ANALYSIS

A copy number variation map of the human genome

Mehdi Zarrei[1], Jeffrey R. MacDonald[1], Daniele Merico[1] and Stephen W. Scherer[1,2]

Abstract | A major contribution to the genome variability among individuals comes from deletions and duplications — collectively termed copy number variations (CNVs) — which alter the diploid status of DNA. These alterations may have no phenotypic effect, account for adaptive traits or can underlie disease. We have compiled published high-quality data on healthy individuals of various ethnicities to construct an updated CNV map of the human genome. Depending on the level of stringency of the map, we estimated that 4.8–9.5% of the genome contributes to CNV and found approximately 100 genes that can be completely deleted without producing apparent phenotypic consequences. This map will aid the interpretation of new CNV findings for both clinical and research applications.

Copy number variation
(CNV). A genomic segment of at least 50 bp that differs in copy number based on the comparison of two or more genomes.

Unbalanced rearrangements
Genomic variants that involve loss (deletion) or gain (duplication) of segments of the genome.

[1]*The Centre for Applied Genomics, and Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario M5G 1L7, Canada.*
[2]*McLaughlin Centre and Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5G 1L7, Canada.*
Correspondence to S.W.S.
e-mail: stephen.scherer@sickkids.ca
doi:10.1038/nrg3871
Published online 3 February 2015

The DNA sequence along human chromosomes is constantly changing, and this process enables humans to evolve and adapt[1–3]. We have long been aware of genetic variation at either size extreme — cytogenetically recognizable segments[4] and single-nucleotide polymorphisms (SNPs)[5,6]. About 10 years ago, scientists began to recognize abundant variation of an intermediate size class known as structural variation[7–10]. Within this class, copy number variation (CNV), which involves unbalanced rearrangements that increase or decrease the DNA content, accounts for the largest component by far[9,11]. We now typically define the size of CNVs as larger than 50 bp[12], whereas smaller elements are known as insertions or deletions (indels). These structural variations encompass more polymorphic base pairs than SNPs by an order of magnitude[11,13–17].

There is a continuous spectrum of phenotypic effects of CNV, from adaptive traits to embryonic lethality[2,18]. The threshold over which the effect of a CNV may be associated with a disease is defined by clinicians and is arbitrary[19]. However, a trait can be adaptive or maladaptive in different environmental contexts; for example, CNVs in the gene encoding α-amylase enable adaptation to dietary starch consumption[20]. Maladaptive CNVs have been associated with conditions such as autism[21,22], schizophrenia[23], Crohn's disease[24,25], rheumatoid arthritis[24], type 1 diabetes[24], obesity[26] and numerous developmental diseases[15,27–29]. A special class of CNVs with rearrangement breakpoints mediated by non-contiguous stretches of repetitive DNA along chromosomes are grouped under the term genomic disorders[30]; mechanistically (and, in many cases, phenotypically), these are the best characterized imbalances in the genome[31]. Essentially, pathogenic CNVs continue to be described in different classes of disease[15]. Clinical geneticists need to discriminate pathogenic or high-risk variants from benign variants in patients, and they do so with information about CNVs found in apparently healthy individuals, which can include parents, siblings and population controls[14,15,27,32]. A major challenge in the field is that there is a growing number of CNVs (known as 'variants of unknown significance') that are suspected to be involved in disease susceptibility but for which additional population-level data are required[33–35].

The Database of Genomic Variants (DGV) was created in 2004 (REF. 7) as a comprehensive catalogue of human CNV and structural variation among 'control' individuals. Working with the community and other major databases[35,36], DGV is continually updated and curated[12]. In many cases, 'healthy controls' have been defined in relation to a particular disease study, and other possible conditions may not have been ruled out. There are also studies that draw attention to the abundance of somatic CNVs in healthy tissues[37], the influence of age on CNV representation in populations[38] and the remaining technical challenges in the characterization of complex CNVs[1,39,40]. Notwithstanding these limitations and with a decade of accumulated data and experience[12], we believe that we are now in a position to use the abundance of existing information to create a human CNV map (BOX 1) to document the extent and characteristics of this form of variation among healthy populations.

© 2015 Macmillan Publishers Limited. All rights reserved

Box 1 | **Copy number variation map**

The copy number variation (CNV) map of the human genome catalogues benign CNVs among presumably healthy individuals of various ethnicities. The current map includes microscopic and submicroscopic variants from 50 bp to 3 Mb. We generated separate CNV maps for deletions and duplications, and a CNV map for both together. Duplications tend to be under less negative selective pressure than deletions because they usually have a milder phenotypic effect.

The CNVs that do not necessarily cause pathogenic effects were first described at the genome-wide level in 2004 (REFS 7,8). We have used the term CNV[1] and CNV map[9] since 2006, but we also acknowledge the term 'chromosome imbalance', which was coined more than 50 years ago[94]. This term has mainly been used to define the dosage-sensitive regions of the genome that are associated with phenotypic abnormalities[27,95–98].

Our CNV map was developed from data in the Database of Genomic Variants (DGV). Over the past decade, DGV has provided a publicly accessible, comprehensive curated catalogue of CNVs and structural variations that are found in the genomes of control individuals from worldwide populations. The content from DGV is regularly incorporated into other large-scale genome reference databases.

The new CNV map from this study is presented as a standalone track in DGV.

High-resolution CNV maps have been constructed to study the impact of CNVs on complex human traits, including diseases[9,14,24,41]. The aim of this Analysis is to create a CNV map of the human genome for variations that are not associated with adverse phenotypes. We performed a meta-analysis for CNVs (deletions and duplications only) on publicly available high-quality data that were generated by systematic analyses of selected publications. Balanced structural variations, such as translocations and inversions (which are often benign[1]), and mitochondrial chromosome variants are ascertained differently and are not part of this Analysis. This is the first iteration of a genomic map that is intended for clinicians and scientists for diagnostic and research purposes. We documented the genomic losses (deletions) and gains (duplications) independently and also merged the two versions to generate a consensus map that represents all variations. We start with an overview of various methodologies and studies that are aimed at CNV detection, and we then describe our approach for generating the CNV map. We present the structure and properties of the CNV map, and discuss the impact of copy number variable regions (CNVRs) on genes and gene function.

Database of Genomic Variants
(DGV). A curated catalogue of copy number and structural variations in the human genomes of healthy control individuals.

Copy number variable regions
(CNVRs). Regions containing at least two copy number variations that overlap and that may have different breakpoints.

Next-generation sequencing
(NGS). A high-throughput DNA sequencing technology that typically generates shorter reads than Sanger sequencing-based methods and that can sequence billions of bases in parallel. NGS minimizes the need for fragment cloning.

Comparative genomic hybridization
(CGH). An array-based technique that interrogates the genome for signs of deletion or duplication in relation to a reference.

SNP-based arrays
Single-nucleotide polymorphism (SNP)-based microarrays that contain SNP probes to genotype human DNA at the single-base level. However, through dosage signals in adjacent regions, they can be used to recognize copy number variations.

CNV discoveries

Microarrays and next-generation sequencing (NGS) are now the primary approaches used for CNV detection. The first studies to assay CNVs primarily used microarrays with large-insert clones (known as bacterial artificial chromosomes (BACs))[7] and oligonucleotide arrays[8]. Subsequent microarrays have included both comparative genomic hybridization (CGH)[9,10,14,41] and SNP-based arrays. With the advent of NGS technologies, new tools can detect CNVs from these data, and the use of these approaches has been particularly widespread during the past few years[42–44].

Currently, no single discovery strategy can capture the entire spectrum of structural variations in the genome[11,16,43], and ascertainment depends largely on the platforms and algorithms used[40,45]. Numerous platforms have a lack of probes in regions of the genome that contain segmental duplications[46,47] — which predispose to structural variation — and are thus unable to recognize many CNVs that have been implicated in human disease (for example, the β-defensin cluster in psoriasis and Crohn's disease[48]). The array-based detection methods are suitable for studying quantitative variants[39]. Estimates of the total CNV content, as derived from earlier platforms such as BAC CGH, are typically inflated owing to low resolution and overestimation of the breakpoints[49]. These platforms miss many variants that are small. Deletions are generally easier to identify, and methods vary in their ability to detect deletions or duplications; for example, more duplications are missed by SNP-based array platforms and NGS approaches[16,40] than by array CGH. Phase I of the 1000 Genomes Project has probably missed variants that are private, or recurrent but rare[44].

The size distribution of variants called by different platforms varies[40]. Compared with array-based detection, studies that use sequencing-based methods, particularly NGS, call smaller variants and are biased towards the detection of deletions[16,43]. Sequencing-based approaches have both high sensitivity and the ability to provide accurate sequence-level breakpoint resolution, whereas array-based approaches have a limited resolution capacity. The highest-resolution array was used by Conrad *et al.*[14], in which the minimum threshold for CNV detection was 450 bp. Duplications are more readily detected by array CGH than sequencing, and this method is more sensitive for detecting small differences in copy number[40].

