Cytogenet Genome Res 123:17–26 (2008)
DOI: 10.1159/000184688

Cytogenetic and Genome Research

Human genes involved in copy number variation: mechanisms of origin, functional effects and implications for disease

A.J. de Smith[a] R.G. Walters[a] P. Froguel[a, b] A.I. Blakemore[a]

[a]Section of Genomic Medicine, Imperial College London, Hammersmith Hospital, London (UK)
[b]CNRS 8090-Institute of Biology, Pasteur Institute, Lille (France)

Accepted in revised form for publication by H. Kehrer-Sawatzki and D.N. Cooper, 4 August 2008.

Abstract. Copy number variants (CNVs) overlap over 7000 genes, many of which are pivotal in biological pathways. The implications of this are profound, with consequences for evolutionary studies, population genetics, gene function and human phenotype, including elucidation of genetic susceptibility to major common diseases, the heritability of which has thus far defied full explanation. Even though this research is still in its infancy, CNVs have already been associated with a number of monogenic, syndromic and complex diseases: the development of high throughput and high resolution techniques for CNV screening is likely to bring further new insights into the contribution of copy number variation to common diseases. Amongst genes overlapped by CNVs, significant enrichments for certain gene ontology categories have been identified, including those related to immune responses and interactions with the environment. Genes in both of these categories are thought to be important in evolutionary adaptation and to be particular targets of natural selection. Thus, a full appreciation of copy number variation may be important for our understanding of human evolution.

The recent appreciation of copy number variation as a frequent and widespread feature of human genomes has the potential to revolutionise our understanding of clinical and population genetics and genetic epidemiology. Changes in copy number are important for all types of human disease: genomic disorders, monogenic diseases, infections, autoimmunity, cancer and other complex disorders, and should now be considered in study design for all aspects of human genetic analysis.

Until recently it was assumed that healthy individuals carry two copies of every autosomal gene, with one copy inherited on the maternally-inherited chromosome, and the other inherited paternally. Any deviation from this was assumed to be phenotypically harmful: it was initially predicted that any two 'normal' individuals are approximately 99.9% similar genetically, with single nucleotide polymorphisms (SNPs) providing the main source of human genetic variation (Altshuler et al., 2000). This view has been completely overturned in the last few years, since the overcoming of two major problems which previously delayed or prevented detection of common CNVs. The most critical was the resolution gap between microscopy-based and molecular biological technologies – resolved by the advent of microarray-based comparative genome hybridisation (array CGH) in 1998 (Pinkel et al., 1998). The second issue was that there has historically been much greater research effort directed toward analysis of disease than to in-depth analyses of 'normal' genomes; this may have led to a selection bias such that aberrations with detrimental phenotypic effects were sought, whereas relatively benign ones were not looked

Higher Education Funding Council for England (HEFCE), UK

Request reprints from Alexandra I.F. Blakemore
Hammersmith Hospital Campus, Imperial College London
Du Cane Road, London W12 0NN (UK)
telephone: +44 207 594 6511; fax: +44 207 594 6543
e-mail: a.blakemore@imperial.ac.uk

1424–8581/08/1234–0017$24.50/0

Accessible online at:
www.karger.com/cgr

for and, therefore, not found. Recent technological advances, however, have revealed this copy number variation to be much more common than previously suspected, with apparently healthy individuals varying from each other in a large number of genomic regions.

Large chromosomal aberrations, containing many genes, have long been known to cause developmental abnormalities and, on a smaller scale, genomic disorders can be caused by microdeletions or duplications (<10 Mb). In some cases, the copy number change of only a single gene is thought to underlie such diseases, as exemplified by the duplication of *PMP22* in Charcot-Marie-Tooth disease type 1A (CMT1A) (Roa et al., 1991) and the deletion of *RAI1* in Smith-Magenis syndrome (SMS) (Slager et al., 2003), both of which came to light only after extensive investigation of affected individuals. Before 2004, however, only a handful of genes were known to vary in copy number in apparently healthy individuals. For example, both deletions and duplications in the HLA class III genes encoding the complement components C2 and C4 are relatively common in healthy individuals (reviewed in Campbell et al., 1986), although they have also been known to associate with systemic lupus erythematosus (SLE) for many years (Agnello et al., 1972; Hauptmann et al., 1974). Other gene clusters such as the betadefensin antimicrobial cluster (Hollox et al., 2003) and olfactory receptor genes (Trask et al., 1998) were also known to vary in copy number, but such phenomena were thought to be anomalies.

Some genes varying in copy number are associated with phenotypic effects in combination with particular environmental exposures. The *GSTM1* gene, for example, is found in only half of the Caucasian population, with the other half carrying homozygous deletions that are thought to increase susceptibility to some cancers, possibly due to a reduction in metabolism of certain carcinogenic compounds in the environment (reviewed in Rebbeck, 1997). Conversely, individuals carrying three copies of the gene have 'ultra rapid' *GSTM1* activity (McLellan et al., 1997), which may confer protection against certain cancers by increasing the breakdown of harmful environmental toxins.

Several genes encoding drug-metabolising enzymes were also known to be variable in copy number, which, in some cases, could lead to different responses to certain drugs. The cytochrome P450 gene *CYP2D6* metabolises up to 20% of all drugs in clinical use, and the rate of metabolism varies depending on the levels of expression of this enzyme in the liver (reviewed in Schaeffeler et al., 2003). As there are extensive polymorphisms of this gene, including whole gene deletions (Sachse et al., 1997), these expression levels are highly variable (Zanger et al., 2001), resulting in some people being ultra-rapid metabolisers, and others, particularly those with deletions, being poor metabolisers who are at higher risk of adverse effects after drug therapy. A similar effect is seen with other cytochrome P450 genes: people who smoke are more likely than non-smokers to carry a duplicated version of *CYP2A6,* the product of which metabolises nicotine to the inactive cotinine. Individuals with higher enzyme activity have less nicotine in their blood for the same number of cigarettes smoked, encouraging them to smoke more (Rao et al., 2000).

Despite these specific examples, it was not generally appreciated that a wide range of genes could be affected by CNVs, without obvious phenotypic effect. This was partly a result of our initial ignorance of how widespread and frequent such variation was. Following two seminal publications (Iafrate et al., 2004; Sebat et al., 2004), a number of studies have been published, revealing a wealth of CNV loci in the genome, overlapping many genes (Tuzun et al., 2005; Conrad et al., 2006; Hinds et al., 2006; McCarroll et al., 2006; Mills et al., 2006; Redon et al., 2006; de Smith et al., 2007; Korbel et al., 2007; Wong et al., 2007). In our investigation alone, a 2-stage custom array CGH analysis of just 50 unrelated healthy French Caucasian males revealed over 2500 genes overlapped by CNVs – 1284 of these genes were previously unknown to vary in copy number (de Smith et al., 2007). This illustrates that a substantial proportion of copy number variation remains to be discovered, and that CNVs are abundant in apparently healthy individuals. Such high-resolution CGH studies will continue to be useful in refining mapping of known CNVs and also in identification of smaller CNVs: it is clear that many more variant regions remain to be identified and characterised.

