Chapter 18
Detecting Copy Number Changes and Structural Rearrangements Using DNA Sequencing

Venkatraman E. Seshan

Abstract Chromosomal abnormalities in the form of deletions, duplications, inversions and translocations are common in cancer. These changes feed the oncogenic process by affecting genes that are involved in tumor growth. Next generation sequencing has aided our ability to study these changes at very high resolution. In this chapter we will describe the nature of these data and the information contained in them for the detection of the structural changes. We will present the circular binary segmentation algorithm for the segmentation of array based copy number data and adapt it to NGS data. We will also present a method for the detection of somatic structural rearrangement. We will illustrate these methods using data from breast cancer cell line (tumor) along with its blood (normal) counterpart generated by the cancer cell-line encyclopedia project.

18.1 Introduction

The flow of genetic information in cells [3, Chap. 5] occurs primarily through the transcription of DNA into RNA which is then translated into proteins that carry out the cellular functions. This is stated as *DNA makes RNA, RNA makes proteins, proteins make us* [18] and referred to as the central dogma of molecular biology [8]. This implies that changes to DNA can have an effect on the biological processes. These changes in DNA can be mutations as well as structural changes. In humans, autosomal chromosomes appear in pairs, one from each parent, and thus have two copies of every gene; the allosomes (sex chromosomes) are XX in females (two copies of X) and XY in males (one copy each of X and Y). Gains and losses of all or parts of chromosomes are known as copy number changes

V.E. Seshan (✉)
Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 307 East 63rd St., 3rd Floor, New York, NY 10065, USA
e-mail: seshanv@mskcc.org

S. Datta and D. Nettleton (eds.), *Statistical Analysis of Next Generation Sequencing Data*, Frontiers in Probability and the Statistical Sciences, DOI 10.1007/978-3-319-07212-8_18,

and are implicated in many diseases. These changes could either be germline (inherited) or somatic (acquired). Examples of germline changes are 3 copies of chromosome 21 (copy number gain) resulting in Down's syndrome [15, Chap. 5] or single X (copy number loss) resulting in Turner's syndrome [15, Chap. 5]. Somatic changes are very common in cancer, where a gene is gained and it promotes growth (oncogene, e.g., ERBB2 (HER2/Neu) in breast cancer [14]), or a gene is deleted and the ability to control growth is lost (tumor suppressor gene, e.g., PTEN in prostate cancer [38]). Other changes to DNA such as the Philadelphia chromosome [23], a reciprocal translocation between chromosomes 9 and 22, is another type of structural change implicated in cancer (chronic myelogenous leukemia or acute lymphocytic leukemia). Thus, studying copy number changes and other structural rearrangements is important for understanding the oncogenic process.

Karyotyping, which is the study of the number and appearance of chromosomes, was the earliest method used for detecting chromosomal aberrations and provides information at a low resolution. The development of comparative genomic hybridization (CGH) [13, 20] allowed measurement of copy number changes over the entire genome and enabled it to be localized to a chromosome at an improved resolution of 10 to 20 megabase. High throughput methods such as BAC (bacterial artificial chromosome), aCGH (array comparative genomic hybridization) and SNP (single nucleotide polymorphism) arrays, based on the microarray technology have systematically increased the resolution and thereby our ability to detect gains and losses of smaller chromosomal regions; see [27] for a review of array CGH technology. Whereas a karyotype assay can clearly show trisomy of chromosome 21, the loss of PTEN cannot be readily visualized in a Affymetrix SNP 6.0 array with over 1.8 million markers. Thus sound analytic methods are required for the large volume of noisy data generated by the high throughput methods.

The analysis of copy number data is composed of two parts—the identification of regions of gains and losses in each subject followed by combining this information across samples to identify recurrent events associated with cancer. Several methods have been proposed for the per sample analysis of copy number data which can be characterized as "smoothing and thresholding" or "segmentation" methods. A comprehensive comparison of the performance of several of these methods was done by [16]. Overall, segmentation methods were found to be most suitable for the per sample analysis of copy number data. The principal concept behind the segmentation methods is that since copy number for a cell is integer valued, gains and losses are discrete events and thus along a chromosome the gain or loss induces a jump discontinuity. Note that the tissue sample being assayed is a collection of cells all of which will not have the same changes. However, the distinct clones that make up the tissue sample is expected to be far fewer than the number of cells and hence the average copy number will have the form of a step function. We formulated this as a change point problem to develop the circular binary segmentation (CBS) algorithm [25, 36] which is one of the widely used methods. GISTIC [4], GRIN [28] and RAE [34] are frequently used algorithms to combine the copy number changes detected in the per sample analysis in order to identify recurrent events and implicated genes.

Next generation sequencing (NGS) of genomic DNA enables us to obtain information on somatic mutations and structural changes. The structural changes include copy number gains and losses as well as rearrangements such as inversions and translocations. [Note: Inversions and translocations are explained in Sect. 18.3.1] Several algorithms such as BreakDancer [6], CNVnator [1], CNVseq [40], CREST [37], SegSeq [7], seqCBS [33], SVminer [12] are currently used for obtaining structural change information from NGS data. In the following sections we will describe the CBS algorithm, adapt it to sequencing data, and demonstrate it using cell line data. We will finish the chapter by presenting a simplified summary of the procedure for identifying other structural variations.

18.2 Background

In this section we will describe the design and techniques used to generate the data that are to be analyzed. The first step in the process of obtaining the data is the generation of a library of genomic DNA composed of short DNA fragments, typically 100 to 500 nucleotides long, from the sample of interest. This library can encompass the entire genome (whole genome sequencing) or selected genomic regions (targeted sequencing). The creation of the library in either case begins with generating DNA fragments by randomly breaking the entire genome using a technique such as sonication. The fragments are then sorted by molecular weight to enable the selection of fragments of the desired length. In targeted sequencing an additional selection process is employed where the DNA is hybridized to arrays with probes that are designed to capture DNA fragments that cover the genomic regions of interest. A specific case of targeted sequencing is whole exome sequencing where the genomic regions selected are all the exons (coding regions) of all known genes (approximately 20,000). Custom gene panels [11] that cover a smaller collection of genes known to be most commonly associated with cancer are also currently in use. The regions in targeted sequencing span a small fraction of the whole genome, 1–2% in the case of whole exome and even less for custom panels, allowing for high coverage of the target.

The library that is generated is then sequenced to obtain reads, which are the strings of bases or nucleotides, that make up a part of the fragment. Sequencing can either be single-end or paired-end where the DNA fragments are sequenced (read) from either one end or both ends, respectively. Read length, which is the number of nucleotides sequenced, can be specified in the instrument for an experiment. The reads are then mapped to a reference genome to obtain positional information on where the reads, and hence the fragments, come from, *i.e.*, their locations. These locations follow a probability distribution that is influenced by factors such as the GC content and mappability. The data used for identifying structural changes are various characteristics of the reads such as their locations and fragment size.

In cancer research, the principal goal of DNA sequencing is to identify changes in DNA (mutations and structural) acquired by the tumor. Hence, typically, both tumor and normal cells are sequenced. The sequencing of normal cells will help identify any germline events, for example, BRCA1 mutation, that may be present. In paired tumor-normal sequencing, the comparison of the tumor to its matched normal will benefit from the canceling out of the influence of the technical factors that affect sequencing. Running them in the same batch would additionally ensure that batch effects are minimized. Although the use of paired tumor-normal samples is ideal for identifying changes that are specific to the tumor, it may not always be feasible. For instance, the analysis of archival tumors in which only tissue samples from the tumor are available will need an external pool of normal samples to identify tumor specific changes. However, large scale copy number polymorphisms have been seen in the germline [31] and Redon et al. (2006) [29] created a first-generation copy number variation (CNV) map from copy number profiling of the HapMap samples. Thus, the comparison of tumor data to an external normal needs to account for technical artifacts that may not cancel as well as inherited copy number events.

