nature genetics

The population genetics of structural variation

Donald F Conrad & Matthew E Hurles

Population genetics is central to our understanding of human variation, and by linking medical and evolutionary themes, it enables us to understand the origins and impacts of our genomic differences. Despite current limitations in our knowledge of the locations, sizes and mutational origins of structural variants, our characterization of their population genetics is developing apace, bringing new insights into recent human adaptation, genome biology and disease. We summarize recent dramatic advances, describe the diverse mutational origins of chromosomal rearrangements and argue that their complexity necessitates a re-evaluation of existing population genetic methods.

Although it has long been appreciated that the human genome contains a size continuum of genomic variation ranging from single-nucleotide changes to large (>3 Mb), microscopically visible karyotypic alterations, only recently has the abundance of structural variation between these two size extremes been appreciated. Structural variation has been defined as genomic alteration involving segments of DNA longer than 1 kb[1]. These segments can be deleted, duplicated, inserted, inverted in orientation or translocated.

Genomic variants of all sizes and types can contribute to genetic disease, and all are potential substrates for natural selection resulting in phenotypic differences between individuals, populations and species. Investigating the medical and evolutionary impact of structural variation requires that we understand the distribution of such variation within a species and the factors influencing that variation: in other words, the population genetics of structural variation.

The general factors influencing the distribution of variation within a species are common to all classes of variant and include mutation, selection, genetic drift, recombination, migration and population demography[2]. Although mutational mechanisms are sufficiently common across species such that structural variation is likely to be a general feature of all genomes (to a greater or lesser extent), here we confine ourselves to variation within the human genome.

Each genetic variant has its own specific evolutionary history, but it is through the analysis of many variants that the general properties of a class of variation can be elucidated. Population genetics is concerned with both variant-specific histories and the general properties of variation, both of which are pertinent to medical and evolutionary issues. Variants that seem to be population genetic outliers relative to a background of 'normal' variation are often enriched for medically relevant mutations. For example, the frequency at which an alpha-globin gene is deleted within a population varies between geographical locations to an unusually high degree. This is because the deletion confers both resistance to malarial infection and susceptibility to mild thalassemia: thus, it has increased in frequency in regions in which malaria is endemic[3] but remains at low frequency in the absence of malaria.

Over the past two decades, SNPs[4,5], microsatellites[6] and minisatellites[7] have been characterized extensively in different human populations, and as a result of population genetic analyses, we have learned much about the recent origin, dispersals and demography of our species and the different mutational and recombinational[8] processes generating and shuffling such variation. Moreover, we can use this information to begin to identify variants conferring risk to common diseases[4]. It has also been possible to identify specific variants that have conferred a selective advantage to our ancestors[9].

Population-genetic studies of SNPs, microsatellites and minisatellites have been a two-step process owing to economic constraints. They involve a discovery phase in which variants are identified in a limited set of individuals, followed by a targeted genotyping phase in which these variants are genotyped in diverse populations. This strategy also holds true for structural variation and introduces significant complications and biases for population genetic analyses.

Most structural variants have been discovered only in the past two years, and as a result, the population genetics of structural variation is very much in its infancy. Having outlined the importance of a population genetic perspective above, we now explore what we presently know about the distribution of structural variation and look toward the interesting questions that we can begin to address.

Current state of structural variation studies

Structural variation encapsulates a heterogeneous mix of variants arising by different mutational mechanisms. This heterogeneity necessitates further subclassification. Structural variants are typically subdivided into those that result in a change in DNA dosage (copy number variants (CNVs)) and those that do not (inversions and balanced translocations). Moreover, loci with variable copy numbers have a direction of change, deletion or duplication and can be biallelic or multiallelic. Thus, biallelic deletion loci have a diploid copy number of 0, 1 or 2, representing the three possible genotypes, whereas biallelic duplications generally have a diploid copy number of 2, 3 or 4 (**Fig. 1**). Multiallelic CNVs can result from deletions and duplications at the same locus and frequently involve tandemly repeated arrays of duplicated sequences. In the case of the gene *FCGR3B*, multiallelic copy number variation results in a diploid copy

Donald F. Conrad is at the Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA, and Matthew E. Hurles is at the Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK.
e-mail: meh@sanger.ac.uk

