University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Li, J;Lupat, R;Amarasinghe, KC;Thompson, ER;Doyle, MA;Ryland, GL;Tothill, RW;Halgamuge, SK;Campbell, IG;Gorringe, KL

Title:
CONTRA: Copy number analysis for targeted resequencing

Date:
2012-05-01

Citation:
Li, J., Lupat, R., Amarasinghe, K. C., Thompson, E. R., Doyle, M. A., Ryland, G. L., Tothill, R. W., Halgamuge, S. K., Campbell, I. G. & Gorringe, K. L. (2012). CONTRA: Copy number analysis for targeted resequencing. Bioinformatics, 28 (10), pp.1307-1313. https://doi.org/10.1093/bioinformatics/bts146.

Persistent Link:
https://hdl.handle.net/11343/230620

License:
CC BY-NC

BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 10 2012, pages 1307–1313
doi:10.1093/bioinformatics/bts146

Genome analysis Advance Access publication April 2, 2012

CONTRA: copy number analysis for targeted resequencing

Jason Li[1,*], Richard Lupat[1,2], Kaushalya C. Amarasinghe[3], Ella R. Thompson[2], Maria A. Doyle[1], Georgina L. Ryland[2], Richard W. Tothill[4], Saman K. Halgamuge[3], Ian G. Campbell[2,5,6] and Kylie L. Gorringe[2,5,6]

[1]Bioinformatics Core Facility, [2]Victorian Breast Cancer Research Consortium Cancer Genetics Laboratory, Peter MacCallum Cancer Centre, VIC 3002, [3]Department of Mechanical Engineering, University of Melbourne, Parkville, VIC 3010, [4]Molecular Genomics Core Facility, Peter MacCallum Cancer Centre, VIC 3002, [5]Sir Peter MacCallum Department of Oncology and [6]Department of Pathology, University of Melbourne, Parkville, VIC 3010, Australia

Associate Editor: Alex Bateman

ABSTRACT

Motivation: In light of the increasing adoption of targeted resequencing (TR) as a cost-effective strategy to identify disease-causing variants, a robust method for copy number variation (CNV) analysis is needed to maximize the value of this promising technology.

Results: We present a method for CNV detection for TR data, including whole-exome capture data. Our method calls copy number gains and losses for each target region based on normalized depth of coverage. Our key strategies include the use of base-level log-ratios to remove GC-content bias, correction for an imbalanced library size effect on log-ratios, and the estimation of log-ratio variations via binning and interpolation. Our methods are made available via CONTRA (COpy Number Targeted Resequencing Analysis), a software package that takes standard alignment formats (BAM/SAM) and outputs in variant call format (VCF4.0), for easy integration with other next-generation sequencing analysis packages. We assessed our methods using samples from seven different target enrichment assays, and evaluated our results using simulated data and real germline data with known CNV genotypes.

Availability and implementation: Source code and sample data are freely available under GNU license (GPLv3) at http://contra-cnv.sourceforge.net/

Contact: Jason.Li@petermac.org

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

Received on April 5, 2011; revised on February 27, 2012; accepted on March 23, 2012

1 INTRODUCTION

Targeted resequencing (TR), including whole-exome sequencing, is becoming widely adopted as a cost-effective way to interrogate specific genomic regions across a large number of samples, a technique particularly useful for the study of genetic causes of cancer and other diseases. A number of studies have demonstrated success in the application of TR to the identification of disease-causing variants, including variants associated with a rare Mendelian disorder. Freeman–Sheldon syndrome (Ng *et al*., 2009), inherited mutations for breast and ovarian cancer (Walsh *et al*., 2010) and a single non-sense mutation that causes a syndromic form of cleft palate (Johnston *et al*., 2010). The primary objective of TR, as in above studies, is the detection of single-nucleotide variants (SNVs) and short (<50 bp) insertions and deletions (indels) within the targeted regions. Inherent limitations on sequence alignment of short reads prohibit the detection of larger indels and, therefore, many potential disease-causing copy number variations (CNVs) are not accessible from TR data. While whole-genome sequencing and single-nucleotide polymorphism (SNP) genotyping microarrays are more appropriate tools for genome-wide CNV interrogations, it is imperative that robust CNV analysis methods be developed for TR data, in order to maximize the utility of the rapidly increasing amount of TR data that are being generated globally.

