Quantitative Biology 2018, 6(2): 99–112
https://doi.org/10.1007/s40484-018-0137-6

REVIEW

Copy number variation related disease genes

Chaima Aouiche, Xuequn Shang* and Bolin Chen

School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
* Correspondence: shang@nwpu.edu.cn

Received September 10, 2017; Revised December 13, 2017; Accepted January 23, 2018

***Background*: One of the most important and challenging issues in biomedicine and genomics is how to identify disease related genes. Datasets from high-throughput biotechnologies have been widely used to overcome this issue from various perspectives, *e.g.*, epigenomics, genomics, transcriptomics, proteomics, metabolomics. At the genomic level, copy number variations (CNVs) have been recognized as critical genetic variations, which contribute significantly to genomic diversity. They have been associated with both common and complex diseases, and thus have a large influence on a variety of Mendelian and somatic genetic disorders.**

***Results*: In this review, based on a variety of complex diseases, we give an overview about the critical role of using CNVs for identifying disease related genes, and discuss on details the different high-throughput and sequencing methods applied for CNV detection. Some limitations and challenges concerning CNV are also highlighted.**

***Conclusions*: Reliable detection of CNVs will not only allow discriminating driver mutations for various diseases, but also helps to develop personalized medicine when integrating it with other genomic features.**

Keywords: CNV; disease gene; complex disease; targeted approach; genome-wide approach; whole exome sequencing

Author summary: In this review, we introduce a type of critical genetic variations at the genomic level, copy number variations (CNVs), which contribute significantly to genomic diversity and have proven to be related with a variety of common and rare complex diseases, phenotypes and genetic syndromes as well. Detecting those CNVs plays important roles in these genetic diseases through: (i) identifying disease related genes, their loci and breakpoints, (ii) allowing discrimination of driver mutation for pathogenesis or diagnosis of complex diseases, and (iii) helping to develop personalized medicines. CNV and its integration with other genomic features will help us understand disease susceptibility and pathogenesis from various perspectives.

INTRODUCTION

Many genetic diseases are not recognized as the result of dysfunction of a single gene [1], but rather related to variations and mutations of multiple genes or their interplays, such as: (i) over- and under-expression of multiple genes [2], (ii) duplicating-/removing of copies of several genes and also (iii) hypo- and hyper-methylation of multiple genes. Thus, identification of mutated genes responsible for specific diseases still remains a challenging issue [3]. Reliable detection of these genomic variations and mutated genes is fundamentally important for us to understand the mechanism of many disease or genetic disorder, such as cancers, diabetes, neuropsychiatric disorders [4,5], birth defects, autoimmune disorders, autism and even susceptibility to HIV.

Nowadays, with the development of high-throughput genomic technologies, it has become easy and cost-effective to comprehensively characterize various complex diseases by using a wide range of genomic datasets, epigenomic datasets, transcriptomic datasets, proteomic datasets, and metabolomic datasets [6]. Specifically: (i) single-nucleotide polymorphism (SNP), copy number variation (CNV), loss of heterozygosity (LOH), genomic rearrangement are datasets at the genome level; (ii) DNA methylation, histone modification, chromatin accessibility, transcription factor (TF) binding and micro RNA

(miRNA) are datasets at the epigenome level; (iii) gene expression and alter-native splicing are datasets at the transcriptome level; (iv) protein expression and post-translational modification are datasets at the proteome level; and (v) metabolite profiling is the dataset at the metabolome level.

Copy number variation (CNV) is one of the most important human genetic variations, which consists of not only sequence variants but also structural variants within populations. Although many genetic variants do not cause overt diseases, they influence disease susceptibility or drug response. Therefore, these CNVs have drawn attention of some scientists and had been recognized as novel genetic variations related to the genomic disease.

The aim of this review is to give insights into the important role of CNVs in the identification of disease related genes. The rest of the paper is organized as follows. In the section of Copy Number Variation Overview, we give a brief overview about the definition of CNVs and their relation with genomic diseases and detection methods. Then, we highlight the most significant disease genes determined using CNV methods. Finally we explain the successful use of this genetic variant in diseases, especially its integration with other genomics data, which will definitely be helpful for identifying new disease genes.

COPY NUMBER VARIATION OVERVIEW

Along with SNPs (which has been among the most abundant genetic variation in humans), CNVs have attracted many attentions, since they refer to a type of intermediate-scale structural variants (SVs) in the genome. A lot of definitions have been given by many researchers to CNV as follows:

- CNVs are DNA segments presenting at variable copy numbers and contribute to a substantial proportion of the variation in a genome owing to their large size [7,8].
- CNVs refer to large-scale (>1 kb) chromosomal copy number changes, *e.g.*, amplifications or deletions compared to a reference genome [9].
- CNVs are deletions or duplications of size (>1 kb) genomic area [10].
- CNVs are inherited or *de novo* structural variations, including all kinds of genomic variations larger than 1 kb, such as insertions, deletions and duplications [11,12].

Briefly, CNVs are defined as either the gain (duplication) or loss (deletion) of a stretch of DNA as compared with a reference genome. They are characterized by the break point loci (starting and ending points), single copy length and number of copies, and they may range in size from a kilobase to several megabase or even an entire chromosome. As depicted in Figure 1.

Additionally, CNVs involve more genomic sequences than SNPs and have potentially greater effects, including alteration of gene dosage, disruption of genes or perturbation of their expression levels. Moreover, CNVs is shown to be enriched in genes also involved in immune responses, cell–cell signaling, and retrovirus- and transposition-related protein coding genes [13]. Thus, based on

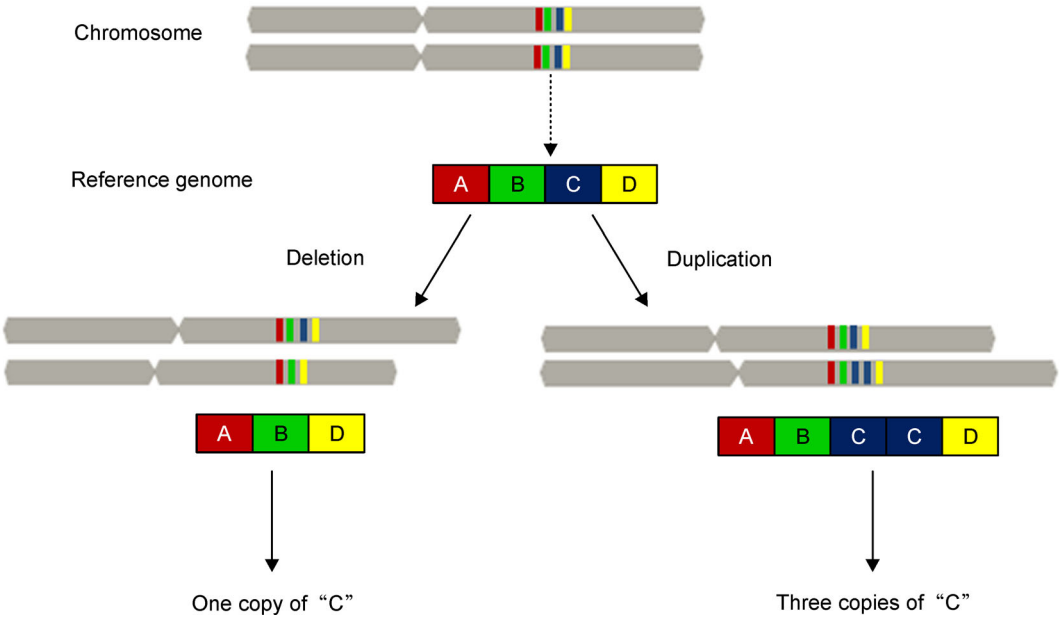

Figure 1. Copy number variations (CNVs). This figure illustrates the typical mechanism of CNV which relatively include duplication and deletion of "C" locus compared to the reference genome (include A, B, C, and D) of the depicted chromosome.

the large portions of human CNVs that have been reported in the Database of Genomic Variants (DGV) [14–16], and their mild effect on multiple gene functions, many CNVs have been associated with disease susceptibility and severity, while the majority continues to be benign. For example, a duplication within the CCL3L1 (C–C motif chemokine ligand 3 like 1) gene is involved in HIV susceptibility and developing AIDS [17], a deletion within the *IRGM* (immunity related GTPase M) gene is linked with Crohn's disease [18], a CNV located within the *TSPAN8* (tetraspanin 8) gene is associated with type 2 diabetes [19]. Similarly, a lower copy number of *FCGR3*B predisposes to immunologically related glomerulonephritis in humans and rats [20] and a higher *EGFR* copy number is linked to non-small cell lung cancer [21].