Unlike karyotying, both sequencing and array based measurement of copy numbers query the DNA fragments directly and do not contain information on individual cells. This introduces an identifiability problem as follows. Let us suppose that a global change in copy number has occurred in which every single chromosome in the cell is duplicated resulting in a total copy number of four. Whole genome duplication such as these and aneuploidy in general are common in cancer [9]. In terms of information contained in the DNA, a tissue with cells of this kind is indistinguishable from a tissue of normal cells. In general, both the array based and sequencing approaches to copy numbers can only provide relative copy number changes and require external information to resolve the relative numbers into absolute copy numbers. The ABSOLUTE algorithm [5] provides a method to use the ploidy (which is the average copy number) and tumor purity to obtain the absolute copy numbers.

In the next section, we will present a method for analyzing copy numbers from matched tumor-normal sequencing data. Furthermore, the changes identified will be based on the relative copy numbers and thus gains and losses will be relative to the average copy number of the tumor.

18.3 Methods

The read data generated from DNA sequencing contains not only information on the nucleotides that make up the subject's genome but also the relative abundance of a locus as well as distances between loci. These additional elements can be leveraged to detect structural changes to the DNA. In the following subsections we will develop a method to obtain the copy number profile from the relative abundance measure.

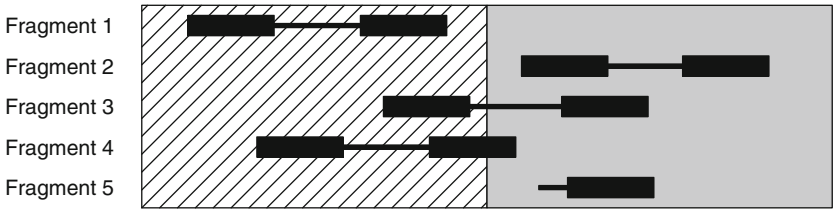

Fig. 18.1 Different types of fragments in a paired end DNA sequencing data

18.3.1 Structural Change Information in NGS Data

In the background section, we described how the reads data that are to be used for identifying structural changes are generated. We will now describe the information contained in these data using Fig. 18.1 which shows a portion of the tumor genome and five fragments from it.

The displayed portion of the tumor genome consists of a striped and a shaded part both of which are contiguous in the germline but the transition from the striped to the shaded does not occur in the germline and thus represents a structural change boundary. The fragments shown are from paired end reads where the thick rectangles are the reads and the thin one connecting them is the inferred intermediate region once the reads are mapped.

In the germline, both the striped and the shaded regions will appear exactly twice in a cell provided they are not polymorphic but in a tumor cell they appear more than twice if the region is gained and fewer than twice if it is lost. The transition corresponds to a translocation if both the regions have the same orientation as in the reference genome and an inversion if their orientations are opposite. The translocations can be intra- or inter-chromosomal depending on whether both regions come from the same chromosome and different chromosomes, respectively.

The fragments shown in the figure are read pairs for which at least one of the two ends is mapped to either the striped or the shaded genomic region. The top three fragments have both ends mapped and the bottom two have only one end mapped. Note that, although both reads of Fragment 4 are shown, only the read contained in the striped region will be mapped using a standard alignment procedure and the other end would require a partial read mapping algorithm such as CREST [37] to be mapped. Unmappable reads, such as the mate pair of Fragment 5, can occur if the read contains repetitive elements that are not uniquely identifiable.

A region that is gained in the tumor will contribute more fragments to the tumor reads and one that is lost will contribute fewer fragments. So the counts of the reads within a region, namely its abundance measure, is related to the copy number. Since the reads in Fragments 1, 2 and 5 are completely contained within the striped and shaded regions, they only contribute to the abundance measure. Since the two ends of Fragments 3 and 4 are mapped to the two regions, not only do they contribute to the abundance measure, they can also inform directly on the possible location of a

structural change. Zhao et al. (2013) [41] provide a review of various computational tools available for CNV detection that use one or a combination of these features.

In a targeted sequencing experiment, a read pair will contain the location of a structural transition, only if that transition occurs near a targeted genomic region which enables such a fragment to be captured. So a targeted sequencing experiment is unlikely to detect translocations and inversions unless the regions where such events could occur are specifically targeted, for example, the translocation in Philadelphia chromosome. Hence de novo structural rearrangements are rarely identifiable in targeted sequencing. The abundance measure however, is available and effective for copy number profiling both in whole genome and targeted sequencing. We will describe our method based on abundance measure (read-depth) in detail.

18.3.2 Circular Binary Segmentation

Let $X_1, X_2, \ldots$ be a sequence of random variables. A change-point is an index ν such that the random variables $X_1, \ldots, X_\nu$ have a common distribution F_0 and $X_{\nu+1}, \ldots$ have a different distribution F_1 (until the next change-point or the end of the sequence). For the copy number problem using data from array CGH the X_is are the log-transformed normalized intensities (or log-ratios) of the markers which are ordered by the position along the chromosome and thus is a finite sequence of length m. Since the copy number of a cell is integer valued and the tumor consists of far fewer distinct clones than cells, it is appropriate to view the locations where the copy number changes to be the change-points that we wish to detect.

The test statistic introduced in the CBS algorithm [25] to detect the change-points is the maximal t-statistic given by:

$$T = \max_{1 \le i < j \le m} \left\{ \hat{\sigma}_{ij} \sqrt{\frac{1}{j-i} + \frac{1}{m-j+i}} \right\}^{-1} \left| \frac{S_j - S_i}{j-i} - \frac{S_m - S_j + S_i}{m-j+i} \right|$$

where $S_i = X_1 + \cdots + X_i$ is the partial sum and $\hat{\sigma}_{ij}^2$ is the mean-squared error given by

$$\hat{\sigma}_{ij}^2 = \frac{1}{m-2} \left[\sum_1^m X_i^2 - (S_j - S_i)^2/(j-i) - (S_m - S_j + S_i)^2/(m-j+i) \right].$$

The motivation for this test statistic is as follows. If the X_is are normally distributed with a common variance then the change-points correspond to a change in mean. Suppose the change-points are fixed at i and j then the optimal statistic to test the equality of the means of the two sets $\{X_{i+1}, \ldots, X_j\}$ and $\{X_1, \ldots, X_i, X_{j+1}, \ldots, X_m\}$ is the t-statistic. Because the change-points are unknown we obtain our test statistic by maximizing the t-statistic over all possible i and j. Note that $j = m$ corresponds to

the case of a single change-point. The null hypothesis of no change-points is rejected if the p-value of the test statistic is below the significance threshold. Since the log-ratio data may not be normally distributed the CBS procedure was made robust by using a permutation reference distribution. The algorithm begins by testing for the presence of change-points in whole chromosomes. If the null hypothesis of no change-points is rejected, then the change-points that are detected will segment the chromosome into two (test detects one change-point) or three contiguous regions (test detects two change-points). The test procedure is applied recursively to each of the regions until no change-points are detected in any of them.

In comparative studies, the CBS algorithm was found to perform well consistently [39] and had the best operational characteristics [16] amongst several methods for analyzing copy number data. However, since the test statistic is maximizing over both i and j the computing time grew as the square of the number of markers which made the analysis burdensome as the resolution of arrays increased. To address this, [36] developed a faster CBS algorithm using tail probability approximations of Gaussian random fields as well as sequential testing. These and additional algorithmic improvements have made the use of this procedure routine for the analysis of array based copy number data.

18.3.3 Adapting CBS to NGS Data

In a sequencing experiment, the DNA fragments are sampled randomly and thus, a region that has a higher copy number contributes a larger number of fragments than a region with a lower copy number. The locations that the reads are mapped to is a function of several factors such as sequence composition and fragment size. Although the distribution of the locations of the mapped reads is non-uniform, the ratio of the read counts between tumor and normal will be proportional to the tumor to normal copy number ratio. Two additional scaling factors are needed for the read count ratio to reflect the true copy number ratio. The first is the ratio of total number of reads in the tumor and normal, which adjusts for the fact that tumors are often sequenced to a higher coverage than normal. The second factor depends on the purity and ploidy of the tumor. Thus the read count ratio data enables us to detect the regions of copy number change but will only give us a relative copy number. For instance, suppose we are interested in knowing whether the ERBB2 gene is gained (relative to the average copy number of the tumor) in a breast cancer sample; we can count the fragments that map to this gene in the tumor and normal samples and compare that ratio to the ratio of total number of fragments mapped in the two samples.

