Patterns of genic intolerance of rare copy number variation in 59,898 human exomes

HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2017 February 17.

Published in final edited form as:

Nat Genet. 2016 October ; 48(10): 1107–1111. doi:10.1038/ng.3638.

Patterns of genic intolerance of rare copy number variation in 59,898 human exomes

Douglas M. Ruderfer[1,2,3], **Tymor Hamamsy**[1], **Monkol Lek**[3,4], **Konrad J. Karczewski**[3,4], **David Kavanagh**[1,2], **Kaitlin E. Samocha**[3,4], **Exome Aggregation Consortium**[5], **Mark J. Daly**[3,4], **Daniel G. MacArthur**[3,4], **Menachem Fromer**[1,2,3,4,*], and **Shaun M. Purcell**[1,2,3,4,6,*]

[1]Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA

[2]Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA

[3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA

[4]Analytic and Translational Genetics Unit, Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

[6]Department of Psychiatry, Brigham & Women's Hospital, Harvard Medical School, Boston, MA, 02115, USA

Abstract

Copy number variation (CNV) impacting protein-coding genes contributes significantly to human diversity and disease. Here we characterized the rates and properties of rare genic CNV (<0.5% frequency) in exome-sequencing data from nearly 60,000 individuals in the Exome Aggregation Consortium (ExAC). On average, individuals possessed 0.81 deleted and 1.75 duplicated genes, and most (70%) carried at least one rare genic CNV. For every gene, we empirically estimated an index of relative *intolerance* to CNVs that demonstrated moderate correlation with measures of genic constraint based on single-nucleotide variation (SNV) and was independently correlated with measures of evolutionary conservation. For individuals with schizophrenia, genes impacted by CNVs were more intolerant than in controls. ExAC CNV data constitutes a critical component of an integrated database spanning the spectrum of human genetic variation, aiding the interpretation of personal genomes as well as population-based disease studies. These data are freely available for download and visualization online.

Corresponding Authors. Douglas Ruderfer (douglas.ruderfer@mssm.edu), Shaun Purcell (shaun.purcell@mssm.edu).
[5]A list of members and affiliations appears in the Supplementary Note
*These authors contributed equally to this work

Author Contributions

DMR, MF and SMP designed the study. ML, KJK and DGM handled sample and data management. DMR, TH, KES, MF and SMP contributed to statistical analyses. DK, DMR and KJK designed and implemented website visualizations. DMR, MJD, DGM, MF and SMP contributed to primary interpretations. DMR, MF and SMP performed the primary drafting of the manuscript. All authors contributed to, read and approved the final manuscript.

Competing financial interests

The authors declare no competing financial interests.

Introduction

Copy number variation (CNV) – in particular a gain or loss of coding sequence – is known to contribute substantially to phenotypic diversity and disease[1,2]. Large CNVs (deletions or duplications) were initially discovered from cytogenetic studies of individuals with Down syndrome and intellectual disability[3–5]. Technological advances in surveying changes in genetic dosage, along with the sequencing of the human genome, have led to improved resolution for detection of CNVs and other forms of structural variation[6,7], better understanding of CNV mechanism[8], and the further implication of CNVs in various diseases[2,9–11]. Still, the ability to ascribe pathogenicity to a particular CNV remains limited[12].

Genotyping arrays have allowed for cost-effective strategies to detect CNVs in large samples but will typically detect only relatively large CNVs[13,14,15]. Conversely, whole-genome sequencing provides a comprehensive assessment of CNV (and other structural variation), but costs[9] currently limit its widespread application. It has recently been demonstrated that CNVs can be detected from exome sequencing, using information on relative read-depth to infer chromosomal gains and losses that impact targeted genes[16,17]. Unlike arrays, exome sequencing can potentially resolve genic CNVs to the level of a single exon. Although still crude in comparison to whole-genome sequencing, exome sequencing data can map smaller genic CNVs (<30kb) that may be undetected by arrays but still impact disease risk[18]. Most crucially, exome sequencing data already exist across multiple large studies and have been compiled under the auspices of the Exome Aggregation Consortium (ExAC, see URLs, *Lek et al.*). Here, we leveraged this large (N ~ 60,000) resource to better characterize the rates and properties of rare CNVs, with population frequencies on the order of 10^{-2} to as low as 10^{-5}. We constructed the ExAC CNV dataset using a previously developed method (XHMM[17]). Specifically, for each autosomal gene, we used sequencing read depth for an individual to calculate the posterior probability of being diploid across that gene (i.e., normal copy number state) versus deleted, or duplicated. Importantly, this approach identifies genes for which we are unable to confidently assess copy number for a given individual. It also flags genes that are only partially impacted by CNV (i.e., some exons are diploid) versus full genic deletion or duplication.

Evolutionary theory predicts that negative selection will result in deleterious mutations being rarer on average than neutral mutations, which has been demonstrated for single nucleotide variants (SNVs)[19,20] and CNVs[21]. Although large CNVs that impact many genes are likely to be deleterious[22], certain genes will be more sensitive to (i.e., intolerant of) dosage changes and thus have fewer CNVs. In this work, we leverage the tens of thousands of exome samples in ExAC to estimate genic frequencies for rare CNV. We then calibrate those empirical frequencies by expected rates of CNV to derive for each gene a measure of relative intolerance to CNVs – that is, a trend of showing fewer CNVs than expected. We show how the estimated CNV intolerance values are related to measures derived from SNV and to

URLs
ExAC web browser (exac.broadinstitute.org), genes implicated in recessive disorders (research.nhgri.nih.gov/CGD).

evolutionary measures of genic constraint. We conclude that considering CNV intolerance can be used to predict the likelihood of a genic CNV being deleterious, and we demonstrate how genic intolerance can be employed in the analysis of disease studies.

