

ARTICLES

Global variation in copy number in the human genome

Richard Redon¹, Shumpei Ishikawa^{2,3}, Karen R. Fitch⁴, Lars Feuk^{5,6}, George H. Perry⁷, T. Daniel Andrews¹, Heike Fiegler¹, Michael H. Shapero⁴, Andrew R. Carson^{5,6}, Wenwei Chen⁴, Eun Kyung Cho⁷, Stephanie Dallaire⁷, Jennifer L. Freeman⁷, Juan R. González⁸, Mònica Gratacòs⁸, Jing Huang⁴, Dimitrios Kalaitzopoulos¹, Daisuke Komura³, Jeffrey R. MacDonald⁵, Christian R. Marshall^{5,6}, Rui Mei⁴, Lyndal Montgomery¹, Kunihiro Nishimura², Kohji Okamura^{5,6}, Fan Shen⁴, Martin J. Somerville⁹, Joelle Tchinda⁷, Armand Valsesia¹, Cara Woodwark¹, Fengtang Yang¹, Junjun Zhang⁵, Tatiana Zerjal¹, Jane Zhang⁴, Lluís Armengol⁸, Donald F. Conrad¹⁰, Xavier Estivill^{8,11}, Chris Tyler-Smith¹, Nigel P. Carter¹, Hiroyuki Aburatani^{2,12}, Charles Lee^{7,13}, Keith W. Jones⁴, Stephen W. Scherer^{5,6} & Matthew E. Hurles¹

Copy number variation (CNV) of DNA sequences is functionally significant but has yet to be fully ascertained. We have constructed a first-generation CNV map of the human genome through the study of 270 individuals from four populations with ancestry in Europe, Africa or Asia (the HapMap collection). DNA from these individuals was screened for CNV using two complementary technologies: single-nucleotide polymorphism (SNP) genotyping arrays, and clone-based comparative genomic hybridization. A total of 1,447 copy number variable regions (CNVRs), which can encompass overlapping or adjacent gains or losses, covering 360 megabases (12% of the genome) were identified in these populations. These CNVRs contained hundreds of genes, disease loci, functional elements and segmental duplications. Notably, the CNVRs encompassed more nucleotide content per genome than SNPs, underscoring the importance of CNV in genetic diversity and evolution. The data obtained delineate linkage disequilibrium patterns for many CNVs, and reveal marked variation in copy number among populations. We also demonstrate the utility of this resource for genetic disease studies.

Genetic variation in the human genome takes many forms, ranging from large, microscopically visible chromosome anomalies to single-nucleotide changes. Recently, multiple studies have discovered an abundance of submicroscopic copy number variation of DNA segments ranging from kilobases (kb) to megabases (Mb) in size^{1–8}. Deletions, insertions, duplications and complex multi-site variants⁹, collectively termed copy number variations (CNVs) or copy number polymorphisms (CNPs), are found in all humans¹⁰ and other mammals examined¹¹. We defined a CNV as a DNA segment that is 1 kb or larger and present at variable copy number in comparison with a reference genome¹⁰. A CNV can be simple in structure, such as tandem duplication, or may involve complex gains or losses of homologous sequences at multiple sites in the genome (Supplementary Fig. 1).

An early association of CNV with a phenotype was described 70 yr ago, with the duplication of the *Bar* gene in *Drosophila melanogaster* being shown to cause the Bar eye phenotype¹². CNVs influence gene expression, phenotypic variation and adaptation by disrupting genes and altering gene dosage^{7,13–15}, and can cause disease, as in microdeletion or microduplication disorders^{16–18}, or confer risk to complex disease traits such as HIV-1 infection and glomerulonephritis^{19,20}. CNVs often represent an appreciable minority of causative alleles

at genes at which other types of mutation are strongly associated with specific diseases: CHARGE syndrome²¹ and Parkinson's and Alzheimer's disease^{22,23}. Furthermore, CNVs can influence gene expression indirectly through position effects, predispose to deleterious genetic changes, or provide substrates for chromosomal change in evolution^{10,11,17,24}.

In this study, we investigated genome-wide characteristics of CNV in four populations with different ancestry, and classified CNVs into different types according to their complexity and whether copies have been gained or lost (Supplementary Fig. 1). To maximize the utility of these data and the potential for integration of CNVs with SNPs for genetic studies, we performed experiments with the International HapMap DNA and cell-line collection²⁵ derived from apparently healthy individuals. The result is the first comprehensive map of copy number variation in the human genome, which provides an important resource for studies of genome structure and human disease.

Two platforms for assessing genome-wide CNV

The HapMap collection comprises four populations: 30 parent–offspring trios of the Yoruba from Nigeria (YRI), 30 parent–offspring trios of European descent from Utah, USA (CEU), 45 unrelated

¹The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. ²Genome Science, and ³Dependable and High Performance Computing, Research Center for Advanced Science and Technology, University of Tokyo, 4-6-1 Komaba Meguro, Tokyo 153-8904, Japan. ⁴Affymetrix, Inc., Santa Clara, California 95051, USA. ⁵The Centre for Applied Genomics and Program in Genetics and Genomic Biology, The Hospital for Sick Children, MaRS Centre–East Tower, 101 College Street, Room 14-701, Toronto, Ontario M5G 1L7, Canada. ⁶Department of Molecular and Medical Genetics, Faculty of Medicine, University of Toronto M5S 1A8, Canada. ⁷Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. ⁸Genes and Disease Program, Center for Genomic Regulation, Charles Darwin s/n, Barcelona Biomedical Research Park, 08003 Barcelona, Catalonia, Spain. ⁹Departments of Medical Genetics and Pediatrics, University of Alberta, Edmonton, Alberta T6G 2H7, Canada. ¹⁰Department of Human Genetics, University of Chicago, 920 East 58th Street, Chicago, Illinois 60637, USA. ¹¹Pompeu Fabra University, Charles Darwin s/n, and National Genotyping Centre (CeGen), Passeig Marítim 37-49, Barcelona Biomedical Research Park, 08003 Barcelona, Catalonia, Spain. ¹²Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan. ¹³Harvard Medical School, Boston, Massachusetts 02115, USA.

Japanese from Tokyo, Japan (JPT) and 45 unrelated Han Chinese from Beijing, China (CHB). Genomic DNA from Epstein–Barr-virus-transformed lymphoblastoid cell-lines was used.

Two technology platforms were used to assess CNV (Fig. 1): (1) comparative analysis of hybridization intensities on Affymetrix GeneChip Human Mapping 500K early access arrays (500K EA), in which 474,642 SNPs were analysed; and (2) comparative genomic hybridization with a Whole Genome TilePath (WGTP) array that comprises 26,574 large-insert clones representing 93.7% of the euchromatic portion of the human genome²⁶.

Stringent quality control criteria were set for each platform and experiments were repeated for 82 individuals on the WGTP and 15 individuals on the 500K EA platforms. The quality of the final data sets was assessed by the standard deviation among \log_2 ratios of autosomal probes (after normalization and filtering for cell-line artefacts), which for the WGTP platform was 0.047 (Supplementary Fig. 2) and for the 500K EA platform was 0.220, both of which are improvements on published data^{8,27}.

The different nature of the two data sets required the development of distinct algorithms to identify CNVs. In essence, these algorithms segment a continuous distribution of intensity ratios into discrete regions of CNV. To train the threshold parameters, we attempted to validate experimentally 203 CNVs that had been defined with varying degrees of confidence in two well-characterized genomes^{4,5,7} (NA10851 and NA15510). By performing technical replicate experiments on both platforms we assessed the proportion of CNV calls that were false positives for different algorithm parameters across a

set of experiments representing the spectrum of data quality. The threshold parameters for both algorithms were set to achieve an average false-positive rate per experiment beneath 5% (Methods; see also Supplementary Methods, Supplementary Tables 1–4 and refs 26, 28).

Because all DNAs were derived from lymphoblastoid cell lines, we differentiated somatic artefacts (such as culture-induced rearrangements and aneuploidies) from germline CNVs. We karyotyped all available 268 HapMap cell lines (Supplementary Table 5) and sought evidence for chromosomal abnormalities in the WGTP and 500K EA intensity data. We identified 30 cell lines with unusual chromosomal constitutions (Supplementary Table 5 and Supplementary Fig. 3), and removed the aberrant chromosomes from further analyses. Chromosomes 9, 12 and X seemed to be particularly prone to trisomy. For a cell line with mosaic trisomy of chromosome 12, we confirmed by array comparative genomic hybridization that this trisomy was not apparent in blood DNA from the same individual (Supplementary Fig. 4). Furthermore, we sought signals of somatic deletions within the SNP genotypes of HapMap trios. A somatic deletion in a parental genome manifests as a cluster of SNPs at which alleles present in the offspring are not found in either parent⁵. We assessed all of our preliminary CNV calls in 120 trio parents and found that 17 (of 4,758) fell in genomic regions that harbour highly significant clusters of HapMap Phase II SNP genotypes compatible with a somatic deletion in a parental genome (Supplementary Table 5A, Supplementary Fig. 5 and Supplementary Note). These putative cell-line artefacts were removed from further analyses. Extrapolating

