OPINION

Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability

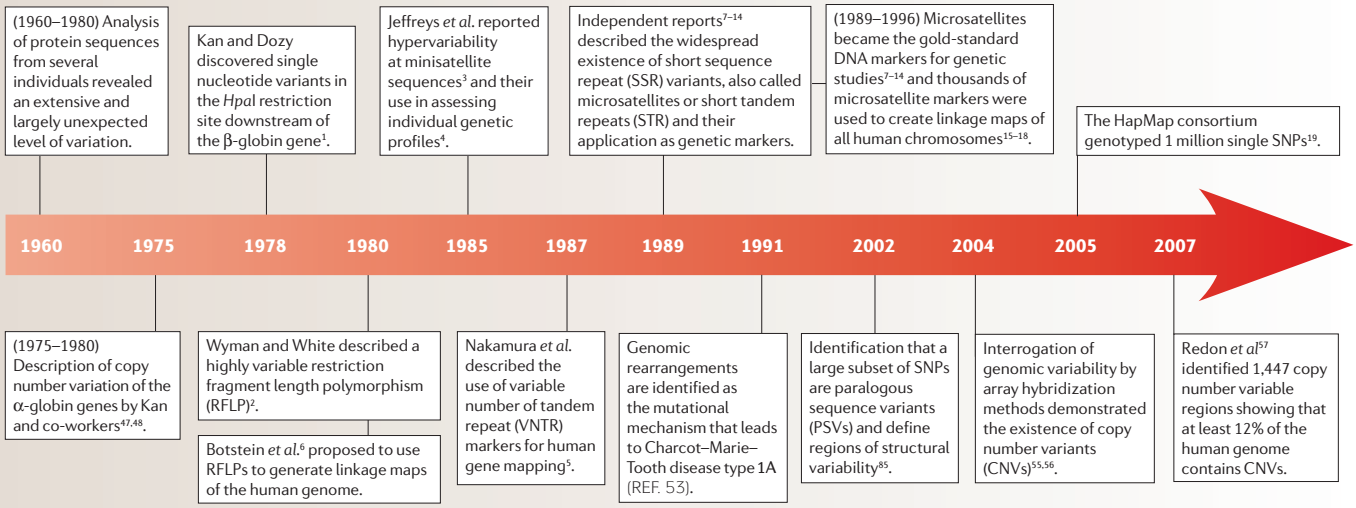
Jacques S. Beckmann, Xavier Estivill and Stylianos E. Antonarakis

Abstract | A considerable and unanticipated plasticity of the human genome, manifested as inter-individual copy number variation, has been discovered. These structural changes constitute a major source of inter-individual genetic variation that could explain variable penetrance of inherited (Mendelian and polygenic) diseases and variation in the phenotypic expression of aneuploidies and sporadic traits, and might represent a major factor in the aetiology of complex, multifactorial traits. For these reasons, an effort should be made to discover all common and rare copy number variants (CNVs) in the human population. This will also enable systematic exploration of both SNPs and CNVs in association studies to identify the genomic contributors to the common disorders and complex traits.

The availability of genetic markers has led to extraordinary progress in human genetics in the past 25 years, including the elucidation of the molecular genetic basis of many Mendelian disorders or traits. Several landmarks mark this progress. Whereas studies of protein polymorphisms in the 1960s predicted a limited amount of genomic variation, analysis of DNA sequences from several individuals revealed an extensive and largely unexpected level of variation (TIMELINE). In 1978, Kan and Dozy were the first to discover single nucleotide variants in the *Hpa*I restriction site that lies downstream of the β-globin gene (*HBB*)[1]. In 1980, Wyman and White described a highly variable restriction fragment length polymorphism (RFLP)[2] and, in 1985, Jeffreys *et al.* reported the minisatellites[3,4], soon followed by the description of variable number of tandem repeats (VNTRs)[5]. These advances paved the way to a systematic genetic exploration of the human genome[6]. As early as 1989, a number of independent reports[7–14] described the widespread existence of short sequence repeat (SSR) variants, also called microsatellites or short tandem repeats (STR). In the 1990s, thousands of such markers were used to create linkage maps of all human chromosomes[15–18]. These multiallelic markers were used extensively by investigators worldwide to map disease-causing mutations in more than 2,000 genes. In 2005, the International HapMap Project genotyped one million SNPs[19] and, in a second phase, about 4 million of the nearly 12 million SNPs that were deposited in public databases (see the NCBI Single Nucleotide Polymorphism database). These variants constitute the major source of inter-individual genetic and phenotypic variation. SNPs and SSRs have found additional uses in forensic studies, discovery of loss of heterozygosity in cancers[20], population genetic studies[21,22], discovery of uniparental disomies[23], elucidation of the origin of aneuploidies[24], diagnostic studies, evolutionary analyses and association studies for polygenic complex traits. This list is not exhaustive, and the literature is continuously expanding with elegant experimental uses of genetic variation. To date, SNPs are the variant type of choice for association studies in common diseases and complex traits. The results of this massive application to case–control studies are just around the corner[25], as testified by the recent impressive publications on age-related macular degeneration[26–29], diabetes[30–37], obesity[38,39], cardiovascular diseases[40,41], prostate cancer[42,43] and breast cancer[44–46].

In addition to these by now 'conventional' genetic markers, it has been known since the 1980s that the human genome also contains another abundant source of polymorphism, one that involves deletions, insertions, duplications and complex rearrangements of genomic

Timeline | **Landmarks in the study of human genetic variation**



regions of 1 kb in length or larger. For example, Goossens *et al.* showed that in a fraction of cases the α-globin loci are triplicated[47]; normally, they are present in two copies per haploid genome, although in some instances they carry deletions[48]. Subsequently, the number of X-linked pigment genes was found to vary among individuals[49,50]. The Rhesus blood group gene *RHD* is another familiar example of a long-established deletion polymorphism[51]. Importantly, the existence of intra-chromosomal duplicons[52] (regions of the genome with identical sequence) predicted the possibility of genomic rearrangements as a result of non-allelic homologous recombination (NAHR) between these regions. In 1991, this course of events was identified as the mutational mechanism that leads to Charcot–Marie–Tooth neuropathy type 1A (REF. 53). Numerous reports describing genomic rearrangements have since followed[54]. In 2004, the interrogation of genomic variability by array hybridization methods clearly demonstrated the existence of copy number variants[55,56]. Intense analysis of this type of genomic variability followed, and the current conservative estimate from studies in a few hundred individuals is that at least 10% of the genome is subject to copy number variation[57,58] (see the Database of Genomic Variants and the UCSC Genome Bioinformatics web site).

