

Detection of Copy Number Variations (CNVs) Based on the Coverage Depth from the Next Generation Sequencing Data

Yanming Feng, David Chen, and Lee-Jun C. Wong

Abstract Intragenic copy number variations (CNVs) in the human genome are significant contributors to the inherited genetic disorders. Currently the most established methods to detect CNVs are array comparative genomic hybridization (aCGH) and MPLA. With the fast adaption of next generation sequencing (NGS) in the clinical sequencing, increasing interest has been attributed to the detection of CNV from NGS data. In this chapter, we describe an easy-to-implement strategy to detect and visualize exonic CNVs from captured NGS data, as well as the confirmation. We also discuss the specificity and sensitivity of this strategy.

Keywords Exonic deletion • aCGH • Molecular diagnosis • Copy number variation • CNV • Next generation sequencing • NGS

1 Introduction

Intragenic copy number variations (CNVs) in the human genome are significant contributors to the inherited genetic disorders [10, 14]. It has been reported that approximately 12% of the human genome has CNV [11]. The pathogenicity of CNVs is variable, and the role of some pathogenic CNVs is still unknown. Intragenic CNVs involving genes matching the clinical phenotype are most likely pathogenic due to the change in gene dosage (whole gene deletion/duplication) or the disruption of the gene (out-of-frame exonic deletion/duplication). In clinical settings, if only a heterozygous pathogenic variant is identified in the candidate

Y. Feng, Ph. D. (✉) • D. Chen
Baylor Genetics, 2450 Holcombe BVLD, Houston, TX 77021, USA
e-mail: yanmingf@bcm.edu

L.-J.C. Wong
Baylor Genetics, 2450 Holcombe BVLD, Houston, TX 77021, USA
Department of Molecular and Human Genetics, Baylor College of Medicine,
One Baylor Plaza, Houston, TX 77030, USA

gene for an autosomal recessive disorder by sequencing, and the phenotype is consistent with the disease gene, search for the second mutant allele shifts to the identification of intragenic deletions or duplications.

Exon targeted array comparative genomic hybridization (aCGH) is currently the most commonly used approach for the detection of exonic CNVs [17, 18]. Since backbone probes throughout the genome are included in the exon targeted array, resolutions are ranging from a few hundred bases to kilobases (Kb), to megabases (Mb), and even the entire chromosome [17, 18]. Multiplex ligation-dependent probe amplification (MLPA) is another commonly used method for CNV detection. However, specially designed probes for individual exons are required, thus, it is difficult to use MLPA for large scale CNV analysis. Various methodologies for CNV detection may not be readily available to some clinical laboratories or the assays developed by individual laboratories may not include a complete set of genes or exons of interest for technical or commercial reasons.

In recent years, next generation sequencing (NGS) technologies have been widely used in the clinical practice of molecular diagnosis of human genetic diseases [5, 15, 16, 23]. Since 85% of all known mutations are located in the coding regions and the intron/exon junctions [4], capture-based target gene enrichment followed by NGS analysis has been a cost effective way to identify point mutations and small indels that are less than 20 bp in the target genes. NGS with consistently deep coverage of individual target exons can potentially provide an opportunity for concurrent detection of copy number changes and point mutations in patients with inherited disorders.

2 Strategies for NGS Based CNV Detection

NGS based CNV detection strategies can be divided into four categories based on sequence reads and coverage depth: (1) Paired-end mapping [2, 7, 8]; (2) Split-read [20]; (3) Depth of coverage [1, 3, 12, 19, 21]; and (4) Assembly based [9, 13, 22].

2.1 Paired-End Mapping Method

Paired-end mapping (PEM) methods require paired-end reads. The distance of paired-end reads is predetermined. If the distance of a pair of mapped reads is significantly larger than the distribution of the predetermined distance, a possible insertion may be identified. If shorter, a possible deletion can be identified. Some programs have been developed using paired-end mapping method, such as PEMer, BreakDancer, and Variation Hunter [2, 7, 8].

