gnomAD-SV

An open resource of structural variation for medical and population genetics

Ryan L. Collins[1-3,*], Harrison Brand[1,2,4,*], Konrad J. Karczewski[1,2], Xuefang Zhao[1,2,4], Jessica Alföldi[1,2], Amit V. Khera[1,2], Laurent C. Francioli[1,2,5], Laura D. Gauthier[1,6], Harold Wang[1,2], Nicholas A. Watts[1,2], Matthew Solomonson[1,2], Anne O'Donnell-Luria[1,2], Alexander Baumann[6], Ruchi Munshi[6], Chelsea Lowther[1,2,4], Mark Walker[1,6], Christopher Whelan[6,10], Yongqing Huang[6], Ted Brookings[6], Ted Sharpe[6], Matthew R. Stone[1,2], Elise Valkanas[1-3], Jack Fu[1,2,4], Grace Tiao[1,2], Kristen M. Laricchia[1,2], Christine Stevens[1], Namrata Gupta[1], Lauren Margolin[1], The Genome Aggregation Database (gnomAD) Production Team[7], The gnomAD Consortium[7], John A. Spertus[8], Kent D. Taylor[9,10], Henry J. Lin[10,11], Stephen S. Rich[12], Wendy Post[13], Yii-Der Ida Chen[9,10], Jerome I. Rotter[9,10], Chad Nusbaum[1,†], Anthony Philippakis[6], Eric Lander[1,14,15], Stacey Gabriel[1], Benjamin M. Neale[1-3,16], Sekar Kathiresan[1,2,5,17], Mark J. Daly[1-3,16], Eric Banks[6], Daniel G. MacArthur[1-3,5], Michael E. Talkowski[1-4,16]

* These authors contributed equally to this work

1. Program in Medical and Population Genetics, Broad Institute of Harvard and Massachusetts Institute of Technology (M.I.T.), Cambridge, MA; 2. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA; 3. Division of Medical Sciences, Harvard Medical School, Boston, MA; 4. Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA; 5. Department of Medicine, Harvard Medical School, Boston, MA; 6. Data Science Platform, Broad Institute of Harvard and M.I.T., Cambridge, MA; 7. Group authors are enumerated at the end of this document; 8. Department of Biomedical and Health Informatics, Saint Luke's Mid America Heart Institute, University of Missouri, Kansas City, MO; 9. Institute for Translational Genomics and Population Sciences, LABioMed, Torrance, CA; 10. Department of Pediatrics at Harbor-UCLA Medical Center, Torrance, CA; 11. Institute for Translational Genomics and Population Sciences, LABioMed at Harbor-UCLA Medical Center, Torrance, CA, USA; 12. Center for Public Health Genomics, University of Virginia School of Medicine, Richmond, VA; 13. Johns Hopkins University School of Medicine, Baltimore, MD; 14. Department of Systems Biology, Harvard Medical School, Boston, MA; 15. Division of Health Sciences and Technology, M.I.T., Cambridge, MA; 16. Stanley Center for Psychiatric Research, Broad Institute of Harvard and M.I.T., Cambridge, MA; 17. Division of Cardiology, Massachusetts General Hospital, Boston, MA; † Current address: Cellarity Inc., Cambridge MA

