BIOINFORMATICS REVIEW

Vol. 28 no. 21 2012, pages 2711–2718
doi:10.1093/bioinformatics/bts535

Sequence analysis

Advance Access publication August 31, 2012

Statistical challenges associated with detecting copy number variations with next-generation sequencing

Shu Mei Teo[1,2,3], Yudi Pawitan[3], Chee Seng Ku[3], Kee Seng Chia[1,2] and Agus Salim[1,*]

[1]Saw Swee Hock School of Public Health, National University of Singapore, Singapore 117597, [2]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore 117456 and [3]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm 17177, Sweden

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Analysing next-generation sequencing (NGS) data for copy number variations (CNVs) detection is a relatively new and challenging field, with no accepted standard protocols or quality control measures so far. There are by now several algorithms developed for each of the four broad methods for CNV detection using NGS, namely the depth of coverage (DOC), read-pair, split-read and assembly-based methods. However, because of the complexity of the genome and the short read lengths from NGS technology, there are still many challenges associated with the analysis of NGS data for CNVs, no matter which method or algorithm is used.

Results: In this review, we describe and discuss areas of potential biases in CNV detection for each of the four methods. In particular, we focus on issues pertaining to (i) mappability, (ii) GC-content bias, (iii) quality control measures of reads and (iv) difficulty in identifying duplications. To gain insights to some of the issues discussed, we also download real data from the 1000 Genomes Project and analyse its DOC data. We show examples of how reads in repeated regions can affect CNV detection, demonstrate current GC-correction algorithms, investigate sensitivity of DOC algorithm before and after quality control of reads and discuss reasons for which duplications are harder to detect than deletions.

Contact: g0801862@nus.edu.sg or agus_salim@nuhs.edu.sg

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

Received on May 30, 2012; revised on August 1, 2012; accepted on August 25, 2012

1 INTRODUCTION

Copy number variations (CNVs) are an important and abundant source of variation in the human genome, encompassing a greater proportion of the genome, as compared with single-nucleotide polymorphisms (SNPs); an estimated 1.2% of a single genome differs from the reference human genome when considering CNVs, as compared with 0.1% by SNPs (Pang *et al.*, 2010). In the past several years, SNP arrays and array comparative hybridization (aCGH) are commonly used for detection of CNVs, albeit with relatively low resolution, especially in terms of breakpoint determination. Sanger sequencing of paired reads, often seen as the gold standard for CNV detection, is able to detect CNVs with higher accuracy and resolution, to detect balanced rearrangements such as inversions and translocations and to detect CNVs in regions where probe density of other platforms, such as SNP arrays, is low. However, the technique is not feasible for a large number of genomes because of time and budget constraints. Next-generation sequencing (NGS) or also known as high-throughput sequencing attempts to combine the benefits of array technology and sequencing. The biggest advantage of NGS over traditional Sanger sequencing is the ability to sequence millions of reads in a single run at a comparatively inexpensive cost (Metzker, 2010). However, because of the complexity of the genome and the short read lengths (usually 35–400 bp) from NGS technology, there are still many challenges associated with the analysis of NGS data for CNVs, no matter which method or algorithm is used.

The growing popularity and success of NGS are evident from large-scale projects such as the 1000 Genomes Project (http://www.1000genomes.org/), which aims to sequence at least 1000 individuals from different populations around the world to construct a detailed map of genetic variations in the human genome (The 1000 Genomes Project Consortium, 2010). Thus far, in its pilot phase, the project has identified ~15 million SNPs, 1 million short indels and >20 000 structural variations (SVs), most of which were previously unreported (~61% of deletions and 89% of duplications are novel). The average SV size detected by the study was 8 kb, approximately four times smaller than a recent SV detection study using tiling CGH array (Conrad *et al.*, 2010). SVs include dosage-altering variants such as CNVs (usually defined as deletions and insertions larger than 1 kb) and shorter indels, as well as dosage-neutral variants such as inversions and translocations.

Nevertheless, current analytical methodologies to analyse NGS data for CNVs are not yet mature, and there are no well-established pipelines/protocols/quality control measures. Broadly, there are four methods for CNV detection using NGS data, namely (i) depth of coverage [DOC, or also known as read-depth (RD) methods], (ii) paired-end mapping (PEM, or also known as read-pair methods), (iii) split-read- (SR) and (iv) assembly-based (AS) methods (Alkan *et al.*, 2011). The different methods are usually complementary to one another, as the underlying concepts excel at detecting certain types of variants, and a large proportion of discovered variants remain unique to a particular approach (Alkan *et al.*, 2011). For example, in the 1000 Genomes Project CNV analysis, the group applied various variations of the four methods, with a total of 36 call sets with

[*]To whom correspondence should be addressed.

vastly varying degrees of false discovery rates (FDR < 10–89%), as well as notable differences in terms of genomic regions ascertained, CNV size range and breakpoint precision among the different methods (Mills *et al.*, 2011). This review article highlights and investigates the challenges encountered when analysing NGS data for CNVs. In particular, we focus on issues pertaining to (i) mappability, (ii) GC-content bias, (iii) quality control measures of reads and (iv) difficulty in identifying duplications. As the characteristics of CNVs in germline and tumour cells are different, we caution that this review focuses largely on CNVs in the germline, and issues unique to tumour CNVs (also known as copy number alterations) are not discussed.

