BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 6 2014, pages 768–774
doi:10.1093/bioinformatics/btt611

Sequence analysis Advance Access publication November 4, 2013

WaveCNV: allele-specific copy number alterations in primary tumors and xenograft models from next-generation sequencing

Carson Holt[1,†], Bojan Losic[1,†,‡], Deepa Pai[1], Zhen Zhao[1], Quang Trinh[1], Sujata Syam[1], Niloofar Arshadi[1], Gun Ho Jang[1], Johar Ali[1], Tim Beck[1], John McPherson[1] and Lakshmi B. Muthuswamy[1,2,*]

[1]Ontario Institute for Cancer Research, Toronto, ON, M5G 0A3, Canada and [2]Department of Medical Biophysics, University of Toronto, Toronto, ON, M5G 2M9, Canada

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Copy number variations (CNVs) are a major source of genomic variability and are especially significant in cancer. Until recently microarray technologies have been used to characterize CNVs in genomes. However, advances in next-generation sequencing technology offer significant opportunities to deduce copy number directly from genome sequencing data. Unfortunately cancer genomes differ from normal genomes in several aspects that make them far less amenable to copy number detection. For example, cancer genomes are often aneuploid and an admixture of diploid/non-tumor cell fractions. Also patient-derived xenograft models can be laden with mouse contamination that strongly affects accurate assignment of copy number. Hence, there is a need to develop analytical tools that can take into account cancer-specific parameters for detecting CNVs directly from genome sequencing data.

Results: We have developed WaveCNV, a software package to identify copy number alterations by detecting breakpoints of CNVs using translation-invariant discrete wavelet transforms and assign digitized copy numbers to each event using next-generation sequencing data. We also assign alleles specifying the chromosomal ratio following duplication/loss. We verified copy number calls using both microarray (correlation coefficient 0.97) and quantitative polymerase chain reaction (correlation coefficient 0.94) and found them to be highly concordant. We demonstrate its utility in pancreatic primary and xenograft sequencing data.

Availability and implementation: Source code and executables are available at https://github.com/WaveCNV. The segmentation algorithm is implemented in MATLAB, and copy number assignment is implemented Perl.

Contact: lakshmi.muthuswamy@gmail.com

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

Received on June 24, 2013; revised on October 1, 2013; accepted on October 21, 2013

1 INTRODUCTION

DNA copy number variations (CNVs) are associated with a wide range of diseases including cancer where detection of copy number alterations has led to guided-therapeutic interventions. For example, amplification of the ERBB2 locus is used to identify patients for trastuzumab treatment. Although Comparative Genome Hybridization (CGH), microarrays have an intrinsic kilobase (kb) resolution for CNV detection, the advent of high-throughput next-generation sequencing (NGS) technologies offers us the potential to probe genomic structural variation at base-pair level. However, with the increase in signal resolution comes a substantially increased noise signature and the problem of how to remove false positives. Recent efforts by various groups (Abyzov *et al*., 2011; Ivakhno *et al*., 2010; Kim *et al*., 2010; Klambauer *et al*., 2012; Magi *et al*., 2011; Medvedev *et al*., 2009; Miller *et al*., 2011; Waszak *et al*., 2010; Xie and Tammi, 2009; Yoon *et al*., 2009) have attempted to mitigate the noise by carrying out a smoothing (binning) of the sequencing read depth on scales of tens to hundreds of base pairs and examining this smoothed read depth. The smoothing process is performed on a set, arbitrary, scale, which can smooth-out physically interesting features of a signal. This is of significant concern for cancer genomes, which are known to have unstable genomes that constantly evolve. Smoothing methods also assume that the noise signature of the signal is overwhelmingly concentrated on a single base-pair genomic scale (high frequency) and ignores the possibility of strong long-range (low-frequency), systemic, correlated noise that may increase the false-positive rate of any detection algorithm. Another recent effort, Varbin (Baslan *et al*., 2012) uses a variable binning approach to take into account an uneven distribution of mappable reads. Although this method is suitable for low or sparse coverage as illustrated in single cell sequencing (Navin *et al*., 2011), it does not fully harness the available base–pair-scale genomic resolution.

Assignment of digitized copy number to genomic segments in tumors is further complicated in cancer genomes due to a number of sample-specific confounding factors. For example, primary tumor tissues may contain low tumor cellularity due to an admixture of diploid/non-tumor cell fraction in patient samples, including pancreatic cancer where tumor cellularity can vary from 5 to 80%, thus making the detection of cancer driver mutations difficult (Biankin *et al*., 2012). In addition to primary tumors, patient-derived samples grown in mouse

[*]To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
[‡]Present address: Icahn Institute for Genomics and Multiscale Biology, Graduate School of Biological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029-6574.