The independent elements in a sequencing experiment are the DNA fragments which are represented by a read pair, if both ends are mapped, and a single read, if only one end is mapped. If the abundance measure is calculated at the nucleotide level, then a DNA fragment contributes to the read count of all the positions within the read as well as all those in its mate pair. This induces a serial correlation in

the abundance measure data indexed by genomic position. Read pairs that span a structural transition, such as Fragment 3, can induce a longer range correlation. In order to obtain copy number data that are independent, we need that each fragment be counted towards only one abundance measure data point.

A deterministic approach to ensure that each fragment is counted towards one copy number data point only, is to represent each fragment by its mid-point. This presents a problem for fragments where only one end is mapped as well as those fragments with both ends mapped that are not consistent with the lengths of the fragments selected for sequencing. Such fragments can be removed from the copy number calculations and since they typically represent a small fraction of the reads, it is expected to have minor effect on the copy number profile. Alternately, we can include them as follows: for fragments with only one end mapped, use the midpoint of the read; for fragments with both ends mapped, pick one of the reads at random and pick its midpoint. In targeted sequencing, we expect only one of the two reads in a read pair that needs such probabilistic assignment, to be near a target interval and can choose the midpoint of that read to represent it. We will calculate the abundance measure for copy number profiling from these positional data. Note that under this data representation, the average number of fragments per position will be the average coverage divided by the read length for single-end sequencing and average coverage divided by twice the read length for paired-end sequencing. For example, in an experiment with $50\times$ coverage using 2×75bp sequencing this translates to a read count of $1/3$ fragments per base. Since this number is small, we will require that the data be binned to aggregate information and provide reliable copy number profile. We recommend a bin size that gives an average bin count of 25 or higher which for this example will result in a bin size of 100 bases.

A final feature of the data that requires attention is specific to targeted sequencing where capture technique is used to enrich DNA fragments from genomic regions of interest. Although these capture technologies have high specificity, it is not perfect, i.e., the library being sequenced will consist of DNA fragments that are not on target. This will lead to a large number of bins, far exceeding the bins that cover the regions being targeted, with very low counts (typically singletons). Since these bins are spread over the entire genome, the data from them will have an undue influence on the copy number profile and should be discarded prior to analysis. We address this by using primarily bins that span the regions of interest with target intervals enlarged to allow for fragment overhang.

With these preliminaries in place, let N_1 and N_0 be the total number of mapped fragments for the tumor and normal samples, respectively. Let (n_{1i}, n_{0i}) be the tumor and normal fragment counts for the i^{th} bin, and m be the number of bins. Similar to the log-ratios from copy number arrays we define the copy number data used for the segmentation as $X_i = \log_2[(1+n_{1i})/(1+n_{0i})] - \log_2(N_1/N_0)$, where the 1 is added to address bins with zero counts. The average fragment counts for bins within a region of constant copy number is proportionally increased or decreased and thus the log-ratio has a constant mean. However, the variability of fragment count is proportional to the average and thus we expect the variability of the log-ratio

to be inversely proportional to the average fragment count. While the test statistic shown in Sect. 18.3.2 is adequate, a weighted version of the statistic will be more appropriate.

Suppose $\{Y_1,\ldots,Y_k\}$ and $\{Z_1,\ldots,Z_l\}$ are two sets of random variables where Y_is have mean μ and variance σ_i^2 and Z_js have mean θ and variance τ_j^2. Then the minimum variance estimate of the difference in means $\mu-\theta$ is the difference in weighted average with weights given by the inverse of the variances, i.e., $(\sum \sigma_i^{-2} Y_i/\sum \sigma_i^{-2}) - (\sum \tau_j^{-2} Z_j/\sum \tau_j^{-2})$. Thus the optimal statistic for testing the hypothesis $\mu=\theta$ is the weighted t-statistic based on this difference in weighted average. The maximal t-statistic we will use for change-point detection will be the maximum over all i and j of the weighted t-statistic suggested by the minimum variance estimate. Note that we need to know the parameters σ_i^2 and τ_j^2, at least up to a constant, to obtain the weighted t-statistic.

For the fragment count data, we expect the variance of the counts to be proportional to the mean. The proportionality constant is 1 if the counts have a Poisson distribution and the relationship holds for distributions with extra variation such as negative binomial. So the variance of the log of the counts will be inversely proportional to the mean counts and thus the weight will be proportional to counts. Note that the tumor counts in the log ratio is affected by gains and losses and which can influence the weights. Thus we recommend using only the normal counts for the weights which is consistent with the null hypothesis of no change. In order to dampen the effect of bins with very large counts we suggest that the weights grow as the logarithm of the counts. In the next section we will present an example of the copy number analysis of sequencing data to demonstrate all these.

An alternate approach to the analysis is to use a variance stabilizing transformation. Anscombe (1948) [2] showed that for a Poisson random variable X, the transformation $\sqrt{X+3/8}$ is variance stabilizing, if the rate parameter is large enough (≥ 5). However, in order to allow for extra variation if we posit that the count data are distributed as negative binomial, then the variance stabilizing transformation is either $\sinh^{-1}\left[\sqrt{(X+3/8)/(k-3/4)}\right]$ or $\log(X+k/2)$ where k is the dispersion parameter. Ignoring the transformation's dependence on the dispersion parameter k, one can define the copy number data as $\sqrt{n_{1i}+3/8} - \sqrt{n_{0i}+3/8}$ and segment them using the unweighted CBS algorithm. Note that these data will not be centered at zero and hence the sign of the segment mean does not indicate a gain or a loss from the average tumor copy number. However, the underlying true regions of constant copy number will be the same as in the log-ratio.

18.4 An Example

In this section we will illustrate in detail the steps involved in the analysis of DNA sequencing data for copy number changes using data from a cancer cell line. The data are from the breast cancer cell line HCC1143 and its blood (normal) counterpart

HCC1143BL which are part of the cancer cell line encyclopedia (CCLE) project (http://www.broadinstitute.org/ccle/home). Whole exome sequencing (paired-end 2×75bp) was done for these two samples and the read data, aligned to the HG19_Broad_variant (Human reference GRCh37) reference genome, are available at the Cancer Genomics Hub (https://cghub.ucsc.edu/datasets/data_sets.html). The size of these two data sets are 10.8 Gb and 8.3 Gb, respectively, and the analysis will require powerful computers. Software requirement for this analysis are: *Bioconductor* [10], specifically the *Rsamtools* [22] and *DNAcopy* [32] packages, *Integrative Genomics Viewer* [30], *Picard* [26] and *samtools* [19]. Note that all dependencies of these software should also be available.

We begin with using *samtools* to summarize the data file that was downloaded from CCLE. The summary data (with line numbers added) for the normal sample are:

```
 1  68629600 + 6562518 in total (QC-passed reads +
                                        QC-failed reads)
 2  10054468 + 1517557 duplicates
 3  67842739 + 5593779 mapped (98.85%:85.24%)
 4  68629600 + 6562518 paired in sequencing
 5  34314800 + 3281259 read1
 6  34314800 + 3281259 read2
 7  66854380 + 5314442 properly paired (97.41%:80.98%)
 8  67196156 + 5353240 with itself and mate mapped
 9  646583 + 240539 singletons (0.94%:3.67%)
10  301196 + 35316 with mate mapped to a different chr
11  260127 + 29485 with mate mapped to a different chr
                                                (mapQ>=5)
```

The first line says that there are approximately 75 million reads in total in this sample which are decomposed into those that passed quality control (QC) and those that did not. This QC flag is platform and aligner specific. We will restrict the analysis to only those reads that passed QC (over 90% of the total). Lines 4 through 6 give the breakdown of the reads in Line 1, namely they are paired (Line 4) and that each end contributes half of the reads (Lines 5 and 6). Line 3 gives the number of reads that are mapped to the reference genome among the number listed in Line 1. The reasons the reads are unmapped are varied, such as structural rearrangement as seen in Fragment 4 of Fig. 18.1 or viral DNA mixed in with the sample. Line 7 gives the number of reads from fragments with both ends mapped and the two reads are consistent with the expected fragment sizes and the reads are in the proper direction (5′ to 3′ and vice versa, respectively). Line 8 gives the reads from fragments for which both ends are mapped. This number is larger than the one in Line 7 as it includes improperly paired reads as well. The counts of improperly paired reads with the two ends mapped to two different chromosomes is given in Line 10 and the subset that meets a mapping quality threshold is given in Line 11. Line 9 gives the number of fragments for which only one of the two reads is mapped. Finally, Line 2 gives the numbers of reads that are considered duplicates.