Earlier sequencing studies that aimed to generate a comprehensive survey of structural variants either used targeted sequencing for specific genomic regions or undertook whole-genome sequencing of one individual at a time. The 1000 Genomes Project[50,51] was among the first to perform whole-genome sequencing of multiple samples from many populations. With CNV data accumulated over the past decade, we now provide the first comprehensive meta-analysis.

Evaluation of CNV studies

Study selection. DGV has collected and curated 2,391,408 CNVs (comprising 202,431 CNVRs) that were discovered from 55 studies[12]. We selected a curated collection of CNVs from the entire DGV collection[12,36], all of which were from peer-reviewed publications. We selected recent high-resolution studies, which were designed to maximize discovery and minimize false discoveries. On the basis of the approach used, we split these studies into three primary categories: (i) studies based on sequencing (NGS and Sanger approaches), (ii) assays based on oligonucleotide or SNP arrays and (iii) studies using other methods (for example, fluorescence *in situ* hybridization (FISH), polymerase chain reaction (PCR), multiplex ligation-dependent probe amplification (MLPA) and optical mapping). Genome-wide assessment and accurate breakpoint resolution were the primary factors for inclusion of studies in the curated collection. Approaches based on oligonucleotide

© 2015 Macmillan Publishers Limited. All rights reserved

and SNP arrays were included if the array platform had a minimum of one million probes or if the array was a targeted or custom CNV assay[52]. We discarded data sets from lower-resolution arrays, and if a study used more than one approach or platform, we used only those that met the minimum criteria. We also included custom and targeted CNV assays, as these provided high accuracy and breakpoint resolution, even if the representation was not genome-wide. We discarded other approaches that are not listed above (including PCR, FISH and MLPA) owing to low sensitivity and low resolution; most approaches were not genome-wide, nor were the breakpoints of the variants delineated well enough for inclusion. From the original 55 studies, we thus derived a subset of 26 studies, which are listed in Supplementary information (Figure S1; Table S1). The majority of CNVs in the resulting data set were called by NGS platforms, followed by the array methods. There were three studies that used Sanger sequencing as the primary detection method (see Supplementary information (Table S1)).

Single-study evaluation. We further evaluated the CNVs from each study selected, particularly for false-positive and rare variants called (see Supplementary information (Figure S2)). The variants of each study were stratified on the basis of the intra-study frequency (all frequencies, 1% or greater and 5% or greater). We merged overlapping variants into consensus variants with outermost coordinates. Frequency-based stratification was used to identify studies enriched in singleton CNVs (that is, those detected in a single individual in a single study) or low-frequency variants, which was interpreted as a sign of lower specificity (see Supplementary information (Figures S3–S15)).

The variants in each study were first assessed based on distribution by frequency. We compared the apparent genomic variability when all variants were considered to that after the 1% and 5% frequency filters were applied to each study (see Supplementary information (Figure S3)). In addition, we curated different lists of genes and other genomic elements that were expected to be under negative selection against CNV (see Supplementary information (Tables S2–S5)). These lists included Mendelian disease implication, essentiality[53], the International Standards for Cytogenomic Arrays (ISCA) dosage sensitivity map[28], Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER)[27], and genomic indices of selective pressure and conservation (see Supplementary information (Table S4)). Significantly higher variability for constrained elements in one study than others was interpreted as a sign of lower specificity (see Supplementary information (Figures S7–S15; Table S6)). We observed that all variants in one study overlapped with 14.6% of coding nucleotides of deleted DECIPHER critical genes[27] (see Supplementary information (Figure S7; Table S6)). By applying the 1% filter, the overlap was reduced to 2.5%, and the 5% frequency filter had a minimal effect. We observed a similar pattern for the overlap of ISCA genes (see Supplementary information (Figure S12)) and Online Mendelian Inheritance In Man (OMIM; see Supplementary information (Figure S13)) genes with CNVs from two studies (see Supplementary information (Table S1)).

A total of 23 studies passed this quality control for the final curated collection of studies (see Supplementary information (Table S1)). This collection contains 2,057,368 variants — including 195,084 gains and 1,862,284 losses (see Supplementary information (Figures S16–S23)) — called from 2,647 subjects (1,404 male, 1,219 female and 24 of unknown sex) from diverse ethnicities (see Supplementary information (Figure S16)).

The CNV map

The aim of the CNV map is to document the variability of the human genome in healthy individuals from various populations. To capture the maximum extent of variability, we combined variants from different studies into a single map. Common variants would be detected in different individuals and ethnicities. Some variants were rare, private singletons or false discoveries; therefore, it is important to account for the presence of singletons. Different platforms have different degrees of accuracy for determining the beginning and end of the variants. For example, Sanger sequencing can detect CNV breakpoints with single-nucleotide resolution, whereas the array-based methods have lower precision. Thus, the coordinates of variants detected by different methods may not be identical.

We devised the following strategy to combine the high-confident variants from different studies into merged CNVRs. As CNVs called from different studies can estimate partially different CNV boundaries, we used a CNVR-clustering algorithm to identify sets of variants in which every possible variant pair had at least 50% reciprocal overlap[14]. This requirement ensured that structurally distinct CNVs were not merged at this stage. Clusters were then filtered on the basis of the number of distinct subjects that carry the variant and the number of distinct studies with at least one variant in the cluster. These two indices were used to filter the clusters that were eventually merged into the final consensus CNVRs[9] with the outermost coordinates. An alternative merging strategy that tests the average CNVR boundaries was evaluated to determine whether the boundaries were being significantly overextended by the original process; however, this was not the case (TABLE 1; see Supplementary information).

The filter based on the number of subjects excluded singletons, and variants that are supported by a larger number of subjects are less likely to be false positives. The filter based on the number of studies ensured the exclusion of potential study-specific artefacts. We evaluated 50 pairs of thresholds for the number of subjects and the number of studies (for example, at least one subject and one study, and at least two subjects and one study, and so on), which generated a set of CNVRs for each threshold pair (see Supplementary information (Figures S24–S31; Table S7)). The CNVRs that were recognized with higher stringency have support from a higher number of subjects and studies. The map supported only by at least one

Segmental duplications
(Also known as low-copy repeats). Highly homologous duplicated segments of DNA that are >1 kb in length and that show >90% sequence similarity.

International Standards for Cytogenomic Arrays
(ISCA). A consortium of clinical cytogeneticists who work together to standardize the use of array-based approaches in clinical genetic testing.

Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources
(DECIPHER). A database that documents phenotype information in patients with observed chromosome abnormalities and that aids the interpretation of genomic variants.

DECIPHER critical genes
Genes located in the critical regions that are associated with the 70 syndromes defined in Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER).

© 2015 Macmillan Publishers Limited. All rights reserved

Table 1 | **Summary of copy number variation in the genome based on the inclusive and stringent maps**

Copy number variation measures	All variants		Gains		Losses	
	Inclusive map	Stringent map	Inclusive map	Stringent map	Inclusive map	Stringent map
Total genome variable (%)*	9.5	4.8	3.9	2.3	7.5	3.6
Total genome variable (Mb)	273	136.6	111.5	64.7	215	102.4
Median interval length of CNVRs (bp)	981	1,237	3,334	9,741	956	1,137
Mean interval length of CNVRs (bp)	11,362	11,647	35,581	55,370	9,181	8,883
Number of CNVRs	24,032	11,732	3,132	1,169	23,438	11,530

CNVR, copy number variable region. *Numbers listed are based on the upper boundary size estimates of CNVRs. Average boundary sizes of the total genome include: all variants in the inclusive map (8.8%) and stringent map (4.1%), gains in the inclusive map (3.5%) and stringent map (1.9%), and losses in the inclusive map (6.9%) and stringent map (3.1%).

subject and one study (map with the stringency level 1) included all CNV and CNVRs (that is, every variant reported, including singletons and other variants) (see Supplementary information).

For further study, we selected thresholds of: (i) at least two subjects and one study for each variant (stringency level 2 (inclusive map); see Supplementary information (Table S9)) and (ii) at least two subjects and two studies (stringency level 12 (stringent map); see Supplementary information (Table S10)). Other stringency levels, additional information about the ethnic origin, number of studies and number of variants represented in each CNVR, and an alternative algorithm for construction of maps are presented in Supplementary information (Figures S32–S40; Table S8). This choice was based on previous estimates of genome copy number variability, overlap with reference variants from different platforms and minimal variability of highly constrained genomic elements (see Supplementary information). For example, DECIPHER critical genes did not overlap with CNVR gains, whereas CNVR losses affect 11.2% of the nucleotides of these genes at stringency level 1, and 2% or 0.5% for the inclusive and stringent maps, respectively (see Supplementary information (Figure S26)). The gene sets for essential genes (see Supplementary information (Figure S27)), OMIM genes (see Supplementary information (Figure S30)) and cancer genes (see Supplementary information (Figure S31)) overlapped with both gain and loss maps, and the overlap was significantly less in the inclusive map than at stringency level 1. Moreover, in the same comparison, the proportion of the genome that apparently contains CNVs decreased from 16.8% to 9.5% (a change of almost 43%), and then to 4.8% in the stringent map (see Supplementary information (Figure S24)). The marked change between the map with stringency level 1 and the inclusive map could be explained if a proportion of the singleton variants were false positives and the rest were rare events. Earlier studies have confirmed that a high proportion of the singleton CNVs deposited in the DGV are extremely rare or are due to false discovery[54], which justifies their omission from a CNV map. The consequences of some of these rare variants are likely to be of late onset and recognized as pathogenic at the time of study. By removing these singleton variants, we are confident that we have reduced the number of false positives and potentially rare pathogenic variants.

