Bioinformatics Advance Access published August 20, 2013

Genome analysis

Determination of genomic copy number alteration emphasizing a restriction-site based strategy of genome re-sequencing

Caihong Zheng[1,2], Xuexia Miao[1,2], Yanen Li[3], Ying Huang[4], Jue Ruan[1], Xi Ma[1], Li Wang[5], Chung-I Wu[1,6*], Jun Cai[1*]
[1]Laboratory of Disease Genomics and Personalized Medicine& Center of Computational Biology , Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, 100029, China, [2]University of Chinese Academy of Sciences, Beijing, 100049, China, [3]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA, [4]Regeneron Pharmaceuticals, Inc., Tarrytown, New York, 10591, USA, [5]Department of Epidemiology, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, School of Basic Medicine, Peking Union Medical College, Beijing, 100730, China, [6]Department of Ecology and Evolution, University of Chicago, Chicago, IL, 60637, USA

Associate Editor: Dr. Janet Kelso

ABSTRACT

Motivation: Copy number alteration (CNA) is one type of genomic aberration that is often induced by genome instability and is associated with diseases such as cancer. Determination of the genome-wide CNA profile is an important step in identifying the underlying mutation mechanisms. Genomic data based on next-generation sequencing technology is particularly suitable for determination of high-quality CNA profile. Now is an important time to reevaluate the use of sequencing techniques for CNA analysis, especially with the rapid growth of the different targeted genome and whole-genome sequencing strategies.

Results: In this study, we provide a comparison of resequencing strategies, with regard to their utility, applied to the same hepatocellular carcinoma (HCC) sample for copy number determination. These strategies include whole-genome, exome, and restriction site-associated DNA (RAD) sequencing. The last of these strategies is a targeted sequencing technique that involves cutting the genome with a restriction enzyme and isolating the targeted sequences. Our data demonstrate that RAD sequencing is an efficient and comprehensive strategy that allows the cost-effective determination of CNAs. Further investigation of RAD sequencing data led to the finding that a precise measurement of the allele frequency would be a helpful complement to the read depth for CNA analysis for two reasons. First, knowledge of the allele frequency helps to resolve refined calculations of allele–specific copy numbers, which in turn identify the functionally important CNAs that are under natural selection on the parental alleles. Second, this knowledge enables deconvolution of CNA patterns in complex genomic regions.

Contact: juncai@big.ac.cn
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Genome instability, including nucleotide mutations, chromosomal rearrangements, and DNA dose aberrations, promotes genetic variation. Among these possible genomic aberrations, DNA copy numbers are of high diversity in normal human populations(Sharp *et al.*, 2005) and are associated with various human diseases, such as birth defects (X.-Y. Lu *et al.*, 2008) and neurodegenerative diseases (Walsh *et al.*, 2008; B. Xu *et al.*, 2008). Notably, tumourigenesis involves a process of accumulation of chromosomal mutations that induce the activation of oncogenes, the loss of function of tumour suppressors, and cell proliferation. Copy number aberrations (CNAs), in the form of regions of somatic amplification or deletion, are an important subclass of chromosomal mutation associated with tumourigenesis(Pinkel and Albertson, 2005; Michael R Stratton *et al.*, 2009).

Investigations of CNAs and the biological mechanism through which they occur have shed new light on the human genome variations that exist among individuals with disease or normal phenotypes. Many studies have examined the differences in CNAs between tumour cells and normal cells and have estimated the rearrangement phylogeny of cancer genomes on the basis of their CNAs (Peter J Campbell *et al.*, 2008; Erin D Pleasance *et al.*, 2010; Tao *et al.*, 2011; C D Greenman *et al.*, 2011; Navin *et al.*, 2011). CNAs, as well as single nucleotide polymorphisms (SNPs), in humans have revealed extensive genetic diversity in populations (Sudmant *et al.*, 2010; W. Fu *et al.*, 2010). A framework based on evolutionary genetics has been adopted to understand the disease-causing deleterious CNAs or beneficial CNAs present in human populations (Nozawa *et al.*, 2007; Gregory M Cooper *et al.*, 2007; Perry *et al.*, 2007; K. W. Lee *et al.*, 2011; Elia *et al.*, 2011).

The successful characterisation of copy number profiles is the first step in pinpointing the CNAs that have significant biological roles in disease occurrence and normal phenotypic variation. In the past 15 years, profound advances have been made in the major technologies offering genome-wide scanning of CNAs, such as array-based comparative genomic hybridisation (aCGH) with high-density oligo probes (Solinas-Toldo *et al.*, 1997; Bentz *et al.*, 1998), SNP genotyping arrays (M. Lin *et al.*, 2004), and recently developed next-generation sequencing-based approaches (Shendure and Ji, 2008). Array-based technologies generate analogue fluorescence signals that are prone to noise (Pinto *et al.*, 2011; Haraksingh *et al.*, 2011; Xi *et al.*, 2011), and short-read sequencing platforms provide an effective alternative way of identifying CNAs. Next-generation sequencing technologies provide high-resolution, nucleotide-level digital readouts of the genomic composition of cells (Ding *et al.*, 2010; Garvey, 2010; Mardis, 2011). Especially since the genomes of more and more species have been sequenced and assembled, resequencing has become increasingly more convenient for CNA analysis than the existing array-based technologies, which require significant work to design new sets of probes. Resequencing strategies for genomes, including exome sequencing and whole genome sequencing (WGS), confer benefits in detecting CNAs and rearrangements (Xi *et al.*,

*To whom correspondence should be addressed.

2011; Sathirapongsasuti *et al.*, 2011; Alkan *et al.*, 2011; Medvedev *et al.*, 2010). Both the read depth and the minor allele frequencies (MAFs) of heterozygous loci estimated from the sequencing data are informative in inferring copy numbers. Low minor allele frequencies have been utilised, in addition to read depth, to screen for CNA segments conferring adaptive advantages and to confirm the breakpoints of CNAs (Tao *et al.*, 2011; Nguyen *et al.*, 2006; Xie and Tammi, 2009; Abyzov *et al.*, 2011; Krumm *et al.*, 2012; Simpson *et al.*, 2010).

