nature genetics

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

Douglas M Ruderfer[1–3], 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[1–4,7] & Shaun M Purcell[1–4,6,7]

Copy number variation (CNV) affecting protein-coding genes contributes substantially to human diversity and disease. Here we characterized the rates and properties of rare genic CNVs (<0.5% frequency) in exome sequencing data from nearly 60,000 individuals in the Exome Aggregation Consortium (ExAC) database. 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 affected by CNVs were more intolerant than in controls. The ExAC CNV data constitute a critical component of an integrated database spanning the spectrum of human genetic variation, aiding in the interpretation of personal genomes as well as population-based disease studies. These data are freely available for download and visualization online.

CNV—in particular, 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], a better understanding of the 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[13–15]. 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 affect 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 (<30 kb in length) that may be undetected by arrays but still affect disease risk[18]. Most crucially, exome sequencing data already exist across multiple large studies and have been compiled under the auspices of ExAC (see URLs)[19]. Here we leveraged this large resource (n ~60,000 participants) to better characterize the rates and properties of rare CNVs, with population frequencies on the order of 1×10^{-2} and as low as 1×10^{-5}. We constructed the ExAC CNV data set 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 the individual being diploid across that gene (having normal copy number state) versus deleted or duplicated. Notably, 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 affected by CNV (that is, where some exons are diploid) rather than having 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 SNVs[20,21] and CNVs[22]. Although large CNVs that affect many genes are likely to be deleterious[23], certain genes will be more sensitive to (intolerant of) dosage changes and thus have fewer CNVs. In this work, we leverage the tens of thousands of exome samples in the ExAC database to estimate genic frequencies for rare CNV. We then calibrate these 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 evolutionary measures of genic constraint. We conclude that considering CNV intolerance

[1]Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [2]Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [4]Analytic and Translational Genetics Unit, Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. [5]A list of members and affiliations appears in the **Supplementary Note**. [6]Department of Psychiatry, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [7]These authors contributed equally to this work. Correspondence should be addressed to D.M.R. (douglas.ruderfer@mssm.edu) or S.M.P. (shaun.purcell@mssm.edu).

Received 17 March; accepted 12 July; published online 17 August 2016; doi:10.1038/ng.3638

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 × 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 (Online Methods). We performed quality control and restricted analysis to genes where each CNV was rare (observed in <600 individuals, corresponding to a maximum allele frequency of ~0.5%). CNV quality was assessed using family trios and demonstrated high specificity and sensitivity consistent with previous reports[17] (Online Methods). Additionally, a subset of 10,091 individuals had high-quality CNV calls from genotyping arrays[24], for whom we assessed the comparability of the CNVs called from genotyping arrays and exome sequencing. The set of array-based CNVs was filtered for high confidence on the basis of number of markers (10), length (>100 kb), and frequency (<1%), as described[24]. Of the most confidently called array-based CNVs—those that were longer and intersected the most coding sequences (greater than 20 targets)—78% were also called in the high-confidence set of exome sequencing CNVs (1,307/1,684). Array-based CNVs intersecting fewer targets were less likely to be called in the exome sequencing set (**Supplementary Fig. 1**), such that 62% of array-based CNVs affecting more than three exons and 54% of all array-based CNVs affecting 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 the CNV increased (**Supplementary Fig. 2**). For the concordantly called CNVs, the 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 data set than in the array-based call set (1.28 versus 0.59 CNVs).

The final ExAC CNV data set 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) intersecting at least one of the 19,430 GENCODE autosomal protein-coding genes (**Fig. 1**). The largest group of 17,565 individuals (29%) carried exactly one rare coding CNV, with 12,812 (21%) carrying zero CNVs and 3,730 (6%) carrying more than five CNVs. The mean extent of CNV per individual was 154 kb (median = 35 kb), representing more duplicated genomic content (107 kb) than deleted content (46 kb). The average length of CNV was 73 kb (median = 15 kb), with duplications being 83 kb (median = 20 kb) and deletions being 56 kb (median = 9 kb). Eighty-four percent of CNVs were smaller than 100 kb, which has generally been used as the size threshold for confidently calling CNVs from genotyping arrays; 56% of CNVs were shorter than 20 kb.

