BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 24 2010, pages 3051–3058
doi:10.1093/bioinformatics/btq587

Sequence analysis

Advance Access publication October 21, 2010

CNAseg—a novel framework for identification of copy number changes in cancer from second-generation sequencing data

Sergii Ivakhno[1,2,*], Tom Royce[3], Anthony J. Cox[2], Dirk J. Evers[2], R. Keira Cheetham[2] and Simon Tavaré[1]

[1]Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, [2]Illumina Cambridge, Chesterford Research Park, Little Chesterford, CB10 1XL, UK and [3]Illumina Inc., Corporate Headquarters, 9885 Towne Centre Drive, San Diego, CA 92121, USA

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Copy number abnormalities (CNAs) represent an important type of genetic mutation that can lead to abnormal cell growth and proliferation. New high-throughput sequencing technologies promise comprehensive characterization of CNAs. In contrast to microarrays, where probe design follows a carefully developed protocol, reads represent a random sample from a library and may be prone to representation biases due to GC content and other factors. The discrimination between true and false positive CNAs becomes an important issue.

Results: We present a novel approach, called CNAseg, to identify CNAs from second-generation sequencing data. It uses depth of coverage to estimate copy number states and flowcell-to-flowcell variability in cancer and normal samples to control the false positive rate. We tested the method using the COLO-829 melanoma cell line sequenced to 40-fold coverage. An extensive simulation scheme was developed to recreate different scenarios of copy number changes and depth of coverage by altering a real dataset with spiked-in CNAs. Comparison to alternative approaches using both real and simulated datasets showed that CNAseg achieves superior precision and improved sensitivity estimates.

Availability: The CNAseg package and test data are available at http://www.compbio.group.cam.ac.uk/software.html.

Contact: Sergii.Ivakhno@cancer.org.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 18, 2010 revised on September 23, 2010; accepted on October 14, 2010

1 INTRODUCTION

A wide spectrum of mutation mechanisms contributes to the onset and progression of cancer (Stratton *et al*., 2009). Copy number abnormalities (CNAs) represent an important type of genetic mutation that lead to abnormal cell growth and proliferation (Santarius *et al*., 2010). Amplification of many oncogenes and loss of tumour suppressor genes have been implicated in the development and progression of the cancer phenotype. Different microarray technologies have been successfully used to identify CNAs in cancer (Pinkel *et al*., 2005). However, their resolution is limited by the number and location of probes on the array. Currently available SNP array platforms comprise >1 million probes and have a lower detection limit of 5–10 kb. New sequencing technologies promise comprehensive characterization of copy number profiles in cancer (Shendure *et al*., 2008). In particular, paired-end read mapping (PEM) allows the detection of insertions and deletions ranging from a few base pairs (indels) to tens of megabases. The fact that PEM allows mapping of the direction of reads makes it possible to identify copy-neutral events, such as structural rearrangements. Several methods have been developed to identify indels and structural rearrangements (Chen *et al*., 2009; Chiang *et al*., 2009; Hormozdiari *et al*., 2009; Lee *et al*., 2008, 2009; Ye *et al*., 2009). In contrast, detection of CNAs from PEM data is much less explored in the literature. Although several methods, such as RDXplorer (Yoon *et al*., 2009) and CNV-seq (Xie *et al*., 2009), have been developed to find copy number variation in normal individuals, they have not been tested on cancer data, so it is not clear if they can detect CNAs across the whole range of sizes and magnitudes present in cancer. Here we present a method, CNAseg, developed specifically for identification of CNAs in cancer from second-generation sequencing data.

In contrast to microarrays, where probe design follows a carefully developed protocol, reads represent a random sampling from a library and may be prone to representation biases due to GC content and other physico-chemical characteristics. Variability could also arise from misalignment caused by mutation hotspots in a tumour. Consequently, discrimination between true and false positives becomes an important issue. A limitation of the published approaches is that (in their current versions) they do not facilitate assessment of the false positive error rate of detected CNAs in cancer genomes. We have developed a novel framework for the identification of CNA events that uses flowcell-to-flowcell variability to estimate the false positive rate and the depth of coverage to finalize copy number calls. In addition to dataset-specific thresholding, our method has other advantages. It uses the Skellam distribution to compare read depth in tumour and control samples, which allows the use of smaller window sizes for copy number estimation and leads to greater sensitivity in pinpointing breakpoints for small CNAs. By comparing the method to alternative approaches using real and simulated datasets, we show that CNAseg achieves superior precision and improved sensitivity estimates.

*To whom correspondence should be addressed.

