Published in The Lancet Volume 374, Issue 9686, 25 July 2009-31 July 2009, Pages 340-350

http://www.thelancet.com/

Genomic copy number variation, human health, and disease

Louise V Wain PhD[a], Prof John AL Armour PhD[b] and Martin D Tobin PhD[a]*

[a]Department of Health Sciences and Department of Genetics, University of Leicester, Leicester, UK

[b]Institute of Genetics and School of Biology, University of Nottingham, Nottingham, UK

*Correspondence to: Martin D Tobin, MRC Clinician Scientist Fellow, Department of Health Sciences and Department of Genetics, University of Leicester, Leicester LE1 7RH, UK

Summary

Despite the long recognised effects of chromosomal structural abnormalities and completion of the Human Genome Project, much of the structural variation in the genome has gone unrecognised until recently. Deletions and duplications of DNA strands of between a few hundred bp and several million bp—collectively referred to as copy number variants—are now known to be widespread. Since 2007, rigorous and adequately powered genome-wide association studies based on single nucleotide polymorphisms have yielded replicated associations to several common diseases. Some copy number variants explain rare, previously uncharacterised disorders, and they are now expected to explain some of the genetic contribution to common diseases. We review efforts to map copy number variants and discuss present and future prospects for assessment of their relation to human health and disease.

Introduction

Human genomic variation in the form of single nucleotide substitutions has been catalogued in incredible detail since the completion of the draft human genome sequence in 2000. The HapMap project, which by 2007 had documented more than 3·1 million single nucleotide polymorphisms (SNPs) and their inter-relation, has underpinned subsequent successful genome-wide association studies.[1] While findings from these studies were showing novel SNP associations for a range of common diseases,[2 and 3] detailed efforts were underway to comprehensively map submicroscopic structural variants across the human genome.[4, 5, 6, 7, 8, 9, 10, 11 and 12] Use of emerging molecular biology technologies has shown that the extent of structural genomic variation is much greater than was previously suspected.[12, 13 and 14]

In view of the established role of large deletions, duplications, and other structural rearrangements in specific diseases termed genomic disorders,[15 and 16] these discoveries have excited interest about the contribution of intermediate-scale structural genomic variation to disease risk. Structural variation caused by a variable number of copies of a particular DNA segment are referred to as copy number variants.[17] Much of the scientific infrastructure to support genome-wide association studies of copy number variation has already been established to undertake genome-wide SNP-based association studies including consortia of studies to achieve large total sample sizes.[2] Genome-wide SNP association studies could be supplemented with association studies of copy number variation by use of existing SNP data or by the undertaking of appropriate further assays.[2]

The potential role of copy number variation in complex diseases,[18] including susceptibility to autism,[19, 20, 21, 22 and 23] schizophrenia,[24, 25, 26 and 27] Crohn's disease,[28] psoriasis,[29] systemic lupus erythematosus,[30, 31 and 32] amyotrophic lateral sclerosis,[33] and HIV-1,[34] is supported by a growing amount of published work. Such variation has also been associated with vertical transmission of HIV-1,[34 and 35] and the progression and response to treatment of HIV-1/AIDS.[36] Although some of these associations could be false positives, others have had either technical validation or replication of findings in different study populations, or both.

These validated associations are probably a direct or an indirect result of changes to the copy number of the relevant genomic sequence.

Understanding of the disease mechanisms of a range of common diseases could soon be informed by the discovery of copy number variants associated with predisposition to disease. However, many challenges lie ahead for the reliable detection of such associations. The location, size, and boundaries (breakpoints) of these variants documented in public databases have been very imprecise, although the new generation of maps of copy number variants are providing much improvement.[14 and 37] Scalable methods to characterise copy number variation in association studies have been inexact. New genome-wide SNP arrays have advanced coverage and resolution, but the potential for misclassification remains much higher for variations in copy number than for SNPs.[14, 38 and 39] Furthermore, studies of this variation are constrained by the same issues of power, differential bias, and extreme multiple-testing that are encountered by SNP-based association studies. In this Review, we introduce copy number variation, examine advances in our understanding, and discuss the inferences that can be reasonably drawn from association studies of copy number variation.

Definition

For nearly all genes in the human genome, we inherit one copy from each parent, so that we have two copies in the nucleus of every diploid cell. However, the copy number per genome varies for some genes. For example, a cluster of several β-defensin genes shows common copy number variation of between two and seven copies per diploid cell, and occasionally copy numbers are as high as ten or 11.[40] Similarly, the salivary amylase gene, *AMY1*, varies in copy number from two to 15 with a mean of seven in the European–American population.[41]

The simplest type of copy number variation is the presence or absence of a gene (figure 1A)—eg, in Europe, the rhesus-negative allele at the main antigen locus for the rhesus blood group is commonly caused by complete deletion of the *RHD* gene.[42] An individual's (diploid) genome could therefore contain two, one, or zero copies of *RHD*, with the zero copies corresponding to rhesus-negative and absence of D antigen expression. Generally, genomic segments with variable copy number could encompass parts of genes, reside entirely

outside genes or, in the case of larger variants, include several known genes. A simple duplication of a genomic segment could result in diploid copy numbers of two, three, or four (figure 1B), and successive rounds of duplication could produce a wide range of diploid copy numbers, known as multiallelic copy number variants (figure 1C). These deletions and duplications result in variants that range from a few hundred to several million bp.

Variable copy number can affect SNP genotyping. A deletion might cause contiguous SNPs to show loss of heterozygosity because hemizygous genotypes are judged to be homozygous. For example, if the major allele (A) is present on one chromosome and the homologous chromosomal location is deleted, then only allele A is detected and the genotype is called AA. This misrepresentation can cause deviation from Hardy-Weinberg equilibrium—in which the expected frequencies of genotypes for any variant can be predicted from the allele frequencies—and apparent inconsistencies with mendelian transmission in family data. This occurrence caused many SNPs in regions of copy number variation to be excluded from the earliest genome-wide genotyping arrays.[8, 11, 14 and 38]

Copy number variation is a subset of the broad occurrence of genomic structural variation, which is any variation exceeding that detected for SNPs or microsatellite variation (blocks of two–six bp repeated ten–100 times). The range of structural genomic variation includes rearrangements such as inversions, which change the orientation of a DNA segment, and balanced translocations, in which a DNA segment is reciprocally exchanged between two chromosomes. These rearrangements change the spatial organisation of DNA but do not result in any net gain or loss of sequence. Copy number variation is not simple to define either rigidly or rationally. For example, tandemly repeated minisatellites, insertions of mobile genetic elements, and highly repeated tandem arrays of genes (eg, ribosomal RNA genes) share common features with copy number variants, but are generally considered separately.

Somewhat arbitrarily, copy number variation usually refers to DNA sequences of about 1000 bp (1 kb) or more. The Database of Genomic Variants[4] (panel 1) records structural variants of 1 kb or more of DNA, but regards those in the 100–1000 bp range to be in a separate

category of indel. This artificial division within a continuous range has been historically defined by the size limits of the methods used to detect copy number variation, including sequencing. Although extreme differences are often clear, no hard and fast criterion can distinguish copy number variants from large-scale structural genomic variation that has been implicated in syndromic anomalies and diseases, which are referred to as genomic disorders (figure 2).

Researchers investigating the potential effect of copy number variation on human health and disease can probably learn from the study of genomic disorders (panel 2). Some of these disorders have phenotypes that are dependent on gene dosage, which is the number of functioning gene copies and determines the amount of gene product. Because genomic disorders are caused by de-novo (rather than inherited) deletions, insertions, or other chromosomal rearrangements, these structural variants contribute to non-heritable components of disease risk. The flanking genomic architecture and mechanisms generating genomic disorders and some large copy number variants might be similar,[13 and 16] although evidence suggests that most common copy number variants are inherited and therefore caused by ancestral structural mutations.[14]

Inferences about the mechanisms by which copy number variation arises have often been made from the structure of alleles in the population, rather than from direct observation of new mutations. Large copy number variants tend to be flanked by segmental duplications,[4 and 49] which are segments of genomic DNA of at least 1 kb in length and 90% sequence similarity that are duplicated throughout the genome, either in tandem array or on different chromosomes. This variation occurred early in human evolution and is now widespread,

whereas copy number variation is thought to have originated more recently. New copy number variation can arise from non-allelic homologous recombination in which recombination takes place between regions that are highly similar in sequence but not allelic, such as non-cognate segmental duplications, resulting in loss or gain of regions adjacent to the recombination event.[46, 50 and 51] Nevertheless, variation can also arise even when there is no sequence similarity between flanking sequences, via simple non-homologous end-joining to repair double-stranded DNA breaks that can lead to loss or gain of nucleotides from imprecise repair. Additionally, detailed sequence analysis has been used for direct examination of germline (sperm) DNA to examine the rates and processes by which copy number variation is generated.[52, 53, 54, 55, 56 and 57] The possibility of somatic mosaicism of copy number variation[58] could complicate analyses.

