Review

Copy number variation: New insights in genome diversity

Jennifer L. Freeman,[1,2] George H. Perry,[1,3] Lars Feuk,[4] Richard Redon,[5] Steven A. McCarroll,[6] David M. Altshuler,[6] Hiroyuki Aburatani,[7] Keith W. Jones,[8] Chris Tyler-Smith,[5] Matthew E. Hurles,[5] Nigel P. Carter,[5] Stephen W. Scherer,[4] and Charles Lee[1,2,9]

[1]*Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA;* [2]*Harvard Medical School, Boston, Massachusetts 02115, USA;* [3]*School of Human Evolution and Social Change, Arizona State University, Tempe, Arizona 85287, USA;* [4]*Department of Genetics and Genomic Biology, The Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada;* [5]*The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom;* [6]*Program in Medical and Population Genetics, Broad Institute of Harvard University and Massachusetts Institute of Technology, Cambridge, Massachusetts 02141, USA;* [7]*Genome Science Division, University of Tokyo, Tokyo, 153-8904 Japan;* [8]*Molecular Genetics Division, Affymetrix, Inc., Santa Clara, California 95051, USA*

DNA copy number variation has long been associated with specific chromosomal rearrangements and genomic disorders, but its ubiquity in mammalian genomes was not fully realized until recently. Although our understanding of the extent of this variation is still developing, it seems likely that, at least in humans, copy number variants (CNVs) account for a substantial amount of genetic variation. Since many CNVs include genes that result in differential levels of gene expression, CNVs may account for a significant proportion of normal phenotypic variation. Current efforts are directed toward a more comprehensive cataloging and characterization of CNVs that will provide the basis for determining how genomic diversity impacts biological function, evolution, and common human diseases.

Genomic variability can be present in many forms, including single nucleotide polymorphisms (SNPs), variable number of tandem repeats (VNTRs; e.g., mini- and microsatellites), presence/absence of transposable elements (e.g., *Alu* elements), and structural alterations (e.g., deletions, duplications, and inversions). Until recently, SNPs were thought to be the predominant form of genomic variation and to account for much normal phenotypic variation (International SNP Map Working Group 2001; The International HapMap Consortium 2005). However, two groups recently reported the widespread presence of copy number variation in normal individuals (Iafrate et al. 2004; Sebat et al. 2004), and these observations have since been replicated and expanded (e.g., de Vries et al. 2005; Schoumans et al. 2005; Sharp et al. 2005; Tuzun et al. 2005; Tyson et al. 2005; Conrad et al. 2006; Hinds et al. 2006; McCarroll et al. 2006; Repping et al. 2006). With this accumulation of information, it now seems appropriate to review our current understanding of copy number variation and its significance in human phenotypic variation (including disease resistance and susceptibility) and to discuss possible future directions for studies in this field.

CNVs in normal individuals

For this review, we prefer to use the term "variant" instead of "polymorphism" when referring to copy number changes. The frequencies of most copy number variants (CNVs) have not yet been well defined in human populations, and "polymorphism" is a term that is usually reserved for genetic variants that have a minor allele frequency of ≥1% in a given population. In our preferred nomenclature, a CNV represents a copy number change involving a DNA fragment that is ~1 kilobases (kb) or larger (Feuk et al. 2006a). At a recent workshop ("The effects of genomic structural variation on gene expression and human disease," The Wellcome Trust Sanger Institute, Hinxton, UK; November 27–28, 2005), it was suggested that CNVs not include those variants that arise from the insertion/deletion of transposable elements (e.g., ~6-kb KpnI repeats) to minimize the complexity of future CNV analyses. The term CNV therefore encompasses previously introduced terms such as large-scale copy number variants (LCVs; Iafrate et al. 2004), copy number polymorphisms (CNPs; Sebat et al. 2004), and intermediate-sized variants (ISVs; Tuzun et al. 2005), but not retroposon insertions. Table 1 lists some of the terminology currently used in the CNV literature.

Large duplications and deletions have been known for some time to be present within the human genome, initially from cytogenetic observations (e.g., Jacobs et al. 1959, 1978, 1992; Edwards et al. 1960; Patau et al. 1960; Coco and Penchaszadeh 1982), but their frequency was presumed to be low and for the most part directly related either to tandemly repeated genes or to specific genetic disorders (e.g., Lupski 1998; Ji et al. 2000; Inoue and Lupski 2002; Stankiewicz and Lupski 2002). In addition, they were often localized to repeat-rich regions such as telomeres, centromeres, and heterochromatin (e.g., Giglio et al. 2001).

A limited number of studies reported the presence of specific large duplications and deletions that were not apparently related to disease (e.g., Barber et al. 1998; Engelen et al. 2000). For example, a deleted region originally thought to be associated with ovarian cancer was later found to also be present in healthy in-

[9]Corresponding author.
E-mail clee@rics.bwh.harvard.edu; fax (617) 264-6861.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.3677206

Table 1. Selected terms in the CNV literature

Term	Definition	Reference
Structural variant	A genomic alteration (e.g., a CNV, an inversion) that involves segments of DNA >1 kb	Feuk et al. (2006a)
Copy number variant (CNV)	A duplication or deletion event involving >1 kb of DNA	
Duplicon	A duplicated genomic segment >1 kb in length with >90% similarity between copies	
Indel	Variation from insertion or deletion event involving <1 kb of DNA	
Intermediate-sized structural variant (ISV)	A structural variant that is ~8 kb to 40 kb in size. This can refer to a CNV or a balanced structural rearrangement (e.g., an inversion)	Tuzun et al. (2005)
Low copy repeat (LCR)	Similar to segmental duplication	Lupski (1998)
Multisite variant (MSV)	Complex polymorphic variation that is neither a PSV nor a SNP	Fredman et al. (2004)
Paralogous sequence variant (PSV)	Sequence difference between duplicated copies (paralogs)	Eichler (2001)
Segmental duplication	Duplicated region ranging from 1 kb upward with a sequence identity of >90%	Eichler (2001)
Interchromosomal	Duplications distributed among nonhomologous chromosomes	
Intrachromosomal	Duplications restricted to a single chromosome	
Single nucleotide polymorphism (SNP)	Base substitution involving only a single nucleotide; ~10 million are thought to be present in the human genome at >1%, leading to an average of one SNP difference per 1250 bases between randomly chosen individuals	The International HapMap Consortium (2003)

dividuals (Lin et al. 2000). Large duplication and deletion variants of portions of gene families/clusters, including olfactory receptors (Trask et al. 1998), major histocompatibility complex (MHC) class III genes (Ghanem et al. 1988), the β-defensin antimicrobial gene cluster (Hollox et al. 2003), and genes at the amylase locus (Groot et al. 1991), were also reported. Moreover, duplications and/or deletions were identified at a golgin-related gene downstream of the promyelocytic leukemia gene (Gilles et al. 2000) and the α7-nicotinic acetylcholine receptor gene (Riley et al. 2002). These and other studies provided the initial evidence that large duplication and deletion events, even if they contained genes, did not necessarily result in the presentation of early onset, highly penetrant genomic disorders or diseases (Buckland 2003).

Recent advancements in technology have facilitated a shift from locus-specific studies to genome-wide assessments of genetic variation. In 2004, two groups independently described the widespread presence of CNVs in the genomes of healthy people with no obvious genetic disorders (Iafrate et al. 2004; Sebat et al. 2004). Iafrate et al. (2004) used a bacterial artificial chromosome (BAC)-based array, with clones chosen at ~1-megabase (Mb) intervals throughout the human genome, together with a technique called array-based comparative genomic hybridization (array CGH; Solinas-Toldo et al. 1997; Pinkel et al. 1998). In this study, the investigators identified >200 loci that contained genomic imbalances among 39 unrelated, healthy individuals representing five populations. The most commonly observed CNV encompassed the amylase locus at chromosome region 1p13.3, and high-resolution fiber FISH analyses showed that this region varied between 150 kb and 425 kb among different individuals. Sebat et al. (2004) amplified BglII-fragments from the genomes of 20 individuals representing nine populations and hybridized these DNAs to a microarray platform containing oligonucleotides spaced at 35-kb intervals throughout the genome (ROMA technique; Lucito et al. 2003). In this study, 76 CNVs with a median size of 222 kb and an average size of 465 kb were identified when a CNV cut-off criterion of three consecutive oligonucleotides was used. On average, 11 CNVs (Sebat et al. 2004) or 12.4 CNVs (Iafrate et al. 2004) were detected in each person with these array CGH assays. Initial commentaries (Carter 2004; Cheung 2004; Buckley et al. 2005) noted the low overlap that appeared to exist between the data sets of these two studies. However, when the CNVs were mapped onto the same build of the human genome, more overlap of the two data sets could be appreciated with CNVs of larger size and frequency (Table 2). Because of the small number of individuals examined and the limited resolution of both array platforms, it seems that the number of CNVs identified by these two studies was an underestimation of the true number of CNVs in humans (Buckley et al. 2005).

Following these two initial studies, Tuzun et al. (2005) used an in silico strategy to compare two human genomes at the DNA sequence level. One genome was represented by the reference human genome sequence (National Center for Biotechnology Information, NCBI build 35). Approximately 67% of this reference sequence originated from a single DNA library (the RPCI-11 BAC library) derived from a single anonymous male. The second genome was in the form of pairs of end-sequence reads from >500,000 fosmid clones of the G248 DNA library. This DNA library was derived from an anonymous North American female of European ancestry. Since the sizes of fosmid clones are tightly regulated at ~40 kb, the investigators reasoned that pairs of end sequences for a given fosmid clone should align to the reference sequence with ~40-kb spacing. Significant deviation of the alignment spacing (i.e., <32 kb or >48 kb) would suggest the presence of a CNV at that locus. Using this criterion, Tuzun et al. (2005) identified 241 CNVs, with most in the size range of 8 kb to 40 kb. More than 80% of these CNVs had not been identified previously, and most were below the expected resolution of the array platforms used in the initial CNV discovery studies (Iafrate et al. 2004; Sebat et al. 2004).

This in silico approach has an added advantage over array-based CNV discovery studies in being capable of detecting other structural genomic variants, namely inversions. These would be detected by consistent discrepancies in the aligned orientation of multiple paired end sequences. In this manner, the investigators identified 56 inversion breakpoints in addition to the 241 CNVs. Together, this suggested the presence of almost 300 putative sites of structural variation when comparing the genomes of two individuals by this method.

