose to use the Bayesian Information Criterion (28) (BIC) as the merging and stopping criterion. The BIC, which is widely used as a model selection criterion in statistics, consists of two terms: One term is the negative log likelihood, which measures how well the model fits the data; and the second term is the penalty for the model complexity, which prevents model overfitting. The model complexity is measured by the product of the number of variables in the model and the logarithm of the number of observations (*Materials and Methods* and *SI Appendix*). For CNV detection, the number of variables is equal to the number of breakpoints plus one. Generally, the model with a smaller BIC is preferred. To accommodate more flexibility to BIC-seq, we introduce a parameter λ in the penalty term of the BIC. This parameter allows the user to control the smoothness (the number of breakpoints) of the final profile generated by BIC-seq and reflects the user's prior belief about the CNV profile. The larger this parameter is, the smoother the final profile will be.

Given an initial configuration of bins, BIC-seq attempts to reduce the overall BIC by merging the appropriate neighboring bins. Briefly, for each pair of neighboring bins, BIC-seq calculates the BIC difference as if the pair were merged. Then, BIC-seq identifies the pair of bins with the smallest BIC difference and merges the bin pair if this BIC difference is less than zero. The above process is repeated until the overall BIC cannot be further reduced by merging pairs of neighboring bins. After merging bin pairs, BIC-seq also attempts to merge three or more neighboring bins if some of them can be merged (*SI Appendix*). By using a red-black tree (29) to efficiently store BIC differences, it is computationally efficient (*SI Appendix*). For instance, if the initial bins are chosen as equally spaced bins of 100 bp, it only takes BIC-seq a few seconds to finish the merging process for a chromosome on a single-processor computer. Based on the segments obtained from the bin merging process, the copy ratios between tumor and normal genome of each segment can then be easily calculated (*SI Appendix*).

BIC-seq can be easily extended to the multisample case by modifying the log likelihood term and the penalty term in the

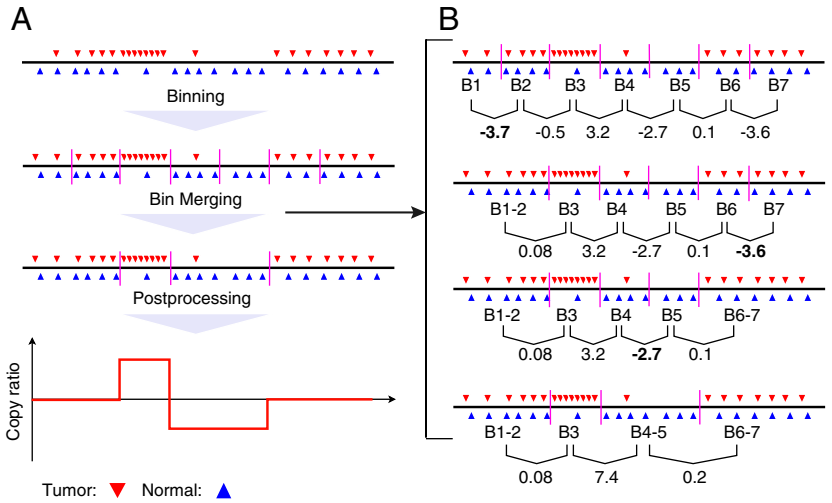


Fig. 1. A schema for BIC-seq. (*A*) The dataflow of BIC-seq. First, the short reads are aligned to the reference genome and the outliers are removed. Then, the short reads are binned into small bins (e.g., 10 bp bins), and the initial bins are iteratively merged using the BIC. The vertical purple bars in the plot are the boundaries between neighboring bins. Lastly, copy ratios are calculated based on the segmentation given by BIC-seq. (*B*) A bin-merging procedure based on the BIC. We demonstrate the procedure for $\lambda = 1$. Given a list of initial bins, BIC-seq first calculates the BIC differences between the current configuration and all possible configurations in which two adjacent bins are merged. In the plot, the numbers under the bin pairs are their corresponding BIC differences. Then, BIC-seq identifies the pair with the smallest BIC difference. If this BIC difference is less than zero, the corresponding bin pair will be merged; otherwise, BIC-seq will stop merging bin pairs. In this example, the bin pair B1 and B2 have the smallest BIC difference (–3.7) and they are merged, giving a new bin B1-2. BIC-seq then updates the BIC differences for the bin pairs. As shown in the plot, we only need to update the BIC difference for the bin pair B1-2 and B3, because all other BIC differences remain the same as before the merging of B1 and B2. This fact holds in general and we used it to expedite BIC-seq (*SI Appendix*). The above process is then repeated until the BIC cannot be improved further—i.e., until no BIC difference is less than zero. After the merging of bin pairs, BIC-seq also tries to merge three or more neighboring bins if their merging can improve the BIC (*SI Appendix*). For this example, the merging of three or more neighboring bins cannot improve the BIC.

definition of the BIC (*Materials and Methods* and *SI Appendix*). In addition to giving point estimates of the breakpoints, it is also desirable to assign confidence to the point estimate. Here, we developed a Bayesian model and a corresponding Gibbs sampler (30, 31) to assign credible intervals to the breakpoints given by BIC-seq (*Materials and Methods* and *SI Appendix*). This error estimation for CNV boundaries is helpful in selecting candidate genes and other genomic elements for further examination.

After the segmentation of the genome, one can apply a copy ratio and/or a *p*-value threshold to get a set of CNV candidates. To give a false discovery rate (FDR) estimate of this CNV call set, we pool the tumor and normal reads together, randomly sample with replacement from the pooled data as tumor/normal reads, and run BIC-seq on the resulting pseudotumor/normal data (*SI Appendix*). CNVs are called with the same criterion as used in the original dataset. Because the randomly sampled data are generated under the null hypothesis of no CNV in the genome, any CNV call from the pseudotumor/normal data would be a false positive. The ratio between the numbers of CNV calls from the resampled data and the original data is then an FDR estimate. A more stable FDR estimate can be obtained by repeating this resampling procedure many times and taking the mean as the final FDR estimate.

BIC-seq has a single parameter λ that the user can specify, based on the desired level of confidence in the CNV calls and the genome coverage (also see section *Sequencing Coverage and Statistical Power*). For all levels of coverage, a copy ratio threshold should first be used (e.g., log2 ratio >0.2 or < -0.2) to remove low-amplitude changes that are likely to a result of random fluctuations. Then, for low-coverage data (<1×), λ should be small (e.g., 1 or 1.2). To further reduce FDR, one can apply a *p*-value threshold to remove the less significant, small CNVs. For medium coverage (e.g., 2–5×), a larger λ (e.g., 2) should work well without the additional *p*-value filtering. For high-coverage (e.g., 10–30×), $\lambda = 4$ should give very confident calls while still detecting many small CNVs (e.g., 100–1,000 bp). *SI Appendix*, Fig. S4 confirms that $\lambda = 4$ has good sensitivity at detecting CNVs as small as 500 bp. We evaluated the effect of λ on the performance of BIC-seq in a simulation study, which showed that both true positive rate and FDR increase as λ gets smaller (*SI Appendix*, Fig. S3). Simulation studies also confirm that BIC-seq is more sensitive than SegSeq for a given FDR level (*SI Appendix*, Fig. S5).

We also note that the merging process of BIC-seq is equivalent to performing a series of likelihood ratio tests. Thus, the choice of the tuning parameter λ is equivalent to setting a type I error rate (a type I error in a merging step is the event that two neighboring bins should be merged but fail to merge). The likelihood ratio test statistic asymptotically follows a χ^2 distribution, which allows us to provide a more intuitive way of specifying the tuning parameter λ—i.e., the user can specify the expected number of type I errors in the merging process (*SI Appendix*).

Copy Number Profiling of Human Solid Tumor (GBM) by BIC-seq. We applied BIC-seq to profiles from a cancer patient, sequenced as part of the Cancer Genome Atlas (TCGA) project (*Materials and Methods*). Tumor and matched control (blood) DNA were obtained from the same individual, a 53-y-old woman who was diagnosed with primary GBM, and sequenced on the Applied Biosystems SOLiD (Sequencing by Oligonucleotide Ligation and Detection) platform. We obtained 833 million (10×) and 603 million (7×) uniquely aligned 35-bp single-end reads for the tumor and its matched normal genome, respectively. To be conservative, we set the tuning parameter $\lambda = 4$; a smaller value results in identification of smaller feature sizes at the expense of lower confidence level. The initial bins were chosen as 10-bp bins, which was a reasonable compromise between resolution and memory usage for this dataset. After filtering for segments with copy ratios greater (tumor/matched normal) than 1.5 or less than 0.5, we found 306 putative somatic CNVs, ranging from 10 bp to 5.7 Mb with a median size of 59 kb. Because deletions or insertions that are about the length of the sequenced read cannot be distinguished from each other by a read-depth method, we focused on CNVs larger than the read length. After removing 15 such CNVs, we obtained 291 putative CNVs (Fig. 2), which covered 88.51 Mb (2.95%) of the human genome. Among the 291 somatic CNVs, 192 (73.56%) showed at least a partial overlap with 1,926 reference sequence (Refseq) genes (including intronic sequences) and 170 (58.42%) with coding sequences (Fig. 2). We further compared these CNVs with cancer genes listed in Futreal et al. (32) and found that 19 out of 291 somatic CNVs showed overlap with 22 out of 288 cancer genes. These 22 genes include the well-known cancer-related genes *EGFR* and *CDKN2A*, whose amplification and deletions have been frequently observed in GBM (33, 34).

We compared these CNVs with those detected by two microarray platforms, Agilent 244K comparative genomic hybridization (CGH) microarrays and Affymetrix SNP 6.0 arrays (see *Materials and Methods*). When we plotted the log2 ratios of the 291 putative CNVs as estimated by array platforms and sequencing (*SI Appendix*, Fig. S6), we found that a large fraction of the BIC-seq ratios were larger than the array ratios (which cluster around zero), indicating that these CNVs detected by BIC-seq were missed by arrays. The slopes of the linear median regression lines are 0.557 and 0.554 with SDs of 0.110 and 0.101. To determine whether CNV size is related to the difference in the platform performance, we classified the 291 CNVs into two sets according to their sizes, one set consisting of 184 CNVs larger than 15 kb and the other consisting of 107 CNVs less than 15 kb. We found that, for the large CNVs, the log2 copy ratios given by BIC-seq and two microarray-based platforms are close, especially between the log2 copy ratios given by BIC-seq and the Agilent array (*SI Appendix*, Fig. S7 *A* and *B* and Fig. S8). But for the small CNVs, the log2 ratios were dramatically different between the platforms. (*SI Appendix*, Fig. S7 *C* and *D*).

