Copy Number Variations and Chronic Diseases

6

August N. Blackburn and Donna M. Lehman

6.1 An Introduction to Copy Number Variation

Genetic variation ranges from single nucleotide changes to large chromosome level events. The most well characterized form of variation is single nucleotide variants with a minor allele frequency (MAF) > 1 %, referred to as single nucleotide polymorphisms (SNP). SNPs have been a workhorse of human genetics due to ease and reproducibility of genotyping. In contrast, structural variation encompasses variants with a broad range of sizes and complexity and has historically lagged behind the progress achieved using SNPs due to difficulty of genotyping. Structural variation is generally defined by 5 groups of variants: deletions, insertions, duplications, translocations, and inversions. In this chapter we focus on a subset of structural variation comprised of deletions and duplications, colloquially termed copy number variation (CNV).

A.N. Blackburn
Department of Cellular and Structural Biology, University of Texas Health Science Center, 7703 Floyd Curl Drive, San Antonio, TX 78229, USA
e-mail: BlackburnA@uthscsa.edu

D.M. Lehman (✉)
Departments of Medicine and Cellular and Structural Biology, University of Texas Health Science Center, 7703 Floyd Curl Drive, San Antonio, TX 78229, USA
e-mail: lehman@uthscsa.edu

Deletions are simply the absence of sequence when compared to a reference genome. When compared to a reference, the sequences flanking a deletion of reference sequence are continuous and juxtaposed in direct orientation. Deletions are unambiguous, having a single genomic location, although in some cases exact breakpoint locations are ambiguous if there is high sequence similarity, such as repetitive DNA, at the sights flanking the deletion. Duplication refers to sequences which share high sequence homology (>90 %) that are found in greater than two copies in the genome. Duplications are often broken into tandem and dispersed categories. Dispersed duplications are often further broken down into inter-chromosomal and intra-chromosomal categories. Intra-chromosomal duplications are further distinguished as being either in direct or inverted orientation with respect to each other.

Originally copy number variation referred only to variation larger than 1 Kb. However this size limitation has largely been disregarded due to its arbitrary nature when compared to the observed spectrum of variant sizes. Very small duplications and deletions (<50 bp) are often referred to as INDELs. Although they are part of a continuous spectrum of sizes of CNVs, these smaller variants are often distinguished from larger variants because they are almost exclusively identified using sequencing. Generally, the term copy number variation is reserved for submicroscopic events, therefore excluding monosomy and trisomy, but does include variants on

R. Duggirala et al. (eds.), *Genome Mapping and Genomics in Human and Non-Human Primates*, Genome Mapping and Genomics in Animals 5, DOI 10.1007/978-3-662-46306-2_6,

the scale of 1–5 Mb even though these are assessable by microscopic techniques such as fluorescent in situ hybridization (FISH).

6.2 The Brief History of Copy Number Variation in the Human Genome

The account of the role of CNV in human disease is appropriately centered on the publication of the human genome. Access to a reference sequence enabled new technologies which subsequently powered the discoveries described below. Examples of CNV and human disease were present prior to the availability of the reference sequence. However, the extent of CNV was unknown, and global investigation of CNVs contribution to human disease was not feasible. Thus, the publication and availability of the reference human genome represents a major inflection point in the rate which new discoveries regarding CNVs were made.

6.2.1 Prior to an Available Reference Human Genome

In Tajima's description of the use of the D statistic to test the Neutral Theory of molecular evolution, he postulates that insertions and deletions may be common in the genome of *Drosophila melanogaster* (Tajima 1989). If this observation is extended to reflect our understanding of genomes as a group at the time, it can be viewed as an early prediction that CNV, albeit small CNVs, would be common in the human genome. Twenty years later we are reaching a consensus understanding of the catalog of human genetic variation at the population level, and discovering that CNV is yet even more common than had been anticipated. The functional effects of these variants at all levels of biology, from cellular processes to disease susceptibility, are still being determined. However, even prior to an available reference sequence we could draw from known examples that CNV plays an important role in human disease.

Prior to the publication of the human genome, examples of CNVs in human disease etiology were known for single gene disorders. The role of deletions at the alpha gene locus in α-thalassemias was known since the 1970s (Ottolenghi et al. 1974; Higgs et al. 1979). In the early 1990s we learned that reciprocal deletions and duplications at 17p11.2 are involved in hereditary neuropathy with liability to pressure palsies (HNPP) and Charcot-Marie-Tooth neuropathy type 1A (CMT1A) respectively (Lupski et al. 1991; Chance et al. 1993). Examples such as these provided early evidence that CNV plays a role in human disease, yet there was no basic reference sequence for the human genome, and the extent of CNV between human genomes remained unknown.

6.2.2 Publication of the Human Genome

The publication of the first drafts of the human genome was a milestone event in human history (Lander et al. 2001; Venter et al. 2001). These first drafts were known to be incomplete because structural variation such as large segmental duplications, a course definition including both polymorphic and evolutionarily fixed duplication events generally defined as sequences with ≥90 % homology, complicated the process of sequencing and assembly. However, the next draft of the human reference sequence, published in 2004 by the International Human Genome Sequencing Consortium (IHGSC), drastically reduced the number of gaps in the human reference sequence (IHGSC 2004). A major discovery during the process of completing the human reference sequence was that an abundance of insertions and deletions, identified by observing length differences when using paired-end fosmid reads, appeared to represent polymorphisms (IHGSC 2004).

6.2.3 Early Discoveries of Genome-Wide Copy Number Variation

In 2002, Bailey et al. used computational methods and available data from sequencing efforts by IHGSC and Celera Genomics to identify segmental duplications in the human genome (Bailey et al. 2002). Interestingly, they observed patterns of both interchromosomal and intrachromosomal segmental duplications, and showed that unannotated segmental duplications created falsely identified SNPs in the public SNP database (dbSNP) (Bailey et al. 2002). Although this work showed that segmental duplications were present, what remained unclear is whether these segmental duplications were polymorphic or fixed in the human population. In the months prior to the 2004 publication by the IHGSC, two independent groups reported observing large-scale copy number variation which appeared to represent polymorphisms (Iafrate et al. 2004; IHGSC 2004; Sebat et al. 2004). Sebat et al. (2004) identified 221 copy number changes representing 76 copy number polymorphisms (CNP) in 20 individuals of geographically diverse ancestry using representational oligonucleotide microarray analysis (ROMA) (Sebat et al. 2004). Iafrate et al. (2004) identified 255 loci which appeared to be polymorphic in 55 unrelated individuals using array-based comparative genomic hybridization (array CGH). These regions were significantly enriched for overlapping segmental duplications and regions of gaps in the human genome sequence (Iafrate et al. 2004). These early studies, taken together with observations from the 2004 publication by the IHGSC provided evidence that structural variation, namely copy number variation (CNV), was common in the human genome.