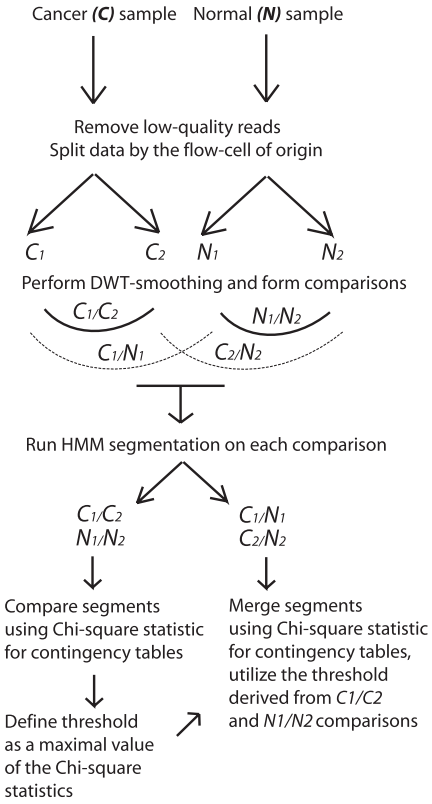

Fig. 1. Diagram showing the workflow of the CNAseg approach. The algorithm is composed of several distinct steps that preprocess and de-noise the data using a discrete wavelet transform, segment them with an HMM, approximate the false discovery rate by using a χ^2 statistic for the intra-dataset comparison and identify segments of distinct copy number states in the cancer genome.

2 METHODS

2.1 CNAseg framework

The CNAseg algorithm comprises several steps: the QC-filtered read counts are de-noised using a discrete wavelet transform, segmented with an HMM and followed by a segment merging step, where inter-lane variability in both cancer and normal samples is used to estimate the merging threshold (Fig. 1).

2.2 Data pre-processing and discrete wavelet transform smoothing

The Illumina Genome Analyzer (GA) can generate more than 60 GB of 50–100 bp paired-end reads from both sides of an insert spanning 100–300 bp. Sequencing on a small number of flow cells can therefore achieve 40× coverage of cancer and reference genomes. The GA pipeline (Bentley *et al*., 2008) was used for image processing, alignment and generation of output files in BAM format (Li *et al*., 2009a). Two BAM files for the cancer and matched normal reads are used by CNAseg for CNA detection. Before segmentation, aligned reads undergo several pre-processing steps to remove low-quality reads. First, poor-quality reads with low alignment score are discarded in order to ensure robust signal with small number of misalignment artefacts. For reads to pass quality filtering in the ELAND alignment, they should exceed a threshold of five for a single read alignment score and one for a paired alignment score. Next, reads that passed both thresholds are used to derive counts in non-overlapping windows. In contrast to methods relying on normal approximation, where windows need to have some pre-specified lower bound, CNAseg does not impose restrictions on the window size. To ensure fast running time of the method, we use a window of size 50 bp. We derive a correction factor for GC content using LOWESS regression for tumour and normal counts in 10 KB windows. This correction is applied to the dataset before splitting the reads by the flowcell of origin.

Different families of repetitive sequences can cause difficulties for alignment algorithms and create regions of poor alignability. Since breakpoints of many structural variants occur within repeats, low read counts created in this way may complicate estimation of the correct copy number state by the segmentation algorithm. Additional variability in read counts may be introduced by clusters of point mutations in cancer genomes or indels, although gapped alignment and indel-detection software alleviate this problem. To improve estimation of relative copy numbers, CNAseg uses an undecimated discrete wavelet transform (DWT) to smooth the count data. The method first calculates the wavelet transform of the noisy signal, then shrinks the noisy wavelet coefficients and finally computes the inverse transform from the modified coefficients (Nason *et al*., 2008; Percival *et al*., 2005). The number of decomposition levels depends on the length of the vector x of window counts for each chromosome and is calculated as $\lfloor \log_2(x) \rfloor$. In section 3, we show that this approach selectively smoothes out regions of low alignability while preserving signal elsewhere. The smoothing step is especially relevant to ensure that HMM does not over-segment the data in the regions of low alignability, since at low counts more mixture components will be required to represent the data.

2.3 HMM segmentation

CNAseg performs a segmentation of the read counts using a Hidden Markov Model (HMM). Counts of reads in windows of predefined size from the normal and cancer genomes are obtained, and then normalized so that both genomes have the same total number of reads. This is achieved by rescaling and rounding the counts from the experiment with the smaller number of reads. These normalized counts act as observed variables in the CNAseg model, while hidden variables represent the relative copy number states of the cancer genome. CNAseg segments the genome using individual 'bins' rather than chromosomes: large chromosomes are divided into two or three equally spaced bins and segmentation is carried out over these bins for improved memory handling.

The HMM uses the differences in the read counts for the segmentation. We have found that a suitable emission distribution for the HMM is provided by the Skellam distribution (Skellam, 1946). One way to describe this distribution is as follows (Karlis *et al*., 2005): define random variables X_n and X_c by $X_n = W_n + W, X_c = W_c + W$ where $W_n \sim$ Poisson(μ_n), $W_c \sim$ Poisson(μ_c), W is a positive random integer-valued variable and W_c and W_n are independent. Then $Z = X_c - X_n = W_c - W_n$ has the Skellam distribution with mean $\mu_c - \mu_n$ and variance $\mu_c + \mu_n$.