Results

Characterizing CNV calls from exome-sequencing data

Read depth information from targeted exome-sequencing of 60,642 individuals was analyzed using XHMM[17]. Briefly, XHMM removes systematic individual, batch, and target effects (artifact or common copy number polymorphism) by use of principal component analysis on the entire read-depth matrix (60,642 individuals by 219,437 targets). A hidden Markov model applied per individual to the normalized data is used to call CNVs at exon-level resolution and estimate genic copy number probabilities (see Online Methods). We performed quality control and restricted analysis to genes where each CNV is rare (observed in < 600 individuals, corresponding to a maximum allele frequency of ~0.5%). CNV quality was assessed using trios and demonstrated high specificity and sensitivity consistent with previous reports[17] (see Online Methods). Additionally, a subset of 10,091 individuals had high quality CNV calls from genotyping arrays[23], for whom we assessed the comparability of CNVs called from genotyping arrays versus exome-sequencing. The set of array-based CNVs were filtered for high confidence based on number of markers (10), length (>100kb) and frequency (<1%), as described[23]. For the most confidently called array-based CNVs, those longer and intersecting the most coding sequence (greater than 20 targets), 78% were also called in the high-confidence set of exome-sequencing CNV (1,307/1,684). Array-based CNV intersecting fewer targets were less likely to be called in the exome-sequencing set (Supplementary Figure 1), such that 62% of array-based CNVs hitting more than 3 exons and 54% of all array-based CNVs hitting at least one GENCODE protein-coding exon (3,200/5,927) were called in the exome-sequencing set. In comparison, of 12,947 CNVs in the exome-sequencing set, 3,268 (25%) were seen in the array-based call set, with this overlap increasing as the number of targets encompassed by CNV increased (Supplementary Figure 2). For the concordantly called CNVs, array-based calls encompassed more exons 70% of the time, however, on average 83% of the exons were included in calls from both technologies (median = 93%). Individuals carried on average 2.2 times more CNVs in the exome-sequencing dataset compared to the array-based call set (1.28 to 0.59).

The final ExAC CNV dataset consisted of 59,898 individuals and 126,771 CNVs overlapping GENCODE autosomal protein-coding genes. On average, individuals carried 2.1 high-confidence, rare CNVs (0.82 deletions, 1.29 duplications) hitting at least 1 of the 19,430 GENCODE autosomal protein coding genes (Figure 1). The largest group of 17,565 (29%) individuals carried exactly 1 rare coding CNV, with 12,812 (21%) carrying zero CNVs, and 3,730 (6%) carrying greater than 5. The mean extent of CNV per individual was 154kb (median = 35kb) representing more duplicated genomic content (107kb) than deleted (46kb). The average length of CNV was 73kb (median = 15kb), with duplications being 83kb (median = 20kb) and deletions being 56kb (median = 9kb). 84% of CNVs were smaller than 100kb, which has generally been used as the size threshold for confidently called CNV from genotyping arrays; 56% of CNVs were shorter than 20kb.

Seventy percent of individuals had at least one gene impacted by a rare CNV (37% had at least one deleted gene, 54% had at least one duplicated gene), with an average of 0.81 deleted genes and 1.75 duplicated genes per individual across the dataset (Figure 2, Table 1). Sixteen percent of CNVs were greater than 100 kb, averaging 79 kb (59 kb for deletions, 91 kb for duplications) and 13 exons (9.7 exons for deletions, 15 exons for duplications). CNV rates varied by population: individuals of African descent had the highest rate, similar to that seen in SNV[24]; however, these rates were significantly confounded by variables such as batch and overall read depth, complicating the interpretation of this finding (Online Methods, Supplementary Table 1, Supplementary Figure 3–4). As previously reported[25], we identified a significant increase of CNV rate in females, after adjusting for read depth, cohort, and 10 principal components of ancestry (mean female CNV rate 1.74, mean male CNV rate 1.49, $p = 1.14\times10^{-10}$, Supplementary Table 1).

On average, each gene was deleted in 3.1 individuals and duplicated in 6.6. Most of the protein-coding genome harbored population-level rare variation in copy number, with only 1,872 genes having no CNVs detected (6,578 genes without deletions, 3,038 genes without duplications). 55% of all CNVs overlapped only a single gene (65% of deletions, 48% of duplications). Of these single-gene CNVs, most (62%) were partial-gene CNVs (Figure 2, Table 1), with some exons deleted or duplicated but also with some exons confidently assigned as diploid (see Online Methods).

A measure of genic intolerance to CNVs

To quantify the effect of genic CNV, we defined genes that harbored fewer CNVs than expected as being more “intolerant”. We expect that CNVs in intolerant genes, when they do occur, will be more likely to have deleterious effects, analogous to genic constraint scores based on SNVs[26,27] (*Lek et al. companion paper*). However, it is not straightforward to model genic CNV rates expected under neutrality in a direct manner, as can be done for SNVs using trinucleotide mutation rates and the gene’s known sequence. To derive expected values, we therefore fit a linear regression model for the observed CNV rate per gene based on gene length, coding sequence length, number of targets, GC content, sequence complexity, genomic localization within pairs of segmental duplications, and sequencing read depth (see Online Methods, Supplementary Table 2, Supplementary Figure 5). Intolerances scores were calculated as the normalized and winsorized model residuals, negated such that higher positive values indicate greater intolerance (a lower than expected rate of CNVs for that gene). As defined, CNV intolerance scores are therefore independent of the predictor variables used in the linear regression (Supplementary Figure 6).

