PERSPECTIVE

nature genetics

Challenges and standards in integrating surveys of structural variation

Stephen W Scherer, Charles Lee, Ewan Birney, David M Altshuler, Evan E Eichler, Nigel P Carter, Matthew E Hurles & Lars Feuk

There has been an explosion of data describing newly recognized structural variants in the human genome. In the flurry of reporting, there has been no standard approach to collecting the data, assessing its quality or describing identified features. This risks becoming a rampant problem, in particular with respect to surveys of copy number variation and their application to disease studies. Here, we consider the challenges in characterizing and documenting genomic structural variants. From this, we derive recommendations for standards to be adopted, with the aim of ensuring the accurate presentation of this form of genetic variation to facilitate ongoing research.

Structural variation in the genome refers to cytogenetically visible and (more commonly) submicroscopic variants, including deletions, insertions, duplications and large-scale copy number variants — collectively termed copy number variations (CNVs) — as well as inversions and translocations (**Box 1**)[1–3]. Genome scanning technologies are now commonplace in many laboratories, allowing new structural variation to be recognized from general population surveys[4–12] or studies of diseases[13–21]. In fact, the Database of Genomic Variants[4,22] (see list of databases in **Table 1**) already contains entries (mainly CNVs) covering some 538 Mb (18.8% of the euchromatic genome) derived from the study of fewer than 1,000 genomes from individuals with no obvious disease phenotype.

This first round of observations came from several studies, each using a different technology platform and data processing algorithms, with different degrees of pre- and postexperimental standardization and validation. As a result, the data vary in quality and often have both high false-positive and false-negative rates. There is the very real possibility of the entire human genome soon being presented as 'structurally variant' in one form or another, based solely on studies of nondisease samples, which would be a distortion. It will be important for all future applications of structural variation information that the scope and detail of variants in the general population be accurately cataloged. In particular, medical genetics research — investigating structural variation profiles in individuals or clinical cohorts — will need a reliable foundation against which to interpret possible pathogenic findings in cytogenomic (**Fig. 1**), linkage and genome-wide association studies[21,23–25].

The field of genomic structural variation, however, is on the cusp of change. Pioneering approaches, often fragmented or fraught with technical limitations, are being supplanted by new technologies that afford much higher resolution screening of the genome at lower cost. We anticipate that, in the next year, the quantity of structural variation data will increase by orders of magnitude owing to microarray-based experiments alone, not to mention the plethora soon to flow from clone-end[6,26] or whole-genome sequencing experiments[27–30]. Many of these studies will survey nondisease samples for structural variation discovery to create control databases. Moreover, in little more than two years from the first description of global CNV distribution[4,5], the field is poised to make structural variation analyses standard in the design of all studies of the genetic basis of phenotypic variation. At this inflection point, we examine what is known about genomic structural variation, and consider perspectives and simple standards designed to safeguard integrity and maximize data utility for the immediate future.

Challenges in characterizing structural variants

Research into structural variation is currently at a state of development comparable to that of the earliest SNP studies. Initiatives to discover and characterize simpler structural variants — such as small insertions, deletions (indels) and balanced inversions — is likely to yield results in proportion to investment, as was the case for SNPs[31–33]. However, for larger and particularly for more complex structural variants, there are additional confounding factors. To provide a framework for discussion of prospective standards, we group into five categories the major issues

Stephen W. Scherer and Lars Feuk are at The Centre for Applied Genomics and Program in Genetics and Genomic Biology, The Hospital for Sick Children, 14th Floor, Toronto Medical Discovery Tower, MaRS Discovery District, 101 College Street, Room 14-701, Ontario M5G 1L7, Canada. Stephen W. Scherer is in the Department of Molecular and Medical Genetics, University of Toronto, Toronto, Ontario M5G 1L7, Canada. Charles Lee is in the Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. Ewan Birney is at the European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. Nigel P. Carter and Matthew E. Hurles are at the Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. David M. Altshuler is in the Program in Medical and Population Genetics, Broad Institute of Harvard University and the Massachusetts Institute of Technology, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. Evan E. Eichler is in the Department of Genome Sciences and Howard Hughes Medical Institute, University of Washington School of Medicine, Seattle, Washington 98195, USA.
e-mail: steve@genet.sickkids.on.ca

Published online 27 June 2007; doi:10.1038/ng2093

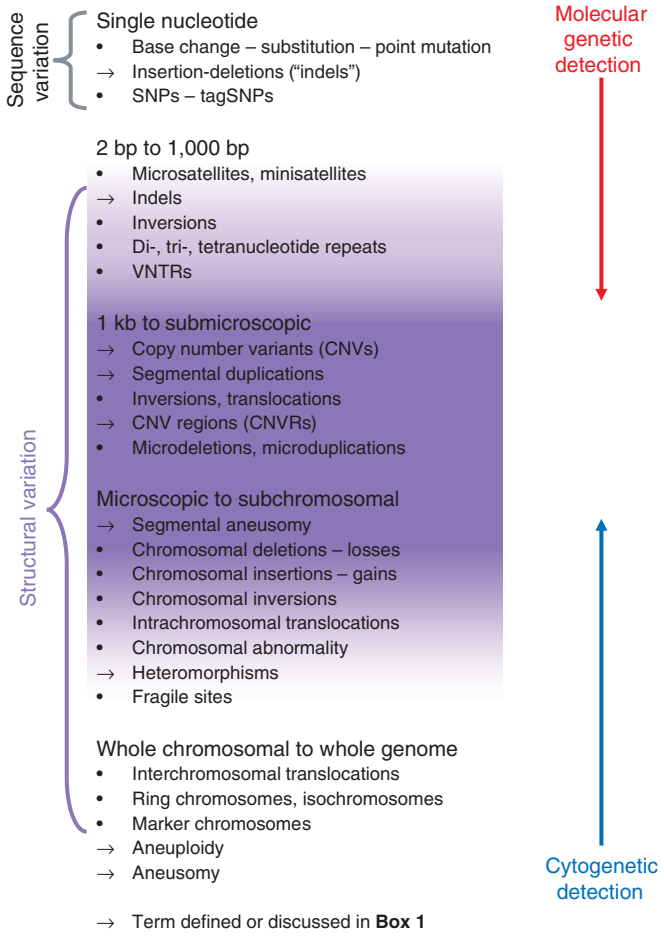

Figure 1 Lexicon of genomic variation. Descriptors of variation began in the realm of cytogenetics, followed by those from the field of molecular genetics and, most recently, by technologies such as those described in this perspective, which bridge the gap for detection of genomic variants (sometimes called cytogenomics[55]). The designation of the category '1 kb to submicroscopic' is somewhat arbitrary at both ends, but is used for operational definition. In a broad sense, structural variation has been used to refer to genomic segments both smaller and larger than the narrower operational definition, as illustrated by the large bracket. The focus of recent discoveries has been the subgroup in the midrange (indicated with strong highlighting), but the gradation of shading illustrates that the biological boundaries may really encompass some forms of variation previously recognized from either cytogenetic or molecular genetic approaches. At the molecular level, SNPs can be identified that are representative of the underlying haplotype structure (tagSNPs). As structural variation becomes better integrated with the existing SNP-based linkage disequilibrium maps, it is likely that presence or absence of many structural variants will simply be inferred by typing selected SNPs[11,25,73].