However, sequencing biases can be introduced from many sources and may negatively impact the precise measurement of read depth and allele frequency throughout the genome. Uneven sampling in DNA fragmentation during library preparation, PCR bias due to the GC-content of fragments, and errors in the mapping of reads are all sources of variability in the characterisation of genomic regions. In addition to these sources of bias, exome sequencing has an excessive allele bias due to the unbalanced sampling of the two alleles that can result from differences in probe affinity (Asan *et al.*, 2011). Unlike deep-coverage (>20X) WGS data, low-coverage (1X~3X) WGS data fail to provide important information regarding the allele frequencies of heterozygous loci. Efforts have been made to overcome these shortcomings in the sequencing data, and both GC correction and hidden Markov models have been integrated into the statistical tools used to detect CNAs; however, the success of these efforts has been limited (Xie and Tammi, 2009; Abyzov *et al.*, 2011; Krumm *et al.*, 2012; Yoon *et al.*, 2009). Restriction site-associated DNA (RAD) sequencing is a targeted sequencing technique that involves cutting genomic DNA with at least one restriction enzyme and isolating the target sequences from entire genomes. RAD sequencing was developed to efficiently identify single nucleotide polymorphisms, map QTLs, and measure the genetic structure of natural populations of various non-model organisms (Baird *et al.*, 2008; Robinson *et al.*, 2012; P A Hohenlohe *et al.*, 2010; Davey *et al.*, 2011). Intuitively, the majority of CNA segments would be covered by densely packed genome-wide RAD reads. Hence, the RAD sequencing strategy holds promise for CNA analysis. Herein, we proposed this wet-lab strategy, RAD sequencing, to cover approximately 5% of the genome of a hepatocellular carcinoma (HCC) sample for optimal CNA analysis. Using the same sample, we systematically compared the performance of RAD sequencing with that of the most commonly used alternative strategies in constructing the CNA profile. From this study, there is sufficient evidence to support the notion that the RAD sequencing strategy, which allows a precise measurement of both the read depth and allele frequency, is a comprehensive solution to the problem of CNA characterisation.

2 METHODS

2.1 Sample Information

We recruited a female patient who had a chronic Hepatitis B Virus (HBV) infection and who was diagnosed with HCC at the age of 35. The tumour section was determined to be grade II to III HCC with prominent clear cell components. The adjacent non-tumour tissue was dissected as a control.

2.2 RAD Sequencing

Fragments adjacent to restriction sites were collected based on the method designed by N A. Baird et al(Baird *et al.*, 2008). 2ug of genomic DNA from the tumour sample was digested for 30 min at 37°C using 1 µl EcoRI-HF in a 50-µl reaction volume. A modified P1 adapter with a single 5'-AATT-3' cohesive end was added to the samples at the same molarity as the tags in a 20 µl T4 ligase reaction system. Restriction digest by an enzyme with a palindromic target sequence enabled the ligation of the adapter to the upstream and downstream sequences flanking the restriction sites. Following sonication, the fraction of the sonicated ligation products corresponding to a fragment size range of 250-350 bps was retrieved. The DNA fragments were then end-repaired, ligated to modified illumina P2 adapters, and amplified. To increase the sequencing specificity for sequences adjacent to restriction sites, we adjusted the sequence of the sequencing primer to ACACTCTTTCCCTACACGACGCTCTTCCGATCTAATTG by adding 5'-AATTG-3' to its 3' end. In RAD sequencing, 100 bps single-end reads were generated using a Hiseq 2000. We designed a compact reference based on the framework of the human NCBI36/hg18 genome assemble which consisted of groups of 100 bps genomic sequences flanking the known EcoRI restriction sites. The BWA software was used to align the raw RAD sequencing data. Total 80X RAD data for the tumour sample were subjected to downstream analysis.

2.3 ExomeSequencing and WGS Sequencing

For exome sequencing, an Agilent SureSelect Human All Exon Kit was used to capture 1.22% (37 Mb) of the human genome and the reads were sequenced with SOLiD(Tao *et al.*, 2011). We utilized the BWA program for the alignment of SOLiD reads to the exome regions of NCBI36/hg18 whole genome assembly.56X uniquely mappable reads were obtained. Whole-genome reads for the same tumour sample were sequenced according to manufacturer's standard protocols on two platforms: Solexa and SOLiD (Tao *et al.*, 2011).The reads were mapped to the NCBI36/hg18 whole genome assembly. We utilized the BWA program for the alignment of both SOLiD reads and Solexa reads with the default parameters. Total 20X uniquely mappable reads for the tumour sample were subjected to downstream analysis. WGS data (20X) were also generated from the adjacent non-tumour sample for use as the control (Tao *et al.*, 2011).

2.4 Statistics describing the precise measurement of read depth profiles

We adopted the mean value of the windowed coefficient of variation as a measure of the dispersion of the read depth signals. The windowed coefficient of variation (WCV), a normalised measure of the dispersion of a probability distribution, is defined as the ratio of the standard deviation σ to the mean µ in a window:

$$WCV = \sigma_w / \mu_w ,$$

Another statistic used in our analysis is the absolute difference of Gaussians (ADOG), which is defined as:

$$f(x;\mu_1,\mu_2,\sigma) = \left| \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right) - \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma^2}\right) \right| ,$$

The integral of the ADOG function reflected the visibility of the edges and was used to measure the edge sharpness distinguishing the CNA breakpoints. In practice, the parameters of two distributions were estimated with the read depths of all the loci in the 5M-length window upstream and downstream to each breakpoint, respectively. The integral value of ADOG for each breakpoint was computed based on the estimated parameters.

2.5 Estimating read depth profiles and minor allele frequency

We defined N consecutive restriction sites as a unit and calculated the read counts in the unit to represent the read depth profile for comparisons across different sequencing strategies. Here, low-coverage WGS data covered the whole genome at a depth of approximately 3X and were collected from two lanes of Solexa sequencing. The read depth profiles were generated with various N values from RAD and WGS deep-sequencing data. Box-plots of the windowed coefficient of variation (WCV) showed that a smooth and stable measurement of the dispersion of the read depth profiles was achieved when we chose N values of no less than 10 (Supplementary Figure S1). Therefore, 10 consecutive restriction sites were treated as a unit in

this study. We took two-step method to define the region without genomic alteration under the two assumptions that the percentage of genome altered might be larger than 50% and there are some balance amplifications of both alleles. Firstly, we chose candidate regions where the heterozygous sites had MAF of 0.5. Next, the probability distribution function of read depth was estimated for loci within the candidate regions and then the read depth with maximum probability value was determined as the value taken to normalize the data. In our data, this value was almost the same as the median value of read depths for loci within the candidate regions.