One consistent feature of CNVs that was noted in these three CNV studies (Iafrate et al. 2004; Sebat et al. 2004; Tuzun et al. 2005) was the preponderance of CNVs near known segmental duplications, significantly more often than expected by chance alone. Segmental duplications (also referred by some as low copy repeats or LCRs; Lupski 1998) can be defined as dupli-

Table 2. Comparison of CNVs identified by Sebat et al. (2004) to Iafrate et al. (2004) based on the number of individuals and the size of the CNVs

	Number of CNVs in Sebat et al. (2004)	Number also detected in Iafrate et al. (2004)	Percentage
No. of individuals			
1 or more (of 20)	76	15	20%
2 or more (of 20)	30	9	30%
3 or more (of 20)	15	7	47%
4 or more (of 20)	10	5	50%
Size of CNV			
All sizes	76	15	20%
At least 400 kb	27	10	37%
At least 1 Mb	11	5	45%

cated DNA fragments that are >1 kb and found either on the same chromosome or on different, nonhomologous chromosomes (Bailey et al. 2002; Lupski and Stankiewicz 2005). Segmental duplications need not vary in copy number, but if they do vary among individuals, they may also be considered CNVs (Feuk et al. 2006a).

Since a significant portion of CNVs was identified in regions containing known segmental duplications, Sharp et al. (2005) reasoned that a custom array, containing DNA clones targeting these known duplicated regions of the human genome (which are also speculated to serve as potential rearrangement hotspots), might be useful in the rapid identification of CNVs. Forty-seven unrelated individuals representing seven different populations were assessed with this targeted array platform, resulting in the identification of 119 CNVs, of which only 39% had been described previously. Moreover, Sharp et al. (2005) concluded that the sharing of CNVs among several populations meant that these specific genomic imbalances either predated the dispersal of modern humans out of Africa or recurred independently in different populations.

Haploinsufficiency is a condition that results when one copy of a dosage-sensitive gene has been deleted and results in developmental delay or impairment. Likewise, the term haplosufficiency may be a term that could be used to describe genomic deletions that do not result in developmental delay or impairment and can be found in healthy and apparently normal individuals. Recently, three CNV discovery studies that specifically interrogated human genomes for such deletion variants were published concurrently (Conrad et al. 2006; Hinds et al. 2006; McCarroll et al. 2006). Two of these studies (Conrad et al. 2006; McCarroll et al. 2006) relied on available SNP data generated from the International HapMap Project (The International HapMap Consortium 2005). The International HapMap Project was established to study human genetic variation in a cohort of 269 individuals from four populations (The International HapMap Consortium 2003). The first population sample consists of 90 individuals from 30 parent–offspring trios from a U.S. population (in Utah) with Northern and Western European ancestry collected by the Center d'Etude du Polymorphisme Humain (CEPH). The second population sample is from the Yoruban people of Ibadan, Nigeria, also consisting of 90 individuals from 30 parent–offspring trios. The third population sample is 45 unrelated Han Chinese from Beijing, China, and the fourth population sample consists of 44 unrelated Japanese from Tokyo, Japan. Phase I of the HapMap Project provided a SNP genotype at ~5-kb resolution in each of these 269 samples studied for a total of 1.2 million SNPs (The International HapMap Consortium 2005). Phase II has now genotyped an additional 4.6 million SNPs to produce a current total of 5.8 million SNPs (http://www.hapmap.org).

Since SNP data are abundant and available at high spatial resolution across the human genome (The International HapMap Consortium 2005), Conrad et al. (2006) and McCarroll et al. (2006) reasoned that these SNP data might be used to discover underlying CNVs, if the underlying CNVs affected the results of SNP genotyping assays. McCarroll et al. (2006) hypothesized that deletion variants could leave at least three kinds of "footprints" in SNP data: (1) the identification of a run of null genotypes in a given individual, (2) the identification of contiguous genomic regions with SNP allele frequencies that deviated from expected Hardy-Weinberg equilibrium ratios, and (3) the recognition of runs of SNP genotyping results that did not fit expected Mendelian inheritance patterns in parent–offspring trios. In total, McCarroll et al. (2006) detected 541 deletion variants, ranging in size from 1 kb to 745 kb. Of the 541 deletions detected, 120 were observed as homozygous deletions (i.e., both copies of the genomic region were absent) in multiple, unrelated individuals. Ten of these homozygous deletions were relatively common and removed one or more exons of genes often involved in activities such as steroid metabolism, olfaction, and drug metabolism.

Conrad et al. (2006) focused exclusively on Mendelian inheritance inconsistencies. Numerous deletion variants (586) were identified, ranging from 300 bp to 1.2 Mb in size. Conrad et al. (2006) reported that the deletion CNV regions identified were relatively gene-poor, implying that many gene-containing deletions were subject to purifying selection. Despite this genome-wide trend, many individual genes are nonetheless affected by deletions. They found 92 genes that were completely deleted and another 109 genes that had portions of their coding sequences deleted (though the majority of these deletions were observed in only one trio and therefore may represent rare variants).

Of the 326 deletions that McCarroll et al. (2006) identified only from Mendelian inconsistencies, the overlap with the Conrad et al. (2006) data set was only 61.7% (201/326) (Fig. 1), which reflects, in part, the fact that Conrad et al. (2006) and McCarroll et al. (2006) used different criteria for defining Mendelian inconsistency (i.e., runs of genotypes that included at least two Mendelian inconsistencies and no heterozygous genotypes in single parent–offspring pairs versus runs of SNPs that showed similar patterns of Mendelian inconsistency across an entire HapMap population sample, respectively). Part of the incomplete overlap of these data sets may also be attributed to the estimated 15% false-positive rate of the two studies, based on confirmation studies on ~100 loci using independent experimental approaches (Conrad et al. 2006; McCarroll et al. 2006).

In the third study specifically identifying deletion variants, Hinds et al. (2006) hybridized DNA samples, from 24 unrelated individuals in a polymorphism discovery resource, to a high-density oligonucleotide array. This resulted in the identification of 215 potential deletion variants ranging from 70 bp to 10 kb. A subset of 100 PCR-confirmed deletions was further characterized, with 41 of the deletions found to be present among the 24 individuals with an allelic frequency of ≥10%. Forty-three deletions overlapped transcripts, and two deletions spanned exons. The deletions were then typed in a sample of 71 individuals who had previously been genotyped for ~1.6 million genome-wide SNPs (Hinds et al. 2005), enabling comparison of the two data sets. The

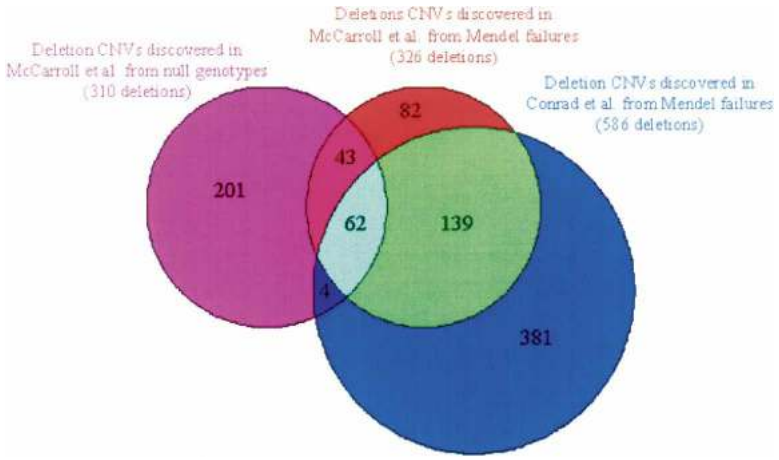

Figure 1. Comparison of overlapping CNVs identified by Conrad et al. (2006) and McCarroll et al. (2006). Conrad et al. (2006) identified a total of 586 deletions based on deviations from expected Mendelian inheritance patterns. McCarroll et al. (2006) identified deviations from Mendelian expectations in addition to null genotypes and deviations from Hardy-Weinberg equilibrium to identify a total of 541 deletions. When overlapping data were compared: (1) 139 deletions were identified only by Mendelian inheritance inconsistency in both studies, (2) 62 deletions were identified by Mendelian inheritance inconsistency and null genotypes by McCarroll et al. (2006) and by Mendelian inheritance inconsistency by Conrad et al. (2006), and (3) four deletions were detected only by null genotypes by McCarroll et al. (2006) but by Mendelian inheritance inconsistency by Conrad et al. (2006).

common deletions were found to be in linkage disequilibrium (LD: nonrandom pattern of alleles at different loci found together, more or less often than expected based on their frequencies) with surrounding SNPs, and the investigators therefore concluded that deletion variants and SNPs may often share similar evolutionary histories. This finding was similar to an observation made by McCarroll et al. (2006) in which many common deletion variants were in LD with nearby SNPs.

Clearly, every CNV discovery study has its own bias toward specific types and sizes of CNVs. For example, although the fine-scale approach of Hinds et al. (2006) was capable of detecting deletions of a wide variety of sizes, their analysis avoided repetitive regions (e.g., segmental duplications) that may be more likely to be associated with larger size CNVs (additional discussion below). Currently, the average size of all CNVs cataloged in the Database of Genomic Variants (http://projects.tcag.ca/variation) is ~118 kb, but the median size is ~18 kb. This discrepancy in mean and median CNV sizes may be due in part to the fact that more than half of the CNV entries now originate from the three recent deletion studies (Conrad et al. 2006; Hinds et al. 2006; McCarroll et al. 2006), which primarily report smaller CNVs; the majority being <10 kb (Eichler 2006). For CNVs detected by lower-resolution, BAC array-based methods, it is unclear what portion of the CNV-containing clone actually varies in copy number. With BAC array-based CGH methods, a BAC clone that shows copy number variation could entirely encompass a smaller CNV, overlap a CNV, or be totally within a CNV that is actually larger than the BAC clone itself. Because of this ambiguity, the size of the entire BAC clone is used in lieu of the actual size of the CNV.

One could speculate that larger CNVs (especially deletion variants) may be subject to increased selection pressures. Along with differences in mutation rates, this could affect the overall size distribution of human CNVs. Furthermore, the possibility that larger CNVs tend to represent multi-copy duplications is consistent with earlier observations that large segmental duplications are more likely to be tolerated by a genome than are deletions of similar sizes (i.e., >100 kb) (Lindsley et al. 1972; Brewer et al. 1999).