currently curbing progress in this field. Data quality, which has impact throughout these other issues, is discussed in the subsequent subsection. The majority of the discussion pertains to the variants classed as CNVs, as these represent the predominant form studied to date. Our comments also mostly target issues related to whole-genome discovery surveys.

Terminology. The newly recognized domain of structural variation is blurring the distinction between traditional cytogenetic and molecular analyses, as it fills the (albeit narrowing) gap between the limits of resolution of these earlier approaches to genetic variation (**Fig. 1**). Terminology established within each camp is sometimes unwieldy in the crossover (**Box 1**). Moreover, there is no standard nomenclature for structural variants that fall between those that can be classified by naming systems established from the cytogenetic[34,35] or mutation literature[36] (for example, indels). For some terms, such as CNV, there is added complication because they are being used regularly as a descriptor in both control and disease studies, but with different meaning. Different classes of CNVs are described in Redon *et al.*[11] and in **Supplementary Figure 1** online. Nomenclature for genes encompassed by structural variants also needs to be considered, but no rules have yet been established.

Annotating complex structural variants. Many structural variants are large in size, flanked by or encompassing complex repetitive DNA sequences. They may be unbalanced in content or highly polymorphic, characteristics that pose significant challenges for detection and analysis. There are many complexities associated with classifying and characterizing CNVs (**Supplementary Figs. 1**, **2** and **3** online). As the precise rearrangement breakpoints are usually not resolved (because of coincidence with large repeats or because of low resolution coverage of assays), it is typically not possible to determine whether the underlying variants are identical by descent or represent independent events in close proximity to one another. Regions of high sequence identity may also cause cross-hybridization on comparative genome hybridization (CGH) platforms, leading to CNV calls in regions that are not actually variable (**Supplementary Fig. 3**). Determining the meiotic and mitotic characteristics of these variants — such as the *de novo* mutation rate, stability and level of mosaicism — can also be confounded not only by the complex nature of the underlying sequences but by technical and comparative limitations, including the source of the DNA (described below).

Technological limitations. At present, no single approach identifies all types of structural variation. Current scans of genome-wide structural variation are screening or discovery assays, and not definitive tests. In our hands, the testing of a single sample by different platforms and 'call' algorithms can lead to substantially different CNV call rates, owing to differing sensitivity, specificity, probe density and type of probe used (**Table 2** and **Supplementary Table 1** online). This matter is underscored by the relatively small degree of overlap among published datasets[2,37], even when assessing identical samples[7,9–11]. The progress on CNV discovery to date is largely due to the availability of numerous microarray platforms, which detect quantitative imbalances. In contrast, there is currently no high-throughput, cost-effective method to scan the genome for inversions or translocations. Short of comparing 'finished' sequence assemblies from independent sources[38,39], it can take a multitude of approaches to identify, validate and sequence the compendium of structural variation comprehensively (**Table 3** and **Supplementary Table 2** online). Other issues, such as relative costs of arrays and reagents and availability of specialized equipment, often limit access to the most appropriate experiments.

Characteristics of reference and test samples. Identification of variation requires comparison to either a reference DNA source[4,5,11,40,41], a reference dataset[11] or a reference genome sequence[6,39,42], which has implications for experimental design and interpretation of results[43]. For example, at present, no standardized 'reference' control DNA has been adopted for laboratory experiments, and in some cases, 'pools' of samples or datasets are used to represent an averaged genome (**Table 2**). This lack of standard reference genomes can complicate both the designation of relative copy-number differences among samples from different projects and the standardization of databases (**Table 1**) that contain information about structural variants. Specifically, if in a single experiment it is impossible to distinguish a loss in the test sample from a gain in the reference sample, then two different studies may report the same CNV as a relative gain or loss (duplication or deletion), respectively. Moreover, using pools of DNA or their intensity outputs as hybridization controls or in comparative intensity analysis (**Table 2**) may lead to a decreased

Box 1 Terminology

Terms that are part of the current vocabulary for structural variation are in bold type below, set into the context of some key definitions and related comments.

Structural variant. Structural variant is the umbrella term to encompass a group of genomic alterations involving segments of DNA typically larger than 1 kb, and which can be microscopic or submicroscopic[1]. We use the term as a neutral descriptor with nothing implied about frequency, association with disease or phenotype, or lack thereof. This definition of size, though perhaps somewhat arbitrary, was undertaken to accommodate this significant class of variation that spans the gap between small variants (such as variable number of tandem repeats (VNTRs)) detected with molecular genetic assays and those recognized microscopically on karyotypes. The structural variation may be quantitative (**copy number variants** comprising **deletions**, **insertions** and **duplications**) and/or positional (**translocations**) or orientational (**inversions**).

Copy number variation/variant (CNV). We use these terms to refer to a DNA segment of at least 1 kb in size, for which copy number differences have been observed in the comparison of two or more genomes. Without further annotation, CNV carries no implication of relative frequency or phenotypic effect. These quantitative variants can be genomic copy number **gains** (**insertions** or **duplications**) or **losses** (**deletions** or **null genotypes**) relative to a designated **reference genome sequence**. A **copy number polymorphism (CNP)** is a CNV that occurs in more than 1% of the population.

CNV locus or **CNV region (CNVR)**. Merging of independently ascertained, but overlapping, genomic segments creates the representation of a **CNV locus** (that is, a segment at a fixed chromosomal position); the accumulation of data gradually will reveal the true underlying structure of the variant segment. In some cases, this will be a discrete cassette of DNA; in others, it will be a multiplex arrangement of variant units in close proximity, forming a **CNV region**[11]. A given variable segment can be detected with multiple clones in a single array or by different arrays in different studies, and its borders gradually fine-tuned with targeted assays. By their very nature, these segments may have different forms among the individuals used for their discovery.

