A peer-reviewed version of this preprint was published in PeerJ on 17 March 2015.

View the peer-reviewed version (peerj.com/articles/836), which is the preferred citable publication unless you specifically need to cite this preprint.

Smith SD, Kawash JK, Grigoriev A. 2015. GROM-RD: resolving genomic biases to improve read depth detection of copy number variants. PeerJ 3:e836 https://doi.org/10.7717/peerj.836

GROM-RD: Resolving Genomic Biases to Improve Read Depth Detection of Copy Number Variants

Sean D. Smith[1], Joseph K. Kawash[1] and Andrey Grigoriev[1,*]

[1]Department of Biology, Center for Computational and Integrative Biology, Rutgers University, Camden, New Jersey 08102, USA

[*]Corresponding author: andrey.grigoriev@rutgers.edu

PeerJ PrePrints

Abstract

Amplifications or deletions of genome segments, known as copy number variants (CNVs), have been associated with many diseases. Read depth analysis of next-generation sequencing (NGS) is an essential method of detecting CNVs. However, genome read coverage is frequently distorted by various biases of NGS platforms, which reduce predictive capabilities of existing approaches. Additionally, the use of read depth tools has been somewhat hindered by imprecise breakpoint identification. We developed GROM-RD, an algorithm that analyzes multiple biases in read coverage to detect CNVs in NGS data. We found non-uniform variance across distinct GC regions after using existing GC bias correction methods and developed a novel approach to normalize such variance. Although complex and repetitive genome segments complicate CNV detection, GROM-RD adjusts for repeat bias and uses a two-pipeline masking approach to detect CNVs in complex and repetitive segments while improving sensitivity in less complicated regions. To overcome a typical weakness of RD methods, GROM-RD employs a CNV search using size-varying overlapping windows to improve breakpoint resolution. We compared our method to two widely used programs based on read depth methods, CNVnator and RDXplorer, and observed improved CNV detection and breakpoint accuracy for GROM-RD. GROM-RD is available at http://grigoriev.rutgers.edu/software/

1. Introduction

Copy number variants (CNVs) have been linked to several diseases including cancer (Berger et al. 2011; Campbell et al. 2010; Stephens et al. 2009), schizophrenia (Stefansson et al. 2009), and autism (Marshall et al. 2008). Compared to single nucleotide polymorphisms (SNPs), structural variants (or SVs, which include CNVs, insertions, inversions, and translocations) account for more differences between human genomes (Baker 2012) in terms of the number of nucleotides and potentially have a greater impact on phenotypic variation (Korbel et al. 2007). Modern sequencing technologies, often identified as next-generation sequencing (NGS), have enabled higher resolution of CNVs compared to older methods such as array comparative genome hybridization (aCGH) and fosmid paired-end sequencing (Korbel et al. 2007). NGS produces sequenced reads, either single- or paired-end, that are mapped to a reference genome. Several strategies have been developed to detect SVs. Paired-read methods search for clusters

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.663v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2014, publ: 4 Dec 2014

of discordant (aberrant insert size or orientation) read pairs. Split-read methods map previously unmapped reads by splitting the reads. Read depth (RD) methods identify CNVs by detecting regions of low or high read coverage. *De novo* methods assemble reads into contigs, particularly useful for detecting insertions. Each detection strategy has advantages and disadvantages, and they complement each other by detecting SVs not found or not detectable using the other strategies. For example, RD does not depend on paired reads for finding SVs and is able to detect CNVs with mutated or rough breakpoints that may not be detectable with paired or split reads, but RD is unable to detect insertions, translocations, and inversions.

Several whole genome sequencing (WGS) RD methods, CNV-seq (Xie & Tammi 2009), SegSeq (Chiang et al. 2009), rSW-seq (Kim et al. 2010), CNAseg (Ivakhno et al. 2010), and CNAnorm (Gusnanto et al. 2012), require a control sample. Other WGS RD methods, such as JointSLM (Magi et al. 2011) and cn.MOPS (Klambauer et al. 2012), require multiple samples. Often multiple samples or a suitable control are not available. Whole exome sequencing (WES) RD methods, including ExomeCNV (Sathirapongsasuti et al. 2011), CONTRA (Li et al. 2012), EXCAVATOR (Magi et al. 2013), CoNIFER (Krumm et al. 2012), and XHMM (Fromer et al. 2012) are limited to detection in coding regions of the genome (Sims et al. 2014). WGS RD methods that do not require a control include FREEC (Boeva et al. 2011), ReadDepth (Miller et al. 2011), CNVnator (Abyzov et al. 2011), and RDXplorer (Yoon et al. 2009).

Detecting CNVs is complicated by GC bias of NGS technologies, whereby read coverage varies depending on the GC content of the genome region. Existing RD methods reduce GC bias by GC bin mean normalization (CNVnator and RDXplorer), polynomial fitting (FREEC), and LOESS regression (ReadDepth). However, these methods do not consider differences in read depth variance with GC content, which may exist after GC bias correction. Complex and repetitive regions are challenging for all CNV detection methods including RD. Complex regions near telomeres and centromeres are known to be SV hotspots (Mills et al. 2011) and sequencing bias has been observed in repeat regions (Ross et al. 2013). However, RD methods have not been tailored for the difficulties of complex and repetitive regions. Additionally, RD methods suffer from low breakpoint resolution.