2 FOUR CLASSES OF METHODS FOR CNV DETECTION USING NGS

We describe each of the four methods for CNV detection using NGS data, namely (i) DOC, (ii) PEM, (iii) SR and (iv) AS methods. Except for the latter, the other three classes of methods require first mapping the sequenced reads to a known reference genome. We summarize a list of commonly used software for CNV detection using NGS data in Table 1. Readers may refer to seqanswers website: http://seqanswers.com/wiki/Software for a more comprehensive list.

The underlying concept of identifying CNVs using DOC is similar to that of using intensity data: a lower than expected DOC/intensity indicates deletion and a higher than expected DOC/intensity indicates duplication. Most DOC methods count the number of reads that fall in each pre-specified window of a certain size (Abyzov *et al.*, 2011; Xie *et al.*, 2009; Yoon *et al.*, 2009). The algorithm relies heavily on the assumption that the sequencing process is uniform, i.e. the number of reads mapped to a region is assumed to follow a Poisson distribution and is proportional to the number of copies. However, certain biases such as GC-content and mappability cause this assumption to be unrealistic. Regions of the genome may be

Table 1. Commonly used software for CNV detection using NGS data

Programme	Reference	Comments
DOC		
CNVnator[a]	Abyzov *et al.*, 2011	Uses mean shift approach on fixed window GC-content-adjusted read counts.
Rdxplorer[a]	Yoon *et al.*, 2009	Uses event-wise testing on fixed window GC-content-adjusted read counts.
SeqCBS	Shen and Zhang, 2012	Gives approximate confidence intervals for assessing confidence in the segmentation.
CNVseq	Xie *et al.*, 2009	Uses ratios between reads from target and reference genome.
SegSeq	Chiang *et al.*, 2009	Segments genomes of a tumour and matched normal sample by a sliding fixed size window. Boundary is refined after change point is called.
ExomeCNV	Sathirapongsasuti *et al.*, 2011	For exome sequencing data. Uses read count ratio to detect CNVs, and B allele frequencies to detect LOH.
Control-FREEC	Boeva *et al.*, 2012	Uses total coverage and B allele frequencies of SNPs to call CNVs and LOH.
PEM		
Variation Hunter[a]	Hormozdiari *et al.*, 2009	Based on maximum parsimony. Uses soft clustering.
BreakDancer[a]	Chen *et al.*, 2009	Consist of two complementary algorithms: BreakDancerMax predicts insertions, deletions, inversions and inter- and intra-chromosomal translocations; BreakDancerMini predicts small indels.
PEMer[a]	Korbel *et al.* 2009	Clusters long and short events separately. Confidence value for each SV. Built in database and simulation programme.
SR		
Pindel[a]	Ye *et al.*, 2009	Uses pattern growth algorithm. Identifies breakpoints of large deletions and medium sized insertions.
Assembly based		
Cortex[a]	Iqbal *et al.*, 2011	Capable of assembling multiple genomes simultaneously.
SOAPdenovo[a]	Li *et al.*, 2010	Claims faster computation time and longer contig size and assembly accuracy when compared with earlier methods such as ABySS and velvet.
Velvet	Zerbino *et al.*, 2008	—
ABySS	Simpson *et al.*, 2009	—
Combination/others		
Genome STRiP[a]	Handsaker *et al.*, 2011	Combines DOC, PEM and distribution of evidence across samples and within a genomic locus.
HYDRA	Quinlan *et al.*, 2010	DOC + PEM
ABI tools	McKernan *et al.*, 2009	CBS
Spanner[a]	Mills *et al.*, 2011	Uses PEM and able to find tandem duplications.
SVDetect	Zeitouni *et al.*, 2010	DOC + PEM Competible with SoLiD and Illumina paired-end reads.

[a] used in 1000 Genomes Project.

over- or under-sampled regardless of the copy number of the region, often resulting in spurious signals. Most DOC algorithms correct for GC-content bias before detecting CNVs (Abyzov *et al*., 2011; Yoon *et al*., 2009), whereas there are others that use ratios between reads from the target and reference genome and claim to mitigate the need for GC-correction if the two data sets are prepared in the same way (Xie *et al*., 2009). Other algorithms also exploit SNP heterozygosity information or also known as 'B allele frequency' to call CNVs and loss of heterozygosity (LOH) regions (Boeva *et al*., 2012; Sathirapongsasuti *et al*., 2011). DOC algorithms usually detect large CNVs and are unable to detect copy neutral events such as inversions and translocations. Single-end or paired-end data may be used for this analysis.