Using 60 trios from the International HapMap Project (International HapMap 3 Consortium (IHM3C) et al. 2003) (30 European trios and 30 Yoruba trios) Conrad et al. (2006) identified 586 regions which had observed SNP genotypes consistent with genotyping artifacts caused by deletions. Concurrently, McCarroll and his collegues reported identifying 541 deletion variants in 269 HapMap individuals by identifying in SNP genotyping data the footprints of segregating deletions, such as Hardy-Weinberg disequilibrium, Mendelian inconsistency, and clusters of null genotypes (McCarroll et al. 2006). The SNPs that produce such errors are commonly disregarded in human genetic studies, so this study group performed additional molecular assays to confirm the presence of many of these deletion variants. Interestingly, they observed that common deletions were often in linkage disequilibrium with nearby SNPs, which was an early indication that CNVs could be tagged and indirectly assayed in genome wide association studies using SNPs (McCarroll et al. 2006).

In 2005, Tuzun et al. applied the paired-end mapping strategy developed for finishing the human genome sequence to a fosmid library from a second genome and discovered 139 insertions, 102 deletions, and 56 inversions when compared to the reference sequence (IHGSC 2004; Tuzun et al. 2005). Although most of the variants they identified were novel, they replicated the observation that CNVs were enriched for regions of segmental duplication. The approach used by Tuzun et al. (2005) was a significant advancement that was able to detect variants with a higher resolution than the CNVs discovered by Iafrate et al. (2004) and Sebat et al. (2004). Subsequently, Korbel et al. (2007) applied the paired-end sequencing approach to massively parallel shotgun sequencing on the 454 platform (Korbel et al. 2007). Using conservative thresholds for variant calling, they identified 1,297 Structural Variants (SV) events, the majority of which are CNVs, with an estimated breakpoint resolution of 644 base pairs (bp). The combination of paired-end sequencing and massively parallel sequencing technologies was a major contribution that set the foundation for future studies to move from identifying CNVs by comparing a handful of genomes to characterizing CNV at the population level.

With the exception of being enriched in regions of segmental duplication, poor consistency was observed between various methodological techniques for CNV identification, suggesting either high rates of Type I or Type II errors. Since many of these studies had performed conformational assays and estimated

false discovery rates (FDR), it could be deduced that many variants remained undiscovered, likely due to methodological biases and conservative interpretation of results. Using a custom high-resolution array based comparative genomic hybridization (aCGH) approach to target previously identified CNV regions at ~1 Kb resolution, Perry et al. (2008) observed that all previous studies they investigated had overestimated the actual size of a substantial portion of the CNVs. Nonetheless, the methods developed for these early studies, and the results of the studies themselves, set the foundation for building a comprehensive understanding of structural variation at the population level, identifying mechanisms of formation, and discovering the role of CNVs in human disease.

6.3 High-Throughput Methods for Discovery and Genotyping of Copy Number Variation

Two general groups of technologies have primarily been used for CNV discovery and genotyping: array-based hybridization, and high-throughput sequencing. Array-based hybridization methods are affordable and are often superior for detecting very large deletions and duplications. However, array-based hybridization often misses smaller variants. High throughput sequencing is much more expensive than array-based hybridization, but is superior for detecting smaller variants, defining CNV boundaries, and for determining absolute copy number of high copy number duplications. Despite improvements in the methods which are currently available, the field is still in need of more comprehensive methods for CNV discovery and genotyping, and more accurate methods for CNV imputation in samples for which these technologies cannot be applied.

6.3.1 Array Based Methods

Microarrays are a group of technologies which rely on hybridization of prepared DNA samples to oligonucleotides designed to represent specific locations of the genome. Therefore, microarrays are a technology enabled by the availability of the reference sequence. There are two basic types of microarrays which are commonly used for CNV discovery and genotyping, aCGH and SNP genotyping microarrays (Alkan et al. 2011).

In aCGH two samples are fluorescently labeled and competitively hybridized to oligonucleotide arrays (Pinkel et al. 1998). Copy number variable regions are represented by imbalance in fluorescent intensity. As a QC measure the experiments are often repeated with swapped dyes. Since either of the two samples can carry copy number differences, well characterized reference genomes are preferred for comparison. Oligonucleotides are designed to uniquely identify specific locations along the genome. Signal intensities are normalized and converted to $\log_2$ ratio, a measurement representative of copy number. To identify CNVs, various algorithms can be applied that segment the genome into regions which appear to differ from the average, which is presumed to represent a copy number of 2. Deletions and duplications are detected as multiple consecutive probes which present similar decreases or increases in $\log_2$ ratio, respectively.

SNP-arrays also produce a measurement of signal intensity by comparing the hybridization intensities across samples (Peiffer et al. 2006). This measurement is known to have a lower signal to noise ratio than aCGH, but is still powerful enough to be useful. The relative intensity of each allele is informative for identifying copy number variation (Peiffer et al. 2006). If the SNP alleles are arbitrarily labeled A and B, the ratio of signal intensity of B to the sum of intensities of A and B, termed B-allele ratio, is informative of copy number. In the normal copy number state of 2, B-allele ratio will fall into three clusters: 0, 0.5, and 1, representing homozygous AA, heterozygous AB, and homozygous BB respectively. However, in the case of deletion the cluster at 0.5 will be lost indicating a loss of heterozygosity (LOH). Similarly, when there is copy number gain the cluster at 0.5 will split into two clusters of 0.33 and 0.66 representing the AAB and ABB genotypes respectively (Peiffer et al. 2006). Additional

patterns of B-allele ratios are apparent for somatic copy number variation and other defined copy number states (Alkan et al. 2011). SNP arrays can also detect copy-neutral loss of heterozygosity, which is indicative of uniparental disomy or identity by descent (Alkan et al. 2011). Further, the application of SNP arrays to CNV calling benefits from the availability of SNP genotypes which can be used for phasing, tagging, and other imputation directed purposes which we will cover in Sect. 6.3.4.

Multiple statistical approaches have been implemented to identify CNVs from aCGH and SNP arrays, the most popular of which have been versions of circular binary segmentation and hidden Markov models (Olshen et al. 2004; Colella et al. 2007; Venkatraman and Olshen 2007; Wang et al. 2007; Coin et al. 2010). Comparative analyses of these algorithms indicate that using multiple algorithms to identify CNVs should be the preferred approach (Winchester et al. 2009; Dellinger et al. 2010; Pinto et al. 2011). Array-based approaches are known to be subject to variation in local DNA concentration that is correlated with GC content, which is often observed as "waviness" of $\log_2$ ratios for markers along the chromosome, and generally requires additional normalization procedures (Diskin et al. 2008). In a recent review, Pinto et al. (2011) showed that newer arrays tended to perform much better than legacy versions, that algorithms tended to perform best on the platforms they were designed for, and that current approaches generally underestimate CNV size, which is a shift from the observation of size overestimation reported by Perry et al. (2008). Overall, this suggests that many of the technical artifacts of CNV discovery have been addressed on newer chips and software pipelines.

CNV discovery differs from CNV genotyping in that in the discovery phase there is no a priori knowledge of the location of CNVs. Once copy number variable regions have been identified, common CNVs can often be genotyped more accurately by comparing marker intensities between samples within the region of interest. This approach has been implemented to perform association testing (Barnes et al. 2008; Welcome Trust Case Control Consortium (WTCCC) et al. 2010). Haplotype structure, determined from SNP genotypes, has also been used to improve CNV genotyping procedures (Coin et al. 2010). Taken together, these implementations indicate that, when available, added information available across samples improves CNV genotyping accuracy.