CNV is also associated with a range of neurodevelopmental disorders [22], including autism [23], schizophrenia [24], and depression [25]. Besides neuropsychiatric diseases, CNV have found to be linked with other disease types, including heart disease [26], obesity [27], cancer [28] and it has also been implicated in altered lifespan [29]. Furthermore, CNVs are also linked with extensive phenotypic traits in domestic animals including pigs [30], sheep [31], chicken [32,33], dogs [34], and cattle [35,36] amongst others.

CNV DETECTION METHODS

Although, CNV studies have developed considerably over time, little is known about how CNVs influence the phenotype of many rare and common complex diseases. To investigate this issue, various CNV detection methods have been developed. These methods can be categorized into two groups [37] (see Table 1 for details):

1. Genome-wide approaches, in which the entire genome is scanned for detecting CNVs.

(a) Microarray-based methods [38] such as array comparative genomic hybridization (aCGH) (Figure 2) and single nucleotide polymorphism (SNP) arrays [39].

(b) Karyotyping and fluorescence in situ hybridization (FISH) [40].

(c) Synthetic high-density oligonucleotide arrays [41].

(d) Deep sequencing platforms [42].

(e) NanoString's digital detection technology [43].

(f) Next-generation sequencing (NGS) [44] such as whole genome sequencing (WGS) and whole exome sequencing (WES).

However, genome-wide approach CNV analyses are not efficient for the validations of a small set of known CNVs. Targeted approaches are more efficient for that purpose.

2. Targeted approaches CNVs include:

(a) Quantitative polymerase chain reaction (qPCR) or southern hybridization for single target screening [45] (Figure 2).

(b) Multiplex ligation-dependent probe amplification (MLPA) [46].

(c) Multiplex amplifiable probe hybridization (MAPH) [47].

(d) Multiplex amplicon quantification [48].

In this regard, a more extensive commonly and high-throughput methods have been used, especially in the context of CNV related to human genomes. Starting from targeted CNV screening and validation, various methods have been applied including qPCR, paralogue ratio test (PRT), and molecular copy-number counting (MCC). qPCR compares the threshold cycles of a target versus reference sequence. PRT uses a single pair of primers to exploit sequence similarities between the elements of test and reference locus [49]. While MCC uses PCR to count the number of molecules in DNA aliquots [50]. Additionally, multiplex PCR-based approaches such as MAPH, MLPA, MAQ, quantitative multiplex PCR of

Table 1 Dataset: targeted and genome wide approaches

Approach	Techniques	
Genome-wide approach	Microarray-based	aCGH
		SNP-array
	Karyotyping and fluorescence in situ hybridization (FISH)	
	Synthetic high-density oligonucleotide arrays	
	Deep sequencing platforms	
	NanoString's digital detection technology	
	NGS-based analyses	Whole-genome sequencing (WGS)
		Whole-exome sequencing (WES)
Targeted approach	Quantitative polymerase chain reaction (qPCR)	
	Multiplex ligation-dependent probe amplification (MLPA)	
	Multiplex amplifiable probe hybridization (MAPH)	
	Multiplex amplicon quantification (MAQ)	

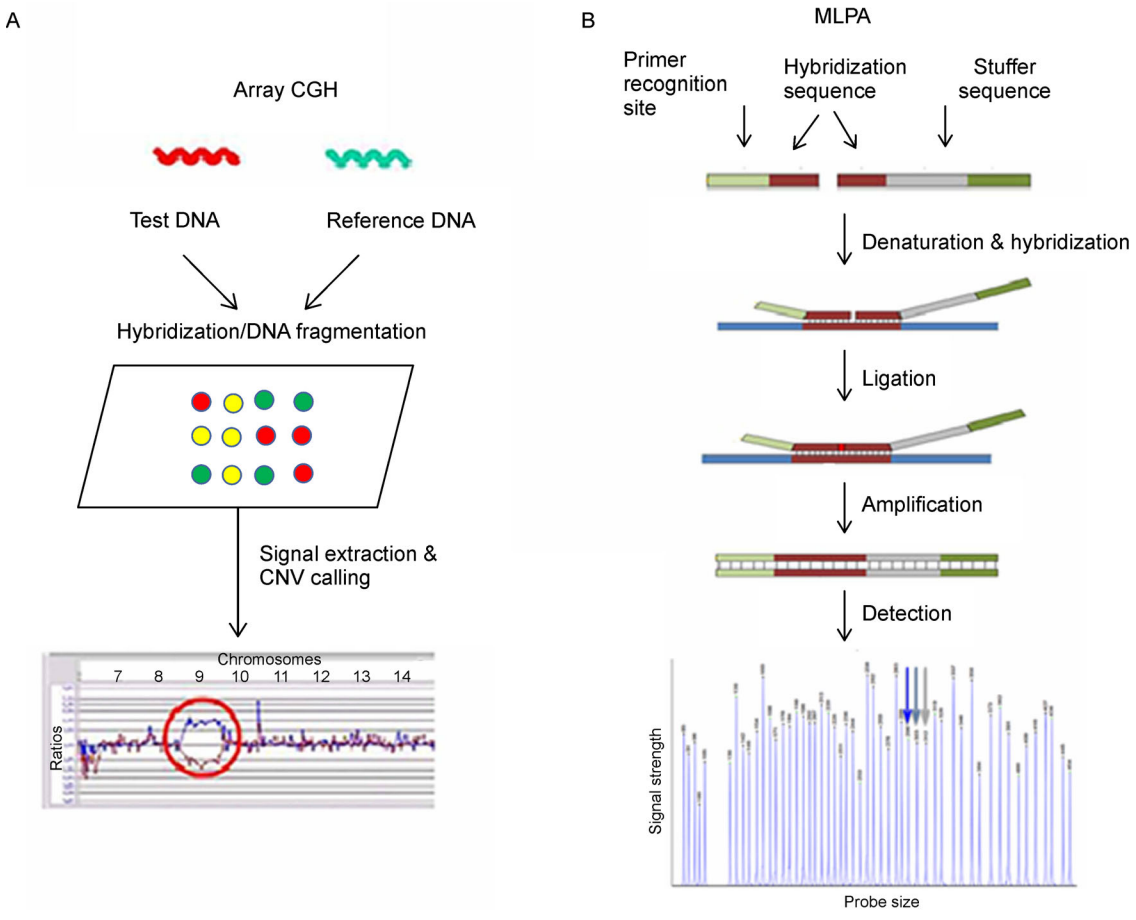

Figure 2. Copy number variation detection methods. (A) Example of genome-wide approach aCGH, test and reference DNA samples are labeled and hybridized onto the whole-genome microarray. Signal intensity ratios are calculated and the copy number differences between the test and reference genome are plotted. (B) Example of targeted approach multiplex ligation-dependent probe amplification (MLPA), each MLPA derived probe oligonucleotide has a different stuffer sequence. The two parts of each probe were hybridized to adjacent target sequences, then ligated and amplified by PCR primer pair. Because of the different lengths of the stuffer sequences, the amplification products of different MLPA probes can be separated, identified, and quantified by capillary electrophoresis.

short fluorescent fragments (QMPSF) and multiplex PCR-based real-time invader assay (mPCR-RETINA), have also been successfully used [51].

On the other hand and from high-throughput perspective, many high-resolution array platforms have been used extensively for CNV detection, which range from cytogenetic technologies such as karyotyping and fluorescence *in situ* hybridization (FISH) to more accurate arrays such as CGH and SNP arrays. CGH arrays considered to be a reliable method, since it can measure the fluorescence ratio along the length of each chromosome and identify novel regions of interest in the test sample. This method has the highest sensitivity and specificity [52], but gives relatively low resolution in CNV detection. Similarly, SNP arrays are more commonly used for CNV analysis, since they provide high resolution of CNVs based on hybridization intensities from custom and non-custom probes and require less sample DNA than CGH [53]. However, the main bias of SNP arrays on CNV detection is the low SNP coverage of the genomic regions.

In the same context of array analysis, a suite of algorithms has been used including but not limited to: CBS [54], GLAD [55], ITALICS [56], CRLMM [57], HMM, PennCNV [58], ParseCNV [59] and R.GADA [60]. Each of these methods has distinctive features and the most of them incorporated log R ratio (LRR) and B-allele frequency (BAF) for reliable CNV identification.