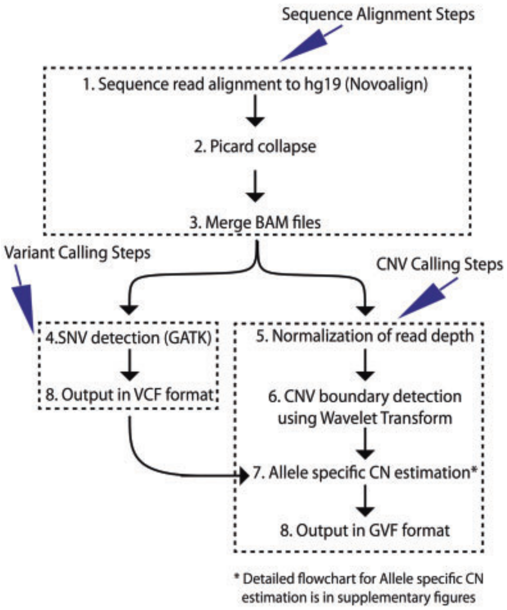
Sequence Alignment Steps
1. Sequence read alignment to hg19 (Novoalign)
2. Picard collapse
3. Merge BAM files
Variant Calling Steps
CNV Calling Steps
4.SNV detection (GATK)
8. Output in VCF format
5. Normalization of read depth
6. CNV boundary detection using Wavelet Transform
7. Allele specific CN estimation*
8. Output in GVF format
* Detailed flowchart for Allele specific CN estimation is in supplementary figures

Fig. 1. Flow chart showing the analysis procedure

xenograft (PDX) models are being increasingly used in pre-clinical settings to understand tumor biology and therapy response (Huynh *et al.*, 2011; Morton and Houghton, 2007). Assignment of digitized copy number to CNVs in these models becomes increasingly difficult due to mouse contamination of the tumor samples that introduces noise in the sequencing coverage as well as allele frequencies for SNVs (both of which are integral to CNV calling methods). Although algorithms such as qpure (Song *et al.*, 2012), genoCN (Sun *et al.*, 2009), ASCAT (Van Loo *et al.*, 2010) and ABSOLUTE (Carter *et al.*, 2012) model for stromal contamination and ploidy estimation on SNP array data, they do not function for genome sequencing data. Also these methods cannot correct for the additional effects of xenograft mouse contamination.

To fill this need, we have developed WaveCNV, a tool that uses DNA sequencing data to model for complex cancer genomes. The algorithm estimates ploidy, tumor cellularity in primary tumors, mouse content in xenograft models and assigns digitized copy numbers and alleles to indicate which parental chromosome pair was affected by each copy number event. Also to overcome limitations associated with binning-based approaches, we use the well-established theory of wavelets to take full advantage of the genomic resolution available in sequencing data. Figure 1 illustrates the overall flowchart of data generation and copy number modeling.

2 METHODS

2.1 Segmentation algorithm

Wavelet theory is used both for denoising of the depth of coverage in NGS data (which is inherently multiscale and carries non-uniform coverage signal) and to identify rapid transitions corresponding to CNV breakpoints. The wavelet transform (Mallat, 2008) breaks a given signal into different frequency components with a resolution matched to its intrinsic scale and can thus claim fundamental advantages over traditional Fourier methods in detecting sharp localized discontinuities as observed in copy number alterations. This specific property of wavelet transform is crucial in analyzing signals, specifically NGS coverage data, where size of copy number alterations can vary from base pair to length of a chromosomal arm. We give a brief description here, while the mathematical details are provided in Supplementary Materials S.1 and S.2.

We first select the wavelet basis function by using the inherent nature of copy number alterations that a genomic region with a read depth f is likely to make digitized step transitions and hence choose the simplest of all wavelets, a step function or the Haar wavelet. Given our choice of the Haar basis, we use a translation-invariant discrete wavelet transformation on the normalized read depth (Coifman and Donoho, 1995) to obtain detailed signal frequency and scale information—encapsulated by the approximation and detail coefficients. The approximation coefficients will contain both the low-frequency component (feature sizes of the order of a few kilobases) and a high-frequency component unique to sequencing data (feature sizes less than a kilobase). The *detail coefficients* will contain an exclusively high-frequency component, which is more likely to have significant noise but also possibly important small-scale insertions and deletions. We scan across scales of interest by successively iterating the decomposition of signal f, with successive *approximation* coefficients being decomposed in turn. This results in the signal being broken down into many lower genomic-resolution components starting from a small scale.