6.3.2 Sequencing Based Methods

Sequencing approaches to CNV discovery can be summarized into 4 approaches: split read, paired-end read, read depth, and de novo assembly approaches (Alkan et al. 2011). For the purpose of this chapter we will describe the benefits of de novo assembly in Sect. 6.3.3. Split-read approaches seek to identify variation that is captured within a single contiguous sequence read. By computationally "splitting" the alignment of a read to a reference sequence, split read approaches can find small deletions, insertions, and duplications. For split read approaches the upper bound on the size of sequence insertions and duplications that can be identified is the length of the read minus the sequence needed to map the read uniquely to a position in the genome, because the inserted or duplicated sequence must be contained within the length of the read. Given that most whole genome sequencing (WGS) approaches produce small reads, split read approaches are generally only able to detect very small insertions and duplications. In theory, split read approaches could detect very large deletions as long as there is a read that gaps the deletion breakpoints. However, in practice split read approaches are more effective for small deletions. In 2006, Mills et al. used a split read approach to identify 415,436 INDEL polymorphisms ranging in length between 1 and 9,989 base pairs using sequencing reads from 36 individuals (Mills et al. 2006). The overwhelming majority of these variants were 1–10 base pairs in length, and they observed little overlap with deletions identified by Conrad et al. (2006) and McCarroll et al. (2006), consistent with the observation of poor overlap between studies at the time.

For the paired-end read sequencing method, a DNA library is prepared such that the length of the DNA fragments to be sequenced fit into a tight distribution. Various approaches are available to produce libraries of different size distributions. Each DNA fragment is then sequenced from both ends. Each read, one from each end of the fragment, are then mapped back to the genome. The distance between read-pairs for regions not carrying large CNVs will fall into a tight distribution indicative of the distribution of sizes of the DNA fragments in the library prep. Deletions and duplications can be identified by abnormalities in the distance between read-pairs when they are mapped back to the reference. Deletions will create read-pairs that map further apart, and insertions will create read-pairs that map closer together than expected based on the distribution of the DNA library. As with a split-read approach there is an upper bound on the size of insertion/duplication that can be detected because the duplication has to be carried by the DNA molecule being sequenced. It is worth noting that paired-end sequencing can also detect inversion and novel sequence insertions by using one read as an anchor. As mentioned earlier, in the discovery of germ-line CNV, the paired-end read approach was first applied to fosmid libraries (Tuzun et al. 2005), and was later combined with WGS (Korbel et al. 2007).

In the read-depth approach, the coverage of the genome by sequencing reads is assumed to be uniformly distributed. Therefore regions with a loss or gain of genetic material are represented by loss or gain in the number of sequence reads. This approach was first applied to germ-line variants by Yoon et al. (2009). This approach is more effective and accurate with higher read-depth and is superior to split-read and paired-end read approaches for identifying large duplication events. Additionally, this method is superior to array based technologies for determining absolute copy number of high copy number duplications. However, similar to aCGH this approach requires correction for genomic "waviness" due to local GC content (Yoon et al. 2009). A comprehensive assessment of the platforms and computational strategies currently available using the read depth approach has recently been conducted (Magi et al. 2012).

6.3.3 Comprehensive Discovery and Genotyping

The methods described in Sects. 6.3.1 and 6.3.2 are not comprehensive. Identifying variants using array based hybridization is dependent on probe hybridization at the locus of the variant. Thus arrays with lower probe densities generally do not detect smaller variants. Arrays have poorer breakpoint definition than sequencing methods. All of the sequencing based approaches presented are powerful, but each is dependent on aligning reads to a reference sequence. There are situations where aligning reads to a reference sequence is not sufficient, such as the identification of unique sequence insertions, or variant calling in regions where short reads cannot be uniquely mapped. This suggests that an alternative approach may be necessary to genotype some CNVs.

De novo assembly followed by genome comparison is argued to be the most likely route to a comprehensive approach for variant discovery and genotyping because this approach is not dependent on aligning reads to a reference sequence (Alkan et al. 2011). This argument is very compelling, but current technologies produce short reads which limit the feasibility of this approach. However, this is a promising route to a truly comprehensive discovery and genotyping assuming technologies can be developed which produce extremely long reads, in the area of 100–200 Kb, with high accuracy.

In 2010, Pang et al. used and compared multiple approaches, including de novo assembly and comparison of genomes to identify CNVs in HuRef (Venter et al. 2001) DNA (Pang et al. 2010). Overall, de novo assembly was the most comprehensive method for CNV identification, but did miss known CNVs identified using other techniques. Each method used had its own distribution of sizes of variants in which it performed best, as expected based on the methods described above. CNVs between 1 and 10 Kb

had the highest proportion of overlap of variants detected between technologies (Pang et al. 2010).

In the absence of a technology which produces long reads with high accuracy, there is growing momentum toward considering various forms of data in combined models. By combining read-depth with high resolution aCGH developed a method for correcting the reference copy number biases in aCGH alluded to in Sect. 6.3.1 (Ju et al. 2010). This approach was then applied to identify common Asian copy number variants with high accuracy (Park et al. 2010). The 1000 genomes project has also taken the approach of combining multiple lines of evidence for CNV identification (Mills et al. 2011). More specifically, Mills et al. combined results from multiple algorithms representing split read, read-pair, read-depth, assembly, and a combination read-pair/read-depth approach to identify CNVs in low coverage sequencing and trio sequencing data generated using three different sequencing platforms.

6.3.4 Imputation of CNVs

Comprehensive variant discovery through WGS is currently prohibitively expensive, which has motivated the development of methods to impute unobserved genotypes in samples using a framework of known genotypes (Howie et al. 2012). The feasibility of imputing di-allelic CNVs was demonstrated using data generated with SNP genotyping platforms and HapMap samples (International HapMap 3 Consortium (IHM3C) et al. 2010; Surakka et al. 2010). Not surprisingly, imputation performs more effectively with population-specific reference panels, especially for polymorphisms with lower Minor Allele Frequencies (MAFs) (IHM3C et al. 2010; Surakka et al. 2010). Since 2010, major strides have been made toward better computational approaches for imputation (Li et al. 2010; Howie et al. 2011, 2012), and toward more comprehensive reference panels (Mills et al. 2011; 1000 Genomes Project Consortium et al. 2012). In population samplings, imputation is currently limited to simple forms of CNV with higher MAFs. However, examples are beginning to indicate that complex regions of the genome containing CNV may be amenable to imputation as well (Boettger et al. 2012). Additionally, in pedigrees, where phase can be determined with high accuracy for entire chromosomes, imputation of complex regions should also be achievable.

6.4 Mechanisms of CNV Formation and Mutation Rates

Mechanisms of CNV formation can be broken into two broad categories: those which involve long homologous sequences, such as non-allelic homologous recombination (NAHR), and those which involve non-homologous repair (NHR), which often entail micro-homology at the breakpoint sites (Hastings et al. 2009).