Duplicates are fragments for which the two ends, when mapped, give the same start and end locations and (nearly) identical read sequences. Since it is very unlikely that two identical DNA fragments are generated during the original DNA preparation, these are considered to have risen at the PCR amplification step where some fragments can get overamplified. Thus, duplicate reads do not provide independent information on the DNA of the sample and hence, only the read pair with the best read qualities is kept and the rest are removed from further analysis. We accomplish the deduplication step using the *Picard* software (MarkDuplicates option) which unlike *samtools* can also remove inter-chromosomal duplicates.

In summary, these data come from approximately 34 million fragments of which 5 million are potential PCR duplicates resulting in 29 million fragments of usable data. The pairs that are not proper (the excess of Line 8 over Line 7), especially, the ones with the mate mapped to a different chromosome (Lines 10 and 11), are the informative ones for non copy number structural changes (translocations and inversions). Additionally, information in the mate pair of the singletons in line 9 can potentially be extracted using partial read alignment for use in detecting structural variations. A similar breakdown of the summary data of the tumor file tells us that there are approximately 34 million usable fragments in that sample. The target enrichment intervals used for the whole exome sequencing is available in the CGHUB website (https://cghub.ucsc.edu/datasets/whole_exome_agilent_1.1_refseq_plus_3_boosters_plus_10bp_padding_minus_mito.Homo_sapiens_assembly19.targets.interval_list.tsv). There are a total of 36.6 million bases in these intervals (31.8 million if the targets labeled `new_exome_1.1_content` are excluded) which results in an expected count of 1 fragment per base in the target region.

Note that for variant (somatic mutation) detection, it is customary to conduct indel realignment and recalibration of quality score steps on the sequence data using *GATK* [21]. The copy number analysis can be performed after these steps and can benefit from them, particularly if read quality is accounted for in the analysis since the influence of poor quality reads can be eliminated. The quality recalibration step is valuable for identifying other structural variations reliably.

Once the data have been deduplicated, we extract the properly paired reads from both the tumor and normal samples. Since the data are from a cancer cell that originated in a female, we only extracted the reads that mapped to the 22 autosomes and the X chromosome which resulted in 28.5 and 33.5 million fragments, respectively, for tumor and normal. The number of reads mapped to the Y chromosome is approximately 16,000 for both the tumor and the normal which is reassuringly negligible. The densities of the fragment lengths for the tumor and normal samples are shown in Fig. 18.2. Fragments with lengths smaller that 76 or larger than 750 were not included in this figure for visual clarity. Although the fragments not included in the density plot can provide alternate information on structural changes, their contribution to the abundance measure is minimal as they represent 0.49% and 0.66% of normal and tumor fragments, respectively. The fragment lengths of the normal sample (median 163) are slightly larger than that for

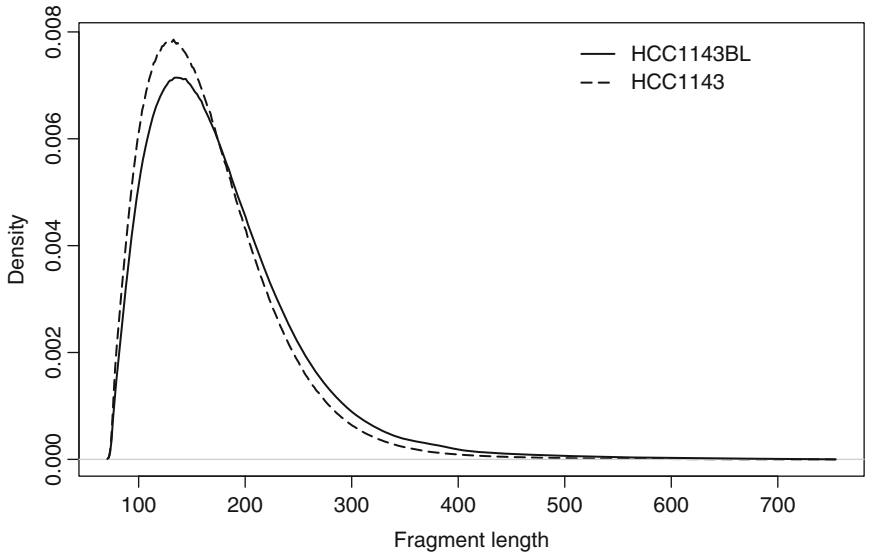

Fig. 18.2 The distribution of fragment lengths in tumor (*dashed line*) and normal (*solid line*) samples. The densities were generated using fragments whose lengths are between 76 and 750 bases

Table 18.1 The target intervals in the TP53 gene

Chr	Start	End	Width	Target
17	7572915	7573020	106	Target_128140
17	7573915	7574045	131	Target_128141
17	7576841	7576938	98	Target_128142
17	7577007	7577167	161	Target_128143
17	7577487	7577620	134	Target_128144
17	7578165	7578301	137	Target_128145
17	7578359	7578566	208	Target_128146
17	7579300	7579602	303	Target_128147
17	7579688	7579733	46	Target_128148
17	7579827	7579924	98	Target_128149

the tumor sample (median 154), and a vast majority of fragments (93.8% of normal and 96.6% of tumor) are fewer than 300 bases in length.

In order to provide further insight into the nature of targeted sequencing data, we will take an in depth look at the well known cancer gene TP53. This gene spans a 10 kilobase region on chromosome 17 with target intervals of different widths which are shown in Table 18.1. A figure of the data from this region for the normal sample generated using *Integrative Genomics Viewer* is in Fig 18.3. In the top part of the figure, the chromosome is shown with the region of interest in p13.1 highlighted in red and the genomic positions in bases. Below that are the genes in that region and the exons. The gene display is packed to show various forms of the gene present in RefSeq (http://www.ncbi.nlm.nih.gov/refseq/); the tall blue rectangles are the exons

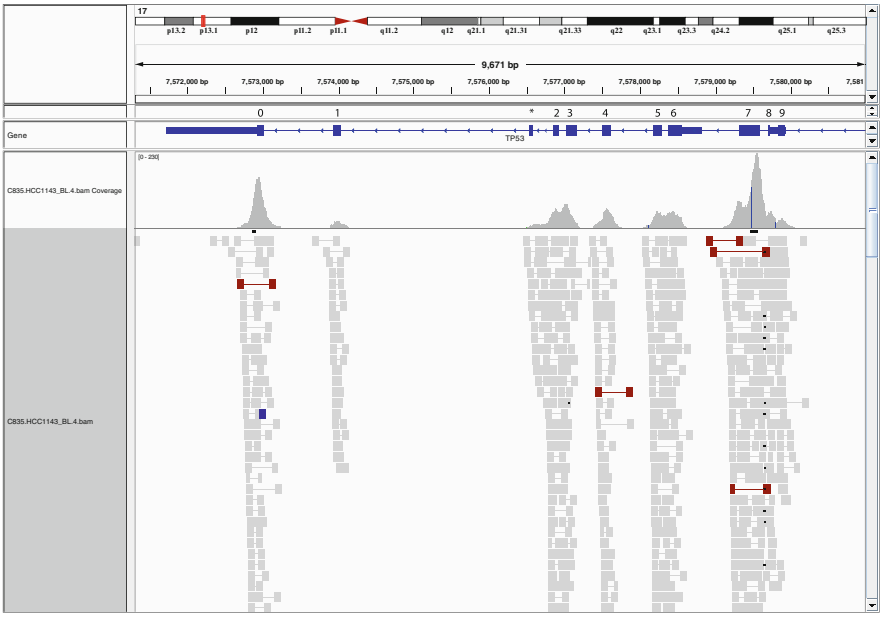

Fig. 18.3 The normal sample coverage plot for TP53 as obtained from the *Integrative Genomics Viewer*. The tall *blue rectangles* are the exons targeted in sequencing

and the shorter ones are start and end of untranslated regions (UTR). The labels for the target intervals in Table 18.1 were added to the figure generated by IGV. The labels are 0 to 9 for the 10 intervals in the table, and the third rectangle is labeled with a star as it does not appear to be a target interval in this sequencing experiment. In the bottom three-quarters of the figure, the coverage histogram is shown in the upper part and a stacked display of individual reads in the lower part.