Heterozygous sites were extracted from the deep-sequencing WGS data obtained from the adjacent non-tumour sample according to the following criteria: (1) the candidate sites were germ-line mutation sites that were included in the SNP database dbSNP130; (2) the total read depths of the sites were not lower than 20; (3) the estimated minor allele frequencies of the sites in the adjacent non-tumour sample were not less than 0.3; (4) the interval between two candidate sites was more than 10bps. The minor allele frequency (MAF) of each heterozygous site was estimated using the read depth of each allele from RAD, deep WGS, and exome sequencing data from the tumour sample. Usually, the MAF baseline was below 0.5 because of marginal effects of the integer read depth ratio. The MAF baseline was approximately 0.4 for the WGS and 0.45 for the RAD sequencing.

2.6 Detection of copy number alterations

We used the JointSLM algorithm, which employs two independent stochastic processes by means of the Shifting Level Model (SLM), to split the read depth profiles into distinct segments (Magi *et al.*, 2011). Then, the candidate CNAs were identified after merging neighbouring segments with similar copy numbers using the FastCall algorithm (Benelli *et al.*, 2010). The above computational process is well established and widely used. The same process was applied to all sequencing data. Three non-CNA regions that were supported by both the read depth and minor allele frequency values obtained from all of the sequencing strategies were validated via qPCR to calibrate the baselines of the read depth profiles. Deletion and amplification were defined according to whether the changes in read depth values observed for the regions corresponded to the gain or loss of at least 20% relative to the baseline. We calculated the average minor allele frequency of each CNA based on all the heterozygous sites within the CNA region. The criterion that the average MAF differed from the frequency baseline by 0.05 was used to qualify the fidelity of the CNA segments.

2.7 Defining low- or ultra-mappable regions and abnormal mappability

We simulated 2×100 paired-end reads, covering the human genome at approximately 50X, to assess the mappability of a diploid genome. The read counts in a 4-kb window were calculated and calibrated by dividing the median value across the whole chromosome. Segmentation of the diploid genome was performed by the standard pipeline combining the JointSLM and FastCall algorithms. The low- and ultra-mappable regions were delimited as segments with outlier read depth values (20% change) compared to the baseline of 1.0. The CNAs in which more than 50% of the sequence spanned regions that were low- or ultra-mappable were defined as those that were associated with abnormal mappability.

2.8 Genotyping validation of allele frequencies by Sequenom

The Sequenom™ platform was employed to screen approximately 58 heterozygous sites with differential allele frequencies identified across different sequencing strategies in order to assess the precision of MAF estimation from sequencing data. The frequency measurement at each site was replicated three times to account for technical variation associated with the method. Genotyping primers and extension probes were on-line designed with default parameters and were confirmed by in-silico PCR amplification prior to synthesis and standard purification. Mass spectrometric genotyping using TypePLEX chemistries was conducted on genomic DNA from the tumour sample. The heights of the raw spectral peaks were quantified, and the mutant allele percentage was determined using the default settings of the MassARRAY Typer 4.0 Analyzer.

2.9 qPCR validation of segments occurring in CNAs

Primers were designed using Primer3 with default settings to limit the amplicon length to 150–200 bps. Each primer was shown to amplify a unique site in the genome via in-silico PCR. qPCR experiments were conducted using MaximaTM SYBR Green qPCR Master Mix in triplicate reactions according to the product manual. The reactions were amplified on a BioRad CFX96TM real-time system. The resultant crossing thresholds (Ct) were analysed using the ΔΔCt method. Two DNA templates derived from the blood of two healthy female donors were adopted as the controls to calibrate the ΔΔCt values in the experiments. We assumed that all q-PCR segments were diploid for the DNA templates from the healthy donors.

The false positives were identified using the formula:

$$\min_{i=1,2}\left|\overline{CopyNumber}-qPCR_i\right| > 2\theta \text{ , where } \overline{CopyNumber}$$

was the estimated copy number using read depths based on deep WGS, low WGS, or RAD sequencing data, and qPCR was the value obtained in two qPCR validations. The threshold θ was defined as the maximum distance between the estimated copy numbers across multiple sequencing strategies and the validated copy numbers via qPCR in the non-CNA segments.

3 RESULTS

3.1 Overall CNA profiles from different resequencing strategies

We applied RAD sequencing on a hepatocellular carcinoma (HCC) sample to screen for significant copy number aberrations during tumourigenesis with EcoRI enzyme. Single-end reads of 100 bps were obtained and mapped to the densely packed reference, consisting of groups of 100 bps genomic sequences flanking the restriction sites. 96.4% of the expected restriction sites were covered by real reads. The distance spanned by 80% of the neighbouring pairs of EcoRI restriction sites was within the range of 1 kb to 10 kb. Ideally, the read counts of pairs of 5' directional and 3' directional genomic fragments that result from enzymatic digestion should be equal if there were no CNA breakpoints flanking the restriction sites. Therefore a filtering system was used to efficiently eliminate partial sequencing biases that would negatively impact the CNA analysis. We considered the restriction sites where the ratio of read counts between the two adjacent fragments was less than 1.5 to be valid for subsequent analysis. As a result of this filtering step, 519,656 valid RAD restriction sites were identified among a total of 778,114 sites obtained by EcoRI digestion, and these were dispersed across the whole genome.

A total of 20X whole-genome sequencing data and 56X exome capture sequencing data were simultaneously generated from the same HCC sample as well as 80X RAD sequencing data. The data from the different resequencing strategies provided an opportunity to conduct a comparative study of their utility in CNA analysis. We took advantage of the read depth and MAF measurements for each chromosomal locus to delineate the CNA profiles from deep WGS, low WGS, exome sequencing, and RAD sequencing data. The CNA profiles composed of read depth profiles and MAF profiles were described in Figure 1. The snapshots of the read depth and MAF profiles provided an intuitive overview and visible comparison of CNA estimation across the different sequencing strategies.