Published online 27 June 2007; doi:10.1038/ng2042

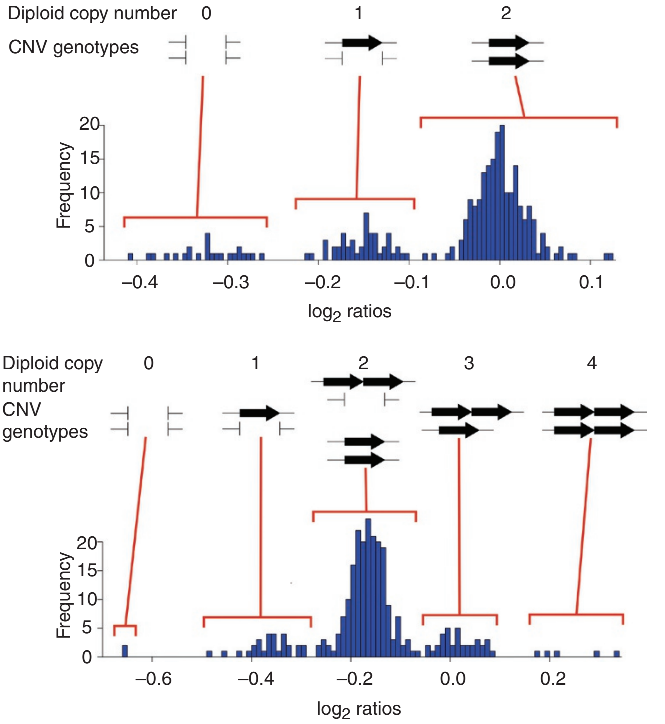

Figure 1 Diploid copy numbers, corresponding CNV genotypes and the underlying quantitative data from an array CGH experiment. Top, biallelic CNV; bottom, multiallelic CNV (data from ref. 11). Note that there is not a 1:1 mapping of diploid copy numbers to CNV genotypes for the multiallelic CNV.

number of 0, 1, 2, 3 or 4 (refs. 10,11), but for the Y-linked gene *TSPY*, the copy number in males ranges from 23–64 (ref. 12). The complexity of structural variation is further underlined by the existence at some loci of alleles that differ by multiple structural changes[13,14].

Demonstrating heritability is the *sine qua non* of all genetic studies, and it is only recently that the heritability of large numbers of structural variants has been demonstrated[11,15]. Observing mendelian inheritance of markers in pedigrees is the traditional method for assessing the heritability of genetic variants, but the frequent inability to attribute numbers of copies to each allele (a diploid copy number of 2 could represent either a 1/1 or 2/0 genotype; see **Fig. 1**) can create something of a problem; however, treating CNV data as quantitative traits (**Supplementary Fig. 1** online) allows the heritability of all types of CNV to be demonstrated[15].

Perhaps the most comprehensive catalog of known structural variation is the Database of Genomic Variants (DGV; http://projects.tcag.ca/variation/), which currently contains results from 37 publications, representing a bevy of experimental and analytical approaches to detecting structural variation. Combining information from different experiments in a meaningful way is challenging: choice of technique, genome assembly and reference sample(s) all frustrate meta-analysis of existing structural variation data. At the time of writing, there were 3,966 entries in the DGV (3,889 CNVs and 77 inversions or inversion breakpoints; see **Fig. 2**) at 2,191 loci, covering a staggering 405 Mb (14%) of the genome. The size distribution of CNV loci in the DGV ranges from 1 kb to 3.89 Mb, with a median of 103 kb. Almost certainly there are a nontrivial number of false positives in the DGV, and individual variants do not come with any measure of validity. Moreover, the sensitivity of the technology is such that when using large-insert clones as microarray probes, a CNV can be detected even if only a minority of the clone is copy number variable, and as a result, the size of a CNV can be overestimated.

Current technologies allow assessment of medium-to-large structural variation across almost all of the euchromatic human genome[16]. CNVs detected thus far are not randomly distributed across the genome but are preferentially clustered near centromeres and telomeres, regions known to be enriched with segmental duplications[11,17].

Thus far, a limited number of populations have been represented in genome-wide CNV studies. Although the populations sampled by the International HapMap Project[4] (European ancestry, Yoruba from Nigeria, Han Chinese, Japanese) are the most thoroughly characterized with respect to CNV[11,15,18,19], several studies have typed small samples from additional populations such as Native Americans and Pacific Islanders[17,20,21]. Although the HapMap samples seem to be representative of global SNP variation[5], there will be a benefit to sampling structural variation from a broader set of populations. Careful planning and description of population sampling will greatly improve the utility of future data sets of genome-wide structural variation.

Clearly, these are the early stages of structural genomic research (**Fig. 2**). Based on genome comparisons[22] and analysis of small indels[23,24] and large polymorphic deletions[18], it is evident that the length distribution of copy number variation is approximately exponential, with many small variants and few large ones. Small structural variants (1–10 kb) are the most underascertained, as they are difficult to discover with most existing platforms. Owing to the experimental difficulties of detecting balanced rearrangements, this class of variation is also largely unstudied. Cytogenetic work has estimated that a balanced translocation is formed in at least 1 of 2,000 concepti[25], and structural variation in subtelomeres is also known to be extensive[26]. Thus far, the most polymorphic inversions have been identified by comparison of pairs of genomes characterized in detail[27–29]. As the number of genomes screened for inversions increases, we should expect to see a rapid increase in the number of known inverted sequences.