We have developed GROM-RD, a control-free WGS RD algorithm with several improvements and novel features compared to existing RD algorithms, such as excessive coverage masking, GC bias mean and variance normalization, GC weighting, dinucleotide repeat bias detection and adjustment, and a size-varying sliding window CNV search. These features address weaknesses in existing RD methods and biases in genomic sequencing that limit CNV sensitivity, specificity, and breakpoint accuracy, as evidenced by comparison of our algorithm to two most commonly used control-free WGS RD tools, RDXplorer (Yoon et al. 2009) and CNVnator (Abyzov et al. 2011). GROM-RD showed improved predictive capabilities and breakpoint resolution for CNVs, as well as excellent scalability for different NGS datasets, both simulated and real.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.663v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2014, publ: 4 Dec 2014

2. Methods

GROM-RD outputs a union set from two pipelines that differ based on the inclusion or exclusion of a pre-filtering step, excessive coverage masking (Fig. 1). Each step from Fig. 1 will be described in the following subsections.

1. Excessive Coverage Masking (Stable Region CNV Detection)
2. GC Weighting
3. GC Bias Normalization
4. Dinucleotide Repeat Bias Normalization
5. Sliding Window CNV Search

Figure 1. GROM-RD Pipeline Summary. Two iterations of the pipeline are combined into a union set of CNV predictions. For the first iteration (step 1 included), CNV detection in stable regions is improved by masking regions of excessive coverage. Without masking (step 1 excluded), CNVs are detected in complex and repetitive regions that are characterized by excessive coverage.

2.1 Excessive Coverage Masking

Abnormal read coverage has been reported in centromere and telomere regions (Rausch et al. 2012). Similarly, we observed excessive read coverage in certain regions, particularly near centromeres (Fig. 2). This might be due to complex and repetitive segments, which are common in the human genome and can complicate CNV detection. Such high read coverage may result in false positives and also reduce CNV sensitivity in less complex regions. GROM-RD uses a two-pipeline approach to detect CNVs in complex and repetitive segments and improve sensitivity in less complicated regions. In the first pipeline, we mask clusters of blocks (10,000 base segments) with high read coverage (default: >2x chromosome average) and run GROM-RD on the masked genome. A cluster is defined as a section of the genome where >25% of the blocks have high read coverage and a minimum of four blocks have high read coverage. High coverage regions have been shown to have a high concentration of SVs (Mills et al. 2011). Thus, in the second pipeline, we run GROM-RD on the unmasked genome. GROM-RD outputs a union set of predicted CNVs from the two pipelines. Many false positives may be produced from spikes in read coverage, particularly for the unmasked genome. Thus during later steps in the pipeline, read coverage greater than twice the chromosome average is adjusted (described in section 2.3).

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.663v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2014, publ: 4 Dec 2014

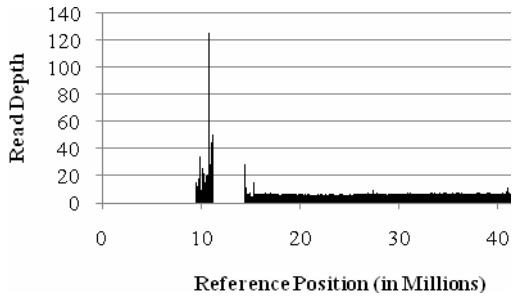

Figure 2. Read depth variation in chromosome 22 of NA12156 (Illumina low-coverage paired-end read dataset aligned with BWA to human reference hg19, 1000 Genomes Project) (Abecasis et al. 2010). Read depth was averaged for 10,000 base regions. Clusters of high read depth occurred near the centromere (~10-15 million base region).

2.2 GC Weighting

Variation in the GC content of genome regions affects read coverage produced by NGS platforms. A post-sequencing approach used by many RD algorithms, such as CNVnator and RDXplorer, is to bin genome regions by GC content and adjust the average read depth of each bin to the average read depth of the genome, referred to as GC bias normalization. Here we discuss the first step of this approach, calculating GC content of genome regions. RD algorithms often divide a chromosome into regions, referred to as windows, of a fixed size and estimate read depth in each window by counting reads within the window. GC content for a window is calculated from the proportion of reference sequence G and C bases within the window. Previous studies (Aird et al. 2011; Benjamini & Speed 2012; Bentley et al. 2008) have identified PCR bias as the main contributor to GC bias in NGS. Thus, reference bases outside a window may affect read coverage within a window, especially for long reads and paired-end reads. Benjamini and Speed (Benjamini & Speed 2012) showed a higher correlation between GC content and read depth when considering the GC content of the entire PCR-replicated DNA fragment rather than the sequenced segment. Based on these observations, we developed a novel GC weighting method to consider all bases within an average insert size. To maximize sensitivity, we do not calculate GC weighting for a window of bases, instead GC weighting is calculated for each base i as $h_i = \sum w_j a_j / \sum w_j$, where j is a base that may affect read depth for base i, w_j is the weight of base j and is equivalent to the sum of average inserts that overlap base j and base i, and a_j is 1 if base j is a G or C and 0 otherwise. For single-end reads, the insert size is equivalent to read length.