We then use de-noised approximation coefficients to define boundaries where there is a transition from one copy number state to another. Detection of breakpoints is achieved by asking when the coefficients of the maximal scale intersect those of the finest scale as given in Equation (1). For reasons of economy and because the CNV distribution is largely unknown, we examine the intersections between the approximation coefficients at entropy scale (aL^*) and the partial autocorrelation scale (aP) (Supplementary Material S.2).

$$a'_{L^p} - a'_{L^*} \equiv \mathrm{sgn}\left[\left(\frac{1}{2^{L_p}}\right)a_p - \left(\frac{1}{2L^*}\right)a_{L^*}\right] \tag{1}$$

The main point in the approach is that by examining the zero crossings of this special function in Equation (1), we should have an extremely low false-negative rate owing to the inherent sensitivity of the Haar wavelets to abrupt changes in the signal at this wide range of scales. One can show that this procedure is equivalent to searching for local maxima of the squared modulus of the dominant wavelet coefficients in the signal (Legarreta *et al.*, 2005). Figure 2 illustrates clearly that the major features of the signal discontinuities are captured by the wavelet transformed and de-noised signal, but with subtle differences in that the detail coefficients are extremely sensitive to steep-gradient features and miss gradual read depth changes that are instead captured by the approximation coefficients.

2.2 Allele-specific copy number estimation

After the breakpoints are detected using our segmentation algorithm, we assign digital copy numbers to each segment. Our basic method is similar to copy number models applied to microarray data (Sun *et al.*, 2009; Van Loo *et al.*, 2010; Wang *et al.*, 2007) with additional layers of complexity added to the model due to tumor-specific confounding factors (Supplementary Materials S.5–S.11 and methods below). We use sequencing coverage modeled as a Poisson distribution and minor allele frequency (MAF) modeled as a binomial distribution to assign digitized copy numbers to each CNV event. We also assign alleles to each copy number event describing the parental chromosome ratio following each duplication or loss. For example, a three-copy region might have an allele of 1:2 (one copy from the first parental chromosome and two copies from the other parental chromosome), whereas allele 0:3 would also be possible (three copies of one parental chromosome and complete loss of the

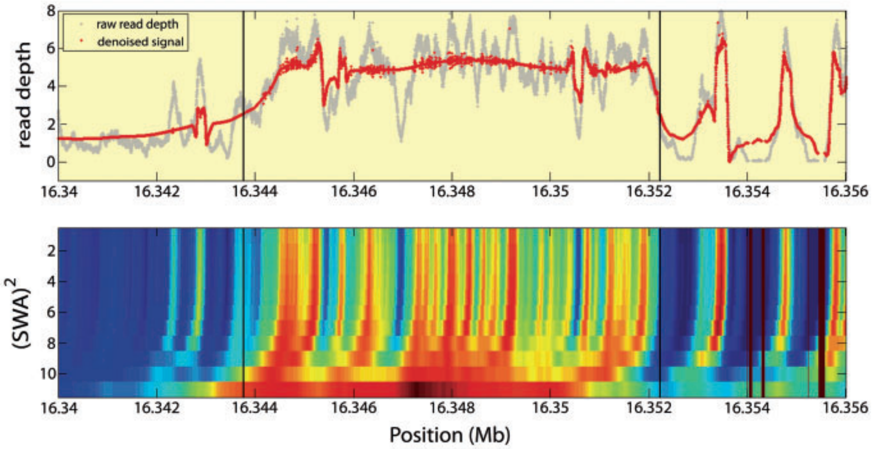

Fig. 2. Detection of signal discontinuities using wavelet transformed and de-noised signal over a 16 kb region. Top panel shows the raw read depth (gray) and the denoised signal (red). Bottom panel illustrates copy number break points where the coefficient of the maximal scale intersects those of the finest scale. The y-axis is the squared approximation wavelet coefficient, and x-axis is the genomic position in megabases

other). Alleles are assigned based on MAF distribution within the CNV event, which will be specific to chromosomal balance (e.g. a 1:2 allele would produce MAF distribution peaks at 0.33 for SNVs on one chromosome and 0.66 for SNVs on the other chromosome). Allelic assignment is possible in cancer because somatic duplication/loss events are recent, so linkage among SNVs is not expected to break down as it does in germline CNVs. The allele assignments in WaveCNV can be used to associate CNVs with SNVs/indels that appear to be preferentially gained or lost.

In addition to modeling for basic coverage and MAF, we also model for aneuploidy, normal/diploid contamination of primary tumor samples, mouse contamination of human tumors grown in xenograft and we perform auto-correction of systematic sequencing biases using matched normal/control samples. For validation purposes, we used WaveCNV to identify CNV events in human pancreatic cancer samples. Sequencing data were aligned using Novoalign (Novovcraft, Inc.) and processed using Genome Analysis Tool Kit to identify SNVs and minor allele frequencies (see Fig. 1 and Supplementary Material S.3 for data generation and S.4 for data pre-processing).