CNV detection methods have been developed for whole-genome sequencing and incorporate three main aspects: the estimation of copy number breakpoint locations by segmentation using depth of coverage (DOC) (Campbell *et al*., 2008; Ivakhno *et al*., 2010; Medvedev *et al*., 2009), the incorporation of paired-end or mate pair information to enhance detection accuracies (A.Abyzov *et al*., submitted for publication; Medvedev *et al*., 2010; Miller *et al*., 2011) and the reduction of representation biases due to GC-content and other physio-chemical characteristics (Aird *et al*., 2011; Boeva *et al*., 2011; Chiang *et al*., 2009). Segmentation and bias reduction have been applied to TR data on large target regions (~40 kb) (Walsh *et al*., 2010). However, the size of a target region in most TR projects is typically small, ranging from 100 to 200 bp as in exon or whole-exome capture. In addition, genomic distribution of target regions is often sparse and uneven due to the size of intronic segments and the arbitrary locations of the genes of interest. The small size, sparseness and non-contiguous nature of target regions pose challenges to the application of existing CNV methods on TR data. Even if the DOC is high in the target regions, data resolution would be too low to make reasonable segmentation and bias reduction on a whole-genome scale. It has also been reported that the underlying assumptions made for CNV estimation in whole-genome sequencing fail to hold in the exome sequencing setting (Sathirapongsasuti *et al*., 2011). More specifically, the assumptions that genome-wide read depth is normally distributed, and that segmentation search space is continuous, fail to apply for exon capture data.

To date, a limited number of methods have been published for the CNV analysis of TR data. Exome CNV (Sathirapongsasuti *et al*., 2011) was designed specifically for whole-exome capture.

*To whom correspondence should be addressed.

© The Author(s) 2012. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

This method is based on data observed from six human samples captured using a single-exome capture platform, and involved the modeling of log-ratios using the Geary–Hinkley transformation for which a normally distributed exon-level DOC is assumed. The method, however, did not address a number of factors biasing the log-ratios, including the discrepancy in total sequence read count between the case and the control samples and the percentage of on-target reads. Other limitations include the lack of assessment of other capture platforms and the lack of a strategy to create a pseudo-control in the absence of a matched control sample.

We have developed a CNV detection tool called CONTRA (COpy Number Targeted Resequencing Analysis) for small-region TR, including data derived from exome capture. Using base-level log-ratios, copy number gain and loss of each region is inferred, with significance estimated based on the null distribution of log-ratios (Fig. 1). Our method was tested on human and mouse samples derived from seven different capture platforms, and it was evaluated using both simulated TR data and real exome data derived from well-studied HapMap individuals. CONTRA includes a module to efficiently create a pseudo-control from multiple samples. The software package interfaces with standard next-generation sequencing (NGS) formats for easy integration into analysis pipelines, taking BAM files as input and generating variant call format (VCF) files as output. CONTRA runs on Unix/Linux and Mac OS and is publicly available under GNU General Public License (GPLv3).

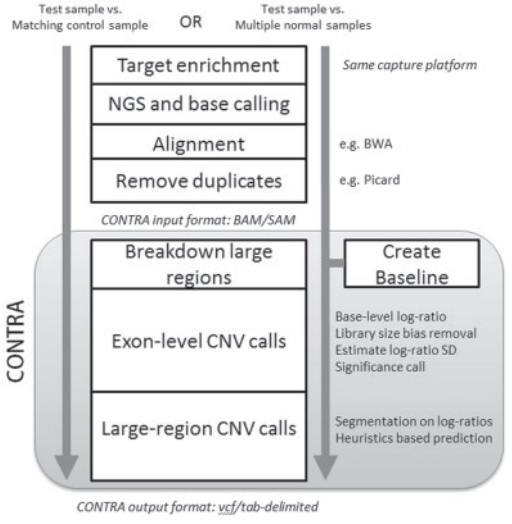


Fig. 1. CONTRA workflow. Either a matched control sample (left arrow) or a pool of normal samples for creating a baseline control (right arrow) must be present.

2 METHODS

2.1 Exome and custom exon capture data

As summarized in Table 1, Illumina GAIIx and HiSeq short-read data derived from various target enrichment platforms were assessed to develop our method of analysis. These include 56 captures across 7 enrichment platforms and include both human and mouse.

2.2 Sequence analysis

Sequence reads were aligned to the reference genome assemblies (HG19 and MM9) using BWA (Li and Durbin, 2009). Local realignment around indels and base quality score recalibration were performed using the Genome Analysis Tool Kit (GATK) software (McKenna *et al*., 2010), with duplicate reads removed using Picard (http://picard.sourceforge.net).