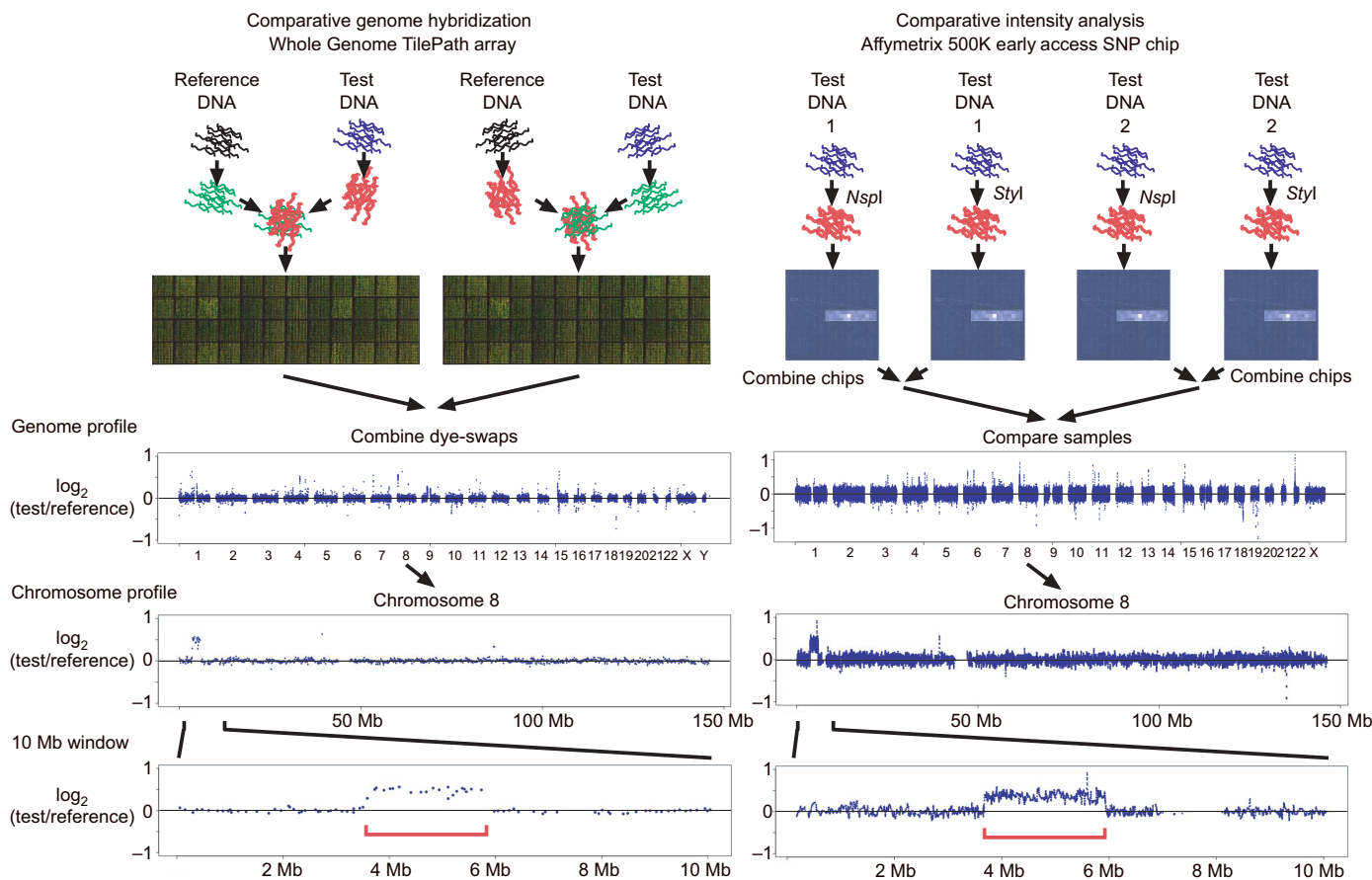


Figure 1 | Protocol outline for two CNV detection platforms. The experimental procedures for comparative genome hybridization on the WGTP array and comparative intensity analysis on the 500K EA platform are shown schematically (see Supplementary Methods for details), for a comparison of two male genomes (NA10851 and NA19007). The genome

profile shows the \log_2 ratio of copy number in these two genomes chromosome-by-chromosome. The 500K EA data are smoothed over a five-probe window. Below the genome profiles are expanded plots of chromosome 8, and a 10-Mb window containing a large duplication in NA19007 identified on both platforms (indicated by the red bracket).

this analysis to the entire HapMap collection suggests that less than 0.5% of the deletions we observed were likely to have been somatic artefacts.

The quality of resultant CNV calls was assessed in additional ways^{26,28}. Technical replicate experiments (triplicates for ten individuals) demonstrated that CNV calls are highly replicable (Supplementary Table 6), and that noisier experiments are characterized by higher false-negative rates, rather than higher false-positive rates (Supplementary Fig. 2). Heritability of CNVs within trios was investigated at 67 biallelic CNVs at which CNV genotypes could be inferred (Fig. 2; see also Supplementary Table 7). Of 12,060 biallelic CNV genotypes, only ~0.2% exhibited mendelian discordance, which probably reflects the genotyping error rate rather than the rate of *de novo* events at these loci. Additional locus-specific experimental validation was performed on subsets of CNVs (Supplementary Table 4). CNVs called in only a single individual (singleton CNVs) are more likely to be false positives compared with CNVs identified in several individuals. We attempted to validate 50 singleton CNVs called on only one platform (25 from each platform) and 14 singleton CNVs called on both platforms. All 14 singleton CNVs replicated by both platforms were verified as true positives, whereas 38 out of 50 of CNVs called by only one platform were similarly confirmed (false-positive rate of 24%). Extrapolating these validation rates across the entire data set suggests that only 8% (24% multiplied by the frequency of singleton CNVs called on only one platform) of the CNV regions we identify (see below) are likely to be false positives.

A genome-wide map of copy number variation

The average number of CNVs detected per experiment was 70 and 24 for the WGTP and 500K EA platforms, respectively (Supplementary Tables 8–10). Owing to the nature of the comparative analysis, each WGTP experiment detects CNVs in both test and reference genomes, whereas each 500K EA experiment detects CNV in a single genome. The median size of CNVs from the two platforms was 228 kb (WGTP) and 81 kb (500K EA), and the mean size was 341 kb and 206 kb, respectively. Consequently, the average length of the genome shown to be copy number variable in a single experiment is 24 Mb and 5 Mb on the WGTP and 500K EA platforms, respectively. The larger median size of the WGTP CNVs partially reflects inevitable overestimation of CNV boundaries on a platform comprising large-insert clones, as CNV encompassing only a fraction of a clone can be detected, but will be reported as if the whole clone was involved.

By merging overlapping CNVs identified in each individual, we delineated a minimal set of discrete copy number variable regions (CNVRs) among the 270 samples (Fig. 3; see also Supplementary Table 11). We identified 913 CNVRs on the WGTP platform and 980 CNVRs on the 500K EA platform and mapped their genomic distribution (Fig. 4). Approximately half of these CNVRs were called in more than one individual and 43% of all CNVs identified on one platform were replicated on the other. Combining the data resulted in a total of 1,447 discrete CNVRs, covering 12% (~360 Mb) of the human genome. Using locus-specific quantitative assays on a subset of regions we validated 173 (12%) of these CNVRs (Supplementary Tables 4 and 12). A minority (30%) of these 1,447 CNVRs overlapped