Insertion/deletion ('indel'). Indel is a collective abbreviation to describe relative gain or loss of a segment of one or more nucleotides in a genomic sequence. It allows the designation of a difference between genomes in situations where the direction of sequence change cannot be inferred: for example, when a reference or ancestral sequence has not been defined. It has typically been used to denote relatively small-scale variants (particularly those smaller than 1 kb); however, we do not propose any size restriction for its use.

Segmental duplication (also called **low-copy repeat (LCR)** or **duplicon**). 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[44,64,65]. These segments can also be CNVs. The duplicated blocks predispose to **nonallelic homologous recombination**.

Human genome reference assembly. The standard reference DNA sequence (or assembly) of the human genome[66] that is regularly curated (successive updates named '**builds**'). The assembly is derived mostly (>60%) of DNA from a **bacterial artificial chromosome (BAC)** library made from a single donor, with the rest of the sequence originating from a mosaic of other sources. The current assembly covers most of the euchromatic regions of the human genome, but there are still some gaps remaining, and many of these co-locate with segmental duplications and/or CNVs.

Aneuploidy, aneusomy and heteromorphism. These terms have origins in classical cytogenetics and describe structural variants at the largest end of the scale. **Aneuploidy** is the state of having an abnormal number of chromosomes. Similarly, **segmental aneusomy**, in reference to a portion of a chromosome, implies abnormality. **Heteromorphism** (literally, 'different form') has come to imply normal variation, or an atypical chromosome form not associated with an abnormal phenotype. Such large-scale variants are often the basis for dysfunction owing to dosage imbalance (such as for **segmental aneusomy syndromes**[67]), but may also be part of normal functional variation.

Minor-allele frequency versus **altered copy-number frequency**. The **minor allele** is the less common allele at a polymorphic locus. The use of this term is complicated when a locus is multiallelic. Locke *et al.*[9] proposed use of **altered copy-number frequency** because measurements of copy number are on diploid samples and screening methods do not necessarily distinguish the two independent alleles. Redon *et al.*[11] adopted the convention of assuming that the minor allele is the derived allele; thus, deletions have a minor allele of lower copy number and duplications have a minor allele of higher copy number.

Nonmendelian inheritance (also called **mendelian incompatibilities** or **mendelian inconsistencies**). These terms refers to transmission from parent(s) to offspring in a manner that does not conform to expectations of classical allelic segregation. (Avoid the term 'mendelian errors'.) Evidence in family studies ('**trios**' in the HapMap data) of apparent **nonmendelian inheritance** for a genomic segment indicates that copy number variation may be involved[7,10].

power to detect variants in highly polymorphic regions of the genome. In these regions, the pool will represent an intermediate between the polymorphic and nonpolymorphic states, resulting in smaller relative difference in intensity than a nonpolymorphic single reference would yield. In terms of annotating variants, the relative nature of CNV determination can pose a problem, as it leads to an overestimation of regions with both apparent gains and losses.

Ultimately, the underlying sequence characteristics of any newly identified structural variant will be compared to the human genome reference assembly. The latest release from the US National Center for Biotechnology Information (NCBI), called Build 36, is a mosaic of some 708 different sources[1], and covers mainly the euchromatic portion of the genome, with some 302 known gaps (http://www.ncbi.nlm.nih.gov/). Concomitance of incomplete or falsely merged regions of the reference assembly with the position of structural variants can confound comparisons of one against the other[44,45]. Moreover, as many technologies use the NCBI reference sequence to guide product development, structural variants residing in the unannotated segments of the human genome may be missed (**Supplementary Fig. 2**). Test samples can also be from a mix of untransformed or transformed tissues, all impacting on interpretation[11,46]. Finally, samples used to discover structural variants from control populations may have little or no genetic (for example, parent of origin) information or phenotypic assessment protocols attached to them. So, despite common presumptions, any variant described by such studies is not necessarily either neutral or benign.

Table 1 Databases

Center for Information Biology Gene Expression Database (CIBEX)	http://cibex.nig.ac.jp/index.jsp
Coriell Cell Repositories NIGMS Human Genetic Cell Repository	http://locus.umdnj.edu/nigms/
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/
Ensembl	http://www.ensembl.org
Gene Expression Omnibus (GEO)	http://www.ncbi.nlm.nih.gov/geo/
Human Gene Nomenclature Committee (HGNC) Database	http://www.gene.ucl.ac.uk/nomenclature/
Human Segmental Duplication Database	http://projects.tcag.ca/humandup/
Human Structural Variation Database	http://humanparalogy.gs.washington.edu/structuralvariation/
NCBI Single Nucleotide Polymorphism Database (dbSNP)	http://www.ncbi.nlm.nih.gov/projects/SNP/
Segmental Duplication Database	http://humanparalogy.gs.washington.edu
UCSC Genome Browser	http://genome.ucsc.edu/

Database issues. The main sources of information for human structural variation are the Database of Genomic Variants and the Human Structural Variation Database. Both are currently limited, in that variants are simply represented as they are described in publications and overlaid on the current reference assembly, without precise location of most breakpoints. There are some unpublished data at these sites, but so far there is no active effort to standardize CNV calling or characteristics through reexamination of the original primary data. Moreover, as the human reference assembly is updated in subsequent assemblies, sites of apparent structural variation can disappear and reappear, presenting a challenge for database management. Although Ensembl and UCSC Genome Browser display data from the Database of Genomic Variants, there is currently no standard requirement to submit published structural variants to any database. Further, there is no system for naming structural variants with unique accession numbers, and surprisingly, only a proportion of studies post their raw or underlying data, and full method of interpretation, for public access.

There are also many challenges in the layout and visualization of the data. For example, it is current practice to display structural variants using estimates of start- and end-points when the breakpoint(s) are suboptimally resolved. When there are two or more overlapping variants originating from the same study, they are sometimes grouped together even if they are not identical[11], and misgrouping can occur, particularly near segmental duplications. Moreover, as the number of surveys continues to grow, the CNVs discovered will become more redundant.

