Structural Variation of the Human Genome

Andrew J. Sharp, Ze Cheng, and Evan E. Eichler

Department of Genome Sciences, University of Washington, Howard Hughes Medical Institute, Seattle, Washington 98195; email: eee@gs.washington.edu

Annu. Rev. Genomics Hum. Genet. 2006. 7:407–42

First published online as a Review in Advance on June 16, 2006

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

This article's doi: 10.1146/annurev.genom.7.080505.115618

1527-8204/06/0922-0407$20.00

Key Words

polymorphism, rearrangement, insertion, deletion, inversion

Abstract

There is growing appreciation that the human genome contains significant numbers of structural rearrangements, such as insertions, deletions, inversions, and large tandem repeats. Recent studies have defined approximately 5% of the human genome as structurally variant in the normal population, involving more than 800 independent genes. We present a detailed review of the various structural rearrangements identified to date in humans, with particular reference to their influence on human phenotypic variation. Our current knowledge of the extent of human structural variation shows that the human genome is a highly dynamic structure that shows significant large-scale variation from the currently published genome reference sequence.

BACKGROUND

Since the first identification of large-scale genetic variation, the Bar duplication as visualized in *Drosophila* polytene chromosomes (22), there have been tremendous advances in the detection and understanding of genetic variation. Conventional cytogenetics, utilizing light microscopy coupled with high-resolution chromosome banding techniques (122), can be regarded as the earliest form of whole-genome polymorphism screening, and quickly led to the identification of a number of visible chromosomal variations, or "heteromorphisms," in humans (61). Since then, molecular techniques, in particular the advent of high-throughput sequencing technologies, have revolutionized our understanding of the spectrum of variation in the human genome, from single base pair changes at one extreme, to multimegabase cytogenetic alterations at the other (**Table 1**). However, despite these advances, an intermediate level of human variation, one with resolution below that of the light microscope, but above that of most sequencing-based methodologies, has until very recently gone unnoticed. This review focuses on the rapidly emerging field of human structural variation, specifically submicroscopic rearrangements between 500 bp and 5 Mb in size. We discuss the current scope and limitations of our knowledge, including the types, causes, and consequences of human structural polymorphism, and likely directions of future research.

SEGMENTAL DUPLICATIONS AND STRUCTURAL REARRANGEMENT

Central to an understanding of structural variation is the segmental duplication architecture of the human genome. Segmental duplications (also termed low copy repeats) are blocks of DNA ranging from 1–400 kb in length that occur at multiple sites within the genome and typically share a high level (>90%) of sequence identity (40). Both in situ hybridization and in silico analyses show that ~5% of the human genome is composed of duplicated sequences (10, 29, 30, 126), which can be broadly classified based on their chromosomal distribution. Although some blocks of sequence may be duplicated to multiple locations within a single chromosome (termed intrachromosomal duplication), others are located on nonhomologous chromosomes (inter- or transchromosomal duplication). One notable feature of segmental duplications is their tendency to cluster within pericentromeric and, to a lesser extent, subtelomeric regions (78, 126). Importantly, unlike tandem duplications, they are often interspersed throughout the genome, although the molecular mechanism responsible for this interspersed architecture is poorly understood.

The interspersed nature and high sequence homology of segmental duplications has major implications for human disease, evolution, and, most notably for this review, structural variation, as they provide a substrate for structural rearrangements via nonallelic homologous recombination (NAHR) (**Figure 1**). Many studies have noted a significant association between the location of segmental duplications and regions of chromosomal rearrangement (6, 12, 57, 63, 87, 115, 123, 124, 136), and segmental duplications have been implicated as the probable basis of more than 25 recurrent genomic disorders (58). Molecular studies show that the presence of large, highly homologous flanking repeats predisposes chromosomal regions to rearrangement by NAHR, resulting in the deletion, duplication, or inversion of the intervening sequence. Unequal crossovers between directly oriented repeats on homologous chromosomes can produce reciprocal duplication and deletion products (26), whereas mispairing between inverted repeats results in inversion of intervening sequences (130) (**Figure 1**). Thus, it is likely that many structural variations are not random events, but result from a predisposition to rearrangement due to the duplication

Table 1 The spectrum of variation in the human genome

Variation	Rearrangement type	Size range[a]	References
Single base-pair changes	Single nucleotide polymorphisms, point mutations	1 bp	(3)
Small insertions/deletions	Binary insertion/deletion events of short sequences (majority <10 bp in size)	1–50 bp	(18, 143)
Short tandem repeats	Microsatellites and other simple repeats	1–500 bp	(39)
Fine-scale structural variation	Deletions, duplications, tandem repeats, inversions	50 bp to 5 kb	(34, 54, 87)
Retroelement insertions	SINEs, LINEs, LTRs, ERVs[b]	300 bp to 10 kb	(17)
Intermediate-scale structural variation	Deletions, duplications, tandem repeats, inversions	5 kb to 50 kb	(34, 54, 87, 136)
Large-scale structural variation	Deletions, duplications, large tandem repeats, inversions	50 kb to 5 Mb	(34, 35, 54, 57, 87, 123, 124, 136)
Chromosomal variation	Euchromatic variants, large cytogenetically visible deletions, duplications, translocations, inversions, and aneuploidy	~5 Mb to entire chromosomes	(61, 62)

[a]Size ranges quoted are indicative only of the scale of each type of rearrangement, and are not definitive.
[b]SINE, short interspersed element; LINE, long interspersed element; LTR, long terminal repeat; ERV, endogenous repeat virus.

architecture of the genome, which acts as a catalyst for chromosomal instability. Indeed, detailed analyses of many sites of structural variation often show an intimate association between the location of segmental duplications and sites of polymorphic rearrangement (124, 136; A. Sharp & E. Eichler, unpublished data), suggesting that segmental duplications frequently mediate polymorphic rearrangement of intervening sequences via NAHR, and in addition they are often variable in copy number (48, 124).

TYPES OF STRUCTURAL VARIATION IN THE HUMAN GENOME

Insertions and Deletions

Insertion and deletion events represent the most frequent type of structural variation in the human genome (136), and also the best characterized (**Figure 2**). Some of the earliest mapped human genetic traits, such as color blindness and the Rhesus factor, were shown to result from this type of rearrangement more than 50 years after their initial discovery (36, 141). According to the Human Genome Mutation database, 5% of all mutations associated with simple Mendelian genetic diseases are currently attributed to submicroscopic insertion or deletions (7). Insertion/deletion polymorphisms of several genes with functions in metabolism influence a variety of common phenotypes. A number of drug detoxification enzymes show this type of polymorphism, with some being homozygously deleted in as many as 30% of individuals of certain ethnicity. Copy number changes of cytochrome P450 drug-metabolizing enzymes, such as *CYP2D6*, are associated with variability in metabolism of tricyclic antidepressants and antipsychotic drugs (24), and are also risk factors for laryngyal and lung cancers (1), whereas homozygous deletions of the glutathione S-transferase genes (*GSTT1* and *GSTM1*) are associated with altered risk for a variety of cancers (49, 97).

As insertions/deletions essentially represent the gain or loss of genetic material, a variety of different techniques have been developed to screen for these events based on the associated change in DNA copy number. To date, eight independent genome-wide studies

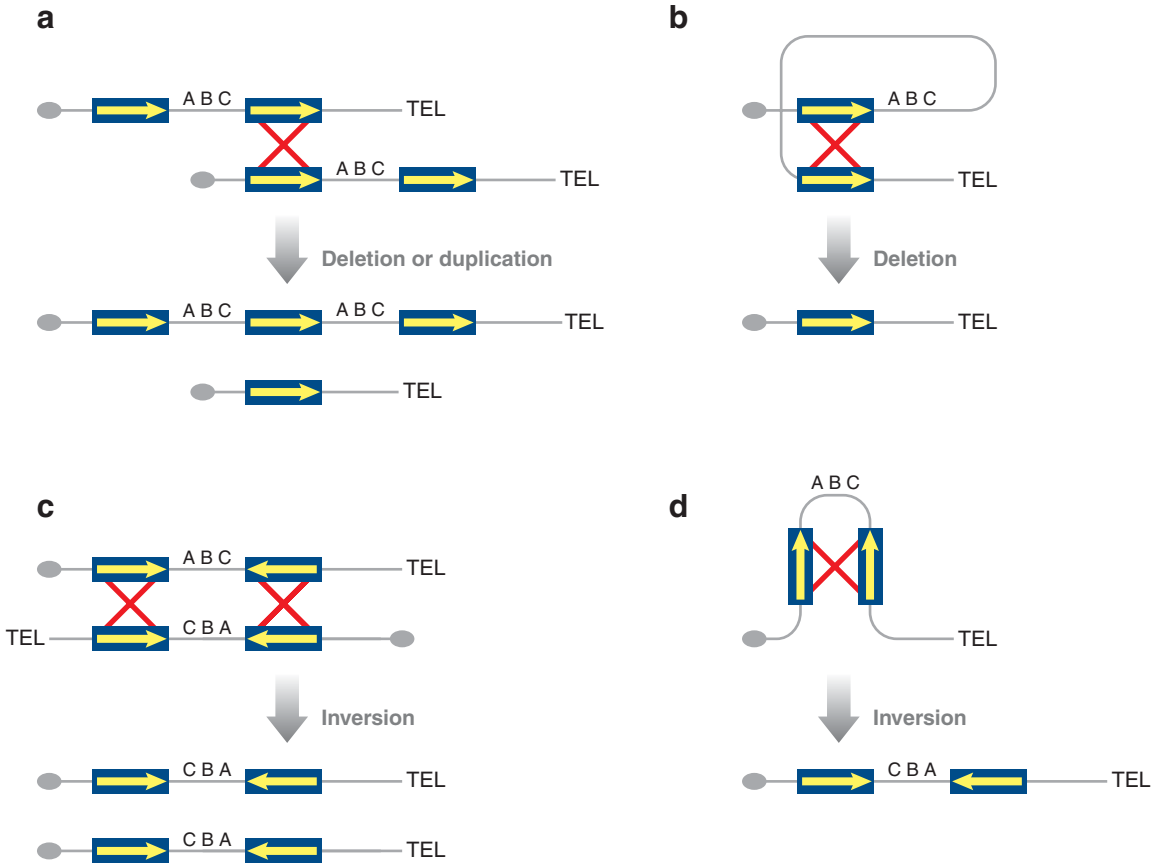

Figure 1

Interspersed segmental duplications provide a substrate for genomic rearrangement via nonallelic homologous recombination (NAHR). (*a* and *b*) Interchromosomal, intrachromosomal, or intrachromatid NAHR between directly orientated repeats causes deletion and/or duplication of the intervening sequence. (*c* and *d*) Interchromosomal, intrachromosomal, or intrachromatid NAHR between inverted repeats causes inversion of the intervening sequence. Repeat sequences are depicted as blue boxes, with their orientation indicated by yellow arrows, and recombination is shown by red crosses. Adapted in part from References 40 and 63.

of insertions and/or deletions have been reported; five using a variety of array-based experimental approaches, and three others using computational methods. One of the earliest studies was reported by Sebat et al. (123), who utilized high-density representational oligonucleotide microarray analysis (ROMA) to identify 76 genomic regions subject to copy number polymorphism, with an average resolution of ~100 kb. In the same year, Iafrate et al. (57) reported the use of Bacterial Artificial Chromosome (BAC) arrays consisting of large-insert clones spaced at ~1 Mb intervals throughout the genome to identify 255 loci that showed variations in copy number. Similar BAC-based array studies, either targeted to regions rich in segmental duplications (124) or using whole-genome tiling arrays (35), have now resulted in a relatively comprehensive first-generation map of large-scale (>50 kb) copy number variation in the human genome, totaling nearly 400

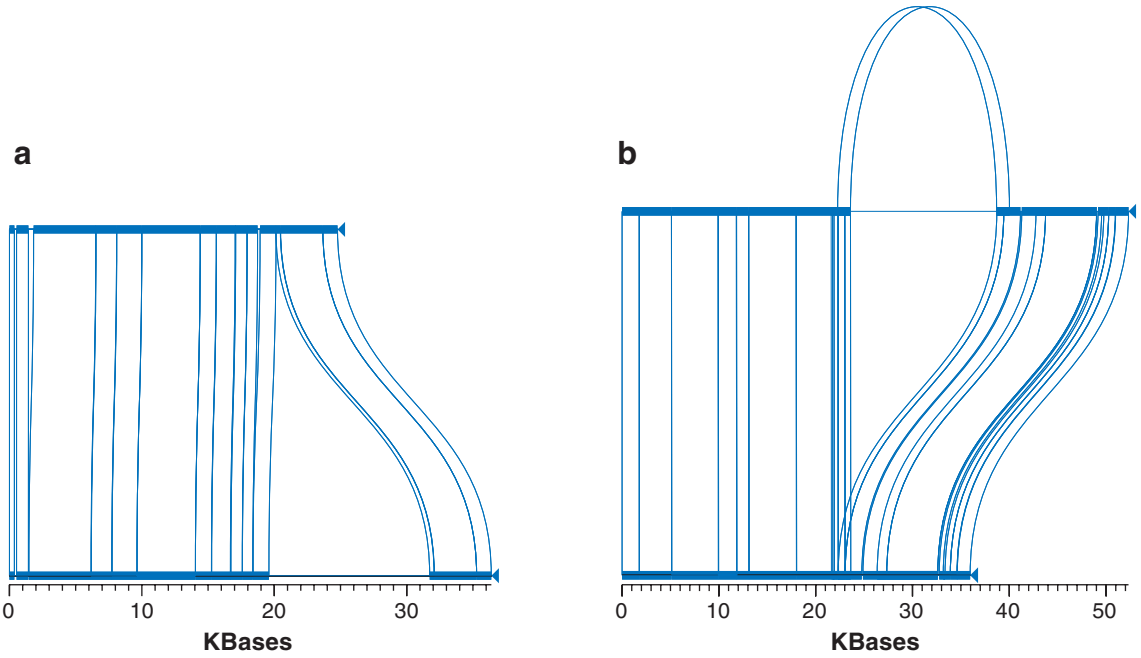

Figure 2

Visualization of insertion and deletion events in the human genome. Miropeats (99) comparison of the sequences of (*a*) fosmid WIBR2–647I01 showing a 12-kb insertion, and (*b*) fosmid WIBR2–1263I16 showing a 14-kb deletion (16), with the corresponding regions of hg17 of the human genome reference assembly (chr15:64,167,000–64,192,000 and chr19:56,805,000–56,857,000, respectively). Lines connect regions of identity both within and between the human genome assembly (*top*) and the fosmid sequence (*bottom*), allowing the structure to be visualized. (*a*) BLAT analysis shows that the 12-kb region inserted in fosmid WIBR2–647I01 is a novel sequence not represented in the human genome reference assembly, and would thus be undetectable by techniques other than paired-end sequence mapping. (*b*) Note the presence of homologous sequences (comprising a 300-bp GT-rich repeat and a 400-bp duplicated sequence) precisely flanking the deletion region, consistent with the hypothesis that the duplication architecture of the human genome predisposes to structural rearrangement by nonallelic homologous recombination.

independent sites covering ~100 Mb of genomic sequence. (Follow the Supplemental Material link from the Annual Reviews home page at **http://www.annualreviews.org** to see **Supplemental Table 1**.)