2.3 Creating a robust baseline from multiple samples as the control

Creating a baseline from multiple samples is essential for population and family studies where a matched control is not available. The baseline should capture the technical variation of a platform, but not variations due to CNVs or copy number polymorphisms (CNPs) in the samples. Therefore, the selection of samples should target subjects from a different background (e.g. not genetically related), so that specific CNPs can be diluted out. For baseline creation, we first define library size as:

$$L_s = N_s \times \text{read length} \times \text{percentage on target} \quad (1)$$

where L_s is the library size of sample s and N_s is the total number of short reads for sample s. Base-level coverage is then computed for each targeted base:

$$d_b = c_b \times \overline{L}/L_s \quad (2)$$

where d_b and c_b are the adjusted and raw coverage of sample s at targeted base b respectively, and $\overline{L}$ the geometric mean of all L_s and the control set.

Our definition of library size takes read length into account so that samples sequenced with different read lengths can be pooled together in the control set. We also observed that although the percentage of on-target reads is mostly stable across samples, it can vary quite significantly in some cases

Table 1. A summary of the samples that have been assessed against our characterization of depth-of-coverage and log-ratios in TR data

Species	Manufacturer	Target enrichment platform	Platform alias used	Technology	No. of samples	Sequencing	SE/PE	Type
Human	Roche NimbleGen	Sequence Capture 2.1 M Exome Array	SeqCap Array	Array based	10	GAIIx	SE and PE	Normal blood DNA
		EZ Exome Library v2.0	EZ Exome v2	Solution based	10	HiSeq or GAIIx	PE	
		EZ Exome Library v1.0	EZ Exome v1		10	GAIIx		
	Agilent	SureSelect All Exon 50 Mb	SureSelect v1		10			
		SureSelect All Exon v.2	SureSelect v2		5			
		SureSelect Custom Exon Capture	Custom Capture		10		SE	Tumor versus normal
Mouse		SureSelect Mouse All Exon	SureSelect Mouse		6		PE	

due to experimental conditions and reduced capture efficiency. Such technical variation is removed by incorporating percentage on target in Equation (1).

A robust average across the samples in the control set is then calculated as a trimmed mean of d_b at each targeted base (denoted as $\overline{d_b}$). The removal of outliers (10% both ends) aims to take out CNPs that are specific to a small number of individuals. If L_s is consistent across samples, the variance of $\overline{d}$ is inversely proportional to the size of the control set, greatly improving the stability of downstream log-ratio analysis when the size is large.

2.4 Algorithm for exon/small-region CNV detection

2.4.1 Base-level log-ratios Using either a matched control or a robust baseline, the first step of our method is to compute base-level log-ratios. This is internally achieved by the following steps: (i) coverage profiles of the samples, or bedgraphs, are created using BEDTools (Quinlan and Hall, 2010); (ii) regions with raw coverage lower than a predefined threshold are excluded from analysis (our default requires at least 10 bp with $c_b > 10$); (iii) coverage is scaled by Equation (2), with $\overline{L}$ being the geometric mean of library size between the case and the control; and (iv) log-ratios are computed for each base using the adjusted coverage. Region-level log-ratios (RLRs) are then computed by taking the mean of base-level log-ratios in the region. The use of geometric mean for scaling has been discussed in the context of RNA-seq analysis (Robinson *et al*., 2010). The reason for using base-level log-ratios, as opposed to RLRs, is to maximally remove GC-content effect on coverage, a bias that has been observed in many second-generation sequencing data, in particular Illumina (Aird *et al*., 2011).

2.4.2 Library-size bias correction Log-ratios are linearly dependent on log-coverage when the library sizes (Equation 1) between case and control are unequal (Fig. 2C). To remove the bias, a straight line is fitted between log-ratio and log-coverage using all RLRs. The fitted line corresponds to copy number neutrality, and is subtracted from each RLR in the correction step.