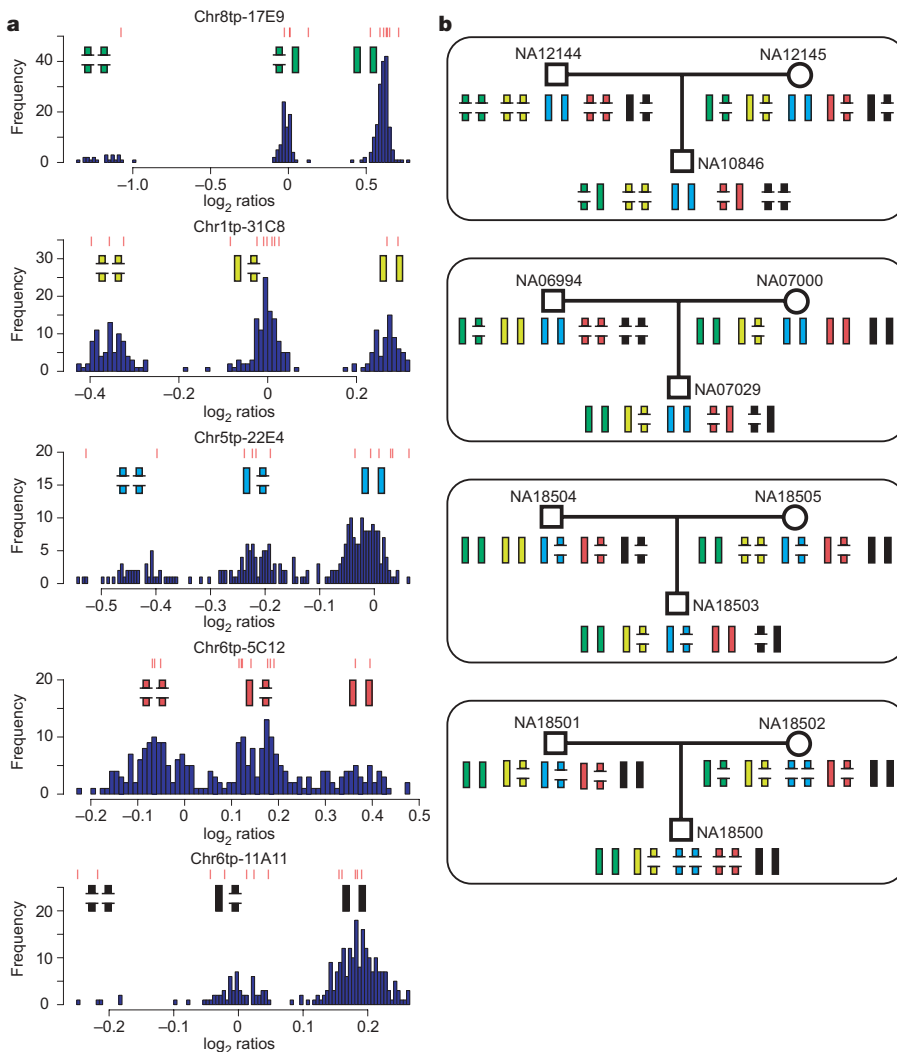


Figure 2 | Heritability of five CNVs in four HapMap trios. **a**, The distribution of WGTP \log_2 ratios at five CNVs with genotype information. Each histogram of \log_2 ratios in 270 HapMap individuals exhibits three clusters, each corresponding to a genotype of a biallelic CNV, with the two alleles depicted by broken and complete bars, representing lower and higher copy number alleles, respectively. Red lines above each histogram denote \log_2 ratios in the 12 individuals represented in **b**. **b**, Mendelian inheritance of five CNVs in four parent-offspring trios. The individual CNVs were genotyped from WGTP clones: green, Chr8tp-17E9; yellow, Chr1tp-31C8; blue, Chr5tp-22E4; red, Chr6tp-5C12; black, Chr6tp-11A11.

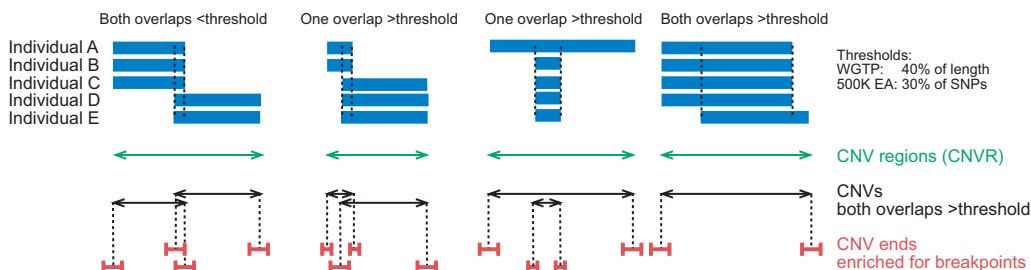


Figure 3 | Defining CNVRs, CNVs and CNV ends. Overlapping CNVs called in five individuals are shown schematically for four loci (in blue); dashed lines indicate overlap. Copy number variable regions (CNVRs) represent the union of overlapping CNVs (in green). Independent juxtaposed CNVs (in black) are identified by requiring that only individual-specific CNVs that overlap by more than a threshold proportion be merged. Intervals

encompassing CNV breakpoints (in red) are defined using platform-dependent criteria (Supplementary Methods), and contain a significant paucity of recombination hotspots^{76,77} (Supplementary Table 13), which results from the enrichment of segmental duplications within which fewer inferred recombination hotspots reside.

those identified in previous studies^{1-3,5-8,29}. Combining different classes of experimental replication revealed that 957 (66%) of the 1,447 CNVRs detected here have been replicated on both WGTP and 500K EA platforms, or with a locus-specific assay, or in another individual, or in a previous study (Supplementary Table 12). Whole-genome views of CNV show that although common, large-scale CNV is distributed in a heterogeneous manner throughout the genome (Supplementary Fig. 6), no large stretches of the genome are exempt from CNV (Fig. 4), and the proportion of any given chromosome susceptible to CNV varies from 6% to 19% (Supplementary Fig. 7).

and emphasizes the need for ongoing characterization of these genomic regions.

Gaps within the reference human genome assembly have an extremely high likelihood of being associated with CNVs; out of the 345 gaps in the build 35 assembly, 48% (164 out of 345) are flanked or overlapped by CNVRs. This finding highlights the complexity in generating a reference sequence in regions of structural dynamism

Comparing the CNVRs identified on the two platforms reveals that the WGTP and 500K EA platforms largely complement one another. The 500K EA platform is better at detecting smaller CNVs (Supplementary Fig. 8), whereas the WGTP platform has more power to detect CNVs in duplicated genomic regions (Supplementary Table 13) where 500K EA coverage is poorer³⁰.

Some CNVRs encompass two or more independent juxtaposed CNVs. For example, a small deletion found in one individual overlapping a much larger duplication in another individual was merged into a single CNVR, despite these representing distinct events. To delineate independent CNVs (CNV events) we applied more stringent merging criteria to separate juxtaposed CNVs (Fig. 3), and identified 1,116 and 1,203 CNVs on the WGTP and 500K EA

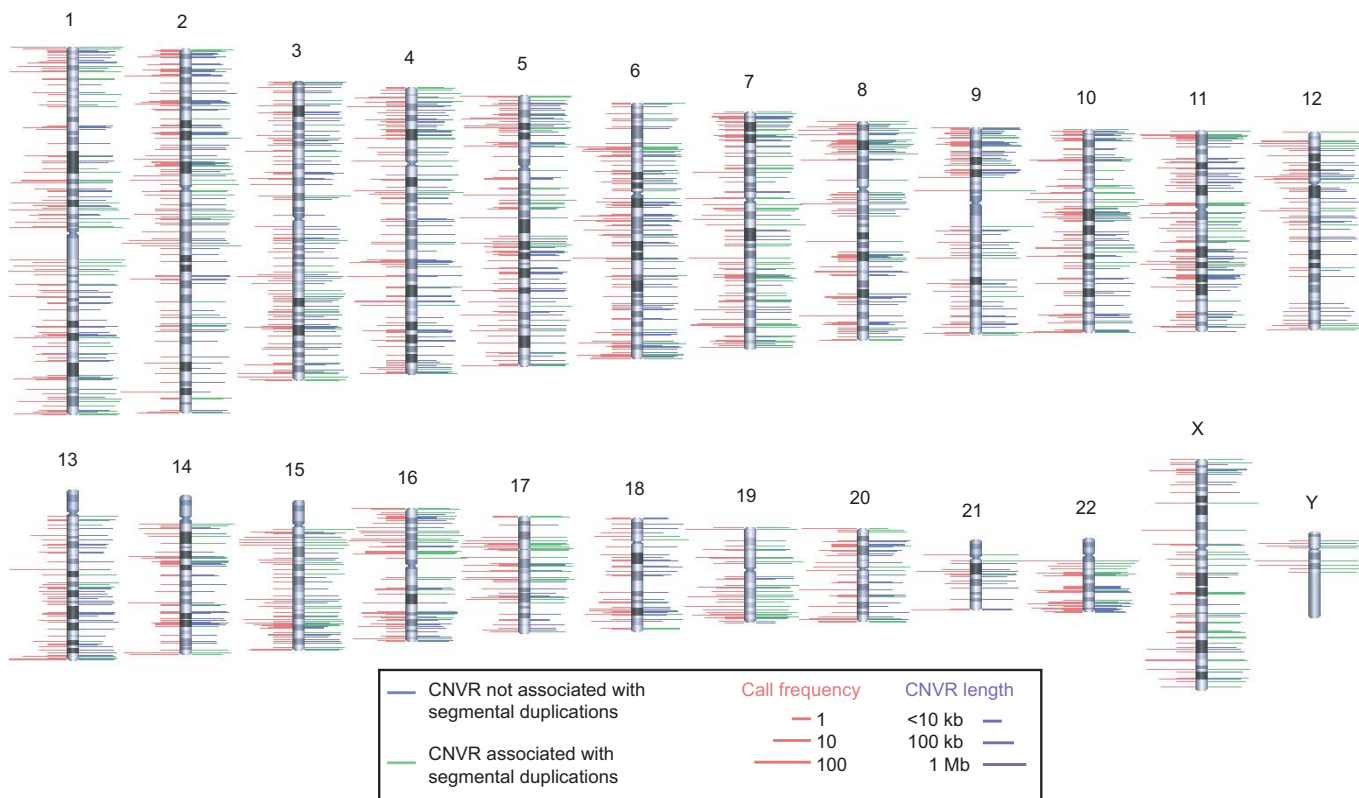


Figure 4 | Genomic distribution of CNVRs. The chromosomal locations of 1,447 CNVRs are indicated by lines to either side of ideograms. Green lines denote CNVRs associated with segmental duplications; blue lines denote CNVRs not associated with segmental duplications. The length of right-hand side lines represents the size of each CNVR. The length of left-hand side lines indicates the frequency that a CNVR is detected (minor call frequency

among 270 HapMap samples). When both platforms identify a CNVR, the maximum call frequency of the two is shown. For clarity, the dynamic range of length and frequency are log transformed (see scale bars). All data can be viewed at the Database of Genomic Variants (<http://projects.tcag.ca/variation/>).

platforms, respectively (Fig. 5; see also Supplementary Table 11). We classified these CNVs into five types: (1) deletions; (2) duplications; (3) deletions and duplications at the same locus; (4) multi-allelic loci; and (5) complex loci whose precise nature was difficult to discern. Owing to the inherently relative nature of these comparative data, it was impossible to determine unambiguously the ancestral state for most CNVs, and hence whether they are deletions or duplications. Here we adopted the convention of assuming that the minor allele is the derived allele³¹, thus deletions have a minor allele of lower copy number and duplications have a minor allele of higher copy number. Approximately equal numbers of deletions and duplications were identified on the WGTP platform, whereas deletions outnumbered duplications by approximately 2:1 on the 500K EA platform. In addition, 33 homozygous deletions (relative to the reference sequence) identified on the 500K EA platform were experimentally validated with locus-specific assays (Supplementary Table 14). Most (27 out of 33) of these have not been observed in a previous genome-wide survey of deletions⁷.