NAHR between segmental duplications and Variable Number of Tandem Repeats (VNTR) shrinkage/expansion produce copy number variants with overlapping, but distinct, size distributions (Conrad et al. 2010). Using arrayCGH data with highly accurate breakpoint resolution, Conrad et al. (2010) investigated mechanisms of formation for CNVs genotyped in 450 individuals and determined that NAHR between segmental duplications contributed more frequency for larger variants than VNTR shrinkage/expansion, which had a greater relative contribution to formation of smaller CNVs. Interestingly, formation of duplications appeared more likely to be sequence dependent than formation of deletions, yet without knowledge of the exact sequence at the breakpoints the precise mechanisms of formation for those which could not be attributed to one of these two mechanisms remained unclear (Conrad et al. 2010).

WGS has provided information for those mechanisms of formation that requires knowledge of the exact sequence at CNV breakpoints. Mills et al. (2011) investigated sequencing data generated during the pilot phase of the 1000 genomes project observed that micro-homology/homology between 2 and 376 bases were present in the sequence flanking 70.8 and 89.6 % of deletions and insertion/duplications respectively.

Interestingly, for tandem duplications, duplication size was linearly correlated with the length of homologous sequence flanking the duplication. Mobile element insertions (MEI) were the predominant mechanisms of formation of insertion/duplications, while non-homologous repair (NHR) mechanisms such as micro-homology mediated break induced repair (MMBIR) were the predominant mechanism of deletion formation. NAHR was the second most predominant mechanism of formation for both insertion/duplications and deletions, making up a substantial portion of both. NAHR and NHR contribute to variants across the spectrum of CNV sizes, yet VNTR-mediated events were enriched for smaller events, which was consistent with the aCGH study by Conrad et al. (2010). Among MEI mediated duplications, there are enrichments of variants at 300 bp and 6 Kb, representing *Alu* and long interspersed elements (LINEs). It is important to note that very large duplications and deletions, greater than 100 Kb for deletions and 10 Kb for duplications are likely underrepresented in this study due to difficulty detecting CNVs beyond these limits using sequencing methods.

Among mechanisms of formation, NAHR between segmental duplications is of high clinical relevance. Through this mechanism, a large portion ($\sim$10 %) of the genome is predisposed to recurrent mutational events (Mefford and Eichler 2009). Recurrent copy number variants resulting from this mechanism are often large enough with sufficient shared genetic material that they can be presumed to exert similar effects on phenotypes of interest. Although NAHR is an important mechanism for recurrent mutation the effect of other mechanisms of formation should not be discounted. There are examples of Mendelian disorders which show that additional mechanisms, which normally mediate non-recurrent CNV mutations, produce similar phenotypic effects as the observed CNVs generated through NAHR between segmental duplications. For example, NAHR between segmental duplications is recognized to contribute to 99 % of CMT1A and HNPP cases, yet Zhang et al. (2010) identified 17 unique CNVs in this same region formed by additional mechanisms that produced phenotypic effects consistent with CMT1A and HNPP (Zhang et al. 2010).

6.5 Common CNVs and Disease

In recent years there has been much discussion over the role of common and rare variants in complex trait variation, with strong arguments being presented in support of both (Gibson 2011). The common disease common variant (CDCV) hypothesis has been tested through GWAS, the hallmark of which was published by the Wellcome Trust Case Control Consortium (WTCCC) in 2007. GWASs have identified over 1,000 SNPs which are associated with human disease-related phenotypes (Hindorff et al. 2009). However, these associations only explain a small portion of the additive heritability of the majority of traits investigated (Gibson 2011).

Given this observation, one may hypothesize that common CNVs, which we will refer to as copy number polymorphisms (CNPs), accounts for a portion of this "missing heritability". However, as discussed above CNPs are generally well tagged by SNPs, and therefore have already been indirectly interrogated through GWAS and are unlikely to explain the observed "missing heritability" (Hinds et al. 2006; McCarroll et al. 2006; Conrad et al. 2010). This observation was confirmed through direct interrogation of 3,432 CNPs in eight disease traits by the WTCCC, in which all significantly associated CNPs were tagged by SNPs already detected in GWAS (WTCCC et al. 2010). CNPs generally make more compelling candidates for functional alleles than SNPs because of their size and increased likelihood to overlap genes. Yet due to LD, proof of functionality requires additional information in the form of biological assays. In addition, associated CNPs are also subject to the possibility of synthetic association similar to those observed with GWAS using SNPs (Dickson et al. 2010).

Despite the observation that common CNVs do not appear to account for a large portion of the missing additive heritability of common complex disorders, there are common structural variants

which are associated with complex disease phenotypes. A hallmark example is a study in which the authors hypothesized and confirmed that copy number variation at the gene CCL3L1 is associated with risk for HIV/AIDS susceptibility (Gonzalez et al. 2005). Further they showed that CCL3L1 copy number is highly population differentiated with higher CCL3L1 being more prevalent in Africans than non-Africans (Gonzalez et al. 2005). Among CNVs investigated by Conrad et al. (2010) this variant was the most highly population differentiated CNV overlapping gene exons.

A second hallmark example comes from age-related macular degeneration. In 2005, three groups independently identified a common SNP coding variant, Y402H, in complement factor H (CFH) which was strongly associated with risk for age-related macular degeneration (AMD) (Edwards et al. 2005; Haines et al. 2005; Klein et al. 2005). The population attributable risk of this variant to AMD was independently estimated to be 43 and 50 % (Edwards et al. 2005; Haines et al. 2005). This work strongly implicated the complement system in age-related macular degeneration. The following year, Hughes et al. was investigating SNPs in the complex region containing CFH and the related receptor genes CFHR1, CFHR2, CFHR3, CFHR4, and CFHR5. During their investigation they discovered a deletion of CFHR1 and CFHR3 on a common haplotype in Europeans that conferred reduced risk (odds ratio of 0.4) for age-related macular degeneration. Using multiple techniques they found that the deletion was 84,682 bases and flanked by two nearly identical 29 Kb segmental duplications implicating NAHR as the mechanism of formation.

A compelling, largely untested mechanism which fits into the CDCV hypothesis is the role of CNPs in mediating recurrent mutational events. As described above, segmental duplication predisposes $\sim$10 % of the genome to recurrent mutation. As we just described, there are known examples in which complex disease risk CNVs were formed by NAHR between segmental duplications. Further as will be presented in Sect. 6.6, there are known examples in which complex disorders are caused by NAHR between segmental duplications which are polymorphic. It is therefore possible, depending on the size of the gap between the duplicated sequences, and the haplotype structure of the population, for these duplications to be carried on separate haplotype blocks. If both locations are polymorphic, this mechanism represents a form of potential epistatic interaction which has not yet been tested.

6.6 De Novo and Low MAF Variants

De novo and low MAF CNVs are known to play a role in multiple complex disorders. Large rare CNVs are enriched in patients with schizophrenia (Malhotra et al. 2011; Walsh et al. 2008; Xu et al. 2008), bipolar disorder (Malhotra et al. 2011), autism spectrum disorders (Sebat et al. 2007; Mefford and Eichler 2009; Pinto et al. 2010; Levy et al. 2011; Sanders et al. 2011), congenital heart disease (Soemedi et al. 2012), and developmental delay (Cooper et al. 2011). NAHR between segmental duplications is known to be a major driving force of de novo and low MAF CNVs across the size spectrum of CNVs, including those identified in these enrichments. Recurrent deletions mediated by NAHR between segmental duplications are hypothesized to predispose to disease (Sharp et al. 2006; Cooper et al. 2011). Micro-deletion syndromes can provide an example through which NAHR derived rare CNVs affect human disease. They are clinically heterogeneous phenotypes which are associated with specific recurrent deletions mediated by NAHR between segmental duplications. In general there is a correlation between the size of deletions and the severity of the phenotypes observed in individuals carrying these deletions due to increased likelihood for large deletions to overlap genes or create effects on gene expression, as discussed in Sect. 6.8. Since currently known micro-deletion syndromes are caused by very large deletions, it is likely that the micro-deletion syndromes represent the tail-end of a continuous distribution of phenotypic

effects that are caused by deletions resulting from NAHR between segmental duplications.