2.3 GC Bias Normalization

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.663v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2014, publ: 4 Dec 2014

As referred to previously, "GC bias" in this context denotes variation in read coverage produced by NGS platforms as a result of variation in the GC content of genome regions. Many RD algorithms, such as CNVnator and RDXplorer, bin genome regions (windows) by GC content and adjust the average read depth of each bin to the average read depth of the genome:

$$r_{i,norm} = r_i \, m / m_{GC} \qquad (1)$$

where $r_{i,norm}$ is the read coverage of a window after normalization, r_i is the read coverage of window i prior to normalization, m is the global mean read coverage of all windows in the genome, and m_{GC} is the mean read coverage of all windows with similar GC content (Yoon et al. 2009). Although this method normalizes the read depth means across the GC bins, we found differences in variance after GC bias correction (Fig. 3). From this observation, we expect methods using this approach to over-predict CNVs when a GC region has high variance and under-predict CNVs when a GC region has low variance.

We use a quantile normalization approach to correct for variance across bins of GC weighted bases (Lin et al. 2004). For this approach, we rank bases in each bin based on read depth and calculate a rank proportion p_i for each base i using

$$\begin{aligned} p_i &= R_i / n && \text{if } 2R_i \leq n \\ p_i &= (n - R_i) / n && \text{if } 2R_i > n \end{aligned} \qquad (2)$$

where R_i is the read depth rank for base i and n is a count of bases with a particular GC weighting. When R_i is 0 (for $2R_i \leq n$) or n-R_i is 0 (for $2R_i > n$), the numerator in Equation 2 is set to 0.5. Subsequently, p_i is converted to standard deviation units, x_i, using a pre-computed normal distribution table. Note when n is identical for all GC bins, each bin distribution will have identical statistical properties, including mean and variance, after quantile normalization. Statistical properties of quantile normalized distributions may vary across GC bins when n varies, however this effect is negligible when n is large. GROM-RD requires a GC bin to have at least 100 bases. GROM-RD does not produce a normalized read depth as in Equation 1 because it is not necessary for further analysis. Instead, read depth in standard deviation units is used. As mentioned previously in section 2.1, to reduce false positives, read coverage greater than twice the chromosome average is adjusted by averaging the rank of the observed read coverage and the rank of read coverage equivalent to twice the chromosome average read coverage. CNVs may occur in low mapping quality regions, however, read coverage distributions tend to differ between low mapping quality and high mapping quality regions. To compensate for variation of read coverage distributions with mapping quality, GROM-RD calculates the average mapping quality for each window and creates

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.663v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2014, publ: 4 Dec 2014

separate distributions for low mapping quality (default: <5) and high mapping quality windows. The nature of the read depth distribution for NGS data has not been clearly defined. A rank-based approach does not assume a specific distribution and is less affected by outliers when compared to parametric methods.

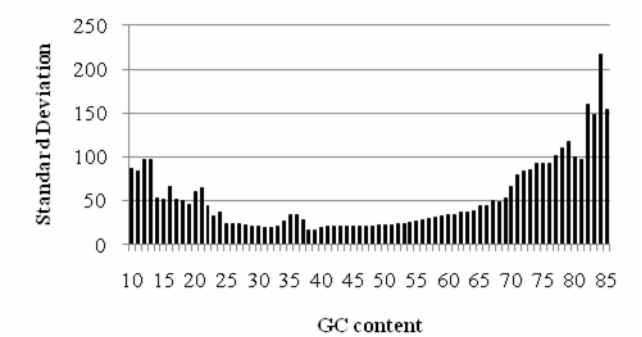

Figure 3. Standard deviation after GC bias normalization. Data produced from chromosome 19 of NA12878 (Illumina high-coverage paired-end read dataset aligned with BWA to human reference hg18, 1000 Genomes Project) (Abecasis et al. 2010) using 100-base non-overlapping windows. Reads were assigned to a window if the read center was within the window. After correcting for GC bias using a common approach, the standard deviation varies with GC content. This negatively impacts further analysis by CNV detection algorithms.

2.4 Dinucleotide Repeat Bias Normalization

Repeat bias has been observed with NGS technologies (Ross et al. 2013). We found similar repeat biases in our investigations. Additionally, these biases may vary with sequencing technology and genomes. For instance, we observed decreased coverage for AT repeats in human (Fig. 4) but not for other genomes (data not shown). We found that dinucleotide repeats as short as 20 bases affected coverage. GROM-RD detects dinucleotide repeat biases and uses a quantile normalization method in the respective genomic regions. Dinucleotide repeats with average read coverage that is more than 1.5 standard deviations below the genome average read coverage, and vice versa (genome coverage more than 1.5 standard deviations above dinucleotide coverage), are considered biased. For a biased dinucleotide repeat, we use a quantile normalization approach similar to our GC bias normalization, except R_i is the read depth rank of occurrence i of a particular dinucleotide repeat. From this we obtain read depth in standard deviation units for each biased dinucleotide repeat occurrence. As we move further from a repeat, GROM-RD creates separate sample distributions in 10 base increments to adjust for the decreasing influence of repeat bias. Thus, we bin bases by distance from the repeat, in contrast to binning by GC weighting as described in section 2.2. Repeat bias normalization is applied within a distance of half-insert size from biased dinucleotide repeats. For genomic regions with dinucleote repeat bias, dinucleotide repeat bias normalization replaces GC bias normalization. To our knowledge, GROM-RD is the first RD method to specifically adjust for repeat bias.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.663v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2014, publ: 4 Dec 2014