Essential genes
Orthologues of mouse genes for which homozygous loss-of-function mutations cause embryonic or neonatal lethality. They are necessary for cellular viability and organism development. They are evolutionarily more conserved than non-essential genes.

Properties of the CNV map

Our inclusive CNV map of the human genome comprises 3,132 CNVR gains and 23,438 CNVR losses (TABLE 1). At this level of stringency, 9.5% of the human genome seems to be involved with losses or gains (7.5% and 3.9%, respectively; see Supplementary information (Figures S41,S42). Earlier studies estimated this statistic as 12%[9] or 13%[55] for all CNVs, or less than 5% of the genome for the large-scale (>50 kb) variants[54]. Based on validated CNVs, Conrad *et al.*[14] reported that 3.7% of the human genome is copy number variable, which is slightly lower than the estimate from our current stringent map (4.8%), and this difference is probably due to the fact that only 40 samples of European and African ancestry were analysed. The higher percentage of apparent variability in our inclusive map could be explained by the fact that more samples from a wider range of ethnic backgrounds were included. Moreover, the present meta-analysis included variants that were derived from sequencing-based approaches, which capture smaller variants than those called in the array-based approaches[14].

The stringent map found variants that are more common, with fewer population-restricted variants. Its more stringent requirement for variant calling reduced the platform biases, which account for about 30% of discordant variant calls from the same method (platform and algorithm)[40]. However, it did remove a proportion of valid, rare variants from the map. The total genomic variability for this map was 4.8% (3.6% for losses and 2.3% for gains), which is 50% less than that of the inclusive map. The number of CNVs varies inversely with length of the variant region[50]. We found most CNVRs to be 300–3,000 bp, with losses slightly smaller than gains (TABLE 1; see Supplementary information (Figures S43,S44)).

CNVRs were found to be unevenly distributed in the genome[56] and among chromosomes (FIG. 1). From the inclusive map, the proportion of any given chromosome that is susceptible to CNV varied from 1.1% to 16.4% for gains and from 4.3% to 19.2% for losses. For gains, chromosome 22 and the Y chromosome showed the highest proportion of variability, followed by chromosomes 16, 9 and 15, which were all above the genome average (FIG. 1a). Chromosomes 3 and 18 showed the lowest proportion of variability in this map (FIG. 1a). On the stringent map of gains, a similar

© 2015 Macmillan Publishers Limited. All rights reserved

Copy number stable
(CNS). Pertaining to regions of the genome without any detected copy number variation in healthy individuals.

pattern was observed with the exception that the entire Y chromosome seemed to be copy number stable (CNS). This is because it is difficult to design probes for the Y chromosome owing to the highly repetitive nature of its sequence, and many commercially available arrays do not assay it[57]. For losses, the highest proportion of variable sequence was in chromosomes 19 and 22 and the Y chromosome on both inclusive and stringent maps (FIG. 1b). The lowest proportion was found in chromosomes 5 and 8 for the inclusive map, and chromosome 18 for the stringent map. This pattern was also seen for gains and losses that were combined together (FIG. 1c). We also observed that CNVs were unevenly distributed along the chromosomes. The pericentromeric regions had a higher proportion of CNVs[58], as did the subtelomeric regions[9,14,50,58] (FIG. 2), albeit with a lower than expected confidence assignment owing to the often complex nature of the sequence in these regions and the difficulty in properly assaying them.

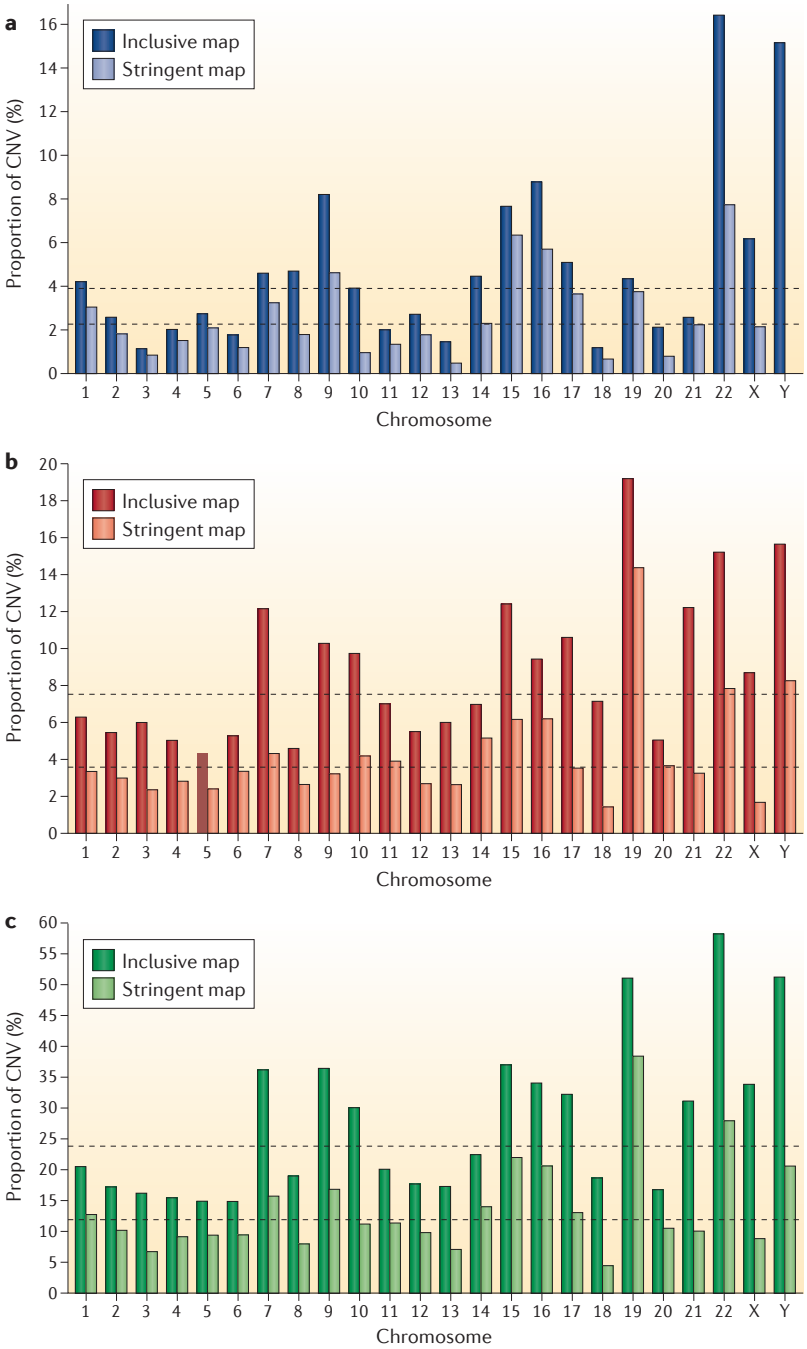


Figure 1 | **Proportion of each human chromosome that is copy number variable based on the inclusive and stringent maps.** Copy number variation (CNV) gains for each chromosome (part **a**), CNV losses for each chromosome (part **b**), and CNV gains and losses for each chromosome (part **c**) are shown. The horizontal dashed lines indicate the genome average for the inclusive map (upper line) and the stringent map (lower line). Assembly gaps, including centromeres and heterochromatic regions, were not included in the calculation of variable proportions. The *y* axes indicate the percentage of nucleotides of each chromosome that may be involved in CNV.

Functional impact of CNV

Different genomic elements are expected to be under different degrees of constraint for variation in copy number. We focused our assessment on the variability of gene exonic sequence. However, we also investigated the variability of transcripts[59–61], the coding DNA sequence and introns (see Supplementary information (Figures S45–S50; Tables S11–S13)). Genes were organized into different sets on the basis of: (i) the distinction between protein-coding and non-coding, (ii) their known phenotypic outcome in humans and mice, or Mendelian disease implication, and (iii) their genomic conservation and evolutionary constraint indices. We also assessed genomic sequences that were grouped on the basis of: (i) their conservation, (ii) their role in gene expression regulation (core promoters, enhancers and CpG islands) and (iii) their structural criteria (centromeres, telomeres, segmental duplications and repeats).