Existing technologies used to survey genome-wide copy number variation have limited the ability to characterize the breakpoints of a CNV as resolution is sacrificed for coverage, and consequently, breakpoints for a given CNV typically can be mapped with a resolution of only 10–100 kb[11]. Without sequencing-level resolution, it is difficult to establish whether two alleles with indistinguishable structures stem from the same or different ancestral mutation events. Resolving this ambiguity facilitates the incorporation of structural variants into standard genetic analyses, which use the genotype as the core currency. Analysis methods for quantitative data (for example, array-based comparative genome hybridization (CGH)) typically identify CNVs as outliers against a background of invariant loci in the same genomes; however, the resultant set of CNV 'calls' cannot be considered as a reliable proxy for genotypes. At a minority of CNVs, the quantitative data can be used to cluster individuals into discrete classes that for biallelic CNVs correspond to the three possible genotypes (**Fig. 1**); however, for multiallelic CNVs, which constitute a sizeable fraction of large CNVs[11], it is not possibly to translate the diploid copy number into a genotype. The prospect of targeted assays for previously identified CNVs promises to dramatically increase the proportion of biallelic CNVs that can be genotyped unambiguously[30].

The ancestral state of a variant is of great importance in population genetics, as it establishes the direction of change and is usually assigned on the basis of comparisons to closely related species. For structural variation, this is complicated by the fact that many sites of structural variation in the human genome are also structurally variable in the chimpanzee genome[31]; however, if ancestral states could be determined for large numbers of structural variants (by analyzing their haplotypic background in humans[12,32] or by studying more outgroup species), subsequent population genetic analysis would be greatly facilitated.

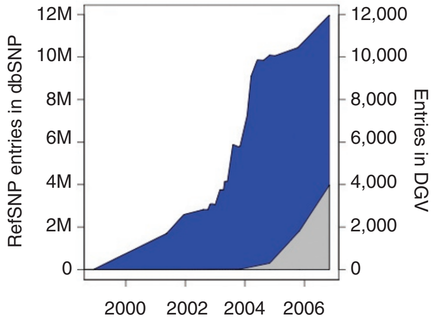

Figure 2 Cumulative number of RefSNP entries in dbSNP and cumulative number of variant loci in the Database of Genomic Variants, plotted as a function of time. Axes have been scaled differently to enhance visualization. RefSNP entries: left axis, blue (M = million). DGV entries: right axis, gray.

Biases in ascertained structural variants

Population geneticists recognize that the ascertainment scheme of variants in a discovery phase strongly influences the inferences drawn from data gathered during a subsequent targeted population-screening phase[33]. For example, to minimize the effort wasted on genotyping monomorphic markers, the ascertainment of SNPs included in Phase I of the HapMap was strongly biased toward SNPs observed more than once in a small discovery panel[4]. This strongly skews the site frequency spectrum toward common variants, and as a result, it biases estimates of linkage disequilibrium (LD) and many other population genetic statistics[34]. Similarly, any such two-phase study of structural variation will need to be corrected for ascertainment-induced biases, but documenting the ascertainment in detail facilitates these corrections.

Currently ascertained structural variants (CNVs in particular) have additional biases. First, only the largest variants have been discovered thus far; second, deletions are typically easier to detect than duplications; and third, there are biases in genomic location owing to incomplete genomic coverage in many surveys. Many of these biases differ markedly between different surveys for structural variation. Therefore, making general inferences about structural variation from current data is fraught with complications. In particular, the size of a CNV is highly correlated with many other features of the CNV, so the present skew toward larger variants could result in misleading inferences if they are considered representative of all CNVs. To give one example, longer CNVs are much more likely to be associated with segmental duplications than shorter CNVs[11,18,27], so the role of segmental duplications in generating all CNVs may be overestimated from the known CNVs.

The dependence on a reference genome assembly for data analysis (for example, fosmid paired-end analysis) or experimental design (for example, array CGH) also introduces biases. For instance, polymorphic sequences that are deleted in the reference genome assembly will not be detected by current array-based methodologies. Although the reference genome assembly is derived from many individuals, these contributions are not representative of global diversity. As approximately half the remaining gaps in the current genome assembly are associated with CNVs[11], the continued refinement of the genome assembly should yield improved understanding of structural variation.

Mutation dynamics of structural variation

The mutational processes that lead to structural variation are diverse and, perhaps unsurprisingly given the low-resolution mapping of most structural variation breakpoints, poorly characterized. Many studies investigating recurrent rearrangements that cause 'genomic disorders' have identified breakpoints embedded within highly similar duplicated sequences[35] (including both dispersed repetitive elements (for example, *Alu* sequences) and segmental duplications). This has led to an appreciation of the role of meiotic nonallelic homologous recombination (NAHR) in the genesis of many large rearrangements. NAHR between direct repeats causes deletions and duplications, NAHR between inverted repeats produces inversions and NAHR between repeats on different chromosomes leads to translocations. Moreover, these NAHR events can occur at rates of up to 10^{-4} per generation[36]; microsatellites and SNPs typically have mutation rates of ~10^{-3} and ~10^{-8} per generation, respectively. Segmental duplications are also enriched around CNVs[11] and inversions[27], thus implicating NAHR in the genesis of some structural variants.