SUMMARY

Structural variants (SVs) rearrange the linear and three-dimensional organization of the genome, which can have profound consequences in evolution, diversity, and disease. As national biobanks, human disease association studies, and clinical genetic testing are increasingly reliant on whole-genome sequencing, population references for small variants (*i.e.*, SNVs & indels) in protein-coding genes, such as the Genome Aggregation Database (gnomAD), have become integral for the evaluation and interpretation of genomic variation. However, no comparable large-scale reference maps for SVs exist to date. Here, we constructed a reference atlas of SVs from deep whole-genome sequencing (WGS) of 14,891 individuals across diverse global populations (54% non-European) as a component of gnomAD. We discovered a rich landscape of 498,257 unique SVs, including 5,729 multi-breakpoint complex SVs across 13 mutational subclasses, and examples of localized chromosome shattering, like chromothripsis, in the general population. The mutation rates and densities of SVs were non-uniform across chromosomes and SV classes. We discovered strong correlations between constraint against predicted loss-of-function (pLoF) SNVs and rare SVs that both disrupt and duplicate protein-coding genes, suggesting that existing per-gene metrics of pLoF SNV constraint do not simply reflect haploinsufficiency, but appear to capture a gene's general sensitivity to dosage alterations. The average genome in gnomAD-SV harbored 8,202 SVs, and approximately eight genes altered by rare SVs. When incorporating these data with pLoF SNVs, we estimate that SVs comprise at least 25% of all rare pLoF events per genome. We observed large (≥1Mb), rare SVs in 3.1% of genomes (~1:32 individuals), and a clinically reportable pathogenic incidental finding from SVs in 0.24% of genomes (~1:417 individuals). We also estimated the prevalence of previously reported pathogenic recurrent CNVs associated with genomic disorders, which highlighted differences in frequencies across populations and confirmed that WGS-based analyses can readily recapitulate these clinically important variants. In total, gnomAD-SV includes at least one CNV covering 57% of the genome, while the remaining 43% is significantly enriched for CNVs found in tumors and individuals with developmental disorders. However, current sample sizes remain markedly underpowered to establish estimates of SV constraint on the level of individual genes or noncoding loci. The gnomAD-SV resources have been integrated into the gnomAD browser (https://gnomad.broadinstitute.org), where users can freely explore this dataset without restrictions on reuse, which will have broad utility in population genetics, disease association, and diagnostic screening.

INTRODUCTION

Structural variants (SVs) are genomic rearrangements that alter segments of DNA ≥50 bp. By virtue of their size and abundance,[1] SVs represent an important mutational force shaping genome evolution and function,[2,3] and a significant contributor to germline and somatic disease.[4-6] The profound impact of SVs is partially attributable to the varied mechanisms by which intra- and inter-chromosomal rearrangements can alter linear and three-dimensional genome structure, which can disrupt protein-coding sequences and/or *cis*-regulatory architecture.[5,7-9] Genomic rearrangements can be grouped into distinct mutational classes, including "unbalanced" SVs associated with gains or losses of DNA (*e.g.*, copy-number variants [CNVs]), and "balanced" SVs that rearrange genomic segments without corresponding dosage alterations (*e.g.,* inversions & translocations) (**Figure 1a**).[10] Other common forms of SVs include mobile elements that insert themselves throughout the genome,[11] and multiallelic CNVs (MCNVs) that exist at high copy states.[12] Beyond these canonical classes, more exotic species of complex SVs exist in all individuals.[13,14] These variants do not conform to a single canonical class, and instead involve two or more SV signatures from a single mutational event interleaved within the same allele. Complex SVs can range from CNV-flanked inversions (*e.g.*, dupINVdup) to rare instances of localized chromosome shattering, such as chromothripsis.[8,15] The variant spectrum of germline SVs in all humans is therefore diverse, as is their influence on genome structure and function.

While SVs alter more nucleotides per genome than single nucleotide variants (SNVs) and small insertion/deletion variants (indels; <50 bp),[1] surprisingly little is known about their mutational spectra, patterns of natural selection, and functional impact on a global scale. The paucity of population-scale characterization of SVs is primarily attributable to the technical challenges of their ascertainment and the limited availability of whole-genome sequencing (WGS) datasets. Whereas gold-standard methods for profiling SNVs and indels are well-established, such with as the Genome Analysis Toolkit (GATK),[16] the uniform detection of SVs from short-read WGS has presented a much greater challenge. Analyses of SVs require specialized computational methods that simultaneously consider multiple SV signatures, and even high-coverage short-read WGS fails to capture a significant component of the variant spectrum accessible to more expensive niche data types such as long-read WGS, optical mapping, or strand-specific sequencing.[17] Current population references of SVs from WGS are thus restricted to the 1000

Genomes Project (N=2,504; 4-8X sequence coverage) or smaller European-centric cohorts.[1,18] This stands in stark contrast to references for coding SNVs from resources such as the Exome Aggregation Consortium (ExAC),[19] and its second iteration, the Genome Aggregation Database (gnomAD), which have jointly analyzed data from >140,000 individuals.[20] These references have transformed most aspects of medical and population genetics research, including the definition of genes constrained against predicted loss-of-function (pLoF) variation,[19,21] and have become integral in the clinical interpretation of small coding variants.[22] Therefore, as short-read WGS becomes the prevailing platform for large-scale human disease studies, and is likely to eventually displace conventional technologies in diagnostic screening, there is a critical need for similar resources of SVs across diverse global populations.