Seventy percent of individuals had at least one gene affected 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 data set (**Fig. 2** and **Table 1**). Sixteen percent of CNVs were greater than 100 kb in length, averaging

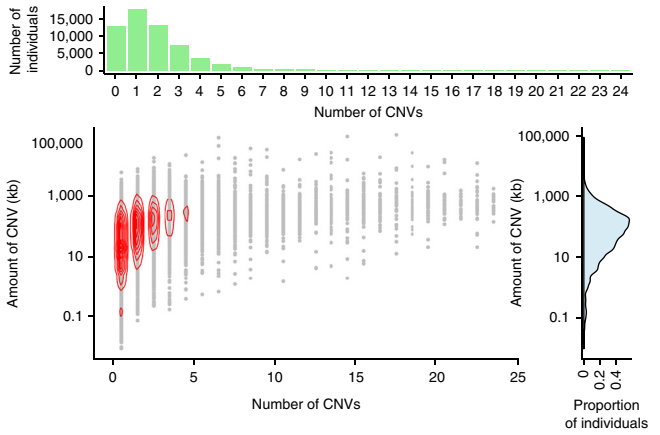

Figure 1 Distribution of the number and amount of CNV across 59,898 exome-sequenced individuals. Plots include a histogram of the number of CNVs per individual (top), a 2D density plot of CNV number and amount (bottom), and a density plot of the amount of CNV per individual (right).

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 for SNV[25]. However, these rates were significantly confounded by variables such as batch and overall read depth, complicating the interpretation of this finding (Online Methods, **Supplementary Figs. 3** and **4**, and **Supplementary Table 1**). As previously reported[26], we identified a significantly higher CNV rate in females, after adjusting for read depth, cohort, and ten principal components of ancestry (mean female CNV rate = 1.74, mean male CNV rate = 1.49, $P = 1.14 \times 10^{-10}$; **Supplementary Table 1**).

On average, each gene was deleted in 3.1 individuals and duplicated in 6.6 individuals. 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). Fifty-five percent 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 (**Fig. 2** and **Table 1**), with some exons deleted or duplicated but also with some exons confidently assigned as diploid (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 predictions from genic constraint scores based on SNVs[19,27,28]. 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 a 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 (Online Methods, **Supplementary Fig. 5**, and **Supplementary Table 2**). Intolerance scores were calculated as 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 Fig. 6**).

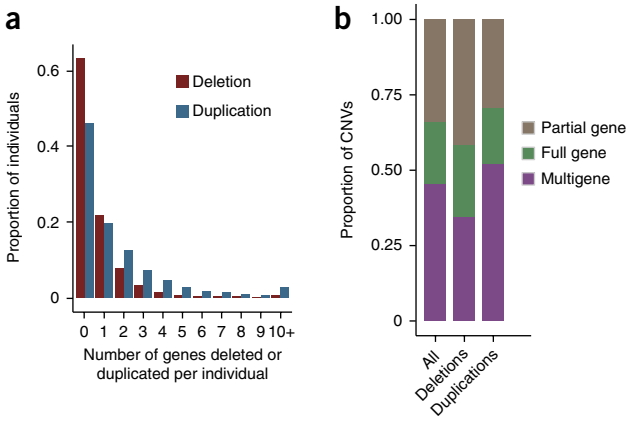

Figure 2 Genic summary of rare deletions and duplications in the ExAC sample. (**a**) Proportion of individuals having from zero to ten or more genes deleted (red) or duplicated (blue). (**b**) Proportion of CNVs that affect multiple genes (multigene), affect the entirety of a single gene (full gene), or partially disrupt a single gene (partial gene). The two rightmost bars show these proportions separately for deletions and duplications.