Alternative approaches have now expanded the scope of copy number variation in the genome, providing resolution well below that afforded by BAC arrays. Computational strategies based on the mapping of fosmid paired-end sequences (136) can detect insertions and deletions down to ~5 kb in size, whereas the use of high-density SNP data, based on clusters of genotyping errors and non-Mendelian transmissions as a signature of deletion events (34, 87), can detect deletions as small as a few hundred base pairs. One further study using ultra-high-density oligonucleotide arrays detected deletion events as small as 70 bp (54). These data suggest that the number of insertion/deletion events in the human genome increases exponentially with decreasing size, a conclusion supported by a study of small events <50 bp (18, 143). Importantly, consistent with a long-standing theory that deleting genetic material is often

selectively disadvantageous (101), several observations suggest that many deletions are subject to negative selection. Both genic regions and the X chromosome, which exists in a haploid state in males, contain an underrepresentation of deletions (34). In addition, there is an apparent excess of rare deletions in comparison to SNPs (54), consistent with the greater action of purifying selection on deletions. Comparable data for insertions, however, are currently lacking.

Tandem Repeats

A number of sites of human structural variation are composed of variable numbers of serially repeated cassettes, some of which may have repeat units several hundred kilobases in size (**Figure 3**). Although many such repeats are not transcriptionally active, a growing list of genes that show wide variation in copy number, such as *AMY1A*, *GSTM1*, and the α- and β-defensins, reside in this type of tandemly repeated array. The underlying

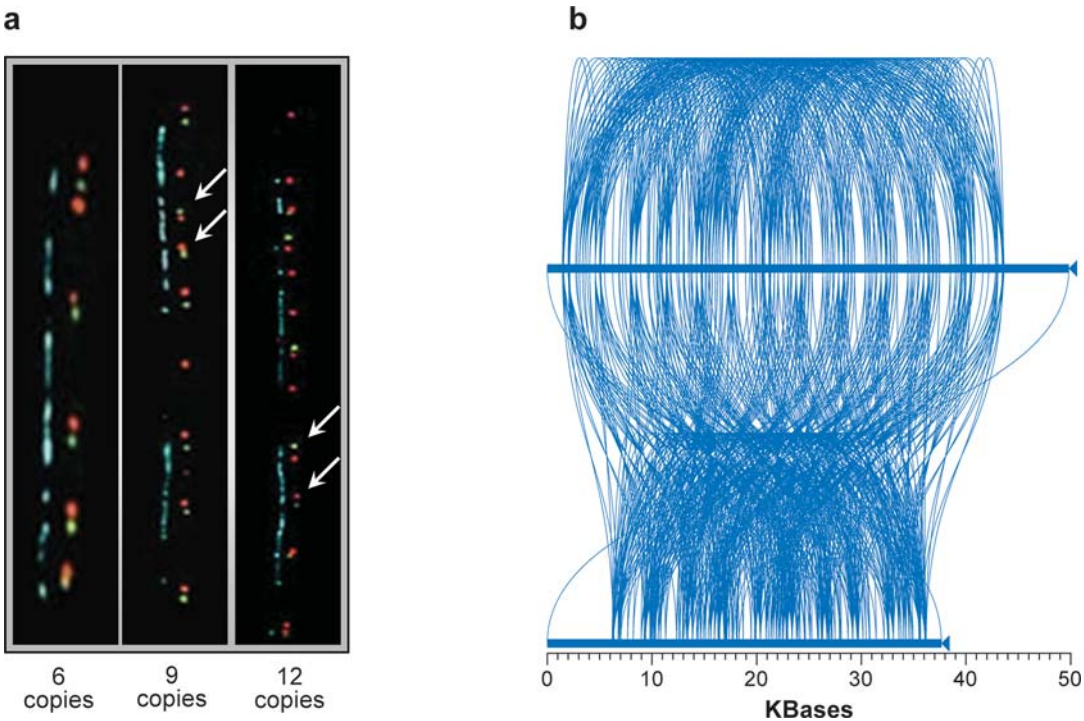

Figure 3

Visualization of tandem repeat structures in the human genome. (*a*) Fiber fluorescent in situ hybridization (FISH) demonstrates variable numbers of tandem repeats of the *AMY1A* gene on chromosome 1. High-resolution fiber FISH was performed on stretched DNA fibers using the BAC RP11–259N12 co-hybridized with a 5′ amylase gene probe (*green*) and a 3′ amylase gene probe (*red*), revealing variable numbers of *AMY1A* tandem gene copies. Six, nine, and twelve gene signals were observed on different chromosomes. Note that the orientation of some gene copies appears inverted within the tandem array (*arrowed*), suggesting complexity induced by multiple rearrangements. The approximate length of the polymorphic region was estimated to vary from 150 kb to 425 kb in these three individuals (adapted from Reference 57). Reproduced with permission from Nature Publishing Group. (*b*) Miropeats (99) comparison of the sequence of fosmid WIBR2–1701E24 (16) with the corresponding region of hg17 of the human genome reference assembly (chr10:124,327,000–124,377,000). Lines connect regions of identity both within and between the human genome assembly (*top*) and the fosmid sequence (*bottom*), allowing the repeat structure to be visualized. This region contains a tandem motif with a repeat size of ~4 kb, which shows variation in copy number from ~8 to ~12 between the two individuals sequenced.

repetitive architecture of these regions likely acts as a substrate for NAHR events, which presumably mediates the expansion and contraction of these repeat arrays.

Cluster analysis of the human genome reference assembly shows that 183 genes contained within the RefSeq database (**http://www.ncbi.nlm.nih.gov/RefSeq/**) have one or more identical copies within a proximity of <1 Mb (mean separation of copies 171 kb), with many arranged either in direct tandem orientations (e.g., *AMY1A*) or contained in larger repeated cassettes (e.g., α- and β-defensins) (A. Sharp & Z. Cheng, unpublished data). Of these 183 genes, 85 (46%) have already been identified as sites of structural polymorphism, some of which contribute to human phenotypic variation. This represents a ~10-fold enrichment ($\chi^2 > 400$, $p < 10^{-40}$) when compared with the ~1300 genes that map to sites of structural variation genome wide, or ~5% of the total human complement of 24,652 genes. (Follow the Supplemental Material link from the Annual Reviews home page at **http://www.annualreviews.org** to see **Supplemental Table 2**.) This strongly indicates that genomic regions with a repetitive architecture are highly prone to rearrangement, a predisposition that makes this particular class of structural variants excellent candidates for sites of recurrent rearrangement. Consistent with this notion, many such loci show multiple different alleles within the population, and evidence from microsatellite analysis of the β-defensin cluster suggests that different copies of the repeat unit may have undergone independent expansion, indicating recurrent rearrangement.

The wide variation in copy number seen at many repetitive arrays makes the development of accurate genotyping assays for these sites technically difficult. Furthermore, the multiple alleles seen at sites of tandem array can result in apparently complex transmission patterns within pedigrees, which, until allele-specific investigations are performed, may appear non-Mendelian (56).

Inversions

Unlike other types of structural variation, inversions are generally presumed to be balanced rearrangements that represent a change in the order or orientation of a DNA segment, and are not thought to be associated with either the gain or loss of genetic material (**Figure 4**). This fact, combined with the increasing use of array-based technologies, which can only detect changes in DNA copy number, has meant that this type of structural variation has remained relatively poorly studied. Prior to 2005, only a handful of polymorphic submicroscopic inversions had been identified in humans (20, 50, 51, 52, 94, 130). The completion of a high-quality human genome assembly paved the way for the development of computational, sequence-based methodologies that can detect fine-scale genomic variation, including balanced rearrangements such as inversions.

Two such studies were recently published. Tuzun et al. (136) compared paired-end sequence data from a high-density fosmid library against the human reference assembly to identify structural variations between these two genomes. In addition to identifying numerous insertion/deletion events, this analysis also identified 56 polymorphic inversions, ranging in size from ~5 kb to 1.9 Mb. Importantly, three quarters of these inversion breakpoints mapped to sites of segmental duplication, suggesting that most inversions within the human genome might be mediated by the presence of flanking repeat structures. Note that many of the inversions identified by Tuzun et al. (92, 136) were accompanied by the gain or loss of material either at, or close to, the breakpoints, demonstrating for the first time that inversions are often not balanced events.

Feuk et al. (47) utilized a different comparative genomics approach. Taking advantage of the draft sequence of the chimpanzee, they performed a cross-species comparison to identify regions of inversion between the human and chimpanzee genomes. Although not

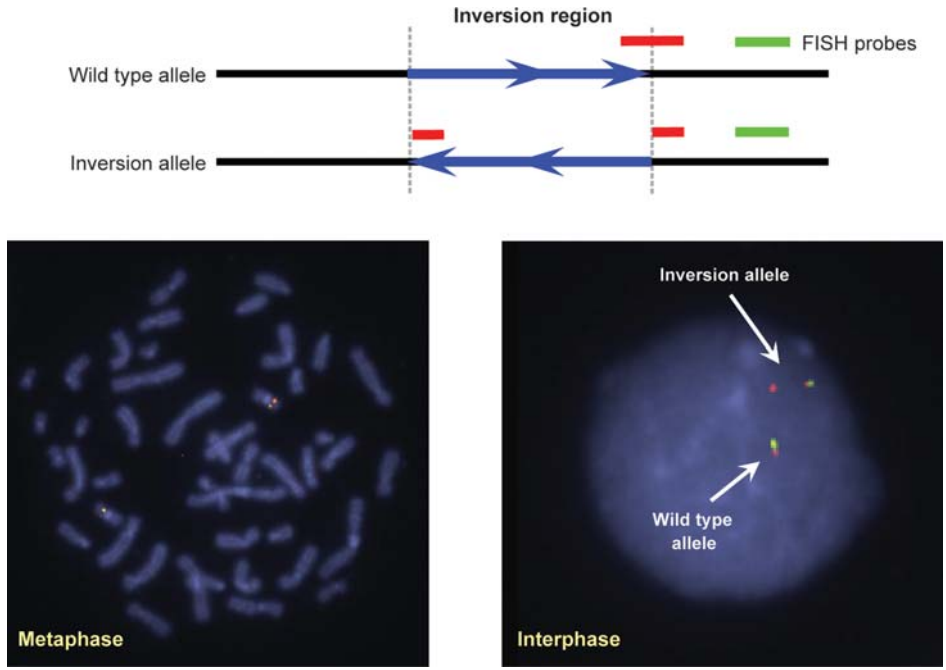

Figure 4

Fluorescent in situ hybridization (FISH) confirmation of a heterozygous 1-Mb inversion. Mapping fosmid paired-end sequences against the human genome reference assembly identified a putative inversion at 16p11.2 (136). Fosmid FISH probes either flanking (*green*) or spanning (*red*) one inversion breakpoint were used. As the inversion is below the resolution of conventional cytogenetics, in metaphase nuclei signals from the two probes are not resolvable. However, in interphase nuclei the inversion results in the splitting of the breakpoint-spanning probe (*red*), seen as dual signals on one allele.

primarily designed to isolate inversions that were polymorphic in the human lineage, 3 of the 23 sites that were verified experimentally fell into this category, with minor allele frequencies ranging from 5–48%. By extrapolation, this suggests that of the 1576 putative human-chimpanzee cross-species inversions detected, 100–200 might also show similar polymorphism in humans, although there are several caveats to this estimate. As observed by Tuzun et al., ~75% of the larger (>25 kb) inversions identified by Feuk et al. were flanked on one or both sides by segmental duplications. Although this included several with pairs of flanking inverted duplications, in other cases the duplications were apparently unrelated.

A significant excess of the interspecific inversions identified by Feuk et al. mapped to the X chromosome. As previous studies have shown the human X to be highly enriched for large inverted repeat structures, which could theoretically catalyze inversion events (142), it is tempting to speculate a mechanistic link between these two observations. Alternatively, two other hypotheses may be put forward to account for this apparent X chromosome bias.

The first is that during male meiosis, pairing between the X and Y is limited to the pseudoautosomal regions. This lack of pairing leads to an elevated rate of X chromosome inversions events specifically in the male germline, such as those seen at the Hemophilia A (113) and Incontinentia Pigmenti loci (5).

The second is that the apparent excess of X-linked inversions may simply be an artifact resulting from a lower quality assembly of

the chimpanzee X chromosome. As the individual sequenced for the chimpanzee genome project was male, with only a single X and Y, this resulted in an ~50% reduction in the number of random shotgun sequences generated to assemble the X compared with the autosomes (31). Because the method used by Feuk et al. depends on a high-quality genome for accurate identification of rearrangements, this reduced read depth and corresponding drop in accuracy of the assembly on the sex chromosomes may explain the apparent X chromosome bias.

However, both of these latter hypotheses also predict a concomitant elevation in the rate of inversion on the Y chromosome, which was not observed by Feuk et al. Furthermore, 20% of the inversions detected by Tuzun et al. using the high-quality human genome assembly also localized to the X chromosome. Considering that the X represents only 5.3% of the total sequence analyzed, this corresponds to a ~fourfold enrichment ($p = 0.04$), indicating that the apparent elevated rate of inversion on the human X chromosome may indeed result from its unique higher-order architecture.

Note that both of these strategies used by Tuzun et al. and Feuk et al. rely on the human genome assembly, which represents a composite haploid sequence of each chromosome. Without high-quality data from other individual genomes, their power to detect polymorphisms is severely limited by the minimal sample size. Given the difficulty of ascertaining submicroscopic inversions, it is likely that the prevalence of these rearrangements is currently underestimated, and that many other undiscovered inversion polymorphisms exist within the human genome.

Perhaps the most notable inversion polymorphism found to date is that reported by Stefansson et al. (131). While generating a detailed physical map of 17q21.31, a region shown by prior studies to have strong linkage disequilibrium (LD) with two highly divergent haplotypes in Europeans, they identified a 900-kb inversion polymorphism. Further analysis showed that chromosomes with these two haplotypes, designated H1 and H2, correlated perfectly with the alternate orientations of this inversion and had diverged ~3 million years ago (mya). As inversions are known to suppress local recombination, this rearrangement explains the divergent haplotype structure at this locus, and also suggests a possible strategy for the genome-wide identification of further inversions in the human lineage. Intriguingly, Stefansson et al. also observed that the haplotype structure of this inversion indicated it has undergone positive selection since its occurrence. Detailed studies showed a small but significant increase in fertility in female carriers of the inversion, explaining how its frequency increased rapidly in Europeans, despite emerging only relatively recently. The exact mechanism by which this inversion causes elevated fertility is unclear, but it may result from the significantly higher genome-wide recombination rate that was observed in these individuals, leading to reduced rates of nondisjunction, an effect that has been noted for some inversions in *Drosophila* (120).