Thus, it appears from recent CNV studies that CNVs are a substantial source of genomic variation among humans. Currently, 1237 CNVs covering an estimated 143 Mb of genomic sequence have been identified (http://projects.tcag.ca/variation; http://paralogy.gs.washington.edu/structuralvariation; Nadeau and Lee 2006). Although it is difficult to compare such different data sets directly, the proportion of nucleotides that differ in copy number between two haploid genomes may be at least as large as the proportion that differs by SNPs. However, one must bear in mind that, for most studies, only a fraction of the putative CNVs have actually been validated by alternate methods or by their presence in multiple, unrelated individuals, and therefore the true number of CNVs in humans is likely to be less than the sum of the data currently being published.

Potential mechanisms of CNV formation

CNVs often occur in regions reported to contain, or be flanked by, large homologous repeats or segmental duplications (Fig. 2; Fredman et al. 2004; Iafrate et al. 2004; Sharp et al. 2005; Tuzun et al. 2005). Segmental duplications could arise by tandem repetition of a DNA segment followed by subsequent rearrangements that place the duplicated copies at different chromosomal loci. Alternatively, segmental duplications could arise via a duplicative transposition-like process: copying a genomic fragment while transposing it from one location to another (Eichler 2001).

CNVs that are associated with segmental duplications may be susceptible to structural chromosomal rearrangements via non-allelic homologous recombination (NAHR) mechanisms (Lupski 1998). NAHR is a process (Fig. 3) whereby segmental duplications on the same chromosome can facilitate copy number changes of the segmental duplicated regions along with intervening sequences (Inoue and Lupski 2002). In addition to the formation of CNVs in normal individuals, NAHR may also result in large structural polymorphisms and chromosomal rearrangements that directly lead to genomic instability or to early onset, highly penetrant disorders (Lupski 1998; Ji et al. 2000; Bailey et al. 2002, 2004; Stankiewicz and Lupski 2002; Scherer et al. 2003; Eichler et al. 2004; Shaw and Lupski 2004; Lupski and Stankiewicz 2005).

Not all CNVs, however, appear to be associated with segmental duplications. It is possible that subsets of CNVs, not associated with segmental duplications, may be formed or maintained by non-homology-based mutational mechanisms (Fig. 3; Shaw and Lupski 2004). Certain CNVs may be found to be associated with non-β DNA structures (DNA regions that differ in structure from the canonical right-handed β-helical duplex, including left-handed Z-DNA and cruciforms). Such DNA structures are believed to promote chromosomal rearrangements (Kurahashi and Emanuel 2001; Bacolla et al. 2004) and may also theoretically contribute to the genesis and maintenance of

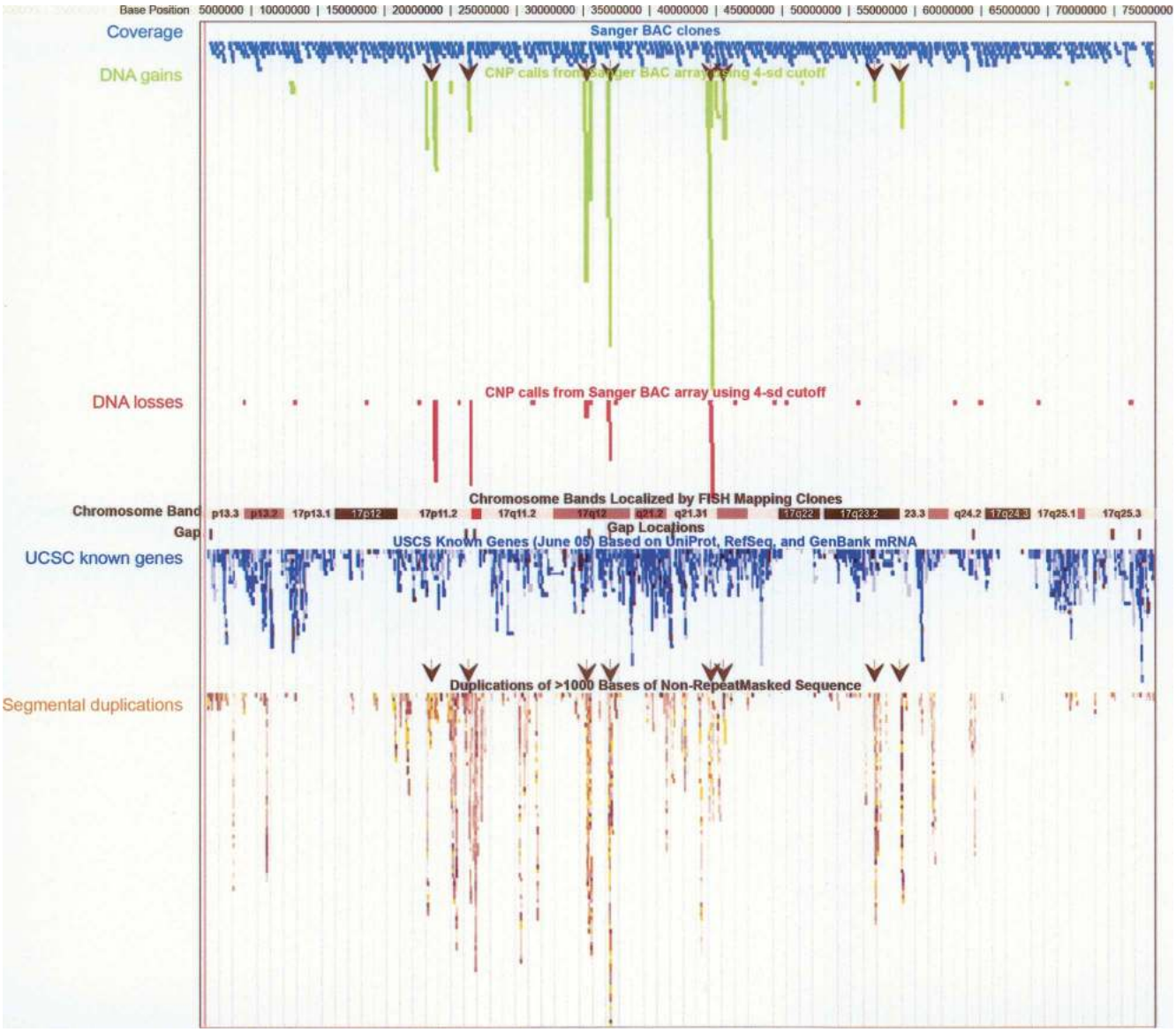

Figure 2. Copy number variation is associated with segmental duplications on chromosome 17. One hundred DNA samples from the HapMap collection were analyzed by CGH on a whole-genome tiling path microarray composed of 27,000 large-insert clones. The coverage of chromosome 17 by the array is displayed in blue (*top* panel). (Green bars) Frequencies of DNA gains, (red bars) frequencies of DNA losses. Gene density (blue) and presence of segmental duplications along chromosome 17 (orange) are reported in *bottom* panels. (Black arrows) Hotspots of DNA copy number variation along the chromosome, which all occur in regions containing or flanked by blocks of segmental duplications.

certain CNVs. Indeed, our understanding of the differential fragility of DNA sequences and mechanisms of non-homologous end-joining repair of double-strand breaks would be greatly improved by future large-scale sequencing efforts and definition of CNV breakpoints.

There may be a relationship between the size of a given CNV and its associated mutational mechanism(s). For example, data from at least two studies have shown that larger CNVs are more frequently associated with segmental duplications than are smaller CNVs (Fig. 4), although the effects of ascertainment biases remain unclear. In addition, there may be differential selection pressures exerted on deletion versus duplication events due to discrepancies in the way genomes tolerate gains and losses of genetic material. Nevertheless, it seems that among the smaller known CNVs, non-homology-driven mutational mechanisms may dominate.

Clinical implications and health

Large duplications and deletions have been known for some time to be related to the presentation of specific genetic disorders (Table 3), presumably as a result of copy number changes involving dosage-sensitive developmental genes. This has led to the establishment of genetic diagnostic tests for certain, well-characterized microdeletion and microduplication syndromes (e.g., Angelman syndrome, DiGeorge syndrome, Charcot-Marie-Tooth disease, etc.). If a de novo chromosomal aberration is recognized in a patient with a constitutional genetic abnormality

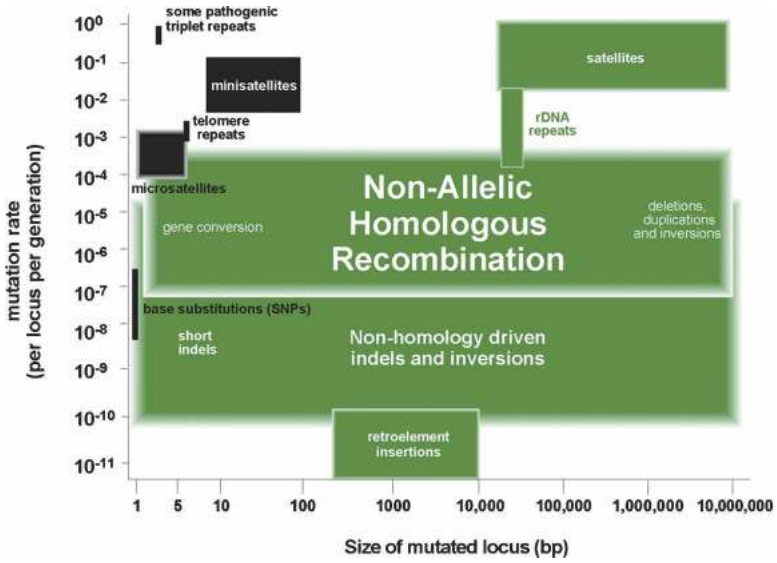

Figure 3. Different classes of mutation operating in the human genome. The range of mutation rates and size of mutated locus are plotted for each class of mutation. (Green highlights) Mutation processes associated with structural variation. On rare occasions, minisatellite alleles can differ in size by >1 kb.

(i.e., follow-up studies fail to reveal a similar chromosomal aberration in either of the two parents, and non-paternity has been excluded) and the aberration is not one of the dozen or so well known common chromosomal polymorphisms (e.g., inversion on chromosome 9; de la Chapelle et al. 1974; Lee 2005), the aberration is assumed to be the cause of the clinically recognized abnormal phenotype.

In many ways, the gold standard for clinical cytogenetic testing still remains the GTG-banded karyotype, where a genome-wide analysis usually identifies chromosomal rearrangements/aberrations of 3–5 Mb and larger. However, with the advent of higher resolution, genome-wide assays (e.g., array-based CGH), many more subtle genomic aberrations are being discovered in patients referred for genetic testing. Along with this improved resolution of testing comes the difficulty of interpreting the increasing number of genomic imbalances identified with each sample. To assist with accurate clinical diagnostic interpretations of genome-wide, high-resolution array CGH testing, the Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER, http://www.sanger.ac.uk/PostGenomics/decipher) has been established and is now comprehensively collecting array CGH data and corresponding clinical information from patients referred for genetic testing. The goal of this database is to help improve medical care while facilitating research on the genetic etiology of submicroscopic chromosomal imbalances.