2.3 Estimation of minimum detectable CNV length

Given that coverage is modeled as a Poisson distribution, the variance for the median coverage can be approximated after adapting Raikov's theorem using the equation:

$$V = c_e/n' \tag{2}$$

where c_e is the expected segment median coverage and n' is the number of independent data points in the region (See Supplementary Material S.5). Variance is thus a function of both coverage and segment length, and a relationship can be derived to identify the minimum segment length required to identify a copy number event to a specified confidence threshold (See Supplementary Materials S.5 and S.7).

The length of all segments must then satisfy the following relationship to be detectable:

$$n' > \frac{c_i \alpha^2}{(d - 0.5)^2} \tag{3}$$

where c_i is the average expected median coverage on the region of interest, d is the difference in coverage from the neighboring segment and α is a selected threshold factor (3.890592 for 0.01%). This relationship specifies that events become detectable with either deeper coverage or longer segments, and low copy events are more easily distinguished than high copy events. Such information is invaluable because it allows us to determine the minimum sequencing coverage required before even beginning an experiment. This can be especially useful when sequencing tumor samples with diploid/normal fraction contamination that dilutes apparent separation between copy number levels. For example, the smallest events that could be identified in a primary tumor sample sequenced with 101 base pair reads and having a cellularity of 0.20 would be ~7 kb in length at 30× coverage and ~2 kb at 100× coverage. We also use this relationship to simplify our calling algorithm and improve run times by merging short segments before calculating fits to each copy number model.

2.4 Estimation of mouse contamination in xenograft models

Human derived tumors are commonly grown as xenografts in mice to facilitate continued study of the tumor's biology or increase total tumor content of low cellularity tumor types. When using these xenografted samples with NGS, mouse DNA contamination of the human-derived tumors can introduce confounding factors into both coverage and MAF, which can falsely alter the apparent copy number. The overall effect of this contamination becomes more extreme as the mouse content of the sample increases. One approach to removing mouse contamination used by tools like Xenome (Conway *et al.*, 2012) is to try and directly identify non-human sequencing reads and remove them upstream of any data processing. However, there are many conserved regions of high sequence identity between human and mouse for which sequencing reads cannot be separated in this way. Unfortunately these regions of conservation are primarily concentrated in gene coding regions (which are of main interest in cancer analysis). We thus take another approach that could be used in complement with tools like Xenome. We adjust expected coverage higher and shift expected MAF values based on the estimated amount of mouse contamination in the region (The mouse is assumed to come from an inbred line, so it will be diploid and homozygous for most mouse-specific SNVs).

Figure 3A and B show a two copy 30-megabase region observed in chromosome 1 of a human pancreatic cancer cell line that is expected to be free of mouse DNA contamination. The observed MAF distribution for the cell line (Fig. 3B red line) is centered around the MAF value of 0.5 and closely matches the calculated expected MAF distribution (Fig. 3B blue line). For the exact same two copy 30-megabase region from the same human tumor sample grown in xenograft, the observed distribution

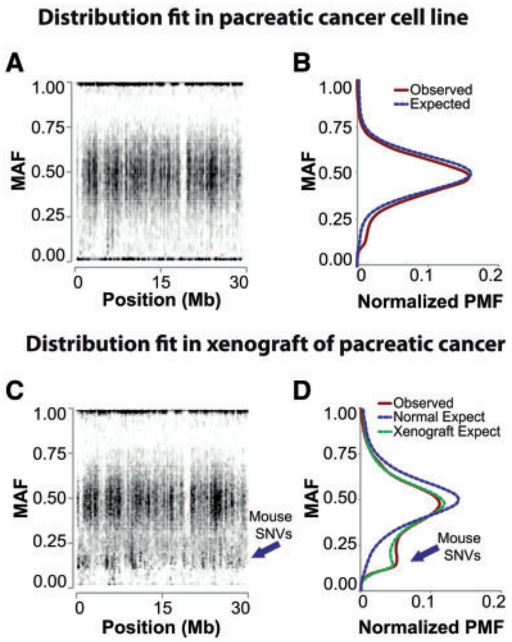

Fig. 3. MAF distribution of SNVs in a 30 Mb region of chr1. (**A**) MAF density in a pancreatic cancer cell line; (**B**) observed (red) and normal fitted expect (blue) distribution curves of MAF for pancreatic cancer cell line; (**C**) MAF density in a pancreatic xenograft model; (**D**) observed (red), normal fitted expect (blue) and expect with mouse contamination (green) for pancreatic xenograft model

of the MAF (Fig. 3D red line) is centered at 0.47, below the expected value of 0.5 (Fig. 3D blue line). There is also an observable band of data introduced by the mouse contamination around 0.16 (Fig. 3C and D blue arrows). Owing to the multi-modality of the MAF distribution, the observation deviates significantly from the expected distribution curve.

