BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 4 2012, pages 470–478
doi:10.1093/bioinformatics/btr707

Genome analysis

Advance Access publication December 23, 2011

Read count approach for DNA copy number variants detection

Alberto Magi[1,*], Lorenzo Tattini[2], Tommaso Pippucci[3], Francesca Torricelli[4] and Matteo Benelli[2,4,5]

[1]Faculty of Medicine, [2]Center for the Study of Complex Dynamics (CSDC), University of Florence, Florence 50019, [3]Department of Gyneacological, Obstetric and Paediatric Sciences, Medical Genetics Unit, University of Bologna, Bologna 40138 and [4]Laboratory Department, Diagnostic Genetic Unit, Careggi Hospital, Florence 5014 and [5]I.N.F.N, Sezione di Firenze, Florence 50100, Italy

Associate Editor: Alex Bateman

ABSTRACT

Motivation: The advent of high-throughput sequencing technologies is revolutionizing our ability in discovering and genotyping DNA copy number variants (CNVs). Read count-based approaches are able to detect CNV regions with an unprecedented resolution. Although this computational strategy has been recently introduced in literature, much work has been already done for the preparation, normalization and analysis of this kind of data.

Results: Here we face the many aspects that cover the detection of CNVs by using read count approach. We first study the characteristics and systematic biases of read count distributions, focusing on the normalization methods designed for removing these biases. Subsequently, we compare the algorithms designed to detect the boundaries of CNVs and we investigate the ability of read count data to predict the exact number of DNA copy. Finally, we review the tools publicly available for analysing read count data. To better understand the state of the art of read count approaches, we compare the performance of the three most widely used sequencing technologies (Illumina Genome Analyzer, Roche 454 and Life Technologies SOLiD) in all the analyses that we perform.

Contact: albertomagi@gmail.com

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

Received on June 30, 2011; revised on October 18, 2011; accepted on December 19, 2011

1 INTRODUCTION

Human genomes are characterized by genetic variants that range from the single base pair to large chromosomal events. Recent studies have clearly shown that human genomes differ more as a consequence of structural variants than of single-base pair differences (Conrad *et al*., 2010; Durbin *et al*., 2010; Iafrate *et al*., 2004; Kidd *et al*., 2008; McCarroll *et al*., 2008; Pang *et al*., 2010; Redon *et al*., 2006; Sebat *et al*., 2004; Tuzun *et al*., 2005). Structural variants (SVs) are operationally defined as genomic events >50 bp (Alkan *et al*., 2011) that include copy number variants (CNVs) and balanced rearrangements such as inversions and translocations. With the sequencing of human genomes now becoming routine, the challenge is to discover the full extent of structural variations and understand its effect on human diseases, complex traits and evolution. The last few years have seen the emergence of several high-throughput sequencing (HTS) platforms that are based on various implementations of cyclic-array sequencing (Bentley *et al*., 2008; McKernan *et al*., 2009; Wheeler *et al*., 2008). The commercial products that are based on this sequencing technology include the Roche's 454, the Illumina's Genome Analyzer (GA), and the Life Technologies's (LT) SOLiD. Although these platforms are quite diverse in sequencing biochemistry as well as in how the array is generated, all of them allow to sequence millions of short sequences (reads) simultaneously and are able to sequence a full human genome per week at a cost 200-fold less than previous methods.

The advent of HTS platforms has opened many opportunities for the study of genomic variants. The first HTS-based approach to detect SVs was based on paired-end read mapping (PEM), which identifies insertions and deletions by comparing the distance between mapped read pairs to the average insert size of the genomic library. Although this method is able to identify deletions <1 kb with high sensitivity, it does not allow for the discovery of insertions larger than the average insert size of the library and of the exact borders of SVs in complex genomic regions rich in segmental duplication (Dalca *et al*., 2010; Medvedev *et al*., 2009). An altenative HTS-based approach is based on split-read (SR) methods that allows to detect deletions and small insertions on the basis of a split sequence-read signature: the alignment to the genome is broken and a continuous stretch of gaps in the read indicates a deletion or in the reference indicates an insertion. Although SRs approach can be devised to detect a wide range of SV classes with exact breakpoint resolution, it is currently reliable only in the unique regions of the genome. In this scenario, a very promising approach for the identification of SVs using HTS technologies consists in measuring the number of reads aligned to the human reference genome (Dalca *et al*., 2010). Assuming the sequencing process is uniform, the number of reads mapping to a region is expected to be proportional to the number of times the region appears in the DNA sample. Following this assumption, the copy number of any genomic region can be estimated by counting the number of reads [read counts (RCs)] aligned to that particular region. Campbell *et al*. (2008) and Chiang *et al*. (2009) were the firsts to use this approach to detect copy number alterations between tumour and healthy samples of the same individual, while Yoon *et al*. (2009) proposed to use RC data to look for genomic regions that differ in copy number between normal individuals of the 1000 Genomes Project (Durbin *et al*., 2010). At present, few algorithms for RC analysis have been packaged

*To whom correspondence should be addressed.

into pipelines and are publicly available, including RDxplorer (Yoon *et al*., 2009), ReadDepth (Miller *et al*., 2011), CNAseg (Ivakhno *et al*., 2010), CNV-seq (Xie and Tammi, 2009), JointSLM (Magi *et al*., 2011) and CNVnator (Abyzov *et al*., 2011) (see Table 1 and Supplementary Material). The analysis pipeline implemented in these packages for discovering CNVs is conceptually derived from array-CGH (aCGH) data analysis and can be divided into four fundamental steps:

- RC data preparation.
- Data normalization.
- CNV regions identification.
- Copy number estimation.

Data preparation consists in filtering and counting the number of mapped reads in non-overlapping genomic windows of length W. Once the RCs have been estimated, the first transformation applied to the data, referred to as normalization, adjusts the individual RC to appropriately mitigate systematic biases so that meaningful biological comparisons can be made. Normalized RCs are then sorted according to genomic position and statistical methods are applied to detect the boundaries of the regions with changed copy number. The last step of the analysis pipeline consists in estimating the DNA copy number of each region within breakpoints. In the present article, we face the many aspects that cover the detection of CNVs by using RCs approach. We study all the steps necessary to infer the copy number of a genomic segment: RC estimation, RC normalization, CNV regions detection and copy number estimation. To better understand the state of the art of RC approach, we compare the performance of the three most used sequencing technologies (Illumina GA, Roche 454 and Life Technologies SOLiD) by using both high (20–40×) and low coverage (4–6×) sequencing data generated by the 1000 genomes project consortium (see Table 2 and Supplementary Material for more details).