Functional consequences

If copy number variation affects entire genes, especially those with important effects on biological function, it would naturally be expected to affect susceptibility to disease. The number of copies of the chemokine gene *CCL3L1* varies within and between populations; in Europe, common variation is from zero to four copies, whereas in Africa, copy numbers as high as ten have been recorded. In fact, the copy number of *CCL3L1* is associated with chemokine concentrations, the proportion of CD4+ cells expressing the chemokine receptor CCR5, viral loads, and the rate of AIDS progression (low copy number is associated with susceptibility to HIV-1 infection).[34 and 36] So far, evidence is consistent with a gradation of risk corresponding to *CCL3L1* copy number. Similarly the copy number of *AMY1* is correlated with salivary amylase protein concentration and dietary starch consumption across populations sampled from Africa, Asia, and Europe.[41] Of particular clinical interest is *CYP2D6*, which affects the metabolism of about 50% of drugs including tricyclic antidepressants, selective serotonin reuptake inhibitors, antipsychotics, antiarrhythmics, β blockers, opioid analgesics, antiemetics, and antihistamines. The highly polymorphic copy number of *CYP2D6* affects expression of the cytochrome P450 enzyme CYP2D6, and contributes substantially to the variation in drug metabolism between individuals.[59]

By contrast, the number of copies of the X chromosome green photopigment gene, *OPN1MW*, is highly variable but only the first gene copy seems to be expressed. Second and

subsequent copies of *OPN1MW* on the X chromosome are postulated to be located too far from the regulatory region (termed the locus control region) to be activated, so that *OPN1MW* duplications do not affect colour vision. Rather, defects in colour vision could be caused by hybrid red–green (*OPN1LW–OPN1MW*) genes, gene deletions, or sequence variation.[60]

The clinical consequences of copy number variation might not always be in direct proportion to gene dosage.[61] The functional effect could crucially depend on whether the copy number variant changes the sequence or relative location of specific segments of genomic DNA that act as enhancers or suppressors of gene expression.[62] Studies examining the link between gene disruption and gene expression[26 and 61] have been limited by substantial uncertainty about boundaries of variants; further studies based on updated maps of variation are needed before any conclusions can be made.[14]

The functional effects of copy number variation can depend on the phenotypic and cellular context, and on the environmental background. For example, a 20 kb deletion in *IRGM* is associated with a reduced expression of the gene and protein in HeLa (immortalised cervical cancer) cells and hepatocellular carcinoma cells, but with increased expression in colon carcinoma cells and human bronchus smooth muscle cells.[28] A rhesus-negative blood group might be regarded as an entirely healthy variation, but rhesus haemolytic disease of newborn children can happen when the fetus is rhesus positive and the mother is sensitised to the D antigen. Even increased CYP2D6 enzyme activity caused by duplications and multiplications of *CYP2D6* can lead to ultrarapid metabolism, and reduced or enhanced drug efficacy and toxic effects, dependent on whether the drug is inactivated or activated by CYP2D6.[59] Drugs that are activated by CYP2D6 include tamoxifen and codeine. Extreme caution is warranted for prescription of codeine to mothers who are breastfeeding because morphine—a potent metabolite of codeine—might be present in high concentrations in the breastmilk of those with ultrarapid metabolism who are taking codeine.[63]

Redon and colleagues[12] showed that genes within regions of copy number variation were over-represented by Gene Ontology categories related to several environmental sensor functions, such as cell adhesion, sensory perception, chemical stimuli, and neurophysiological processes. This bias could be caused by positive or balancing selection

for different copy numbers at these environmental sensor genes, indicating variable selection pressure. Alternatively, copy number variation in these gene categories might be better tolerated than in other gene categories for which variation might cause harm and therefore would be quickly removed by selection.[64]

Mapping of copy number variation

Discovery of copy number variation has been accelerated by systematic genome-wide application of array comparative genomic hybridisation, next generation sequencing, and SNP genotyping arrays. Nevertheless, several copy number variants were identified between 10 and 20 years ago by standard methods of molecular genetic analysis at individual genetic loci, including Southern blotting, cytogenetic methods, and PCR-based approaches.

Cytogenetic studies have used techniques such as fluorescent in-situ hybridisation to define two major structural variants on chromosome 8 (p23.1): the so-called euchromatic variant caused by high copy number of a repeat containing β-defensin genes,[40] and common inversion of nearly the whole chromosome band that is carried in a heterozygous state by about a quarter of European people and about a third of Japanese people.[65, 66, 67 and 68] PCR approaches, including quantitative real-time PCR, were originally used to verify copy number. However, subsequent advances have mainly focused on definition of the nature and extent of copy number variation across the entire genome. Quantitative PCR and related locus-specific molecular methods are still of particular use for molecular validation and detailed studies of putative regions of variation, to follow up evidence from genome-wide screens.

Use of array comparative genomic hybridisation is well known for detection of changes in copy number between normal cells and tumour cells. The method uses an array of probes that represent the genome and differentially labelled test and reference DNA samples are jointly hybridised to the array (figure 3); at each probe, the ratio of the labels is used to identify comparative deletions or duplications of DNA. Arrays of either bacterial artificial chromosome (BAC) clones (DNA constructs based on bacterial plasmids into which fragments of genomic DNA 80–200 kb from across the genome can be inserted and

replicated) or oligonucleotides are often used as targets. Redon and colleagues[12] used more than 25 000 large-insert clones, effectively covering almost all the human genome.

However, data from array comparative genomic hybridisation are often much harder to assess than a sequence trace. Even when changes to the BAC signal ratios indicate a real copy number variant, a deletion in the test sample would be indistinguishable from a duplication in the reference sample. Furthermore, although the technique is sensitive to copy number changes of regions as small as 20 kb, the reported start and end coordinates of the variant are actually those of the fragment of DNA inserted into the BAC clone, and therefore the size of the copy number variant is often overestimated.

By contrast, sequencing-based approaches can map copy number variants with much greater accuracy than can array comparative genomic hybridisation, and can detect inversions or translocations. Sequencing studies compare a segment of reference DNA with the paired ends of DNA sequences from a test sample prepared directly from site-selected genomic DNA[69] or from fosmid cloning (figure 4).[7 and 13] For fosmid cloning, genomic DNA is inserted into a fosmid cloning vector and transfected into cells for propagation; only a restricted size range (32–48 kb) of DNA inserts can be efficiently cloned. The expense of sequencing-based approaches at present prohibits their application on a genome-wide scale across many individuals, but sequencing has played an important part in development of maps of copy number variation.

Scalable mapping of copy number variation with improved cost-effectiveness can now be undertaken with the latest generation of genotyping arrays. Early arrays were not ideal since they excluded many SNPs that deviated from three distinct genotype clusters, Hardy-Weinberg equilibrium, or mendelian inheritance. Since these deviations can be explained by copy number variants, early arrays had a paucity of SNP probes in regions of these copy number variants.[14 and 38] However, new hybrid arrays include non-polymorphic probes to detect copy number variation (widely referred to as non-SNP probes or copy number variant

probes) and probes for many more SNPs, such as those purposely excluded from the early arrays, to address this problem.[14 and 38] Approaches to identify copy number variants from genome-wide SNP arrays differ. All methods use quantitative data from SNP probes,[70, 71, 72, 73 and 74] and some also use information from non-polymorphic probes and maps of copy number variation (panel 3).[23 and 75]

Developed maps of copy number variation

An assay of 270 HapMap individuals on an Affymetrix 6.0 array has been used to produce a map of copy number variation of much higher resolution (about 2 kb) than have maps derived from BAC arrays. This map cast doubt about previous assumptions of the frequency, size, and origin of copy number variants.[14] For example, copy number variant sizes were between five and 15 times smaller than were shown by BAC arrays[12] and almost half were reported in multiple unrelated individuals.[14] Most copy number variants were inherited, and although some were less easily tagged than SNPs of an equivalent frequency, common copy number variants tended to show strong linkage disequilibrium with a HapMap SNP. Several such variants that had been previously judged as complex, were explained by two separate simple copy number variants co-occurring on homologous chromosomes (copy number measurements are for the diploid genome).[14]

The knowledge gained from this map[14] has several important implications. First, if most copy number variants are stably inherited at an appreciable frequency in the population, most common copy number variants might be expected to have an individually modest effect on disease risk. Second, most small copy number variants with frequency greater than 1% are also diallelic and in Hardy-Weinberg equilibrium, and thus are analogous to SNPs. Third, some such copy number variants associated with disease might also show association with a nearby SNP. The corollary is that some findings from genome-wide association studies could be explained by copy number variation, as can the signal of association with Crohn's disease from SNPs in the region of *IRGM*.[28]

Association studies

Many years before the surge of interest in the association between copy number variation and disease, some associations were identified by candidate gene studies with low-throughput molecular methods to measure copy number. Examples include association of the α-globin copy number variant with α-thalassaemia and malaria resistance,[79 and 80] and the association between *RHD* gene deletion and haemolytic disease of newborn children.[42 and 81] The link between deficiencies in the complement factor 4 gene (*C4A/C4B*) and systemic lupus erythematosus is also well established;[82] Yang and co-workers[32] confirmed that low copy number of these genes is a risk factor. Regions containing genes of the immune system, especially the innate immune and complement systems, tend to be copy number variable and have been the subject of many candidate gene association studies.