To overcome the issues driven by array-based techniques, studies turn to adopt the new approach of NGS, which has rapidly emerged as a viable option to identify CNVs in human diseases. This approach confers a number of critical advantages including higher coverage

and resolution, more precise detection of breakpoints, and higher capability to identify smaller CNVs [61,62].

In general, three main approaches have been used in NGS technologies: (i) read count, (ii) paired-end and (iii) assembly [63] as shown in Figure 3 plus two additional strategies including split read (SR) and combinatorial of these four methods. In the read depth (RD) approach a sliding window is used to count the number of short reads, and then these read count values are used to identify CNV regions. RD-based methods can detect the exact number of copy numbers, large insertions, CNVs in complex genomic region classes and can be applied to both WGS and WES data. However, they cannot detect precise breakpoints, inversions and translocations events. Paired-end (PE) approach or paired-end mapping (PEM) identifies genomics aberration based on the distances between a pair of paired-end reads and not single-end reads. Also, PEM is able to identify efficiently inversions and translocations but unable to detect CNVs in low complexity regions. In the assembly (AS) approach overlapping short reads (contigs) are used to assemble the genomics regions, and CNV regions are detected by comparing these assembled contigs to the reference genome. On the other hand, SR methods rely on the only unique mapping information. Since they can only split the incompletely mapped reads of read pairs into multiple fragments, and then the start and end fragments of each split read will be aligned to the reference genome to assign insertion or deletion events. For every approach, a diverse set of popular methods and tools have been developed such as CNV-seq [64], FREEC/Control-FREEC [65], CNVnator [66], SegSeq [67], eventwise testing (EWT) [68], BreakDancer [69], ExomeCNV [70], XHMM [71], ExoCNVTest [72], GPHMM [73], CLImAT [74], Cortex assembler [75] and Magnolya [76]. Although there has been great progress in each category, none of the methods and tools could comprehensively detect all types of CNVs. Thereafter, a combinatorial approach has been used attempting to increase the performance in detecting CNVs more reliably.

Furthermore, several well-established methods have also been used to find recurrent copy number aberration (RCNA) or somatic copy number alteration (SCNA) from a cohort of tumor patients. Among these methods GISTIC (genomic identification of significant targets in cancer) [77], GISTIC 2.0 [78], JIS-TIC [79], NN-SSVD (non-negative sparse singular value decomposition) [80], DiNAMIC (discovering copy number aberrations manifested in cancer) [81] and PLA (piecewise-constant and low-rank approximation) [82], have extensively applied. All of these methods were focused on the identification of driver aberrations that was proved to be crucial for many diseases progression, unlike passenger events that have no functional effect. GISTIC can identify significant

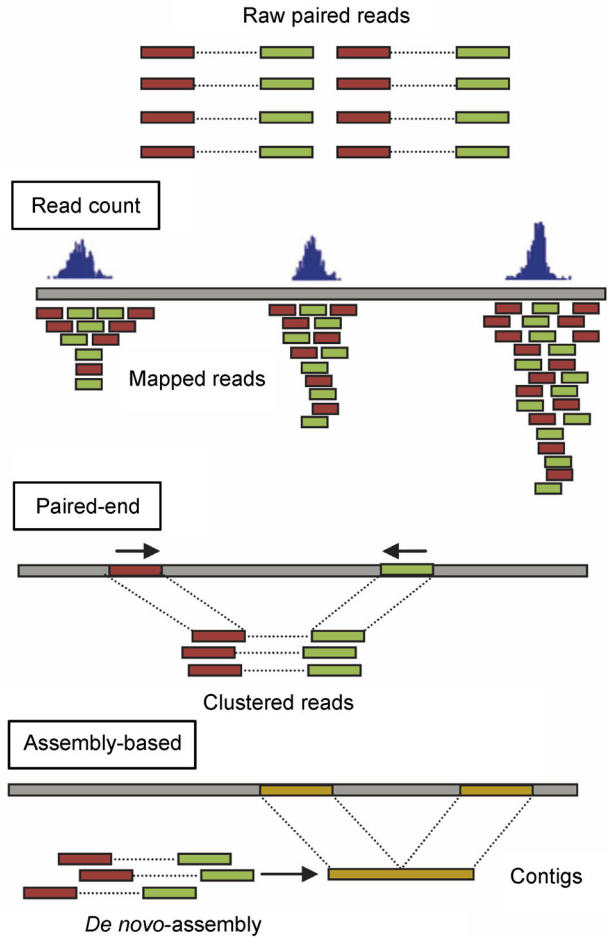

Figure 3. CNV detected approaches from NGS. This figure illustrate the NGS typical methods used to detect CNV.

driver SCNA by evaluating the frequency and amplitude of observed events based on G-score (the product of frequency and average amplitude) and a greedy peeling-off. However, GISTIC 2.0, a revised version of GISTIC discovers recurrent CNVs based on G-score (the negative logarithm of the likelihood of both frequency and amplitude of each aberration region) and an arbitrated peeling-off. Similarly, JISTIC which is an improved tool of GISTIC algorithm can detect more significant sub-regions within large aberrant regions. In addition, The RCNA regions of DiNAMIC are detected by using peeling method tailored to the inner cyclic shift procedure from various input-data types (continuous, continuous segmented or discrete segmented). While PLA detects RCNAs by the sample frequency of the low-rank component from multi-sample data, NN-SSVD have the ability to discover RCNAs in complex patterns based on low-rank approximation component of only one layer.

DISEASES RELATED COPY NUMBER VARIATION

In this review, we summarize a serious of CNV (common/rare) related diseases and their associated disease genes. The details of those information are illustrated in the below Tables 2−9 and the following subsections.

Immune response and inflammation

Various studies have confirmed the significant impact of CNVs on immune response and inflammation. As initial estimates of the Online Mendelian Inheritance in Man (OMIM) and the Gene Ontology (GO) analysis, large portions of genes and exons as well were linked to CNV and code proteins involved in the immune response and inflammation. For example low copy numbers for the gene *CCL3L1* were associated with an accelerated rate of HIV progression and developing AIDS (Table 2) [83].

Table 2 Immune response and inflammation

Disease/phenotype	Implicated genes /loci	Technique
Immune response and inflammation	Low copy number of *CCL3L1*	GO analysis

Syndromes, schizophrenia, mental retardation and autism spectrum disorder

CNVs have been associated with diseases, through (i) dosage of a single gene [84,85], (ii) a contiguous set of genes (*e.g.*, Williams-Beuren syndrome [86,87], DiGeorge syndrome [88], Smith-Magenis syndrome [89], Potocki-Lupski syndrome [90]) or (iii) allele combinations in the case of complex diseases.

Recent studies in (i) syndromes, (ii) schizophrenia [91,92], (iii) mental retardation [93,94] and (iv) autism spectrum disorder [95] not only detected multiple disease related genes but also led to the description of variable phenotypes, novel microdeletion and microduplication syndromes [96–99].

Firstly, in the context of syndromes, various popular classic examples have been identified including (i) the 15q11-q13 deletion associated with Prader-Willi and Angelman syndromes [100], (ii) the 17p11 deletion associated with Smith-Magenis syndrome [101], (iii) the 7q11 deletion associated with Williams-Beuren syndrome [102], and (iv) the 22q11 deletions associated with velocardiofacial syndrome (Table 3)[103].

Secondly, alterations in the following three regions were associated with both schizophrenia and mental retardation: (i) 1q21.1, (ii) 15q11.2 and (iii) 15q13.3. While deletions at all three loci were linked to schizophrenia, related psychoses [94,95], as well as 22q11.2 deletion syndrome (22qDS), which is identifiable genetic alteration that has been associated only with schizophrenia [104].

Thirdly, duplication of the entire 15q11-q13.3 and both deletions and duplications of band 1q21.1 were associated with mental retardation.

Finally, duplication of the entire 15q11-q13.3 region was shown to cause autism spectrum disorder (ASD) [95]. Similarly, duplication of band 1q21.1 was also identified in patients with ASD.

These examples suggest that a simple alteration at any given chromosome position in the human genomes can cause some disorders and phenotypes.