Losses contributed more than gains to the overall variation that was associated with various gene classes (FIG. 3), which is probably due to the underlying detection biases. The exons of all genes (as defined by RefSeq[62]) were more variable than the genome average (that is, the background) (FIG. 3). The exons of non-coding genes had the highest proportion of copy number variable sequence, which is higher than the exons of protein-coding genes. By contrast, exons of many of the constrained gene sets — particularly those associated with diseases (OMIM, ISCA and cancer) or other fitness-altering phenotypes — were less variable than the genome average (FIGS 3,4). Cancer-related genes[43,63] and particularly DECIPHER critical genes were under strong negative selection. The latter were the least affected by deletions. After DECIPHER critical gene exons, the least variable for both gains and losses were genes that were predicted by the genic intolerance score[64] and the haploinsufficiency index[65] (see Supplementary information). Exons in OMIM genes had relatively less overlap with CNVRs (gains or losses)[9,18,66,67] in both the inclusive map and the stringent map, and they were as constrained as the ISCA genes (FIGS 3,4). The variability of exons in essential genes was comparable to that of the OMIM genes. Exons in dominant genes showed no significant differences in variability compared

© 2015 Macmillan Publishers Limited. All rights reserved

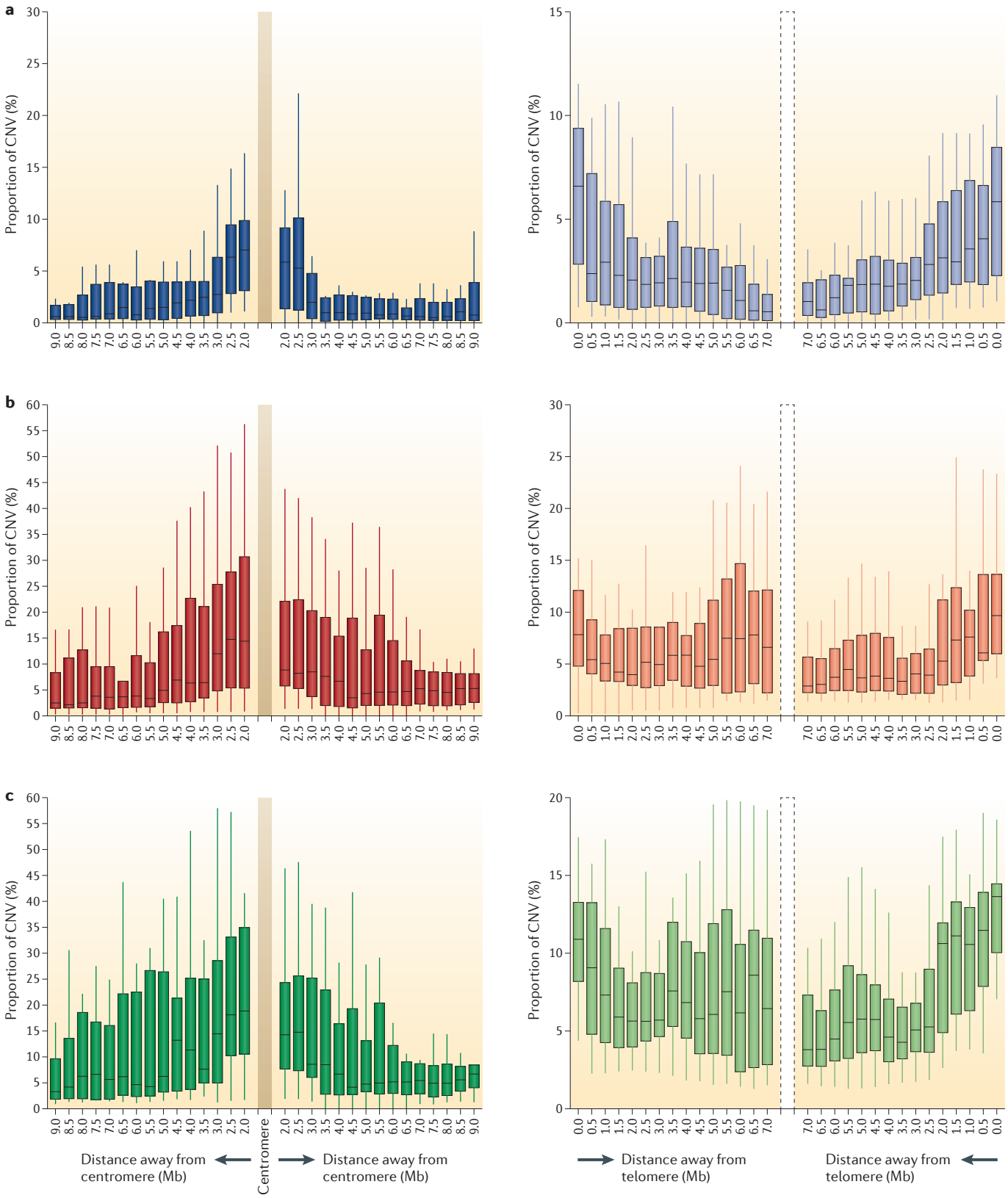


Figure 2 | **Distribution of copy number variable regions in pericentromeric and subtelomeric regions of human chromosomes.** Copy number variable region (CNVR) gains (part **a**), CNVR losses (part **b**), and CNVR gains and losses in the inclusive map (part **c**) are shown for pericentromeric regions (left panels) and subtelomeric regions (right panels). The *y* axes indicate the percentage of nucleotides in each window that may involve copy number variation. We used sliding windows of 5 Mb with steps of 0.5 Mb for this purpose (see Supplementary information (Figures S57,S58) for the stringent map).

© 2015 Macmillan Publishers Limited. All rights reserved

with those in recessive genes. An earlier study showed that dominant and recessive genes have relatively few CNVRs[43]. Long intergenic non-coding RNAs (lincRNAs) were enriched in CNVRs. Promoters were enriched in CNVRs (FIG. 3) compared with the entire genome[14,43]. This might be related to previously observed indel enrichment at both ends of genes[68]. Enhancers were impoverished in CNVRs (FIG. 3). PhastCons elements were found in relatively few CNVRs (FIG. 3), as were ultra-conserved elements, which are under strong selection against mutation[9,63,69]. CpG islands were more variable than the genome background[14].

Previous studies have indicated a positive correlation between CNVs and gene density[63,67,70]. Our analysis indicated that this is not universal. Deletions are biased away from certain categories of genes owing to negative selection[9,71], whereas duplications are less likely to be pathogenic and are often under positive selection, which drives evolution of many gene families[66]. In particular, the common deletions affect the protein-coding genes less than the rare variations[51].

Segmental duplications and CNVs

The role of segmental duplications in genomic stability and disease has previously been reviewed[72,73]. These elements are enriched at pericentromeric and subtelomeric regions, and their correlation with CNVRs has been shown[7,9,10,14] (see Supplementary information (Figures S51–S56)). Chromosomes vary in this respect — chromosome 19 showed the highest correlation with the evolutionarily older segmental duplications (with low sequence identity of around 90%), and chromosome 14 showed the highest correlation with the evolutionarily recent segmental duplications (with high sequence similarity of around 99%) (see Supplementary information (Figure S56)). Approximately 60% and 45% of nucleotides of segmental duplications were CNVs in the inclusive and stringent maps, respectively (see Supplementary information (Figure S51)). The substantial overlap and association between CNVs and segmental duplications highlights the dynamic nature of these regions, which blurs the distinction between CNV[7,9] and ancestral duplications that are presumed to be fixed in a specific population[72]. The segmental duplications with low sequence identity showed slightly higher genome-wide correlation with the CNVRs than the elements with high sequence similarity, which probably reflects a technical bias for better probe coverage in the low-identity repeats rather than a true mechanistic correlation[40] (see Supplementary information (Figure S56)).

Homozygous deleted genes

We generated a null CNV map of the human genome by compiling all CNVs in the DGV that were identified as being homozygously deleted and restricting them to regions that were shared with CNVR losses in the inclusive and stringent maps (see Supplementary information (Tables S14,S15)). These CNVRs were mildly, but not significantly, enriched in genes with paralogues compared with stringent loss CNVRs (see Supplementary information (Figures S59,S60)). The null CNVs were limited to 0.75% of the genome in the stringent map (see Supplementary information (Table S16)). From the inclusive map, we identified 107 human protein-coding genes (99 genes were from the stringent map), for which at least 85% of the exons were deleted in the null CNVRs (see Supplementary information (Table S18)). These genes would seem to be non-essential, as they could be missing from the genomes of apparently healthy individuals. These genes have more gene family members (see Supplementary information (Tables S17,S18)), which is common among duplicated genes[74], and are more prevalent in segmental duplications than all genes on average (see Supplementary information). The function of non-essential genes may be redundant (14 have paralogues), or they may be related to late-onset phenotypic consequences that do not substantially reduce the fitness. Some of the non-essential genes from our list are described in OMIM as age-related phenotypes (UDP glucuronosyltransferase 2 family polypeptide B17 (*UGT2B17*) for osteoporosis[75]), Rh-negative blood type (*RHD*[76]), delayed AIDS from HIV-1 infection (*KIR3DL1* (killer cell immunoglobulin-like receptor, three domains, long cytoplasmic tail, 1)[77]), susceptibility to multiple sclerosis (the major histocompatibility complex gene *HLA-DRB1* (REF. 78)), autoimmune disease (*HLA-DQA1* (REF. 79) for coeliac disease), and placental function and fetal wellbeing (pregnancy-specific beta-1-glycoprotein 1 (*PSG1*)[80]). Interestingly, our list of non-essential genes is non-overlapping with the loss-of-function genes reported by MacArthur *et al.*[81].