FCGR3B encodes a variant of the type III IgG receptor. The gene has variable copy number, and low copy number is associated with glomerulonephritis in systemic lupus erythematosus.[30] Further studies have shown that low *FCGR3B* copy number is associated with a range of systemic autoimmune diseases, including systemic lupus erythematosus, polyangiitis, and Wegener's granulomatosis.[31] Sequence variation in the complement factor H and related genes can increase the risk of age-related macular degeneration, whereas deletion of the related genes *CFHR1* and *CFHR3* protects against the disorder[83] and predisposes individuals to atypical haemolytic-uraemic syndrome.[84] However, the effect of the deletion is modified by other risk factors, including risk alleles present on non-deletion chromosomes. Once these other factors are accounted for, the protective effect of the deletion on age-related macular degeneration might only be modest.[85]

A unit of seven β-defensin genes has variable copy number on chromosome 8.[40] These genes, including *DEFB4*, encode antimicrobial peptides, but findings from additional studies have suggested that the genes have diverse functions such as chemokine activity.[86] Low copy number has been associated with Crohn's colitis,[87] and high copy number with predisposition to psoriasis.[29]

As mentioned above, the chemokine genes *CCL3L1* and *CCL4L1* are both present on a unit of variable copy number,[88] which shows interesting differences in variation between European (generally zero–four copies) and African populations (four–ten copies).[34] A large study including more than 1500 individuals with HIV-1 indicated that infection and progression to

AIDS were both associated with lower than average copies of the unit,[34] but the result was not confirmed by a subsequent study of infection in adolescents.[89] Copy number variation of *CCL3L1* and *CCL4L1* has also been associated with systemic lupus erythematosus[90] and rheumatoid arthritis.[91]

Genome-wide approaches have provided aetiological insights, particularly for autism and schizophrenia. In autism, large-scale structural abnormalities have long been recognised, including inherited duplications (15)(q11;q13). Strong evidence that rare copy number variants are causally related to familial and sporadic autism has come from studies that use a range of molecular approaches such as representational oligonucleotide microarray analysis,[21] array comparative genomic hybridisation with BAC clones,[19] SNP arrays,[22] and new generation hybrid arrays containing SNP probes and non-polymorphic probes.[20 and 23] Most of the copy number variants detected in these studies are rare—perhaps unique to an individual or family—and are therefore difficult to replicate in independent studies. However, autism has been associated with uncommon de-novo and inherited deletions and duplications of a 593 kb genomic segment on chromosome 16 (p11.2).[19, 20 and 23] Weiss and colleagues[23] have validated and replicated this association, and noted that developmental regression, which is common in late-onset autism (about 40% of cases), was absent in the subgroup of cases with deletions or duplications of the 593 kb genomic segment. Several rare copy number variants also predispose to schizophrenia[24, 25 and 27] including del(1)(q21.1) of about 1·4 Mb,[24 and 25] del(15)(q11.2) of about 470 kb,[24] del(15)(q13.3) of about 1·6 Mb,[24 and 25] and del(22)(q11.2),[25 and 27] which was the locus responsible for velo-cardio-facial syndrome.[92]

Many candidate gene associations relate to multiallelic copy number variants, which could indicate either an increased biological propensity for multiallelic variants to affect function, or a tendency for multiallelic variants to be selected as candidates for association studies. Artifactual associations might be more probable with multiallelic than with diallelic copy number variants because of the technical difficulties with accurate genotyping.[66]

Design of association studies

The examples of genome-wide association studies mentioned previously reported large rare copy number variants, which might identify important subgroups of some common diseases.[12, 21 and 27] Crucially, such studies have often used genome-wide arrays with inadequate resolution to study small common copy number variants and filtering strategies that tend to exclude such variants.

Consequently, different strategies are needed to study common copy number variants. Almost all the factors that are relevant to SNP-based association studies are also important for the design of association studies of copy number variation. Therefore, we consider some aspects of the design, analysis, and reporting of association studies of copy number variation from an epidemiological perspective. SNP association studies[2, 93, 94, 95 and 96] and copy number variation association studies[39 and 95] have been reviewed in detail previously.

Adequate statistical power to detect the generally modest effects expected for common copy number variants is only likely to be achieved through large sample sizes. If common copy number variants exert effects on disease risk that are as modest as those often recorded for common SNPs, then at least several thousand cases and controls will be needed, and study size will need to increase if typing of copy number variation has an appreciable measurement error.[97]

Do copy number variants need to be measured directly (panel 3) or can they be measured adequately through linkage disequilibrium (tagging) from nearby SNPs? Estimation of the proportion of copy number variants across the genome that can be effectively tagged by neighbouring SNPs has varied substantially, which is unsurprising in view of differences in the accuracy of genotyping of copy number variation and the density of SNP genotyping.[9, 10, 11 and 12]

Most common copy number variants (allele frequency >5%) detected by McCarroll and colleagues[14] from an Affymetrix 6.0 array were in perfect linkage disequilibrium with a HapMap SNP. Although copy number variants were less well captured than were SNPs of an equivalent allele frequency, the mean pairwise r^2 as a function of distance was similar for both, suggesting that this reduced tagging was caused by a paucity of SNPs in the relevant regions, rather than recurrent mutations. Even on first-generation SNP arrays, 40–50% of

common copy number variants (allele frequency >5%) were tagged (r^2>0·8) and this proportion was increased with the newest arrays (about 65% for the Illumina 1M array).[14]

Of the common copy number variants detected with the Illumina Human 1M array by Cooper and colleagues,[38] about 50% were tagged by the newest arrays (54% for the Illumina 1M array). These estimates need to be refined, ideally with copy number variants ascertained from independent approaches, but evidence suggests that SNP arrays can capture a substantial proportion of simple deletions through linkage disequilibrium with neighbouring SNPs. However, the remainder of simple and almost all complex multiallelic copy number variants will not be adequately captured by tag SNPs. Comprehensive association studies will therefore need such variants to be measured directly (panel 3).

Statistical models to test association are usually straightforward compared with the complex statistical approaches used by calling algorithms for copy number variants. An additive genetic model is usually assumed in SNP association tests,[2] and this model would be an appropriate starting point for association studies of copy number variation to test dose dependency. Such a model assumes that the diploid copy number has an additive relation to the log odds ratio of disease, which is equivalent to a multiplicative effect on the odds ratio. Primary association tests do not need to be adjusted for a wide range of covariates in association studies of copy number variants or SNPs, since associations can be masked if the relevant covariate is an intermediate on the causal pathway. Furthermore, phenotypic and lifestyle characteristics do not confound associations between a disease and a genotype because the genotype is randomly assigned at meiosis (mendelian randomisation). Presentation of findings adjusted in this way leaves scope for data-derived selection of the most noteworthy model and can be misleading.[98]

Many issues that seem unique to studies of copy number variation are, in fact, special cases of problems encountered generally in epidemiological studies, such as misclassification.[93] For example, overlapping but non-identical copy number variants might be collectively assigned into one exposure category, on which tests of association between variation and disease would be based. This classification is correct when the variant breakpoints differ only because of measurement error; however, if the differences in breakpoints represent

separate biological events with differing phenotypic consequences, the resulting misclassification will bias the findings of any association test.