Cancers

It has been specifically reported that an accurate CNVs

Table 3 Syndromes, schizophrenia, mental retardation and autism spectrum disorder

Disease/phenotype		Implicated genes/loci	Technique
Syndromes	Prader-Willi and Angelman syndromes	Deletion at 15q11-q13	High-density arrays
	Smith-Magenis syndrome	Deletion at 17p11.2	High-density arrays
	Potocki-Lupski syndrome	Duplication at 17p11.2	High-density arrays
	Williams-Beuren syndrome	Deletion at 7q11.23	High-density arrays
	7q11 duplication syndrome	7q11.23 duplication	High-density arrays
	Velocardiofacial syndrome	Deletion at 22q11	High-density arrays
Schizophrenia	Microduplication 22q11.2	Deletion at 1q21.1, 15q11.2, 15q13.3	High-density arrays
		22q11.2 duplication	High-density arrays
Mental retardation		Duplication at 15q11-q13.3	High-density arrays
		Duplication and deletion of 1q21.2	High-density arrays
ASD		Duplication at 15q11-q13.3	High-density arrays
		Duplication at 1q21.1	High-density arrays

detection is an essential part of cancer genome analysis, which holds great promise to improve cancer prognosis and treatment decision. Therefore, significant effort has found associations between somatic CNVs and cancers, based on their oncogene activation and tumor suppressor gene inactivation caused by copy number amplification and heterozygous/homozygous deletion respectively.

Generally, there are three kinds of CNV variations: (i) Germline CNVs, (ii) somatic CNVs and (iii) inherited CNVs. Among these variations, somatic CNVs have been successfully associated with cancer. For example, Walters *et al*. [105] predicted an amplified copy number of CHD7-PVT1 likely to have a relative effect in tumor genesis of small cell lung cancer. Fanciulli *et al*. have also found associations between somatic CNVs and another kinds of cancers such as prostate and colorectal cancers [83].

In the context of prostate cancer, a landscape of CNVs with the clinical/pathological endpoints of metastasis were observed including that of Barbieri *et al*. [106] and the Cancer Genome Atlas (TCGA) prostate cancer cohort. The related CNVs include: (i) genomic deletion on chromosomes at 6p, 8p, 13q and 16p, (ii) genomic duplication at 7q and 8q, and (iii) focal alterations spanning PTEN, RB1, and tumor protein p53 (TP53) among others.

Additionally, in the same context of the cancer genome, another disease called oral cavity squamous cell carcinoma (OSCC), including (i) cigarette smoking, (ii) alcohol consumption, and (iii) betel quid chewing, have caused many genomic aberrations and widespread genomic instability, especially in eastern and west countries [65,107,108]. Therefore, various studies on OSCC have detected multiple mutations related to genes like: *TP53*, *NOTCH1*, *CASP8*, *FAT1*, *CDKN2A*, *HRAS*, *USP9X* [109,110] and multiple CNVs events (Table 4) [109,111,112], such as deletions at 3p, 8p, 9p, 18q and duplication at 3q, 5p, 7p, 8q, 11q, and 20q.

Furthermore, various methods have also been applied to detect alterations in OSCC heterogeneous patient's samples. Among these methods, an approach called Ultra-deep targeted sequencing (UDT-Seq) successfully identified new pathogenic CNVs, like: (i) PIK3CA duplica-tion [113], (ii) FGFR1 duplication [114], and (iii) deletions of *PTEN*, *RB1*, *SMAD4*, and *TP53* [114–116].

A suite of studies related to CNVs and their roles in lung cancer have also been discussed. For example, alterations in chromosome regions at 3q26.2-q29, 3p26.3-p11.1, 17p13.3-p11.2 and 9p13.3-p13.2 have been deemed as the main predictors for lung cancer. Moreover, an integrative analysis of transcriptional profile and CNV of lung cancer have captured more significant CNV driven genes.

Cardiovascular disease

The association of CNVs with cardiovascular diseases has also been early demonstrated. In which, many aberrant

Table 4 Cancers

Disease/phenotype	Implicated genes/loci	Technique
Small cell lung cancer (SCLC)	*CHD7-PVT1*	WGS and WES
Prostate cancer	Deletion at 6p, 8p, 13q and 16p	WGS and WES
	Duplication at 7 and 8q	WGS and WES
	Focal alteration of phosphatase *PTEN*, *RB1*, *TP53*	WGS and WES
OSCC	Deletions at 3p, 8p, 9q, 18q	UDT-Seq
	Duplication at 3q, 5p, 7p, 8q, 11q, 20q	UDT-Seq
	Duplication of *PIK3CA*	UDT-Seq
	Duplication of *FGFR1*	UDT-Seq
	Deletions of *PTEN*, *RB1*, *SMAD4*, and *TP53*	UDT-Seq
Lung cancer ASD	Alteration at 3q26.2-q29.2, 3p26.3-p11.1, 17p13.3-p11.2, 9p13.3-p13.2	UDT-Seq

Table 5 Cardiovascular diseases

Disease/phenotype	Implicated genes/loci	Technique
Ventricular tachycardia	Deletion of Calsequestrin gene at 21p13.2-1p13.1	aCGH
	Duplication of Calsequestrin gene at 21p13.2-1p13.1	aCGH
Hypertrophic cardiomyopathy	Deletion of Myosin, light polypeptide3, alkali at 3p21.31	aCGH
	Deletion of histo-compatibility complex, class2, DR1 at 6p21.32	aCGH
Immune disease and cardiomyopathy	Duplication of histo-compatibility complex, class2, DR1 at 6p21.32	aCGH

CNVs loci and related disease genes have been discovered to strong genetic components often single-gene or "monogenic" disorder (Table 5) [117] such as:

(i) Ventricular tachycardia disease: deletion and duplication of calsequestrin gene at 21p13.2-1p13.1.

(ii) Hypertrophic cardiomyopathy disease: deletion of myosin, light polypeptide 3, alkali at the band 3p21.31.

(iii) Immune disease and cardiomyopathy: deletion and duplication of major histo-compatibility complex, class II, *DR1* in 6p21.32.

Neuropsychiatric disease

A great number of studies have also demonstrated the role of CNVs in the etiology of several neuropsychiatric disorders [73,117]. Argyrophilic grain disease (AGD) is an example of this genetic disorder. Thus, to identify CNVs related to AGD. They first used aCGH (180k platform) (Table 6) alone and then adopted the same 180K aCGH platform with an extra 400 independent samples and revealed no rare CNVs was significant. However, they highlighted a 40 kb microdeletion at 17p13.2 that includes the *CTNS* gene which causes cystinosis disorder, and a 65 kb deletion that includes the *SHPK* gene [119].

Autoimmune disorders

Another genetic disorder called autoimmune disorders have been associated with multiple CNVs rather than single CNV, which is useful for understanding the pathogenesis and discovering new drug targets [120–123]. Several studies have reported this association by discovering several genes such as (Table 7):

(i) Systemic lupus erythematosus (SLE) [122]: Fcy receptors located at 6p21, complement component 4 (C4) at 1q23 positions, *RABGAP1L*, and deletion at 10q21.3 deletion.

(ii) Psoriasis and Crohn's disease (CD): *ITP* and b-defensin genes.

(iii) Rheumatoid arthritis (RA): *VPREB1* at 22q11 region.

(iv) Ankylosing spondylitis (AS) [76]: deletion of *HHAT* (1q32.2), *HLA-DPB1* (6p21.3), *PRKRA* (2q31.2), *EEF1DP3* (13q13.1) and 16p13.3.

Psoriasis

In contrast to most complex diseases, the role of common CNV in the pathogenesis of psoriasis has been well addressed across multiple studies [123–126]. Particularly, a promising association between psoriasis and a 32.2 kb deletion of *LCE3B* and *LCE3C* genes have been identified extensively in (i) in European populations [123], and (ii) subsequently replicated in a Chinese cohort [126], and then confirmed extremely by ExoCNV Test exome sequencing method (Table 8).

Huntington's disease

While it is known that gene deletion and duplication can affect neurological disease, a new study was also able to investigate an association between CNV and variable adult age of onset (AAO) of Huntington's disease (HD). As a result, CNV of *SLC2A3* has been finally observed between 1 copy (heterozygous deletion) and 3 copies (heterozygous duplication) in HD [127], whereas many genes and loci were related extensively to neurodegenerative disorders such as the triplication of alpha-synuclein or large deletions of the parkin gene causing Parkinson's disease [128].