2.4.3 Modeling log-ratio variation against log-coverage We observed that the corrected RLRs are normally distributed for regions with similar coverage. This observation has been validated across different platforms and across a spectrum of coverage (Supplementary Fig. S1). We, therefore, model the RLRs using a normal distribution:

$$\mathrm{RLR} \sim \mathrm{N}\left(\mu_d, \sigma_d\right) \tag{3}$$

where subscript d corresponds to the adjusted coverage of the region of interest, signifying that the distribution is dependent on coverage. The distribution mean, μ_d is estimated by the mean of RLRs, which is close to zero after library size bias correction. The SD, σ_d decreases with increasing coverage (Fig. 3), and is estimated using the following procedure: (i) regions are binned based on their similarity in log-coverage; (ii) an empirical SD of RLRs, $\hat{\sigma}$ is obtained for each bin; (iii) linear interpolation is used between adjacent bins to make $\hat{\sigma}$ function of d; and (iv) σ_d is then estimated by $\hat{\sigma}(d)$. We discuss other ways of estimating σ_d in Supplementary Material.

2.4.4 Computing significance Based on Equation (3), a two-tailed P-value is computed for each region and is adjusted to reduce false discovery rates by applying the Benjamini–Hochberg multiple test correction (Benjamini and Hochberg, 1995). We also allow the user to set some arbitrary thresholds on raw read counts on case and/or control under which significance is not called due to the fact that sequencing data are extremely noisy and unpredictable at low read counts. This is useful when the normal control is expected to be diploid but has regions with very few or no reads, probably due to capture or sequencing artifacts. Excluding these regions is a recommended filtering step.

2.5 A heuristic approach for predicting large CNV

For detecting large CNVs spanning multiple target regions, we first perform circular binary segmentation (Olshen *et al*., 2004) on RLRs, using different parameters to achieve different resolutions of segmentation. Starting from the coarsest resolution (largest regions), we apply three criteria for calling a segment significant: (i) the segment has a log-ratio >0.3 (gain) or <-0.3 (loss); (ii) at least half of the regions covered by the segment must have been called significant in region-level CNV predictions; and (iii) the CNV direction (i.e. gain or loss) must be consistent between A and B. After all, the segments at a given resolution are processed, the same criteria are re-applied to the segments at a higher resolution (shorter segments). If a segment at the higher resolution overlaps with one that has been called significant at a lower resolution, the segment at the higher resolution is ignored; i.e. larger segments take higher precedence.

2.6 Simulated TR data

We simulated Illumina paired-end short reads using Chromosome 20 of the human reference assembly (hg19). The data incorporate a degrading quality score profile toward the end of the reads and substitution errors that are present in actual llumina Genome Analyzer IIx data. Our simulated data had a read length of 100 bp and a median insert size of 200 bp. We generated a control dataset and case datasets at different median coverage levels. The data cover the 4743 exons in Chromosome 20, as used by the Agilent SureSelect

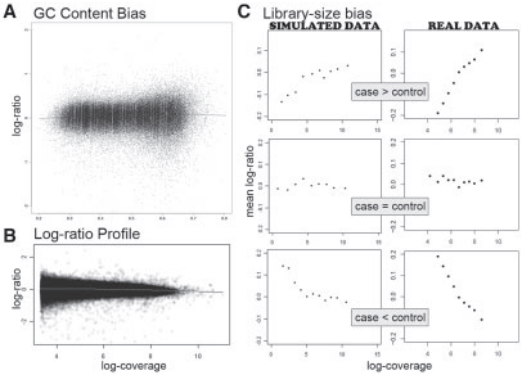


Fig. 2. Characteristics of base-level log-ratios. (**A**) Log-ratio versus GC-content; (**B**) log-ratio versus $\log_2$-coverage derived from two normal samples; and (**C**) effect of imbalanced library-size on log-ratios, for both simulated negative binomial data (left) and real data (right). The data points represent copy number neutrality. Top: library size of case sample is two times that of control; middle: equal size; bottom: case is half of control.

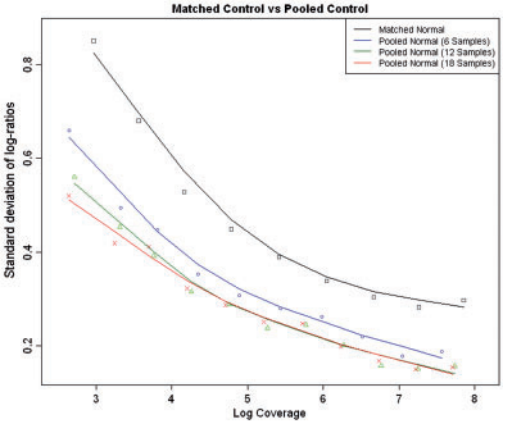


Fig. 3. Comparison of log-ratio variations between matched control and pooled controls of varying number of samples, plotting log-ratio SD against $\log_2$ coverage. The same case sample has been used throughout. Control sample(s) are subset/superset of others.