Intolerance scores based only on deletions were highly correlated to those based only on duplications ($r = 0.37$, $p << 10^{-20}$) and both scores correlated highly with the combined score ($r = 0.7$ for deletions, $r = 0.89$ for duplications, the difference reflecting the greater number of duplications). A complementary approach to predict haploinsufficiency[28] that compared genes sensitive to gene loss to those where having a single copy resulted in no discernable phenotype demonstrated significant correlation with CNV intolerance scores ($r = 0.12$, $p = 2\times10^{-36}$). CNV intolerance scores were also significantly correlated with a measure of genic constraint based on missense SNVs[26] ($r = 0.2$, $p = 2\times10^{-137}$) derived from

the ExAC sample (*Lek et al.* companion paper), this effect being stronger for deletions (r = 0.23, p = 2×10^{-176}) compared to duplications (r = 0.14, p = 1×10^{-63}). This correlation was consistent across the distribution of scores showing an increase of CNV intolerance score as both SNV scores (based on either missense or LoF variants) increased (Supplementary Figure 7). Similarly, CNV intolerance scores also correlated with an index of haploinsufficiency ("pLI", *Lek et al.* companion paper) based on loss-of-function variants (nonsense and canonical splice site SNVs) derived from this sample (all CNV: r = 0.18, p = 6×10^{-110}, deletions: r = 0.23, p = 1×10^{-176}, duplications: r = 0.11, p = 1×10^{-39}). Unlike for SNV-based scores, CNV intolerance scores will be correlated across multiple genes hit by larger CNVs. We therefore calculated CNV intolerance scores from CNVs that only hit a single gene and identified similar correlations with pLI (r = 0.22 deletions, r = 0.06 duplications). While single-gene CNVs are likely more individually informative for quantifying intolerance, the sole use of these CNVs in creating the scores would reduce the number of events by half. We therefore use the all CNV scores going forward but provide both scores online (see URLs).

CNV intolerance scores were also associated with an independent measure of evolutionary constraint, GERP[29]. Genes with higher mean per-base GERP scores (calculated including introns) tended to have higher CNV intolerance scores (r = 0.13, p = 5×10^{-46}). In a joint linear regression of genic GERP score on CNV intolerance and SNV constraint scores, all terms were independently and positively associated with genic GERP scores (CNV intolerance p = 3×10^{-33}; SNV missense constraint p = 6×10^{-27}; SNV LoF constraint p = 3×10^{-5}), suggesting that both CNV and SNV-based scores contribute non-redundant information regarding the potential deleteriousness of genic CNVs.

Characterizing CNV tolerant and intolerant genes

For a particular gene, intolerance of genetic variation such as CNV implies higher functional importance of that gene (*Lek et al.* companion paper). We thus considered the relationship between the intolerance of a gene to CNV and its expression across 27 tissues[30], focusing on the 7,754 genes that are highly expressed in at least one of those tissues (but not all of them). We found that for the majority of tissues (n=17), the highly expressed genes indeed had significantly higher intolerance scores compared to all other genes within this subset (Figure 3a). Notably, genes highly expressed in the brain showed the most intolerance to CNV. Tissues expressing genes that are more intolerant of CNVs also tended to show relatively fewer genes with homozygous loss-of-function SNVs and short indels ("complete knockouts") in a recent survey of the Icelandic population[31] (Spearman's rho = 0.45, p = 0.019) (Supplementary Table 3). Genes highly expressed in three tissues - duodenum, liver, and pancreas - demonstrated significantly lower intolerance scores (i.e., greater tolerance) than average genes, raising the hypothesis of greater robustness to dosage changes in those tissues.

Genes previously defined as haploinsufficient[28] or essential[32] showed higher CNV intolerance scores compared to all genes (p = 2×10^{-25} and 2×10^{-12}, respectively, Supplementary Table 4). In contrast, genes implicated in recessive disorders (see URLs) and those with no identifiable phenotype in mice[15] tended to show greater tolerance to CNV (p =

0.007 and 0.009, respectively, Supplementary Table 4). With the exception of the recessive disorder genes, similar overall results were recently obtained in an analysis of a large dataset of CNVs from genotyping arrays[15] (Supplementary Table 4). Applying generic geneset enrichment analysis to the most and least CNV intolerant genes (top/bottom 5%, 787 genes each, Figure 3b), intolerant genes were significantly enriched in Gene Ontology (GO) sets related to neuronal and axon development and synapse organization and assembly, consistent with the aforementioned higher intolerance of genes that are highly expressed in brain tissue (GO:0048666 Neuron Development $p = 2\times10^{-6}$, GO:0050808 Synapse Organization $p = 6\times10^{-6}$, Supplementary Tables S5–S8).

Application to disease: CNV intolerance and schizophrenia

ExAC-derived genic CNV intolerance scores can be used alongside other genic annotations in disease association studies. As a proof-of-principle, we set aside a single case/control study present in ExAC [4,793 schizophrenia (SCZ) cases and 6,102 controls[33]] and calculated intolerance scores in the remaining 47,787 individuals as described above. As previously reported[23], this sample of SCZ cases showed a higher number of genes affected by CNVs compared to controls (2.12 versus 1.78, $p = 1\times10^{-10}$). Over and above the number of genes hit, cases carried a higher mean intolerance across all genes hit by CNVs compared to controls (−1.35 versus −1.42, $p = 0.007$). (Note that, as expected, genes for which we observe any CNV in a given sample in fact tend to be more tolerant, thus both groups have negative means). Further, cases carried a greater normalized intolerance (see Online Methods) of CNVs than controls (0.44 versus 0.33, $p = 1\times10^{-11}$). To assess the independent information contained in the CNV intolerance score, we calculated the normalized mean SNV-based constraint score for each individual and tested whether these scores correlated with disease status. We identified significant increased constraint in schizophrenia cases compared to controls from the missense constraint score ($p=4\times10^{-4}$), loss-of-function constraint score ($p=2\times10^{-4}$), and pLI ($p=8\times10^{-8}$). In a joint test of all scores from independent annotations, the CNV intolerance scores remains the most significant predictor (CNV: $p=6\times10^{-7}$, missense: p=0.17, pLI: p=0.004). This suggests that it will be beneficial to develop disease risk-association testing frameworks that jointly consider the type of CNV with respect to their genic intolerance scores, as well as the number of deleted or duplicated genes.