Selection of the appropriate number of HMM states is the second step in the HMM design after specifying an appropriate emission distribution. Ideally, we want each copy number in the cancer genome to be represented by an HMM state. However, variability in read depth, losses and gains in the matched normal genome and genomic instability in cancer make it hard to estimate the copy number state unambiguously. In addition, underrepresented copy number states may lead to singularity problems during HMM convergence. We use an approximation based on k-means clustering to determine the putative number of copy number states in cancer and normal genomes: read counts in each window are clustered between 2 and 7 clusters. The best partitioning is selected using the Calinski–Harabasz (CH) pseudo F-statistic (Calinski *et al*., 1974), which is based on the ratio of within-cluster sum of squares to between-cluster sum of squares. The CH selection criterion showed superior performance in the comparison of 30 techniques for choosing the appropriate number of clusters (Milligan *et al*., 1985).

2.4 Segment merging

HMM segmentation identifies segments of contiguous windows with similar read counts. Unfortunately, these segments will not only represent distinct copy number states, but also capture local variability in read depth arising from different noise sources. In practice, a strategy for merging segments with the same copy number is required to achieve more accurate CNA calls. Although the problem of segment merging is not unique to sequence data [a similar strategy was used for segmentation of microarray data (Willenbrock *et al.*, 2005)], merging count data requires the development of different approaches.

A desirable method should merge segments with the same copy number state and preserve segments with different copy number states. We proceed as follows. Consider two segments and suppose that in the first the normal sample has a counts, the cancer sample c counts, whereas in the second the normal sample has b counts, the cancer sample d counts. We want to merge segments for which $c/a \approx d/b$, for which a statistic based on the difference $ad-bc$ is suitable. Viewing the four counts in the form of a 2×2 contingency table, we see that the Pearson χ^2 statistic, a multiple of $(ad-bc)^2$, is appropriate. Low values of the statistic indicate the absence of significant differences between cancer and normal ratios for adjacent segments, suggesting that they represent the same copy number state. Applied in an iterative manner, such an approach produces an empirical distribution of goodness-of-fit statistics for adjacent segments. Given the threshold θ the following algorithm merges HMM segments:

Algorithm 1

(1) Compare the cancer and normal counts from adjacent segments using Pearson's χ^2 statistic, repeating for segments $j = 1, 2, \ldots, m-1$.

(2) Select the pair that produced the smallest χ^2 statistic.

(3) Merge these cancer and normal segments and recompute the χ^2 statistics with the adjacent left and right segments.

(4) Iterate until no statistics exceed the selected threshold θ.

2.5 Deriving the merging threshold

The CNAseg implementation requires specification of the threshold θ for the χ^2 statistic during the segment merging. We achieve this by exploiting intra-flowcell variability in tumour and normal data. In particular, if we segment the data with Algorithm 1, but this time comparing different flowcells within tumour or normal samples (intra-type comparison), rather than comparing tumour with normal samples (inter-type comparison), then the resulting HMM segments should only represent technical differences at the post-library preparation stage. These differences will also be reflected in values of the χ^2 statistics after performing the segment-merging step. The segmentation of reads from matched tumour and normal samples that provides higher statistics than those derived from intra-sample comparison (tumour–tumour/normal–normal) will usually suggest the presence of true copy number differences between samples for particular segments. Such a flowcell splitting approach utilizing intra-flowcell variability forms the basis for computing a merging threshold based on the χ^2 statistic. The CNAseg method is composed of the following steps:

(1) Split tumour and normal pre-processed read counts into two roughly equal groups by assigning flowcells to one or other group. Let $\mathcal{C}_1$ and $\mathcal{C}_2$ represent the counts in the cancer groups and $\mathcal{N}_1$ and $\mathcal{N}_2$ counts in the corresponding normal groups.

(2) Perform independent HMM segmentation with the Skellam emission distribution using two comparisons by combining groups as follows: (i) $\mathcal{C}_1 \bigcup \mathcal{C}_2$ versus $\mathcal{N}_1 \bigcup \mathcal{N}_2$ and (ii) $\mathcal{C}_1 \bigcup \mathcal{N}_1$ versus $\mathcal{C}_2 \bigcup \mathcal{N}_2$.

(3) Compute the χ^2 statistics between adjacent segments for the comparison (ii). Use these comparisons to set the merging threshold θ as the maximal value of the statistics in these two comparisons.

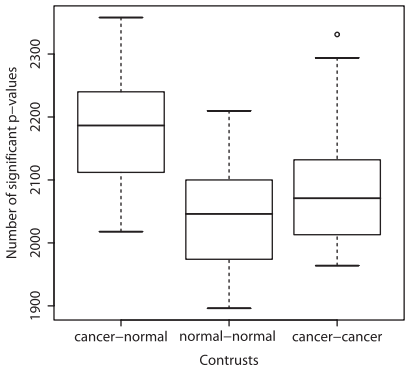

Fig. 2. The number of significant P-values in intra- and inter-sample flowcell comparisons. Adjacent 10 kb segments between two flowcells were compared using a χ^2 test.

(4) Merge HMM segments from the comparisons (i) using Algorithm 1. Use the threshold derived in the previous step to end the merging process.

(5) Estimate the approximate copy number state by comparing the log ratio of QC-filtered and DWT-smoothed read counts in the remaining segments.

Splitting reads into two comparisons using the flowcell attribute is based on the premise that reads from each flowcell have roughly equal noise patterns along the genome. However, the noise will depend on the experimental design during sample preparation: usually cancer and normal samples will be prepared separately and it can be argued that some of the between-sample variation will not be reflected in the inter-flowcell comparisons for normal samples.