Published CNV data are documented on the TCAG Database of Genomic Variants (http://projects.tcag.ca/variation), and by November 2008 this contained 6225 CNV loci, covering an estimated 28.8% of the genome. These loci consist of 19,792 separate CNVs, of which almost half overlap genes (Table 1). Considering the extent of this recently discovered genomic variation, and the fact that many biologically important and potentially disease-causing genes are indeed involved, there are likely to be significant phenotypic consequences for long-term health.

Origins of CNVs

The mechanisms through which CNVs are formed have, as yet, been only partially elucidated. It is apparent that there are different subgroups of CNV in the genome, some with stable regions flanked by consistent boundaries, others with more complex, overlapping copy number variant regions (CNVRs), and these different variant types may arise through different processes or combinations of processes. For example, a proportion of variants have been associated with regions of segmental duplication (Sharp et al., 2005; Redon et al., 2006), and these low copy repeats are thought to propagate CNVs in the genome through a process known as non-allelic homologous recombination (NAHR). Only ~50% of reported variant sequences have been found to overlap segmental duplications (Cooper et al., 2007), and, therefore, other non-homology-based mechanisms, such as non-homologous end joining (NHEJ), have been implicated in the formation of some CNVs (Korbel et al., 2007; Perry et al., 2008). Another process, initiated by errors in DNA replication rather than in recombination, has recently been proposed to play a role in the formation of some copy

number changes. Fork stalling and template switching ('FoSTeS') has been suggested to explain tandem duplications and other complex rearrangements of the *PLP1* gene leading to Pelizaeus-Merzbacher disease (PMD) (Lee et al., 2007); such mechanisms might also account for the formation of CNV duplications, and possibly deletions, especially in complex regions where CNV generation is not readily explained by NAHR or NHEJ.

In a recent investigation into the origins of common deletion variants, we suggested a mechanism of CNV formation involving Alu repeats that has not been described previously. These elements are significantly enriched at the breakpoints of this class of deletion, and for nine out of 40 deletion breakpoints, the Alu poly-A tail ends precisely at the breakpoint junction (de Smith et al., 2008). This suggests the Alu poly-A sequence may be involved in the formation of some CNVs, possibly due to these sequences being prone to single or double-strand breakage, initiating the process of NHEJ. Alus are thought to correlate with gene-rich and GC-rich regions of the genome (International Human Genome Sequencing Consortium, 2001), and it may be possible that these repeats play a greater role in the formation of gene CNVs than non-gene variants. This is consistent with the fact that the CNVs found to have an enrichment of Alus at their breakpoints were initially discovered using a custom CGH array biased towards genes (de Smith et al., 2007).

Different types of genes may, therefore, be more or less prone to copy number variation depending on the type and number of nearby repeat elements (e.g. Alus and segmental duplications). In turn, this could have an effect on the underlying mechanisms of formation that are involved. The structure of a gene could also affect the type of CNV they overlap, for example some very repetitive genes may be more likely to contain complex CNVRs, whereas others may be less repetitive and only contain ancient stable CNVs – a large number of CNVs may have arisen many generations in the past and persisted stably in the population since then. For instance, on the basis of sequencing data, linkage disequilibrium and imputation of extended haplotypes, it was deduced that each of 20 deletion CNVs found in multiple unrelated individuals had a single unique historical origin, and had been stably inherited over many generations (de Smith et al., 2008).

It is clear from family studies that de novo copy number changes may occur during meiotic processes. However, copy number differences have recently been reported between monozygotic twins (Bruder et al., 2008) and there are also early reports of CNVs occurring between different tissues in the same individual (Piotrowski et al., 2008). Thus it is clear that genomic rearrangements leading to CNVs do not occur solely during meiosis. The relative importance of the various mechanisms which might give rise to CNVs, either in meiotic or somatic tissues, is completely unknown; nor is there any indication of the extent of somatic CNV mosaicism or of the rate of accumulation of such de novo CNVs during a person's lifetime. These issues may be very important for particular phenotypes, particularly with respect to cancers and diseases of ageing.

Characteristics of copy number variant genes

Genomic biases of copy number variants

The distribution of CNVs in the genome is thought to be non-random, and several 'hot' spots of variation have been identified: for example, 250 regions of 1 Mb DNA sequence have been found in which >50% bases are within variants (Cooper et al., 2007). In particular, CNVs occur more frequently towards centromeres and telomeres (Nguyen et al., 2006), perhaps because of the repetitive nature of these genomic regions. Equally, the relationship between the size of chromosomes and the number of variants contained within them is not straightforward, with some chromosomes having a relatively high proportion of copy number variation and others having a low proportion relative to their size. Chromosome 18, for instance, has only ~19.8% copy number variant sequence, while chromosomes 16, 17, 19, and 22 each have >41% copy number variant sequence (http://projects.tcag.ca/variation). This may reflect interchromosomal differences in genomic structure, for example chromosome 19 is rich in segmental duplications and tandemly clustered gene families (Grimwood et al., 2004), which may make it more prone to copy number variation. Alternatively, variations in selective pressures may lead to more or less variable chromosomes, for example the Y chromosome, which is thought to be subject to relaxed selection (reviewed in Charlesworth and Charlesworth, 2000), has a particularly high proportion of copy number variant sequence (~41.8%).

A significant relationship has also been reported between genes and CNV regions, with gene-rich genomic regions enriched with CNVs, and vice versa. Cooper et al. (2007) found an enrichment of CNVs in the most gene-rich regions of the genome (>30% cf. 21% genome-wide average), and an enrichment of exon density in the most CNV-rich genomic regions (>2.7% cf. 2.1% genome average). Analysis of CNVs in the rat genome also revealed that CNVs overlapped more genes than expected based on random distribution, suggesting this phenomenon is not limited to the human genome (Guryev et al., 2008). Data from other workers, however, suggest that there is a bias against structural variants overlapping genes (Redon et al., 2006; Korbel et al., 2007). Some of these discrepancies may reflect differences in the technological approaches used to find CNVs, for example some arrays may contain probes biased towards genes. Future studies using higher resolution technologies and approaches with reduced bias for particular genomic features, such as whole genome sequencing, should resolve this controversy.