2 METHODS

2.1 Data preparation

RC method belongs to the category of resequencing approaches, and the first fundamental step of this analysis consists in mapping the set of short reads against a reference genome by means of short read aligners (see Supplementary Material). Once short reads have been aligned to the reference genome, we need to perform a series of preparation steps before RC estimation. These steps include:

- Removal of duplicated sequences.
- Removal or flagging of sequences with low mapping quality (MQ).
- Choice of the best window/bin size.

The main purpose of removing duplicates is to mitigate the effects of PCR amplification bias introduced during library construction. All the analyses performed in this article have been made on aligned data with duplicated reads removed by means of rmdup command of samtools (Li *et al*., 2009) (Supplementary Material). After duplicated reads removal, low MQ sequences need to be taken into consideration: sequences with low MQ score usually fall in repetitive regions of the reference genome or have low base quality. For these reasons, Magi *et al*. (2011) and Yoon *et al*. (2009) removed all the reads with MQ < 30. Conversely, Abyzov *et al*. (2011) used the reads with MQ score equal to zero to classify CNVs called in duplicated or retrotransposon regions of the reference genome. Finally, the best window/bin size needs to be estimated. Yoon *et al*. (2009) and Magi *et al*. (2011) chose a window size of 100 bp for the high coverage data of the 1000 genomes project (20–40× coverage) because a larger window size would provide less precision in defining the breakpoints of CNVs and because at 30× coverage, the distribution of RCs of 100 bp windows are well approximated by a normal distribution, while RCs in smaller window sizes are not. Ivakhno *et al*. (2010) used a bin size of 50 bp for the analysis of the COLO-829 malignant melanoma cell line sequenced at 40× coverage. Abyzov *et al*. (2011) found that the optimal bin size, and thus breakpoint resolution accuracy, scales inversely with the coverage, resulting in ∼100 bp bins for 20–30× coverage, ∼500 bp bins for 4–6× coverage and ∼30 bp bins for 100× coverage. At present, only two papers have introduced a method to automatically estimate the best window size. Miller *et al*. (2011) propose to estimate the best window size by modelling RCs by means of a negative binomial distribution. They generate a negative binomial distribution with mean $\mu=\lambda$ (with $\lambda=N*W/G$, where N is the total number of read, W is the bin size and G is the genome size), and size parameter $\theta=\lambda/(i-1)$, where i is the index of dispersion. Then they generate distributions using the expected number of reads with copy number of one, two and three, and choose a threshold value for gains and losses that minimizes the number of bins that are misclassified. The FDR can then be calculated as the number of misclassified bins divided by the total number of bins. Xie and Tammi (2009) calculate the best possible resolution (i.e. the best possible window size) according to a preset value for significance level p and CNV detection threshold r (Supplementary Material).

2.2 RC biases and normalization

RCs data are affected by two main sources of bias: the local GC content and the genomic mappability. The correlation between read coverage and DNA

Table 1. Summary of the publicly available tools for the analysis of RC data (see Supplementary Materials for more details)

Tool	Reference	Platform	Input	Sample	Output
RDxplorer	Yoon *et al*. (2009)	R/Java	.bam	One sample	P/T
ReadDepth	Miller *et al*. (2011)	R	.bam	One sample	T
CNAseg	Ivakhno *et al*. (2010)	R	.bam/RC data	Two sample	T
CNV-seq	Xie and Tammi (2009)	R	tab-delimited	Two sample	P
JointSLM	Magi *et al*. (2011)	R	RC data	One sample	P/T
CNVnator	Abyzov *et al*. (2011)	C++	.bam	One sample	T/V

P, output as a plot of the detected alterations; T, output as a tab-delimited file with all the detected variants; V, Command visualization.

Table 2. Summary of the experiments used in the analyses

Sample	Platform	Read length	Milions of reads	Coverage
NA19240	Illumina GA	35	2738.03	39.1×
	Roche 454	200	45.05	3.7×
	LT SOLID	25	2550.73	2.6×
NA12878	Illumina GA	35	2510.6	35.8×
	Roche 454	200	187.9	15.3×
	LT SOLID	25	1775.37	18.1×
NA11830	Illumina GA	35	149.96	2.1×
	LT SOLiD	25	269.17	2.7×
NA11840	LT SOLiD	25	316.1	3.2×
	Roche 454	200	17.48	1.4×
NA12043	Illumina GA	35	163.25	2.3×
	Roche 454	200	13.44	1.1×

All the data were first used by Durbin *et al.* (2010).

GC content has been reported in several papers: Harismendy *et al.* (2009) analysed human sequences generated by the Roche 454, Illumina GA and the LT SOLiD technologies for the same 260 kb in four individuals concluding that read depth of coverage decreases with increasing AT content for all the three platforms. Similar results were found by Dohm *et al.* (2008) and Hillier *et al.* (2008). Mappability bias (Miller *et al.*, 2011) is due to the fact that the genome contains many repetitive elements and aligning reads to these positions leads to ambiguous mapping. In order to minimize the effect of these sources of variation and make data comparable within and between samples, RCs need to be normalized. At present, there are two approaches for correcting RC data for sequencing biases due to local GC content. Chiang *et al.* (2009) proposed to mitigate the dependence between local GC content and RC by using the ratio of the number of reads in tumour DNA and its paired normal DNA, processed at the same time. Yoon *et al.* (2009) proposed to adjust the RCs by using the observed deviation in coverage for a given GC percentage. In practice, for all the GC percentages (0, 1, 2, ... , 100%) they determined the deviation of coverage from the genome average and then corrected each RC according to the following formula:

$$\overline{\mathrm{RC}_i}=\mathrm{RC}_i\cdot\frac{m}{m_{\mathrm{GC}}}, \tag{1}$$

where RC_i are read counts of the i-th window, m_{GC} is the median RC of all the windows that have the same GC percentage as the i-th window, and m is the overall median of all the windows.

Mappability normalization has been faced in two different papers. Miller *et al.* (2011) proposed to correct for mappability by multiplying the number of reads in a given bin by the inverse of the percent mappability in that region, whrease Ivakhno *et al.* (2010) proposed to use an undecimated discrete wavelet transform (DWT) to smooth RCs in the regions of low alignability. Here we propose to correct RC for mappability bias with a novel normalization scheme inspired by the GC content normalization of Yoon *et al.* (2009): RCs are corrected by using the observed deviation in coverage for a given mappability score. Each RC is corrected by the following formula:

$$\overline{RC_i}=\mathrm{RC}_i\cdot\frac{m}{m_{\mathrm{MAP}}}. \tag{2}$$

where RC_i are RCs of the i-th window, m_{MAP} is the median RC of all the windows that have the same mappability score as the i-th window, and m is the overall median of all the windows.