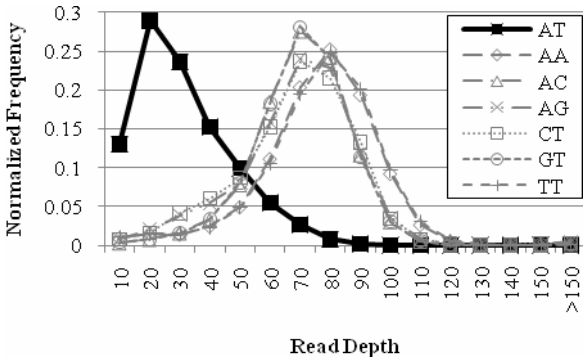

Figure 4. Example of dinucleotide repeat bias in a human genome. AT repeats had lower coverage compared to other dinucleotide repeats for human genome NA12878 (Illumina high-coverage paired-end read dataset aligned with BWA to human reference hg18, 1000 Genomes Project). Dinucleotide repeats less than 20 bases were filtered. Dinucleotide combinations with less than 50 occurrences in the genome are not shown.

2.5 Sliding Window CNV Search

RD methods typically suffer from reduced breakpoint resolution compared to other methods, such as split-read. One reason for low resolution is fixed-size, non-overlapping windows. We employ sliding windows that sequentially increase in one-base increments to improve breakpoint resolution. Fixed-size, non-overlapping windows also reduce sensitivity when CNVs start or end near the center of a non-overlapping window. Using sliding windows, GROM-RD is equally sensitive to CNVs regardless of start or end points. Additionally, by creating distributions for incremental window sizes, GROM-RD improves sensitivity on a range of CNV sizes.

As described in the previous sections, GROM-RD normalizes GC bias or, if necessary, dinucleotide repeat bias for each base. However, we do not expect to find one base deletions or duplications, instead GROM-RD combines normalized bases into windows by averaging standard deviation units of all bases in a window. Since the means and variances of the bases have been normalized with respect to GC bias or dinucleotide repeat bias, GC and dinucleotide bias are not associated with the windows.

For each window size, we sample a set of windows from the dataset and obtain a read depth mean and standard deviation. Then, we identify base positions with read coverage $\geq 1.3 r_{ave,h}$ or $\leq 0.70 r_{ave,h}$ (for diploids) as potential breakpoints, where $r_{ave,h}$ is the average read depth for bases windows with h weighted GC content. At a potential breakpoint j, we calculate a z-score, z, based on a sample distribution of read depths for the minimum window size, w_{min} (default=100), and the read depth of a window i having size w_{min} and beginning at j.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.663v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2014, publ: 4 Dec 2014

Several parameters affect calling CNVs as outlined below (and they can potentially be modified by a user). A CNV is called if $z < \alpha$, (default: $\alpha=1x10^{-6}$). We increase the window size in one-base increments and recalculate z to either extend or detect a CNV until a maximum window size w_{max} (default=10,000) is reached. If no CNV has been detected, we move to the next potential breakpoint and repeat our statistical testing. Attempts to extend or detect a CNV will end before reaching w_{max} if less than half the bases have extreme read coverage (≥ 1.3 or $\leq 0.70 r_{ave,h}$ for diploids). If a CNV was found and w_{max} has been reached, we try to extend the CNV by sliding a window of size w_{max} and recalculating z. Attempts to extend a CNV continue until thresholds related to read coverage and distance from the CNV end breakpoint have been reached.

Results

Datasets

To test GROM-RD's performance, we used both simulated (with known SVs) and experimental (with a large number of validated SVs) datasets for a human genome (Table 1). We first compared our approach with two commonly used RD algorithms, CNVnator and RDXplorer, on a simulated dataset. We used RSVSim (Bartenhagen & Dugas 2013) to simulate 10,000 deletions and duplications ranging from 500 to 10,000 bases using the most recent human reference genome (hg19). Deletions were heterozygous (1 copy number) and duplications ranged from 3 to 10 copy numbers. RSVSim biased SVs to certain types of repeat regions and corresponding mechanisms of formation, such as non-allelic homologous recombination, based on several studies (Chen et al. 2008; Lam et al. 2010; Mills et al. 2011; Ou et al. 2011; Pang et al. 2013). We then used pIRS (Hu et al. 2012) to simulate 100-base Illumina paired-end reads with 500 base inserts and read coverage above ten. pIRS is designed to simulate Illumina base-calling error profiles and GC bias. The simulated reads were mapped to human reference genome hg19 using BWA (Li & Durbin 2009).