This submicroscopic structural type of variation has been termed copy number variation (CNV) or copy number polymorphism (CNP). Although a typical SNP affects only one single nucleotide pair, their genomic abundance (over 10 million) makes them the most frequent source of polymorphic changes. By contrast, CNVs are far less numerous but can affect from one kilobase to several megabases of DNA per event, adding up to a significant fraction of the genome[57–59].

The discovery of extensive copy number variation in the genomes of normal individuals provides new hypotheses to account for the phenotypic variability among inherited (Mendelian and polygenic) disorders and aneuploid syndromes. Many aspects of the importance of the considerable plasticity of the human genome have previously been discussed[57,58,60–63], but their impact on the myriad of phenotypic traits and genetic diseases remains to be elucidated[54,64–67]. Here we provide examples to show how CNVs might account for phenotypic variability in genetic disease.

Dominant traits and penetrance

Penetrance, the fraction of individuals with a particular genotype that show the associated phenotype, is an important aspect of all genetic disorders. Penetrance is a complicated issue in genetic counselling, clinical practice, mapping of disease loci, positional cloning and association studies for complex, multifactorial and polygenic disorders. Remarkably, many dominant disorders have variable penetrance, such as tuberous sclerosis, neurofibromatosis type 1 or breast cancer due to mutations in *BRCA1* or *BRCA2*. We propose that CNV could be one determinant to explain reduced penetrance of some disease-causing mutations. Consider the pedigree in FIG. 1a: individual II-4 manifests the disease phenotype because she has only one copy of the normal allele. By contrast, although individual II-2 receives the mutant allele from his affected mother, he is not affected, because he inherited a duplicated (compensatory) allele from his father (the father is healthy because a higher copy number is not harmful). So, for a dominant loss-of-function mutation, a CNV gain in *trans* (on the non-carrier chromosome) could rescue the phenotype. Note that individual II-2 can transmit the mutant allele to his offspring, which could — depending on the status of its other allele — manifest the phenotype. In the above example, we assumed that extra copies of a CNV result in increased gene expression. This is not always the case — in ~10% of cases, negative correlations between CNV and levels of gene expression were reported[68]. Consider for instance, a loss-of-function mutation in the context of an expanding CNV that reduces gene expression. In this case, individual II-2 could be affected whereas II-4 would be an asymptomatic carrier.

The mechanism described above might not apply to gain-of-function or dominant-negative mutations; however, one could argue that the effect of these mutations could be diminished in the presence of two copies of normal alleles. Similarly, a hypomorphic allele (with a low level of expression, for example) could be masked if the gene was duplicated within

a CNV; its effect could only be manifested in a state of a single allele on each chromosome. These arguments could also apply to variations in the severity of the resulting phenotype. Reduced penetrance has been observed for several diseases that result from CNV, including DiGeorge syndrome and its reciprocal duplication syndrome, as well as speech problems in patients with duplication of the Williams syndrome region.

An estimated 365 Mb of DNA were found to have a variable number of copies in the genomic DNA of lymphoblastoid cell lines that were derived from 270 HapMap individuals[57]. Approximately 15% of genes within these CNVs are known to underlie Mendelian monogenic disease phenotypes (285 out of the 1961 genes listed in the OMIM Morbid Map). We predict that the overall penetrance and/or variations in the severity of the phenotype of dominant disorders that are caused by mutations in genes located within CNVs will be modified compared with the penetrance of dominant traits that are caused by mutations in genes that do not map in CNV regions.

For example, it has been shown that an amino-acid variant (Y402H) in the complement factor H and membrane cofactor (*CFH*) gene predisposes to age-related macular degeneration[26–29,69]. As *CFH* is included in a CNV[57], it is conceivable that the risk that is conferred by one such change (Y402H) is modified by variability in copy number at the *CFH* gene or the nearby genomic region that includes the *CFHR1* and *CFHR3* genes. In fact, Hughes *et al.* have reported that a *CFHR1* and *CFHR3* deletion haplotype is protective against age-related macular degeneration[70]. Similarly, atypical haemolytic–uraemic syndrome might be modified by copy number variation at this locus, thereby explaining the variability in the clinical presentation of the disorder[71]. Because both gains and losses of genomic sequences affect the region in which these genes lie[57], studies that combine both SNPs and CNVs might further clarify the contribution of these two types of genomic variation to these diseases.

CNV contribution to trisomy phenotypes

Despite the availability of the sequence of the euchromatic portion of the human and other mammalian genomes, and the ongoing functional annotation, the genes and other functional elements that are responsible for the various phenotypes of the common trisomies remain largely unknown. Several attempts to define minimum triplicated regions that underlie the trisomy-associated phenotypes have not yielded unequivocal results[72]. It now seems logical to consider polymorphic triplicated regions (that is, CNVs) in normal individuals as contributing to the phenotypic characteristics of trisomies. For example, CNVs involve 3.5 Mb on chromosome 21, 10.1 Mb on chromosome 13 and 6.5 Mb on chromosome 18 — respectively, 10.1%, 10.6% and 8.6% of the sequenced nucleotides of these chromosomes (see the Database of Genomic Variants). Regions of these chromosomes that vary in copy number might harbour genes or other functional genomic elements that, in three copies, are insufficient *per se* to cause the various phenotypes of trisomy 21, 13 or 18, because these micro-triplications are present in normal individuals and are therefore free of any gene-dosage-dependent phenotypic consequences. As nearly 50% of CNVs have been detected only once[57], disclosure of the contribution of common CNVs to the modulation of the phenotype of trisomies will require the study of larger cohorts.