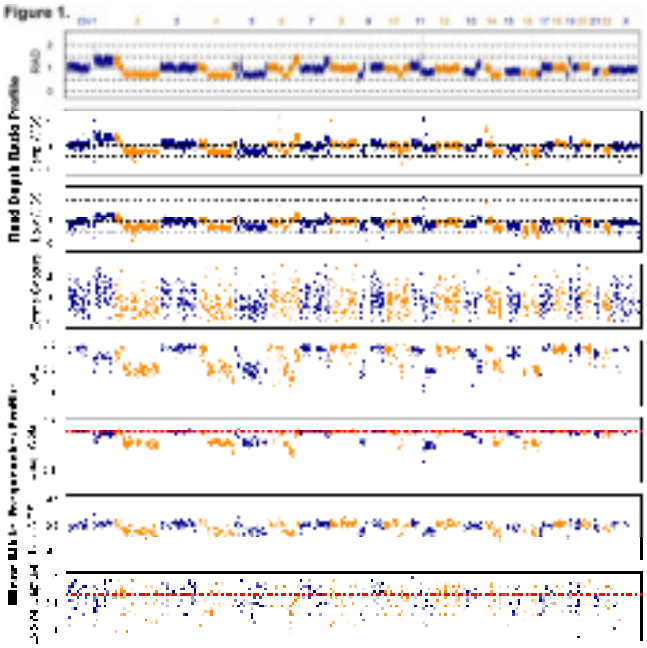

Fig. 1. Overall read depth and minor allele frequency (MAF) profiles in an HCC sample across different resequencing strategies. In the read depth profiles, each locus is represented by the average read count and calibrated by the respective median value. 10 consecutive restriction sites were treated as a unit. In MAF profiles, each locus represents the regionally averaged frequencies of 5 heterozygous sites identified by RAD and exome sequencing and 100 heterozygous sites identified by WGS.

3.2 Comparison of the precise measurements of read depth and minor allele frequency

Intuitively, we would achieve high accuracy in CNA detection if the scatter diagram was compact and the edges of the CNA segments were sharp. Here, we applied statistical and experimental validations to precisely evaluate and compare the read depth and MAF measurements obtained from the resequencing data. The first statistic applied was the windowed coefficient of variation, the mean value of which represents the dispersion of the read depth profile across each chromosome. Another statistic applied was the absolute difference of Gaussians (ADOG) function, which is regarded as an indicator of the sharpness of the segment boundary at the CNA breakpoint. Experimentally, the allele frequencies of 58 randomly selected heterozygous sites were validated via Sequenom genotyping to evaluate the accuracy of the MAF estimation made on the basis of the sequencing data.

The results of the comparisons are described in Figure 2. We balanced the datasets of RAD-seq in the total number of reads by randomly sampling reads in the experiments when comparing with exome sequencing and WGS data. The raw data and the balanced datasets supporting the subplots shown in Figure 2 are summarised in Supplementary Tables S1, S2 and S3. In Figures 2A and 2B, box-plots of two statistics measured the dispersion of the segments and the sharpness of the breakpoints in the read depth profiles. The mean values of the windowed coefficient of variation revealed that the datasets provided by both the RAD sequencing and deep WGS strategies had lower dispersion in their read depth profiles than was found in data generated by the low WGS strategy (Figure 2A). RAD sequencing was found to be the most effective strategy for obtaining sharpened edges when integer values of the absolute difference between two probability distributions of read depth flanking 16 common breakpoints were measured, even when total coverage of reads was reduced to a low level (Figure 2B). The data indicated that exome sequencing provided insufficient measurement of the required variables (Figure 2A and 2B). The percentile plot shown in Figure 2C depicts the deviation of the estimated allele frequency values from the corresponding frequencies obtained by independent genotyping analysis for 58 heterozygous sites. RAD sequencing provided over 60,000 heterozygous sites (corresponding to 6 percent of the sites characterised by deep WGS) where minor allele frequencies were scored for CNA analysis. The results of the genotyping analysis demonstrated that the estimated allele frequencies of approximately 70% of the validated heterozygous sites identified by 80X RAD sequencing were within the error margin of the Sequenom genotyping system. The percentage did not reduce too much when RAD reads were sampled to be 50X and 20X. Therefore, sufficient evidence supports the conclusion that RAD sequencing gave a precise measurement of read depth and minor allele frequency and achieved a comparable result to that of the deep WGS strategy for CNA analysis.

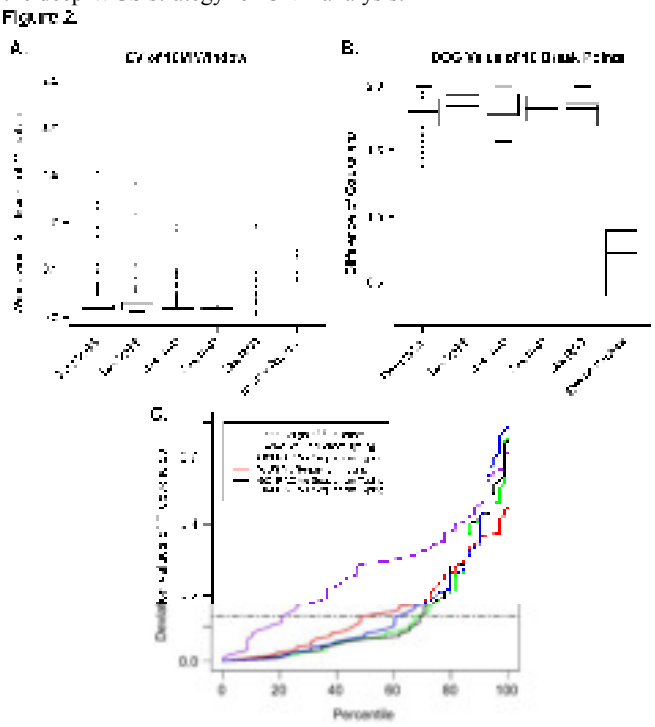

Fig. 2. Comparisons of the performance of read depth and minor allele frequency-based estimation of CNAs from datasets of RAD-seq, WGS, and exome sequencing. 80X RAD-seq was comparable with low coverage WGS with the same number of reads. 50X RAD-seq was balanced dataset equal to exome sequencing in total number of reads. And 20X RAD-seq was also compared with the same coverage of deep WGS. A. A box-plot of the 10 Mb windowed coefficient of variation measuring the dispersion of segments in the read depth profiles; B. A box-plot of the integral values of the ADOG function, reflecting the visibility and sharpness of the edges of 16 common CNA breakpoints; C. A percentile plot of the deviations away from the allele frequencies of 58 heterozygous sites measured independently via Sequenom analysis. The black dashed line indicates the error margin of the Sequenom genotyping system.