All Exon v2 capture platform. The case sample contained 311 deletions and duplications ranging from 20 bp to 10 kb, including full and partial exon deletions and duplications, as well as variations spanning multiple exons (Supplementary Table S3). Our scripts for generating the simulated data are available via the CONTRA website.

2.7 Samples from HapMap/1000 genomes project

Five human individuals that have been studied in both the HapMap (www.hapmap.org) and the 1000 Genomes project (www.1000genomes.org) were selected for evaluating the performance of CONTRA. The selection was made such that: (i) they are members of the CEU population, which is the more studied group in HapMap; (ii) exome sequencing was performed in the same Genome Centre (Beijing Genome Institute); (iii) exome capture was performed using the same assay (NimbleGen V2); and (iv) they are all the same gender (male). The HapMap sample IDs of the selected individuals are NA11893, NA12347, NA12413, NA12775 and NA12827. One individual (NA12546) was initially selected but was later excluded due to having too few CNV regions with a coverage >10x. The exome sequencing data (.bam files) were obtained from the 1000 genomes project website. The CNV genotype profiles were obtained from the HapMap website.

3 RESULTS

3.1 Characteristics of depth-of-coverage and log-ratios in target enrichment platforms

3.1.1 Depth-of-coverage We observed a large degree of variation of DOC in all the target enrichment platforms that we interrogated (Table 1). DOC variation across exons within a sample is highlighted in Figures 4A and 4B. The instability of DOC along a chromosome differentiates TR data from whole-genome sequence data, justifying the need for new developments of specialized CNV methods. A previous study (Nord *et al*., 2011) was successful in explaining and correcting the variation by using two factors, GC-content and bait-distance biases. However, their data were derived from a custom capture design targeting a few but very large regions (∼40 kb). For the off-the-shelf exome capture platforms that target much smaller regions (∼200 bp), we failed to observe sufficient correlations between DOC and GC-content or bait distance (Fig. 4C and D) that would allow similar corrections. While there is a trend of GC-content correlation, as suggested by the lowest line in Figure 4C, the correlation is weak, with DOC spreading a wide range of values at any given GC-content. Similar observations apply to the first 300 bp of the bait-distance plot (Fig. 4D), after which no correlation is observed at all (flat line).

Despite the huge variation of DOC, we observed a consistency of coverage profiles between samples captured by the same platform, a key characteristic of DOC that enables us to carry out CNV analysis for TR data. This observation has been reported for one capture platform in an independent study (Sathirapongsasuti *et al*., 2011). Here, we present similar observations on other platforms, and contrast them against the poor correlations between samples captured by different platforms (Fig. 5A). Although all capture platforms exhibit consistency in DOC profiles, the array-based platform (SeqCap Array) clearly exhibits a much larger variation of coverage, whereas SureSelect v1 and EZ Exome v2 are shown to have least variation, making them the most stable platforms for CNV analysis (Fig. 5B).

3.1.2 Log-ratios The characteristics of DOC as discussed above suggest that having a control derived from the same platform would largely reduce technical variation. We show that the base-level log-ratios between case and control are independent of GC-content (Fig. 2A), but have variations dependent on coverage (Figs 2B and 3). Also, we observed that log-ratios are linearly dependent on log-coverage when the library sizes (Equation 1) between case and control are unequal (Fig. 2C). This bias exists when the coverage data follow a Poisson (over-dispersed) or a negative binomial distribution (Fig. 5B suggests our data are negative binomial).

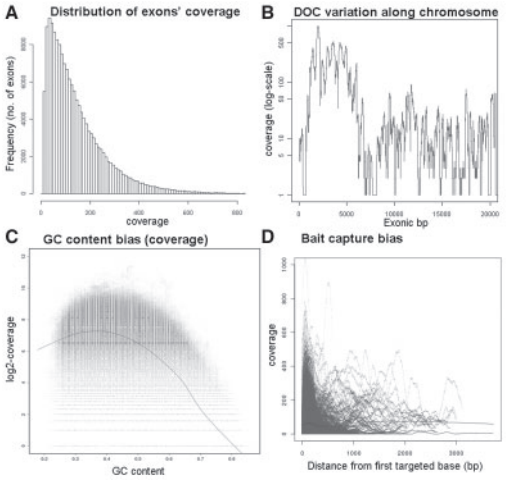


Fig. 4. Variation of DOC in TR. (**A**) Histogram of exon-level coverages in a single sample; (**B**) coverage profile along a chromosome (showing first 20 k targeted bp of Chromosome 1); (**C**) coverage versus GC-content; and (**D**) coverage versus distance from the first targeted base.