In our CNV calling algorithm, we adjust the expected MAF frequencies to take confounding factors caused by the aligning mouse reads into account by adding an independent distribution peak for mouse-derived SNVs as well as modeling for the degree that MAF peaks will be shifted by mouse reads (mouse-derived SNVs will be two copy homozygous for inbred lines). Our improved expected distribution seen in Figure 3D (green line), clearly matches the observed distribution (red line) better than the standard expect (blue line). We also alter expected coverage for the segments by estimating the quantity of mouse reads that will align (these values are fixed into WaveCNV, but can also be supplied as a BAM file if mouse was sequenced independently).

Based on kernel density estimation of mouse-expected coverage, the average mouse contamination of this particular xenograft was 21% of the total DNA content of the sample. We validated the estimated mouse contamination using qPCR. Two target loci were chosen such that one of them maps uniquely to human and another to the mouse genome. The values from TaqMan® qPCR analysis were used to calculate the relative absolute quantity between human and mouse probes, which demonstrated a 27% mouse contamination in the pancreatic xenograft compared with the tumor cell line derived from the same tumor. Thus these two alternate approaches ascertain the estimation of mouse contamination using our model within an acceptable margin of error.

2.5 Estimation of cellularity in primary tumors

Normal/diploid cell contamination of primary tumors complicates CNV calling by diluting signal from the tumor cells and reducing the amount of observed coverage separating copy number levels as well as altering the expected minor allele frequencies at each copy number level. Corrections for shift in coverage and MAF can be obtained if you know the cellularity of a sample. Previously qpure (Song *et al.*, 2012) has attempted to estimate cellularity using a relationship for the shift in MAF in the single outermost peak of loss of heterozygosity (LOH) events. Notably, however, they found that the relationship they use does not hold linear for values <20% cellularity. We followed an approach similar to theirs by using the shift in MAF for LOH events to estimate cellularity; however, we make use of LOH events at multiple copy number levels and derived a relationship that does hold linear even at low cellularity:

$$\frac{1}{M_{LOH}} = \left(\frac{T}{1-T}\right)N + 2 \tag{4}$$

where T is the tumor cellularity, N is the copy number of the region, and M_{LOH} is the left-most central MAF peak for the region at copy number N. The slope of the relationship is therefore a function of the cellularity T. Also because the y intercept of the relationship is always fixed at 2 (reciprocal of MAF 0.5), we can fit N to the proper copy number for complex aneuploidy events.

Table 1. Experimental validation of cellularity estimates

Mixed tumor fraction	WaveCNV estimate
0.05	0.043
0.10	0.088
0.15	0.155
0.20	0.236
0.40	0.403
0.60	0.602
1.00	1.00

Note: The table shows WaveCNV-derived cellularity estimates for a dilution series of diploid/normal contamination mixed into a pancreatic cancer cell line model.

Supplementary Figure S4, panel A clearly shows the outer most MAF peaks for copy numbers 1–3 of a patient-derived pancreatic cancer primary tumor sample. As shown in Supplementary Figure S4, panel B, when the MAF values from LOH peaks are used with Equation (4), the slope allows us to derive the cellularity of the sample. The resulting slope 0.611 ($R^2 = 0.99876$) corresponds to a cellularity of 0.38 for this tumor sample.

We further validated our model using a dilution series of pancreatic tumor cells derived from a primary tumor cell line mixed with increasing quantities of diploid cells derived from matched normal. Table 1 shows a convincing validation of tumor content estimation for these samples ranging from 5 to 100% cellularity. Estimates match well with expected values even for low cellularities, demonstrating the effectiveness of our method.

Identifying genomic mutational landscape has been difficult in tumor genomes where tumor content is <20%. However, using sequencing data, it may now be possible to use low cellularity tumors to detect mutational landscape if coverage is sufficient [overall coverage determines the minimum length of detectable copy number events according to Equation (3) earlier mentioned in the text].

2.6 Estimation of ploidy

One of the most difficult aspects of assigning digital copy number values to a sample is in determining what the expected coverage or copy neutral coverage would be. Many algorithms assume that the majority of a sample is diploid and any gains and losses are determined based on normalizing the coverage of each chromosome using this assumption.

This becomes problematic especially for tumor samples where the majority of the genome is often not expected to be diploid.

We have developed a procedure for identifying the base coverage corresponding to a one-copy shift that can be used to determine the ploidy of a given sample. Because multiples of the optimal value for the base coverage should correlate with the observed coverage for all segments of the genome, we perform an iterative search for a value that generates a genome-wide maximum coverage likelihood while simultaneously generating the best fit to an MAF as measured by residual sum of squares (rss). This conveniently happens at the point of maximum separation between normalized curves of coverage likelihood and rss. The overall procedure for selecting a base coverage is further detailed in the Supplementary Material S.10.