Genomic imbalances that appear to be inherited from a phenotypically normal parent are usually considered to be clinically less significant (Shaw-Smith et al. 2004; de Vries et al. 2005; Tyson et al. 2005). Consider an example where an apparently healthy individual carries a certain copy number change along with other genetic variant(s) that compensate for that genomic imbalance. Another person having the same genomic imbalance may not have inherited the additional compensatory genetic variant(s), leading to a different and possibly clinical phenotype. Such scenarios underlie the growing lack of confidence for interpreting the clinical consequences of genomic imbalances. This is further exacerbated by the fact that genomic imbalances identified by array CGH represent cumulative and not allele-specific CNV values. Thus, true inheritance patterns of CNVs could be masked by array CGH results (Fig. 5). Clearly, accurate interpretations of CNV inheritance patterns will be greatly facilitated with the development of locus-specific and allele-specific quantitative assays for evaluating DNA copy number. Ultimately, clinical diagnostic interpretations should be based on a more holistic view of the genome whereby the phenotypic consequences of an imbalance incorporates the genotype and state of all alleles of a given CNV, neighboring DNAs, and other influencing genomic regions (e.g., enhancers, repressors, etc.). Until such comprehensive information is available for each patient, caution should continue to be

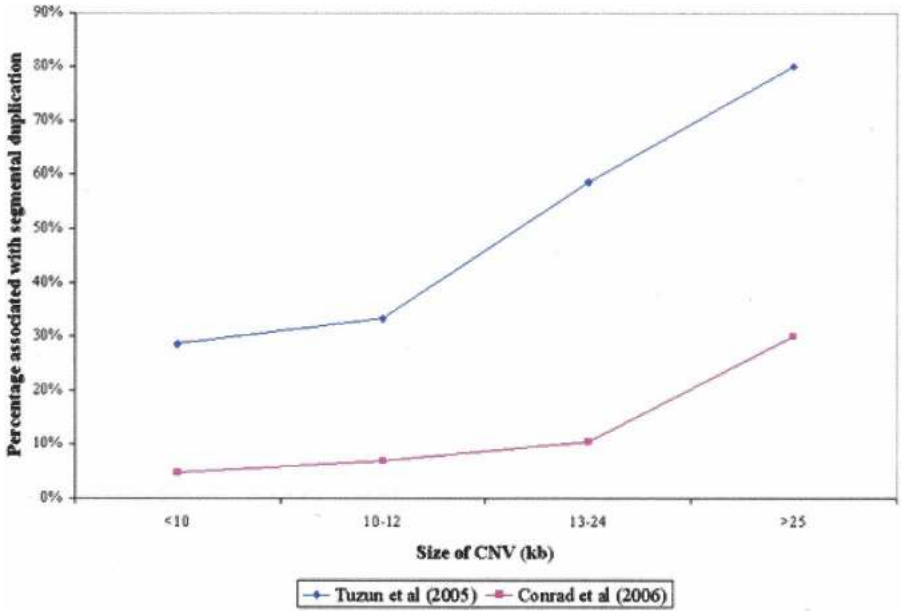

Figure 4. The positive correlation between size of CNVs and likelihood of association with segmental duplication. This correlation is noted by both the Conrad et al. (2006) and Tuzun et al. (2005) studies. The lower proportion of segmental duplication-associated CNVs in the Conrad et al. (2006) data relates to the greater difficulty in detecting CNVs in regions of segmental duplication when analyzing SNP genotyping data as opposed to fosmid end sequence mapping. The CNV size classes were chosen so as to obtain approximately equal numbers of CNVs in each class for the smaller data set.

Table 3. Examples of disorders caused by genomic imbalances and CNVs identified in regions associated with these disorders[a]

Chromosomal location	Disease phenotype associated with region	Reference(s)	Studies showing CNVs in vicinity of these loci	Known gene(s) in region
5p15	Cri du chat syndrome	Zhang et al. (2005)		
5q13.2	Spinal muscular atrophy (SMA)	Campbell et al. (1997)	Iafrate et al. (2004); Sebat et al. (2004); de Vries et al. (2005); Sharp et al. (2005)	*BIRC1, GTF2H2, SERF1A, SERF1B, SMN1, SMN2*
7q11.23	Williams-Beuren syndrome	Ewart et al. (1993); Osborne et al. (2001); Scherer et al. (2003)		
8q12	CHARGE syndrome	Vissers et al. (2005)		
11p15.4	Charcot-Marie-tooth disease type 4B2	Senderek et al. (2003)	Iafrate et al. (2004)	*ADM, SBF2*
15q11–13	Prader-Willi and Angelman syndrome	Ledbetter et al. (1982); Williams et al. (1989)	Iafrate et al. (2004); de Vries et al. (2005); Conrad et al. (2006); McCarroll et al. (2006)	*ATP10A, OCA2, OR4M2, OR4N4, UBE3A*
17p11.2	Smith-Magenis syndrome	Juyal et al. (1996); Lupski (1998)	Tuzun et al. (2005)	*ATPAF2, COPS3, DRG2, MED9, NT5M, RAI1, SMCR8, SREBF1*
17p12	Charcot-Marie-tooth disease type 1A	Lupski (1998)	de Vries et al. (2005); McCarroll et al. (2006); Sharp et al. (2005)	*COX10, HS3ST3A1, PMP22, TEKT3, ZNF286*
21q21	Alzheimer disease	Rovelet-Lecrux et al. (2006)		
22q11.2	DiGeorge/Velocardiofacial syndrome	Carlson et al. (1997); Edelmann et al. (1999)	Sharp et al. (2005); Conrad et al. (2006); McCarroll et al. (2006)	*GGT2, GNB1L, HIC2*
Xq22.2	Pelizaeus-Merzbacher disease	Woodward et al. (2005)		

[a]In some of these same regions, CNVs have also been identified among non-affected individuals.

exercised when trying to interpret the inheritance and clinical significance of copy number variants. Some possible mechanisms by which the same CNV could have differential effects on phenotypic traits and gene expression have been recently reviewed by Feuk et al. (2006b).

CNVs that do not directly result in early onset, highly penetrant genomic disorders may consequently be considered to be neutral in function, but afterward shown to play a role in later onset genomic disorders or common diseases. Analyses of the functional attributes of currently known CNVs reveal a remarkable enrichment for genes that are relevant to molecular–environmental interactions and influence our response to specific environmental stimuli (Sebat et al. 2004; Tuzun et al. 2005; Feuk et al. 2006a; Nguyen et al. 2006). These include, but are not limited to, processes involving drug detoxification (e.g., glutathione-*S*-transferase, cytochrome P450 genes, and carboxylesterase gene families), immune response and inflammation (e.g., leukocyte immunoglobulin-like receptor, defensin, and *APOBEC* gene families), surface integrity (e.g., late epidermal cornified envelope and mucin gene families), and surface antigens (e.g., galectin, melanoma antigen gene, and rhesus blood group gene families). Likewise, some CNVs encompass genes that may contribute to interindividual variation in drug responses (Ouahchi et al. 2006), as well as in immune defense and disease resistance/susceptibility among humans. For example, interindividual and interpopulation differences in the copy number of the gene encoding CCL3L1, a human immunodeficiency virus-1 (HIV-1)–suppressive chemokine and ligand for the HIV coreceptor CCR5, were recently reported (Gonzalez et al. 2005). Individuals with a lower-than-average number of CCL3L1 copies had lower levels of CCR5-CCL3L1 complexes, leaving more CCR5 available for HIV entry and hence increasing their susceptibility to HIV/AIDS. Most recently, Aitman et al. (2006) discovered copy number variation of the *Fcgr3* gene in rats, which predisposed those animals carrying fewer gene copies to develop a condition similar to glomerulonephritis in humans. *Fcgr3* encodes for a transmembrane receptor found on the cell surfaces of macrophages that when activated results in phagocytosis and cytotoxicity. The duplicated (paralogous) *Fcgr3*-rs gene appears to have an inhibitory effect on *Fcgr3* such that loss of *Fcgr3-rs* leads to an increased immune response and, in some cases, possibly autoimmunity. The orthologous gene in humans varies in copy number from 0 to 4, and association studies revealed that a lower copy number of the *Fcgr3* ortholog (*FCGR3B*) in humans is an independent risk factor predisposing those individuals to immunologically related glomerulonephritis.

One obvious way by which CNVs result in human phenotypic diversity is by altering transcriptional levels (and presumably subsequent translational levels) of the genes that are in variable copy number. Such a correlation has already been demonstrated for certain CNV genes at the transcriptional (Hollox et al. 2003; Aldred et al. 2005; McCarroll et al. 2006) and translational (Gonzalez et al. 2005; Linzmeier and Ganz 2005) levels. Studies correlating mRNA and protein levels with genomic copy number of CNV genes need to always consider that some CNVs may have phenotypic effects that are apparent only in certain tissues and/or stages of development. Experimental approaches may also be required to distinguish between the effects of CNVs themselves and any regulatory SNPs with which they may be in strong LD (e.g., Stranger et al. 2005).