To test the accuracy of the copy ratio estimates given by BIC-seq for small CNVs (less than 15 kb), we selected 16 CNVs for which good primers could be designed, ranging from 110 bp to 14 kb, from the 107 small CNVs for qPCR validation. The experi-

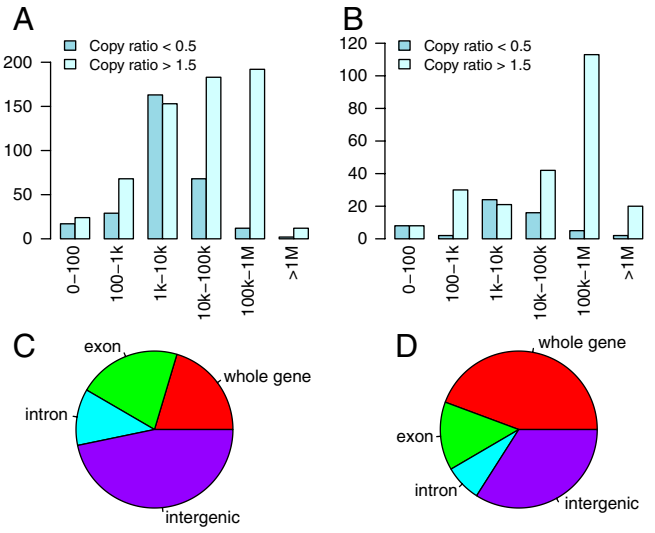


Fig. 2. CNVs detected by BIC-seq in the GBM genome. (*A* and *B*) The distribution of putative CNVs detected in GBM with tuning parameter $\lambda = 2$ and 4, respectively. Here, the *x* axis notes the CNV sizes. (*C* and *D*) Overlaps of GBM CNVs with Refseq genes for $\lambda = 2$ and $\lambda = 4$. Intergenic, no overlap with any gene; whole gene, covering an entire gene; exon, overlapping with at least one exon but not covering an entire gene; Intron, overlapping with introns but not exons.

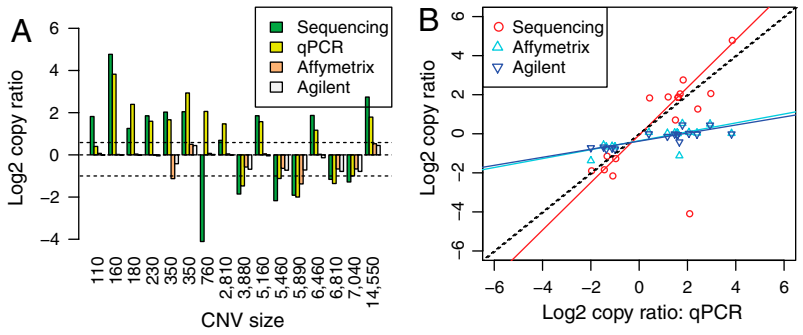


Fig. 3. Experimental validation of 16 BIC-seq CNVs. (*A*) Bar plot of the log2 copy ratios estimated from the four platforms. The two dashed lines correspond to copy ratios 1.5 and 0.5, respectively. (*B*) Scatter plot of the log2 copy ratios given by the sequencing data and two array-based platforms versus the log2 copy ratios given by qPCR. The red, cyan, and blue solid lines are the fitted linear median regression models using the log2 copy ratios given by qPCR as a predictor and the log2 copy ratios given by sequencing data and two array-based platforms as responses, respectively. The slopes of the linear model for sequencing, Affymetrix, and Agilent are 1.21 (SD 0.20), 0.23 (SD 0.06), and 0.21 (SD 0.05), showing that copy ratios given by sequencing data are more accurate than that given by array platforms.

mental validation confirmed 14 out of 16 CNV calls (true positive 87.5%; Fig. 3), assuming that a copy gain or loss is a true positive if its copy ratio estimate given by qPCR is greater than 1.5 or less than 0.5, respectively. Fig. 4 shows two examples of validated CNVs discovered by BIC-seq but missed by array platforms. To further investigate the accuracy of copy ratio estimates, we fitted three linear median regression models, using the log2 copy ratios given by qPCR as the predictor and the log2 copy ratios given by BIC-seq and the two array platforms as the responses. For BIC-seq ratios, the estimated slope of the linear model is not significantly different from one (1.21 with SD 0.20) and the estimated intercept is not significantly different from zero (−0.062

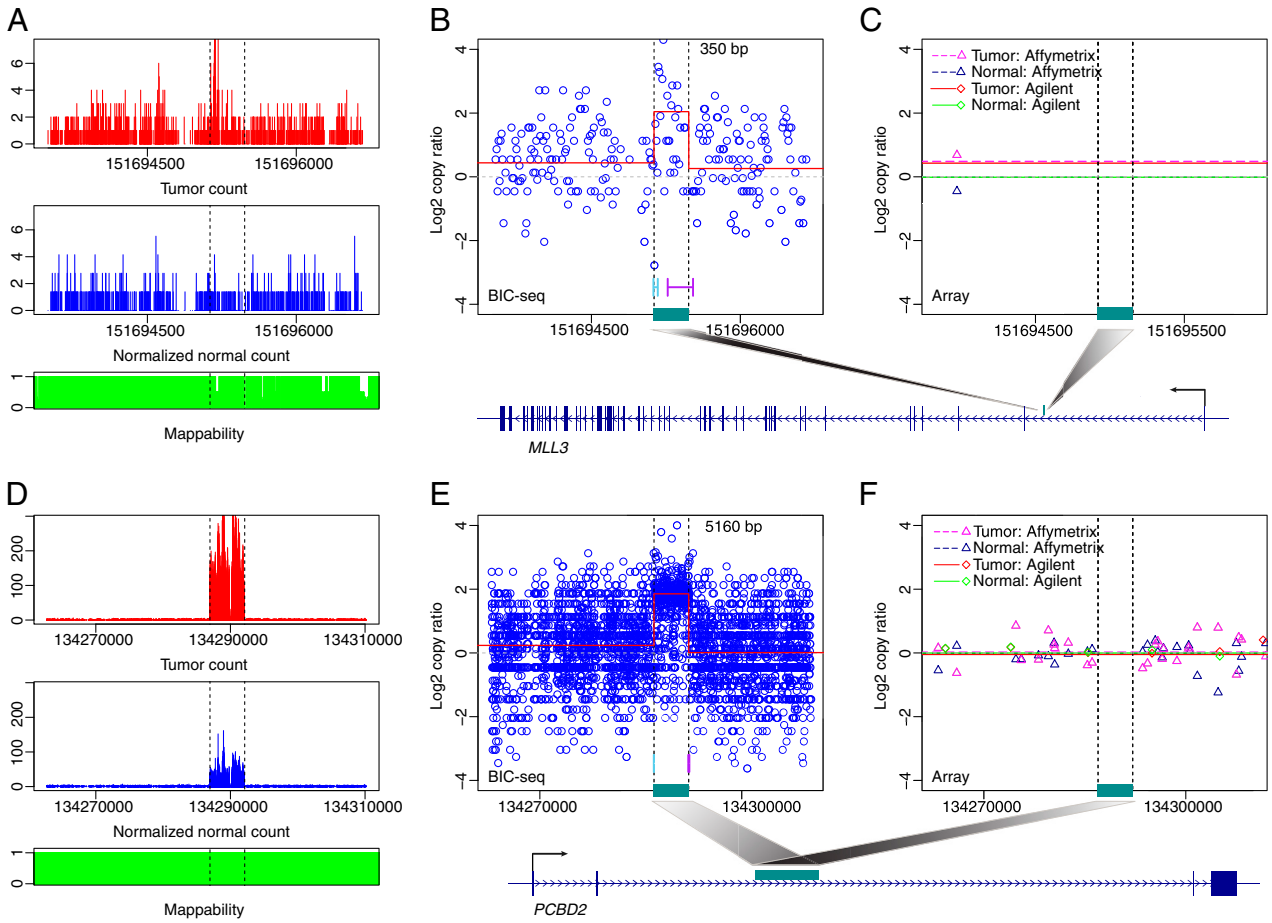


Fig. 4. Two qPCR validated CNVs that were missed by the array-based platforms. (*A*–*C*) A 350-bp focal CNV. (*A*) The distribution of tumor (*Top*) and normal (*Bottom*) reads near the identified CNV at nucleotide resolution. The normal reads are rescaled such that the summation of the transformed values is the same as the total tumor read count. (*B*) The local profile given by BIC-seq (red line). The circles are the copy ratios calculated based on 10-bp bins. The regions marked by cyan and purple lines are the 95% credible intervals for the left and right breakpoints of the CNV. The CNV overlaps with the intronic region of the gene *MLL3* and the position of the CNV in the gene is marked by the dark cyan bar. (*C*) The profiles given by Affimetrix and Agilent platforms. (*D*–*F*) Another validated CNV missed by the array platforms. The CNV overlapped with the gene *PCBD2*. Both tumor and normal genomes are enriched in this CNV region, but the enrichment magnitude of the tumor genome is even greater than its matched normal genome.

with SD 0.33). However, for Affymetrix ratios, the slope and the intercept of the model are 0.23 (SD 0.06) and −0.36 (SD 0.09); for Agilent ratios, the slope and the intercept are 0.21 (SD 0.05) and −0.38 (SD 0.08). This regression analysis suggests that the copy ratios given by BIC-seq are more concordant with qPCR-based estimates than those given by array platforms.