PEM methods require the reads to be paired (Chen *et al*., 2009; Hormozdiari *et al*., 2009; Korbel *et al*., 2009). The concept is that the fragments of deoxyribonucleic acid (DNA) from which the reads are to be sequenced have a fragment length (or also known as insert size) of a certain distribution. When the sequenced ends of the fragment map to the reference at a distance longer than expected, it is indicative of a deletion in the studied genome. Vice versa, when the sequenced ends of the fragment map to the reference at a distance shorter than expected, it is indicative of an insertion in the studied genome. Based on the patterns from which the paired reads are mapped to the reference, PEM can also detect inversions and translocations (see Xi *et al*., 2011 for a review of the different SV signatures in PEM). For example, if the two ends of a fragment are mapped with a wrong orientation, it could be an indication of an inversion (Feuk, 2010). The size of CNVs detected using PEM is limited by the insert size, and as a result, PEM often detects smaller CNVs.

SR methods focus on pairs of reads where one read is mapped uniquely to the reference, whereas the other read failed to be aligned (Ye *et al*., 2009). The idea is that the location of the unmapped read may span the breakpoint of the CNV. The mapped read is used as an anchor to narrow down the search space for the SR alignment of the unmapped read. SR analysis has the advantage of being able to pinpoint the location of the breakpoints.

AS methods, on the other hand, do not align the reads to a known reference but construct the genome piece-by-piece; this is also known as *de novo* sequencing. Some AS methods still use the reference genome as a guide to resolve repeats. This is known as *comparative assembly* (Pop *et al*., 2004). AS methods can discover new non-reference sequence insertions. AS methods work best for small genomes such as bacterial genomes and are less widely used in NGS sequencing of humans because the short reads from NGS makes assembly in repeat regions difficult (Ye *et al*., 2009). Most AS algorithms for NGS data are extensions of the method described by Pevzner *et al*., 2001, which uses de Bruijn graphs. It is difficult to judge which method is superior, although the methods developed more recently such as SOAPdenovo (Li *et al*., 2010), claims faster computation time and longer contig size and assembly accuracy when compared with earlier methods such as ABySS (Simpson *et al*., 2009) and velvet (Zerbino *et al*., 2008). Cortex (Iqbal *et al*., 2011) is capable of assembling multiple genomes simultaneously.

Some algorithms use a combination of methods for more accurate detection of CNVs. For example, CNVer (Medvedev *et al*., 2010), HYDRA (Quinlan *et al*., 2010) and SVDetect (Zeitouni *et al*., 2010) supplements DOC with PEM information. Genome STRiP combines information from DOC, PEM and SR, as well as other features of sequence data at population level (Handsaker *et al*., 2011). Genome STRiP is one of the highest performing methods used in the 1000 Genomes pilot Project, indicating that there is benefit in combining different approaches (Mills *et al*., 2011).

3 DATA SETS

For the purpose of gaining insights to the issues we are about to discuss, we download sequenced data of individual NA12891 that was deeply sequenced (>20× coverage) by the Illumina Genome Analyzer platform as part of the 1000 Genomes pilot Project (The 1000 Genomes Consortium, 2010). The reads are paired, 36 bases in length and aligned to the human reference build 36 (hg18) using the MAQ aligner (Li *et al*., 2008). The aligned reads are downloaded in BAM format from http://www.1000genomes.org/.

MAQ calculates a Phred-scaled quality score for each read/pair of reads that is equal to minus ten times the common logarithm of the probability that a read is wrongly aligned; a quality score of 30 indicates a 1 in 1000 probability that the read is incorrectly mapped. When a read can be mapped equally well to more than one location, a random position is chosen out of all equally possible positions, and the reads are assigned a quality score of zero; these reads are termed *multi-reads* (Harismendy *et al*., 2009; Treangen *et al*., 2012). Different aligners have different approaches of dealing with multi-reads. For example, the aligner 'micro-read fast alignment search' (mrFAST) reports all suitable positions of multi-reads (Alkan *et al*., 2009).

3.1 Estimating DOC

We estimate DOC by counting the number of reads, based on their start positions, in non-overlapping windows of 100 bases. This is the current strategy of most DOC algorithms (Abyzov *et al*., 2011; Yoon *et al*., 2009).

3.2 Pre-filtering criteria

For DOC calculation, we keep only reads that are flagged properly aligned, termed 'read mapped in proper pair' in Picard (http://picard.sourceforge.net/explain-flags.html), and reads that are not paired (i.e singletons). Approximately 67% of the reads are flagged properly aligned and ~23% are singletons. We exclude reads that are technical duplicates, paired reads where one read in the pair is unmapped and other reads that are not 'mapped in proper pair'. Singletons are reads where only one end of the fragment is sequenced either because of library preparation or sequencing failure of one the reads in a pair (A.Abyzov, personal communication, Nov 2011). It should be differentiated from reads that are paired but where only one of the reads in the pair is mapped to the reference. Singletons are informative and should not be filtered. This is illustrated in Figure 1, which shows obvious signal of decreased DOC using only singleton reads in a region validated to be a deletion by Mills *et al*., 2011.