CNVs in other species and evolution

Another aspect of CNVs that needs to be addressed is whether levels and patterns of copy number variation among humans are similar to those in non-human primates and other organisms. Wide-spread copy number variation has already been documented among inbred strains of laboratory mice (Li et al. 2004; Adams et al. 2005). Li et al. (2004) used a minimal-tiled, whole-genome array platform containing ~19,000 mouse BAC clones

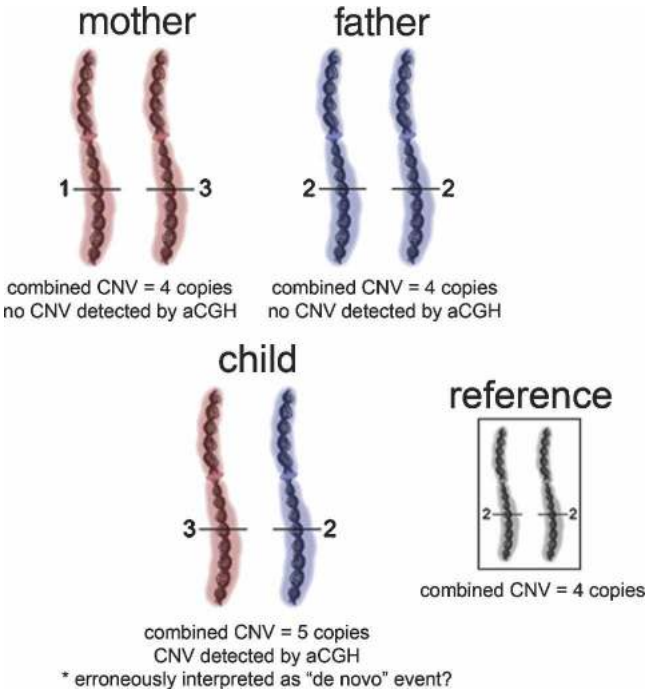

Figure 5. CNV inheritance patterns. To determine whether a CNV may be inherited or is a de novo event, trios including child, mother, and father are assessed. Currently, there is the potential for erroneous interpretation of the trio data since array-based CGH assays calculate copy number additively. For example, a mother who has a one-copy CNV on one chromosome and a three-copy CNV on the homologous chromosome (i.e., four total copies) would have no copy number difference when compared with a reference individual with four total copies (i.e., two copies on each homologous chromosome). In addition, a father who has two copies present on each homologous chromosome (i.e., four total copies) would also have no copy number difference when compared with the reference individual. If the child inherits the maternal chromosome containing three copies and the paternal chromosome containing two copies of the CNV, the child ends up with a total of five copies of the CNV. Upon comparison with the reference, the child could appear to have a de novo CNV. The development of locus-specific and allele-specific quantitative assays will aid in the interpretation of these CNV inheritance patterns.

from the RPCI-23 library (derived from a C57BL/6J strain mouse) to interrogate the genomes of 15 commonly used inbred mouse strains. In total, the investigators identified 346 BAC clones that showed copy number variation among the mouse strains tested when they used the C57BL/6J as a reference. Adams et al. (2005) used a 1-Mb mouse BAC-based array CGH platform to compare genomic DNA from a 129S5 mouse with that from a C57BL/6J mouse and identified a total of 112 CNVs (corresponding to 130 BAC clones of the 2803 clones on their array).

Li et al. (2004) found that ~10% of the CNV-containing BAC clones identified were within 200 kb of known segmental duplications, similar to that observed in humans. This again suggests that NAHR may play a role in the genesis and evolution of specific subsets of CNVs. Interestingly, large-scale deletions may be more tolerated in mice than in humans, especially when the deletion encompasses gene desert regions (Nobrega et al. 2004). Li et al. (2004) also found that unsupervised hierarchical cluster analysis of CNV patterns for each mouse strain led to stratification of the strains in a manner comparable to their known evolutionary history.

It is interesting to speculate on the phenotypic effects of these CNV patterns in different mouse strains. The mouse has long been recognized as a valuable model system for genetic research of human diseases, and sophisticated genetic manipulation studies can provide critical insights into the function of the corresponding genes in humans. Along with SNPs, CNVs may contribute to phenotypic variation among mouse strains and explain why different strains of mice sometimes produce apparently contradicting phenotypes when the same gene is knocked out/mutated. By carefully correlating functional variation with specific CNVs or sets of CNVs in the mouse, it may be possible to begin extrapolating the phenotypic consequences of orthologous CNVs in other organisms, including humans.

Insights into the evolutionary properties of CNVs can be obtained from cross-species comparative studies. Nguyen and colleagues (2006) compared the genes within known human and mouse CNVs and determined that human CNVs were often associated with genes that have relatively elevated ratios of non-synonymous (amino-acid-changing) to synonymous substitution rates. This may be interpreted as evidence for positive selection on CNVs during the evolutionary history of modern humans. Alternatively, this pattern may also include relaxation of selection or the presence of higher levels of purifying selection against CNVs in other gene types and families.

In an effort to understand the evolutionary history and significance of CNVs, chimpanzee (*Pan troglodytes*) CNV regions were recently identified and compared with human CNV regions (Perry et al. 2006). Using the same BAC array CGH platform that Iafrate et al. (2004) employed to identify >200 CNVs among 39 unrelated humans, 331 CNVs were identified among the genomes of 20 wild-born western chimpanzees. Interestingly, 74 of the chimpanzee CNVs occurred in the same regions as known human CNVs, and many of these CNVs were frequent in both species. These loci were also enriched (>20-fold, compared with all clones on the array) for segmental duplications that are shared by both species' genomes. From an evolutionary standpoint, this raises at least two issues. First, CNVs may be discovered in homologous regions of other closely related species, depending on when the ancestral segmental duplications in these regions arose. Second, if NAHR occurs regularly in these regions, the high intraspecific frequency of some CNVs may be the result of multiple recurrences within a species rather than a single ancestral duplication or deletion event followed by an increase in frequency.

Gene duplication is known to be an important long-term evolutionary force, and as suggested for different strains of mice, some lineage-specific copy number differences may contribute to the phenotypic differences among taxa, including those that distinguish humans from chimpanzees and bonobos (*Pan paniscus*) (Ohno 1970; Samonte and Eichler 2002; Locke et al. 2003; Shaw and Lupski 2004; Feuk et al. 2005; Newman et al. 2005; Goidts et al. 2006; Wilson et al. 2006). Two studies have presented data suggesting that the fixation rate of unique duplications and gene-containing duplications on the human lineage was elevated relative to that of the chimpanzee (Fortna et al. 2004; Cheng et al. 2005). Currently, it is unclear whether these results reflect experimental ascertainment biases, different duplication mutation rates, relaxed functional constraint, or human lineage-positive selection for duplications. Regardless, in their analyses, Cheng et al. (2005) and Newman et al. (2005) established an important correlation between differences in lineage-specific copy number and changes in gene expression, a relationship previously inferred in a study of gene expression differences between humans

and chimpanzees (Khaitovich et al. 2004). Detailed experimental efforts (including the generation of accurate finished sequences for some regions of the chimpanzee genome) will be necessary to link the fixation of any human lineage-specific CNVs to significant events in our evolutionary history. Similar studies would likely also be valuable in understanding the evolution of any organism.

Toward a global CNV map of the human genome

An important long-term goal for copy number variation research is to establish a comprehensive atlas of CNVs in the human genome. Such an effort would include correlation to phenotypes, mutational and evolutionary aspects, and behavior with other genomic factors (e.g., epigenetic control, linkage disequilibrium, etc.). Clearly, there are multiple methods for CNV discovery, with advantages and disadvantages for each technique. For example, the fosmid paired end sequence comparison strategy has proven to be an excellent means for CNV discovery (Tuzun et al. 2005), but is limited by the availability of DNA sequence data. The National Human Genome Research Institute of the NIH recently announced their intention to establish DNA libraries from 48 of the HapMap individuals (http://www.genome.gov/18016538) for the purposes of end sequencing as many as 1 million clones from each DNA library for fosmid paired end sequence comparisons. Such work should provide a catalog of structural variants in these representative individuals, including CNVs and balanced rearrangements (e.g., inversions), as well as lead to rapid demarcation of boundaries of specific copy number changes and genomic alterations. However, this strategy may be most ideal for identifying variants in the 8-kb to 40-kb range (since virtually no fosmids have inserts much larger than 40 kb), and it is unclear to what extent cloning artifacts and cloning biases lead to false positives and negative results.

Array-based comparative genomic experiments (e.g., array CGH) have also been shown to be valuable for discovering CNVs. Advantages of array-based CGH approaches include cost effectiveness and rapid screening of numerous individuals with a given platform, but clearly the resolution is limited by the size and the number of elements placed on the array. However, higher resolution arrays are now being assembled that could be used in such CNV discovery studies, including tiling arrays and higher density oligonucleotide arrays (Ishkanian et al. 2004; Dhami et al. 2005; Selzer et al. 2005; Urban et al. 2006). Typical array CGH assays are unable to provide some of the allele-specific CNV information that can be deduced from fosmid paired end sequence comparison strategies, but the array CGH assays do have the potential to identify a larger size range of CNVs. Finally, array CGH assays do not provide data on absolute copy number of a given CNV since the copy number of that CNV is unknown in the reference sample being used in the CGH assay. Hence, a copy number loss detected by array CGH may represent a deletion in the test material or a multi-copy duplication that is simply present in more copies in the reference sample being used.

The Copy Number Variation Project, an international consortium including founding researchers from The Wellcome

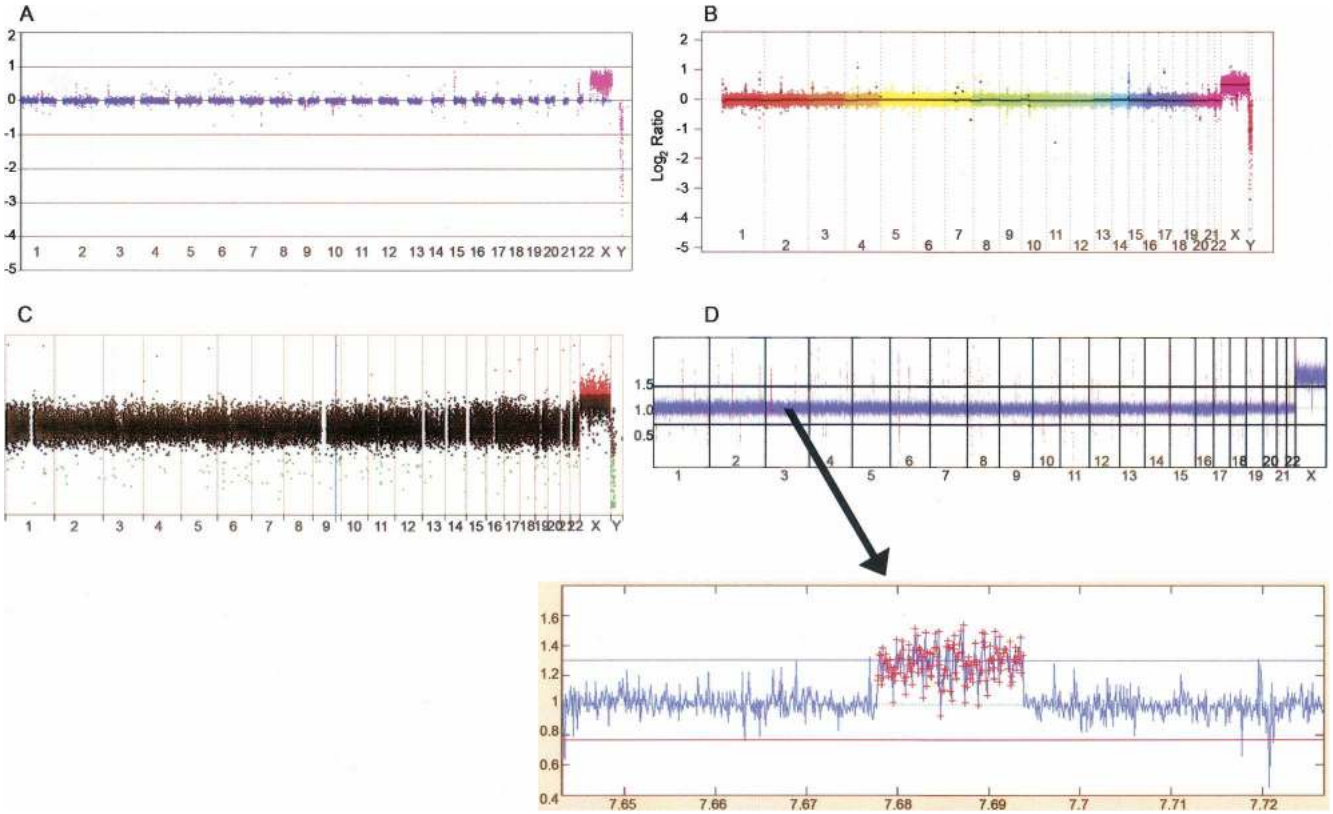

Figure 6. Cross-platform identification and validation of CNVs. *(A)* Array CGH, *(B)* Nimblegen array, *(C)* Agilent array, and *(D)* Affymetrix 500k SNP array platforms all identifying copy number variants in the GM 15510 individual from whom the G248 fosmid DNA library, used in the Tuzun et al. (2005) study, was created.