Intolerance scores based only on deletions were highly correlated with those based only on duplications ($r = 0.37$, $P << 1 \times 10^{-20}$), and both scores were highly correlated with the combined score ($r = 0.7$ for deletions, $r = 0.89$ for duplications; the difference in correlation reflects the greater number of duplications). The results from a complementary approach to predict haploinsufficiency[29] that compared genes sensitive to gene loss to those where having a single copy resulted in no discernable aberrant phenotype demonstrated significant correlation with CNV intolerance scores ($r = 0.12$, $P = 2 \times 10^{-36}$). CNV intolerance scores were also significantly correlated with a measure of genic constraint based on missense SNVs[27] ($r = 0.2$, $P = 2 \times 10^{-137}$) derived from the ExAC sample[19], with this effect being stronger for deletions ($r = 0.23$, $P = 2 \times 10^{-176}$) than for duplications ($r = 0.14$, $P = 1 \times 10^{-63}$). This correlation was consistent across the distribution of scores, with increased CNV intolerance score corresponding to increases in both SNV scores (based on either missense or loss-of-function variants) (**Supplementary Fig. 7**). Similarly, CNV intolerance scores also correlated with an index of haploinsufficiency (pLI)[19] based on loss-of-function variants (nonsense and canonical splice-site SNVs) derived from this sample (all CNVs: $r = 0.18$, $P = 6 \times 10^{-110}$; deletions: $r = 0.23$, $P = 1 \times 10^{-176}$; duplications: $r = 0.11$, $P = 1 \times 10^{-39}$). Unlike for SNV-based scores, CNV intolerance scores will be correlated across multiple genes affected by larger CNVs. We therefore calculated CNV intolerance scores from CNVs that only affected a single gene and identified similar correlations with pLI ($r = 0.22$ for deletions, $r = 0.06$ for duplications). Although 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 scores based on all CNVs going forward but provide both scores for download online (see URLs).

CNV intolerance scores were also associated with an independent measure of evolutionary constraint, GERP[30]. Genes with higher mean per-base GERP scores (calculated including introns) tended to have higher CNV intolerance scores ($r = 0.13$, $P = 5 \times 10^{-46}$). 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 score (CNV intolerance score, $P = 3 \times 10^{-33}$; SNV constraint score from missense variants, $P = 6 \times 10^{-27}$; SNV constraint score from loss-of-function variants, $P = 3 \times 10^{-5}$), suggesting that both CNV- and SNV-based scores contribute non-redundant information regarding the potential deleteriousness of genic CNVs.

Table 1 Number of total genes affected and mean number of gene-level CNVs per individual

	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

The bottom two rows consider only CNVs affecting a single entire gene (single gene) or only part of a gene (partial gene); deletions and duplications are shown separately to the right. Rate is the mean number of CNVs per individual.

Characterizing CNV-tolerant and CNV-intolerant genes

For a particular gene, intolerance of genetic variation such as CNV implies higher functional importance of that gene[19]. We thus considered the relationship between the intolerance of a gene to CNV and its expression across 27 tissues[31], focusing on the 7,754 genes that were 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 than all other genes within this subset (**Fig. 3a**). Notably, genes highly expressed in the brain showed the most intolerance to CNV. Tissues expressing genes that were more intolerant of CNV 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[32] (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 (greater tolerance) than average genes, raising the hypothesis of greater robustness to dosage changes in these tissues.

Genes previously defined as haploinsufficient[29] or essential[33] showed higher CNV intolerance scores in comparison to all genes ($P = 2 \times 10^{-25}$ and 2×10^{-12}, respectively; **Supplementary Table 4**). In contrast, genes implicated in recessive disorders (see URLs) and those corresponding to genes with no identifiable aberrant phenotype when disrupted 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 genes implicated in recessive disorders, similar overall results were recently obtained in an analysis of a large data set of CNVs from genotyping arrays[15] (**Supplementary Table 4**). Applying generic gene set enrichment analysis to the most and least CNV-intolerant genes (top and bottom 5%, 787 genes each; **Fig. 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 to CNV of genes that are highly expressed in brain tissue (neuron development (GO:0048666), $P = 2 \times 10^{-6}$; synapse organization (GO:0050808), $P = 6 \times 10^{-6}$; **Supplementary Tables 5–8**).