Figure 3 | **Copy number variations that involve regulatory elements or exons of specified gene lists.** From the inclusive maps, the proportion of nucleotides from each category involved in copy number variation (CNV) gains (part **a**), CNV losses (part **b**), and CNV gains and losses (part **c**) is shown. The horizontal dashed lines indicate the genome average for the inclusive map. The *y* axes indicate the percentage of total nucleotides of exons of each gene set or genomic element that may contain CNVs. Numbers in parentheses show the total number of genes included in the analysis for each set, and the percentages in parentheses are the percentages of the sequenced genome of the element. We considered a result significant when the one-sided empirical *P* value (calculated by a 10,000 randomized data set) was less than 0.01 (no mark), whereas borderline significance was defined as empirical *P* value between 0.01 and 0.1 (marked with *), and *P* values greater than 0.1 (marked with **) are considered insignificant. The permutation was not performed for cancer and Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER) critical genes for CNV gains, as they did not overlap with any CNVs. See Supplementary information (Tables S2–S5,S8). CGD, Clinical Genomic Database; GI, genic intolerance; HI, haploinsufficiency; ISCA, International Standards for Cytogenomic Arrays; lincRNA, long intergenic non-coding RNA; OMIM, Online Mendelian Inheritance In Man.

Genic intolerance score
An index of intolerance to rare, non-synonymous variation.

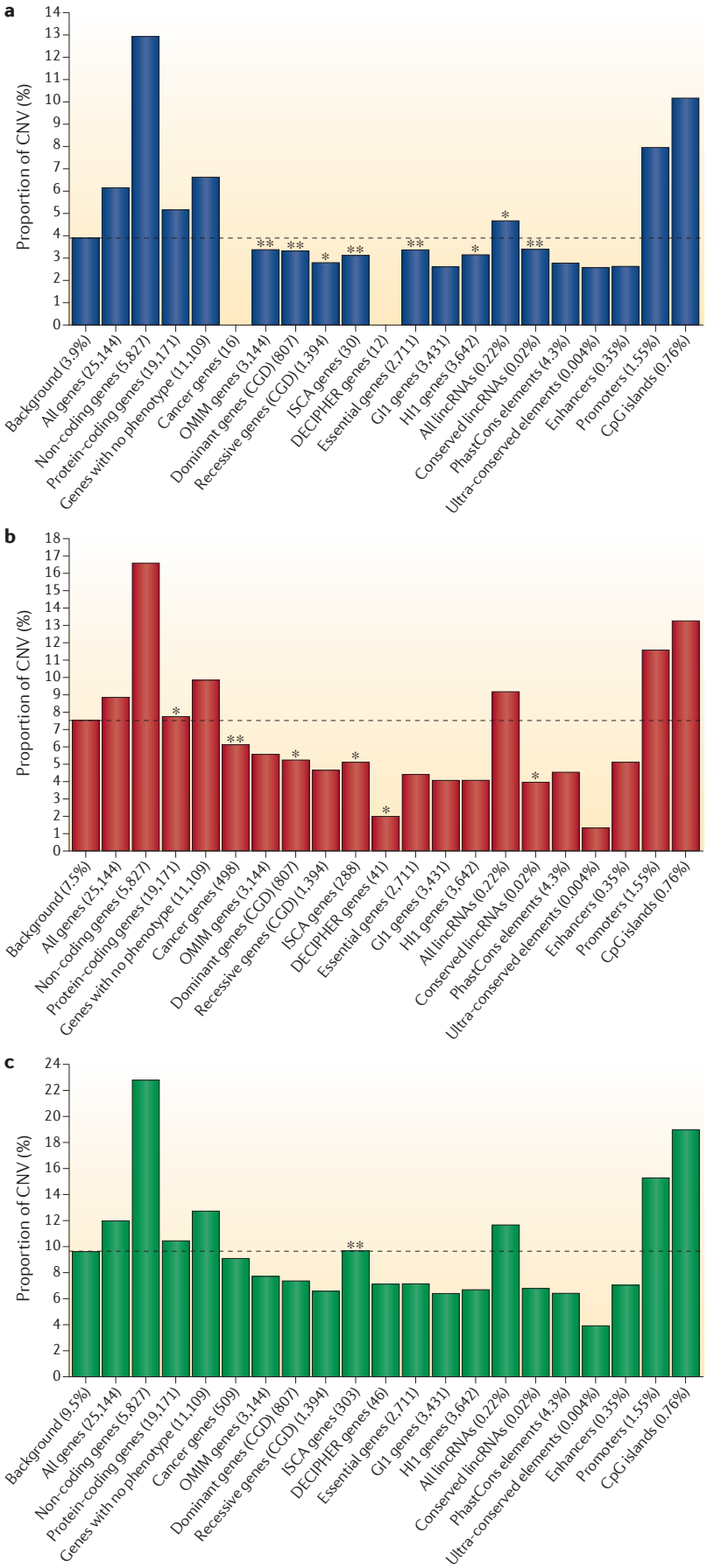
Haploinsufficiency
Reduction in the amount of gene product owing to functional loss of an allele that leads to an abnormal or a disease state.

Long intergenic non-coding RNAs
(lincRNAs). Non-coding RNAs that are thought to be key regulators of diverse cellular processes. Their expression seems to be more tissue-specific than that of coding genes.

PhastCons elements
Evolutionarily conserved elements that were identified by modelling substitution rates in multiple genome alignments.

Ultra-conserved elements
Regions of DNA that are conserved across mammalian genomes and that mostly consist of non-protein-coding regions (that is, regions with little or no evolutionary changes since the divergence of mammals and birds).

© 2015 Macmillan Publishers Limited. All rights reserved



Functional enrichment analysis

We tested Gene Ontology[82] and pathway-based gene sets for enrichment in variable and stable genes. The enriched gene sets are similar for losses (FIG. 5) and gains (see Supplementary information (Figure S61)) in the inclusive and stringent maps; the stringent map showed a more significant enrichment, particularly for functional gene sets in stable regions. Gene-set enrichment results were visualized as a network in order to group highly overlapping gene sets into functional clusters[83]. Xenobiotic, drug and steroid metabolism, immune response[84–86], olfactory receptors[84–87], starch and sucrose metabolism, pregnancy-specific adhesion molecules, endoplasmic reticulum, vesicle and Golgi apparatus were the functions enriched in the CNVRs in both inclusive and stringent maps (FIG. 5; see Supplementary information (Figure S61)). Protein phosphorylation, signal transduction, protein degradation, transcriptional machinery and regulation, intracellular transport, organ and system development, cell differentiation, cell cycle and nucleus compartment were the functions enriched in the CNS regions of the genome. Genes that belong to each functional category are shown in Supplementary information (Table S19). Gene families with members that overlap with segmental duplications were enriched for CNVs; for example, the human salivary amylase genes, which encode enzymes that catalyse the digestion of dietary starch, showed extensive CNV[20].

Discussion and future work

Our CNV map demonstrates and quantitates the extent of heterogeneity within the human genome in terms of CNV. The human genome is continually shaped by ongoing modifications of its structure that are reflected in both healthy individuals and those with disease. The presence of CNVRs in and among functional sequences — including those associated with disease — makes it challenging to discriminate between benign and pathogenic CNVs. The CNS regions of the genome in the inclusive map (that is, of control genomes) may be the most dosage-sensitive regions of the genome, in which CNVs are likely to be associated with disease or reduced fitness. A CNV found in the inclusive and stringent maps that overlaps with a disease-associated gene may reflect reduced penetrance of the variant with respect to the disease phenotype[15,19,32].

This map is particularly valuable, as the highest-resolution sequencing or array-based studies still have limitations that can be overcome by pooling resources from multiple studies. For example, Conrad *et al.*[14] examined the genome with the highest-resolution array-based platform but was still unable to reliably detect variants smaller than 450 bp. Moreover, this study included a limited number of samples for CNV discovery, which captured only two different populations. By contrast, the 1000 Genomes Project studies[44,50] analysed a larger number of samples from a wider range of populations but used low-coverage sequencing, which limited CNV detection; therefore, the majority of CNVs discovered were smaller than 400 bp, and both duplications and larger variations

© 2015 Macmillan Publishers Limited. All rights reserved

were under-represented. By combining CNVs from 23 studies (including the latter three[14,44,50] in the CNV map), we are able to overcome issues seen in individual studies. Although this meta-analysis has generated the most complete map (to date) of CNV in the human genome, variants of 50–500 bp need further characterization with single-nucleotide resolution and by sequence-based approaches, such as NGS[16]. This CNV map can be used as a standard for comparison in studies of the genetic basis of phenotypic variation. It will be a tool for using CNV genome annotations to discriminate pathogenic CNVs from benign CNVs, and for identifying those CNVs that are associated with evolutionary fitness in the population. Moreover, this map can be used to guide the development of CNV-specific genotyping assays.