Trisomy and more than three copies

Most of the phenotypic variability that is associated with trisomies has been attributed to the allelic contribution to trisomy and threshold effect of gene-expression variation[72]. But some of the phenotypic variability might be due to the presence of more than three copies of certain sequences. Consider the pedigree in FIG. 1b: The trisomy 21 individual has, through an error in meiosis, inherited two copies of the chromosomal segment from the mother and one from the father. Because this segment carries a CNV with two copies of the 'red' gene, the total number of copies of the red gene in the trisomy 21 individual is, in fact, five. This genomic imbalance could contribute to her phenotype. Had she inherited two copies of the other maternal allele, she would be trisomic for the whole of chromosome 21, with only three copies of the red gene, and therefore would not develop the phenotype that is associated with the red gene. Note that the mother carries three copies of the red gene, and she is not affected. According to the discussion presented above, the region of the red gene is excluded from contributing to the trisomy phenotype if triplicated because the mother is a normal individual.

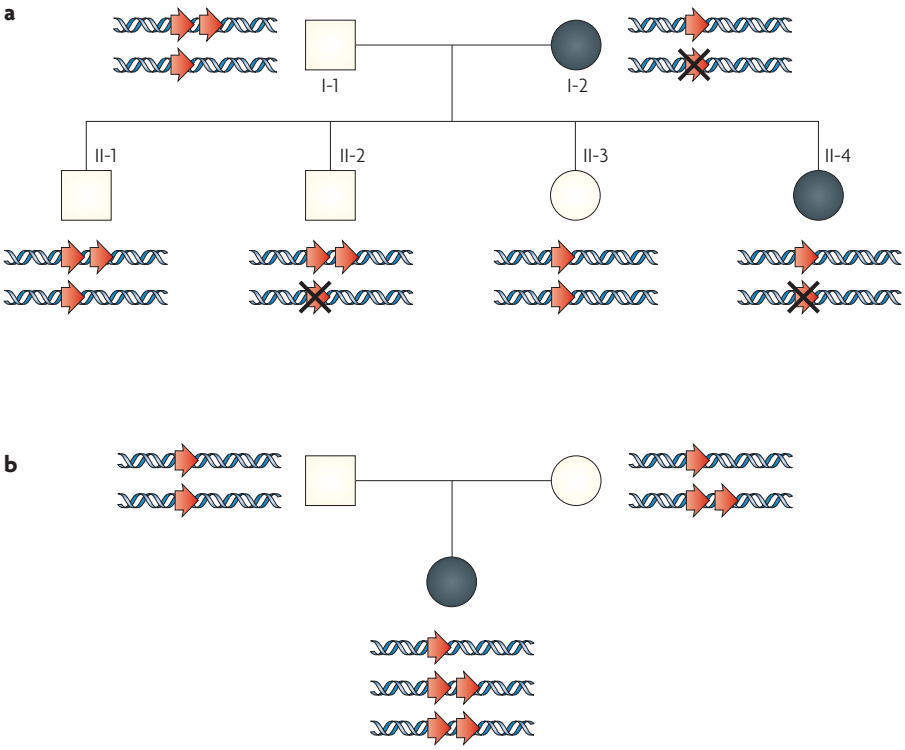


Figure 1 | **Influence of copy number variations (CNVs) on penetrance and variation in severity of the phenotype. a** | A hypothetical example to show how CNVs could be used to determine the penetrance of a dominant mutant allele; see text for details. **b** | A hypothetical example to show how CNVs could modify the phenotypic expression of trisomies; see text for details.

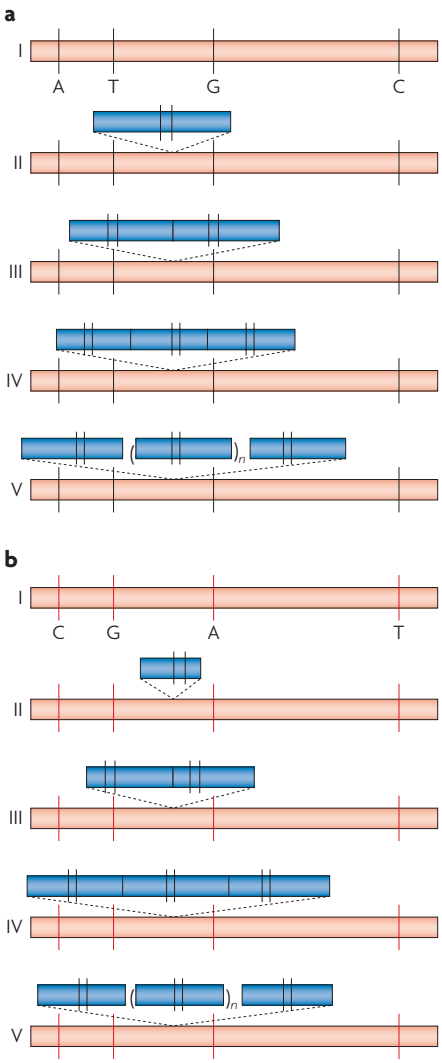


Figure 2 | **Possible haplotype configurations in a copy number variation (CNV)-prone chromosomal region.** A simplified schematic representation of various haplotypic scenarios of a CNV-prone chromosomal region (represented as a horizontal red bar) and its biallelic SNPs (represented as vertical lines), with two on either side of the CNV region (represented as blue bars above the chromosomal region) forming the black (in part **a**) or red (in part **b**) haplotypes, and two that are internal to the CNV. Individuals who are homozygous for a CNV-null allele (I) will not contain the internal SNPs, whereas all other combinations will vary in copy number of the corresponding SNPs. Offspring from a heterozygous I–III combination will inherit either a null or a double dose of the internal SNPs, and thus manifest a non-Mendelian inheritance. Furthermore, recurrence of CNVs in this chromosomal region will blur any linkage disequilibrium between the CNV and its flanking markers. Note also that, although markers that flank rare CNVs might be in perfect linkage disequilibrium with them, the same haplotype will also be found on the most common CNV allele.