Discussion

Here we have presented gene-level frequencies and intolerance scores for CNVs from nearly 60,000 individuals, providing a data-driven means for estimating the likely deleteriousness of genic CNV. Consistent with their relevance to gene function, the current estimates of CNV intolerance show non-random profiles with respect to tissue-specific gene expression patterns, to independent measures of genic constraint, and to risk of disease. We provide summaries of these data at the gene and exon level and detailed QC metrics online.

Limitations of this work include the relative difficulty in ascertaining accurate copy number calls from targeted (exome) short-read sequencing and the inability to accurately call common or more complex variants, along with the rarity of these events that increases the

noise around point estimates of frequency and corresponding intolerance scores. In generating intolerance scores, we attempted to control for gene-to-gene variability in observed CNV rates resulting from factors other than evolutionary selection on the phenotypic consequences of bearing a CNV in that gene, for example, gene size and sequencing coverage. Yet, though we attempted to model the increased rates of CNV proximal to segmental duplications, our incomplete knowledge of CNV mutational mechanisms can add noise and bias to these estimates of intolerance, in particular in regions of known recurrence.

It is also important to note that many ExAC sample participants were ascertained on disease status. Inasmuch as a minority of genes had significantly higher rates of CNVs because of this, then these genes will have slightly deflated intolerance estimates compared to those derived from a phenotypically-screened control sample.

Despite these limitations, the analyses presented here point to the value of more comprehensive assessments of genetic variation. Whether or not a gene tolerates deletion or duplication is most directly estimated by considering the empirical patterns of genic CNV rates in large samples, as performed here. Combination with other measures of genic constraint, including those based on SNVs and evolutionary analyses, is likely to yield better and more general metrics for assessing the likely impact of any type of genic variant, leading to improved interpretation of personal genomes and disease association studies.

Online Methods

CNV calling in exome-sequencing data of 60,642 individuals

XHMM was run as previously described[17]. Briefly, GATK DepthOfCoverage was employed to calculate mean per-base coverage (counting unique fragments based on reads mapping with a quality >20), across 219,437 targets (including 7,439 and 708 on chromosomes X and Y, respectively, and 9 on the mitochondrial genome). To accommodate the variety of exome captures used across the various component projects, these targets were liberally defined as the Illumina ICE v1 targets plus GENCODE v19 coding regions, both padded by 2 bp, from which the unique set of relevant “exome targets” was finalized. A total of 31,769 of these targets were subsequently filtered out before CNV calling: 21,072 for having mean sequencing depth (across all samples) <10×, 8,875 for having low complexity sequence (as defined by RepeatMasker) in >25% of its span, 225 for having GC content <10% or >90%, 1,582 for covering <10 bp, and 15 targets spanning >10 kbp. The resulting sample-by-target read depth matrix was scaled by mean-centering the targets, after which principal component analysis (PCA) of the full matrix was performed; note that with the LAPACK implementation in XHMM, this still required 800 GB of RAM and ~1 month of computation time. For data normalization, the top 388 principal components (those with variance >70% of the mean variance across all components) were removed from the data to account for systematic biases at the target- or sample-level, such as GC content or sequencing batch effects. Subsequently, 3 targets were removed for still having high variance after normalization (standard deviation >50), and sample-level z-scores were calculated (with absolute values capped at 40). CNV were called using the Viterbi hidden Markov model (HMM) algorithm with default XHMM parameters, and XHMM CNV quality scores were

calculated as previously described using the forward-backward HMM algorithm and modifications as previously described. In addition, all called CNV were statistically genotyped across all samples using the same XHMM quality scores and output as a single uniformly-called VCF file.

QC of CNV data

In total, we attempted CNV calling for 60,642 out of the 60,706 (99.9%) ExAC samples, the remainder having either failed calling for low overall read depth or were not included due to upstream data access issues. The CNVs output by XHMM were first frequency filtered to remove common CNVs, i.e., those seen more than 600 times (>1%), defined as overlapping more than 50% of their respective targets. Based on previous work[17], we retained only those CNVs with quality scores greater than or equal to 60. We removed any individual having a CNV count greater than 3 standard deviations above the mean, that is, 24 CNVs (n=775 samples removed). Thus, our final dataset consisted of 59,898 individuals and 126,771 CNVs overlapping GENCODE autosomal protein-coding genes.

Filtering of genes

Of the 20,345 GENCODE v19 genes labeled as protein-coding, we limited our analyses to the set of 19,430 genes occurring on autosomes, where CNVs on sex chromosomes were removed due to technical issues. Next, we removed any gene where half or more of its targets were filtered out during the CNV calling (1,068 genes, see above). We further removed genes having unusually low (<30×) or high (>200×) mean coverage (944 genes). Using data from a recent report on CNV from whole genome sequencing data of 849 genomes sequenced from the 1000 Genomes Project[34], we removed any gene known to be multi-allelic (735 genes). Finally, we removed any gene in which there existed any CNV with frequency greater than 0.5% (1,193 genes). This yielded a final set of 15,734 genes for all subsequent genic analyses.