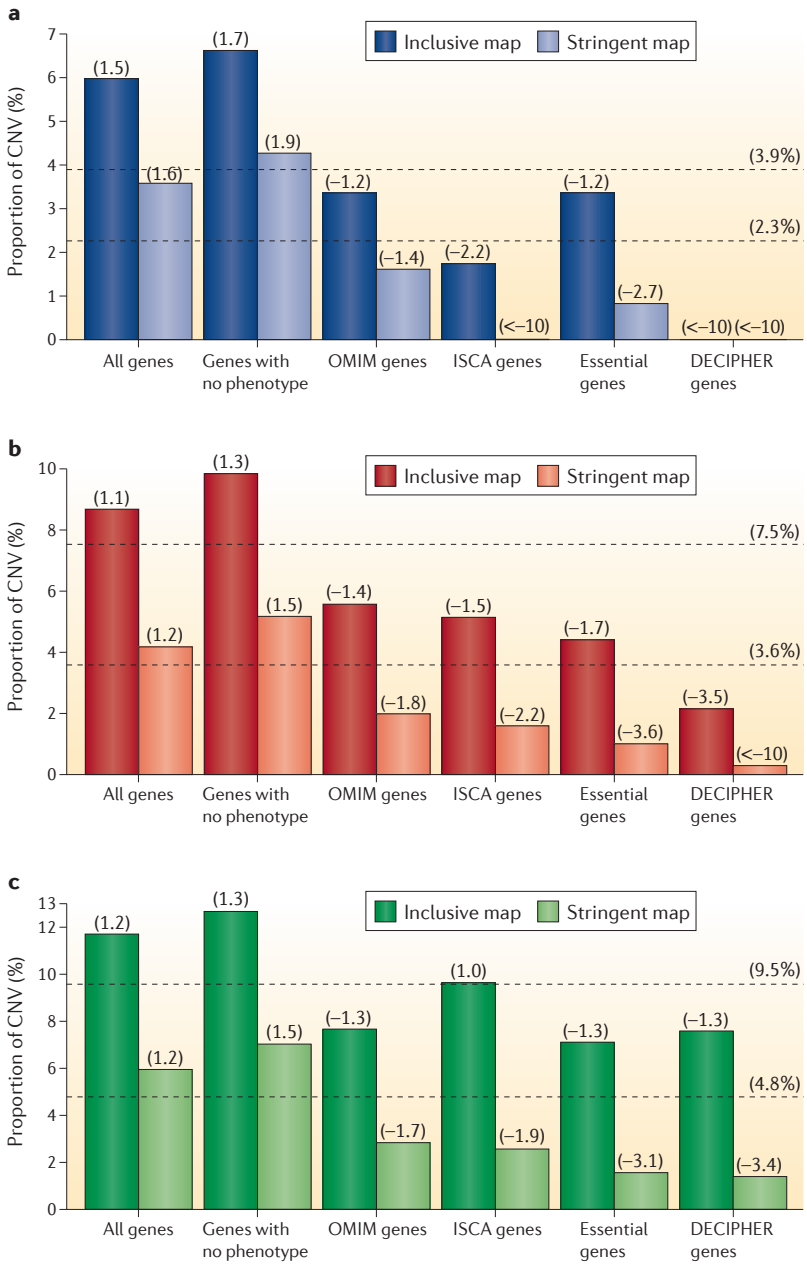


Figure 4 | **Copy number variations that involve genes with or without disease association.** Copy number variation (CNV) gains (part **a**), CNV losses (part **b**), and CNV gains and losses (part **c**) are shown. The *y* axes indicate the percentage of exonic nucleotides for each gene set that may be involved in CNVs. The horizontal dashed lines indicate the genome average that is variable for the inclusive map (upper line) and the stringent map (lower line). The number above each bar is the ratio of variability in designated exons to that of the genome background. The negative sign indicates less variability than the genome background. DECIPHER, Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources; ISCA, International Standards for Cytogenomic Arrays; OMIM, Online Mendelian Inheritance In Man.

Clinical applications of the CNV map. This map can be used as a tool in the investigation of CNVs for medical applications[9,15,32]. To assess the clinical importance of a CNV found in a case subject, the criteria could include: (i) whether it is found among the CNVRs of the CNV map and (ii) whether it overlaps with medically relevant genes. We manually curated a list of 935 medically relevant genes that have been shown to involve CNVs (see Supplementary information (Table S5 and References)). This comprehensive list of disease-associated genes was derived from 407 peer-reviewed publications and includes the ISCA dosage-sensitive map genes, the Baylor College of Medicine 105K oligonucleotide array disorder list, the DECIPHER critical genes and cancer genes. A CNV found in a patient that overlaps with any of these genes while not being found in the CNV map would suggest medical importance; in particular, CNVs that are associated with the more stringent gene sets, such as DECIPHER critical genes and embryonically lethal genes, would have a higher likelihood of pathogenicity. Comparisons between inclusive and stringent maps could allow consideration of conditional phenotypes, such as age-related phenotypes. The genes in which mutations cause late-onset disorders are less conserved than those that are associated with earlier-onset disorders (which have a greater effect on genetic fitness)[88] and might be better filtered by the stringent map.

There are no solid boundaries between what we recognize as traits and diseases, nor between variants that are classified as benign or neutral and variants that are classified as predisposing risk factors or disease-related. Genomic background, including ethnicity, can influence the thresholds. For example, the low copy number of the amylase gene is a disadvantage in Asian populations, particularly in the Japanese, but it has no effect in African populations[20].

Future research direction. This CNV map has been generated with certain constraints. First, information about CNVs in healthy individuals is incomplete. The CNVRs of the genome defined in the current inclusive map are those detected in at least two different independent samples among 2,647 subjects. We will continue to generate maps as new CNV data sets become available but, for Europeans, the CNV map may approach saturation because our testing of 3,000 additional samples from this group[24] only slightly increased the total genomic variability (from 9.5% to 9.7%) (see Supplementary information (Figures S41,S42)).

© 2015 Macmillan Publishers Limited. All rights reserved

All singleton findings were assigned to the CNS regions of the map (see Supplementary information (Figure S24)); however, some of these variants may indeed be true variants because low-frequency variants can show an increased level of population differentiation[51,54,89,90] and because CNVs segregate at different frequencies in different populations[9,14,20]. For example, the CNVs in regions of segmental duplication are more likely to be population stratified than CNVs in regions of unique sequence[45]. Campbell *et al.*[45] showed that CNVs involving bitter taste receptor genes (*TAS2R46* and *TAS2R48*) are more prevalent in non-Africans than in Africans. These authors also found that a CNV overlapping the occludin (*OCLN*) gene is at a lower copy number in East Asian individuals than in African individuals. The chemokine (C-C motif) ligand 3-like 1 (*CCL3L1*) gene is one of the most differentiated exonic CNVs[14]. There can also be sex-influenced modifiers[91] and CNVs that express phenotypes at different ages[92,93], and many other complex phenomena, such as pleiotropy, could be involved[2]. These points highlight the breadth of phenotypic effects and that some of the currently excluded singleton variants may be included at a later time when additional samples are available. Moreover, larger and broader sample collections would enable us to generate population-specific CNV maps, which could capture geographically segregated variants[89], and also to generate maps that include only rare variants. The identification of population-specific CNVs could be important for studying the role of positive and negative selection on these variants, and how they relate to diseases or beneficial traits in specific populations[13,19]. The interaction between CNV size and genomic feature variability needs to be better investigated (see Supplementary information (Figures S62–S66)). For example, certain genomic features may be prone to large gains that encompass their full length, but they are constrained for small gains that alter their structure.

We also note that the copy number level for each CNVR is not specified in the current map, as many of the underlying data sets were not genotyped to capture this information. With improved algorithms and the transition to genome sequencing approaches, we will have a more comprehensive catalogue of genotyped variants that enable us to accurately report the population frequencies at each CNVR.