We observed that several of the validated small CNVs overlapped with cancer-related genes, including *INPP1*, *ADAM12*, *MGMT*, *MLL3*, *PCBD2*, and *AMY2A*, whose copy number alteration may influence cancer development and progression. For example, *INPP1* encodes one of the enzymes involved in the phosphatidylinositol (PI) signaling pathway, which is frequently altered in glioblastoma (33). Thus, in addition to well-known events such as *PIK3CA* mutations, genomic alterations involving the *INPP1* locus may contribute to glioblastoma pathogenesis. Overexpression of *ADAM12* has been reported in a variety of human cancers and the expression level of *ADAM12* mRNA showed a correlation with glioblastoma cell proliferation (35). *MGMT* promoter methylation occurs frequently in glioblastoma and is associated with a better treatment response and a more favorable prognosis, but it is generally believed that *MGMT* is not commonly mutated or deleted (36). However, we observed a focal deletion involving the *MGMT* locus, which suggests that a focal deletion involving this locus may act as an additional inactivating mechanism in glioblastoma. This phenomenon may contribute to the occasional discordance between *MGMT* promoter status and treatment response. Indeed, a recent report suggests that treatment response correlates more closely with *MGMT* mRNA expression level than with *MGMT* promoter methylation (37). The gene *MLL3*, which encodes components of a histone H3 lysine 4 methyltransferase complex (38), is frequently mutated in various cancer types suggesting that *MLL3* may play a role in gliomagenesis (39–41). Homozygous deletion involving *AMY2A* was observed in approximately 20% of primary gastric cancers (42) and the amplification of *PCBD2* was previously observed in mesothelioma (43).

We observed that, in some CNV regions, there is read enrichment in the tumor genome compared to the normal genome, yet both tumor and normal genomes have several magnitude more reads than in their neighboring regions (Fig. 4 *D–F*). This read distribution might be due to a complex genomic configuration (e.g., tumor-specific somatic copy gain occurring within patient-specific germ-line CNVs). This phenomenon is also observed in the genomes of acute myeloid leukemia (AML) patients described below and deserves further investigation.

Copy Number Profiling of Human Hematologic Tumor (AML) by BIC-seq. We further applied BIC-seq on two AML genomes (44, 45). For the first AML genome (44) (AML1), we obtained 1.9 (20×) and 0.8 (9×) billion uniquely aligned 32 bp single-end reads from the tumor and its matched normal genomes, respectively (*SI Appendix*, Fig. S9). Most of the second AML genome (45) (AML2) were paired-end reads (*SI Appendix*). We aligned the two ends of paired-end reads independently and obtained 588 (10×) and 655 (11×) million uniquely aligned 49-bp reads from the tumor and its matched normal genome (*SI Appendix*, Fig. S10). Similar to the analysis of the GBM genome, the initial bins were chosen as 10-bp bins and the tuning parameter λ was set as 4. BIC-seq identified 985 and 143 putative CNVs for AML1 and AML2, respectively. After removing the putative CNVs smaller than the read length, AML1 had 943 putative CNVs with median size 260 bp (ranging from 40 bp to 127 kb), and covered 0.6 Mb (0.02%) of the human genome; AML2 had 107 putative CNVs with median size 1,160 bp (ranging from 50 bp to 19 kb) and covered 0.2 Mb (0.006%) of the human genome. Four hundred seventy-one out of 943 AML1 CNVs overlapped with 479 Refseq genes (*SI Appendix*, Fig. S11) including nine genes in the cancer gene list (32). For example, a 1.1-kb CNV (copy ratio 0.5) overlapped with the coding region of *RUNX1*, which was reported to be involved in AML (46, 47). For AML2 CNVs, 20 out of 107 CNVs overlapped with 19 Refseq genes (*SI Appendix*, Fig. S11) and 2 of 19 genes are on a previously published list of cancer-relevant genes (32).

Between AML1 and AML2, there was one common CNV, which overlapped *ANKRD30BL*. If CNVs smaller than the read length are included, there were three more, involving *PCBD2* and *MAN1A1*. Interestingly, *PCBD2*, which was previously reported to be amplified in a mesothelioma tumor genome (43), is also detected in the GBM genome (Fig. 4 *D–F*). BIC-seq predicts that all three tumor genomes are amplified at this locus, and the CNV sizes are 5,160, 4,520, and 40 bp for GBM, AML1, and AML2, respectively.

Sequencing Coverage and Statistical Power. The statistical power for detecting CNVs is dependent on the sequencing depth. Generally, the greater the number of reads sequenced, the higher the statistical power. However, it is important to characterize the performance of the algorithm as a function of sequencing depth because this is a key consideration in designing a study. The cost of sequencing is roughly proportional to the coverage and one must determine, e.g., whether it is more beneficial to sequence each individual at 20× or to sequence twice the number of individuals at 10×. In addition to the robustness of the algorithm, intrinsic properties of the CNV, such as its size, magnitude of change in copy number, and alignability of the genomic location, can also have substantial influence on the statistical power of CNV detection.

Here, we use simulation to depict the relationship between these factors and the statistical power. We first simulated 100 "tumor" chromosomes based on the human reference chromosome 22 (see *Materials and Methods*). Each of the simulated tumor chromosomes contained 42 CNVs (7 CNV sizes × 6 copy ratios) ranging from 100 bp to 100 kb with varying magnitudes of change. Then, short reads were randomly placed on the 100 template tumor chromosomes and the human reference chromosome 22 using MetaSim (48), and these short reads were aligned back to the reference genome using Bowtie (25). Finally, we applied BIC-seq to evaluate the statistical power for CNV detection under many different scenarios (Fig. 5 *A–C*). The CNV regions were chosen as regions with log2 copy ratios less than −0.2 or greater than 0.2 and further filtered using a *p*-value cutoff such that the estimated FDR is less than 0.01.

Overall, copy loss is much easier to detect than copy gain. For example, the power to detect a two-copy loss CNV is roughly the same as that for detecting four-copy gain CNVs. One-copy gain CNVs have the lowest statistical power, sometimes dramatically; e.g., at 3× coverage (Fig. 5*B*), the power for a 500-bp CNV of two-copy loss is around 90%, but the power for a one-copy gain CNV of the same size is less than 1%.

For comparison, we also tested the performance of a well-known algorithm BreakDancer (19), which is based on paired-ends mapping (PEM). As above, we first simulated 100 tumor chromosomes, each of which contained 42 CNVs with predefined sizes and copy numbers (*Materials and Methods*). Then, we applied MetaSim to generate paired-end reads. The insert size and the SD of the insert size were set as 220 and 20 bp, respectively, as estimated from the AML2 data. Then, we applied BreakDancer on the simulated data to call CNVs. Because BIC-seq needs sequencing data from both tumor and normal genomes to call the somatic CNVs, but BreakDancer only needs sequencing data from tumor genomes to call the SVs, we required the tumor reads for BreakDancer to be at least twice as many as that for BIC-seq to make the comparison fair (*Materials and Methods*). For BreakDancer, we used a lenient criterion to determine whether an SV call is a positive discovery (*Materials and Methods*). As shown in Fig. 5 *D* and *E*, BreakDancer performs well

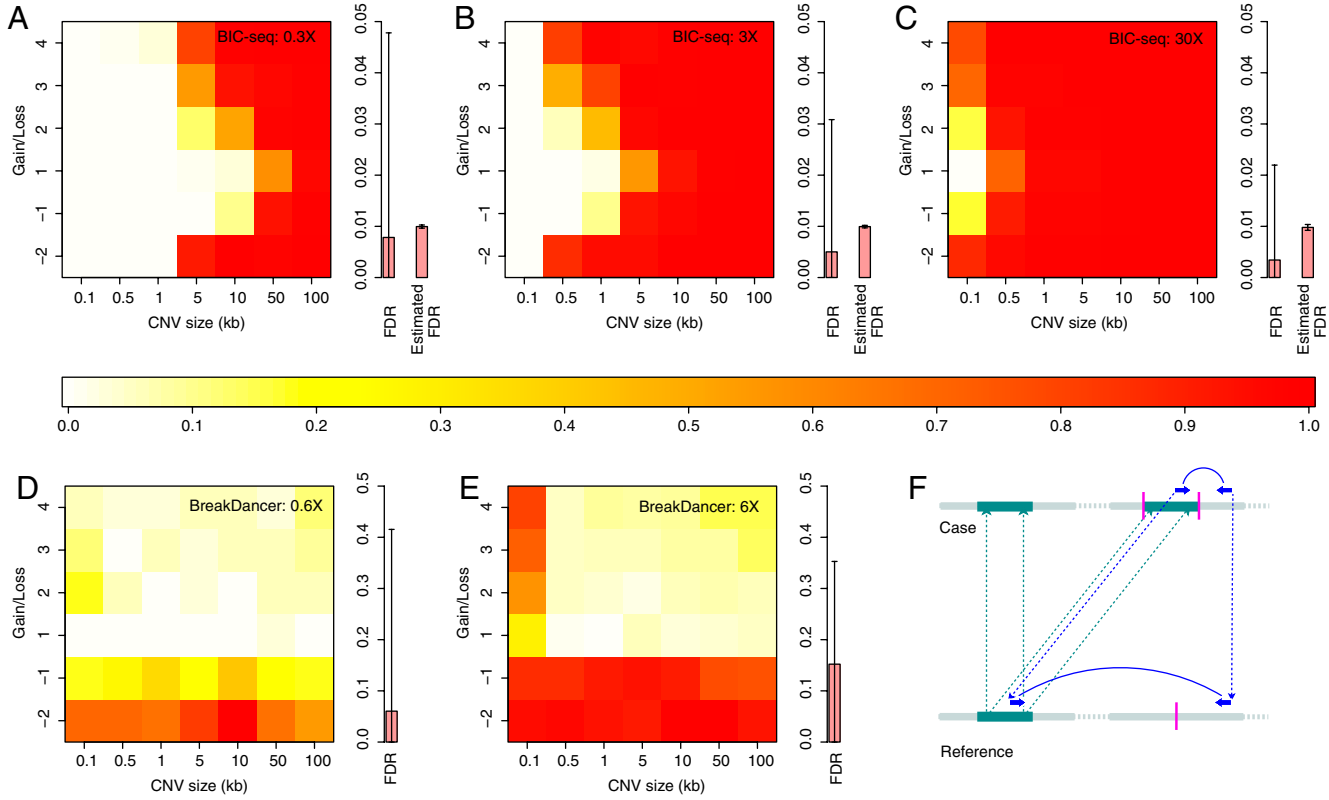


Fig. 5. Statistical power for CNV detection. The colors represent the power of CNV detection under different scenarios, with red regions indicating the greatest power. The *y* axis denotes the number of copies gained or lost. (*A*–*C*) The CNV detection power of BIC-seq at 0.3×, 3×, and 30× coverage, respectively. The bar plot on the right shows the mean FDR of the 100 simulations and the corresponding mean estimated FDR. Here, 0.3× coverage means that both tumor and normal sequencing are 0.3× and similarly for the others. The bars in the box plot are the error bars of the true and the estimated FDRs (2 SD around the mean with lower bound of zero). (*D* and *E*) The SV detection power of BreakDancer at 0.6× and 6× coverage. Because BreakDancer only uses sequencing data from the tumor genome, we doubled the sequencing reads to make the comparison fair (*Materials and Methods*). (*F*) The signature of a PEM resulting from a duplication that would be misleading. A DNA segment (the dark cyan segment in the reference genome and the left dark cyan segment in the case genome) is duplicated and inserted to another position on the same chromosome (the dark cyan segment on the right part of the case genome). The purple bar in the plot represents the breakpoint of the insertion. A paired-end read (the blue arrows) that spans the breakpoint in the case genome is sequenced. When the pair is mapped back to the reference genome, the mapped distance between the two ends is significantly larger than the insert size, which is similar to a paired-end read resulting from a large deletion in the case genome.