2.2 *Split-Read*

Split-read method also need paired-end reads. Unlike paired-end mapping method, in which the break points are not in the reads, the split-read method need one perfectly matched read and one read contains the breakpoint so that this read cannot be perfectly mapped to the reference genomic sequence. This unmatched read is then split into several fragments, and the first and the last fragments are mapped to the reference genomic sequence. The unmatched reads are split into several short fragments too short to be mapped to the genomic reference sequence. This split-read method usually requires long reads. Pindel is a split-read based program [20].

2.3 *Depth of Coverage Based*

The depth of coverage information is embedded in all NGS data, thus, depth of coverage based methods have become the main method for CNV detection. NGS results from both paired-end and single-end reads can be used for coverage depth based methods. Many programs have been developed using the depth of coverage information, such as SegSeq, CNVseq, Rdxplorer, CNVnator, and ExomeCNV [1, 3, 12, 19, 21]. The fundamental hypothesis of the depth of coverage based method is that the coverage is related to the copy number.

2.4 *Assembly Based*

In paired-end mapping, split-read, and depth of coverage methods, the reads need to be mapped to a reference genomic sequence. In contrast, the assembly-based method does not need a reference genome to map the reads. Instead, the reads are assembled without a reference genomic sequence. The assembled sequence is then compared to the genome sequence. The difference usually contains the structural variation information, including CNV. Velvet, ABySS and SOAPdenovo are all assembly based method [9, 13, 22].

Each of these strategies has its own strength and weakness, and maybe adopted for different purposes. The paired-end mapping based methods and the split-read methods can indicate the location of the CNV so it is easier to find the breakpoint. However, they cannot determine the exact copy number. These two methods also require paired-end reads. The depth of coverage (DOC) based method does not need additional specific algorithm because DOC information is already embedded in all NGS data. This is an important advantageous point because in clinical settings, the major NGS approach is captured based, either target panel or whole exome sequencing (WES). Thus, DOC strategy is readily applicable. The assembly based method

is different from the other three in that it does not need reference genome sequence for mapping. However, it does need long reads with continuous coverage, thus, both the data collection and processing are time and cost consuming. The method is the least commonly used.

3 Procedures to Detect CNVs Based on Depth of Coverage

3.1 *Reference Samples*

Most DOC based CNV detection methods share the similar principle that is to compare the average coverage depth of a test sample to DOC of a reference. DOC of a reference is usually the mean or medium DOC of a group of samples that are analyzed in the same batch. An ideal reference is with the lowest coefficients of variations in the coverage depth. A few factors may contribute to variations. One is that there are intrinsic CNVs in the reference samples. These CNVs could be present in any samples depending on their allele frequencies. These are most likely benign. The others are rare, clinically significant CNVs that may be associated with disease phenotypes. We can select reference samples that do not contain CNVs in the genes of interest. These samples may be available publicly or in the individual laboratories that have validated the reference samples by a second method, such as aCGH. Still, variations in coverage depth maybe due to batch effects, including sample quality, sample processing, sample or exon specific differences, as well as instrumentation, technical, and other experimental variations. These types of variations are usually characterized and minimized during validation steps, although they cannot be completely removed. The reference DOC file can be generated by averaging DOC from a group of samples that do not contain CNVs in regions of interests. Since pathogenic CNVs are rare, to further minimize variations from batch effects, in routine practice, NGS results of at least 20 samples performed under the same conditions as the testing samples are grouped to generate the reference file.

3.2 *DOC Based CNV Detection Using Exon as Sliding Window*

Unlike whole genome sequencing, in which sequence data are continuous, the fundamental elements of capture based NGS are exons. Capture probes are designed for individual exons as regions of interests. NGS reads are grouped by exons and are not continuous due to the interruption of introns that are not captured and sequenced. Therefore, it is most reasonable to use exon as the sliding window.

3.3 Normalization of the Depth of Sequence Read to the Total Amount of DNA Loaded to Sequencing Machine

The amount of DNA template loaded to the sequencing machine naturally determines the total sequence reads generated, thus, it also affects the depth of coverage of individual exons. Although the loading amount of DNA template is carefully controlled for each sample, variation among different samples is inevitable. For CNV detection, accurate quantification of the number of sequence read is critical because the read depth is what CNV detection based upon. The DOC in the NGS data is not only determined by the copy number, but also by the amount of total target DNA loaded unto the flow cell and sequenced. Thus, before the DOC of testing sample and reference sample is compared, the total coverage of each individual sample is normalized for equal loading of total DNA template, which is determined by the total mapped reads.