A well-known micro-deletion syndrome is Koolen syndrome involving chromosome 17q21.31, for which carriers present with mental retardation, hypotonia, recognizable facial characteristics, and other heterogeneous phenotypes (Koolen et al. 2006, 2008; Sharp et al. 2006). Flanking segmental duplications predisposing to deletion fall on an inversion polymorphism shown to be under positive selection in Europeans (Stefansson et al. 2005), and has been shown to be a genetic determinate of meiotic recombination rates (Stefansson et al. 2005; Chowdhury et al. 2009; Fledel-Alon et al. 2011). Further investigation of this region has delineated at least 9 unique common haplotypes determined by inversion and duplication status of two unique sequences (Boettger et al. 2012; Steinberg et al. 2012). Segmental duplications containing the gene KANSL1 have independently derived and risen to high population frequencies in two unique instances suggesting positive selective pressure for increased copy number of KANSL1 (Boettger et al. 2012; Steinberg et al. 2012). The duplicated sequences are in direct orientation in only one of these two distinct segmental duplications events, and therefore only one of the two segmental duplication events predisposes to 17q21.31 micro-deletion syndrome. The haplotype carrying this segmental duplication is only observed at an appreciable frequency in Caucasian individuals (Boettger et al. 2012; Steinberg et al. 2012). This example and others, such as the complexity of the regions harboring CFH, CFHR1, and CFHR3 genes, suggest that similar complexity can be expected to underlie currently unidentified regions responsible for the heritable component of complex disease.

Taken together with the enrichment of large de novo and low MAF CNVs in complex disorders it is likely that additional micro-deletions will account for a portion of the apparent missing heritability of complex traits. As much as 5 % of schizophrenia and autism have been attributed to rare copy number variation at only a half dozen genomic locations (Gibson 2011). The strong known role of rare and de novo copy number variation is cited as support of rare variation in the etiology of common complex diseases (Gibson 2011).

As an example, micro-deletion at the 17q12 locus causes renal cyst and diabetes syndrome, also referred to as maturity onset diabetes of the young 5 (MODY5) (Nagamani et al. 2010). The effect of this deletion is sufficiently strong such that we were able to predict diabetes status in 2 of 3 related women carrying a ~1.44 Mb deletion in this region (Blackburn et al. 2013). Interestingly, the age of onset of the women with diabetes were 17 and 22.4 years respectively, representing the tail end of the distribution, while one woman was diabetes free at age 31, indicating incomplete penetrance (Blackburn et al. 2013). This observation fits the described scenario in which currently identified micro-deletion syndromes represent the tail-end of a continuous distribution of phenotypic effects, and supports the hypothesis that recurrent micro-deletions account for a portion of the observed phenotypic variation in complex disorders.

6.7 Population Studies of CNV

CNV has been investigated in multiple populations including Caucasians (Conrad et al. 2010; International HapMap 3 Consortium (IHM3C) et al. 2010; Mills et al. 2006, 2011), Asians (IHM3C et al. 2010; Ku et al. 2010; Park et al. 2010; Lou et al. 2011; Mills et al. 2011), Africans (IHM3C et al. 2010; Mills et al. 2011; Wineinger et al. 2011), and admixed populations such as Mexican Americans (IHM3C et al. 2010; Itsara et al. 2010; Mills et al. 2011; Blackburn et al. 2013). Following expected results according to population genetics theory, populations which have undergone bottlenecking carry the lowest number of polymorphisms, followed by admixed populations and populations which have not undergone recent bottlenecking such as Africans. Smaller CNVs are more frequent in individual genomes than larger CNVs, which may be attributable to selective forces, but may also be a byproduct of the mechanisms of formation. Deletions overlapping genes are enriched

for lower minor allele frequencies (Conrad et al. 2010; Mills et al. 2011). Additionally, there is an inverse relationship between the size of deletions and their individual minor allele frequencies, which suggests that large deletions are under stronger purifying selection (Blackburn et al. 2013). Interestingly, CNPs in segmental duplication regions appear to be more population differentiated than CNPs in unique regions, and biallelic CNPs show greater population stratification than frequency matched SNPs (Campbell et al. 2011). Taken together these observations suggest that large deletions in regions of segmental duplication generally produce stronger effects and are under stronger selective pressure than SNPs, smaller deletions, and less complex regions of the genome. It is also observed that low MAF CNVs are more likely to be population specific (Mills et al. 2011) which is consistent with an enrichment of rare variants due to recent population expansion. These low MAF variants may contribute significantly to heritability of and ethnic differences in complex disorders. In summary, population genetic studies of CNVs provide evidence that suggest that large deletions and regions of segmental duplication may be especially deleterious and that these are good candidate regions to affect complex disease. Their low MAF may explain why their role has remained undiscovered to date.

6.8 CNV and Gene Expression

The role of gene expression in gene mapping is extensively covered in Chap. 5. Briefly, a mechanism through which disease variants can exert their effect is by affecting gene transcript abundance. Further, the expression quantitative trait loci (eQTL) with the strongest effect sizes have been observed to act primarily in *cis* (Göring et al. 2007). As a result, transcript abundances are of great interest as highly map able endophenotypes, and as a model of disease gene mapping. Given this, it is important to briefly address the role of CNVs in heritable gene expression.

Currently there are only a few comprehensive reports regarding investigating the role of copy number variation in heritable variation in gene expression. Schlattl et al. (2011), used Bacterial Artificial Chromosomes (BAC) arrays and 500 k SNP arrays to ascertain CNVs in 210 unrelated HapMap individuals, and attempted to identify a relative contribution of CNVs and SNPs to gene expression (Stranger et al. 2007). They determined that there was little overlap between eQTL associated with SNPs and those associated with CNVs (Stranger et al. 2007). However, a more comprehensive study by Schlattl et al. used CNV calls from the 1000 genomes project data, and equally high quality gene expression data and concluded that ~48 % of CNV-associated eQTL genes are also identified using SNPs (Schlattl et al. 2011), an observation that is consistent with LD between common CNVs and SNPs. As with SNP associations from GWAS, it often remains unclear whether the CNVs identified are causally related to the expression phenotypes of interest because they could simply be tagging truly causal variants through LD. Schlattl et al. (2011) showed that significant CNV-gene pairs in which the CNV and gene overlap were enriched for positive correlations, strongly suggesting causality. Further, Gamazon et al. (2011) found that SNPs tagging CNVs are significantly enriched for *cis* eQTLs, and are overrepresented in the National Human Genome Research Institute (NHGRI) catalog of GWAS SNPs. Taken together; this evidence suggests that CNVs overlapping genes make very compelling candidate variants in eQTL, QTL, and GWAS regions.