when detecting deletions and small insertions, but it has limited power for detecting insertions larger than 500 bp. The reason is that, for an insertion larger than the insert size, the signature of the PEMs is complex. Some of PEMs may appear to contain a deletion between the pairs instead (Fig. 5*F*), making it difficult for the algorithm to make a prediction. In contrast, BIC-seq performs very well at detecting larger CNVs. For low coverage data (Fig. 5 *A* and *D*), BIC-seq performs well for detecting large deletions (e.g., one-copy loss of size >50 kb) in addition to large insertions, but BreakDancer does poorly because there are few or no supporting PEMs for the SVs. This simulation suggests that BIC-seq and the PEM-based algorithms are complementary methods and that a combination of the two methods would give higher sensitivity, especially for low-coverage data.

In practice, it is important to determine which sequencing depth will allow for the detection of nearly all CNVs. The simulation studies here suggest that read-depth methods should be able to detect almost all CNVs larger than 500 bp at 30× coverage. Combined with a PEM-based method, it should be possible to detect all but very small deletions and insertions. In general, one should first use a read-depth method followed by a PEM-based method to discover large CNVs and a PEM-based method followed by a read-depth method to capture smaller CNVs.

Discussion

We described a statistically rigorous and computationally efficient algorithm for detecting CNVs from high-throughput sequencing data. BIC-seq was able to detect CNVs as small as 40 bp in three human cancer genomes (one GBM and two AMLs) with medium-to-high coverage. This algorithm is fast: Once the reads are aligned, it can resolve CNVs in a 30× genome with 10-bp resolution in approximately 2 h. Subsequent qPCR validation demonstrated that the sequencing-based CNV calls, including the estimated copy number magnitude, are more accurate than those from array platforms. Taking advantage of the discrete sequencing data, we also developed a method to assign error bars for the breakpoints, which would be important for subsequent localization of the breakpoints. To facilitate detection of recurrent CNVs in multiple tumor genomes, we extended BIC-seq to the multisample case. There are several algorithms, such as GISTIC (Genomic Identification of Significant Targets in Cancer) (49), for finding recurrent CNVs in a group of samples. These algorithms combine information from multiple samples after each profile is segmented individually. Multisample BIC-seq, on the other hand, performs multiprofile segmentation directly on the raw data and therefore is more powerful for detecting recurrent small or low-copy-change CNVs that may not be captured on a sample-by-sample segmentation.

Paired-end sequencing with an insert size ranging from 200–300 bp to approximately 3–5 kb is becoming the standard in whole-genome sequencing. The only exceptions are when the DNA is sufficiently degraded that a paired-end library cannot be constructed. PEM is advantageous in several aspects, including allowing one to infer the location of duplicated segments and

to detect balanced rearrangements such as translocations and inversions. However, its performance in genomic regions enriched with repeats or segmental duplications is limited. Further, as shown in the simulation study above, one should combine the PEM algorithms with a read-density approach to improve the sensitivity and specificity of CNV detection. For example, if the read-density-based methods detect a copy loss region, there should be paired-end reads surrounding the detected region whose mapped distances are significantly larger than expected. Assuming sufficient sequencing depth, the copy loss calls that are not supported by such paired-end reads are then likely to be false positives. Likewise, the methods based on PEM have limited power in detecting large insertions, but the read-density-based methods are better at detecting larger CNV events. Thus, one may first apply the read-density-based methods to detect the large copy gain regions, and then use PEM to confirm the existence of the copy gain region or to refine the breakpoint positions.

In the 1,000 Genomes Project (50), the investigators identified thousands of SVs in a normal population based on the consensus of many SV/CNV detection algorithms, but they only focused on deletions. They found that, for deletions, Genome STRiP (Structure in Populations) (51), which combines the read-depth and the paired-end information, gave the lowest false discovery rate. However, this method relies on population-wide information and has limited power for detecting rare CNVs. Furthermore, because this algorithm first uses a PEM-based method to call the candidate CNVs and then uses read depth to remove false positives, it also suffers from the similar problem as BreakDancer—i.e., low sensitivity for detecting large insertions. For the same reason, this algorithm also has limited power for detecting large deletions using low-coverage data, e.g., at 0.3× coverage as shown in Fig. 5 *A* and *D*.

Array CGH methods have brought tremendous progress in characterizing copy number profiles in many diseases. But, as sequencing technology continues to improve, the cost of sequencing is rapidly declining and a multitude of samples will be subject to whole-genome sequencing. Given the power to detect CNVs and other genomic changes at unprecedented resolution, the next algorithmic challenge will be to determine the biological significance of the aberrations and identify the "driving" variants from the "passenger" variants. For more subtle variants, this problem will require delineating the phenotype more precisely and collecting appropriate samples as well as relevant controls. For some phenotypes, it will also require a population-scale sequencing effort to reach an acceptable sensitivity and specificity. In all cases, it will be important to handle the challenges of managing large amounts of data and to apply efficient and statistically powerful algorithms.

Materials and Methods

Data Availability. TCGA accession numbers for the GBM data are TCGA-06-0208-01 (tumor) and TCGA-06-0208-10 (normal). The sequencing data are available from the database of Genotypes and Phenotypes (dbGaP). Because these are individual-level data, authorization is required for data access. The segmented profiles (level III) for the Agilent and Affymetrix data were obtained from the Cancer Genome Atlas data portal (http://tcga-data.nci.nih.gov/tcga/). These profiles had been generated by applying the Circular Binary Segmentation (52) method, one of the leading CNV detection algorithms for microarray platforms.

Outlier Removal. Some genomic positions can have mapped short reads several orders of magnitude more than their neighboring regions. We remove these outliers before further analysis. To determine whether a genomic position s_0 is an outlier, we choose a local window of s_0 and calculate the pth quantile q (e.g., 0.95th quantile) of the read counts at the genomic positions in the local window. If read count n_{s_0} at s_0 is more than m (e.g., $m = 5$) times of the quantile q, this genomic position will be viewed as an outlier position and the read count at this position will be set as mq. Here, the local window is chosen as the window (a, b) such that both the left window (a, s_0) and the right window (s_0, b) contain $w/2$ genomic positions that have at least one read, where w is a parameter determined by the user. In the CNV analysis on the GBM and AML genomes, we set $w = 200$, $m = 5$, and $p = 0.95$.

The Statistical Model and the Segmentation Algorithm. Given a short read R that was mapped to the reference genome, let Y be one if the read is from the tumor genome and zero if from the normal genome, and S be the mapped position of the read. We view the short read R as equivalent to the two-dimensional random variable (Y, S)—i.e., $R = (Y,S)$. Suppose that the joint distribution of $R = (Y,S)$ is $f(y,s)$, and the marginal distribution function of S is $f(s)$. Then, given a set of short reads $R_1 = (y_1,s_1),\ldots,R_n = (y_n,s_n)$ on a reference chromosome c, the joint likelihood of these short reads are $L_n = \prod_{i=1}^{n} f(y_i,s_i) = \prod_{i=1}^{n} q_{s_i}^{y_i}(1-q_{s_i})^{1-y_i} f(s_i)$, where $q_{s_i} = Pr(Y = 1|S = s_i)$ is the probability of a read being a tumor read conditional on the read being mapped to the genomic position s_i.

If the tumor genome is identical to the normal genome, the conditional probabilities q_{s_i} would be the same genome wide. However, if the tumor genome has any somatic CNV regions, the conditional probabilities q_{s_i} will be different for different CNV regions, but for any two points between two consecutive breakpoints, the conditional probabilities q_{s_i} would still be the same. To determine which set of breakpoints is the best, we use the BIC as the criterion, and the set of the breakpoints with the smaller BIC is preferred. The BIC of a model is in general defined as

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

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 profile with m breakpoints, there are $m+1$ parameters $p_0,p_1,\ldots,p_m$, where p_i are the common probabilities q_s for s between two consecutive breakpoint τ_i and τ_{i+1}. We thus define the BIC as

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

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. Note that though the distribution function f is unknown, the term involving f is the same for all models. Thus, we can compare two models' BIC without specifying any parametric form about f. (See *SI Appendix* for the BIC-seq algorithm.)

The above statistical model can be easily extended to the multisample case. Specifically, suppose that we have sequencing data from G pairs of tumor/normal genomes and that there are n_k short reads from the kth pair of the genomes on a reference chromosome c. Let L_{n_k} be the joint likelihood of the kth pair as above. Then, the joint likelihood of the short reads on c from G pairs of tumor/normal genomes is just the product of the individual likelihood L_{n_k}—i.e., $L = \prod_{k=1}^{G} L_{n_k}$. The BIC becomes

$$\mathrm{BIC}(\lambda) = -2\log(L) + (m+1)G\lambda\log(\sum_{k=1}^{G} n_k),$$

where m is the number of breakpoints. Then, we can easily extend the corresponding algorithm for CNV detection to the multisample case.