Assessment of CNV quality in parent-child trios

To assess overall CNV quality, we utilized 241 previously described[35,36] parent-offspring trios from Bulgaria to confirm that apparent *de novo* rates and parent-to-child transmission broadly conformed to expectations of random Mendelian segregation (note that the offspring had a diagnosis of schizophrenia and were not part of the primary ExAC dataset, which included only unrelated individuals). Poor sensitivity would result in severely reduced transmission statistics, while poor specificity would induce many false positive CNV calls and increased rates of *de novo* CNVs. Through reasonable estimates of transmission and *de novo* events, we can infer high specificity and sensitivity of CNV calls overall. Defining CNV transmission as implemented in the Plink/Seq cnv-denovo command[17], we assessed whether the rate of transmission for CNV converged to the expected Mendelian rate of 50% across a range of quality score thresholds. Using the recommended quality score cutoff (SQ >= 60), median per trio CNV transmission rates were at the expected 50%, with the aggregate transmission rate across CNVs in all trios falling to 43% (44% for deletions, 42% for duplications). These rates exclude situations where the offspring's CNV is neither confidently called deleted or duplicated (SQ >= 60) nor confidently called diploid (DQ >= 60). Including these more uncertain events, and conservatively counting them as non-

transmissions, results in aggregate transmission rates of 32%. Nevertheless, these results remain consistent with high specificity as confirmed by a low mean of 0.058 *de novo* CNVs per trio (half of which were over 1 kb and spanning 5 or more exons), which only increases to 0.13 *de novo* CNVs per trio when treating uncertain events in the parents as diploid. Indeed, a comparable *de novo* CNV rate of 0.051 was found in a larger version of this cohort (622 trios) using genotyping arrays[35].

Gene/Exon-specific copy number calls

We defined gene-specific copy number state per individual, assessing the probability of a CNV occurring anywhere between transcription start and end. Specifically, this was performed by defining the genomic intervals spanned by each gene and then using the same sample-by-target matrix of z-scores described in the “CNV calling” section above, in order to statistically genotype these gene regions across all samples. This genotyping procedure yielded a VCF file containing key copy number metrics, including those corresponding to the probability that an individual is confidently diploid for the extent of the gene, or, alternatively, has some deletion or duplication therein. All of these probability-derived metrics were calculated using the forward-backward HMM algorithm modified to efficiently calculate posterior probabilities across all targets in a gene, analogous to genotyping across all targets in a particular called CNV region (as described above). Though XHMM performs exome-wide correction for both regional and individual read depth variability, we found that increased sample read depth is still correlated with increased numbers of CNVs (Supplementary Figure 1). In the absence of large-scale validation efforts and given the focus on CNV that are rare at any particular locus, it is not feasible to easily normalize out this effect. However, we did account for potential confounders, such as gene size and read depth, in calculating gene-specific diploid quality (by defining a threshold of three standard deviations below the mean diploid quality of all individuals). Using this approach, we obtained confidence measures for deletion, duplication, and diploid status for every individual at every gene. We further employed the same strategy to call exon-specific copy number states, again starting with the genic exons and overlapping those with all targets at which read depths were calculated and normalized; note that this typically included a single target per exon, but for a small proportion of exons, this included 2 or more targets, due to the slight differences in the definition of the target regions for CNV calling and the GENCODE exon regions (see “CNV calling” section above). Genic CNV counts derived from this procedure correlated with the number of loss-of-function variants in a gene.

Creating genic CNV intolerance scores

For the 15,734 genes that survived QC, we constructed genic measures of intolerance for all CNVs and separately for deletions and duplications. In the absence of a high-quality mutation model for CNVs, we employed an empirical approach incorporating genomic information. From a set of 9,396 unique pairs of segmental duplications on the same chromosome downloaded from the UCSC Genome Browser, we created a subset of 2,790 non-redundant pairs requiring that the genomic intervals between them were less than 80% overlapping and less than 4Mb in length. We identified a significant increase in the number of CNVs in genes within these regions (Supplementary Figure 5), so we included this a factor in predicting CNV frequency. Ultimately, we calculated genic intolerance from the

residuals of a logistic regression of CNV frequency on gene length, read depth, GC content, sequence complexity, and the number of pairs of segmental duplications the gene is between, along with higher order terms. We next calculated z-scores such that positive values represented a lower frequency of CNV (more intolerance), winsorising the negative tail at 5%.

Stratifying CNV by genic content affected

We stratified CNVs by the number of genes and exons for which they (confidently) affect dosage. Specifically, we defined "single-gene" CNVs as those with a gene-specific confidence score greater than 60 in one of the 15,734 genes that remained after gene QC, but also strictly requiring overlap with only one of the 19,430 GENCODE autosomal protein-coding genes. CNVs overlapping more than one gene were labeled as "multi-gene." Utilizing the exon-level CNV calls, we further refined our single-gene CNVs into three classes: 1) "full" were genes where all exons were confidently called as deleted or duplicated, 2) "ambiguous" were genes with at least one exon confidently called deleted or duplicated but no exons confidently called diploid, or 3) "partial" were genes in which there was at least one exon confidently called deleted or duplicated and at least one exon confidently called as diploid.

Predefined gene sets

We collated three groupings of gene sets to test for enrichment. The first is a set of highly expressed genes from expression data of 27 tissue types (pancreas, liver, duodenum, small intestine, kidney, colon stomach, salivary glands, testis, prostate, skin, esophagus, gall bladder, thyroid gland, heart, adipose tissue, urinary bladder, ovary, adrenal glands, lymph nodes, appendix, lung, bone marrow, placenta, spleen, endometrium, and brain) previously published[30]. We defined highly expressed per tissue as having fragments per kilobase of exon per million fragments mapped (FPKM) greater than 20, but excluding genes that were highly expressed in all tissues. The second is a set of disease-implicated genes collated in a previous paper analyzing a large set of CNVs;[15] these include sets of dominant and recessive disease genes, genes implicated in cancer, haploinsufficient genes, genes essential in mice, genes intolerant to loss of function variants, and genes not related to a specific phenotype in any such database (Supplementary Table 3–4).

Gene set enrichment analysis

We selected the genes at the top and bottom 5% of CNV intolerance score (n=787 each) and ran gene set enrichment analysis using ToppFun[37], which uses a hypergeometric test of gene sets across 18 possible categories, of which we selected 9 categories of pathways (GO molecular, GO biological, GO cellular, Human Phenotype, Mouse Phenotype, Domain, Pathway, Gene Family, and Disease). The most intolerant genes were enriched in GO sets related to neuronal and axon development and synapse organization and assembly. The most tolerant genes were enriched for metallothioneins and myosin filament genes (Figure 2b, Supplementary Tables 5–8).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to acknowledge Eugene Fluder and Khalid Shakir for their help in running XHMM at the large scale required for over 60,000 samples. Work at the Icahn School of Medicine at Mount Sinai was supported by the Institute for Genomics and Multiscale Biology (including computational resources and staff expertise provided by the Department of Scientific Computing) and NIH grants R01-HG005827 and R01-MH099126 (to S.M.P.).