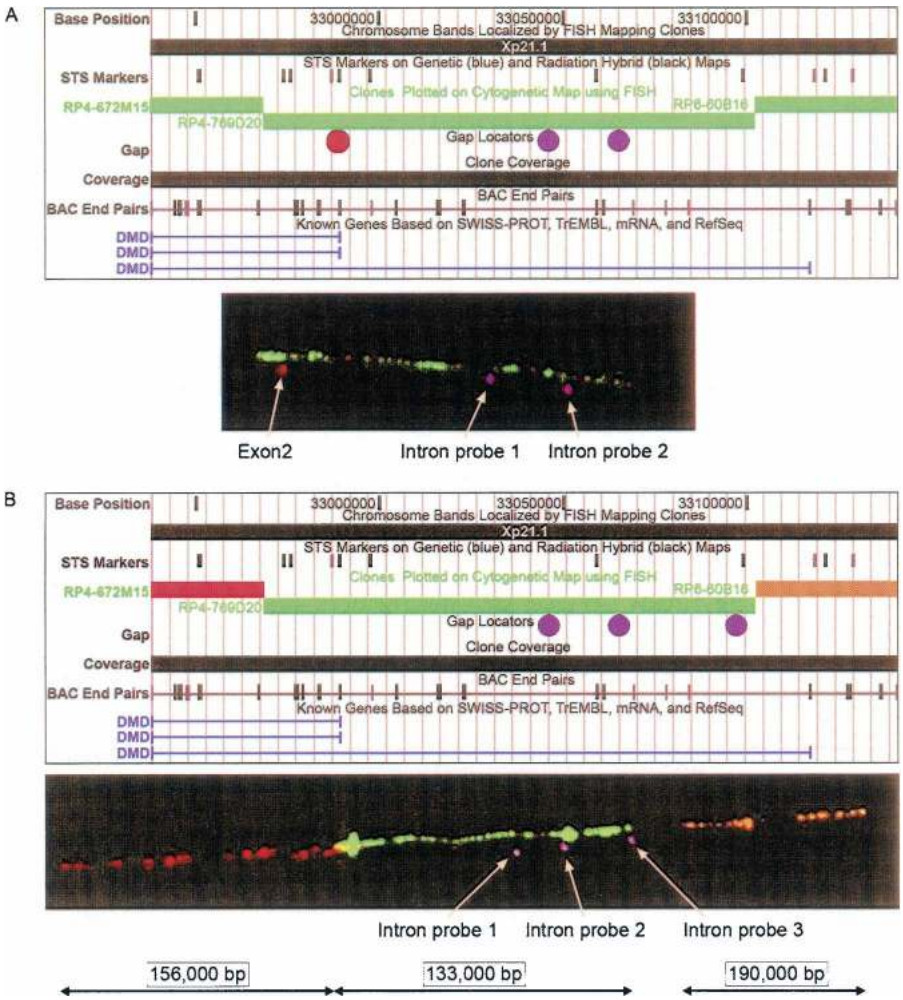

Figure 7. Fiber FISH image of the Dystrophin locus. Copy number variation has been identified at the Dystrophin locus in phenotypically normal humans (Iafrate et al. 2004; Conrad et al. 2006). Deletions at this locus have also been associated with Duchenne muscular dystrophy. Cytogenetic tools such as fiber FISH can be used to study the fine-scale structure of CNVs. *(A)* The genome structure from the UCSC genome browser showing the location of the 1-kb intron (two intron probes, purple dots), the exon (exon 2, red dot), and the three-color fiber FISH image (RP4–769D20, green). The Dystrophin locus CNV overlaps the 5′ end of Dystrophin, including exon 2 (red) and much of intron 1 (first purple dot). *(B)* The genome structure from the UCSC Genome Browser and the location of the 1-kb intron (three intron probes, purple dots) including non-polymorphic flanking BACs (RP4–672M15, red; RP6–60B16, orange) and a four-color fiber FISH image (RP4–769D20, green).

Trust Sanger Institute (Hinxton, United Kingdom), Hospital for Sick Children (Toronto), University of Tokyo (Tokyo), Affymetrix (Santa Clara, CA), and Harvard Medical School/Brigham and Women's Hospital (Boston, MA) aims to discover and characterize CNVs in human populations (http://www.sanger.ac.uk/humgen/cnv) using different technologies (Fig. 6). The initial goal of the consortium is to comprehensively identify CNVs in the 269 samples used for the International HapMap Project. By using the HapMap individuals as a resource for CNV studies, the resulting CNV data can be integrated with available SNP data to broaden our understanding of the genetic variation within an individual and eventually permit subsequent detailed association studies of genetic variation and human diseases.

As the pace of CNV discovery accelerates, we caution that there will be numerous false positives and false negatives, irrespective of the platform used, and a priority will be to minimize these. For example, many CNV discovery studies utilize material from established cell cultures. The use of cell cultures provides an ongoing resource, with the possibility for multiple replicate experiments, follow-up validation studies, and subsequent transcriptional and translational associations and functional assays. However, if CNVs are relatively unstable regions of the genome, it is possible that some small genomic imbalances will arise as a result of the cell culture transformation and propagation, and these genomic imbalances could be erroneously typed as endogenous CNVs. Hence, in CNV discovery studies, validation should be given high priority. Validation (with varying degrees of confidence) might include the observation of the same CNV among multiple individuals using one or more experimental methods (e.g., array CGH, ROMA, fosmid end sequencing, analyses of SNP data sets) or confirmation in the same individual with different technologies (e.g., quantitative PCR, direct sequencing, fluorescence in situ hybridization [FISH], and fiber FISH [Fig. 7]).

Since it now seems likely that CNVs are responsible for extensive differences in interindividual expression of immunological and environmental sensor genes, there is great interest in the possibility that CNVs play a role in the etiology of common diseases such as diabetes, cancer, and heart disease. Their potential relevance to common diseases and complex disorders deserves full investigation and may be accomplished by large-scale studies comprehensively comparing the CNV patterns between carefully phenotyped cohorts. However, while some CNVs may be in LD with flanking SNPs and could be effectively assayed by SNP genotyping (Hinds et al. 2006; McCarroll et al. 2006; Newman et al. 2006), other CNVs may have recurred multiple times independently (Conrad et al. 2006; Perry et al. 2006; Repping et al. 2006) and may not be as readily detectable through SNP-based association studies. Moreover, SNP and STR genotyping within CNV regions may be affected by variations in copy number of the SNP and STR sites themselves. For example, a multisite variant may not be genotyped correctly and almost certainly would not be scored such that the true underlying nature of this variant could be recovered (Fredman et al. 2004). This is worth consideration when moving toward fine-scale linkage and association studies, as unexpected fluctuations of significant scores may occur near or within the CNV region itself. SNP and STR markers in heterozygously deleted CNV regions may be scored as homozygous for the remaining allele, while SNP and STR markers at multicopy CNVs may be scored as homozygous for the most common SNP or STR allele. Indeed, some typing methods have

even made calls of SNPs in homozygously deleted regions. In each case, statistical power may be compromised in or near these regions during linkage and association analyses. In addition, CNV alterations at one or more multiple sites in the genome may themselves introduce genetic and phenotypic heterogeneity, adding additional levels of complexity in genetic disease studies. Direct and accurate genotyping of the CNVs themselves will help to resolve some of these issues, so assessment of suitable large-scale technologies to accomplish this should also be made a priority.

Conclusions

The recent discovery of widespread copy number variation in human and other mammalian genomes provides immediate insights into genetic variability among populations and provides a foundation for studies of the contribution of CNVs to evolution and disease. The published data are still largely rudimentary, but new developments in high-resolution scanning technologies will likely facilitate the establishment of comprehensive CNV maps. It is unlikely that any one technology alone will allow thorough identification of all classes of CNVs, so a priority of future work should focus on verifying primary results, integrating multiple data sources, and assigning population frequencies to these genomic variants.

Acknowledgments

We thank Don Conrad (University of Chicago) for additional data included in Figure 4, Shona Hislop for Figure 5, Shumpei Ishikawa (University of Tokyo) for Figure 6D, John Iafrate (Massachusetts General Hospital, Boston) for Figure 7, and Nancy Voynow for critical reading of this manuscript. Some of the work presented here, from the Copy Number Variation Project, has been supported by grants from Genome Canada/Ontario Genomics Institute and the Canadian Institutes of Health Research (S.W.S.), the Department of Pathology at Brigham and Women's Hospital and the Leukemia and Lymphoma Society (C.L.), and The Wellcome Trust (N.P.C., M.E.H., and C.T.-S.). S.W.S. is an International Scholar of the Howard Hughes Medical Institute.

References

Adams, D.J., Dermitzakis, E.T., Cox, T., Smith, J., Davies, R., Banerjee, R., Bonfield, J., Mullikin, J.C., Chung, Y.J., Rogers, J., et al. 2005. Complex haplotypes, copy number polymorphisms and coding variation in two recently divergent mouse strains. *Nat. Genet.* **37:** 532–536.