If indeed there is a bias against CNVs overlapping genes, this could be due to selective constraints acting against these variants. Alternatively, there may be selection against genomic deletions rather than CNVs as a whole, as the genome is thought to be more tolerant of duplications than deletions (Brewer et al., 1999). Conrad et al. (2006) reported that CNV deletions were biased away from genes, and also found a strong under-representation of SNPs within genes in regions of deletions compared with the genome-wide av-

erage. Consistent with this, a lower proportion of CNV deletions compared with duplications were found to overlap OMIM genes (Redon et al., 2006). Purifying selection may, therefore, be acting against CNVs in genes generally or more specifically against gene deletions.

Gene categories enriched for CNVs

In addition to the apparent non-random distribution of CNVs, there also appear to be certain patterns in the types of genes overlapped by variants. Gene ontology (GO) analyses from various studies have revealed genes variant for copy number to be enriched for a number of categories, including immune responses and responses to external biotic stimuli (Feuk et al., 2006). Other genes involved in interactions with the environment, such as sensory perceptions of smell and chemical stimuli, as well as genes connected to taste and sight, have also been associated with CNVs (Redon et al., 2006; de Smith et al., 2007; Wong et al., 2007). In higher resolution studies, enrichment of genes involved in neurophysiological processes and in brain development has also been noted (de Smith et al., 2007). The protocadherin gene cluster on chromosome 5, for example, is particularly rich in CNVs. These genes are thought to play a role in the generation of combinatorial complexity in brain synaptic connections (Cooper et al., 2007), which hints at an impact of CNVs on the genetics of brain function and intelligence. These variant environment-response and neurodevelopmental genes may be acted upon by natural selection, thus playing a role in adaptability and fitness in response to external pressures (Feuk et al., 2006).

As well as enrichments for certain gene groups, CNVs seem to be biased away from other types of genes. For example, an impoverishment of genes encoding nucleic acid binding proteins or involved in nucleic acid metabolism was found in deletion regions (Conrad et al., 2006). Similarly, genes involved in cell signalling and cell proliferation were also reported to be underrepresented in CNVs (Redon et al., 2006). Such underrepresentation presumably reflects the effects of strong selective pressure, due to the crucial role of these genes in transcriptional regulation and development.

Phenotypic effects of gene copy number changes

Effects of copy number variation on gene expression and phenotype

Copy number variation can affect genes in a variety of ways, and this, in turn, may be reflected in their effects on phenotype. Where whole genes vary in copy number, a simple consequence would be variation in gene expression through dosage effects, whereby gains or losses in copy number would increase or decrease expression levels respectively. It has been estimated that CNVs account for at least 17.7% heritable variation in gene expression (Stranger et al., 2007), and for dosage-sensitive genes, both gain- and loss-of-function events could be phenotypically harmful or beneficial, depending on the gene. Variation in expression due to CNVs may also vary from gene to gene, for example in three commonly deleted genes, *GSTT1, GSTM1* and *UGT2B17,* gene dosage variation was found to explain 88, 75 and 26% of the observed variation in expression levels respectively (McCarroll et al., 2006). Similarly, for the α-synuclein gene *(SNCA)* an almost perfect correlation was found between gene dosage, mRNA and protein levels (Miller et al., 2004). However, not all gene CNVs result in corresponding changes in expression levels, for example in individuals carrying varying copy numbers of α-defensin genes, no relationship was found between total mRNA levels and gene copy number (Aldred et al., 2005). Indeed, Stranger et al. (2007) found that in the significant CNV expression associations, 5–15% of these were representative of negative correlations between copy number and gene expression. The majority of associations were, however, due to the correlation of increased expression with increased gene copy number.

Gene dosage effects are not the only consequences of copy number variation that can affect gene expression. Stranger et al. (2007) found that around half of the effects of CNVs on expression levels were due to disruption of gene coding sequences, such as removal of exons, or by affecting regulatory elements and other functional regions. CNVs that involve part of a gene could lead to the formation of variant proteins through processes of exon shuffling or generation of splice variants. Novel fusion genes may even be produced: for example, Korbel et al. (2007) identified a new gene resulting from the fusion between the coding regions of two olfactory receptor (OR) genes, *OR51A4* and *OR51A2*. The potentially damaging phenotypic effects of such genomic events is shown by a recent study, which identified a gene fusion between *PRSS1* and *PRSS2* that is thought to be the underlying cause of disease in a French family with hereditary pancreatitis (Masson et al., 2008a).

Deletions or duplications lying outside coding sequences may also affect gene expression by changing the location or efficiency of important regulatory elements, including through position effects. There is also evidence that CNVs can affect long-range gene regulation, as Stranger et al. (2007) detected several significant associations >2 Mb away from genes. Deletion of a gene repressor element could lead to increased transcriptional levels of that gene, whereas duplication of sequences downstream from a promoter could lower expression levels due to the altered position of the promoter relative to the gene (Rodriguez-Revenga et al., 2007). For example, a small duplication downstream of the proteolipid protein gene *(PLP1)* silences expression of the gene and has been shown to result in a spastic paraplegia type 2 phenotype similar to that seen with null mutations of the same gene (Lee et al., 2006). Another potential effect of intergenic CNVs could be phenotypic mosaicism due to position effect variegation (Muller, 1930; Schultz, 1936), whereby a gene may be brought into closer proximity with regions of heterochromatin, which can spread into the euchromatin region resulting in the silencing of that gene (reviewed in Kleinjan and van Heyningen, 1998). The spreading of heterochromatin is variable between cells, so that mosaicism in gene silencing results from the gene being expressed in

Table 1. Data from the TCAG Database of Genomic Variants as of November 2008. This shows the number (and % of total) of CNVs and InDels overlapping gene features – gene transcripts, genes where exons are overlapped and OMIM genes – along with the number (and % of total) of each gene feature overlapped by these genomic variants. Genes and OMIM genes are defined as described in the TCAG Database.

Feature	No. of CNVs overlapped by features (% of total CNVs = 19,792)	No. of features overlapped by CNVs (% of total features)	No. of InDels overlapped by features (% of total InDels = 11,336)	No. of features overlapped by InDels (% of total features)
Gene (transcripts) (27,212)	9,319 (47.08%)	7,500 (27.56%)	4,366 (38.51%)	2,266 (8.33%)
Gene (exons) (27,212)	6,895 (34.84%)	7,046 (25.89%)	203 (1.79%)	176 (0.65%)
OMIM genes (3,430)	1,545 (7.81%)	1,373 (40.03%)	584 (5.15%)	520 (15.16%)

some cells but not in others. Tissue-specific effects could also arise through copy number changes of specific regulatory elements or isoform-specific exons, leading to expression changes only in particular isoforms of the gene product.