Presenting structural variation data in relation to the reference assembly can also be problematic[1,39] because the standard browsers were not designed to display these data. This issue notwithstanding, smaller variants (usually <10 kb) are present in NCBI's dbSNP, and a goal of the Human Structural Variation Database is to integrate structural variation data, such as fosmid paired-end sequences[6], with the NCBI human reference sequence (including those regions not represented in the current assembly)[26]. The Database of Genomic Variants will continue to display structural variation data originating from nondisease-defined samples, but stricter criteria for inclusion, as well as assessment and annotation of the quality standards described below, will become critical aspects of the curatorial process.

Content and quality of early studies of structural variants

To assess current practices in collection and validation of discovery data, we review and comment on 12 experimentally diverse and highly cited studies, each undertaken to search for structural variation in the human genome. In **Table 3** and **Supplementary Table 2**, we summarize selected parameters and the strengths and weaknesses of these studies.

Genomes surveyed and reference samples. The number of genomes investigated with each study ranged from one (in sequence comparisons to reference assemblies[6,39]) to 270 (in three studies of the HapMap collection[9–11]). Appropriate attention was given to samples being from unrelated individuals or from families, and ethnic diversity was usually noted. Tissue sources of DNA were heterogeneous, and whether or not they were transformed or cultured was inconsistently documented. Phenotypic information would generally have been unknown, or assumed to be unremarkable (from 'healthy volunteers'), although Iafrate *et al.* included samples with known karyotypic abnormalities as

Table 2 Copy number variants called on the same test sample (NA15510) using different experimental platforms and algorithms[a]

Platform	Method	Reference sample	Analysis tool	CNVs detected	Platform-specific CNVs	Regions in DGV	Average size
Whole-genome tiling path BAC (WTSI)	Clone array CGH[68]	NA10851	CNVfinder[68]	74	38	72	237 kb
Nimblegen 385k array	Oligo CGH	NA10851	CNVfinder	63	18	59	343 kb
Agilent 244k array	Oligo CGH	NA10851	CGH Analytics	42	8	40	74 kb
Affymetrix 500k	Comparative intensity analysis	Pool	GEMCA[69]	24	7	21	316 kb
		Pool	dCHIP[70]	7	0	7	496 kb
		Pool	CNAG[71]	7	1	7	437 kb
Illumina 650Y	Comparative intensity analysis	Pool	QuantiSNP[72]	9	1	9	236 kb
		Pool	Bead studio	5	0	5	523 kb
Sequencing	Fosmid ends[6]	Build 35	Alignment	241	213	N/A	29 kb

[a]The number of CNV regions identified depends on the parameters used for each analysis tool. DGV, Database of Genomic Variants. Of these platforms, the BAC array platform detects the most CNVs because currently it provides the best coverage across segmental duplications[11,68]. The experimental platforms and analysis tools have different availability, application and cost. For example, the Wellcome Trust Sanger Institute (WTSI) BAC array is available by collaboration. Fosmid-end sequencing requires characterization of clone libraries, limiting it to analysis of a small number of samples[26]. The remaining products are sold commercially. Current prices for full processing of commercial arrays are typically in the ~$400–$1,000 range. The regions from the fosmid-end sequencing were already in DGV. Data from the other five platforms is presented here for the first time. The total number of CNVs found in sample NA15510 using these six approaches was 340. **Supplementary Table 1** provides a matrix for overlap of CNV discovery between pairs of platforms. N/A, not applicable.

Table 3 Summary of 12 published surveys (2004–2007) of structural variation content in human genomes[a]

Study	Genomes assayed	Reference sample(s)	A. Array details B. Comparison method	Variants reported No.	 Size	Found in >1 sample	Other validation	Comments, caveats
A. Primary discovery by array CGH								
Oligonucleotide array								
Sebat *et al.* 2004	20 blood DNA, cell lines, sperm	Several	ROMA	76	I–L	41%	FISH, alternate array	One of the first two papers describing global CNVs in the human genome; low coverage in segmental duplication regions and technology not widely adopted
Hinds *et al.* 2005	24 discovery 71 diversity	Reference sequence	Haploid	215	S	67%	Deletion PCR	This array-based approach detects only small (<10 kb) deletions; technology not simple
BAC array								
Iafrate *et al.* 2004	39 healthy, 16 other; blood DNA, cell lines	Pooled	2,632 BAC clones	255	L	40%	FISH, qPCR	One of the first two papers describing global CNVs in the human genome; only 1-Mb resolution arrays used; some false positives due to aberrant clones
Sharp *et al.* 2005	47	1 male	2,194 BAC clones	119	L	55%	FISH, other	BAC arrays targeted at segmental duplications; limited genome coverage
Locke *et al.* 2006	263 HapMap	GM15724	2,194 BAC clones	384	L	67%	Oligo CGH, clone sequencing	Same platform as Sharp *et al.* focused on segmental duplications
Redon *et al.* 2006	270 HapMap	NA10851	(i) 26,574 BAC clones (ii) SNP intensity 500K	1,447	I–L	66%	qPCR	Complementary arrays employed generating the first comprehensive human map of CNVs; introduced concept of CNVRs to deal with resolution issues
Wong *et al.* 2007	95 discovery 10 other	1 male	26,363 BAC clones	3,654	I–L	*	Oligo CGH, qPCR	The comparatively large number of CNVs detected in this study may reflect a high false discovery rate
B. Primary discovery by sequence comparison								
Alignment to reference sequence								
Tuzun *et al.* 2005	1 fosmid library NA15510	Build 35 (hg17)	Fosmid end sequence alignment	297	S–I		BAC array CGH, sequencing, PCR	First study to use clone end sequence mapping to identify variation; identifies many inversion variants; cost per experiment limits broad applicability
Mills *et al.* 2006	36	Build 35 (hg17)	Computational alignment	294,498	S		PCR, sequencing	Alignment of human sequence trace 'reads' to the reference assembly to identify structural variation; no regions >10 kb reported
Khaja *et al.* 2006	1 (Celera R27c assembly)	Build 35 (hg17)	Computational alignment	13,534	S–I		PCR, qPCR, FISH	Alignment of two human genome assemblies to identify variations; small number of genome assemblies available limits application
Study of HapMap trios								
Conrad *et al.* 2006	60 HapMap trios	N/A	Mendelian incompatibilities	587	S–L	61%	Oligo CGH, qPCR	One of the first two genome-wide studies using SNP genotypes to identify deletions
McCarroll *et al.* 2006	269 HapMap	N/A	SNP footprints	541	S–L	51%	FISH, allele-specific fluorescence, deletion PCR, qPCR	Companion study to Conrad *et al.*, using SNP genotypes to identify deletions; these two studies describe simple deletions only