NAHR is not the only mechanism generating structural variation; indeed, even for the largest CNVs, NAHR can account for only a minority of mutational events. Moreover, the smaller the CNV, the less likely that NAHR is involved[11,18,27]. Non–homology based mutation mechanisms must be responsible for the majority of structural variants. Nonhomologous end joining (NHEJ) is an alternative process by which DNA double-strand breaks are repaired, and it is likely that it has a substantial role in generating structural variation. In contrast to NAHR, NHEJ events are rarely recurrent, which suggests that they occur at a much lower rate at a given locus, probably $<10^{-7}$ per generation. It has been suggested that the propensity of a DNA sequence to adopt non-B conformations increases the likelihood of DNA double-strand breaks[37] and, hence, structural variation; however, only in the case of translocations involving palindromic AT-rich repeats on chromosome 22 has this fragility been demonstrated to result in a higher rate of rearrangement at a specific locus[38].

Comparative genomic analysis[39,40] and, to a lesser extent, diversity within species[41] have indicated that some duplications are transpositional in nature (in other words, the additional copy is inserted in a distant genomic location); this is especially prevalent in subtelomeric and pericentromeric regions of the genome[40]. The mechanism(s) of duplicative transposition are not well understood and deserve detailed characterization.

One special class of mutational mechanisms generating smaller structural variants is the random integration of cellular mRNA transcripts by the action of the LINE-1 reverse transcriptase in a process known as retrotransposition. Many of the resultant processed pseudogenes can be transcribed[42], so although this mechanism may not account for a high proportion of all structural variants, it is likely to have a disproportionately large functional impact.

These differences in the mutational mechanisms generating structural variation have important implications for population genetic models of

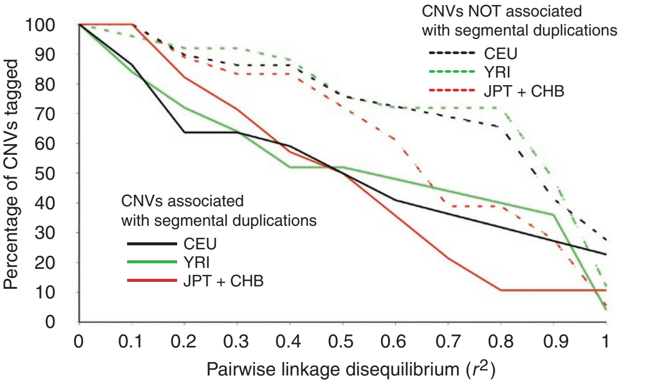

Figure 3 The maximal pairwise LD with a nearby SNP is lower around CNVs that are associated with segmental duplications than around CNVs in single-copy sequences.

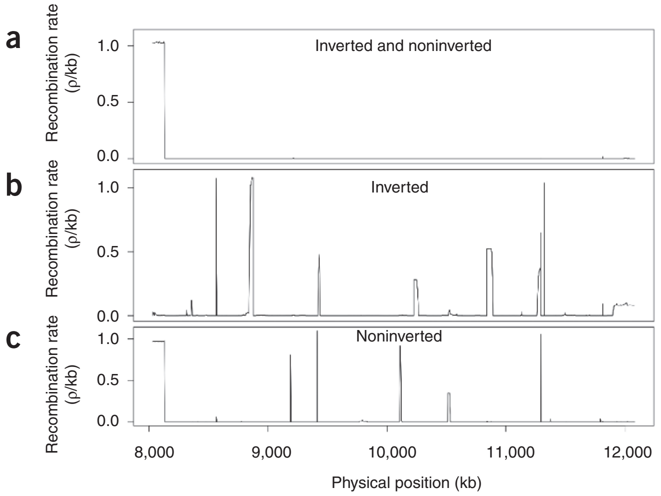

Figure 4 Estimates of fine-scale recombination rate across the 8p23 inversion. Haplotypes from **Supplementary Figure 2** were used to generate estimates of population genetic recombination rate using the method from ref. 50. Phylogenetic analysis of the SNP haplotypes uncovered two primary clades, and clade membership was used as a proxy for inversion status, with the minor allele assumed to be the inverted allele, arbitrarily. (**a**) Estimate of the population-scaled recombination rate ($\rho = 4N_er$) using 20 minor and 20 major (common) alleles. (**b**) Estimate of $4N_er$ using 40 haplotypes of the minor allele. (**c**) Estimate of $4N_er$ using 40 haplotypes of the major allele. As recombination is restricted in inversion heterozygotes, mutations that are private to either inversion background will be in extreme LD when considered at the population level, leading to low estimates of $4N_er$ in **a**. The analysis for **a** was run several times with different sets of minor and major haplotypes, and at most one 'hotspot' was detected visually.

structural variation. Mutational models for SNPs are not appropriate for microsatellites (and vice versa), and the same is true for different mechanisms generating structural variation. The 'infinite sites'[43] and 'infinite alleles'[44] models that are commonly used for modeling SNP variation may well be appropriate for structural variants generated by NHEJ, but the higher rate of NAHR, and the multiallelic nature of some resultant structural variants, suggests that a model closer in nature to the stepwise mutational models[45] used for microsatellites[46] would be more appropriate.