Credible Intervals to Breakpoints. We develop a Gibbs sampler to assign credible intervals to the breakpoints given by BIC-seq. Given a genomic window a, b, assume that there is only one breakpoint τ in the window. Suppose that the reads in the window a, b are $D = (R_{i_1},\ldots,R_{i_2})$. Let p_1, p_2 be the probabilities of a read being a tumor read before and after the breakpoint τ. Then, conditional on the breakpoint τ and the probabilities p_1, p_2, the joint distribution of $D = (R_{i_1},\ldots,R_{i_2})$ is

$$f(D|\tau,p_1,p_2) = \prod_{k=i_1}^{i_2}[p_1^{y_k}(1-p_1)^{1-y_k}I(s_k \le \tau) + p_2^{y_k}(1-p_2)^{1-y_k}I(s_k > \tau)]f(s_k),$$

where $I(\cdot)$ is the indicator function. Put uniform priors on p_1, p_2, and τ. We have that the full conditional distributions of p_1, p_2 are β-distributions and

the full conditional distribution of τ can be easily sampled using the inverse of its cumulative distribution function (*SI Appendix*). Given a breakpoint τ_j predicted by BIC-seq, we can take a, b to be the breakpoints before and after the breakpoint τ_j. However, to expedite the Gibbs sampler, we choose a, b such that there are at most 1,000 normal reads in the intervals (a, τ_j) and (τ_j, b) (*SI Appendix*).

Experimental Validation Using qPCR. All CNVs detected by BIC-seq were validated with qPCR. Primers were constructed internal to the CNV with Primer3 (53). See *SI Appendix*, Table S1 for primer sequences and sizes of qPCR amplification. For all CNVs, reactions were performed in duplicate. The reaction mixture (25 μL) contained 1× Power SYBR green (catalog no. 4367659, Applied Biosystems), 2 μM of each primer, and 10 ng of DNA. Reactions mixtures were run on a 7900HT fast real-time PCR machine (Applied Biosystems) with the following thermal conditions: one cycle of 95 °C for 10 min, 40 cycles of 94 °C for 15 s, 60 °C for 30 s, followed by data collection. Data were initially analyzed with 7900HT Fast System Software at a threshold determination of 0.02 for TCGA-06-0208 tumor and blood. Threshold cycle (C_T) values were exported to Excel for further statistical analysis. The average ΔC_Ts were calculated and compared against the BIC-seq copy ratios, to assess the accuracy of the program.

Simulation. We used chromosome 22 (hg18) as the template to generate tumor chromosomes containing CNVs. The CNV sizes were set as 100 or 500 bp, 1, 5, 10, 50, and 100 kb. The copy number of the CNV segments were chosen as 0, 1, 3, 4, 5, 6, which corresponds to homozygous deletion, heterozygous deletion, one-, two-, three-, and four-copy gain, respectively. The positions of the CNVs were randomly selected and the only constraint was that the CNV region should not contain more than 20% Ns. In each simulation, we generated two tumor chromosomes which contained 42 (equal to 7 CNV sizes × 6 copy ratios) CNVs with different sizes and copies of CNV segments. The two chromosomes are identical except in the CNV regions. For homozygous deletion both copies of the chromosomes had the randomly selected segments deleted, but for heterozygous deletion, only one of the two simulated chromosomes had the deletion. For copy gain regions, if the copy number was even, the two simulated chromosomes would be homologous in the CNV region; but if the copy number was odd, one of the two simulated chromosomes had one more copy than the other chromosome. Duplicated segments were placed next to its original location. After the tumor chromosomes were generated, we used them as templates and applied the sequencing simulator MetaSim (48) to generate 36 bp short reads. The sequencing error model was the empirical error model of Illumina sequencing platform. We generated 600 million short reads from each pair of simulated chromosomes and approximately 57% of them can be uniquely mapped back to chromosome 22, resulting in about 30× coverage of chromosome 22. The short reads from normal genome were generated using chromosome 22 as template. For 0.3× and 3× coverage simulation, we randomly sampled 1% and 10% of the 30× simulated data, respectively. A CNV gain (loss) region is viewed as detected if at least 50% of the region is covered by CNV gain (loss) regions predicted by BIC-seq.

For the comparison of BIC-seq and BreakDancer, we generated another set of tumor chromosomes using chromosome 22 as the template. The synthetic tumor chromosomes contained 42 CNVs whose sizes and copy numbers were set as above. The simulation setup for deletions was the same as before. Duplications, on the other hand, were randomly inserted to the synthetic chromosome rather than being placed next to the original segments. After the synthetic tumor chromosomes were generated, we applied MetaSim to generate 36-bp paired-end reads assuming that the insert sizes are normally distributed with the mean of 220 bp and the SD of 20 bp. The insert size and the SD of the insert size were estimated from AML2 paired-end sequencing data. For each synthetic tumor chromosome, we simulated two sets of paired-end reads, one with 6 million paired-end reads and the other with 0.6 million. We used BWA (27) with the default parameters to map the reads back to the reference genome; the resulting sequence coverage is about 6× for the 6 million dataset and 0.6× for the 0.6 million dataset. BreakDancer was then applied to call the insertions and deletions. For the copy loss, if there is a deletion called by BreakDancer reciprocally overlapping at least 50% with the "true" CNV region, we say that the copy loss is correctly predicted. For copy gain regions, we used a lenient criterion to determine if the region is correctly predicted by BreakDancer. We say a copy gain region is correctly predicted by BreakDancer as long as there is some insertion predicted by BreakDancer that overlaps with any copy of the duplicated segment. Furthermore, sometimes BreakDancer predicts an insertion as a translocation (*SI Appendix*, Fig. S12), which is much larger than the original size of the insertion. These predictions were also counted as correct predictions. If we count these predictions as false positives, the performance of BreakDancer for detecting insertions larger than 500 bp would drop (*SI Appendix*, Fig. S13). For comparison, we also generated normal paired-end reads using chromosome 22 as a template and applied BIC-seq to call the CNVs. To make the comparison fair, BIC-seq is applied only based on half of the tumor paired-end reads. The performance of BIC-seq based on the paired-end reads drops slightly compared to its performance based on the single-end reads, but overall they are very similar (*SI Appendix*, Fig. S14). The common CNVs detected by both algorithms are shown in *SI Appendix*, Fig. S15.

Software. The R-package *BICseq* can be obtained from http://compbio.med.harvard.edu/Supplements/PNAS11.html.

ACKNOWLEDGMENTS. This work was supported by the US National Institutes of Health Grants RC1 HG005482 and R01 GM082798 (to P.J.P.) and U24 CA144025 (to R.K.).

1. Fanciulli M, et al. (2007) FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat Genet* 39:721–723.
2. Sebat J, et al. (2007) Strong association of de novo copy number mutations with autism. *Science* 316:445–449.
3. Stone J, et al. (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 455:237–241.
4. Stefansson H, et al. (2008) Large recurrent microdeletions associated with schizophrenia. *Nature* 455:232–236.
5. Walters R, et al. (2010) A new highly penetrant form of obesity due to deletions on chromosome 16p11. *Nature* 463:671–675.
6. Bochukova E, et al. (2010) Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* 463:666–670.
7. Walter M, et al. (2009) Acquired copy number alterations in adult acute myeloid leukemia genomes. *Proc Natl Acad Sci USA* 106:12950–12955.
8. Bredel M, et al. (2005) High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer Res* 65:4088–4096.
9. Beroukhim R, et al. (2010) The landscape of somatic copy-number alteration across human cancers. *Nature* 463:899–905.
10. Bignell G, et al. (2010) Signatures of mutation and selection in the cancer genome. *Nature* 463:893–898.
11. Redon R, et al. (2006) Global variation in copy number in the human genome. *Nature* 444:444–454.
12. McCarroll S, et al. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* 40:1166–1174.
13. Díaz de Ståhl T, et al. (2008) Profiling of copy number variations (CNVs) in healthy individuals from three ethnic groups using a human genome 32 K BAC-clone-based array. *Hum Mutat* 29:398–408.
14. Zogopoulos G, et al. (2007) Germ-line DNA copy number variation frequencies in a large North American population. *Hum Genet* 122:345–353.
15. Shaikh T, et al. (2009) High-resolution mapping and analysis of copy number variations in the human genome: A data resource for clinical and research applications. *Genome Res* 19:1682–1690.
16. Itsara A, et al. (2009) Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* 84:148–161.
17. Sindi S, Helman E, Bashir A, Raphael B (2009) A geometric approach for classification and comparison of structural variants. *Bioinformatics* 25:i222–230.
18. Hormozdiari F, Alkan C, Eichler E, Sahinalp S (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* 19:1270–1278.
19. Chen K, et al. (2009) BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* 6:677–681.
20. Medvedev P, Stanciu M, Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* 6:S13–S20.
21. Tuzun E, et al. (2005) Fine-scale structural variation of the human genome. *Nat Genet* 37:727–732.
22. Chiang D, et al. (2008) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* 6:99–103.
23. Xie C, Tammi M (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* 10:80.
24. Lai W, Johnson M, Kucherlapati R, Park P (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* 21:3763–3770.
25. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18:1851–1858.
26. Langmead B, Trapnell C, Pop M, Salzberg S (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.
27. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26:589–595.
28. Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464.
29. Bayer R (1972) Symmetric binary B-trees: Data structure and maintenance algorithms. *Acta Inf* 1:290–306.
30. Tanner M, Wong W (1987) The calculation of posterior distributions by data augmentation. *J Am Stat Assoc* 82:528–540.

31. Gelfand A, Smith A (1990) Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc* 85:398–409.
32. Futreal P, et al. (2004) A census of human cancer genes. *Nat Rev Cancer* 4:177–183.
33. McLendon R, et al. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455:1061–1068.
34. Zhu H, et al. (2009) Oncogenic EGFR signaling cooperates with loss of tumor suppressor gene functions in gliomagenesis. *Proc Natl Acad Sci USA* 106:2712–2716.
35. Kodama T, et al. (2004) ADAM12 is selectively overexpressed in human glioblastomas and is associated with glioblastoma cell proliferation and shedding of heparin-binding epidermal growth factor. *Am J Pathol* 165:1743–1753.
36. Esteller M, et al. (2000) Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *N Engl J Med* 343:1350–1354.
37. Kreth S, et al. (2011) O-Methylguanine-DNA methyltransferase (MGMT) mRNA expression predicts outcome in malignant glioma independent of MGMT promoter methylation. *PLoS One* 6:e17156.
38. Ansari K, Mandal S (2010) Mixed lineage leukemia: Roles in gene expression, hormone signaling and mRNA processing. *FEBS J* 277:1790–1804.
39. Parsons D, et al. (2010) The genetic landscape of the childhood cancer medulloblastoma. *Science* 331:435–439.
40. Wang X, et al. (2011) Somatic mutations of the mixed-lineage leukemia 3 (MLL3) gene in primary breast cancers. *Pathol Oncol Res* 17:429–433.
41. Balakrishnan A, et al. (2007) Novel somatic and germline mutations in cancer candidate genes in glioblastoma, melanoma, and pancreatic carcinoma. *Cancer Res* 67:3545–3550.
42. Kang J, Koo S, Kwon K, Park J (2010) AMY2A: A possible tumor-suppressor gene of 1p21. 1 loss in gastric carcinoma. *Int J Oncol* 36:1429–1435.
43. Bueno R, et al. (2010) Second generation sequencing of the mesothelioma tumor genome. *PLoS One* 5:e10612.
44. Ley T, et al. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456:66–72.
45. Mardis E, et al. (2009) Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* 361:1058–1066.
46. Dicker F, et al. (2010) Mutation analysis for RUNX1, MLL-PTD, FLT3-ITD, NPM1 and NRAS in 269 patients with MDS or secondary AML. *Leukemia* 24:1528–1532.
47. Shukla A, Green J (1991) Bowling plexopathy. *N Engl J Med* 324:928–928.
48. Richter D, Ott F, Auch A, Schmid R, Huson D (2008) MetaSim: A sequencing simulator for genomics and metagenomics. *PLoS One* 3:e3373.
49. Beroukhim R, et al. (2007) Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc Natl Acad Sci USA* 104:20007–20012.
50. Mills R, et al. (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* 470:59–65.
51. Handsaker R, Korn J, Nemesh J, McCarroll S (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* 43:269–276.
52. Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5:557–572.
53. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132:365–386.

