Structural variation in the human genome

Lars Feuk, Andrew R. Carson and Stephen W. Scherer

Abstract | The first wave of information from the analysis of the human genome revealed SNPs to be the main source of genetic and phenotypic human variation. However, the advent of genome-scanning technologies has now uncovered an unexpectedly large extent of what we term 'structural variation' in the human genome. This comprises microscopic and, more commonly, submicroscopic variants, which include deletions, duplications and large-scale copy-number variants — collectively termed copy-number variants or copy-number polymorphisms — as well as insertions, inversions and translocations. Rapidly accumulating evidence indicates that structural variants can comprise millions of nucleotides of heterogeneity within every genome, and are likely to make an important contribution to human diversity and disease susceptibility.

Aneuploidy
The presence of an abnormal number of chromosomes within a cell.

Heteromorphism
A microscopically visible region of a chromosome that varies in size, morphology or staining properties. They include euchromatic and non-euchromatic variation, such as satellite–satellite stalk variation and heterochromatic variation (centromeres and other C-band positive regions).

The Centre for Applied Genomics and Program in Genetics and Genomic Biology, The Hospital for Sick Children and Department of Molecular and Medical Genetics, University of Toronto, MaRS Centre — East Tower, 101 College Street, Room 14-701, Ontario M5G 1L7, Canada.
Correspondence to S.W.S.
e-mail:
steve@genet.sickkids.on.ca
doi:10.1038/nrg1767

A striking observation from the analysis of the human genome[1–3] is the extent of DNA-sequence similarity among individuals from around the world: any two humans are thought to be about 99.9% identical in their DNA sequence[4,5]. It is therefore through studies of a small fraction of the genome — which constitutes the genetic variation between individuals — that insights into phenotypic variation and disease susceptibility can be gained.

Decades before the availability of sequencing technology, the first differences observed in our genetic composition were mainly rare changes in the quantity and structure of chromosomes. These included aneuploidies[6–8], rearrangements[9–14] (which were often associated with disease), heteromorphisms[15–19] and fragile sites[20], all of which were large enough to be identified using a microscope. We define such variants, which are ~3 Mb or more in size, as microscopic structural variants (BOX 1). Subsequently, with the advent of molecular biology, and DNA sequencing in particular, smaller and more abundant alterations were observed. Such differences include SNPs, various repetitive elements that involve relatively short DNA sequences (for example, micro- and minisatellites), and small (usually <1 kb) insertions, deletions, inversions and duplications[21]. It was presumed that these small-scale variants constitute most genetic variation; for example, estimates predict that there are at least 10 million SNPs within the human population[22], averaging 1 every 300 nucleotides among the ~3 billion nucleotide base pairs that constitute the genome of an individual.

Significant effort has been invested in characterizing human genetic variation at the karyotype[12,13] and nucleotide[23,24] level, but knowledge about variation in between these two extremes has been less extensive. However, in the past few years, the effective completion of the primary sequence of the human genome has underpinned the creation of new strategies and tools for the efficient assessment of genome composition. In particular, genome-scanning array technologies[25–27] and comparative DNA-sequence analyses[28,29] have begun to reveal DNA variation that involves segments that are smaller than those recognized microscopically, but larger than those that are readily detected by conventional sequence analysis. We define these variants, which range from ~1 kb to 3 Mb in size, as submicroscopic structural variants (BOX 1).

During the past year alone, fast on the heels of two breakthrough studies[26,27], hundreds of submicroscopic copy-number variants (CNVs)[28,30,31] and inversions[28,32,33] have been described in the human genome. These observations lead us to predict that structural genomic variants are as important as SNPs, short tandem repeats (STRs) and other small changes in their contribution to genome variation. Moreover, these types of variant can encompass millions of bases of DNA, containing entire genes and their regulatory regions[26–28,31,32]. Although structural variants in some genomic regions have no obvious phenotypic consequence[26–28,31], others influence gene dosage, which might cause genetic disease, either alone or in combination with other genetic or environmental factors[34].

Box 1 | **Structural variation definitions**

Structural variants are operationally defined as genomic alterations that involve segments of DNA that are larger than 1 kb, and can be microscopic or submicroscopic. Nothing is implied about their frequency, association with disease or phenotype, or lack thereof. Definitions of those types of structural variant that are the main focus of this review are given below; other alterations that can be considered structural variants include heteromorphisms, fragile sites, ring and marker chromosomes, isochromosomes, double minutes, and gene-conversion products. The term structural abnormality is often used if a structural variant is thought to be disease causing or is discovered as part of a disease study. Here we generally refer to smaller (<1 kb) variations or polymorphisms that involve the copy-number change of a segment of DNA as insertions or deletions (indels).

Types of structural variant

Copy-number variant (CNV). A segment of DNA that is 1 kb or larger and is present at a variable copy number in comparison with a reference genome. Classes of CNVs include insertions, deletions and duplications. This definition also includes large-scale copy-number variants, which are variants that involve segments of DNA ≥50 kb, allowing them to be detected by clone-based array comparative genome hybridization (array-CGH).

Copy-number polymorphism. A CNV that occurs in more than 1% of the population. Originally, this definition was used to refer to all CNVs[26].

Segmental duplication or low-copy repeat. A segment of DNA >1 kb in size that occurs in two or more copies per haploid genome, with the different copies sharing >90% sequence identity. They are often variable in copy number and can therefore also be CNVs.

Inversion. A segment of DNA that is reversed in orientation with respect to the rest of the chromosome. Pericentric inversions include the centromere, whereas paracentric inversions do not.

Translocation. A change in position of a chromosomal segment within a genome that involves no change to the total DNA content. Translocations can be intra- or inter-chromosomal.

Segmental uniparental disomy. Uniparental disomy describes the phenomenon in which a pair of homologous chromosomes in a diploid individual is derived from a single parent. With segmental uniparental disomy, only a portion of the chromosome pair is involved.

Here we focus mainly on submicroscopic structural variants, as they are the most recently discovered form of genetic variation in the human genome. In addition, despite their smaller size, their overall potential contribution to human genetic variation and disease might be expected to be higher than for microscopic variants, as they seem to occur at a higher frequency. We discuss the methods that have allowed the identification of submicroscopic structural variants, the recent studies that have begun to reveal the extent of their abundance in the human genome, and the implications of this newly discovered form of variation for studies of human diversity and disease. First, to provide a historic perspective and because larger structural variants might contribute to disease and diversity in a similar way, we provide a brief overview of microscopic structural variation.

Microscopic structural variation

Evidence for the basis of human genetic variation began with the ability to see individual chromosomes under the microscope. The earliest unbanded karyotypes consisted of relatively short, condensed chromosomes that were barely distinguishable from one another. However, aneuploidies, marker chromosomes and gross rearrangements could be discerned[6–9], and variation in Y-chromosome size was noted[35,36]. In solid-stained (unbanded) chromosomes, some heteromorphisms could be recognized, including secondary constrictions, satellite-region variants in acrocentric chromosomes, fragile sites and size variations in certain heterochromatic regions[20,37,38].

With the advent of various chromosome-banding techniques and the ability to work with elongated prometaphase chromosomes, more discrete structural abnormalities became apparent, mostly in disease samples. Reciprocal translocations, deletions, duplications, insertions and inversions could be discerned[10–14,39], and advances in fluorescence *in situ* hybridization (FISH) analysis allowed a more refined characterization of the extent of these variants (FIG. 1). Banding also revealed a much greater variety of heteromorphisms than had previously been suspected, including examples on chromosomes 1, 3, 4, 9, 13–16, 21, 22 and Y (REFS 15–19,40). At this resolution, the most common heteromorphisms that could be detected involved increases in length or inversions of the pericentric heterochromatic region of chromosome 9, at frequencies of about 8% and 1.5%, respectively[19,41]. Variation of this region might be attributed to unequal exchanges that involve repetitive sequences at recombination hotspots near the centromere[42]. In general, cytogenetically detectable heterochromatic variants have been considered clinically benign.

At the karyotype level, structurally abnormal variants — including translocations, inversions, deletions and duplications — are identified less frequently than aneuploidies. However, this might reflect an ascertainment bias, which has led to an underestimate of their numbers, in particular with respect to the submicroscopic structural variants that we discuss below. Bearing this caveat in mind, data that are typically cited indicate that structural abnormalities occur in about 1 out of 375 live births, with a quarter of these being unbalanced[37]. The risk for congenital abnormalities that are associated with such variants is 6.7% for reciprocal translocations and inversions, and 25.6% for all types of marker chromosome[12]. By contrast, 'normal' variants (those that are apparently non-disease-related, including heteromorphisms), are found in the range of about four to six per individual[43].

A recent review provides a comprehensive summary of data on unbalanced chromosome abnormalities, euchromatic variants and their associated phenotypes[44]. However, the literature that relates to how common heteromorphisms might influence non-disease phenotypes or contribute to common complex disease is otherwise sparse.

Identifying submicroscopic structural variants

In recent years, the development of both experimental (wet-laboratory) and computational strategies has allowed human structural genetic variation to be analysed at a higher resolution than the studies described above. These methods assay the genome in either a global (genome-wide) or a targeted manner, with varying degrees of resolution[45] (TABLE 1). Here we discuss those approaches that have had the greatest impact on recent discoveries of submicroscopic variants in the human genome.

Fragile site
A small break or a constriction of a chromosome that can be visualized under special cell-culture conditions. Some fragile sites are universal, others are normal structural variants, and two are associated with mental retardation syndromes (FRAXA and FRAXE).

Isochromosome
A chromosome that has two genetically and morphologically identical arms.

Double minute
Acentric, extra-chromosomally amplified chromatin, which usually contains a particular chromosomal segment or gene; common in cancer cells.

Marker chromosome
(Also known as an extra-structurally abnormal chromosome or 'supernumerary' chromosome.) Chromosomes that are seen in addition to the normal chromosome complement in fluorescence *in situ* hybridization experiments.

