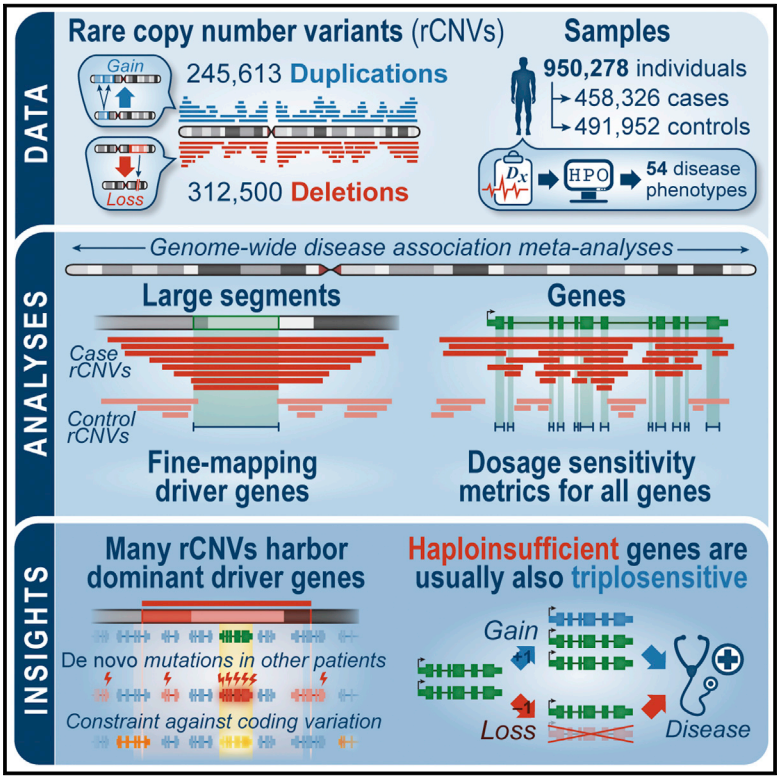
Cell

Resource

A cross-disorder dosage sensitivity map of the human genome

Graphical abstract



Highlights

- Meta-analysis of rare copy-number variants (rCNVs) in nearly one million humans
- Discovered hundreds of rCNV-disease associations across 54 disorders
- Convergence of rCNVs & damaging coding variants at dosage sensitive loci
- Ensemble machine learning identified 3,635 highly dosage sensitive genes

Authors

Ryan L. Collins, Joseph T. Glessner, Eleonora Porcu, ..., Shamil Sunyaev, Harrison Brand, Michael E. Talkowski

Correspondence

rlcollins@g.harvard.edu (R.L.C.), hbrand@broadinstitute.org (H.B.), talkowsk@broadinstitute.org (M.E.T.)

In brief

Harmonizing genomic data from nearly one million individuals yields insights into the properties of rare copy-number variants across disorders and dosage sensitivity predictions for all autosomal protein-coding genes.

Collins et al., 2022, Cell *185*, 1–15
August 4, 2022
https://doi.org/10.1016/j.cell.2022.06.036

CellPress

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

Cell

CellPress

Resource

A cross-disorder dosage sensitivity map of the human genome

Ryan L. Collins,[1,2,3,*] Joseph T. Glessner,[4,5] Eleonora Porcu,[6,7] Maarja Lepamets,[8,9] Rhonda Brandon,[10] Christopher Lauricella,[10] Lide Han,[11] Theodore Morley,[11] Lisa-Marie Niestroj,[12] Jacob Ulirsch,[2,3,13] Selin Everett,[1,2] Daniel P. Howrigan,[1,2,13] Philip M. Boone,[1,2,14,15] Jack Fu,[1,2,14] Konrad J. Karczewski,[1,2,13] Georgios Kellaris,[16,17] Chelsea Lowther,[1,2,14] Diane Lucente,[1] Kiana Mohajeri,[1,2,3] Margit Nõukas,[8,9] Xander Nuttle,[1,2,14] Kaitlin E. Samocha,[1,2,3,18] Mi Trinh,[18] Farid Ullah,[16,17] Urmo Võsa,[8] Epi25 Consortium, Estonian Biobank Research Team, Matthew E. Hurles,[18] Swaroop Aradhya,[10,28] Erica E. Davis,[16,17] Hilary Finucane,[1,2,13] James F. Gusella,[1,2] Aura Janze,[10]

(Author list continued on next page)

[1]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
[2]Program in Medical and Population Genetics, Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA 02142, USA
[3]Division of Medical Sciences and Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
[4]Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[5]Department of Pediatrics, Division of Human Genetics, Perelman School of Medicine, Philadelphia, PA 19104, USA
[6]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland
[7]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
[8]Estonian Genome Centre, Institute of Genomics, University of Tartu, 51010 Tartu, Estonia
[9]Institute of Molecular and Cell Biology, University of Tartu, 51010 Tartu, Estonia
[10]GeneDx, Gaithersburg, MD 20877, USA
[11]Division of Genetic Medicine, Department of Medicine, and Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN 37232, USA
[12]Cologne Center for Genomics, University of Cologne, 51149 Cologne, Germany
[13]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA
[14]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA
[15]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA 02115, USA
[16]Advanced Center for Translational and Genetic Medicine, Stanley Manne Children's Research Institute, Lurie Children's Hospital, Chicago, IL 60611, USA
[17]Departments of Pediatrics and Cell and Developmental Biology, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA
[18]Human Genetics Programme, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10, UK
[19]Avenue 80, Inc., Fort Wayne, IN 46802, USA

(Affiliations continued on next page)

SUMMARY

Rare copy-number variants (rCNVs) include deletions and duplications that occur infrequently in the global human population and can confer substantial risk for disease. In this study, we aimed to quantify the properties of haploinsufficiency (i.e., deletion intolerance) and triplosensitivity (i.e., duplication intolerance) throughout the human genome. We harmonized and meta-analyzed rCNVs from nearly one million individuals to construct a genome-wide catalog of dosage sensitivity across 54 disorders, which defined 163 dosage sensitive segments associated with at least one disorder. These segments were typically gene dense and often harbored dominant dosage sensitive driver genes, which we were able to prioritize using statistical fine-mapping. Finally, we designed an ensemble machine-learning model to predict probabilities of dosage sensitivity (pHaplo & pTriplo) for all autosomal genes, which identified 2,987 haploinsufficient and 1,559 triplosensitive genes, including 648 that were uniquely triplosensitive. This dosage sensitivity resource will provide broad utility for human disease research and clinical genetics.

INTRODUCTION

Deletions and duplications of genomic segments, collectively known as copy-number variants (CNVs), have been recognized as important evolutionary mechanisms for over 50 years (Ohno, 1970); however, examples of CNVs conferring adaptive advantages in humans are scarce (Almarri et al., 2020). Instead, most large CNVs in humans are held at low frequencies across

Nicholas Katsanis,[16,17] Ludmila Matyakhina,[10] Benjamin M. Neale,[1,2,13] David Sanders,[19] Stephanie Warren,[10] Jennelle C. Hodge,[20] Dennis Lal,[12,21,22] Douglas M. Ruderfer,[11,23] Jeanne Meck,[10] Reedik Mägi,[8] Tõnu Esko,[8] Alexandre Reymond,[6] Zoltán Kutalik,[7,24,25] Hakon Hakonarson,[4,5] Shamil Sunyaev,[2,3,26] Harrison Brand,[1,2,14,27,*] and Michael E. Talkowski[1,2,13,14,29,*]

[20]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA
[21]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA
[22]Epilepsy Center, Neurological Institute, Cleveland Clinic, Cleveland, OH 44195, USA
[23]Center for Precision Medicine, Department of Biomedical Informatics, and Department of Psychiatry and Behavioral Sciences, Vanderbilt University Medical Center, Nashville, TN 37232, USA
[24]Center for Primary Care and Public Health, University of Lausanne, 1015 Lausanne, Switzerland
[25]Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland
[26]Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115, USA
[27]Pediatric Surgical Research Laboratories, Massachusetts General Hospital, Boston, MA 02114, USA
[28]Present address: Invitae, San Francisco, CA 94103, USA
[29]Lead contact
*Correspondence: rlcollins@g.harvard.edu (R.L.C.), hbrand@broadinstitute.org (H.B.), talkowsk@broadinstitute.org (M.E.T.)
https://doi.org/10.1016/j.cell.2022.06.036

global populations by strong purifying selection (Collins et al., 2020; Itsara et al., 2010). These rare CNVs (rCNVs; frequency <1%) have been widely associated with Mendelian and complex diseases (Zhang et al., 2009). A subset of disease-associated rCNVs, known as "genomic disorders" (GDs), have been prominent in the literature for decades (Lupski, 2009; Tommerup, 1993). GDs are sites of recurrent rCNVs and are often formed by non-allelic homologous recombination (Carvalho and Lupski, 2016; Sharp et al., 2005). Several dozen GDs have been reported, including "reciprocal" GDs where deletions and duplications of the same locus have been independently associated with disease (Lupski, 2009; Weiss et al., 2008). Many GDs exhibit variable phenotypes, but they collectively comprise one of the most common genetic causes of abnormal neurodevelopment (Girirajan et al., 2012) and can have subtle effects on other traits like height and blood pressure even in the absence of disease (Auwerx et al., 2022; Owen et al., 2018). Some reciprocal GDs have been linked to "mirror" phenotypes, where increased or decreased dosage leads to opposing phenotypes, like microcephaly and macrocephaly in the 16p11.2 GD (Golzio et al., 2012; Jacquemont et al., 2011). These mirror phenotypes suggest that one or more genes or elements within these GDs may be dosage sensitive (DS) "drivers" of their associated phenotypes. Despite their clear impact on morbidity, the pathogenic mechanisms of most rCNVs remain unclear for several reasons. Many large rCNVs encompass millions of nucleotides and dozens of genes, obfuscating their critical driver(s). Large rCNVs can also have myriad indirect consequences, including regulatory, polygenic, or epistatic effects (Girirajan and Eichler, 2010). Finally, CNVs have a lower (≥100-fold) mutational density than single-nucleotide variants (1000 Genomes Project Consortium et al., 2015), meaning that well-powered rCNV association studies require massive sample sizes.

The simplest explanation for the effects of most rCNVs is that they presumably delete or duplicate one or more critical gene(s) driving their associated phenotypes. However, the identification of DS driver genes within rCNVs has been historically challenging. Sensitivity to decreased DNA dosage (i.e., haploinsufficiency) or increased DNA dosage (i.e., triplosensitivity) has been documented for individual genes (Bragin et al., 2014; Riggs et al., 2018), although genome-wide annotations of DS genes and segments remain incomplete. Prior studies have estimated selection against coding CNVs and developed methods to prioritize likely pathogenic CNVs (Aguirre et al., 2019; Huang et al., 2010; Ruderfer et al., 2016), but there are no widely adopted frameworks to evaluate both haploinsufficiency and triplosensitivity for every human gene. Consequently, existing lists of DS loci exceeding genome-wide significance thresholds or meeting robust guidelines for clinical interpretation are limited: for example, the ClinGen Genome Dosage Map includes just 15 triplosensitive (TS) genes (Riggs et al., 2018). Similarly, it remains mostly unknown whether the deletion- and duplication-associated phenotypes for reciprocal GDs are driven by the same bidirectionally DS gene or by two (or more) independent haploinsufficient (HI) and TS genes. Comprehensive maps of bidirectional dosage sensitivity across disorders are therefore critically needed for human disease research and clinical interpretation.

In this study, we meta-analyzed large rCNV data from 950,278 individuals to build a genome-wide catalog of rCNV associations for 54 disease phenotypes. We also integrated these rCNVs with 145 genome annotations to predict the probability of haploinsufficiency (pHaplo) and probability of triplosensitivity (pTriplo) for all protein-coding genes, allowing us to define 3,635 high-confidence DS genes. We provide all maps and metrics derived in this study as an open resource for the community and anticipate that they will have broad utility for human genomic research and medical genetics.

RESULTS

A catalog of large rCNVs across 54 disorders

We aggregated rCNVs ascertained by microarrays from 17 sources, ranging from diagnostic laboratories to national biobanks (Table S1). To account for technical heterogeneity across sources, we developed a strict harmonization procedure to retain large (≥100 kb) sub-chromosomal (≤20 Mb) CNVs appearing at <1% frequency in every source in our dataset and in every global population documented by four genome sequencing-based CNV reference maps (Abel et al., 2020; Almarri et al., 2020; Byrska-Bishop et al., 2021; Collins et al., 2020). This

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

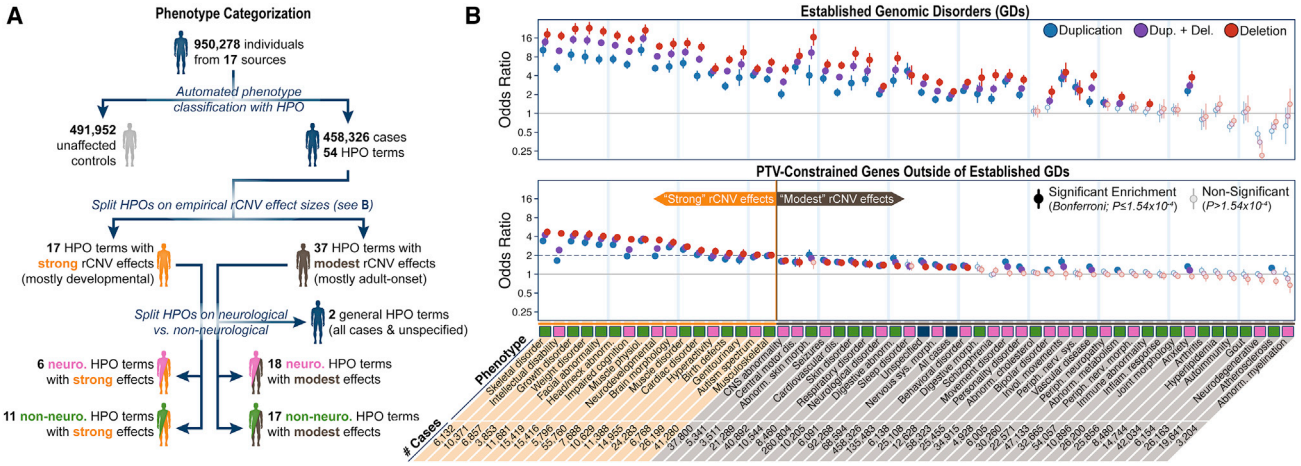


Figure 1. The contribution of rCNVs to 54 disease phenotypes

(A) Phenotype categorization for 950,278 samples using Human Phenotype Ontology.

(B) ORs per phenotype from meta-analyses of rCNVs matching 95 GDs reported in the literature (top) and of rCNVs impacting PTV-constrained genes outside of known GDs (bottom). Error bars indicate 95% confidence intervals.

See also Tables S1 and S2.

process reduced technical variability between sources by nearly two orders of magnitude (Figure S1), with the final harmonized dataset including a total of 558,113 rCNVs in 950,278 individuals. To control for residual heterogeneity, we further grouped sources into seven cohorts based on their technical similarities, such as microarray platform and sample recruitment strategy.

The extent of phenotypic data also varied between sources, ranging from the presence of a single primary phenotype in disease-specific cohorts to phenome-wide metadata collected by many biobanks. Therefore, we first standardized all phenotypes by mapping all available phenotype data per sample onto the Human Phenotype Ontology (Köhler et al., 2019) with a fuzzy keyword-matching approach. We next performed recursive hierarchical clustering to define a minimal set of non-redundant primary phenotypes that each included a minimum of >300 samples in at least three independent cohorts, >3,000 samples in total across all cohorts, and had less than 80% sample overlap with any other phenotype. This process yielded a total of 54 disease phenotypes, including 24 neurological, 28 non-neurological, and two general categories to capture broadly defined or nondescript phenotypes (Figure 1A; Table S2). Although imperfect, this strategy partitioned our dataset into 458,326 samples matching one or more primary disease phenotype (i.e., "cases") and 491,952 samples not matching any of the 54 disease phenotypes (i.e., "controls").

Having curated rCNV and phenotype data from nearly one million individuals, we evaluated the rCNV contributions to the genetic architectures of these 54 phenotypes. We first curated a reference set of 95 established GDs from six prior surveys (Dittwald et al., 2013; Firth et al., 2009; Girirajan et al., 2012; Owen et al., 2018; Riggs et al., 2012; Stefansson et al., 2014) and computed the prevalence of each GD per phenotype in our harmonized rCNV data. As anticipated, most phenotypes (43/54; 89.6%) exhibited significant associations with GD rCNVs after correcting for multiple comparisons ($p \leq 1.54 \times 10^{-4}$), with the strongest effects found in severe pediatric-onset disorders (Figure 1B). We next explored the extent to which rCNVs outside of GDs contributed disease risk by focusing on genes constrained against protein-truncating variants (PTVs) in the general population as defined by metrics like the loss-of-function (LoF) observed-expected upper bound fraction (LOEUF; ≤ 0.38) (Karczewski et al., 2020). We found significant disease risk contributed by rCNVs impacting PTV-constrained genes for most phenotypes (39/54; 72.2%) even after excluding all GDs, although the effect size of constrained gene rCNVs outside of GDs was less than GDs (GD median odds ratio [OR] = 3.38 versus constrained median OR = 1.38). Finally, we used these empirical effect size estimates to partition all phenotypes into two subsets: those with stronger rCNV effects ("strong-effect phenotypes"; n = 17) that had OR $\geq$ 2.0 for deletions of constrained genes outside of established GDs, and those with weaker rCNV effects ("modest-effect phenotypes"; n = 37) that had OR < 2.0 for constrained gene deletions. This simple approach allowed us to subdivide our dataset by approximate phenotypic severity for downstream analyses.

Genome-wide discovery of rCNV-disease associations

Building on decades of seminal studies of CNV in disease, we leveraged our sample size to systematically identify rCNV associations for every phenotype. We divided all 22 autosomes into 200 kb sliding windows in 10 kb steps and performed an association meta-analysis of rCNVs per phenotype for each 200 kb window while accounting for technical differences between cohorts like microarray platform or probe density (Figure 2A; Figure S2). We assessed significance at two thresholds: (1) genome-wide significance corresponding to the number of non-overlapping windows tested ($p = 3.74 \times 10^{-6}$) and (2) Benjamini-Hochberg false discovery rate (FDR) <1%. We further required nominal evidence ($p < 0.05$) in at least two independent cohorts to mitigate ascertainment and/or platform-specific

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

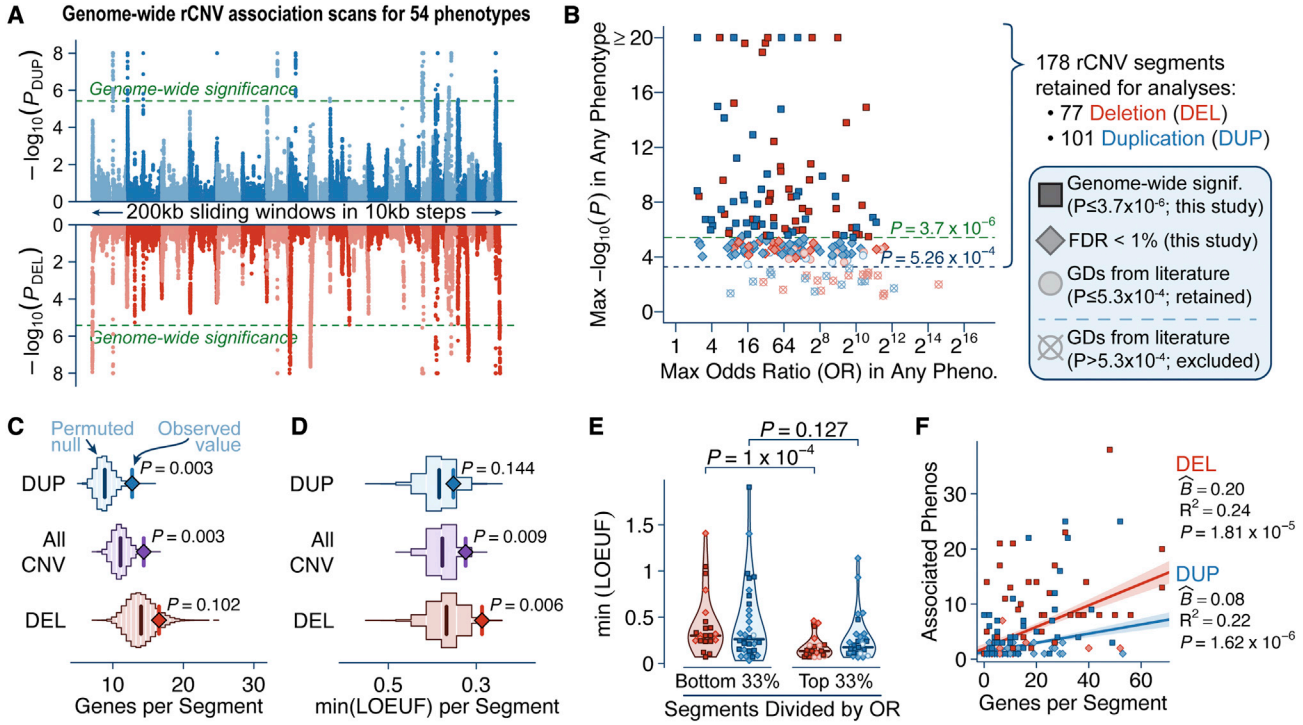


Figure 2. Characteristics of disease-associated rCNV segments

(A) rCNV association statistics for one example phenotype, neurodevelopmental abnormalities (HP:0012759).

(B) Relationship between effect size and strength of association for the 163 segments from our discovery meta-analyses as well as 42 GDs reported in the literature that did not reach FDR < 1% in our discovery analysis.

(C) Our consensus set of 178 disease-relevant rCNV segments overlapped 44% more genes than expected based on 100,000 random permutations (one-tailed permutation test).

(D) rCNV segments overlapped genes under 34% greater constraint against PTVs than expected by permutations (one-tailed permutation test).

(E) Segments in the top third of all effect sizes overlapped genes under stronger constraint than segments in the bottom third of effect sizes (two-tailed Wilcoxon tests).

(F) The number of genes per segment was related to the number of phenotypes associated with each segment. Trend lines are outlier-robust linear fits with 95% confidence intervals.

See also Figures S2, S3, and S5 and Table S3.

signals. Finally, we refined each significant association to the minimal region expected to contain the causal element(s) by adapting a Bayesian algorithm to define the 95% credible interval(s) per associated locus (Wakefield, 2009). In total, this approach discovered 795 rCNV-phenotype associations corresponding to 163 distinct large rCNV "segments" (i.e., genomic intervals where rCNVs were enriched in cases) after clustering individual phenotype-rCNV associations across all phenotypes (median size = 740 kb; 69 deletions & 94 duplications; Figure S3; Table S3).

We next assessed whether we had captured *bona fide* disease associations by cross-examining the 163 significant rCNV segments from our meta-analyses versus the list of 95 GDs that we curated from the literature. Across all literature-based GDs, most (53/95; 55.8%) were recapitulated in our meta-analyses, which was 6.9-fold more than expected by chance based on 100,000 random permutations (Figure S4A; $p < 10^{-5}$, one-tailed permutation test), whereas another 15 GDs exceeded a Bonferroni-corrected significance threshold when testing the 95 GDs selected *a priori* ($p \leq 5.26 \times 10^{-4}$). The remaining 27 literature-based GDs were either exceedingly rare in the population and/or exhibited incompletely penetrant effects too weak to be detected by our meta-analyses (Figure 2B). To remain comprehensive in our subsequent analyses, we constructed a consensus set of 178 disease-associated rCNV segments comprising the 163 segments from the discovery meta-analyses and the 15 additional GDs surpassing $p \leq 5.26 \times 10^{-4}$ in the targeted analyses (Table S4).

A combined analysis of these 178 large rCNV segments revealed that the presence of individual DS genes was a key feature distinguishing such segments from the rest of the genome. Most rCNV segments were gene dense, overlapping 44% more genes than expected based on 100,000 random permutations (Figure 2C; observed = 12.7 genes; expected = 8.6 genes; p = 0.003, one-tailed permutation test). Even after conditioning on their relatively higher gene density, these 178 segments were significantly more likely to overlap phenotype-matched disease genes than expected by random permutations (Figure S4B; Table S5; $p = 3.8 \times 10^{-4}$, one-tailed permutation test) (Amberger et al., 2015) and were enriched for genes under

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

strong mutational constraint: the 77 deletion segments in our consensus set overlapped genes that were 34% more depleted for PTVs in the general population (Karczewski et al., 2020) than expected by chance (Figure 2D; average minimum LOEUF per segment: observed = 0.25, expected = 0.37; p = 0.006, one-tailed permutation test). These trends were even more pronounced when segments were stratified based on their penetrance. In comparison to segments in the bottom third of the effect size distribution ("incompletely penetrant segments"; mean OR = 10.3), segments in the top third of effect sizes ("highly penetrant segments"; mean OR = 444.0) overlapped more genes on average (Figure S4C; $p = 3.34 \times 10^{-5}$, two-tailed Wilcoxon test) and were enriched for genes under stronger constraint (Figure 2E; Figure S4D; $p = 1.46 \times 10^{-4}$, two-tailed Wilcoxon test). This observation remained significant for highly penetrant deletion segments even after accounting for the total number of genes per segment (Figure S4E; p = 0.009, one-tailed permutation test), whereas this was not true for incompletely penetrant deletion segments (p = 0.126). Finally, we discovered a strong relationship between the number of genes per segment and the number of associated phenotypes (Figure 2F; deletions, $p = 1.81 \times 10^{-5}$; duplications, $p = 1.62 \times 10^{-6}$; linear regression); pleiotropic segments (i.e., segments associated with multiple phenotypes) were also more likely to overlap strongly constrained genes compared with segments associated with just one phenotype (Figure S4F; $p = 8.66 \times 10^{-4}$, two-tailed Wilcoxon test). Collectively, these results supported a model where the penetrance of individual rCNVs is influenced by both rCNV size and the properties of individual DS genes within the rCNV. However, we also noted that 6.7% (12/178) of segments did not overlap any annotated protein-coding sequence. The interpretation of these noncoding rCNV segments was more challenging, but we were able to identify a plausible regulatory mechanism for one noncoding segment: hyperactivity-associated noncoding deletions that overlapped a validated recursive splice site within the first intron of *CADM2*, a known common-variant risk locus for behavioral disorders (Ibrahim-Verbaas et al., 2016; Sanchez-Roige et al., 2019; Sibley et al., 2015).