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 **2**

2 CNV identification using the BIC **3**

3 Copy ratio estimation and significance assignment **5**

4 Credible intervals of breakpoints **5**

5 FDR estimate and an alternative way of specifying λ **6**

6 Distribution of short reads on the reference genome **8**

7 Examples of outliers **9**

8 Effect of the tuning parameter λ **9**

9 Comparison with SegSeq **12**

10 Comparison between sequencing and array platforms **12**

11 Variable read lengths in the AML1 and AML2 **16**

12 Comparison with a PEM-based method **19**

13 Primers for PCR validation **20**

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.

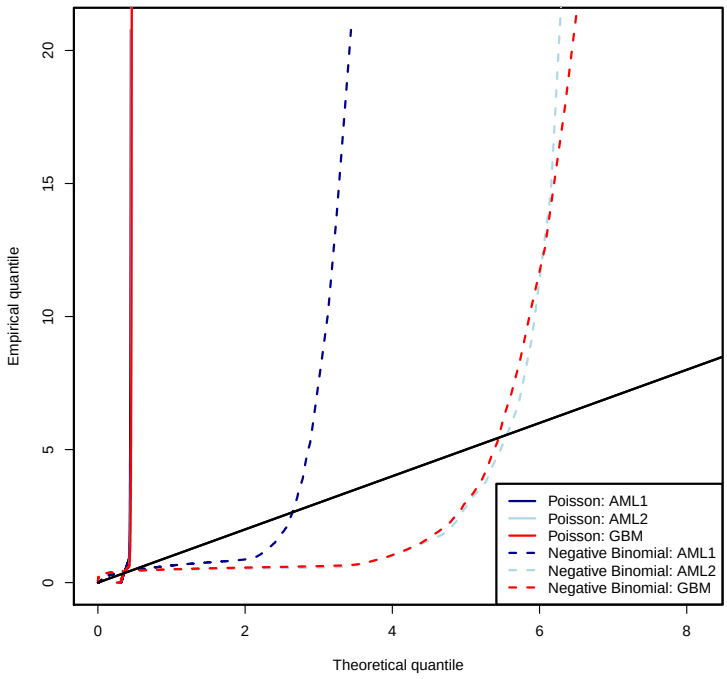


Fig. S1: QQ-plot of read counts in 1 Kb bins. The black line corresponds to the line $y = x$.

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.

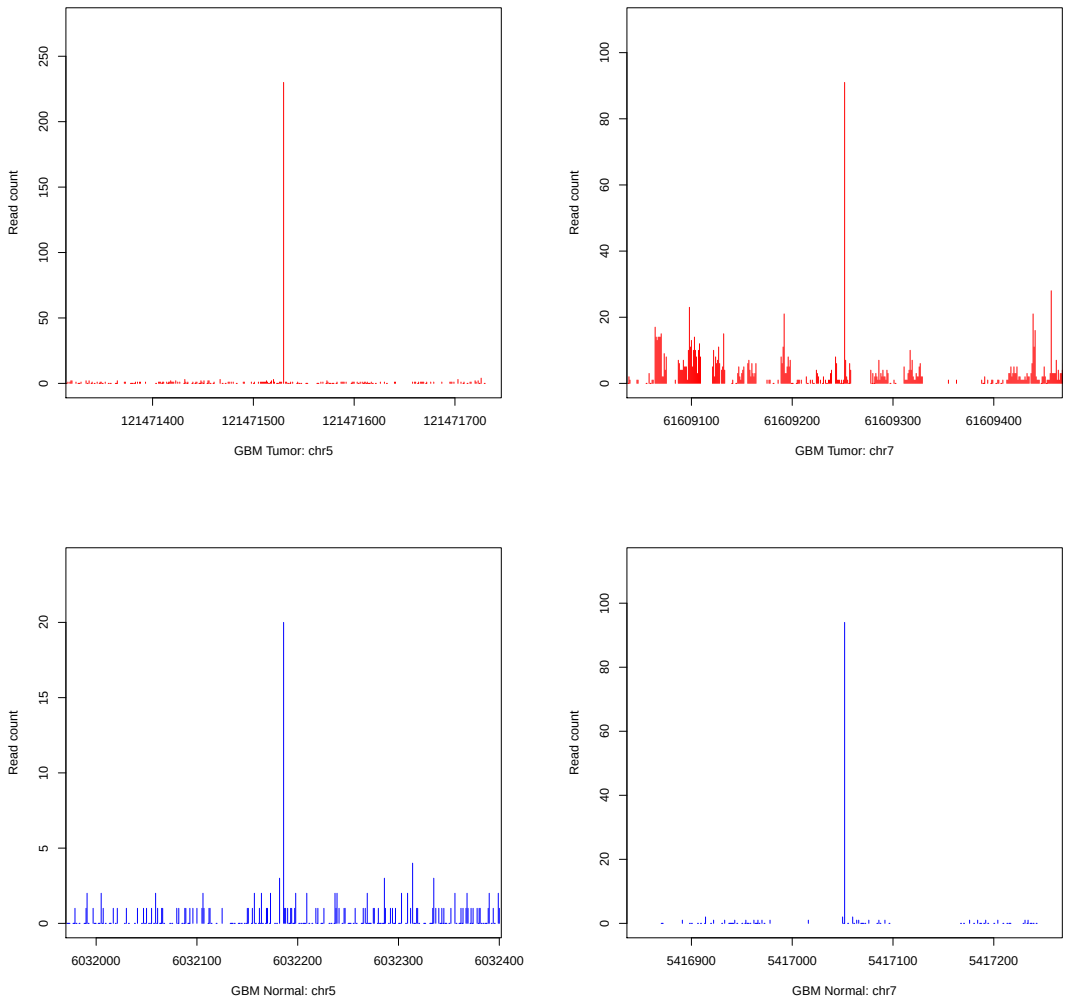


Fig. S2: Examples of genomic positions that have considerably more reads than their neighboring positions.

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

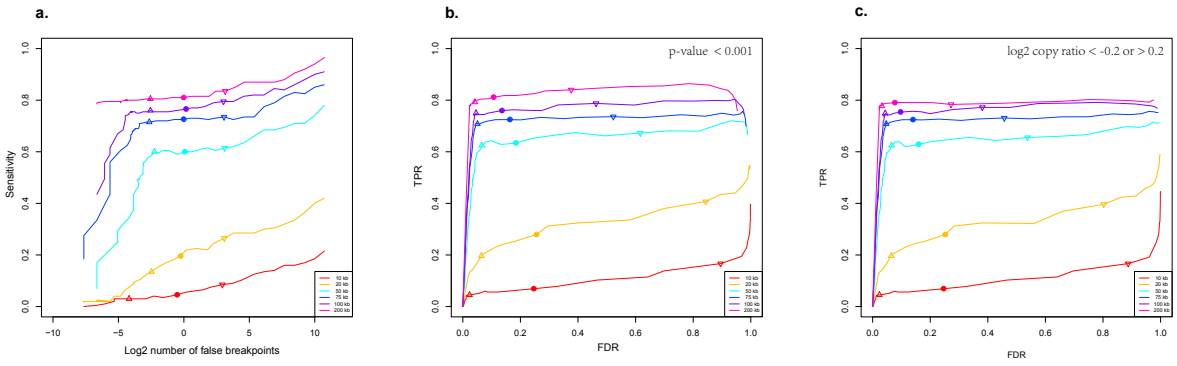


Fig. S3: (**a**) Sensitivity against the log2 number of false breakpoints: the point-up triangles, the bullets and the point-down triangles correspond to $\lambda = 1.2$, 1 and 0.8, respectively. (**b**) The TPR against the FDR: the point-up triangles, the bullets and the point-down triangles correspond 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 $\lambda = 2$.