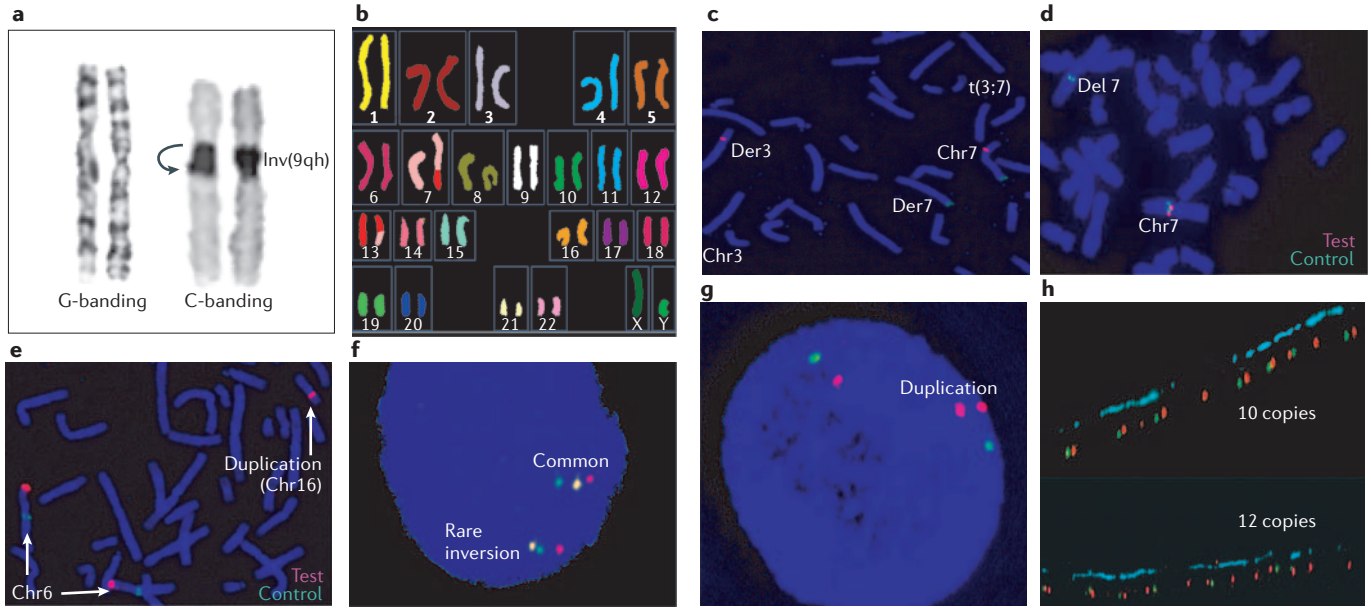

Figure 1 | **Cytogenetic detection and confirmation of structural variants. a** | Giemsa banding (G-banding) involves treating chromosomes with Giemsa stain, which produces a distinct pattern of bands for each chromosome, whereas centromere (C)-banding specifically stains centromeres. These methods can be used to identify an inversion heteromorphism, as shown here for inv(9qh). **b** | Spectral karyotyping is ideal for the identification of rearrangements that involve the exchange of DNA between chromosomes. Differentially labelled DNA probes for all chromosomes are used, making it possible to identify every chromosome in a single hybridization. The example shows the detection of a t(7;13) translocation. **c** | Here a cryptic t(3;7) translocation is detectable only by fluorescence *in situ* hybridization (FISH) that is carried out using metaphase chromosomes. Der indicates derivative chromosomes. **d**,**e** | Copy-number variant decrease (**d**) and increase (**e**) is detected by metaphase FISH. In panel **d** the green control probe is present in two copies on chromosome 7 (chr7), whereas the red probe shows a signal on only one of the homologous chromosome 7 copies. In panel **e** a gain of a hybridization signal on chromosome 16 is seen in addition to the signal that is present on both copies of chromosome 6. **f** | Three-colour FISH that was carried out using interphase nuclei shows that a 700-kb micro-inversion at 7p22 is polymorphic, as indicated by the change in order of BAC clones (which are labelled in different colours) between the two copies of the same chromosome that are present in the nucleus. Reproduced, with permission, from REF. 29 © (2005) Public Library of Science. **g** | Two-colour FISH reveals a large-scale copy-number variant, in this case a duplication. **h** | In high-resolution fibre FISH, probes are hybridized to mechanically stretched chromosome fibres. Here the 5′ and 3′ ends of the α-amylase gene are used as probes and labelled in different colours, allowing copy numbers to be counted directly. The top of the panel shows a chromosome that has 10 copies of this gene, which span ~300 kb. The bottom of the panel shows a different chromosome that has 12 copies, which span ~425 kb. Reproduced, with permission, from *Nature Genetics* REF. 27 © (2004) Macmillan Magazines Ltd.

Secondary constriction
A thin chromatic filament that connects a chromosomal satellite with the rest of the chromosome.

Acrocentric chromosome
A chromosome that has a centromere at or close to one end. Human acrocentric chromosomes are 13, 14, 15, 21 and 22.

Chromosome banding
A method of defining chromosome structure by staining with Giemsa and looking at the banding pattern in the heterochromatin of the centromeric regions.

Fluorescence *in situ* hybridization
A technique in which fluorescently labelled DNA probes are hybridized to interphase cells, metaphase chromosome preparations or DNA fibres, as a means to determine the presence and relative location of target sequences.

Unbalanced rearrangement
A genomic variant that involves gain or loss of DNA, such as deletion and duplication.

Euchromatic variant
A subset of cytogenetic heteromorphisms that involve microscopically visible variations of the euchromatic regions of chromosomes.

Genome-wide, array-based experimental approaches. Currently, the main approaches for identifying unbalanced structural variants are array-based analyses[25–27,30,31,46–53] and quantitative, primarily PCR-based assays[54–60]. Array-based comparative genome hybridization (array-CGH) approaches[25,61] (FIG. 2a) provide the most robust methods for carrying out genome-wide scans to find novel CNVs. These approaches use labelled fragments from a genome of interest, which are competitively hybridized with a second differentially labelled genome to arrays that are spotted with cloned DNA fragments, revealing copy-number differences between the two genomes. Genomic clones (for example, BACs), cDNAs, PCR products and oligonucleotides can all be used as array targets. The use of array-CGH with BACs is particularly popular, owing to the extensive coverage of the genome it provides, the availability of reliable mapping data and ready access to clones. The last of these factors is important both for the array experiments themselves, and for confirmatory FISH experiments.

The use of CGH with arrays that comprise long oligonucleotides (60–100 bp) can improve the detection resolution over that achieved using BACs (which starts from 50 kb to, theoretically, a few kb), and was first implemented in an assay format that is known as representational oligonucleotide microarray analysis (ROMA)[49]. The principle of ROMA is similar to that applied in the use of BAC arrays, but to increase the signal-to-noise ratio, the 'complexity' of the input DNA is reduced by a method called representation or whole-genome sampling[62] (FIG. 2b). Here the DNA that is to be hybridized to the array is treated by restriction digestion and then ligated to adapters, which results in the PCR-based amplification of fragments in a specific size-range. As a result, the amplified DNA makes up a fraction of the entire genomic sequence — that is, it is a representation of the input DNA that has significantly

Table 1 | **Methods for detecting structural variants in the human genome**

Method	Translocation	Inversion	LCV (>50 kb)	CNV indel (1–50 kb)	Small sequence variants (<1 kb)
Genome-wide scans					
Karyotyping	Yes (>3 Mb)	Yes (>3 Mb)	Yes (>3 Mb)	No	No
Clone-based array-CGH	No	No	Yes (>50 kb)	No	No
Oligonucleotide-based array-CGH	No	No	Yes (>35 kb)	Yes (>35 kb)	No
SNP array	No	No	Yes	Yes	Yes (SNPs)
Sequence-assembly comparison	Yes	Yes	Yes	Yes	Yes
Clone paired-end sequencing (fosmid)	Yes	Yes (breakpoints)	Yes (>8 kb of deletions)	Yes (>8 kb of deletions; <40 kb of insertions)	No
Targeted scans					
Microsatellite genotyping	No	No	Yes (deletions)	Yes (deletions)	Yes
MAPH	No	No	Yes	Yes	Yes
MLPA	No	No	Yes	Yes	Yes
QMPSF	No	No	Yes	Yes	Yes
Real-time qPCR	No	No	Yes	Yes	Yes
FISH	Yes	Yes	Yes	Yes	No
Southern blotting	Yes	Yes	Yes	Yes	Yes

Detection limits are shown in parentheses where applicable. The emphasis is on those approaches that are used for the detection of submicroscopic variants, although karyotyping is also shown. For comparison, each technology's ability to detect smaller sequence variants (<1 kb) is also shown. CGH, comparative genome hybridization; CNV, copy-number variant; FISH, fluorescence *in situ* hybridization (including metaphase, interphase and fibre FISH); indel, insertion or deletion; LCV, large-scale CNV; MAPH, multiplex amplifiable probe hybridization; MLPA, multiplex ligation-dependent probe amplification; QMPSF, quantitative multiplex PCR of short fluorescent fragments; qPCR, quantitative PCR.

reduced complexity, which leads to a reduction in background noise. Companies such as NimbleGen and Agilent Technologies have developed other long-oligonucleotide arrays that can be used for direct (non-representational) CGH[51]. The resolution of most available oligonucleotide arrays is in the 30 to 50-kb range, which will increase as higher-resolution arrays become available.

Another variation on the array-based approach is to use the hybridization signal intensities that are obtained from spotted oligonucleotides on Affymetrix SNP arrays[53,63–65]. Here hybridization intensities are compared with average values that are derived from controls, such that deviations from these averages indicate a change in copy number. As well as providing information about copy number, SNP arrays have the added advantage of providing genotype information. For example, they can reveal loss of heterozygosity[66,67], which could provide supporting evidence for the presence of a deletion, or might indicate segmental uniparental disomy[68–70] (which can also be considered as a form of structural variation (BOX 1); also see below).