Although inversions can potentially affect gene expression, either by disrupting coding regions that span the breakpoints or by position effects acting on genes adjacent to the breakpoints (69), most inversions are not associated with alterations in gene copy number and thus may not cause an obvious phenotypic effect. However, there is a growing recognition that several polymorphic inversions confer a predisposition to further chromosomal rearrangement in subsequent generations (**Table 2**). To date, three genomic disorders, each of which is characterized by the deletion of a large genomic segment flanked by highly homologous segmental duplications, apparently occur at increased frequencies when the transmitting parent carries an inversion of the segment that is deleted in the affected offspring. In a study of Angelman syndrome, Gimelli et al. (52) observed that a heterozygous inversion of 15q11-q13 was present in four of six (67%) mothers of patients carrying a type II 15q11-q13 microdeletion of maternal

Table 2 Summary of polymorphic inversions that predispose to further rearrangements

Locus	Cytogenetic location	Population frequency	Size of inversion region	Associated predisposition	Ref.
OR genes	4p16	12%	~6 Mb	t(4;8)(p16;p23) translocation	(51)
Sotos syndrome critical region	5q35	Unknown	2.2 Mb	Deletion of SoS critical region	(140)
Williams-Beuren syndrome critical region	7q11.23	Unknown	1.6 Mb	Deletion of WBS critical region (and atypical WBS phenotype?)	(94)
OR genes	8p23	26%	4.7 Mb	inv dup(8p), +der(8)(pter-p23.1::p23.2-pter) and del(8)(p23.1;p23.2)	(50)
Angelman syndrome critical region	15q11-q13	9%	~4.5 Mb	Deletion of AS critical region	(52)
Proximal Yp	Yp11.2	33%	~4 Mb	*PRKX*/*PRKY* translocation (sex reversal)	(64)

origin, compared with 9% in the normal population. Similarly, a heterozygous inversion of the 1.6-Mb Williams-Beuren syndrome critical region at 7q11.23 is observed in one third of parents who transmit a deletion of this same region to their offspring, a frequency much higher than that in the normal population (94). In Sotos syndrome, which is often caused by microdeletion of a 2.2-Mb region at 5q35, a heterozygous inversion of the critical region was detected in all fathers of the children carrying a paternally derived deletion (140). In each case, inversion of the region between the flanking duplications is thought to result in abnormal meiotic pairing, leading to an increased susceptibility to unequal NAHR. Thus, although these inversions have so far only been associated with an increased susceptibility to deletions at these loci, theoretically they may also confer a propensity to the reciprocal duplication, a hypothesis that has not yet been investigated.

Three further examples of inversions that increase susceptibility to subsequent large-scale chromosomal rearrangement have also been identified. Clusters of olfactory receptor genes at 4p16 and 8p23 are both sites of large common inversions that occur in the heterozygous state at significantly higher frequencies in the transmitting parents of individuals with the recurrent t(4;8)(p16;p23) translocation (51), and inverted duplications, marker chromosomes, and deletions of 8p23 (50), respectively. A ~4-Mb inversion in Yp11.2, present on approximately one third of normal Y chromosomes, is observed in nearly all cases of sex reversal with translocation between the X/Y homologous genes *PRKX* and *PRKY* (64). In all of these cases, the inversions presumably predispose to secondary rearrangement by switching the orientation of large, highly identical stretches of sequence on one chromosome homolog, thus allowing their subsequent alignment during synapsis and hence facilitating illegitimate recombination. Furthermore, consistent with the observations of Tuzun et al. and Feuk et al., the breakpoints of all of these inversions have been localized to pairs of large inverted segmental duplications that apparently mediate these rearrangements. We predict that targeted studies of genomic regions flanked by similar paired inverted repeats will likely reveal many more such rearrangements.

In addition to acting as susceptibility factors for further genomic rearrangement, it has long been recognized that heterozygosity for an inversion can also act as an apparent suppressor of local recombination. Progeny analysis of inversion heterozygotes often shows

reduced map lengths near or spanning an inverted region (133). However, in many cases this phenomenon is thought to arise primarily from a failure to recover many lethal recombinant chromosomes, rather than a direct reduction in recombination frequency per se. Meiotic exchange within the inverted region in a heterozygote frequently results in deleterious aberrations, such as di- or acentric chromosomes in the case of paracentric inversions, or deletions and duplications in the case of pericentric inversions; there are many examples in the literature of such structural chromosome abnormalities in the offspring of inversion carriers. However, detailed studies in mouse show that inversion heterozygosity can alter both the frequency and position of meiotic exchange, resulting in small, but significantly increased, rates of nondisjunction (70).

Euchromatic Variants

During the past 20 years, numerous families have been described in which unbalanced chromosome abnormalities segregate, seemingly without phenotypic effect (15) (summarized at **http://www.som.soton.ac.uk/research/geneticsdiv/anomaly%20register/**). Although the vast majority of these observations are confined to single families, and thus cannot be considered polymorphic, a number of such cytogenetically visible aberrations do occur at appreciable frequencies and have collectively been termed euchromatic variants (EVs). In each case, molecular investigations have revealed the underlying molecular basis of these events. All of these EVs appear to represent extreme expansions of duplicated DNA segments that are highly polymorphic in copy number in the normal population, and can therefore be regarded as part of the continuum of copy number variation in the human genome. Thus, although these EVs are cytogenetically visible events outside the scope of this review, they warrant special mention in the larger context of structural variation.

8p23.1v. A number of studies have shown that this variation represents expansion of a ~240-kb cassette in the β-defensin gene cluster. Whereas cytogenetically normal individuals possess 2–12 copies per diploid genome, chromosomes with 7 or 8 copies of this repeat unit are cytogenetically visible as carriers of this EV (56). This locus is discussed in more detail below.

9p12v/9q12v/9qhv. Approximately 6–8% of individuals in the general population carry enlarged chromosome 9 pericentric heterochromatin (82). Euchromatic variations are rarer, and have been shown by fluorescent in situ hybridization (FISH) to represent the exchange of homologous material between proximal 9p and 9q (96). Indeed, the pericentromeric region of chromosome 9 is notable in the human genome, in that it is composed almost solely of large and highly homologous (>99% identity) intrachromosomal duplications (126). Recent array comparative genomic hybridization (CGH) studies using BAC probes covering these pericentromeric duplications reveal frequent copy number variation in this region (124), suggesting that these duplications are highly dynamic in nature, and that the microscopically visible 9p and 9q EVs probably represent extremes of this distribution of variation.

15q11.2v. The presence of extra euchromatic material in proximal 15q results from amplification of a large duplication cassette containing paralogous pseudogenes of *GABRA5*, *NF1*, and the *IGVH* gene. Although most individuals carry between 1 and 4 copies of this duplication, chromosomes with a cytogenetic elongation at 15q are composed of ~20 tandemly repeated copies (112).

16p11.2v. Similar to the *15q11.2v*, this EV also represents the amplification of a cassette containing paralogous pseudogenes, including the creatinine transporter *SLC6A8*, myosin heavy-chain gene, and the immunoglobulin heavy-chain gene *IGH*.

FISH estimates suggest that normal chromosomes possess 2 copies of this cassette, which expands to ~12 copies on EV chromosomes (14).

An over-riding theme of these EVs is that each is composed of multicopy sequences that have undergone duplicative transposition from other genomic locations during recent evolutionary history. There is a distinct bias for EVs to occur in pericentromeric regions, which are known sites for the preferential integration of segmental duplications (10). Given that the sequences underlying each EV are highly polymorphic at the submicroscopic level (124), this suggests that many recently duplicated pericentromeric sequences may be liable to similar amplification, providing a paradigm for the discovery of novel structural variation. A similar phenomenon is observed at many human subtelomeres, which harbor clusters of olfactory receptor genes that are highly polymorphic in copy number (135). Thus, it is likely that regions that have undergone recent segmental duplication represent a rich source of structural variation in our genome.

Transchromosomal Structural Variation

To date, only a handful of transchromosomal structural polymorphisms have been reported, but, this likely reflects an ascertainment bias. Nearly all current genome-wide screening technologies used to identify structural variation are unable to discern the chromosomal location of an insertion unless targeted follow-up studies using techniques such as FISH are performed; in the absence of such data it is presumed that many of these events are local. The presence of large numbers of fixed interchromosomal segmental duplications in the human genome (10) attests to the fact that such rearrangements have occurred on numerous occasions during recent primate evolution. Until a duplicative transposition event becomes fixed in the human lineage, any segmental duplication is, by definition, a structural variant. As there is much evidence to suggest that the process of duplicative transposition is ongoing within humans, the frequency of transchromosomal structural polymorphism, although much lower than that seen intrachromosomally, is likely still significant (**Figure 5**).

Although computational sequence-based detection approaches, such as the paired-end mapping strategy described by Tuzun et al. (136), could theoretically detect interchromosomal structural variation, these events were specifically filtered from the published data set as a means of reducing the false positive rate. However, careful analysis of the finished sequence of several human chromosomes has revealed a number of transchromosomal polymorphisms. For instance, a polymorphic segment containing the *DUSP22* gene is transposed from 16p11.2 to 6p25 (86). The olfactory receptor (OR) genes are the most striking examples of transchromosomal polymorphism. Nearly half of all human subtelomeres contain one or more OR gene copies, and their distribution and copy number show wide variation both within and between ethnic groups (135). Similarly, members of the zinc-finger and immunoglobulin heavy-chain gene families are also located within multiple human subtelomeres and, like the OR genes, show marked variation in their copy number and distribution (111).

A number of studies have demonstrated that translocation, particularly material from the short arms of the acrocentric chromosomes, is a common occurrence within the human population. Approximately 12% of individuals have appreciable amounts of 15p pericentromeric satellite III DNA sequences translocated onto the short arm of another acrocentric chromosome, usually chromosome 14 (132), suggesting that these sequences are particularly prone to interchromosomal rearrangement. Such events are stably inherited between generations, and, as they apparently consist solely of heterochromatic material, are thought to have little or no phenotypic impact.

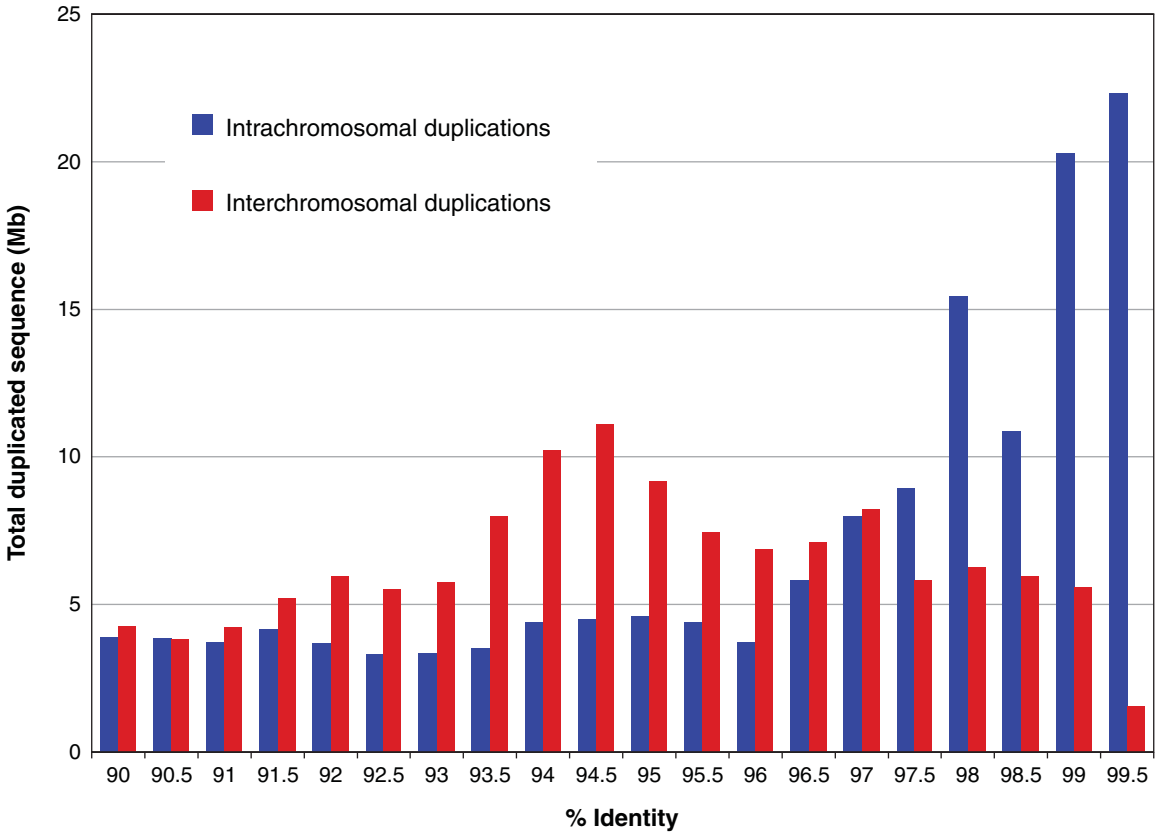

Figure 5

Relative rates of intra- and interchromosomal segmental duplication in the human genome. Total amount of duplicated sequence >90% identity in hg17 is shown in bins of 0.5%. In the absence of gene conversion and assuming a constant rate of substitution, the percent identity of duplicated sequences provides an estimate of their evolutionary age. Overall, the amount of sequence duplicated within homologous chromosomes (intrachromosomal, *blue*, 143 Mb) is slightly greater than that duplicated between chromosomes (interchromosomal, *red*, 128 Mb). However, during recent primate evolution, there has been an increase in the relative amount of intrachromosomal duplication. Considering duplications with >97.5% identity (corresponding to events that occurred within the last ~10 million years), intrachromosomal duplications outweigh interchromosomal events ~threefold (78 Mb versus 25 Mb, respectively), and this trend appears to be accelerating. Until they are fixed within the population, such events are, by definition, structural variants. By this measure, there is probably a considerable excess of intrachromosomal structural polymorphism relative to interchromosomal events in modern humans, although ascertainment biases near centromeres and telomeres cannot be excluded.