3.2 CONTRA—a novel method for CNV detection

Based on our observed characteristics of DOC and log-ratios, we have developed a novel method for CNV detection. Our method uses base-level log-ratios to remove GC-content bias, corrects for imbalanced library size effect by estimating a linear dependency between log-ratios and log-coverage, and uses a robust baseline creation strategy to reduce variations when a control set is available. The relationship between log-ratio variation and DOC is estimated empirically from the data through binning and linear interpolation. Our methods have been implemented as a software tool called CONTRA, a publicly available package that can interface with most other NGS analysis packages via standard formats, BAM/SAM and VCF. CONTRA was implemented in Python and R, and has been tested on 32/64-bit Linux (Redhat/Ubuntu) and Mac OS X. We have also made publicly available pre-calibrated baseline files for the various exome capture platforms, which can be used as a pseudo-control for samples that do not have a matched control.

3.3 Performance assessment based on simulated data

We carried out a comparison between CONTRA and a previously published R package, Exome CNV (Sathirapongsasuti *et al*., 2011), using a simulated dataset. The data contain Illumina paired-end short reads, with a mean coverage of 50x. We simulated 311 deletions and duplications ranging from 20 bp to 10 kb in the case sample and no CNP in the control sample. For a large CNV spanning multiple target regions, we consider it as a true positive if an algorithm calls more

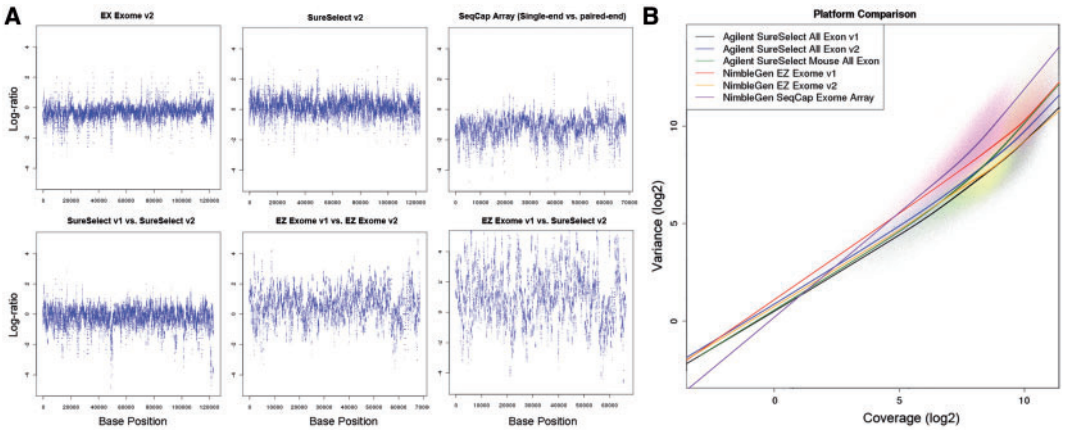


Fig. 5. Coverage correlation between samples. (**A**) Log-ratio versus targeted base position along Chromosome 20, derived from pairs of random samples as indicated in the plot titles. E.g. top-left: log-ratios between two EZ Exome v2 samples; bottom-right: an EZ Exome v1 sample matched against a SureSelect v2 sample. See also Supplementary Figure S4. (**B**) Base-level coverage variance against coverage mean, using six random samples for each platform.

Table 2. CNV detection performance over a 50x coverage simulated dataset, using default algorithmic parameters

Size of variants	No. of instances simulated	CONTRA (%)		ExomeCNV (%)	
		Sensitivity	Specificity	Sensitivity	Specificity
20–50 bp	100	57.0	99.7	8.0	100.0
50–200 bp	100	68.0	100.0	25.0	100.0
Full exons	111	96.4	100.0	62.2	100.0

than half of the regions positive. Both algorithms were run with default parameters.

Both CONTRA and Exome CNV are able to predict large CNVs better than the smaller ones (Table 2), as they were both designed to detect exon- or region-level CNVs. Exome CNV is very conservative in calling a region significant, resulting in low sensitivity and high specificity. Although the algorithm provides an option to relax specificity, no remarkable improvement has been observed in sensitivity, and many regions remain uncalled (neither positive nor negative) due to insufficient coverage.