In SNP-based association studies, over-stringent quality thresholds for SNP genotype calling can cause differential misclassification of genotypes between cases and controls (differential bias).[99] Similarly, in the context of an association study of copy number variation, calling as missing data copy number variant calls that are less than certain could cause or exacerbate differential bias. Incorporation of this uncertainty into the association testing would be preferable, for example by weighting the call according to the posterior probability of the call. The importance of differential bias might decrease as the accuracy of assays for copy number variation improves, but until then, methods that take appropriate account of the uncertainty in calling will be advantageous. One approach to keep differential bias to a minimum is to combine calling and association testing into one statistical procedure.[100]

The choice of statistical threshold for declaring associations to be significant demands careful examination for copy number variation. In genome-wide association studies based on SNP genotypes, artifactual associations tend to be over-represented among the smallest p values.[3] Even when the distribution of p values does not indicate obvious sources of bias, the use of a very low p value threshold does not, on its own, provide any guarantee of causal association.[2] Presently, the number of copy number variants across the human genome is not known, and therefore a correction for multiple testing based on this number is not possible to define. A cautious approach would be to use a threshold for genome-wide association tests of copy number variation similar to those used for SNPs (eg, $p<5\times10^{-7}$).[3] By comparison, a less-stringent threshold would be appropriate for candidate gene studies, provided that the selective reporting of only significant associations can be kept to a minimum.

In either case, results should be interpreted in view of the power of the study and existing evidence. The false-positive report probability can also be a helpful adjunct for reporting of study findings since this value explicitly acknowledges existing evidence and the power of the study.[101] Replication is central to assessment of the robustness of results from genome-wide SNP association studies, and therefore association studies of copy number variation are likely to be similar in this respect. If replication is not feasible, investigators should be

encouraged to report findings fully to allow broad replication efforts. Furthermore, replication studies should ideally be powered to detect a smaller effect size than that reported in the original study.[102]

Future prospects

Rapid advances in our understanding of the nature of copy number variation in the human genome have been made in the past 5 years. The resolution of emerging maps of copy number variation seems to be improving to one probe per 50 bp.[37] Genome-wide and candidate region approaches to detect copy number variation are also improving, together with statistical calling algorithms that can take account of the information provided by maps of copy number variation. We seem to be on the threshold of new discoveries in association studies of copy number variation and of developing our understanding of the contributions of sequence and structural variations to predisposition to common diseases.

Search strategy and selection criteria

We searched PubMed with the terms “copy number variation” and “structural variation” for articles published in English between Jan 1, 2004, and Nov 3, 2008. We received regular updates from PubMed of papers published with these terms. We also searched related articles and followed up articles cited in reference lists. Review articles were sometimes selected in preference to original articles because of space constraints. Precise search terms were used to search PubMed for detailed articles on specific traits and methods. The date of the last search was Nov 30, 2008. Information about relevant database entries was updated during proof-reading (April 30, 2009).

Contributors

LVW, JALA, and MDT reviewed published work and participated in writing of the report. LVW designed the figures.

Conflicts of interest

We declare that we have no conflicts of interest.

Acknowledgments

We thank Ammar Al-Chalabi, Cother Hajat, Ed Hollox, and James R Lupski for providing comments on the report; and anonymous reviewers for their helpful comments and suggestions. MDT holds a Medical Research Council Clinician Scientist Fellowship (G0501942).

Panel 1. Database of Genomic Variants

The Database of Genomic Variants[4] is the most up-to-date repository for results of comprehensive genome-wide screening for copy number variation from peer-reviewed studies. Duplications, insertions, or deletions of more than 1 kb of DNA are classed as copy number variants and those of 100–1000 bp as indels.

In April, 2009, the database contained more than 6558 loci of copy number variants and the database is regularly updated with new studies and perspectives. Data from some of the oldest studies have been removed, but the data are otherwise not subject to editorial screening. Furthermore, interpretation of the database presents challenges even for knowledgeable researchers because the contributing studies use a variety of methods with different resolutions, error rates, and genome coverage.

Crucially, the database can provide inaccurate estimates of variant boundaries and size. For example, array comparative genomic hybridisation is often used to detect changes in copy number but usually overestimates the size of the variant.[14 and 43] Interpretation of variants with close or overlapping boundaries presents additional challenges, especially when their size has been overestimated and the locus could contain more than one underlying copy number variant.

The most satisfactory records of copy number variants will be those for which the structural basis has been defined, ideally based on the DNA sequence, and for which frequencies have been established by population surveys.

Panel 2. Genomic disorders

Genomic disorders are diseases attributable to genomic rearrangements that occur as a result of specific genome architecture.[15] Lupski[16] described four examples that take place on the proximal short arm of chromosome 17. This region is rich in segmental duplications, which are thought to predispose to genomic rearrangements.

Charcot–Marie–Tooth neuropathy 1A (CMT1A) is a well characterised genomic disorder, and most cases result from tandem duplication (17)(p12) of a 1·4 million bp (1·4 Mb) region. Deletion of the same region causes hereditary neuropathy with liability to pressure palsies (HNPP).[44] Although this region contains many genes, the *PMP22* gene that encodes a component of the peripheral nervous system myelin[45] is the dose-sensitive gene responsible for the clinical phenotype.[16]

Deletion (17)(p11.2) of a 4 Mb region that excludes the *PMP22* gene causes Smith–Magenis syndrome, which is characterised by developmental delay and neurobehavioural abnormalities.[46] By contrast, duplication of the same region causes Potocki–Lupski syndrome, which has clinical features distinct from Smith–Magenis syndrome that include infantile hypotonia, failure to thrive, cardiovascular anomalies, and autism spectrum disorder.[47] These duplications and deletions affect the retinoic acid inducible 1 gene (*RAI1*). Smith–Magenis syndrome has also been documented with point mutations in *RAI1*, showing that this gene is responsible for the clinical effects.[48] Smith–Magenis and Potocki–Lupski syndromes have also been documented with deletions and duplications of differing sizes and breakpoints, all of which affect *RAI1*.

Panel 3. Use of genome-wide SNP arrays and hybrid arrays to identify copy number variants

Association studies of copy number variation have been undertaken with commercial genome-wide SNP assays marketed by Affymetrix (Santa Clara, CA, USA)[20, 23 and 27] and Illumina (San Diego, CA, USA),[23, 26 and 33] and with hybrid arrays composed of both SNP probes and non-polymorphic probes.[14, 23 and 38] These studies measure the combined signal (r) from both alleles at a particular SNP, and express the value as a $\log_2$ r ratio between the recorded intensity and the expected intensity; the expected intensity is derived from the average intensity of the genotype clusters (not a reference sample, as in array comparative genomic hybridisation). Additionally, some studies also use the ratio of fluorescence signals between allelic probes, since these ratios would be expected to be 0, 0·5, and 1·0 in the absence of copy number variation. A range of normalisation strategies (eg, adjustment for possible batch effects, adjustment for restriction fragment length, corrections for GC nucleotide content or other variations in genome architecture, and quantile normalisation), and a plethora of calling algorithms for copy number variants have been used for these data.[23, 26, 33, 70, 71, 72, 73, 74 and 76]

Some algorithms are designed to use additional data from non-polymorphic probes that are included in several genome-wide assays, such as Affymetrix 5·0 and 6·0 arrays.[23 and 75] Many of the calling algorithms use a hidden Markov model, which segments the continuous data into several predefined and biologically meaningful discrete states. The data are assumed to be generated by a stochastic process defined by a specific number of hidden states;[77] for example, QuantiSNP assumes six states of differing copy number and genotype combinations.[70]

Generally, association studies based on either genome-wide or candidate region data cannot produce accurate copy number estimates for the large number of samples needed for case–control association studies.[78] Both new genotyping arrays and use of a-priori data from updated maps of copy number variation (about variant locations and breakpoints) have improved accuracy.[14 and 75] However, typing in most published association studies of

copy number variation is subject to substantial measurement error; validation and replication of findings is crucial.