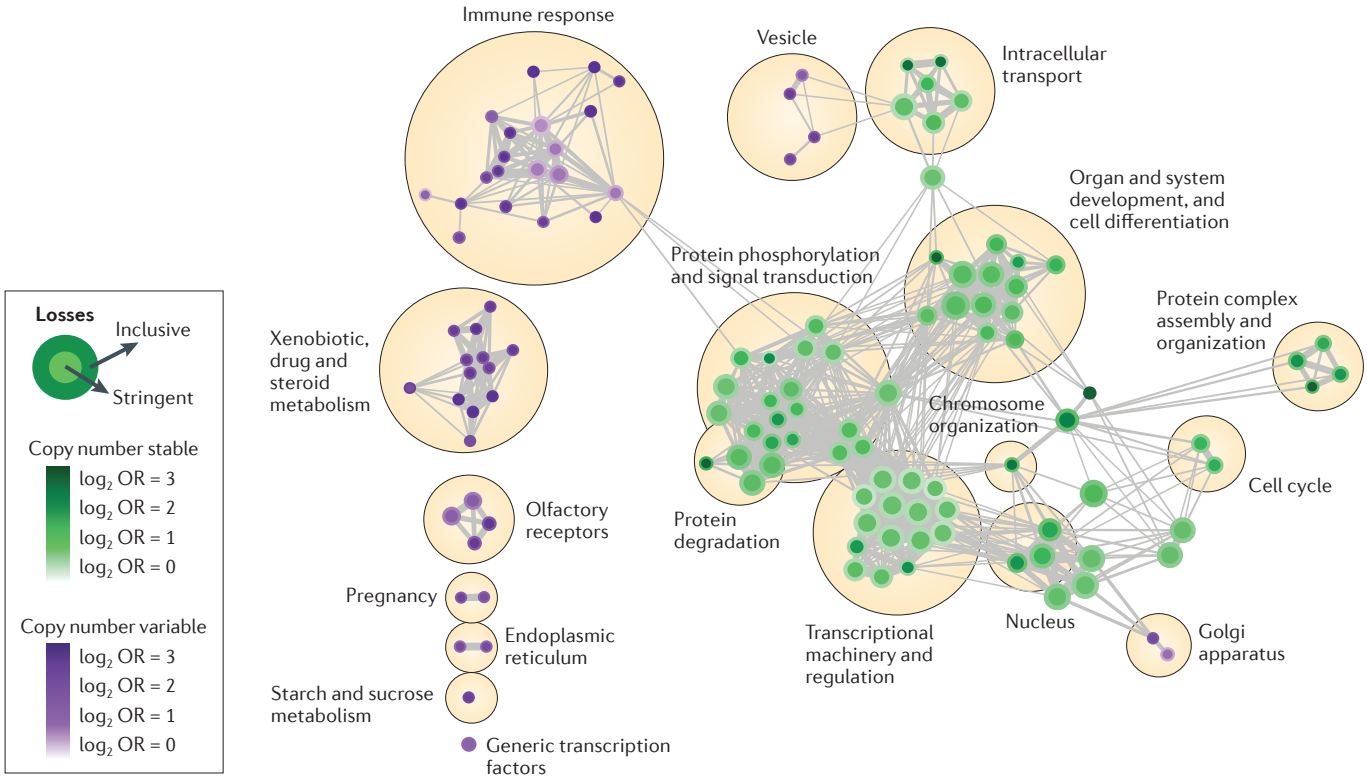


Figure 5 | **Gene function enrichment map for the inclusive and stringent maps of copy number variation losses.** The map was generated using the Enrichment Map Cytoscape plug-in[99]. Node size is proportional to the size of the functional gene set, and the edge thickness is proportional to the overlap between the gene sets. We show only gene sets that are enriched in copy number variable or stable regions with a false discovery rate of ≤10%; only gene sets with a size between 50 and 2,500 genes were analysed. Biologically related gene sets tend to form clusters, which are manually identified and labelled with the appropriate Gene Ontology terms. Green indicates enrichment in the copy number stable genes, and purple indicates the enrichment in copy number variable genes. The border of each circle corresponds to the inclusive map, and the centre corresponds to the stringent map. The intensity of the node colour is proportional to the enrichment score, which is shown separately for the copy number stable and variable regions of the genome. See Supplementary information (Figure S61 for gains; Table S19). OR, odds ratio.

© 2015 Macmillan Publishers Limited. All rights reserved

The DGV curates published variants on an ongoing basis and has become the primary repository for CNVs to be used both in research studies and for the annotation of variants in clinical settings. The DGV documents variants from presumably healthy controls, but the amount of phenotypic documentation available is often limited. For example, a control sample in a cancer study may not have been assessed for health status with respect to blood pressure or other health problems. Moreover, health is not static, and the status of a research participant could change. The DGV does not exclude variations with late-onset phenotypes; therefore, caution is needed in its use for prediction of health outcomes[19]. By contrast, databases such as DECIPHER aim to catalogue the highly penetrant variants that cause overt phenotypes. In our previous work[9,19], we noted some overlap of content in control and disease databases, such as DGV and DECIPHER, respectively, and such overlap requires an explanation either by improved resolution of the CNVRs and more-detailed phenotyping or, more generally, through a better understanding of the phenomenon of penetrance. Ultimately, our latest CNV map of the human genome provides a reference for understanding the effect of genomic imbalance on evolution, health and disease.

1. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nature Rev. Genet.* **7**, 85–97 (2006).
This is a comprehensive review of CNV and structural variation that suggests nomenclature for the newly emerging field.
2. Beckmann, J. S., Estivill, X. & Antonarakis, S. E. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nature Rev. Genet.* **8**, 639–646 (2007).
3. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nature Rev. Genet.* **10**, 551–564 (2009).
4. Jacobs, P. A., Browne, C., Gregson, N., Joyce, C. & White, H. Estimates of the frequency of chromosome abnormalities detectable in unselected newborns using moderate levels of banding. *J. Med. Genet.* **29**, 103–108 (1992).
5. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
6. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
7. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36**, 949–951 (2004).
8. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
References 7 and 8 were the first to describe the genome-wide presence of CNVs in the human genome, which provided the basis for the generation of a CNV map. Reference 7 also shows that segmental duplications can be copy number variable and introduces the idea of a public CNV database.
9. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
This is the first comprehensive study to investigate CNV (larger than 50 kb in size) in multiple samples in order to study its impact on population genetics and genome dynamics.
10. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nature Genet.* **37**, 727–732 (2005).
11. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
This paper reports the first personal genome sequence of an identified individual that was generated using Sanger sequencing and identifies hundreds of thousands of smaller indels in human DNA.
12. MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986–D992 (2014).
13. Conrad, D. F. & Hurles, M. E. The population genetics of structural variation. *Nature Genet.* **39**, S30–S36 (2007).
14. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
This paper provides a second-generation CNV (larger than 450 bp in size) map that was constructed using high-resolution oligonucleotide microarrays, which represents a 'gold-standard' data set for comparisons.
15. Lee, C. & Scherer, S. W. The clinical context of copy number variation in the human genome. *Expert Rev. Mol. Med.* **12**, e8 (2010).
16. Pang, A. W., Macdonald, J. R., Yuen, R. K., Hayes, V. M. & Scherer, S. W. Performance of high-throughput sequencing for the discovery of genetic variation across the complete size spectrum. *G3 (Bethesda)* **4**, 63–65 (2014).
17. Lupski, J. R. Genomic rearrangements and sporadic disease. *Nature Genet.* **39**, S43–S47 (2007).
18. Hurles, M. E., Dermitzakis, E. T. & Tyler-Smith, C. The functional impact of structural variation in humans. *Trends Genet.* **24**, 238–245 (2008).
19. Buchanan, J. A. & Scherer, S. W. Contemplating effects of genomic structural variation. *Genet. Med.* **10**, 639–647 (2008).
20. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nature Genet.* **39**, 1256–1260 (2007).
21. Pinto, D. *et al.* Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* **94**, 677–694 (2014).
22. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
23. Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148**, 1223–1241 (2012).
24. Wellcome Trust Case Control Consortium *et al.* Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713–720 (2010).
25. Cantsilieris, S. & White, S. J. Correlating multiallelic copy number polymorphisms with disease susceptibility. *Hum. Mutat.* **34**, 1–13 (2013).
26. Jacquemont, S. *et al.* Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* **478**, 97–102 (2011).
27. Firth, H. V. *et al.* DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
28. Riggs, E. R. *et al.* Towards an evidence-based process for the clinical interpretation of copy number variation. *Clin. Genet.* **81**, 403–412 (2012).
This paper provides an evidence-based framework for clinical evaluation, which supports or refutes the dosage sensitivity for individual genes and regions.
29. de Vries, B. B. *et al.* Diagnostic genome profiling in mental retardation. *Am. J. Hum. Genet.* **77**, 606–616 (2005).
30. Lupski, J. R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14**, 417–422 (1998).
31. Nuttle, X., Itsara, A., Shendure, J. & Eichler, E. E. Resolving genomic disorder-associated breakpoints within segmental DNA duplications using massively parallel sequencing. *Nature Protoc.* **9**, 1496–1513 (2014).
32. Lee, C., Iafrate, A. J. & Brothman, A. R. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nature Genet.* **39**, S48–S54 (2007).
33. Choy, K. W., Setlur, S. R., Lee, C. & Lau, T. K. The impact of human copy number variation on a new era of genetic testing. *BJOG* **117**, 391–398 (2010).
34. Miller, D. T. *et al.* Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am. J. Hum. Genet.* **86**, 749–764 (2010).
35. de Leeuw, N. *et al.* Diagnostic interpretation of array data using public databases and internet sources. *Hum. Mutat.* **33**, 930–940 (2012).
36. Church, D. M. *et al.* Public data archives for genomic structural variation. *Nature Genet.* **42**, 813–814 (2010).
37. Campbell, I. M. *et al.* Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. *Am. J. Hum. Genet.* **95**, 173–182 (2014).
38. Forsberg, L. A. *et al.* Age-related somatic structural changes in the nuclear genome of human blood cells. *Am. J. Hum. Genet.* **90**, 217–228 (2012).
39. Scherer, S. W. *et al.* Challenges and standards in integrating surveys of structural variation. *Nature Genet.* **39**, S7–S15 (2007).
This paper highlights the challenges in the characterization and documentation of structural variation. The authors propose recommendations that can be adopted for standardizing the presentation of CNVs and structural variations.
40. Pinto, D. *et al.* Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotech.* **29**, 512–520 (2011).
41. Park, H. *et al.* Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nature Genet.* **42**, 400–405 (2010).
42. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
43. Pang, A. W. *et al.* Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* **11**, R52 (2010).
44. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
Small deletions in the human genome have been identified in a large sample collection by the 1000 Genomes Project.
45. Campbell, C. D. *et al.* Population-genetic properties of differentiated human copy-number polymorphisms. *Am. J. Hum. Genet.* **88**, 317–332 (2011).
46. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genet.* **41**, 1061–1067 (2009).
47. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
48. Hollox, E. J. *et al.* Psoriasis is associated with increased β-defensin genomic copy number. *Nature Genet.* **40**, 23–25 (2008).
49. de Smith, A. J. *et al.* Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum. Mol. Genet.* **16**, 2783–2794 (2007).
50. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
51. Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
52. Itsara, A. *et al.* Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.* **84**, 148–161 (2009).