If the presence of at least one DS gene was a common feature of many disease-associated rCNV segments, we reasoned that this trend should be confirmed by damaging point mutations in independent patients with related phenotypes, as has been previously proposed (Coe et al., 2014). To test this hypothesis, we focused on the subset of 93 rCNV segments (48 deletion and 45 duplication) that were associated with at least one neurological phenotype with stronger rCNV effects (see Figure 1A) and overlapped at least one gene. We cross-examined these 93 rCNV segments versus two datasets of damaging *de novo* mutations (DNMs; PTVs and missense) from exome-sequencing studies of developmental disorders (Kaplanis et al., 2020) and autism spectrum disorder (ASD) (Fu et al., 2021), collectively comprising 46,094 parent-child trios with probands affected by broadly defined neurodevelopmental disorders (NDDs; Figure 3A). In both cohorts, we found that the genes in these 93 segments accrued more damaging DNMs in NDD probands than expected after accounting for gene-specific mutation rates (Figures 3B and 3C; Figure S5). These enrichments exhibited clear biases by mutational consequence: PTVs were nominally enriched in deletion segments (p = 0.028, one-tailed permutation test) but not duplication segments (p = 0.488), whereas missense DNMs were enriched in duplication segments (p = 0.002) but not in deletion segments (p = 0.287). Moreover, the distributions of damaging DNMs in NDD probands were highly non-uniform within many rCNV segments (Figures 3D and 3E; Figures S4G and S4H). When focusing on the subset of segments (70/93) with at least two more PTV or missense DNMs than expected based on known mutation rates, we found that 38% of deletion segments (12/32) and 16% of duplication segments (6/38) had their excesses of PTVs or missense DNMs almost entirely ($\geq$90%) concentrated in just one or two genes, and these recurrently mutated genes usually corresponded to established developmental disorder genes (Figures 3F and 3G). To assess how much of this signal was driven by known versus currently unrecognized disease genes, we excluded all 270 genes that had been associated with NDDs by prior DNM analyses in these same cohorts (Fu et al., 2021; Kaplanis et al., 2020). Removing these 270 NDD genes ablated the enrichment of PTV DNMs in deletion segments (p = 0.203, one-tailed permutation test) but surprisingly had no meaningful impact on the enrichment of missense DNMs within duplication segments ($p < 10^{-5}$; Figures 3B and 3C; Figure S5). Collectively, these results indicated that one-quarter (18/70; 26%) of NDD-associated GDs with sufficiently dense DNM data to be evaluated in our analyses harbored one or more dominant driver gene(s) and that many of the dominant driver genes acting through mechanisms other than haploinsufficiency (e.g., triplosensitivity) may be unrecognized at present but could be discovered in future studies.

Fine-mapping individual dosage sensitive genes within large rCNVs

We next sought to identify individual genes enriched for coding rCNVs in cases over controls by conducting exome-wide rCNV association tests. We meta-analyzed exonic rCNVs in cases and controls per phenotype for 17,263 autosomal protein-coding genes using methods similar to our sliding window analyses and assessed significance at two thresholds: exome-wide significance ($p = 2.90 \times 10^{-6}$) and FDR < 1%, again requiring nominal evidence ($p < 0.05$) in at least two independent cohorts. These meta-analyses identified a total of 5,680 significant gene-phenotype associations across 739 unique genes; however, given that many large rCNVs deleted or duplicated multiple adjacent genes, we expected that most of these associated genes were unlikely to be causal but instead simply carried to significance due to their proximity to true causal genes, analogous to the influence of linkage disequilibrium in genome-wide association studies (GWASs) of common variation (Tam et al., 2019). To address this problem, we adapted a Bayesian fine-mapping algorithm to define the 95% credible set of genes at each associated locus separately for deletions and duplications while also prioritizing the most likely causal gene(s) based on their association statistics and 145 gene-level annotations (Kichaev et al., 2014; Wen et al., 2017; Figure 4A; Table S5). Fine-mapping reduced the average number of candidate genes per locus by 48%, resulting in a total of 115 credible sets averaging 4.2 genes each (range: 1–28 genes) (Figures 4B and 4C; Table S6), and prioritized 31 "highly confident" and 90 "confident" genes with

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

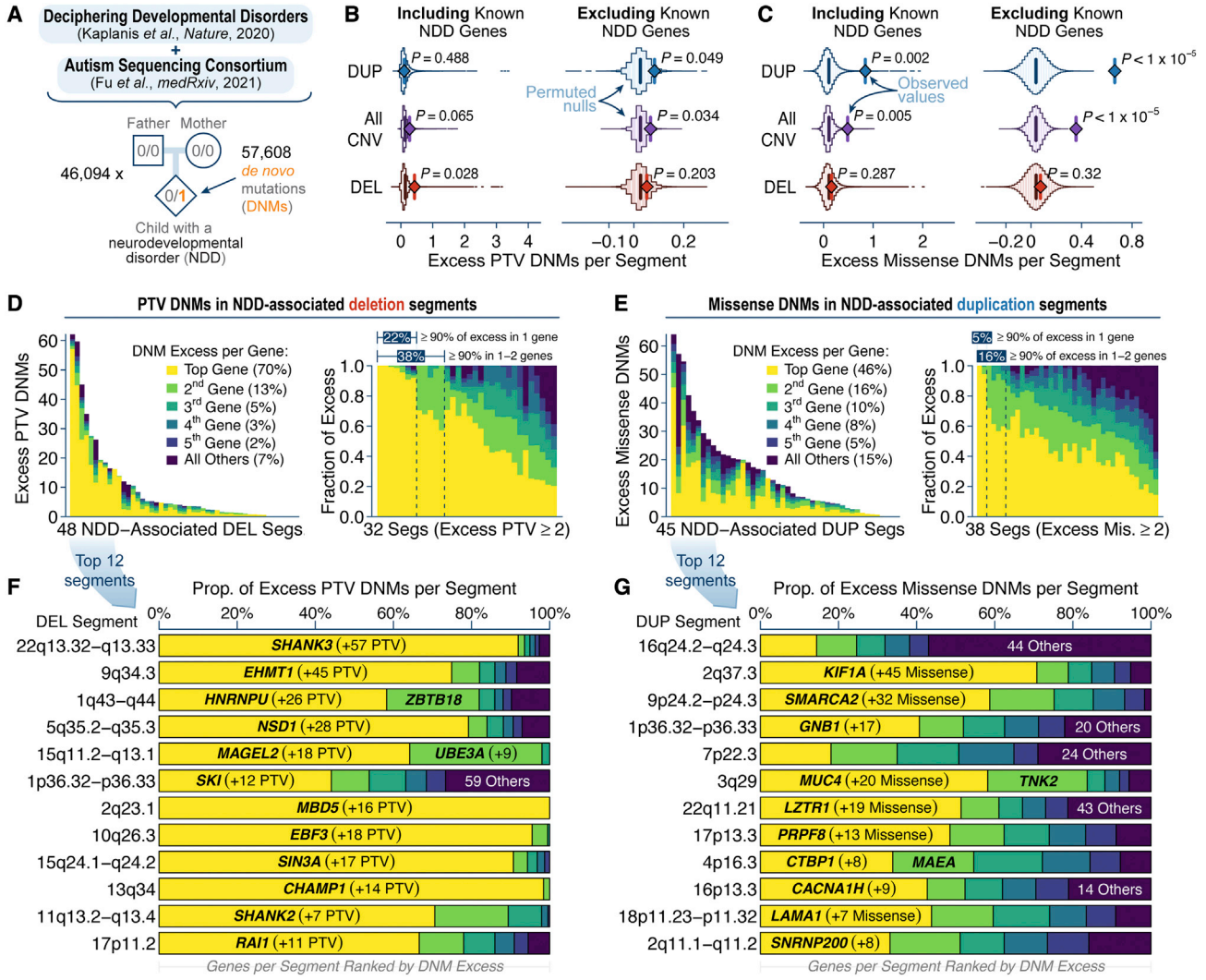


Figure 3. Coding DNMs pinpoint dominant driver genes within rCNVs

(A) We aggregated *de novo* mutations (DNMs) from two exome-sequencing studies of neurodevelopmental disorders (NDDs) (Fu et al., 2021; Kaplanis et al., 2020).

(B) NDD-associated deletion segments were enriched for PTV DNMs beyond expectations based on 100,000 matched permutations adjusted for gene-specific mutation rates (one-tailed permutation tests). This enrichment was ablated after excluding the 270 NDD-associated genes identified in the studies from (A).

(C) Duplication segments associated with NDDs were enriched for missense DNMs, and this enrichment persisted after excluding the 270 known NDD-associated genes (one-tailed permutation tests).

(D and E) The distributions of (D) PTV DNMs per deletion segment and (E) missense DNMs per duplication segment were highly non-uniform. Genes in each segment have been ranked and colored according to their excess PTV DNMs; percentages indicate what fraction of total excess PTV DNMs is attributable to each gene rank across all segments.

(F and G) The deletion (F) and duplication (G) segments with the greatest total excess of PTV or missense DNMs usually featured a single, prominent driver gene accounting for most of that segment's mutational excess.

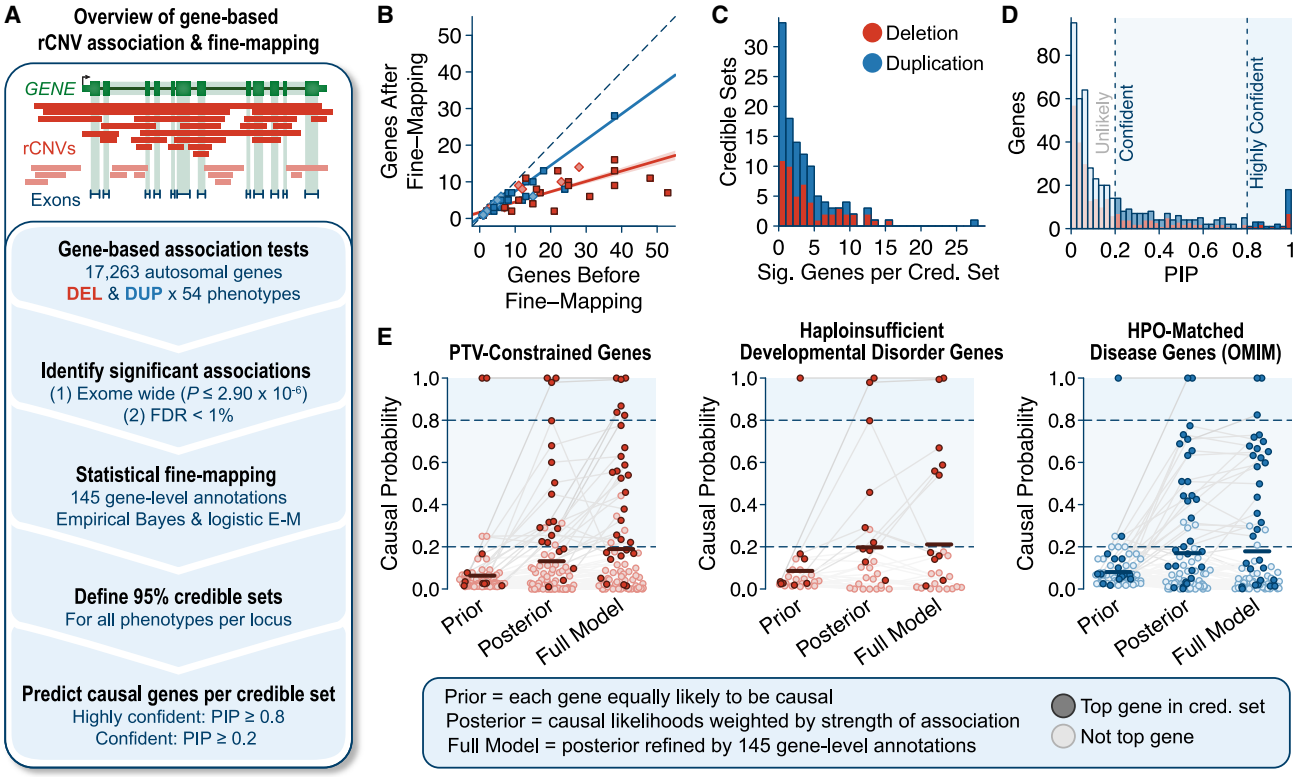
See also Figures S4 and S5.

posterior inclusion probabilities (PIPs) ≥0.8 and 0.2–0.8, respectively, in at least one credible set for one CNV type (Figure 4D). These 121 prioritized genes were enriched in gene sets with existing evidence supporting their disease relevance as plausible driver genes, such as mutational constraint and prior reports of gene-disease associations (Figures 4E and 4F).

These gene-based analyses did not identify any additional disease-associated loci that were not already captured by our sliding window analyses but did uncover 32 more rCNV-phenotype associations at these loci (total n = 829 associations between both approaches). Fine-mapping nominated at least one candidate gene (PIP ≥ 0.2) in 55.2% (90/163) of significant rCNV segments from our sliding window analyses, 48 of which had previously documented roles in disease, like *SHANK3* (PIP = 0.54; Figure S2G) in Phelan-McDermid syndrome (Zhou et al., 2019) or *RAI1* (PIP = 0.56; Figure S2H) in Smith-Magenis

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

CellPress



F

	Top gene in credible set					Not top gene in credible set				
Highly confident (PIP ≥ 0.8)	***ACACA***$^{D+}$	*AP000783.1*	***BCR***$^{D+}$	*C16ORF45*	*CHL1*	(No genes)				
	CNST	***DOCK4***	***EHMT1***$^{D+}$	*GET4*	*KCNT1*$^{D+}$					
	LAT^{D}	***LIMK1***$^{D+}$	*NPHP1*$^{D+}$	***NRXN1***$^{D+}$	*OCA2*$^{D+}$					
	OR2T4	*RABL5*	*RASA3*	*SEPT10*	*ST6GAL2*					
Confident (PIP ≥ 0.2)	*ABAT*^{D}	*ACP1*	*ADAM30*	***ANKRD11***$^{D+}$	***ATAD5***	*ABCC1*$^{D+}$	*ALDH3A1*	*BCL9*	*C21ORF58*	*CDIPT*
	BCL2L11	*CCDC154*	*CD160*	*CDR2*	*CDRT4*	*CECR5*	*CHD1L*	*ELN*$^{D+}$	*EMILIN2*	*FAM150B*
	CECR1	*CHST12*	***COLEC12***	*CRHR1*	*CTBP1*$^{D+}$	*GLOD4*	***INPP5A***	*KANK1*$^{D+}$	***KIF26B***	*MERTK*$^{D+}$
	CTSB^{D}	***CYFIP1***	*DNASE1*^{D}	*DOCK8*$^{D+}$	*ESYT2*	*MKRN3*$^{D+}$	***MYH11***$^{D+}$	***NCAPG2***$^{D+}$	*NIPA1*$^{D+}$	***NIPA2***
	FAM149A	*FAN1*$^{D+}$	***FBXO45***	*GEMIN4*$^{D+}$	*GJA5*$^{D+}$	***OTUD7A***	*PLEKHG4B*	***PMP22***$^{D+}$	*POLR3E*	*QPRT*
	GUCD1	***HDLBP***	***IGF1R***$^{D+}$	*IQCG*	*IRF2*	*REG4*	*RNF115*	*SGK223*	*SLC51A*	*TEFM*
	KIAA0430	*KLF13*$^{D+}$	*LAT*^{D}	***LIMK1***$^{D+}$	*LRRC30*	***TENM3***$^{D+}$	*TRAP1*			
	MAPK1^{D+}	*METRN*	*MPG*	*MUC4*	*MYO19*					
	NIPA2	*NKX6–2*$^{D+}$	*NPAP1*$^{D+}$	***NXN***$^{D+}$	*PCYT1A*$^{D+}$					
	PGPEP1L	*PI4KA*$^{D+}$	***PIAS3***	*PRKAB2*	*PRKCZ*					
	PRSS55	***RAI1***$^{D+}$	*RPS6KA2*	*SDCCAG8*$^{D+}$	*SEZ6L2*					
	SHANK3^{D+}	*SLC12A7*	***SMCHD1***$^{D+}$	*SPECC1*	*SPN*					
	TARSL2	*TDRP*	*TGIF1*$^{D+}$	*TLR3*$^{D+}$	*TXNL4A*$^{D+}$					
	WDR60	*YBEY*								
Unlikely (PIP < 0.2)	***HIRA***	*LRRC72*				157 genes	153 genes	31 genes		

Deletion (DEL) — **Bold** = PTV constrained
Duplication (DUP) — Underline = Missense constrained
DEL & DUP — D = Known disease gene (OMIM)
D+ = Known disease gene for same phenotype (OMIM)

Figure 4. Fine-mapping prioritizes candidate genes in disease-associated rCNVs

(A) Gene-based rCNV-disease association and fine-mapping workflow.

(B) Fine-mapping reduced the number of candidate genes per locus by 48%. Trend lines are outlier-robust linear fits with 95% confidence intervals.

(C) Distribution of credible set sizes after fine-mapping.

(D) Distribution of posterior inclusion probabilities (PIPs) for all genes across all credible sets.

(E) Comparison of per-gene probabilities across three gene sets of interest for each stage of our fine-mapping approach: naive uniform prior, genetics-only posterior, and functionally informed PIP (i.e., "full model").

(F) Summary of fine-mapping for all genes associated with one of 17 phenotypes with stronger rCNV effects (see Figure 1A) stratified by whether the gene had the highest PIP (i.e., "top gene") among all genes in at least one credible set.

See also Figure S2 and Tables S5 and S6.

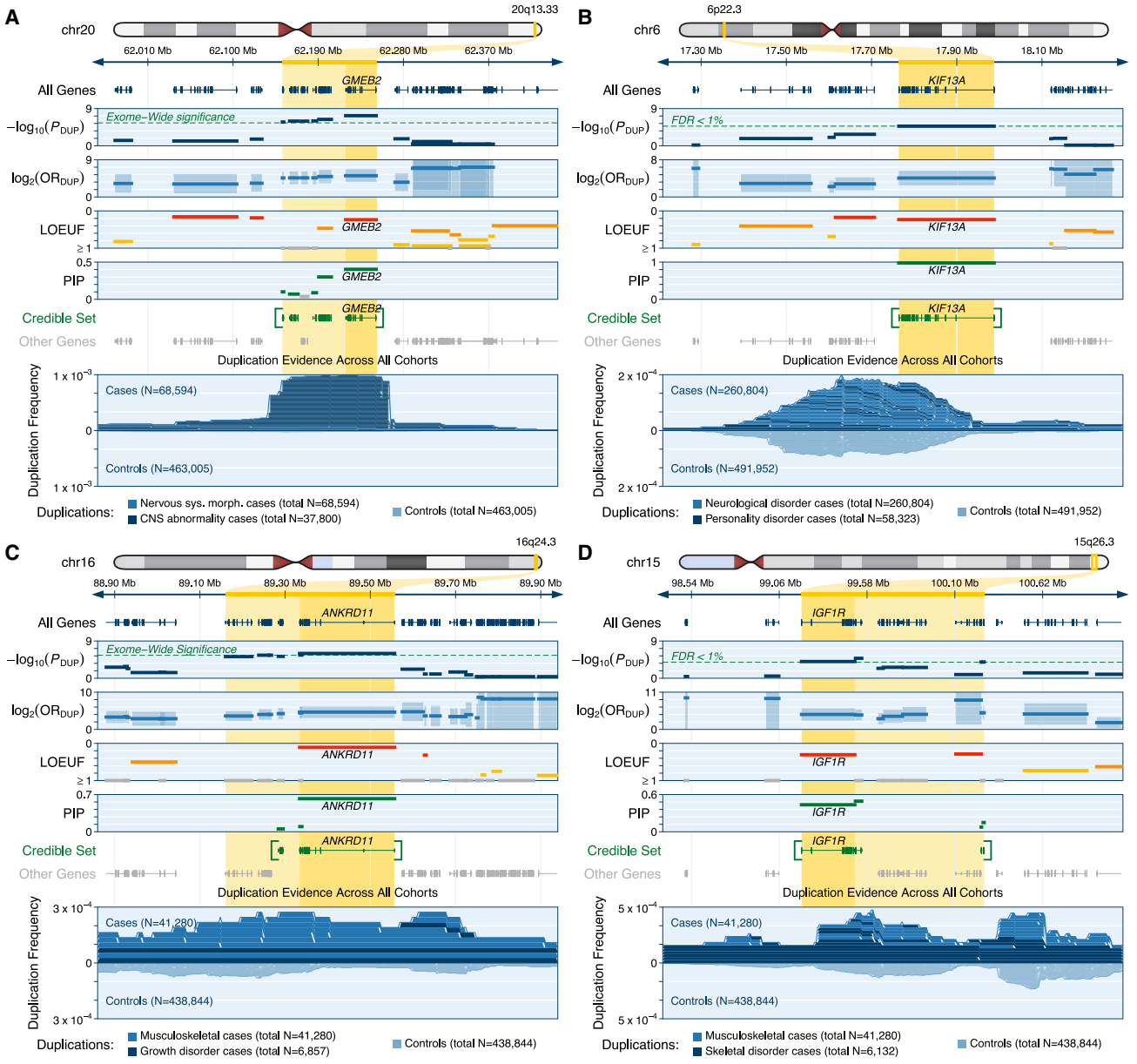


Figure 5. Example triplosensitive disease genes nominated by fine-mapping

(A) We identified a 95% credible set of five genes on chr20 where rare duplications were associated with nervous system abnormalities (OR = 35.2; 95% CI = 11.9–103.5). Fine-mapping prioritized *GMEB2* as the top candidate for this association (PIP = 0.40).

(B) We identified an association ($p = 1.79 \times 10^{-5}$; FDR Q = 0.004) between rare duplications and personality disorders (OR = 16.8; 95% CI = 5.7–49.4) on chr6, which fine-mapping reduced to just one gene, *KIF13A* (PIP = 0.98).

(C) We identified a three-gene credible set on chr16 where rare duplications were associated with growth abnormalities (OR = 25.0; 95% CI = 8.5–73.7). Fine-mapping nominated the known haploinsufficient gene, *ANKRD11* (PIP = 0.62), as the most likely causal gene.

(D) We identified a four-gene credible set on chr15 where rare duplications were associated with skeletal abnormalities (OR = 21.3; 95% CI = 6.1–75.2). Fine-mapping prioritized *IGF1R* (PIP = 0.44), a biologically plausible candidate gene (Abuzzahab et al., 2003), as one of two candidates for this association. For all panels, meta-analysis p values and ORs are provided for the more specific (smaller N) of the two phenotypes listed at the bottom of the panel, and all ORs are represented with 95% CIs.

See also Figure S6.

syndrome (Slager et al., 2003). These analyses also prioritized 14 genes that were mutationally constrained but had no known roles in disease, including candidate TS genes like *GMEB2* in central nervous system abnormalities (PIP = 0.40; OR = 35.2; 95% CI = 11.9–103.5; Figure 5A) and *KIF13A* in personality disorders (PIP = 0.98; OR = 16.8; 95% CI = 5.7–49.4; Figure 5B).