There is preliminary evidence that mutation rates for certain rearrangements differ markedly between apparently healthy individuals[38,47,48]. For example, at several loci, it has become apparent that carriers of heterozygous inversions are more susceptible to meiotic rearrangements involving sequences within the inverted interval[47]. A much greater understanding of the degree to which this is a general phenomenon[8] is needed to discern whether mutation rate polymorphism needs to be integrated into population genetic modeling of structural variation.

Explicit population genetic models are required for hypothesis testing and parameter estimation from samples of genetic variation, especially in relation to selection, demography and recombination. At present, most models do not account for structural variation and as a result are likely to give inaccurate estimates and unreliable inferences in some structurally variable regions of the genome (see below). It will be necessary to develop improved models and analyses to cope with the complexity of genetic variation in these genomic locations. Models that consider alleles with distinct structures differently are required.

Our understanding of the mutational mechanisms generating structural variation would be catalyzed by the high-throughput mapping of thousands of structural variation breakpoints at the nucleotide level, which at present is a laborious multistep process even for small numbers of variants.

LD around structural variants

LD is a term used to describe the nonrandom association between alleles at different loci. LD contains information about demographic history[49], recombination[50,51] and gene conversion[52]; it can be used to infer the action of natural selection[53,54] and is important for the design and analysis of genome-wide association studies[55].

The aim of association mapping is to assay directly or indirectly a large portion of genetic variation in a sample by genotyping a subset of well-characterized, easy-to-assay markers (typically SNPs). Estimation of the extent of LD between SNPs and structural variation is thus crucial and should inform the design of next-generation genome-wide association studies. Existing data suggest that the extent of LD between SNPs and CNVs is lower than LD among SNPs alone[11,15]. There are several reasons for this. First, the enrichment of CNVs around duplicated sequences places them in the most difficult regions to analyze using high-throughput SNP typing technology[56]. As LD decays with increasing distance, lower LD between CNVs and SNPs can result from the reduced density of genotypable SNPs in the vicinity of many CNVs. This results in CNVs associated with segmental duplications being less successfully tagged than CNVs in single-copy regions of the genome (**Fig. 3**). Second, as the mutation rate of some CNVs is higher than that of SNPs, low LD with SNPs could result from recurrent mutation generating allelic diversity. This lower LD has important implications for our prospects of understanding the phenotypic impact of structural variation, as existing indirect association methods will not fare well in the face of allelic diversity[57].

Population genetic analyses of genome-wide variation have shown that 80% of allelic recombination is confined to hotspots covering 10%–20% of the genome[8]. Structural variation is not integrated into the simple models of recombination from which these rate estimates are derived, and it can be expected to decrease the reliability of these estimates in some regions of the genome, especially those harboring common inversions (**Fig. 4** and **Supplementary Fig. 2** online). Experimental data describing local patterns of recombination around structural variants of all sizes are needed to address this issue and would also improve methods for detecting signals of natural selection acting on structural variants.

Population differentiation and population structure

The distribution of genetic variation across populations within a species is shaped by population demography and can be measured in different ways. The F_{ST} family of statistics[58,59] aims to quantify the proportion of variation within and between populations. Studies using diverse marker

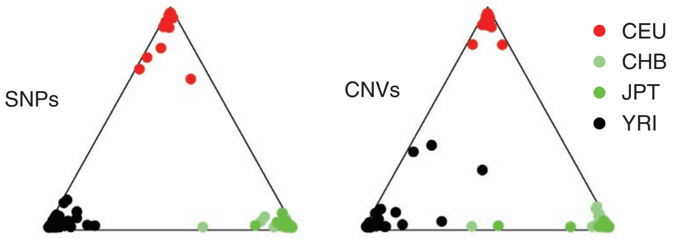

Figure 5 Plots of population structure for 67 CNVs and 67 unlinked SNPs in 210 unrelated HapMap individuals, assuming three ancestral populations. The slightly improved clustering quality from SNP genotypes most likely relates to the lower frequency of missing genotypes in the SNP data.

sets have shown conclusively that humans show little population differentiation relative to other comparable species; typically only 10%–15% of variance occurs between continental groups[60,61]. This feature of human diversity accords with the archaeological and paleontological evidence for a recent common origin in Africa some 50,000 years ago, which affords little time for extensive differentiation[2]. A survey of 67 common CNVs amenable to genotyping estimated that only 11% of the variation was attributable to differences between populations[11]; most of these variants are shared between populations and thus predate the migration out of Africa. Clearly, the distribution of structural variation between populations, like all other forms of genomic variation, is dominated by the recent common ancestry of humans.