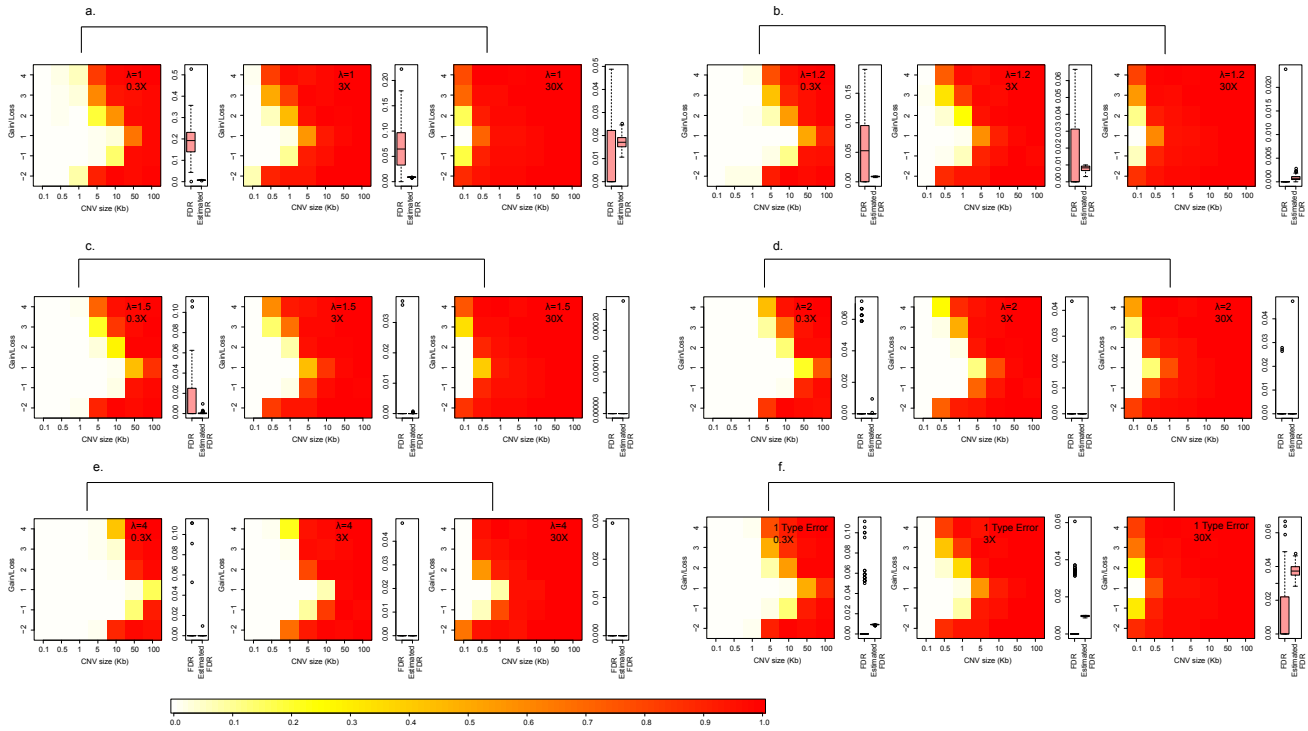


Fig. S4: The CNV detection power of BIC-seq for (**a**) $\lambda = 1$, (**b**) $\lambda = 1.2$, (**c**) $\lambda = 1.5$, (**d**) $\lambda = 2$, (**e**) $\lambda = 4$, (**f**) 1 type I error. The CNVs were chosen as regions with log2 copy ratio greater than 0.2 or less than -0.2.

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

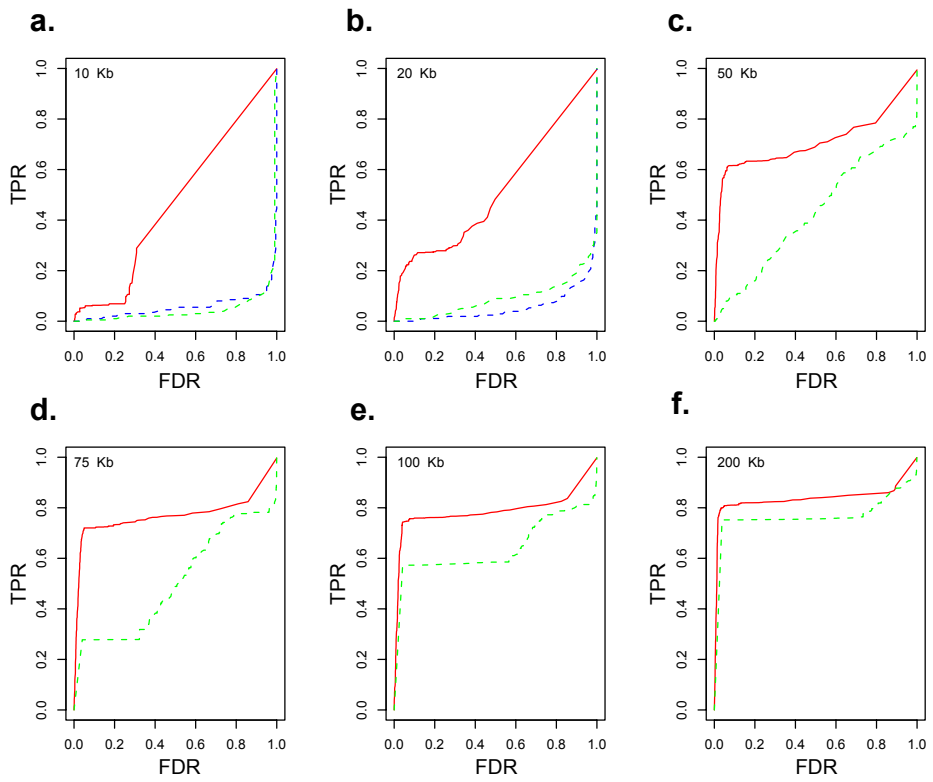


Fig. S5: Comparison of BIC-seq and SegSeq. BIC-seq: red solid lines; SegSeq: blue dashed lines ($w = 50$) and green dashed lines ($w = 400$).

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).

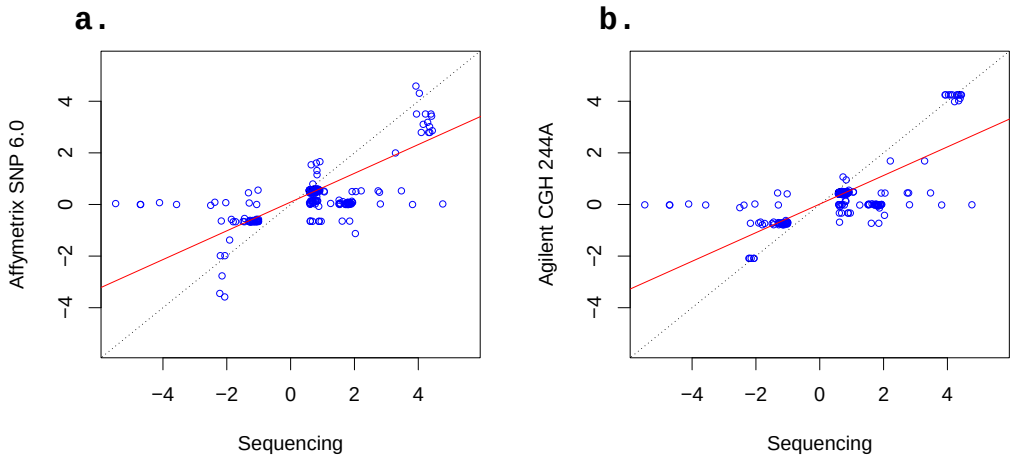


Fig. S6: Comparison of the copy ratio estimates given by BIC-seq and two array platforms, (**a**) Affymetrix SNP 6.0 Array and (**b**) Agilent 244A CGH array. The y-axes are the log2 copy ratio given by the array platforms and the x-axes are the log2 copy ratio given by BIC-seq.

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.

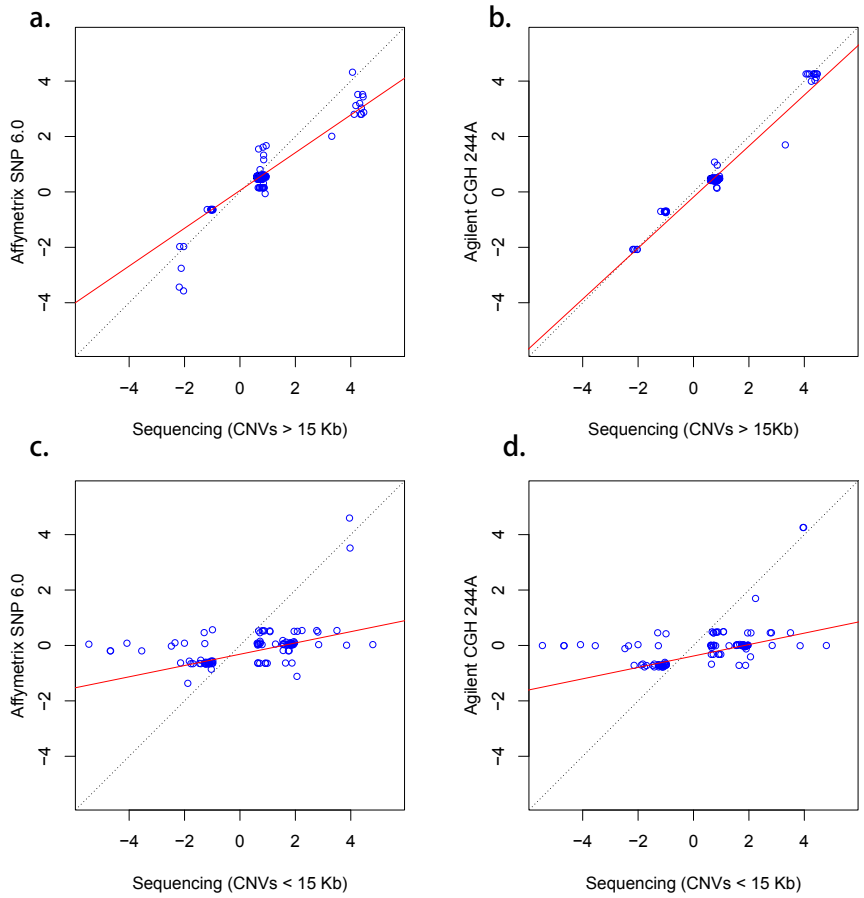


Fig. S7: Comparison of the copy ratio estimates given by BIC-seq and two array platforms; the CNVs given by BIC-seq were classified into two groups, small CNV group (less than 15 Kb) and large CNV group (larger than 15 Kb). The y-axes are the log2 copy ratio given by the array platforms and the x-axes 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.