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

CellPress

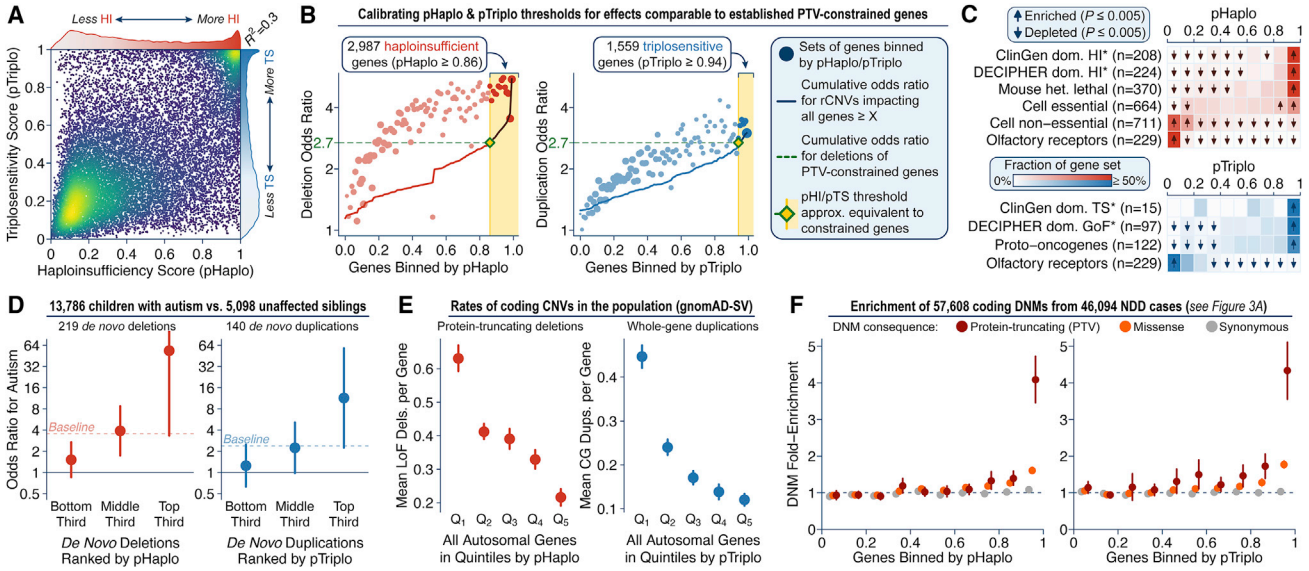


Figure 6. Predicting dosage sensitivity at single-gene resolution

(A) The probabilities of haploinsufficiency (pHaplo) and triplosensitivity (pTriplo) were moderately correlated per gene (Pearson $R^2 = 0.30$; $p < 10^{-100}$).

(B) We calibrated thresholds for pHaplo and pTriplo to define 2,987 HI and 1,559 TS genes where the effect sizes of deletions or duplications were comparable with loss of function of PTV-constrained genes (Karczewski et al., 2020).

(C) We observed clear shifts in the distributions of pHaplo and pTriplo across gene sets with prior biological evidence as being dosage sensitive or insensitive. Asterisks indicate gene sets considered when training our models and are not fully independent test sets.

(D) pHaplo and pTriplo stratified risk for ASD conferred by *de novo* protein-truncating deletions and whole-gene copy-gain (CG) duplications outside of GDs in an independent dataset of 13,786 affected children and 5,098 unaffected siblings (Fu et al., 2021). Baseline indicates the overall OR for all *de novo* deletions or duplications. Error bars correspond to 95% CIs.

(E) pHaplo and pTriplo were inversely correlated with rates of protein-truncating deletions and CG duplications in the general population (Collins et al., 2020). Error bars correspond to 95% CIs.

(F) The top decile of genes when ranked by pHaplo and pTriplo were enriched for damaging DNMs (PTVs and missense) in 46,094 probands affected by NDDs (Fu et al., 2021; Kaplanis et al., 2020). Error bars correspond to 95% CIs.

See also Figure S6 and Table S7.

Finally, we identified seven genes within duplication associations that had previously been shown to be dominant genetic causes of diseases via haploinsufficiency, like *ANKRD11* in growth abnormalities (PIP = 0.62; OR = 25.0; 95% CI = 8.5–73.7; Figure 5C) and *IGF1R* in skeletal abnormalities (PIP = 0.44; OR = 21.3; 95% CI = 6.1–75.2; Figure 5D). Haploinsufficiencies of *ANKRD11* and *IGF1R* are known causes of Cornelia de Lange and insulin growth factor I resistance syndromes (Abuzzahab et al., 2003; Ansari et al., 2014), respectively, but our study suggests that both genes are in fact bidirectionally DS.

Quantifying the dosage sensitivity of all protein-coding genes

We reasoned that a catalog of dosage sensitivity metrics for all genes—even if imperfect—would represent a potentially useful tool for genomics research and clinical genetics and thus developed a two-step procedure to computationally predict the pHaplo and pTriplo for 18,641 autosomal protein-coding genes. We first used an empirical Bayes approach to compute the likelihood that each gene belonged to one of two manually curated sets of likely DS and dosage insensitive genes based on the summary statistics from our gene-level association meta-analyses (Figures S6A–S6D), with slight modifications and optimized parameters (see STAR Methods). We then trained an ensemble of eight machine-learning models to predict these likelihoods from 145 gene-level features. The resulting pHaplo and pTriplo scores (Figure 6A; Table S7) easily separated known DS and dosage insensitive genes with high precision and recall (Figures S6E and S6F). Encouragingly, we found that pTriplo was more effective than pHaplo when classifying genes that were uniquely TS (Figures S6G and S6H) and vice-versa for pHaplo in HI-specific genes, indicating that pTriplo may provide *in silico* support to the challenges of interpreting duplications in clinical genetics. Finally, we computed standardized cutoffs for pHaplo and pTriplo where the average effect sizes of deletions and duplications were as strong as the LoF of genes known to be constrained against PTVs (average OR $\geq$ 2.7) (Karczewski et al., 2020). Applying these cutoffs defined 2,987 HI (pHaplo $\geq$ 0.86) and 1,559 TS (pTriplo $\geq$ 0.94) genes with rCNV effect sizes comparable with LoF of gold-standard PTV-constrained genes (Figure 6B).

We assessed the quality and practical value of our dosage sensitivity scores using five approaches. First, pHaplo and pTriplo were predictive of genes with biological evidence for being HI or TS independent of our model's training criteria (Figure 6C). Second, pHaplo and pTriplo stratified risk for ASD conferred by *de novo* CNVs outside of known GDs in 13,786 affected children and their 5,098 unaffected siblings (Fu et al., 2021;

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

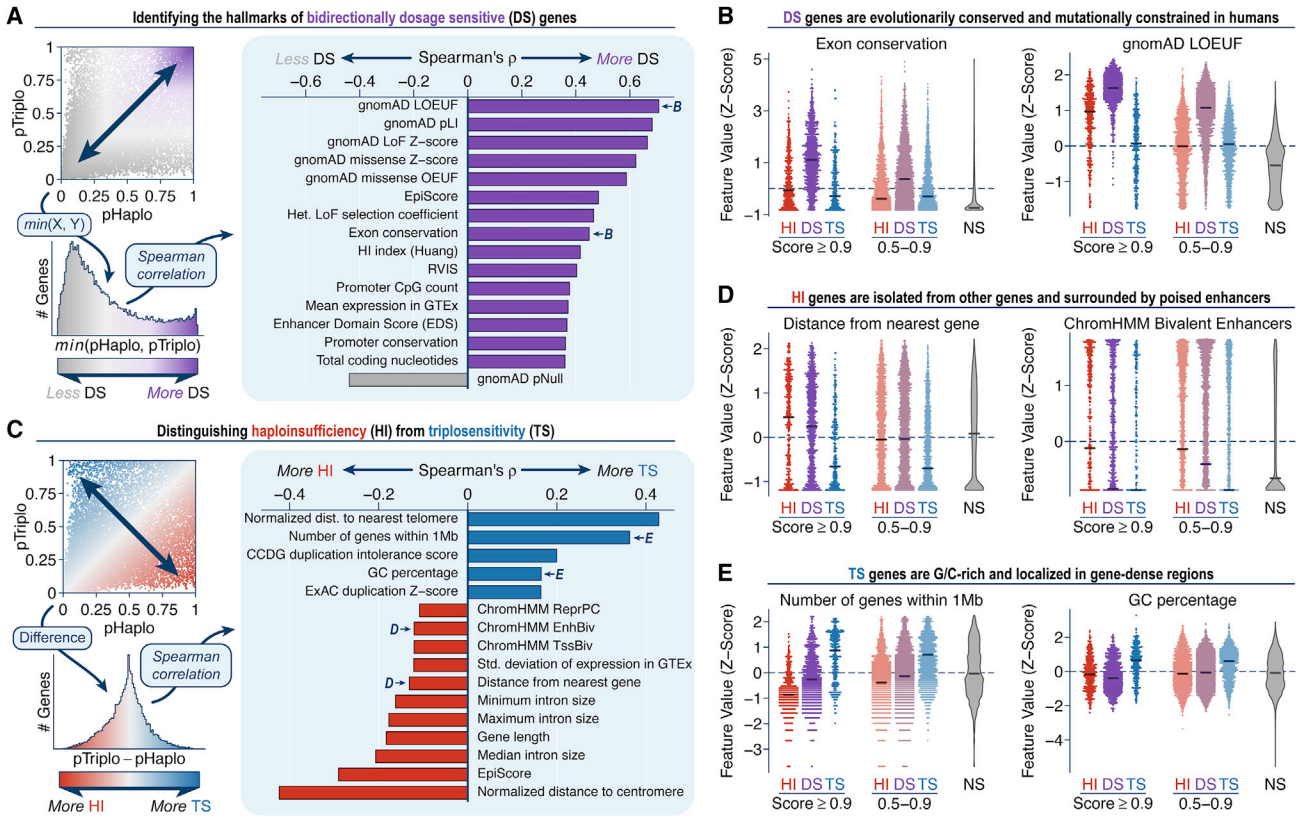


Figure 7. Insights into the biological basis of genic dosage sensitivity

(A) We identified the features most correlated with bidirectionally dosage sensitive (DS) genes by comparing the minimum pHaplo & pTriplo per gene to 145 gene-level features. The 16 gene-level features with the largest absolute correlation coefficients are shown here.

(B) Distributions of selected gene-level features for subsets of genes classified as haploinsufficient (HI), DS, triplosensitive (TS), and not dosage sensitive (NS). For clarity, all features have been transformed into *Z* scores.

(C) We also identified features predictive of genes uniquely HI or TS (but not both) using a Spearman correlation approach similar to (A).

(D and E) See (B).

See also Figure S7.

Figure 6D). Third, both pHaplo and pTriplo were inversely correlated with rates of protein-truncating deletions and whole-gene duplications in a large catalog of CNVs from genome sequencing (Figure 6E; Figures S6I and S6J; Collins et al., 2020). Fourth, genes with high pHaplo and pTriplo had significant excesses of damaging DNMs and chromosomal rearrangements in individuals with NDDs, although we observed no such enrichments for synonymous DNMs nor damaging DNMs in their unaffected siblings (Figure 6F; Figures S6K and S6L; Fu et al., 2021; Kaplanis et al., 2020; Redin et al., 2017). Fifth, both pHaplo and pTriplo were correlated with gene-level constraint metrics derived from point mutations (Figure S7; Table S7).

Satisfied with their technical quality, we leveraged pHaplo and pTriplo to understand the general properties governing the dosage sensitivity of human genes. First, we determine the features underpinning bidirectional dosage sensitivity by evaluating the minimum of pHaplo and pTriplo per gene (Figure 7A), finding that bidirectionally DS genes were defined by their evolutionary constraint above all other features (Figure 7B). Second, we explored the properties distinguishing haploinsufficiency and triplosensitivity by considering the difference of pHaplo and pTriplo scores per gene (Figure 7C). This approach revealed that genes more sensitive to deletion than duplication (i.e., primarily HI genes) tended to be larger, farther from other genes, and with a greater number of poised enhancers in *cis*, all of which are hallmarks of precisely regulated, developmentally critical genes (Figure 7D; Montavon et al., 2011; Ovcharenko et al., 2005). Conversely, genes more sensitive to duplication than deletion (i.e., primarily TS genes) were generally shorter, G/C-rich, and located in gene-dense, highly active regions (Figure 7E). Although preliminary, these analyses represent an initial step toward understanding the principles of genic dosage sensitivity and decoupling the mechanisms of haploinsufficiency from triplosensitivity in the human genome.

DISCUSSION

We have systematically assessed the contribution of rCNVs across 54 human disorders by meta-analyzing a large compendium of biomedical datasets to produce a genome-wide catalog

of standardized rCNV association statistics. This catalog includes a consensus list of 178 DS genomic segments involved in human disease, with a high-confidence subset of 88 achieving strict genome-wide significance. We further showed that a substantial fraction of these segments likely harbors at least one DS driver gene based on enrichments of constrained disease genes and non-uniform distributions of damaging DNMs within rCNV segments. A general framework of approximately one causal gene per phenotype per segment is in agreement with the increased density of constrained genes we observed for pleiotropic rCNVs and comports with existing knowledge for a handful of prominent GDs, such as *CRKL* and *TBX1* associations with kidney and heart abnormalities, respectively, in 22q11.2 GD deletions (Lindsay et al., 2001; Lopez-Rivera et al., 2017). However, the full genetic effects of most rCNVs are likely to be more complex, given the known examples of *cis*-regulatory effects (Franke et al., 2022), gene-gene interactions (Carvalho et al., 2014; Singh et al., 2020), and variable penetrance or expressivity due to secondary variants and polygenic background (Albers et al., 2012; Davies et al., 2020; Girirajan et al., 2012).

A major challenge in disease association studies and clinical genetics is identifying the causal gene(s) or element(s) within large, multi-gene CNVs. In this study, we repurposed GWAS fine-mapping algorithms to statistically prioritize individual genes within large rCNVs across a range of effect sizes and genetic architectures. The patterns we revealed by integrating short variant datasets (e.g., damaging DNMs, mutational constraint) indicated that short variants and rCNVs frequently converge on the same causal genes at disease-associated loci and that this convergence can be indicative of mechanism, as evinced by the CNV direction-specific enrichments of PTV and missense DNMs we uncovered. We expect that the eventual unification of all classes of genetic variation in large-scale sequencing studies and comprehensive association frameworks will catalyze breakthroughs in our understanding of diverse pathogenic mutational consequences beyond haploinsufficiency.

Finally, we leveraged these data to predict the dosage sensitivity of every autosomal protein-coding gene. Over the last decade, there have been multiple efforts to predict the functional and/or pathogenic impact of CNVs (Abel et al., 2020; Aguirre et al., 2019; Huang et al., 2010; Ruderfer et al., 2016). The vast genetic and functional datasets now publicly available have enabled us to extend from previous approaches and explore predictions of bidirectional dosage sensitivity for individual genes using sample sizes considerably larger than prior studies. Our triplosensitivity scores in particular may provide a unique lens when interpreting rare duplications (Riggs et al., 2020) and even for some disease-associated missense variants, for which gain-of-function and LoF consequences are challenging to distinguish *in silico* (Heyne et al., 2020). Although our model did not directly estimate natural selection against dosage changes in genes, they clarified some previously observed patterns underlying selection against CNVs in humans. It has been shown that constraint metrics derived from point mutations can also predict genes intolerant of CNVs (Collins et al., 2020; Ruderfer et al., 2016), but the trends distinguishing HI and TS genes were more intriguing. Our model's prediction that HI genes tend to be larger, farther from other genes, and surrounded by poised enhancers is an archetypical description of critical developmental genes, which are generally separated from other genes due to their intricate *cis*-regulatory networks and complex regulation across tissues and time points (Montavon et al., 2011; Ovcharenko et al., 2005; Siepel et al., 2005). On the other hand, TS genes appeared to typically be small, G/C-rich genes localized to broadly active, gene-dense regions. Although these patterns are preliminary, they nevertheless provide an important foothold for future investigations of dosage sensitivity at sequence resolution and for decoupling the principles of haploinsufficiency and triplosensitivity throughout the human genome.

Limitations of the study

The major trade-off of boosting statistical power by aggregating microarray-based datasets is the low resolution of rCNV inference from microarray technology. Although we implemented methods to mitigate the technology-specific impact of breakpoint precision, our conservative approach undoubtedly failed to capture all modes by which rCNVs might alter genes or influence disease risk. This may be especially true for duplications, the molecular consequences of which are comparatively more diverse than deletions (Hurles et al., 2008). These analyses of canonical CNVs also certainly oversimplifies the features governing complex and multiallelic CNVs, which are known to play roles in some human phenotypes (Sekar et al., 2016; Wirth et al., 2006; Zekavat et al., 2018). Similarly, this study did not consider smaller (<100 kb), common (frequency > 1%), or somatic mosaic rCNVs, which are fertile areas for future research in sequencing datasets. Finally, it was not feasible to search for rCNVs with protective effects due to non-uniform phenotyping across cohorts, but such analyses in the future will be intriguing (Kamitaki et al., 2020).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Cohorts and raw CNV datasets
  - CNV harmonization
  - Phenotype curation
  - Curation of other datasets
  - Global analyses of rCNVs in disease architecture
  - Large segment association meta-analyses
  - Genomic feature analyses of large rCNV segments
  - Gene-based association meta-analyses & fine-mapping
  - Genic dosage sensitivity scoring
  - Secondary analyses of genic dosage sensitivity scores

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2022.06.036.

CONSORTIA

The members of the Estonian Biobank Research Team are Andres Metspalu, Reedik Mägi, Mari Nelis, Lili Milani, and Tõnu Esko.

ACKNOWLEDGMENTS

We thank Leonid Mirny, Jesse Engreitz, Joe Nasser, Elise Valkanas, and Lily Wang for useful discussions and feedback; Claire Redin, Jeremiah Wala, Steven Schumacher, and Stephan Sanders for assistance with data access; and Jake Conway and Matt Stone for advice on software development. We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the SSC principal investigators. These studies were supported by the National Institutes of Health grants MH115957, HD081256, HD099547, NS093200, HD096326, DE026824, DE031261, HD105266, and MH106826. R.L.C. was supported by NHGRI T32HG002295 and NSF GRFP #2017240332. This work was also supported by grants from the Swiss National Science Foundation (31003A_182632 to A.R. and 310030-189147, 32473B-166450 to Z.K.). M.E.T. was supported by Desmond and Ann Heathwood. M.L. and R.M. have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101016775 and are supported by the Estonian Research Council grant PUT (PRG687). M.N. is supported by the Estonian Research Council grant PUT (PRG555). U.V. and T.E. are supported by the Estonian Research Council grant PUT (PRG1291). This research was further supported by the European Union through the European Regional Development Fund (Project no. 2014-2020.4.01.16-0125).

AUTHOR CONTRIBUTIONS

Conceptualization, R.L.C., S.S., H.B., and M.E.T.; methodology, R.L.C., J.U., D.P.H., J.F., K.J.K., K.E.S., M.T., H.F., B.M.N., Z.K., S.S., H.B., and M.E.T.; software, R.L.C.; formal analysis, R.L.C., S.E., and H.B.; resources and data curation, R.L.C., J.T.G., E.P., M.L., R.B., C. Lauricella, L.H., T.M., L.-M.N., S.E., D.P.H., G.K., C. Lowther, K.M., M.N., F.U., U.V., S.A., J.F.G., A.J., L.M., B.M.N., D.S., S.W., J.C.H., D. Lal, D.M.R., J.M., T.E., Z.K., H.H., H.B., and M.E.T.; writing, R.L.C., H.B., and M.E.T.; supervision, project administration, and funding acquisition, R.L.C., D. Lucente, M.E.H., S.A., E.E.D., H.F., J.F.G., A.J., N.K., L.M., B.M.N., D.S., S.W., J.C.H., D. Lal, D.M.R., J.M., R.M., T.E., A.R., Z.K., H.H., S.S., H.B., and M.E.T.

DECLARATION OF INTERESTS

M.E.T. receives research funding and/or reagents from Levo Therapeutics, Microsoft Inc., and Illumina Inc. R.B., C. Lauricella, A.J., L.M., S.W., and J.M. are employees of GeneDx, Inc. S.A. is an employee of Invitae Corp.

Received: December 22, 2020
Revised: March 17, 2022
Accepted: June 20, 2022
Published: July 27, 2022

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Abel, H.J., Larson, D.E., Regier, A.A., Chiang, C., Das, I., Kanchi, K.L., Layer, R.M., Neale, B.M., Salerno, W.J., Reeves, C., et al. (2020). Mapping and characterization of structural variation in 17,795 human genomes. Nature *583*, 83–89.

Abuzzahab, M.J., Schneider, A., Goddard, A., Grigorescu, F., Lautier, C., Keller, E., Kiess, W., Klammt, J., Kratzsch, J., Osgood, D., et al. (2003). IGF-I receptor mutations resulting in intrauterine and postnatal growth retardation. N. Engl. J. Med. *349*, 2211–2222.

Aguirre, M., Rivas, M.A., and Priest, J. (2019). Phenome-wide burden of copy-number variation in the UK Biobank. Am. J. Hum. Genet. *105*, 373–383.

Albers, C.A., Paul, D.S., Schulze, H., Freson, K., Stephens, J.C., Smethurst, P.A., Jolley, J.D., Cvejic, A., Kostadima, M., Bertone, P., et al. (2012). Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. Nat. Genet. *44*, s431–s432.

Almarri, M.A., Bergström, A., Prado-Martinez, J., Yang, F., Fu, B., Dunham, A.S., Chen, Y., Hurles, M.E., Tyler-Smith, C., and Xue, Y. (2020). Population structure, stratification, and introgression of human structural variation. Cell *182*, 189–199.e15.

Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res. *43*, D789–D798.

Annis, A., Pandit, A., LeFaive, J., Taliun, S.G., Fritsche, L., VandeHaar, P., Boehnke, M., Zawistowski, M., Abecasis, G., and Zöllner, S. (2021). False discovery rates for genome-wide association tests in biobanks with thousands of phenotypes. Nat Porfolio. https://doi.org/10.21203/rs.3.rs-873449/v1.

Ansari, M., Poke, G., Ferry, Q., Williamson, K., Aldridge, R., Meynert, A.M., Bengani, H., Chan, C.Y., Kayserili, H., Avci, S., et al. (2014). Genetic heterogeneity in Cornelia de Lange syndrome (CdLS) and CdLS-like phenotypes with observed and predicted levels of mosaicism. J. Med. Genet. *51*, 659–668.

Auwerx, C., Lepamets, M., Sadler, M.C., Patxot, M., Stojanov, M., Baud, D., Mägi, R., Estonian Biobank Research Team, Porcu, E., Reymond, A., et al. (2022). The individual and global impact of copy-number variants on complex human traits. Am. J. Hum. Genet. *109*, 647–668.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. *28*, 45–48.

Bragin, E., Chatzimichali, E.A., Wright, C.F., Hurles, M.E., Firth, H.V., Bevan, A.P., and Swaminathan, G.J. (2014). DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. Nucleic Acids Res. *42*, D993–D1000.

Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2021). High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Preprint at bioRxiv. https://doi.org/10.1101/2021.02.06.430068.

Carvalho, C.M., and Lupski, J.R. (2016). Mechanisms underlying structural variant formation in genomic disorders. Nat. Rev. Genet. *17*, 224–238.

Carvalho, C.M., Vasanth, S., Shinawi, M., Russell, C., Ramocki, M.B., Brown, C.W., Graakjaer, J., Skytte, A.B., Vianna-Morgante, A.M., Krepischi, A.C., et al. (2014). Dosage changes of a segment at 17p13.1 lead to intellectual disability and microcephaly as a result of complex genetic interaction of multiple genes. Am. J. Hum. Genet. *95*, 565–578.

Cassa, C.A., Weghorn, D., Balick, D.J., Jordan, D.M., Nusinow, D., Samocha, K.E., O'Donnell-Luria, A., MacArthur, D.G., Daly, M.J., Beier, D.R., et al. (2017). Estimating the selective effects of heterozygous protein-truncating variants from human exome data. Nat. Genet. *49*, 806–810.

Coe, B.P., Witherspoon, K., Rosenfeld, J.A., van Bon, B.W., Vulto-van Silfhout, A.T., Bosco, P., Friend, K.L., Baker, C., Buono, S., Vissers, L.E., et al. (2014). Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat. Genet. *46*, 1063–1071.

Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451.

Cooper, G.M., Coe, B.P., Girirajan, S., Rosenfeld, J.A., Vu, T.H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., et al. (2011). A copy number variation morbidity map of developmental delay. Nat. Genet. *43*, 838–846.

Cummings, B.B., Karczewski, K.J., Kosmicki, J.A., Seaby, E.G., Watts, N.A., Singer-Berk, M., Mudge, J.M., Karjalainen, J., Satterstrom, F.K., O'Donnell-Luria, A.H., et al. (2020). Transcript expression-aware annotation improves rare variant interpretation. Nature *581*, 452–458.

Davies, R.W., Fiksinski, A.M., Breetvelt, E.J., Williams, N.M., Hooper, S.R., Monfeuga, T., Bassett, A.S., Owen, M.J., Gur, R.E., Morrow, B.E., et al. (2020). Using common genetic variation to examine phenotypic expression and risk prediction in 22q11.2 deletion syndrome. Nat. Med. *26*, 1912–1918.

Dey, R., Schmidt, E.M., Abecasis, G.R., and Lee, S. (2017). A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. Am. J. Hum. Genet. *101*, 37–49.

Dittwald, P., Gambin, T., Szafranski, P., Li, J., Amato, S., Divon, M.Y., Rodríguez Rojas, L.X., Elton, L.E., Scott, D.A., Schaaf, C.P., et al. (2013). NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and mendelizing traits. Genome Res. *23*, 1395–1409.

Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: database of chromosomal imbalance and phenotype in humans using Ensembl resources. Am. J. Hum. Genet. *84*, 524–533.

Franke, M., Daly, A.F., Palmeira, L., Tirosh, A., Stigliano, A., Trifan, E., Faucz, F.R., Abboud, D., Petrossians, P., Tena, J.J., et al. (2022). Duplications disrupt chromatin architecture and rewire GPR101-enhancer communication in X-linked acrogigantism. Am. J. Hum. Genet. *109*, 553–570.

Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773.

Fu, J.M., Satterstrom, F.K., Peng, M., Brand, H., Collins, R.L., Dong, S., Klei, L., Stevens, C.R., Cusick, C., Babadi, M., et al. (2021). Rare coding variation illuminates the allelic architecture, risk genes, cellular expression patterns, and phenotypic context of autism. Preprint at medRxiv. https://doi.org/10.1101/2021.12.20.21267194.

Girirajan, S., and Eichler, E.E. (2010). Phenotypic variability and genetic susceptibility to genomic disorders. Hum. Mol. Genet. *19*, R176–R187.

Girirajan, S., Rosenfeld, J.A., Coe, B.P., Parikh, S., Friedman, N., Goldstein, A., Filipink, R.A., McConnell, J.S., Angle, B., Meschino, W.S., et al. (2012). Phenotypic heterogeneity of genomic disorders and rare copy-number variants. N. Engl. J. Med. *367*, 1321–1331.

Golzio, C., Willer, J., Talkowski, M.E., Oh, E.C., Taniguchi, Y., Jacquemont, S., Reymond, A., Sun, M., Sawa, A., Gusella, J.F., et al. (2012). KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. Nature *485*, 363–367.

GTEx Consortium; Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

Han, X., Chen, S., Flynn, E., Wu, S., Wintner, D., and Shen, Y. (2018). Distinct epigenomic patterns are associated with haploinsufficiency and predict risk genes of developmental disorders. Nat. Commun. *9*, 2138.

Heyne, H.O., Baez-Nieto, D., Iqbal, S., Palmer, D.S., Brunklaus, A., May, P., Epi25 Collaborative, Johannesen, K.M., Lauxmann, S., Lemke, J.R., et al. (2020). Predicting functional effects of missense variants in voltage-gated sodium and calcium channels. Sci. Transl. Med. *12*, eaay6848.

Huang, A.Y., Yu, D., Davis, L.K., Sul, J.H., Tsetsos, F., Ramensky, V., Zelaya, I., Ramos, E.M., Osiecki, L., Chen, J.A., et al. (2017). Rare copy number variants in NRXN1 and CNTN6 increase risk for Tourette syndrome. Neuron *94*, 1101–1111.e7.

Huang, N., Lee, I., Marcotte, E.M., and Hurles, M.E. (2010). Characterising and predicting haploinsufficiency in the human genome. PLoS Genet. *6*, e1001154.

Hurles, M.E., Dermitzakis, E.T., and Tyler-Smith, C. (2008). The functional impact of structural variation in humans. Trends Genet. *24*, 238–245.

Ibrahim-Verbaas, C.A., Bressler, J., Debette, S., Schuur, M., Smith, A.V., Bis, J.C., Davies, G., Trompet, S., Smith, J.A., Wolf, C., et al. (2016). GWAS for executive function and processing speed suggests involvement of the CADM2 gene. Mol. Psychiatry *21*, 189–197.

Itsara, A., Wu, H., Smith, J.D., Nickerson, D.A., Romieu, I., London, S.J., and Eichler, E.E. (2010). De novo rates and selection of large copy number variation. Genome Res. *20*, 1469–1481.

Jacquemont, S., Reymond, A., Zufferey, F., Harewood, L., Walters, R.G., Kutalik, Z., Martinet, D., Shen, Y., Valsesia, A., Beckmann, N.D., et al. (2011). Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature *478*, 97–102.

Kamitaki, N., Sekar, A., Handsaker, R.E., de Rivera, H., Tooley, K., Morris, D.L., Taylor, K.E., Whelan, C.W., Tombleson, P., Loohuis, L.M.O., et al. (2020). Complement genes contribute sex-biased vulnerability in diverse disorders. Nature *582*, 577–581.

Kaplanis, J., Samocha, K.E., Wiel, L., Zhang, Z., Arvai, K.J., Eberhardt, R.Y., Gallone, G., Lelieveld, S.H., Martin, H.C., McRae, J.F., et al. (2020). Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature *586*, 757–762.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

Kichaev, G., Roytman, M., Johnson, R., Eskin, E., Lindström, S., Kraft, P., and Pasaniuc, B. (2017). Improved methods for multi-trait fine mapping of pleiotropic risk loci. Bioinformatics Oxf. Engl. *33*, 248–255.

Kichaev, G., Yang, W.Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. *10*, e1004722.

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J.P., Gargano, M., Harris, N.L., Matentzoglu, N., McMurry, J.A., et al. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. *47*, D1018–D1027.

Lee, C.M., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Gonzalez, J.N., Hinrichs, A.S., Lee, B.T., Nassar, L.R., Powell, C.C., et al. (2020). UCSC Genome Browser enters 20th year. Nucleic Acids Res. *48*, D756–D761.

Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.-L., Alavere, H., Snieder, H., Perola, M., Ng, P.C., Mägi, R., Milani, L., et al. (2015). Cohort profile: Estonian biobank of the Estonian Genome Center, University of Tartu. Int. J. Epidemiol. *44*, 1137–1147.

Li, Y.R., Glessner, J.T., Coe, B.P., Li, J., Mohebnasab, M., Chang, X., Connolly, J., Kao, C., Wei, Z., Bradfield, J., et al. (2020). Rare copy number variants in over 100,000 European ancestry subjects reveal multiple disease associations. Nat. Commun. *11*, 255.

Lindsay, E.A., Vitelli, F., Su, H., Morishima, M., Huynh, T., Pramparo, T., Jurecic, V., Ogunrinu, G., Sutherland, H.F., Scambler, P.J., et al. (2001). Tbx1 haploinsufficieny in the DiGeorge syndrome region causes aortic arch defects in mice. Nature *410*, 97–101.

Lopez-Rivera, E., Liu, Y.P., Verbitsky, M., Anderson, B.R., Capone, V.P., Otto, E.A., Yan, Z., Mitrotti, A., Martino, J., Steers, N.J., et al. (2017). Genetic drivers of kidney defects in the DiGeorge syndrome. N. Engl. J. Med. *376*, 742–754.

Lupski, J.R. (2009). Genomic disorders ten years on. Genome Med. *1*, 42.

Macé, A., Tuke, M.A., Beckmann, J.S., Lin, L., Jacquemont, S., Weedon, M.N., Reymond, A., and Kutalik, Z. (2016). New quality measure for SNP array based CNV detection. Bioinformatics Oxf. Engl. *32*, 3298–3305.

Macé, A., Tuke, M.A., Deelen, P., Kristiansson, K., Mattsson, H., Nõukas, M., Sapkota, Y., Schick, U., Porcu, E., Rüeger, S., et al. (2017). CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. Nat. Commun. *8*, 744.

Marshall, C.R., Howrigan, D.P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D.S., Antaki, D., Shetty, A., Holmans, P.A., Pinto, D., et al. (2016). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nat. Genet. *49*, 27–35.

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., de Laat, W., Spitz, F., and Duboule, D. (2011). A regulatory archipelago controls Hox genes transcription in digits. Cell *147*, 1132–1145.

Morley, T.J., Han, L., Castro, V.M., Morra, J., Perlis, R.H., Cox, N.J., Bastarache, L., and Ruderfer, D.M. (2021). Phenotypic signatures in clinical data enable systematic identification of patients for genetic testing. Nat. Med. *27*, 1097–1104.

Niestroj, L.-M., Perez-Palma, E., Howrigan, D.P., Zhou, Y., Cheng, F., Saarentaus, E., Nürnberg, P., Stevelink, R., Daly, M.J., Palotie, A., et al. (2020). Epilepsy subtype-specific copy number burden observed in a genome-wide study of 17 458 subjects. Brain *143*, 2106–2118.

Ohno, S. (1970). Evolution by Gene Duplication (Springer-Verlag).

Ovcharenko, I., Loots, G.G., Nobrega, M.A., Hardison, R.C., Miller, W., and Stubbs, L. (2005). Evolution and functional classification of vertebrate gene deserts. Genome Res. *15*, 137–145.

Owen, D., Bracher-Smith, M., Kendall, K.M., Rees, E., Einon, M., Escott-Price, V., Owen, M.J., O'Donovan, M.C., and Kirov, G. (2018). Effects of pathogenic CNVs on physical traits in participants of the UK Biobank. BMC Genomics *19*, 867.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., and Goldstein, D.B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. *9*, e1003709.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics Oxf. Engl. *26*, 841–842.

Redin, C., Brand, H., Collins, R.L., Kammin, T., Mitchell, E., Hodge, J.C., Hanscom, C., Pillalamarri, V., Seabra, C.M., Abbott, M.A., et al. (2017). The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. Nat. Genet. *49*, 36–45.

Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al. (2015). ClinGen–the clinical genome resource. N. Engl. J. Med. *372*, 2235–2242.

Riggs, E.R., Andersen, E.F., Cherry, A.M., Kantarci, S., Kearney, H., Patel, A., Raca, G., Ritter, D.I., South, S.T., Thorland, E.C., et al. (2020). Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). Genet. Med. *22*, 245–257.

Riggs, E.R., Church, D.M., Hanson, K., Horner, V.L., Kaminsky, E.B., Kuhn, R.M., Wain, K.E., Williams, E.S., Aradhya, S., Kearney, H.M., et al. (2012). Towards an evidence-based process for the clinical interpretation of copy number variation. Clin. Genet. *81*, 403–412.

Riggs, E.R., Nelson, T., Merz, A., Ackley, T., Bunke, B., Collins, C.D., Collinson, M.N., Fan, Y.S., Goodenberger, M.L., Golden, D.M., et al. (2018). Copy number variant discrepancy resolution using the ClinGen dosage sensitivity map results in updated clinical interpretations in ClinVar. Hum. Mutat. *39*, 1650–1659.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Ruderfer, D.M., Hamamsy, T., Lek, M., Karczewski, K.J., Kavanagh, D., Samocha, K.E., Exome Aggregation Consortium, Daly, M.J., MacArthur, D.G., Fromer, M., et al. (2016). Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. Nat. Genet. *48*, 1107–1111.

Sanchez-Roige, S., Fontanillas, P., Elson, S.L., Gray, J.C., de Wit, H., MacKillop, J., and Palmer, A.A. (2019). Genome-Wide Association Studies of Impulsive Personality Traits (BIS-11 and UPPS-P) and Drug Experimentation in up to 22,861 Adult Research Participants Identify Loci in the CACNA1I and CADM2 genes. J. Neurosci. *39*, 2562–2572.

Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., et al. (2015). Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron *87*, 1215–1233.

Sekar, A., Bialas, A.R., de Rivera, H., Davis, A., Hammond, T.R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., et al. (2016). Schizophrenia risk from complex variation of complement component 4. Nature *530*, 177–183.

Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R., et al. (2005). Segmental duplications and copy-number variation in the human genome. Am. J. Hum. Genet. *77*, 78–88.

Sibley, C.R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., Trabzuni, D., Ryten, M., Weale, M.E., Hardy, J., et al. (2015). Recursive splicing in long vertebrate genes. Nature *521*, 371–375.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

Singh, M.D., Jensen, M., Lasser, M., Huber, E., Yusuff, T., Pizzo, L., Lifschutz, B., Desai, I., Kubina, A., Yennawar, S., et al. (2020). NCBP2 modulates neurodevelopmental defects of the 3q29 deletion in Drosophila and Xenopus laevis models. PLoS Genet. *16*, e1008590.

Slager, R.E., Newton, T.L., Vlangos, C.N., Finucane, B., and Elsea, S.H. (2003). Mutations in rai1 associated with Smith-Magenis syndrome. Nat. Genet. *33*, 466–468.

Stefansson, H., Meyer-Lindenberg, A., Steinberg, S., Magnusdottir, B., Morgen, K., Arnarsdottir, S., Bjornsdottir, G., Walters, G.B., Jonsdottir, G.A., Doyle, O.M., et al. (2014). CNVs conferring risk of autism or schizophrenia affect cognition in controls. Nature *505*, 361–366.

Sweeting, M.J., Sutton, A.J., and Lambert, P.C. (2004). What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat. Med. *23*, 1351–1375.

Talkowski, M.E., Rosenfeld, J.A., Blumenthal, I., Pillalamarri, V., Chiang, C., Heilbut, A., Ernst, C., Hanscom, C., Rossin, E., Lindgren, A.M., et al. (2012). Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. Cell *149*, 525–537.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. *20*, 467–484.

Tommerup, N. (1993). Mendelian cytogenetics. Chromosome rearrangements associated with Mendelian disorders. J. Med. Genet. *30*, 713–727.

Uddin, M., Thiruvahindrapuram, B., Walker, S., Wang, Z., Hu, P., Lamoureux, S., Wei, J., MacDonald, J.R., Pellecchia, G., Lu, C., et al. (2015). A high-resolution copy-number variation resource for clinical and population genetics. Genet. Med. *17*, 747–752.

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package*36*, p. 48.

Vulto-van Silfhout, A.T., Hehir-Kwa, J.Y., van Bon, B.W., Schuurs-Hoeijmakers, J.H., Meader, S., Hellebrekers, C.J., Thoonen, I.J., de Brouwer, A.P., Brunner, H.G., Webber, C., et al. (2013). Clinical significance of de novo and inherited copy-number variation. Hum. Mutat. *34*, 1679–1687.

Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am. J. Hum. Genet. *81*, 208–227.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with P-values. Genet. Epidemiol. *33*, 79–86.

Wang, X., and Goldstein, D.B. (2020). Enhancer domains predict gene pathogenicity and inform gene discovery in complex disease. Am. J. Hum. Genet. *106*, 215–233.

Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A., Green, T., et al. (2008). Association between microdeletion and microduplication at 16p11.2 and autism. N. Engl. J. Med. *358*, 667–675.

Wen, X., Pique-Regi, R., and Luca, F. (2017). Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. PLoS Genet. *13*, e1006646.

Wirth, B., Brichta, L., Schrank, B., Lochmüller, H., Blick, S., Baasner, A., and Heller, R. (2006). Mildly affected patients with spinal muscular atrophy are partially protected by an increased SMN2 copy number. Hum. Genet. *119*, 422–428.

Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., van Kogelenberg, M., King, D.A., Ambridge, K., Barrett, D.M., Bayzetinova, T., et al. (2015). Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet *385*, 1305–1314.

Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhsng, Zhsng, C.Z., Wala, J., Mermel, C.H., et al. (2013). Pan-cancer patterns of somatic copy number alteration. Nat. Genet. *45*, 1134–1140.

Zarrei, M., Burton, C.L., Engchuan, W., Young, E.J., Higginbotham, E.J., MacDonald, J.R., Trost, B., Chan, A.J.S., Walker, S., Lamoureux, S., et al. (2019). A large data resource of genomic copy number variation across neurodevelopmental disorders. npj Genom. Med. *4*, 26.

Zekavat, S.M., Ruotsalainen, S., Handsaker, R.E., Alver, M., Bloom, J., Poterba, T., Seed, C., Ernst, J., Chaffin, M., Engreitz, J., et al. (2018). Deep coverage whole genome sequences and plasma lipoprotein(a) in individuals of European and African ancestries. Nat. Commun. *9*, 2606.

Zhang, F., Gu, W., Hurles, M.E., and Lupski, J.R. (2009). Copy number variation in human health, disease, and evolution. Annu. Rev. Genomics Hum. Genet. *10*, 451–481.

Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. *50*, 1335–1341.

Zhou, Y., Sharma, J., Ke, Q., Landman, R., Yuan, J., Chen, H., Hayden, D.S., Fisher, J.W., 3rd, Jiang, M., Menegas, W., et al. (2019). Atypical behaviour and connectivity in SHANK3-mutant macaques. Nature *570*, 326–331.

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Summary statistics from genome-wide rCNV association meta-analyses	This paper	https://zenodo.org/record/6347673
Gene dosage sensitivity scores	This paper	https://zenodo.org/record/6347673
Gene features curated from existing sources	This paper	https://zenodo.org/record/6347673
De novo mutation data from individuals with developmental disorders	(Kaplanis et al., 2020)	https://doi.org/10.1038/s41586-020-2832-5
De novo mutation and *de novo* CNV data from individuals with autism spectrum disorder and their unaffected siblings	(Fu et al., 2021)	https://doi.org/10.1101/2021.12.20.21267194
Gene reference dataset	Gencode v19, (Frankish et al., 2019)	Gencode: https://www.gencodegenes.org/human/release_19.html
Genome Reference Consortium Human Build 37	Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/
CNV data from BioVU Biobank	(Morley et al., 2021)	https://doi.org/10.1038/s41591-021-01356-z
CNV data from Boston Children's Hospital	(Talkowski et al., 2012)	https://doi.org/10.1016/j.cell.2012.03.028
CNV data from Children's Hospital of Philadelphia	(Li et al., 2020)	https://doi.org/10.1038/s41467-019-13624-1
CNV data from Coe et al.	NCBI dbVar	dbVar: nstd100
CNV data from Cooper et al.	NCBI dbVar	dbVar: nstd54
CNV data from Epi25 Collaborative	(Niestroj et al., 2020)	https://doi.org/10.1093/brain/awaa171
CNV data from Estonian Biobank	(Auwerx et al., 2022)	https://doi.org/10.1016/j.ajhg.2022.02.010
CNV data from Ontario Population Genetics Platform	NCBI dbVar	dbVar: estd212
CNV data from Psychiatric Genetics Consortium	(Marshall et al., 2016)	https://doi.org/10.1038/ng.3725
CNV data from Radboud University Medical Center	NCBI dbVar	dbVar: nstd85
CNV data from SickKids Hospital	NCBI dbVar	dbVar: nstd173
CNV data from the Simons Simplex Collection	(Sanders et al., 2015)	https://doi.org/10.1016/j.neuron.2015.09.016
CNV data from the Cancer Genome Atlas	(Zack et al., 2013)	https://doi.org/10.1038/ng.2760
CNV data from the Genetic Etiology of Tourette Syndrome Consortium	(Huang et al., 2017)	https://doi.org/10.1016/j.neuron.2017.06.010
CNV data from the UK Biobank	(Auwerx et al., 2022)	https://doi.org/10.1016/j.ajhg.2022.02.010
Software and algorithms		
Data processing & analysis scripts	This paper	https://doi.org/10.5281/zenodo.6647918
R v3.6.1	The R Foundation	https://www.r-project.org/
Python 3.7.9	Python Software Foundation	https://www.python.org/
BEDTools v2.30.0	(Quinlan and Hall, 2010)	https://github.com/arq5x/bedtools2/
Docker v20.10.0	Docker	https://www.docker.com/
Functional fine-mapping algorithm	(Kichaev et al., 2014; Wen et al., 2017)	N/A
Metafor v2.1.0	(Viechtbauer, 2010)	https://www.metafor-project.org/
Scikit-Learn v0.23.2	(Pedregosa et al., 2011)	https://scikit-learn.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Michael Talkowski (talkowsk@broadinstitute.org).

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All genome-wide association summary statistics, gene features, and gene dosage sensitivity scores have been deposited at Zenodo and are publicly available as of the date of publication. DOIs are listed in the key resources table. This paper also analyzes publicly available data. These accession numbers for the datasets are listed in the key resources table. The CNV data reported GeneDx and Indiana University clinical testing sites were not consented for public release.
- All original code has been deposited at GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Cohorts and raw CNV datasets

We aggregated microarray-based CNV data across 17 sources, ranging from population-scale biobanks to genetic testing laboratories (Table S1). Where applicable, we provide references below to publications describing the sample recruitment and CNV discovery procedures for each cohort; please refer to these publications for full details for those cohorts. We applied cohort-specific quality control measures to each raw CNV dataset prior to CNV harmonization across all cohorts. Below, we describe each cohort and any cohort-specific steps applied prior to CNV harmonization:

1. BioVU Biobank at Vanderbilt University (BioVU): We obtained CNV calls for 66,360 BioVU participants who had been genotyped from peripheral blood samples on the Illumina MegaEx microarray platform as previously described (Morley et al., 2021). After excluding all related individuals, individuals with known chromosomal abnormalities (*i.e.*, aneuploidies, abnormal karyotypes, or CNVs >50Mb), or individuals with incident blood cancers, we retained 46,967 samples.
2. Boston Children's Hospital (BCH): We obtained CNV calls for 3,591 clinically referred samples from the Children's Hospital of Boston as described in a previous study (Talkowski et al., 2012). We subsequently converted the CNVs from hg18 to GRCh37 coordinates using the UCSC liftOver tool while requiring 50% of each CNV to remap contiguously to GRCh37 (Lee et al., 2019).
3. Children's Hospital of Philadelphia (CHOP): We obtained CNV calls for 342,408 samples analyzed by the Center for Applied Genomics at the Children's Hospital of Philadelphia, a subset of which were described in a recent publication (Li et al., 2020). We subsequently restricted CNVs to quality score $\geq$40, $\geq$10 probes, and $\geq$25kb in size. After CNV filtering, samples with LRR_SD <0.25, >20 CNV calls, or SNP call rate <98% were excluded as outliers, as well as samples genotyped on arrays with <175k SNP probes or samples labeled as cancer or Down's Syndrome patients. Finally, we identified 19 loci with platform-specific artifactual CNV pileups. CNVs covered $\geq$10% by any of these artifact regions were removed from the final CHOP CNV dataset. After filtering, we retained a total of 178,031 samples.
4. Coe et al. (Coe): We obtained CNV calls for 29,083 samples with developmental disorders and 11,256 unaffected control samples from NCBI dbVar accession #nstd100 as described in a previous publication (Coe et al., 2014).
5. Cooper et al. (Cooper): We obtained CNV calls for 8,329 unaffected control samples from NCBI dbVar accession #nstd54 as described in a previous publication (Cooper et al., 2011). While affected samples are also described in Cooper et al., we retained only the unaffected controls as all cases from Cooper et al. were reprocessed and included in Coe et al. above.
6. Epi25 Consortium (Epi25k): We obtained CNV calls for 12,758 samples with epilepsy and 8,478 unaffected control samples from the Epi25 Consortium as described in a recent publication (Niestroj et al., 2020). We subsequently restricted CNVs to $\geq$10 probes and $\geq$25kb in size. After CNV filtering, samples with >25 CNV calls were excluded as outliers, retaining a total of 12,053 cases and 8,173 controls.
7. Estonian Biobank (EstBB): We obtained CNV calls for 201,385 samples from the Estonian Biobank using the CNV detection pipeline described in a previous publication (Auwerx et al., 2022; Leitsalu et al., 2015; Macé et al., 2016). We excluded samples that were not included in SNP imputation and that had genotype calls missing for $\geq$2% of sites when processing the raw microarray data. We also excluded all samples from two genotyping batches based on visual inspection of genotyping intensity parameters, followed by further exclusion of genotyping plates ($\leq$24 samples per plate) that contained >3 samples with >200 CNV calls. We retained unrelated samples that had been linked to Estonian health registries and that had $\leq$50 raw CNV calls. We included CNVs with a quality score $\geq$15, were covered by $\geq$10 probes, and were $\geq$25kb in size. Finally, after pruning related samples and any samples with known malignant cancers or chromosomal disorders (*e.g.*, Down's Syndrome, sex chromosome aneuploidies), we retained a total of 78,842 unrelated samples for subsequent analyses.
8. GeneDX (GDX): We obtained CNV calls for 76,208 clinical samples referred to GeneDX for routine genome-wide microarray testing, 21,932 of which were profiled on Affymetrix CytoScan HD arrays and 54,276 of which were profiled on a custom Agilent SNP arrays or array comparative genomic hybridization (aCGH). Sample types were predominantly postnatal and

consisted of blood, buccal, oral rinse, or DNA provided by the client with most clinical indications including a neurodevelopmental disorder such as intellectual disability, developmental delay, autism, and/or seizures. Postnatal samples were run either on a 180k Agilent array (118,000 copy number probes; 66,000 SNP probes) or the Affymetrix CytoScan HD array. Prenatal samples consisted of cultured or uncultured amniotic fluid or chorionic villus samples or client-provided DNA and were referred for a variety of reasons including abnormal fetal sonographic finding, abnormal noninvasive prenatal test result, abnormal maternal serum screen, previous history of child or relative with chromosomal or genetic disease. Prenatal samples were run either on a 60k Agilent array (42,000 copy number probes and 18,000 SNP probes) or on the CytoScan HD microarray. Data from Agilent arrays were analyzed using Agilent Genomic Workbench 7.0 while data from CytoScan HD arrays were analyzed with the Applied Biosystems Chromosome Analysis Suite. All CNVs were required to be $\geq$20kb and <40Mb in length. Except for the minority of 9,958 samples for which additional CNV call-level metadata was unavailable, all CNVs were further required to not have been annotated as a suspected false positive or mosaic event, have estimated copy numbers $\leq$1.5 for deletions or $\geq$2.5 for duplications, include $\geq$10 probes and have P(CNV) $\leq 10^{-10}$. Following CNV call filtering, we excluded a total of 2,184 samples that either had >10 calls each, were identified as potential biological replicates, were referred for testing due to being a relative of a known carrier of a medically relevant CNV, or had been referred for testing due to advanced maternal age. In total, we retained 74,028 samples for subsequent analyses. Finally, we converted the subset of CNVs provided in hg18 coordinates to GRCh37 using the UCSC liftOver tool while requiring 50% of each CNV to remap contiguously to GRCh37 (Lee et al., 2019).