CNV disease associations

The effect of CNVs on gene expression, and their potentially disruptive effects on gene structure and function, suggests that they are likely to make a considerable contribution to human diseases. Due to the relatively recent discovery of CNVs, however, and current limitations in high throughput techniques, the full extent of CNV disease associations is not yet clear. Nevertheless, from the growing number of instances where such associations have been demonstrated, it is likely that they make a substantial contribution to human disease.

Given the large number of genes which are overlapped by CNVs (Table 1), a significant proportion of biomedically relevant genes are likely to be affected. In our CNV discovery study, for example, almost half the genes intersecting variants were represented in the OMIM database, including genes associated with Mendelian diseases, genomic disorders and common diseases (de Smith et al., 2007). Indeed, many gene copy number changes contribute directly to monogenic diseases. In recessive diseases, hemizygosity due to deletion of a gene, or part of a gene, could unmask a mutation on the other gene copy. Conversely, duplication of a healthy gene copy on one chromosome could theoretically mask the effects of a disease-causing mutation in the gene on the other chromosome, thus rescuing the phenotype. Indeed, it has been predicted that a proportion of the variable penetrance shown by many dominant genetic disorders could be explained by CNVs (Beckmann et al., 2007).

Autosomal dominant early-onset Alzheimer's disease (ADEOAD) is known to be caused by missense mutations in *APP* genes on chromosome 21, but duplication of the *APP* locus has also been found in patients with this disorder (Rovelet-Lecrux et al., 2006). This copy number gain is thought to lead to an abundance of amyloid deposits in the brain. Similarly, triplication of the *SNCA* gene, which leads to a profusion of Lewy bodies, has been associated in patients with autosomal dominant Parkinson disease (Singleton et al., 2003). Neither of these genes are known to be overlapped by variants in healthy individuals, and these copy number gains are, therefore, thought to underlie disease: such features are termed copy number mutations (CNMs). Loss or gain of exonic material could also result in missense mutations or frameshifts: indeed Duchenne muscular dystrophy (DMD) is usually caused by de novo deletions and duplications resulting in frameshifts. If these were discovered for the first time today, they would be known as CNMs. Importantly, most sequencing strategies for identification of mutations causing monogenic disease would miss such variants, which might therefore account for a significant proportion of 'missing' mutations, and complicate genetic counselling.

As well as monogenic diseases, variations in copy number of large genomic regions are the underlying cause of many genomic disorders, and such aberrations can affect the copy numbers of multiple genes. In some cases, the dosage changes of many genes are thought to contribute to phenotype, for instance with the ~1.6-Mb deletion at chromosome 7q11.23 that leads to Williams-Beuren syndrome (Peoples et al., 2000). In other disorders, such as CMT1A and SMS, dosage changes in only a single gene are thought to underlie disease (Roa et al., 1991; Slager et al., 2003). Interpretation of the data derived from clinical investigations into the underlying cause of suspected genomic disorders is frequently complicated by the presence of CNVs. We do not yet have a full appreciation of the normal spectrum of copy number variation, particularly in non-HapMap (The International HapMap Consortium, 2003) population groups, and so it may be very difficult to distinguish between benign CNVs and disease-causing variants. Further complications may be caused by CNVs that overlap larger aberrations, so that different combinations of variants mitigate or worsen phenotypes. This could help to explain, for example, the different manifestations of phenotypes seen in trisomy 21 patients: for example, ~40% patients have congenital heart defects (Freeman et al., 1998) and ~1% develop leukaemia (Zipursky et al., 1992).

The situation with regard to complex disease is even less straightforward. Recent advances in complex disease analysis, using genome-wide SNP association approaches, have highlighted new genes and potential pathogenic pathways (Frayling et al., 2007; Sladek et al., 2007), but the SNP markers found still do not account for the estimated heritability of these disorders. It is likely, therefore, that other genetic factors contribute to common complex disorders, including

rare variants, epigenetic modifications and copy number variation. Although CNVs overlap many biologically important genes, several of which were already associated with disease, this is not of itself evidence that they play a role in disease. Therefore, studies are now underway to identify specific disease-associated CNVs. This has already led to reported disease associations with variants that are relatively common in apparently healthy populations. A substantial proportion of these associations has been found with genes involved in the immune system and in defence against disease. The first example of this was the discovery that low copy number of a frequent CNV including the *FCGR3B* gene is associated with glomerulonephritis in rats and humans (Aitman et al., 2006). This gene plays a key role in regulation of inflammatory and immune responses, especially in the tethering of neutrophils to immune complexes and the clearance of these complexes, and has since been associated with systemic lupus erythematosus (SLE) and other systemic autoimmune disorders, such as ANCA-associated vasculitis (Fanciulli et al., 2007).

As described earlier, variations in copy number of another gene, complement component *C4,* have long been associated with SLE. This gene also plays a role in the clearance of immune complexes, as well as in the activation of complement pathways that act against invading microbes, and in the reduction of the threshold for B lymphocyte activation. It has recently been confirmed that low copy numbers of *C4* increase risk of SLE, while high copy numbers of this gene have a protective role against disease (Yang et al., 2007).

Another gene involved in defence against disease is *CCL3L1,* which is also incorporated within an extremely common and highly polymorphic CNV. This gene has been implicated in susceptibility to and disease progression in HIV, since CCL3L1 is the most effective ligand for CC chemokine receptor 5 (CCR5), which is the major HIV co-receptor, thus it is an important HIV-suppressive chemokine (Menten et al., 2002). Possession of low *CCL3L1* copy number is a major risk factor for HIV, associated with higher viral loads and increased subsequent loss of T-cells (Gonzalez et al., 2005).

Possibly the most intriguing example of a disease-associated CNV that overlaps genes related to the immune system is that of the beta defensin genes, which are candidates for variation in susceptibility to autoimmune and inflammatory disorders, due to their anti-microbial and pro-inflammatory roles. These genes vary greatly in copy number both in humans (Armour et al., 2007) and also macaques, suggesting this is an ancient hotspot for copy number variation (Lee et al., 2008). A large repeat unit at chromosome 8p23.1, including *DEFB4, SPAG11, DEFB103, DEFB104* and *DEFB105* amongst others, is highly variable in copy number, with individuals carrying between 2 and 12 copies per diploid genome. High copy numbers of this unit increase susceptibility to the common inflammatory skin disease psoriasis, consistent with an exaggerated immune response leading to an inflammatory disease (Hollox et al., 2008). Conversely, low copy numbers of the *DEFB4* gene have been associated with colonic Crohn's disease, which is thought to be due to a weakening of the antibacterial barrier in the colonic mucosa due to relative deficiency of beta-defensins (Fellermann et al., 2006). This is the first example of a common CNV that in low copy numbers can lead to one disease, and in high copy numbers may lead to another phenotypically distinct disease.