Targeted, PCR-based experimental approaches. The most robust assays for screening targeted regions of the genome are mainly PCR-based. Perhaps the best established of these is real-time quantitative PCR (qPCR). However, although most protocols for this method work well for scoring individual deletions and duplications[55,71], they are generally not suitable for multiplexing. Alternative methods for the simultaneous interrogation of multiple regions include quantitative multiplex PCR of short fluorescent fragments (QMPSF)[56], multiplex amplifiable probe hybridization (MAPH)[57,58] and multiplex ligation-dependent probe amplification (MLPA)[59], in which copy-number differences for up to 40 regions can be scored in one experiment (FIG. 3). Another approach is to specifically target regions that harbour known segmental duplications (FIG. 4), which are often sites of copy-number variation[27,72]. By targeting the variable nucleotides between two copies of a segmental duplication (called paralogous sequence variants[73]) using a SNP-genotyping method that provides independent fluorescence intensities for the two alleles, it is possible to detect an increase in intensity of one allele compared with the other[72].

Computational approaches. Structural variants can also be identified *in silico* by comparing DNA sequences from different sources. In the simplest approach, two assemblies from unique human DNA sources are aligned to detect differences. One advantage of this method is that all types of variant, including balanced variants, can be detected (TABLE 1). In addition, there is no limit to the resolution, and the variants that are identified can be defined at the nucleotide level. Sequence scaffolds from the Celera Genomics whole-genome shotgun project are available in public databases and can be used for comparison with the human genome reference assembly[74]. There are also two large segments of the human genome — chromosome 7 (REFS 75–77) and the HLA region[78] — for which two near-complete and well-characterized assemblies exist, allowing simple comparative analysis. Comparison of the latest chromosome 7 assemblies reveals 704,297 bases of unmatched sequence at 185 sites between the two assemblies, including two equivalent

Derivative chromosome
An abnormal chromosome consisting of segments of two or more chromosomes joined together as a result of a translocation of other rearrangement.

Segmental duplications
Segments of chromosomal DNA that are >1 kb in size and have >90% inter-copy sequence identity (also called low-copy repeats or duplicons). They have been shown to constitute ~5% of the reference sequence of the human genome, where they are thought to have arisen over the past 35 million years of primate evolution.

Balanced variant
A genomic variant that involves no net loss or gain of genetic material. They include perfect inversions and translocations.

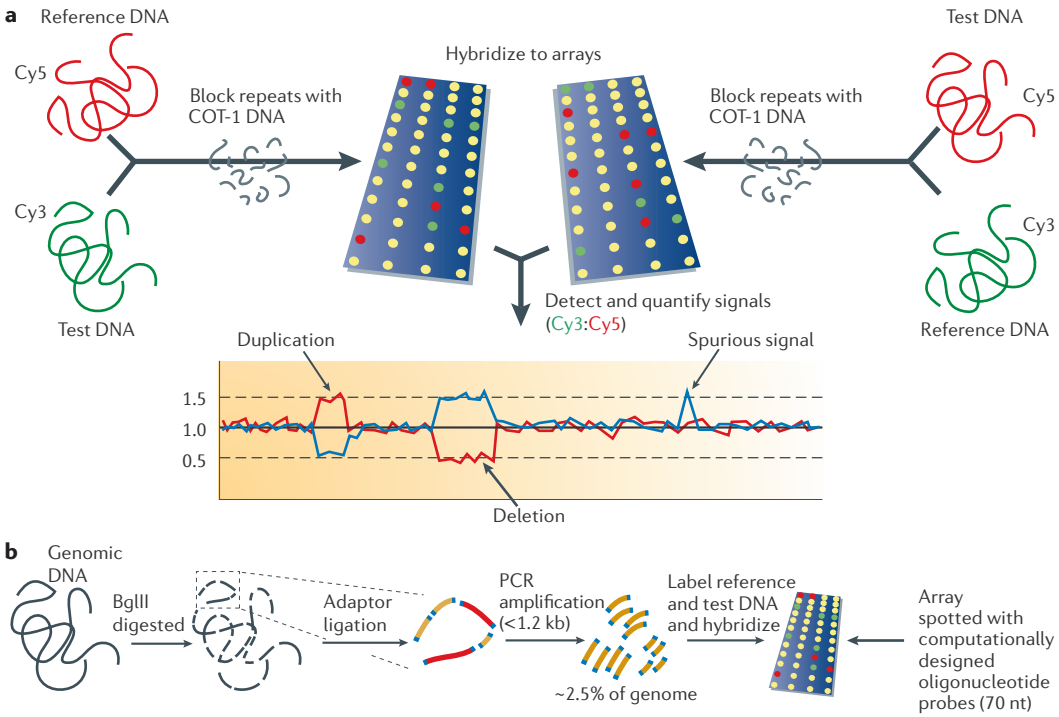

Figure 2 | **Array-based, genome-wide methods for the identification of copy-number variants. a** | In array-based comparative genome hybridization (array-CGH), reference and test DNA samples are differentially labelled with fluorescent tags (Cy5 and Cy3, respectively), and are then hybridized to genomic arrays after repetitive-element binding is blocked using COT-1 DNA. The array can be spotted with one of several DNA sources, including BAC clones, PCR fragments or oligonucleotides. After hybridization, the fluorescence ratio (Cy3:Cy5) is determined, which reveals copy-number differences between the two DNA samples. Typically, array-CGH is carried out using a 'dye-swap' method, in which the initial labelling of the reference and test DNA samples is reversed for a second hybridization (indicated by the left and right sides of the panel). This detects spurious signals for which the reciprocal ratio is not observed. An example output for a dye-swap experiment is shown: the red line represents the original hybridization, whereas the blue line represents the reciprocal, or dye-swapped, hybridization. **b** | Representational oligonucleotide microarray analysis (ROMA) is a variant of array-CGH in which the reference and test DNA samples are made into 'representations' to reduce the sample complexity before hybridization. DNA is digested with a restriction enzyme that has uniformly distributed cleavage sites (BglII is shown here). Adaptors (with PCR primer sites) are then ligated to each fragment, which are amplified by PCR. However, owing to the PCR conditions that are used, only DNA of less than 1.2 kb (yellow) is amplified. Fragments that are greater than this size (red) are lost, therefore reducing the complexity of the DNA that will be hybridized to the array. It is estimated that around 200,000 fragments of DNA are amplified, comprising approximately 2.5% of the human genome[49]. In ROMA, an oligonucleotide array is used, which is spotted with computationally designed 70-nt probes. Each probe is designed to hybridize to one of the fragments in the representation.

genomic segments in an inverted orientation[29,76]. Of these, 23 intervals between 10 kb and 100 kb in size make up ~450 kb. Most of these differences probably represent the incompleteness of the respective assemblies, but a fraction are likely to be due to actual structural variation between the individuals on whom the different assemblies were based.

In a second computational approach, anchor points are derived from sequences at the ends of clones (for example, fosmids) from a genomic library of a selected genome[28]. These anchor points are then aligned to the reference assembly, and the distance between them is compared with the expected size of the clone. Any discrepancy highlights potential insertion or deletion variants. This method, known as the paired-end sequence approach, is also suitable for the detection of some inversions, as end sequences would be in an incorrect orientation with respect to the reference assembly. Although this approach does not provide the same resolution as comparing sequence assemblies, it will remain a viable alternative until a reduction in the cost of generating further genome assemblies of high accuracy is achieved. Alternatively, structural variants can be identified by analysing sequence read-depths from shotgun-sequencing data and comparing this to what is expected from the reference genome — a method that has been used successfully for annotation of segmental duplications in the human genome[79]. Variations of this method will become more relevant when whole-genome shotgun sequencing of multiple human genomes becomes cost-efficient. Finally, comparison of human and primate (in particular chimpanzee) assemblies can highlight interspecies structural variants, and in some cases these genomic sites also show intraspecies polymorphism[29].

COT-1 DNA
DNA that is mainly composed of repetitive sequences. It is produced when short fragments of denatured genomic DNA are re-annealed.

Fosmid
A bacterially propagated phagemid vector system that is suitable for cloning genomic inserts of approximately 40 kb in size.