[a]The studies surveyed were each undertaken to seek and characterize human genomic structural variation without any focus on phenotype association. These represent the earliest pan-genome surveys. A more detailed representation (in chronological order) is documented in **Supplementary Table 2**. Sizes of variants are detailed in **Supplementary Table 2**, but roughly classified here as small (S), intermediate (I), large (L) or a combination. FISH, fluorescence *in situ* hybridization; oligo, oligonucleotide; ROMA, representational oligonucleotide microarray analysis; *800 (22%) found in >2 samples.

controls[4], and Wong *et al.* used some material from cancer programs[41]. Each study used different reference sample(s) for genome comparison. One used pooled DNA[4], three compared to the reference human genome assembly[6,39,42], one made a variety of comparisons[5] and the other CGH approaches each used a different single male reference sample. Future studies will increase the variety of genomes surveyed, and these would benefit from a consensus standard of documented information about their sources. In contrast, a smaller number of reference sequences would facilitate the process of collective documentation.

Primary discovery methods. Table 3 is organized according to the methods used to search for structural variants. The upper portion includes seven studies that employed CGH, each with a different array platform, encompassing a range of probe size, complexity and resolution. One approach[9,40] targeted regions associated with segmental duplications, but the rest spanned the genome, with arrays carrying from 2,000 up to about 26,000 clones in genome tiling-path arrays[11,41]. Redon *et al.*[11] added a second complementary screening strategy based on relative fluorescence intensities with arrays designed originally for SNP genotyping. The lower portion of **Table 3** summarizes five studies with completely different strategies, based on genomic sequence comparisons. These studies used existing data from either the reference human genome sequence[6,39,42] or the HapMap project[7,10] to mine for deletions and other relatively small structural rearrangements. The fosmid-based approach[6] and sequence comparison[39] were able to discern orientational as well as quantitative variants.

Experimental quality controls. Before structural variants can be revealed by genome comparisons, positive data arising from other biological or technical causes need to be filtered. Biological differences that were variously accounted for among these studies include (i) male-female X and Y chromosome dosage differences[9,11,40], (ii) somatic rearrangements of the immunoglobulin genes[5,11], (iii) cell-culture artifacts such as mosaic trisomies[46] and (iv) results of genomic instability of virus-transformed cell lines[11]. Similarly, any variation relative to a reference human genome sequence in the computational approaches must be interpreted in light of the known gaps and potential assembly artifacts[1,6,39].

As these screening strategies are themselves biological, with associated technical artifacts, replication is the most important experimental tool for assessing the validity of observations, and it took many forms among these studies. Within each CGH array, clones were typically in duplicate or triplicate. Interexperimental replication involved ostensibly the same conditions and/or an experimental alternate, such as 'dye-swap' of the two fluorochrome labels between the test and reference samples. The means of dealing with discordant replicates was inconsistent among the studies, and sometimes difficult to discern from the publications. In most studies[4,9,11,40], discordant dye-swap results were eliminated, but in Wong *et al.*[41], only 20% of samples were assayed in both orientations. Within each study, experiments also showed variable background 'noise', and some studies repeated and/or deleted individual assays that did not meet a defined quality threshold. When sources of 'noise' are nonrandom, replication alone will reproducibly yield false positive calls, which argues for replication by diverse methods.

Other controls showed the effectiveness of the respective screening methods. Self-versus-self hybridization was used[4,5,9,40] to estimate somatic effects and/or numbers of false positive calls. Two studies assayed samples with previously characterized imbalances[4,40]. Sharp *et al.*[40] showed the enhanced (11-fold) effectiveness of their targeted 'hot spot' array relative to a genome-wide assay. Redon *et al.*[11] evaluated concordance between their two primary platforms and undertook numerous technical replicates.

Each study defined its own algorithm for 'calling' differences between sample and reference as putative structural variants. As for all screening assays, they were driven to optimize both sensitivity and specificity of the ascertainment, but approaches to this balance differed. Redon *et al.*[11] set parameters in their algorithm to allow fewer than 5% false positive 'calls' per experiment. Other studies set thresholds and assessed numbers of false positives retrospectively. Some reported these type I errors in relation to the number of clones in the array[4,40,41] and others relative to the proportion of positive calls[5,7], prohibiting a direct comparison of specificity among the various studies. Sensitivity was harder to assess, and arguably impossible without knowledge of the true (or at least gold standard–based data) underlying numbers of structural variants. Estimates ranged from 5% false negatives[9] to 50% power to detect 25-kb deletions[7], but sensitivity was generally compromised in favor of specificity.

Structural variants identified. Assay design had a strong impact on the type and size of structural variants detected (**Fig. 1**, **Supplementary Fig. 2** and **Table 2**). All revealed quantitative variation (gains or losses), but three recognized only deletions[7,8,10], and two could also detect evidence of inversions[6,39]. Sizes of variant segments could be as small as

Table 4 Classification of modifiers used for the description of structural variation[a]

Location	Origin	Frequency	Phenotypic Consequence
• Heterochromatic	• Maternal	• Unique	• None
• Euchromatic	• Paternal	• Rare	• Benign
• Centromeric	• Somatic	• Novel	• Unknown
• Telomeric	• Germline	• Polymorphic	• Undefined
• Contiguous	• Constitutional	• Common	• Neutral
• Interchromosomal	• Pedigree-specific	• Fixed	• Phenotype-associated
• Intrachromosomal	• Population-specific		• Quantitative
			• Susceptibility-related
			• With variable penetrance
			• Disease-associated
			• Disease-causing
			• Medically relevant
			• Syndrome-associated
			• Lethal

[a]Examples of modifiers that might be used to enhance basic descriptions of structural variants. **Figure 1** displays the spectrum of terminology used to describe the form of genomic variants, in the context of scale. Terms should be chosen from among those to best reflect the basic relative structure observed (avoiding terms with inherent implications beyond that of form, such as 'chromosomal abnormality'). The basic structural descriptor may then be annotated with modifiers such as these, according to what additional information is known and needed at the time. Nomenclature issues can be complex when there is variable expressivity of a phenotype associated with a structural variant.