We also compared CNVnator, RDXplorer, and GROM-RD on two "gold standard" datasets, one low coverage (NA12156) and the other high coverage (NA12878). Both datasets contain Illumina paired-end reads produced as part of the 1000 Genome Project (Abecasis et al. 2010) and have a large set of experimentally validated and high confidence SVs, commonly referred to as the "gold standard".

Simulation Results

CNVnator, RDXplorer, and GROM-RD prediction results for the simulated dataset are shown in Fig. 5. At least 10% reciprocal overlap between a predicted CNV and a simulated CNV was required for a true positive. Default parameters were used for all algorithms, except for the window (bin) size for CNVnator. We estimated the optimal window size for CNVnator (230 bases) by curve fitting the window size and read coverage combinations (resulting in bin size = $2205x^{-0.941}$, where x is the read

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.663v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2014, publ: 4 Dec 2014

depth) recommended by the program's authors (Abyzov et al. 2011). The default window size for RDXplorer and GROM-RD is 100 bases. For GROM-RD, we found a 100 base-window to be suitable for all datasets tested.

Table 1. Summary of simulated and gold standard datasets.

Dataset	Read Length	Insert Size	Coverage	Reference
Simulation	100	500	11x	hg19
NA12156	100	270	7x	hg19
NA12878	101	400	76x	hg18

For the simulated dataset, GROM-RD had the highest sensitivity and lowest false discovery rate (FDR, or the proportion of predictions that were false positives) for duplications. For deletions, our method also had the lowest FDR and second-best sensitivity after RDXplorer, which showed a very high FDR (0.75) when compared to GROM-RD (0.02). When the FDR is very high, it may be more informative to consider the false positive counts. RDXplorer had 13,457 false positives compared to only 61 false positives for GROM-RD. All methods had lower sensitivity and a higher FDR for deletions than duplications, which may be due to the fact that 3 to 10 copy number changes for duplications should be easier to detect than halved RD deletions.

Gold Standard Results

Prediction results for the gold standard datasets are shown in Table 2. Again, GROM-RD had the highest sensitivity for deletions and duplications in the low coverage (NA12156) dataset and duplications in the high coverage (NA12878) dataset. However, CNVnator found 39 more true deletions (10% of predicted total) than GROM-RD in the high coverage dataset.

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.663v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2014, publ: 4 Dec 2014

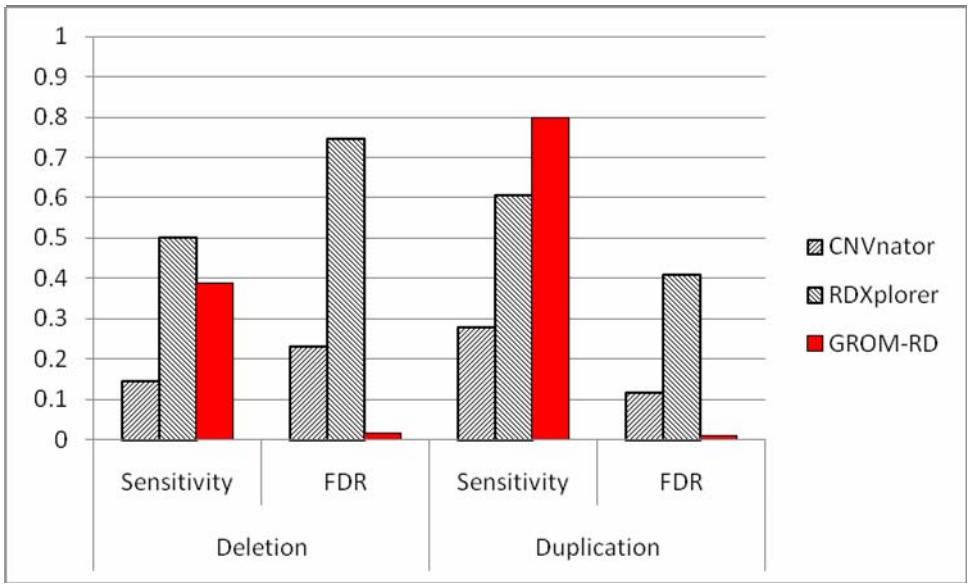

Fig. 5. Sensitivity and FDR for simulated dataset. GROM-RD had the highest sensitivity and lowest FDR for duplications. GROM-RD's sensitivity was lower than RDXplorer's sensitivity for deletions, but GROM-RD had a much lower FDR. Ten thousand deletions and duplications were simulated from human reference hg19 using RSVSim. CNVs were biased to repeat regions. One hundred-base paired-end Illumina reads with 500 base inserts were simulated at 11x coverage using pIRS and mapped to hg19 using BWA.

240

241 **Table 2.** CNV prediction results for gold standard datasets (SN denotes sensitivity, TP - true positives).