In this study, we developed gnomAD-SV, a reference atlas of SVs from deep WGS in ~15,000 samples aggregated as part of gnomAD. Our analyses reveal diverse mutational patterns among SVs, and principles of strong selection acting against reciprocal dosage changes. We also find that SVs contribute approximately 25% of all rare pLoF events currently accessible to short-read WGS in each genome, and that 0.24% of individuals in the general population harbor a clinically reportable, likely pathogenic incidental finding from SVs. These reference maps have been directly incorporated into the gnomAD browser (http://gnomad.broadinstitute.org), which can be mined for new insights into genome biology and will provide an openly accessible resource for interpretation of SVs in diagnostic screening.

RESULTS

SV discovery & genotyping

We analyzed 14,891 samples in gnomAD-SV, of which 14,216 (95.5%) passed all data quality thresholds (**Supplementary Tables 1** and **Supplementary Figure 1**). Samples were aggregated across population genetic and complex disease association studies, and the samples in gnomAD-SV represent a subset of the overall gnomAD project (see **Supplementary Table 2**).[20] All samples were previously aligned to the GRCh37/hg19 human reference assembly. This gnomAD-SV reference included 45.6% European (N=6,484), 34.7% African/African-American (N=4,937), 9.2% East Asian (N=1,304), and 7.8% Latino (N=1,109) samples, as well as 2.7% samples from admixed or other populations (N=382; **Figure 1b**). We discovered and genotyped SVs using a cloud-based version of a multi-algorithm pipeline for Illumina short-read WGS, which has been previously described in a disease association study of autism spectrum disorder (ASD) in 519 quartet families, where molecular assays yielded a 97% validation rate for predicted *de novo* SVs (**Supplementary Figure 2**).[23] In brief, this pipeline integrates four orthogonal signatures of SVs to delineate variants across the size and allele frequency (AF) spectrum accessible to short-read WGS, including six classes of canonical SVs (**Figure 1a**; deletions, duplications, MCNVs, inversions, insertions, translocations) and 13 subclasses of complex SVs (**Figure 2**).[14] We augmented these methods with approaches to account for the technical heterogeneity of aggregated WGS datasets (**Extended Data Figure 1** and **Supplementary Figures 3-4**). In total, these methods discovered 498,257 distinct SVs (**Figure 1c** and **Supplementary Table 3**). Following family-based analyses from 966 parent-child trios included for quality assessment (*e.g.*, *de novo* rates), we pruned all first-degree relatives from further analyses, retaining a total of 12,549 unrelated genomes. Analyses of SVs from short-read WGS also produces thousands of incompletely resolved non-reference breakpoint junctions per genome, sometimes referred to as 'breakends' (BNDs; **Figure 1a**), which can be valuable to document as deviations from reference sequence, but lack interpretable alternate allele structures for biological annotation. Given that these BNDs substantially inflated our variant counts (16.3% of all SVs detected), were enriched in false positives (**Extended Data Figure 2a**),[23] and cannot be interpreted for functional impact, we removed them from our final dataset. All analyses were thus performed on 382,460 unique, completely resolved SVs from 12,549 unrelated genomes (**Supplementary Table 3**).