Application of CNV intolerance to 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 cases and 6,102 controls[34]) and calculated intolerance scores in the remaining 47,787 individuals as described above. As previously reported[24], this sample of schizophrenia cases showed a higher number of genes affected by CNV than the

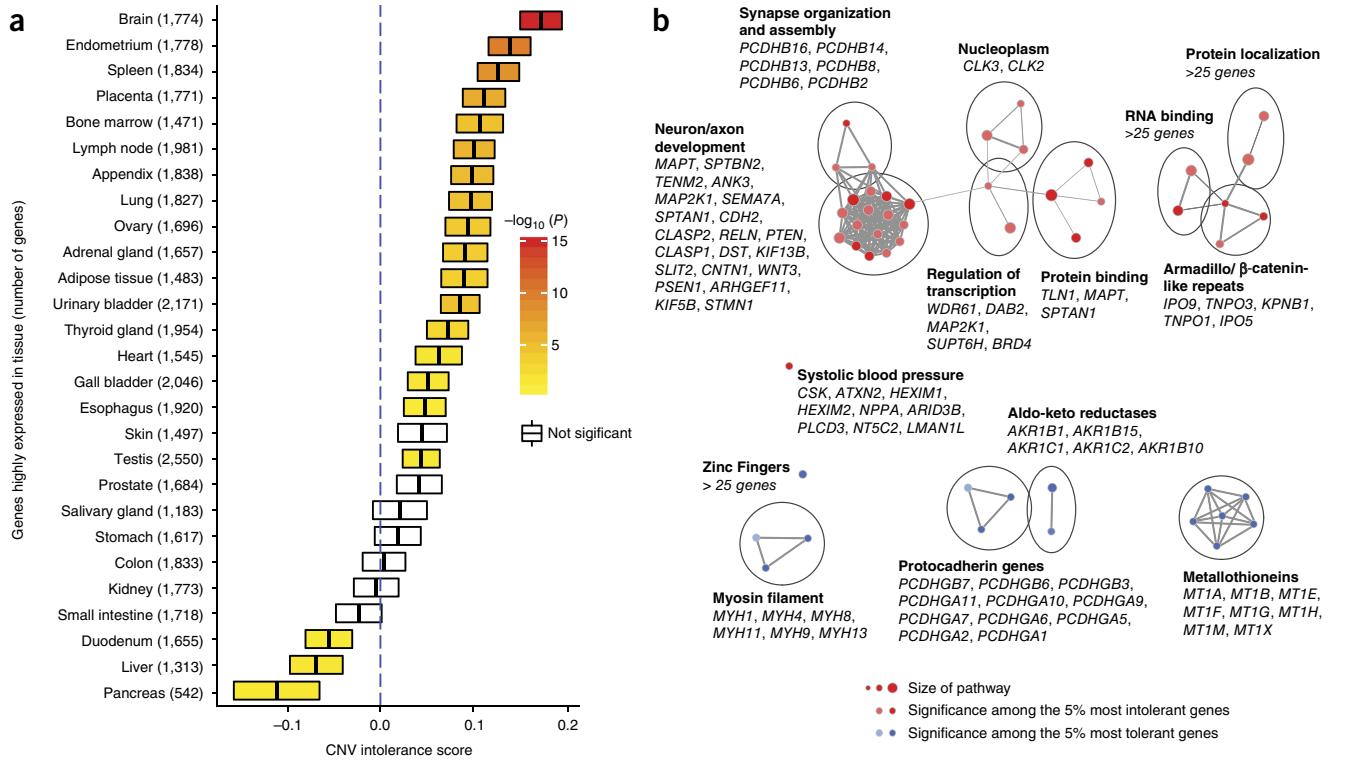

Figure 3 Brain-relevant genes demonstrate the greatest intolerance to dosage changes from CNVs. (**a**) After removing genes highly expressed in all tissues (FPKM >20), 27 tissues[31] were rank-ordered by the mean ExAC CNV intolerance scores for the highly expressed genes in each tissue; mean intolerance scores are indicated by the middle bar in each box, and box width represent s.e.m. Box color denotes the significance of a two-sided *t* test of the difference in intolerance scores between tissue-expressed genes and all others; white bars indicate no significant difference ($P > 0.05$). The 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 the Enrichment Map Cytoscape plugin[35]). Results are based on tests of nine categories of pathways (GO molecular, GO biological, GO cellular, Human Phenotype, Mouse Phenotype, Domain, Pathway, Gene Family, and Disease); only those surpassing Bonferroni-corrected ($P < 0.05$) and false discovery rate (FDR) significance thresholds are shown. Node size represents the number of genes in a pathway, node color represents the significance of enrichment, and the 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 the genes listed were those present in all significant pathways within a group.

controls (2.12 versus 1.78 genes affected by CNV per individual, $P = 1 \times 10^{-10}$). Over and above the number of genes affected, cases carried higher mean intolerance across all genes targeted by CNVs than controls (−1.35 versus −1.42, $P = 0.007$). (Note that, as expected, genes for which we observed any CNV in a given sample in fact tended to be more tolerant; thus, both groups had negative means.) Further, cases carried greater normalized intolerance (Online Methods) for CNVs than controls (0.44 versus 0.33, $P = 1 \times 10^{-11}$). 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 significantly higher constraint in schizophrenia cases than in controls from the constraint score based on missense variants ($P = 4 \times 10^{-4}$), the constraint score based on loss-of-function variants ($P = 2 \times 10^{-4}$), and pLI ($P = 8 \times 10^{-8}$). In a joint test of all scores from independent annotations, the CNV intolerance scores remained the most significant predictor (CNV, $P = 6 \times 10^{-7}$; SNV (missense variants), $P = 0.17$; pLI, $P = 0.004$). This finding suggests that it will be beneficial to develop disease risk association testing frameworks that jointly consider the type of CNVs 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 CNVs. Consistent with the relevance of CNV 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 quality control 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, which 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, although 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 the basis of disease status. Insofar as a minority of genes had significantly higher rates of CNV because of this, these genes will have slightly deflated intolerance estimates in comparison 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 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 of CNV intolerance scores 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.

URLs. Exome Aggregation Consortium (ExAC) web browser, http://exac.broadinstitute.org/; genes implicated in recessive disorders, http://research.nhgri.nih.gov/CGD.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to acknowledge E. Fluder and K. 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.).