Variants have also been shown to overlap cancer-relevant genes, for example a 630-kb deletion region on chromosome 3p21.3 deleted in lung cancer, incorporating three tumour suppressor genes *TUSC2, TUSC4* and *NAT6,* was found to overlap a relatively common deletion CNV in an ostensibly healthy population (Wong et al., 2007). Many other oncogenes and tumour-suppressors are affected by copy number variation, including *LPP, MLLT3, MEN1, APC, VAV2, TNFRSF25, BCAS1* and *HIC2* (Conrad et al., 2006; de Smith et al., 2007; Wong et al., 2007). Studies to determine their consequences for cancer susceptibility are underway, but we already have at least one example of a significant association. A *UGT2B17* deletion variant, found in around 11–12% of healthy subjects, exhibits a significant association with risk of prostate cancer in Caucasians. Increased levels of serum testosterone and other androgens are a risk factor for prostate cancer, and it is thought that deletion of this gene, which is involved in androgen metabolism, may lead to increased serum androgen levels (Park et al., 2006). One interesting question, which has yet to be investigated, is whether there is any relationship between the inherited CNVs in cancer-relevant genomic regions and the incidence of the various genomic losses and gains that occur during cancer progression. Since these changes have considerable prognostic significance, such a relationship may have important consequences for early decisions on therapeutic management.

As described, genes involved in brain development are enriched in CNVs (de Smith et al., 2007), and a proportion of these variants may, therefore, contribute to susceptibility to neurological and psychiatric disorders, such as bipolar disorder (BD) and schizophrenia. Indeed, some of the most important BD and schizophrenia candidate genes, such as *PDE4β, CHRNA7* and *DISC1,* are overlapped by known variants. In a cohort of BD patients, for example, a significant increase was found for the presence of a known CNV overlapping the *GSK3β* gene compared with healthy controls (Lachman et al., 2007). This is a credible candidate gene for BD as it is involved in neuronal cell development, and transgenic mice with *GSK3β* overexpression have been shown to mimic acts of clinical mania, with increased locomotor activity and acoustic startle response (Prickaerts et al., 2006). The *GSK3β* variant has only been documented in two healthy control samples thus far, so it could be described as a rare CNV. Some studies aimed at identification of variants associated with particular diseases, however, have uncovered gene copy number variants that are present only in patients with those diseases, and not in the general population. These variants should, therefore, properly be termed copy number mutations (CNMs), as they are not present at appreciable frequency (>1%) in the general population and

may be the direct cause of disease, rather than acting as susceptibility loci.

In addition to common CNVs playing a role in neurological disorders, a number of de novo CNMs have also been associated with such diseases. Variants incorporating three brain-expressed genes involved in glutamate signalling, *GLUR7, AKAP5* and *CACNG2,* were found only in patients with schizophrenia in one study (Wilson et al., 2006). It is speculated that these genes are differentially expressed during early human embryonic development, and that the development of a normal central nervous system depends on complex regulation of these genes (Wilson et al., 2006). A recent study also indicates that rare de novo variants with high penetrance may underlie schizophrenia in some cases. Xu et al. (2008) found a 10% frequency of novel variants in patients with sporadic schizophrenia, which was eight times higher than in controls. The number of genes overlapped by these variants was relatively small, but GO analysis showed the most enriched categories to be pathways associated with neuronal development (Xu et al., 2008). These results are mirrored by another recent report of an association with rare variants affecting neurodevelopmental genes in cases of schizophrenia (Walsh et al., 2008).

The findings of these studies of schizophrenia have parallels in recent investigations into the genetics of autism, as associations of de novo variants have also been determined with this neurodevelopmental disorder (Sebat et al., 2007; Marshall et al., 2008). Sebat et al. (2007), for example, found the frequency of spontaneous mutations to be 10% in sporadic cases of autism, compared with only 1% in unaffected controls, and a number of genes, including *SHANK3, NLGN4* and *NRXN1,* have been implicated in the aetiology of autism through studies of copy number variation. Interpretation of these data is somewhat complicated, however, by the difficulties inherent in (a) proving that a particular copy number variation is actually de novo (most current methodologies suffer from high false negative, as well as false positive results) and (b) establishing the 'normal' rate of generation of de novo copy number changes.

The distinction between common gene copy number variants and rare CNMs is not always a clear one. A particular example of this are the common copy number gains and losses of the region on chromosome 7 that incorporates the *PRSS1* and *PRSS2* genes, missense mutations in which are known to cause hereditary pancreatitis. This is an autodigestive disease, whereby an activation cascade of pancreatic digestive enzymes is caused by the premature activation of trypsin (Le Marechal et al., 2006). CNVs overlapping the two genes have been found in healthy individuals, but a triplication of this region is associated with disease. It is thought that the increased dosage of the triplicated *PRSS1* and *PRSS2* genes may disrupt the balance between activation and inhibition of trypsin within the pancreas (Le Marechal et al., 2006), which suggests that a duplication of the same locus on both chromosomes could have the same effect. The triplicated region is, therefore, a copy number mutation that underlies disease, whereas a single copy gain on one chromosome (i.e. duplication CNV) could be described as a pre-mutation CNV for the adverse phenotype. The situation is further complicated, however, by the recent identification of a duplication of this locus in four patients with hereditary pancreatitis, which was not found in controls (Masson et al., 2008b): it is possible, therefore, that duplication of this locus may, in some individuals, cause disease but in others act as a benign CNV.

The same disease has also recently been shown to result from another genetic mechanism altogether. As mentioned earlier, Masson et al. (2008a) identified a hybrid *PRSS2/PRSS1* gene, which they describe as having a 'double gain-of-function' effect, with both qualitative and quantitative consequences, in a French family with hereditary pancreatitis. This fusion gene essentially consists of a duplication of half of each gene, thus acting as a 'quantitative' CNM, in addition to a 'qualitative' missense mutation, which has resulted in a highly penetrant phenotype in this family. This appears to be a novel genotype-phenotype relationship.

Very few investigations have so far been carried out to examine the dual effects of CNVs and SNPs. One example, however, is analysis of the complement factor H and membrane cofactor (*CFH)* gene, which contains an amino acid variant that predisposes to age-related macular degeneration (Klein et al., 2005). This gene is contained within a CNV region; thus it is possible that variations in copy number of this gene, or its surrounding genomic region, could modify the risk of disease. In support of this, a haplotype carrying deletions of the nearby *CFHR1* and *CFHR3* genes has been shown to be protective against the disease (Hughes et al., 2006). This example highlights the need to evaluate the contribution of both types of variant to complex phenotypes and disease, as cases like these may only be the tip of the iceberg (Masson et al., 2008a).