3.3 Determination of genomic copy number alterations in an accurate and robust fashion

CNAs in the HCC sample that were not situated within the centromeric or telomeric regions were surveyed. For the 80X RAD sequencing dataset, a total of 36 CNAs were detected from the read depth profile, of which 30 (approximately 83.3%) were confirmed by the MAFs (Table 1, Supplementary Table S4). We trimmed the raw reads of 80X RAD sequencing to a length of 35 bps and surveyed the CNAs based on the trimmed data. 32 of the 37 CNA segments were overlapped with the 36 segments identified from the original raw data (Supplementary Table S5). Moreover, RAD-seq reached a high level of convergence on CNA calling in 80X, 50X and 20X reads (Supplementary Table S9).The RAD sequencing strategy was therefore robust for CNA identification.

The CNAs identified from WGS data were also summarized in Table 1 and Supplementary Table S4 for comparison. The existing CNA calls previously identified using WGS mate-pairs (>3 kb library insert size) and read depth profiles (Tao *et al.*, 2011) were all detected by RAD sequencing. The confirmed CNAs with a length of more than 2 Mb were consistent across the RAD sequencing and deep WGS strategies. As shown by the Venn diagram in Figure 3A, 20 of the CNA regions overlapped across the different sequencing strategies.

However, we also noted that many CNAs were specific to certain sequencing strategies, and low- or high-WGS data contributed 2.5-fold more CNAs, of which less than half were supported by their MAF profiles. Therefore, we performed a simulated analysis and experimental validation to assess the CNA pools that were specific to particular sequencing strategies. Simulated WGS data from the human genome NCBI36/hg18 assembly exhibited many medium-sized spanning regions that were subject to abnormal mappability, which would introduce false-positive CNAs. The data shown in Table 1 revealed that in deep WGS, 11 of 33 CNAs (and 6 of 13 double-confirmed CNAs) less than 1 Mb spanned regions that were determined to be subject to abnormal mappability. In contrast, when the RAD sequencing strategy was used, in which subregions of the human genome were used as the reference and a minimum of mismatched or multiple hits occurred, the mappability of the targeted genomic regions had less influence on the precise identification of CNAs. Almost all of the 36 CNAs detected in the RAD sequencing data were identified in spanning regions that were not subject to abnormal mappability. Experimentally, we randomly selected targets from the pool where the CNAs with distinct copy numbers estimated across different sequencing strategies. qPCR verification was performed in the tumour sample using blood samples from two normal adults as controls. Three non-CNA regions that were supported by both their read depths and their minor allele frequencies from all of the sequencing strategies were also validated to identify and evaluate the false positives in the estimated copy numbers. The differences between the estimated copy numbers and the qPCR values are shown in Figure 3B. The raw data supporting the subplot are summarised in Supplementary Table S6. Using the three non-CNA segments as controls, the marked outliers represent the false positives in the different sequencing strategies (Figure 3B). The qPCR results indicate that RAD sequencing achieved an 89.4% (17 out of 19) validation rate and that there were approximately 50% or 75% false positives derived from deep- or low-WGS, respectively.

Even for the overlapping CNAs shown in the Venn diagram (Figure 3A), the different sequencing strategies estimated the copy numbers with varying degrees of precision. For example, there was an interesting high-copy duplication region located at chr11q. The copy number of this region was determined to be 5.21 via qPCR and was estimated as 3.93, 4,54, and 3.15 from deep-WGS, RAD, and low-WGS data, respectively. RAD sequencing inferred more accurate copy numbers in this ultra-high amplification region.

3.4 Minor allele frequency: an essential element for CNA analysis

The genomes of tumour cells experience enormous CNA mutations. Unlike diploid genomic regions, the relative ratio of two parental copies will not be equal to 1:1 under the selection pressures for clonal expansion in a tumour. This has led to the conclusion that CNA segments with uneven amplification of the two parental copies have functional importance that is strongly associated with tumourigenesis (Tao *et al.*, 2011; Nguyen *et al.*, 2006). Thus, the refined calculation of allele-specific copy numbers would be meaningful following the accurate identification of CNA segments.

Figure 3.

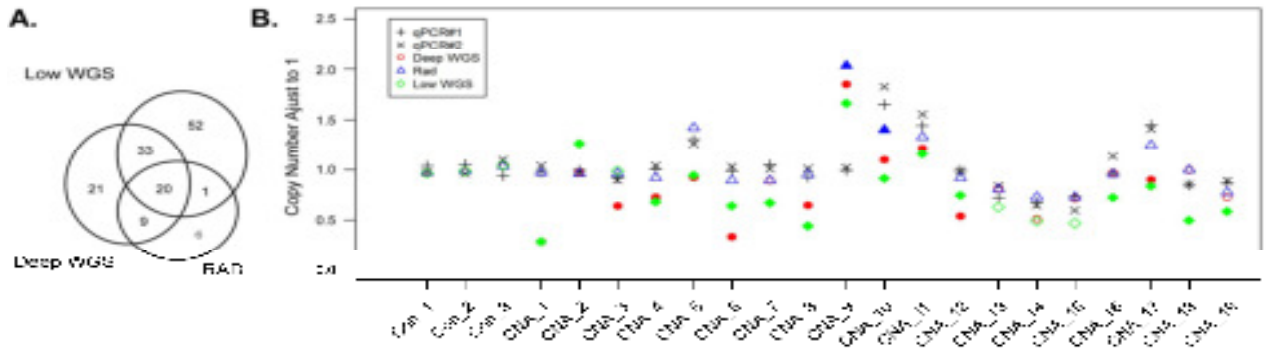

Fig. 3. A. A Venn diagram of detected CNAs across different sequencing strategies. B. qPCR validation results for CNAs. A total of 22 segments were verified, including three non-CNA regions as controls and 19 pending CNA regions. Two primers were designed for each CNA segment and the qPCR values for the two primers were marked with "+" or "x", respectively. The segments supported by two qPCR primers with similar qPCR values were retained. The copy numbers estimated from deep WGS, low WGS and RAD sequencing data are represented by circles, triangles and diamonds, respectively. Outliers confirmed as being false positives were filled in a solid style.