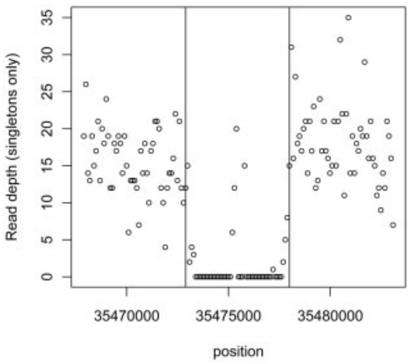

Fig. 1. DOC using singletons only (deletion region in Chromosome 22: 34572901–35478000). This figure shows that singleton reads independently provide informative evidence of a deletion in this region

In the 'Phred score filtered dataset', we further remove 7% of the reads whose mapping quality is <30 (but not zero). Approximately 14% of the reads have a mapping quality of zero. These multi-reads are reads that cannot be uniquely aligned to a single position in the genome, meaning that there exists more than one location where the read can be mapped to equally well. We observe the patterns of multi-reads in regions with known CNVs to investigate how these reads can affect CNV detection.

3.3 Reference CNVs

We use the integer copy numbers for a total of 5037 CNV loci from the studies of Conrad *et al.* (2010) and McCarroll *et al.* (2008) as a reference list. Conrad's experiments were done as follows: first, a set of 20 NimbleGen arrays, each comprising 2.1 million oligonucleotide probes, were used to generate a new map of CNV locations. Subsequently, a customized Agilent CGH-platform composed of 105 000 oligonucleotide probes was used to detect the loci, and the genotypes were estimated for 450 HapMap samples using a Bayesian algorithm with stringent selection for optimal normalization and cluster locations for every locus (see Supplementary Methods in Conrad *et al.*, 2010 for more details). In total, for individual NA12891, there are 517 deletions (copy number <2) and 253 duplications (copy number >2). It should be noted, however, that a true gold standard reference list for CNVs is not available, and this list does not have 100% sensitivity and specificity.

3.4 SNP array intensities

We download SNP array intensities for the Affymetrix 6.0 array for individual NA12891 from the HapMap 3 project raw data database (ftp://ftp.ncbi.nlm.nih.gov/hapmap/raw_data/hapmap3_affy6.0/). We use the PennCNV algorithm (Wang *et al.*, 2007) to obtain log R ratios (LRRs), using samples from the third phase of the HapMap project as the reference panel.

3.5 High-confidence regions

To investigate the reasons for the discordance between the reference regions and DOC data, we plot DOC data for specific regions and observe patterns in the data. To narrow down our search for interesting regions, we limit this analysis to high-confidence regions, which we define as follows: a deletion region from the reference list is considered 'high-confidence' if it also shows an average LRR of $<\log(0.5) \sim -0.7$. A duplication region is considered 'high-confidence' if it shows an average LRR of $>\log(1.5) \sim 0.4$. There are 60 high-confidence deletions and 8 high-confidence duplications. The regions range from 1 to 156 kb, and the number of SNP markers range from 1 to 73.

4 REPEAT REGIONS AND MAPPABILITY ISSUES

NGS technology produces mainly short reads, and this poses a challenge in the alignment to the reference genome because reads that fall in repetitive regions in the genome cannot be mapped unambiguously. Furthermore, mutations or sequencing errors in one or two locations may also cause reads to be mapped wrongly (Li *et al.*, 2008). In the 1000 Genomes trios Project, ~20% of the reference genome was considered inaccessible (defined as regions with many ambiguously placed reads or unexpectedly high or low numbers of aligned reads). The resulting low sensitivity in detecting CNVs in repeated/segmental-duplicated regions is a serious problem because there is an observed enrichment of CNVs in segmental duplicated regions, and many breakpoints lie in duplicated regions (Medvedev *et al.*, 2009). This class of CNVs is one of the most poorly studied variants, as previous technologies for CNV detection such as aCGH and SNP arrays also have problems resolving them.

For AS methods, repeat regions create challenges because if the read length is shorter than the repeat region, it is not straightforward to decipher the original sequence because overlap between the reads or contigs will be ambiguous (Knudsen *et al.*, 2010). For other methods that require mapping to a reference, there are different alignment strategies for dealing with multi-reads, such as (i) discarding the reads, (ii) choosing a position at random out of all equally good match positions and (iii) reporting all possible positions.

The first strategy limits the analysis only to unique regions of the genome, and may miss many CNVs. Moreover, when using DOC methods, excluding multi-reads may cause the identification of false deletions, i.e. regions with a large number of multi-reads will be falsely detected as a deletion if these reads were removed. This is illustrated in Figure 2, which shows a region in Chromosome 20 where several deletions would be falsely identified if multi-reads are excluded. This phenomenon was also observed by Abyzov *et al.* (2011), whose algorithm picked up ten times as many deletions when multi-reads are discarded.