9. Indiana University (IU): Chromosomal microarray analysis was performed on genomic DNA extracted from peripheral blood or buccal swab samples from 1,673 consecutive patients tested in the Indiana University Cytogenetics Laboratory during the period of January 2018 through August 2019. These samples were evaluated using the CytoScan HD Microarray platform and the data were analyzed via Chromosome Analysis Suite software version 3.3 (ThermoFisher Scientific, USA). All non-artifact copy number variants (CNVs) identified in these samples were retained, irrespective of clinical significance. Following CNV detection, we excluded samples derived from buccal swab DNA, samples with known aneuploidies or large runs of homozygosity, and samples with no phenotypic indication specified. After filtering, we retained a total of 1,576 samples.
10. Ontario Population Genomics Platform (Ontario): We obtained CNV calls for 873 healthy population control samples analyzed on the CytoScan-HD platform via NCBI dbVar accession #estd212 as described in a previous publication (Uddin et al., 2015).
11. Psychiatric Genetics Consortium (PGC): We obtained CNV calls for 21,094 samples with schizophrenia and 20,277 unaffected control samples as described in a previous study (Marshall et al., 2016). We subsequently converted the CNVs from hg18 to GRCh37 coordinates using the UCSC liftOver tool while requiring 50% of each CNV to remap contiguously to GRCh37 (Lee et al., 2019).
12. Radboud University Medical Center (RUMC): We obtained CNV calls for 5,531 individuals with intellectual disability analyzed by RUMC via NCBI dbVar accession #nstd85 as described in a previous publication (Vulto-van Silfhout et al., 2013).
13. SickKids Hospital (SickKids): We obtained CNV calls for 2,691 samples with a variety of neurodevelopmental or neuropsychiatric disorders and 1,769 unaffected relatives from NCBI dbVar accession #nstd173 as described in a previous publication (Zarrei et al., 2019). We retained CNVs $\geq$25kb in size. After CNV filtering, samples with >80 CNV calls were excluded as outliers. We identified a single locus on chr12 that had CNVs only appearing in ADHD samples at 2.8% frequency; these CNVs were removed from the callset. Finally, given that all unaffected samples were first-degree relatives of the affected probands, we excluded all unaffected relatives from subsequent analyses. After filtering, we retained a total of 2,689 samples.
14. Simons Simplex Collection (SSC): We obtained CNV calls for 2,795 samples with autism spectrum disorder as described in a previous publication (Sanders et al., 2015). CNVs were filtered on P(CNV) $\leq 10^{-9}$, per recommendation of the authors. We subsequently converted the CNVs from hg18 to GRCh37 coordinates using the UCSC liftOver tool while requiring 50% of each CNV to remap contiguously to GRCh37 (Lee et al., 2019).
15. The Cancer Genome Atlas (TCGA): We obtained CNV calls for 8,670 whole-blood derived normal (*i.e.*, non-tumor) samples from donors with non-hematological cancers as described in a previous publication (Zack et al., 2013). We subsequently restricted CNVs to $\geq$10 probes and $\geq$25kb in size. Finally, to enrich for true germline CNVs and protect against tumor contamination, we further required deletions and duplications to have $\log_2$ intensities $\leq \log_2(0.6)$ and $\geq \log_2(1.45)$, respectively.
16. The Genetic Etiology of Tourette Syndrome Consortium (TSAICG): We obtained CNV calls for 2,434 samples with Tourette Syndrome and 4,093 unaffected controls as described in a previous publication (Huang et al., 2017).
17. UK BioBank (UKBB): We obtained CNV calls for 480,051 samples from the UK BioBank that were genotyped on a custom ThermoFisher Axiom microarray containing approximately 850,000 SNP probes. CNV calls were generated using methods that have been described in previous publications (Auwerx et al., 2022; Macé et al., 2017). We subsequently restricted CNVs on quality score $\geq$17 and $\geq$25kb in size. After CNV filtering, samples with >10 CNV calls were excluded as outliers as well as any samples with known malignant cancers or chromosomal disorders (*e.g.*, Down's Syndrome or sex chromosome aneuploidies), which retained a total of 429,871 samples.

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

CNV harmonization

After collecting and processing CNV calls for each cohort as described above, we next subjected all CNV calls from each cohort to the same set of global filters, as described below:

1. CNV defragmentation: Many microarray-based CNV calling algorithms infrequently introduce fragmentation (*i.e.*, over-segmentation) of large CNV calls. For the purposes of this study, fragmentation of CNV calls has the potential to bias association tests, as single individuals might be counted multiple times for a given locus or gene. Thus, we applied a standardized defragmentation step to raw CNV calls for all studies where this was possible. We defragmented each raw CNV dataset by merging CNVs of the same type (deletion or duplication) found in the same sample if their breakpoints are within ±25% of the size of their corresponding original CNV calls. We applied this process to all cohorts except PGC, TSAICG, BioVU, Ontario, RUMC, and control samples from Coe and Cooper, which were unable to be defragmented due to inadequate sample-level metadata.
2. CNV size & ploidy: We excluded all CNVs not mapped to autosomes from the primary GRCh37 assembly and further required all CNVs to be $\geq$100kb and $\leq$20Mb in size.
3. Maximum frequency from population sequencing: We excluded CNVs that matched reported common CNVs (allele frequency >1%) from any annotated population in four CNV reference catalogs derived from high-coverage genome sequencing (Abel et al., 2020; Almarri et al., 2020; Byrska-Bishop et al., 2021; Collins et al., 2020). We protected against imprecise allele frequency estimates due to populations with small sample sizes by computing the lower bound of the 95% binomial confidence interval for each variant's allele frequency and used this more conservative estimate when restricting to common variants. The version of gnomAD-SV used for this analysis (gnomAD-SV v2.1, non-neuro) included 8,342 samples without known neuropsychiatric disorders as available from the gnomAD website and described in a previous publication (Collins et al., 2020). To ensure sufficient sample size for allele frequency estimates, samples from the 1000 Genomes Project (1KGP) (Byrska-Bishop et al., 2021) and the Human Genome Diversity Panel (HGDP) (Almarri et al., 2020) were evaluated at the level of continental or regional populations rather than subpopulations. To be considered a match, CNVs in our dataset were required to have a matching CNV type (deletion or duplication), $\geq$50% reciprocal overlap by size as calculated by BEDTools (Quinlan and Hall, 2010), and both coordinates from the CNV breakpoint were required to be within ±100kb of the corresponding genome sequencing-based CNV.
4. Maximum frequency in this study: We excluded CNVs that overlapped other CNVs in at least 1% of samples within the same dataset or in any of the other array CNV datasets. When determining the maximum permitted number of overlapping CNV carriers, we computed an adaptive cutoff per cohort corresponding to the upper bound of the 95% binomial confidence interval matching a 1% frequency within each cohort (*i.e.*, the smallest CNV frequency at which we could be 95% confident that the true frequency was >1%). For this filter, we compared each cohort's CNV calls to the raw CNVs (prior to filtering) of each other cohort pairwise in serial and determined matching CNV calls using the same criteria as described for the population sequencing-based frequency filter (as in step 3, above). CNVs with at least 75% reciprocal overlap versus known genomic disorder CNV loci reported by at least two of six established resources (described below) were excluded from frequency filtering, as we reasoned that some of the most common genomic disorders may appear at >1% frequency in cohorts predominantly comprised of severely affected individuals (*e.g.*, GDX, IU, RUMC, *etc.*).
5. Sequence context: We excluded all CNVs with $\geq$50% coverage by somatic hypermutable sites (*e.g.*, T-cell receptor genes, antibody parts, *etc.*), segmental duplications, simple/low-complexity/satellite repeats, or N-masked bases of the GRCh37 reference assembly. Coverage was computed using BEDTools, repeats were obtained from the UCSC Genome Browser, and a list of somatic hypermutable sites was repurposed from a previous study (Collins et al., 2020; Lee et al., 2020; Quinlan and Hall, 2010). CNVs with at least 75% reciprocal overlap versus known genomic disorder CNV loci reported by at least two of six established resources (described below) were excluded from frequency filtering, as many of the most common genomic disorders are mediated by non-allelic homologous recombination (NAHR) between large regions of flanking segmental duplications, and we did not want to artificially remove these clinically important CNVs from our callset based on their proximity to the NAHR-mediating repetitive regions.

After applying the CNV harmonization procedures described above to the raw CNV calls from each cohort, we pooled harmonized CNV data across cohorts into seven matched groups, dubbed "meta-cohorts," to control for technical differences between individual data sources and cohorts. These meta-cohorts represent the basic unit on which all association testing was performed. Individual cohorts were assigned to meta-cohorts based on similarity between sample recruitment strategies, study or cohort design, microarray platforms, and processing pipelines for each callset. Original cohorts were never divided among multiple meta-cohorts except for the UKBB: we found that the UKBB Axiom array design was the best match for a large fraction of clinical samples from the GDX cohort, and thus randomly selected a subset of 13,139 healthy UKBB controls (dubbed "UKBB subset") to pair with GDX while retaining the remaining 416,732 UKBB samples (dubbed "UKBB main") as an independent cohort. Meta-cohort assignments were as follows:

- *Cohort 1*: BCH, Coe, Cooper, IU.
- *Cohort 2*: Epi25k, GDX, Ontario, TSAICG, UKBB (subset).

- *Cohort 3*: PGC, RUMC, SSC, SickKids, TCGA.
- *Cohort 4*: CHOP.
- *Cohort 5*: UKBB (main).
- *Cohort 6*: EstBB.
- *Cohort 7*: BioVU.

Phenotype curation

We applied a standardized phenotype consolidation scheme across all samples to convert all sample-level phenotype information into Human Phenotype Ontology (HPO) terms (Köhler et al., 2019). This process is described below.

First, for each sample, we recursively matched all available phenotype information against HPO keywords based on keyword substrings and assigned the corresponding HPO terms to each sample when a match was found. If no exact matches were found, we iterated through the HPO keywords again while searching for at least 85% fuzzy matching between each phenotypic indication at least 10 characters long and any HPO keyword. We abided by all HPO phenotype definitions with a single exception: we matched all congenital anomalies to HP:0001197 (Abnormality of prenatal development or birth), despite the official description of that term excluding fetal structural anomalies. We made this exception because congenital anomalies were a phenotype of particular interest in this study, and there was no obvious existing HPO code corresponding to these phenotypes collectively. Finally, we assigned any samples that failed to match into any existing HPO codes into an ''unspecified'' category, which can occur in the case of certain named multi-system disorders (*e.g.,* an indication of ''Noonan Syndrome'' could fail to map into our HPO scheme absent any more specific indications about the affected organ systems or other clinical features). Effectively all such ''unspecified'' samples came from two clinical laboratory datasets (GeneDx & CHOP), where we were able to confirm that these samples were generally referred for genetic testing by a physician due to a phenotypic indication with an elevated prior likelihood on the possibility of pathogenic rCNVs, such as intellectual disability, developmental delay, or other syndromic disorders. Thus, we included these samples as a separate phenotype category as they were likely still informative for our analyses despite lacking the appropriate phenotypic information to assign them to any unique HPO term(s).

For the UKBB and EstBB cohorts, all phenotypes were originally encoded as ICD-10 codes, while BioVU phenotypes were encoded in ICD-9. We converted BioVU ICD-9 codes to ICD-10 and uniformly filtered all ICD-10 codes to a smaller subset relevant to this study. This was accomplished by (1) automated filtering to isolate ICD-10 codes with a cohort prevalence of at least 0.01% but no greater than 5%, and (2) manual review by a board-certified physician (P.M.B.) to further identify terms unlikely to have a strong risk component attributable to rare genetic variants. Afterwards, phenotypes for each sample were converted from ICD-10 to plain-text indications using the ICD-10 descriptions for each term. Once converted to plain-text indications, the UKBB, EstBB, and BioVU cohorts were subjected to indication-to-HPO conversion as described above.

Second, after assigning HPO terms to each sample, we reduced the number of HPO terms used in this study to those with $\geq$3,000 samples in total and $\geq$300 samples contributed by three independent cohorts. Next, we compared shared sample memberships between all pairs of related HPO terms to determine the optimal subset of non-redundant terms. If any pair of HPO terms had a Jaccard similarity (*i.e.*, intersection divided by union) >80%, we retained the larger term and dropped the smaller term. If <3,000 samples differed between a pair of hierarchically related HPO terms, we retained the more general (*i.e.*, higher-level) term, and excluded the more specific (*i.e.*, lower-level) term. If both terms were equally high-level and siblings (defined as reciprocally sharing at least 50% of their parent terms), we retained the term with the larger sample size. Finally, we manually reviewed all HPO terms passing the above curation steps and pruned nine HPO terms that were determined to be indistinguishably similar to another retained HPO term of larger size (*e.g.*, autistic behavior [HP:0000729] vs. autism [HP:0000717]; arterial stenosis [HP:0100545] vs. peripheral arterial stenosis [HP:0004950]). In total, this process yielded a condensed hierarchical phenotype classification system with 54 distinct HPO terms, each which had $\geq$3,000 samples in total, $\geq$300 samples from at least three different meta-cohorts and differed from all other HPO terms by $\geq$3,000 distinct samples. The final list of HPO terms and samples per term per cohort is provided in Table S2.

Curation of other datasets

Previously reported genomic disorders

We curated lists of previously reported genomic disorders (GDs), defined as genomic intervals where rare CNVs have been linked to one or more diseases. For this purpose, we integrated lists of GDs from six existing publications and public resources (Dittwald et al., 2013; Firth et al., 2009; Girirajan et al., 2012; Owen et al., 2018; Riggs et al., 2012; Stefansson et al., 2014). We extracted GD coordinates from these references as provided and converted them to GRCh37 using UCSC liftOver where necessary (Lee et al., 2020) before excluding GDs <200kb or >10Mb in size. We only considered GDs from the ClinGen Pathogenic CNV Regions list that were scored at high- or medium-confidence. To integrate all six sources, we first computed the density of reported GD intervals in 10kb windows across the genome using BEDTools separately for deletions and duplications. We next sequentially curated GDs into three groups based on their degree of concordance across all six sources: high confidence ($\geq$4/6 sources), medium confidence (2-3/6 sources), and low confidence (just one source). In each iteration, we excluded GDs from each of the six original sources that had any greater than 20% overlap with a consensus GD in a higher confidence tier. The result of this curation process was to preferentially

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

retain the coordinates for each GD agreed upon by the greatest fraction of sources and drop regions flanking high-confidence GDs where just one or a few sources likely over-reported the GD breakpoints. After overlapping all GDs from the six sources above, we trimmed overlapping segmental duplications overlapping the boundaries of each GD interval, if any were present. Segmental duplications were downloaded from the UCSC Genome Browser in GRCh37 coordinates (Lee et al., 2020).

Predicted NAHR-mediated CNV regions

We defined a set of loci where NAHR-mediated CNVs might be predicted to occur based on the genomic properties of the flanking regions. To build this set of loci, we first defined pairs of segmental duplications meeting all the following criteria:

1. Both members of the pair mapped to the same chromosome;
2. The pair was no closer than 100kb and no farther than 10Mb apart;
3. Both members of the pair were $\geq$1kb in size;
4. The pair featured strict sequence identity (no indels) $\geq$90%;
5. The pair mapped with direct orientation of repeats (*i.e.*, same strand);
6. The total intervening sequence between members of the pair had $\leq$30% coverage by the excluded regions used during CNV harmonization (as described above; *e.g.*, simple repeats); and
7. At least 100kb of the intervening sequence remained after subtracting the excluded regions used during CNV harmonization.

After defining candidate pairs of segmental duplications (above), we collapsed overlapping pairs into predicted NAHR-mediated CNV regions while requiring $\geq$50% reciprocal overlap of intervening sequence per BEDTools as well as both ends of their respective intervals to be within 1Mb of each other (Quinlan and Hall, 2010). For each cluster of segmental duplication pairs, we retained the pair with the smallest intervening (*i.e.*, spanning) distance, and used the innermost coordinates for analysis purposes.

Genes and gene sets

For all analyses in this study, we defined genes strictly as canonical transcripts from autosomal protein-coding genes as recorded in Gencode v19 (Frankish et al., 2019). We extracted canonical transcripts for protein-coding genes according to the official Ensembl definition. We further restricted all canonical transcripts to exons expressed in $\geq$20% of transcripts in at least one human tissue catalogued by the Genotype-Tissue Expression (GTEx) Project (GTEx Consortium et al., 2017). To accomplish this, we computed the per-exon maximum proportion expressed across transcripts ("pext") score across all available tissues (Cummings et al., 2020). Bases from exons with missing pext scores were ignored when calculating mean pext scores per exon, and exons entirely missing pext scores were considered as passing the $\geq$20% requirement. Genes with no exons remaining after exon-level expression filtering were removed outright from all subsequent analyses.

We also considered various subsets of genes throughout this study. A full list of these gene sets is provided in Table S5, along with inclusion criteria and references for external data where necessary. When curating these gene sets, we required unique matches to gene symbols from one of the canonical protein-coding genes curated above.

Four gene sets are particularly prominent throughout our analyses and merit detailed explanation here. We curated sets of genes likely to be DS based on existing evidence by starting from the list of canonical protein-coding autosomal genes we curated for this study and applying various criteria, which is described below and graphically depicted in Figure S6.

- Haploinsufficient genes (N=264): all genes meeting at least three of the following four criteria: (1) constrained against PTVs in gnomAD v2.1 (Karczewski et al., 2020), (2) confirmed autosomal dominant causes of disease via LoF or haploinsufficient mechanisms per ClinGen ("high confidence" only) or DECIPHER/DDG2P ("confirmed" only), (3) intolerant against low-expressor samples in GTEx v7 (GTEx Consortium et al., 2017), or (4) had no loss-of-function deletions documented in 8,342 unrelated samples with no known neuropsychiatric diagnosis in gnomAD v2.1 (hereafter referred to as "gnomAD v2.1 non-neuro") (Collins et al., 2020). Details for each of these criteria are provided in Table S5.
- Triplosensitive genes (N=178): all genes meeting at least three of the following four criteria: (1) constrained against missense mutations in gnomAD v2.1 (Karczewski et al., 2020), (2) confirmed autosomal dominant causes of disease via triplosensitive, gain-of-function, or other unspecified mechanisms per ClinGen (triplosensitive genes of any confidence rating) or DECIPHER/DDG2P ("confirmed" gain-of-function or "other" mechanism only) (Rehm et al., 2015; Wright et al., 2015), (3) intolerant against high-expressor samples in GTEx v7 (GTEx Consortium et al., 2017), or (4) had no whole-gene copy-gain duplications documented in 8,342 unrelated samples from the non-neuro subset of gnomAD v2.1 (Collins et al., 2020). Details for each of these criteria are provided in Table S5.
- Haplosufficient genes (N=228): all genes meeting all of the following four criteria: (1) mutationally tolerant in gnomAD v2.1 (Karczewski et al., 2020), (2) no reported disease associations per OMIM, ClinGen, or DECIPHER/DDG2P (Amberger et al., 2015; Rehm et al., 2015; Wright et al., 2015), (3) high rate of low-expressor outlier samples in GTEx v7 (GTEx Consortium et al., 2017), and (4) had at least one loss-of-function deletion documented in 8,342 unrelated individuals from the non-neuro subset of gnomAD v2.1 (Collins et al., 2020). Details for each of these criteria are provided in Table S5.
- Triploinsensitive genes (N=150): all genes meeting all of the following four criteria: (1) mutationally tolerant in gnomAD v2.1 (Karczewski et al., 2020), (2) no reported disease associations per OMIM, ClinGen, or DECIPHER/DDG2P (Amberger et al., 2015; Rehm et al., 2015; Wright et al., 2015), (3) high rate of high-expressor outlier samples in GTEx v7 (GTEx Consortium

et al., 2017), and (4), had at least one whole-gene copy-gain duplication documented in 8,342 unrelated individuals from the non-neuro subset of gnomAD v2.1 (Collins et al., 2020). Details for each of these criteria are provided in Table S5.

Gene-level features

We compiled a core table of gene-level features to be used in gene-based fine-mapping and genic dosage sensitivity scoring. The full list of features is provided in Table S5. For each of the canonical protein-coding genes curated for this study, we collected a total of 145 gene-level features across five major categories: genomic, expression, chromatin, protein, and mutational constraint. All expression features were derived from GTEx v7 due to it being the last GTEx version native to GRCh37, which was a direct match for the Gencode version used in this study (GTEx Consortium et al., 2017; Frankish et al., 2019). All expression matrices from GTEx v7 were transformed as $\log_{10}(X+1)$ prior to per-gene feature calculation. All chromatin features were based on the Roadmap Epigenomics Project using the expanded 18-state ChromHMM model on 98 tissues as described in a previous publication (Roadmap Epigenomics Consortium et al., 2015). All protein features were extracted from the curated subset of human proteins annotated in the Swissprot database (Bairoch and Apweiler, 2000).

Following annotation of all 145 features per gene, we collapsed correlated annotations to retain the principal components, or "eigenfeatures," that captured at least 99% of inter-gene variance. Prior to principal components analysis (PCA), we normalized the data in two steps. First, all variables were transformed using Box-Cox power transformations across all genes, except for a small subset of features as noted in Table S5. Second, following transformation, all variables were centered (mean=0) and scaled (standard deviation=1) across all genes. We subsequently performed PCA on the transformed & standardized gene feature matrix and used the resulting principal components as features inputs to gene-based fine-mapping and dosage sensitivity scoring.

Global analyses of rCNVs in disease architecture

We assessed the contribution of rCNVs to each of the 54 disease phenotypes considered in our study through two analyses. In the first analysis, we computed the cumulative incidence of all GDs curated from the literature (defined above) per phenotype in each cohort and subsequently conducted a fixed-effect meta-analysis of GD rCNV counts per phenotype across all seven cohorts to define an overall OR and 95% CI separately for deletions and duplications as well as combined across CNV types using the metafor R package (Viechtbauer, 2010). We considered a rCNV as a match for a literature-based GD if the rCNV overlapped at least half of the consensus GD interval defined in our curation process. In the second analysis, we computed the cumulative incidence of all rCNVs overlapping a well-defined set of known PTV-constrained genes (defined in Table S5) per phenotype in each cohort after excluding all CNVs matching known GDs identified in the first analysis. We required each deletion or duplication to overlap $\geq$2% or $\geq$84% of a gene's coding sequence (CDS), respectively, to be counted for our meta-analyses; see below for a description of how these CDS threshold were selected. From these counts, we computed an OR and 95% CI using the same fixed-effect meta-analysis strategy as in the first analysis. We further partitioned phenotypes into two subgroups based on the ORs of deletions of constrained genes estimated in our second analysis: all phenotypes with OR$\geq$2 were assigned to a group of phenotypes with "strong rCNV effects", whereas phenotypes with OR<2 were assigned to a group with "weaker rCNV effects." In both the first and second analyses, we excluded any cohorts from meta-analyses for each phenotype if those cohorts had <300 total cases corresponding to the phenotype in question.

Large segment association meta-analyses

We designed a sliding window framework to scan all autosomes for signals of association between rCNVs and any of the disease phenotypes considered in this study. To accomplish this, we meta-analyzed counts of rCNVs in cases vs. controls in parallel for each phenotype as described below.

Sliding window model design

We generated sliding windows for all autosomes at 200kb resolution with 10kb step size and excluded any windows with $\geq$30% coverage by N-masked sequences or known somatically hypermutable sites (described above). The window size of 200kb was selected to approximately match the median rCNV size for most cohorts after harmonization. Next, we intersected rCNVs against all sliding windows separately for cases and controls for each cohort while requiring rCNVs to overlap at least 50% of each window to be counted. We conducted this procedure once for deletions and once for duplications per phenotype per cohort. After CNVs were tallied in cases and controls for each window, we next compared the ratios of CNV carriers between cases and controls per phenotype per cohort using a one-sided Fisher's exact test. While we did not use these single-cohort association statistics for any assessment of significance, and instead used a fixed-effects meta-analysis as described below, we did use the Fisher's exact test results for designating the single most significant cohort per window for each combination of CNV type and phenotype.

Meta-analysis association test

We combined rCNV association statistics across metacohorts for each sliding window using a fixed-effects meta-analysis implemented using the metafor R package (Viechtbauer, 2010). We applied two key modifications to this meta-analysis. First, given that rCNV counts per window are (a) sparse, (b) zero-inflated, and (c) had unbalanced case & control sample sizes for most phenotypes (*e.g.*, frequently >10- to 100-fold more controls than cases), we implemented an empirical continuity correction as previously proposed (Sweeting et al., 2004). Second, as sample size imbalance between cases and controls can distort test statistics in genetic association studies (Zhou et al., 2018), we applied a saddlepoint approximation correction to the test statistics for each phenotype

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

per CNV type to recalibrate our P-values. Our implementation of this procedure was based on a previously proposed algorithm (Dey et al., 2017). During saddlepoint approximation, we excluded all windows overlapping with one of the known GD we curated from the literature (matching on CNV type) and further winsorized the remaining Z-scores by 0.5% on both tails to protect against outlier windows with extreme effect sizes biasing an accurate re-approximation of the null. Finally, we excluded cohorts from each meta-analysis if they had <300 cases for the phenotype being analyzed.