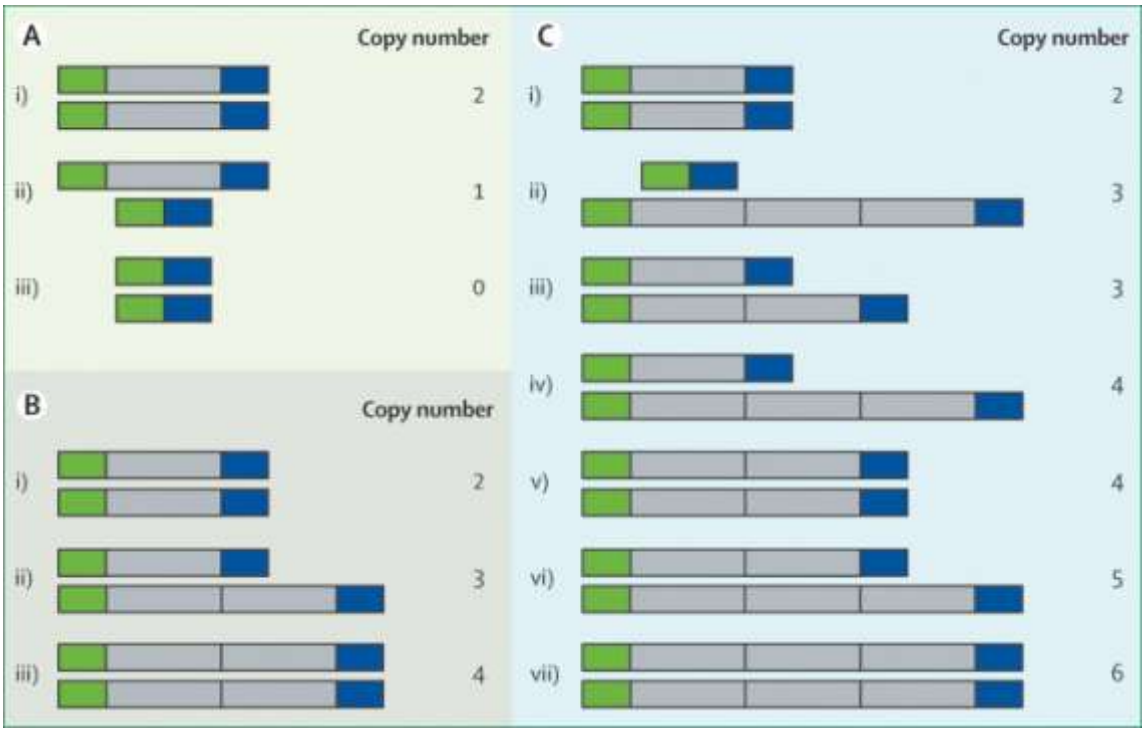

Figure 1. Diallelic and multiallelic copy number variation

Diallelic locus (grey) and flanking loci (green and blue) with variation caused by (A) deletion and (B) duplication, each showing the locus with (i) normal copy number, (ii) heterozygous modification, and (iii) homozygous modification. (C) Multiallelic locus showing (i) normal copy number, (ii) multiple rounds of duplication on one chromosome and a deletion on the homologous chromosome, (iii) duplication on one chromosome and no deletion on the homologous chromosome, (iv) multiple rounds of duplication on one chromosome and no deletion on the homologous chromosome, (v) one round of duplication on each chromosome, (vi) one round of duplication on one chromosome and multiple rounds of duplication on the homologous chromosome, and (vii) multiple rounds of duplication on both chromosomes. Multiallelic assays measure diploid copy number and therefore cannot distinguish between (ii) and (iii), or (iv) and (v). Note that although duplications are usually assumed to be contiguous, this might not always be the case.

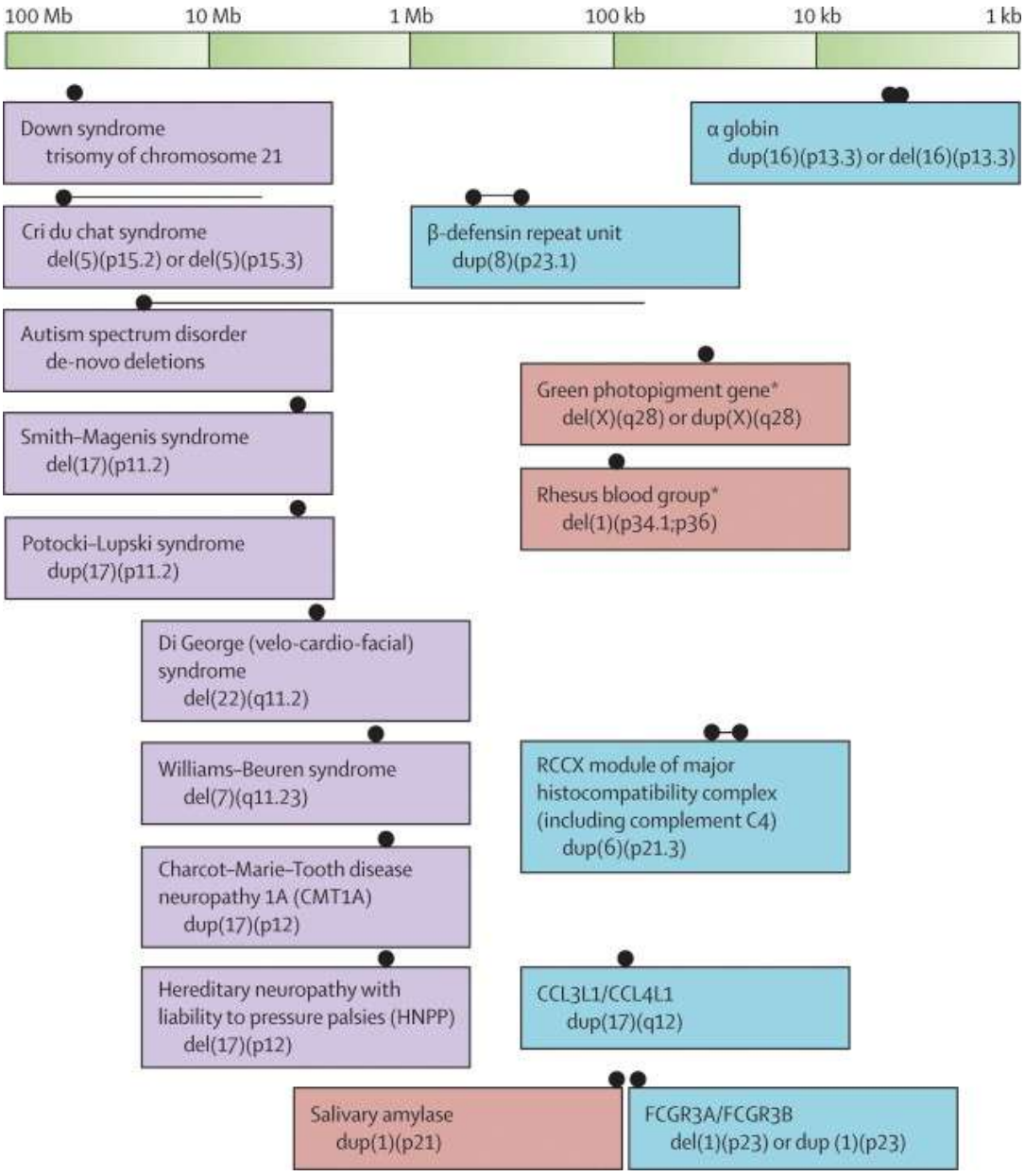

Figure 2. Size range of structural genomic variationStructural variants with obvious clinical effects (purple), underlying normal phenotypic variation (red), and implicated in normal phenotypic variation with evidence for disease association (blue). Black circles and lines above the boxes show the size of the variants on the log scale. No lower limit is defined for autism or Cri du chat syndrome. For some disorders, causes other than structural variation have also been identified—eg, Smith–Magenis and Potocki–Lupski syndromes, rhesus negative blood group, and CMT1A are also caused by sequence variation. The size range shown for β-defensin indicates uncertainty about the true size of the repeat unit. *Variants can also have obvious clinical effects.