Placing a multi-read at random (Strategy 2) is also not ideal: for example, a true deletion may exist in a region where there exist similar sequences elsewhere in the genome, causing multi-reads to be mapped to the deletion region where there is supposed to be less or none, thereby diluting the signal (Fig. 3). This suggests that the alignment strategy of discarding multi-reads or random placement of multi-reads is inadequate for detecting duplications in repeated regions. A better strategy incorporating other alignment methods and other kinds of data is needed to identify these regions.

He *et al.* (2011) developed a new algorithm for tandem CNV reconstruction in repeat-rich regions that considers all locations of possible mappings and uses information on PEM and DOC.

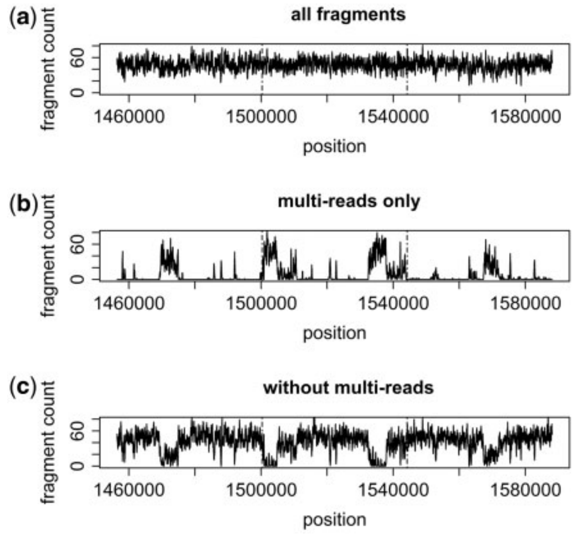

Fig. 2. Fragment count for NA12891, Chromosome 20. (**a**) It uses all fragments and shows a relatively flat DOC, which varies around the average. (**b**) It uses only multi-reads and shows several small regions with spikes in multi-reads. (**c**) It uses all fragments with multi-reads removed; we observe 'holes' or dips in DOC that would be identified as deletions by DOC algorithms. Multi-reads are placed at random out of all equally possible locations

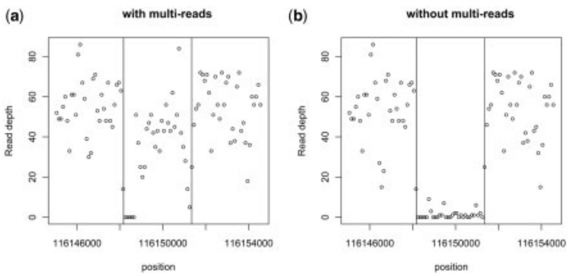

Fig. 3. High-confidence deletion region in Chromomsome 4 (116148170–116151343) not identified by DOC methods in Mills *et al.* (2011). (**a**) Some evidence of deletion is seen when we include all reads. (**b**) Deletion signal becomes more obvious with multi-reads removed

Alkan *et al.* (2009) developed a new alignment method, mrFAST; the aligner maps short sequence reads to a repeat-masked reference genome, meaning that all loci with known high-copy common repeats were first masked before alignment, and reports all mapping locations for multi-reads. It also keeps track of mutation in multi-reads. This method has been shown to be able to predict absolute copy number and multi-copy differences. Sudmant *et al.* (2010) also uses a similar approach to identify and genotype CNVs within segmental duplications. However, these approaches seem to work only for deeply sequenced data (>20×), and more has to be done to extend these methods for lower coverage data (Chiang and McCarroll, 2009).

Longer read lengths from third-generation sequencing (TGS) may partially solve the problems with repeats, but even with a read length of 1 kb, there still remains ~1.5% of the human genome sequence that is non-unique (Schatz *et al.*, 2010).

5 GC-CONTENT

It has been observed that DOC has a unimodal relationship with GC-content (Abyzov *et al.*, 2011; Benjamini *et al.*, 2012; Yoon *et al.*, 2009), where regions with high or low GC-content have decreased DOC. Harismendy *et al.* (2009) also observed that unique sequences present at equimolar quantities in library generation end up being sequenced at vastly different DOC. This bias causes problems in all methods. For example, in PEM or SR methods, a region of low DOC may have insufficient reads for enough evidence to discern the variants at that location. For AS methods, regions with low coverage may also result in insufficient information to infer a continuous sequence (Knudsen *et al.*, 2010). The problem can however be solved by increasing the overall sequence depth. The most affected of four methods by GC-content bias is the DOC method.