The authors of this book chapter, and others, have reported that larger CNVs appear at lower frequencies, which suggest purifying selection (Blackburn et al. 2013). Similarly, it is observed that larger CNVs are more likely to influence the expression of nearby genes, which provides a mechanism through which larger CNVs could be under stronger purifying selection (Schlattl et al. 2011). Currently there remain many aspects of the relationship between CNV, heritable gene expression, and complex disease that remain undetermined. Presumably there is a plethora of

unidentified CNV-associated eQTL to be discovered. Further, we don't know the contribution of dispersed duplications to heritable gene expression at their insertion sites since, for some duplications, the insertion sites currently remain unknown. We also do not know if CNVs which affect the expression of a gene are more likely to affect the expression of a second non-overlapping gene, and if so, what the predominant mechanisms for this effect are. The relative contribution of common and rare variants on gene expression in most human tissues is also unknown at this time.

6.9 Somatic CNVs, Aging, and Cancer

Somatic mosaicism of copy number variation is an understudied aspect of the heritable component of human disease. Initially, these two areas seemed divorced from each other, as somatic events were thought to be stochastic. However, elucidations of the mechanisms which determine copy number mutational events suggest these may be related to each other. As we discussed, some regions of the genome are predisposed to recurrent mutational events. One can now image a scenario in which CNVs predispose a region of the genome harboring a tumor suppressor or oncogene to deletion or duplication through mechanisms outlined in Sect. 6.4 above, the end result being predisposition to the specific recurrent somatic mutational events observed in cancer. Multiple lines of evidence already strongly suggest that mutations in genes regulating DNA repair predispose to cancer phenotypes. However, little work has been done to identify whether the somatic events observed in cancer are themselves heritable, and if so what the genetic determinates of this heritable component are. The high heritability estimates of some cancers and the observed recurrent causal mutation events might suggest that the recurrent mutational events themselves are heritable, although to our knowledge this hypothesis has not been directly tested.

In 2008, two studies reported results indicating somatic mosaicism of copy number variation. Bruder et al. (2008) reported observing discordant CNVs between monozygotic twins, a clear indication of somatic mosaicism. Piotrowski et al. (2008) reported observing copy number differences between otherwise healthy differentiated tissues. However, three sets of twins studied using WGS did not appear to harbor discordant copy number variation (Baranzini et al. 2010). These observations seemed to be at odds. However, improved methods for detection of somatic mosaicism from SNP array and array-CGH data have been developed (Gonzalez et al. 2011), which is beginning to lead to a more refined understanding of somatic structural changes (Forsberg et al. 2012). The primary observations thus far are that somatic structural changes increase with age and that there appears to be self-removal of these aberrant cells in blood (Forsberg et al. 2012). Both of these observations may potentially explain the apparent discrepancies observed in the previous studies. Interestingly, these observations are consistent with late age of onset somatic diseases such as cancer.

6.10 Final Remarks

Technological advances following the publication of the human genome have allowed us to begin to investigate copy number variation in human populations in a genome-wide fashion. Early studies investigating copy number variation showed poor overlap of identified variants between studies, but provided important methods which are now commonly used in the field. Comprehensive methods for CNV discovery and genotyping are a necessity for thorough investigation, and these methods are still in development. Genome wide association studies of copy number variation have provided examples of variants that fit the CDCV hypothesis; however the observed associations are not sufficient to account for the estimated additive heritability of complex disorders. Rare and de novo CNVs have been strongly associated with multiple complex

disorders, and have provided evidence for recurrent mutation as a mechanism of disease. Although little work has been done to elucidate the relationship between CNV and heritable gene expression, early investigations from this area of research indicate that CNVs which overlap genes make especially enticing functional variant candidates in complex disease loci. The role of heritable predisposition to somatic mosaicism of CNV in complex disease is a wholly unstudied research area which is empirically promising based on observations from studies in cancer and the mechanisms of formation of CNVs. Initial studies indicate that somatic mosaicism of CNV is likely ripe with undiscovered disease mechanisms. In summary, the field of investigation of the role of CNV in common complex disorders is immature, yet early work indicating that CNV is a major source of genetic and heritable phenotypic variation between individuals suggests that those willing to investigate these more complicated regions of the genome in complex diseases should be prepared for interesting discoveries.

References

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073. doi: 10.1038/nature09534

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65. doi:10.1038/nature11632

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12:363–376. doi:10.1038/nrg2958

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002) Recent segmental duplications in the human genome. Science 297:1003–1007

Baranzini SE, Mudge J, van Velkinburgh JC, Khankhanian P, Khrebtukova I, Miller NA, Zhang L, Farmer AD, Bell CJ, Kim RW, May GD, Woodward JE, Caillier SJ, McElroy JP, Gomez R, Pando MJ, Clendenen LE, Ganusova EE, Schilkey FD, Ramaraj T, Khan OA, Huntley JJ, Luo S, Kwok PY, Wu TD, Schroth GP, Oksenberg JR, Hauser SL, Kingsmore SF (2010) Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. Nature 464(7293):1351–1356. doi:10.1038/nature08990

Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, Hurles ME (2008) A robust statistical method for case-control association testing with copy number variation. Nat Genet 40:1245–1252. doi:10.1038/ng.206

Blackburn A, Göring HH, Dean A, Carless MA, Dyer T, Kumar S, Fowler S, Curran JE, Almasy L, Mahaney M, Comuzzie A, Duggirala R, Blangero J, Lehman DM (2013) Utilizing extended pedigree information for discovery and confirmation of copy number variable regions among Mexican Americans. Eur J Hum Genet 21:404–409. doi:10.1038/ejhg.2012.188

Boettger LM, Handsaker RE, Zody MC, McCarroll SA (2012) Structural haplotypes and recent evolution of the human 17q21.31 region. Nat Genet 44:881–885. doi:10.1038/ng.2334

Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, Diaz de Ståhl T, Menzel U, Sandgren J, von Tell D, Poplawski A, Crowley M, Crasto C, Partridge EC, Tiwari H, Allison DB, Komorowski J, van Ommen GJ, Boomsma DI, Pedersen NL, den Dunnen JT, Wirdefeldt K, Dumanski JP (2008) Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet 82:763–771. doi:10.1016/j.ajhg.2007.12.011

Campbell CD, Sampas N, Tsalenko A, Sudmant PH, Kidd JM, Malig M, Vu TH, Vives L, Tsang P, Bruhn L, Eichler EE (2011) Population-genetic properties of differentiated human copy-number polymorphisms. Am J Hum Genet 88:317–332. doi:10.1016/j.ajhg.2011.02.004

Chance PF, Alderson MK, Leppig KA, Lensch MW, Matsunami N, Smith B, Swanson PD, Odelberg SJ, Disteche CM, Bird TD (1993) DNA deletion associated with hereditary neuropathy with liability to pressure palsies. Cell 72:143–151

Chowdhury R, Bois PR, Feingold E, Sherman SL, Cheung VG (2009) Genetic analysis of variation in human meiotic recombination. PLoS Genet 5(9): e1000648. doi:10.1371/journal.pgen.1000648