While the number of SVs per genome in gnomAD-SV using the integration of multiple algorithms (n=8,202) is a marked increase from publicly accessible references from short-read WGS, such as the 1000 Genomes Project (3,441 SVs per genome from ~7X coverage WGS) and the GTEx project (3,658 SVs per genome from ~50X coverage WGS),[1,24] it is far lower than estimates of the total SVs per genome from recent long-read WGS analyses (24,825 per genome from 40X long-read coverage).[17] In the absence of gold-standard SV benchmarking methods, we evaluated the technical qualities of the gnomAD-SV callset using five orthogonal approaches summarized in **Extended Data Figure 2**, **Supplementary Figures 5-7**, and **Supplementary Table 4**. Briefly, we assessed Mendelian inheritance in 966 parent-child trios (2,898 genomes). Almost all SVs that violate Mendelian transmission patterns represent algorithmic false positives or false negatives in the child and/or parents, and thus provide a proxy for the performance of SV detection and genotyping accuracy. Here, we observed an average Mendelian violation rate of 4.2% per trio (**Extended Data Figure 2a**). We found 97.8% sensitivity to detect large CNVs (>40 kb) previously reported from microarrays in 1,893 individuals.[25] As another proxy for genotyping accuracy, we calculated that 87% of SVs across all populations were in Hardy-Weinberg Equilibrium, although this is an imperfect metric given the potential confounding assumptions and population genetic forces that may not hold true for all SV sites (**Extended Data Figure 2b**). We also leveraged Pacific Biosciences long-read WGS[17] in four individuals and found long-read support for up to 88.1% of SVs predicted from short-read WGS. The AFs from gnomAD-SV were correlated with variants also observed in the 1000

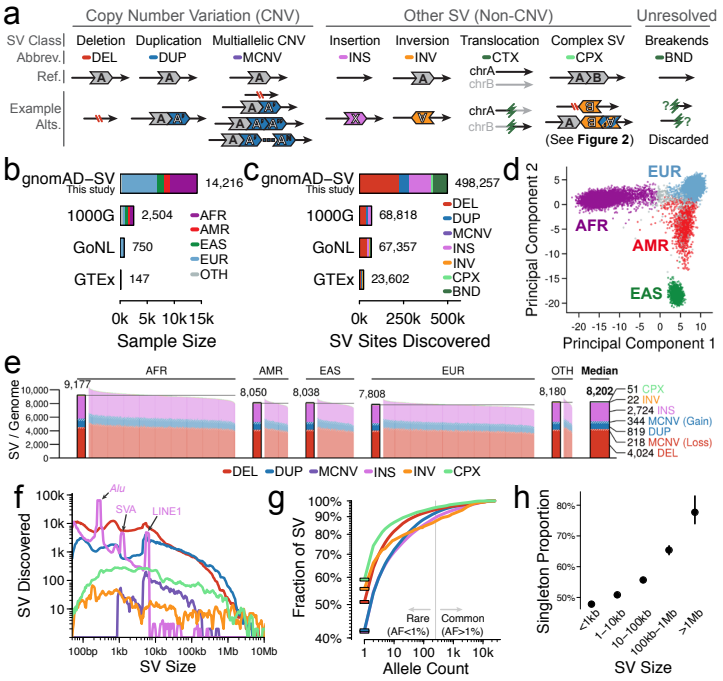


Figure 1 | Properties of SVs across human populations

*(**a**) SVs were catalogued across seven mutational classes. Complex SVs were further categorized into 13 subclasses (see **Figure 2**). We also catalogued unresolved non-reference sequence junctions, or breakends (BNDs), but they were excluded from all analyses. (**b**) After sample quality control, we processed 14,216 samples from four major continental populations: African (AFR), Latino (AMR), East Asian (EAS), and European (EUR). A small subset of samples came from admixed or other populations (OTH). Three publicly available WGS-based datasets of SVs are included for comparison (1000 Genomes Project [1000G]; Genome of the Netherlands Project [GoNL]; Genotype-Tissue Expression Project [GTEx]).[1,18,24] (**c**) We discovered 498,257 SVs (also see **Supplementary Table 3**), and provide counts from prior studies for comparison.[1,18,24] (**d**) A principal component analysis of SV genotypes separated samples along axes corresponding to genetic ancestry. (**e**) The median genome harbored 8,202 SVs (also see **Extended Data Figure 3**). (**f**) Most SVs were small. Expected insertion peaks are marked at ~300bp, ~2.1kb, and ~6kb, corresponding to three classes of mobile element insertion (Alu, SINE-VNTR-Alu [SVA], and LINE1). (**g**) Most SVs were rare (AF<1%), and 46% of SVs were singletons (solid bars). (**h**) AFs were inversely correlated with SV size (also see **Extended Data Figure 4**).*

Genomes Project (R^2=0.67; **Extended Data Figure 2c** and **Supplementary Figures 6-7**),[1] though 87% of SVs in gnomAD-SV were novel compared to the 1000 Genomes Project, reflecting the increase in scale and sensitivity of the current dataset.