To test this idea, we split reads by their flow cells of origin and compared counts between two flowcells for adjacent 10 kb windows using the χ^2 test. The distribution of log odds ratios for each flow cell comparison was obtained independently for normal-to-normal, cancer-to-cancer and cancer-to-normal comparisons. Chromosome 10 from the COLO-829 cell line (see Section 3) was reported to have the smallest number of focal CNAs and was used in the comparison to ensure that a high odds ratio between adjacent windows can be attributed to noise patterns rather than CNAs in cancer. After adjusting for CNA events in cancer (by removing HMM states with absolute mean log ratio >0.6, which corresponds to the putative gain of three copies), we found that the cancer-to-normal comparison provides more significant P-values (uncorrected for multiple testing) than the other two comparisons, reflecting in part technical factors due to library preparation (Fig. 2). The Welch Two Sample t-test comparing the number of significant P-values for flowcell comparisons between cancer-normal and normal–normal or cancer–cancer was 8.6×10^{-6} (df = 38.8) and 4.9×10^{-5} (df = 70.3), respectively, but 0.06 (df = 45.2) for normal–normal to cancer–cancer comparisons. To make sure that CNAseg can at least indirectly capture this between-sample library variation, we developed an algorithm for adjusting the threshold value as follows:

(1) Estimate the ratio of median values of χ^2 statistics for between-sample and within-sample comparisons for each bin.

(2) Find the second quantile ρ of this median's ratio distribution. Since the magnitude of the ratio correlates with the CNA burden, the second quantile of the distribution will represent a bin with no or few CNAs and where differences in χ^2 statistics are mostly attributable to between-sample technical variation.

(3) Increase the final threshold θ in proportion to the magnitude of the ratio ρ. This approach should not only correct for any intra-sample

variability, but also adjust θ in cases where normal samples have higher variance.

3 RESULTS

3.1 Datasets included in this study

For the purpose of algorithm testing and comparison, we used the genome of the COLO-829 malignant melanoma cell line (European Genome-Phenome Archive accession number EGAS00000000052) sequenced to 40-fold coverage (Pleasance *et al*., 2010). COLO-829BL, a lymphoblastoid line derived from the same patient, was used as the control sample and was sequenced to 32-fold coverage. In addition, COLO-829 has been profiled with Affymetix SNP 6.0 arrays for independent confirmation of CNAs found through read depth analysis.

3.2 Specification of DWT-smoothing parameters

ELAND (Cox *et al*., 2006) does not place reads that have multiple mappings to the genome; these are normally repetitive sequences. This strategy decreases the false positive rate of the alignment, but can introduce unevenness in the read depth. Correspondingly, the 'low alignability' regions are often underrepresented with reads in both tumour and reference genomes. This creates difficulties in estimating correct copy number calls due to low read counts. In cancer genomes, the task of correcting local differences in read depth due to alignability is exacerbated by the presence of multiple losses and gains, which also alter read counts. The correction of alignability-induced read depth variability therefore should preserve read depth variation due to genuine CNA events.

The alignability property requires unambiguous specification and the following method was used to calculate it:

(1) Generate all possible 32mer single reads and align them to the reference genome with ELAND.

(2) Count the number of unique alignments covering each base pair.

(3) The uniquely aligned stretch of the genome will therefore have scores around 32, while highly repetitive regions will have near-zero scores.

We found that regions with zero alignability are mostly very small (up to 100 bp) (Fig. 3a), which falls short of the size of typical CNA events (1 kb–10 Mb). CNAseg applies an undecimated DWT to smooth read counts in the regions of low alignability. To explore its impact on read depth in the regions of zero alignability and elsewhere, we estimated read counts in these two regions before and after applying DWT. DWT adjustment increased the number of counts for regions with zero alignability, while preserving the number of counts in the rest of the genome (Fig. 3b and c). The smoothing effect of DWT reduces unevenness in read depth and decreases the number of non-CNA induced HMM segments. This reduction has two effects on the performance of CNAseg. First, a large number of HMM segments increases the chance of obtaining false positive CNAs in the final segmentation results. Second, the merging process with a large number of segments becomes computationally intensive, hence the speed-up in CNAseg running time due to the smaller number of starting segments (Supplementary Fig. S6).