To investigate mechanisms of CNV formation, we studied the sequence context of sites of CNV. Non-allelic homologous recombination can generate rearrangements as a result of recombination between highly similar duplicated sequences^{32,33}. Segmental duplica-

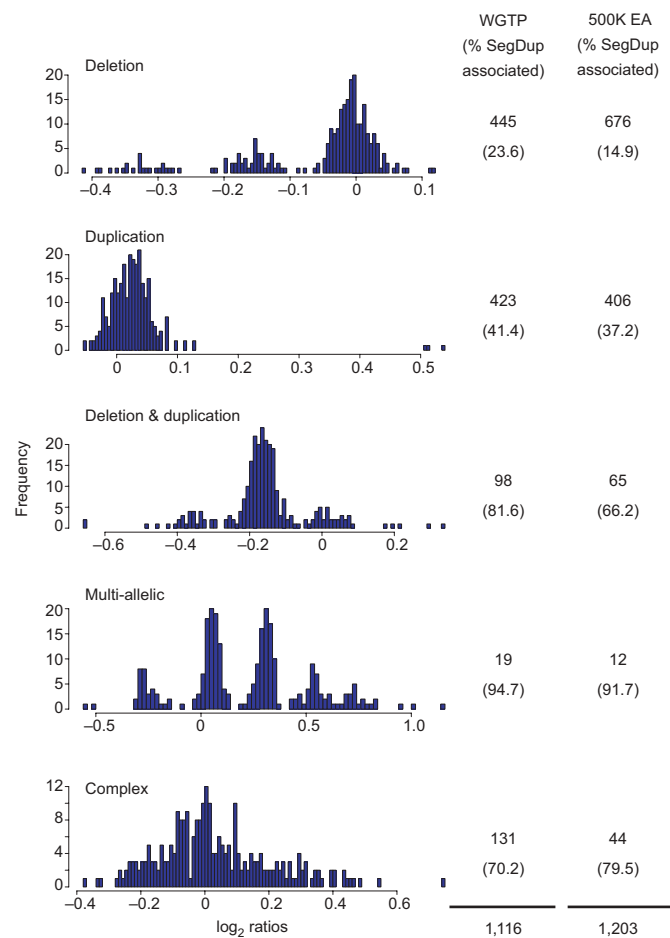


Figure 5 | Classes of CNVs. CNVs identified from WGTP and 500K EA platforms can be classified from the population distribution of \log_2 ratios (exemplified with WGTP data) into five different types (see text). Biallelic CNVs (deletions and duplications) can be genotyped if the clusters representing different genotypes are sufficiently distinct. The numbers of each class of CNV identified on WGTP and 500K EA platforms are given, along with the proportion of those CNVs that overlap segmental duplications. The overall proportion of CNVs overlapping segmental duplications was 20% and 34% on the 500K EA and WGTP platforms, respectively.

tions are defined as sequences in the reference genome assembly sharing $>90\%$ sequence similarity over >1 kb with another genomic location^{34,35}. We found that 24% of the 1,447 CNVs were associated with segmental duplications, a significant enrichment ($P < 0.05$). This association results from two factors: (1) rearrangements generated by non-allelic homologous recombination; and (2) not all annotated segmental duplications are fixed in humans, but are, in fact, CNVs. This latter point highlights the essentially arbitrary nature of defining segmental duplications on the basis of a single genome sequence (albeit derived from several individuals).

The likelihood of a CNV being associated with segmental duplications depended on its length and its classification: multi-allelic CNVs, complex CNVs and loci at which both deletions and duplications occurred were markedly enriched for segmental duplications (Fig. 5; see also Supplementary Fig. 9). This is not surprising given the role that non-allelic homologous recombination has been shown to have in generating complex structural variation³⁶, arrays of tandem duplications that vary in size³⁷, and reciprocal deletions and duplications³⁸.

The likelihood of a segmental duplication being associated with a CNV was greater for intrachromosomal duplications than for interchromosomal duplications, and was highly correlated with increasing sequence similarity to its duplicated copy (Supplementary Fig. 10). Non-allelic homologous recombination is known to operate mainly on intrachromosomal segmental duplications and to require 97–100% sequence similarity between duplicated copies^{33,39}.

This role for non-allelic homologous recombination in generating CNVs in duplicated regions of the genome is supported by the enrichment of segmental duplications within intervals that probably contain the breakpoints of the CNV (Fig. 3). We identified 88 CNVs from the 500K EA platform and 53 CNVs from the WGTP platform that contain a pair of segmental duplications, one at either end. These pairs of segmental duplications were biased towards high ($>97\%$) sequence similarity, and were more frequently associated with the longest CNVs (Supplementary Fig. 11). In addition to segmental duplications, there are other types of sequence homologies that can promote non-allelic homologous recombination, for example, dispersed repetitive elements, such as *Alu* elements⁴⁰. We performed an exhaustive search for sequence homology of all kinds⁴¹ and identified 121 CNVs from the 500K EA platform and 223 on the WGTP platform that contain lengths of perfect sequence identity longer than 100 bp between either end of the CNV.

Genomic impact of CNV

Deletions are known to be biased away from genes⁵, as a result of selection. In contrast, the selective pressures on duplications are poorly understood; the existence of gene families pays testament to positive selection acting on some gene duplications over longer-term evolution⁴². We identified the different classes of functional sequence that fell within CNVs, and tested whether they were significantly enriched or impoverished within these CNVs compared to the entire genome (Table 1; see also Supplementary Table 13 and Supplementary Methods).

Table 1 | Functional sequences within CNVs

Functional sequence	WGTP CNVs	500K EA CNVs	Merged CNVs
RefSeq genes	2,561	1,139†	2,908†
OMIM genes	251	112†	285
Ultra-conserved elements	48†	16†	50†
Conserved non-coding elements	116,678*	55,937*	130,353*
Non-coding RNAs	57	29†	67

Statistical significance of the enrichment or paucity of functional sequences within CNVs was assessed by randomly permuting the genomic location of autosomal CNVs (Supplementary Methods). Significant observations are shown in bold. Note that both conserved non-coding elements⁴³ and CNVs are biased away from genes, so an enrichment of conserved non-coding elements in CNVs is not unexpected.

* Significant ($P < 0.05$) enrichment.

† Significant ($P < 0.05$) paucity.

It is not possible to define precisely the breakpoints of CNVRs; therefore, some of these functional sequences might flank rather than be encompassed by CNVRs. We observed a significant paucity of all functional sequences (with the exception of conserved non-coding sequences⁴³) in CNVRs detected on the 500K EA platform, which provided the highest resolution breakpoint mapping (Table 1). Thus, CNVs are preferentially located outside of genes and ultra-conserved elements in the human genome⁴⁴. We attempted to validate experimentally 11 CNVs containing 12 ultra-conserved elements. Although all but two of the CNVs validated, only two ultra-conserved elements actually fell within these CNVs (Supplementary Table 13B), so the selection against CNV at ultra-conserved elements is likely to be even stronger than this analysis would suggest. Nevertheless, thousands of putatively functional sequences, including known disease-related genes, flank or fall within these CNVs: over half (58%) of the 1,447 CNVRs overlap known RefSeq genes, and more than 99% overlap conserved non-coding sequences⁴³.

We examined whether deletions or duplications are equally likely to encompass these different classes of functional sequences. We observed that a significantly lower proportion of deletions than duplications (identified on the 500K EA platform) overlap with the Online Mendelian Inheritance in Man (OMIM) database of disease-related genes ($P = 0.017$, chi-squared) and RefSeq genes ($P = 1.7 \times 10^{-9}$). Thus, deletions are biased away from genes with respect to duplications. The same trend was observed with ultra-conserved elements but their number is too small to provide statistical significance.

If deletions are under stronger purifying selection (which removes deleterious variants from the population) than duplications^{8,45}, then deletions should, on average, be both less frequent and smaller than duplications. Although, on average, deletions were almost threefold shorter than duplications (43 kb versus 120 kb from 500K EA), we detected no significant difference in the frequencies with which deletions and duplications were called ($P > 0.05$ using G -test for independence⁴⁶ on WGTP data). We note that our length analysis could be confounded if long duplications arise more frequently than long deletions, whereas our frequency analysis could be confounded if the power to detect duplications was lower as a result of the smaller relative change in copy number (3:2 versus 2:1).

We identified functional categories of genes that were enriched within CNVs using the Gene Ontology (GO) database (Supplementary Table 15). The most enriched GO category among genes overlapped by the 1,447 CNVRs was cell adhesion. Other highly enriched categories include sensory perception of smell and of chemical stimulus. Notably, neurophysiological processes were also a highly enriched GO category. The most highly enriched GO categories within CNVRs overlapped appreciably with those identified in a previous analysis of genes in CNVs¹⁴. Genes found in segmental duplications are known to be biased in terms of GO categories³; however, an enrichment of cell adhesion genes was also observed within CNVRs not associated with segmental duplications. We also investigated functional categories that are under-represented within CNVRs, as these might reveal classes of genes that are more likely to be dosage sensitive. We noted an impoverishment of GO categories relating to cell signalling, cell proliferation and numerous kinase- and phosphorylation-related categories. The impoverishment of these gene functions within CNVs probably reflects purifying selection acting both against the altered copy number of cell-signalling molecules vital for development and of dosage-sensitive oncogenes or tumour suppressor genes⁴⁷ that could predispose to early-onset tumorigenesis.