AUTHOR CONTRIBUTIONS

D.M.R., M.F., and S.M.P. designed the study. M.L., K.J.K., and D.G.M. handled sample and data management. D.M.R., T.H., K.E.S., M.F., and S.M.P. contributed to statistical analyses. D.K., D.M.R., and K.J.K. designed and implemented website visualizations. D.M.R., M.J.D., D.G.M., M.F., and S.M.P. contributed to primary interpretations. D.M.R., M.F., and S.M.P. performed the primary drafting of the manuscript. All authors contributed to, read, and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. 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).
2. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
3. Jacobs, P.A., Baikie, A.G., Court Brown, W.M. & Strong, J.A. The somatic chromosomes in mongolism. *Lancet* **1**, 710 (1959).
4. Lejeune, J., Turpin, R. & Gautier, M. Chromosomic diagnosis of mongolism. *Arch. Fr. Pediatr.* **16**, 962–963 (1959).
5. Jacobs, P.A., Matsuura, J.S., Mayer, M. & Newlands, I.M. A cytogenetic survey of an institution for the mentally retarded: I. chromosome abnormalities. *Clin. Genet.* **13**, 37–60 (1978).
6. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
7. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
8. Hastings, P.J., Lupski, J.R., Rosenberg, S.M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10**, 551–564 (2009).
9. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
10. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
11. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
12. Buchanan, J.A. & Scherer, S.W. Contemplating effects of genomic structural variation. *Genet. Med.* **10**, 639–647 (2008).
13. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
14. McCarroll, S.A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).
15. Zarrei, M., MacDonald, J.R., Merico, D. & Scherer, S.W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **16**, 172–183 (2015).
16. Plagnol, V. *et al.* A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* **28**, 2747–2754 (2012).
17. Fromer, M. *et al.* Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* **91**, 597–607 (2012).
18. Poultney, C.S. *et al.* Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. *Am. J. Hum. Genet.* **93**, 607–619 (2013).
19. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* http://dx.doi.org/10.1038/nature19057 (2016).
20. Fu, W. & Akey, J.M. Selection and adaptation in the human genome. *Annu. Rev. Genomics Hum. Genet.* **14**, 467–489 (2013).
21. Purcell, S.M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
22. 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).
23. Need, A.C. *et al.* A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS Genet.* **5**, e1000373 (2009).
24. Szatkiewicz, J.P. *et al.* Copy number variation in schizophrenia in Sweden. *Mol. Psychiatry* **19**, 762–773 (2014).
25. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
26. Desachy, G. *et al.* Increased female autosomal burden of rare copy number variants in human populations and in autism families. *Mol. Psychiatry* **20**, 170–175 (2015).
27. Samocha, K.E. *et al.* A framework for the interpretation of *de novo* mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
28. 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).
29. Huang, N., Lee, I., Marcotte, E.M. & Hurles, M.E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* **6**, e1001154 (2010).
30. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
31. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* **13**, 397–406 (2014).
32. Sulem, P. *et al.* Identification of a large set of rare complete human knockouts. *Nat. Genet.* **47**, 448–452 (2015).
33. 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).
34. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45**, 1150–1159 (2013).
35. 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).

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 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 capture methods 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) over >25% of its span, 225 for having GC content <10% or >90%, 1,582 for covering <10 bp, and 15 for spanning >10 kb. The resulting sample × target read depth matrix was scaled by mean-centering the targets, after which principal-component analysis 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, three targets were removed for still having high variance after normalization (s.d. >50), and sample-level z scores were calculated (with absolute values capped at 40). CNVs were called using the Viterbi 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 CNVs were statistically genotyped across all samples using the same XHMM quality scores and output as a single uniformly called VCF file.

Quality control of CNV data. In total, we attempted CNV calling for 60,642 of the 60,706 (99.9%) ExAC samples, with the remainder either having failed calling for low overall read depth or not included because of upstream data access issues. The CNVs output by XHMM were first frequency filtered to remove common CNVs, that is, those seen more than 600 times (>1%), defined as overlapping more than 50% of their respective targets. On the basis of previous work[17], we retained only CNVs with quality scores greater than or equal to 60. We removed any individual having a CNV count greater than 3 s.d. above the mean, that is, 24 CNVs (n = 775 samples removed). Thus, our final data set 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, with CNVs on sex chromosomes removed because of technical issues. Next, we removed any gene where half or more of its targets were filtered out during CNV calling (1,068 genes). 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 for 849 genomes sequenced from the 1000 Genomes Project[36], we removed any gene known to be multiallelic (735 genes). Finally, we removed any gene in which there existed any CNV with a frequency greater than 0.5% (1,193 genes). This filtering 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 used 241 previously described[37,38] 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 data set, which included only unrelated individuals). Poor sensitivity would result in severely reduced transmission statistics, whereas poor specificity would induce many false positive CNV calls and increased rates of *de novo* CNV. Through reasonable estimates of transmission and *de novo* events, we could infer high specificity and sensitivity for the CNV calls overall. Defining CNV transmission as implemented in the PLINK/Seq cnv-denovo command[17], we assessed whether the rate of transmission for CNVs 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 excluded situations where an offspring's CNV was neither confidently called as deleted or duplicated (SQ ≥ 60) nor confidently called as diploid (DQ ≥ 60). Including these more uncertain events and conservatively counting them as non-transmissions resulted 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 in length and spanned five or more exons), which only increased 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[38].