MF is now an employee at Verily Life Sciences.

References

1. Lupski JR. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet. 1998; 14:417–422. [PubMed: 9820031]
2. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006; 7:85–97. [PubMed: 16418744]
3. Jacobs PA, Baikie AG, Court Brown WM, Strong JA. The somatic chromosomes in mongolism. Lancet. 1959; 1:710. [PubMed: 13642857]
4. Lejeune J, Turpin R, Gautier M. Chromosomic diagnosis of mongolism. Arch Fr Pediatr. 1959; 16:962–963. [PubMed: 14415503]
5. Jacobs PA, Matsuura JS, Mayer M, Newlands IM. A cytogenetic survey of an institution for the mentally retarded: I. Chromosome abnormalities. Clin Genet. 1978; 13:37–60. [PubMed: 146575]
6. Sudmant PH, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015; 526:75–81. [PubMed: 26432246]
7. Kidd JM, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008; 453:56–64. [PubMed: 18451855]
8. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009; 10:551–564. [PubMed: 19597530]
9. Conrad DF, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010; 464:704–712. [PubMed: 19812545]
10. Iafrate AJ, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004; 36:949–951. [PubMed: 15286789]
11. Sebat J, et al. Large-scale copy number polymorphism in the human genome. Science. 2004; 305:525–528. [PubMed: 15273396]
12. Buchanan JA, Scherer SW. Contemplating effects of genomic structural variation. Genet Med. 2008; 10:639–647. [PubMed: 18978673]
13. Redon R, et al. Global variation in copy number in the human genome. Nature. 2006; 444:444–454. [PubMed: 17122850]
14. McCarroll SA, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. 2008; 40:1166–1174. [PubMed: 18776908]
15. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015; 16:172–183. [PubMed: 25645873]
16. Plagnol V, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics. 2012; 28:2747–2754. [PubMed: 22942019]
17. Fromer M, et al. Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth. American Journal of Human Genetics. 2012; 91:597–607. [PubMed: 23040492]
18. Poultney CS, et al. Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. Am J Hum Genet. 2013; 93:607–619. [PubMed: 24094742]

19. Fu W, Akey JM. Selection and adaptation in the human genome. Annu Rev Genomics Hum Genet. 2013; 14:467–489. [PubMed: 23834317]
20. Purcell SM, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014; 506:185-+. [PubMed: 24463508]
21. Itsara A, et al. Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet. 2009; 84:148–161. [PubMed: 19166990]
22. Need AC, et al. A genome-wide investigation of SNPs and CNVs in schizophrenia. PLoS Genet. 2009; 5:e1000373. [PubMed: 19197363]
23. Szatkiewicz JP, et al. Copy number variation in schizophrenia in Sweden. Molecular Psychiatry. 2014; 19:762–773. [PubMed: 24776740]
24. MacArthur DG, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012; 335:823–828. [PubMed: 22344438]
25. Desachy G, et al. Increased female autosomal burden of rare copy number variants in human populations and in autism families. Mol Psychiatry. 2015; 20:170–175. [PubMed: 25582617]
26. Samocha KE, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014; 46:944–950. [PubMed: 25086666]
27. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 2013; 9:e1003709. [PubMed: 23990802]
28. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. PLoS Genet. 2010; 6:e1001154. [PubMed: 20976243]
29. Cooper GM, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005; 15:901–913. [PubMed: 15965027]
30. Fagerberg L, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol Cell Proteomics. 2014; 13:397–406. [PubMed: 24309898]
31. Sulem P, et al. Identification of a large set of rare complete human knockouts. Nat Genet. 2015; 47:448–452. [PubMed: 25807282]
32. Ye YN, Hua ZG, Huang J, Rao N, Guo FB. CEG: a database of essential gene clusters. BMC Genomics. 2013; 14:769. [PubMed: 24209780]
33. Ripke S, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nature Genetics. 2013; 45:1150-U282. [PubMed: 23974872]
34. Handsaker RE, et al. Large multiallelic copy number variations in humans. Nat Genet. 2015; 47:296–303. [PubMed: 25621458]
35. Kirov G, et al. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. Molecular Psychiatry. 2012; 17:142–153. [PubMed: 22083728]
36. Fromer M, et al. De novo mutations in schizophrenia implicate synaptic networks. Nature. 2014; 506:179-+. [PubMed: 24463507]
37. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009; 37:W305–W311. [PubMed: 19465376]
38. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One. 2010; 5:e13984. [PubMed: 21085593]