Table 1. Identified and confirmed CNA segments. Simulated deep WGS data exhibited medium-size spanning regions subject to abnormal mappability. The numbers of CNAs located in these regions are shown.

	Size of CNAs	All the CNAs regions			CNAs regions under abberated mappability		
		Deep WGS	Low WGS	RAD	Deep WGS	Low WGS	RAD
CNAs detected from coverage profile	[0,1M]	33	35	9	11	7	1
	(1M,2M]	12	18	1	1	3	0
	>2M	38	53	26	3	2	1
	Total	83	106	36	15	12	2
Double-confirmed CNAs by MAFs	[0,1M]	13	3	8	6	1	0
	(1M,2M]	4	5	1	0	0	0
	>2M	26	25	21	2	0	1
	Total	43	33	30	8	1	1

It is the minor allele frequency that distinguishes the uneven copy numbers of the parental alleles. Based on MAF data from RAD sequencing, we confirmed a total of 30 CNA regions as being under allele selection where the corresponding MAF was far from the baseline of 0.5 (Table 1 and Supplementary Table S4). A chr5q deletion with a MAF of 0.28 and a 1.5-fold read depth reduction was one such important CNA identified, and this CNA is known to act as a driver of tumourigenesis. First, the deleted allele of the chr5q deletion forms a C5orf51-CPEB4 fusion gene, which has been validated by PCR on cDNA(Tao *et al.*, 2011). Second, a tumour-driving somatic nonsynonymous mutation in cyclin G1 (CCNG1), which is a target of P53, was in the region spanned by chr5q, but was not located in the deleted allele of chr5q. Therefore, uneven loss of the parental allele copies results in the gain of a new fusion gene accompanied by the driver mutation CCNG1 and therefore promotes the selection of cells for clonal outgrowth in the tumour.

Genetic instability transforms the genome from its diploid state to a chaotic karyotype. Complex patterns of genomic architecture indicate mixed temporal sequences of rearrangements. Unlocking the nature and mechanisms of the CNA events that occur will give us valuable insights into the mechanisms of tumourigenesis. MAFs are able to identify "hidden" CNA events and track the history of CNA occurrences that cannot be resolved by read depth profiles alone. For example, there was an interesting observation at chr6q where a complex genomic composition existed. We divided the chr6q region into three segments, referred to as segment I, segment II, and segment III (Figure 4A). The copy number of segment II was at the baseline, while segments I and II had 0.75-fold and 1.5-fold copy numbers relative to the baseline, respectively. Outwardly, the read depth profile indicated two separate segments with abnormal copy numbers in the chr6q region where two CNA events occurred independently. Correspondingly, segment I, segment II and segment III had average minor allele frequencies of 0.28, 0.20 and 0.38, respectively, following MAF correction under the rule that the MAF of the baseline should be 0.5 (Figure 4B). The MAF data suggested an inconsistent fact, namely, that the frequency deviated strongly from the 0.5 ratio in segment II where the copy number was at the baseline. All of the patterns of two-component CNA events occurring in chr6q region were examined and rejected because two CNA events could not explain how the values of both the copy number and the MAF had arisen (Figure 4C). The data suggested that there were at least three sequential non-separable CNA events in the chr6q region. The most likely explanation is that the tumour genome has a tetraploid baseline, a deletion of one allele copy in segments I and II accompanied by a duplication of another allele copy in segments II and III, and an aneuploid duplication in segment III, which leads to the complex genomic composition of chr6q.

Figure 4.

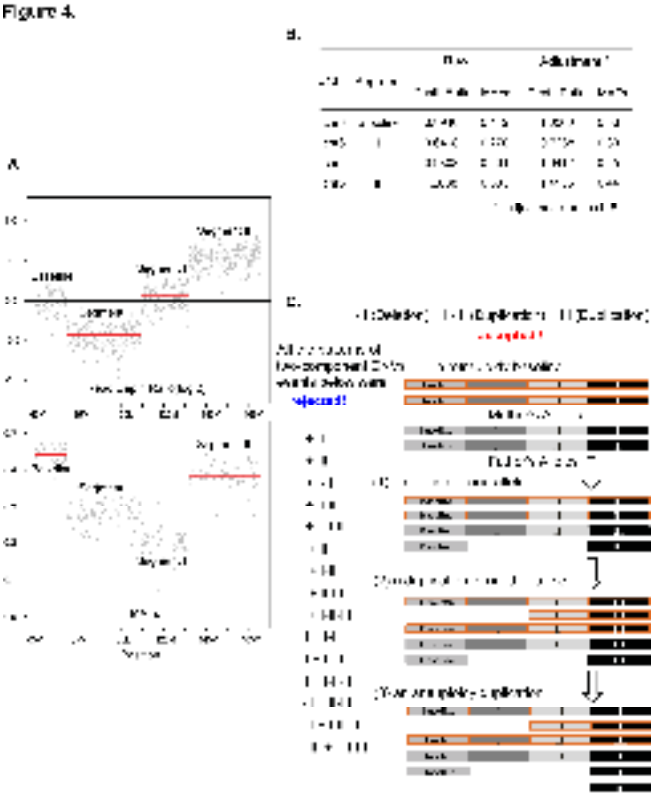

Fig. 4. Disclosing the "hidden" CNA events at chr6q by integrating MAFs with read depth information. A. A partial, enlarged view of the read depth and MAF profiles for the chr6p CNA region, together with a definition of the different segments analysed. B. The corrected minor allele frequencies and the ratios of read depth for each segment. C. Description of the possible CNA patterns. All of the patterns involving two-component CNA events failed to explain the distinct values of read depth ratio and MAF observed for the three segments. An assembly consisting of three CNA events is given as a potential explanation of the origin of the genomic variation observed at chr6q.