Gene- and exon-specific copy number calls. We defined gene-specific copy number state for each individual, assessing the probability of a CNV occurring anywhere between the transcriptional start and end sites of a gene. Specifically, this analysis was performed by defining the genomic interval spanned by each gene and then using the sample × target matrix of z scores 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 called 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. Although XHMM performs exome-wide correction for both regional and individual read depth variability, we found that increased sample read depth was still correlated with increased numbers of CNVs (**Supplementary Fig. 1**). In the absence of large-scale validation efforts and given the focus on CNVs that were rare at any particular locus, it was 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 for diploid quality of 3 s.d. below the mean for 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 genic exons and overlapping these with all targets at which read depth was calculated and normalized; note that this analysis typically included a single target per exon, but for a small proportion of exons two or more targets were included because of the slight difference in the definition of target regions for CNV calling and GENCODE exonic regions. Genic CNV counts derived from this procedure correlated with the number of loss-of-function variants in a gene (**Supplementary Fig. 7**).

Creating genic CNV intolerance scores. For the 15,734 genes that survived quality control, 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 be less than 80% overlapping and less than 4 Mb in length. We identified a significant increase in the number of CNVs in genes within these regions (**Supplementary Fig. 1**), so we included this 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 located between, along with higher-order terms. We next calculated z scores such that positive values represent a lower frequency of CNV (more intolerance), winsorizing the negative tail at 5%.

Stratifying CNVs by genic content affected. We stratified CNVs by the number of genes and exons for which they were called (confidently) to affect dosage. Specifically, we defined single-gene CNVs as those with a gene-specific

doi:10.1038/ng.3638

confidence score greater than 60 in one of the 15,734 genes that remained after gene quality control, but we also strictly required overlap with only one of the 19,430 GENCODE autosomal protein-coding genes. CNVs overlapping more than one gene were labeled as multigene CNVs. Using the exon-level CNV calls, we further refined our single-gene CNVs into three classes: (i) full-gene CNVs corresponded to genes where all exons were confidently called as deleted or duplicated; (ii) ambiguous CNVs corresponded to genes with at least one exon confidently called as deleted or duplicated but no exons confidently called as diploid; and (iii) partial-gene CNVs corresponded to genes in which there was at least one exon confidently called as 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 was 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[31]. We defined highly expressed genes for each tissue as those having FPKM values greater than 20 but excluded genes that were highly expressed in all tissues. The second was a set of disease-implicated genes collated in a previous study analyzing a large set of CNVs[15]; these included sets of genes involved in dominant and recessive diseases, genes implicated in cancer, haploinsufficient genes, genes whose counterparts are essential in mice, genes intolerant to loss-of-function variants, and genes not related to a specific phenotype in any such database (**Supplementary Tables 3** and **4**).

Gene set enrichment analysis. We selected the genes with the top and bottom 5% of CNV intolerance scores ($n = 787$ each set) and ran gene set enrichment analysis with ToppFun[39], 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 CNV-intolerant genes were enriched in GO sets related to neuronal and axon development and synapse organization and assembly. The most CNV-tolerant genes were enriched for metallothioneins and myosin filament genes (**Fig. 2b** and **Supplementary Tables 5–8**).

36. Handsaker, R.E. *et al.* Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).
37. Kirov, G. *et al. De novo* CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol. Psychiatry* **17**, 142–153 (2012).
38. Fromer, M. *et al. De novo* mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
39. Chen, J., Bardes, E.E., Aronow, B.J. & Jegga, A.G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305–W311 (2009).