We validated our procedure using a triploid pancreatic tumor sample and its diploid matched control/normal. The expected coverage median for the diploid matched normal genome was determined to be 38 using Gaussian kernel density estimation (Fig. 4A). A search through the base coverage candidate space (Fig. 4B) using normalized coverage likelihood (red line) and normalized rss fit for MAF (blue line) reveals that maximum separation (yellow line) occurs at coverage 19.73. Given that the kernel-derived genome median coverage is 38, a base coverage of 19.73 would give a correct ploidy estimate of two for the genome. When the same procedure is applied to the triploid tumor sample (Fig. 4C and D) the base coverage is calculated to be 9.84 and the expected median coverage is 28, giving a correct ploidy estimate of three for the sample.

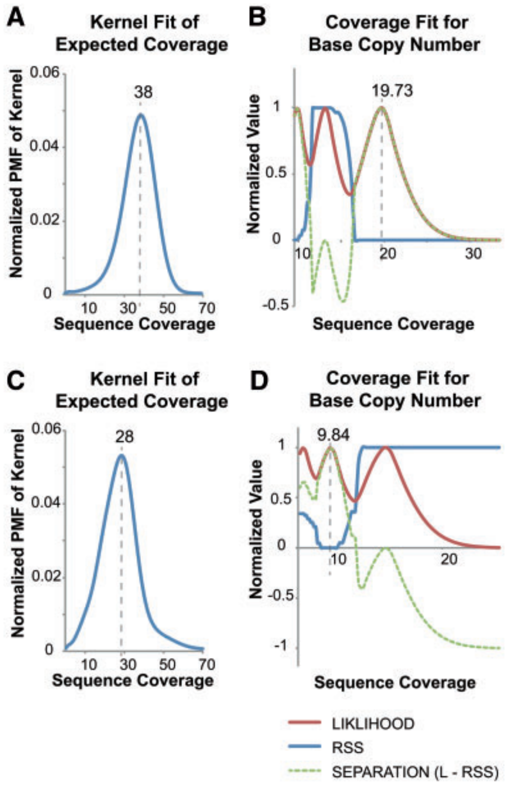

Fig. 4. Modeling for aneuploidy. (**A**) The expected segment median coverage for a diploid genome is estimated using kernel density estimation. This value then serves to define a range for estimating the sample base coverage (coverage of copy number 1). (**B**) The normalized likelihood of the observed coverage (red line) as well as the normalized residual sum of squares value (rss) for all MAF distribution fits (blue line) are calculated for each candidate base coverage (assuming ploidy range 1–4). The base coverage that produces the maximum separation between likelihood and rss (yellow line) is then selected. (**C** and **D**) show the expected segment median coverage and the base coverage selected for a triploid genome

2.7 Matched normal corrects for coverage bias and germline events

Because WaveCNV is a somatic CNV caller, as CNVs are assigned to the tumor sample it can simultaneously assign copy numbers to the same segment in the diploid matched control (sequenced together with the tumor). This allows WaveCNV to determine if given losses, gains and LOH events are in fact somatic or germline events. Furthermore, the matched normal also allows WaveCNV to correct for anomalies present in the reference sequence including systematic variance in coverage, high repeat regions, unsequencable regions with consistently missing coverage and so forth. Supplementary Figure S5 clearly demonstrates the decrease in variance for genomic coverage when a matched control based correction is applied (blue line) as opposed to the standard coverage distribution (red line). Final somatic calls are highlighted in the output report to distinguish them from other copy number calls, thus allowing researchers to immediately focus on the events most likely to be important to tumor progression. Further details for matched normal/control-based correction are found in the Supplementary Materials S.11.

3 RESULTS

We identified 764 somatic copy number aberrations in pancreatic cancer genome sequencing data using WaveCNV. The size of CNV events varied from 284 bp to 33 Mb. Supplementary Table S4 lists these events along with their verification status using alternate platforms, and Supplementary Figure S7 illustrates the size distribution of those events.

3.1 Ascertainment of somatic CNVs using microarray and qPCR technologies

CNVs identified using WaveCNV were verified using three alternate technologies: Nimblegen, 2.1 million CGH tiling array, Illumina 1 million Omni-quad SNP array and verification of 80 CNV loci with copy number varying from 0 to 30 using qPCR. We find a high correlation between CN estimated from qPCR method and WaveCNV as shown in Figure 5A. We fit linear regression and found that regression coefficient (0.94) with $P < 2\text{e-}16$. With the regression coefficient close to 1, it confirms that our CN model used in WaveCNV algorithm is able to predict accurately a wide range of copy numbers. We also compared CNVs from the whole genome with two different array-based platforms, Illumina Omni 1 M quad and Nimblegen 2.1 M array CGH. Invariably, most of the array platforms have lower dynamic range compared with sequencing that results in approximate digitized CN. Hence we compared CN from sequencing to the median intensity signal of the probes covering the region from array platforms as shown in Figure 5B and C. We find a high concordance between array platforms and WaveCNV. The weighted Pearson correlation coefficients are calculated to be 0.86 for Illumina array and 0.97 for Nimblegen array with weights proportional to the length of the segment.