In rarer cases, acrocentric sequences also integrate into other chromosomes, potentially affecting their patterns of meiotic segregation (109). While conducting a FISH screen for subtelomeric abnormalities in couples with recurrent miscarriage, Cockwell et al. (33) serendipitously observed that several carried cryptic translocation of centromeric and pericentromeric sequences between their acrocentric chromosomes. Visible only with the use of FISH probes, two cases had a rearranged chromosome 13, with the centromeric region derived from the short arm, centromere, and proximal long arm of

chromosome 15. Two further cases carried a der (22), with the 22 centromeres essentially replaced by sequences from the centromeres of chromosomes 14 and 13/21. While these rearrangements were identified in ~4% of their study population, this may well be an underestimate due to the nonspecificity of FISH probes used, which are unable to detect centromere exchanges between chromosomes 13 and 21, or 14 and 22. Although not proven to be the cause of recurrent miscarriage, the frequency of these abnormalities was significantly increased compared with controls, and it seems highly likely that the hybrid centromeres of these derivative chromosomes would make them prone to abnormal segregation leading to the formation of nonviable zygotes.

FEATURES OF STRUCTURAL VARIATIONS

Relationship with Genes

Structural variation creates abundant diversity of the genic complement in the human population. Studies published to date (34, 35, 54, 57, 87, 123, 124, 136) show that the coding regions of some 800–1800 known genes are subject to variation in copy number through deletion, duplication, or tandem rearrangement. (Follow the Supplemental Material link from the Annual Reviews home page at **http://www.annualreviews.org** to see **Supplemental Tables 1 and 3**.) The uncertainty in this estimate is largely due to the use of BAC array CGH technologies in a number of studies, a technique that provides relatively poor precision on the boundaries of insertion/deletion events. Tandem arrangements in particular can often result in wide variation of gene copy number at certain loci, including the *AMY1A*, defensin, and cytochrome P450 gene clusters. A growing number of common structural rearrangements that alter gene copy number have been implicated in various human phenotypes (24) (**Table 3**). Given that only a fraction of structural polymorphisms have been investigated to date, it seems likely that this type of variation plays a fundamental role in influencing human disease susceptibility.

Structural variations may influence gene expression in various ways. Insertion, deletion, or tandem repetition can add or remove entire copies of genes, leading to a concomitant change in gene dosage, which for genes that are dosage sensitive will influence carrier phenotype (haplo-sensitive, triplo-sensitive, or imprinted). However, because of selective constraints, it is likely that most genes involved in common polymorphic rearrangements are reasonably tolerant of changes in copy number, and are associated with more subtle alterations in phenotype. Insertions, deletions, or inversions involving only part of a gene can potentially result in the formation of variant proteins through exon shuffling, the creation of splice variants, or even novel fusion genes, although the majority of such events are likely nonfunctional unless an open reading frame is maintained. Structural variations outside of coding regions can also lead to changes in gene expression through position effects that might alter the location or effect of essential regulatory elements, such as that seen at the red/green pigment genes (36), or through changes in local chromatin structure (69). Finally, rather than affecting phenotype through direct alterations in gene copy number, deletion polymorphisms may also act by revealing recessive mutations on the single remaining haplotype.

Note, however, that not all variations in gene copy number necessarily result in concomitant changes in protein levels. For example, studies of individuals carrying different copy numbers of α-defensin genes have shown no relationship between total *DEFA1/DEFA3* mRNA levels and gene copy number, suggesting the existence of additional *trans*-acting factors that mediate mRNA levels at this locus (2). This phenomenon seems to be highly variable at different loci. Similar studies of other structurally variant genes show that between 26% and 88% of the

Table 3 Summary of common genic structural variations with known phenotypic effect

Gene name(s)	Locus	Population frequency	Diploid copies	Size of variant segment	Associated phenotype	Ref.
GSTM1	1p13.3	>3%	1–3	18 kb	Altered enzyme activity	(88)
RHD	1p36.11	15–20%	0–2	~60 kb	Rhesus blood group sensitivity	(141)
SMN2	5q13.2	~60%	1–4	500 kb	Altered severity of spinal muscular atrophy	(46)
CYP21A2	6p21.32	1.6%	2–3	35 kb	Congenital adrenal hyperplasia	(73)
LPA	6q25.3	94%	2–38	5.5 kb	Altered coronary heart disease risk	(74)
α-Defensin gene cluster	8p23.1	~90%	4–14	19 kb	Immune system function	(2, 79)
β-Defensin gene cluster	8p23.1	~90%	2–12	240 kb	Immune system function	(56)
IGHG1 region	14q32.33	12–74%	1–6	5–170 kb	Immune system function?	(106, 116)
CCL3-L1/CCL4-L1	17q12	51%/27%	0–14	>2 kb	Susceptibility to and progression of HIV infection, susceptibility to Kawasaki disease	(25, 134)
CYP2A6	19q13.2	1.7%	2–3	7 kb	Altered nicotine metabolism	(107)
IGL	22q11.22	28–85%	2–7	5.4 kb	Altered Igκ:Igλ in B lymphocytes	(138)
GSTT1	22q11.23	20%	0–2	>50 kb	Altered susceptibility to toxins and cancer	(49, 97)
CYP2D6	22q13.1	1–29%	0–13	Undefined	Altered drug metabolism, increased cancer susceptibility	(1)
OPN1LW/OPN1MW	Xq28	75%	0–4/0–7	15 kb/13 kb	Defective color vision	(91)
Testis-specific genes (*DAZ*, *BPY*, *RBM* families)	Yq11.2	3.2%	0–1	1.6 Mb	Low-penetrance spermatogenic failure	(110)

observed variation in expression levels can be explained by gene copy number in different cases (56, 87), whereas for the α-synuclein gene there is apparently an almost perfect correlation between gene dosage, mRNA, and protein levels (89).

Although many genes are located within sites of structural variation, it has been hypothesized that many structural rearrangements, particularly deletions, may be subject to purifying selection and thus deleterious. Global analysis of the genic content of large deletion polymorphisms indicates that this is the case, with transcribed regions being significantly underrepresented in deletions compared with the genome average (34). Furthermore, there appear to be biases in the types of genes found within structurally variant regions. Transcripts encoding nucleic acid binding proteins or those involved in nucleic acid metabolism are underrepresented in deletion regions (34), possibly reflecting the dosage-sensitive nature that many such genes have because of their fundamental role in

transcriptional regulation and development. Conversely, other classes of genes, particularly those involved in molecular and environmental interaction, appear to be enriched in structurally dynamic regions. Genes involved in inflammation and immune response, drug detoxification, metabolism, surface integrity, signal transduction, and sensory perception are all overrepresented in regions of structural variation (34, 136). Although these genes may not be essential for viability, they are generally involved in responses to environment stimuli, suggesting that structural polymorphisms play an important role in evolutionary adaptation. Consistent with this notion, some genes that show variations in copy number overlap with those identified in studies of positive selection (32) and lineage-specific expansion and contraction (43).

Structural Variation, Segmental Duplications, and the Human Genome Assembly

Recent genome-wide analyses document an intimate association between the location of structural polymorphisms and both segmental duplications and sequence assembly gaps. Segmental duplications are clearly hot spots of chromosomal rearrangement, both in polymorphism and between species (6, 12, 80), and it is estimated that there is an ~4–10-fold enrichment for segmental duplications at sites of human structural variations >10 kb in size (57, 87, 123, 124, 136). Current evidence suggests that this is, in part, due to the propensity for regions flanked by duplications to undergo nonallelic homologous recombination, but also that many segmental duplications are highly polymorphic in copy number.

The presence of recent (i.e., highly homologous) segmental duplications is the most important predictor of gap location in euchromatic sequences, with approximately half of all current sequence assembly gaps occurring as duplicated sequence (41). This is due, in part, to the high-sequence identity of these regions, as most paralogs have >95% identity, hindering the discrimination of multiple distinct copies and resulting in the inadvertent collapse of their sequence to one location (9). However, the co-occurrence of large-scale structural variation between chromosomal haplotypes with many segmental duplications further compounds the difficulty of assembling regions, making accurate contig assembly impossible without specialized strategies to traverse across these rearrangements. This results in the formation of sequence gaps if two structurally different haplotypes have been incorporated into the assembly (41).

Specialized strategies will be required to facilitate gap closure at many sites of structural variation. Sequencing and assembly of single haplotypes, facilitated by construction of contigs from hydatidiform moles or monochromosomal somatic cell hybrids, would eliminate problems that are associated with large-scale structural polymorphisms and segmental duplications in conventional diploid libraries. For example, accurate sequencing and assembly of the human Y chromosome, which is highly enriched for segmental duplications and palindromic repeats, and contains a number of large, common structural variations, was achieved largely because the sequence came from one male donor, and hence a single Y chromosome (129). Other complex structurally variant sites, such as the spinal muscular atrophy (SMA) locus at 5q13.3, have required an excessive redundancy of clones to be sequenced to enable a single continuous haplotype to be reconstructed (118).

Indeed, the SMA locus represents one of the few large structurally variant regions of the genome for which more than one allele has been sequenced and assembled. The region of variation covers at least 1–2 Mb, contains the highest density of segmental duplications on chromosome 5, and shows a complex, highly variable structure, which required considerable effort to sequence and assemble the two individual haplotypes (118). The two sequenced alleles, designated SMAvar1 and

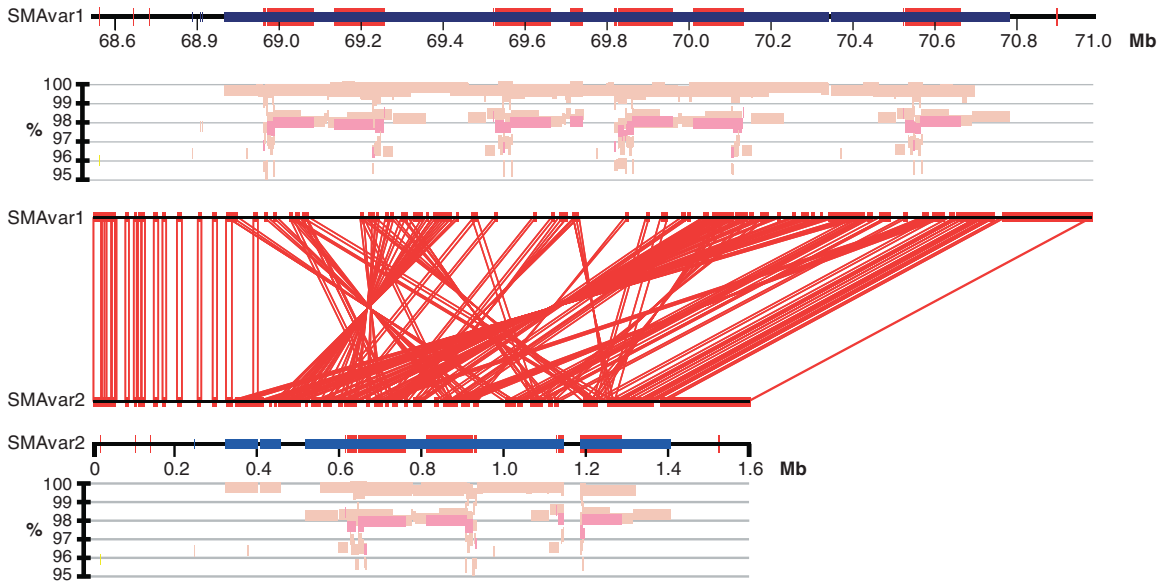

Figure 6

Extensive allelic structural variation at the duplication-rich spinal muscular atrophy (SMA) locus on 5q13.3. Diagram of the SMA region showing both SMAvar1, the published variant, and SMAvar2, the alternative RPC11 variant. (*top*) The duplication pattern for SMAvar1 is shown, with interchromosomal (*red*) and intrachromosomal (*blue*) duplications indicated. Segmental duplications (>95%, >1 kb) are depicted as a function of percent identity (below the horizontal line), with different colors corresponding to the location of the pairwise alignment on different human chromosomes (*light pink* = Chr5; *dark pink* = Chr6). (*center*) A comparison of the interhaplotype structure between the two variants using Miropeats (99), showing complex rearrangements between the two alleles, including multiple inversions and insertion/deletion events. (*bottom*) Duplication pattern for SMAvar2 (adapted from Reference 118). Reproduced with permission from Nature Publishing Group.

SMAvar2, are highly divergent (**Figure 6**), and suggest that a hypothetical ancestral haplotype underwent multiple large rearrangements (>400 kb) and small insertion/deletion events to produce these structures. Considering the extent of variation between these two sequenced alleles, it seems likely that multiple alternate configurations of the SMA region exist in the human population, and suggests that similar extensive structural variation between haplotypes may be a common feature in regions of intense segmental duplication. Data from the sequencing of alleles that differ from the human reference sequence in smaller regions of structural variation captured within fosmid clones support this hypothesis (136; E. Eichler, unpublished data). Thus, large-scale efforts to sequence sites of structural variation taken from numerous individuals will be necessary to fully understand the extent of variation in the human genome, especially in regions composed of numerous high-copy segmental duplications.

Recurrence Rates of Structural Variations and Linkage Disequilibrium

The growing number of genes that show variation in copy number suggests that many structural polymorphisms likely play important roles in common disease. A fundamental aspect of assessing the phenotypic consequences of common structural variation is the determination of their relative frequency in test and control populations. Traditionally,

this type of association study commonly used single nucleotide polymorphisms (SNPs) to detect correlations between a gene or region of interest and a phenotype. To facilitate these studies, recent research has focused on developing a genome-wide LD map, or HapMap, to define chromosomal regions carrying shared combinations of alleles (haplotypes) as a result of common ancestry (3). Given the availability of this haplotype map and the many high-throughput methods available for SNP genotyping, there is considerable interest as to whether structural polymorphisms also show LD with other surrounding markers such as SNPs. That is, can SNP markers be used as surrogates to accurately genotype structural variations, or are specific assays tailored to each structural variation required?

Central to this question is the recurrence rate of structural variation, i.e., do such mutational events recur on different genetic backgrounds? Specifically, if structural variants are typically due to recurrent mutations, then they will occur on multiple haplotype backgrounds, and will show little or no LD with flanking SNPs. Conversely, if a structural variant occurred only once during human evolution on a single ancient haplotype, it will likely show association with nearby genetic markers, which could then be used as informative proxies in disease-association studies.

Using the wealth of data collected on the occurrence of large deletions and duplications at the Xp21 Duchenne Muscular Dytstrophy (DMD) locus, and extrapolating it to the entire human genome, van Ommen (139) attempted to estimate how often de novo structural rearrangements occur, calculating a genome-wide rate for deletion events of ~0.12 per generation, and ~0.02 insertion events per generation. However, although providing a reasonable baseline rate for the frequency of this type of rearrangement in unique regions of the genome, it should be noted that the segmental duplication content of the DMD locus upon which this estimate was based is well below the genome average (0.004% versus ~5%, respectively) (10). As structural variations occur preferentially within regions rich in segmental duplications (57, 58, 87, 123, 124, 136), the true figure for de novo rearrangement may be much higher due to the possibility of recurrent mutations resulting from nonallelic homologous recombination. Furthermore, this estimate sheds little light on whether structural rearrangements are recurrent.