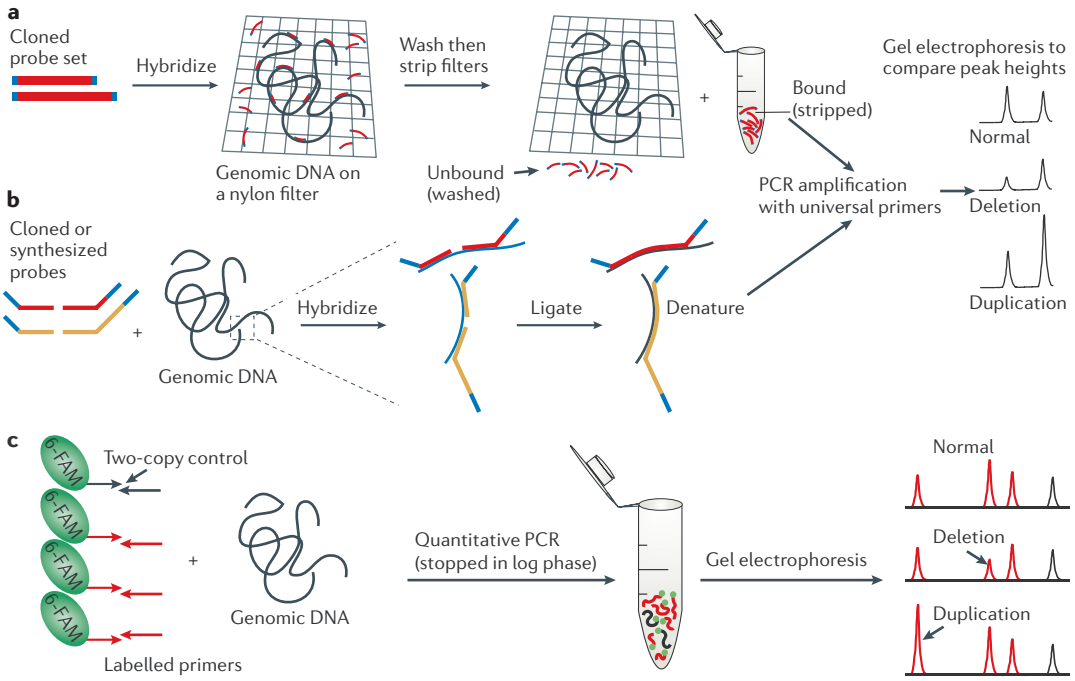

Figure 3 | **Multiplex PCR-based methods for the identification of copy-number variants. a** | In multiplex amplifiable probe hybridization (MAPH), probes of different sizes (red) are cloned into vectors and amplified by PCR such that each is flanked by the same primer site (blue). The probes are then hybridized to genomic DNA that has been fixed to a membrane. After rigorous washing to remove unbound probes, the probes are stripped from the membranes. The amount of probe that is present at this stage is proportional to its copy number in the target genomic DNA. Probes are then amplified by a universal primer pair and size-separated by gel electrophoresis. Changes in peak heights, relative to controls, can be detected to indicate deletions or duplications. **b** | Multiplex ligation-dependent probe amplification (MLPA) uses 2 probes for each target region (probes for 2 regions are shown in red and yellow), which hybridize adjacent to each other. All probe pairs are flanked by universal primer sites (blue). Following hybridization to genomic DNA, ligation is carried out to join the two primers together, such that the number of ligated primers is proportional to the target copy number. After denaturation, PCR amplification is carried out to amplify the probes that have been ligated. As well as having a universal primer site, one of these probes also has a 'stuffer' sequence, which allows each probe set to produce fragments of a different size. Size separation by gel electrophoresis is carried out as with MAPH to detect deletions and duplications. **c** | Quantitative multiplex PCR of short fluorescent fragments (QMPSF) is a quantitative technique that uses labelled primers for the target region (labelled here with the fluorescent moiety 6-FAM) in PCR amplifications. This PCR is multiplexed using control primers that are targeted to a region of known copy number. Gel electrophoresis separates the products by size, and, by comparing ratios with the control, the relative copy number for each target region is assessed.

Validation of structural variants. Ideally, any finding of structural variation using the array-based, PCR-based or computational techniques outlined above should be confirmed using an independent method. This is particularly true in the case of CNVs, as neither the PCR-based nor the hybridization-based methods that are currently used to identify them give exact information about the boundaries of the variants that are identified or their locations in the genome. Secondary confirmation using FISH is particularly useful, as it is unique in providing data both on copy number and on chromosome position (FIG. 1).

The importance of a reference-genome assembly. All the methods described above rely on comparison to a 'reference' genome to define a structural variant — a fact that has implications for experimental design and interpretation of results. For karyotypic analysis, the well-established International System for Human Cytogenetic Nomenclature identifies each chromosome band and sub-band in standardized 400-, 550-, and 850-band preparations for comparison[80]. However, for CGH experiments and PCR-based methods, no single DNA source has yet been adopted as a standardized control, although pooled samples (which represent an 'averaged' genome) are sometimes used in CGH experiments[27,81]. The lack of a standard reference genome can complicate both the designation of relative copy-number changes between samples from different projects and the standardization of databases that contain information on structural variants.

Ultimately, the underlying DNA sequence of any newly identified structural variant will be compared to the human genome reference assembly, which itself is a hybrid that reflects the hierarchical mapping and sequencing strategy that was used in its generation[1,3]. Although most of this assembly (66%) was derived from a single BAC library, the reference assembly contains sequences from an extra 41 BAC, PAC, cosmid and fosmid

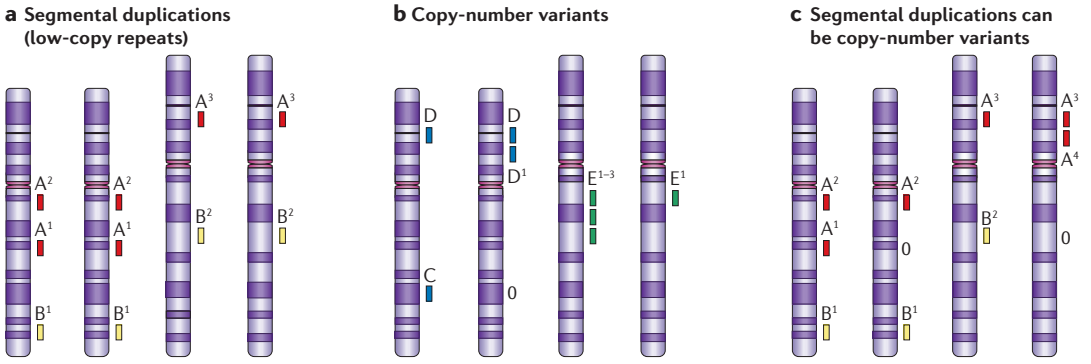

Figure 4 | **The complexity of segmental duplications and copy-number variants. a** | Segmental duplications can be duplicated in tandem or transposed to new locations in the genome, and often comprise complex blocks of repetitive DNA[79]. This complexity can cause problems with assembling genome sequences and therefore segmental duplications are often found near gaps or problematic regions of the human genome reference sequence[82,83]. Two groups of segmental duplications are shown. Group A is present in three copies; A^1 and A^2 are on one chromosome and A^3 is on another. Group B is present in two copies; B^1 and B^2 are on different chromosomes. **b** | Three groups of copy-number variants (CNVs) are illustrated. The left chromosome in each pair represents a reference DNA sequence (which could be the reference-genome assembly or a control-reference DNA sample, as used in array-based comparative genome hybridization). CNV 'C' represents a deletion (or decrease) in copy number. Relative to the reference genome, CNV 'D' represents a duplication (or an increase in copy number, which is named D^1). CNV 'E' is present in three copies in the reference sequence, but in this case, only once on the homologous chromosome. **c** | Segmental duplications have an increased tendency to vary in copy number, which is due to their repetitive nature. In these cases they can also be categorized as copy-number variants. Segmental duplication A^1 has a copy-number decrease compared with the reference, whereas A^3 is increased by one copy (named A^4) on one homologous chromosome. Importantly, in this example, although there is a decrease and increase in copy number of this segmental duplication at different loci there would be no net gain of copy number at the genome level (that is, there would still be 6 copies of segmental duplication A). In this situation, fluorescence *in situ* hybridization would detect the variable distribution of repeat A along the chromosomes, but quantitative methods would not. In the second example, B^2 is deleted on the homologous chromosome. It can be assumed in the cases that are grouped in the 'C' category that the extra copy changes would most often be newer events and would be specific to the human lineage. They could be *de novo* in origin or inherited, but would not yet be fixed within the human genome. Moreover, these would often not be annotated in the reference sequence of the human genome.

libraries (32.1% of the total) and 706 non-standard clone sources (for example, phage clones, comprising 1.9% of the total), which were derived from different individuals (see online supplementary information S1 (figure)). The most recent analysis also indicates that some 341 physical gaps remain[3,82,83], many of which overlap with newly identified structural variants[27,31]. Moreover, we compared the NCBI reference assembly of the human genome to the Celera assembly, and showed that 18.7 Mb of euchromatic sequence is present only in the latter; there are also sequences present in the most recent build of the public assembly (NCBI Build 35) that are not present in the Celera assembly (R. Khaja, personal communication). These differences could be due to cloning artefacts or assembly errors, or could represent structural variation between the multiple sources of chromosomal DNA that are used in each assembly. The apparently missing sequence is unlikely to appear as a 'target' on any microarray, and might also be missed in computational comparisons if the correct databases are not examined. This means that the true content of structural variation in the human genome will be underestimated by studies that do not consider these issues.

In our opinion, the goal for a finalized reference assembly should be that it encompasses the longest chromosomal sequences, including polymorphic regions that might be absent in some individuals. Therefore, the integration of Celera (and other) genomic sequences within the publicly available reference human assembly should be carried out if it produces a more complete sequence. In addition, efforts for 'fixing' errors and filling the remaining gaps should continue. The most complete assembly would allow for more simple comparisons to test for the presence or absence of sequences (including a decrease in copy number), or for alternative orientations of matching sequences. When known, population frequencies of structural variants (and all other variants) should be assigned to the complete reference assembly, and this data would ideally be centralized in the widely used genome browsers that are annotated with genes and other genomic features.

Submicroscopic structural variants

Before 2004 only a few dozen reasonably well-defined, non-disease-associated, submicroscopic structural variants and heteromorphisms had been documented in the human genome. These were mostly insertion–deletion polymorphisms[39,84–88] and subtelomeric cryptic translocations[89]. Since 2004, however, on the basis of a database of about 100 human genomes (see the Database of Genomic Variants web page), more than 600 submicroscopic structural variants, comprising at least 104 Mb of DNA, have been described in the literature[26–28,31,72]. Many of these variants have been observed in more than one

Cryptic translocation
A translocation, particularly one that involves DNA near telomeres, that is too small to be detectable by traditional chromosome-banding analysis.