Table 6 Neuropsychiatric disease

Disease/phenotype	Implicated genes/loci	Technique
AGD	Microdeletion of CTNS at 17p13	aCGH
Cystinosis	Deletion of SHPK	aCGH

Table 7 Autoimmune disorders

Disease/phenotype	Implicated genes/loci	Technique
Systemic lupus erythematosus (SLE)	FCy receptor at 6p21,	GWA
	C4 at 1q23, *RABGAP1L*,	GWA
	Deletion at 10q21	GWA
Psoriasis and Crohns disease (CD)	*ITP*, b-defensin	GWA
Rheumatoid arthritis (RA)	*VPREB1* at 22q.11	GWA
Ankylosing spondylitis (AS)	Deletion of *HHAT* (1q32.2),	GWA
	HLA-DPB1 (6p21.3), *PRKRA* (2q31.2),	GWA
	EEF1DP3 (13q13.1) and 16p13.3	GWA

Table 8 Psoriasis

Disease/phenotype	Implicated genes/loci	Technique
Psoriasis	32.2 kb deletion of *LCE3B-LCE3C*	WES, ExoCNV

Table 9 Huntington's disease (HD)

Disease/phenotype	Implicated genes/loci
Huntington's disease (HD)	SLC2A3

CONCLUSIONS

Nowadays, with advances in high-throughput genomic technologies, various genomic datasets have been significantly reported from various biological levels. In this review, we have focused on CNVs which belong to the genomic level. An overview of the definition of CNVs, CNVs related diseases/phenotypes and detections methods has been firstly introduced. Then, based on a number of complex diseases, the critical role of CNVs (whether rare or common CNVs) have been summarized for the identification of disease related genes.

To date, a large number of genetic diseases and phenotypes have been associated with CNVs. CNVs play important roles in these genetic diseases through: (i) identifying multiple genes whether existed genes or newly discovered ones, (ii) allowing discrimination of driver mutation for pathogenesis or diagnosis of complex diseases and (iii) helping to develop personalized medicines.

Various methods have been discovered to validate and detect the reliability and the accuracy of CNVs. These methods are: (i) cytogenetics and karyotyping methods, (ii) microarrays based methods, (iii) next generation sequencing methods and (iv) third generation approaches as well. Each of these approaches has advantages and disadvantages, which are: coverage biases, batch effects, poor sensitivity and precision as well as higher effective resolution and less complex data analysis. However, accurate detection of small CNVs specifically and their precise boundaries from massively amount of data using these methods is still a challenge, which largely due to the complexities of tumor samples. Thus, the validity and reliability detection of CNVs will improve quickly as genotyping technologies advance, which will support the required replication.

In addition, the role of CNVs in genetic syndromes has long been recognized, with recurrent microdeletion/microduplications detected in syndromes, such as Prader–Willi, Smith–Magenis and Williams–Beuren. However, with the increased clinical use of array-based CNV analysis, the list of CNVs associated with disease phenotypes has continued to grow. This has led to the discovery of many new microdeletion and microduplication syndromes. These novel syndromes and the ever-expanding of CNVs associated with disease phenotypes, highlight the significant involvement of CNVs in genetic diseases.

An important issue that has also to be reported in the context of CNVs was called missing heritability [129]. This issue has been studied in order to estimate the heritability of common diseases. Missing heritability in genome wide association studies, which was identified as the failure to account for a considerable fraction of heritability by the variants detected is also still a challenging issue in human genetics. For solving this puzzle, a number of CNV based methods have been proposed. However, none of them have accurately accounted for missing heritability due to the conflict raised from rare and common genetic variants.

In conclusion, CNV alone will not meet a great advancement enough and will not be a worthy endeavor enough without its integration with other genomic datasets. Multiple data integrations have proved to be successful, which include those datasets such as gene expression, DNA methylation, protein-protein interaction (PPI), metabolism pathways and Gene Ontology. These integrations will help us understand disease susceptibility and pathogenesis from various perspectives.

AUTHOR'S CONTRIBUTIONS

Bolin Chen initiated this review work; Chaima Aouiche and Bolin Chen discussed the whole outline and designed the review topics; Chaima Aouiche wrote the paper; Chaima Aouiche, Xuequn Shang and Bolin Chen revised the manuscript for many times throughly. All authors have read and approved the final manuscript.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Nos. 61602386 and 61332014), the Natural Science Foundation of Shaanxi Province (No. 2017JQ6008), and the top university visiting foundation for excellent youth scholars of Northwestern Polytechnical University.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Chaima Aouiche, Xuequn Shang and Bolin Chen declare that they have no conflict of interests.

This article is a review article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Schadt, E. E. (2009) Molecular networks as sensors and drivers of common human diseases. Nature, 461, 218–223
2. Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M. and Barabási, A. L. (2007) The human disease network. Proc. Natl. Acad. Sci. USA, 104, 8685–8690
3. Davies, R. J., Miller, R. and Coleman, N. (2005) Colorectal cancer screening: prospects for molecular stool analysis. Nat. Rev. Cancer, 5, 199–209
4. Beckmann, J. S., Estivill, X. and Antonarakis, S. E. (2007) Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat. Rev. Genet., 8, 639–646
5. Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J., Urashima, M., *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. Nature, 463, 899–905
6. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. and Kim, D. (2015) Methods of integrating data to uncover genotype-phenotype interactions. Nat. Rev. Genet., 16, 85–97
7. Ionita-Laza, I., Rogers, A. J., Lange, C., Raby, B. A. and Lee, C. (2009) Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. Genomics, 93, 22–26
8. 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.* (2006) Global variation in copy number in the human genome. Nature, 444, 444–454
9. Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., Hurles, M. E., *et al.* (2006) Copy number variation: new insights in genome diversity. Genome Res., 16, 949–961
10. Stankiewicz, P. and Lupski, J. R. (2010) Structural variation in the human genome and its role in disease. Annu. Rev. Med., 61, 437–455
11. Feuk, L., Carson, A. R. and Scherer, S. W. (2006) Structural variation in the human genome. Nat. Rev. Genet., 7, 85–97
12. Eichler, E. E., Nickerson, D. A., Altshuler, D., Bowcock, A. M., Brooks, L. D., Carter, N. P., Church, D. M., Felsenfeld, A., Guyer, M., Lee, C., *et al.* (2007) Completing the map of human genetic variation. Nature, 447, 161–165
13. Li, W. and Olivier, M. (2013) Current analysis platforms and methods for detecting copy number variation. Physiol. Genomics, 45, 1–16
14. Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W. and Lee, C. (2004) Detection of large-scale variation in the human genome. Nat. Genet., 36, 949–951
15. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., *et al.* (2004) Large-scale copy number polymorphism in the human genome. Science, 305, 525–528
16. Zhang, J., Feuk, L., Duggan, G. E., Khaja, R. and Scherer, S. W. (2006) Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet. Genome Res., 115, 205–214
17. Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R., Freedman, B., Marlon P. , Quinones, M., Bamshad, M., *et al.* (2005) The influence of *CCL3L1* gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science, 307, 1434–1440
18. McCarroll, S. A., Huett, A., Kuballa, P., Chilewski, S. D., Landry, A., Goyette, P., Zody, M. C., Hall, J. L., Brant, S. R., Cho, J. H., *et al.* (2008) Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. Nat. Genet., 40, 1107–1112
19. Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D. F., Giannoulatou, E., *et al.* (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature, 464, 713–720
20. Aitman, T. J., Dong, R., Vyse, T. J., Norsworthy, P. J., Johnson, M. D., Smith, J., Mangion, J., Roberton-Lowe, C., Marshall, A. J., Petretto, E., *et al.* (2006) Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans. Nature, 439, 851–855
21. Cappuzzo, F., Hirsch, F. R., Rossi, E., Bartolini, S., Ceresoli, G. L., Bemis, L., Haney, J., Witta, S., Danenberg, K., Domenichini, I., *et al.* (2005) Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non-small-cell lung cancer. J. Natl. Cancer Inst., 97, 643–655
22. Glessner, J. T., Connolly, J. J. and Hakonarson, H. (2012) Rare genomic deletions and duplications and their role in neurodevelopmental disorders. Curr. Top. Behav. Neurosci., 12, 345–360
23. Glessner, J. T., Wang, K., Cai, G., Korvatska, O., Kim, C. E., Wood, S., Zhang, H., Estes, A., Brune, C. W., Bradfield, J. P., *et al.* (2009) Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature, 459, 569–573
24. Glessner, J. T., Reilly, M. P., Kim, C. E., Takahashi, N., Albano, A., Hou, C., Bradfield, J. P., Zhang, H., Sleiman, P. M., Flory, J. H., *et al.* (2010) Strong synaptic transmission impact by copy number variations in schizophrenia. Proc. Natl. Acad. Sci. USA, 107, 10584–10589
25. Glessner, J. T., Wang, K., Sleiman, P. M., Zhang, H., Kim, C. E., Flory, J. H., Bradfield, J. P., Imielinski, M., Frackelton, E. C., Qiu, H., *et al.* (2010) Duplication of the *SLIT3* locus on 5q35.1 predisposes to major depressive disorder. PLoS One, 5, e15463
26. Goldmuntz, E., Paluru, P., Glessner, J., Hakonarson, H., Biegel, J. A., White, P. S., Gai, X. and Shaikh, T. H. (2011) Microdeletions and microduplications in patients with congenital heart disease and multiple congenital anomalies. Congenit. Heart Dis., 6, 592–602
27. Glessner, J. T., Bradfield, J. P., Wang, K., Takahashi, N., Zhang, H., Sleiman, P. M., Mentch, F. D., Kim, C. E., Hou, C., Thomas, K. A., *et al.* (2010) A genome-wide study reveals copy number variants exclusive to childhood obesity cases. Am. J. Hum. Genet., 87, 661–666
28. Kuusisto, K. M., Akinrinade, O., Vihinen, M., Kankuri-