DISCUSSION

Array is a comprehensive and commercial method for CNV calling. However, Xi R. et al. demonstrated that the copy ratios given by sequencing data are more accurate than that given by array platforms (Xi *et al.*, 2011). And our in-house data on MCF7 cell lines also showed direct plots in which array was less sensitive than WGS to some segments with copy number alterations (Supplementary Figure S3, unpublished data). We re-factored WGS into array-

type format in essence and proposed a strategy of RAD sequencing aiming to combine the advantage of array with sequencing for CNA analysis. In summary, RAD sequencing has its unique characteristics. It reduces the mapping errors and noises in WGS. At the same time, it formats the fluorescence intensity in array to numbers of reads, as digital PCR was derived from qPCR. High noise-signal ratio and spatial biases in arrays are avoided. RAD sequencing as well as other NGS could detect multi-copy amplification, which arrays were not good at due to saturation of probes, and then give an estimation on copy number in a broader range for a chromosomal region. In addition, the restricted enzyme is flexible to alter the density of restriction sites without an upper limit but the density of SNP array is limited with the number of all heterozygous sites. It is also worth to highlighting that RAD sequencing is efficient without customizing specific probes of arrays for CNA analysis on new species.

However, we applied the RAD sequencing approach on an HCC tumor as an example to interpret the effectiveness of CNA calling and recognize there are some limits for the method. Firstly, the size of CNA captured with an RAD will depend heavily on the density of restriction sites. The EcoRI enzyme, with its 6- nucleotide recognition sequence, cut the genome into fragments within the range of 1 kb to 10 kb. As a result of treating 10 consecutive restriction sites as a unit, the valid resolution of CNAs was limited to approximately 150 kb. Flexibly, an approach was taken for the selection of an alternative enzyme to ensure the controllable resolution of CNA segments. Supplementary Figure S2 showed the length distributions of restriction fragments for enzymes with different recognition sequences. For enzymes with 4- or 5-nucleotide recognition sites, the distance spanned by pairs of neighbouring restriction sites was shorter and the achieved CNA resolution would reach the 10 kb level. The use of multiple enzymes to co-digest the genomic DNA would, therefore, be a possible strategy for increasing the sensitivity of CNA detection by RAD sequencing. Secondly, we did a direct comparison on the cost between RAD sequencing and SNP array. The cost of RAD sequencing is higher than that of arrays, although it is competitive with other sequencing methods. We calculated the cost of 50X RAD sequencing with ~1.8M restriction sites (co-digested by two enzymes) and 100bp-length reads (Supplementary Table S7). It is 1.5 times of that of SNP 6.0.

Our deep WGS data were mixed with the reads generated at an earlier stage from two sequencing platforms, SOLiD and Solexa. We called the CNAs using the same pipeline for the SOLiD reads and the Solexa reads independently to evaluate the consistency across platforms. 894 Mb of the 1150 Mb CNA regions identified in the SOLiD data overlapped with the 1158 Mb CNA regions identified in the Solexa data (Supplementary Table S8). In addition, the sequencing quality was reproducible across the Solexa and SOLiD platforms (Tao *et al.*, 2011; Zhou *et al.*, 2011). Deep sequencing data from the Solexa platform would admittedly provide a superior dataset, but the mixture of reads from two platforms had few negative effects on CNA determination. Meanwhile, although it provides a powerful dataset, deep WGS is costly and is associated with a heavy computational and storage load. In fact, deep WGS was not ideally perfect for CNA analysis as we expected. The number of spanning regions that were subject to abnormal mappability was one of the negative factors that caused inaccurate estimation of read depth and MAFs. In the simulated deep WGS data, approximately 2.97% (85.2 Mb of 2865.4 Mb) of the regions of the diploid genome had mapped read counts that were outside the range of 1.2- to 0.8-fold of the baseline. These regions increased the noise in the genomic variability analysis and increased the number of false-positive CNAs.

The use of MAF profiles provided a multitude of benefits in our study as follows: (1) MAF profiles confirmed the CNAs identified on the basis of read depth profiles; (2) they allowed a refined calculation of the allele-specific copy number and defined imbalanced parental alleles and driver CNAs that may be under selection for clonal expansion in tumourigenesis; and (3) the precise estimation of frequencies enabled the resolution of CNA events in complex genomic variations. These benefits demonstrate that the allele frequency profile of heterozygous sites is a wonderful complement to read depth for CNA analysis with sequencing data. But suitable and sensitive statistical approaches incorporating read depths and allele frequencies are required to be developped. We decoded the CNA composition in an example of a complex genomic region at Chr6q by integrating MAF and read depth information. Notably, Carter et.al have recently quantified the absolute copy numbers of chromosomes and tumour purity directly by analysis of somatic DNA alterations (S. L. Carter et al., 2012). However, efforts are needed to build a mathematical inference framework for deep analysis of CNAs, not only to determine the nature and mechanism of CNA events occurring in complex genomic regions but also to infer tumour purity and actual DNA ploidy.

CONCLUSIONS

CNA is an important type of genetic mutation associated with disease. Current strategies to investigate the CNA profile can be classified into two major categories: the nucleotide probe hybridization-based strategies and the sequencing-based strategies. The latter take advantage of precise information regarding read depth, read content, and allele frequencies and are therefore more comprehensive in their determination of copy numbers. However, the amount of sequencing data, the PCR amplification efficiency, the GC density, and the mappability of raw reads introduce complexities that can interfere with high-quality CNA analysis. Here, a comparison among deep WGS, low WGS, exome sequencing, and RAD sequencing strategies was performed.

The RAD sequencing strategy took full advantage of its precise measurement of allele frequency and read depth to make amplification and deletion calls from sequencing data as effectively as deep WGS. Importantly, two characteristics of RAD sequencing should be emphasized. First, the reads were mapped to a small subset of the whole genome, which was composed of flanking regions adjacent to the restriction sites. This yielded a minimum of mismatched or multiple-mapping hits and resulted in high quality mapping of reads in general. Second, the 5' directional and 3' directional sequences adjacent to the restriction sites were simultaneously sequenced because of the employment of a palindromic enzyme. This enabled us to introduce a filtering strategy, namely, that there should be no significant disparity between the read counts of the upstream and downstream sequences adjacent to the restriction sites. In this way, sequencing biases, such as unbalanced

amplifications, were significantly reduced. In RAD sequencing, regionally averaged minor allele frequencies (MAFs) at heterozygous sites, in addition to the read depth profile, offered reliable estimation of the copy numbers. A precise minor allele frequency was an essential element, not only for the refined calculation of allele–specific copy numbers but also in decoding multiple CNA patterns in complex genomic segments.