2.3 CNVs detection algorithms

Once the RCs have been corrected for local GC content and mappability, the data that we obtain are mathematically very similar to the signal obtained from aCGH experiments. Deletions or duplications are identified as decrease or increase of RC across multiple consecutive windows. Moreover, like aCGH data, RC signals are affected by noise caused by mapping errors and random fluctuations in genome coverage. For these reasons, the events in RC data can be detected using the same algorithmic approaches that have been used for aCGH data. At present, few statistical methods have been developed and tested for the detection of CNVs from RC data. Some of these algorithms come from microarray literature while others have been tailored for this kind of data. Campbell *et al.* (2008) and Miller *et al.* (2011) used the circular binary segmentation algorithm (CBS) (Olshen *et al.*, 2005), originally developed for genomic hybridization microarray data, and both applied it to sequencing data generated by the Illumina GA platform. Magi *et al.* (2011) extended the shifting level model (SLM) (Magi *et al.*, 2010) algorithm to detect recurrent CNVs across multiple samples and tested its performance on Illumina high coverage data from 1000 genomes project samples. Abyzov *et al.* (2011) exploited a mean-shift algorithm (MSB), previously applied to the analysis of aCGH data, to partition RC signals to the end of detecting CNVs. MSB was tested on data produced by Illumina and SOLiD platforms. Yoon *et al.* (2009) developed a new method based on significance testing (EWT) that works on intervals of data points: EWT searches the entire genome for specific classes of small events that meet criteria of statistical significance, and then clusters of small events are grouped into larger events. EWT was applied to high coverage Illumina data produced by the 1000 genome project consortium. Xie and Tammi (2009) used a sliding window approach, named CNV-seq, to analyse the ratios between RCs from two individuals (Normal and Tumour). The observed ratios are assessed by the computation of the probability of a random occurrence, given no CNV. Xie and Tammi (2009) tested their algorithm on shotgun sequencing data of Dr Craig Venter and Dr James D. Watson generated by Sanger and 454 platforms, respectively. Ivakhno *et al.* (2010) introduced CNASeg, an HMM-based algorithm to segment the RC data, followed by a segment merging step. The CNASeg method was originally applied to cancer sequencing data generated by the Illumina GA platform. While the first four statistical methods require only one sample at once, the approaches of Xie and Tammi (2009) and Ivakhno *et al.* (2010) need the sequencing data of two samples. The EWT and CNV-seq methods are sliding window approaches that converts RC data into a statistic (t-statistic for CNV-seq and Z-score for EWT) and infer altered regions by using the distribution of that statistic. Conversely, the CBS, SLM, MSB and CNASeg algorithms are segmentation methods that allow to split RC data into segments, each containing the same number of DNA copies. Segments with an altered DNA copy number are detected by means of a simple threshold method (Campbell *et al.*, 2008; Ivakhno *et al.*, 2010; Magi *et al.*, 2011; Miller *et al.*, 2011).

2.4 Copy number estimation

The statistical methods introduced in the previous section allow for the identification of genomic regions with altered DNA copy counts by detecting the border of consecutive windows with increased or reduced RCs. Once the limits of the altered region has been detected, the estimation of DNA copy number (genotyping) for those regions must be performed. Campbell *et al.* (2008), Yoon *et al.* (2009) and Magi *et al.* (2011) estimated DNA copy number by rounding the median of the RCs (normalized to copy number 2) of each detected region to the nearest integer, while Abyzov *et al.* (2011) assigned copy number to each genomic region by calculating its RC signal normalized to the genomic average for the region of the same length. These simple estimation strategies follow the assumption that the sequencing process is uniform and consequently the number of read that maps to a genomic region is expected to be proportional to the number of times the regions appears in the DNA sample: a genomic region that has been deleted (duplicated) will have less (more) reads mapping to it than a region not deleted (duplicated).

3 RESULTS

3.1 Data filtering and bin size estimation

To understand the effect of removing sequences with low MQ on RC distributions, we calculated the signal to noise ratio (SNR) for different MQ threshold and different windows size W (see Supplementary Material for more details) and the results are reported in Supplementary Figure S1. The SNR has been calculated by means of the following formula:

$$\mathrm{SNR} = \frac{m}{\sigma^2}, \qquad (3)$$

where m is the median value of the normalized RC of genomic regions predicted as two copies while σ^2 is the variance of the normalized RC of regions predicted to be one copy by McCarroll *et al.* (2008). The results of these analyses show that filtering out reads with low MQ slightly affects the SNR of RC data. These results suggest not to remove low MQ reads, since they can be used for subsequent analysis as in Abyzov *et al.* (2011).

In order to investigate the performance of the two bin size estimation methods proposed by Xie and Tammi (2009) and Miller *et al.* (2011), we simulated sequencing data for different coverage for the three HTS platforms (see Supplementary Material for more details). The results of these analyses are reported in Supplementary Figure S2. As expected, the larger is the coverage of the sequencing data and the smaller is the bin size predicted by the two methods. The method proposed by Xie and Tammi (2009) estimates bins of similar size for the Illumina and SOLiD platforms. This is due to the fact that the Xie and Tammi estimation procedure does not take into account the overdispersion of RC distributions. Conversely, the approach proposed by Miller *et al.* (2011) allows for a better estimation of the bin size for the three HTS technologies, taking advantage of the use of the index of dispersion. However, the bin sizes predicted by the two methods are not optimal: for Illumina platform, at 30× coverage, the Miller method estimate a bin size of 1000 bp, while at 5× coverage it estimates a bin size of 6600 bp. These estimates are at least 10 times higher than those reported by Yoon *et al.* (2009), Magi *et al.* (2011) and Abyzov *et al.* (2011): for instance, by using a bin size of 100 bp for 20–30× coverage data, the JointSLM and EWT algorithms were able to detect CNV regions as small as 500 bp with a true positive rate >0.8 and with a minor fraction of false positive events of this size (Magi *et al.*, 2011). The use of the bin sizes estimated by the method of Miller *et al.* (2011) (i.e. 1000 bp for high coverage) does not allow for the detection of CNVs as small as 500 bp leading to a loss of resolution and accuracy in the discovery of genomic variants.