Aspects of the data seen in the figure are:

- In order to achieve target coverage, the capture probes must be designed such that either end of the fragment falls on the target interval. This leads the coverage to extend beyond the target intervals (overhang on all target intervals).
- Overlap of fragments leads to non-uniform coverage within a target interval. This is attributable to varying widths of the fragments as well as tiling of capture probes. (Notice the bimodality of the coverage histogram for the eighth target.)
- Targets need not achieve the same average coverage as seen in the intervals labeled 0 and 7 having much higher coverage than the rest and the interval labeled 1 having a low coverage. Possible reasons for this are capture probe efficiency and interval characteristics such as size and GC content.

The figure provides several pieces of information on the individual reads. It color codes fragments in red to indicate that they are too wide compared to expected width and blue to indicate that these fragments are narrower than read length.

Such fragments are suggestive of insertions and deletions respectively. Other colors are used to indicate the two ends mapping to different parts of the genome which are inconsistent with the expected fragment lengths (see http://www.broadinstitute.org/software/igv/AlignmentData for details).

While a majority of fragments will be on target, there is a non-negligible proportion of fragments that are off-target and they can influence copy number computations which we will illustrate now. We obtained all the fragments (279,702 for normal and 339,030 for tumor) that are mapped to a 10 megabase region on chromosome 17 beginning at the 40 megabase mark. We binned the fragments by their midpoints into consecutive bins of length 100 bases where the genomic position a bin represents is its midpoint. We obtained the number of fragments in each bin for both normal and tumor samples. Of the 100,000 possible bins in the region, 13,297 had a nonzero count for at least one of the two samples. We expanded the target intervals in this region by 100 base pairs in both directions and derived the bins that intersect with the intervals. Of the 13,297 bins, only 6,835 of them do and hence are expected to have nonzero counts. However, bins with very small counts in the normal sample are inconsistent with the desired high coverage of the targets and thus are candidates for removal. There are 784 bins with fragment count of 2 or lower. Of the 6,439 bins that do not intersect with the target intervals 553 have fragment counts in the normal sample of at least 10, far more than the small number expected due to off-target fragments. Therefore, we included them in the copy number analysis. This results in 6,604 bins that are to be used in the copy number analysis and 6,693 bins to be discarded. The discarded bins account for just 5,807 fragments in the normal and 8,582 in the tumor (less than 3%). Fig. 18.4 shows the log-ratio computed as the ratio of scaled fragment counts where the grey and red points correspond to the included and discarded bins, respectively. Note that the red points form a band around zero with a significant presence near 1 and -1, which are the bins with one fragment in the tumor sample and zero in the normal, and vice versa. Despite the clear gain visible at the 46 megabase location, the loss in the 40–41.3 megabase region and focal loss around 42.7 megabase, the large number of red points in those regions will have an adverse effect on the copy number analysis, demonstrating the utility of pruning these bins. For the whole genome, binning the data results in 1,723,210 bins with nonzero counts in either sample of which 1,039,881 are to be discarded using the same consideration; they account for less than 4% of the total fragment count which is far lower than that expected from target efficiency.

The final piece of information needed for applying weighted CBS to the data are the weights assigned to the bins. The optimal weight for a bin is proportional to the variance of the fragment count for that bin which is a function of the unobserved rate parameter. The fragment counts which are the estimates of the rates are also very skewed thus using weights proportional to counts will make a handful of bins with large counts highly influential. Thus, we chose weights proportional to the logarithm of bin counts assigning greater weights to bins with large counts but protects against undue influence of bins with extreme counts. Although the optimal weights for the weighted t-statistic will depend on the mean counts of both the normal and the tumor

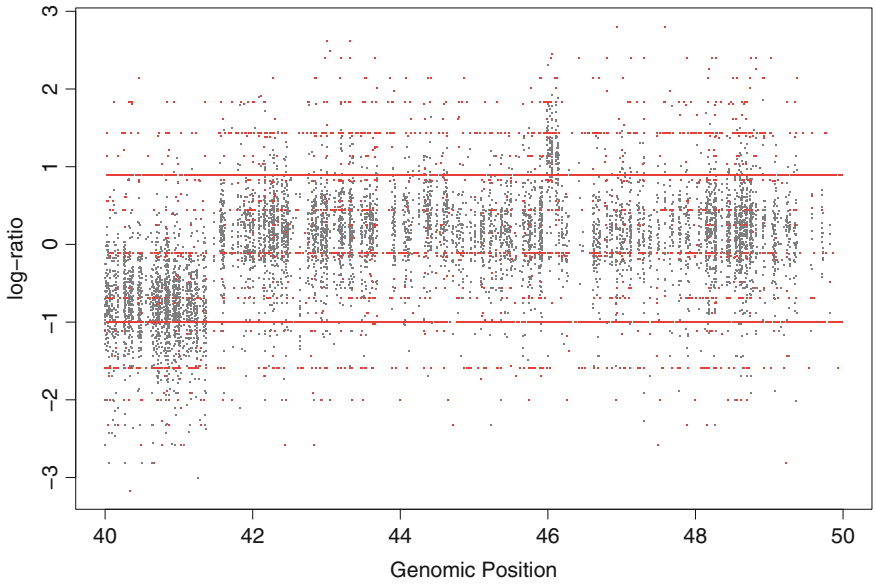

Fig. 18.4 The copy number log-ratio plot of the 40–50 megabase region on chromosome 17. The bins included in the analysis are in *grey* and the ones excluded are in *red*

samples, the tumor counts can change dramatically due to gains and losses. Thus, a more suitable choice of weights is to use the logarithm of just the normal counts (or median of several normal samples, if available).

Using the *DNAcopy* package, we segmented the logarithm of the ratio of scaled fragment counts for the bins to be used in the analysis. In Fig. 18.5, we show the whole genome copy number profile for this sample. The points are the log-ratio of the bins which are shown in alternate shades of grey to indicate different chromosomes. The algorithm segmented the genome into 419 regions with constant copy number which are shown as blue lines drawn at the level of the segment mean. The number of segments vary between chromosomes with the largest number (44) in chromosome 1 and the smallest number (7) in chromosome 22. The figure also shows the segmentation results from a SNP array analysis as red lines. Note that the SNP array data are not necessarily in the same scale and thus the red and blue lines may not overlap. Furthermore, since the SNP array probes cover the genome more uniformly than the targeted exome sequencing, there are far more red segments. However, the two sets of results show remarkable concordance except for chromosomes 2 and X, where the systematic large gap between the blue and red lines suggests that the cells used for exome sequencing have one fewer copy of these two chromosomes compared to the cells used for the SNP array.