To account for heterogeneity in microarray platforms across cohorts, we conditionally excluded individual cohorts from meta-analyses on a locus-specific basis. Cohorts were excluded if <90% of their control samples were processed on microarray platforms with <10 probes per 200kb window. This threshold of <10 probes was selected to correspond to the minimum number of probes required for CNV quality control in most of the datasets included in our analyses. We applied these criteria exclusively to control samples (rather than both cases and controls) because all control samples were genotyped on microarrays at least as dense—and often denser—than their cohort's respective cases. Therefore, a low probe density in cases would lead to false negatives (*i.e.*, the erroneous absence of CNVs in affected cases at a particular locus), as opposed to the potential for false positives driven by low probe density in controls. This rationale has been used previously by others in large-scale CNV association analyses (Coe et al., 2014; Cooper et al., 2011). After applying these criteria, a total of 264,034 distinct windows remained for our association meta-analyses.

Each phenotype and CNV type were meta-analyzed separately for a total of two genome-wide meta-analyses per phenotype. Further, for each combination of phenotype and CNV type, we conducted two versions of the same meta-analysis: once while including all seven cohorts, and once while conditionally excluding the single most significant cohort per window, as described above. We designated the results from the model including all seven cohorts as the "primary" statistics and designated the results from the conditional exclusion model as the "secondary" statistics for purposes of defining genome-wide significance, described below.

Assessing genome-wide significance

After conducting genome-wide meta-analyses for each combination of phenotype and CNV type, we next wanted to control false discovery rate (FDR) at an approximate equivalent to genome-wide significance. Estimating the number of independent tests performed across all windows, phenotypes, and CNV types is difficult due to numerous necessary assumptions, such as the independence of samples between phenotypes or the local correlation structure of CNV counts between neighboring windows. Instead, we assessed a "genome-wide" significance threshold of $P \leq 3.74 \times 10^{-6}$, which corresponds to a Bonferroni correction if applied to the number of non-overlapping 200kb windows with sufficient probe density to be tested in our analysis (effective N=13,339.6 independent window tested).

We considered a window to be genome-wide significant if its primary P-value exceeded the Bonferroni-corrected genome-wide significance threshold ($P \leq 3.74 \times 10^{-6}$), and it additionally satisfied at least one of the following two criteria: (1) its secondary P-value (as described above) was also nominally significant (P<0.05); and/or (2) at least two separate cohorts were nominally significant (P<0.05) per single-cohort Fisher's exact tests. In addition to defining windows at genome-wide significance, we also applied the Benjamini-Hochberg FDR correction procedure to all windows per phenotype and considered a window to be "FDR significant" at FDR Q < 0.01 if it also satisfied at least one of the two secondary criteria applied to our genome-wide significant windows above.

We empirically assessed the calibration of our genome-wide significant primary P-value threshold using a permutation-based approach similar to recent proposals for conventional GWAS of Biobank-scale datasets (Annis et al., 2021). The goal of this permutation strategy was to determine the most extreme P-value we would expect to observe by chance alone if the rCNVs in our analyses were distributed randomly with respect to sample phenotype, which would correspond to the P-value threshold at which we could assess our empirical meta-analysis results and expect effectively no false positive associations genome-wide for any one phenotype. To accomplish this, we repeated the following three steps 50 times for each CNV type and phenotype:

1. Permuted phenotypes labels for all rCNVs within each cohort while matching on CNV type (deletion or duplication);
2. Reran full sliding window rCNV association meta-analysis for each phenotype and CNV type combination using identical parameters and filters as in our discovery analysis from our empirical (non-permuted) dataset; and
3. Computed the most extreme P-value observed genome-wide for each CNV type and phenotype.

For each phenotype, we computed the median value from step 3 across all 50 permutations and reported this median value as the most extreme P-value expected to be seen by chance alone for that phenotype given our data. Finally, we compared these empirical P-value thresholds to our approximated Bonferroni-equivalent threshold to determine how appropriate the Bonferroni approximation was for our datasets and study design (Figures S2B and S2C).

Association refinement and annotation

For each locus where we identified a significant association between rCNVs and one or more phenotypes, we aimed to identify the minimal interval(s) that contained the causal element(s) with 95% confidence. This procedure was performed separately for deletions and duplications as follows:

First, all genome-wide and FDR significant windows per phenotype were collapsed into nonredundant blocks by merging all windows within ±100kb of any significant window. Next, for each block, we executed the following five steps:

1. Computed an approximate Bayes factor (ABF) for each window following the procedure specified by Dr. Wakefield in prior publications (Wakefield, 2007, 2009). To estimate the null variance of true causal loci, we used Bayesian model averaging

across three priors: (i) the average of all significant windows, (ii) the most significant window per block, and (iii) previously reported GDs, as described above. We estimated null variance separately for the subset of phenotypes with stronger rCNV effects and for those with weaker rCNV effects (see Figure 1B);

2. Identified the minimal set of windows that captured at least 95% of the total ABF for the block (*i.e.*, defined the 95% credible set) by ranking all n windows in descending order according to their ABF and subsequently computing the minimal set of k windows as follows:

$$95\%\ \text{credible set} = \left\{ k \in n : \frac{\sum_{i=1}^{k} \Lambda_i}{\sum_{i=1}^{n} \Lambda_i} \geq 0.95 \right\}$$

Where Λ_i is the ABF per window i;

3. Merged all windows in the 95% credible set to define the set of intervals comprising the 95% credible set, and extended each 95% credible set to include all other windows sharing at least 80% of all rCNVs (specifically, windows for which the intersection of rCNVs overlapping both windows divided by the union of rCNVs from either window exceeded 80%);
4. Clustered any overlapping credible intervals of the same CNV type (deletion or duplication) from different phenotypes while padding each credible interval by ±100kb;
5. Redefine the 95% credible set of windows jointly across all phenotypes per clustered locus while using the mean posterior probability across all phenotypes per window.

The result of this process was one 95% credible set for each genome-wide or FDR significant locus identified by our sliding window meta-analysis. After refinement, we computed two summary statistics per phenotype associated with each credible set:

1. Case and control rCNV frequencies, which were computed as the mean rCNV carrier rate across all windows in the 95% credible set;
2. Pooled effect sizes, which were computed as the inverse variance-weighted mean across all windows in the 95% credible set.

We computed two additional summary statistics for each rCNV segment (*i.e.*, credible sets defined jointly across all phenotypes):

1. Case and control rCNV frequencies, which were computed as the weighted mean rCNV carrier rate per 95% credible set, where the weights corresponded to the square root of the sample size for each phenotype;
2. Pooled effect sizes, which were computed as the inverse variance-weighted mean across all significant phenotypes per segment.

Following our sliding window meta-analysis and association refinement, we annotated each significant rCNV segment based on overlap with exons from protein-coding genes and overlap with previously reported GDs. We also attempted to infer the mechanism (NAHR vs. non-NAHR) of each rCNV segment by comparing all case rCNVs with 50% reciprocal overlap versus the credible interval(s) per segment versus the predicted NAHR intervals curated above. Segments were annotated as predicted NAHR if $\geq$20% of case rCNVs had $\geq$50% reciprocal overlap with at least one predicted NAHR interval. We manually polished all *in silico* NAHR predictions by comparing the segment coordinates to the GRCh37 segmental duplication, repeat masker, and assembly gap tracks in the UCSC Genome Browser (Lee et al., 2020) and revised any apparent discrepancies *post hoc*.

Genomic feature analyses of large rCNV segments

We assessed disease-associated rCNV segments for enrichments of various genomic and gene-level features using two different permutation-based approaches. First, for features based on genomic coordinate (*e.g.*, previously reported GDs), we conducted 100,000 random permutations while matching on rCNV segment size using a custom implementation of BEDTools shuffle (Quinlan and Hall, 2010). Second, for all gene-based features (*e.g.*, proportion of constrained genes, number of disease genes), we conducted 100,000 random permutations while matching on number of genes per rCNV segment. In this gene-matched permutation approach, we randomly seeded each permuted segment by selecting one gene at random from the list of all autosomal protein-coding genes considered in our analyses and iteratively added genes based on minimal linear distance (*i.e.*, closest neighboring gene first) until the total number of genes was matched to the number overlapped by the original rCNV segment. In this way, we generated randomly permuted lists of rCNV segments matched exactly on the number of genes per segment. In both the size-matched and gene-matched permutation schemes, we applied the same two restrictions: (1) we did not allow shuffled segments of the same CNV type to overlap, and (2) we did not consider any regions not covered by at least one 200kb window assessed in our sliding window meta-analyses.

For most features, we assessed enrichment by computing the average value across all rCNV segments for all 100,000 permutations and comparing the distribution of permuted values to the empirically observed average value of the rCNV segments from our disease-association meta-analyses. The exception was for all features relating to DNMs from prior exome sequencing studies of

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

NDDs and ASD (Fu et al., 2021; Kaplanis et al., 2020). For DNM-related features, we computed the excess DNMs per segment by subtracting the expected number of DNMs per segment extrapolated based on gene-specific mutation rates (Karczewski et al., 2020), and further normalized excess PTV or missense DNMs against synonymous DNMs by residualizing against an outlier-robust linear fit of excess of PTV or missense DNMs predicted from the excess of synonymous DNMs. This step was necessary to control for rare instances where we observed regional miscalibration of gene-based mutation rates *vis a vis* the observed DNM data.

Gene-based association meta-analyses & fine-mapping

We designed a gene-based meta-analysis model to identify individual genes where rCNVs were associated with one or more diseases, as detailed below.

Gene-based model design

Our gene-based association test compared counts of gene-overlapping rCNVs in cases vs. controls per cohort across all 54 phenotypes considered in this study. We begin by enumerating all autosomal protein-coding genes that passed our curation procedure as described above. We further excluded genes if their canonical transcript was $\geq$30% covered by any of the excluded regions during CNV harmonization (*e.g.*, somatically hypermutable sites, simple repeats, *etc.*). After all filtering, we retained 170,428 exons from 17,263 genes for these analyses.

We next empirically optimized the minimum overlap to require between an rCNV and a gene's total coding sequence (CDS) for association testing. To do this, we annotated every rCNVs from a sample matching at least one phenotype with stronger rCNV effects (see Figure 1) for overlap with any gene from our curated lists of haploinsufficient (for deletions) or triplosensitive (for duplications) genes, as defined above. For this analysis, we excluded all genes overlapping a known NAHR-mediated genomic disorder previously reported in the literature, since the CDS overlap fractions are not informative for these loci as many NAHR-mediated CNVs share nearly identical breakpoints. Where overlap was identified, we further annotated the fraction of total CDS from that gene covered by the rCNV. Next, we meta-analyzed the number of rCNVs in cases versus controls for a range of potential minimum CDS thresholds, stepping from 1% to 100% in increments of 1%. At each potential threshold, we computed association statistics using the same fixed-effects meta-analysis as for our sliding window analyses. For deletions, we observed an essentially monotonic inverse relationship between significance of association and effect size, consistent with larger deletions being less frequent but having stronger effects, as expected. Therefore, we selected the minimum CDS threshold for deletions that maximized statistical power, which corresponded to $\geq$2% CDS. For duplications, we observed a non-monotonic relationship between effect size and power, presumably due to a stronger association emerging for truly triplosensitive genes at higher CDS thresholds where the duplication CNVs might more reliably result in whole-gene copy gain. Given this, we selected the minimum CDS threshold greater that yielded maximum statistical power for duplications, bounded on a lower limit of 50% CDS to enrich for true whole-gene copy-gain events; this procedure yielded an optimal CDS cutoff of $\geq$84% for duplications.

Next, to conduct exome-wide association tests, we tallied genic rCNVs in cases vs. controls for each CNV type, cohort, and phenotype. For each rCNV-gene pair, we computed the CDS fraction from the gene overlapped by the rCNV and considered a rCNV to overlap a gene based on its CDS overlap conditional on its CNV type using the optimized cutoffs determined above ($\geq$2% for deletions and $\geq$84% for duplications). As in our sliding window analyses, we computed per-cohort association statistics with one-sided Fisher's exact tests for each gene, CNV type, and phenotype, but only used these statistics for determining the individually most significant cohort per gene for each combination of CNV type and phenotype.

Meta-analysis association test

We conducted exome-wide association meta-analyses for each phenotype and CNV type using identical methods as for our sliding window analyses described above: a fixed-effects model with empirical continuity correction and saddlepoint approximation to update the null distribution used for inference (Dey et al., 2017; Sweeting et al., 2004). We also conditionally excluded cohorts on a per-gene basis due to low probe density in control samples; however, given that our minimum CNV size was 100kb, we expanded the boundaries of each gene to be at least 100kb in size when computing probe density.

Assessing exome-wide significance

We assessed our gene-based rCNV association tests at two thresholds: (1) a Bonferroni-equivalent of exome-wide significance of $P \leq 2.90 \times 10^{-6}$ (*i.e.*, adjusting for 17,230 genes with sufficient probe density to be tested in at least two independent cohorts), and (2) a Benjamini-Hochberg corrected FDR $Q < 0.01$. In practice, we considered a gene to be exome-wide or FDR significant if its primary P- or Q-value exceeded the corresponding significance threshold, and it additionally satisfied at least one of the following two criteria: (1) its secondary P-value (as described above) was also nominally significant ($P<0.05$); and/or (2) at least two separate cohorts were nominally significant ($P<0.05$) per single-cohort Fisher's exact tests.

Fine-mapping genic associations

Lastly, we aimed to prioritize individual likely causal genes and to define the minimal set of genes per block that most confidently explains each phenotypic association. To accomplish this, we first clustered all significant genes per phenotype into "gene blocks" by computing the rCNV covariance between all pairs of genes, defined as the intersection of rCNVs impacting both genes divided by the union of all rCNVs impacting either gene. We constructed gene blocks by clustering all genes with rCNV covariance $\geq$20%, subset to gene blocks that contained at least one exome-wide or FDR significant gene, and subsequently padded each gene block with all other genes within ±200kb of that block even if those additional genes did not meet the $\geq$20% rCNV covariance criterion. We next adapted several established fine-mapping algorithms from conventional GWAS to our rCNV data as follows:

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

1. We transformed all association statistics per phenotype for all genes per block—including non-significant genes—into approximate Bayes factors (ABFs) following the same procedure as in our sliding window analysis (Wakefield, 2007, 2009). For these calculations, we used an empirical Bayes approach to estimate null variance based on the observed variances for (i) all exome-wide and FDR significant genes pooled across all phenotypes, (ii) the most significant gene per gene block, and (iii) the curated lists of gold-standard haploinsufficient and triplosensitive genes for deletions and duplications, respectively. We applied Bayesian model averaging across these different null variance estimates separately for developmental and adult-onset phenotypes.
2. We calculated posterior inclusion probabilities (PIPs) for each gene per phenotype per block based on the ABFs from [1] while imposing a flat prior (*i.e.*, when assuming that each gene per block is equally likely to be causal) using the same formulation as previously proposed (Wakefield, 2007, 2009).
3. We developed an adaptation of E-M algorithms described in two previous studies (Kichaev et al., 2017; Wen et al., 2017) to iteratively update gene priors for each block based on functional enrichments in a logit regression model. For this purpose, we used the gene-level principal components derived from all 145 gene features described earlier and optimized regression hyperparameters using a grid search for lasso/ridge mixture coefficient and regularization weight. We assessed convergence of this E-M process based on the root mean squared error of both PIPs and feature coefficients falling below 10^{-8} between subsequent iterations of the E-M process. In each iteration, we used the max PIP for genes associated with multiple phenotypes. When training the logit model, we excluded genes corresponding to NAHR-mediated (*i.e.*, recurrent) CNV regions as defined earlier; due to identical CNV breakpoints in most CNV-carrying samples for these regions, the genes often featured nearly identical association statistics and therefore were uninformative for prioritizing features associated with stronger local associations. In each training step, we also supplemented the genes from our empirically defined gene blocks with the curated lists of gold-standard haploinsufficient/haploinsufficient or triplosensitive/triploinsensitive genes when fine-mapping deletions or duplications, respectively; these genes were introduced into our training set with PIP=1 for haploinsufficient and triplosensitive genes and PIP=0 for haplosufficient and triploinsensitive genes.
4. After refining PIPs for each gene based on our functional E-M procedure from [3], we clustered overlapping gene blocks across phenotypes if two or more gene blocks from different phenotypes shared at least one gene that was significant (at either exome-wide or FDR<0.01 thresholds) in both phenotypes. For gene blocks clustered across multiple phenotypes, we computed the unweighted mean of all PIPs for each gene, and re-scaled PIPs such that the sum of all PIPs after clustering was equal to 1.
5. Finally, we computed a 95% credible set of genes per block based on the cumulative sum of PIPs for genes ranked by causal likelihood. We only retained exome-wide and FDR significant genes after calculating each credible set (*i.e.*, genes that were not originally exome-wide significant were included when calculating the 95% credible set but were removed from the credible set after calculation).

For downstream analyses, we considered any gene with PIP $\geq$ 0.2 to be a "confident" candidate gene, and any gene with PIP $\geq$ 0.8 to be a "very confident" candidate gene.

Genic dosage sensitivity scoring

We developed a model to predict haploinsufficiency and triplosensitivity for each protein-coding gene across all 22 autosomes, as detailed below.

Parameter selection & optimization

We assigned six parameters used in model training as follows:

- CDS overlap per gene: the goal of our model was to predict the consequences of whole-gene loss or gain (*i.e.*, 100% of CDS deleted or duplicated). However, given the possibilities of CNV breakpoint imprecision, transcript misannotation, and other technical factors, we required $\geq$84% CDS overlap for a CNV to count as gene-disruptive for both deletions and duplications to match the effects observed for duplications in our previous CDS overlap optimization analyses.
- Phenotypes: we restricted all gene scoring analyses to the subset of phenotypes with stronger rCNV effects (see Figure 1B).
- rCNV mechanism: we restricted all analyses to rCNVs that featured <50% reciprocal overlap versus any predicted NAHR interval, as defined above.
- Maximum number of genes per rCNV: given that our dosage sensitivity model aimed to quantify the effects of rCNVs on individual genes, we assigned a limit on the maximum number of genes an rCNV could disrupt in order to be considered by our model. To determine the optimal threshold of genes per rCNV, we determined the point at which we preferentially retained as many "informative" rCNVs (*i.e.*, those that overlapped at least one curated gold-standard dosage sensitive or insensitive gene but did not overlap genes from both categories) as possible while depleting for "uninformative" rCNVs (*i.e.*, those that overlapped both sensitive and insensitive genes and may therefore cause inflated effect sizes for the insensitive genes neighboring sensitive genes). To determine this threshold, we computed the size distributions of informative rCNVs and uninformative rCNVs in terms of number of genes per rCNV and subsequently approximated the probability distribution functions (PDFs) of both categories. We selected a cutoff where the difference in PDFs between informative and uninformative rCNVs equaled

zero; i.e., the point at which increasing our threshold by +1 gene per rCNV would result in a relative enrichment of uninformative rCNVs compared to informative rCNVs. We computed these thresholds separately for deletions and duplications in all seven cohorts and retained the average threshold across both CNV types and all cohorts as the final cutoff to use in our models ($\leq$37 genes per rCNV).

- Maximum effect size standard error per gene: lastly, we wanted to exclude genes from model training that had insufficient rCNV evidence to derive precise effect size estimates. Given that precision of effect size estimates varied between genes based on the distribution of rCNVs across all seven cohorts and their corresponding sample sizes, we set a uniform criterion by requiring genes to have a standard error <1 for the log-odds ratio (log-OR) estimated by our fixed-effect meta-analyses in order to be included in our dosage sensitivity model training.
- Gene exclusion criteria: we only considered genes that could be accurately modeled when estimating priors and when training the dosage sensitivity classifier. Specifically, we excluded genes that met either of the following two criteria. First, we excluded genes that had insufficient rCNV evidence to derive precise effect size estimates. Given that precision of effect size estimates varied between genes based on the distribution of rCNVs across all seven cohorts and their corresponding sample sizes, we set a uniform criterion by requiring genes to have a standard error <1 for the log-OR estimated by our fixed-effect meta-analyses. Second, we excluded any genes that overlapped a NAHR-mediated GD, as curated above. Note that we re-introduced any genes from our curated positive or negative training gene sets when training the dosage sensitivity classifier even if they met one of these two exclusion criteria.

Empirical Bayes estimation of priors

Our model required estimating three priors: (1-2) the CNV effect sizes expected for dosage sensitive and insensitive genes, and (3) the expected fraction of all genes in the genome that are truly dosage sensitive. To estimate the expected effect sizes, we computed the median log-OR across all genes in each training gene set of likely dosage sensitive or insensitive genes. To estimate the expected fraction of truly dosage sensitive genes, we averaged across three statistics derived from existing data: (i) 16.3% of all genes that are known to be constrained against PTVs in gnomAD v2.1(Karczewski et al., 2020); (ii) 21.7% of all genes have at least one disease association reported in OMIM (Amberger et al., 2015); (iii) 2.9% of all genes are known to feature dominant haploinsufficiency per ClinGen and/or DDG2P (Bragin et al., 2014; Rehm et al., 2015). Averaging across all three of these values produced a prior expectation of 13.7% of all genes being truly dosage sensitive.

Bayesian dosage sensitivity likelihoods

We next computed the likelihood ratio that each gene was dosage sensitive versus insensitive based on the CNV effect size empirically observed in our dataset. For each gene, g, with observed log-OR $\widehat{\theta}_g$ and standard error $\widehat{V}_g$, we specified null and alternative hypotheses as follows:

$$\begin{cases} H_0 \ : \ \widehat{\theta}_g \leq \theta_0 \\ H_1 \ : \ \widehat{\theta}_g \geq \theta_1 \end{cases}$$

Where θ_0 and θ_1 are the effect size priors for dosage insensitive and sensitive genes, respectively, computed on a CNV type-specific basis as detailed above. From these hypotheses, we computed the Bayes factor per gene as:

$$BF_g = \frac{\mathrm{P}(\widehat{\theta}_g|H_0)}{\mathrm{P}(\widehat{\theta}_g|H_1)} = \frac{1 - N(\widehat{\theta}_g - \theta_0, 1)}{N(\widehat{\theta}_g - \theta_1, \widehat{V}_g)}$$

Where $N(m, v)$ denotes the normal distribution. We assumed unit variance (v) for H_0. Subsequently, we computed the Bayesian false discovery probability (BFDP) per gene based on equation (2) from (Wakefield, 2007) as:

$$BFDP_g = \mathrm{P}(H_0|\widehat{\theta}_g) = \frac{PO \times BF_g}{PO \times BF_g + 1}$$

Where the prior odds (PO) of each gene being dosage insensitive was:

$$PO = \frac{\pi_0}{1 - \pi_0}$$

Where $\pi_0 = 1 - \pi_1 = 0.863$, given the prior of $\pi_1 = 0.137$ as estimated above.

Finally, we manually reassigned BFDP values of 0 or 1 to any gene in our curated training sets of likely dosage sensitive or insensitive gene, respectively.

Ensemble machine learning strategy & composite architectures

Using the BFDP values calculated for each gene described above, we next trained and applied an ensemble machine learning framework to predict the probability of a gene being haploinsufficient or triplosensitive based on 145 gene-level features. This process is described below and was conducted independently for deletions to predict haploinsufficiency and for duplications to predict triplosensitivity.

We first partitioned all genes into 10 subsets. Given that many gene-level features and our CNV evidence exhibited autocorrelation among neighboring genes, we did not randomly assign genes into these 10 subsets. Instead, we split all chromosomes by arm (*i.e.*,

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

p or q), and grouped chromosome arms into 10 subsets while balancing the average number of genes per subset. Each of these 10 subsets therefore had roughly the same number of genes but the data for each subset was entirely independent from the other 9 subsets, as no rCNVs in our dataset were allowed to span centromeres dividing p and q arms.

We next used 10-fold cross-validation (CV) to train a dosage sensitivity classifier based on these 10 subsets. For each CV fold, we trained a classifier on 9/10 subsets and computed test accuracy on the 10th held-out subset. In each model training step, we attempted to predict per-gene BFDPs from the principal components of the 145 gene-level features as described earlier, and rounded BFDPs to binary values where necessary for each model's architecture. Genes were excluded from training based on the criteria specified above. Following CV, we selected the best-fit model across all CVs based on the square root of mean squared error of predicted and actual BFDPs for the 10% of genes held out for testing. We predicted scores for all genes irrespective of training exclusion criteria. As a final step, we computed standardized scores for each gene as the transformation of predicted BFDPs based on the normal cumulative distribution function of predicted BFDPs across all 10 subsets, which scaled all scores to adhere to the range of [0, 1].