Aitman, T.J., Dong, R., Vyse, T.J., Norsworthy, P.J., Johnson, M.D., Smith, J., Manglon, J., Roberton-Lowe, C., Marshall, A.J., Petretto, E., et al. 2006. Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans. *Nature* **16:** 851–855.

Aldred, P.M., Hollox, E.J., and Armour, J.A. 2005. Copy number polymorphism and expression level variation of the human α-defensin genes *DEFA1* and *DEFA3*. *Hum. Mol. Genet.* **14:** 2045–2052.

Bacolla, A., Jaworski, A., Larson, J.E., Jakupciak, J.P., Chuzhanova, N., Abeysinghe, S.S., O'Connell, C.D., Cooper, D.N., and Wells, R.D. 2004. Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc. Natl. Acad. Sci.* **101:** 14162–14167.

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., and Eichler, E.E. 2002. Recent segmental duplications in the human genome. *Science* **297:** 1003–1007.

Bailey, J.A., Baertsch, R., Kent, W.J., Haussler, D., and Eichler, E.E. 2004. Hotspots of mammalian chromosomal evolution. *Genome Biol.* **5:** R23.

Barber, J.C.K., Joyce, C.A., Collinson, M.N., Nicholson, J.C., Willatt, L.R., Dyson, H.M., Bateman, M.S., Green, A.J., Yates, J.R.W., and Dennis, N.R. 1998. Duplication of 8p23.1: A cytogenetic anomaly with no established clinical significance. *J. Med. Genet.* **35:** 491–496.

Brewer, C., Holloway, S., Zawalnyski, P., Schinzel, A., and FitzPatrick, D. 1999. A chromosomal duplication map of malformations: Regions of suspected haplo- and triplolethality—and tolerance of segmental aneuploidy—in humans. *Am. J. Hum. Genet.* **64:** 1702–1708.

Buckland, P.R. 2003. Polymorphically duplicated genes: Their relevance to phenotypic variation in humans. *Ann. Med.* **35:** 308–315.

Buckley, P.G., Mantripragada, K.K., Piotrowski, A., de Stahl, T.D., and Dumanski, J.P. 2005. Copy-number polymorphisms: Mining the tip of an iceberg. *Trends Genet.* **21:** 315–317.

Campbell, L., Potter, A., Ignatius, J., Dubowitz, V., and Davies, K. 1997. Genomic variation and gene conversion in spinal muscular atrophy: Implications for disease process and clinical phenotype. *Am. J. Hum. Genet.* **61:** 40–50.

Carlson, C., Sirotkin, H., Pandita, R., Goldberg, R., McKie, J., Wadey, R., Patanjali, S.R., Weissman, S.M., Anyane-Yeboa, K., Warburton, D., et al. 1997. Molecular definition of 22q11 deletions in 151 velocardiofacial syndrome patients. *Am. J. Hum. Genet.* **61:** 620–629.

Carter, N.P. 2004. As normal as normal can be? *Nat. Genet.* **36:** 931–932.

Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., Church, D., DeJong, P., Wilson, R.K., Pääbo, S., et al. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437:** 88–93.

Cheung, V.G. 2004. Polymorphic landscape of the human genome. *Eur. J. Hum. Genet.* **13:** 133–135.

Coco, R. and Penchaszadeh, V.B. 1982. Cytogenetic findings in 200 children with mental retardation and multiple congenital anomalies of unknown cause. *Am. J. Med. Genet.* **12:** 155–173.

Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E., and Pritchard, J.K. 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38:** 75–81.

de la Chapelle, A., Schroder, J., Stenstrand, K., Fellman, J., Herva, R., Saarni, M., Anttolainen, I., Tallila, I., Tervila, L., Husa, L., et al. 1974. Pericentric inversions of human chromosomes 9 and 10. *Am. J. Hum. Genet.* **26:** 746–765.

de Vries, B.B., Pfundt, R., Leisink, M., Koolen, D.A., Vissers, L.E., Janssen, I.M., van Reijmersdal, S., Nillesen, W.M., Huys, E.H., de Leeuw, N., et al. 2005. Diagnostic genome profiling in mental retardation. *Am. J. Hum. Genet.* **77:** 606–616.

Dhami, P., Coffey, A.J., Abbs, S., Vermeesch, J.R., Dumanski, J.P., Woodward, K.J., Andrews, R.M., Langford, C., and Vetrie, D. 2005. Exon array CGH: Detection of copy-number changes at the resolution of individual exons in the human genome. *Am. J. Hum. Genet.* **76:** 750–762.

Edelmann, L., Pandita, R.K., Spiteri, E., Funke, B., Goldberg, R., Palanisamy, N., Chaganti, R.S.K., Magenis, E., Shprintzen, R.J., and Morrow, B.E. 1999. A common molecular basis for rearrangement disorders on chromosome 22q11. *Hum. Mol. Genet.* **8:** 1157–1167.

Edwards, J.H., Harnden, D.G., Cameron, A.H., Crosse, V.M., and Wolff, O.H. 1960. A new trisomic syndrome. *Lancet* **1:** 787–790.

Eichler, E.E. 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* **17:** 661–669.

———. 2006. Widening the spectrum of human genetic variation. *Nat. Genet.* **38:** 9–11.

Eichler, E.E., Clark, R.A., and She, X. 2004. An assessment of the sequence gaps: Unfinished business in a finished human genome. *Nat. Rev. Genet.* **5:** 345–354.

Engelen, J.J.M., Moog, U., Evers, J.L.H., Dassen, H., Albrechts, J.C.M., and Hamers, A.J.H. 2000. Duplication of chromosome region 8p23.1-p23.3: A benign variant? *Am. J. Med. Genet.* **91:** 18–21.

Ewart, A.K., Morris, C.A., Atkinson, D., Jin, W., Sternes, K., Spallone, P., Stock, A.D., Leppert, M., and Keating, M.T. 1993. Hemizygosity at the elastin locus in a developmental disorder, Williams syndrome. *Nat. Genet.* **5:** 11–16.

Feuk, L., MacDonald, J.R., Tang, T., Carson, A.R., Li, M., Rao, G., Khaja, R., and Scherer, S.W. 2005. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet.* **1:** 489–498.

Feuk, L., Carson, A.R., and Scherer, S.W. 2006a. Structural variation in the human genome. *Nat. Rev. Genet.* **7:** 85–97.

Feuk, L., Marshall, C.R., Wintle, R.F., and Scherer, S.W. 2006b. Structural variants: Changing the landscape of chromosomes and design of disease studies. *Hum. Mol. Genet.* **15:** R1–R10.

Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., Brenton, M., Hink, R., Burgers, S., Hernandez-Boussard, T., et al. 2004. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol.* **2:** E207.

Fredman, D., White, S.J., Potter, S., Eichler, E.E., Den Dunnen, J.T., and Brookes, A.J. 2004. Complex SNP-related sequence variation in