We evaluated the performance of CONTRA at a lower coverage level (35x) across different numbers of bins. With the reduction in coverage, sensitivity for the smaller CNVs drops by ∼5%, while that for large CNVs remains unchanged. When the number of bins is relatively small (<10 bins), sensitivity is dependent on the bin size (Supplementary Table S1). This dependency, however, diminishes as the number of bins approaches optimum (between 10 and 40; Supplementary Fig. S2).

3.4 Performance on samples with known CNV

We applied CONTRA on the exome data of 5 healthy human individuals that have been studied in both the International HapMap Project (www.hapmap.org) and the 1000 Genomes Project (www.1000genomes.org). The CNV genotypes of these individuals, thoroughly profiled in the HapMap project, were used as the 'known truth' in our performance evaluation. A robust baseline is constructed from all five samples, plus an additional HapMap sample (see

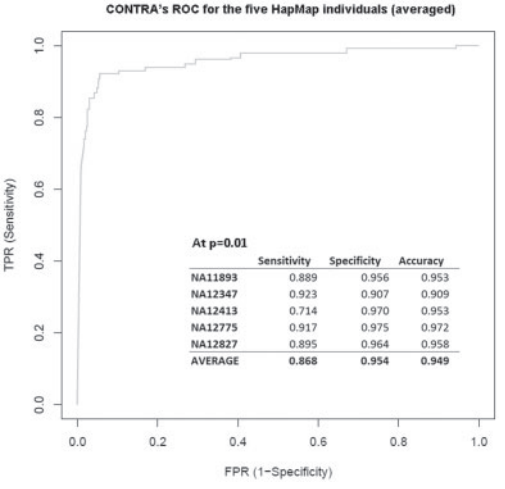


Fig. 6. Receiver operating characteristics (ROC) curve for the HapMap samples, generated by varying CONTRA's *P*-value threshold. The middle table shows sensitivities and specificities for each individual sample at a *P*-value of 0.01.

Section 2), to serve as the control. For each sample, a region is considered a real CNV if its HapMap copy number is not 2 and at least four of the remaining five samples have copy number equal to 2 for that region. Existing methods such as Exome CNV cannot be applied to this HapMap dataset due to the lack of a function to create a pseudo-control from multiple germline samples.

CONTRA achieved an average sensitivity of 86.8% and specificity of 95.4% with $P=0.01$. The trade-off between true positive rates (sensitivity) and false positive rates (1 − specificity) is summarized in Figure 6. The performance of CONTRA is comparable to the reported performance of Exome CNV on a melanoma sample with matched control (sensitivity 87%; specificity 95%; Sathirapongsasuti *et al.*, 2011). Given that the number of exons is high in a whole-exome capture, a higher specificity (e.g. 99%) is often preferred. In this case, a more stringent *P*-value is to be used in exchange for lower sensitivity rates. As will be further discussed, this

limitation is rooted in the coverage variation of target enrichment assays.

4 DISCUSSION

The technical variation of coverage across five human and one mouse exome capture platforms were compared. The one platform that exhibits a distinct coverage distribution is the older, array-based capture platform (SeqCap Array), with a slope larger than all other platforms in the (log) mean-variance plot (Fig. 5B). Its association with larger DOC variations is also apparent from the chromosome-wide log-ratio plots (Fig. 5A), making it unsuitable for CNV and other DOC-based analyses. Other solution-based capture platforms are comparable, with EZ Exome v2 being the most stable in our sample cohort. Our comparison of log-ratio plots between single-end and paired-end SeqCap Array data shows no remarkable differences. However, we have no access to further single-end data to make similar comparisons for other platforms. Given the reduction in sequencing cost, it is expected that sequencing data will be predominantly paired-end.

The use of a matched control versus that of a control set was also compared in terms of log-ratio variation (Fig. 3). As expected, using multiple samples to create a baseline coverage helps to reduce DOC variations, and in turn log-ratio variations, improving CNV detection sensitivity. As shown in Figure 3, increasing the number of samples in the control set would reduce the SD of log-ratios at any given coverage, until a minimum level of SD is reached. Therefore, even when a matched control is available, a secondary analysis using multiple unmatched normal samples as the control can be performed (in additional to a paired analysis) to improve the detection of true positives.