Insights into population genetics & genome biology

The properties of SVs in the gnomAD-SV dataset followed expectations from human demographic history,[26] with the top two principal components projecting samples onto well-established axes according to population structure (**Figure 1d**). African/African-American samples exhibited the greatest genetic diversity (median 9,177 SVs per genome compared to 7,888 per non-African genome) (**Figure 1e**), and East Asian genomes featured the highest levels of homozygosity (median 1,582 homozygous SVs per East Asian genome compared to 1,475 per non-East Asian genome) (**Extended Data Figure 3a-d**). Most SVs were small (median SV size=374 bp; **Figure 1f**) and rare (AF<1%; 92% of SVs; **Figure 1g**). Nearly half of all SVs (46.4%) were singletons (*i.e.*, only one allele observed across all samples), and the singleton proportion varied by SV class and was strongly dependent on SV size (**Figure 1h** and **Extended Data Figure 4**). We completely resolved 5,729 complex SVs across 13 mutational subclasses, of which 4,341 (75.8%) involved inverted segments (**Figure 2**), confirming prior predictions that most inversion variation accessible to short-read WGS is comprised of complex SVs rather than canonical inversions.[1,27] Among canonical SVs, deletions were collectively more rare than other classes ($P < 1x10^{-100}$; one-sided Wilcoxon Test; **Supplementary Figure 8**). However, complex SVs were rarer than all canonical classes, including deletions ($P < 1x10^{-100}$; one-sided Wilcoxon Test), suggesting that purifying selection on SVs is likely strongest against loss of genomic content and extensive structural rearrangement.

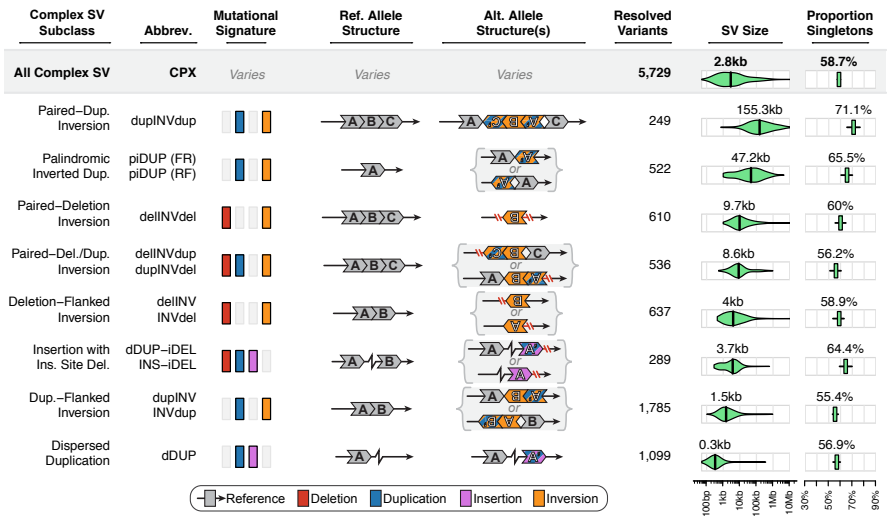


Figure 2 | Complex SVs are abundant in the human genome

*We discovered and fully resolved 5,729 complex SVs across 13 distinct mutational subclasses, 75.8% of which involved at least one inversion. Each subclass is detailed here, including their mutational signatures, non-reference allele structures, abundance, sizes, and allele frequencies. For clarity, five pairs of subclasses have been collapsed into single rows due to mirrored or highly similar alternate allele structures (e.g., delINV vs INVdel). Two highly complex SVs that did not conform to any subclass are not included in this table (see **Extended Data Figure 8**).*