© 2015 Macmillan Publishers Limited. All rights reserved

53. Ye, Y. N., Hua, Z. G., Huang, J., Rao, N. & Guo, F. B. CEG: a database of essential gene clusters. *BMC Genomics* **14**, 769 (2013).
54. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genet.* **40**, 1166–1174 (2008).
55. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61**, 437–455 (2010).
56. Makino, T., McLysaght, A. & Kawata, M. Genome-wide deserts for copy number variation in vertebrates. *Nature Commun.* **4**, 2283 (2013).
57. Yuen, R. K. *et al.* Development of a high-resolution Y-chromosome microarray for improved male infertility diagnosis. *Fertil. Steril.* **101**, 1079–1085.e3 (2014).
58. Wong, L. P. *et al.* Deep whole-genome sequencing of 100 southeast Asian Malays. *Am. J. Hum. Genet.* **92**, 52–66 (2013).
59. ENCODE Project Consortium *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
60. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
61. Gerstein, M. B. *et al.* Comparative analysis of the transcriptome across distant species. *Nature* **512**, 445–448 (2014).
62. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).
63. Johansson, A. C. & Feuk, L. Characterization of copy number-stable regions in the human genome. *Hum. Mutat.* **32**, 947–955 (2011).
This paper defines a list of dosage-sensitive regions of the human genome and correlates them with the rare and *de novo* CNVs identified in patients with intellectual disability or autism.
64. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
65. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* **6**, e1001154 (2010).
This important study uses functional features to determine haploinsufficiency scores for human protein-coding genes and their likelihood to be involved in disease.
66. Hurles, M. Gene duplication: the genomic trade in spare parts. *PLoS Biol.* **2**, E206 (2004).
67. Nguyen, D. Q., Webber, C. & Ponting, C. P. Bias of selection on human copy-number variants. *PLoS Genet.* **2**, e20 (2006).
68. Ng, P. C. *et al.* Genetic variation in an individual human exome. *PLoS Genet.* **4**, e1000160 (2008).
69. Katzman, S. *et al.* Human genome ultraconserved elements are ultraselected. *Science* **317**, 915 (2007).
70. Nguyen, D. Q. *et al.* Reduced purifying selection prevails over positive selection in human copy number variant evolution. *Genome Res.* **18**, 1711–1723 (2008).
71. Uddin, M. *et al.* Brain-expressed exons under purifying selection are enriched for *de novo* mutations in autism spectrum disorder. *Nature Genet.* **46**, 742–747 (2014).
72. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
This paper provides the first map of segmental duplications in the human genome, which includes an analysis of their relationship to genes and genetic diseases.
73. Mefford, H. C. & Eichler, E. E. Duplication hotspots, rare genomic disorders, and common disease. *Curr. Opin. Genet. Dev.* **19**, 196–204 (2009).
74. Woods, S. *et al.* Duplication and retention biases of essential and non-essential genes revealed by systematic knockdown analyses. *PLoS Genet.* **9**, e1003330 (2013).
75. Yang, T. L. *et al.* Genome-wide copy-number-variation study identified a susceptibility gene, *UGT2B17*, for osteoporosis. *Am. J. Hum. Genet.* **83**, 663–674 (2008).
76. Szulman, A., Nardozza, L. M., Barreto, J. A., Araujo Junior, E. & Moron, A. F. Investigation of pseudogenes *RHD*Ψ and *RHD-CE-D* hybrid gene in D-negative blood donors by the real time PCR method. *Transfus Apher Sci.* **47**, 289–293 (2012).
77. Jiang, Y. *et al.* KIR3DS1/L1 and HLA-Bw4-80I are associated with HIV disease progression among HIV typical progressors and long-term nonprogressors. *BMC Infect. Dis.* **13**, 405 (2013).
78. International Multiple Sclerosis Genetics Consortium *et al.* Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.* **357**, 851–862 (2007).
79. Hadithi, M. *et al.* Accuracy of serologic tests and HLA-DQ typing for diagnosing celiac disease. *Ann. Intern. Med.* **147**, 294–302 (2007).
80. Bartels, I. & Lindemann, A. Maternal levels of pregnancy-specific β_1-glycoprotein (SP-1) are elevated in pregnancies affected by Down's syndrome. *Hum. Genet.* **80**, 46–48 (1988).
81. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
82. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
83. Aguiar, D., Halldorsson, B. V., Morrow, E. M. & Istrail, S. DELISHUS: an efficient and exact algorithm for genome-wide detection of deletion polymorphism in autism. *Bioinformatics* **28**, i154–i162 (2012).
84. Cooper, G. M., Nickerson, D. A. & Eichler, E. E. Mutational and selective effects on copy-number variants in the human genome. *Nature Genet.* **39**, S22–S29 (2007).
85. McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009).
86. Iskow, R. C., Gokcumen, O. & Lee, C. Exploring the role of copy number variants in human adaptation. *Trends Genet.* **28**, 245–257 (2012).
87. Hasin-Brumshtein, Y., Lancet, D. & Olender, T. Human olfaction: from genomic variation to phenotypic diversity. *Trends Genet.* **25**, 178–184 (2009).
88. Blekhman, R. *et al.* Natural selection on genes that underlie human disease susceptibility. *Curr. Biol.* **18**, 883–889 (2008).
89. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
90. Nelson, M. R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100–104 (2012).
91. Lionel, A. C. *et al.* Disruption of the *ASTN2*/*TRIM32* locus at 9q33.1 is a risk factor in males for autism spectrum disorders, ADHD and other neurodevelopmental phenotypes. *Hum. Mol. Genet.* **23**, 2752–2768 (2014).
92. Kuningas, M. *et al.* Large common deletions associate with mortality at old age. *Hum. Mol. Genet.* **20**, 4290–4296 (2011).
93. Costain, G. *et al.* Adult neuropsychiatric expression and familial segregation of 2q13 duplications. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **165B**, 337–344 (2014).
94. Castoldi, G. L., Ricci, N., Punturieri, E. & Bosi, L. Chromosomal imbalance in plasmacytoma. *Lancet* **1**, 829 (1963).
95. Cattanach, B. M. Snaker: a dominant abnormality caused by chromosomal imbalance. *Z. Vererbungsl.* **96**, 275–284 (1965).
96. Sparkes, R. S. Genetic abnormalities: the consequences of chromosome imbalance. *Science* **235**, 916a (1987).
97. Epstein, C. J. The consequences of chromosome imbalance. *Am. J. Med. Genet. Suppl.* **7**, 31–37 (1990).
98. Bejjani, B. A., Theisen, A. P., Ballif, B. C. & Shaffer, L. G. Array-based comparative genomic hybridization in clinical diagnosis. *Expert Rev. Mol. Diagn.* **5**, 421–429 (2005).
99. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* **5**, e13984 (2010).

Acknowledgements
The authors thank R. Ziman and G. Pellecchia for computational support, as well as J. Buchanan, J. Stavropoulos, C. Marshall, R. Yuen, B. Thiruvahindrapuram, M. Uddin, M. Mohammed and L. Feuk for discussions. They thank The Centre for Applied Genomics Science and Technology Innovation Centre (funded by Genome Canada and the Ontario Genomics Institute) for computational support. The Database of Genomic Variants and our research are supported by grants from Genome Canada, the Canada Foundation of Innovation, the Canadian Institute for Advanced Research, the government of Ontario, the Canadian Institutes of Health Research (CIHR), The Hospital for Sick Children, and the University of Toronto McLaughlin Centre. S.W.S. holds the GlaxoSmithKline–CIHR Endowed Chair in Genome Sciences at The Hospital for Sick Children and the University of Toronto.

Competing interests statement
The authors declare no competing interests.

SUPPLEMENTARY INFORMATION
See online article: Supplementary information | Table S1 | Table S3 | Table S4 | Table S5 | Table S6 | Table S7 | Table S8 | Table S9 | Table S10 | Table S11 | Table S13 | Table S15 | Table S17 | Table S18 | Table S19

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

© 2015 Macmillan Publishers Limited. All rights reserved