However, the red gene region, although it does not contribute to the phenotype in three copies, could contribute if present in five copies, either directly or as a modifier of the phenotype. This hypothesis can be evaluated by association studies and quantitative molecular methods of the copy number of DNA elements and certain phenotypic characteristics, and CNVs should be considered as potential modifiers of the phenotypic variability of trisomies (or aneuploidies in general).

CNVs and genomic disorders

The pathogenic role of CNVs has been known since the elucidation of the aetiology of Charcot–Marie–Tooth neuropathy type 1 (REF. 53) and hereditary neuropathy with liability to pressure palsies[73]. Point mutations in the disease-causing genes within these CNVs have been uncovered. However, most patients with genomic disorders have genomic copy number abnormalities rather than point mutations, suggesting that the frequency of *de novo* events for these types of changes, and hence their population prevalence, differ significantly in favour of chromosomal rearrangements[54].

Specific efforts are underway to uncover genomic changes that are involved in cases of clinical abnormalities (see the DECIPHER web site). Although copy number changes were initially documented through the study of inherited diseases, we now know that CNVs cover approximately 12% of the human genome, potentially altering gene dosage, disrupting genes or perturbing regulation of their expression, even at long-range distances; thus, a considerable number of apparently Mendelian disorders might be due to CNVs. And just as CNVs can affect monogenic traits (including monogenic forms of common disorders, see REFS 74–76), CNVs are also likely to underlie the aetiology of common disorders as a result of variability in gene dosage[77,78].

CNVs and complex traits

Several complex disorders have already been associated with CNVs. For example, susceptibility to HIV-1, lupus and Crohn disease have been found associated with CNVs involving the *CCL3L1* (REF. 79), *FCGR3B*[64,80] and *C4* (REF. 81), and *DEFB4* (REF. 82) genes, respectively. The identification of rare CNVs involved in susceptibility to other common disorders, such as chronic pancreatitis, autism spectrum disorders, Alzheimer disease or Parkinson disease, is likely to enhance the identification of the molecular basis of inherited monogenic forms of these diseases[74–76,83]. In addition, the preponderance and overall chromosomal dispersion of CNVs[57,58] might also impact on the inter-individual differences in drug response[84], as well as susceptibility to infection[79] or cancer[60], either directly or by modulating penetrance or variability in the expression of the trait examined.

SNPs are generally considered in the context of genetic association studies as powerful markers for the mapping of loci that underlie phenotypic variation; these SNPs are mostly proxies for the causal variants with which they are in linkage disequilibrium. However, the use of SNPs in association studies in CNV-related cases will fail to identify the causative genomic regions, because SNPs in these regions do not fulfil the criteria for Mendelian inheritance or Hardy–Weinberg equilibrium in the studied samples; this is especially true in complex and multiallelic CNVs, probably owing to recurrent expansion or contraction[57]. Furthermore, as many as 20% of the SNPs deposited in NCBI are paralogous sequence variants (PSVs)[85], and therefore are a resource for structural variation studies. Redon *et al.* found that biallelic CNVs can be tagged by SNPs[57]. However, this is particularly difficult for CNVs that are complex or that have multiple alleles. This is likely to be due to *de novo* events that might lead to many occurrences of the same or similar CNV alleles (FIG. 2). This is clearly the case for the complex CNVs (containing several variable copies of the *CCL3L1* gene) that are associated with predisposition to HIV-1 infection[57,79] (X.E., unpublished observations). We propose that genome-wide association studies should be systematically complemented by high-density array comparative genomic hybridization (aCGH) studies, which could capture the genetic information of CNVs (FIG. 3).

CNVs, SNPs and genetic information

The study of SNPs has revealed that any two randomly selected human genomes differ by 0.1% (see the International HapMap Project web site). Remarkably, the study of CNVs revised this estimate: in fact, two randomly selected genomes differ by at least 1%, and most of this difference is due to CNVs[57]. Both types of variants, SNPs and CNVs, are widely dispersed throughout the genome, although their genomic prevalence may

differ by two orders of magnitude in favour of SNPs. Yet, because common CNVs affect as much as one-tenth of the human genome[57], any variation in copy number will affect a wide spectrum of genomic sequences (from the kilobase up to the megabase range), and possibly many genes. By contrast, although SNPs are present throughout the entire genome, they involve only a single nucleotide at a time. Therefore, it might be that only a minority of SNPs, those that affect functional elements, have a causative role in phenotype. Considering the higher mutation rate for CNVs versus point mutations[86], and the fact that common CNVs span at least one-tenth of the human genome[57], these types of mutational events are likely to be important players in the aetiology of common disease traits and sporadic birth defects. Furthermore, because of their nature, a significant fraction of CNVs are likely to have functional, causal consequences rather than being linked only to the disorder or trait.

Cataloguing millions of common human SNPs has already yielded important insights into human chromosomal architecture and evolution. It has revealed a block-like structure of linkage disequilibrium, as well as the existence of areas of low or high recombination rate, leading to the identification of so-called tagging SNPs — SNPs that can be used to predict with high probability the alleles at other co-segregating 'tagged' SNPs[19]. We are only beginning to identify and catalogue human CNVs, and our perception of their impact is constrained by the currently unknown nature of their variability and the technology limitations for CNV analysis. Indeed, one might anticipate that common CNVs are likely to affect the recombination landscape in their vicinity, and thus the haplotypic relationships to common genetic markers in these regions. It is therefore legitimate to assume that the abundance of CNVs and their impact on chromosomal haplotypic architecture will also be reflected in the informativeness of syntenic SNPs. For this reason, a common CNV is not likely to be adequately covered by SNPs, as markers that map within the CNV (FIG. 2) have most probably been excluded from the HapMap SNP markers because of their departure from Mendelian inheritance or Hardy–Weinberg equilibrium. SNP selection and current SNP maps are (by design) biased for markers that show Mendelian inheritance. In multiallelic, recurrently occurring CNVs, even flanking SNPs will fail as markers whenever recurrent CNVs fall in different haplotypes, as defined by their respective tagging SNPs (FIG. 2).