Taken together, we focused on the enzyme digestion-targeted sequencing technique, RAD sequencing, and have demonstrated that it is a new method for the characterization of CNAs that provides reliable data for the inference of genomic copy numbers. It appears that there is a great potential for the application of RAD sequencing to a variety of areas of research. In population evolution research and GWAS studies especially, extensive characterization of CNAs in numerous individuals is required to define the genetic variations among populations.

ACKNOWLEDGEMENTS

We thank Dr. ShuangliMi for proofreading of the manuscript. We appreciate the helpful discussion with Dr. Yu Wang, Dr. Yong Tao and Dr. YaliHou.

Funding: This work was supported by National Basic Research Program of China Grant [2012CB316505]; and National Natural Science Foundation of China Grant [31171265] to J. C.

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 research, 21, 974–84.

Alkan,C. et al. (2011) Genome structural variation discovery and genotyping. Nat Reviews Genet, 12, 363–76.

Asan et al. (2011) Comprehensive comparison of three commercial human whole-exome capture platforms. Genome biology, 12, R95.

Baird,N.A. et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS ONE, 3, e3376.

Benelli,M. et al. (2010) A very fast and accurate method for calling aberrations in array-CGH data. Biostatistics, 11, 515–8.

Bentz,M. et al. (1998) Minimal sizes of deletions detected by comparative genomic hybridization. GENES, CHROMOSOMES & CANCER, 21, 172–5.

Campbell,Peter J et al. (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nature genetics, 40, 722–9.

Carter,S.L. et al. (2012) Absolute quantification of somatic DNA alterations in human cancer. Nature biotechnology, 30, 413–21.

Cooper,Gregory M et al. (2007) Mutational and selective effects on copy-number variants in the human genome. Nature genetics, 39, S22–9.

Davey,J.W. et al. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature reviews. Genetics, 12, 499–510.

Ding,L. et al. (2010) Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature, 464, 999–1005.

Elia,J. et al. (2011) Genome-wide copy number variation study associates metabotropic glutamate receptor gene networks with attention deficit hyperactivity disorder. Nat Genet, 44, 78–84.

Fu,W. et al. (2010) Identification of copy number variation hotspots in human populations. Am J Hum Genet, 87, 494–504.

Garvey,C. (2010) A decade and genome of change. Genome biology, 11, 120.

Greenman,C D et al. (2011) Estimation of rearrangement phylogeny for cancer genomes. Genome research.

Haraksingh,R.R. et al. (2011) Genome-Wide Mapping of Copy Number Variation in Humans: Comparative Analysis of High Resolution Array Platforms. PLoS ONE, 6, e27859.

Hohenlohe,P A et al. (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet, 6, e1000862.

Krumm,N. et al. (2012) Copy number variation detection and genotyping from exome sequence data. Genome research, 22, 1525–32.

Lee,K.W. et al. (2011) Genome wide association studies (GWAS) and copy number variation (CNV) studies of the major psychoses: what have we learnt? Neurosci Biobehav Rev, 36, 556–571.

Lin,M. et al. (2004) dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics, 20, 1233–40.

Lu,X.-Y. et al. (2008) Genomic imbalances in neonates with birth defects: high detection rates by using chromosomal microarray analysis. Pediatrics, 122, 1310–8.

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

Mardis,E.R. (2011) A decade's perspective on DNA sequencing technology. Nature, 470, 198–203.

Medvedev,P. et al. (2010) Detecting copy number variation with mated short reads. Genome research, 20, 1613–22.

Navin,N. et al. (2011) Tumour evolution inferred by single-cell sequencing. Nature, 472, 90–4.

Nguyen,D.-Q. et al. (2006) Bias of selection on human copy-number variants. PLoS Genet, 2, e20.

Nozawa,M. et al. (2007) Genomic drift and copy number variation of sensory receptor genes in humans. PNAS, 104, 20421–6.

Perry,G.H. et al. (2007) Diet and the evolution of human amylase gene copy number variation. Nature genetics, 39, 1256–60.

Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. Nature genetics, 37 Suppl, S11–7.

Pinto,D. et al. (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nature biotechnology, 29, 512–20.

Pleasance,Erin D et al. (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. Nature, 463, 191–6.

Robinson,M.D. et al. (2012) Copy-number-aware differential analysis of quantitative DNA sequencing data. Genome research, 22, 2489–96.

Sathirapongsasuti,J.F. et al. (2011) Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics, 27, 2648–54.

Sharp,A.J. et al. (2005) Segmental duplications and copy-number variation in the human genome. Am J Hum Genet, 77, 78–88.

Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. Nature biotechnology, 26, 1135–45.

Simpson,J.T. et al. (2010) Copy number variant detection in inbred strains from short read sequence data. Bioinformatics, 26, 565–7.

Solinas-Toldo,S. et al. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. GENES, CHROMOSOMES & CANCER, 20, 399–407.

Stratton,Michael R et al. (2009) The cancer genome. Nature, 458, 719–24.

Sudmant,P.H. et al. (2010) Diversity of human copy number variation and multicopy genes. Science, 330, 641–6.

Tao,Y. et al. (2011) Rapid growth of a hepatocellular carcinoma and the driving mutations revealed by cell-population genetic analysis of whole-genome data. PNAS, 108, 12042–7.

Walsh,T. et al. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science, 320, 539–43.

Xi,R. et al. (2011) Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. PNAS, 108, E1128–36.

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.

Xu,B. et al. (2008) Strong association of de novo copy number mutations with sporadic schizophrenia. Nature genetics, 40, 880–5.

Yoon,S. et al. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome research, 19, 1586–92.

Zhou,R. et al. (2011) Population genetics in nonmodel organisms: II. natural selection in marginal habitats revealed by deep sequencing on dual platforms. MBE, 28, 2833–42.