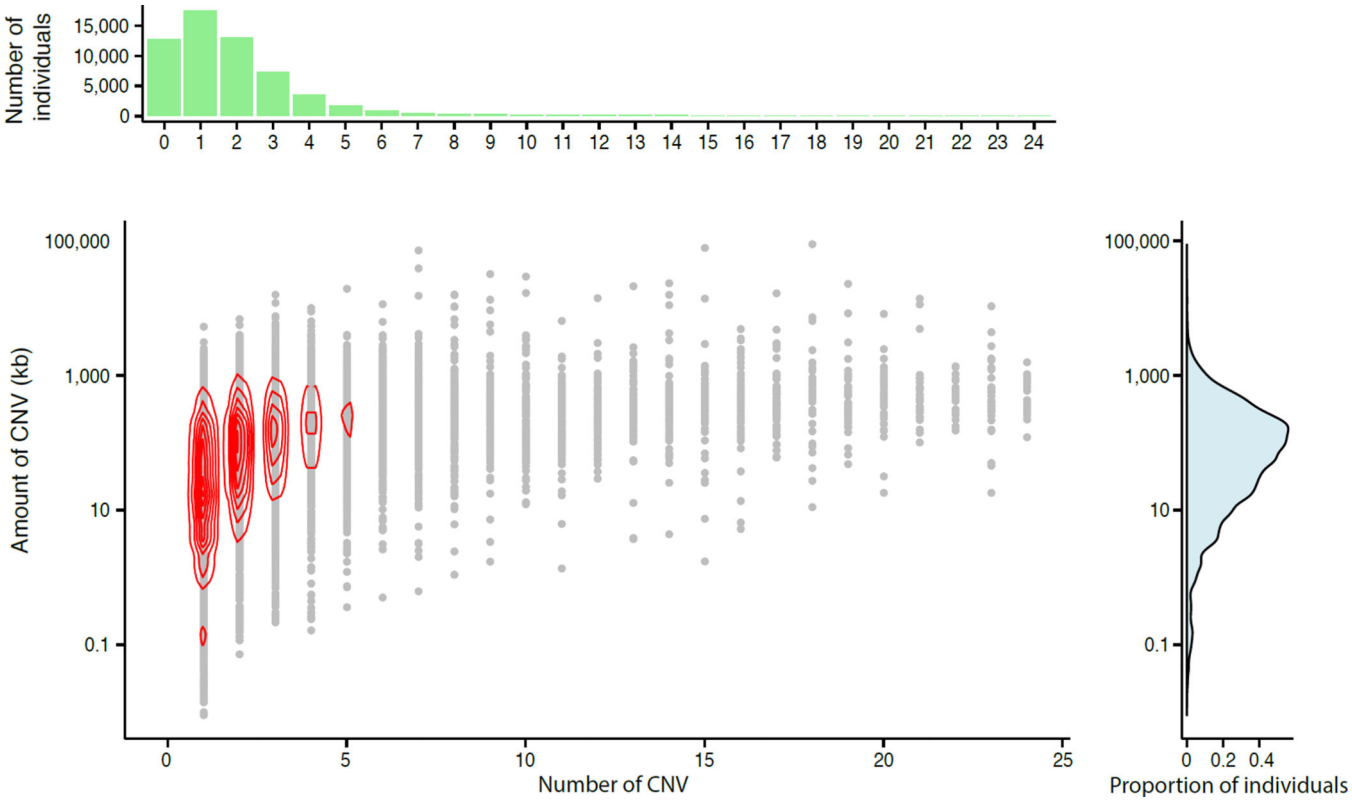

Figure 1.

Distribution of number and amount (in kb) of CNV across 59,898 exome-sequenced individuals. Including histogram of number of CNVs per individual (top), two-dimensional density plot of CNV number and amount (middle), and density plot of amount of CNV per individual (right).

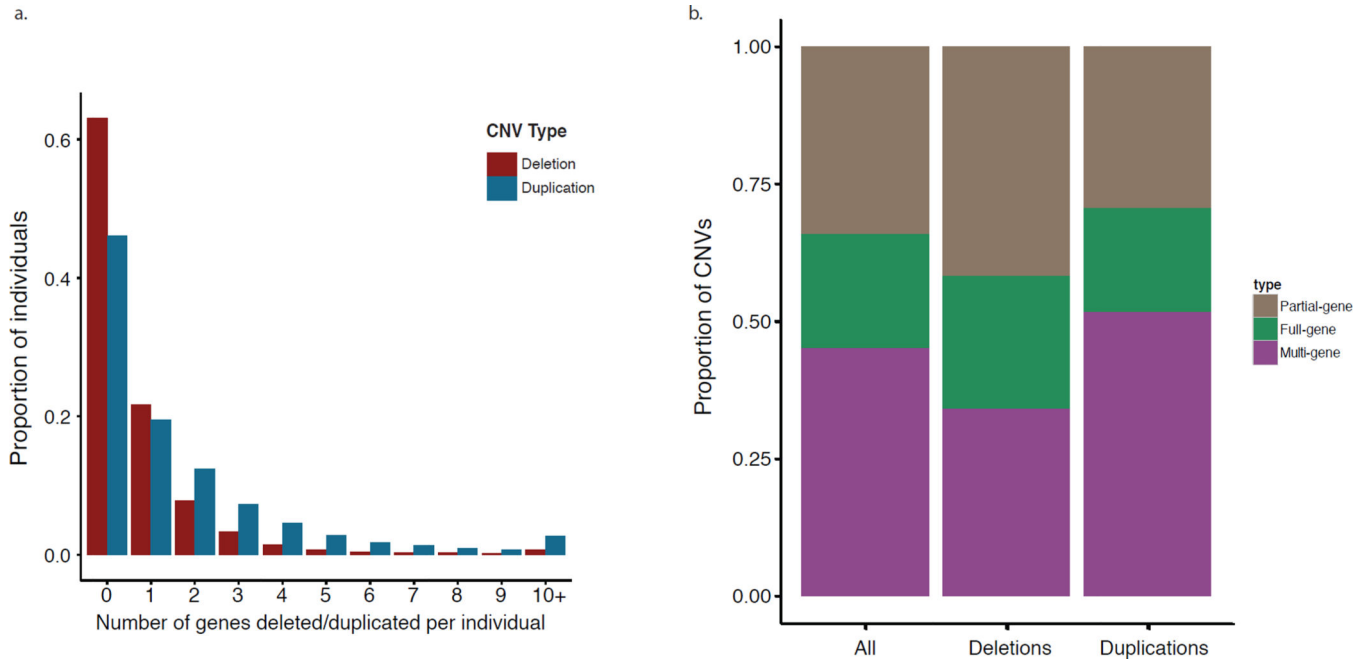

Figure 2. Genic summary of rare deletions and duplications in ExAC sample

a. Proportion of individuals having from 0 to 10 or more genes deleted (red) or duplicated (blue). **b**. Proportion of CNV that affect multiple genes (multi-gene), impact the entirety of a single gene (full-gene), or partially disrupt a single gene (partial-gene). The two rightmost bars split these proportions for deletion and duplications, respectively.