It is likely, therefore, that variations in gene copy number play an important role in human health, with some common gene CNVs increasing susceptibility to certain complex diseases, whereas so called CNMs are the direct cause of Mendelian-like diseases. Current estimates indicate that copy number variants overlap more of the human genome than SNPs (current estimates of ~29% bp cf. 0.4% (http://projects.tcag.ca/variation)); however, it remains to be proven whether CNVs are the main source of inter-individual differences in physiological phenotype, as well as overall fitness and disease susceptibility. The interaction between functionally-relevant SNPs and CNVs adds a further level of complexity to investigations.

Evolutionary implications of CNVs

In addition to the possible effects on phenotype, the potential for gene or domain duplication as a result of copy number variation suggests a possible role in human evolution; while conserved genomic sequences represent loci of central biological importance, regions of genomic variability may also be important in evolving species, as it is genetic variation upon which natural selection acts. Copy number variation in the form of gene duplication, for example,

has long been known as a source of evolutionary change (Bridges, 1935; Muller, 1936), generating protein diversity, increases in gene dosage and evolution of new functions (Ohno et al., 1968), and is consequently thought to be one of the primary mechanisms in the proliferation of primate species (Dumas et al., 2007).

GO analysis of regions of the genome shown to be subject to recent selection has revealed several enriched categories, including chemosensory perception and olfaction, acquired and innate immunity, gametogenesis, spermatogenesis, fertilisation, and vitamin transport (Voight et al., 2006). As already discussed, many of these categories are also enriched for copy number variation, suggesting a relationship between evolution and CNVs. For example, genes involved in chemosensation and immune response are over-represented amongst those overlapped by CNVs: these genes are important in adaptation to novel environmental niches (Nguyen et al., 2006). An enrichment was also found for genes involved in fertility and reproduction (de Smith et al., 2007), which are thought to be subject to rapid adaptive evolution in primates due to sexual competition and defence against pathogens (Voight et al., 2006).

CNVs may, thus, provide a substantial proportion of the genetic variability which is the substrate for natural selection, increasing genetic plasticity so that organisms can evolve more quickly in response to novel external pressures, and thereby playing an important role in their evolutionary fitness and adaptability (discussed in Feuk et al., 2006). An investigation into copy number variation in the chimpanzee genome revealed many variants that overlap with human CNVs, and found similar enrichment for immunity and environmental response-related genes, which suggests these variants could be retained in the genomes of both species by natural selection (Perry et al., 2006). Alternatively, because they apparently lack immediate phenotypic impact, CNVs in these genes may accumulate to a greater extent than in critical genes, in which variants may have more adverse phenotypic effects. An additional possibility is that the same regions of human and chimpanzee genomes are vulnerable to the mechanisms which lead to CNV formation, so that CNVs tend to arise at similar locations.

There is evidence of positive selection acting on some CNV genes, in that there is overlap between genes identified in studies of positive selection and copy number variation. Indeed, a positive correlation was recently determined between copy numbers of the salivary amylase gene *(AMY1)* and protein expression levels, and mean gene copy numbers were found to be greater in 'high-starch' populations, compared to populations with 'low-starch' diets that consist mainly of meat, fruit, honey and milk (Perry et al., 2007). Given the relatively recent change to high-starch diets, it is suggested that positive selection has acted on pre-existing CNVs to increase *AMY1* copy number in populations which have adopted a high-starch diet but that, in the absence of such selection, copy number has evolved neutrally.

Nevertheless, it is clear that, in many instances, CNVs have the potential to be deleterious. An important question, therefore, is why there is so much copy number variation in the genome when this has potentially adverse phenotypic effects. For instance, why is there so much variation in the copy number of *DEFB4* when there are deleterious effects associated with both too few and too many copies of the gene? It seems most likely that CNVs are an unavoidable consequence of errors in genetic recombination and repair, which may particularly afflict regions containing many repetitive elements or in which there are tandem duplications of regions of the genome. Where CNVs affect important genes they are presumably subject to strongly negative (purifying) selection; however, phenotypically neutral CNVs may persist and accumulate, so that when a population is subsequently exposed to a change in the environment, some of this accumulated genetic variation may include CNVs that fortuitously impart a selective advantage, which thus becomes a substrate for natural selection.

Future perspectives

Recent CNV studies have revealed a previously cryptic level of genetic variation at a much higher level than SNPs, and have also improved our knowledge of disease aetiology, which may ultimately contribute towards the treatment of certain diseases. The full extent of copy number variation in the human genome remains to be elucidated, but by using technologies with ever-increasing resolution, and studying a larger number of samples, including individuals of a wider range of ethnicities, this situation should be ameliorated in the near future. Improved mapping capacity has already demonstrated that the size of many CNVs in the TCAG database of genomic variants is inflated (de Smith et al., 2007; Kidd et al., 2008; Perry et al., 2008). Using high-resolution oligonucleotide arrays, Perry et al. (2008) reduced the likely amount of CNV sequence by >50% for 876 loci, and estimated that 88% (1020/1153) of loci they investigated were smaller than documented. The number of known small-sized variants, however, is likely to increase with the advent of whole genome sequencing studies.

The high level of copy number variation uncovered so far, and the large number of genes involved, suggests that these variants are likely to contribute not only to specific single-gene disorders but also to common complex diseases, such as obesity and type II diabetes, which are thought to be caused by interactions between multiple genetic loci and the environment. Copy number variants are also likely to play a role in susceptibility to certain cancers in at least two ways: inherited variants may decrease expression of tumour suppressor genes, or may increase expression of oncogenes to a level that may subsequently require only a single additional somatic mutation to lead to cancer. Additionally, germline and somatic variants may cause genomic fragility that could increase the likelihood of an individual developing further genomic rearrangements that may result in tumour development, or affect disease progression.

At the time of writing, however, only a handful of studies have associated specific CNVs with human diseases, and as yet there have been no intensive genome-wide surveys for

variants associated with common diseases. CNVs may act in conjunction with single nucleotide mutations, and possibly SNPs, to produce an adverse phenotype so that future studies need to take an integrated approach, taking into account genetic variants at both the copy number and nucleotide level, along with environmental factors. This must now be a priority, and will require the development of high-throughput methodologies for replication of phase 1 genome-wide association data in large sample cohorts, rapidly establishing absolute copy number of particular CNVs in several thousands of individuals. It may be that some common variants initially appear benign, but in combination with other genetic and environmental factors, have deleterious effects resulting in disease.