Tammilehto, M., Laasanen, S. L. and Schleutker, J. (2013) copy number variation analysis in familial *BRCA1/2*-negative Finnish breast and ovarian cancer PLoS One, 8, e71802

29. Glessner, J. T., Smith, A. V., Panossian, S., Kim, C. E., Takahashi, N., Thomas, K. A., Wang, F., Seidler, K., Harris, T. B., Launer, L. J., *et al.* (2013) Copy number variations in alternative splicing gene networks impact lifespan. PLoS One, 8, e53846
30. Johansson Moller, M., Chaudhary, R., Hellmén, E., Höyheim, B., Chowdhary, B. and Andersson, L. (1996) Pigs with the dominant white coat color phenotype carry a duplication of the *KIT* gene encoding the mast/stem cell growth factor receptor. Mamm. Genome, 7, 822–830
31. Norris, B. J. and Whan, V. A. (2008) A gene duplication affecting expression of the ovine *ASIP* gene is responsible for white and black sheep. Genome Res., 18, 1282–1293
32. Wright, D., Boije, H., Meadows, J. R. S., Bed'hom, B., Gourichon, D., Vieaud, A., Tixier-Boichard, M., Rubin, C. J., Imsland, F., Hallböök, F., *et al.* (2009) Copy number variation in intron 1 of *SOX5* causes the Pea-comb phenotype in chickens. PLoS Genet., 5, e1000512
33. Dorshorst, B., Harun-Or-Rashid, M., Bagherpoor, A. J., Rubin, C. J., Ashwell, C., Gourichon, D., Tixier-Boichard, M., Hallböök, F. and Andersson, L. (2015) A genomic duplication is associated with ectopic eomesodermin expression in the embryonic chicken comb and two duplex-comb phenotypes. PLoS Genet., 11, e1004947
34. Salmon Hillbertz, N. H. C., Isaksson, M., Karlsson, E. K., Hellmén, E., Pielberg, G. R., Savolainen, P., Wade, C. M., von Euler, H., Gustafson, U., Hedhammar, A., *et al.* (2007) Duplication of *FGF3*, *FGF4*, *FGF19* and *ORAOV1* causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. Nat. Genet., 39, 1318–1320
35. Drögemüller, C., Distl, O. and Leeb, T. (2001) Partial deletion of the bovine *ED1* gene causes anhidrotic ectodermal dysplasia in cattle. Genome Res., 11, 1699–1705
36. Capitan, A., Allais-Bonnet, A., Pinton, A., Marquant-Le Guienne, B., Le Bourhis, D., Grohs, C., Bouet, S., Clément, L., Salas-Cortes, L., Venot, E., *et al.* (2012) A 3.7 Mb deletion encompassing ZEB2 causes a novel polled and multisystemic syndrome in the progeny of a somatic mosaic bull. PLoS One, 7, e49084
37. Aten, E., White, S. J., Kalf, M. E., Vossen, R. H., Thygesen, H. H., Ruivenkamp, C. A., Kriek, M., Breuning, M. H. and den Dunnen, J. T. (2008) Methods to detect CNVs in the human genome. Cytogenet. Genome Res., 123, 313–321
38. Kim, T. M., Yim, S. H. and Chung, Y. J. (2008) Copy number variations in the human genome: potential source for individual diversity and disease association studies. Genomics Inform., 6, 1–7
39. Carter, N. P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. Nat. Genet., 39, S16–S21
40. Buysse, K., Delle Chiaie, B., Van Coster, R., Loeys, B., De Paepe, A., Mortier, G., Speleman, F., Menten, B. (2009) Challenges for CNV interpretation in clinical molecular karyotyping: lessons learned from a 1,001 sample experience. Eur. J. Med. Gene., 52, 398–403
41. Lucito, R., Healy, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., Rodgers, L., Brady, A., Sebat, J., Troge, J., *et al.* (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Res., 13, 2291–2305
42. Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J., Zhao, X., Carter, S. L., Russ, C., Nusbaum, C., Meyerson, M. and Lander, E. S. (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat. Methods, 6, 99–103
43. Geiss, G. K., Bumgarner, R. E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D. L., Fell, H. P., Ferree, S., George, R. D., Grogan, T., *et al.* (2008) Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat. Biotechnol., 26, 317–325
44. Abel, H. J. and Duncavage, E. J. (2013) Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. Cancer Genet., 206, 432–440
45. Weksberg, R., Hughes, S., Moldovan, L., Bassett, A. S., Chow, E. W. and Squire, J. A. (2005) A method for accurate detection of genomic microdeletions using real-time quantitative PCR. BMC Genomics, 6, 180
46. Schouten, J. P., McElgunn, C. J., Waaijer, R., Zwijnenburg, D., Diepvens, F. and Pals, G. (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res., 30, e57
47. Armour, J. A., Sismani, C., Patsalis, P. C. and Cross, G. (2000) Measurement of locus copy number by hybridisation with amplifiable probes. Nucleic Acids Res., 28, 605–609
48. Kumps, C., Van Roy, N., Heyrman, L., Goossens, D., Del-Favero, J., Noguera, R., Vandesompele, J., Speleman, F. and De Preter, K. (2010) Multiplex amplicon quantification (MAQ), a fast and efficient method for the simultaneous detection of copy number alterations in neuroblastoma. BMC Genomics, 11, 298
49. Fernandez-Jimenez, N., Castellanos-Rubio, A., Plaza-Izurieta, L., Gutierrez, G., Irastorza, I., Castaño, L., Vitoria, J. C. and Bilbao, J. R. (2011) Accuracy in copy number calling by qPCR and PRT: a matter of DNA. PLoS One, 6, e28910
50. Daser, A., Thangavelu, M., Pannell, R., Forster, A., Sparrow, L., Chung, G., Dear, P. H. and Rabbitts, T. H. (2006) Interrogation of genomes by molecular copy-number counting (MCC). Nat. Methods, 3, 447–453
51. Ceulemans, S., van der Ven, K. and Del-Favero, J. (2012) Targeted screening and validation of copy number variations. Methods Mol. Biol., 838, 311–328
52. Haraksingh, R. R., Abyzov, A., Gerstein, M., Urban, A. E. and Snyder, M. (2011) Genome-wide mapping of copy number variation in humans: comparative analysis of high resolution array platforms. PLoS One, 6, e27859
53. Oldridge, D. A., Banerjee, S., Setlur, S. R., Sboner, A. and Demichelis, F. (2010) Optimizing copy number variation analysis using genome-wide short sequence oligonucleotide arrays. Nucleic Acids Res., 38, 3275–3286

54. Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics, 5, 557–572
55. Hupé, P., Stransky, N., Thiery, J. P., Radvanyi, F. and Barillot, E. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics, 20, 3413–3422
56. Rigaill, G., Hupé, P., Almeida, A., La Rosa, P., Meyniel, J. P., Decraene, C. and Barillot, E. (2008) ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays. Bioinformatics, 24, 768–774
57. Scharpf, R. B., Ruczinski, I., Carvalho, B., Doan, B., Chakravarti, A. and Irizarry, R. A. (2011) A multilevel model to address batch effects in copy number estimation using SNP arrays. Biostatistics, 12, 33–50
58. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., Hakonarson, H. and Bucan, M. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res., 17, 1665–1674
59. Glessner, J. T., Li, J. and Hakonarson, H. (2013) ParseCNV integrative copy number variation association software with quality tracking. Nucleic Acids Res., 41, e64
60. Pique-Regi, R., Cáceres, A. and González, J. R. (2010) R-Gada: a fast and flexible pipeline for copy number analysis in association studies. BMC Bioinformatics, 11, 380
61. Alkan, C., Coe, B. P. and Eichler, E. E. (2011) Genome structural variation discovery and genotyping. Nat. Rev. Genet., 12, 363–376
62. Meyerson, M., Gabriel, S. and Getz, G. (2010) Advances in understanding cancer genomes through second-generation sequencing. Nat. Rev. Genet., 11, 685–696
63. Zhao, M., Wang, Q., Wang, Q., Jia, P. and Zhao, Z. (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: fea-tures and perspectives. BMC Bioinformatics, 14 (Suppl 11), S1
64. 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
65. Boeva, V., Zinovyev, A., Bleakley, K., Vert, J. P., Janoueix-Lerosey, I., Delattre, O. and Barillot, E. (2011) Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics, 27, 268–269
66. Abyzov, A., Urban, A. E., Snyder, M. and Gerstein, M. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res., 21, 974–984
67. Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J. T., Zhao, X., Carter, S. L., Russ, C., Nusbaum, C., Meyerson, M. and Lander, E. S. (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat. Methods, 6, 99–103
68. Yoon, S., Xuan, Z., Makarov, V., Ye, K. and Sebat, J. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res., 19, 1586–1592
69. 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.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods, 6, 677–681
70. Sathirapongsasuti, J. F., Lee, H., Horst, B. A., Brunner, G., Cochran, A. J., Binder, S., Quackenbush, J. and Nelson, S. F. (2011) Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics, 27, 2648–2654
71. Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., Handsaker, R. E., McCarroll, S. A., O'Donovan, M. C., Owen, M. J., *et al.* (2012) Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am. J. Hum. Genet., 91, 597–607
72. Coin, L. J., Cao, D., Ren, J., Zuo, X., Sun, L., Yang, S., Zhang, X., Cui, Y., Li, Y., Jin, X., *et al.* (2012) An exome sequencing pipeline for identifying and genotyping common CNVs associated with disease with application to psoriasis. Bioinformatics, 28, i370–i374
73. Li, A., Liu, Z., Lezon-Geyda, K., Sarkar, S., Lannin, D., Schulz, V., Krop, I., Winer, E., Harris, L. and Tuck, D. (2011) GPHMM: an integrated hidden Markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome SNP arrays. Nucleic Acids Res., 39, 4928–4941
74. Yu, Z., Liu, Y., Shen, Y., Wang, M. and Li, A. (2014) CLImAT: accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data. Bioinformatics, 30, 2576–2583
75. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. and McVean, G. (2012) *De novo* assembly and genotyping of variants using colored de Bruijn graphs. Nat. Genet., 44, 226–232
76. Nijkamp, J. F., van den Broek, M. A., Geertman, J. M., Reinders, M. J., Daran, J. M. and de Ridder, D. (2012) *De novo* detection of copy number variation by co-assembly. Bioinformatics, 28, 3195–3202
77. Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J. C., Huang, J. H., Alexander, S., *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc. Natl. Acad. Sci. USA, 104, 20007–20012
78. Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R. and Getz, G. (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol., 12, R41
79. Sanchez-Garcia, F., Akavia, U. D., Mozes, E. and Pe'er, D. (2010) JISTIC: identification of significant targets in cancer. BMC Bioinformatics, 11, 189
80. Walter, V., Nobel, A.B., and Wright, F.A. (2011) DiNAMIC: A method to identify recurrent DNA copy number aberrations in tumors. Bioinformatics, 27, 678–685
81. Zhou, X., Liu, J., Wan, X. and Yu, W. (2014) Piecewise-constant and low-rank approximation for identification of recurrent copy number variations. Bioinformatics, 30, 1943–1949

82. Xi, J. and Li, A. (2016) Discovering recurrent copy number aberrations in complex patterns via non-negative sparse singular value decomposition. IEEE/ACM Trans. Comp. Biol. Bioinfo., (TCBB)., 13, 656–668
83. Fanciulli, M., Petretto, E. and Aitman, T. J. (2010) Gene copy number variation and common human disease. Clin. Genet., 77, 201–213
84. Aldred, P. M., Hollox, E. J. and Armour, J. A. (2005) Copy number polymorphism and expression level variation of the human alpha-defensin genes *DEFA1* and *DEFA3*. Hum. Mol. Genet., 14, 2045–2052
85. Breunis, W. B., van Mirre, E., Bruin, M., Geissler, J., de Boer, M., Peters, M., Roos, D., de Haas, M., Koene, H. R. and Kuijpers, T. W. (2008) Copy number variation of the activating *FCGR2C* gene predisposes to idiopathic thrombocytopenic purpura. Blood, 111, 1029–1038
86. Bayés, M., Magano, L. F., Rivera, N., Flores, R. and Pérez Jurado, L. A. (2003) Mutational mechanisms of Williams-Beuren syndrome deletions. Am. J. Hum. Genet., 73, 131–151
87. Marshall, C. R., Young, E. J., Pani, A. M., Freckmann, M. L., Lacassie, Y., Howald, C., Fitzgerald, K. K., Peippo, M., Morris, C. A., Shane, K., *et al.* (2008) Infantile spasms is associated with deletion of the *MAGI2* gene on chromosome 7q11.23-q21.11. Am. J. Hum. Genet., 83, 106–111
88. Baldini, A. (2004) DiGeorge syndrome: an update. Curr. Opin. Cardiol., 19, 201–204
89. Bi, W., Yan, J., Stankiewicz, P., Park, S. S., Walz, K., Boerkoel, C. F., Potocki, L., Shaffer, L. G., Devriendt, K., Nowaczyk, M. J., *et al.* (2002) Genes in a refined Smith-Magenis syndrome critical deletion interval on chromosome 17p11.2 and the syntenic region of the mouse. Genome Res., 12, 713–728
90. Potocki, L., Chen, K. S., Park, S. S., Osterholm, D. E., Withers, M. A., Kimonis, V., Summers, A. M., Meschino, W. S., Anyane-Yeboa, K., Kashork, C. D., *et al.* (2000) Molecular mechanism for duplication 17p11.2 – the homologous recombination reciprocal of the Smith-Magenis microdeletion. Nat. Genet., 24, 84–87
91. Stone, J. L., O'Donovan, M. C., Gurling, H., Kirov, G. K., Blackwood, D. H. R., Corvin, A., Craddock, N. J., Gill, M., Hultman, C. M., Lichtenstein, P., *et al.* (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature, 455, 237–241
92. Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O. P., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer–Voskamp, J. E., *et al.* (2008) Large recurrent microdeletions associated with schizophrenia. Nature, 455, 232–236
93. de Vries, B. B., Pfundt, R., Leisink, M., Koolen, D. A., Vissers, L. E., Janssen, I. M., Reijmersdal, S., Nillesen, W. M., Huys, E. H., Leeuw, N., *et al.* (2005) Diagnostic genome profiling in mental retardation. Am. J. Hum. Genet., 77, 606–616
94. Friedman, J. M., Baross, Á., Delaney, A. D., Ally, A., Arbour, L., Asano, J., Bailey, D. K., Barber, S., Birch, P., Brown-John, M., *et al.* (2006) Oligonucleotide microarray analysis of genomic imbalance in children with mental retardation. Am. J. Hum. Genet., 79, 500–513
95. Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., *et al.* (2008) Structural variation of chromosomes in autism spectrum disorder. Am. J. Hum. Genet., 82, 477–488
96. Koolen, D. A., Sharp, A. J., Hurst, J. A., Firth, H. V., Knight, S. J., Goldenberg, A., Saugier-Veber, P., Pfundt, R., Vissers, L. E., Destrée, A., *et al.* (2008) Clinical and molecular delineation of the 17q21.31 microdeletion syndrome. J. Med. Genet., 45, 710–720
97. Sharp, A. J., Mefford, H. C., Li, K., Baker, C., Skinner, C., Stevenson, R. E., Schroer, R. J., Novara, F., De Gregori, M., Ciccone, R., *et al.* (2008) A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nat. Genet., 40, 322–328
98. Mefford, H. C., Sharp, A. J., Baker, C., Itsara, A., Jiang, Z., Buysse, K., Huang, S., Maloney, V. K., Crolla, J. A., Baralle, D., *et al.* (2008) Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. N. Engl. J. Med., 359, 1685–1699
99. Shaffer, L. G., Theisen, A., Bejjani, B. A., Ballif, B. C., Aylsworth, A. S., Lim, C., McDonald, M., Ellison, J. W., Kostiner, D., Saitta, S., *et al.* (2007) The discovery of microdeletion syndromes in the post-genomic era: review of the methodology and characterization of a new 1q41q42 microdeletion syndrome. Genet. Med., 9, 607–616
100. Butler, M. G., Meaney, F. J., Palmer, C. G., Opitz, J. M. and Reynolds, J. F. (1986) Clinical and cytogenetic survey of 39 individuals with Prader-Labhart-Willi syndrome. Am. J. Med. Genet., 23, 793–809
101. Chen, K. S., Manian, P., Koeuth, T., Potocki, L., Zhao, Q., Chinault, A. C., Lee, C. C. and Lupski, J. R. (1997) Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. Nat. Genet., 17, 154–163
102. Pérez Jurado, L. A., Peoples, R., Kaplan, P., Hamel, B. C. and Francke, U. (1996) Molecular definition of the chromosome 7 deletion in Williams syndrome and parent-of-origin effects on growth. Am. J. Hum. Genet., 59, 781–792
103. Edelmann, L., Pandita, R. K., Spiteri, E., Funke, B., Goldberg, R., Palanisamy, N., Chaganti, R. S., Magenis, E., Shprintzen, R. J. and Morrow, B. E. (1999) A common molecular basis for rearrangement disorders on chromosome 22q11. Hum. Mol. Genet., 8, 1157–1167
104. Bassett, A. S. and Chow, E. W. C. (2008) Schizophrenia and 22q11.2 deletion syndrome. Curr. Psychiatry Rep., 10, 148–157
105. Walters, R., Jacquemont, S., Valsesia, A., de Smith, A.J., Martinet, D., Andersson, J., Falchi, M., Chen, F., Andrieux, J., Lobbens, S., *et al.* (2010) A new highly penetrant form of obesity due to deletions on chromosome 16p11. 363 Nature. 463, 671–675
106. Barbieri, C. E., Baca, S. C., Lawrence, M. S., Demichelis, F., Blattner, M., Theurillat, J. P., White, T. A., Stojanov, P., Van Allen, E., Stransky, N., *et al.* (2012) Exome sequencing identifies