DOC algorithms rely heavily on the assumption that the sequencing process is uniform, so that the DOC can be assumed to be proportional to the copy number. However, when there are biases that cause sequencing depth to differ for reasons other than the change in copy number, it makes differentiating true deletions/duplication from under/over-sampled regions in the genome difficult. Previous published algorithms correct for GC-content by adjusting the DOC in the window using the GC-content of the window (Abyzov *et al.*, 2011; Yoon *et al.*, 2009). This method of correction may be inadequate as the choice of bin size is often arbitrary. Moreover, several studies have observed that it is the GC-content of the full DNA fragments, not only the reads, that causes most of the bias (Benjamini *et al.*, 2012).

A recently developed algorithm for GC correction considers the GC-content of the fragment and can produce base pair resolution predictions of GC-content bias (Benjamini *et al.*, 2012). We applied the method on this data set but observed an increase in overall variance of DOC after correction. Hence, we did not use the results of this correction for subsequent comparisons (see Supplementary Materials for more details).

We download the GC-content per five bases from the University of California, Santa Cruz genome bioinformatics website: http://hgdownload.cse.ucsc.edu/goldenPath/hg18/gc5Base/. We correct for GC-content bias in a similar fashion as described by Yoon *et al.* (2009). The GC-corrected DOC was calculated using the following equation:

$$\mathrm{RD}^{i}_{\mathrm{corrected}} = \mathrm{RD}_{\mathrm{global}} \times \mathrm{RD}^{i}_{\mathrm{raw}} / \mathrm{RD}_{\mathrm{gc}},$$

where i is the bin index, $\mathrm{RD}_{\mathrm{global}}$ is the average DOC over all bins in the chromosome (we used a trim mean, omitting 5% of bins from both extremes), $\mathrm{RD}^{i}_{\mathrm{raw}}$ is the DOC for the i^{th} window before correction and $\mathrm{RD}_{\mathrm{gc}}$ is the median DOC of all windows with the same GC-content. As there are few windows with GC <20 or >75, for robustness, we set the lower/upper limits for GC in a window to 20 and 75, respectively. Figure 4 (left) plots the DOC of the windows versus the GC percentage of the windows. We observe a similar unimodal relationship between DOC and GC-content, as reported by previous articles. In AT-rich regions, coverage increases with increasing GC, and in GC-rich regions, coverage decreases with increasing GC. The peak coverage can be different for different data sets and different chromosomes but is usually located between

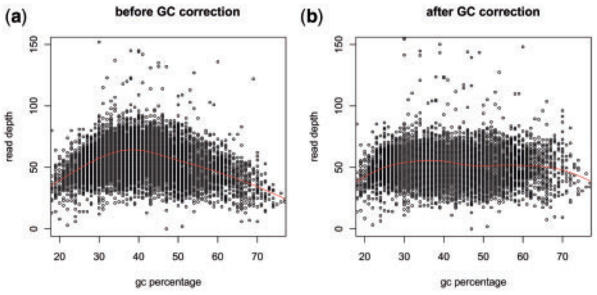

Fig. 4. Read depth versus GC percentage before and after correction

0.35 and 0.5 GC. Figure 4 (right) shows that GC-content bias is removed after correction. However, it is worth noting that even though GC-content bias is removed, the variance in DOC remains rather large, meaning that not all local variations in DOC are associated with GC-content and thus cannot be removed by the GC-correction.

The cause of GC-content bias is speculated to be largely because of polymerase chain reaction (PCR) amplification step in NGS (Aird *et al*., 2011; Benjamini *et al*., 2012). As PCR amplification is not required in TGS, bias observed in DOC because of PCR may be resolved (Schadt *et al*., 2010). The longer read lengths of TGS will also improve challenges caused by the short read lengths of NGS. However, as TGS technology is still new, it is premature to comment on its performance, and too soon to judge whether TGS can fulfil its promises of advancement over NGS.

6 PHRED-SCORE FILTERING

There has been little documentation on how read mapping quality affects CNV calling. Most algorithms state a default filtering criteria without any substantial evidence for the choice. For example, PEM algorithm, BreakDancer, uses a default filter of mapping quality >10 (Chen *et al*., 2009), whereas DOC algorithm, Rdxplorer, uses a default filter of mapping quality >30 (Yoon *et al*., 2009).

7 COMPARISONS

We perform sensitivity analysis to investigate the effects of GC-correction and Phred-score filtering. We compare three methods: (i) GC-corrected and Phred-score filtered, (ii) GC-corrected but not filtered by Phred-score and (iii) Phred-score filtered but GC-uncorrected. For each CNV in the reference list, we use the *t*-statistic to determine whether the DOC in the region is significantly increased/decreased. For each deletion region *i*, we calculate the *t*-statistic as such:

$$t_i = \frac{\bar{x}_i - c\hat{\mu}_g}{\hat{\sigma}_i/\sqrt{n_i}},$$

where $\bar{x}_i$ is the average DOC in the region, $\hat{\mu}_g$ is the global average DOC for the chromosome where the region lies, $\hat{\sigma}_i$ is the standard deviation of the DOC in region *i*, n_i is the number of windows in the region and *c* is a constant that we vary from 0.5 to 0.8. For each duplication region *j*, the *t*-statistic is calculated in a similar fashion:

$$t_j = \frac{k\hat{\mu}_g - \bar{x}_j}{\hat{\sigma}_j/\sqrt{n_j}},$$

where *k* varies from 1.2 to 1.6. For each set of tests, we account for multiple comparisons using the FDR. A region is identifiable if the FDR is <0.01.