3.2 RC distribution and biases

The detection of CNVs using RC analysis is based on the assumption that the reads are randomly and independently sampled from any location of the test genome with equal probability. Under this assumption, the distribution of the count of reads that map into a window of the reference genome should be Poissonian. However, Bentley *et al.* (2008) and Yoon *et al.* (2009) have previously reported that RCs by Illumina GA follow a Poisson distribution with a slight overdispersion. In order to study the properties of RC distribution, we analysed high and low coverage sequencing data generated by the 1000 genomes project consortium and we used different values of W for the three HTS platforms (see Supplementary Material for more details). The results of these analyses are summarized in Figure 1 and Supplementary Figure S3 and clearly show that RC data can be modelled by means of a negative binomial distribution. According with the results of Bentley *et al.* (2008) and Yoon *et al.* (2009), we found that RC distribution for Illumina and SOLiD platforms exhibit an index of dispersion (ratio between variance and mean) largely greater than one. Conversely, 454 platform produces the RC data distribution with the lower ratio between variance and mean (Supplementary Tables S1 and S2).

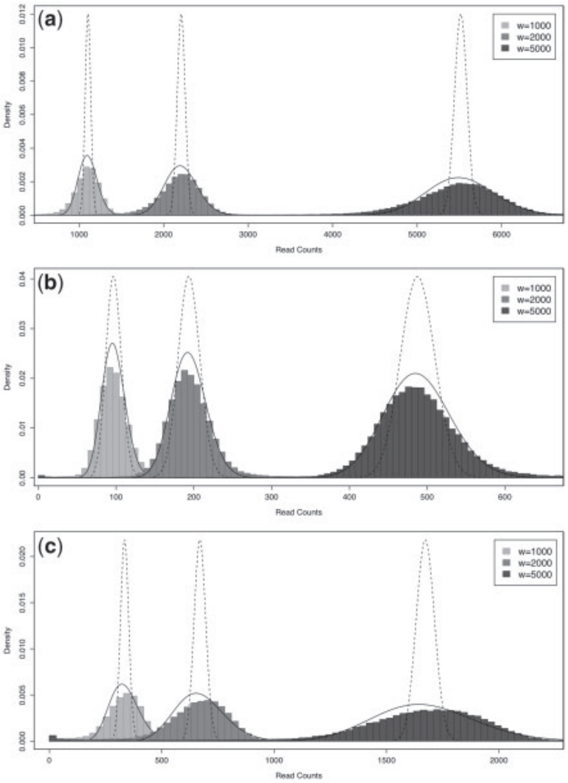

Fig. 1. RC distributions for the three HTS platforms at high-sequencing coverage. The plot reports the histogram of the RC distributions for the three HTS platforms [Illumina GA (**a**), Roche 454 (**b**) and LT SOLiD (**c**)] for three different window size ($W = 1000$ bp, $W = 2000$ bp and $W = 5000$ bp). For each window size is also reported the Poisson distribution (dashed line) with mean = μ and the negative binomial distribution (solid line) with mean = μ and variance = σ^2.

The overdispersion of RC data distributions can be accounted for to three main sources:

- The existence of genomic regions of duplications and deletions (CNVs).
- The correlation between read coverage and the DNA local GC content.
- The correlation between read coverage and the mappability (i.e. the inability to map reads into repetitive regions of the genome).

The effect of CNV regions on the distribution of RC data are reported in Figure 2a, d and g, in Supplementary Figure 4a,d and g and in Supplementary Tables S3 and S4. The current estimation of the

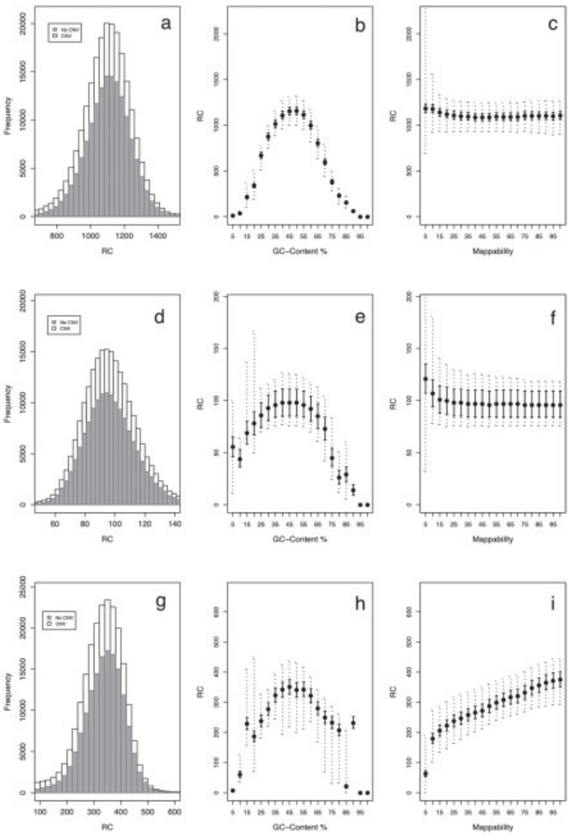

Fig. 2. Sources of Overdispersion of RC distribution for high sequencing coverage. (**a**, **d**, **g**) Histograms of the RC with (white bar) and without (grey bar) CNVs regions for Illumina, 454 and SOLiD platforms, respectively. (**b**, **e**, **h**) Correlation between RC data and GC content for the three HTS platforms (GA, 454 and SOLiD). (**c**, **f**, **i**) Correlation between RC data and region mappability for the three HTS platforms (GA, 454 and SOLiD). Upper border of the dashed lines represent the 90th percentile of the normalized RCs, while the lower border represent the 10th percentile. Upper border and lower border of the solid lines represent the 90th percentile and the 10th percentile, respectively, of the poisson distribution with mean = mean value of the RCs. To obtain the histograms without CNV (grey bar) of Figure 1a, d and g, we removed from RC data all the CNV regions that belong to the Database of Genomic Variants and the CNV regions previously identified by McCarrol *et al.*

fraction of the genome subject to variation of the copy number is ~3.7% (Conrad *et al.*, 2010). This means that a considerable amount of genomic regions contains an average number of reads smaller or greater than the global average number of reads. The removal of genomic regions with known CNVs reduces the index of dispersion of RC data distribution for all the three HTS platforms for high and low coverage data.