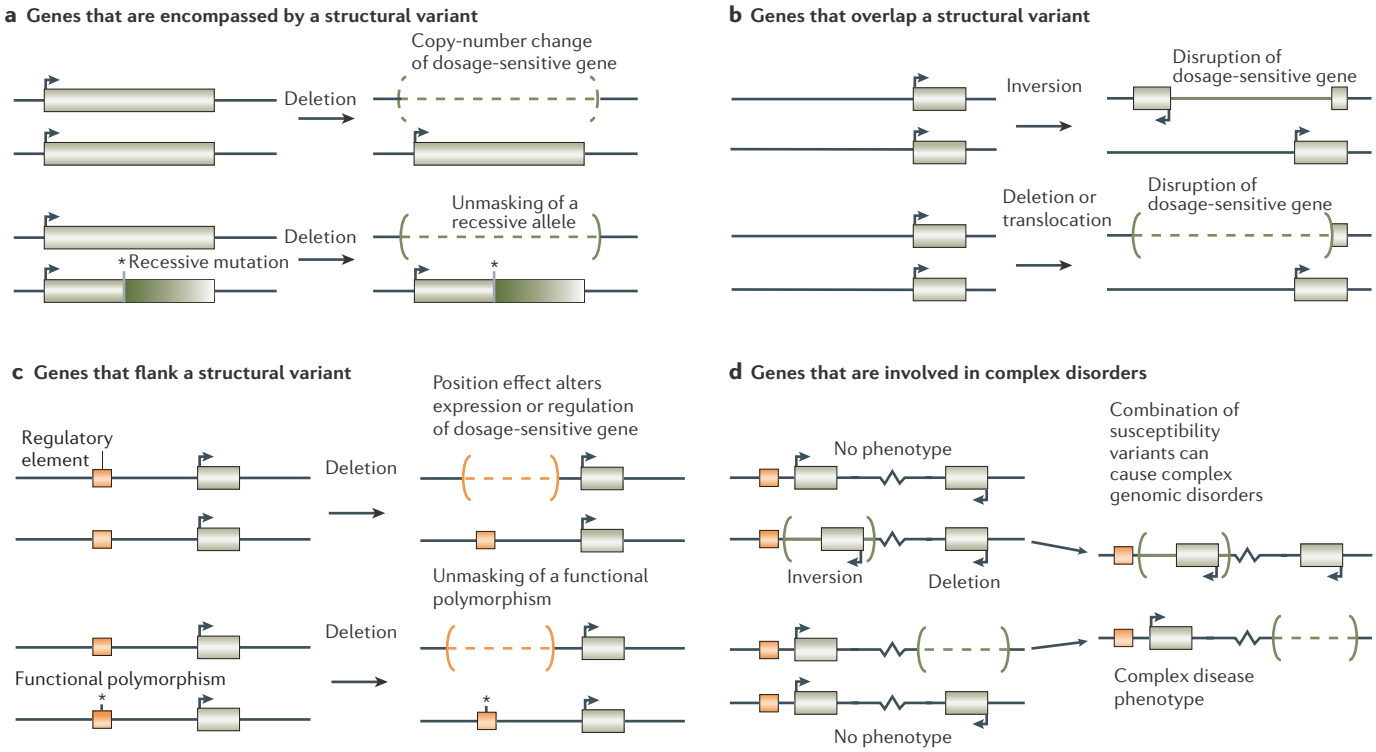

Figure 5 | **Influence of structural variants on phenotype.** Structural variants can be benign, can have subtle influences on phenotypes (for example, they can modify drug response), can predispose to or cause disease in the current generation (for example, owing to inversion, translocation or microdeletion that involves a disease-associated gene), or can predispose to disease in the next generation[34]. On the basis of their proximity to structural variants, genes might be influenced in several ways. **a** | Dosage-sensitive genes that are encompassed by a structural variant can cause disease through a duplication or deletion event (upper panel; a deletion is shown here). Dosage-insensitive genes can also cause disease if a deletion of the gene unmasks a recessive mutation on the homologous chromosome (lower panel). **b** | Genes that overlap structural variants can be disrupted directly by inversion (upper panel), translocation or deletion (lower panel), or copy-number variant breakpoints (not shown), which leads to the reduced expression of dosage-sensitive genes. Breakpoints that disrupt gene structures can also lead to the formation of new transcripts through gene fusion or exon shuffling (not shown). **c** | Structural variants that are located at a distance from dosage-sensitive genes can affect expression through position effects. An example is shown in the upper panel. A deletion of important regulatory elements can alter gene expression; similar effects could result from inversion or translocation of such elements. Alternatively, deletion of a functional element could unmask a functional polymorphism within an effector (lower panel), which could have consequences for gene function. **d** | Structural variants can function as susceptibility alleles, where a combination of several genetic factors are required to produce the phenotype. This is illustrated by an example of two structural variants that, individually, do not produce a phenotype. However, in combination they contribute to a complex disease state.

sample or are supported by more than one line of experimental evidence. Some, however, are 'singletons' in need of confirmation. At two meetings in 2005 ('Structural Variation in the Human Genome' in Toronto and the American Society for Human Genetics annual meeting in Salt Lake City), hundreds of new variants were described, and there are currently new reports of such variants being discovered every month; numerous deletion CNVs have also been found through analysis of data from The International HapMap Project[24]. Several online resources now provide information about the structural variants that have been identified and their role in disease.

Copy-number variants. Five recent studies[26–28,31,72] have provided most of the data on CNVs. In all cases, DNA from individuals with apparent 'non-disease phenotypes' was examined. The first study used quantitative SNP genotyping, in which the ratios of the two alleles in paralogous sequence variants are used to identify copy-number changes. Seventeen regions known to contain segmental duplications were targeted, and 28% of these either varied in copy number or were implicated in possible gene-conversion events[72]. This was consistent with previous observations that segmental duplications are associated with high rates of non-allelic homologous recombination (NAHR), making these sequences more susceptible to rearrangements in general[34].

Three other studies used array-CGH (ROMA[26] or clone[27,31] arrays) to identify CNVs in control individuals. These experiments targeted the entire genome, but with incomplete resolution. In the Iafrate *et al.* study, for example, 1 BAC clone every 1 Mb throughout the genome was represented on their clone array, covering an estimated 12% of euchromatic sequence[27]. Each of

Gene conversion
A process in which one sequence directs the sequence conversion of a partner allele or paralogous sequence into its own form.

these 3 studies detected about 12 CNVs per genome. The size of the variants could not be determined with accuracy as non-contiguous genomic regions were assayed, although one study did estimate a median size of 200 kb (REF. 26). In the most recent of these studies, 61% of the variants identified by Sharp *et al.* had not previously been described[31]. This result was not surprising, given the small number of genomes that have been studied so far and the limited resolution of previous detection methods, which suggests that numerous other CNVs are still to be discovered.

In contrast to the studies described above, Tuzun *et al.*[28] compared the human genome reference sequence with representative fragments of another genome using the fosmid paired-end sequence approach. Technical constraints confined the size of detectable insertions to the range of about 8–40 kb and deletions to >8 kb in size, probably leading to an underestimation of the number of variants. Nonetheless, 241 differences, including 56 inversion breakpoints, were identified between the two genomes.

We have compiled information on 563 apparently unique CNVs by comparing the data from published studies, including the five papers described above (L.F., A.R.C. and S.W.S., unpublished observations). Of these variants, 264 (47%) were described as copy-number losses, 242 (41%) as copy-number gains, and 57 (10%) as variants that can be either gains or losses. These values require validation, however, as one of the studies from which information was taken did not describe any CNVs that behaved as both gains and losses of DNA[26]. Moreover, the fosmid-end sequencing approach used by Tuzun and colleagues[28] does not readily distinguish gains and losses at the same site. Interestingly, a high percentage of the variants (25–50%, depending on the study) were found to be in close proximity to segmental duplications. In at least one case (involving the defensin locus at 8p23.1) the CNV, when defined at the molecular level, underlies a cytogenetically detectable euchromatic variant (or heteromorphism) at that chromosomal site[90], which provides a precedent for resolving the basis of other known heteromorphisms.

Inversions. Several inversions have been identified because of their involvement in human disease. Examples include the recurrent 400-kb inversion of the factor VIII gene that is found in 40% of patients with haemophilia A[91], as well as smaller inversions that affect the idunorate 2-sulphatase (*IDS*) gene in Hunter syndrome[92] and the emerin gene in Emery–Dreifuss muscular dystrophy[93]. However, less is known about inversion variants in the general population, as until recently there has not been a robust method for detecting balanced, submicroscopic variants of this type. The examples that have been identified have mainly been found in studies of human disease in cases in which the inversion variants have no detectable effect in parents, but increase the risk of a disease-associated CNV in the offspring. For example, about one-third of parents of patients with Williams–Beuren syndrome have a 1.5-Mb inversion at 7q11.23 (REF. 94), and about half of the parents of patients with Angelman syndrome carry an inversion of 4 Mb at 15q12 (REF. 95). The frequency of these variants in the general population is 5% and 9%, respectively. In another example, Sotos syndrome in Japanese patients is predominantly a microdeletion syndrome[96], and most fathers of these patients carry a 1.9-Mb inversion variant at 5q35 that predisposes to the disease in their offspring[33]. In each of these examples, the inversion breakpoints coincide with segmental duplications.

One of the few recurrent constitutional translocations in the human genome is mediated by an inversion. At two separate loci, 4p16 and 8p23, clusters of segmentally duplicated olfactory-receptor genes mediate polymorphic inversions (at frequencies of 12.5% and 26%, respectively)[97,98]. These loci are involved in the recurrent t(4;8)(p16;p23) translocation, and parents of individuals with the translocations were shown to be heterozygous carriers of inversions at both 4p16 and 8p23. The parents do not show any associated phenotype, whereas offspring who carry the translocation show phenotypes that range from Wolf–Hirschhorn syndrome to a milder spectrum of dysmorphic features[98]. These examples highlight the importance of identifying inversion variants in the general population, as they might lead to an increased risk for further disease-causing structural variation to occur in the offspring of carriers.

Mapping and sequencing studies have also identified a 900-kb inversion polymorphism on chromosome 17q21.31, which is apparently under positive selection in the European population (see below)[32]. Additionally, the Tuzun *et al.* study identified 56 putative inversion breakpoints in one individual[28]. Moreover, while carrying out comparative DNA-sequence analysis on human and chimpanzee assemblies, we identified three other polymorphic inversions in the human genome, and numerous other putative examples that await experimental confirmation[29]. Taken together, the last two studies indicate that inversion variants might be a much more common feature of our genome than was previously realized[28,29]. In all of these three studies a high proportion (~50%) of identified inversion breakpoints were associated with segmental duplications, further highlighting the propensity of these regions to mediate structural changes[28,29,32].

Other structural variants. Other types of submicroscopic structural variation also exist in the human genome, including cryptic translocations and segmental uniparental disomy (UPD). We mention these two types of variant in particular because, similar to CNVs and inversions, there are increasing reports of their occurrence, particularly in the case of segmental UPDs[68–70]. These types of variant are usually balanced, and as such cannot be identified by the array-based and PCR-based methods that are used to detect copy-number variations. Cryptic translocations, however, are amenable to detection by FISH[89] and by carrying out sequence-assembly comparisons. Segmental UPDs can be detected using microsatellites or SNP data, which are particularly effective for demonstrating stretches of homozygosity[66,67,70]. However, the genotyping of parents and/or the use of quantitative methods are still needed to distinguish UPD from a deletion.

Constitutional translocations
Chromosome abnormalities that occur before fertilization (during meiosis) or early in embryogenesis (during mitosis), such that essentially all cells in the individual harbour the same abnormality.