Table 2. Sensitivity of Phred-score filtered and unfiltered data sets, and GC-corrected and non-GC-corrected data sets

Dataset	Deletion				Duplication				
c/k	0.5	0.6	0.7	0.8	1.6	1.5	1.4	1.3	1.2
Phred-score filtered + GC-corrected	0.31	0.66	0.83	0.89	0.07	0.12	0.15	0.2	0.24
Phred-score unfiltered	0.31	0.66	0.82	0.89	0.09	0.12	0.16	0.21	0.25
GC-uncorrected	0.32	0.65	0.81	0.89	0.08	0.11	0.15	0.2	0.25

Table 2 shows that there are little differences in sensitivities for all three methods, suggesting that both GC-correction and Phred-score filtering do not seem to be crucial in the sensitivity of detection of CNVs. It should be noted however that this analysis does not investigate the specificity of CNV detection. Overall, GC-correction and Phred-score filtering lowers the variance of DOC, indicating the potential of minimizing the number of false positive regions identified. However, this is a simple and limited analysis, and further studies are needed to discern the benefits of GC-correction and filtering by Phred-score.

8 INSERTIONS ARE HARDER TO DETECT THAN DELETIONS

For all methods, identifying duplications has been acknowledged as more challenging as compared with identifying deletions. With regards to PEM methods, the bias against detection of insertions is because PEM detects insertions when the mapped reads are at a distance shorter than the fragment length, and hence, it is unable to detect insertions larger than the insert size of the reference library, or more specifically the length upper bound of an insertion detected is the average fragment length minus the length of the reads (Hormozdiari *et al*., 2009; Wang *et al*., 2008). This is evident in detection of CNVs using PEM of the diploid Asian 'YH' genome, where 2441 deletions were identified as compared with 33 duplications (Wang *et al*., 2008).

In DOC methods, we observe that the sensitivity of detecting deletions and duplications is ~89 and 25%, respectively, for the best case scenarios (Table 2). This observation is similar to that observed in Abyzov *et al*. (2011), who estimated that ~90% of deletions identified by aCGH or SNP arrays are DOC accessible, whereas only 20–30% of duplications are DOC accessible. This may be because of the lack of sensitivity of DOC methods in distinguishing a change in number of copies from N to $N+1$, especially if N is large. For example, suppose a sequence is repeated twice in the reference genome ($N=2$) at locations

A and B, although the studied genome has an additional copy ($N = 3$). Then, assuming an average of 20× coverage, locations A and B would have an average of 60 reads shared among both locations (following strategy of random placement of multi-reads), meaning an average of 30 reads at both A and B, a 50% increase in DOC. If we increase N to 5 in the reference and 6 in the studies genome, then each repeated location in the reference would have an average of $120/5 = 24$ reads, only 20% more than the average, and likely to be undetectable because of the high variance in DOC.

In the list of high-confidence regions (see section 'High-Confidence Regions'), all 60 deletions can be found in at least one release set from Mills *et al.* (2011), but four of the regions were not detected by DOC methods. When we plotted the read depth in these regions, we observed that two regions have obvious decreased DOC (figure not shown) and should have been detected, whereas the other two were not detected most likely because of the presence of multi-reads diluting the deletion signal (see Fig. 3).

On the other hand, all eight duplications are not identified in any of the release sets (see Table S1 and Figs S1–8 in Supplementary Materials). This is partly because of the fact that most release sets in Mills *et al.* (2011) focus mainly on deletions, with few sets reporting duplications/insertions. Even then, of the 8 regions, only Regions 2 and 5 show distinct elevated DOC; these regions have little or no multi-reads. Among the other six regions that do not show obvious increase in DOC, four of them overlap with known segmental duplication regions (segmental duplicated regions as defined in http://humanparalogy.gs.washington.edu/). This is also supported by the presence of multi-reads in these regions; neither keeping nor removing multi-reads result in strong DOC signal of the presence of duplication.

9 DISCUSSION

NGS, with its ability to perform massive parallel sequencing in a single run, is becoming increasingly popular. This brings with it an unprecedented opportunity to sequence many genomes at a relatively inexpensive cost (as compared with using Sanger sequencing). However, with billions of reads generated per individual, the sheer and exponentially increasing amount of data demands for better bioinformatics support and computers with larger storage and higher computing powers. No less important than the production of the data is the information technology infrastructure, and bioinformatics team needed to analyse it, with speculations that the costs associated with handling, storing and analysis of the data could be more than the production of the data.