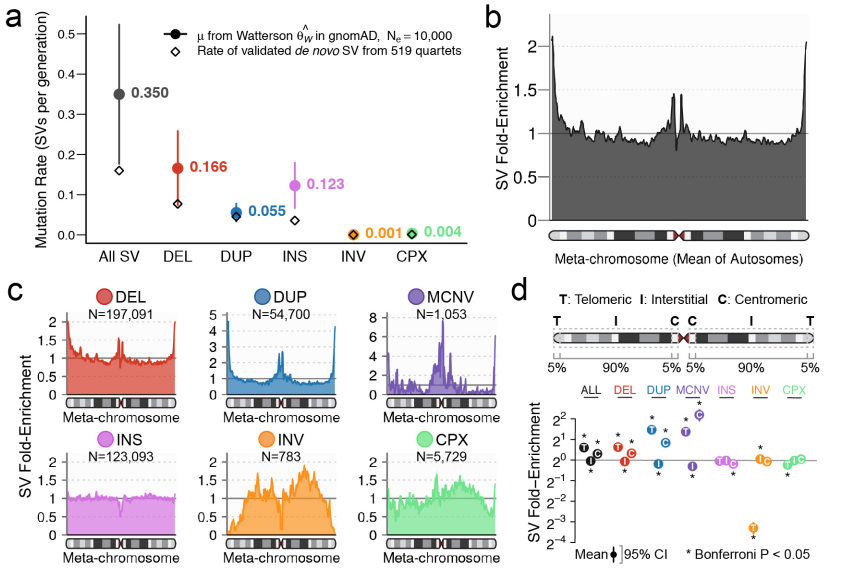


Figure 3 | Genome-wide mutational patterns of SVs

*(**a**) We estimated the mutation rate (μ) for each SV class using the Watterston estimator,[28] projecting an average of 0.35 new SVs per generation. Bars represent 95% confidence intervals, and we provide rates of molecularly validated* de novo *SVs from 519 quartet families for comparison.[23] (**b**) SVs were non-uniformly distributed across the genome. Shown here is the smoothed enrichment of SVs per 100 kb window across the average of all autosomes normalized by chromosome arm length (a "meta-chromosome"; also see **Supplementary Figure 9**). (**c**) The distribution of SVs along the meta-chromosome was dependent on variant class. (**d**) Biallelic CNVs were predominantly enriched at telomeres, MCNVs were predominantly enriched at centromeres, and canonical and complex inversions were depleted near telomeres. P-values computed using a t-test; bars correspond to 95% confidence intervals (CIs).*

Mutation rates for SVs have remained difficult to estimate due to technical limitations of SV discovery from WGS, and the frequent use of cell line-derived DNA rather than whole blood in population studies.[1] Using the Watterson Estimator,[28] we projected a mean mutation rate of 0.35 *de novo* SVs per generation in regions of the genome accessible to short-read WGS (95% confidence interval: 0.18-0.52 SV/generation), or roughly one new SV every 2-6 live births, with mutation rates varying markedly by SV class (**Figure 3a**). While this method estimates mutation rates from variation aggregated across unrelated individuals, we previously demonstrated comparable rates from molecularly validated *de novo* SVs in WGS analyses of 519 quartet families.[23] However, our calculations certainly underestimate the true mutation rates for SVs given the reduced sensitivity of short-read WGS in repetitive and low-complexity sequences that can mediate their formation.[29] We anticipate that emerging long-read WGS and assembly methods will greatly increase future estimates of SV mutation rates and clarify their associated mechanisms. Despite the limitations of short-read WGS in repetitive sequence, it was notable that the density of SVs in this study was significantly enriched near centromeres and telomeres (**Figure 3b** and **Supplementary Figure 9**). This trend was strongly dependent on SV class: biallelic deletions and duplications were predominantly enriched at telomeres, whereas MCNVs were preferentially enriched near centromeres (**Figure 3c-d**). Conversely, inversions and complex SVs were apparently depleted in telomeres, although these variants might be more susceptible to false negatives than CNVs due to local repeat structures. These analyses indicate that the processes influencing SV mutation rates and mechanisms of formation vary by SV class and chromosomal context.