The small proportion of variation that can be attributed to differences between populations contains signals of genetic relatedness. A simple method for analyzing these signals is to cluster individuals into an optimal number of populations without regard to their geographical origin[62]. The four HapMap populations can be clustered into three groups that reflect their continent of origin with high confidence using genotypes at only 67 common autosomal CNVs. The CNV-based clustering is qualitatively similar to that obtained for 67 common autosomal SNPs (**Fig. 5**) and is sufficient to assign correctly 209/210 individuals to their known continent of origin[11].

Selection can distort the population distribution of a given variant such that it is markedly more (or less) differentiated than the average. Thus, identifying unusual patterns of population differentiation should highlight structural variants that have been under recent selective pressures. Studies of individual disease-related loci have identified some notable structural variants with unusually high levels of population differentiation[3,63,64], and a recent genome-wide CNV survey replicated many of these findings and identified additional outliers that may have been under recent population-specific selection (see below for specific examples)[11]. Only a minority of CNVs can be genotyped with high confidence in existing data sets, yet measures of population differentiation such as F_{ST} rely on qualitative genotypes. Thus, to quantify the population differentiation of all forms of CNV, it has been necessary to adapt these traditional statistics to the underlying quantitative data[11].

Selection on structural variation

Existing CNV maps are not compatible with a model in which structural variation is distributed randomly across the genome. In addition to broad-scale mutational biases toward subtelomeric and pericentromeric regions, there is substantial evidence that CNVs are biased away from functional sequences of all classes[11,18,65]. The simplest explanation for such observations is that as a class, CNVs are slightly or moderately deleterious and that selection acts against (or 'purifies') changes in copy number of functional sequences. Further support for the action of purifying selection may be apparent in the site frequency spectrum of CNVs recorded in recent population surveys, which seems to be skewed toward rare variants[66]. Ascertainment of CNV by CGH is complicated by incomplete power and a nontrivial false positive rate, which makes formal analysis of the frequency spectrum extremely challenging.

Karyotypic analyses of individuals with segmental aneuploidy suggests that the genome is more tolerant of duplication than deletion[67]. This finding has been confirmed by recent higher-resolution techniques: deletions seem to be biased away from OMIM genes and RefSeq genes compared with duplications. When comparing large-scale (>10 kb) copy number variation ascertained with the same platform, duplications show a much larger median length (120 kb versus 43 kb) and higher average frequency than deletions do[11,15]. Balanced changes should, in principle, be less deleterious, although such rearrangements may disrupt genes directly (if the breakpoint occurs within a gene) or indirectly (through position effect), although informative data on this in humans are extremely scarce.

There are several notable large-scale differences in gene copy number between humans and chimpanzees, some of which may have become fixed as humans adapted to their changing environments[68,69]. There are many circumstantial claims of natural selection acting on existing structural variants, the majority of which involve genes mediating innate or acquired immunity. A number of deletions (including those in genes such as globins and *SLC4A1*) are found only in areas of the world where malaria is endemic. Extreme population differentiation has been noted for the *CCL3L1* polymorphism[11,63], which influences human susceptibility to HIV infection. Characterization of a recently discovered 1-Mb inversion at 17q21 has demonstrated unusual patterns of divergence between the two inversion alleles[64] and has been adduced as a signal of natural selection acting on the derived inversion haplotype.

Most studies of CNVs have detected an enrichment of genes involved in sensory perception, immune response and cell adhesion within polymorphic sequences[65]. This observation has been used to argue for the action of positive selection[65]. We must think carefully about invoking such forces. On one hand, LD-based signatures of recent positive selection are enriched within certain gene ontology classes[70]. However, genes overrepresented in CNV are also enriched within segmental duplications[71], which themselves show elevated structural dynamism. An enrichment of these classes within structural variants may reflect, in part, mutational biases and perhaps the genomic 'fossils' of past selective events that acted on gene copy number.

Balancing selection might also explain why some classes of genes are enriched for structural variation, but early genome-wide surveys have suggested that ancient balancing selection is rare within human populations[72,73]. In the long term, it seems that gene duplication is often a more stable evolutionary strategy than balancing selection for accommodating similar but differentiated gene functions, but the possibility that recent balancing selection is more common merits further investigation.