In Fig. 18.6, we present a 25 megabase region on chromosome 17 to highlight the results. Note that while the exome segments (blue) and SNP segments (red) are similar, there are some locations where they differ. There is a small region

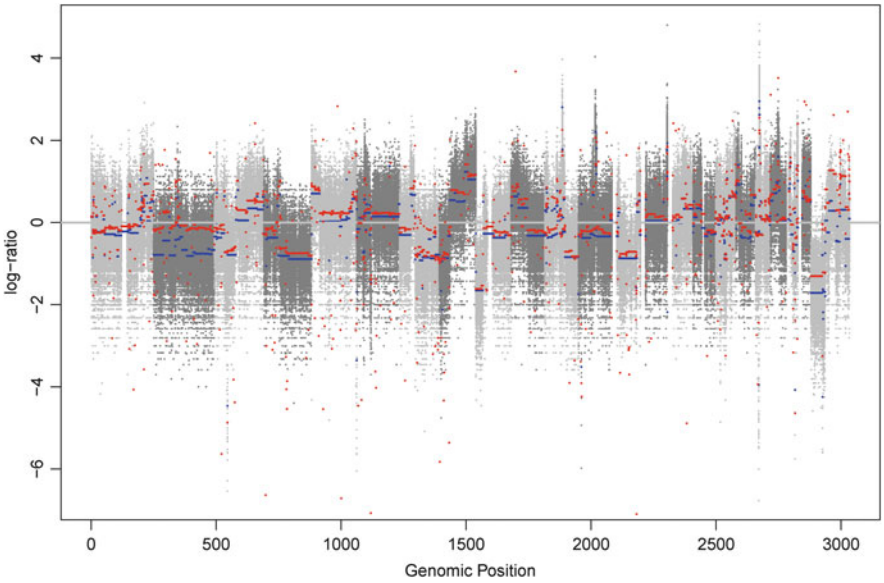

Fig. 18.5 The copy number profile of the whole genome. The chromosomes are colored in alternate shades of *grey*. The *blue lines* are regions of constant copy number identified from exome sequencing data. The *red lines* are the regions from SNP array data

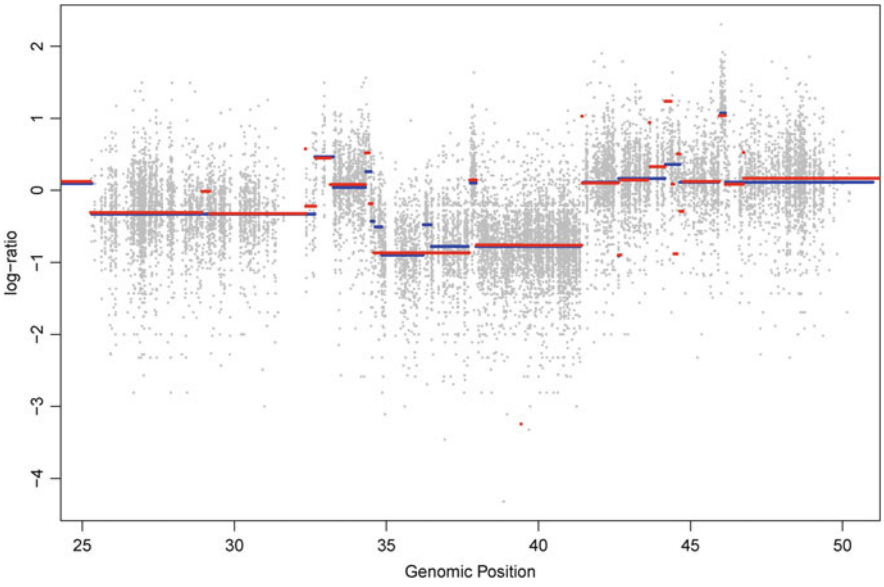

Fig. 18.6 The copy number profile of the 25–40 megabase region on chromosome 17. The *blue lines* are the segment means from exome sequencing data. The *red lines* from SNP array data

around 29 megabase and several small regions around 44 megabase the SNP array identifies that are not seen in the exome data. This may be attributable to the lack of data since the exome target intervals do not span the genome uniformly. In order to ascertain this, we reviewed the intervals identified by the SNP array and compared it to the target interval. The interval around 29 megabase spanned from positions 28,931,871 to 29,187,109 which is a 255 kilobase region. There are 17 target intervals in the whole exome sequencing in the last one third of this region starting from positions 29,111,193 and ending at 29,185,353 that covered just 4,676 bases. Likewise, the seven regions identified around 44 megabase in the SNP array covered an area that is 1.1 megabase long but were target poor for exome sequencing in that the target intervals only spanned 23 kilobases. The region between 35 and 38 megabase shows three segments for the exome where as just one for the SNP. This could be either due to higher resolution of exome data or the cell lines not being static.

It is common practice to undo small changes that do not meet a magnitude threshold. This occurs when a gentle wave in the data due to technical artifact looks like a change in mean. This step was not applied in the results presented as the goal was to present the full results. The overarching message from this analysis is that DNA sequencing, in particular targeted sequencing, can be successfully used to obtain copy number profiles.

18.5 Other Structural Variations

DNA sequencing can be used to identify other structural variants such as inversions and translocations. As seen in Sect. 18.2, the informative fragments for identifying these are those of Type 3 in Fig. 18.1. These are fragments that have high quality reads on both ends that are reliably mapped to the genome but are not properly paired. The improper pairing can occur due to an inter-chromosomal translocation, where the two reads are mapped to two different chromosomes, an intra-chromosomal translocation, where the reads from the two ends are mapped to the same chromosome but are directed away from each other rather than towards each other, or an inversion, where the reads from the two ends are mapped in the same direction. In all cases, the inferred fragment size is far larger than the expected fragment size. [Note: A proper pair can result in a large fragment size when there is a deletion in between the two reads; Fragments of type 4 in Fig. 18.1 can also be used for identifying these structural variations provided partial alignments can be done.]

The “bam” files used in this step have been deduplicated, realigned and their base quality scores recalibrated. The first step in identifying inversions and translocations is to extract all the improperly paired fragments where both reads are mapped to chromosome 1 through X and pass the instrument’s quality control. There are 158,433 such fragments in the normal sample and 145,858 in the tumor. Note that these counts are just 0.5% of the total number of fragments in the sample. This is

to be expected since the striped-shaded region junction (seen in Fig. 18.1) needed for these structural events are uncommon as most fragments are interior fragments (of Types 1 and 2). Additionally fragments of Types 4 and 5 will be unmapped using standard alignment software. Although a read may pass quality control as determined by the sequencing machine, the mapping quality of the read may not be high enough to provide valid information. Thus, we will use the mapping quality filter of 20 (possible error in alignment of 1%) to restrict the analysis to high quality reads. This reduced the number of fragments where both ends are mapped with a quality greater than 20 to 108,055 for the normal and 98,385 for the tumor. Note that there are more improperly paired fragments in the normal than the tumor. This might be due to the sequence similarity between different regions in the genome and hence mapping may not be unique and absolute.

A single fragment suggesting a structural variation is not a proof of it. The more the number of fragments indicating a structural variation the stronger the evidence. However, a somatic structural change acquired in the tumor will not be present in the germline. Hence one must verify that any structural variant identified in the tumor is present only in the tumor and not the germline. We begin this by counting the number of fragments from both tumor and normal samples that are in a neighborhood of every fragment. In the Example section earlier, we noted that most fragments are between 75 and 300 bases long. Thus, we define the neighborhood of a fragment to be within 1,000 bases of the starting location of the reads from both ends. Note that the neighborhood of each fragment will contain itself and hence the minimum fragment count is 1. Of the 206,440 combined fragments, only 3,042 have fragment counts greater than 1. Furthermore, if a fragment has several other fragments in its neighborhood, then all of them have this fragment in their neighborhoods as well. In fact, they cluster strongly and the 3,042 fragments with neighborhood count of more than one reduce to a far smaller number of clusters.

In Fig. 18.7, we display the fragment counts in the tumor plotted against the counts in the normal. The scatterplot shows that, in this data, there is a strong relationship between the tumor and normal counts suggesting that most of the suggested changes are present in both tumor and normal cells. In order to identify possible tumor specific changes we restricted ourselves to the fragments for which the normal count in the neighborhood is at most 3 and conducted a Binomial test for the hypothesis that the proportion of tumor counts out of the total is greater than 0.5. Table 18.2 lists the three clusters of fragments that are significant after adjusting for multiple comparison. The table gives the chromosome to which the fragments are mapped, the median start location of the first and second read, and the number of fragments in tumor and normal.