We investigated the relationship between RC and GC content (Supplementary Materials) for all the three HTS platforms and according with the results of Harismendy *et al.* (2009), Dohm *et al.* (2008) and Hillier *et al.* (2008) we observed that RC is maximum for values of GC content between 35% and 60% while it decreases at both extremes. In particular, we observed that GC content bias is larger for GA and SOLiD platforms, while it is smaller for the ROCHE data (see Figures 2b, e and h and Supplementary Fig. S4). This is confirmed by the statistics reported in Supplementary Tables S3 and S4, where the percent variation between the raw index of dispersion and the GC index of dispersion is much larger for GA and SOLiD data than for the Roche data. The analysis of regional mappability (Figure 2c, f and i and Supplementary Fig. S4) show a strong correlation between RC data and genome mappability: the RC distribution for high mappability score is closer to Poissonian than genomic regions with low mappability (low mappability regions show large RC overdispersion). Moreover, for GA and 454, the mappability has little effect on the mean number of aligned reads at each bin of mappability score, while, for SOLiD platform, the mappability strongly affect the RC mean value. Also these results are confirmed by the statistics reported in Supplementary Tables S3 and S4: the percent variation between the raw index of dispersion and the mappability index of dispersion is very large for the SOLiD platform (65% and 83% for high coverage and about 60% for low coverage) while it is comparable to other source of variation for GA and 454 platforms. This can also explain why SOLiD platform shows the highest value of the index of dispersion.

3.3 RC data normalization

In order to evaluate the performance of the five normalization approaches described in Section 2, we applied them to the high and low coverage data generated by the 1000 genomes project consortium and the results of these comparisons are reported in Figure 3 and Supplementary Figure S5 for GC content and in Figure 4 and Supplementary Figure S6 for mappability. The ratio approach proposed by Chiang *et al.* (2009) (see Figure 3c, f and i and Supplementary Figure S5) is not able to remove the GC content effect for all the three HTS platforms, while it performs very well in mitigating the mappability bias also in the case of the SOLiD platform where the mappability effect is very strong (see Figure 4e, j and o and Supplementary Fig. S6). The GC content normalization approach proposed by Yoon *et al.* (2009) (see Figure 3b, e and h and Supplementary Fig. S5) is able to properly remove the GC content effect for all the three HTS platforms. The comparison between the other three mappability normalization methods clearly show that the approaches proposed by Miller *et al.* (2011) (see Figure 4c, h and m and Supplementary Fig. S6) and Ivakhno *et al.* (2010) (see Figure 4d, i and n and Supplementary Fig. S4) are not able to completely correct the non-linear bias produced by genomic regions with low mappability. These analyses also demonstrate that the approach introduced by Miller *et al.* (2011) generates additional biases and has the disadvantage that much data are discarded since RCs with extremely low mappability (< 25%) are filtered out to prevent overcorrection. The additional bias generated by the Miller *et al.* (2011) method is due to the assumption that RC is proportional to the percent of mappability: multiplying RC data by the inverse of percent mappability leads to overcorrection. Conversely, the mappability normalization scheme proposed in this article (Figure 4b, g and l and Supplementary Fig. S6) allows to correct this bias without filtering out RC data. Moreover, also in the case of the highly biased SOLiD data the median method permits correction of RC mean value. The results of all these analyses indicate that the normalization method by Yoon *et al.* (2009) and the median method are the best strategies to remove GC content and mappability biases, respectively. Moreover,

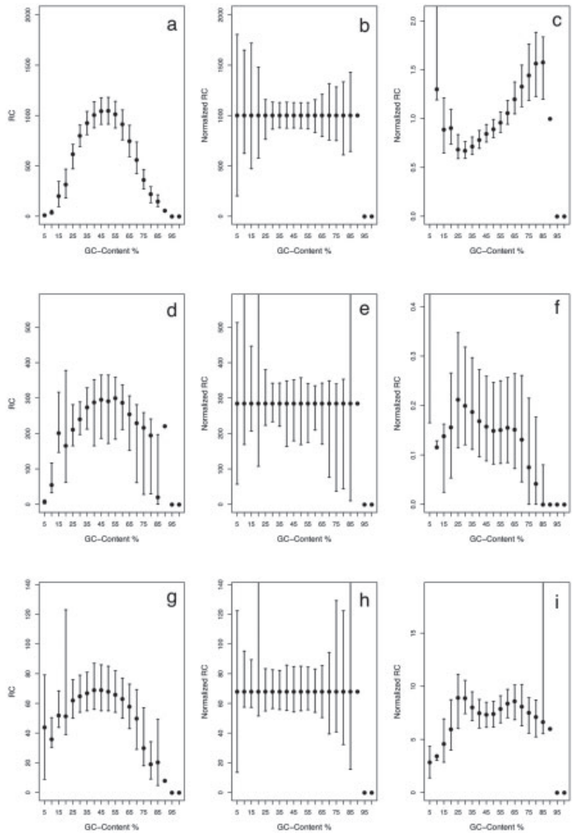

Fig. 3. Local GC content normalization methods. Effect of median normalization of Yoon *et al.* and ratio normalization of Chiang *et al.* on the local GC content for high sequencing coverage data. (**a**, **d**, **g**) Dependencies between RC data and local GC content for the three HTS platforms [GA (a), 454 (d), SOLiD (g)]. (**b**, **e**, **h**) Effect of the median method proposed by Yoon *et al.* for the removal of the local GC content bias for the three HTS platforms [GA (b), 454 (e), SOLiD (h)]. (**c**, **f**, **i**) Effect of ratio normalization proposed by Chiang *et al.* for the mitigation of the local GC content bias for the three HTS platforms [GA (c), 454 (f), SOLiD (i)]. Upper border of the dashed lines represent the 90th percentile of the normalized RCs, whereas the lower border represent the 10th percentile.

also when we need to compare the copy number alteration of two samples by means of a ratio approach (for example tumour DNA and its paired normal DNA), it is recommended to correct data with the Yoon *et al.* normalization scheme and the median mappability scheme.