Each structural variant requires detailed characterization to fully resolve its evolutionary history. Having robust genotyping assays for specific rearrangements would facilitate this characterization. Qualitative assays that are targeted to the breakpoints of structural variants have significant advantages over quantitative assays[27], including the possibility that they can be applied to balanced rearrangements such as inversions[74]. With such genotyping assays in hand, it should be possible to estimate the age of structural variants (as has been possible for other allelic variants[75]) and integrate them into their surrounding haplotypes, which will provide much informative data on patterns of selection[54]. It is worth noting that existing haplotype-based tests[9,54,70] for selection often assume that a variant does not perturb neighboring sites of variation, so these methods often need to be adapted to take into account the size of a structural variant[11].

Future directions

We are currently observing the birth of a new subdiscipline in the population genetics of structural variation. Although the same questions can be asked about all types of variation, the theoretical and experimental tools required for investigating structural variation will inevitably require some adaptation. We emphasize that the mutational complexity of structural variation precludes a one-size-fits-all approach to modeling structural variation.

Population genetics has the power to provide insights into the demographic history of populations, selective pressures acting on genetic variation and mutational processes generating diversity. We see the future of the population genetics of structural variation as making substantial contributions to the latter two areas. By virtue of their number and simplicity, SNPs and microsatellites will remain the markers of choice for illuminating population demographic histories.

Clearly, our current knowledge of the locations, frequencies and types of structural variation in the human genome is rudimentary, but we anticipate rapid growth in the discovery of novel variants, especially smaller ones, over the next few years. Integrating structural variation detection within new sequencing technologies will be critical to take advantage of the coming era of human genome–wide resequencing.

We end by highlighting two important future challenges: (i) identifying structural variants that have facilitated recent human adaptation to novel environmental pressures and (ii) using our understanding of the population genetics of structural variation to identify structural variants influencing disease risk. These two challenges epitomize the benefits to evolutionary and medical genetics of an improved understanding of the population genetics of structural variation.

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

ACKNOWLEDGMENTS
The authors are grateful to G. Coop and C. Tyler-Smith for their comments on an earlier manuscript and to D. Andrews for data processing.

COMPETING INTERESTS STATEMENT
The authors declare no competing financial interests.