	NA12156 (low coverage)					
	Deletion			Duplication		
Algorithm	Sensitivity	True Positives	Other	Sensitivity	True Positives	Other
CNVnator	0.16	92	578	0.15	37	290
RDXplorer	0.10	56	416	0.08	20	799
GROM-RD	0.39	224	747	0.18	45	455
	NA12878 (high coverage)					
	Deletion			Duplication		
Algorithm	Sensitivity	True Positives	Other	Sensitivity	True Positives	Other
CNVnator	0.79	391	27597	0.15	34	975
RDXplorer	0.23	117	1650	0.10	22	794
GROM-RD	0.71	352	5395	0.20	45	1464

242

243 True positives indicate at least 10% reciprocal overlap between a predicted CNV and the gold standard.
244 CNV predictions not overlapping the gold standard were labeled "Other". Default parameters were used
245 for all algorithms, except for the window size for CNVnator. Using the previously described curve
246 fitting for CNVnator, we estimated 350 and 100 base windows for the low coverage (NA12156) and
247 high coverage (NA12878) datasets, respectively. We note that implementation of the dinucleotide repeat

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.663v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2014, publ: 4 Dec 2014

bias adjustment reduced GROM-RD's NA12156 and NA12878 deletion predictions by 10 and 48%, respectively, while not losing any true positive predictions. Additionally, when employing the two-pipeline approach for excessive coverage masking, deletion and duplication sensitivity increased 7 and 25%, respectively, for the low coverage gold standard dataset and 4 and 15% for the high coverage gold standard dataset.

Breakpoint Accuracy

Breakpoint accuracy is one of the traditional weaknesses of the RD methods and improvements in this area can help in narrowing down CNV borders and facilitate subsequent validation experiments. CNVnator, RDXplorer, and GROM-RD breakpoint accuracy on the simulated and gold standard datasets is summarized in Table 3. GROM-RD had the lowest deletion and duplication breakpoint error for all datasets.

Table 3. Mean breakpoint error for simulated and gold standard datasets. Lowest error for each measurement is bolded. GROM-RD had the lowest deletion (Del) and duplication (Dup) breakpoint error for all datasets.

	Simulation		NA12156		NA12878	
Algorithm	Del	Dup	Del	Dup	Del	Dup
CNVnator	278	303	4426	42507	2846	23729
RDXplorer	270	147	6267	35941	8454	27122
GROM-RD	**128**	**91**	**2538**	**29587**	**2025**	**13536**

Algorithm Metrics

Run times for the algorithms on the gold standard datasets are provided in Table 4. We tested all three programs on a single CPU (Intel Xeon E31270, 3.4 GHz) on a Linux workstation with 16 GB RAM memory. Standard BAM files were used as input. In contrast to other tools, GROM-RD's run time is relatively insensitive to read coverage with a 9-fold increase in coverage resulting in only a 20% increase in run time. GROM-RD is written in C, uses standard BAM files as input, is able to utilize paired or single reads, and is available at http://grigoriev.rutgers.edu/software/

Table 4. Run times (in minutes) on gold standard datasets. *RDXplorer outputs very large files, low I/O throughput may have affected the run time for this dataset significantly.

Algorithm	Low coverage (NA12156)	High coverage (NA12878)
CNVnator	61	206
RDXplorer	347	4378*
GROM-RD	124	149

Discussion

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.663v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2014, publ: 4 Dec 2014

We developed a novel RD approach for detecting CNVs in NGS data. Many RD algorithms, such as CNVnator and RDXplorer, correct GC bias by binning genome regions based on GC content and normalizing the read depth mean of each bin to the global average. However, read depth variance tends to vary with GC content after normalizing the means (Fig. 3). GROM-RD normalizes variance by using a quantile normalization approach to convert read depth to standard deviation units. As a result, our method produces fewer false positives overall. GROM-RD, CNVnator, and RDXplorer were tested on a simulated and two gold standard datasets. GROM-RD performed well on the simulated data having the highest sensitivity and lowest FDR. Although RDXplorer had a somewhat higher sensitivity for deletions compared to GROM-RD, it came at the expense of extreme overprediction: RDXplorer had a very high FDR resulting in 13,457 false positives compared to only 61 false positives for GROM-RD. GROM-RD had the highest sensitivity for deletions and duplications on the low coverage gold standard dataset and for duplications on the high coverage gold standard dataset. For deletions in the high coverage dataset, GROM-RD had comparable sensitivity (0.71) to CNVnator (0.79). GROM-RD's dinucleotide repeat bias normalization reduced GROM-RD's deletion predictions by 10% and 48% on the low and high coverage datasets, respectively, without reducing true positives, suggesting an improvement in specificity. As expected, duplication predictions were not affected by dinucleotide repeat bias normalization. Compared to one pipeline with no excessive coverage masking, our two pipeline approach with excessive coverage masking increased deletion and duplication sensitivity 7 and 25%, respectively, for the low coverage gold standard dataset and 4 and 15% for the high coverage gold standard dataset.

Often RD algorithms analyze read depth in non-overlapping windows with a fixed size. A read is placed in a window if the read's center (CNVnator) or start (RDXplorer) occurs in the window. Fixed-size, non-overlapping windows result in low breakpoint resolution. GROM-RD utilizes sliding windows with sizes varying in one-base increments to improve breakpoint accuracy. For all datasets, GROM-RD had the lowest deletion and duplication breakpoint error, thus improving this common weakness of RD methods.