The main difference between CONTRA and Exome CNV is their approach to calling a region significant. Exome CNV models the log-ratios using a Geary–Hinkley transformation (Sathirapongsasuti *et al*., 2011), based on an approximation that DOC has a normal distribution of equal mean and variance. This model makes the assumption that DOC follows a Poisson distribution, which converges to normal with sufficient DOC. There are two limitations of this approach: (i) low DOC regions are not properly addressed and (ii) a Poisson model fails to capture the increasing variance at high DOC, a characteristic that is evident from the mean-variance plot of our data (Fig. 5B) and that has been discussed in the context of RNA-seq and a negative binomial model (Robinson *et al*., 2010). CONTRA, on the other hand, was developed based on empirical relationships between log-ratios and DOC that have been observed to be consistent across multiple target enrichment platforms. It relies on the case sample being largely copy number neutral, through which a null distribution of log-ratios is estimated and outlying events (CNVs) detected. This assumption has been demonstrated to hold sufficiently well for normal individuals in our evaluation using HapMap data. For non-diploid samples, the limitation of CONTRA is a reduced sensitivity in CNV detection when a large proportion of the target regions are associated with a copy number change, such as in a small custom capture of tumor material (Supplementary Materials). However, this limitation is eased in the case of whole-exome capture, where a large proportion of copy number neutral regions can be achieved. Furthermore, this limitation does not apply to family and population studies, where CNV events are expected to be rare.

The inherent limitation of TR data, regardless of analysis methods, lies in the high variation of DOC. Such a high variation makes the detection of a single copy number gain a very difficult task, as apparent from Figure 2B, where the cloud of null data spreads across the log-ratio $= 0.58$ line (1.5 fold-change). The detection of hemizygous deletions would perform better as it corresponds to a -2 fold-change (log-ratio $= -1$), but would be challenging for regions with coverage less than $\sim$30x.

In addition to CNV calls at the region (exon) level, CONTRA offers a function to predict large CNVs spanning multiple regions. However, care must be given to the interpretation of these large predictions. A large CNV segment often consists of a percentage of regions that have inconclusive CNV status (high *P*-values). They may correspond to either copy number neutral regions or low coverage regions. The strategy employed by CONTRA is to set a threshold on this percentage, above which no CNV call is made. On the other hand, the strategy used by Exome CNV is to add up DOC across the regions and then assess the overall log-ratio (Sathirapongsasuti *et al*., 2011), heavily biasing toward those regions with high coverage. Neither of these methods adequately addresses the sparseness and non-contiguous nature of target regions, an inherent limitation of TR data that restricts practical prediction to be made at the region level. For this reason, while TR (or exome sequencing) is appropriate for predicting novel single-exon CNVs and screening for known CNVs, other technologies such as genotyping microarrays must be used for the accurate predictions of larger CNVs.

For the upstream processing of sequence data, CONTRA imposes no requirements on the methods used, but it is recommended that multi-mapped reads and PCR duplicates be removed to reduce signal noise in DOC. Downstream analysis after CNV detection includes the removal of known CNPs from candidate CNV regions using public resources, such as the Copy Number Variation Project (http://www.sanger.ac.uk/humgen/cnv). This is particularly useful when a pooled control strategy is used.

5 CONCLUSION

Targeted resequencing data across seven capture platforms have been assessed in terms of coverage and log-ratio variations. We have developed a method for the detection of CNV based on empirical relationships between log-ratio and coverage. CONTRA outperforms an existing algorithm based on a simulated dataset, and is particularly suitable for population and family studies. Our methods are available as a software package, CONTRA, via http://contra-cnv.sourceforge.net

Funding: Peter MacCallum Cancer Foundation Endowment Fund, the Victorian Breast Cancer Research Consortium and Australian Research Council (grant DP1096296).

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.*, vol. 21, pp. 974–984.

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

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **57**, 289–300.

Boeva,V. *et al.* (2011) Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, **27**, 268–269.

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*, **6**, 99–103.

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.

Johnston,J.J. *et al.* (2010) Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. *Am. J. Hum. Genet.*, **86**, 743–748.

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

McKenna,A. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

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

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

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

Ng,S.B. *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272–276.

Nord,A. *et al.* (2011) Accurate and exact CNV identification from targeted high-throughput sequence data. *BMC Genomics*, **12**, 184.Olshen,A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

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

Walsh,T. *et al.* (2010) Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. In *Proceedings of the National Academy of Sciences*, Vol. 107, pp. 12629–12633.