Even in the case of a recent and rare CNV, with which all the flanking tagging SNPs will be in complete linkage disequilibrium, these marker alleles will usually be poor predictors for this CNV, as they will also be found on the common CNV allele, and therefore be of little use in association studies (FIG. 2). Several such CNVs (common and rare) that are likely to be predisposing have been associated with some disorders such as HIV-1 and AIDS susceptibility[79], hereditary pancreatitis[75], Crohn disease[82] and systemic lupus erythematosus[64,80,81]. A detailed analysis of the genomic architecture and genetic characteristics of these regions should provide important clues about the use of CNVs and SNPs in the discovery of molecular causes of complex disorders and traits.

SNP-based case–control studies have limited power when the causal variation is distributed over different chromosomal backgrounds; that is, no single tagging SNP allele or haplotype sufficiently discriminates affected subjects from control subjects. This is in contrast to causal CNVs, the recurrence of which is less of an issue because one need not follow specific haplotypes of each CNV but, rather, relate the overall gene dosage to the phenotype (for example, using $\log_2$ ratios) (FIG. 3). This is illustrated in two recent reports showing that a fraction of autism risk loci is likely to involve rare CNVs rather than SNPs[83,87]. These studies also show how CNVs could rapidly lead to the identification of the genes that carry disease-causing mutations, many of which arise *de novo*[87].

SNP and CNV patterns, together with CNV dosage and environmental factors, could combine to produce the phenotype of a trait or disease (FIG. 3). In support of the need of a combined SNP and CNV genotyping approach, Stranger *et al.*[68] recently examined RNA levels in lymphoblastoid cell lines from 210 unrelated HapMap individuals and studied CNVs using BAC arrays[57]; they concluded that as much as 18% of the detected genetic variation in expression of around 15,000 genes could be explained by variability in CNVs. Furthermore, comparing the efficiency of SNPs or CNVs in detecting expression QTLs through association studies, they reported that fewer than 20% of the detected CNV associations overlapped with SNP associations, demonstrating that these two approaches interrogate different parts of the genome and arguing in favour of combining both SNP and CNV analysis in such investigations. This study represents the first attempt to evaluate the impact of both SNPs and CNVs on gene expression; however, the CNV molecular definition must be refined before general conclusions about CNVs with respect to the effect of SNPs on phenotype can be made.

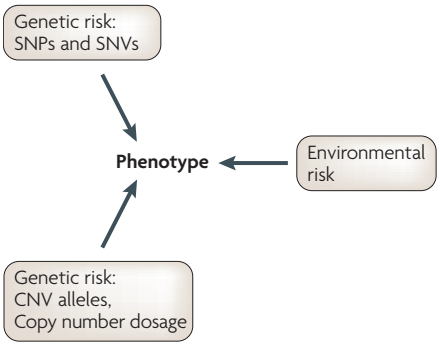


Figure 3 | **The genetic and environmental risks combined confer the total risk for a complex phenotype.** The genetic risk could be subdivided into that contributed by the SNP alleles, and that contributed by copy number variation (CNV alleles or copy number dosage). SNV, single nucleotide variant.

Conclusions

The recent demonstration of the considerable plasticity of the human genome might help to explain phenotypic discrepancies in the penetrance of genetic traits and/or in the severity of the resulting phenotype; it might also provide new leads for the detection of the molecular basis of common complex disorders.

CNVs could offer some of the missing clues to the genetic enigma of complex disorders, as they could harbour genes or other functional elements that, either by copy number alterations or deleterious point mutations, cause or predispose to these phenotypes. Full exploration of the impact of CNVs on phenotype will necessitate CGH arrays with higher resolutions and alternative methods such as multiplex ligation-dependent probe amplification (MLPA)[88], multiplex amplification and probe hybridization (MAPH)[89,90] or quantitative multiplex PCR of short fluorescent fragment (QMPSF)[91] for selected regions of the genome, allowing the detection of small CNVs while providing precise quantitative

estimates of copy numbers. In view of the growing awareness of the need to capture both types of variation in genome-wide studies, the genotyping industry is currently adapting genotyping platforms to allow for simultaneous monitoring of SNPs and known CNVs. It is therefore highly likely that in the foreseeable future high-quality diagnostic tools for a systematic exploration of both types of variations will be produced including the emerging cost-effective ultra high-throughput genome sequencing technologies.

Currently, our knowledge of CNVs is still incomplete, and higher-resolution whole-genome tiling arrays are needed to capture CNVs of smaller size; that is, in the range of several kilobases. Meanwhile, granted the availability of high-density, whole-genome SNP genotyping arrays that extract a considerable amount of the genetic information (about 80%), investigators should be encouraged to report chromosomal regions that contain a series of consecutive markers[82] that are not in Hardy–Weinberg equilibrium, or that depart from Mendelian transmission[92,93], in cases or controls. This information could not only pinpoint additional CNVs but also facilitate the identification of chromosomal regions that are associated with a given disease phenotype.

Note added in proof

In an elegant recent review, Lupski estimated that the *de novo* locus-specific mutation rates for rearrangements are between 100- and 10,000-fold greater than those for point mutations[94].

Glossary

Aneuploidy
Having more or less than the typical chromosome number (46 for humans).

Array comparative genomic hybridization
A technology in which sampled and reference DNA are differentially labelled and hybridized on BAC or oligonucleotide microarrays to show copy number differences between the sampled genomes.

Association study
A population-based genetic study that examines whether a marker allele segregates with a phenotype (such as disease occurrence or a quantitative trait) at a significantly higher rate than would be predicted by chance alone. This is ascertained by genotyping variants in both affected and unaffected or control individuals.

BAC
A DNA construct, derived from a fertility plasmid (or F-plasmid), which usually carries an insert of 100–300 kb. Complete genomic libraries cloned in BACs (or PACs, which are produced from P1-plasmids) have been useful in constructing arrays for array comparative genomic hybridization experiments.

Copy number variant or polymorphism
A structural genomic variant that results in confined copy number changes in a specific chromosomal region. If its population allele frequency is less than 1%, it is referred to as a variant; if its frequency exceeds 1%, the term polymorphism is used.

Duplicon
A duplication, or portion thereof, of genomic sequence that shows a high level of sequence identity (over 90%) to another region of a reference genome. Also sometimes referred to as a low copy repeat or segmental duplication.