Coin LJ, Asher JE, Walters RG, Moustafa JS, de Smith AJ, Sladek R, Balding DJ, Froguel P, Blakemore AI (2010) cnvHap: an integrative population and haplotype-based multiplatform model of SNPs and CNVs. Nat Methods 7:541–546. doi:10.1038/nmeth.1466

Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J (2007) QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35:2013–2025

Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38:75–81

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J; Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME (2010) Origins and functional impact of copy number variation in the human genome. Nature 464:704–712. doi: 10.1038/nature08516

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, Abdel-Hamid H, Bader P, McCracken E, Niyazov D, Leppig K, Thiese H, Hummel M, Alexander N, Gorski J, Kussmann J, Shashi V, Johnson K, Rehder C, Ballif BC, Shaffer LG, Eichler EE (2011) A copy number variation morbidity map of developmental delay. Nat Genet 43:838–846. doi:10.1038/ng.909

Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, Li YJ (2010) Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res 38:e105. doi:10.1093/nar/gkq040

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. PLoS Biol 8:e1000294. doi:10.1371/journal.pbio.1000294

Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, Bucan M, Maris JM, Wang K (2008) Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res 36:e126. doi:10.1093/nar/gkn556

Edwards AO, Ritter R 3rd, Abel KJ, Manning A, Panhuysen C, Farrer LA (2005) Complement factor H polymorphism and age-related macular degeneration. Science 308:421–424

Fledel-Alon A, Leffler EM, Guan Y, Stephens M, Coop G, Przeworski M (2011) Variation in human recombination rates and its genetic determinants. PLoS ONE 6:e20321. doi:10.1371/journal.pone.0020321

Forsberg LA, Rasi C, Razzaghian HR, Pakalapati G, Waite L, Thilbeault KS, Ronowicz A, Wineinger NE, Tiwari HK, Boomsma D, Westerman MP, Harris JR, Lyle R, Essand M, Eriksson F, Assimes TL, Iribarren C, Strachan E, O'Hanlon TP, Rider LG, Miller FW, Giedraitis V, Lannfelt L, Ingelsson M, Piotrowski A, Pedersen NL, Absher D, Dumanski JP (2012) Age-related somatic structural changes in the nuclear genome of human blood cells. Am J Hum Genet 90:217–228. doi:10.1016/j.ajhg.2011.12.009

Gamazon ER, Nicolae DL, Cox NJ (2011) A study of CNVs as trait-associated polymorphisms and as expression quantitative trait loci. PLoS Genet 7:e1001292. doi:10.1371/journal.pgen.1001292

Gibson G (2011) Rare and common variants: twenty arguments. Nat Rev Genet 13:135–145. doi:10.1038/nrg3118

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'connell RJ, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307:1434–1440

González JR, Rodríguez-Santiago B, Cáceres A, Pique-Regi R, Rothman N, Chanock SJ, Armengol L, Pérez-Jurado LA (2011) A fast and accurate method to detect allelic genomic imbalances underlying mosaic rearrangements using SNP array data. BMC Bioinf 12:166. doi:10.1186/1471-2105-12-166

Göring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JB, Abraham LJ, Rainwater DL, Comuzzie AG, Mahaney MC, Almasy L, MacCluer JW, Kissebah AH, Collier GR, Moses EK, Blangero J (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet 39:1208–1216

Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, Schnetz-Boutaud N, Agarwal A, Postel EA, Pericak-Vance MA (2005) Complement factor H variant increases the risk of age-related macular degeneration. Science 308:419–421

Hastings PJ, Ira G, Lupski JR (2009) A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet 5:e1000327. doi:10.1371/journal.pgen.1000327

Higgs DR, Pressley L, Old JM, Hunt DM, Clegg JB, Weatherall DJ, Serjeant GR (1979) Negro alpha-thalassaemia is caused by deletion of a single alpha-globin gene. Lancet 2:272–276

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367. doi:10.1073/pnas.0903103106

Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet 38:82–85

Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44:955–959. doi:10.1038/ng.2354

Howie B, Marchini J, Stephens M (2011) Genotype imputation with thousands of genomes. G3 (Bethesda) 1:457–470. doi:10.1534/g3.111.001198

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36:949–951

International HapMap 3 Consortium, Altshuler DM, Gibbs RA et al (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58. doi: 10.1038/nature09298

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945

Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, Eichler EE (2010) De novo rates and selection of large copy number variation. Genome Res 20:1469–1481. doi:10.1101/gr.107680.110

Ju YS, Hong D, Kim S, Park SS, Kim S, Lee S, Park H, Kim JI, Seo JS (2010) Reference-unbiased copy number variant analysis using CGH microarrays. Nucleic Acids Res 38:e190. doi:10.1093/nar/gkq730

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389

Koolen DA, Sharp AJ, Hurst JA et al (2008) Clinical and molecular delineation of the 17q21.31 microdeletion syndrome. J Med Genet 45:710–720. doi:10.1136/jmg.2008.058701

Koolen DA, Vissers LE, Pfundt R, de Leeuw N, Knight SJ, Regan R, Kooy RF, Reyniers E, Romano C, Fichera M, Schinzel A, Baumer A, Anderlid BM, Schoumans J, Knoers NV, van Kessel AG, Sistermans EA, Veltman JA, Brunner HG, de Vries BB (2006) A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. Nat Genet 38:999–1001

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318:420–426

Ku CS, Pawitan Y, Sim X, Ong RT, Seielstad M, Lee EJ, Teo YY, Chia KS, Salim A (2010) Genomic copy number variations in three Southeast Asian populations. Hum Mutat 31:851–857. doi:10.1002/humu.21287

Lander ES, Linton LM, Birren B et al (2001) International human genome sequencing consortium initial sequencing and analysis of the human genome. Nature 409:860–921

Levy D, Ronemus M, Yamrom B, Lee YH, Leotta A, Kendall J, Marks S, Lakshmi B, Pai D, Ye K, Buja A, Krieger A, Yoon S, Troge J, Rodgers L, Iossifov I, Wigler M (2011) Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron 70:886–897. doi:10.1016/j.neuron.2011.05.015

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34:816–834. doi:10.1002/gepi.20533

Lou H, Li S, Yang Y, Kang L, Zhang X, Jin W, Wu B, Jin L, Xu S (2011) A map of copy number variations in Chinese populations. PLoS ONE 6:e27341. doi:10.1371/journal.pone.0027341

Lupski JR, de Oca-Luna RM, Slaugenhaupt S, Pentao L, Guzzetta V, Trask BJ, Saucedo-Cardenas O, Barker DF, Killian JM, Garcia CA, Chakravarti A, Patel PI (1991) DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell 66:219–232

Magi A, Tattini L, Pippucci T, Torricelli F, Benelli M (2012) Read count approach for DNA copy number variants detection. Bioinformatics 28:470–478. doi:10.1093/bioinformatics/btr707

Malhotra D, McCarthy S, Michaelson JJ, Vacic V, Burdick KE, Yoon S, Cichon S, Corvin A, Gary S, Gershon ES, Gill M, Karayiorgou M, Kelsoe JR, Krastoshevsky O, Krause V, Leibenluft E, Levy DL, Makarov V, Bhandari A, Malhotra AK, McMahon FJ, Nöthen MM, Potash JB, Rietschel M, Schulze TG, Sebat J (2011) High frequencies of de novo CNVs in bipolar disorder and schizophrenia. Neuron 72:951–963. doi:10.1016/j.neuron.2011.11.007

McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM; International HapMap Consortium (2006) Common deletion polymorphisms in the human genome. Nat Genet 38:86–92

Mefford HC, Eichler EE (2009) Duplication hotspots, rare genomic disorders, and common disease. Curr Opin Genet Dev 19:196–204. doi:10.1016/j.gde.2009.04.003

Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16:1182–1190

Mills RE, Walter K, Stewart C et al 1000 Genomes Project (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470:59–65. doi: 10.1038/nature09708

Nagamani SC, Erez A, Shen J, Li C, Roeder E, Cox S, Karaviti L, Pearson M, Kang SH, Sahoo T, Lalani SR, Stankiewicz P, Sutton VR, Cheung SW (2010) Clinical spectrum associated with recurrent genomic rearrangements in chromosome 17q12. Eur J Hum Genet 18:278–284. doi:10.1038/ejhg.2009.174

Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5:557–572

Ottolenghi S, Lanyon WG, Paul J, Williamson R, Weatherall DJ, Clegg JB, Pritchard J, Pootrakul S, Boon WH (1974) The severe form of alpha thalassaemia is caused by a haemoglobin gene deletion. Nature 1974 251:389–392

Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, Park H, Hurles ME, Lee C, Venter JC, Kirkness EF, Levy S, Feuk L, Scherer SW (2010) Towards a comprehensive structural variation map of an individual human genome. Genome Biol 11:R52. doi:10.1186/gb-2010-11-5-r52

Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, Kim S, Lee S, Suh D, Hong D, Kang HP, Yoo YJ, Shin JY, Kim HJ, Yavartanoo M, Chang YW, Ha JS, Chong W, Hwang GR, Darvishi K, Kim H, Yang SJ, Yang KS, Kim H, Hurles ME, Scherer SW, Carter NP, Tyler-Smith C, Lee C, Seo JS (2010) Discovery of common Asian copy number variants using integrated high-resolution array

CGH and massively parallel DNA sequencing. Nat Genet 42:400–405. doi:10.1038/ng.555

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16:1136–1148

Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA, Park HS, Kim JI, Seo JS, Yakhini Z, Laderman S, Bruhn L, Lee C (2008) The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet 82:685–695. doi:10.1016/j.ajhg.2007.12.010

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20:207–211

Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, Macdonald JR, Mills R, Prasad A, Noonan K, Gribble S, Prigmore E, Donahoe PK, Smith RS, Park JH, Hurles ME, Carter NP, Lee C, Scherer SW, Feuk L (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol 29:512–520. doi:10.1038/nbt.1852

Pinto D, Pagnamenta AT, Klei L et al (2010) Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466:368–372. doi:10.1038/nature09146

Piotrowski A, Bruder CE, Andersson R, Diaz de Ståhl T, Menzel U, Sandgren J, Poplawski A, von Tell D, Crasto C, Bogdan A, Bartoszewski R, Bebok Z, Krzyzanowski M, Jankowski Z, Partridge EC, Komorowski J, Dumanski JP (2008) Somatic mosaicism for copy number variation in differentiated human tissues. Hum Mutat 29:1118–1124. doi:10.1002/humu.20815

Sanders SJ, Ercan-Sencicek AG, Hus V et al (2011) Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron 70:863–885. doi:10.1016/j.neuron.2011.05.002

Schlattl A, Anders S, Waszak SM, Huber W, Korbel JO (2011) Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. Genome Res 21:2004–2013. doi:10.1101/gr.122614.111

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YH, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimäki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W, Warburton D, King MC, Skuse D, Geschwind DH, Gilliam TC, Ye K, Wigler M (2007) Strong association of de novo copy number mutations with autism. Science 316:445–449

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525–528

Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, Fitzpatrick CA, Segraves R, Richmond TA, Guiver C, Albertson DG, Pinkel D, Eis PS, Schwartz S, Knight SJ, Eichler EE (2006) Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nat Genet 38:1038–1042

Soemedi R, Wilson IJ, Bentham J et al (2012) Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. Am J Hum Genet 91:489–501. doi:10.1016/j.ajhg.2012.08.003

Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, Desnica N, Hicks A, Gylfason A, Gudbjartsson DF, Jonsdottir GM, Sainz J, Agnarsson K, Birgisdottir B, Ghosh S, Olafsdottir A, Cazier JB, Kristjansson K, Frigge ML, Thorgeirsson TE, Gulcher JR, Kong A, Stefansson K (2005) A common inversion under selection in Europeans. Nat Genet 37:129–137

Steinberg KM, Antonacci F, Sudmant PH, Kidd JM, Campbell CD, Vives L, Malig M, Scheinfeldt L, Beggs W, Ibrahim M, Lema G, Nyambo TB, Omar SA, Bodo JM, Froment A, Donnelly MP, Kidd KK, Tishkoff SA, Eichler EE (2012) Structural diversity and African origin of the 17q21.31 inversion polymorphism. Nat Genet 44:872–880. doi:10.1038/ng.2335

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848–853

Surakka I, Kristiansson K, Anttila V, Inouye M, Barnes C, Moutsianas L, Salomaa V, Daly M, Palotie A, Peltonen L, Ripatti S (2010) Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. Genome Res 20:1344–1351. doi:10.1101/gr.106534.110

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

The International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE (2005) Fine-scale structural variation of the human genome. Nat Genet 37:727–732

Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23:657–663

Venter JC, Adams MD, Myers EW et al (2001) The sequence of the human genome. Science 291:1304–1351

Walsh T, McClellan JM, McCarthy SE et al (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320:539–543. doi:10.1126/science.1155174

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, 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

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678

Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, 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. doi: 10.1038/nature08979

Winchester L, Yau C, Ragoussis J (2009) Comparing CNV detection methods for SNP arrays. Brief Funct Genomic Proteomics 8:353–366. doi:10.1093/bfgp/elp017

Wineinger NE, Pajewski NM, Kennedy RE, Wojczynski MK, Vaughan LK, Hunt SC, Gu CC, Rao DC, Lorier R, Broeckel U, Arnett DK, Tiwari HK (2011) Characterization of autosomal copy-number variation in African Americans: the HyperGEN study. Eur J Hum Genet 19:1271–1275. doi:10.1038/ejhg.2011.115

Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, Karayiorgou M (2008) Strong association of de novo copy number mutations with sporadic schizophrenia. Nat Genet 40:880–885. doi:10.1038/ng.162

Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19:1586–1592. doi:10.1101/gr.092981.109

Zhang F, Seeman P, Liu P, Weterman MA, Gonzaga-Jauregui C, Towne CF, Batish SD, De Vriendt E, De Jonghe P, Rautenstrauss B, Krause KH, Khajavi M, Posadka J, Vandenberghe A, Palau F, Van Maldergem L, Baas F, Timmerman V, Lupski JR (2010) Mechanisms for nonrecurrent genomic rearrangements associated with CMT1A or HNPP: rare CNVs as a cause for missing heritability. Am J Hum Genet 86:892–903. doi:10.1016/j.ajhg.2010.05.001