recurrent *SPOP*, *FOXA1* and *MED12* mutations in prostate cancer. Nat. Genet., 44, 685–689

107. Kerdpon, D., Sriplung, H. and Kietthubthew, S. (2001) Expression of p53 in oral squamous cell carcinoma and its association with risk habits in southern Thailand. Oral Oncol., 37, 553–557
108. Topcu, Z., Chiba, I., Fujieda, M., Shibata, T., Ariyoshi, N., Yamazaki, H., Sevgican, F., Muthumala, M., Kobayashi, H. and Kamataki, T. (2002) *CYP2A6* gene deletion reduces oral cancer risk in betel quid chewers in Sri Lanka. Carcinogenesis, 23, 595–598
109. India Project Team of the International Cancer Genome Consortium. (2013) Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. Nat Commun, 4, 2873
110. Pickering, C. R., Zhang, J., Yoo, S. Y., Bengtsson, L., Moorthy, S., Neskey, D. M., Zhao, M., Ortega Alves, M. V., Chang, K., Drummond, J., *et al.* (2013) Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. Cancer Discov., 3, 770–781
111. Stransky, N., Egloff, A. M., Tward, A. D., Kostic, A. D., Cibulskis, K., Sivachenko, A., Kryukov, G. V., Lawrence, M. S., Sougnez, C., McKenna, A., *et al.* (2011) The mutational landscape of head and neck squamous cell carcinoma. Science, 333, 1157–1160
112. Salahshourifar, I., Vincent-Chong, V. K., Kallarakkal, T. G. and Zain, R. B. (2014) Genomic DNA copy number alterations from precursor oral lesions to oral squamous cell carcinoma. Oral Oncol., 50, 404–412
113. Murugan, A. K., Munirajan, A. K. and Tsuchida, N. (2013) Genetic deregulation of the *PIK3CA* oncogene in oral cancer. Cancer Lett., 338, 193–203
114. Freier, K., Schwaenen, C., Sticht, C., Flechtenmacher, C., Mühling, J., Hofele, C., Radlwimmer, B., Lichter, P. and Joos, S. (2007) Recurrent *FGFR1* amplification and high *FGFR1* protein expression in oral squamous cell carcinoma (OSCC). Oral Oncol., 43, 60–66
115. Martín-Ezquerra, G., Salgado, R., Toll, A., Gilaberte, M., Baró, T., Alameda Quitllet, F., Yébenes, M., Solé, F., Garcia-Muret, M., Espinet, B., *et al.* (2010) Multiple genetic copy number alterations in oral squamous cell carcinoma: study of *MYC*, *TP53*, *CCDN1*, *EGFR* and *ERBB2* status in primary and metastatic tumours. Br. J. Dermatol., 163, 1028–1035
116. Mendes, R. A. (2012) Oncogenic pathways in the development of oral cancer. J. Carcinog. Mutagen., 3, 2
117. Lee, J. A. and Lupski, J. R. (2006) Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. Neuron, 52, 103–121
118. Cook, E. H. Jr and Scherer, S. W. (2008) Copy-number variations associated with neuropsychiatric conditions. Nature, 455, 919–923
119. Kalatzis, V. and Antignac, C. (2002) Cystinosis: from gene to disease. Nephrol. Dial. Transplant., 17, 1883–1886
120. Stahl, E. A., Raychaudhuri, S., Remmers, E. F., Xie, G., Eyre, S., Thomson, B. P., Li, Y., Kurreeman, F. A., Zhernakova, A., Hinks, A., *et al.* (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet., 42, 508–514
121. Jung, S. H., Yim, S. H., Hu, H. J., Lee, K. H., Lee, J. H., Sheen, D. H., Lim, M. K., Kim, S. Y., Park, S. W., Kim, S. H., *et al.* (2014) Genome-wide copy number variation analysis identifies deletion variants associated with ankylosing spondylitis. Arthritis Rheumatol., 66, 2103–2112
122. Kim, J. H., Jung, S. H., Bae, J. S., Lee, H. S., Yim, S. H., Park, S. Y., Bang, S. Y., Hu, H. J., Shin, H. D., Bae, S. C., *et al.* (2013) Deletion variants of *RABGAP1L*, 10q21.3, and C4 are associated with the risk of systemic lupus erythematosus in Korean women. Arthritis Rheum., 65, 1055–1063
123. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., *et al.* (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature, 506, 376–381
124. de Cid, R., Riveira-Munoz, E., Zeeuwen, P. L., Robarge, J., Liao, W., Dannhauser, E. N., Giardina, E., Stuart, P. E., Nair, R., Helms, C., *et al.* (2009) Deletion of the late cornified envelope *LCE3B* and *LCE3C* genes as a susceptibility factor for psoriasis. Nat. Genet., 41, 211–215
125. Hüffmeier, U., Bergboer, J. G., Becker, T., Armour, J. A., Traupe, H., Estivill, X., Riveira-Munoz, E., Mössner, R., Reich, K., Kurrat, W., *et al.* (2010) Replication of *LCE3C-LCE3B* CNV as a risk factor for psoriasis and analysis of interaction with other genetic risk factors. J. Invest. Dermatol., 130, 979–984
126. Xu, L., Li, Y., Zhang, X., Sun, H., Sun, D., Jia, X., Shen, C., Zhou, J., Ji, G., Liu, P., *et al.* (2011) Deletion of *LCE3C* and *LCE3B* genes is associated with psoriasis in a northern Chinese population. Br. J. Dermatol., 165, 882–887
127. Veal, C. D., Reekie, K. E., Lorentzen, J. C., Gregersen, P. K., Padyukov, L. and Brookes, A. J. (2014) A 129-kb deletion on chromosome 12 confers substantial protection against rheumatoid arthritis, implicating the gene *SLC2A3*. Hum. Mutat., 35, 248–256
128. Singleton, A. B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., Hulihan, M., Peuralinna, T., Dutra, A., Nussbaum, R., *et al.* (2003) α-synuclein locus triplication causes Parkinson's disease. Science, 302, 841–841
129. Nagao, Y. (2015) Copy number variations play important roles in heredity of common diseases: a novel method to calculate heritability of a polymorphism. Sci. Rep., 5, 17156