Genomic disorder
A disorder that results from the gain, loss or re-orientation of a genomic region that often contains dosage-sensitive gene(s). The result is a genomic rearrangement (such as duplication, deletion and inversion). Segmental duplications are often involved in the rearrangement event through non-allelic homologous recombination.

Haplotype Block
A chromosomal region in which groups of alleles at different genetic loci are inherited together more often than would be expected by chance. Adjacent blocks are separated by recombination hotspots (short regions with high recombination rates).

Hardy–Weinberg equilibrium
The binomial distribution of genotypes in a population, such that frequencies of genotypes AA, Aa and aa will be p^2, 2pq and q^2, respectively, where p is the frequency of allele A, and q is the frequency of allele a.

High-resolution tiling path CGH arrays
Arrays for comparative genomic hybridization (CGH) that offer a resolution in the order of bases to kilobases. The arrays currently use BACs or long oligonucleotides.

Hypomorphic
Describes an allele that carries a mutation that causes a partial loss of gene function.

Linkage disequilibrium
A measure of whether alleles at two loci coexist within gametes in a population in a nonrandom fashion. Alleles that are in linkage disequilibrium are found together on the same haplotype more often than would be expected by chance.

Microsatellite
A class of repetitive DNA sequences, scattered throughout the genome, that are made up of tandemly organized repeats of 2–8 nucleotides in length. They can be highly polymorphic and are frequently used as molecular markers in population genetics studies.

Minisatellites
Regions of DNA in which repeat units of 7–100 bp are arranged in tandem arrays of 0.5–30 kb long.

Multiplex amplification and probe hybridization
A technique in which, following hybridization to immobilized samples of nucleic acid sequences, amplification of each oligonucleotide probe yields a product of unique size. The copy number of target sequences is reflected in the relative intensities of the amplification products.

Multiplex ligation-dependent probe amplification
A technique involving the ligation of two adjacent annealing oligonucleotides followed by quantitative PCR amplification of the ligated products, allowing the detection of deletions, duplications and trisomies, and characterization of chromosomal aberrations in copy number or sequence and SNP or mutation detection.

Non-allelic homologous recombination
Recombination between non-allelic paralogous segmental duplications (also known as low copy repeats); a major mechanism leading to deletions, duplications and inversions, as well as complex structural polymorphism and rearrangements in the human genome.

Paralogous sequence variants
Genetic changes that are not due to polymorphism but to nucleotide mismatches from paralogous copies of duplicated sequences of the genome. About 20% of the SNPs deposited in databases are not true SNPs but paralogous sequence variants.

Penetrance
The extent to which a given genotype manifests itself in a given phenotype. The penetrance of some genotypes for some diseases is age-related, complicating the determination of true penetrance.

Quantitative multiplex PCR of short fluorescent fragments
Semi-quantitative, high-throughput analysis of targeted genomic alterations using locus-specific primers.

Restriction fragment length polymorphism
A fragment length variant of a DNA sequence that is generated through the gain or loss of a site for a restriction enzyme.

Tagging SNPs
SNPs that are correlated with and therefore can serve as proxies for a set of variants with which they are in linkage disequilibrium.

Ultra high-throughput sequencing
A compendium of new sequencing technologies with a common aim to accelerate (from years to days or hours) and reduce the cost (from millions to thousands or hundreds of dollars) of sequencing.

Uniparental disomy
A state wherein both homologues (alleles) at a locus derive from the same parent. Uniparental disomy of some chromosomal segments generates characteristic syndromes.

Variable number of tandem repeat locus
A locus that contains a variable number of short tandemly repeated DNA sequences that vary in length and are highly polymorphic.

Whole-genome tiling array
A high-density oligonucleotide array that represents the majority of DNA sequences of an organism's genome.

Jacques S. Beckmann is at the Department of Medical Genetics, University of Lausanne and Centre Hospitalier Universitaire Vaudois, 2 Avenue Pierre Decker, 1011 Lausanne, Switzerland.

Xavier Estivill is at the Genes and Disease Program, Center for Genomic Regulation (CRG), National Genotyping Center (CeGen), CIBERESP and Pompeu Fabra University (UPF), Charles Darwin s/n, PRBB building, E-08003 Barcelona, Catalonia, Spain.

Stylianos E. Antonarakis is at the Department of Genetic Medicine and Development, University of Geneva Medical School, and University Hospital of Geneva, 1 rue Michel-Servet, 1211 Geneva, Switzerland.

Correspondence to J.S.B., X.E. or S.E.A.
e-mails: Jacques.Beckmann@chuv.ch; Xavier.Estivill@crg.es; Stylianos.Antonarakis@medecine.unige.ch