Copy number variation of medical relevance

In the absence of phenotypic information for HapMap donors, our data are most relevant for highlighting variable regions of the genome that warrant consideration in disease studies, rather than for immediate application to clinical diagnostics.

We found that 285 out of 1,961 (14.5%) genes in the OMIM morbid map overlapped with CNVs (Supplementary Table 16). We observed numerous examples of possible relevance to both mendelian and complex diseases. For example, the breakpoint region(s) for 12 of 25 loci involved in genomic disorders (which cause 33 different diseases) such as DiGeorge and Williams–Beuren syndromes³⁹ were found to be highly polymorphic (Supplementary Table 17). CNVs were also identified within the regions commonly deleted in DiGeorge, Smith–Magenis, Williams–Beuren, Prader–Willi and Angelman syndromes, which may be relevant for discerning uncharacterized or atypical cases. We also found CNVs at the spinal muscular atrophy and nephronophthisis loci, as expected, as these diseases are recessive in nature with relatively high carrier frequencies³³. Finally, 39 CNVs were found to reside within 500 kb of the ends of 36 chromosomal arms, which is relevant when assessing subtelomeric rearrangements in disease.

We found CNVs in genes already known to be responsible for complex traits, including *CCL3L1* and *FCGR3B*^{19,20}. Some new observations were also documented. Two CEU samples (mother and offspring) manifested a gain of CNV-95 involving the first six exons of *DISC1*, which is disrupted in schizophrenia⁴⁸. CNV-575, encompassing the LPA apolipoprotein A gene, demonstrated population variability, which may influence susceptibility to atherosclerosis. The *CRYBB2-CRYBB3* β -crystallin genes in CNV-1367 were observed as gains and losses in copy number in CEU and YRI samples. However, only gains were detected in Asians, leading us to speculate that variability in crystallin copy number may be linked to population differences of onset of age-related cataracts⁴⁹. Following a similar rationale, we highlight CNV-507 for possible involvement in sarcoidosis owing to its proximity to the *BTNL2* gene⁵⁰, and CNV-505 in psoriasis susceptibility, because it covers the 6p21.3 (*PSORS1*) susceptibility locus⁵¹.

We also highlight challenges in resolving genotype–phenotype correlations in complex CNVRs and how CNV detection can delineate unstable genomic regions (details in Supplementary Note including Supplementary Fig. 12). We identified patients with congenital cardiac defects and lens abnormalities⁵² that share a deleted or duplicated ~ 1 -Mb region of 1q21.1 containing pertinent candidate genes. Duplication of the same interval was observed in other cases with mental retardation²⁹. Proband can inherit the disease-associated rearrangements from unaffected parents, which underscores the variable penetrance of some diseases resulting from dosage effects^{53,54}. We found that this locus is highly duplicated, polymorphically inverted, contains assembly gaps, and is flanked by segmental duplications of variable copy number, all features being increasingly observed in CNV regions of the human genome.

Imprint of CNV on SNP genotypes

Deletions perturb patterns of marker genotypes within pedigrees⁵⁴ and these patterns highlight the location of such deletions. SNP genotype patterns characteristic of deletions are enrichment of the following: null genotypes in homozygous deletions; mendelian discrepancies in families; and Hardy–Weinberg disequilibrium within a population^{5,7}. Duplications can similarly lead to misinterpretation of marker genotypes^{53,55}, although their impact on high-density SNP maps is poorly understood.

We characterized the patterns of Phase I HapMap SNP genotypes within deletions and duplications on a chromosome-by-chromosome basis to take account of regional biases, and found that most classes of aberrant SNP genotypes were significantly enriched in both deletions and duplications on most chromosomes (Supplementary Fig. 13). We replicated the patterns of SNP genotypes within deletions (described above), and demonstrated that duplications also impact significantly on SNP genotypes. The spectrum of SNP failures enriched within duplications distinguishes them from deletions (Supplementary Fig. 13 and Supplementary Table 18). Most notably, SNPs exhibiting mendelian inconsistencies are more common within

deletions than duplications, whereas the opposite is true for SNPs in Hardy–Weinberg disequilibrium and SNPs with missing genotypes.

Cell-line artefacts also have an impact on SNP genotypes. For example, the partial deletion of 17p in NA12056 causes a distinct paucity of heterozygous SNP genotypes in that individual (one-quarter that of the population average) over several megabases of chromosome 17 (HapMap release 20).

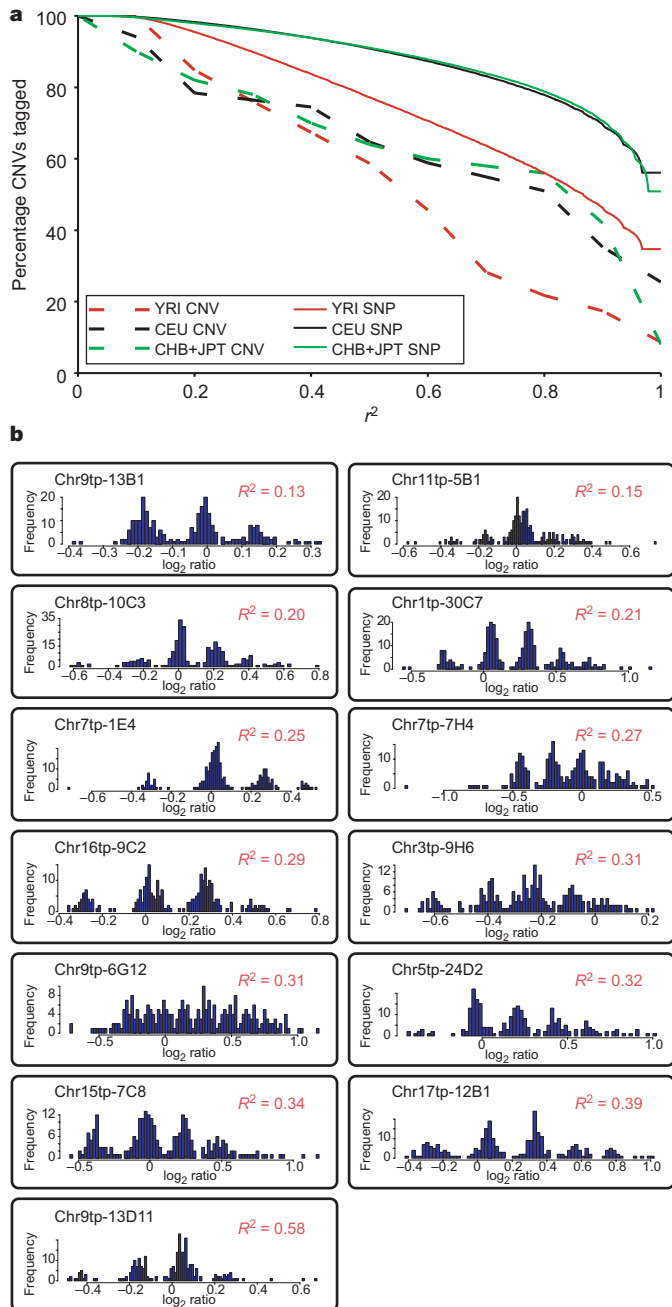


Figure 6 | Patterns of linkage disequilibrium between CNVs and SNPs. **a**, The proportion of variants that are tagged by a nearby proxy SNP (from Phase I HapMap) increases as the pairwise linkage disequilibrium (r^2) required for a proxy SNP is relaxed. This cumulative distribution is shown for both Phase I HapMap SNPs and for 65 biallelic CNVs. **b**, Histograms of the \log_2 ratios among all HapMap individuals are shown for 13 multi-allelic CNVs. The maximal squared Pearson correlation coefficient (R^2) observed at a neighbouring Phase I HapMap SNP—which is highly correlated with pairwise linkage disequilibrium (r^2) at biallelic CNVs (Supplementary Fig. 15)—is given for each CNV.

Linkage disequilibrium around CNVs

Indirect methods to identify causative variants, such as co-segregation of linked markers in families and genetic association with markers in linkage disequilibrium with the causative variant, are considered to be blind to the nature of the underlying mutation⁵⁶. This raises the question of whether SNP-based whole-genome association studies have the same power to detect disease-related CNVs as for disease-related SNPs. This question can be addressed by considering the maximal pairwise linkage disequilibrium (r^2) between a particular variant (CNV or SNP) and any of its neighbouring polymorphic markers. If a neighbouring marker is in high linkage disequilibrium (r^2 close to 1) with the variant of interest, that variant is ‘tagged’ by the neighbouring marker; the genotype of the variant of interest can be predicted with high probability by knowing the genotype at the ‘tagging’ marker.

Recent studies of linkage disequilibrium around CNVs have produced conflicting evidence as to the degree to which CNVs are ‘tagged’ by neighbouring SNPs^{6–8}. Here we performed a balanced comparison between the linkage disequilibrium properties of biallelic CNVs and Phase I HapMap SNPs by considering CNVs irrespective of their genomic location, and by analysing CNV genotypes (Supplementary Table 7 and Supplementary Fig. 14) of the same frequency and quality as SNP genotypes^{5,7,25} (see Methods for details). We quantified pairwise linkage disequilibrium around 65 biallelic CNVs using the same three analysis panels as for the Phase I HapMap (CEU, YRI and JPT plus CHB). Comparing the proportion of variants tagged by a neighbouring SNP with an arbitrary threshold of $r^2 > 0.8$ shows that whereas 75–80% of Phase I SNPs in non-African populations were tagged, only 51% of CNVs were tagged in the same populations (Fig. 6a and Supplementary Table 19). In the YRI, both SNPs and CNVs exhibited lower linkage disequilibrium, with only 22% of CNVs being tagged with $r^2 > 0.8$.