1 bp with computational alignments[39,42] (though many of these were smaller than our defining size threshold of 1 kb[1]). Small deletions were detected through haploid hybridization (70 bp–10 kb)[8] or oligonucleotide (SNP) footprints (1–404 kb)[7] (1–745 kb)[10], and the fosmid approach revealed variants in the range of library inserts (40 kb)[6]. Array methods approached the larger end of the spectrum for CNVs (collectively, about 50 kb–1 Mb)[4,5,9,11,40,41]. BAC clone probes tend to initially overestimate the apparent size of variants, as the clones may be large relative to the variant segment(s) they harbor, and the more sensitive the platform, the greater the overestimation[11,47]. Oligonucleotide arrays, on the other hand, approach the boundaries of variable segments from within, and should provide more accurate size estimates as long as the region has sufficient probe density.

The architecture of a variant region can influence its apparent size. Independently discrete genomic segments whose borders overlap can form a variable region characterized as much larger than its component variants, or containing complex rearrangements of smaller independently variable elements (**Supplementary Figs. 1** and **3**). As a result, the basis for definitions of overlap, variants, variant regions, merged variants, locations and so forth have been discretionary and varied. The field is probably ready for functional consensus in this area.

The earliest surveys reported about 100 variants or regions[4,5]; more recently, Wong *et al.* reported a disproportionate 3,654 CNVs, from which only 800 were considered 'high frequency' and more likely to be true positives[41]. Sequence comparisons flagged many more thousands of sites[39,42], albeit ones that were much smaller and often reflected sequence assembly artifacts. Each of the 12 studies in **Table 3** added a majority of apparently new variant loci, though as the catalog of genomic structural variants accumulates, the number of such new additions will eventually plateau.

Validation of putative structural variants. We reemphasize that the discovery strategies in **Table 3** are screening tests, which draw attention to genome segments with an increased probability of harboring true structural variation. Eventually, comprehensive sequence data will document the breadth and detail of each variable region and individual variant, as illustrated by fosmid insert sequence data[6] and direct sequence assembly comparisons[39]. In the meantime, various validation strategies have been applied to subsets of putative variants in each of the discovery reports. These included (i) FISH of metaphase, interphase or fiber chromosomes using various clones or PCR-amplified molecules; (ii) PCR or quantitative PCR (qPCR) for allele loss or quantitative variation; (iii) multiple ascertainment, whereby considerable weight was given to whether or not a putative variant was seen in more than one individual or had been reported in previous studies; (iv) array CGH to validate computational screening results[6,7] or for finer resolution of BAC-screening results by oligonucleotide arrays[9,41]; (v) sequence analysis of fosmid inserts to confirm calls and to assess some discordant ones[6,9]; (vi) allele-specific fluorescence intensities[10] and (vii) familial clustering[41].

These assays were variously applied to subsets of data, and outcomes were used effectively in some studies[7,10,11] to further evaluate the sensitivity and specificity and/or error rates of the primary screening methods. The proportion of putative variant loci that have been individually validated by means other than multiple ascertainments remains small, presumably due to the technical challenges of the confirmatory tests. All studies provided some information about the frequency of each putative structural variant or region, both as an argument for validation and to characterize the findings. A growing consensus in the field is for more validation of variants using two or more technologies.

Recommendations for standards

Based on our enumeration of the challenges facing this new field and a thorough review of published experimental designs, we provide four broad guidelines that follow the natural progression of experimentation as an initial step toward the development of standards. As the field matures, these guidelines should serve as precursors to stricter standards that undergo regular and comprehensive vetting by the community[48]. We are struck by the resemblance to issues raised by the MIAME (minimum information about a microarray experiment) standards[49], as well as by Lander and Kruglyak[50], with recommendations to find the right balance of stringency and value judgment to avoid as much error as possible without delaying discovery. The latter paper's recommendations for modifiers (suggestive, significant, highly significant and confirmed) might well be adapted for the statistical annotation of structural variants in databases.

In their current form, the recommended standards could also serve as a checklist for reviewers and editors as they assess manuscripts that report structural variation data. Moreover, as more structural variation data are reported and the nature of the variants becomes better understood, curators of databases would be at greater liberty to accept or reject complete or partial datasets according to established quality thresholds.

1. Describing the sample. The study should report the origin of each sample (for example, new or from a repository) and all of its characteristics, including the source (for example, blood, cell line, tissue) and karyotypic status, as well as the age, sex, ethnicity and phenotype (disease or nondisease features) of the donor. For surveys aiming to capture structural variation from the general population for control databases, there should be particular emphasis on detailing the extent of phenotype investigation. The study should also accurately document the genetic relationship of samples and any manipulation of the samples such as cell-culturing conditions or whole genome amplification, including protocols for extracting and labeling samples. Previous publications using the sample and all associated aliases should be listed.

2. Reporting experiments. Upon publication, the researchers must declare all aspects of the experimental design and results, including the experimental platform (for example, all clone or sequence identifiers used in arrays), technical procedures, data extraction and processing protocols, the version of the reference genome sequence used for comparison or annotation, and all validation results. The information must be made available in a format that enables unambiguous interpretation, replication of the experiment and the opportunity for other researchers to reanalyze the data to verify the conclusions[48,49]. For example, many array CGH experiments are performed using different test and reference samples, a variable number of spot replicates and differential use of dye-swap replicates. These methodological details affect the interpretation of the data and inferences regarding the presence or absence of a particular structural variant. Most existing new structural variation data are being generated using microarrays; therefore, suitable repositories include the Gene Expression Omnibus (GEO)[51], ArrayExpress[52] and CIBEX[53] databases. As more sequence data emerge in structural-variation discovery initiatives, it is important that the underlying sequences and traces be made publicly available. Similarly, methodological differences exist in alignment algorithms; in addition to simple lists of sequence differences between assemblies or traces, the underlying alignments from which these events were called should be available.

3. Quality control. All studies should apply stringent criteria to ensure an accurate empirical estimation of the performance of the detection protocol used. Ideally, the parameters of the detection should be calibrated using a limited set of test data to achieve an acceptable level of false positive among the regions that are called. There are several metrics for this estimation, for example, the false discovery rate[54]. Parameters should be set to maximize screening specificity (minimize false positive calls) without undue compromise to sensitivity. To

simplify this process, we recommend that all studies include at least one (and preferably more) standard control sample to be used as a reference for comparison. Initially, we propose sample NA15510 from the US National Institute of General Medical Sciences (NIGMS) Human Genetic Cell Repository, as it has already been characterized using a number of platforms (**Table 2**), and is also now being sequenced. A second reference sample could be NA10851, as it has also been characterized extensively[11].