3.4 CNV regions identification

To test the ability of different algorithms in detecting CNVs of different size, we made an intensive simulation based on synthetic chromosomes generated from RCs of the individuals NA12878 and NA19240 for high coverage sequencing data and the individuals NA11840, NA11830 and NA12043 for low coverage sequencing data. The principal aim of these simulations is to evaluate and compare the capability of each algorithm in detecting sudden shifts in the mean value of the signal as a function of the width of the shift. For this purpose, we have built a benchmark synthetic dataset generated by using GC- and mappability-adjusted RC data with the same normalization scheme: GC correction was performed by means of the Yoon *et al*. (2009) normalization scheme, while the mappability bias was corrected by means of the median normalization scheme. Each synthetic chromosome was generated by sampling RC data windows from genomic regions previously predicted as two-copy and one-copy by McCarroll *et al*. (2008) to simulate both normal copy count and altered regions (see Supplementary Material for more details). We compared the performance of the six algorithms described in Section 2: two sliding window methods (CNVseq, EWT) and four segmentation algorithms (CBS, SLM, MSB and the HMM of CNASeg). To evaluate the performance of the six algorithms, we used two different strategies. To test the capability of each algorithm in discovering CNVs, we used the approach previously used by Yoon *et al*. (2009) and Magi *et al*. (2011): a detected segment is considered a true positive (TP) if there is any overlap between the detected segment and the synthetic altered region, and is considered a false positive (FP) if there is no overlap with a synthetic altered region. To understand the accuracy of the six methods in detecting CNVs at the boundaries (breakpoints detection), we computed the receiver operating characteristic (ROC) curve as in Lai *et al*. (2010) and we calculated the area under the ROC curve (AUC). The results of all the simulations are summarized in Figure 5 and Supplementary Figure S7–S38. Globally, the algorithms that ensure the best results in terms of both sensitivity and specificity are the EWT and the SLM methods. The CBS and MSB algorithms obtain good results in detecting alterations made of a large number of windows ($N = 50$ and $N = 100$), while their performance reduces for alterations made of a small number of windows ($N = 5$ and $N = 10$). The HMM algorithm of the CNASeg package performs well on RCs generated from high coverage sequencing data while leads to modest results for low coverage data. The CNVseq method gives poor results for both high and low coverage data with all the bin size we simulated. The CNVs detection analyses (Supplementary Figures S31–S38) show that for high coverage data from GA platform the EWT and SLM algorithms are able to detect genomic alterations as small as 500 bp and 1 kb, respectively, with a TPR >0.8 and with a minor number FP events. On the same data, the CBS, MSB and CNASeg methods enable the detection of CNVs as small as 2–5 kb with a TPR >0.8 and small number of FP events. For low coverage data (Illumina GA platform), the EWT and SLM methods are able to discover genomic alterations as small as 5 kb, while CBS, MSB and CNASeg detect CNVs as small as 25 kb. The CNVseq algorithm identifies a very large number of FP events for both high and low coverage sequencing data making its use difficult for the detection of genomic regions involved in CNVs. The ROC curves reported in Supplementary Figures S11–S30 show that segmentation algorithms (SLM, CBS, MSB and the HMM of CNASeg) have low FPR at the expense of low TPR, while smoothing algorithms (EWT and CNVseq) have high TPR at the expense of high FPR. Moreover, as reported in Lai *et al*. (2010), ROC curves are informative in understanding how an algorithm performs in estimating the boundary of the altered region. When the algorithm over-estimates the boundary, FPR increases while TPR remains fixed. Conversely, when an algorithm under-estimates the boundary, TPR decreases while FPR remains fixed. Bearing this in mind, the ROC curves reported in Supplementary Figures S11–S30 suggest that segmentation methods (SLM, CBS, MSB and the HMM of CNASeg) tend to under-estimate the boundaries of CNV, while the EWT algorithm tend to over-estimate the boundaries of the CNV

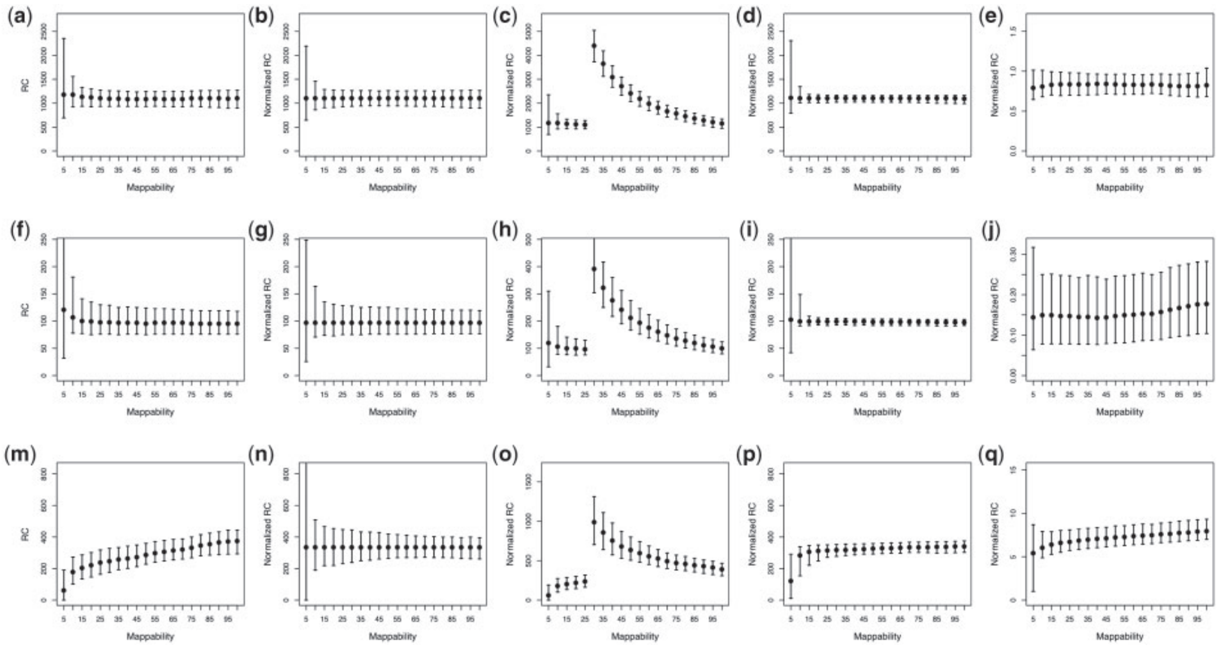

Fig. 4. Mappability normalization. Comparison between the four mappability normalization methods (median correction, Miller *et al.*, Ivakhno *et al.* and ratio normalization) on the high sequencing coverage data of the three HTS platforms. (**a**, **f**, **k**) Dependencies between RC data and region mappability for the three HTS platforms [GA (a), 454 (f), SOLiD (k)]. (**b**, **g**, **l**) Results of the median correction method for the removal of the region mappability bias for the three HTS platforms [GA (b), 454 (g), SOLiD (l)]. (**c**, **h**, **m**) Results of the Miller *et al.* correction method for the removal of the region mappability bias for the three HTS platforms [GA (c), 454 (h), SOLiD (m)]. (**d**, **i**, **n**) Results of the Ivakhno *et al.* wavelet-based correction method for the removal of the region mappability bias for the three HTS platforms [GA (d), 454 (i), SOLiD (n)]. (**e, j, o**) Results of the ratio correction method for the removal of the region mappability bias for the three HTS platforms [GA (e), 454 (j), SOLiD (o)]. Upper border of the dashed lines represent the 90th percentile of the normalized RCs, whereas the lower border represent the 10th percentile.