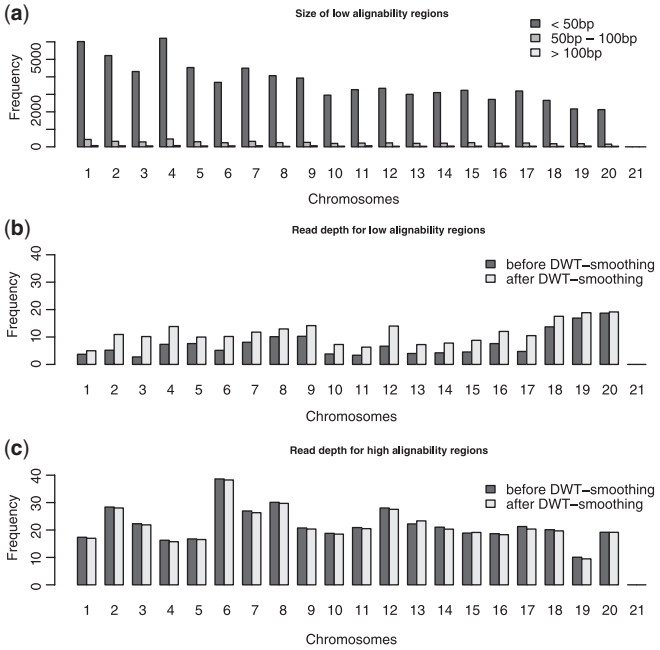

Fig. 3. DWT-based correction of read depth in zero alignability regions. (**a**) Size of contiguous segments with zero alignability. Read depth, estimated as a number of counts in 100 bp windows, in (**b**) low alignability regions and (**c**) high alignability regions, before and after DWT-based correction.

Other alignment algorithms have different strategies for dealing with alignment in repetitive regions: BWA (Li *et al*., 2009b) places short reads randomly across the multiple equally best positions while mrFAST attempts to align all reads that have multiple hits (Alkan *et al*., 2009). Simulation results and BWA realignment of the COLO-829 data confirmed that DWT adjustment of BWA-aligned reads does not introduce biases into read counts in regions of normal alignability (Supplementary Figs S1–S3 and section 'Relevance of spline-correction for identification of recurrent CNAs'). mrFAST introduces more biases (Supplementary Fig. S4); however, this method was specifically developed to improve the identification of CNVs. Since CNAseg is designed for identification of CNAs in cancer genomes through use of a matched reference sample, the fact that it might have low discriminatory ability in the areas of segmental duplication should not impact the error rate of CNA detection.

3.3 CNAseg segmentation of COLO-829 genome and algorithm comparison

DWT-corrected read counts were segmented with CNAseg for CNA identification. Segmentation results from two CNA detection methods were used for comparative purpose. First, the segmentation results reported by cnv-seq (Xie *et al*., 2009) were used. Second, the RDXplorer package for detection of CNV events (Yoon *et al*., 2009) was independently applied to the COLO-829 genome. CNAseg identified 182 CNA events, cnv-seq reported 854 CNAs and RDXplorer produced 920 CNAs. As always with algorithmic comparison on real biological datasets, it is hard to assess the error rate considering that many CNAs were not independently validated. Among PCR-confirmed structural rearrangements were 25 deletions, which we have used to assess sensitivity of the method. We found

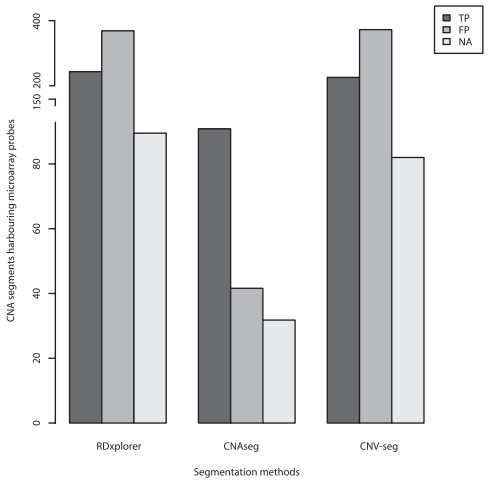

Fig. 4. Assessment of error rate in CNA regions using Affymetrix SNP array probes. Concordance of log ratio values with copy number calls was tested for three different methods. TP (true positive)—concordance in sequence and array-based copy number calls; FP (false positive)—discrepancy in sequence and array-based copy number calls; NA (information not available)—information cannot be accurately assessed (<2 probes spanning CNA region).

that all methods identified most of them (20, 21, 21, respectively, of which 19 were shared), suggesting that SV-supported deletions can be identified with high accuracy.

The precision of the algorithm is much harder to evaluate when only partial confirmation is available, therefore we used indirect evidence to approximate the false positive rate of the methods. First, data from Affymetrix SNP 6.0 arrays were used to confirm independently copy number calls of the sequence segmentation algorithms. The percentage of CNAs that harboured microarray probes was 76% (138/182) for HMMseq, 67% (616/920) for RDXplorer and 69% (589/854) for CNV-seq. Taking the means of absolute log ratios of probe intensities falling into CNA regions and setting a threshold of 0.2 for calling true positive CNAs (to ensure high stringency in selecting putatively false positive CNAs), we found that CNAseg has a much lower proportion of putatively false positive hits (Fisher's exact test *P*-values of 0.023) (Fig. 4).

Features of the alignment scores from paired-end read data can lend additional support to the validity of CNA calls. ELAND provides two alignment scores (single and paired alignment) to reflect confidence in alignment of each read independently and as a pair. Mismatched bases, reduced quality bases and multiple mappings downweight the single alignment score. A pair that maps with a size and/or orientation different from expected will have a zero paired alignment score. Misalignment can arise from the presence of short 2–10 bp indels or clusters of point mutations, which impede alignment of some reads. Consequently, such regions will contain reads with low single alignment score due to indels or mutation clusters. On the other hand, reads spanning genuine breakpoints around the areas of losses and gains will have anomalous insert sizes, which will be reflected in reduced paired alignment scores. Based on these two scores, we devised the following strategy for assessing the propensity of CNAs being false positives:

- For all reads mapped within 300 bp of a breakpoint, we calculate the percentage of reads with zero single alignment score and independently the percentage of reads with zero paired alignment score.
- Next, we select 100 regions with clusters of point mutations and/or indels, generated by the CASAVA (Consensus Assessment of Sequence and Variation) pipeline (Illumina *et al*., 2009) and compute the same statistics for reads in 5 kb intervals.