We considered three explanations for these observations of lower apparent linkage disequilibrium around CNVs than SNPs. First, some duplications might represent transposition events that would generate linkage disequilibrium around the (unknown) acceptor locus but not the donor locus. One of the genotyped CNVs is known to be a duplicative transposition⁵⁷, but evidence from *de novo* pathogenic duplications strongly suggests a preference for tandem, rather than dispersed, duplications, regardless of whether the duplication is caused by non-allelic homologous recombination⁵⁸. Second, some CNVs might undergo recurrent mutations or reversions, especially tandem duplications which are mechanistically prone to unequal crossing over, causing reversions back to a single copy¹². However, duplications were not in lower linkage disequilibrium with flanking SNPs than were deletions. Finally, we considered that CNVs might occur preferentially in genomic regions with lower densities of SNP genotypes in HapMap Phase I. We found that CNVs are enriched within segmentally duplicated regions of the genome, in which there is a paucity of genotyped SNPs owing to technical difficulties²⁵. Thus, the strongest factor decreasing apparent linkage disequilibrium around biallelic CNVs is not that linkage disequilibrium around such CNVs is necessarily lower, but that there is, on average, lower coverage of these structurally dynamic regions of the genome by SNPs genotyped in Phase I of the HapMap project.

We investigated whether the copy number of multi-allelic CNVs could be predicted reliably by nearby SNPs. We treated diploid genome copy number at multi-allelic CNVs as a quantitative trait, and asked which nearby SNPs are most predictive of this trait, and how strongly predictive these SNPs are. We identified 13 multi-allelic CNVs in which the quantitative WGTP data clearly clustered into discrete diploid genome copy numbers and quantified the predictive ability of neighbouring SNPs using the square of Pearson’s correlation coefficient (R^2) (Supplementary Fig. 15). We found that diploid copy number of multi-allelic CNVs is poorly predicted by neighbouring SNPs (Fig. 6b). It may be that combining information from

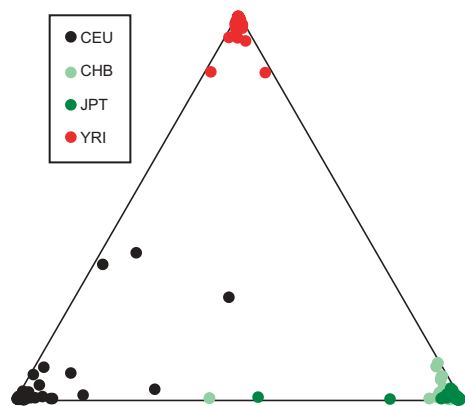


Figure 7 | Population clustering from CNV genotypes. A triangle plot showing the clustering of 210 unrelated HapMap individuals assuming three ancestral populations ($k = 3$). The proximity of an individual to each apex of the triangle indicates the proportion of that genome that is estimated to have ancestry in each of the three inferred ancestral populations. The clustering together of most individuals from the same population near a common apex indicates the clear discrimination between populations obtained through this analysis. The clustering was qualitatively similar to that obtained previously with a similar number of biallelic *Alu* insertion polymorphisms on different African, European and Asian population samples⁶⁰.

several SNPs could provide greater power for predicting diploid genome copy number at these loci.

Population genetics of copy number variation

In contrast to other classes of human genetic variation, the population genetics of copy number variation remains unexplored. The distribution of copy number variation within and among different populations is shaped by mutation, selection and demographic history. A range of polymorphisms, including SNPs²⁵, microsatellites⁵⁹ and *Alu* insertion variants⁶⁰, has been used to investigate population structure. To demonstrate the utility of copy number variation genotypes for population genetic inference we performed population clustering⁶¹ on 67 genotyped biallelic CNVs. We obtained the optimal clustering with the assumption of three ancestral populations, with the African, European and Asian populations clearly differentiated (Fig. 7). Population differentiation of individual variants is commonly estimated by the statistic F_{ST} , which varies from 0 (undifferentiated) to 1 (population-specific)⁶². The average F_{ST} for the same 67 autosomal CNVs was 0.11, very similar to that observed for all autosomal Phase I HapMap SNPs (0.13)²⁵.

Recent population-specific positive selection elevates population differentiation. To explore population differentiation at all CNVs, we devised a statistic, V_{ST} , that estimates population differentiation based on the quantitative intensity data and varies from 0 to 1, similar to F_{ST} (Supplementary Fig. 16). Estimating V_{ST} for all clones on the

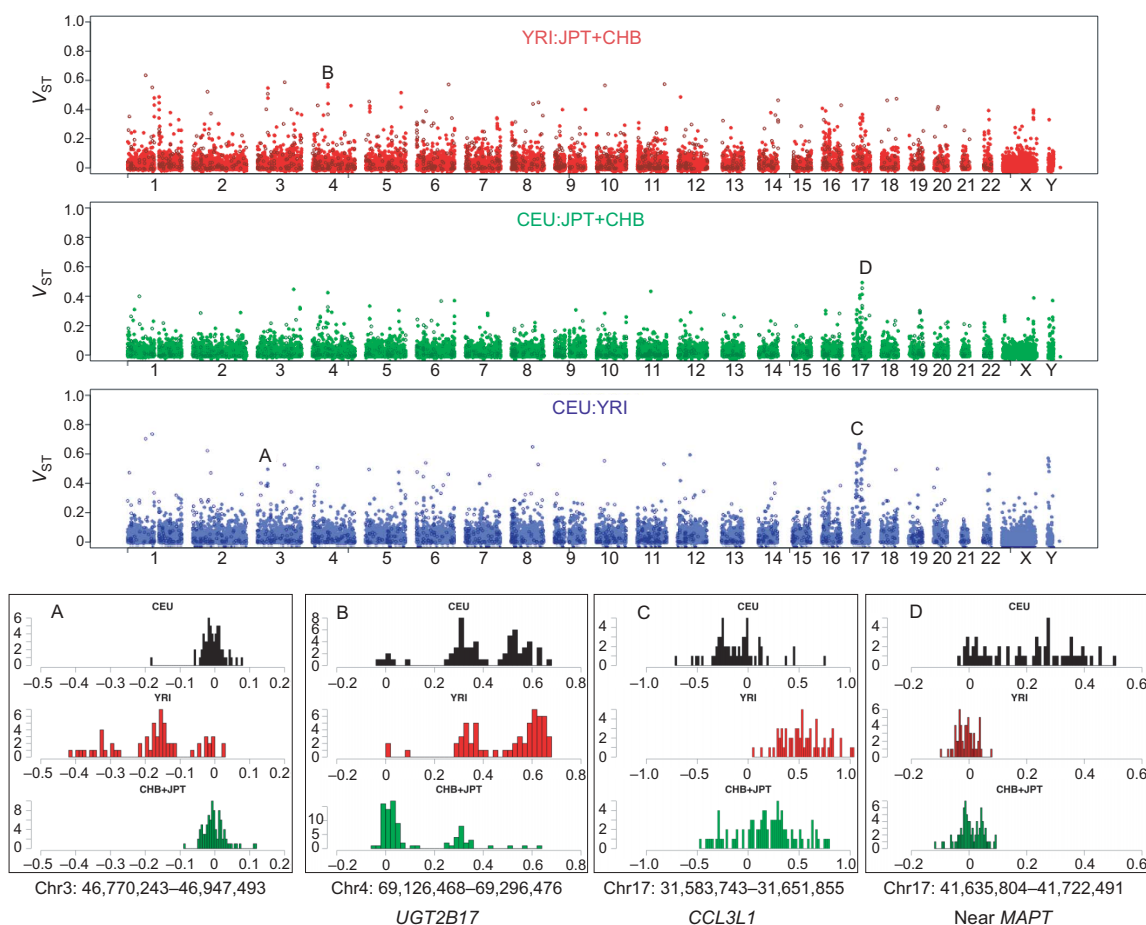


Figure 8 | Population differentiation for copy number variation. Population differentiation, estimated by V_{ST} , for each of the three population pairwise comparisons is plotted along each chromosome. For each pairwise comparison, the V_{ST} values for all clones on the WGTP platform are shown in the lighter colour with filled circles, with V_{ST} values of CNVs detected on the 500K EA platform superimposed in a darker shade with unfilled circles. Histograms showing the distributions of \log_2 ratios (on

the WGTP platform) among the unrelated individuals in each population are plotted for four example CNVs exhibiting high population differentiation, labelled A–D. y axis values indicate absolute frequency. Each example histogram is labelled with the chromosome coordinates of the WGTP clone, and flanking/encompassed genes are given for those CNVs mentioned in the text.

WGTP array and all CNVs on the 500K EA array revealed a number of outliers with levels of population differentiation suggestive of population-specific selective pressures (Fig. 8; see also Supplementary Table 20). Among these outliers were two CNVs previously demonstrated to have elevated population differentiation^{7,19}: *UGT2B17* is a gene encoding a UDP-glucuronosyl transferase with roles in androgen metabolism and xenobiotic conjugation^{63,64}, and *CCL3L1* is a chemokine-encoding multi-copy gene at which greater copy numbers protect against HIV-1 infection¹⁹.