In addition to calibrating the parameters used for CNV calling, the quality of the total set of variants called across the entire sample set should be assessed. This requires unbiased sampling of the putative variants to be validated: that is, not just assessing those called most frequently, but ensuring representation of the entire frequency distribution. Good examples from the different experimental approaches outlined in **Table 3** include validation of singleton and nonsingleton error rates[11], estimation of fosmid read-pair error rates by sequencing the fosmid[6] and estimation of error rates using a secondary technology such as oligonucleotide arrays[7]. It should no longer be considered sufficient to estimate the error rates by extrapolating from self-self experiments, without confirming that the estimated error rates were indeed correct and investigating how individual experimental error rates translate into study-wide error rates.

4. Describing structural variants. The study should thoroughly report characteristics of the structural variants, including sequence content (start and end points or complete sequence content with appropriate annotation), and population frequency and distribution (if known), including samples and assays used to determine these parameters. A future challenge will be to develop standards for defining CNV regions (CNVRs)—merging data from different individuals and different surveys into a single set of CNVRs. The ideal situation would be that each 'called' CNVR has an audit trail of both the experimental data and the processing of the data to the final call. Robust documentation of standardized CNVRs in databases will require specific rules to be established, and although their description is beyond the scope of this Perspective, the writing of it will stimulate future discussion. For CNVs and CNVRs, the definitions and criteria used by Redon *et al.*[11] offer a good framework to build on (also see **Supplementary Fig. 1**). The current limitations in breakpoint resolution make it difficult to assign specific accession numbers to CNVs. However, once structural variants are described with boundaries mapped at nucleotide resolution, identifiers should be assigned using a nomenclature similar to that currently used for SNPs.

Summary and the future

Many of the issues confronting the field of structural variation will be resolved as advances in technology allow robust and economical analysis of structural variants at the nucleotide level in multiple genomes. Such techniques will include 'tiling path'-coverage oligonucleotide arrays, paired-end sequence relationship comparisons, and partial or complete sequence assembly comparisons. The ultimate standard will be sequence resolution of all structural variation in a defined set of reference individuals to establish a benchmark for genotyping platforms. We do not foresee that any one approach will capture all genetic variation reliably, nor, for at least a few more years, will a single strategy predominate over microarray-based approaches. Therefore, the main challenges from this point onward will surely include managing a huge data volume, integrating information from various discovery platforms and discerning phenotypic implications. New issues will arise, such as how to best annotate structural variation data in individual diploid genome assemblies (arising from personalized sequencing projects), as well as how to put haplotypes of structural variants (with or without SNPs) into context with respect to the latest human reference sequence. Structural variation data should also assist SNP, linkage disequilibrium and gene expression determination, but new database tools will be required to fully interpret the data.

Structural variation discoveries offer the potential to bridge a long-standing gap between cytogenetic and sequence-based investigations, and unify our understanding of genetic variation. Interestingly, at the onset of writing, we tried to sidestep the topic of terminology (and nomenclature), but kept returning to it in some way or another as we worked to define and distill the breadth of issues before us. In fact, it was the issue of terminology that highlighted the extreme heterogeneity in data being published, with the related strengths, caveats and differences in the studies being attributable in part to the different backgrounds of the researchers involved.

An equally intricate issue for data integration in the future will be categorizing structural variants in terms of whether they are 'normal', 'disease-causing' or 'phenotype-associated', as these designations can be part of a continuous range[1,24,55,56]. In **Table 4**, we put forward ideas of annotation modifiers that will assist in maximizing the utility of structural variation information. Molecular cytogeneticists have always been faced with this dilemma and its particular implications in the prenatal or diagnostic setting. Now, with the ability to readily recognize submicroscopic and sequence-level variation, the question of how to differentiate benign and disease-associated structural changes will be increasingly important. There are already well defined examples in which the presence of a structural variant correlates directly with a syndrome or phenotype, such as the many dosage-related microdeletions and duplications that cause genomic disorders[57–63] (also see the DECIPHER database). Family-based studies can demonstrate whether a change is *de novo* or has been inherited and, in the latter case, whether there are likely to be associated phenotypic consequences (noting there are numerous examples of variable expression of phenotype and disease in inherited chromosomal rearrangements)[1,21,55]. Otherwise, large population studies and control and disease reference databases will provide the best source of information about a structural variant's frequency and likelihood of causing a phenotypic outcome.