doi:10.1038/nrg2149

1. Kan, Y. W. & Dozy, A. M. Polymorphism of DNA sequence adjacent to human β-globin structural gene: relationship to sickle mutation. *Proc. Natl Acad. Sci. USA* **75**, 5631–5635 (1978).
2. Wyman, A. R. & White, R. A highly polymorphic locus in human DNA. *Proc. Natl Acad. Sci. USA* **77**, 6754–6758 (1980).
3. Jeffreys, A. J., Wilson, V. & Thein, S. L. Hypervariable 'minisatellite' regions in human DNA. *Nature* **314**, 67–73 (1985).
4. Jeffreys, A. J., Wilson, V. & Thein, S. L. Individual-specific 'fingerprints' of human DNA. *Nature* **316**, 76–79 (1985).
5. Nakamura, Y. *et al.* Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235**, 1616–1622 (1987).
6. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
7. Litt, M. & Luty, J. A. A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.* **44**, 397–401 (1989).
8. Weber, J. L. & May, P. E. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**, 388–396 (1989).
9. Tautz, D. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res.* **17**, 6463–6471 (1989).
10. Smeets, H. J., Brunner, H. G., Ropers, H. H. & Wieringa, B. Use of variable simple sequence motifs as genetic markers: application to study of myotonic dystrophy. *Hum. Genet.* **83**, 245–251 (1989).
11. Williamson, R. *et al.* Report of the DNA committee and catalogues of cloned and mapped genes and DNA polymorphisms. *Cytogenet. Cell Genet.* **55**, 457–778 (1990).
12. Economou, E. P., Bergen, A. W., Warren, A. C. & Antonarakis, S. E. The polydeoxyadenylate tract of Alu repetitive elements is polymorphic in the human genome. *Proc. Natl Acad. Sci. USA* **87**, 2951–2954 (1990).
13. Kashi, Y. *et al.* Large restriction fragments containing poly-TG are highly polymorphic in a variety of vertebrates. *Nucleic Acids Res.* **18**, 1129–1132 (1990).
14. Beckmann, J. S. & Weber, J. L. Survey of human and rat microsatellites. *Genomics* **12**, 627–631 (1992).
15. Weissenbach, J. *et al.* A second-generation linkage map of the human genome. *Nature* **359**, 794–801 (1992).
16. Murray, J. C. *et al.* A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). *Science* **265**, 2049–2054 (1994).
17. Gyapay, G. *et al.* The 1993–94 Genethon human genetic linkage map. *Nature Genet.* **7**, 246–339 (1994).
18. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154 (1996).
19. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
20. Dutt, A. & Beroukhim, R. Single nucleotide polymorphism array analysis of cancer. *Curr. Opin. Oncol.* **19**, 43–49 (2007).
21. Varilo, T. & Peltonen, L. Isolates and their potential use in complex gene mapping efforts. *Curr. Opin. Genet. Dev.* **14**, 316–323 (2004).
22. Weir, B. S., Anderson, A. D. & Hepler, A. B. Genetic relatedness analysis: modern data and new challenges. *Nature Rev. Genet.* **7**, 771–780 (2006).
23. Engel, E. Uniparental disomy revisited: the first twelve years. *Am. J. Med. Genet.* **46**, 670–674 (1993).
24. Antonarakis, S. E. Parental origin of the extra chromosome in trisomy 21 as indicated by analysis of DNA polymorphisms. Down Syndrome Collaborative Group. *N. Engl. J. Med.* **324**, 872–876 (1991).
25. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
26. Edwards, A. O. *et al.* Complement factor H polymorphism and age-related macular degeneration. *Science* **308**, 421–424 (2005).
27. Hageman, G. S. *et al.* A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. *Proc. Natl Acad. Sci. USA* **102**, 7227–7232 (2005).
28. Haines, J. L. *et al.* Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308**, 419–421 (2005).
29. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
30. Bottini, N. *et al.* A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nature Genet.* **36**, 337–338 (2004).
31. Smyth, D. J. *et al.* A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nature Genet.* **38**, 617–619 (2006).
32. Grant, S. F. *et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature Genet.* **38**, 320–323 (2006).
33. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
34. Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).
35. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
36. Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
37. Todd, J. A. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genet.* **53**, 1884–1889 (2007).
38. Frayling, T. M. *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).
39. Dina, C. *et al.* Variation in FTO contributes to childhood obesity and severe adult obesity. *Nature Genet.* **39**, 724–726 (2007).
40. Helgadottir, A. *et al.* A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316**, 1491–1493 (2007).
41. McPherson, R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488–1491 (2007).
42. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nature Genet.* **39**, 631–637 (2007).
43. Camp, N. J. *et al.* Compelling evidence for a prostate cancer gene at 22q12. 3 by the International Consortium for Prostate Cancer Genetics. *Hum. Mol. Genet.* **16**, 1271–1278 (2007).
44. Hunter, D. J. *et al.* A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nature Genet.* **39**, 870–874 (2007).
45. Easton, D. F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
46. Stacey, S. N. *et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nature Genet.* **39**, 865–869 (2007).
47. Goossens, M. *et al.* Triplicated α-globin loci in humans. *Proc. Natl Acad. Sci. USA* **77**, 518–521 (1980).
48. Kan, Y. W. *et al.* Deletion of α-globin genes in haemoglobin-H disease demonstrates multiple α-globin structural loci. *Nature* **255**, 255–256 (1975).
49. Vollrath, D., Nathans, J. & Davis, R. W. Tandem array of human visual pigment genes at Xq28. *Science* **240**, 1669–1672 (1988).
50. Drummond-Borg, M., Deeb, S. S. & Motulsky, A. G. Molecular patterns of X chromosome-linked color vision genes among 134 men of European ancestry. *Proc. Natl Acad. Sci. USA* **86**, 983–987 (1989).
51. Wagner, F. F. & Flegel, W. A. *RHD* gene deletion occurred in the Rhesus box. *Blood* **95**, 3662–3668 (2000).
52. Ji, Y., Eichler, E. E., Schwartz, S. & Nicholls, R. D. Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res.* **10**, 597–610 (2000).
53. Lupski, J. R. *et al.* DNA duplication associated with Charcot–Marie–Tooth disease type 1A. *Cell* **66**, 219–232 (1991).
54. Lee, J. A. & Lupski, J. R. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron* **52**, 103–121 (2006).
55. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36**, 949–951 (2004).
56. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
57. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
58. Wong, K. K. *et al.* A comprehensive analysis of common copy-number variations in the human genome. *Am. J. Hum. Genet.* **80**, 91–104 (2007).
59. Eichler, E. E. *et al.* Completing the map of human genetic variation. *Nature* **447**, 161–165 (2007).
60. Cho, E. K. *et al.* Array-based comparative genomic hybridization and copy number variation in cancer research. *Cytogenet. Genome Res.* **115**, 262–272 (2006).
61. Freeman, J. L. *et al.* Copy number variation: new insights in genome diversity. *Genome Res.* **16**, 949–961 (2006).
62. Khaja, R. *et al.* Genome assembly comparison identifies structural variants in the human genome. *Nature Genet.* **38**, 1413–1418 (2006).
63. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nature Rev. Genet.* **7**, 85–97 (2006).
64. Aitman, T. J. *et al.* Copy number polymorphism in *FCGR3* predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851–855 (2006).
65. Antonarakis, S. E. & Beckmann, J. S. Mendelian disorders deserve more attention. *Nature Rev. Genet.* **7**, 277–282 (2006).
66. Feuk, L., Marshall, C. R., Wintle, R. F. & Scherer, S. W. Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum. Mol. Genet.* **15**, R57–R66 (2006).
67. Vissers, L. E., Veltman, J. A., van Kessel, A. G. & Brunner, H. G. Identification of disease genes by whole genome CGH arrays. *Hum. Mol. Genet.* **14**, R215–R223 (2005).
68. Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
69. Thakkinstian, A. *et al.* Systematic review and meta-analysis of the association between complement factor H Y402H polymorphisms and age-related macular degeneration. *Hum. Mol. Genet.* **15**, 2784–2790 (2006).
70. Hughes, A. E. *et al.* A common *CFH* haplotype, with deletion of *CFHR1* and *CFHR3*, is associated with lower risk of age-related macular degeneration. *Nature Genet.* **38**, 1173–1177 (2006).