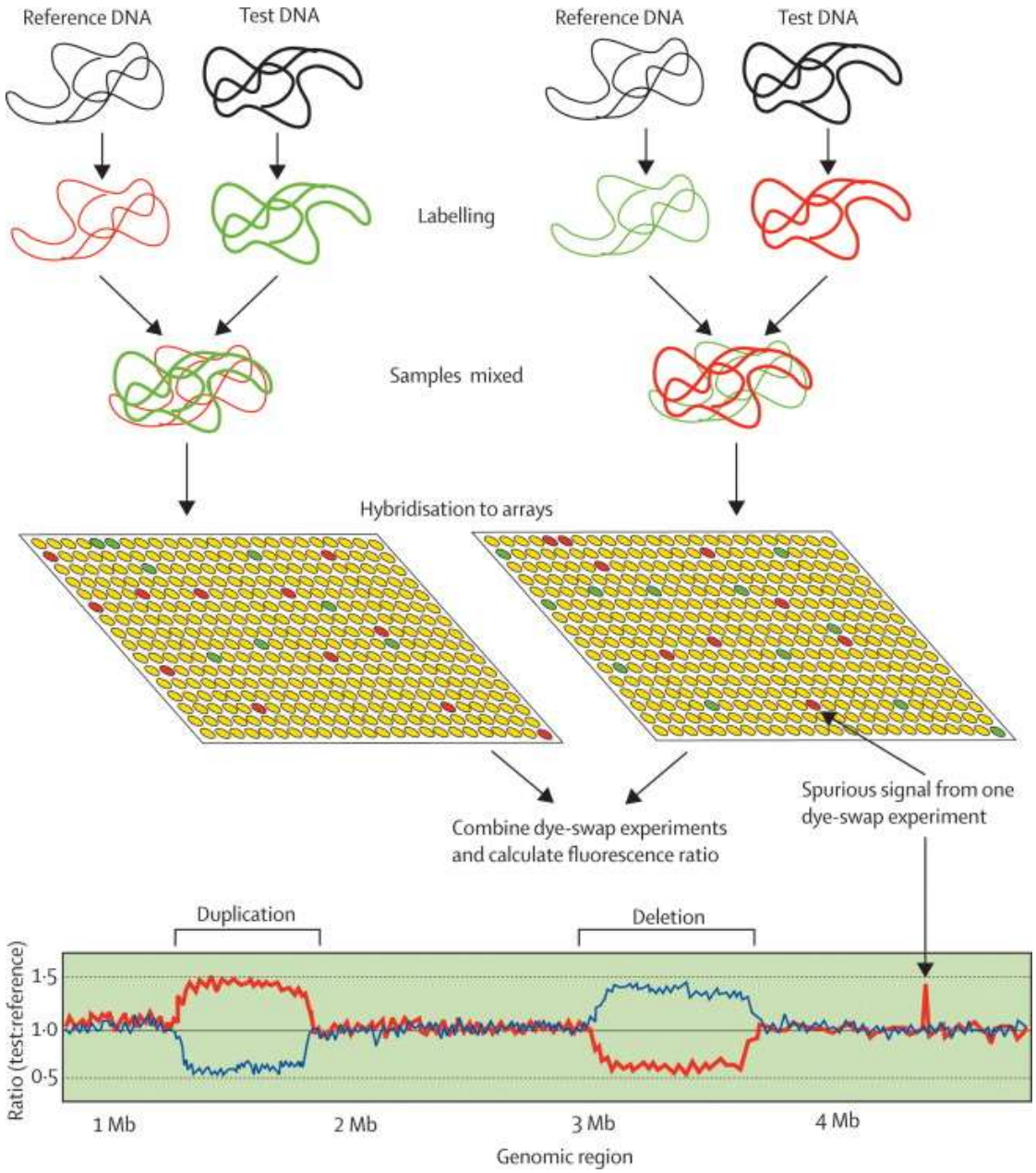

Figure 3. Array comparative genomic hybridisationTest and reference DNA samples are differentially labelled, mixed, and passed over a target array of probes (eg, BAC clones or oligonucleotides) containing DNA fragments from across the whole human genome. The experiment is often repeated with reversal of the test and reference dyes to detect dye effects or identify spurious signals. DNA samples hybridise with their corresponding probe, and the ratio of fluorescence from each probe (test:reference) is used to detect regions that vary in copy number between the test and the reference sample (red line: original hybridisation; blue line: dye-swapped hybridisation). Equal copy number for both the test and reference DNA is identified by equal binding, resulting in a ratio of one. Duplication in a genomic region of the test sample is identified by an increased ratio, and a deletion by a decreased ratio, but a deletion in the test sample is indistinguishable from a duplication in the reference sample. These ratios are usually converted to $\log_2$ scale for further analysis. Adapted from Feuk and colleagues[17] with permission from *Nature Reviews Genetics*.

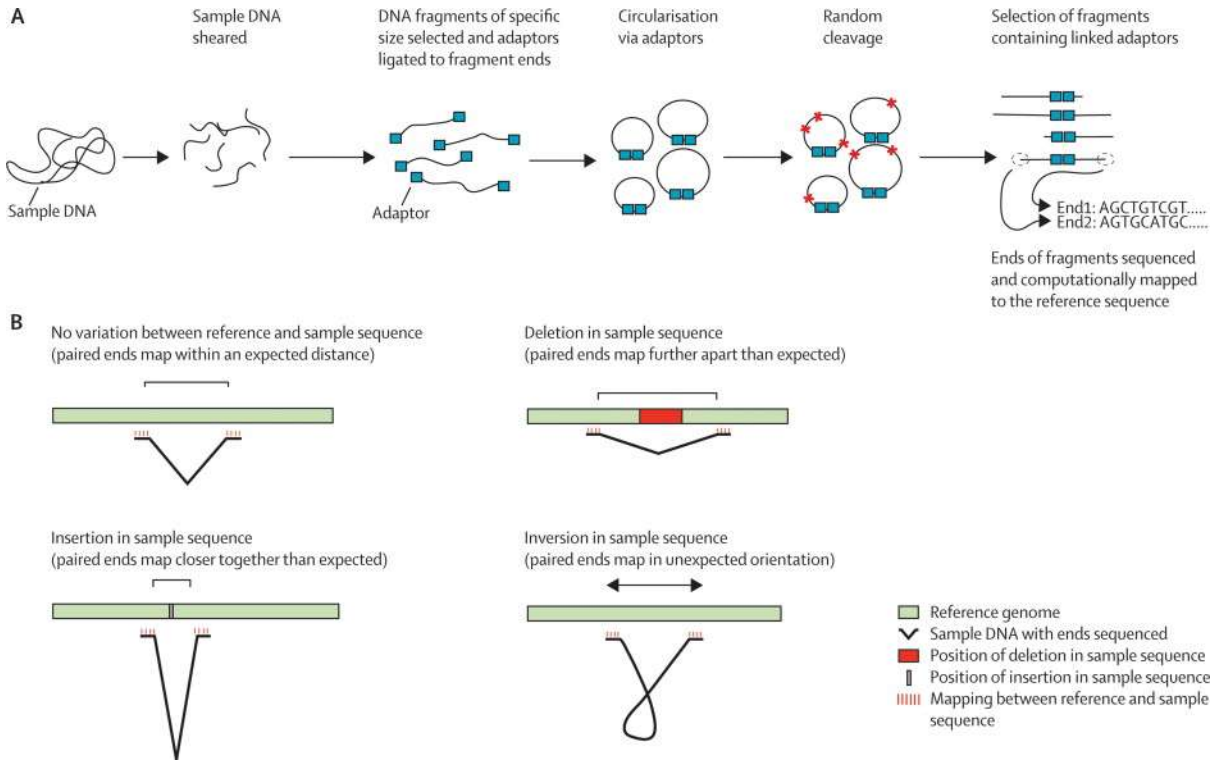

Figure 4. End-pair sequencing to identify copy number variation between a sample and reference genome sequence(A) Preparation and sequencing of sample DNA and (B) example results.

References

1 KA Frazer, DG Ballinger and DR Cox *et al.*, A second generation human haplotype map of over 3·1 million SNPs, *Nature* **449** (2007), pp. 851–861.
2 MI McCarthy, GR Abecasis and LR Cardon *et al.*, Genome-wide association studies for complex traits: consensus, uncertainty and challenges, *Nat Rev Genet* **9** (2008), pp. 356–369.
3 The Wellcome Trust Case Control Consortium, Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls, *Nature* **447** (2007), pp. 661–678.
4 AJ Iafrate, L Feuk and MN Rivera *et al.*, Detection of large-scale variation in the human genome, *Nat Genet* **36** (2004), pp. 949–951.
5 J Sebat, B Lakshmi and J Troge *et al.*, Large-scale copy number polymorphism in the human genome, *Science* **305** (2004), pp. 525–528.
6 AJ Sharp, S Hansen and RR Selzer *et al.*, Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome, *Nat Genet* **38** (2006), pp. 1038–1042.
7 E Tuzun, AJ Sharp and JA Bailey *et al.*, Fine-scale structural variation of the human genome, *Nat Genet* **37** (2005), pp. 727–732.
8 DF Conrad, TD Andrews, NP Carter, ME Hurles and JK Pritchard, A high-resolution survey of deletion polymorphism in the human genome, *Nat Genet* **38** (2006), pp. 75–81.
9 DA Hinds, AP Kloek, M Jen, X Chen and KA Frazer, Common deletions and SNPs are in linkage disequilibrium in the human genome, *Nat Genet* **38** (2006), pp. 82–85. 10 DP Locke, AJ Sharp and SA McCarroll *et al.*, Linkage disequilibrium and heritability of copy number polymorphisms within duplicated regions of the human genome, *Am J Hum Genet* **79** (2006), pp. 275–290.
11 SA McCarroll, TN Hadnott and GH Perry *et al.*, Common deletion polymorphisms in the human genome, *Nat Genet* **38** (2006), pp. 86–92.
12 R Redon, S Ishikawa and KR Fitch *et al.*, Global variation in copy number in the human genome, *Nature* **444** (2006), pp. 444–454. 13 JM Kidd, GM Cooper and WF Donahue *et al.*, Mapping and sequencing of structural variation from eight human genomes, *Nature* **453** (2008), pp. 56–64.