Two recent studies directly address this question. By identifying deletion loci in the same populations characterized during the HapMap project, McCarroll et al. (87) and Hinds et al. (54) examined patterns of LD between deletions, ranging from 70 bp to several hundred kilobases in size, and SNPs in their flanking regions. Both studies found that the majority of deletions examined showed strong LD with flanking SNP markers, with some variants having perfect SNP proxies. Furthermore, by examining a number of different ethnic groups, McCarroll et al. found that these deletions were associated with the same SNP alleles in each population, indicating that these represented ancestral mutation that occurred before the divergence of human populations. These observations apparently suggest that, at least in the majority of cases, the identification of tagging SNPs will enable deletion polymorphisms to be effectively genotyped through the use of these neighboring markers, and as such the phenotypic effects of deletion variants could be assayed using conventional LD-based association studies, without the need for specific copy number assays.

Note that not every deletion investigated in these two studies showed significant LD with surrounding markers, and, intriguingly, the results from a third concurrent publication suggest the opposing view that some large genomic deletion polymorphisms may be recurrent events. Using SNP genotyping data generated as part of the HapMap project, Conrad et al. (34) observed that some deletions that occurred in a single region were often present on different haplotype

backgrounds, suggesting that these were recurrent events that may have occurred in multiple founders. Further data to support this hypothesis came from studies using high-density oligonucleotide arrays, which showed that these polymorphisms were often complex and exhibited multiple different breakpoint alleles in different pedigrees, an observation consistent with the recurrent rearrangement hypothesis.

There are also a number of major caveats to the data presented by McCarroll et al. and Hinds et al. First, regions rich in segmental duplications are difficult to study using SNP-based approaches, as SNP density at these locations is generally poor or inaccurate due to the presence of paralogous sequence variants, which are often incorrectly annotated as SNPs (10, 45). This results in reduced power to detect structural variations by SNP-based approaches in sites of segmental duplication, and reduced power to determine patterns of LD in these regions. As the study of many human genomic disorders clearly demonstrates that the presence of segmental duplications is a strong predisposing factor to recurrent rearrangement (58), studies that use haplotype data as their primary means of detecting structural variations likely suffer an ascertainment bias for polymorphisms that show strong LD.

Second, by focusing on a subset of sites that are not fully representative (see below) of the wide spectrum of human structural variation, these studies may have presented a biased view of LD around copy number variants. For example, McCarrol et al. (87) and Hinds et al. (54) both studied regions of simple haploid deletion, and largely avoided sites of complex, multicopy, or tandem rearrangement that might be predicted to have a higher frequency of variation.

Third, the assay design used by Hinds et al. relied on PCR amplification of genomic fragments ~10 kb in size. This meant that regions flanked by large segmental duplications or other repeat structures would go largely undetected due to the requirement to design unique PCR primers for each amplicon. Furthermore, although Hinds et al. identified a number of deletions that showed multiple different breakpoint alleles, a hallmark likely associated with recurrent rearrangements, these were excluded from subsequent LD analysis (A. Kloek & K. Frazer, personal communication).

Therefore, although significant LD clearly exists at many sites of structural variation in the human genome, it is plausible that a different picture may emerge from other genomic regions with different properties. Sites that exhibit highly labile genomic architecture, such as the tandemly repeated cassettes seen at the *AMY1A* and α- and β-defensin loci, are excellent candidates to undergo recurrent rearrangement. Similarly, sites that show both insertion and deletion alleles suggestive of reciprocal exchange events (125), or those that exhibit a variety of breakpoints in different individuals, are also excellent candidates. Until a full and unbiased assessment is made with particular reference to regions with features that may render them susceptible to increased frequencies of rearrangement, the conclusions drawn from studies published to date should not be extended genome wide (42).

Finally, global analyses suggest that certain genomic features predispose to rearrangement, further suggesting that certain sites might be prone to increased frequencies of rearrangement. Approximately one third of all human segmental duplications terminate within *Alu* repeats that were active during primate evolution, indicating that *Alu-Alu*-mediated recombination plays a significant role in the proliferation of recent segmental duplications (8, 11). Other sites of segmental duplication occur within DNA that has physical properties similar to those of "fragile sites" where genetic rearrangements frequently occur—specifically, a decreased helix stability and an increased DNA flexibility (144). It is likely that loci with similar properties might have an increased propensity to undergo recurrent rearrangement, suggesting

a rationale for the identification of recurrent structural variation sites.

Somatic Variation

The extent of structural variation in the human genome that has now been uncovered raises the question of whether these variants are limited to the germline. Although somatic mutation at the nucleotide level is a relatively common event (53), the mechanisms underlying this type of mutation differ from those leading to larger-scale rearrangements, and the data available to date regarding large structural variants are relatively limited. Array-based studies of the human genome suggest that somatic variation, at least within the limits of current methodologies, is not a widespread phenomenon, being limited to the immunoglobulin gene clusters at 2q11, 22q11.2, and 14q32.3, which are known to undergo somatic V(D)J-type recombination (123; D. Locke, A. Sharp & E. Eichler, unpublished data). However, these studies can only detect changes that occur in the majority of cells within a tissue sample under investigation.

Evidence from the study of genomic disorders does show that NAHR can occur mitotically. In numerous cases, somatic mosaicism for duplication-mediated rearrangements has been identified (67, 77, 145). There is also evidence suggesting that NAHR plays a significant role in loss of heterozygosity, chromosomal instability, and amplifications in tumorigenesis (16, 114, 117, 137). In addition, some limited cytogenetic analysis suggests that copy number at the 8p23.1 β-defensin gene cluster may show amplification in some cells (J. Barber, personal communication). Thus, although wide-scale somatic structural variation of the human genome seems unlikely, limited changes at specific sites may occur. If confirmed, any such sites of somatic change would also be excellent candidates for sites of independent recurrent rearrangement in the germline.

METHODS FOR THE STUDY OF STRUCTURAL VARIATION

The limited resolution of karyotype analysis using the light microscope (~3–5 Mb) meant that, until recent times, the vast majority of structural variation present in our genomes went unrecognized. Building on early isotopic labeling techniques (95), the development of FISH (102), and subsequently stretched-fiber FISH (98), for the first time allowed specific sequences to be localized within the nucleus with high resolution, although its labor-intensive nature precludes its use as a genome-wide screening tool.

Undoubtedly, the major technological breakthrough in the field of structural variation was the development of microarrays (104). Composed of thousands of locus-specific probes immobilized onto a solid substrate, typically a glass slide, microarrays allow entire genomes to be compared at very high resolution. Initially composed of large-insert clones spaced throughout the genome (103), BAC array CGH has now developed to a point where tiling-path arrays allow virtually the entire human genome to be analyzed for the gain or loss of material in a single experiment (35, 60). However, because of the large insert size of the probes (typically ~150 kb), the resolution of BAC arrays is limited to genomic alterations in excess of ~50 kb. Despite this, studies using BAC arrays (35, 57, 124) have led to the identification of several hundred sites of structural variation (**Supplemental Table 1**), indicating that many variations in humans are large-scale events. The ability to comprehensively screen the genome of a large number of individuals has effectively led to widespread use of this platform.

Other types of microarrays can potentially achieve much higher resolution. Both cloned cDNAs (105) and single-stranded PCR products (38, 83) used as probes on microarrays are effective for measuring DNA copy number alterations at the single-exon level. Perhaps the most promising alternative is the development of oligonucleotide arrays, which utilize

synthetic probes 25–75 bp in length (21). This type of array has the advantage of being rapidly synthesized with very high uniformity and density (currently several hundred thousand features per slide), and customized to target virtually any region of interest with high resolution. However, due to the variability in individual probe performance, multiple oligonucleotides are generally required to detect structural copy number variants reliably at a genome-wide level (123). Although these approaches are attractive, high-density oligonucleotide arrays necessary to cover the entire genome are currently extremely costly. Alternative oligonucleotide arrays, designed primarily for SNP genotyping, have also been utilized for copy number measurements with reduced resolution (108), whereas other designs have relied on reducing the complexity of the genome by prior digestion with restriction enzymes and subsequent fragment amplification (81, 123).

However, the advent of novel computational methods which take advantage of the large amounts of publicly available sequence data promises to further revolutionize the detection of structural variation. One such methodology, reported by Tuzun et al. (136), utilized over 1.1 million paired-end sequences from a high-density fosmid library (59). Given the physical properties of the fosmid vector, which tightly limits the cloned insert to a size of ~40 kb, mapping these paired-end reads against the reference assembly identified numerous structural rearrangements between these two genomes with a resolution of ~8 kb. This approach has the advantage of essentially cloning any structural variation identified, thus allowing it to be fully sequenced and precisely characterized, and, in contrast to array-based approaches, is also able to detect balanced rearrangements such as inversions. Like many other technologies, paired-end mapping suffers from reduced power to map structural variation in regions of near-perfect sequence identity where end sequences do not map uniquely. Nevertheless, this strategy is unique among current technologies in being able to detect and map novel sequences not represented in the reference assembly. Variant regions that are polymorphically deleted from the human genome reference are essentially undetectable by all other techniques, as they rely on the reference genome sequence as their starting point for any assay.

Two recent studies (34, 87) utilized a different form of sequence data—namely the availability of high-density SNP genotypes generated by the International HapMap Project (3) for detecting structural variation. Both McCarrol et al. and Conrad et al. used transmission patterns of SNP genotypes within parent-offspring trios, reasoning that SNPs contained within hemizygous deletion regions would be incorrectly genotyped, and would manifest as deviations from Mendelian inheritance in carrier individuals. McCarrol et al. further mined these SNP data for detecting deletions, using apparent deviations from Hardy-Weinberg equilibrium and clusters of null genotypes, the latter being a signature of homozygous deletion events. Given an appropriate SNP density, these techniques can accurately identify rearrangements less than 1 kb in size, but are currently limited specifically to the detection of deletions.

A third methodology for detecting structural rearrangements was described by Feuk et al. (47), who utilized the sequence assembly from chimpanzee to perform a cross-species comparison with the human genome as a means of identifying inversions between the two species. Although this method was aimed primarily at identifying interspecific rearrangements, several polymorphic human inversions were discovered, suggesting that many evolutionarily recent rearrangements are not yet fixed within the human population.

Different techniques for the genome-wide detection of structural rearrangements, and their respective advantages and disadvantages, are summarized in **Table 4**. Although these are all useful for discovering structural rearrangements, there is currently a need for more targeted assays capable of genotyping

Table 4 Comparison of different techniques for the genome-wide detection of structural rearrangements

Techniques	Advantages	Disadvantages	References
1. Array-based approaches	Can be easily targeted to any study population of interest	Indirect assay does not resolve complexity of events Results are always relative to the reference individual(s) used Unable to detect balanced rearrangements	
Large-insert clone (e.g., BAC) array CGH	Can screen entire genome in a single experiment with complete coverage Relatively inexpensive for large sample populations	Relatively low resolution (max ~50 kb for BAC array CGH) Rearrangement breakpoints are poorly defined	(35, 57, 103, 124)
Oligo- or polynucleotide array CGH/ROMA	Custom designs can target any region at ultra-high density	Probes may lack necessary specificity in nonunique regions (e.g., segmental duplications) ROMA resolution limited by restriction sites Commercially available arrays very expensive	(38, 81, 83, 123)
SNP oligonucleotide microarray	Can also detect loss of heterozygosity and uniparental disomy	Limited by availability of informative SNPs Typically exclude repeat regions and significant portions of segmental duplication	(23, 54, 108)
2. Sequence-based approaches	Provide a more direct assay to resolve underlying complexity of events	Costs for generation of sequence data can be very high Analyses limited by the availability of relevant sequence data	
Paired-end sequence mapping	Allows structural variants to be cloned and sequenced Can detect balanced rearrangements, such as inversions, and novel sequence not represented in reference assembly	Limited by the availability of high-density paired-end sequence data from different individuals Unable to identify insertions greater than the insert size of the vector used Costs for library production and end sequencing are very high	(92, 136)
SNP genotype analysis	Can rapidly utilize SNP data from any population	Limited by the availability and density of SNP data (SNP density is reduced in segmental duplications) Only able to detect deletions with current techniques	(34, 87)
Comparative sequence analysis	Can detect balanced rearrangements	Limited by the availability of sequence assemblies Complex regions prone to assembly errors leading to potentially high false positive rates	(47)

structural variants in a cost-effective and relatively high-throughput fashion. Two such techniques, namely multiplex ligation-dependent probe amplification (MLPA) (119) and multiplex amplifiable probe hybridization (MAPH) (55), were recently developed. Both methods rely on the hybridization of multiple locus-specific probes to their target sequences, which are then simultaneously amplified using fluorescently tagged universal primers, and the amount of each resulting product is quantified by capillary electrophoresis. In this way, up to 50 independent loci can be genotyped for copy number in a single reaction, which is relatively rapid and scalable for the study of large populations.

A CASE STUDY OF STRUCTURAL VARIATION: THE DEFENSIN GENE CLUSTER

The defensin gene cluster at 8p23 represents one of the most intensively studied regions of structural variation in the human genome, and makes an excellent example of this type of polymorphism. Defensins are a family of secreted antimicrobial peptides that are thought to be a component of the host defense mechanism (121). The family is subdivided into α-, β-, and θ-defensins, most of which are located in large gene clusters on human chromosomes 8 and 20. To date, 6 human α-defensins (HNP-1–4, HD-5, and HD-6) and 11 human β-defensins (HBD-1–6 and HBD-25–29) have been identified, all of which are thought to be involved in innate immunity. Apparently benign, cytogenetically visible alterations at 8p23.1 were first reported in the early 1990s (75), and, although some cases of duplication at this locus were associated with significant pathology (68, 71), over the next decade it became clear that the vast majority of individuals carrying additional 8p23.1 material did not have any identifiable phenotypic abnormalities (13, 93). As such, it was generally referred to as an EV with no established clinical significance.