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

1. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
2. Jobling, M.A., Hurles, M.E. & Tyler-Smith, C. *Human Evolutionary Genetics: Origins, Peoples and Disease* (Garland Science, New York, 2004).
3. Flint, J. *et al.* High frequencies of alpha-thalassaemia are the result of natural selection by malaria. *Nature* **321**, 744–750 (1986).
4. IHMC. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
5. Conrad, D.F. *et al.* A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet.* **38**, 1251–1260 (2006).
6. Bowcock, A.M. *et al.* High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455–457 (1994).
7. Armour, J.A.L. *et al.* Minisatellite diversity supports a recent African origin for modern humans. *Nat. Genet.* **13**, 154–160 (1996).
8. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–324 (2005).
9. Sabeti, P.C. *et al.* Positive natural selection in the human lineage. *Science* **312**, 1614–1620 (2006).
10. Aitman, T.J. *et al.* Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851–855 (2006).
11. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
12. Repping, S. *et al.* High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nat. Genet.* **38**, 463–467 (2006).
13. Schmutz, J. *et al.* The DNA sequence and comparative analysis of human chromosome 5. *Nature* **431**, 268–274 (2004).
14. Fernandes, S. *et al.* A large AZFc deletion removes DAZ3/DAZ4 and nearby genes from men in Y haplogroup N. *Am. J. Hum. Genet.* **74**, 180–187 (2004).
15. Locke, D.P. *et al.* Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**, 275–290 (2006).
16. Fiegler, H. *et al.* Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Res.* **16**, 1566–1574 (2006).
17. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
18. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38**, 75–81 (2006).
19. McCarroll, S.A. *et al.* Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**, 86–92 (2006).
20. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
21. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
22. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
23. Mills, R.E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16**, 1182–1190 (2006).
24. Weber, J.L. *et al.* Human diallelic insertion/deletion polymorphisms. *Am. J. Hum. Genet.* **71**, 854–862 (2002).
25. Warburton, D. De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. *Am. J. Hum. Genet.* **49**, 995–1013 (1991).
26. Linardopoulou, E.V. *et al.* Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* **437**, 94–100 (2005).
27. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
28. Feuk, L. *et al.* Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet* **1**, e56 (2005).
29. Khaja, R. *et al.* Genome assembly comparison identifies structural variants in the human genome. *Nat. Genet.* **38**, 1413–1418 (2006).
30. Newman, T.L. *et al.* High-throughput genotyping of intermediate-size structural variation. *Hum. Mol. Genet.* **15**, 1159–1167 (2006).
31. Perry, G.H. *et al.* Hotspots for copy number variation in chimpanzees and humans. *Proc. Natl. Acad. Sci. USA* **103**, 8006–8011 (2006).
32. Jobling, M.A. *et al.* Recurrent duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males. *Hum. Mol. Genet.* **5**, 1767–1775 (1996).
33. Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H. & Nielsen, R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* **15**, 1496–1502 (2005).
34. Nielsen, R. & Signorovitch, J. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theor. Popul. Biol.* **63**, 245–255 (2003).
35. Stankiewicz, P. & Lupski, J.R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).
36. Shaffer, L.G. & Lupski, J.R. Molecular mechanisms for constitutional chromosomal rearrangements in humans. *Annu. Rev. Genet.* **34**, 297–329 (2000).
37. Bacolla, A. *et al.* Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc. Natl. Acad. Sci. USA* **101**, 14162–14167 (2004).
38. Kurahashi, H. & Emanuel, B.S. Unexpectedly high rate of de novo constitutional t(11;22) translocations in sperm from normal males. *Nat. Genet.* **29**, 139–140 (2001).
39. Johnson, M.E. *et al.* Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proc. Natl. Acad. Sci. USA* **103**, 17626–17631 (2006).
40. Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**, 88–93 (2005).
41. Wong, Z., Royle, N.J. & Jeffreys, A.J. A novel human DNA polymorphism resulting from transfer of DNA from chromosome 6 to chromosome 16. *Genomics* **7**, 222–234 (1990).
42. Balakirev, E.S. & Ayala, F.J. Pseudogenes: are they "junk" or functional DNA? *Annu. Rev. Genet.* **37**, 123–151 (2003).
43. Kimura, M. The rate of molecular evolution considered from the standpoint of population genetics. *Proc. Natl. Acad. Sci. USA* **63**, 1181–1188 (1969).
44. Kimura, M. & Crow, J.F. The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738 (1964).
45. Ohta, T. & Kimura, M. A model of mutation appropriate to estimate the number of electrophoretically detectable molecules in a finite population. *Genet. Res.* **22**, 201–204 (1973).
46. Valdes, A.M., Slatkin, M. & Freimer, N.B. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* **133**, 737–749 (1993).
47. Bayes, M., Magano, L.F., Rivera, N., Flores, R. & Perez Jurado, L.A. Mutational mechanisms of Williams-Beuren syndrome deletions. *Am. J. Hum. Genet.* **73**, 131–151 (2003).
48. Han, L.L., Keller, M.P., Navidi, W., Chance, P.F. & Arnheim, N. Unequal exchange at the Charcot-Marie-Tooth disease type 1A recombination hot-spot is not elevated above the genome average rate. *Hum. Mol. Genet.* **9**, 1881–1889 (2000).
49. Voight, B.F. *et al.* Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc. Natl. Acad. Sci. USA* **102**, 18508–18513 (2005).
50. McVean, G.A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
51. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
52. Andolfatto, P. & Nordborg, M. The effect of gene conversion on intralocus associations. *Genetics* **148**, 1397–1399 (1998).
53. Hudson, R.R., Bailey, K., Skarecky, D., Kwiatowski, J. & Ayala, F.J. Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. *Genetics* **136**, 1329–1340 (1994).
54. Sabeti, P.C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
55. Zondervan, K.T. & Cardon, L.R. The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* **5**, 89–100 (2004).
56. Fredman, D. *et al.* Complex SNP-related sequence variation in segmental genome duplications. *Nat. Genet.* **36**, 861–866 (2004).
57. Pritchard, J.K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
58. Wright, S. The genetical structure of populations. *Ann. Eugen.* **15**, 323–354 (1951).

59. Nei, M. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. USA* **70**, 3321–3323 (1973).
60. Watkins, W.S. *et al.* Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. *Genome Res.* **13**, 1607–1618 (2003).
61. Barbujani, G., Magagni, A., Minch, E. & Cavalli-Sforza, L.L. An apportionment of human DNA diversity. *Proc. Natl. Acad. Sci. USA* **94**, 4516–4519 (1997).
62. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
63. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440 (2005).
64. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
65. Nguyen, D.Q., Webber, C. & Ponting, C.P. Bias of selection on human copy-number variants. *PLoS Genet.* **2**, e20 (2006).
66. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. & Frazer, K.A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* **38**, 82–85 (2006).
67. Brewer, C., Holloway, S., Zawalnyski, P., Schinzel, A. & FitzPatrick, D. 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 (1999).
68. Johnson, M.E. *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519 (2001).
69. Popesco, M.C. *et al.* Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. *Science* **313**, 1304–1307 (2006).
70. Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
71. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
72. Przeworski, M., Hudson, R.R. & Di Rienzo, A. Adjusting the focus on human variation. *Trends Genet.* **16**, 296–302 (2000).
73. Bubb, K.L. *et al.* Scan of human genome reveals no new Loci under ancient balancing selection. *Genetics* **173**, 2165–2177 (2006).
74. Turner, D.J. *et al.* Assaying chromosomal inversions by single-molecule haplotyping. *Nat. Methods* **3**, 439–445 (2006).
75. Slatkin, M. & Rannala, B. Estimating allele age. *Annu. Rev. Genomics Hum. Genet.* **1**, 225–249 (2000).