14 SA McCarroll, FG Kuruvilla and JM Korn *et al.*, Integrated detection and population-genetic analysis of SNPs and copy number variation, *Nat Genet* **40** (2008), pp. 1166–1174.
15 JR Lupski, Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits, *Trends Genet* **14** (1998), pp. 417–422.
16 JR Lupski, Genomic rearrangements and sporadic disease, *Nat Genet* **39** (2007), pp. S43–S47.
17 L Feuk, AR Carson and SW Scherer, Structural variation in the human genome, *Nat Rev Genet* **7** (2006), pp. 85–97.
18 JA Buchanan and SW Scherer, Contemplating effects of genomic structural variation, *Genet Med* **10** (2008), pp. 639–647.
19 RA Kumar, S KaraMohamed and J Sudi *et al.*, Recurrent 16p11.2 microdeletions in autism, *Hum Mol Genet* **17** (2008), pp. 628–638.
20 CR Marshall, A Noor and JB Vincent *et al.*, Structural variation of chromosomes in autism spectrum disorder, *Am J Hum Genet* **82** (2008), pp. 477–488.
21 J Sebat, B Lakshmi and D Malhotra *et al.*, Strong association of de novo copy number mutations with autism, *Science* **316** (2007), pp. 445–449.
22 P Szatmari, AD Paterson and L Zwaigenbaum *et al.*, Mapping autism risk loci using genetic linkage and chromosomal rearrangements, *Nat Genet* **39** (2007), pp. 319–328.
23 LA Weiss, Y Shen and JM Korn *et al.*, Association between microdeletion and microduplication at 16p11.2 and autism, *N Engl J Med* **358** (2008), pp. 667–675.
24 H Stefansson, D Rujescu and S Cichon *et al.*, Large recurrent microdeletions associated with schizophrenia, *Nature* **455** (2008), pp. 232–236.
25 The International Schizophrenia Consortium, Rare chromosomal deletions and duplications increase risk of schizophrenia, *Nature* **455** (2008), pp. 237–241.
26 T Walsh, JM McClellan and SE McCarthy *et al.*, Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia, *Science* **320** (2008), pp. 539–543.
27 B Xu, JL Roos, S Levy, EJ van Rensburg, JA Gogos and M Karayiorgou, Strong association of de novo copy number mutations with sporadic schizophrenia, *Nat Genet* **30** (2008), p. 30.
28 SA McCarroll, A Huett and P Kuballa *et al.*, Deletion polymorphism upstream of *IRGM* associated with altered IRGM expression and Crohn's disease, *Nat Genet* **24** (2008), p. 24.
29 EJ Hollox, U Huffmeier and PL Zeeuwen *et al.*, Psoriasis is associated with increased beta-defensin genomic copy number, *Nat Genet* **40** (2008), pp. 23–25.

30 TJ Aitman, R Dong and TJ Vyse *et al.*, Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans, *Nature* **439** (2006), pp. 851–855.
31 M Fanciulli, PJ Norsworthy and E Petretto *et al.*, *FCGR3B* copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity, *Nat Genet* **39** (2007), pp. 721–723.
32 Y Yang, EK Chung and YL Wu *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** (2007), pp. 1037–1054.
33 HM Blauw, JH Veldink and MA van Es *et al.*, Copy number variation in sporadic amyotrophic lateral sclerosis: a genome-wide screen, *Lancet Neurol* **7** (2008), pp. 319–326.
34 E Gonzalez, H Kulkarni and H Bolivar *et al.*, The influence of *CCL3L1* gene-containing segmental duplications on HIV-1/AIDS susceptibility, *Science* **307** (2005), pp. 1434–1440. 35 L Kuhn, DB Schramm and S Donninger *et al.*, African infants' *CCL3* gene copies influence perinatal HIV transmission in the absence of maternal nevirapine, *AIDS* **21** (2007), pp. 1753–1761.
36 SK Ahuja, H Kulkarni and G Catano *et al.*, *CCL3L1-CCR5* genotype influences durability of immune recovery during antiretroviral therapy of HIV-1-infected individuals, *Nat Med* **14** (2008), pp. 413–420.
37 Conrad DF, Pinto L, Feuk L, et al. A comprehensive map of common copy number variation at 50bp resolution, and resulting biological insights. 58th Annual Meeting of The American Society of Human Genetics; Philadelphia, PA, USA; Nov 11–15, 2008. Abstr 151.
38 GM Cooper, T Zerr, JM Kidd, EE Eichler and DA Nickerson, Systematic assessment of copy number variant detection via genome-wide SNP genotyping, *Nat Genet* **40** (2008), pp. 1199–1203.
39 SA McCarroll, Extending genome-wide association studies to copy number variation, *Hum Mol Genet* **17** (2008), pp. R135–R142.
40 EJ Hollox, JA Armour and JC Barber, Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster, *Am J Hum Genet* **73** (2003), pp. 591–600.
41 GH Perry, NJ Dominy and KG Claw *et al.*, Diet and the evolution of human amylase gene copy number variation, *Nat Genet* **39** (2007), pp. 1256–1260.

42 ND Avent, PG Martin and SS Armstrong-Fisher *et al.*, Evidence of genetic diversity underlying Rh D-, weak D (Du), and partial D phenotypes as determined by multiplex polymerase chain reaction analysis of the *RHD* gene, *Blood* **89** (1997), pp. 2568–2577.
43 GH Perry, A Ben-Dor and A Tsalenko *et al.*, The fine-scale and complex architecture of human copy number variation, *Am J Hum Genet* **24** (2008), p. 24.
44 K Inoue, K Dewar and N Katsanis *et al.*, The 1·4-Mb *CMT1A* duplication/*HNPP* deletion genomic region reveals unique genome architectural features and provides insights into the recent evolution of new genes, *Genome Res* **11** (2001), pp. 1018–1033.
45 CA Wise, CA Garcia and SN Davis *et al.*, Molecular analyses of unrelated Charcot-Marie-Tooth (CMT) disease patients suggest a high frequency of the *CMTIA* duplication, *Am J Hum Genet* **53** (1993), pp. 853–863.
46 KS Chen, P Manian and T Koeuth *et al.*, Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome, *Nat Genet* **17** (1997), pp. 154–163.
47 L Potocki, W Bi and D Treadwell-Deering *et al.*, Characterization of Potocki-Lupski syndrome (dup(17)(p11.2p11.2)) and delineation of a dosage-sensitive critical interval that can convey an autism phenotype, *Am J Hum Genet* **80** (2007), pp. 633–649.
48 SH Elsea and S Girirajan, Smith-Magenis syndrome, *Eur J Hum Genet* **16** (2008), pp. 412–421.
49 AJ Sharp, DP Locke and SD McGrath *et al.*, Segmental duplications and copy number variation in the human genome, *Am J Hum Genet* **77** (2005), pp. 78–88.
50 L Pentao, CA Wise, AC Chinault, PI Patel and JR Lupski, Charcot-Marie-Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1·5 Mb monomer unit, *Nat Genet* **2** (1992), pp. 292–300.
51 P Stankiewicz and JR Lupski, Genome architecture, rearrangements and genomic disorders, *Trends Genet* **18** (2002), pp. 74–82.
52 K Holloway, VE Lawson and AJ Jeffreys, Allelic recombination and de novo deletions in sperm in the human beta-globin gene region, *Hum Mol Genet* **15** (2006), pp. 1099–1111.
53 KW Lam and AJ Jeffreys, Processes of copy number change in human DNA: the dynamics of {alpha}-globin gene deletion, *Proc Natl Acad Sci USA* **103** (2006), pp. 8921–8927.
54 KW Lam and AJ Jeffreys, Processes of de novo duplication of human alpha-globin genes, *Proc Natl Acad Sci USA* **104** (2007), pp. 10950–10955.