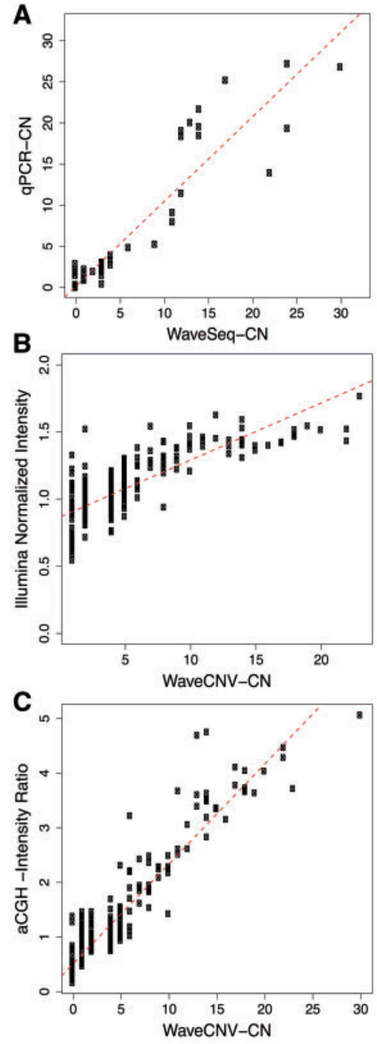

Fig. 5. Validation of copy number calls using three methods. (**A**) Verification of 80 CNV loci by qPCR on a pancreatic cancer genome. Copy numbers from qPCR were estimated based on threshold cycle (Ct) values. The Pearson correlation coefficient is 0.94. (**B**) Verification of 473 somatic CNVs on the whole-genome using Illumina Human Omni 1Million microarray. Shown here is the concordance between intensity ratios in microarray to WaveCNV CN. The Pearson correlation coefficient is 0.86. (**C**) Verification of 468 somatic CNVs on the whole genome using Nimblegen 2.1 Million aCGH microarray. Shown here is the concordance between aCGH intensities ratio in microarray to WaveCNV CN. The Pearson correlation coefficient is 0.97

3.2 Algorithm performance comparison

We additionally compared our results to the sequencing-based CNV calling algorithms CNVnator (Abyzov *et al.*, 2011) and OncoSNP-SEQ (Yau, 2013). We used base pair level congruency between algorithm calls to compare matches. We define congruency to be the average of sensitivity (the fraction of a reference feature predicted) and specificity (the fraction of a prediction overlapping a reference feature). In all cases the reference is the algorithm we are comparing with.

Table 2 shows the comparative statistics between the three algorithms. Comparing copy number events observed in WaveCNV with CNVnator, we observed an overall congruency of 93% (95% in gains and 92% in losses). When comparing WaveCNV to OncoSNP-SEQ (a cancer-specific CNV caller), we see an overall congruency of 80% (87% for amplifications and 80% for deletions). The lower match for OncoSNP-SEQ is primarily due to our sample coverage being lower than that recommended for accurate OncoSNP-SEQ performance (our sample was sequenced to 30×, whereas OncoSNP-SEQ requires a minimum of 60×).

Our concordance with CNVnator is one of the highest reported so far between any two programs for sequencing data thereby supporting the validity of our algorithm. There are key additional features that are unique to our algorithm, which are critical for cancer genomes. WaveCNV successfully combines the read depth distribution, MAF and reference-based normalization of tumor with matched normal to estimate ploidy of the genome and corrects for mouse contamination with the additional benefits of copy number allele assignments and LOH detection. We have a well-defined mechanism to control for detectable event sizes at different levels of sequencing coverage and tumor sample cellularity. Also although WaveCNV can assign copy numbers to any segment within the genome, the primary focus of cancer research is on the somatic changes and somatic CNVs are identified in our output by integrating matched normal/controls into the copy number analysis.