RD algorithms are complementary to and have some advantages compared to other CNV detection methods. For instance, RD algorithms may be able to detect CNVs with rough breakpoints and duplications with few uniquely mapped reads that paired- and split-read methods may have difficulty detecting. However, RD methods frequently have low breakpoint resolution. Our results suggested that GROM-RD was able to improve RD sensitivity, specificity, and breakpoint accuracy compared to CNVnator and RDXplorer, the two most frequently used RD algorithms. Additionally, GROM-RD had a short run time that was relatively insensitive to read coverage indicating excellent scalability of the method for different datasets.

Acknowledgements

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.663v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2014, publ: 4 Dec 2014

We thank Kevin Abbey and Sulbha Choudhari of Rutgers University for excellent technical help and advice throughout the development and testing process.

Funding: This work was in part supported by the NSF grant DBI-1126052 to AG.

Conflict of Interest: none declared.

References

Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, and McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061-1073.

Abyzov A, Urban AE, Snyder M, and Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research* 21:974-984.

Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, and Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 12:R18.

Baker M. 2012. Structural variation: the genome's hidden architecture. *Nature Methods* 9:133-137.

Bartenhagen C, and Dugas M. 2013. RSVSim: an R/Bioconductor package for the simulation of structural variations. *Bioinformatics* 29:1679-1681.

Benjamini Y, and Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research* 40:e72.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara ECM, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.663v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2014, publ: 4 Dec 2014

M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, and Smith AJ. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53-59.

Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, Onofrio R, Carter SL, Park K, Habegger L, Ambrogio L, Fennell T, Parkin M, Saksena G, Voet D, Ramos AH, Pugh TJ, Wilkinson J, Fisher S, Winckler W, Mahan S, Ardlie K, Baldwin J, Simons JW, Kitabayashi N, MacDonald TY, Kantoff PW, Chin L, Gabriel SB, Gerstein MB, Golub TR, Meyerson M, Tewari A, Lander ES, Getz G, Rubin MA, and Garraway LA. 2011. The genomic complexity of primary human prostate cancer. *Nature* 470:214-220.

Boeva V, Zinovyev A, Bleakley K, Vert JP, Janoueix-Lerosey I, Delattre O, and Barillot E. 2011. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* 27:268-269.

Campbell PJ, Yachida S, Mudie LJ, Stephens PJ, Pleasance ED, Stebbings LA, Morsberger LA, Latimer C, McLaren S, Lin ML, McBride DJ, Varela I, Nik-Zainal SA, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, Griffin CA, Burton J, Swerdlow H, Quail MA, Stratton MR, Iacobuzio-Donahue C, and Futreal PA. 2010. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* 467:1109-1113.

Chen W, Kalscheuer V, Tzschach A, Menzel C, Ullmann R, Schulz MH, Erdogan F, Li N, Kijas Z, Arkesteijn G, Pajares IL, Goetz-Sothmann M, Heinrich U, Rost I, Dufke A, Grasshoff U, Glaeser B, Vingron M, and Ropers HH. 2008. Mapping translocation breakpoints by next-generation sequencing. *Genome Research* 18:1143-1149.

Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, and Lander ES. 2009. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods* 6:99-103.

Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, Handsaker RE, McCarroll SA, O'Donovan MC, Owen MJ, Kirov G, Sullivan PF, Hultman CM, Sklar P, and Purcell SM. 2012. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *American Journal of Human Genetics* 91:597-607.

Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, and Berri S. 2012. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics* 28:40-47.

Hu X, Yuan J, Shi Y, Lu J, Liu B, Li Z, Chen Y, Mu D, Zhang H, Li N, Yue Z, Bai F, Li H, and Fan W. 2012. pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics* 28:1533-1535.

Ivakhno S, Royce T, Cox AJ, Evers DJ, Cheetham RK, and Tavare S. 2010. CNAseg--a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics* 26:3051-3058.

Kim TM, Luquette LJ, Xi R, and Park PJ. 2010. rSW-seq: algorithm for detection of copy number alterations in deep sequencing data. *BMC Bioinformatics* 11:432.

Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, and Hochreiter S. 2012. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Research* 40:e69.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, and Snyder M. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318:420-426.

Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, Quinlan AR, Nickerson DA, and Eichler EE. 2012. Copy number variation detection and genotyping from exome sequence data. *Genome Research* 22:1525-1532.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.663v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2014, publ: 4 Dec 2014

Lam HY, Mu XJ, Stutz AM, Tanzer A, Cayting PD, Snyder M, Kim PM, Korbel JO, and Gerstein MB. 2010. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature Biotechnology* 28:47-55.

Li H, and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754-1760.

Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, Tothill RW, Halgamuge SK, Campbell IG, and Gorringe KL. 2012. CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* 28:1307-1313.

Lin CH, Lee GB, Fu LM, and Chen SH. 2004. Integrated optical-fiber capillary electrophoresis microchips with novel spin-on-glass surface modification. *Biosens Bioelectron* 20:83-90.

Magi A, Benelli M, Yoon S, Roviello F, and Torricelli F. 2011. Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. *Nucleic Acids Research* 39:e65.

Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, Mangano E, Battaglia C, Bonora E, Kurg A, Seri M, Magini P, Giusti B, Romeo G, Pippucci T, De Bellis G, Abbate R, and Gensini GF. 2013. EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol* 14:R120.

Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, Shago M, Moessner R, Pinto D, Ren Y, Thiruvahindrapduram B, Fiebig A, Schreiber S, Friedman J, Ketelaars CE, Vos YJ, Ficicioglu C, Kirkpatrick S, Nicolson R, Sloman L, Summers A, Gibbons CA, Teebi A, Chitayat D, Weksberg R, Thompson A, Vardy C, Crosbie V, Luscombe S, Baatjes R, Zwaigenbaum L, Roberts W, Fernandez B, Szatmari P, and Scherer SW. 2008. Structural variation of chromosomes in autism spectrum disorder. *American Journal of Human Genetics* 82:477-488.

Miller CA, Hampton O, Coarfa C, and Milosavljevic A. 2011. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One* 6:e16327.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, Fu Y, Grubert F, Hajirasouliha I, Hormozdiari F, Iakoucheva LM, Iqbal Z, Kang S, Kidd JM, Konkel MK, Korn J, Khurana E, Kural D, Lam HY, Leng J, Li R, Li Y, Lin CY, Luo R, Mu XJ, Nemesh J, Peckham HE, Rausch T, Scally A, Shi X, Stromberg MP, Stutz AM, Urban AE, Walker JA, Wu J, Zhang Y, Zhang ZD, Batzer MA, Ding L, Marth GT, McVean G, Sebat J, Snyder M, Wang J, Ye K, Eichler EE, Gerstein MB, Hurles ME, Lee C, McCarroll SA, Korbel JO, and Genomes P. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* 470:59-65.

Ou Z, Stankiewicz P, Xia Z, Breman AM, Dawson B, Wiszniewska J, Szafranski P, Cooper ML, Rao M, Shao L, South ST, Coleman K, Fernhoff PM, Deray MJ, Rosengren S, Roeder ER, Enciso VB, Chinault AC, Patel A, Kang SH, Shaw CA, Lupski JR, and Cheung SW. 2011. Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. *Genome Research* 21:33-46.

Pang AW, Migita O, Macdonald JR, Feuk L, and Scherer SW. 2013. Mechanisms of formation of structural variation in a fully sequenced human genome. *Human Mutation* 34:345-354.

Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, and Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28:i333-i339.

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, and Jaffe DB. 2013. Characterizing and measuring bias in sequence data. *Genome Biol* 14:R51.

Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S, Quackenbush J, and Nelson SF. 2011. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* 27:2648-2654.

Sims D, Sudbery I, Ilott NE, Heger A, and Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews: Genetics* 15:121-132.

Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, Werge T, Pietilainen OP, Mors O, Mortensen PB, Sigurdsson E, Gustafsson O, Nyegaard M, Tuulio-Henriksson A, Ingason A, Hansen T, Suvisaari J, Lonnqvist J, Paunio T, Borglum AD, Hartmann A, Fink-Jensen A, Nordentoft M, Hougaard D,

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.663v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2014, publ: 4 Dec 2014

Norgaard-Pedersen B, Bottcher Y, Olesen J, Breuer R, Moller HJ, Giegling I, Rasmussen HB, Timm S, Mattheisen M, Bitter I, Rethelyi JM, Magnusdottir BB, Sigmundsson T, Olason P, Masson G, Gulcher JR, Haraldsson M, Fossdal R, Thorgeirsson TE, Thorsteinsdottir U, Ruggeri M, Tosato S, Franke B, Strengman E, Kiemeney LA, Genetic R, Outcome in P, Melle I, Djurovic S, Abramova L, Kaleda V, Sanjuan J, de Frutos R, Bramon E, Vassos E, Fraser G, Ettinger U, Picchioni M, Walker N, Toulopoulou T, Need AC, Ge D, Yoon JL, Shianna KV, Freimer NB, Cantor RM, Murray R, Kong A, Golimbet V, Carracedo A, Arango C, Costas J, Jonsson EG, Terenius L, Agartz I, Petursson H, Nothen MM, Rietschel M, Matthews PM, Muglia P, Peltonen L, St Clair D, Goldstein DB, Stefansson K, and Collier DA. 2009. Common variants conferring risk of schizophrenia. *Nature* 460:744-747.

Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, Stebbings LA, Leroy C, Edkins S, Mudie LJ, Greenman CD, Jia M, Latimer C, Teague JW, Lau KW, Burton J, Quail MA, Swerdlow H, Churcher C, Natrajan R, Sieuwerts AM, Martens JW, Silver DP, Langerod A, Russnes HE, Foekens JA, Reis-Filho JS, van 't Veer L, Richardson AL, Borresen-Dale AL, Campbell PJ, Futreal PA, and Stratton MR. 2009. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* 462:1005-1010.

Xie C, and Tammi MT. 2009. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* 10:80.

Yoon S, Xuan Z, Makarov V, Ye K, and Sebat J. 2009. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research* 19:1586-1592.

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.663v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2014, publ: 4 Dec 2014