3.4 Generation of Reference File

The reference file of DOC of exons is essentially the average DOC of a group of selected samples performed in the same NGS batch. There are two important values in the reference file that is used for exon based CNV detection algorithm. One is the mean value (μ) of the first normalized DOC of an exon, which is later used for the testing sample normalization/comparison. The other is the standard deviation (σ) of this mean value, from which the coefficient of variation (CV) is obtained. CV is an indicator of the quality of the reference file.

3.5 Normalization of DOC of the Testing Sample (Second Normalization)

Unlike reference samples, in order to detect CNVs, the DOC of the testing sample is normalized twice. First, it is similar to reference samples, the DOC of each exon in the testing sample is normalized against the total mapped reads. The normalized DOC of an exon is then normalized again to the mean DOC (μ) of the corresponding exon. The mean DOC (μ) is the average DOC of a specific individual exon in the reference file.

3.6 Detection and Visualization of CNVs

Ideally, the final normalized DOC of an exon with normal copy number is 1 or around 1. The secondary normalized DOC is 0.5 for exons with heterozygous deletion, and 0 for homozygous deletion. Duplication with a total of 3copies, the

Index	Gene	CDS	Average Reads	Normalized	Reference	Norm/Ref	copies	Standard Deviation	CV	CNV Call
948	PHKB	1	185.19	1.67E-05	1.46E-05	1.1474937	2.294987	2.91E-06	0.199643	-
949	PHKB	1b	1024.4	9.25E-05	0.0001054	0.8779857	1.755971	1.21E-05	0.115065	-
950	PHKB	2	589.23	5.32E-05	0.0001076	0.4947187	1.040827	1.86E-05	0.172962	del
951	PHKB	3	533.32	4.82E-05	0.000107	0.450048	0.946845	1.71E-05	0.159316	del
952	PHKB	4	640.32	5.78E-05	0.0001245	0.4644381	0.977121	1.83E-05	0.1471	del
953	PHKB	5	488.39	4.41E-05	9.67E-05	0.4564205	0.960252	1.78E-05	0.184325	del
954	PHKB	6	520.8	4.70E-05	8.73E-05	0.5387064	1.133372	1.90E-05	0.217116	del
955	PHKB	7	231.45	2.09E-05	5.31E-05	0.3934672	0.827807	1.10E-05	0.206516	del
956	PHKB	8	566.67	5.12E-05	0.0001069	0.4786702	1.007063	2.34E-05	0.218864	del
957	PHKB	9	455.19	4.11E-05	9.07E-05	0.4535737	0.954263	1.66E-05	0.1836	del
958	PHKB	10	565.74	5.11E-05	0.0001095	0.466699	0.981877	1.77E-05	0.161279	del
959	PHKB	11	782.62	7.07E-05	6.94E-05	1.0190121	2.038024	1.60E-05	0.229942	-
960	PHKB	12	845.94	7.64E-05	8.03E-05	0.9520423	1.904085	1.83E-05	0.228602	-
961	PHKB	13	1113.9	0.0001006	0.0001156	0.8705247	1.741049	1.72E-05	0.148717	-
962	PHKB	14	1179.4	0.0001065	0.0001188	0.8965683	1.793137	1.58E-05	0.133099	-
963	PHKB	15	829.32	7.49E-05	7.53E-05	0.9951613	1.990323	1.00E-05	0.132983	-
***	***	***	***	***	***	***	***	***	***	***