Implications for phenotype and disease

Molecular mechanisms. Structural variants can lead to phenotypic variation or disease in several ways (FIG. 5). In the simplest cases, structural variants can affect gene dosage directly (in the case of CNVs), or can indirectly alter gene expression through position effects. As well as these potentially disease-causing changes in gene expression, the presence of a structural variant might also predispose to further, potentially harmful structural changes, as we have described for inversions and segmental duplications.

Our analysis of 639 breakpoints of structural variants (including CNVs and inversions) from the Database of Genomic Variants indicates that 235 (37%) of these overlap with known segmental duplications (L.F. and S.W.S., unpublished observations). Presumably, in these cases, as has been shown for genomic disorders[34,94,95], otherwise benign structural polymorphisms could predispose a locus to disease-related rearrangement. In the case of polymorphic inversions, a reduced recombination frequency, resulting from the different orientation of chromosomal segments, might increase the chance of misalignment between non-allelic segmental duplications. Therefore, carriers of the inversion might be at higher risk of *de novo* deletion or other chromosomal rearrangement during meiosis[34].

As well as directly causing or predisposing to disease, structural variants might function as susceptibility alleles in complex genetic disease. Although some large variants might seem to be benign and are prevalent in certain populations, in combination with other genetic and environmental factors, including SNPs and other CNVs or inversions, they might contribute to a disease phenotype.

Determining phenotypic effects. Care is needed in categorizing variants in terms of whether they are 'normal' or 'disease-causing', as the two designations can be part of a dynamic range. Molecular cytogeneticists have always been faced with this dilemma, particularly in the prenatal or diagnostic setting. Now, with the ability to readily recognize submicroscopic variation, the question of how to differentiate benign and disease-associated structural abnormalities will be increasingly important. In general, rearrangements of heterochromatic regions, such as various translocations between the Y chromosome and acrocentric chromosomes, are found to have no clinical consequence, whereas those that involve euchromatic regions are more likely to disrupt genes and/or regulatory elements. However, the gene density of the specific affected region should also be considered[99]. In addition, balanced rearrangements are often benign, whereas unbalanced changes that lead to loss of genes are more likely to have a phenotypic effect.

There are examples in which the presence of a structural variant correlates directly with disease, such as the many dosage-related microdeletions and duplications that cause genomic disorders[34]. Family-based studies can demonstrate whether a change is *de novo* or has been inherited and, when inherited, whether there are likely to be associated phenotypic consequences (note that there are numerous examples of variable expression of phenotype and disease in inherited chromosomal rearrangements[100]). Otherwise, large population studies and reference databases provide the best source of information about a variant's frequency and its likelihood of causing a phenotypic outcome.

The accurate identification of significant phenotypic association can require comprehensive control studies that use multiple experimental methods — points that are highlighted by several recent studies[32,101–103]. For example, one study analysed 105 autism kindreds for deletions using microsatellite genotyping followed by FISH confirmation[101]. In 12 families, null alleles were identified at 4 marker sites, and these were shown to result from deletions that range in size from 5 to >260 kb. Some of these deletions were complex in nature, involving non-contiguous rearrangements. Deletions at three of the loci were shown to be specific to autism kindreds, whereas one was found in all populations screened. This indicates that both cases and controls are required to assess disease association, as some variants might not be specific for a certain phenotype or might contribute to a phenotype only when in combination with other variants.

In a second study that uses array-CGH and FISH, Gribble *et al.*[102] examined 10 patients with learning disabilities and dysmorphism who had constitutional, *de novo*, apparently balanced translocations. In 3 of the 10 patients they identified complex multiple rearrangements — including deletions, inversions and insertions — at or near one or both of the translocation breakpoints. So, although the initial identification of balanced rearrangements is often straightforward, it is important to look for further imbalances or alterations surrounding these rearrangements that might be associated with disease.

Fitness effects of structural variants. There is emerging evidence that structural variants might contribute to the phenotypic variation that has a role in determining fitness, with potential evolutionary implications. One example of this concerns the study, mentioned above, that identified a 900-kb inversion polymorphism in the European population[32]. A common allele at this position is in the opposite orientation to the (rare) allele that is represented in the reference-genome sequence. Copy-number variation was observed in the segmental duplications that flank the inversion, which indicates that this is a highly dynamic region of the genome. Genotyping tens of thousands of samples from around the world provided evidence that the inversion variant is undergoing positive selection in the Icelandic population, such that carrier females have more children and have higher recombination rates than non-carriers. However, it is unclear how the inversion variant or other variants on the inversion haplotype mediate this effect on fitness. This study not only highlights how comprehensive control studies can identify associations of variants with specific phenotypes, but might also reveal selective pressures on variants in certain populations.

Of the CNVs that have been identified so far, ~293 (41%) encompass 1 or more known genes, and

Position effect
A change in the expression of a gene that is produced by changing its location within a genome.

Genomic disorders
A group of human diseases that are caused by recurrent genomic rearrangements of unstable genomic regions. These give rise to phenotypes as a result of abnormal gene dosage within the rearranged genomic region. Segmental duplications are often involved in the rearrangement event.

Haplotype
A tightly linked group of genetic markers, which tend to be inherited as a unit because of their close proximity.

a total of 663 genes are subject to dosage differences that are due to these variants. We carried out a Gene Ontology (GO) analysis that showed a statistically significant enrichment of genes that are involved in immune responses and responses to biotic stimuli (L.F., A.R.C. and S.W.S., unpublished observations). Breaking down these GO categories, there is a particular enrichment of genes that are involved in general defence responses, including defence responses to bacteria, responses to external biotic stimuli, xenobiotic metabolism and regulation of cell organization and biogenesis. These observed enrichments indicate that genes involved in structural variation might have roles in the adaptability and fitness of an organism in response to external pressures. In general, these genes are thought to be more 'plastic', having a greater potential to evolve quickly. This implies that structural variation might be important for the dynamics of gene and organismal evolution.

Finally, in a study of individuals from around the world using qPCR and expression studies, Gonzalez *et al.*[103] demonstrated significant inter-individual and inter-population differences in CNVs that encompass functional *CCL3L1* chemokine receptor genes (and pseudogenes at the same locus). Carrying a *CCL3L1* copy number that is lower than the population average is associated with markedly increased HIV susceptibility. Again this highlights how a variant that is not overtly disease-causing can function as a susceptibility allele that is involved in a complex phenotype.

Perspective, predictions and future studies

From the mounting evidence about CNVs alone, it is already clear that the contribution of structural variation to the overall heterogeneity of the human genome is considerable. Initial studies, which have scanned only a portion of the genome, suggest that a minimum of 12 CNVs reside in each of our genomes. With extrapolation to an entire genome and the consideration of segmentally duplicated intervals, which have not yet been adequately analysed, we anticipate that some 100 CNVs per individual, each >50 kb in size, will be identified when compared to the reference sequence. In addition to these large CNVs, a significant number of intermediate-sized CNVs and inversions (8 to 40 kb) are being identified, as are numerous smaller structural variants (1 to 8 kb). Considering the total nucleotide content that is contained within these structural variants, it is likely that they will contribute an equal amount to the overall variation within the genome as SNPs. The 99.9% genome-sequence identity that is often proposed to exist between humans might therefore be considered an overestimate if a stricter definition of identity that takes structural variants into account is used. Moreover, recent analysis of the human and chimpanzee genomes indicate that segmental duplication events have had a greater effect on altering the genome than single base-pair changes, suggesting that it will be important to study structural variants from an evolutionary perspective[104,105].

As with any new discovery that involves genetic variation, an extensive initial effort will be required to catalogue its extent in diverse populations. No single method will provide the means to detect the total complement of genomic structural variation. Even the highest-resolution analysis — genome re-sequencing — would resolve only a proportion of structural variation (mostly smaller-scale insertion and deletion changes). A significant amount of information would be lost owing to the resistance of structural variants to proper assembly, misinterpretation of hemizygosity as homozygosity, or because of the characteristics of the human DNA reference sequence. As described earlier, array-CGH, qPCR, and fosmid paired-end analysis also have limits in resolution or robustness. Optical mapping of restriction-enzyme digested chromosomes promises to allow direct size estimations to be made, adding another tool to the repertoire[106]. So far, however, this method has been applied only to small genomes, but will provide a valuable alternative to existing methods once it is optimized for more complex mammalian genomes.

To better understand the occurrence of structural variation in the human genome, a consortium of investigators, including ourselves, has been set up to examine the CNV content of the 270 DNA samples that form the basis of The International HapMap Project[24]. This consortium is carrying out CGH using tiling-path arrays from different suppliers, as well as the Affymetrix 500K SNP chips and various long-oligonucleotide arrays. Inversions in these samples will also be characterized, with the goal of publicly releasing all information and contributing to the structural variation reference databases. This type of study, along with others elsewhere[107–109], will also facilitate the integration of structural variation with the existing SNP-based information about linkage disequilibrium and haplotype structure in these samples. This should greatly facilitate future studies in large patient and control cohorts that are aimed at correlating genetic variation with disease. Several other studies are using the platforms that are described in TABLE 1 to test directly for disease-associated variants, in some cases within multigeneration families[110–112].

Together, these initiatives and others will provide better insights into the nature of the population history, natural selection and degree of randomness involved in the occurrence of structural variants in the human genome. Our preliminary data indicate that most of the CNVs and inversions that have been studied so far follow Mendelian patterns of inheritance (N. Carter, M. Hurles, K. Jones, C. Lee, S.W.S., unpublished observations); however, no detailed study has thoroughly assessed the overall stability of structural variants from generation to generation, or their rate of emergence. On the basis of studies of the human dystrophin gene, one study predicts a 1 in 8 chance for deletions or 1 in 50 chance for duplications of a new variant arising *de novo* in an individual[113].

Finally, an important challenge in the characterization of structural variants in the human genome will be to allow the comparison of data from different sources. This should include the sharing of raw data files to facilitate standardized integration into reference databases, use of common controls and the use of consistent nomenclature

Gene Ontology
A project that provides sets of controlled vocabularies that have been developed to help describe and categorize genes. They describe the molecular function, biological process and cellular localization of gene products.