71. Fremeaux-Bacchi, V. *et al.* The development of atypical haemolytic–uraemic syndrome is influenced by susceptibility factors in factor H and membrane cofactor protein: evidence from two independent cohorts. *J. Med. Genet.* **42**, 852–856 (2005).
72. Antonarakis, S. E., Lyle, R., Dermitzakis, E. T., Reymond, A. & Deutsch, S. Chromosome 21 and Down syndrome: from genomics to pathophysiology. *Nature Rev. Genet.* **5**, 725–738 (2004).
73. Boerkoel, C. F., Inoue, K., Reiter, L. T., Warner, L. E. & Lupski, J. R. Molecular mechanisms for *CMT1A* duplication and HNPP deletion. *Ann. NY Acad. Sci.* **883**, 22–35 (1999).
74. Singleton, A. B. *et al.* α-Synuclein locus triplication causes Parkinson's disease. *Science* **302**, 841 (2003).
75. Le Marechal, C. *et al.* Hereditary pancreatitis caused by triplication of the trypsinogen locus. *Nature Genet.* **38**, 1372–1374 (2006).
76. Rovelet-Lecrux, A. *et al.* APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nature Genet.* **38**, 24–26 (2006).
77. Knight, J. C. Regulatory polymorphisms underlying complex disease traits. *J. Mol. Med.* **83**, 97–109 (2005).
78. Stranger, B. E. & Dermitzakis, E. T. From DNA to RNA to disease and back: the 'central dogma' of regulatory disease variation. *Hum. Genomics* **2**, 383–390 (2006).
79. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440 (2005).
80. Fanciulli, M. *et al. FCGR3B* copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature Genet.* **39**, 721–723 (2007).
81. Yang, Y. *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).
82. Fellermann, K. *et al.* A chromosome 8 gene-cluster polymorphism with low human β-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am. J. Hum. Genet.* **79**, 439–448 (2006).
83. Szatmari, P. *et al.* Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nature Genet.* **39**, 319–328 (2007).
84. Ouahchi, K., Lindeman, N. & Lee, C. Copy number variants and pharmacogenomics. *Pharmacogenomics* **7**, 25–29 (2006).
85. Estivill, X. *et al.* Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum. Mol. Genet.* **11**, 1987–1995 (2002).
86. Inoue, K. & Lupski, J. R. Molecular mechanisms for genomic disorders. *Annu. Rev. Genomics Hum. Genet.* **3**, 199–242 (2002).
87. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
88. Schouten, J. P. *et al.* Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.* **30**, e57 (2002).
89. Armour, J. A., Sismani, C., Patsalis, P. C. & Cross, G. Measurement of locus copy number by hybridisation with amplifiable probes. *Nucleic Acids Res.* **28**, 605–609 (2000).
90. Sellner, L. N. & Taylor, G. R. MLPA and MAPH: new techniques for detection of gene deletions. *Hum. Mutat.* **23**, 413–419 (2004).
91. Saugier-Veber, P. *et al.* Simple detection of genomic microdeletions and microduplications using QMPSF in patients with idiopathic mental retardation. *Eur. J. Hum. Genet.* **14**, 1009–1017 (2006).
92. McCarroll, S. A. *et al.* Common deletion polymorphisms in the human genome. *Nature Genet.* **38**, 86–92 (2006).
93. Conrad, D. F. *et al.* A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genet.* **38**, 1251–1260 (2006).
94. Lupski, J. R. Genomic rearrangements and sporadic disease. *Nature Genet.* **39**, S43–S47 (2007).

Acknowledgements

We thank all the past and present members of our laboratories and clinics for ideas, debates and discussions. We thank J. Lupski for critical reading of the manuscript. Work in J.S.B.'s laboratory is funded by grants from the SNF (Swiss National Science Foundation) and the University of Lausanne, Switzerland. The laboratory of X.E. is supported by: the Departament d'Educació i Universitats and the Departament de Salut of the Catalan Autonomous Government (Generalitat de Catalunya); the Ministry of Health and the Ministry of Education and Science of the Spanish Government; the European Union Sixth Framework Programme; and Genoma España. S.E.A.'s laboratory is supported by the SNF, European Union, US National Institutes of Health and the Lejeune and ChildCare Foundations (France).

Competing interests statement

The authors declare no competing financial interests.

DATABASES

The following terms in this article are linked online to:

Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

CCL3L1 | *CFH* | *CFHR1* | *CFHR3* | *C4* | *DEFB4* | *FCGR3B* | *HBB* | *RHD*

OMIM: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

Alzheimer disease | atypical haemolytic–uraemic syndrome | *BRCA1* | *BRCA2* | hereditary neuropathy with liability to pressure palsies | neurofibromatosis type 1 | Parkinson disease | tuberous sclerosis

FURTHER INFORMATION

Centre for Genomic Regulation: http://pasteur.crg.es

Database of Genomic Variants: http://projects.tcag.ca/variation

DECIPHER: http://www.sanger.ac.uk/PostGenomics/decipher/

International HapMap Project: http://www.hapmap.org

NCBI Single Nucleotide Polymorphism: http://www.ncbi.nlm.nih.gov/SNP

OMIM Morbid Map: http://www.ncbi.nlm.nih.gov/Omim/getmorbid.cgi

UCSC Genome Bioinformatics: http://genome.ucsc.edu

Access to this links box is available online.