Constraint against SVs in protein-coding genes

By virtue of their size and mutational diversity, SVs can have varied consequences on protein-coding sequence (**Figure 4a** and **Supplementary Figure 10**). All classes of SVs can result in pLoF, either by deletion of coding nucleotides or alteration of open-reading frames, and many

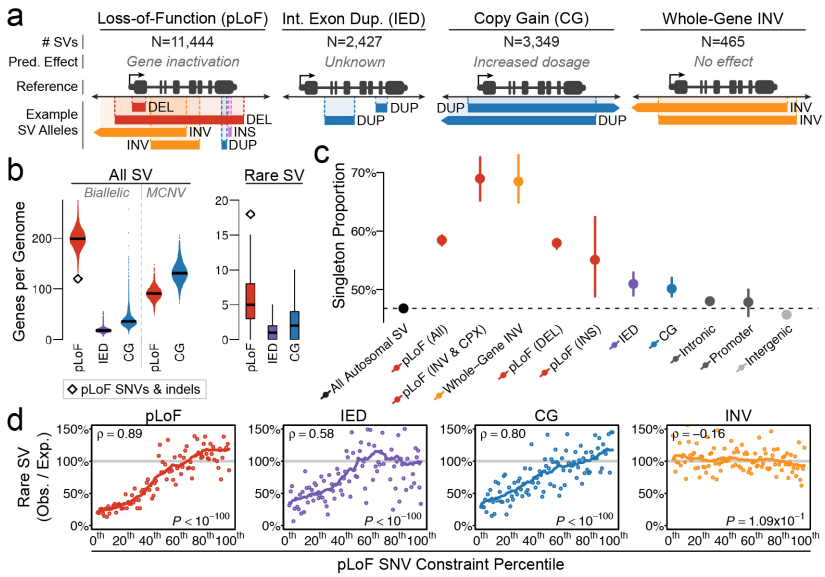


Figure 4 | Pervasive selection against SVs in genes mirrors patterns observed from coding point mutations

*(**a**) Four categories of gene-overlapping SVs, with counts of SVs in gnomAD-SV. (**b**) Distributions of genes altered by SVs per genome. (**c**) Autosomal SVs that overlap genes were enriched for singleton variants (a proxy for the strength of selection[20]) above baseline of all SVs genome-wide, and explicitly intergenic SVs (also see **Extended Data Figure 5c-d**). Bars indicate 100-fold bootstrapped 95% confidence intervals. (**d**) We evaluated the relationship of constraint against pLoF SNVs versus the four categories of gene-overlapping SVs from (a).[20] Each point represents the total of ~175 genes, which have been ranked by SNV constraint. Correlations were assessed with a Spearman test. Solid lines represent 21-bin rolling means. See **Supplementary Figure 10** for comparisons to missense SNV constraint.*

SVs can duplicate or invert coding and noncoding loci.[30] Coding duplications can result in copy-gain of entire genes (CG) or duplication of a subset of exons contained within a gene, referred to here as intragenic exonic duplication (IED). The average genome in gnomAD-SV harbored 253 genes altered by biallelic SVs (199 pLoF, 18 IED, and 36 CG), of which 24 were predicted to be completely inactivated by homozygous biallelic pLoF SVs (**Figure 4b** and **Extended Data Figure 2e-h**). When restricted to rare (AF<1%) SVs, the average genome harbored 8 altered genes (5 pLoF, 1 IED, and 2 CG), effectively all of which were in the heterozygous state. By comparison, prior analyses estimated 120 pLoF SNV/indels per genome, of which 18 were rare,[19] suggesting that up to 25% of all rare pLoF events per genome are likely to result from SVs. We found signals of pervasive selection, such as the proportion of singleton variants,[20] against all classes of SVs that overlap genes, including intronic SVs and pLoF SVs as small as single-exon deletions (**Figure 4c** and **Extended Data Figure 5a-d**). While further methods development will continue to refine these annotations, these data suggest that SVs represent a substantial fraction of all gene-altering variants per genome.