Determination of CNV disease associations may also enable earlier diagnosis based on an individual's genetic constitution, or may contribute to elucidation of pathogenic pathways, allowing identification of disease sub-types amenable to particular therapeutic approaches. Similarly, by testing for disease susceptibility loci, individuals may be able to determine their individual risk for certain diseases, enabling them to alter their lifestyle based on this information. Any CNV that affects clinical phenotype, disease prognosis or response to particular therapies will be of particular interest in our movement towards the era of personalised medicine.

References

Agnello V, De Bracco MM, Kunkel HG: Hereditary c2 deficiency with some manifestations of systemic lupus erythematosus. J Immunol 108: 837–840 (1972).

Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, et al: Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans. Nature 439:851–855 (2006).

Aldred PM, Hollox EJ, Armour JA: Copy number polymorphism and expression level variation of the human alpha-defensin genes *DEFA1* and *DEFA3*. Hum Mol Genet 14:2045–2052 (2005).

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES: An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407: 513–516 (2000).

Armour JA, Palla R, Zeeuwen PL, den Heijer M, Schalkwijk J, Hollox EJ: Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. Nucleic Acids Res 35:e19 (2007).

Beckmann JS, Estivill X, Antonarakis SE: Copy number variants and genetic traits: Closer to the resolution of phenotypic to genotypic variability. Nat Rev Genet 8:639–646 (2007).

Brewer C, Holloway S, Zawalnyski P, Schinzel A, FitzPatrick D: A chromosomal duplication map of malformations: Regions of suspected haplo- and triplolethality – and tolerance of segmental aneuploidy – in humans. Am J Hum Genet 64: 1702–1708 (1999).

Bridges C: Salivary chromosome maps. J Hered 26: 60–64 (1935).

Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, et al: Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet 82:763–771 (2008).

Campbell RD, Carroll MC, Porter RR: The molecular genetics of components of complement. Adv Immunol 38:203–244 (1986).

Charlesworth B, Charlesworth D: The degeneration of Y chromosomes. Phil Trans R Soc London 355:1563–1572 (2000).

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

Cooper GM, Nickerson DA, Eichler EE: Mutational and selective effects on copy-number variants in the human genome. Nat Genet 39:S22–29 (2007).

de Smith AJ, Tsalenko A, Sampas N, Scheffer A, Yamada NA, et al: Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: Implications for association studies of complex diseases. Hum Mol Genet 16:2783–2794 (2007).

de Smith AJ, Walters RG, Coin LJM, Steinfeld I, Yakhini Z, et al: Small deletion variants have stable breakpoints commonly associated with Alu elements. PLoS ONE 3:e3104 (2008).

Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, Sikela JM: Gene copy number variation spanning 60 million years of human and primate evolution. Genome Res 17: 1266–1277 (2007).

Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, et al: *FCGR3B* copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 39:721–723 (2007).

Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, et al: A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. Am J Hum Genet 79:439–448 (2006).

Feuk L, Carson AR, Scherer SW: Structural variation in the human genome. Nat Rev Genet 7: 85–97 (2006).

Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, et al: A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316:889–894 (2007).

Freeman SB, Taft LF, Dooley KJ, Allran K, Sherman SL, et al: Population-based study of congenital heart defects in Down syndrome. Am J Med Genet 80:213–217 (1998).

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, et al: The influence of *CCL3L1* gene-containing segmental duplications on HIV1/AIDS susceptibility. Science 307:1434–1440 (2005).

Grimwood J, Gordon LA, Olsen A, Terry A, Schmutz J, et al: The DNA sequence and biology of human chromosome 19. Nature 428:529–535 (2004).

Guryev V, Saar K, Adamovic T, Verheul M, van Heesch SA, et al: Distribution and functional impact of DNA copy number variation in the rat. Nat Genet 40:538–545 (2008).

Hauptmann G, Grosshans E, Heid E: Lupus erythematosus syndrome and complete deficiency of the fourth component of complement. Boll Ist Sieroter Milan 53:suppl:228 (1974).

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

Hollox EJ, Armour JA, Barber JC: Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster. Am J Hum Genet 73: 591–600 (2003).

Hollox EJ, Huffmeier U, Zeeuwen PL, Palla R, Lascorz J, et al: Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet 40:23–25 (2008).

Hughes AE, Orr N, Esfandiary H, Diaz-Torres M, Goodship T, Chakravarthy U: A common CFH haplotype, with deletion of *CFHR1* and *CFHR3*, is associated with lower risk of age-related macular degeneration. Nat Genet 38:1173–1177 (2006).

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

International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. Nature 409:860–921 (2001).

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al: Mapping and sequencing of structural variation from eight human genomes. Nature 453:56–64 (2008).

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al: Complement factor H polymorphism in age-related macular degeneration. Science 308: 385–389 (2005).

Kleinjan DJ, van Heyningen V: Position effect in human genetic disease. Hum Mol Genet 7: 1611–1618 (1998).

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, et al: Paired-end mapping reveals extensive structural variation in the human genome. Science 318:420–426 (2007).

Lachman HM, Pedrosa E, Petruolo OA, Cockerham M, Papolos A, et al: Increase in *GSK3beta* gene copy number variation in bipolar disorder. Am J Med Genet B Neuropsychiatr Genet 144:259–265 (2007).

Le Marechal C, Masson E, Chen JM, Morel F, Ruszniewski P, Levy P, Ferec C: Hereditary pancreatitis caused by triplication of the trypsinogen locus. Nat Genet 38:1372–1374 (2006).