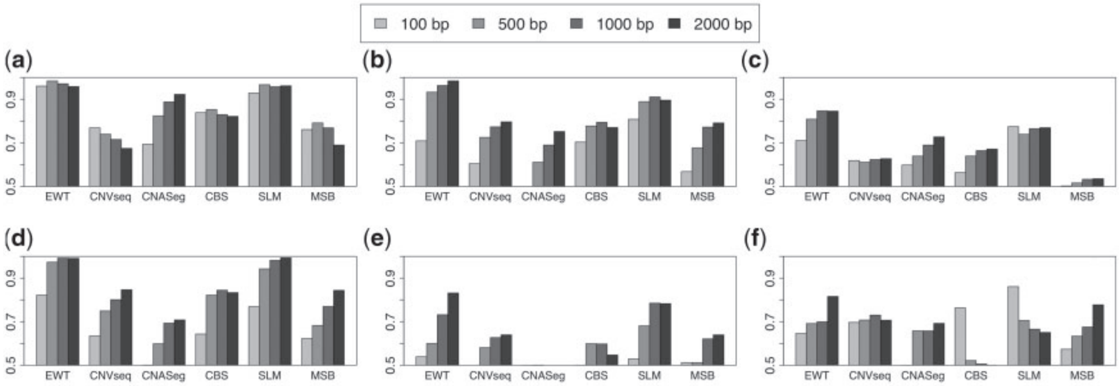

Fig. 5. Comparison between CNV detection algorithms. Six CNV detection algorithms (EWT, CNVseq, CNAseg, CBS, SLM and MSB) are compared in the analysis of synthetic chromosomes. Synthetic chromosomes are obtained from high and low coverage sequencing data of the three HTS platforms for bin size of different length ($W = 100$ bp, $W = 500$ bp, $W = 1000$ bp and $W = 2000$ bp). For each algorithm, the area under the curve is averaged across 1000 simulations. Each bar of the plot is obtained averaging the performance of each algorithm for all the alteration widths simulated ($N = 5$, $N = 10$, $N = 20$, $N = 50$, $N = 100$). (**a**) Illumina GA high coverage. (**b**) Roche 454 high coverage. (**c**) LT SOLiD high coverage. (**d**) Illumina GA low coverage. (**e**) Roche 454 low coverage. (**f**) LT SOLiD low coverage.

regions. According to the AUC results, the CNVseq algorithm obtain poor results in terms of both sensitivity and specificity. All these results reflect the algorithmic nature of each method: as reported in aCGH literature, smoothing approaches allow for the detection of small genomic events but with low breakpoint resolution, while segmentation strategies are more suited for the identification of larger events with high breakpoint resolution. The barplots of Figure 5 and Supplementary Figures S7–S10 show that the Illumina platform outperforms the other two sequencing technologies in the detection of alterations made of small number of windows for all the bin sizes we studied. Moreover, we found that the detection of DNA alterations in samples with low sequencing coverage is identical to analysing samples with high sequencing coverage if a bin size is larger. In particular, we found that the Illumina platform obtains high accuracy for bin size W = 100 bp at high coverage and W = 500 bp at low coverage, while the 454 and SOLiD platforms give good results for bin size W = 500 bp for high coverage and W = 1000 bp for low coverage.

3.5 Copy number estimation

To study the relationship between DNA copy number and RCs data, we examined several broad genomic regions that were previously reported to have copy numbers equal to 0, 1, 2, 3 and 4 by McCarroll *et al.* (2008) (Supplementary Material). We analysed RC data of these regions for different window sizes for the high and low coverage sequencing data generated with the three HTS platforms. The results of all these analyses are reported in Figure 6 (high coverage data) and Supplementary Figure S11 (low coverage data) and clearly reflect the overdispersion of the RC distributions generated by the three sequencing technologies. For GA and 454 platforms, we observed an excellent agreement between mean RCs and DNA copy number for high coverage data for all the bin sizes we analysed, while for SOLiD platform we found that RC are not well correlated with validated DNA copy number. For low coverage data, we found a smaller correlation coefficient for all the HTS platforms. However, also in this case GA and 454 technologies better predict the absolute value of DNA copy number.

4 DISCUSSION AND CONCLUSION

The use of RC of sequences aligned to a reference genome is, at present, the most powerful method to accurately predict absolute copy numbers of genomic regions. Although this computational strategy has been recently introduced in literature, much work has been already done for the preparation, normalization and analysis of this kind of data. Normalization methods allow to remove systematic biases due to local GC content and region mappability: the results reported here clearly show that the best way to remove local GC content bias is the Yoon *et al.* approach, while the best scheme to correct mappability bias is the median method proposed in the present article. CNVs detection algorithms can be exploited to detect CNVs with an unprecedented resolution that in the best case reaches the order of hundreds of base pair. In all the simulations, we performed we found that the EWT and SLM algorithms give the best results in terms of both sensitivity and specificity. The resolution of CNV detection can be improved by increasing the SNR of RC signals: reducing the sequencing error rate or increasing the coverage of the sequencing experiments will improve the performance of statistical methods in detecting small shifts in the signals. Automatic strategies for bin size calculation fail in estimating the optimal bin size whatever the coverage. Although the method proposed by Miller *et al.* models RC data by means of a negative binomial distribution, it overestimates the bin size leading to a loss of resolution accuracy. After an intensive simulation on synthetic data, we found that the best way to choose the optimal bin size is following the suggestion of Abyzov *et al.*: 100 bp window for 20–30× coverage, 500 bp window for 4–6× coverage and 50 bp window for 100× coverage. The comparison between the three HTS platforms clearly shows