Fig. 1 Example of DOC normalization and CNV call from NGS data. *Norm/Ref* is the final normalized coverage. CV is the coefficient of variation of the reference. The values in the CNV Call column are the automatic CNV calls based on the Norm/Ref value and the CV value

normalized DOC is 1.5. However due to the technical variation and various genomic properties, the final normalized DOC is in a range. Different exons have different variations. We developed a combo CNV detection and visual checking algorithm, which includes automatic CNV detection from the statistical aspect, and a visualization method for visual checking. To balance the sensitivity and specificity and avoid false negatives, we have these settings: (1) if the normalized value is less than $1-1.5CV$, it is scored as a deletion; (2) if the normalized value is greater than $1 + 1.5CV$, it is scored as duplication; (3) if the normalized value is in between, then, it is considered normal. An example of heterozygous deletion of E2-E10 of *PHKB* is shown in Fig. 1, in which each exon captured and sequenced is normalized and CNV is scored as described above. In this figure, column Norm/Ref is the final normalized DOC. Column CV is the coefficient of variation of the reference. The values in the CNV Call column are automatic CNV calls based on the Norm/Ref value and the CV value.

We have also generated a custom UCSC track file from the normalized DOC and the genomic coordinates. This file can be uploaded to UCSC genome browser to visualize the results. One advantage of the customized track file is that multiple samples can be visually simultaneously and compared. An example is shown in Fig. 2, in which four custom tracks in the figure represent the final normalized DOC of *PHKB* exons of four different samples, including one positive sample in the blue box, which has *PHKB* E2-E10 heterozygous deletion. Each vertical bar represents an exon. The height of the bar is the copy number of this exon. Exons in the red box are exons with only one copy (heterozygous deletion).

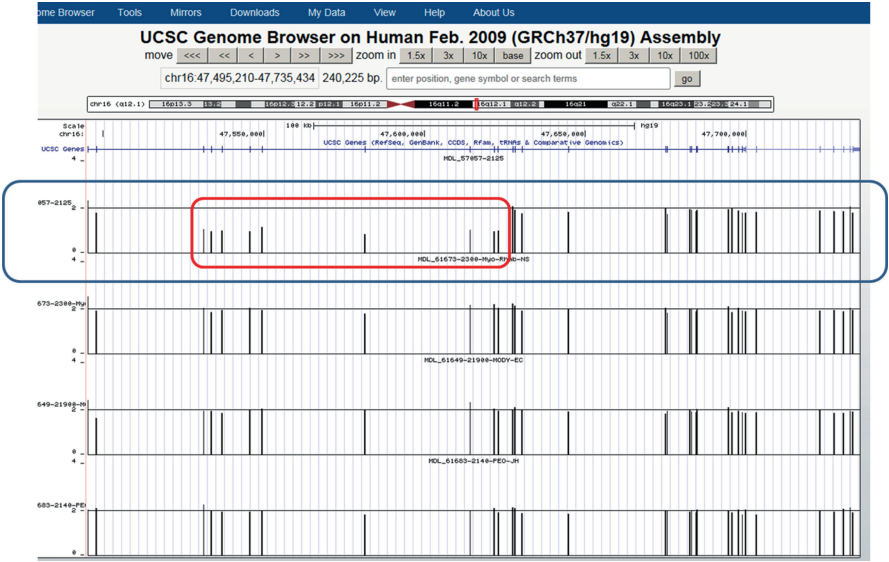


Fig. 2 Example of visualization of CNV call using UCSC genome browser with custom track. Four samples including one positive sample *circled in blue* and three negative samples are displayed together, along with the UCSC genes track which shows the genes and exons information. Nine exons with heterozygous deletions *circled in red* have half of the bar height of other normal exons

4 Confirmation of CNV

CNVs detected by DOC from NGS data can be confirmed by a second method, such as MPLA, aCGH or long range PCR (LR-PCR). High density aCGH is often used because it has the ability to reveal the boundary of the CNVs, if the breakpoints are not in the targeted exons. MLPA and LR-PCR are fast and cost effective ways to confirm exonic small CNVs and concurrent CNVs. Figure 3 is an example of the aCGH confirmation of a heterozygous deletion of E2-E10 of *PHKB* identified by using coverage based NGS data.