segmental genome duplications. *Nat. Genet.* **36:** 861–866.
Ghanem, N., Uring-Lambert, B., Abbai, M., Hauptmann, G., Lefranc, M.P., and Lefranc, G. 1988. Polymorphism of MHC class III genes: Definition of restriction fragment linkage groups and evidence for frequent deletions and duplications. *Hum. Genet.* **79:** 209–218.
Giglio, S., Broman, K.W., Matsumoto, N., Calvari, V., Gimelli, G., Neumann, T., Ohashi, H., Voullaire, L., Larizza, D., Giorda, R., et al. 2001. Olfactory receptor-gene clusters, genomic-inversion polymorphisms and common chromosome rearrangements. *Am. J. Hum. Genet.* **68:** 874–883.
Gilles, F., Goy, A., Remache, Y., Manova, K., and Zelenetz, A.D. 2000. Cloning and characterization of golgin-related gene from the large-scale polymorphism linked to PML gene. *Genomics* **70:** 364–374.
Goidts, V., Armengol, L., Schempp, W., Conroy, J., Nowak, N., Muller, S., Cooper, D.N., Estivill, X., Enard, W., Szamalek, J.M., et al. 2006. Identification of large-scale human-specific copy number differences by inter-species array comparative genomic hybridization. *Hum. Genet.* **119:** 185–198.
Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R.J., Freedman, B.J., Quinones, M.P., Bamshed, M.J., et al. 2005. The influence of *CCL3L1* gene—containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307:** 1434–1440.
Groot, P.C., Mager, W.H., and Frants, R.F. 1991. Interpretation of polymorphic DNA patterns in the human α-amylase multigene family. *Genomics* **10:** 779–785.
Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A., and Cox, D.R. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* **307:** 1072–1079.
Hinds, D.A., Kloek, A.P., Jen, M., Chen, X., and Frazer, K.A. 2006. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* **38:** 82–85.
Hollox, E.J., Armour, J.A.L., and Barber, J.C.K. 2003. Extensive normal copy number variation of a β-Defensin antimicrobial-gene cluster. *Am. J. Hum. Genet.* **73:** 591–600.
Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. 2004. Detection of large-scale variation in the human genome. *Nat. Genet.* **36:** 949–951.
Inoue, K. and Lupski, J.R. 2002. Molecular mechanisms for genomic disorders. *Annu. Rev. Genomics Hum. Genet.* **3:** 199–242.
The International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426:** 789–796.
———. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.
International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928–933.
Ishkanian, A.S., Malloff, C.A., Watson, S.K., deLeeuw, R.J., Chi, B., Coe, B.P., Snijders, A., Albertson, D.G., Pinkel, D., Marra, M.A., et al. 2004. A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet.* **36:** 299–303.
Jacobs, P.A., Baikie, A.G., Court Brown, W.M., and Strong, J.A. 1959. The somatic chromosomes in mongolism. *Lancet* **1:** 710.
Jacobs, P.A., Matsuura, J.S., Mayer, M., and Newlands, I.M. 1978. A cytogenetic survey of an institution for the mentally retarded: I. Chromosome abnormalities. *Clin. Genet.* **13:** 37–60.
Jacobs, P.A., Browne, C., Gregson, N., Joyce, C., and White, H. 1992. Estimates of the frequency of chromosome abnormalities detectable in unselected newborns using moderate levels of banding. *J. Med. Genet.* **29:** 103–108.
Ji, Y., Eichler, E.E., Schwartz, S., and Nicholls, R.D. 2000. Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res.* **10:** 597–610.
Juyal, R.C., Figuera, L.E., Hauge, X., Elsea, S.H., Lupski, J.R., Greenberg, F., Baldini, A., and Patel, P.I. 1996. Molecular analyses of 17p11.2 deletions in 62 Smith-Magenis syndrome patients. *Am. J. Hum. Genet.* **58:** 998–1007.
Khaitovich, P., Muetzel, B., She, X., Lachmann, M., Hellman, I., Dietzsch, J., Steigele, S., Do, H.H., Weiss, G., Enard, W., et al. 2004. Regional patterns of gene expression in human and chimpanzee brains. *Genome Res.* **14:** 1462–1473.
Kurahashi, H. and Emanuel, B.S. 2001. Long AT-rich palindromes and the constitutional t(11;22) breakpoint. *Hum. Mol. Genet.* **10:** 2605–2617.
Ledbetter, D.H., Mascarello, J.T., Riccardi, V.M., Harper, V.D., Airhart, S.D., and Strobel, R.J. 1982. Chromosome 15 abnormalities and the Prader-Willi syndrome: A follow-up report of 40 cases. *Am. J. Hum. Genet.* **34:** 278–285.
Lee, C. 2005. Vive la difference! *Nat. Genet.* **37:** 660–661.
Li, J., Jiang, T., Mao, J.H., Balmain, A., Peterson, L., Harris, C., Rao, P.H., Havlak, P., Gibbs, R., and Cai, W.W. 2004. Genomic segmental polymorphisms in inbred mouse strains. *Nat. Genet.* **36:** 952–954.
Lin, H., Pizer, E., and Morin, P.J. 2000. A frequent deletion polymorphism on chromosome 22q13 identified by representational difference analysis of ovarian cancer. *Genomics* **69:** 391–394.
Lindsley, D.L., Sandler, L., Baker, B.S., Carpenter, A.T.C., Denell, R.E., Hall, J.C., Jacobs, P.A., Gabor Miklos, G.L., Davis, B.K., Gethmann, R.C., et al. 1972. Segmental aneuploidy and the genetic gross structure of the *Drosophila* genome. *Genetics* **71:** 157–184.
Linzmeier, R.M. and 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:** 423–430.
Locke, D.P., Segraves, R., Carbone, L., Archidiacono, N., Albertson, D.G., Pinkel, D., and Eichler, E.E. 2003. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res.* **13:** 347–357.
Lucito, R., Healy, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., Rodgers, L., Brandy, A., Sebat, J., Troge, J., et al. 2003. Representational oligonucleotide microarray analysis: A high-resolution method to detect genome copy number variation. *Genome Res.* **13:** 2291–2305.
Lupski, J.R. 1998. Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14:** 417–422.
Lupski, J.R. and Stankiewicz, P. 2005. Genomic disorders: Molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet.* **1:** 627–633.
McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P., Zodi, M.C., Barrett, J., Dallaire, S., Gabriel, S.B., Lee, C., Daly, M.J., et al. 2006. Common deletion variants in the human genome. *Nat. Genet.* **38:** 86–92.
Nadeau, J.H. and Lee, C. 2006. Copies count. *Nature* **439:** 798–799.
Newman, T.L., Tuzun, E., Morrison, V.A., Hayden, K.E., Ventura, M., McGrath, S.D., Rocchi, M., and Eichler, E.E. 2005. A genome-wide survey of structural variation between human and chimpanzee. *Genome Res.* **15:** 1344–1356.
Newman, T.L., Rieder, M.J., Morrison, V.A., Sharp, A.J., Smith, J.D., Sprague, L.J., Kaul, R., Carlson, C.S., Olson, M.V., Nickerson, D.A., et al. 2006. High-throughput genotyping of intermediate-size structural variation. *Hum. Mol. Genet.* **15:** 1159–1167.
Nobrega, M.A., Zhu, Y., Plajzer-Frick, I., Afzal, V., and Rubin, E.M. 2004. Megabase deletions of gene deserts result in viable mice. *Nature* **431:** 988–993.
Nguyen, D.Q., Webber, C., and Ponting, C.P. 2006. Bias of selection on human copy number variants. *PLoS Genet.* **2:** 198–207.
Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, Berlin.
Osborne, L.R., Li, M., Pober, B., Chitayat, D., Bodurtha, J., Mandel, A., Costa, T., Grebe, T., Cox, S., Tsui, L.C., et al. 2001. A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nat. Genet.* **29:** 321–325.
Ouahchi, K., Lindeman, N., and Lee, C. 2006. Copy number variants and pharmacogenomics. *Pharmacogenomics* **7:** 25–29.
Patau, K., Smith, D.W., Therman, E., Inhorn, S.L., and Wagner, H.P. 1960. Multiple congenital anomaly caused by an extra autosome. *Lancet* **1:** 790–793.
Perry, G.H., Tchinda, J., McGrath, S.D., Zhang, J., Picker, S.R., Caceres, A.M., Iafrate, A.J., Tyler-Smith, C., Scherer, S.W., Eichler, E.E., et al. 2006. Hotspots for copy number variation in chimpanzees and humans. *Proc. Natl. Acad. Sci.* **103:** 8006–8011.
Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20:** 207–211.
Repping, S., van Daalen, S.K., Brown, L.G., Korver, C.M., Lange, J., Marszalek, J.D., Pyntikova, T., van der Veen, F., Skaletsky, H., Page, D.C., et al. 2006. High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nat. Genet.* **38:** 463–467.
Riley, B., Williamson, M., Collier, D., Wilkie, H., and Makoff, A. 2002. A 3-Mb map of a large segmental duplication overlapping the α7-nicotinic acetylcholine receptor gene (*CHRNA7*) at human 15q13-q14. *Genomics* **79:** 197–209.
Rovelet-Lecrux, A., Hannequin, D., Raux, G., Le Meur, N., Laquerriere, A., Vital, A., Dumanchin, C., Feuillette, S., Brice, A., Vercelletto, M., et al. 2006. *APP* locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat. Genet.* **38:** 24–26.
Samonte, R.V. and Eichler, E.E. 2002. Segmental duplications and the evolution of the primate genome. *Nat. Rev. Genet.* **3:** 65–72.

Scherer, S.W., Cheung, J., MacDonald, J.R., Osborne, L.R., Nakabayashi, K., Herbrick, J.A., Carson, A.R., Parker-Katiraee, L., Skaug, J., Khaja, R., et al. 2003. Human chromosome 7: DNA sequence and biology. *Science* **300:** 767–772.

Schoumans, J., Ruivenkamp, C., Holmberg, E., Kyllerman, M., Anderlid, B.M., and Nordenskjold, M. 2005. Detection of chromosomal imbalances in children with idiopathic mental retardation by array based comparative genomic hybridization (array CGH). *J. Med. Genet.* **42:** 699–705.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305:** 525–528.

Selzer, R.R., Richmond, T.A., Pofahl, N.J., Green, R.D., Eis, P.S., Nair, P., Brothman, A.R., and Stallings, R.L. 2005. Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosomes Cancer* **44:** 305–319.

Senderek, J., Bergmann, C., Weber, S., Ketelsen, U.P., Schorle, H., Rudnik-Schoneborn, S., Buttner, R., Buchheim, E., and Zerres, K. 2003. Mutation of the *SBF2* gene, encoding a novel member of the myotubularin family, in Charcot-Marie-Tooth neuropathy type 4B2/11p15. *Hum. Mol. Genet.* **12:** 349–356.

Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R., et al. 2005. Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77:** 78–88.

Shaw, C.J. and Lupski, J.R. 2004. Implications of human genome architecture for rearrangement-based disorders: The genomic basis of disease. *Hum. Mol. Genet.* **13:** R57–R64.

Shaw-Smith, C., Redon, R., Rickman, L., Rio, M., Willatt, L., Fiegler, H., Firth, H., Sanlaville, D., Winter, R., Colleaux, L., et al. 2004. Microarray based comparative genomic hybridization (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J. Med. Genet.* **41:** 241–248.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., Cremer, T., and Lichter, P. 1997. Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. *Genes Chromosomes Cancer* **20:** 399–407.

Stankiewicz, P. and Lupski, J.R. 2002. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18:** 74–82.

Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S.E., Tavare, S., et al. 2005. Genome-wide associations of gene expression variation in humans. *PLoS Genet.* **1:** 695–704.

Trask, B.J., Friedman, C., Martin-Gallardo, A., Rowen, L., Akinbami, C., Blankenship, J., Collins, C., Giorgi, D., Iandonato, S., Johnson, F., 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.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., et al. 2005. Fine-scale structural variation of the human genome. *Nat. Genet.* **37:** 727–732.

Tyson, C., Harvard, C., Locker, R., Friedman, J.M., Langlois, S., Lewis, M.E.S., Van Allen, M., Somerville, M., Arbour, L., Clarke, L., et al. 2005. Submicroscopic deletions and duplications in individuals with intellectual disability detected by array-CGH. *Am. J. Med. Genet.* **139A:** 173–185.

Urban, A.E., Korbel, J.O., Selzer, R., Richmond, T., Cubells, J.F., Hacker, A., Popescu, G.V., Green, R., Emanuel, B.S., Gerstein, M.B., et al. 2006. High resolution mapping of DNA copy alterations in human chromosome 22 using high density tiling oligonucleotide arrays. *Proc. Natl. Acad. Sci.* **103:** 4534–4539.

Vissers, L.E., Veltman, J.A., van Kessel, A.G., and Brunner, H.G. 2005. Identification of disease genes by whole genome CGH arrays. *Hum. Mol. Genet.* **14:** R215–R223.

Williams, C.A., Gray, B.A., Hendrickson, J.E., Stone, J.W., and Cantu, E.S. 1989. Incidence of 15q deletions in the Angelman syndrome: A survey of twelve affected persons. *Am. J. Med. Genet.* **32:** 339–345.

Wilson, G.M., Flibotte, S., Missirlis, P.I., Marra, M.A., Jones, S., Thornton, K., Clark, A.G., and Holt, R.A. 2006. Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and gorilla. *Genome Res.* **16:** 173–181.

Woodward, K.J., Cundall, M., Sperle, K., Sistermans, E.A., Ross, M., Howell, G., Gribble, S.M., Burford, D.C., Carter, N.P., Hobson, D.L., et al. 2005. Heterogeneous duplications in patients with Pelizaeus-Merzbacher disease suggest a mechanism of coupled homologous and nonhomologous recombination. *Am. J. Hum. Genet.* **77:** 966–987.

Zhang, X., Snijders, A., Segraves, R., Zhang, X., Niebuhr, A., Albertson, D., Yang, H., Gray, J., Niebuhr, E., Bolund, L., et al. 2005. High-resolution mapping of genotype-phenotype relationships in Cri du Chat syndrome using array comparative genomic hybridization. *Am. J. Hum. Genet.* **76:** 312–326.