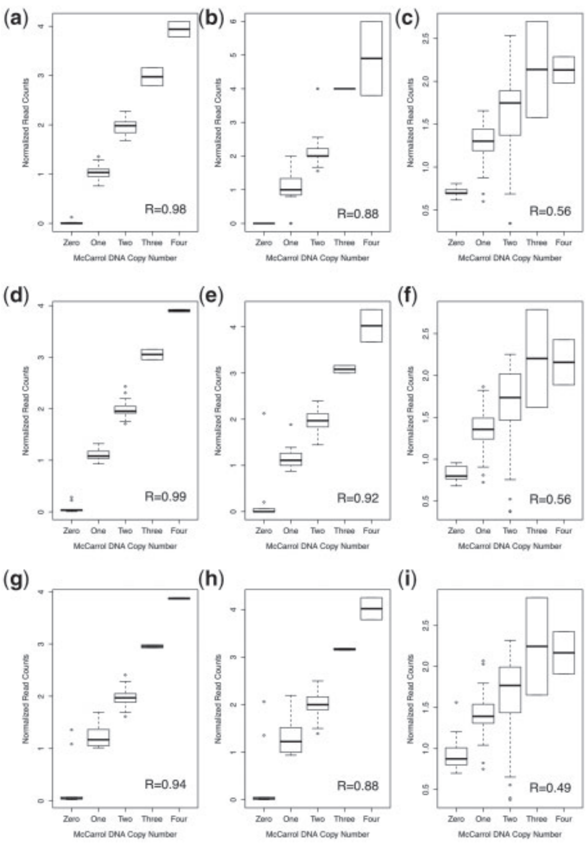

Fig. 6. Correlation between DNA copy number and normalized RC. DNA copy number of genomic regions previously genotyped by McCarroll *et al.* are compared with the normalized RCs of the high coverage sequencing data produced by the 1000 genomes project for the three HTS platforms. (**a**, **d**, **g**) GA platform. (**b**, **e**, **h**) 454 platform. (**c**, **f**, **i**) SOLiD platform. The analyses have been performed for three different window sizes W. (a–c) $W = 100$ bp. (d–f) $W = 1000$ bp. (g–i) $W = 2000$ bp. R is the Pearson's correlation coefficient.

that the Illumina platform allow to obtain the best level of accuracy and resolution in the discovery of genomic regions involved in CNV and the prediction of absolute DNA copy number. Although RC approach is the only sequencing-based method to accurately predict absolute DNA copy numbers, it has distinct advantages and disadvantages over other approaches in detecting certain classes of SVs and the breakpoint resolution is often poor with respect to other sequencing-based methods. The analyses employed with different approaches [PEM, SR (Mills *et al.*, 2006) and RC] on the same data in the framework of the 1000 Genomes Project show that RC-based approach can better ascertain CNVs in segmental duplication than PEM-based methods. Conversely, RC methods mostly miss CNVs consisting entirely of a single retrotransposon (LINE, SVA or HERV-K) that are easily detected by PEM and SR approaches. Additionally, RC analysis is not able to detect balanced rearrangements that can be instead discovered by PEM and split read methods. The comparison with the CNVs identified by microarray technologies shows that the great majority of the calls overlapped between the two methods (Alkan *et al.*, 2011; Magi *et al.*, 2011; Yoon *et al.*, 2009). Despite the great overlap between RC and microarray calls, RC is better than microarray at detecting smaller events and does not suffer from oversaturation at high copy counts, allowing a more accurate estimation of very high copy counts. While the sections discussed above describe the great progress achieved over the last 3 years in using RC data to discover CNVs, much work remains. When RC data are used to analyse tumour samples, statistical approach to infer copy number should take into account cellularity and tumoural heterogeneity and for this reason we would need a more sophisticated approach similar to CGHcall (van de Wiel *et al*., 2007) or FastCall (Benelli *et al*., 2010) instead of using the simple rounding to the closest integer. Finally, the breakpoint resolution of RC methods is often poor with respect to other sequencing-based approaches. However, given the approximate CNV breakpoint detected by an RC approach, the detection precision can be brought to 1 base resolution by refining the breakpoint by means of SRs techniques. The ultimate way to accurately detect all forms of genomic structural variants is *de novo* assembly (Levy *et al*., 2010; Li *et al*., 2011). Nevertheless, assembly approaches are still in their early stages and are capable to type structural variants only if the sequence reads are long and accurate enough to allow *de novo* assembly. Moreover, assembly algorithms have been shown to collapse in highly repeated and highly duplicated genomic regions (Alkan *et al*., 2011). Albeit assembly approaches show a lot of potential (facilitating the pair-wise genome comparison) their application as routine methods still need further efforts, in both computational and technological developments.

ACKNOWLEDGEMENT

We thank the 1000 Genomes Consortium that provided all the data used in this study (http://www.1000genomes.org).

Funding: This study was supported by the project “Gene expression profile and therapeutic implication in gastric cancer. From the clinical overview to the translational research” of the Istituto Toscano Tumori (ITT).

Conflict of Interest: none declared.

REFERENCES

Abyzov,A. *et al.* (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, **21**, 974–984.

Alkan,C. *et al.* (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet*., **12**, 363–376.

Benelli,M. *et al.* (2010) A very fast and accurate method for calling aberrations in array-CGH data. *Biostatistics*, **11**, 515–518.

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

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

Chiang,D.Y. *et al.* (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods*, **9**, 99–103.

Conrad,D.F. *et al*. (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.

Dalca,A.V. and Brudno,M. (2010) Genome variation discovery with high-throughput sequencing data. *Brief. Bioinform*, **11**, 3–14.

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

Durbin,R.M. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **7319**, 1061–1073.

Harismendy,O. *et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.*, **10**, R32.

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

Iafrate,A.J. *et al*. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.

Ivakhno,S. *et al.* (2010) CNAseg-a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics*, **26**, 3051–3058.

Kidd,J.M. *et al*. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64.

Lai,W.R.R. *et al.* (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array-CGH data. *Bioinformatics*, **21**, 3763–3770.

Levy,S. *et al*. (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.

Li,H. *et al.*; 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,Y. *et al*. (2011) Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nat. Biotechnol.*, **29** 723–730.

Magi,A. *et al.* (2010) A shifting level model algorithm that identifies aberrations in array-CGH data. *Biostatistics*, **11**, 265–280.

Magi,A. *et al.* (2011) Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. *Nucleic Acids Res.*, **39**, e65.

McCarroll,S. *et al*. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.*, **40**, 1166–1174.

McKernan,K.J. *et al*. (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.*, **19**, 1527–1541.

Medvedev,P. *et al.* (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods*, **6**, S13–S20.

Miller,C.A. *et al.* (2011) ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One*, **6**, e16327.

Mills,R.E. *et al.* (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.,* **9**, 1182–1190.

Olshen,A.B. *et al.* (2005) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Pang,A.W. *et al.* (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.*, **11**, R52.

Redon,R. *et al*. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Sebat,J. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.

Tuzun,E. *et al*. (2005) Fine-scale structural variation of the human genome. *Nat. Genet.*, **37**, 727–732.

van de Wiel,M.A. *et al.* (2007) CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics*, **23**, 892–894.

Wheeler,D.A. *et al*. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.

Xie,C. and Tammi,M.T. (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinf.*, **6**, 80.

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