We applied this CV framework identically for eight different regression and machine learning models: a logistic classifier, a support-vector machine, a random forest, a latent discriminant analysis, a naïve Bayes classifier, a fully connected feed-forward neural net, a gradient-boosted decision tree, and a K-nearest neighbors classifier. All models were implemented in Python with the SciKitLearn library (Pedregosa et al., 2011). We optimized the hyperparameters for each model using a grid search with five-fold CV over a dictionary of selected available hyperparameters. After the eight models had been trained and applied to predict scores for all genes, we computed the receiver-operating characteristic (ROC) for each model versus their corresponding dosage sensitive and insensitive training genes and determined the ROC-optimal score cutoff as the minimized Euclidean distance between the ROC and hypothetically perfect performance (*i.e.*, 100% true positives and 0% false positives). Finally, we computed an ensemble score for each gene as the average of all eight models weighted by their accuracies corresponding to their model-specific ROC-optimal classification thresholds.

To evaluate model performance, we computed ROCs and precision-recall curves (PRCs) for all nine models, including the eight independent architectures and the final ensemble model. We used our training gene sets to define performance for both ROC and PRC analysis and derived an overall rank of architecture performance based on the harmonic mean of the area under ROCs and PRCs for both deletions and duplications. To further evaluate the specificity of our deletion- and duplication-derived scores for predicting haploinsufficiency and triplosensitivity, respectively, we defined subsets of genes from our curated dosage sensitive gene sets that were only present in one of the two sets (*e.g.*, curated haploinsufficient training genes that were not also present in the curated triplosensitive training gene set) and subsequently repeated ROC and PRC analyses as described above. After model evaluation, we designated the ensemble scores as our final metrics for use in all subsequent analyses and dubbed these metrics the "probability of haploinsufficiency" (pHaplo) and the "probability of triplosensitivity" (pTriplo) for deletions and duplications, respectively.

Secondary analyses of genic dosage sensitivity scores

Calibration of high-confidence cutoffs for dosage sensitivity scores

We determined cutoffs for pHaplo and pTriplo to isolate subsets of high-confidence dosage sensitive genes by comparing the empirical rCNV effect sizes observed in our study to a gold-standard set of genes constrained against PTVs as per gnomAD v2.1.1 (see Figure 6B; Table S5; Karczewski et al., 2020). To determine these cutoffs, we first computed the empirical log-OR of deletions of these gold-standard constrained genes in our rCNV dataset. We next divided all genes based on their pHaplo or pTriplo into 100 fixed-width bins in increments of ±0.01. For each bin, we computed the marginal log-OR of rare deletions or duplications impacting at least one gene per bin. We also computed the cumulative log-OR of all rare deletions or duplications impacting at least one gene with a pHaplo or pTriplo score at least as large as the smallest score in that bin. We assigned pHaplo and pTriplo thresholds to the lowest score such that these cumulative log-ORs were at least as strong as the log-OR we observed for deletions of gold-standard constrained genes.

Comparisons to de novo *CNVs in autism families*

We assessed the utility of our pHaplo and pTriplo scores in interpreting the disease risk contributed by *de novo* CNVs (dnCNVs) from a recent exome sequencing study of ASD (Fu et al., 2021). For these comparisons, we subset the dnCNV dataset to include the independent subset of 13,786 children affected by ASD and their 5,098 unaffected siblings that were not already present in our rCNV dataset. We converted CNV coordinates from hg38 to GRCh37 using UCSC liftOver (Lee et al., 2020) and subsequently excluded all dnCNVs meeting any of the following criteria: (1) $\geq$20Mb in size, (2) $\geq$50% overlap at the level of individual genes versus any of the known GDs curated from the literature, and (3) <5% or <100% CDS overlap of at least one gene for deletions or duplications, respectively. We next scored all dnCNVs based on the highest pHaplo or pTriplo score of any gene with a predicted gene-truncating deletion or whole-gene duplication, respectively. After scoring each dnCNV with pHaplo or pTriplo, we next stratified all dnCNVs into thirds by score per CNV class and computed the empirical OR per third of carrying a qualifying dnCNV between affected and unaffected children. We performed this comparison in thirds due to the sparsity of dnCNV data in unaffected siblings after filtering causing any finer subdivisions to lead to unstable effect size estimates.

Please cite this article in press as: Collins et al., A cross-disorder dosage sensitivity map of the human genome, Cell (2022), https://doi.org/10.1016/j.cell.2022.06.036

Score comparisons to gnomAD-SV

We compared our pHaplo and pTriplo scores to the rates of gene-disrupting CNVs in the general population as documented by large-scale genome sequencing. To accomplish this, we counted the number of gene-truncating deletions or whole-gene copy-gain duplications from the subset of 8,342 unrelated samples with no known neurological phenotypes from the publicly available gnomAD-SV v2.1 dataset (Collins et al., 2020). We next separated genes into 10 fixed-width bins across the range of pHaplo and pTriplo scores in increments of ±0.1. For each bin of genes, we computed the average number of protein-truncating deletions or copy-gain duplications. We also computed a normalized value adjusting for 21 gene-level genomic features, like gene length and CDS length (see Table S5 for details). To compute an expected number of CNVs in gnomAD-SV per gene, we first normalized and decomposed all 21 genomic features into their top 15 principal components (PCs) as described earlier. We next fit a negative binomial regression model to predict the number of protein-truncating deletions or copy-gain duplications in gnomAD based on these 15 PCs. To reduce the confounding effects of negative selection for constrained genes, we restricted to the subset of 5,578 likely unconstrained genes where we did not expect strong negative selection against coding variants (see Table S5 for details) before fitting this negative binomial regression model. We applied the fit model to predict an expected number of CNVs for all genes and computed a binwise observed:expected ratio by summing the total number of observed CNVs across all genes per pHaplo or pTriplo bin and dividing by the sum of expected CNV counts.

Supplemental figures

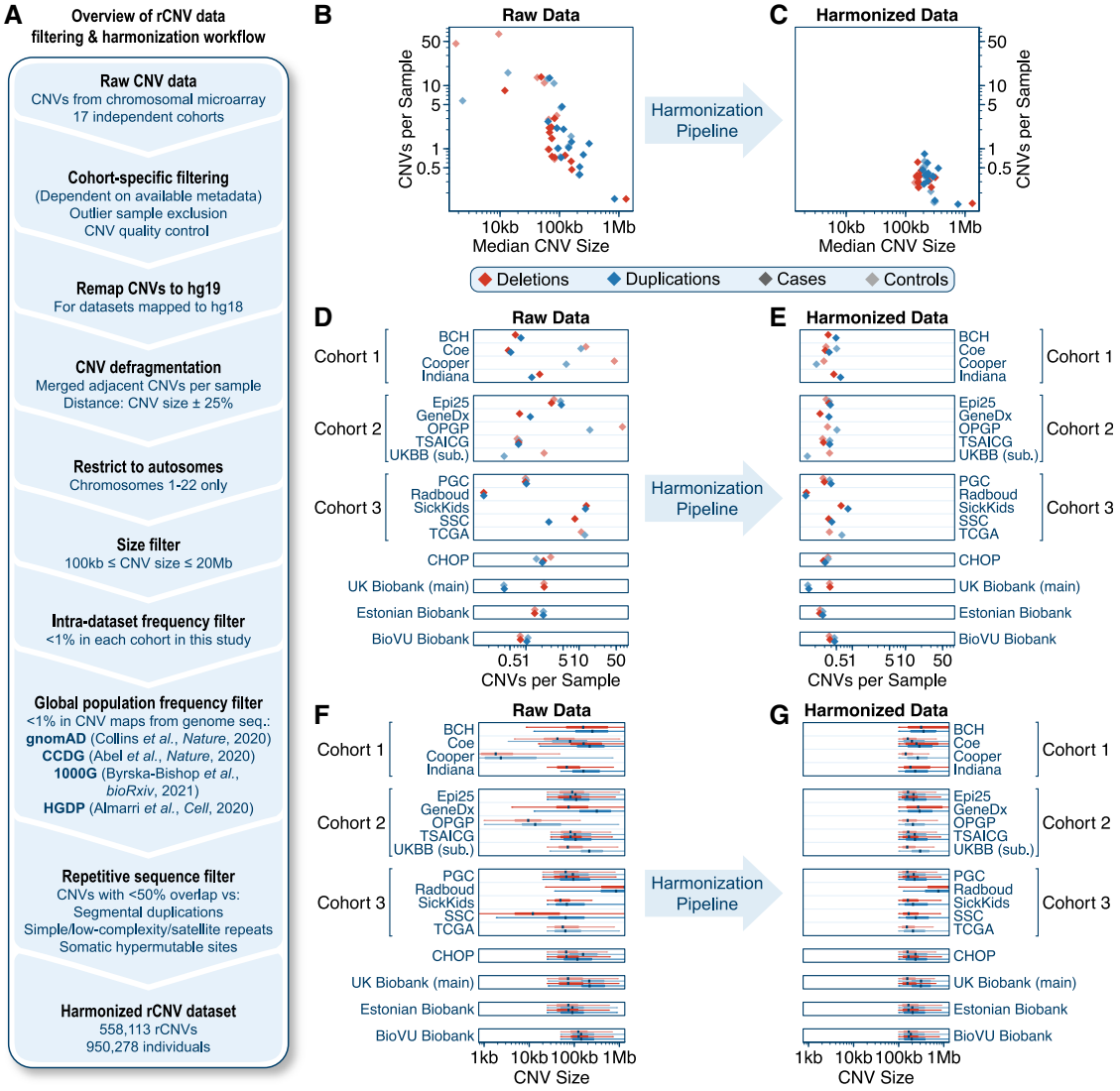


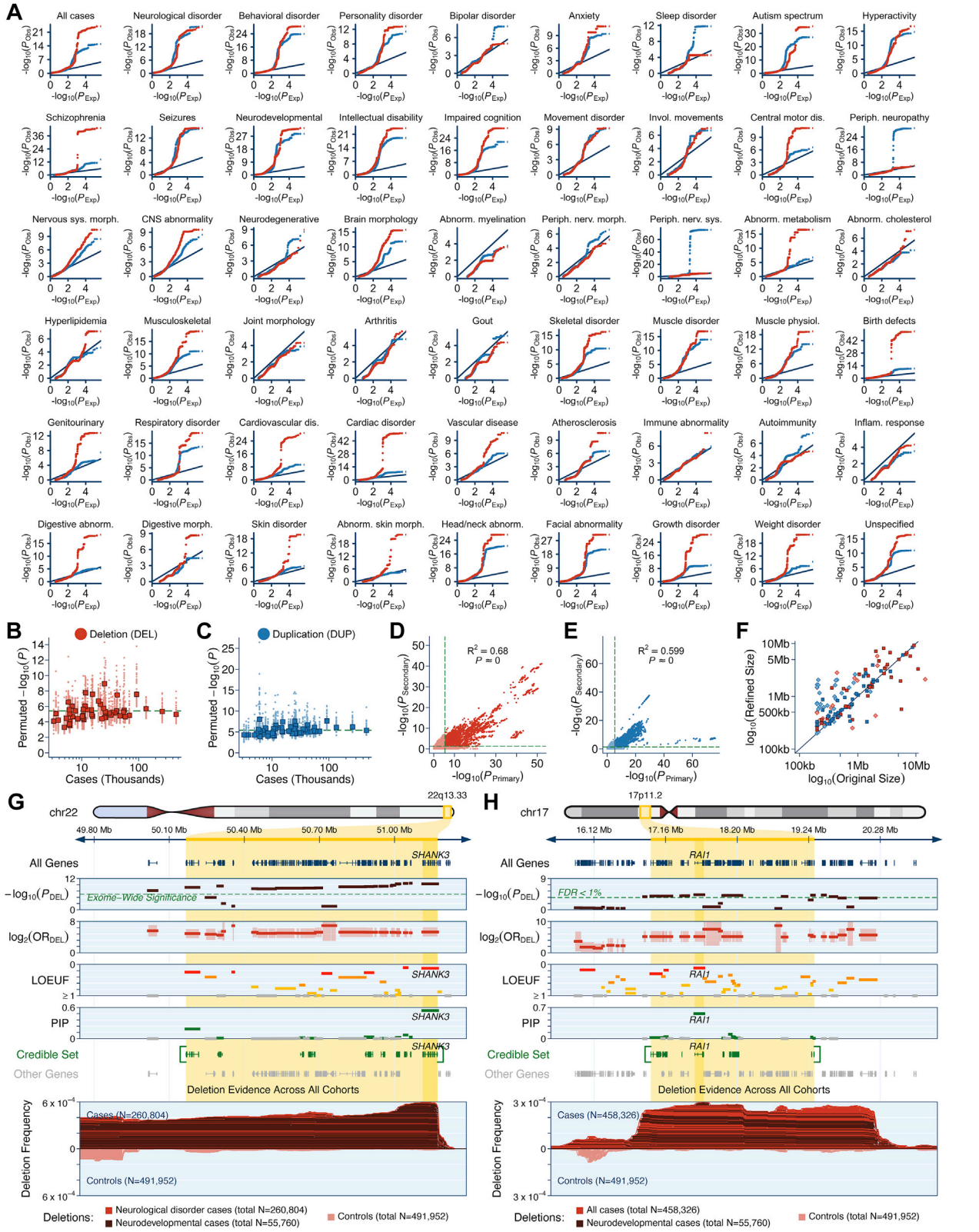
Figure S1. rCNV harmonization, related to Figure 1

We aggregated & harmonized rCNV data from 950,278 individuals across 17 sources (Table S1).

(A) Overview of rCNV harmonization pipeline.

(B, D, and F) Selected summary metrics for raw CNV data from each source prior to harmonizing. Sources are grouped by meta-analysis cohort.

(C, E, and G) Summary metrics after harmonization.



(legend on next page)

Figure S2. Technical details of rCNV association meta-analyses, related to Figures 2, 4, and 5

We conducted genome-wide rCNV association meta-analyses for 54 disease phenotypes across all autosomes using two strategies: (i) 200 kb sliding windows in 10 kb steps and (ii) collapsing gene-based tests for 17,263 autosomal protein-coding genes.

(A) Quantile-quantile plots for meta-analysis test statistics for deletions (red) and duplications (blue) per phenotype from our sliding window analyses. Solid black lines denote the expected distribution if the null hypothesis (i.e., no association) was true for all windows.

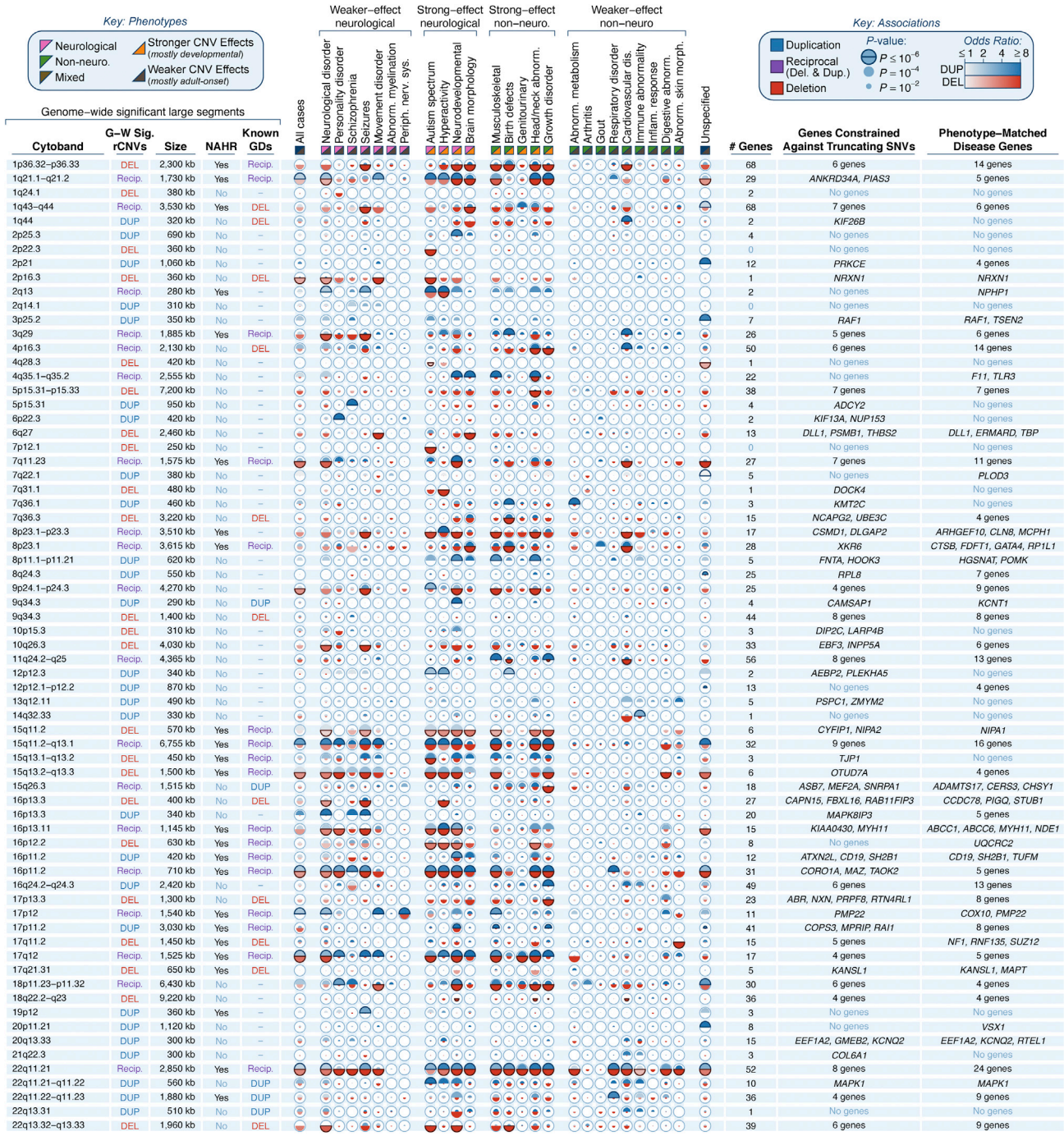
(B and C) We assessed the calibration of our approximated genome-wide significance threshold for our sliding window analyses ($p = 3.74 \times 10^{-6}$; dashed horizontal line) against empirical p value thresholds derived from 50 random phenotype permutations for deletions (B) and duplications (C) as described in the STAR Methods. Small, light points represent the empirical p value threshold corresponding to strictly zero false discoveries across all tested windows for a single permutation. Large squares are median p value thresholds per phenotype across all 50 permutations.

(D and E) Relationship of primary and secondary p values across all sliding window meta-analyses for deletions (E) and duplications (F). Secondary p values reflect the meta-analysis outcome after excluding the most significant cohort per window. Dashed lines indicate thresholds for genome-wide significance, and dark shaded points indicate windows surpassing both thresholds. R^2 and p values are provided from a Pearson correlation test.

(F) Relationship of original significant segment size versus the size of the 95% credible interval after refinement for each of the 163 disease-associated large rCNV segments discovered by our sliding window meta-analyses.

(G) We identified a significant association between rare deletions and neurodevelopmental abnormalities (OR = 32.1; 95% CI = 30.6–33.8) that included a total of 38 genes on chr22. Gene-based fine-mapping prioritized *SHANK3* (PIP = 0.54) as the most likely driver gene for this association, which matches its known role as the haploinsufficient cause of Phelan-McDermid Syndrome (Zhou et al., 2019). Meta-analysis p values and ORs are provided for the more specific (smaller N) of the two phenotypes listed at the bottom of the panel.

(H) We also identified an association between rare deletions at 17q11.2 and neurodevelopmental abnormalities (OR = 38.9; 95% CI = 36.6–41.3) involving a total of 52 genes. Fine-mapping prioritized *RAI1* (PIP = 0.56) as the predicted driver gene, which matches its known role as the dominant cause of Smith-Magenis syndrome associated with 17q11.2 deletions (Slager et al., 2003).



Cytoband	G–W Sig. rCNVs	Size	NAHR	Known GDs	# Genes	Genes Constrained Against Truncating SNVs	Phenotype–Matched Disease Genes
1p36.32–p36.33	DEL	2,300 kb	Yes	Recip.	68	6 genes	14 genes
1q21.1–q21.2	Recip.	1,730 kb	Yes	Recip.	29	*ANKRD34A, PIAS3*	5 genes
1q24.1	DEL	380 kb	No	–	2	No genes	No genes
1q43–q44	Recip.	3,530 kb	Yes	DEL	68	7 genes	6 genes
1q44	DUP	320 kb	No	DEL	2	*KIF26B*	No genes
2p25.3	DUP	690 kb	No	–	4	No genes	No genes
2p22.3	DEL	360 kb	No	–	0	No genes	No genes
2p21	DUP	1,060 kb	No	–	12	*PRKCE*	4 genes
2p16.3	DEL	360 kb	No	DEL	1	*NRXN1*	*NRXN1*
2q13	Recip.	280 kb	Yes	–	2	No genes	*NPHP1*
2q14.1	DUP	310 kb	No	–	0	No genes	No genes
3p25.2	DUP	350 kb	No	–	7	*RAF1*	*RAF1, TSEN2*
3q29	Recip.	1,885 kb	Yes	Recip.	26	5 genes	6 genes
4p16.3	Recip.	2,130 kb	No	DEL	50	6 genes	14 genes
4q28.3	DEL	420 kb	No	–	1	No genes	No genes
4q35.1–q35.2	Recip.	2,555 kb	No	–	22	No genes	*F11, TLR3*
5p15.31–p15.33	DEL	7,200 kb	No	–	38	7 genes	7 genes
5p15.31	DUP	950 kb	No	–	4	*ADCY2*	No genes
6p22.3	DUP	420 kb	No	–	2	*KIF13A, NUP153*	No genes
6q27	DEL	2,460 kb	No	–	13	*DLL1, PSMB1, THBS2*	*DLL1, ERMARD, TBP*
7p12.1	DEL	250 kb	No	–	0	No genes	No genes
7q11.23	Recip.	1,575 kb	Yes	Recip.	27	7 genes	11 genes
7q22.1	DUP	380 kb	No	–	5	No genes	*PLOD3*
7q31.1	DEL	480 kb	No	–	1	*DOCK4*	No genes
7q36.1	DUP	460 kb	No	–	3	*KMT2C*	No genes
7q36.3	DEL	3,220 kb	No	DEL	15	*NCAPG2, UBE3C*	4 genes
8p23.1–p23.3	Recip.	3,510 kb	Yes	–	17	*CSMD1, DLGAP2*	*ARHGEF10, CLN8, MCPH1*
8p23.1	Recip.	3,615 kb	Yes	Recip.	28	*XKR6*	*CTSB, FDFT1, GATA4, RP1L1*
8p11.1–p11.21	DUP	620 kb	No	–	5	*FNTA, HOOK3*	*HGSNAT, POMK*
8q24.3	DUP	550 kb	No	–	25	*RPL8*	7 genes
9p24.1–p24.3	Recip.	4,270 kb	No	–	25	4 genes	9 genes
9q34.3	DUP	290 kb	No	DUP	4	*CAMSAP1*	*KCNT1*
9q34.3	DEL	1,400 kb	No	DEL	44	8 genes	8 genes
10p15.3	DEL	310 kb	No	–	3	*DIP2C, LARP4B*	No genes
10q26.3	DEL	4,030 kb	No	–	33	*EBF3, INPP5A*	6 genes
11q24.2–q25	Recip.	4,365 kb	No	–	56	8 genes	13 genes
12p12.3	DUP	340 kb	No	–	2	*AEBP2, PLEKHA5*	No genes
12p12.1–p12.2	DUP	870 kb	No	–	13	No genes	4 genes
13q12.11	DUP	490 kb	No	–	5	*PSPC1, ZMYM2*	No genes
14q32.33	DUP	330 kb	No	–	1	No genes	No genes
15q11.2	DEL	570 kb	Yes	Recip.	6	*CYFIP1, NIPA2*	*NIPA1*
15q11.2–q13.1	Recip.	6,755 kb	Yes	Recip.	32	9 genes	16 genes
15q13.1–q13.2	DEL	450 kb	Yes	Recip.	3	*TJP1*	No genes
15q13.2–q13.3	DEL	1,500 kb	Yes	Recip.	6	*OTUD7A*	4 genes
15q26.3	Recip.	1,515 kb	No	DUP	18	*ASB7, MEF2A, SNRPA1*	*ADAMTS17, CERS3, CHSY1*
16p13.3	DEL	400 kb	No	DEL	27	*CAPN15, FBXL16, RAB11FIP3*	*CCDC78, PIGQ, STUB1*
16p13.3	DUP	340 kb	No	–	20	*MAPK8IP3*	5 genes
16p13.11	Recip.	1,145 kb	Yes	Recip.	15	*KIAA0430, MYH11*	*ABCC1, ABCC6, MYH11, NDE1*
16p12.2	DEL	630 kb	Yes	Recip.	8	No genes	*UQCRC2*
16p11.2	DUP	420 kb	Yes	Recip.	12	*ATXN2L, CD19, SH2B1*	*CD19, SH2B1, TUFM*
16p11.2	Recip.	710 kb	Yes	Recip.	31	*CORO1A, MAZ, TAOK2*	5 genes
16q24.2–q24.3	DUP	2,420 kb	No	–	49	6 genes	13 genes
17p13.3	DEL	1,300 kb	No	DEL	23	*ABR, NXN, PRPF8, RTN4RL1*	8 genes
17p12	Recip.	1,540 kb	Yes	Recip.	11	*PMP22*	*COX10, PMP22*
17p11.2	DUP	3,030 kb	Yes	Recip.	41	*COPS3, MPRIP, RAI1*	8 genes
17q11.2	DEL	1,450 kb	Yes	DEL	15	5 genes	*NF1, RNF135, SUZ12*
17q12	Recip.	1,525 kb	Yes	Recip.	17	4 genes	5 genes
17q21.31	DEL	650 kb	Yes	DEL	5	*KANSL1*	*KANSL1, MAPT*
18p11.23–p11.32	Recip.	6,430 kb	No	–	30	6 genes	4 genes
18q22.2–q23	DEL	9,220 kb	No	–	36	4 genes	4 genes
19p12	DUP	360 kb	Yes	–	3	No genes	No genes
20p11.21	DUP	1,120 kb	No	–	8	No genes	*VSX1*
20q13.33	DUP	300 kb	No	–	15	*EEF1A2, GMEB2, KCNQ2*	*EEF1A2, KCNQ2, RTEL1*
21q22.3	DUP	300 kb	No	–	3	*COL6A1*	No genes
22q11.21	Recip.	2,850 kb	Yes	Recip.	52	8 genes	24 genes
22q11.21–q11.22	DUP	560 kb	No	DUP	10	*MAPK1*	*MAPK1*
22q11.22–q11.23	DUP	1,880 kb	Yes	DUP	36	4 genes	9 genes
22q13.31	DUP	510 kb	No	DUP	1	No genes	No genes
22q13.32–q13.33	DEL	1,960 kb	No	DEL	39	6 genes	9 genes

Figure S3. Summary of all genome-wide significant rCNV segments, related to Figure 2

We identified 88 unique rCNV segments associated with at least one phenotype at genome-wide significance ($p \leq 3.74 \times 10^{-6}$). Details for each of the 88 segments are summarized here and are provided in Table S3. Overlapping segments have been merged into single rows for clarity, such as for reciprocal genomic disorders (GDs) where both deletions and duplications were associated with at least one phenotype. For each locus, we provide basic information including meta-analysis association statistics for a representative subset of 27 phenotypes, and genic content partitioned by PTV constraint (Karczewski et al., 2020) and previously reported associations with the same phenotypes (Amberger et al., 2015). Association statistics are represented as one semicircle each for duplication (blue) and deletion (red) shaded by the effect size estimate with radii scaled proportionally to the $-\log_{10}$ p value. Sample sizes vary per phenotype; see Table S2 for details.