4 DISCUSSION

We have developed a computational algorithm to detect CNV boundaries from whole-genome sequencing data and assigned digitized copy number by modeling for sample-specific confounding factor such as aneuploidy, normal/diploid contamination of primary tumors and mouse contamination in xenograft models. The segmentation algorithm based on wavelet transform provides a unique opportunity to probe the genome in any

Table 2. WaveCNV comparison to other algorithms

Algorithm	Events	Gains	Losses	Total basepair gains	Total basepair losses	Congruency gains	Congruency losses	Congruency all
WaveCNV	764	359	405	312 922 439	567 442 194	–	–	–
CNVnator	3658	829	2829	319 703 400	622 106 100	0.95	0.92	0.93
OncoSNP	1423	567	856	260 783 488	912 819 293	0.87	0.80	0.80

Note: This table shows the base pair level congruency in copy number alterations called by WaveCNV compared with CNVnator and OncoSNP-SEQ.

spatial genomic scale. Although the first part of the algorithm identifies all discontinuities, the second part of WaveCNV provides a statistical framework to assign CN and merge neighboring events carrying the same copy number. This corrects for false shearing of copy number events that may arise due to poor quality of sequencing data.

A key component of WaveCNV is the matched–normal-based copy number correction. Being aware of the diploid control ensures that any systemic artifacts that may appear in both tumor and normal genomes, including platform-specific biases, unsequenceable regions and so forth, are effectively removed or corrected for. This resulted in a high concordance between somatic CN calls from our algorithm in sequencing data to both microarray data and qPCR.

Xenograft models for many types of primary tumors have increasingly become useful tools to understand cancer biology and to test therapeutic targets. Our model estimates mouse contamination and the reported allele and copy number reflects the correction for mouse contamination. The mouse contamination estimate matches well with our mouse-specific qPCR data. On the same note, most directly sequenced primary tumor samples contain stromal contamination, and our algorithm can quantify and model for the presence of contaminating diploid cells in that sequencing data.

5 CONCLUSION

Our segmentation algorithm is unique from its methodology perspective, and can potentially improve the boundary assignments on the smaller CNV events found via whole-genome sequencing. In addition, the assignment of specific alleles to copy number losses/gains can give researchers the ability to explore relationships between selected sequence mutations and structural variation. For example, in pancreatic cancer a KRAS activating point mutation is often coupled with duplication events, thus amplifying the effect of this oncogene. Being able to identify similar correlations based on reports from our algorithm could prove useful in prioritizing-specific genes for further study.

ACKNOWLEDGEMENTS

The authors acknowledge support from the Ontario Institute for Cancer Research, Pancreatic Cancer Genome initiative. They thank the OICR genome production sequencing group. They thank Dr Lincoln Stein (OICR) for valuable comments on the manuscript and Dr Christopher Yau (University of Oxford) for discussions on running OncoSNP-SEQ.

Funding: New Investigator award, Ontario Institute for Cancer Research, supported by Ontario Ministry of Research and innovation.

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.

Baslan,T. *et al.* (2012) Genome-wide copy number analysis of single cells. *Nat. Protoc.*, **7**, 1024–1041.

Biankin,A.V. *et al.* (2012) Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature*, **491**, 399–405.

Carter,S.L. *et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, **30**, 413–421.

Coifman,R.R. and Donoho,D.L. (1995) Translation-invariant de-noising. In: Antoniadis,A. and Oppenheim,G. (eds) *Wavelets and Statistics*. Springer-Verlag, New York.

Conway,T. *et al.* (2012) Xenome—a tool for classifying reads from xenograft samples. *Bioinformatics*, **28**, i172–i178.

Huynh,A.S. *et al.* (2011) Development of an orthotopic human pancreatic cancer xenograft model using ultrasound guided injection of cells. *PLoS One*, **6**, e20330.

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.

Kim,T.M. *et al.* (2010) rSW-seq: algorithm for detection of copy number alterations in deep sequencing data. *BMC Bioinformatics*, **11**, 432.

Klambauer,G. *et al.* (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.*, **40**, e69.

Legarreta,I.R. *et al.* (2005) A comparison of continuous wavelet transform and modulus maxima analysis of characteristic ECG features. *Comput. Cardiol.*, **32**, 755–758.

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

Mallat,S. (2008) *A Wavelet Tour of Signal Processing*. 3rd edn. The Sparse Way. Academic Press.

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

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

Morton,C.L. and Houghton,P.J. (2007) Establishment of human tumor xenografts in immunodeficient mice. *Nat. Protoc.*, **2**, 247–250.

Navin,N. *et al.* (2011) Tumour evolution inferred by single-cell sequencing. *Nature*, **472**, 90–94.

Song,S. *et al.* (2012) qpure: a tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles. *PLoS One*, **7**, e45835.

Sun,W. *et al.* (2009) Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Res.*, **37**, 5365–5377.

Van Loo,P. *et al.* (2010) Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA*, **107**, 16910–16915.

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.

Waszak,S.M. *et al.* (2010) Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity. *PLoS Comput. Biol.*, **6**, e1000988.

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.

Yau,C. (2013) OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinformatics*, **29**, 2482–2484.

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