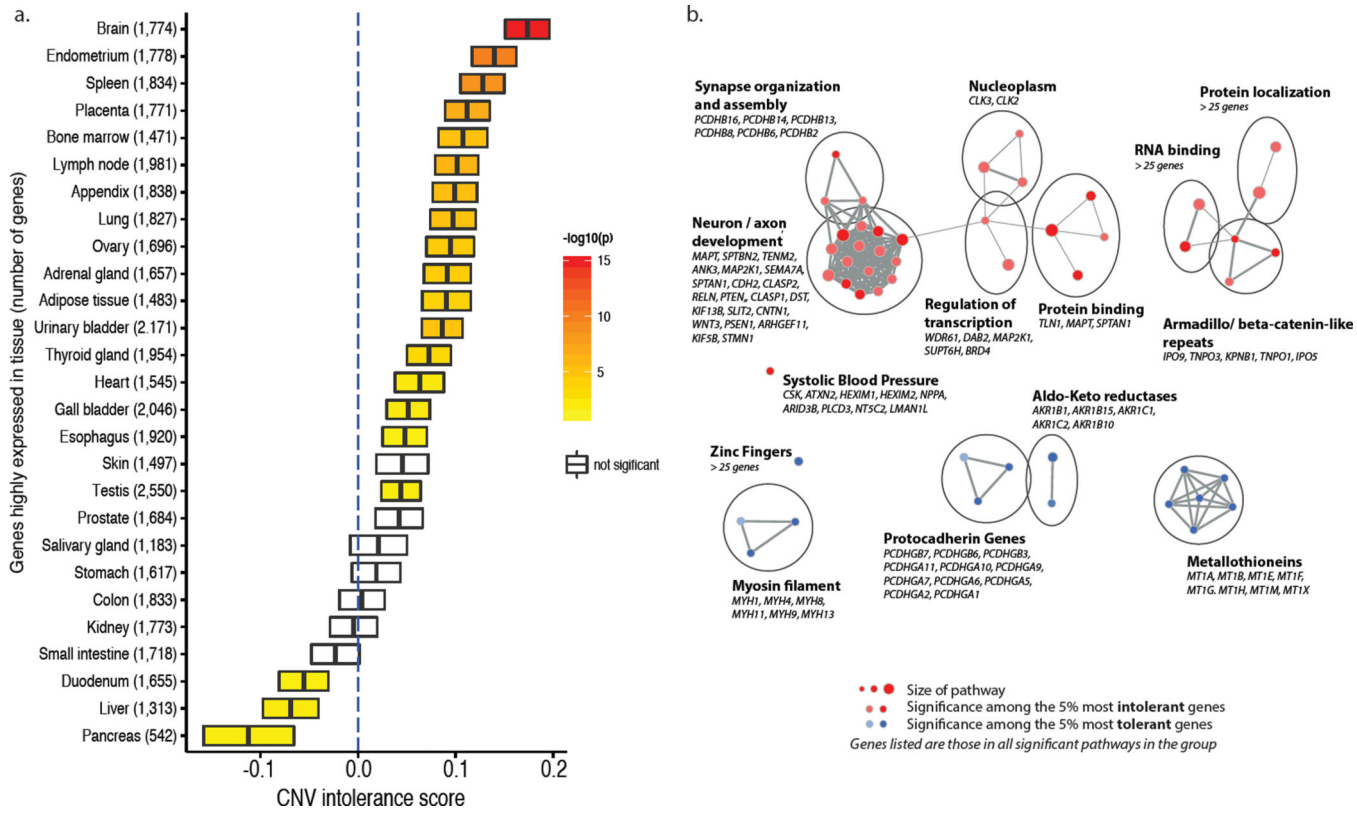

Figure 3. Brain relevant genes demonstrate greatest intolerance to dosage changes from CNVs

a. After removing genes highly expressed in all tissues (FPKM > 20), 27 tissues[30] were rank-ordered by the mean ExAC CNV intolerance scores for the highly expressed genes in each tissue; mean and standard error of mean intolerance score are indicated by bold line and box width, respectively. Box color denotes significance of two-sided t-test of difference of intolerance scores between tissue-expressed genes and all others; white bars indicate no significant difference ($p > 0.05$). Vertical dashed blue line marks the mean CNV intolerance score for all genes. **b**. Network diagrams of pathways significantly enriched for the 5% most CNV-intolerant (red) and CNV-tolerant (blue) genes [created using Enrichment Map Cytoscape plug-in[38]]. Results are based on tests of 9 categories of pathways (GO molecular, GO biological, GO cellular, Human Phenotype, Mouse Phenotype, Domain, Pathway, Gene Family, and Disease); only those surpassing Bonferroni ($p < 0.05$) and FDR significance are shown. Node size represents number of genes in a pathway, color represents significance of enrichment, and thickness of a pairwise edge corresponds to the proportion of genes overlapping between the corresponding pair of gene sets. Groupings were manually assigned a label, and genes listed are those present in all significant pathways within a group.

Table 1

Number of total genes impacted (N), and mean number of gene-level CNV per individual (rate). The bottom two rows consider only CNV affecting a single entire gene (single-gene) or only part of a gene (partial-gene); second and third columns separately split out deletions and duplications.

	All		*Deletions*		*Duplications*	
Genes (n=15,734)	**N**	**Rate**	**N**	**Rate**	**N**	**Rate**
All	13,862	2.565	9,156	0.817	12,696	1.747
Single-gene	7,159	0.881	4,723	0.399	5,268	0.481
Partial-gene	4,886	0.543	3,358	0.251	3,435	0.292