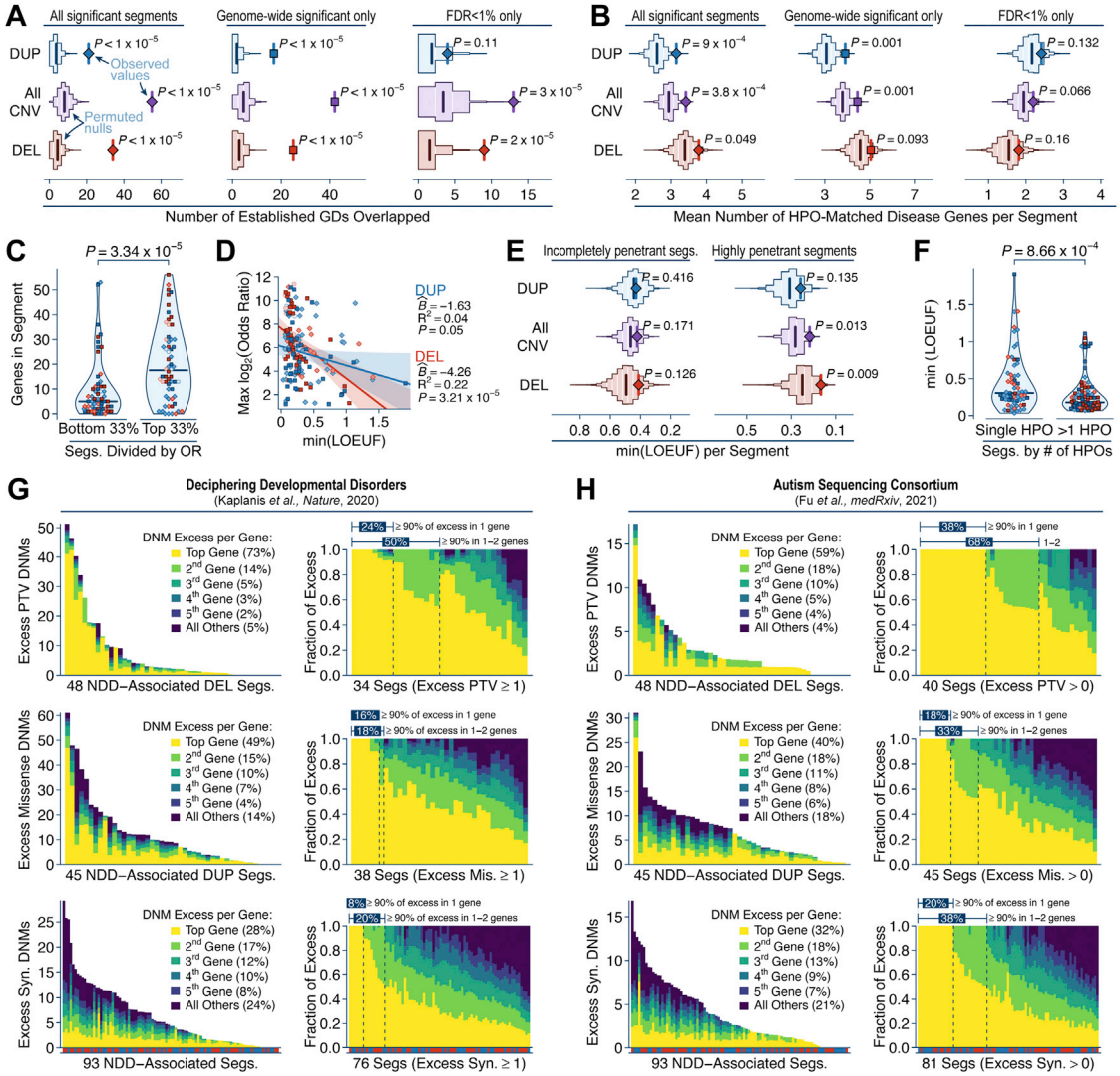


Figure S4. Properties of disease-associated rCNV segments, related to Figures 2 and 3

(A) The disease-associated rCNV segments from our discovery meta-analyses overlapped a set of 95 GDs reported in the literature 6.9-fold more often than expected by chance based on 100,000 random sets of size-matched segments (one-tailed permutation tests).

(B) Significant rCNV segments from our discovery meta-analyses overlapped more disease genes previously implicated in the same phenotypes (Amberger et al., 2015) than expected by chance based on 100,000 random permutations matched on number of genes per segment (one-tailed permutation tests).

(C) Segments in the top third of all effect sizes (mean odds ratio [OR] = 444.0; "highly penetrant segments") overlapped significantly more genes than segments in the bottom third of effect sizes (mean OR = 10.3; "incompletely penetrant segments") (two-tailed Wilcoxon test). For (C)–(F), dark squares indicate genome-wide significance; medium diamonds indicate FDR < 0.01; and light circles indicate GDs curated from the literature that did not reach FDR < 0.01 in our meta-analyses (see Figure 2B).

(D) Segment effect size was correlated with the magnitude of constraint (i.e., LOEUF) of the most constrained gene within each segment (i.e., lowest LOEUF value among all genes per segment) (Karczewski et al., 2020). Lines represent outlier-robust linear regression, and shaded areas are 95% confidence intervals.

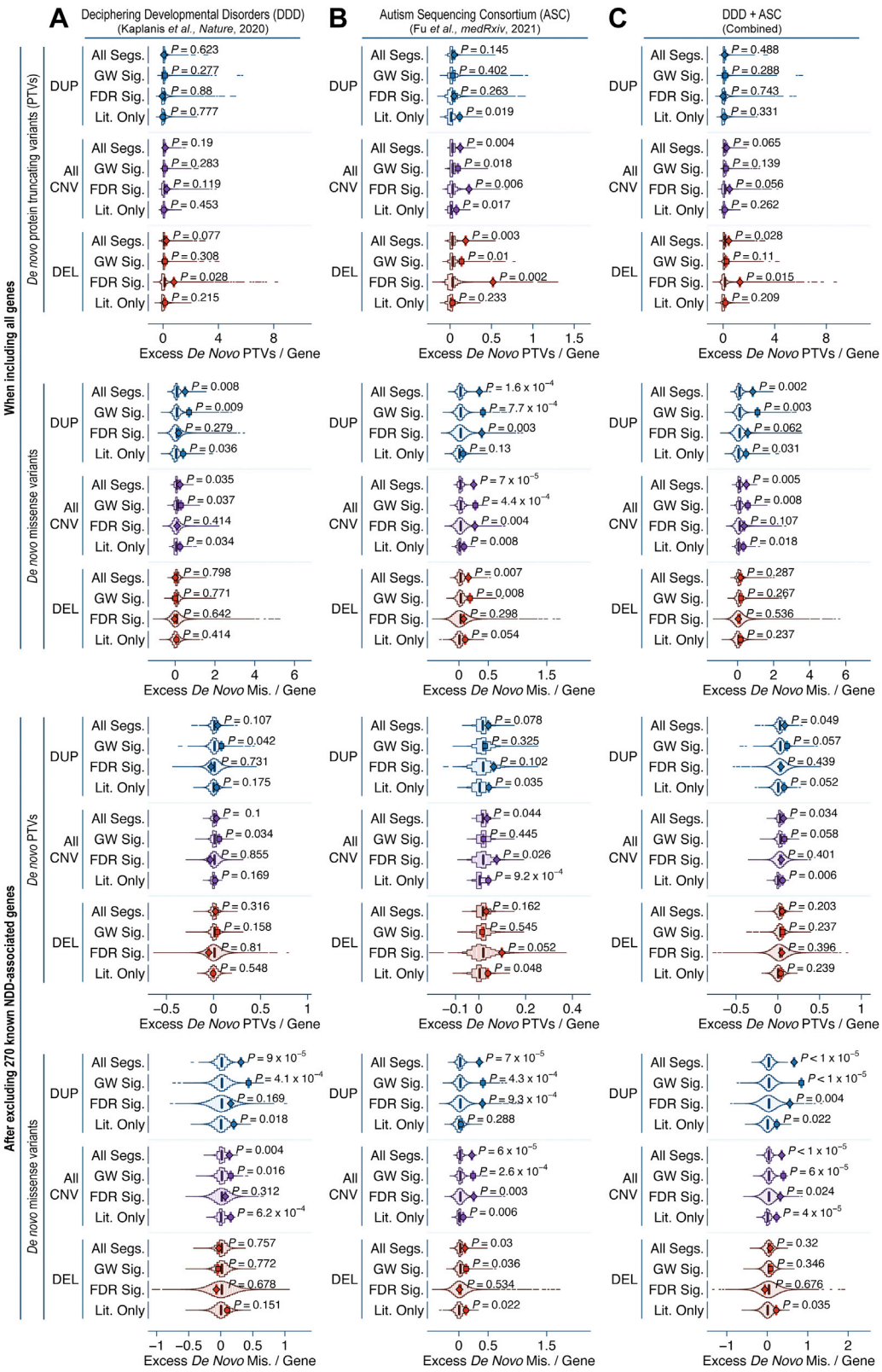
(E) Highly penetrant segments overlapped genes under stronger mutational constraint than expected by chance based on permutations matched on the number of genes per segment, whereas we did not observe the same trend for incompletely penetrant segments (one-tailed permutation tests).

(F) Segments associated with multiple phenotypes in our meta-analyses overlapped genes under greater mutational constraint as compared to segments associated with one phenotype (two-tailed Wilcoxon test).

(G and H) We evaluated the distributions of DNMs from exome-sequencing studies of (G) developmental disorders (Kaplanis et al., 2020) and (H) ASD (Fu et al., 2021) to better understand the genetic architecture of the 93 large rCNV segments (48 deletion & 45 duplication) from our consensus set that overlapped at least one gene and were also associated with at least one strong-effect neurological phenotype (see Figure 1A). In each panel, the left-most columns provide the distributions of total excess DNMs per segment. Genes within each segment were ranked based on their mutation rate-adjusted (Karczewski et al., 2020) excess of DNMs per consequence (protein-truncating variants [PTVs], missense [Mis.], and synonymous [Syn.]) before being tallied in order. Red and blue tabs beneath

(legend continued on next page)

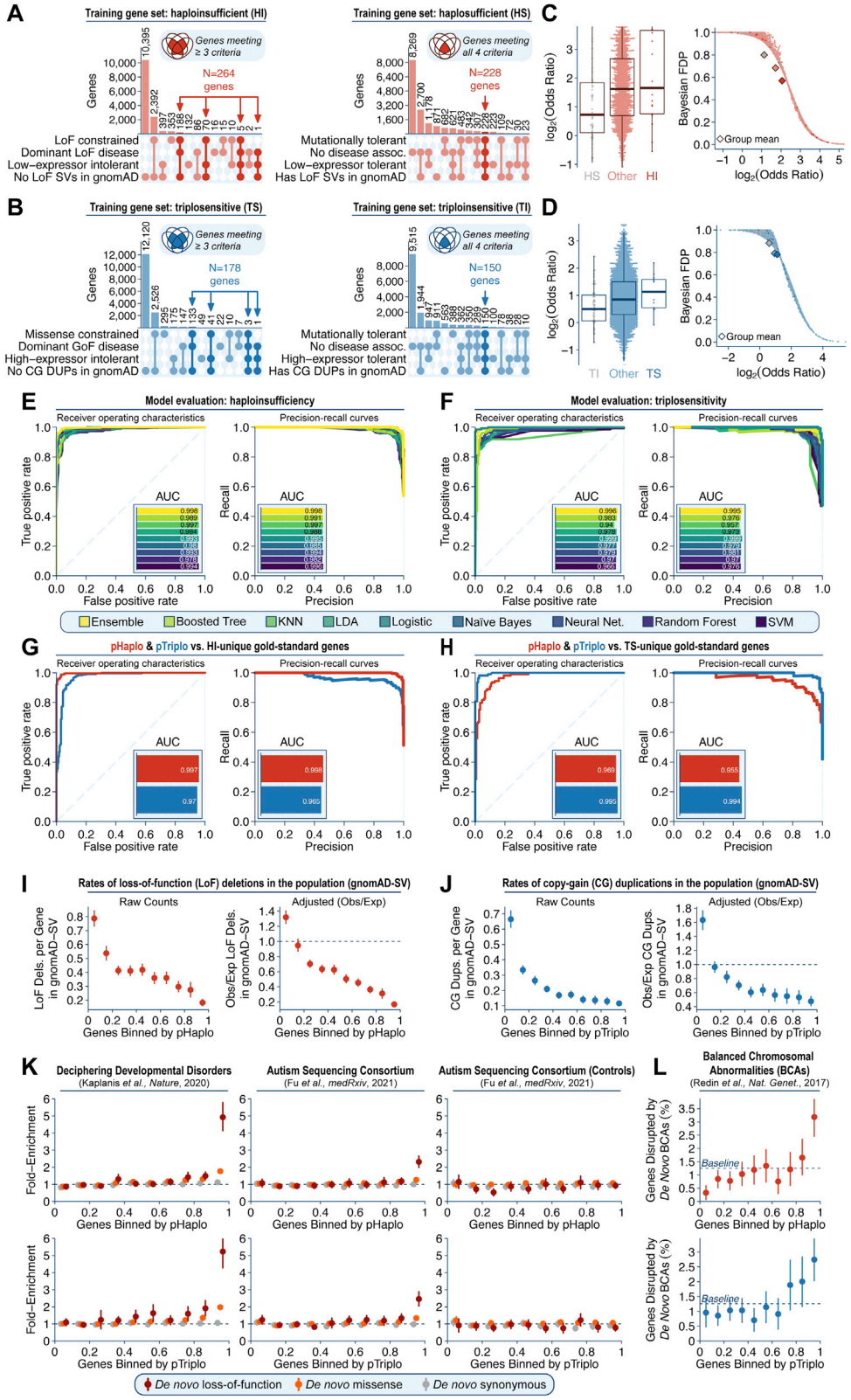
each segment in the synonymous DNM panels (bottom row) indicate deletion or duplication segments, respectively. The right-most columns in each panel present the cumulative distributions of excess DNMs per segment; i.e., data from the first or third columns divided by the total number of *excess* DNMs per segment for all segments meeting the minimum DNM excess criteria listed on the x axis. Vertical dashed lines indicate what fraction of segments have at least 90% of their total DNM excess concentrated in the top one or two genes, respectively.



(legend on next page)

Figure S5. Permutation tests of DNMs in NDD-associated rCNV segments, related to Figure 3

The genes within NDD-associated rCNV segments were significantly enriched for excesses of damaging *de novo* mutations (DNMs) identified in exome-sequencing studies of (A) 31,058 individuals with developmental disorders (Kaplanis et al., 2020), (B) 15,036 individuals with ASD (Fu et al., 2021), or (C) a combined analysis of both datasets (total n = 46,094 individuals with broadly defined neurodevelopmental disorders [NDDs]). Each panel provides the results from 100,000-fold one-tailed permutation tests matched on the number of genes per segment for PTVs and missense variants. Counts of DNMs were residualized against gene-specific mutation rates for each mutational consequence (Karczewski et al., 2020), and subsequently normalized against the rate of synonymous DNMs to control for any mutation rate miscalibration. All analyses presented here were restricted to the set of 93 rCNV segments that overlapped at least one gene and were also associated with at least one strong-effect neurological phenotype (see Figure 1A). The top two rows of panels present the results when considering all genes, whereas the bottom two rows of panels present the same analysis after excluding all 270 genes that were significantly associated with either developmental disorders and/or ASD at exome-wide significance in either of the two exome-sequencing studies (Fu et al., 2021; Kaplanis et al., 2020).



(legend on next page)

Figure S6. Dosage-sensitivity model development and quality assessment, related to Figure 6

(A and B) We defined training sets of genes that were likely (A) haploinsufficient (HI) & haplosufficient (HS) and (B) triplosensitive (TS) & triploinsensitive (TI) as all genes meeting at least three (for TS & HI) or four (for TI & HS) of the four criteria listed per category.

(C and D) We computed case:control ORs using the same meta-analysis approach as for our gene-based association analyses, with minor modifications for optimized parameters (see STAR Methods). We subsequently transformed these ORs into Bayesian false discovery probabilities (BFDPs), which were used to train eight machine-learning architectures to predict probabilities of dosage sensitivity for every gene.

(E and F) For our final scores, we used an ensemble average (yellow) across all eight machine-learning models (other colors), which outperformed any individual model alone.

(G and H) We assessed the specificity of our final scores, pHaplo and pTriplo, against the non-overlapping subsets of our curated likely dosage-sensitive training genes that were (G) only HI or (H) only TS.

(I) Left: average counts of loss-of-function (LoF) deletions in gnomAD-SV v2.1 (Collins et al., 2020) for all genes binned into deciles by pHaplo. Right: same data from left panel after adjustment for per-gene expected number of LoF deletions using a negative binomial regression as previously described (Collins et al., 2020).

(J) Comparisons of whole-gene copy-gain duplications from gnomAD-SV v2.1 versus pTriplo, presented identically to (G).

(K) We evaluated the relationship of pHaplo and pTriplo versus enrichments of coding DNMs from two exome-sequencing studies of developmental disorders (Kaplanis et al., 2020) and ASD (Fu et al., 2021). Genes were assigned to ten fixed-width bins based on pHaplo or pTriplo score, and the ratio of total observed DNMs to expected DNMs was computed per gene while adjusting for gene-specific mutation rates (Karczewski et al., 2020). We observed enrichments of PTV and missense DNMs in affected individuals but not in their unaffected siblings (right-most panels; labeled "controls").

(L) Genes with higher pHaplo and pTriplo scores were more likely to have been disrupted by *de novo* balanced chromosomal abnormalities among 273 individuals with congenital anomalies (Redin et al., 2017). In all panels, error bars correspond to 95% confidence intervals based on 1,000-fold bootstrapping.

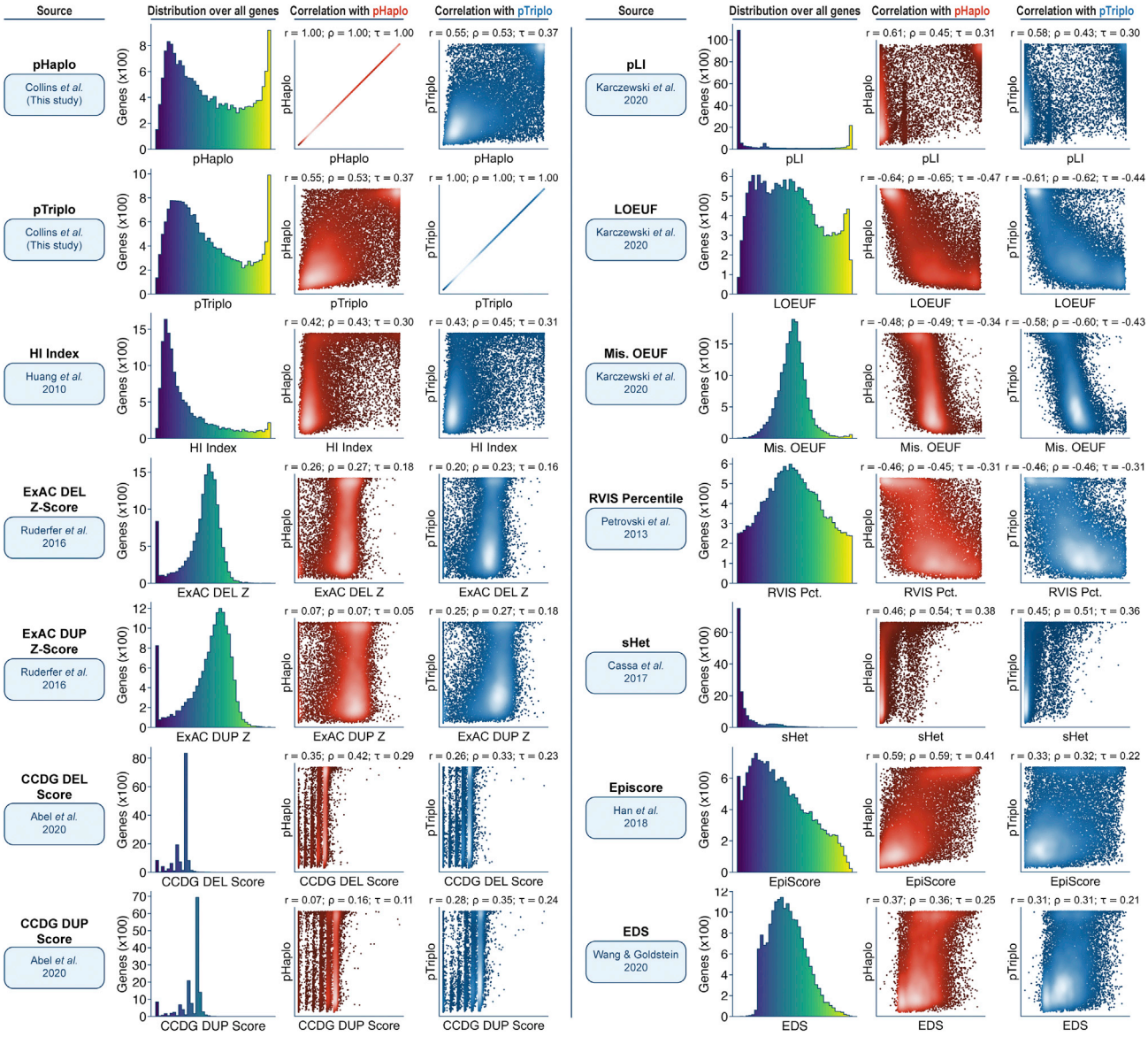


Figure S7. Comparison of genic dosage-sensitivity scores to existing metrics, related to Figures 6 and 7

Comparisons of pHaplo and pTriplo against five existing gene-level constraint metrics derived from analyses of short variants (pLI, LOEUF, missense OEUF [Mis. OEUF], RVIS Percentile, sHet) (Cassa et al., 2017; Karczewski et al., 2020; Petrovski et al., 2013), five metrics derived from CNVs (HI index, ExAC DEL *Z* score, ExAC DUP *Z* score, CCDG DEL score, CCDG DUP score) (Abel et al., 2020; Huang et al., 2010; Ruderfer et al., 2016), and two metrics produced by other methods (Episcore, EDS) (Han et al., 2018; Wang and Goldstein, 2020). For each metric, we provide the distribution of scores for the 18,641 autosomal protein-coding genes we considered when developing pHaplo and pTriplo. We also provide scatterplot comparisons of each metric against pHaplo (red) and pTriplo (blue). Pearson's r, Spearman's rho, and Kendall's tau are provided as measures of correlation for each comparison.