Early molecular studies hinted at the underlying nature of this cytogenetic variation, indicating the presence of multiple and variable copies of genes encoding the different types of α-defensin proteins on chromosome 8 (85). More detailed molecular investigations of this EV showed that, in addition to copy number polymorphisms in the α-defensin genes, there was also significant structural variation in the adjacent β-defensin gene cluster. Three of the β-defensin genes, *DEFB4*, *DEFB103*, and *DEFB104*, together with a fourth gene *SPAG11*, which is thought to play a role in sperm maturation, are contained within a repeat unit ~240 kb in length. MAPH and semiquantitative FISH analysis showed that this entire repeat unit is highly polymorphic in copy number, with individuals possessing between 2–12 copies per diploid genome. The study of individuals who carry the 8p23.1 EV revealed that this represents chromosomes with seven or eight copies of this repeat unit, thus identifying the molecular basis of this cytogenetically visible additional chromatin. Interestingly, segregation analysis in different pedigrees using microsatellite markers indicated that different copies of the repeat unit have undergone independent expansion, suggesting that this region may be subject to frequent and recurrent rearrangement. Furthermore, analysis of RNA from different individuals by semiquantitative reverse transcriptase-PCR (RT-PCR) showed a significant correlation between genomic copy number and levels of *DEFB4* mRNA (56).

Subsequent detailed analysis of the neighboring α-defensin locus also revealed a similar theme. The two α-defensin genes *DEFA1* and *DEFA3*, together with the θ-defensin pseudogene *DEFT1*, lie within a 19-kb cassette, which occurs in a tandem array exhibiting variable copy number. Additional gene copies also occur in an adjacent 9.5-kb partial repeat, such that total copy number of these α-defensins ranges from 4–14 in the human population. Analysis of closely related primate species indicates that *DEFA3* is specific to the human lineage, and is completely absent in 10–25% of the normal population, suggesting that this gene has recently evolved and is not yet fully fixed within the human lineage (2, 79). Furthermore, the organization, location, and dosage of *DEFA1* and *DEFA3* in humans are all independently variable within the repeat structure, suggesting that unequal recombination has resulted in their reorganization between different alleles during human evolution. Although the peptide encoded by *DEFA3* appears to have a lower potency than *DEFA1* against various common bacterial pathogens (44), it is tempting to speculate that its emergence in humans may represent a novel gain of function for this gene family by duplication and subsequent divergence.

Do these structural variations in the defensin genes have any phenotypic influence? Defensins clearly form an important part of the immune system, as defensin-deficient individuals are prone to a variety of recurrent infections. Although partial or complete deficiency of HNP-3, encoded by *DEFA3*, is relatively common (2), the small number of reported patients who lack all three α-defensins (HNP-1, HNP-2, and HNP-3) suffer from frequent common bacterial infections. Although no disease-association studies with defensin copy number have yet been reported, some evidence suggests that high levels of α-defensin proteins may slow tumor proliferation (90). However, SNP-association studies indicate *DEFB1* influences susceptibility to both asthma (76) and opportunistic oral infections (66). *DEFB4* has also been suggested as a modifier locus for cystic fibrosis (CF) because of its efficacy against *P. aeruginosa*, a major cause of morbidity in CF patients (128). Thus, it is likely that the extensive variation in defensin copy number seen in humans and other primates plays a significant role in mediating immune phenotype.

CONCLUSIONS AND PERSPECTIVES

Since the first recognition of structural variation by early cytogeneticists, there has been an explosion in the information regarding levels of polymorphism known to exist in the human genome. Currently published data have defined more than 1500 independent sites representing ~100–150 Mb of the human genome as structurally variant (a complete summary of all published structural variations is shown in **Supplemental Table 1**) (34, 35, 54, 57, 87, 123, 124, 136). However, comparing these data sets shows a relatively small amount of overlap between different studies, suggesting that only a fraction of the total amount of structural variation may have been ascertained to date (**Figure 7**).

Although the number of SNPs in the genome far outweighs large-scale variation events, our best estimates indicate that the contribution of structural rearrangements to human genetic polymorphism is significant. Given a de novo single-base mutation frequency of 2×10^{-8} per generation (72), it can be estimated that approximately 120 novel single-base changes occur per individual per generation. By this same measure, assuming that insertion/deletion events have a mean size of ~15 kb (136) and a total de novo rate of ~0.14 per generation (139), structural variation results in approximately 2100 nucleotides of novel polymorphism per individual, i.e., over an order of magnitude higher than that conferred by SNPs. Similarly, assuming a SNP frequency of 1/1200 bp between any two individuals (3), approximately 2.5×10^6 single base pairs of sequence differ between any two individuals. Although the incidence of human structural variation is currently unknown, probably the best estimate to date comes from the study of Tuzun et al. (136). They suggest that ~250 copy number variations with a mean size of 15 kb, or a total 3.75×10^6 base pairs of structural polymorphism, exist between any two diploid genomes, again a greater amount than that conferred by SNPs but at a smaller number of loci.

It is interesting that these figures are proportionately similar to the amount of genetic difference between humans and chimpanzees conferred by SNPs (1.2%) and large segmental duplications (2.7%), respectively, suggesting that large-scale variations play a significant role in both intra- and interspecific variation (28, 31). The estimated 115 Mb of large-scale structural variation between the two species includes several hundred genes that have been gained or lost between the two lineages. Although the biological or evolutionary importance of most of these differences is not known, it is intriguing that several of these structural variants are associated with new or rapidly evolving gene families

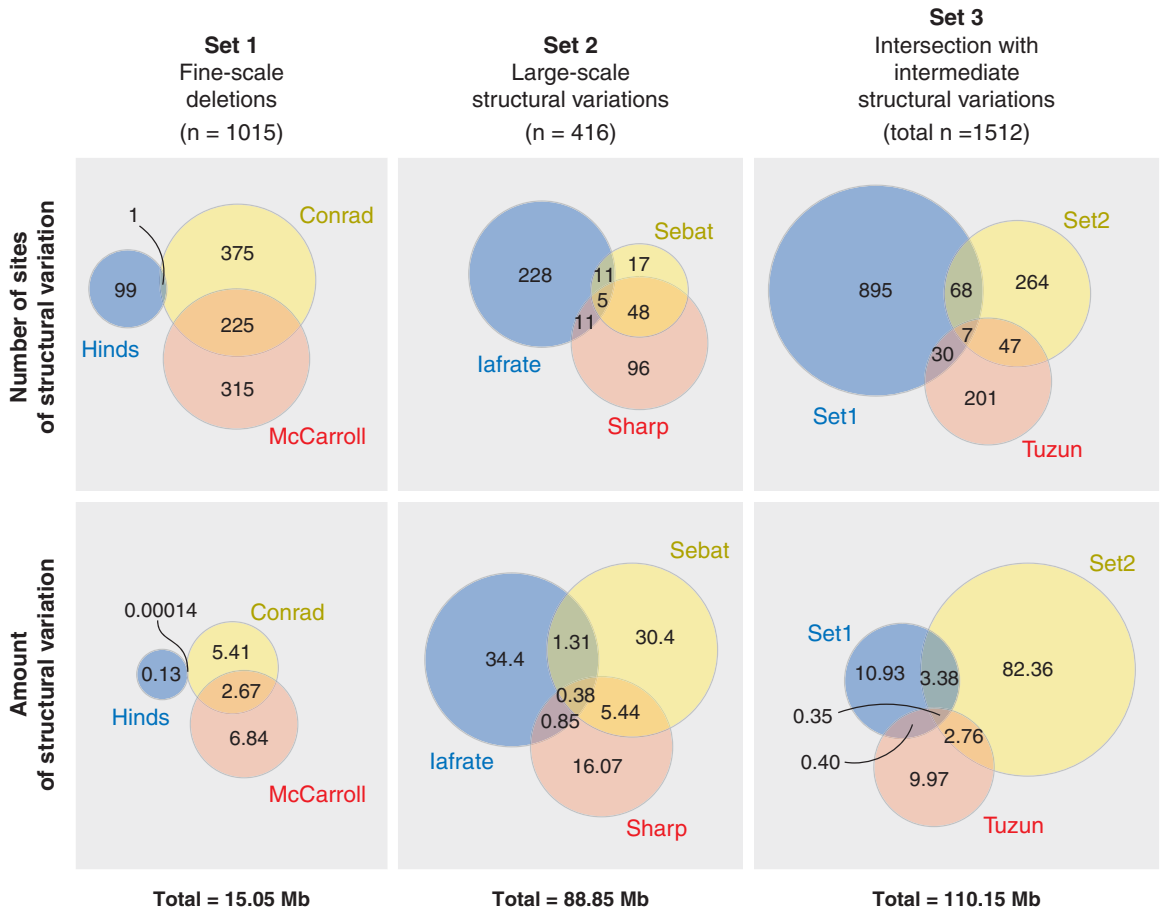

Figure 7

Intersection of studies of structural human genome variation. The relatively small amount of overlap between different studies with similar resolution suggests either that only a fraction of the structural variation in the human genome has been ascertained to date, and/or a high false positive rate in these studies. (*a–c*) Venn diagram comparing the number of intersecting sites having a minimum of a 100-bp overlap. (*a*) Total nonredundant set of 1015 fine-scale deletions (34, 54, 87). (*b*) Total of 416 large-scale copy number variations (57, 123, 124). (*c*) Total of 1512 structural variants, showing the intersection of fine-scale deletions (*a*) and large-scale copy number variations (*b*) with intermediate structural variations (136). (*d–f*) The same data sets shown in (*a–c*), but comparing the number of intersecting structurally variant base pairs within the human genome (hg16). (*d*) Total of 15.05 Mb of deleted base pairs. (*e*) Total of 88.85 Mb of large-scale copy number variations. (*f*) Total of 110.15 Mb. Differences in the resolution and specific methodologies used in each study likely explain why only 25% of the deletion sites are shared, even though many of the same samples were analyzed (34, 87). Web browsers that include these sites along with other published studies of structural variation may be found at **http://humanparalogy.gs.washington.edu/structuralvariation** and **http://projects.tcag.ca/variation/**. Slight differences in the expected totals are due to mapping inconsistencies that arose when converting the sites to the same build (hg16). Only validated sites are considered for the Hinds et al. data set (54).

(19, 65, 100) or genes involved in immunity (31, 92), further suggesting that structural variation plays an important role in environmental adaptation.

These statistics illustrate the significant role that structural variation likely plays in defining phenotypic differences. The importance of large deletions, duplications, and inversions in overt genetic disease has been recognized for decades. However, to date only a relatively small number of common phenotypes has been ascribed to structural polymorphisms (classically a polymorphism is defined as having a frequency of $\geq$1% for the minor allele), largely because of the difficulties in identifying these type of rearrangements until recently (**Table 3**). Given the emerging picture of a highly dynamic genome rich in structural variation and a growing list of genes that are associated with these rearrangements, it seems almost inevitable that many of these polymorphisms play important roles in a wide range of common diseases and phenotypes (24). We predict that the investigation of structural variations in relationship to human disease will likely be a rich source of discovery in the future. A major challenge for such work will be the development of high-throughput assays capable of accurately genotyping loci that are highly variable in copy number and show numerous different alleles and haplotypes. Although surrogate "tagging SNPs" could be used at many loci that show strong patterns of LD, it remains to be determined if SNPs will be viable markers for all structural rearrangements. Concerted efforts to determine the extent of LD surrounding the full spectrum of structural variation will likely resolve this issue. Candidate gene approaches based on the known functions of individual genes, or previous linkage or association studies indicating the involvement of a structurally variant region in a specific disease, will aid the definition of phenotypes that might be influenced by a given rearrangement. Looking further ahead, it is likely that some structural variations will also affect other genomic elements, such as microRNAs (4), other noncoding RNAs (27), and conserved nongenic sequences (37), whose significance is currently poorly understood.

To facilitate disease-association studies, a major goal of future research must be to generate an accurate, comprehensive, and centralized genome-wide map of human structural variation. Currently, a number of privately maintained databases exist, such as The Database of Genomic Variants (**http://projects.tcag.ca/variation/**) and the Human Structural Variation Database (**http://humanparalogy.gs.washington.edu/structuralvariation/**), which attempts to curate the growing list of structural variations. However, neither of these resources is fully comprehensive or accurate, in part due to a number of significant problems with current data. First, published data sets have reported the locations of structural variations in several different builds of the human genome assembly, leading to significant mapping discrepancies when attempting to merge or update data from one assembly to another. This problem is compounded by the strong tendency for structural variations to occur in regions rich in both segmental duplications, which are often associated with assembly errors, and also by gaps in the genome assembly. Second, it should be acknowledged that the increasing use of global assays and large test populations to detect structural variation has resulted in published data sets that include a significant false positive rate. There is an increasing danger that without appropriate quality controls and validation, as more studies are published many sites within the genome will be incorrectly annotated as containing structural polymorphisms. One method to address this problem, at least in part, is by accumulating and cross-validating data from multiple studies and assay types, allowing those sites that are reported as being structurally variant by different studies to be defined with high confidence.

To this end, we advocate that a centralized effort should be made to catalog human structural variation in a publicly available

repository, similar to that which currently exists for single nucleotide variation. Data from a number of published structural variation studies have already been incorporated into tracks within recent builds (hg16 and hg17) of the UCSC Genome Browser (**http://genome.ucsc.edu/**). Integration of data in this setting allows structural variation data from multiple studies to be displayed and compared against virtually any other form of genome annotation in an intuitive manner that can facilitate further research into the causes and consequences of structural polymorphism (**Figure 8**).

We also predict a continuing trend toward the identification of progressively smaller genomic rearrangements. However, despite such advances, one major challenge facing researchers will be the accurate definition of the specific sequence changes underlying structural variations, particularly in the case of large-scale events that will require intensive investigation to resolve individual alleles (86, 118). Because of the limitations of most current detection methods, the precise breakpoints and sequence content associated with most sites of structural variation identified to date are poorly defined. These data are vital for understanding the underlying causes of structural rearrangements. Many such regions, particularly those that are composed of highly duplicated sequence, undergo numerous and complex rearrangements, which in some cases are impossible to resolve without complete sequencing across the variant region (**Figure 6**). Large-scale resequencing efforts targeted to structurally variant loci will likely be necessary to give meaningful insight into the spectrum of human variation.

Additionally, in the case of insertion polymorphisms, without the benefit of additional studies using FISH, the chromosomal location of an inserted sequence is generally ambiguous. Although it seems likely that most insertions are local or tandem events, it is currently unclear whether all polymorphic duplications in the human genome are local to their source of origin, or if a proportion are interspersed to other intra- or transchromosomal locations. Detailed mapping and/or sequencing studies will be necessary to resolve this ambiguity.

There is also a requirement to develop techniques that can identify novel polymorphic sequences that are not currently represented in the human genome reference assembly. As most methodologies for detecting structural variation are based on the sequence present in current builds of the human genome, they cannot detect any novel sequences absent from this reference. One solution to this problem is the use of paired-end mapping strategies utilizing large-insert clone libraries derived from numerous donors (136). Of all the techniques currently in use, this strategy is the only one that allows the presence of novel genomic sequences to be identified and mapped within the reference assembly, and the novel insert fully sequenced. With the promise of significant advances in sequencing technology (84, 127), such approaches will likely become increasingly tenable.