In Fig. 18.8, we show the copy number profiles, obtained using the abundance measure data, for these two regions. The top and bottom row of figures correspond to chromosomes 21 and 14, respectively. The first figure in each row shows the entire region where the start locations of the respective reads are marked by a vertical line. For chromosome 14, the two lines at position 105.412 megabase appear as one due to their closeness. The second and third figures in the top

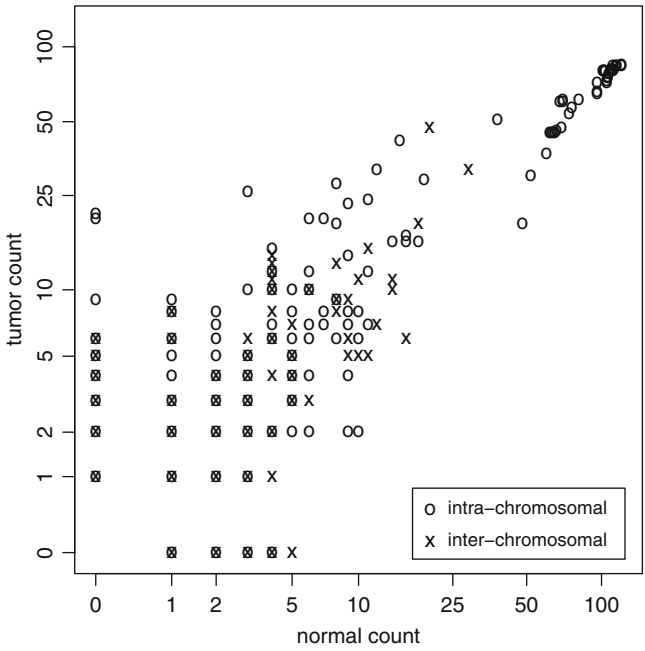

Fig. 18.7 The number of fragments in the neighborhood of an improperly paired fragment that belong to tumor and normal samples

Table 18.2 The details of the three clusters of fragments identified as present in tumor only

First read		Second read		Fragment count	
Chromosome	Location	Chromosome	Location	Tumor	Normal
21	43,246,325	21	47,347,121	21	0
14	106,471,416	14	107,282,893	20	0
14	105,412,008	14	105,412,453	26	3

row show the read locations of the first and second read are close to breakpoints identified in the copy number segmentation in the previous section and thus this rearrangement is consistent with copy number data. The second figure in the bottom row corresponds to the read location of 106.471 megabase in the second row of Table. 18.2. While this location is close to a break point, its companion is close to the end of chromosome with just two target intervals in its vicinity, and thus no additional information on the structural change is available. The third figure in the bottom row shows the two read locations in the third row and the two points in the interval between them that are seen in the figure are below the majority of the points in their vicinity. This suggests a small deletion since the two locations are just 445 bases apart. However the copy number segmentation does not pick them up as the magnitude of the change is within the noise of the copy number ratio. In all,

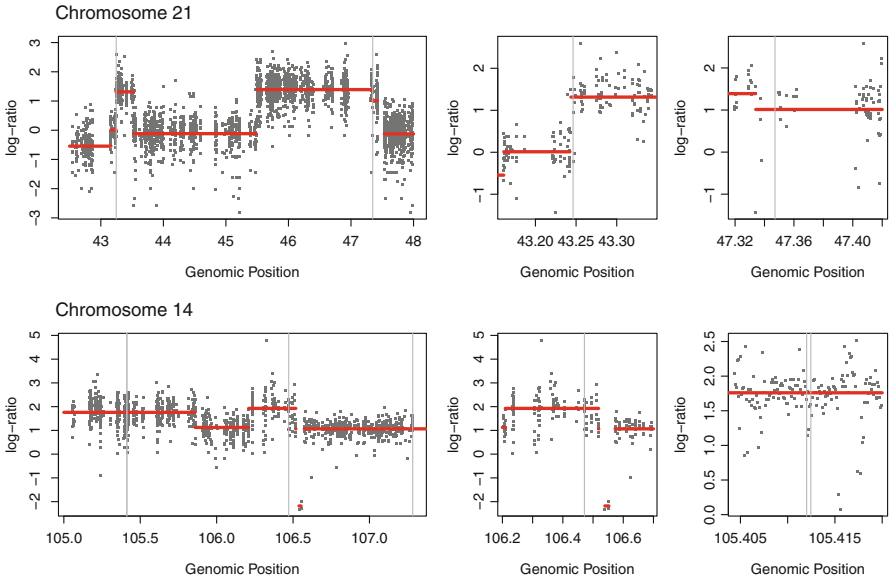

Fig. 18.8 The copy number profiles of the regions with plausible structural variants listed in Table 18.2

the presence of breakpoints in the copy number profile near the read locations of the plausible structural variations, lend support to their presence.

Although there are extensive structural rearrangements present in this cell line (the spectral karyotype of these cells is at http://www.path.cam.ac.uk/~pawefish/BreastCellLineDescriptions/HCC1143.html), we identified just 3 of them and none of the inter-chromosomal ones. Our inability to detect such an event is primarily due to fact that these data rose from a targeted sequencing and hence has large gaps in information. In order for targeted sequencing to be able to detect inversions and translocations, the junction (the striped-shaded region boundary in Fig. 18.1) should be close to a target interval and the capture probe should fully reside within the striped or the shaded region. This makes the likelihood of a fragment that contains an inversion or a translocation event being captured and sequenced very low. Thus whole genome sequencing is more apt for identifying structural rearrangements as it does not select for specific fragments to be sequenced and is thus far more likely to contain fragments with such events.

18.6 Summary

DNA sequencing, in particular targeted sequencing, is widely used in cancer research with the primary purpose of identifying somatic mutations. In this chapter, we adapted the Circular Binary Segmentation algorithm for the analysis of copy

numbers using DNA sequencing data. We showed using a whole exome sequencing dataset that copy number profile can be obtained from it. Despite the target intervals covering less than 2% of the genome, this profile is highly concordant with profile obtained from SNP array with whole genome coverage. The high coverage used in exome sequencing has the potential to identify intragenic changes such as deletion of a few exons which may not be feasible with current whole genome arrays.

DNA sequencing also provides information on polymorphic sites (SNPs) within the target intervals which in turn provides allele specific copy number information. We adapted CBS to obtain parent specific copy number profile from SNP array data [24] which can in turn be adapted to sequencing data. Similarly, the ASCAT algorithm, developed by Van Loo et al. (2010) [35], for the analysis of allele specific copy numbers can also be applied in the sequencing context. Such an analysis can provide additional information such as copy neutral loss of heterozygosity or uniparental disomy which enhances our understanding of the oncogenic process.

DNA sequencing contains three types of information - copy number, genotype and structural rearrangement. The methods we described treat these separately. However, since these data elements are not orthogonal to each other, there is potential to borrow information from all three types of data to develop a unified method to detect these structural variations. Other considerations such as the optimal bin size and the choice of filtering parameters and their effect should be studied for existing methods as well as those being developed.

The purpose behind studying structural variations is their impact on gene expression and their consequences. There is a positive correlation between copy numbers and gene expression. Likewise, the *bcr-abl* fusion protein provides a powerful example for the consequences of translocations. However, a comprehensive catalog of all possible events will require several tens of thousands of samples [17]. Thus careful consideration of the design of these experiments is essential. As we noted, targeted sequencing may not provide information on structural rearrangements but the high coverage that they can achieve to detect somatic mutations will be prohibitively resource intense for whole genome sequencing. Additionally, fusion transcripts are best detected using RNA sequencing. These aspects present a vibrant area for future research on how best to combine different sequencing methodologies to extract the information in a sample. A related problem is how best data from multiple samples can be combined to identify the affected biological processes and pathways and how they can be prioritized for further study.

Finally, the volume of data from these experiments are immense and will require efficient software for processing them. This presents a venue for the development of efficient methods and algorithms. In summary, DNA sequencing provides a wealth of data which can add to our knowledge with further research and proper analytic tools.

Acknowledgements The author thanks Arshi Arora for her programming assistance in processing the bam files and Drs. Glenn Heller, Jennifer Levine and Ronglai Shen for their valuable comments and suggestions. Supported by grants from the National Cancer Institute (CA163251, CA008748) and the Susan G. Komen for the Cure Foundation (IIR12221291).