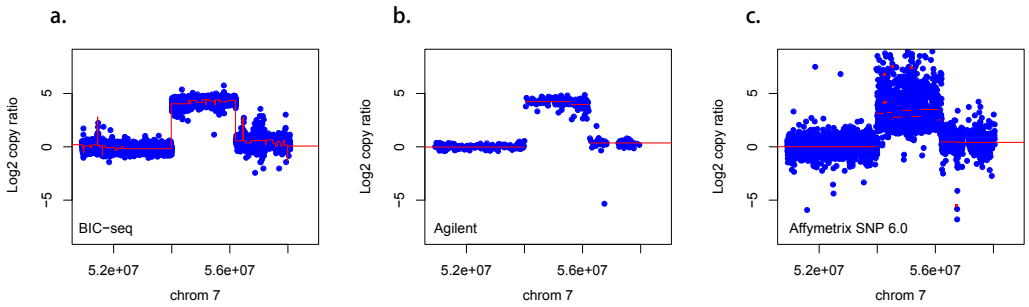
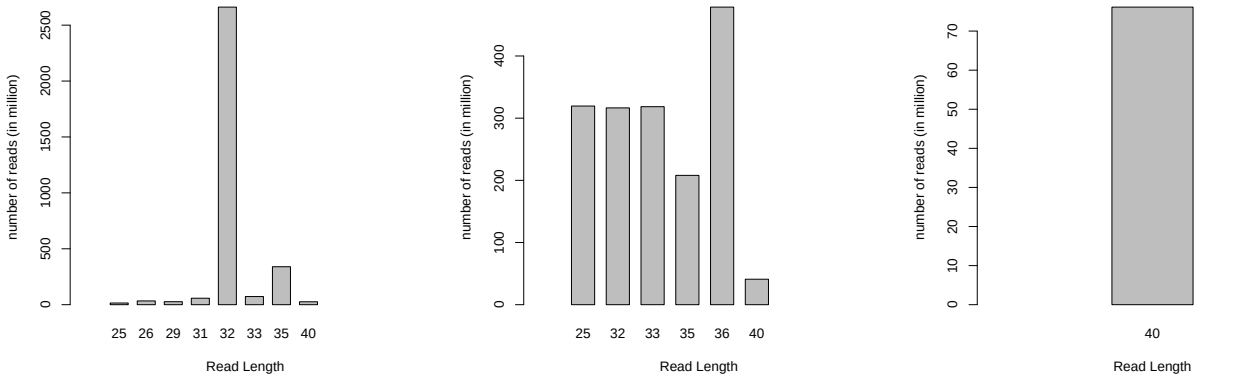


Fig. S8: A large CNV region that was identified by all three platforms.

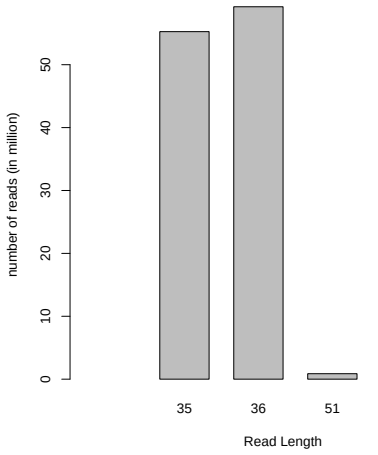
11 Variable read lengths in the AML1 and AML2



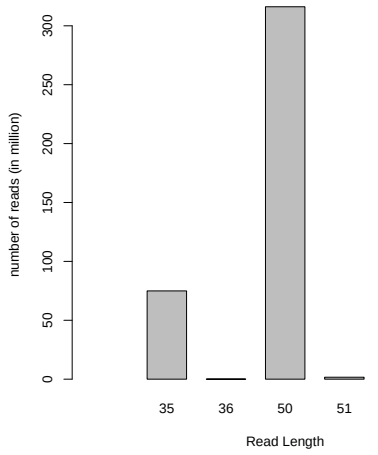
(a) Tumor: single-end (b) Normal: single-end (c) Tumor: paired-end

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

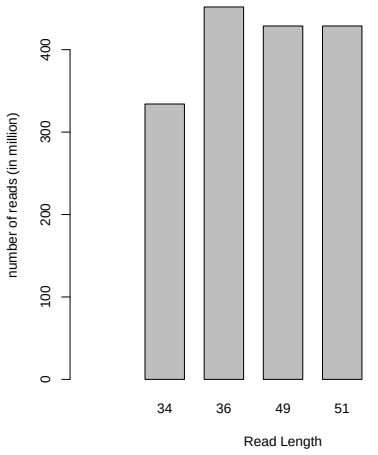
Figure S9 shows the distribution of single-end reads and pair-end reads of AML1. The AML1 normal genome has no paired-end reads. As there are many more single-end reads than the paired-end reads, we only used the single-end reads for CNV detection. Since most of the AML1 tumor single-end reads are 32 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.



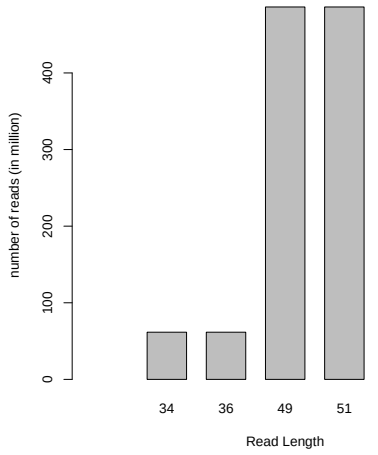
(a) Tumor: single-end



(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.

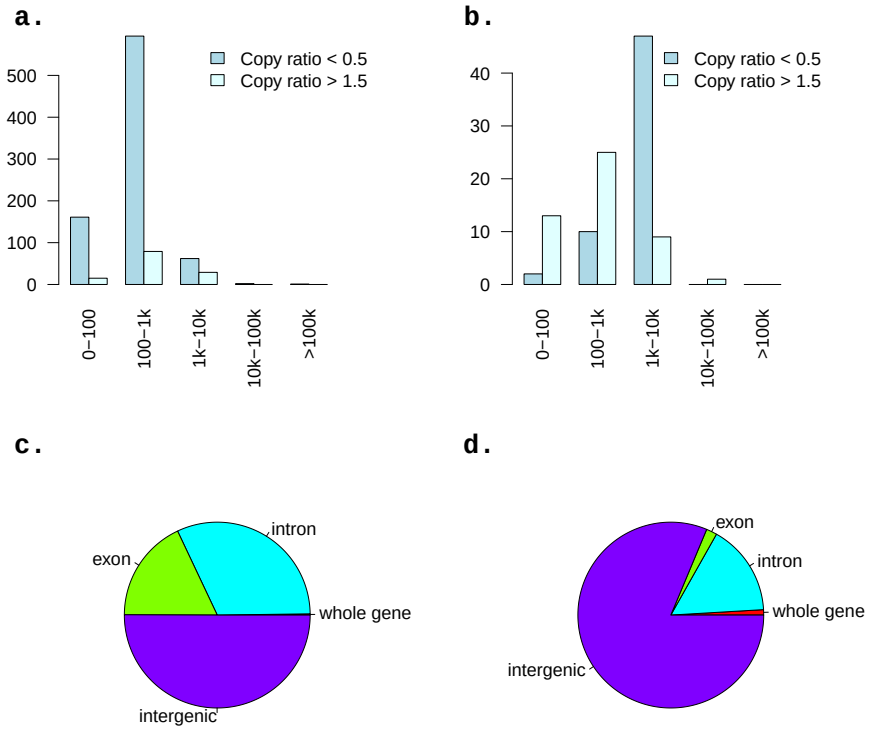


Fig. S11: CNVs detected by BIC-seq in two AML genomes. (**a**,**b**) The distribution of putative CNVs detected in AML1 and AML2 ($\lambda = 4$). Here, the labels on the x-axes demonstrated the size ranges of the CNVs in the corresponding groups. (**c**, **d**) Overlaps of AML1 and AML2 CNVs with Refseq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

12 Comparison with a PEM-based method

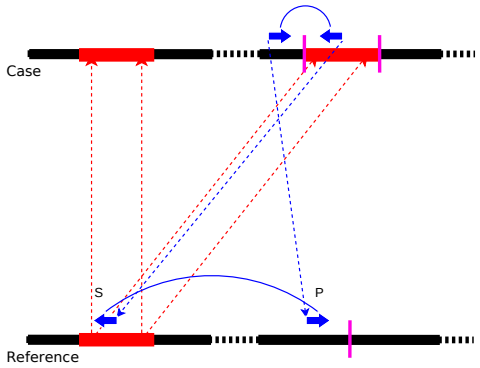


Fig. S12: An insertion that is predicted as a large translocation by BreakDancer

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.

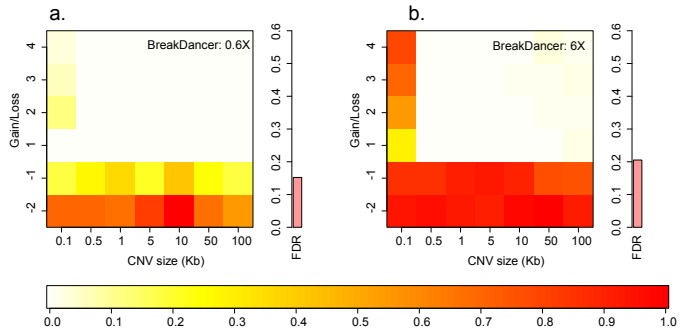


Fig. S13: The SV detection power of BreakDancer. (**a**) 0.6x. (**b**) 6x.

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

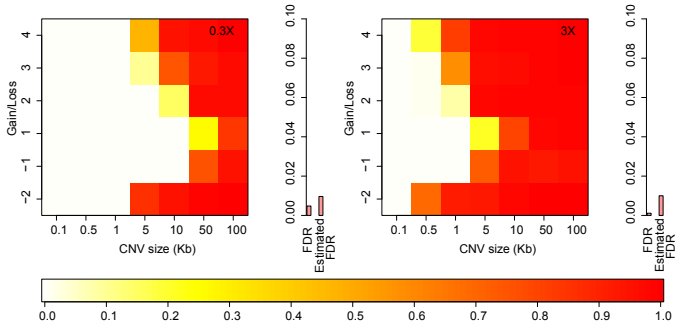


Fig. S14: The CNV detection power of BIC-seq on paired-end data. (**a**) 0.3X. (**b**) 3X.

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.

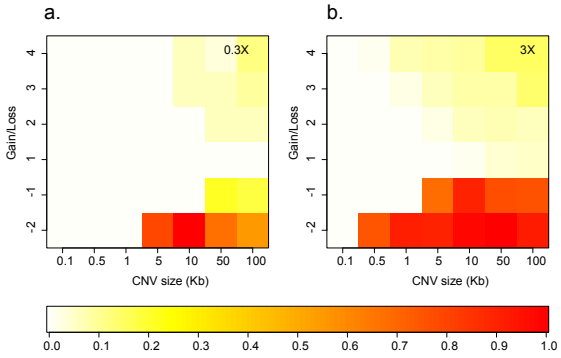


Fig. S15: Common CNVs detected by BIC-seq and BreakDancer. The color bar bar indicates the percentage of the true CNVs detected by both algorithms. Clearly, the deletions are highly overlapping but not amplifications for both 0.3X and 3X coverages.

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.