Not all regions that have been under recent positive selection exhibit elevated population differentiation⁶⁵. To detect other CNVs that may have recently been under positive selection, we identified CNVRs that fell within 124 out of 752 (16%) genomic locations previously shown⁶⁶ to exhibit haplotype patterns suggestive of a partial selective sweep (Supplementary Table 21). Two of these overlapping CNVs also fell within the set exhibiting highest population differentiation, shown in Supplementary Table 20. One of these selection-associated CNVs is a duplication specific to the CEU (Fig. 8) and lies near to the *MAPT* gene, which is associated with a set of neurodegenerative disorders known as 'tauopathies'⁶⁷. Both the *MAPT* gene and the duplication lie within a chromosomal region that has recently been shown to have a complex evolutionary history, characterized by a common chromosomal inversion, deep divergence between inverted haplotypes and recent positive selection in European populations⁶⁸. We adapted methods used to identify partial selective sweeps at SNPs^{66,69} to estimate relative extended haplotype homozygosity values (REHH) on either flank of the 67 CNVs for which we extracted genotype information (Supplementary Methods). We identified no convincing signals ($P < 0.01$ on both flanks) of positive selection on any CNV in any one population, although there were weaker signals ($P < 0.05$) apparent for some CNVs (Supplementary Fig. 17 and Supplementary Table 21).

Discussion

Our map of copy number variation in the human genome demonstrates the ubiquity and complexity of this form of genomic variation. The abundance of functional sequences of all types both within and flanking areas of copy number variation suggests that the contribution of CNVs to phenotypic variation is likely to be appreciable. This prediction is underscored by the impact of copy number variation on variation in gene expression (B. Stranger and E. Dermitzakis, personal communication).

CNV assessment should now become standard in the design of all studies of the genetic basis of phenotypic variation, including disease susceptibility. Similarly important will be CNV annotation in all future genome assemblies. The identification of CNVs causing severe, sporadic diseases has been hampered by an inability to distinguish between normal and causative variants. Our CNV map, in tandem with the DECIPHER (<http://www.sanger.ac.uk/PostGenomics/decipher/>) project's sharing of copy number information on patients with rare, severe phenotypes should advance progress in this area. For mendelian genetic diseases, our data contain numerous known and candidate recessive disease alleles, and existing linkage data can be used to prioritise CNVs for further investigation.

Genetic association studies are the predominant strategy for identifying haplotypes conferring risk for complex genetic diseases. Such studies are typically based on SNP genotyping, either within candidate loci or genome-wide⁵⁶. Our analysis of linkage disequilibrium between CNVs and SNPs gives us limited optimism that CNVs influencing risk to complex disease will be detected by such approaches. The tag SNPs that we have identified (Supplementary Table 19) for specific CNVs can be used as proxies for these CNVs. Moreover, CNV-specific genotyping assays can be developed for CNVs for which tag SNPs are not readily identifiable but whose proximity to candidate genes warrants further characterization. Finally, we see great merit in mining CNV information from quantitative SNP gen-

otyping data, and in enriching future generations of genome-wide SNP genotyping platforms by targeting untaggable CNVs.

The overall utility of any map depends on its coverage and its completeness. Extrapolation based on existing data suggests that smaller deletions (<20 kb) are much more frequent than larger deletions (>20 kb)⁵, and the same may be true for duplications. Although we have generated the most complete CNV map yet described, given our lower power to detect smaller CNVs a substantial fraction of copy number variation >1 kb in size in these individuals remains to be characterized. No single available technology will capture all variation. Smaller rearrangements are amenable to detection using technologies such as sequence assembly comparisons⁷⁰, paired-end sequence relationships⁴, sequence trace analysis⁷¹ and higher-resolution tiling arrays⁵. Ultimately, it is desirable to know the precise chromosomal location and sequence content of each and every CNV. At present, generating this information requires the use of multiple experimental methods. However, in the future, the comparison of independently assembled whole-genome sequences could provide a definitive solution.

Our study of human genetic variation ties together cytogenetics, submicroscopic copy number variation and single nucleotide polymorphisms, providing a framework for future genetic studies. This framework will need to be supported by continual refinement of the reference genome sequence and robust nomenclature and databasing for structural variation—both facilitated through international collaboration—to enable further unravelling of the complexity of human genomic variation.

METHODS

CNV determination on WGTP and 500K EA arrays. The experimental methods and algorithms used for CNV calling using the WGTP and the 500K EA platforms are described in the Supplementary Methods and elsewhere^{26,28}.

500K EA and WGTP data quality assessment. In order to estimate false-positive and false-negative rates on both platforms, we used quantitative PCR to test experimentally CNVs called from replicate experiments with NA15510 (ref. 4; Supplementary Tables 1–4). With 500K EA, the average proportion of CNV calls from this sample that were false positive was 2.3% (0.33 out of 15 CNV calls), and the false-negative percentage was 24% (3.3 validated CNVs not called in any one replicate per 38 total validated CNVs)²⁸. With WGTP, we found an average false-positive proportion of 5% (3.4 out of 68.2) and a false-negative percentage of 37.8% (58 non-called out of 154 tested)²⁶. Reproducibility was also assessed by analysing ten HapMap DNAs in triplicate (Supplementary Table 6). For the 500K EA platform, on average, 80% of CNVs were called in all three replicates, 10% were called twice, and 10% were called only once. For the WGTP platform, using the replicate with the lowest number of calls as the baseline, on average 73% of CNVs were called in three experiments, 14% were called twice, and 13% were called once²⁶.

Population genetic and statistical analyses. Sixty-seven non-redundant biallelic CNVs suitable for genotyping were identified on the WGTP and 500K EA platforms. Two procedures were used to cluster intensity ratios into discreet copy number genotypes: K-means and partitioning around medoids (PAM)⁷². Pairwise linkage disequilibrium (r^2) was estimated between 65 biallelic CNVs and all filtered non-redundant Phase I HapMap SNPs within 500 kb of the CNV borders using HaploView⁷³. Population clustering was performed using STRUCTURE⁶¹ and population-specific CNVs were estimated using F_{ST} ⁶² and the new statistic V_{ST} (Supplementary Fig. 16). V_{ST} is calculated by considering $(V_T - V_S)/V_T$, where V_T is the variance in \log_2 ratios apparent among all unrelated individuals and V_S is the average variance within each population, weighted for population size. For REHH analysis^{69,74}, we treated each genotyped CNV as a SNP located at the CNV end, and used Phase I HapMap SNPs 500 kb upstream and downstream of the CNV using the program Sweep (<http://www.broad.mit.edu/mpg/sweep/resources.html>). See Supplementary Methods for more details.

Data release. The raw data from the 500K EA as well as the 500K commercial arrays are posted at the Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>), with accession numbers GSE5013 and GSE5173. WGTP data are posted at ArrayExpress (<http://www.ebi.ac.uk/arrayexpress/>) with accession number E-TABM-107, and at the Wellcome Trust Sanger Institute (<http://www.sanger.ac.uk/humgen/cnv/data/>). CNV calls have been released at the Database of Genomic Variants (<http://projects.tcag.ca/variation/>) integrated with all other CNV data.

Received 13 June; accepted 10 October 2006.

1. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36**, 949–951 (2004).
2. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
3. Sharp, A. J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
4. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nature Genet.* **37**, 727–732 (2005).
5. Conrad, D. F., Andrews, T. D., Carter, N. P., Hurler, M. E. & Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. *Nature Genet.* **38**, 75–81 (2006).
6. Hinds, D. A., Kloek, A. P., Jen, M., Chen, X. & Frazer, K. A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nature Genet.* **38**, 82–85 (2006).
7. McCarroll, S. A. *et al.* Common deletion polymorphisms in the human genome. *Nature Genet.* **38**, 86–92 (2006).
8. Locke, D. P. *et al.* Linkage disequilibrium and heritability of CNPs within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**, 275–290 (2006).
9. Fredman, D. *et al.* Complex SNP-related sequence variation in segmental genome duplications. *Nature Genet.* **36**, 861–866 (2004).
10. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nature Rev. Genet.* **7**, 85–97 (2006).
11. Freeman, J. L. *et al.* Copy number variation: new insights in genome diversity. *Genome Res.* **16**, 949–961 (2006).
12. Bridges, C. B. The Bar 'gene': a duplication. *Science* **83**, 210–211 (1936).
13. Buckland, P. R. Polymorphically duplicated genes: their relevance to phenotypic variation in humans. *Ann. Med.* **35**, 308–315 (2003).
14. Nguyen, D. Q., Webber, C. & Ponting, C. P. Bias of selection on human copy-number variants. *PLoS Genet.* **2**, e20 (2006).
15. Repping, S. *et al.* High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nature Genet.* **38**, 463–467 (2006).
16. Inoue, K. & Lupski, J. R. Molecular mechanisms for genomic disorders. *Annu. Rev. Genomics Hum. Genet.* **3**, 199–242 (2002).
17. Lupski, J. R. & Stankiewicz, P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet.* **1**, e49 (2005).
18. Shaw-Smith, C. *et al.* Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J. Med. Genet.* **41**, 241–248 (2004).
19. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440 (2005).
20. Aitman, T. J. *et al.* Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851–855 (2006).
21. Jongmans, M. C. *et al.* CHARGE syndrome: the phenotypic spectrum of mutations in the CHD7 gene. *J. Med. Genet.* **43**, 306–314 (2006).
22. Singleton, A. B. *et al.* α -Synuclein locus triplication causes Parkinson's disease. *Science* **302**, 841 (2003).
23. Rovelet-Lecrux, A. *et al.* APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nature Genet.* **38**, 24–26 (2006).
24. Feuk, L., Marshall, C. R., Wintle, R. F. & Scherer, S. W. Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum. Mol. Genet.* **15** (Suppl. 1), R57–R66 (2006).
25. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
26. Fiegler, H. *et al.* Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Res.* **16**, 1566–1574 (2006).
27. Ylstra, B., van den Ijssel, P., Carvalho, B., Brakenhoff, R. H. & Meijer, G. A. BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res.* **34**, 445–450 (2006).
28. Komura, D. *et al.* Genome-wide detection of human copy number variations using high density DNA oligonucleotide arrays. *Genome Res.* **16**, 1575–1584 (2006).
29. de Vries, B. B. *et al.* Diagnostic genome profiling in mental retardation. *Am. J. Hum. Genet.* **77**, 606–616 (2005).
30. Wirtenberger, M., Hemminki, K., Chen, B. & Burwinkel, B. SNP microarray analysis for genome-wide detection of crossover regions. *Hum. Genet.* **117**, 389–397 (2005).
31. Marth, G. T., Czabarka, E., Murvai, J. & Sherry, S. T. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351–372 (2004).
32. Hurler, M. How homologous recombination generates a mutable genome. *Hum. Genomics* **2**, 179–186 (2005).
33. Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).
34. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
35. Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* **4**, R25 (2003).
36. Aradhya, S. *et al.* Multiple pathogenic and benign genomic rearrangements occur at a 35 kb duplication involving the NEMO and LAGE2 genes. *Hum. Mol. Genet.* **10**, 2557–2567 (2001).
37. Neitz, M. & Neitz, J. Numbers and ratios of visual pigment genes for normal red-green color vision. *Science* **267**, 1013–1016 (1995).
38. Chance, P. F. *et al.* Two Autosomal-dominant neuropathies result from reciprocal DNA duplication/deletion of a region on chromosome-17. *Hum. Mol. Genet.* **3**, 223–228 (1994).
39. Lupski, J. R. & Stankiewicz, P. (eds) *Genomic Disorders: the Genomic Basis of Disease* (Humana Press, Totawa, New Jersey, 2006).
40. Deininger, P. L. & Batzer, M. A. Alu repeats and human disease. *Mol. Genet. Metab.* **67**, 183–193 (1999).
41. Kurtz, S. *et al.* REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* **29**, 4633–4642 (2001).
42. Hurler, M. Gene duplication: the genomic trade in spare parts. *PLoS Biol.* **2**, E206 (2004).
43. Drake, J. A. *et al.* Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature Genet.* **38**, 223–227 (2006).
44. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
45. Brewer, C., Holloway, S., Zawalynski, 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).
46. Sokal, R. R. & Rohlf, J. F. *Biometry* (W. H. Freeman and Co, New York, 1994).
47. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* **4**, 177–183 (2004).
48. Millar, J. K. *et al.* DISC1 and PDE4B are interacting genetic factors in schizophrenia that regulate cAMP signaling. *Science* **310**, 1187–1191 (2005).
49. Sasaki, K., Sasaki, H., Jonasson, F., Kojima, M. & Cheng, H. M. Racial differences of lens transparency properties with aging and prevalence of age-related cataract applying a WHO classification system. *Ophthalmic Res.* **36**, 332–340 (2004).
50. Valentonyte, R. *et al.* Sarcoidosis is associated with a truncating splice site mutation in BTNL2. *Nature Genet.* **37**, 357–364 (2005).
51. Orru, S., Giurelli, E., Carcassi, C., Casula, M. & Contu, L. Mapping of the major psoriasis-susceptibility locus (PSORS1) in a 70-Kb interval around the corneodesmosin gene (CDSN). *Am. J. Hum. Genet.* **76**, 164–171 (2005).
52. Christiansen, J. *et al.* Chromosome 1q21.1 contiguous gene deletion is associated with congenital heart disease. *Circ. Res.* **94**, 1429–1435 (2004).
53. Lupski, J. R. *et al.* DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell* **66**, 219–232 (1991).
54. Chance, P. F. *et al.* DNA deletion associated with hereditary neuropathy with liability to pressure palsies. *Cell* **72**, 143–151 (1993).
55. Matise, T. C. *et al.* Detection of tandem duplications and implications for linkage analysis. *Am. J. Hum. Genet.* **54**, 1110–1121 (1994).
56. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* **6**, 95–108 (2005).
57. 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).
58. Woodward, K. J. *et al.* Heterogeneous duplications in patients with Pelizaeus-Merzbacher disease suggest a mechanism of coupled homologous and nonhomologous recombination. *Am. J. Hum. Genet.* **77**, 966–987 (2005).
59. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
60. Bamshad, M. J. *et al.* Human population genetic structure and inference of group membership. *Am. J. Hum. Genet.* **72**, 578–589 (2003).
61. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
62. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution Int. J. Org. Evol.* **38**, 1358–1370 (1984).
63. Turgeon, D., Carrier, J. S., Chouinard, S. & Belanger, A. Glucuronidation activity of the UGT2B17 enzyme toward xenobiotics. *Drug Metab. Dispos.* **31**, 670–676 (2003).
64. Wilson, W. III. *et al.* Characterization of a common deletion polymorphism of the UGT2B17 gene linked to UGT2B15. *Genomics* **84**, 707–714 (2004).
65. Xue, Y. *et al.* Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *Am. J. Hum. Genet.* **78**, 659–670 (2006).
66. 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).
67. Rademakers, R., Cruts, M. & van Broeckhoven, C. The role of tau (MAPT) in frontotemporal dementia and related tauopathies. *Hum. Mutat.* **24**, 277–295 (2004).
68. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nature Genet.* **37**, 129–137 (2005).
69. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
70. Khaja, R. *et al.* Genome assembly comparison identifies structural variants in the human genome. *Nature Genet.* **38**, 1413–1418 (2006).
71. Bennett, E. A., Coleman, L. E., Tsui, C., Pittard, W. S. & Devine, S. E. Natural genetic variation caused by transposable elements in humans. *Genetics* **168**, 933–951 (2004).

72. Kaufman, L. & Rousseeuw, P. J. *Finding Groups in Data: an Introduction to Cluster Analysis* (Wiley, New York, 1990).
73. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
74. Walsh, E. C. *et al.* Searching for signals of evolutionary selection in 168 genes related to immune function. *Hum. Genet.* **119**, 92–102 (2006).
75. Dermitzakis, E. T. *et al.* Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* **14**, 852–859 (2004).
76. Petes, T. D. Meiotic recombination hot spots and cold spots. *Nature Rev. Genet.* **2**, 360–369 (2001).
77. 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).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Altshuler, C. Bird, Y. Chen, M. Daly, C. Fahey, A. M. Joseph-George, Y. He, K. Hirose, Z. Hu, V. Jayanth, C. Langford, M. Li, C. Lu, G. Liu, Z. Liu, H. Meguro, L. Pantano, T. Paton, I. Pe'er, S. Pullenayegum, Y. Qi, S. Russell, M. Schachowsky, M. Shago, K. Shiina and Y. Xue for advice, sharing data, cell lines, technical assistance or bioinformatics support. The Centre for Applied Genomics at The Hospital for Sick Children and the Microarray Facility of the Wellcome Trust Sanger Institute are acknowledged for database support and array printing, respectively. We also thank J. R. Lupski and J. Pritchard for comments on earlier versions of the manuscript. The research was supported by The Wellcome Trust (M.E.H., N.P.C., C.T.S.), Canada Foundation of Innovation and Ontario

Innovation Trust (S.W.S.), Canadian Institutes of Health Research (CIHR) (S.W.S.), Genome Canada/Ontario Genomics Institute (S.W.S.), the McLaughlin Centre for Molecular Medicine (S.W.S.), Ontario Ministry of Research and Innovation (S.W.S.), the Hospital for Sick Children Foundation (S.W.S.), The Department of Pathology, Brigham and Women's Hospital (C.L.), The Leukemia and Lymphoma Society (C.L.), the Core Research for Evolutional Science and Technology (CREST) from the Japan Science and Technology Agency, Grant-in-Aid for Scientific Research (H.A.), Grants-in-Aid for Young Scientists and Scientific Research on Priority Areas 'Applied Genomics' from the Ministry of Education, Culture, Sports, Science and Technology of Japan (S.I.), the 'Departament d'Universitats Recerca i Societat de la Informació' (X.E.) and the Genoma España and Genome Canada joint R+D+I projects (X.E. and S.W.S.), the Generalitat de Catalunya (Departament d'Universitats and Departament de Salut), the 'Fondo de Investigaciones Sanitarias de la Seguridad Social, Instituto de Salud Carlos III', and the National Genotyping Center Barcelona Node (CeGen) supported by Genoma España (X.E.), and a Packard Fellowship to J. K. Pritchard (D.C.). L.F. is supported by a fellowship from CIHR; R.R. by a Sanger Institute Postdoctoral Fellowship; and A.C. from the Natural Science and Engineering Research Council. S.W.S. is an Investigator of the CIHR and International Scholar of the Howard Hughes Medical Institute.

Author Contributions R.R., S.I., K.R.F., L.F. and G.H.P. are first authors with equal contribution; C.T.-S., N.P.C., H.A., C.L., K.W.J., S.W.S. and M.E.H. are principal investigators with equal contribution.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to M.E.H. (meh@sanger.ac.uk) or S.W.S. (steve@genet.sickkids.on.ca).