The full extent of the amount of structural variation present in the human genome is only just beginning to be appreciated. The availability of a high-quality genome assembly and dense SNP maps, in combination with novel computational strategies to take advantage of these data, and the advent of techniques such as array CGH, now allows the study of an additional type and level of human variation, one between that visualized with the light microscope at one extreme and detected by sequencing-based methodologies at the other. Given the high gene content of many structural variations and the classes of transcript that are enriched in these regions, it seems almost inevitable that structural polymorphisms will play an important and exciting role in dictating human phenotype.

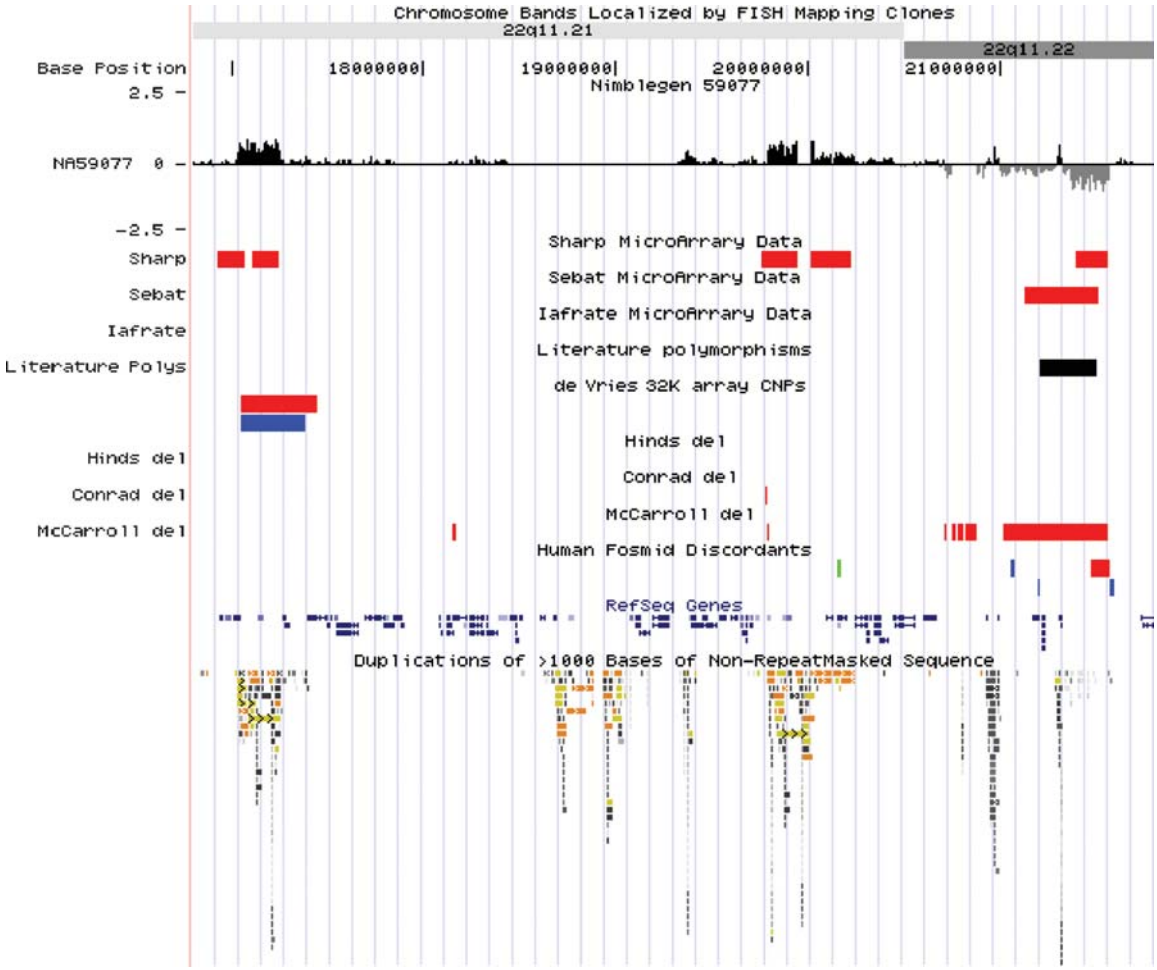

Figure 8

Screenshot taken from a custom genome browser showing a comprehensive view of known human structural variation and segmental duplication content within a 5-Mb region of proximal chromosome 22q (A. Sharp, Z. Cheng & E. Eichler, unpublished data). Regions reported as deletions are shown in red, insertions in blue, and inversions in green. Direct comparison of data from multiple studies allows high confidence to be assigned to sites identified using multiple different methodologies. Note the tendency for structural variations to occur in regions of high segmental duplication content. Custom tracks (from *top* to *bottom*) are as follows: *Nimblegen 59,077*: probe log2 ratio data from a high-density custom oligo array in individual NA59077, used as part of the HapMap project (A. Sharp & E. Eichler, unpublished data). Regions showing consistent positive signals correspond to insertions in this individual compared with the reference sample, whereas consistent negative signals correspond to deletions. Sharp MicroArray data (124); Sebat MicroArray data (123); Iafrate MicroArray Data (57); Literature polymorphisms: location of the IGL gene cluster, which was previously reported as structurally variant (138); de Vries 32k array CNPs: data on transmitted variants (35); Hinds del (54); Conrad del (34); McCarroll del (87); Human Fosmid Discordants (136). Also shown is the physical location in base pairs (hg16), cytogenetic band, genes, and segmental duplications >1 kb and >90% identity. A subset of this data can be viewed on the publicly available University of California at Santa Cruz Genome Browser (**http://genome.ucsc.edu/**).

ACKNOWLEDGMENTS

This work was supported by a fellowship from Merck Research Laboratories (A.J.S.) and the Howard Hughes Medical Institute (E.E.E).

LITERATURE CITED

1. Agundez JA, Gallardo L, Ledesma MC, Lozano L, Rodriguez-Lescure A, et al. 2001. Functionally active duplications of the *CYP2D6* gene are more prevalent among larynx and lung cancer patients. *Oncology* 61:59–63
2. Aldred PM, Hollox EJ, Armour JA. 2005. Copy number polymorphism and expression level variation of the human α-defensin genes *DEFA1* and *DEFA3*. *Hum. Mol. Genet.* 14:2045–52
3. International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437:1299–320
4. Alvarez-Garcia I, Miska EA. 2005. MicroRNA functions in animal development and human disease. *Development* 132:4653–62
5. Aradhya S, Woffendin H, Jakins T, Bardaro T, Esposito T, et al. 2001. A recurrent deletion in the ubiquitously expressed *NEMO* (*IKK-γ*) gene accounts for the vast majority of *incontinentia pigmenti* mutations. *Hum. Mol. Genet.* 10:2171–79
6. Armengol L, Pujana MA, Cheung J, Scherer SW, Estivill X. 2003. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum. Mol. Genet.* 12:2201–8
7. Armour JA, Barton DE, Cockburn DJ, Taylor GR. 2002. The detection of large deletions or duplications in genomic DNA. *Hum. Mutat.* 20:325–37
8. Babcock M, Pavlicek A, Spiteri E, Kashork CD, Ioshikhes I, et al. 2003. Shuffling of genes within low-copy repeats on 22q11 (LCR22) by *Alu*-mediated recombination events during evolution. *Genome Res.* 13:2519–32
9. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001. Segmental duplications: organization and impact within the current Human Genome Project assembly. *Genome Res.* 11:1005–17
10. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, et al. 2002. Recent segmental duplications in the human genome. *Science* 297:1003–7
11. Bailey JA, Liu G, Eichler EE. 2003. An *Alu* transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* 73:823–34
12. Bailey JA, Baertsch R, Kent WJ, Haussler D, Eichler EE. 2004. Hotspots of mammalian chromosomal evolution. *Genome Biol.* 5:R23
13. Barber JC, Joyce CA, Collinson MN, Nicholson JC, Willatt LR, et al. 1998. Duplication of 8p23.1: a cytogenetic anomaly with no established significance. *J. Med. Genet.* 35:491–96
14. Barber JC, Reed CJ, Dahoun SP, Joyce CA. 1999. Amplification of a pseudogene cassette underlies euchromatic variation of 16p at the cytogenetic level. *Hum. Genet.* 104:211–18
15. Barber JC. 2005. Directly transmitted unbalanced chromosome abnormalities and euchromatic variants. *J. Med. Genet.* 42:609–29
16. Barbouti A, Stankiewicz P, Nusbaum C, Cuomo C, Cook A, et al. 2004. The breakpoint region of the most common isochromosome, i(17q), in human neoplasia is characterized by a complex genomic architecture with large palindromic low-copy repeats. *Am. J. Hum. Genet.* 74:1–10

17. Bennett EA, Coleman LE, Tsui C, Pittard WS, Devine SE. 2004. Natural genetic variation caused by transposable elements in humans. *Genetics* 168:933–51
18. Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA. 2005. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet.* 14:59–69
19. Birtle Z, Goodstadt L, Ponting C. 2005. Duplication and positive selection among hominin-specific *PRAME* genes. *BMC Genomics* 6:120
20. Bondeson ML, Dahl N, Malmgren H, Kleijer WJ, Tonnesen T, et al. 1995. Inversion of the *IDS* gene resulting from recombination with *IDS*-related sequences is a common cause of the Hunter syndrome. *Hum. Mol. Genet.* 4:615–21
21. Brennan C, Zhang Y, Leo C, Feng B, Cauwels C, et al. 2004. High-resolution global profiling of genomic alterations with long oligonucleotide microarray. *Cancer Res.* 64:4744–48
22. Bridges CB 1936. The *Bar* "gene" duplication. *Science* 83:210–11
23. Bruce S, Leinonen R, Lindgren CM, Kivinen K, Dahlman-Wright K, et al. 2005. Global analysis of uniparental disomy using high density genotyping arrays. *J. Med. Genet.* 42:847–51
24. Buckland PR. 2003. Polymorphically duplicated genes: their relevance to phenotypic variation in humans. *Ann. Med.* 35:308–15
25. Burns JC, Shimizu C, Gonzalez E, Kulkarni H, Patel S, et al. 2005. Genetic variations in the receptor-ligand pair *CCR5* and *CCL3L1* are important determinants of susceptibility to Kawasaki disease. *J. Infect. Dis.* 15:192:344–49
26. Chance PF, Abbas N, Lensch MW, Pentao L, Roa BB, et al. 1994. Two autosomal dominant neuropathies result from reciprocal DNA duplication/deletion of a region on chromosome 17. *Hum. Mol. Genet.* 3:223–28
27. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 308:1149–54
28. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, et al. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 437:88–93
29. Cheung J, Estivill X, Khaja R, MacDonald JR, Lau K, et al. 2003. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* 4:R25
30. Cheung VG, Nowak N, Jang W, Kirsch IR, Zhao S, et al. 2001. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* 409:953–58
31. Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87
32. Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302:1960–63
33. Cockwell AE, Jacobs PA, Beal SJ, Crolla JA. 2003. A study of cryptic terminal chromosome rearrangements in recurrent miscarriage couples detects unsuspected acrocentric pericentromeric abnormalities. *Hum. Genet.* 112:298–302
34. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* 38:75–78
35. de Vries BB, Pfundt R, Leisink M, Koolen DA, Vissers LE, et al. 2005. Diagnostic genome profiling in mental retardation. *Am. J. Hum. Genet.* 77:606–16
36. Deeb SS. 2005. The molecular basis of variation in human color vision. *Clin. Genet.* 67:369–77
37. Dermitzakis ET, Reymond A, Antonarakis SE. 2005. Conserved non-genic sequences—an unexpected feature of mammalian genomes. *Nat. Rev. Genet.* 6:151–57

38. Dhami P, Coffey AJ, Abbs S, Vermeesch JR, Dumanski JP, et al. 2005. Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. *Am. J. Hum. Genet.* 76:750–62
39. Dib C, Faure S, Fizames C, Samson D, Drouot N, et al. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380:152–54
40. Eichler EE. 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* 17:661–69
41. Eichler EE, Clark RA, She X. 2004. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat. Rev. Genet.* 5:345–54
42. Eichler EE. 2006. Widening the spectrum of human genetic variation. *Nat. Genet.* 38:9–11
43. Emes RD, Goodstadt L, Winter EE, Ponting CP. 2003. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum. Mol. Genet.* 12:701–9
44. Ericksen B, Wu Z, Lu W, Lehrer RI. 2005. Antibacterial activity and specificity of the six human α-defensins. *Antimicrob. Agents Chemother.* 49:269–75
45. Estivill X, Cheung J, Pujana MA, Nakabayashi K, Scherer SW, Tsui LC. 2002. Chromosomal regions containing high density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum. Mol. Genet.* 11:1987–95
46. Feldkotter M, Schwarzer V, Wirth R, Wienker TF, Wirth B. 2002. Quantitative analyses of *SMN1* and *SMN2* based on real-time lightCycler PCR: fast and highly reliable carrier testing and prediction of severity of spinal muscular atrophy. *Am. J. Hum. Genet.* 70:358–68
47. Feuk L, MacDonald JR, Tang T, Carson AR, Li M, et al. 2005. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet.* 1:e56
48. Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ. 2004. Complex SNP-related sequence variation in segmental genome duplications. *Nat. Genet.* 36:861–66
49. Garcia-Closas M, Malats N, Silverman D, Dosemeci M, Kogevinas M, et al. 2005. *NAT2* slow acetylation, *GSTM1* null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. *Lancet* 366:649–59
50. Giglio S, Broman KW, Matsumoto N, Calvari V, Gimelli G, et al. 2001. Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am. J. Hum. Genet.* 68:874–83
51. Giglio S, Calvari V, Gregato G, Gimelli G, Camanini S, et al. 2002. Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. *Am. J. Hum. Genet.* 71:276–85
52. Gimelli G, Pujana MA, Patricelli MG, Russo S, Giardino D, et al. 2003. Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. *Hum. Mol. Genet.* 12:849–58
53. Gottlieb B, Beitel LK, Trifiro MA. 2001. Somatic mosaicism and variable expressivity. *Trends Genet.* 17:79–82
54. Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. 2006. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* 38:82–85.
55. Hollox EJ, Atia T, Cross G, Parkin T, Armour JA. 2002. High throughput screening of human subtelomeric DNA for copy number changes using multiplex amplifiable probe hybridisation (MAPH). *J. Med. Genet.* 39:790–95
56. Hollox EJ, Armour JA, Barber JC. 2003. Extensive normal copy number variation of a β-*defensin* antimicrobial-gene cluster. *Am. J. Hum. Genet.* 73:591–600

57. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al. 2004. Detection of large-scale variation in the human genome. *Nat. Genet.* 36:949–51
58. Inoue K, Lupski JR. 2002. Molecular mechanisms for genomic disorders. *Annu. Rev. Genomics Hum. Genet.* 3:199–242
59. International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431:931–45
60. Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, et al. 2004. A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet.* 36:299–303
61. Jacobs PA. 1977. Human chromosome heteromorphisms (variants). *Prog. Med. Genet.* 2:251–74
62. Jacobs PA, Browne C, Gregson N, Joyce C, White H. 1992. Estimates of the frequency of chromosome abnormalities detectable in unselected newborns using moderate levels of banding. *J. Med. Genet.* 29:103–8
63. Ji Y, Eichler EE, Schwartz S, Nicholls RD. 2000. Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res.* 10:597–610
64. Jobling MA, Williams GA, Schiebel GA, Pandya GA, McElreavey GA, et al. 1998. A selective difference between human Y-chromosomal DNA haplotypes. *Curr. Biol.* 8:1391–94
65. Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, et al. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature* 413:514–19
66. Jurevic RJ, Bai M, Chadwick RB, White TC, Dale BA, et al. 2003. Single-nucleotide polymorphisms (SNPs) in human β-*defensin* 1: high-throughput SNP assays and association with Candida carriage in type I diabetics and nondiabetic controls. *J. Clin. Microbiol.* 41(1):90–96
67. Juyal RC, Kuwano A, Kondo I, Zara F, Baldini A, Patel PI. 1996. Mosaicism for del(17)(p11.2p11.2) underlying the Smith-Magenis syndrome. *Am. J. Med. Genet.* 66:193–96
68. Kennedy SJ, Teebi AS, Adatia I, Teshima I. 2001. Inherited duplication, dup (8)(p23.1p23.1) pat, in a father and daughter with congenital heart defects. *Am. J. Med. Genet.* 104:79–80
69. Kleinjan DK, van Heyningen V. 1998. Position effect in human disease. *Hum. Mol. Genet.* 7:1611–18
70. Koehler KE, Millie EA, Cherry JP, Schrump SE, Hassold TJ. 2004. Meiotic exchange and segregation in female mice heterozygous for paracentric inversions. *Genetics* 166:1199–214
71. Kondoh T. 2001. Clinical manifestations of Coffin-Lowry syndrome associated with de novo 8p23 duplication. *Am. J. Hum. Genet.* 69:A646
72. Kondrashov AS. 2002. Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. *Hum. Mutat.* 21:12–27
73. Koppens PF, Hoogenboezem T, Degenhart HJ. 2002. Duplication of the *CYP21A2* gene complicates mutation analysis of steroid 21-hydroxylase deficiency: characteristics of three unusual haplotypes. *Hum. Genet.* 111:405–10
74. Kraft HG, Lingenhel A, Kochl S, Hoppichler F, Kronenberg F, Abe A, et al. 1996. Apolipoprotein(a) kringle IV repeat number predicts risk for coronary heart disease. *Arterioscler. Thromb. Vasc. Biol.* 16:713–19
75. Krasikov N. 1992. Benign variant 8p23.1? *Am. J. Hum. Genet.* 53:A568
76. Levy H, Raby BA, Lake S, Tantisira KG, Kwiatkowski D, et al. 2005. Association of defensin ß-1 gene polymorphisms with asthma. *J. Allergy Clin. Immunol.* 115:252–58

77. Liehr T, Rautenstrauss B, Grehl H, Bathke KD, Ekici A, et al. 1996. Mosaicism for the Charcot-Marie-Tooth disease type 1A duplication suggests somatic reversion. *Hum. Genet.* 98:22–28
78. Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. 2005. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* 437:94–100
79. Linzmeier RM, Ganz T. 2005. Human defensin gene copy number polymorphisms: comprehensive analysis of independent variation in α- and β-*defensin* regions at 8p22-p23. *Genomics* 86:423–30
80. Locke DP, Segraves R, Carbone L, Archidiacono N, Albertson DG, et al. 2003. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res.* 13:347–57
81. Lucito R, West J, Reiner A, Alexander J, Esposito D, et al. 2000. Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. *Genome Res.* 10:1726–36
82. Madan K, Bobrow M. 1974. Structural variation in chromosome no. 9. *Ann. Genet* 17:81–86
83. Mantripragada KK, Buckley PG, Jarbo C, Menzel U, Dumanski JP. 2003. Development of *NF2* gene specific, strictly sequence defined diagnostic microarray for deletion detection. *J. Mol. Med.* 81:443–51
84. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–80
85. Mars WM, Patmasiriwat P, Maity T, Huff V, Weil MM, Saunders GF. 1995. Inheritance of unequal numbers of the genes encoding the human neutrophil defensins HP-1 and HP-3. *J. Biol. Chem.* 270:30371–76
86. Martin J, Han C, Gordon LA, Terry A, Prabhakar S, et al. 2004. The sequence and analysis of duplication-rich human chromosome 16. *Nature* 432:988–94
87. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, et al. 2006. Common deletion polymorphisms in the human genome. *Nat. Genet.* 38:86–92
88. McLellan RA, Oscarson M, Alexandrie AK, Seidegard J, Evans DA, et al. 1997. Characterization of a human glutathione S-transferase μ cluster containing a duplicated *GSTM1* gene that causes ultrarapid enzyme activity. *Mol. Pharmacol.* 52:958–65
89. Miller DW, Hague SM, Clarimon J, Baptista M, Gwinn-Hardy K, et al. 2004. α-synuclein in blood and brain from familial Parkinson disease with *SNCA* locus triplication. *Neurology* 62:1835–38
90. Muller CA, Markovic-Lipkovski J, Klatt T, Gamper J, Schwarz G, et al. 2002. Human α-defensins HNPs-1, -2, and -3 in renal cell carcinoma: influences on tumor cell proliferation. *Am. J. Pathol.* 160:1311–24
91. Neitz M, Neitz J. 1995. Numbers and ratios of visual pigment genes for normal red-green color vision. *Science* 267:1013–16
92. Newman TL, Tuzun E, Morrison VA, Hayden KE, Ventura M, et al. 2005. A genome-wide survey of structural variation between human and chimpanzee. *Genome Res.* 15:1344–56
93. O'Malley DP, Storto RD. 1999. Confirmation of the chromosome 8p23.1 euchromatic duplication as a variant with no clinical manifestations. *Prenat. Diagn.* 19:183–84
94. Osborne LR, Li M, Pober B, Chitayat D, Bodurtha J, et al. 2001. A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nat. Genet.* 29:321–25

95. Pardo D, Luciani JM, Stahl A. 1975. Localization of the genes of 28S and 18S RNA in human somatic chromosomes by in situ hybridization. *Ann. Genet.* 18:105–9
96. Park JP, Wojiski SA, Spellman RA, Rhodes CH, Mohandas TK. 1998. Human chromosome 9 pericentric homologies: implications for chromosome 9 heteromorphisms. *Cytogenet. Cell Genet.* 82:192–94
97. Parl FF. 2005. Glutathione S-transferase genotypes and cancer risk. *Cancer Lett.* 221:123–29
98. Parra I, Windle B. 1993. High resolution visual mapping of stretched DNA by fluorescent hybridization. *Nat. Genet.* 5:17–21
99. Parsons JD. 1995. Miropeats: graphical DNA sequence comparisons. *Comput. Appl. Biosci.* 11:615–19
100. Paulding CA, Ruvolo M, Haber DA. 2003. The *Tre2* (*USP6*) oncogene is a hominoid-specific gene. *Proc. Natl. Acad. Sci. USA* 100:2507–11
101. Petrov DA. 2001. Evolution of genome size: new approaches to an old problem. *Trends Genet.* 17:23–28
102. Pinkel D, Straume T, Gray JW. 1986. Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. *Proc. Natl. Acad. Sci. USA* 83:2934–38
103. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* 20:207–11
104. Pinkel D, Albertson DG. 2005. Comparative genomic hybridization. *Annu. Rev. Genomics Hum. Genet.* 6:331–54
105. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, et al. 1999. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* 23:41–46
106. Rabbani H, Pan Q, Kondo N, Smith CI, Hammarstrom L. 1996. Duplications and deletions of the human *IGHC* locus: evolutionary implications. *Immunogenetics* 45:136–41
107. Rao Y, Hoffmann E, Zia M, Bodin L, Zeman M, et al. 2000. Duplications and defects in the *CYP2A6* gene: identification, genotyping, and in vivo effects on smoking. *Mol. Pharmacol.* 58:747–55
108. Rauch A, Ruschendorf F, Huang J, Trautmann U, Becker C, et al. 2004. Molecular karyotyping using an SNP array for genomewide genotyping. *J. Med. Genet.* 41:916–22
109. Reddy KS, Sulcova V. 1998. The mobile nature of acrocentric elements illustrated by three unusual chromosome variants. *Hum. Genet.* 102:653–62
110. Repping S, Skaletsky H, Brown L, van Daalen SK, Korver CM, et al. 2003. Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. *Nat. Genet.* 35:247–51
111. Riethman H, Ambrosini A, Castaneda C, Finklestein J, Hu XL, et al. 2004. Mapping and initial analysis of human subtelomeric sequence assemblies. *Genome Res.* 14:18–28
112. Ritchie RJ, Mattei MG, Lalande M. 1998. A large polymorphic repeat in the pericentromeric region of human chromosome 15q contains three partial gene duplications. *Hum. Mol. Genet.* 7:1253–60
113. Rossiter JP, Young M, Kimberland ML, Hutter P, Ketterling RP, et al. 1994. Factor VIII gene inversions causing severe hemophilia A originate almost exclusively in male germ cells. *Hum. Mol. Genet.* 3:1035–39
114. Saglio G, Storlazzi CT, Giugliano E, Surace C, Anelli L, et al. 2002. A 76-kb duplicon maps close to the *BCR* gene on chromosome 22 and the *ABL* gene on chromosome 9: possible involvement in the genesis of the Philadelphia chromosome translocation. *Proc. Natl. Acad. Sci. USA* 99:9882–87

115. Samonte RV, Eichler EE. 2002. Segmental duplications and the evolution of the primate genome. *Nat. Rev. Genet.* 3:65–72
116. Sasso EH, Buckner JH, Suzuki LA. 1995. Ethnic differences of polymorphism of an immunoglobulin *VH3* gene. *J. Clin. Invest.* 96:1591–1600
117. Sawyer JR, Tricot G, Lukacs JL, Binz RL, Tian E, et al. 2005. Genomic instability in multiple myeloma: evidence for jumping segmental duplications of chromosome arm 1q. *Genes Chromosomes Cancer* 42:95–106
118. Schmutz J, Martin J, Terry A, Couronne O, Grimwood J, et al. 2004. The DNA sequence and comparative analysis of human chromosome 5. *Nature* 431:268–74
119. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, et al. 2002. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.* 30:e57
120. Schultz J, Redfield H. 1951. Interchromosomal effects on crossing over in Drosophila. *Cold Spring Harbor Symp. Quant. Biol.* 16:175–97
121. Schutte BC, McCray PB Jr. 2002. β-defensins in lung host defense. *Annu. Rev. Physiol.* 64:709–48
122. Seabright M. 1971. A rapid banding technique for human chromosomes. *Lancet* 2:971–72
123. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* 305:525–28
124. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, et al. 2005. Segmental duplications and copy number variation in the human genome. *Am. J. Hum. Genet.* 77:78–88
125. Shaw CJ, Bi W, Lupski JR. 2002. Genetic proof of unequal meiotic crossovers in reciprocal deletion and duplication of 17p11.2. *Am. J. Hum. Genet.* 71:1072–81
126. She X, Horvath JE, Jiang Z, Liu G, Furey TS, et al. 2004. The structure and evolution of centromeric transition regions within the human genome. *Nature* 430:857–64
127. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, et al. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309:1728–32
128. Singh PK, Jia HP, Wiles K, Hesselberth J, Liu L, et al. 1998. Production of β-defensins by human airway epithelia. *Proc. Natl. Acad. Sci. USA* 95:14961–66
129. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, et al. 2003. The male specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423:825–37
130. Small K, Iber J, Warren ST. 1997. *Emerin* deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nat. Genet.* 16:96–99
131. Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, et al. 2005. A common inversion under selection in Europeans. *Nat. Genet.* 37:129–37
132. Stergianou K, Gould CP, Waters JJ, Hulten MA, et al. 1993. A DA/DAPI positive human 14p heteromorphism defined by fluorescence in-situ hybridisation using chromosome 15-specific probes D15Z1 (satellite III) and p-TRA-25 (alphoid). *Hereditas* 119:105–10
133. Sturtevant AH, Beadle GW. 1936. The relations of inversions in the X chromosome of *Drosophila melanogaster* to crossing over and disjunction. *Genetics* 21:554–604
134. Townson JR, Barcellos LF, Nibbs RJ. 2002. Gene copy number regulates the production of the human chemokine CCL3-L1. *Eur. J. Imm.* 32:3016–26
135. Trask BJ, Friedman C, Martin-Gallardo A, Rowen L, Akinbami C, et al. 1998. Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum. Mol. Genet.* 7:13–26
136. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al. 2005. Fine-scale structural variation of the human genome. *Nat. Genet.* 37:727–32

137. van Dartel M, Hulsebos TJ. 2004. Amplification and overexpression of genes in 17p11.2 ~ p12 in osteosarcoma. *Cancer Genet. Cytogenet.* 153:77–80
138. van der Burg M, Barendregt BH, van Gastel-Mol EJ, Tumkaya T, Langerak AW, van Dongen JJ. 2002. Unraveling of the polymorphic C2–C3 amplification and the Ke + Oz- polymorphism in the human *Ig* locus. *J. Immunol.* 169:271–76
139. van Ommen GJ. 2005. Frequency of new copy number variation in humans. *Nat. Genet.* 37:333–34
140. Visser R, Shimokawa O, Harada N, Kinoshita A, Ohta T, et al. 2005. 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
141. Wagner FF, Flegel WA. 2000. *RHD* gene deletion occurred in the Rhesus box. *Blood* 95:3662–68
142. Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G. 2004. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res.* 14:1861–69
143. Weber JL, David D, Heil J, Fan Y, Zhao C, Marth G. 2002. Human diallelic insertion/deletion polymorphisms. *Am. J. Hum. Genet.* 71:854–62
144. Zhou Y, Mishra B. 2005. Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. *Proc. Natl. Acad. Sci. USA* 102:4051–56
145. Zori RT, Lupski JR, Heju Z, Greenberg F, Killian JM, et al. 1993. Clinical, cytogenetic, and molecular evidence for an infant with Smith-Magenis syndrome born from a mother having a mosaic 17p11.2p12 deletion. *Am. J. Med. Genet.* 47:504–11

Annual Review of Genomics and Human Genetics

Volume 7, 2006

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* chapters may be found at http://genom.annualreviews.org/