Optical mapping
A technology that uses *in situ* restriction digests of individual DNA molecules from genomic DNA to produce detailed optical-restriction maps of genomes.

Tiling-path array
An array that contains a set of clones that represents the sequence of a chromosome or a portion of a genome with minimum overlap.

for the different types of variant that are outlined in this review. Furthermore, initial studies are largely based on samples from repositories with little attached phenotypic information. To fully understand the contribution of this type of genomic variation to subtle phenotypes, disease development and human diversity, it will be important to examine controls and clinical samples from well-defined longitudinal studies. Such data will affect the understanding of our experimental findings, clinical outcomes and ultimately ourselves.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
4. Przeworski, M., Hudson, R. R. & Di Rienzo, A. Adjusting the focus on human variation. *Trends Genet.* **16**, 296–302 (2000).
5. Reich, D. E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genet.* **32**, 135–142 (2002).
6. Jacobs, P. A., Baikie, A. G., Court Brown, W. M. & Strong, J. A. The somatic chromosomes in mongolism. *Lancet* **1**, 710 (1959).
7. Edwards, J. H., Harnden, D. G., Cameron, A. H., Crosse, V. M. & Wolff, O. H. A new trisomic syndrome. *Lancet* **1**, 787–790 (1960).
8. Patau, K., Smith, D. W., Therman, E., Inhorn, S. L. & Wagner, H. P. Multiple congenital anomaly caused by an extra autosome. *Lancet* **1**, 790–793 (1960).
9. Bobrow, M., Joness, L. F. & Clarke, G. A complex chromosomal rearrangement with formation of a ring 4. *J. Med. Genet.* **8**, 235–239 (1971).
10. Jacobs, P. A., Matsuura, J. S., Mayer, M. & Newlands, I. M. A cytogenetic survey of an institution for the mentally retarded: I. Chromosome abnormalities. *Clin. Genet.* **13**, 37–60 (1978).
11. Coco, R. & Penchaszadeh, V. B. Cytogenetic findings in 200 children with mental retardation and multiple congenital anomalies of unknown cause. *Am. J. Med. Genet.* **12**, 155–173 (1982).
12. Warburton, D. *De novo* balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. *Am. J. Hum. Genet.* **49**, 995–1013 (1991).
A comprehensive survey that describes the prevalence of microscopic structural variants and their relevance to clinical diagnostics.
13. Jacobs, P. A., Browne, C., Gregson, N., Joyce, C. & White, H. Estimates of the frequency of chromosome abnormalities detectable in unselected newborns using moderate levels of banding. *J. Med. Genet.* **29**, 103–108 (1992).
A landmark study of the frequency of chromosomal abnormalities that affect newborns.
14. Kim, S. S., Jung, S. C., Kim, H. J., Moon, H. R. & Lee, J. S. Chromosome abnormalities in a referred population for suspected chromosomal aberrations: a report of 4117 cases. *J. Korean Med. Sci.* **14**, 373–376 (1999).
15. Benyush, V. A., Luckash, V. G. & Shtannikov, A. V. Quantitative analysis of C-bands based on optical density profiles in human chromosomes. *Hum. Genet.* **39**, 169–175 (1977).
16. Maegenis, R. E., Donlon, T. A. & Wyandt, H. E. Giemsa-11 staining of chromosome 1: a newly described heteromorphism. *Science* **202**, 64–65 (1978).
17. Verma, R. S., Rodriguez, J. & Dosik, H. The clinical significance of pericentric inversion of the human Y chromosome: a rare 'third' type of heteromorphism. *J. Hered.* **73**, 236–238 (1982).
18. Hsu, L. Y., Benn, P. A., Tannenbaum, H. L., Perlis, T. E. & Carlson, A. D. Chromosomal polymorphisms of 1, 9, 16, and Y in 4 major ethnic groups: a large prenatal study. *Am. J. Med. Genet.* **26**, 95–101 (1987).
19. Verma, R. S., Dosik, H. & Lubs, H. A. Size and pericentric inversion heteromorphisms of secondary constriction regions (h) of chromosomes 1, 9, and 16 as detected by CBG technique in Caucasians: classification, frequencies, and incidence. *Am. J. Med. Genet.* **2**, 331–339 (1978).
20. Lubs, H. A. A marker X chromosome. *Am. J. Hum. Genet.* **21**, 231–244 (1969).
21. Wright, A. F. in *Nature Encyclopedia of the Human Genome 2* 959–968 (Nature Publishing Group, London, 2003).
22. Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. *Nature Genet.* **27**, 234–236 (2001).
23. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
24. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
25. Solinas-Toldo, S. *et al.* Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer* **20**, 399–407 (1997).
26. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
This and reference 27 were the first papers to describe the global presence and distribution of CNVs in the human genome.
27. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36**, 949–951 (2004).
28. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nature Genet* **37**, 727–732 (2005).
The first description of a clone-end sequencing strategy to discover mainly intermediate-sized structural variants in the human genome.
29. Feuk, L. *et al.* Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet.* **1**, e56 (2005).
30. Dhami, P. *et al.* Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. *Am. J. Hum. Genet.* **76**, 750–762 (2005).
31. Sharp, A. J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
32. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nature Genet.* **37**, 129–137 (2005).
This reports the discovery and characterization of a 900-kb inversion polymorphism on chromosome 17. This variant was found to be under positive selection in Europeans, as determined by population-based screening.
33. Visser, R. *et al.* Identification of a 3.0-kb major recombination hotspot in patients with sotos syndrome who carry a common 1.9-Mb microdeletion. *Am. J. Hum. Genet.* **76**, 52–67 (2005).
34. Inoue, K. & Lupski, J. R. Molecular mechanisms for genomic disorders. *Annu. Rev. Genomics Hum. Genet.* **3**, 199–242 (2002).
An outstanding review of the mechanisms behind genomic disorders, including their association with segmental duplications, and non-allelic homologous recombination.
35. Gripenberg, U. Size variation and orientation of the human Y chromosome. *Chromosoma* **15**, 618–629 (1964).
36. Nielsen, J. & Sillesen, I. Incidence of chromosome aberrations among 11148 newborn children. *Humangenetik* **30**, 1–12 (1975).
37. Nussbaum, R. L., McInnes, R. R. & Willard, H. F. *Thompson & Thompson Genetics in Medicine* (W.B. Saunders, 2004).
38. McKinlay Gardner, R. J. & Sutherland, G. R. *Chromosome Abnormalities and Genetic Counseling* (Oxford Univ. Press, USA, 2003).
39. Barber, J. C. *et al.* Duplication of 8p23.1: a cytogenetic anomaly with no established clinical significance. *J. Med. Genet.* **35**, 491–496 (1998).
40. Babu, A. & Verma, R. S. Heteromorphic variants of human chromosome 4. *Cytogenet. Cell Genet.* **41**, 60–61 (1986).
41. Verma, R. S. *Heterochromatin: Molecular and Structural Aspects* (Cambridge Univ. Press, New York, 1988).
42. Starke, H. *et al.* Homologous sequences at human chromosome 9 bands p12 and q13–21.1 are involved in different patterns of pericentric rearrangements. *Eur. J. Hum. Genet.* **10**, 790–800 (2002).
43. Wyandt, H. E. & Tonk, V. S. *Atlas of Human Chromosome Heteromorphisms* (Kluwer Academic, Netherlands, 2004).
44. Barber, J. C. Directly transmitted unbalanced chromosome abnormalities and euchromatic variants. *J. Med. Genet.* **42**, 609–629 (2005).
45. Speicher, M. R. & Carter, N. P. The new cytogenetics: blurring the boundaries with molecular biology. *Nature Rev. Genet.* **6**, 782–792 (2005).
46. Locke, D. P. *et al.* BAC microarray analysis of 15q11–q13 rearrangements and the impact of segmental duplications. *J. Med. Genet.* **41**, 175–182 (2004).
47. 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** (Suppl. 2), R215–R223 (2005).
48. Mantripragada, K. K. *et al.* DNA copy-number analysis of the 22q11 deletion-syndrome region using array-CGH with genomic and PCR-based targets. *Int. J. Mol. Med.* **13**, 273–279 (2004).
49. Lucito, R. *et al.* Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res.* **13**, 2291–2305 (2003).
50. Pollack, J. R. *et al.* Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genet.* **23**, 41–46 (1999).
51. Barrett, M. T. *et al.* Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc. Natl Acad. Sci. USA* **101**, 17765–17770 (2004).
52. Brennan, C. *et al.* High-resolution global profiling of genomic alterations with long oligonucleotide microarray. *Cancer Res.* **64**, 4744–4748 (2004).
53. Zhao, X. *et al.* An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.* **64**, 3060–3071 (2004).
54. Heid, C. A., Stevens, J., Livak, K. J. & Williams, P. M. Real time quantitative PCR. *Genome Res.* **6**, 986–994 (1996).
55. Bieche, I. *et al.* Novel approach to quantitative polymerase chain reaction using real-time detection: application to the detection of gene amplification in breast cancer. *Int. J. Cancer* **78**, 661–666 (1998).
56. Charbonnier, F. *et al.* Detection of exon deletions and duplications of the mismatch repair genes in hereditary nonpolyposis colorectal cancer families using multiplex polymerase chain reaction of short fluorescent fragments. *Cancer Res.* **60**, 2760–2763 (2000).
57. 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).
58. Hollox, E. J., Akrami, S. M. & Armour, J. A. DNA copy number analysis by MAPH: molecular diagnostic applications. *Expert Rev. Mol. Diagn.* **2**, 370–378 (2002).
59. Schouten, J. P. *et al.* Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.* **30**, e57 (2002).
60. White, S. J. *et al.* Two-color multiplex ligation-dependent probe amplification: detecting genomic rearrangements in hereditary multiple exostoses. *Hum. Mutat.* **24**, 86–92 (2004).
61. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet.* **20**, 207–211 (1998).
62. Kennedy, G. C. *et al.* Large-scale genotyping of complex DNA. *Nature Biotechnol.* **21**, 1233–1237 (2003).