5 Sensitivity and Specificity

The sensitivity and specificity of NGS coverage based CNV detection was described previously [6]. In this paper, 12 validation samples were performed both NGS and aCGH, and the CNV detection results were compared. The total number of exons included in the comparison is 25,608. The sensitivity for the detection of deletion is 100% (9/9), but only 66.7% (2/3) for duplication. The specificity for the

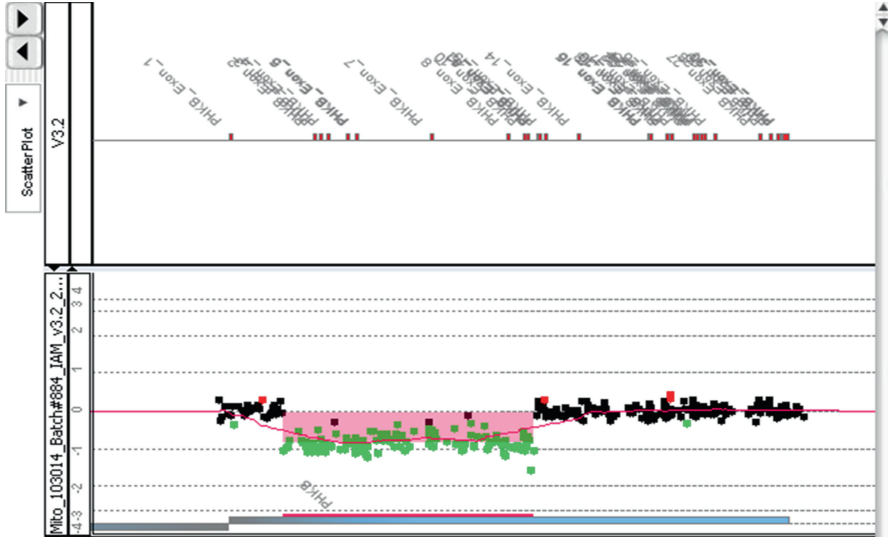


Fig. 3 aCGH confirmation of heterozygous deletion of PHKB E2-E10. Probes in green shows 1 copy of E2 to E10. Since there are probes in the intron region, aCGH usually can provide more information on the boundary of the deletion

detection of deletion and duplication is 99.92% and 99.86%, respectively. NGS coverage based CNV analysis is able to detect all deletions confirmed by aCGH at the single exon level without any false negative. The false positive rate of NGS based method is much higher for duplications (94.7%) than deletions (68.9%). The positive predicative value of duplication detection is only 5.3% (2/38). Even though all copy number losses detected by aCGH have been detected by NGS based analysis, the positive predicative rate is only 31% (9/29). This implies that all deletions detected by NGS based method require further confirmation with a second method, if the approach is to be used for clinical diagnostic purpose. In contrast, the negative predicative values for both deletions and duplications are 100%. This would suggest that a testing sample can be considered negative if the NGS based CNV analysis is negative.

6 Challenges and Issues

The most decisive step in the captured based NGS is the hybridization during the library preparation, which is affected by technical conditions and DNA properties. One outstanding factor is the GC content of the DNA. High GC content DNA is usually captured not as consistently as DNA with normal GC content. Some algorithms have been developed to correct the effect of GC content. However, our experience indicates that DNA with high GC content is more sensitive to subtle changes in experimental conditions during the hybridization step than DNA with normal GC

content. So far, no good algorithms are able to take this into account effectively. Exons with high GC content often show high coefficient of variations (CV). Fortunately, overall only less than 2% of all exons have high CV that CNVs cannot be determined reliably.

Another issue is the effect of homologous regions and pseudogenes on capture and sequencing coverage depth. Due to the presence of off-target high homologous sequences, the NGS data alignment software (aligner) sometimes cannot differentiate them to map sequences correctly. Therefore, the DOC may be distorted.

7 Future

Currently, the NGS based CNV detection algorithms have made great progress in the clinical utility in the diagnosis of inherited Mendelian diseases, in which the copy number of DNA is an integer, for example, 0, 1, 2 or 3. However, clinical utility of NGS has been gradually expanded to the detection of somatic mutations, in which the fraction of pathogenic variants is not present at 0%, 50% or 100%, as that is generally true for Mendelian mutations. It will be challenging and meaningful to investigate the performance of CNV detection in this situation.

Availability A script for the detection of CNVs used in this chapter was developed in Ruby and is available at <https://sourceforge.net/projects/cnvanalysis>.