Analysing NGS data for structural variants is a relatively new and challenging field, with no standard protocols or quality control measures. The four methods of CNV detection are complementary. Comparing DOC, PEM and SR methods used in the 1000 Genomes Project, each approach uniquely discovered 30–60% of the CNVs (Abyzov *et al.*, 2011). These three methods require first mapping the sequenced reads to a reference genome. As the mapped reads are used in all downstream analysis, this first step of alignment is extremely crucial. As has been shown in the article, how the aligner or subsequent algorithm deals with reads in repeat regions is important for detecting variants that lie in these regions. Currently, the problem of CNV detection in repeated regions is still not completely solved.

Using real data from the 1000 Genomes Project, this article highlights and investigates challenges associated with current methodologies and areas of potential biases encountered when analysing NGS data for CNVs. In particular, we focus on issues pertaining to (i) mappability, (ii) GC-content bias, (iii) quality control measures of reads and (iv) difficulty in identifying duplications. We feel this is a timely critical review that would aid researchers in a much needed well-validated pipeline for the analysis of structural variants.

Funding: This work is supported by the Swedish Science Council and National University of Singapore Start-up Grant No. R-186-000-103-133. S.M.T. acknowledges support from the National University of Singapore Graduate School for Integrative Sciences and Engineering Scholarship.

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.

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

Aird,D. *et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.

Benjamini,Y. and Speed,T.P. (2012) Summarizing and correction for GC-content bias in high throughput sequencing. *Nucleic Acids Res.*, **40**, e72. doi:10.1093/nar/gks001.

Boeva,V. *et al.* (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, **28**, 423–425.

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

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

Chiang,D.Y. and McCarroll,S.A. (2009) Mapping duplicated sequences. *Nat. Biotechnol.*, **27**, 1001–1002.

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

Feuk,L. (2010) Inversion variants in the human genome: role in disease and genome architecture. *Genome Med.*, **2**, 11.

Handsaker,R.E. *et al.* (2011) Discovery and genotyping of genome structural polymorphism by sequencing on population scale. *Nat. Genet.*, **43**, 269–276.

Harismendy,O. *et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.*, **10**, R32. doi: 10.1186/gb-2009-10-3-r32.

He,D. *et al.* (2011) Efficient algorithms for tandem copy number variation reconstruction in repeat-rich regions. *Bioinformatics*, **27**, 1513–1520.

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

Iqbal,Z. *et al.* (2011) *De novo* assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.*, **44**, 226–232.

Korbel,J. *et al.* (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.*, **10**, R23.

Knudsen,B. *et al.* (2010) A computer simulator for assessing different challenges and strategies of *de novo* sequence assembly. *Genes*, **1**, 263–282.

Li,H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Li,R. *et al.* (2010) *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.*, **20**, 265–272.

McCarroll,S.A. *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 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 (11 Suppl.)**, S13–S20.

Medvedev,P. *et al.* (2010) Detecting copy number variation with mated short reads. *Genome Res.*, **20**, 1613–1622.

Metzker,M.L. (2010) Sequencing technologies—the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

Mills,R.E. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

Pang,A.W. *et al.* (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.*, **11**, R52. doi: 10.1186/gb-2010-11-5-r52.

Pevzner,P.A. *et al.* (2001) An Eulerian path approach to DNA fragment assembly. *Proc. Natl Acad. Sci. USA*, **98**, 9748–9753.

Pop,M. *et al.* (2004) Comparative genome assembly. *Brief. Bioinformatics*, **5**, 237–248.

Quinlan,A.R. *et al.* (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res.*, **20**, 623–635.

Sathirapongsasuti,J.F. *et al.* (2011) Exome sequencing-based copy number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*, **27**, 2648–2654.

Schadt,E.E. *et al.* (2010) A window into third-generation sequencing. *Hum. Mol. Genet.*, **19**, R227–R240.

Schatz,M.C. *et al.* (2010) Assembly of large genomes using second generation sequencing. *Genome Res.*, **20**, 1165–1173.

Shen,J.J. and Zhang,N. (2012) Change-point model on nonhomogeneous poisson processes with application in copy number profiling by next-generation DNA sequencing. *Ann. Appl. Stat.*, **6**, 476–496.

Simpson,J.T. *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.*, **19**, 1117–1123.

Sudmant,P.H. *et al.* (2010) Diversity of human copy number variation and multicopy genes. *Science*, **330**, 641–646.

The 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Treangen,T.J. and Salzberg,S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, **13**, 36–46.

Wang,K. *et al.* (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, **17**, 1665–1674.

Wang,J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.

Xi,R. *et al.* (2011) Detecting structural variations in the human genome using next generation sequencing. *Brief. Funct. Genomics*, **9**, 405–415.

Xie,C. and Tammi,M.T. (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 break points 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.

Zeitouni,B. *et al.* (2010) SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, **26**, 1895–1896.

Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.