Notwithstanding the challenges, we believe that the recommendations presented here offer necessary first steps toward standardization of many of the variables that, if ignored, will impede progress. At the same time, we recognize that consensus is important, and that standards require time to mature before adoption and implementation[48]. With some ground rules now set, it is also our intention to continue discussions with the genomic structural variation research community at the most relevant meeting opportunities.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank Dr. Janet Buchanan for assistance in manuscript preparation and D. Pinto, C. Marshall, R. Redon, I. Ragoussis and A. Carson for sharing ideas and unpublished data. The work is supported by Genome Canada/Ontario Genomics Institute, The Centre for Applied Genomics, the Canadian Institutes of Health Research (CIHR), the McLaughlin Centre for Molecular Medicine, the Canadian Institute of Advanced Research and the Hospital for Sick Children Foundation. M.E.H. and N.P.C. are supported by the Wellcome Trust. L.F. is supported by CIHR and S.W.S. is an Investigator of CIHR and holds the GlaxoSmithKline/CIHR Pathfinder Chair in Genetics and Genomics at the Hospital for Sick Children and the University of Toronto.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
2. Freeman, J.L. *et al.* Copy number variation: new insights in genome diversity. *Genome Res.* **16**, 949–961 (2006).
3. Sharp, A.J., Cheng, Z. & Eichler, E.E. Structural variation of the human genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 407–442 (2006).
4. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
5. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
6. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
7. 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. *Nat. Genet.* **38**, 75–81 (2006).
8. 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. *Nat. Genet.* **38**, 82–85 (2006).
9. Locke, D.P. *et al.* Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**, 275–290 (2006).
10. McCarroll, S.A. *et al.* Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**, 86–92 (2006).
11. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
12. Simon-Sanchez, J. *et al.* Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. *Hum. Mol. Genet.* **16**, 1–14 (2007).
13. Vissers, L.E. *et al.* Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities. *Am. J. Hum. Genet.* **73**, 1261–1270 (2003).
14. 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).
15. 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).
16. de Vries, B.B. *et al.* Diagnostic genome profiling in mental retardation. *Am. J. Hum. Genet.* **77**, 606–616 (2005).
17. Koolen, D.A. *et al.* A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat. Genet.* **38**, 999–1001 (2006).
18. Sharp, A.J. *et al.* Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat. Genet.* **38**, 1038–1042 (2006).
19. Shaw-Smith, C. *et al.* Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat. Genet.* **38**, 1032–1037 (2006).
20. Urban, A.E. *et al.* High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **103**, 4534–4539 (2006).
21. Szatmari, P. *et al.* Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat. Genet.* **39**, 319–328 (2007).
22. Zhang, J., Feuk, L., Duggan, G.E., Khaja, R. & Scherer, S.W. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet. Genome Res.* **115**, 205–214 (2006).
23. Cooper, G.M., Nickerson, D.A. & Eichler, E.E. Mutational and selective effects on copy-number variants in the human genome. *Nat. Genet.* **39**, S22–S29 (2007).
24. Lee, C., Iafrate, A.J. & Brothman, A.R. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat. Genet.* **39**, S48–S54 (2007).
25. McCarroll, S.A. & Altshuler, D.M. Copy-number variation and association studies of human disease. *Nat. Genet.* **39**, S37–S42 (2007).
26. Eichler, E.E. *et al.* Completing the map of human genetic variation. *Nature* **447**, 161–165 (2007).
27. Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* **5**, 335–344 (2004).
28. Bennett, S.T., Barnes, C., Cox, A., Davies, L. & Brown, C. Toward the $1,000 human genome. *Pharmacogenomics* **6**, 373–382 (2005).
29. Bentley, D.R. Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* **16**, 545–552 (2006).
30. Service, R.F. Gene sequencing. The race for the $1000 genome. *Science* **311**, 1544–1546 (2006).
31. Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).
32. Mullikin, J.C. *et al.* An SNP map of human chromosome 22. *Nature* **407**, 516–520 (2000).
33. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
34. Report of the Standing Committee on Human Cytogenetic Nomenclature, ISCN 1985. An International System for Human Cytogenetic Nomenclature. *Birth Defects Orig. Artic. Ser.* **21**, 1–117 (1985).
35. Heim, S. Genetic nomenclature: ISCN and ISGN. *Pediatr. Hematol. Oncol.* **13**, iii (1996).
36. den Dunnen, J.T. & Antonarakis, S.E. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum. Mutat.* **15**, 7–12 (2000).
37. Eichler, E.E. Widening the spectrum of human genetic variation. *Nat. Genet.* **38**, 9–11 (2006).
38. Istrail, S. *et al.* Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl. Acad. Sci. USA* **101**, 1916–1921 (2004).
39. Khaja, R. *et al.* Genome assembly comparison identifies structural variants in the human genome. *Nat. Genet.* **38**, 1413–1418 (2006).
40. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
41. Wong, K.K. *et al.* A comprehensive analysis of common copy-number variations in the human genome. *Am. J. Hum. Genet.* **80**, 91–104 (2007).
42. Mills, R.E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16**, 1182–1190 (2006).
43. Carter, N.P. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.* **39**, S16–S21 (2007).
44. Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* **4**, R25 (2003).
45. Bailey, J.A. & Eichler, E.E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet.* **7**, 552–564 (2006).
46. Risin, S., Hopwood, V.L. & Pathak, S. Trisomy 12 in Epstein-Barr virus-transformed lymphoblastoid cell lines of normal individuals and patients with nonhematologic malignancies. *Cancer Genet. Cytogenet.* **60**, 164–169 (1992).
47. Carson, A.R., Feuk, L., Mohammed, M. & Scherer, S.W. Strategies for the detection of copy number and other structural variants in the human genome. *Hum. Genomics* **2**, 403–414 (2006).
48. Burgoon, L.D. The need for standards, not guidelines, in biological data reporting and sharing. *Nat. Biotechnol.* **24**, 1369–1373 (2006).
49. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).
50. Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* **11**, 241–247 (1995).
51. Barrett, T. *et al.* NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.* **35**, D760–D765 (2007).
52. Parkinson, H. *et al.* ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* **35**, D747–D750 (2007).
53. Ikeo, K., Ishi-i, J., Tamura, T., Gojobori, T. & Tateno, Y. CIBEX: center for information biology gene expression database. *C. R. Biol.* **326**, 1079–1082 (2003).
54. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological)* **57**, 289–300 (1995).
55. Feuk, L., Marshall, C.R., Wintle, R.F. & Scherer, S.W. Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum. Mol. Genet.* **15** (special no. 1), R57–R66 (2006).
56. Lee, J.A. & Lupski, J.R. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron* **52**, 103–121 (2006).
57. Lupski, J.R. *et al.* DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell* **66**, 219–232 (1991).
58. Ewart, A.K. *et al.* Hemizygosity at the elastin locus in a developmental disorder, Williams syndrome. *Nat. Genet.* **5**, 11–16 (1993).
59. Chance, P.F. *et al.* Two autosomal dominant neuropathies result from reciprocal DNA duplication/deletion of a region on chromosome 17. *Hum. Mol. Genet.* **3**, 223–228 (1994).
60. Chen, K.S. *et al.* Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. *Nat. Genet.* **17**, 154–163 (1997).
61. Small, K., Iber, J. & Warren, S.T. Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nat. Genet.* **16**, 96–99 (1997).
62. Potocki, L. *et al.* Molecular mechanism for duplication 17p11.2— the homologous recombination reciprocal of the Smith-Magenis microdeletion. *Nat. Genet.* **24**, 84–87 (2000).
63. Kurotaki, N. *et al.* Haploinsufficiency of *NSD1* causes Sotos syndrome. *Nat. Genet.* **30**, 365–366 (2002).
64. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. & Eichler, E.E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
65. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
66. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
67. Budarf, M.L. & Emanuel, B.S. Progress in the autosomal segmental aneusomy syndromes (SASs): single or multi-locus disorders? *Hum. Mol. Genet.* **6**, 1657–1665 (1997).
68. Fiegler, H. *et al.* Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Res.* **16**, 1566–1574 (2006).
69. Komura, D. *et al.* Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res.* **16**, 1575–1584 (2006).
70. Lin, M. *et al.* dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics* **20**, 1233–1240 (2004).
71. Nannya, Y. *et al.* A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res.* **65**, 6071–6079 (2005).
72. Colella, S. *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**, 2013–2025 (2007).
73. Conrad, D.F. & Hurles, M.E. The population genetics of structural variation. *Nat. Genet.* **39**, S30–S36 (2007).