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. 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. Methods.* **6**(9), 677–681 (2009)
3. 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. Methods.* **6**(1), 99–103 (2009)
4. Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloglu, A., Ozen, S., Sanjad, S., et al.: Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **106**(45), 19096–19101 (2009)
5. Cui, H., Li, F., Chen, D., Wang, G., Truong, C.K., Enns, G.M., Graham, B., Milone, M., Landsverk, M.L., Wang, J., et al.: Comprehensive next-generation sequence analyses of the entire mitochondrial genome reveal new insights into the molecular diagnosis of mitochondrial DNA disorders. *Genet. Med.* **15**(5), 388–394 (2013)
6. Feng, Y., Chen, D., Wang, G.L., Zhang, V.W., Wong, L.J.: Improved molecular diagnosis by the detection of exonic deletions with target gene capture and deep sequencing. *Genet. Med.* **17**(2), 99–107 (2015)

7. Hormozdiari, F., Alkan, C., Eichler, E.E., Sahinalp, S.C.: Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.* **19**(7), 1270–1278 (2009)
8. Korb, J.O., Abyzov, A., Mu, X.J., Carriero, N., Cayting, P., Zhang, Z., Snyder, M., Gerstein, M.B.: PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.* **10**(2), R23 (2009)
9. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., et al.: De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**(2), 265–272 (2010)
10. Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K., et al.: Mapping copy number variation by population-scale genome sequencing. *Nature*. **470**(7332), 59–65 (2011)
11. 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)
12. Sathirapongsasuti, J.F., Lee, H., Horst, B.A., Brunner, G., Cochran, A.J., Binder, S., Quackenbush, J., Nelson, S.F.: Exome sequencing-based copy-number variation and loss of heterozygosity detection: exomeCNV. *Bioinformatics*. **27**(19), 2648–2654 (2011)
13. Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I.: ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**(6), 1117–1123 (2009)
14. Stankiewicz, P., Lupski, J.R.: Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61**(1), 437–455 (2010)
15. Tang, S., Wang, J., Zhang, V.W., Li, F.Y., Landsverk, M., Cui, H., Truong, C.K., Wang, G., Chen, L.C., Graham, B., et al.: Transition to next generation analysis of the whole mitochondrial genome: a summary of molecular defects. *Hum. Mutat.* **34**(6), 882–893 (2013)
16. Wang, J., Cui, H., Lee, N.C., Hwu, W.L., Chien, Y.H., Craigen, W.J., Wong, L.J., Zhang, V.W.: Clinical application of massively parallel sequencing in the molecular diagnosis of glycogen storage diseases of genetically heterogeneous origin. *Genet. Med.* **15**(2), 106–114 (2013)
17. Wang, J., Zhan, H., Li, F.Y., Pursley, A.N., Schmitt, E.S., Wong, L.J.: Targeted array CGH as a valuable molecular diagnostic approach: experience in the diagnosis of mitochondrial and metabolic disorders. *Mol. Genet. Metab.* **106**(2), 221–230 (2012)
18. Wong, L.J., Dimmock, D., Geraghty, M.T., Quan, R., Lichter-Konecki, U., Wang, J., Brundage, E.K., Scaglia, F., Chinault, A.C.: Utility of oligonucleotide array-based comparative genomic hybridization for detection of target gene deletions. *Clin. Chem.* **54**(7), 1141–1148 (2008)
19. Xie, C., Tammi, M.T.: CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinf.* **10**, 80 (2009)
20. Ye, K., Schulz, M.H., Long, Q., Apweiler, R., Ning, Z.: Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. **25**(21), 2865–2871 (2009)
21. Yoon, S., Xuan, Z., Makarov, V., Ye, K., Sebat, J.: Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* **19**(9), 1586–1592 (2009)
22. Zerbino, D.R., Birney, E.: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**(5), 821–829 (2008)
23. Zhang, W., Cui, H., Wong, L.J.: Application of next generation sequencing to molecular diagnosis of inherited diseases. *Top. Curr. Chem.* **336**, 19–45 (2012)

Next Generation Sequencing Based Clinical Molecular
Diagnosis of Human Genetic Disorders

Wong, L.-J.C. (Ed.)

2017, VIII, 364 p. 23 illus., 17 illus. in color., Hardcover

ISBN: 978-3-319-56416-6