63. Slater, S. R. *et al.* High-resolution identification of chromosomal abnormalities using oligonucleotide arrays containing 116,204 SNPs. *Am. J. Hum. Genet.* **77**, 709–726 (2005).
64. Huang, J. *et al.* Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum. Genomics* **1**, 287–299 (2004).
65. Bignell, G. R. *et al.* High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.* **14**, 287–295 (2004).
66. Lindblad-Toh, K. *et al.* Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nature Biotechnol.* **18**, 1001–1005 (2000).
67. Mei, R. *et al.* Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res.* **10**, 1126–1137 (2000).
68. Altug-Teber, O. *et al.* A rapid microarray based whole genome analysis for detection of uniparental disomy. *Hum. Mutat.* **26**, 153–159 (2005).
69. Raghavan, M. *et al.* Genome-wide single nucleotide polymorphism analysis reveals frequent partial uniparental disomy due to somatic recombination in acute myeloid leukemias. *Cancer Res.* **65**, 375–378 (2005).
70. Bruce, S. K. *et al.* Global analysis of uniparental disomy using high-density genotyping arrays. *J. Med. Genet.* **42**, 847–851 (2005).
71. Ponchel, F. *et al.* Real-time PCR based on SYBR-Green I fluorescence: An alternative to the TaqMan assay for a relative quantification of gene rearrangements, gene amplifications and micro gene deletions. *BMC Biotechnol* **3**, 18 (2003).
72. Fredman, D. *et al.* Complex SNP-related sequence variation in segmental genome duplications. *Nature Genet.* **36**, 861–866 (2004).
73. 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).
74. Istrail, S. *et al.* Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl Acad. Sci. USA* **101**, 1916–1921 (2004).
75. Scherer, S. W. *et al.* Human chromosome 7: DNA sequence and biology. *Science* **300**, 767–772 (2003).
76. Scherer, S. W. & Green, E. D. Human chromosome 7 circa 2004: a model for structural and functional studies of the human genome. *Hum. Mol. Genet.* **13** (Spec. No. 2), R303–R313 (2004).
77. Hillier, L. W. *et al.* The DNA sequence of human chromosome 7. *Nature* **424**, 157–164 (2003).
78. Beck, S. & Trowsdale, J. The human major histocompatability complex: lessons from the DNA sequence. *Annu. Rev. Genomics Hum. Genet.* **1**, 117–137 (2000).
79. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
The first global map of segmental duplications in the human genome, including an analysis of their relationship to genes and genetic diseases.
80. An International System for Human Cytogenetic Nomenclature ISCN 1985. Report of the Standing Committee on Human Cytogenetic Nomenclature. *Birth Defects Orig. Artic. Ser.* **21**, 1–117 (1985).
81. Shaw-Smith, C. *et al.* Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J. Med. Genet.* **41**, 241–248 (2004).
82. Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* **4**, R25 (2003).
83. Eichler, E. E., Clark, R. A. & She, X. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nature Rev. Genet.* **5**, 345–354 (2004).
84. Robledo, R. *et al.* A 9.1-kb gap in the genome reference map is shown to be a stable deletion/insertion polymorphism of ancestral origin. *Genomics* **80**, 585–592 (2002).
85. Ghanem, N. *et al.* Polymorphism of MHC class III genes: definition of restriction fragment linkage groups and evidence for frequent deletions and duplications. *Hum. Genet.* **79**, 209–218 (1988).
86. Groot, P. C., Mager, W. H. & Frants, R. R. Interpretation of polymorphic DNA patterns in the human α-amylase multigene family. *Genomics* **10**, 779–785 (1991).
87. Buckland, P. R. Polymorphically duplicated genes: their relevance to phenotypic variation in humans. *Ann. Med.* **35**, 308–315 (2003).
88. Hollox, E. J., Armour, J. A. & Barber, J. C. Extensive normal copy number variation of a β-defensin antimicrobial-gene cluster. *Am. J. Hum. Genet.* **73**, 591–600 (2003).
89. Knight, S. J. & Flint, J. Perfect endings: a review of subtelomeric probes and their use in clinical diagnosis. *J. Med. Genet.* **37**, 401–409 (2000).
90. Barber, J. C. *et al.* Duplications and copy number variants of 8p23.1 are cytogenetically indistinguishable but distinct at the molecular level. *Eur. J. Hum. Genet.* **13**, 1131–1136 (2005).
91. Lakich, D., Kazazian, H. H. Jr, Antonarakis, S. E. & Gitschier, J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nature Genet.* **5**, 236–41 (1993).
92. Bondeson, M. L. *et al.* Inversion of the *IDS* gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. *Hum. Mol. Genet.* **4**, 615–621 (1995).
93. Small, K., Iber, J. & Warren, S. T. Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nature Genet.* **16**, 96–99 (1997).
94. Osborne, L. R. *et al.* A 1.5 million-base pair inversion polymorphism in families with Williams–Beuren syndrome. *Nature Genet.* **29**, 321–5 (2001).
A study of structural variants at the Williams–Beuren locus, describing a 1.5-Mb polymorphic micro-inversion that predisposes to subsequent disease-causing deletions.
95. Gimelli, G. *et al.* Genomic inversions of human chromosome 15q11–q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. *Hum. Mol. Genet.* **12**, 849–858 (2003).
96. Kurotaki, N. *et al.* Fifty microdeletions among 112 cases of Sotos syndrome: low copy repeats possibly mediate the common deletion. *Hum. Mutat.* **22**, 378–387 (2003).
97. Giglio, S. *et al.* Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am. J. Hum. Genet.* **68**, 874–883 (2001).
98. Giglio, S. *et al.* Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. *Am. J. Hum. Genet.* **71**, 276–285 (2002).
99. Nobrega, M. A., Zhu, Y., Plajzer-Frick, I., Afzal, V. & Rubin, E. M. Megabase deletions of gene deserts result in viable mice. *Nature* **431**, 988–993 (2004).
100. Ravnan, J. B. *et al.* Subtelomere FISH analysis of 11,688 cases: An evaluation of the frequency and pattern of subtelomere rearrangements in individuals with developmental disabilities. *J. Med. Genet.* 30 September 2005 (doi:10.1136/jmg.2005.036350).
101. Yu, C. E. *et al.* Presence of large deletions in kindreds with autism. *Am. J. Hum. Genet.* **71**, 100–115 (2002).
A thorough analysis of microsatellite markers behaving in a non-Mendelian manner in autism kindreds, which led to the discovery of novel microdeletions.
102. Gribble, S. M. *et al.* The complex nature of constitutional *de novo* apparently balanced translocations in patients presenting with abnormal phenotypes. *J. Med. Genet.* **42**, 8–16 (2005).
103. Gonzalez, E. *et al.* The influence of *CCL3L1* gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440 (2005).
104. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
105. Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**, 88–93 (2005).
106. Aston, C., Mishra, B. & Schwartz, D. C. Optical mapping and its potential for large-scale sequencing projects. *Trends Biotechnol.* **17**, 297–302 (1999).
107. Hinds, D. A., Kloek, A. P., Jen, M., Chen, X. & Frazer, K. A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nature Genet.* **38**, 82–85 (2006).
108. Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. *Nature Genet.* **38**, 75–81 (2006).
109. McCarroll, S. A. *et al.* Common deletion polymorphisms in the human genome. *Nature. Genet.* **38**, 86–92 (2006).
110. Le Caignec, C. *et al.* Detection of genomic imbalances by array based comparative genomic hybridisation in fetuses with multiple malformations. *J. Med. Genet.* **42**, 121–128 (2005).
111. Jobanputra, V. *et al.* Application of ROMA (representational oligonucleotide microarray analysis) to patients with cytogenetic rearrangements. *Genet. Med.* **7**, 111–118 (2005).
112. Bejjani, B. A. *et al.* Use of targeted array-based CGH for the clinical diagnosis of chromosomal imbalance: is less more? *Am. J. Med. Genet. A* **134**, 259–267 (2005).
113. van Ommen, G. J. Frequency of new copy number variation in humans. *Nature Genet.* **37**, 333–334 (2005).

Acknowledgements
The authors thank R. Khaja, J. MacDonald, J. Zhang and M. Shago (from The Centre for Applied Genomics, Hospital for Sick Children), N. Carter and M. Hurles (Wellcome Trust Sanger Institute), K. Jones (Affymetrix), and C. Lee (Brigham and Woman's Hospital, Harvard Medical School) for discussions. The work was supported from grants from Genome Canada/Ontario Genomics Institute, the Canadian Institutes of Health Research (CIHR), and the McLaughlin Centre for Molecular Medicine. A.R.C. is supported by the Natural Science and Engineering Research Council and L.F. is supported by the Swedish Medical Research Council. S.W.S. is an Investigator of CIHR and International Scholar of the Howard Hughes Medical Institute.

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 | factor VIII | *IDS*
OMIM: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
Angelman syndrome | Hunter syndrome | Sotos syndrome | Williams–Beuren syndrome | Wolf–Hirschhorn syndrome

FURTHER INFORMATION
Affymetrix: http://www.affymetrix.com/index.affx
Agilent Technologies: http://www.home.agilent.com/cgi-bin/pub/agilent/intl_bus/home.jsp?COUNTRY_CODE=US&LANGUAGE_CODE=eng&cmpid=4533
Celera Genomics: http://www.celera.com
Chromosome Abnormality Database: http://www.ukcad.org.uk/cocoon/ukcad/
Database of Chromosomal Imbalance and Phenotypes in Humans using Ensembl Resources (DECIPHER): http://www.sanger.ac.uk/PostGenomics/decipher
Database of Genomic Variants: http://projects.tcag.ca/variation
Human Genome Segmental Duplication Database: http://projects.tcag.ca/humandup
Human Structural Variation Database: http://humanparalogy.gs.washington.edu/structuralvariation
Mendelian Cytogenetics Network Online Database: http://www.mcndb.org
Mitelman Database of Chromosome Aberrations in Cancer: http://cgap.nci.nih.gov/Chromosomes/Mitelman
NIGMS Human Genetic Cell Repository: http://locus.umdnj.edu/nigms
NimbleGen: http://www.nimblegen.com
The European Collection of Cell Cultures (ECACC): http://www.ecacc.org.uk
The International HapMap Project: www.hapmap.org

SUPPLEMENTARY INFORMATION
See online article: S1 (figure)
Access to this links box is available online