Analysis of single to paired alignment score ratios from these indel/mutation enriched reads showed that on average they are much higher for RDXplorer than the ratios obtained from 100 5 kb regions selected at random (Supplementary Fig. S7). Although means of the score ratios for reads aligned to CNAseg and RDxplorer breakpoints were both smaller than for reads aligned to validated SV breakpoints, the significance level was much higher for RDxplorer than CNAseg (*t*-test *P*-values 2.2×10^{-9} and 4.3×10^{-4}, respectively). We used the 95th percentile from indel/mutation enriched ratios as a threshold for calling false positive CNAs. Since our definition of false positive rate is not based on biological validation, the high percentile cut-off helped us ensure higher accuracy in identifying potentially erroneous CNA calls.

3.4 Simulation strategy for evaluating algorithm performance

Simulated data provide a *de facto* standard for comparing different algorithms in the absence of extensively validated biological datasets. Simulation of cancer genome sequencing data is exacerbated by multiple confounding factors that can alter read depth in regions of constant copy number. Such situations give rise to a mixture of over and underdispersed count distributions, which are difficult to simulate *de novo*. Instead, we utilized read depth in the normal COLO-829BL sample to guide our simulation process as follows:

(1) Duplicate pre-processed read counts from the COLO-829BL data, and let the duplicates form vectors $\mathcal{C}$ of pseudo-cancer and $\mathcal{N}$ of pseudo-normal genomes.

(2) Estimate the mean and variance of CNAseg windows along 1 kb genome intervals of each vector $\mathcal{C}$ and $\mathcal{N}$ and derive corresponding mean and variance sets.

(3) Add random negative binomial noise to each CNAseg window of $\mathcal{C}$ and $\mathcal{N}$ using previously derived mean and variance values amounting to 30% of read depth.

(4) Simulate different CNA events in the pseudo-cancer dataset by altering the following parameters:
   - The length of CNA events, varied from 1 kb to 100 kb;
   - CNA types, including heterozygous deletions, complete losses and 1, 2 and 3 copy number gains.

(5) In addition to the actual read depth reported for COLO-829BL ($\sim 32\times$), simulate read depths of $22\times$ and $9\times$ by randomly removing a proportion of reads.

(6) Spike in CNAs at predetermined positions of the pseudo-cancer genome $\mathcal{C}$.

Varying the amount of noise allows us to explore robustness of the algorithm to outliers. By adding noise amounting to 30% of

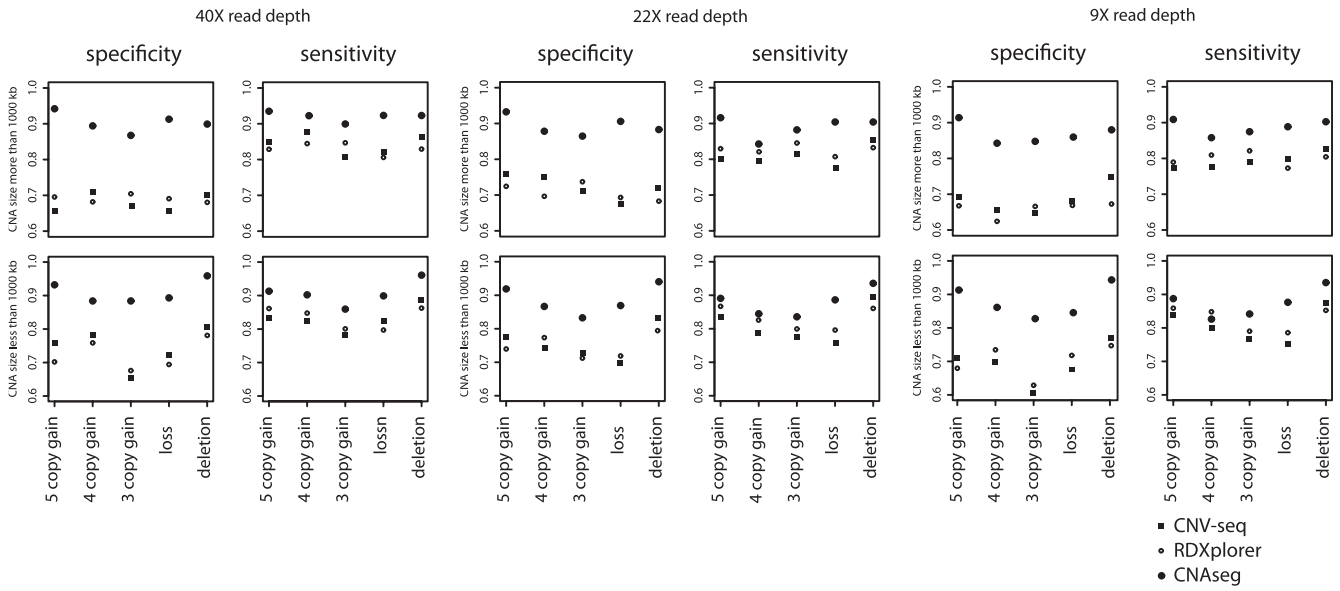

Fig. 5. Error rate comparison of CNAseg, RDXplorer and CNV-seq on the simulated dataset. The sensitivity and specificity estimates are provided for different values of read depth. An overlap of five windows (250 bp with the default window size) was allowed to call CNA a true positive.