55 JA Lee, CM Carvalho and JR Lupski, A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders, *Cell* **131** (2007), pp. 1235–1247.
56 SJ Lindsay, M Khajavi, JR Lupski and ME Hurles, A chromosomal rearrangement hotspot can be identified from population genetic variation and is coincident with a hotspot for allelic recombination, *Am J Hum Genet* **79** (2006), pp. 890–902.
57 DJ Turner, M Miretti and D Rajan *et al.*, Germline rates of de novo meiotic deletions and duplications causing several genomic disorders, *Nat Genet* **40** (2008), pp. 90–95. 58
58 Bruder, A Piotrowski and AA Gijsbers *et al.*, Phenotypically concordant and discordant monozygotic twins display different DNA copy number-variation profiles, *Am J Hum Genet* **14** (2008), p. 14.
59 M Ingelman-Sundberg, SC Sim, A Gomez and C Rodriguez-Antona, Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects, *Pharmacol Ther* **116** (2007), pp. 496–526.
60 SS Deeb, The molecular basis of variation in human color vision, *Clin Genet* **67** (2005), pp. 369–377.
61 BE Stranger, MS Forrest and M Dunning *et al.*, Relative impact of nucleotide and copy number variation on gene expression phenotypes, *Science* **315** (2007), pp. 848–853.
62 ME Hurles, ET Dermitzakis and C Tyler-Smith, The functional impact of structural variation in humans, *Trends Genet* **28** (2008), p. 28.
63 G Koren, J Cairns, D Chitayat, A Gaedigk and SJ Leeder, Pharmacogenetics of morphine poisoning in a breastfed neonate of a codeine-prescribed mother, *Lancet* **368** (2006), p. 704.
64 DQ Nguyen, CP Webber, J Hehir-Kwa, R Pfundt, J Veltman and CP Ponting, Reduced purifying selection prevails over positive selection in human copy number variant evolution, *Genome Res* **7** (2008), p. 7.
65 S Giglio, KW Broman and N Matsumoto *et al.*, Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements, *Am J Hum Genet* **68** (2001), pp. 874–883.
66 S Giglio, V Calvari and G Gregato *et al.*, Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation, *Am J Hum Genet* **71** (2002), pp. 276–285.

67 H Sugawara, N Harada and T Ida *et al.*, Complex low-copy repeats associated with a common polymorphic inversion at human chromosome 8p23, *Genomics* **82** (2003), pp. 238–244.
68 KW Broman, N Matsumoto and S Giglio *et al.*, Common long human inversion polymorphism on chromosome 8p. In: DR Goldstein, Editor, *Science and statistics: a festschrift for Terry Speed*, Institute of Mathematical Statistics, Beachwood, OH (2003), pp. 237–245.
69 JO Korbel, AE Urban and JP Affourtit *et al.*, Paired-end mapping reveals extensive structural variation in the human genome, *Science* **318** (2007), pp. 420–426.
70 S Colella, C Yau and JM Taylor *et al.*, QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data, *Nucleic Acids Res* **35** (2007), pp. 2013–2025.
71 D Komura, F Shen and S Ishikawa *et al.*, Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays, *Genome Res* **16** (2006), pp. 1575–1584.
72 T Laframboise, D Harrington and BA Weir, PLASQ: a generalized linear model-based procedure to determine allelic dosage in cancer cells from SNP array data, *Biostatistics* **8** (2007), pp. 323–336.
73 Y Nannya, M Sanada and K Nakazaki *et al.*, A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays, *Cancer Res* **65** (2005), pp. 6071–6079.
74 K Wang, M Li and D Hadley *et al.*, PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data, *Genome Res* **5** (2007), p. 5.
75 JM Korn, FG Kuruvilla and SA McCarroll *et al.*, Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs, *Nat Genet* **40** (2008), pp. 1253–1260.
76 X Zhao, C Li and JG Paez *et al.*, An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays, *Cancer Res* **64** (2004), pp. 3060–3071.

77 N Day, A Hemmaplardh, RE Thurman, JA Stamatoyannopoulos and WS Noble, Unsupervised segmentation of continuous genomic data, *Bioinformatics* **23** (2007), pp. 1424–1426.
78 SA McCarroll, Copy number analysis goes more than skin deep, *Nat Genet* **40** (2008), pp. 5–6.
79 J Flint, RM Harding, AJ Boyce and JB Clegg, The population genetics of the haemoglobinopathies, *Baillieres Clin Haematol* **11** (1998), pp. 1–51.
80 J Flint, AV Hill and DK Bowden *et al.*, High frequencies of alpha-thalassaemia are the result of natural selection by malaria, *Nature* **321** (1986), pp. 744–750.
81 ND Avent and ME Reid, The Rh blood group system: a review, *Blood* **95** (2000), pp. 375–387.
82 AH Fielder, MJ Walport and JR Batchelor *et al.*, Family study of the major histocompatibility complex in patients with systemic lupus erythematosus: importance of null alleles of C4A and C4B in determining disease susceptibility, *Br Med J (Clin Res Ed)* **286** (1983), pp. 425–428.
83 AE Hughes, N Orr, H Esfandiary, M Diaz-Torres, T Goodship and U Chakravarthy, A common *CFH* haplotype, with deletion of *CFHR1* and *CFHR3*, is associated with lower risk of age-related macular degeneration, *Nat Genet* **38** (2006), pp. 1173–1177.
84 PF Zipfel, M Edey and S Heinen *et al.*, Deletion of complement factor H-related genes *CFHR1* and *CFHR3* is associated with atypical hemolytic uremic syndrome, *PLoS Genet* **3** (2007), p. e41.
85 KL Spencer, MA Hauser, LM Olson, S Schmidt, WK Scott and P Gallins *et al.*, Deletion of *CFHR3* and *CFHR1* genes in age-related macular degeneration, *Hum Mol Genet* **17** (2008), pp. 971–977.
86 F Niyonsaba, H Ogawa and I Nagaoka, Human beta-defensin-2 functions as a chemotactic agent for tumour necrosis factor-alpha-treated human neutrophils, *Immunology* **111** (2004), pp. 273–281.
87 K Fellermann, DE Stange and E Schaeffeler *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** (2006), pp. 439–448.
88 JR Townson, LF Barcellos and RJ Nibbs, Gene copy number regulates the production of the human chemokine CCL3-L1, *Eur J Immunol* **32** (2002), pp. 3016–3026. 89 W Shao, J Tang

and W Song *et al.*, *CCL3L1* and *CCL4L1*: variable gene copy number in adolescents with and without human immunodeficiency virus type 1 (HIV-1) infection, *Genes Immun* **8** (2007), pp. 224–231.
90 M Mamtani, B Rovin and R Brey *et al.*, *CCL3L1* gene-containing segmental duplications and polymorphisms in *CCR5* affect risk of systemic lupus erythematosus, *Ann Rheum Dis* **67** (2008), pp. 1076–1083.
91 C McKinney, ME Merriman and PT Chapman *et al.*, Evidence for an influence of chemokine ligand 3-like 1 (*CCL3L1*) gene copy number on susceptibility to rheumatoid arthritis, *Ann Rheum Dis* **67** (2008), pp. 409–413.
92 NM Williams, MC O'Donovan and MJ Owen, Chromosome 22 deletion syndrome and schizophrenia, *Int Rev Neurobiol* **73** (2006), pp. 1–27.
93 DJ Balding, A tutorial on statistical methods for population association studies, *Nat Rev Genet* **7** (2006), pp. 781–791.
94 AT Hattersley and MI McCarthy, What makes a good genetic association study?, *Lancet* **366** (2005), pp. 1315–1323.
95 SW Scherer, C Lee and E Birney *et al.*, Challenges and standards in integrating surveys of structural variation, *Nat Genet* **39** (2007), pp. S7–15.
96 KT Zondervan and LR Cardon, Designing candidate gene and genome-wide case-control association studies, *Nat Protoc* **2** (2007), pp. 2492–2501.
97 PR Burton, AL Hansell and I Fortier *et al.*, Size matters: just how big is BIG? Quantifying realistic sample size requirements for human genome epidemiology, *Int J Epidemiol* **38** (2009), pp. 263–273.
98 GD Smith, DA Lawlor, R Harbord, N Timpson, I Day and S Ebrahim, Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology, *PLoS Med* **4** (2007), p. e352.
99 DG Clayton, NM Walker and DJ Smyth *et al.*, Population structure, differential bias and genomic control in a large-scale, case-control association study, *Nat Genet* **37** (2005), pp. 1243–1246.
100 C Barnes, V Plagnol and T Fitzgerald *et al.*, A robust statistical method for case-control association testing with copy number variation, *Nat Genet* **40** (2008), pp. 1245–1252.

101 S Wacholder, S Chanock, M Garcia-Closas, L El Ghormli and N Rothman, Assessing the probability that a positive report is false: an approach for molecular epidemiology studies, *J Natl Cancer Inst* **96** (2004), pp. 434–442.
102 KE Lohmueller, CL Pearce and M Pike *et al.*, Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease, *Nat Genet* **33** (2003), pp. 177–182.