Lee JA, Madrid RE, Sperle K, Ritterson CM, Hobson GM, et al: Spastic paraplegia type 2 associated with axonal neuropathy and apparent *PLP1* position effect. Annals Neurol 59:398–403 (2006).
Lee JA, Carvalho CM, Lupski JR: A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131:1235–1247 (2007).
Lee AS, Gutierrez-Arcelus M, Perry GH, Vallender EJ, Johnson WE, et al: Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. Hum Mol Genet 17:1127–1136 (2008).
Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, et al: Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet 82: 477–488 (2008).
Masson E, Le Marechal C, Delcenserie R, Chen JM, Ferec C: Hereditary pancreatitis caused by a double gain-of-function trypsinogen mutation. Hum Genet 123:521–529 (2008a).
Masson E, Le Marechal C, Chandak GR, Lamoril J, Bezieau S, et al: Trypsinogen copy number mutations in patients with idiopathic chronic pancreatitis. Clin Gastroenterol Hepatol 6:82–88 (2008b).
McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, et al: Common deletion polymorphisms in the human genome. Nat Genet 38: 86–92 (2006).
McLellan RA, Oscarson M, Alexandrie AK, Seidegard J, Evans DA, et al: Characterization of a human glutathione *S*-transferase mu cluster containing a duplicated *GSTM1* gene that causes ultrarapid enzyme activity. Mol Pharmacol 52:958–965 (1997).
Menten P, Wuyts A, Van Damme J: Macrophage inflammatory protein-1. Cytokine Growth Fact Rev 13:455–481 (2002).
Miller DW, Hague SM, Clarimon J, Baptista M, Gwinn-Hardy K, et al: Alpha-synuclein in blood and brain from familial Parkinson disease with *SNCA* locus triplication. Neurology 62:1835–1838 (2004).
Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, et al: An initial map of insertion and deletion (indel) variation in the human genome. Genome Res 16:1182–1190 (2006).
Muller HJ: Types of visible variations induced by X-rays in *Drosophila*. J Genet 22:299–335 (1930).
Muller HJ: Bar duplication. Science 83:528–530 (1936).
Nguyen DQ, Webber C, Ponting CP: Bias of selection on human copy-number variants. PLoS Genet 2:e20 (2006).
Ohno S, Wolf U, Atkin NB: Evolution from fish to mammals by gene duplication. Hereditas 59: 169–187 (1968).
Park J, Chen L, Ratnashinge L, Sellers TA, Tanner JP, et al: Deletion polymorphism of UDP-glucuronosyltransferase 2B17 and risk of prostate cancer in African American and Caucasian men. Cancer Epidemiol Biomarkers Prev 15: 1473–1478 (2006).
Peoples R, Franke Y, Wang YK, Perez-Jurado L, Paperna T, et al: A physical map, including a BAC/PAC clone contig, of the Williams-Beuren syndrome–deletion region at 7q11.23. Am J Hum Genet 66:47–68 (2000).
Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, et al: Hotspots for copy number variation in chimpanzees and humans. Proc Natl Acad Sci USA 103:8006–8011 (2006).
Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, et al: Diet and the evolution of human amylase gene copy number variation. Nat Genet 39: 1256–1260 (2007).
Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, et al: The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet 82:685–695 (2008).
Pinkel D, Segraves R, Sudar D, Clark S, Poole I, et al: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20:207–211 (1998).
Piotrowski A, Bruder CE, Andersson R, de Stahl TD, Menzel U, et al: Somatic mosaicism for copy number variation in differentiated human tissues. Hum Mutat 29:1118–1124 (2008).
Prickaerts J, Moechars D, Cryns K, Lenaerts I, van Craenendonck H, et al: Transgenic mice overexpressing glycogen synthase kinase 3beta: A putative model of hyperactivity and mania. J Neurosci 26:9022–9029 (2006).
Rao Y, Hoffmann E, Zia M, Bodin L, Zeman M, et al: Duplications and defects in the *CYP2A6* gene: Identification, genotyping, and in vivo effects on smoking. Mol Pharmacol 58:747–755 (2000).
Rebbeck TR: Molecular epidemiology of the human glutathione S-transferase genotypes *GSTM1* and *GSTT1* in cancer susceptibility. Cancer Epidemiol Biomarkers Prev 6:733–743 (1997).
Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al: Global variation in copy number in the human genome. Nature 444:444–454 (2006).
Roa BB, Garcia CA, Lupski JR: Charcot-Marie-Tooth disease type 1A: Molecular mechanisms of gene dosage and point mutation underlying a common inherited peripheral neuropathy. Int J Neurol 25–26:97–107 (1991).
Rodriguez-Revenga L, Mila M, Rosenberg C, Lamb A, Lee C: Structural variation in the human genome: The impact of copy number variants on clinical diagnosis. Genet Med 9:600–606 (2007).
Rovelet-Lecrux A, Hannequin D, Raux G, Le Meur N, Laquerriere A, et al: *APP* locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. Nat Genet 38:24–26 (2006).
Sachse C, Brockmoller J, Bauer S, Roots I: Cytochrome P450 2D6 variants in a Caucasian population: Allele frequencies and phenotypic consequences. Am J Hum Genet 60:284–295 (1997).
Schaeffeler E, Schwab M, Eichelbaum M, Zanger UM: *CYP2D6* genotyping strategy based on gene copy number determination by TaqMan real-time PCR. Hum Mutat 22:476–485 (2003).
Schultz J: Variegation in *Drosophila* and the inert chromosome regions. Proc Natl Acad Sci USA 22:27–33 (1936).
Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al: Large-scale copy number polymorphism in the human genome. Science 305:525–528 (2004).
Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al: Strong association of de novo copy number mutations with autism. Science 316: 445–449 (2007).
Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, et al: Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77:78–88 (2005).
Singleton AB, Farrer M, Johnson J, Singleton A, Hague S, et al: Alpha-synuclein locus triplication causes Parkinson's disease. Science 302: 841 (2003).
Sladek R, Rocheleau G, Rung J, Dina C, Shen L, et al: A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445: 881–885 (2007).
Slager RE, Newton TL, Vlangos CN, Finucane B, Elsea SH: Mutations in *RAI1* associated with Smith-Magenis syndrome. Nat Genet 33:466–468 (2003).
Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al: Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848–853 (2007).
The International HapMap Consortium: The International HapMap Project. Nature 426:789–796 (2003).
Trask BJ, Friedman C, Martin-Gallardo A, Rowen L, Akinbami C, et al: Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. Hum Mol Genet 7:13–26 (1998).
Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al: Fine-scale structural variation of the human genome. Nat Genet 37:727–732 (2005).
Voight BF, Kudaravalli S, Wen X, Pritchard JK: A map of recent positive selection in the human genome. PLoS Biol 4:e72 (2006).
Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, et al: Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320:539–543 (2008).
Wilson GM, Flibotte S, Chopra V, Melnyk BL, Honer WG, Holt RA: DNA copy-number analysis in bipolar disorder and schizophrenia reveals aberrations in genes involved in glutamate signaling. Hum Mol Genet 15:743–749 (2006).
Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, et al: A comprehensive analysis of common copy-number variations in the human genome. Am J Hum Genet 80:91–104 (2007).
Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, Karayiorgou M: Strong association of de novo copy number mutations with sporadic schizophrenia. Nat Genet 40:880–885 (2008).
Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, et al: Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): Low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. Am J Hum Genet 80:1037–1054 (2007).
Zanger UM, Fischer J, Raimundo S, Stuven T, Evert BO, et al: Comprehensive analysis of the genetic factors determining expression and function of hepatic CYP2D6. Pharmacogenetics 11:573–585 (2001).
Zipursky A, Poon A, Doyle J: Leukemia in Down syndrome: A review. Pediatr Hematol Oncol 9: 139–149 (1992).