References

[1] Abyzov, A., Urban, A.E., Snyder, M., Gerstein, M.: Cnvnator: an approach to discover, genotype, and characterize typical and atypical cnvs from family and population genome sequencing. Genome Res. **21**(6), 974–984 (2011)

[2] Anscombe, F.J.: The transformation of poisson, binomial and negative-binomial data. Biometrika **35**(3–4), 246–254 (1948)

[3] Berg, J.M., Tymoczko, J.L., Stryer, L.: Biochemistry. Seventh Edition. W. H. Freeman & Company, New York (2011)

[4] Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J.C., Huang, J.H., Alexander, S., et al.: Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc. Natl. Acad. Sci. **104**(50), 20,007–20,012 (2007)

[5] Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A., et al.: Absolute quantification of somatic dna alterations in human cancer. Nat. Biotechnol. **30**(5), 413–421 (2012)

[6] Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., et al.: Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Meth. **6**(9), 677–681 (2009)

[7] Chiang, D.Y., Getz, G., Jaffe, D.B., O'Kelly, M.J., Zhao, X., Carter, S.L., Russ, C., Nusbaum, C., Meyerson, M., Lander, E.S.: High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat. Meth. **6**(1), 99–103 (2009)

[8] Crick, F., et al.: Central dogma of molecular biology. Nature **227**(5258), 561–563 (1970)

[9] Ganem, N.J., Storchova, Z., Pellman, D.: Tetraploidy, aneuploidy and cancer. Curr. Opin. Genet. Dev. **17**(2), 157–162 (2007)

[10] Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. **5**(10), R80 (2004). URL http://www.bioconductor.org/

[11] Han, S.W., Kim, H.P., Shin, J.Y., Jeong, E.G., Lee, W.C., Lee, K.H., Won, J.K., Kim, T.Y., Oh, D.Y., Im, S.A., et al.: Targeted sequencing of cancer-related genes in colorectal cancer using next-generation sequencing. PloS One **8**(5), e64,271 (2013)

[12] Hayes, M., Pyon, Y.S., Li, J.: A model-based clustering method for genomic structural variant prediction and genotyping using paired-end sequencing data. PloS One **7**(12), e52,881 (2012)

[13] Kallioniemi, A., Kallioniemi, O.P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F., Pinkel, D.: Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science **258**(5083), 818–821 (1992)

[14] King, C.R., Kraus, M.H., Aaronson, S.A.: Amplification of a novel v-erbb-related gene in a human mammary carcinoma. Science **229**(4717), 974–976 (1985)

[15] Kumar, V., Abbas, A.K., Fausto, N., Aster, J.C.: Robbins and cotran pathologic basis of disease, Professional Edition: Expert Consult-Online. Elsevier Health Sciences, New York (2009)

[16] Lai, W.R., Johnson, M.D., Kucherlapati, R., Park, P.J.: Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data. Bioinformatics **21**(19), 3763–3770 (2005)

[17] Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., Getz, G.: Discovery and saturation analysis of cancer genes across 21 tumour types. Nature **505**, 495–501 (2014)

[18] Leavitt, S.A., et al.: Deciphering the genetic code: Marshall Nirenberg. http://history.nih.gov/exhibits/nirenberg/glossary.htm (2010)

[19] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al.: The sequence alignment/map format and samtools. Bioinformatics **25**(16), 2078–2079 (2009). URL http://samtools.sourceforge.net/

[20] du Manoir, S., Speicher, M.R., Joos, S., Schröck, E., Popp, S., Döhner, H., Kovacs, G., Robert-Nicoud, M., Lichter, P., Cremer, T.: Detection of complete and partial chromosome gains and losses by comparative genomic in situ hybridization. Human Genet. **90**(6), 590–610 (1993)

[21] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al.: The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. Genome Res. **20**(9), 1297–1303 (2010). URL http://www.broadinstitute.org/gatk/

[22] Morgan, M., Pagès, H., Obenchain, V.: Rsamtools: Binary alignment (BAM), variant call (BCF), or tabix file import. URL http://bioconductor.org/packages/release/bioc/html/Rsamtools.html

[23] Nowell, P.C., Hungerford, D.A.: A minute chromosome in chronic granulocytic leukemia. Science **132**, 1497–1501 (1960)

[24] Olshen, A.B., Bengtsson, H., Neuvial, P., Spellman, P.T., Olshen, R.A., Seshan, V.E.: Parent-specific copy number in paired tumor–normal studies using circular binary segmentation. Bioinformatics **27**(15), 2038–2046 (2011)

[25] Olshen, A.B., Venkatraman, E., Lucito, R., Wigler, M.: Circular binary segmentation for the analysis of array-based dna copy number data. Biostatistics **5**(4), 557–572 (2004)

[26] Picard: Command-line utilities that manipulate SAM files. URL http://picard.sourceforge.net/index.shtml

[27] Pinkel, D., Albertson, D.G.: Array comparative genomic hybridization and its applications in cancer. Nat. Genet. **37**, S11–S17 (2005)

[28] Pounds, S., Cheng, C., Li, S., Liu, Z., Zhang, J., Mullighan, C.: A genomic random interval model for statistical analysis of genomic lesion data. Bioinformatics **29**(17), 2088–2095 (2013)

[29] Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al.: Global variation in copy number in the human genome. Nature **444**(7118), 444–454 (2006)

[30] Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P.: Integrative genomics viewer. Nat. Biotechnol. **29**(1), 24–26 (2011). URL http://www.broadinstitute.org/igv/

[31] Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., et al.: Large-scale copy number polymorphism in the human genome. Science **305**(5683), 525–528 (2004)

[32] Seshan, V.E., Olshen, A.: DNAcopy: DNA copy number data analysis. URL http://bioconductor.org/packages/release/bioc/html/DNAcopy.html

[33] Shen, J.J., Zhang, N.R.: Change-point model on non-homogeneous poisson processes with application in copy number proling by next-generation dna sequencing. https://statistics.stanford.edu/sites/default/files/BIO%20257.pdf (2011)

[34] Taylor, B.S., Barretina, J., Socci, N.D., DeCarolis, P., Ladanyi, M., Meyerson, M., Singer, S., Sander, C.: Functional copy-number alterations in cancer. PLoS One **3**(9), e3179 (2008)

[35] Van Loo, P., Nordgard, S.H., Lingjærde, O.C., Russnes, H.G., Rye, I.H., Sun, W., Weigman, V.J., Marynen, P., Zetterberg, A., Naume, B., et al.: Allele-specific copy number analysis of tumors. Proc. Natl. Acad. Sci. **107**(39), 16,910–16,915 (2010)

[36] Venkatraman, E., Olshen, A.B.: A faster circular binary segmentation algorithm for the analysis of array cgh data. Bioinformatics **23**(6), 657–663 (2007)

[37] Wang, J., Mullighan, C.G., Easton, J., Roberts, S., Heatley, S.L., Ma, J., Rusch, M.C., Chen, K., Harris, C.C., Ding, L., et al.: Crest maps somatic structural variation in cancer genomes with base-pair resolution. Nat. Meth. **8**(8), 652–654 (2011)

[38] Wang, S., Gao, J., Lei, Q., Rozengurt, N., Pritchard, C., Jiao, J., Thomas, G.V., Li, G., Roy-Burman, P., Nelson, P.S., et al.: Prostate-specific deletion of the murine pten tumor suppressor gene leads to metastatic prostate cancer. Canc. cell **4**(3), 209–221 (2003)

[39] Willenbrock, H., Fridlyand, J.: A comparison study: applying segmentation to array cgh data for downstream analyses. Bioinformatics **21**(22), 4084–4091 (2005)
[40] Xie, C., Tammi, M.T.: Cnv-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinform. **10**(1), 80 (2009)
[41] Zhao, M., Wang, Q., Wang, Q., Jia, P., Zhao, Z.: Computational tools for copy number variation (cnv) detection using next-generation sequencing data: features and perspectives. BMC Bioinform. **14**(Suppl 11), S1 (2013)