read depth, we ensure that it remains relatively close to the values necessary for one copy loss or gain (40–50%), thereby ensuring conservative testing of the false positive error rate.

Testing the algorithm on the simulated dataset shows that CNAseg provides high sensitivity and specificity at all read depths (Fig. 5). In contrast, both RDXplorer and CNV-seq had much lower specificity estimates due to a large number of reported segments. Further results of algorithm comparison are provided in the Supplementary Material (section 'Notes on the comparison of segmentation algorithms', Supplementary Fig. S5). We attribute the higher number of reported CNAs and false positive error rate for RDXplorer and CNV-seq to the fact that both methods attempt to detect CNVs with high sensitivity by focusing on per-window-based CNV detection. Although this strategy might be beneficial when calling CNVs, it is less appropriate for identifying CNAs that have a much wider size distribution. For example, the median length of focal CNAs in a survey of 3131 cancer samples was reported to be 1.8 Mb (range of 0.5 kb–85 Mb), much larger than the average CNV size (Beroukhim *et al*., 2010).

3.5 Properties of flow cell variation

Successful application of CNAseg requires good empirical understanding of the flowcell-dependent read depth variation. We first explored the sequence properties of the reads from different flowcells to see if they can explain flowcell-to-flowcell read depth variability. One such property is GC content. Different factors during library preparation and sequencing may alter the relationship between read depth and local GC content and influence the patterns of read depth variability along the genome. The correlation between GC content and read depth has been described before (Dohm *et al*., 2008; Hillier *et al*., 2008); less investigated are observations that the actual value of the correlation coefficient and its sign may vary between library preparations (Chiang *et al*., 2009). We found that the GC content-read depth correlation varies between distinct flow cells. When comparing regression coefficients from GC content to read depth regression for different flow cells, we found that the variance (for the set of regression coefficients) is higher than when reads were assigned to flow cells randomly (P-value = 0.083 from the F-test to compare two variances), suggesting that flow cell loading and PCR amplifications may also introduce slight variation into the final read depth estimates (Fig. 6). Such variations, although not statistically significant, can be high enough to produce significant test statistics during comparison of read counts from intra-sample flowcells.

4 DISCUSSION

Here, we present a novel framework, CNAseg, for identification of copy number aberrations in cancer from second-generation sequencing data. It utilizes read depth variability between different flowcells from cancer and matched normal samples to control the false positive rate. By comparing to alternative segmentation methods, we showed that CNAseg significantly increases the precision of CNA detection at a comparable sensitivity. Although specific to Illumina reads, this general approach could readily be extended to other sequencing technologies.

Here, we use flowcell-to-flowcell variability to disentangle genuine CNA events from noise, but in principle the CNAseg framework can be modified to allow for other sources of variation. Now that a human genome can be sequenced to high coverage on one flowcell, lane-to-lane variability could be used instead for identifying technical variation. Another important future extension to the CNAseg framework is incorporation of explicit modelling of normal tissue admixture in cancer samples and heterogeneity of cancer cells in a tumour. Both extensions will be necessary for making accurate copy number calls and understanding tumour evolution.

Compared with microarrays, second-generation sequencing provides much higher resolution for detecting CNA events. In

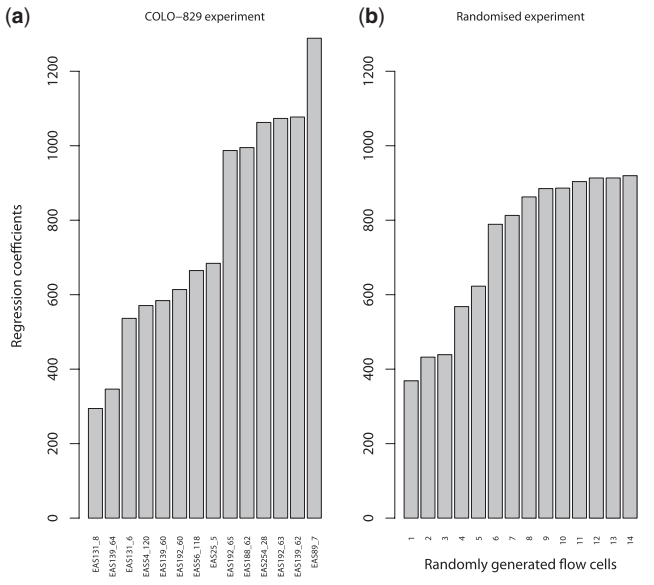

Fig. 6. (**a**) Magnitude of regression coefficients between read depth for 5000 kb window derived independently for reads from each flow cell and GC content. (**b**) Alternative coefficients when reads are assigned randomly to 'virtual' flow cells, but with preservation of the total number of reads from each flow cell.

addition, sequencing using paired-end reads gives more information about molecular mechanisms leading to copy number changes. A number of algorithms utilize anomalous insert sizes and orientation of PEM reads to identify structural variation (SV) in normal individuals (Chen *et al*., 2009; Chiang *et al*., 2009; Lee *et al*., 2008, 2009; Ye *et al*., 2009). Detection of SV abnormalities in cancer is an important topic in its own right, as positioning of breakpoints can aid detection of fusion genes and suggest SV-based mechanism of gene inactivation (Bignell *et al*., 2007; Campbell *et al*., 2008; Hampton *et al*., 2007; Stephens *et al*., 2009; Zhao *et al*., 2009). Our approach uses read depth to determine copy number changes in cancer and complements SV-detecting approaches.

The problem of using SV as the sole source of information in detecting CNAs is compounded by the large sizes of many CNAs compared with CNVs, and also by the variety of molecular mechanisms (and hence PEM properties) that can lead to copy number changes with complex breakpoints. Consequently, the placement of breakpoints using SV data will not comprehensively segment cancer genome. However, SV information is poised to improve detection of CNA events once better understanding of mechanisms behind genomic instability in cancer becomes available. For example, knowledge of SVs can be used to build a heterogeneous HMM where the probability of transitions between states is influenced by SV locations.

ACKNOWLEDGEMENT

We acknowledge the assistance of Dr Phil Stephens of The Wellcome Trust Sanger Institute, UK, in providing the raw COLO-829 Affymetix SNP 6.0 array data.

Funding: Cancer Research UK, Illumina Inc.

Conflict of Interest: All authors at Illumina (see affiliations) are employees of Illumina Inc., a public company that develops and markets systems for genetic analysis.

REFERENCES

Alkan,F. *et al.* (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genet.*, **41**, 1061–1067.

Bentley,D. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Beroukhim, R. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.

Bignell,G. *et al.* (2007) Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res.*, **17**, 1296–1303.

Calinski,R. *et al.* (1974) A dendrite method for cluster analysis. *Commun. Stat.*, **3**, 1–27.

Campbell,P. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.

Chen,K. *et al.* (2009) BreakDancer: an algorithm for high resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.

Chiang,D. *et al.* (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods*, **6**, 99–103.

Cox,A. (2006) Multiple inexact pattern matching. *European Patent*, **EP1704506**.

Dohm, J. *et al.* (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.

Hampton,O. *et al.* (2007) A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome Res.*, **19**, 167–177.

Hillier,L. *et al.* (2008) Whole-genome sequencing and variant discovery in C. elegans. *Nat. Methods*, **5**, 183–188.

Hormozdiari,F. *et al.* (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.*, **19**, 1270–1278.

Illumina LTD. (2009) Complete Secondary Analysis Workflow for the Genome Analyzer. *Technical Note: Illumina Systems and Software*. Available at http://www.illumina.com/Documents/products/technotes/technote_casava_secondary_analysis.pdf7 (last accessed date November 1, 2010).

Karlis,D. and Ntzoufras,I. (2005) Bayesian analysis of the differences of count data. *Stat. Med.*, **25**, 1885–1905.

Lee,S. *et al.* (2008) A robust framework for detecting structural variations in a genome. *Bioinformatics*, **24**, i59–i67.

Lee,S. *et al.* (2009) MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods*, **6**, 473–474.

Li,H. *et al.* (2009a) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*,**24**, 2078–2079.

Li,H. *et al.* (2009b) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Milligan,G. *et al.* (1985) An examination of procedures of determining the number of clusters in a data set. *Psychometrika*, **448**, 159–179.

Nason,G. (2005) *Wavelet Methods in Statistics with R (Use R)*. Springer, New York, USA.

Percival,D. and Walden, A. (2005) *Wavelet Methods for Time Series Analysis*. Cambridge University Press, New York, USA.

Pinkel,D. and Albertson,D. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.*, (Suppl. 1), 11–17.

Pleasance,E. *et al.* (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**, 191–196.

Santarius,T. *et al.* (2010) A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer*, **10**, 59–64.

Shendure,J. *et al.* (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.

Skellam,J. (1946) The frequency distribution of the difference between two Poisson variates belonging to different populations. *J. R. Stat. Soc. Ser. A*, **109**, 296.

Stephens,P. *et al.* (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **7276**, 1005–1010.

Stratton,M. *et al.* (2009) The cancer genome. *Nature*, **7239**, 719–724.

Willenbrock,H. and Fridlyand,J. (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, **21**, 4084–4091.

Xie,C. *et al.* (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, **10**, 80.

Ye,K. *et al.* (2009) Pindel: a pattern growth approach to detect breakpoints of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.

Yoon,S. *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592 .

Zhao,Q. *et al.* (2009) A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Proc. Natl Acad. Sci. USA*, **106**, 1886–1891.