Metrics that quantitatively estimate the strength of selection on functional variation per gene, such as the probability of LoF intolerance (pLI), have become a core resource in human genetics.[19,21] No comparable metrics exist for SVs due to small variant counts by comparison to SNVs. To gain some insight into this problem, we estimated the number of rare SVs expected per gene while adjusting for gene length, exon-intron structure, and genomic context (see **Methods**). This model is imperfect, as expectations can be influenced by many known and unknown covariates, and SVs are too sparse to derive precise gene-level estimates of SV constraint at current sample sizes. Nevertheless, the results from this model displayed several clear and informative patterns. We found strong concordance between pLoF constraint metrics from gnomAD exome analyses and the depletion of rare pLoF SVs across 100 bins of 175 genes each, ordered by SNV constraint (**Figure 4d**; Spearman's rho=0.89).[20] This result was also true of missense constraint, as expected given the strong correlation of missense and pLoF constraint (**Supplementary Figure 11**). We also discovered a comparable positive correlation between CG from rare SVs and pLoF constraint from SNVs (rho=0.80). A weaker, yet significant correlation was detected for IED as well (rho=0.58). By contrast, there was no correlation between pLoF constraint and rare inversions of entire genes without directly disrupting their open reading frames (rho=-0.16), despite canonical and complex inversions appearing under particularly strong selection based on other metrics, such as the proportion of singleton SVs. Intriguingly, we found evidence for strong selection against noncoding inversions involving two or more recombination hotspots, which might suggest that large inversions influence meiotic mechanics in the general population (**Extended Data Figure 6**). When we cross-examined these relationships by using variant frequency distributions as a proxy for the strength of selection, we found the expected trend of an inverse correlation between proportion of singleton SVs and SNV constraint across all functional categories of SVs (**Extended Data Figure 5f**). These comparisons confirm that selection against multiple classes of gene-altering SVs is consistent with patterns observed for SNVs and indels. They further suggest that constraint metrics like pLI, which are derived from pLoF point mutations alone, underlie a general correspondence between haploinsufficiency and triplosensitivity, *on average*, for a large fraction of genes in the genome. Furthermore, these results imply that many highly constrained genes are not simply sensitive to pLoF, but intolerant to increased dosage and structural alterations more broadly.

Relevance to disease association & clinical genetics

Most large-scale disease association studies of SVs have relied upon chromosomal microarrays (CMA), which are limited to detection of large CNVs and have not had reliable reference resources to restrict analyses to ultra-rare variants.[31] We evaluated gnomAD-SV as a filtering tool for previously published CMA-based association studies (N>10,000 samples) that have identified a significant contribution of large CNVs to developmental disorders (DDs),[32] ASD,[25] schizophrenia,[33] and cancer (**Extended Data Figure 7**).[34] Filtering based on gnomAD-SV AFs magnified the previously reported associations of rare genic CNVs in DDs, ASD, and cancer, with less pronounced differences between schizophrenia cases and controls at ultra-rare AFs, consistent with ex-

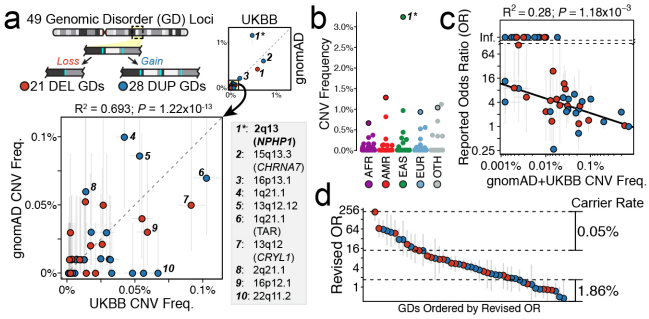


Figure 5 | Using gnomAD-SV to refine estimates of genomic disorder frequencies and penetrance at sequence resolution

*(**a**) Comparison of carrier frequencies for 49 putatively disease-associated deletions (red) and duplications (blue) at genomic disorder (GD) loci between gnomAD-SV and microarray analyses in the UK BioBank (UKBB).[37] Grey bars indicate binomial 95% confidence intervals. Duplications of* NPHP1*, where WGS-derived CNV frequencies significantly differed from UKBB estimates, are marked with an asterisk. (**b**) GD CNV frequencies were comparable across populations in gnomAD-SV, except for duplications at 2q13 (*NPHP1*), where the frequency in East Asian samples was up to 5-fold greater than other populations (2q13 duplications marked with solid black outlines). (**c**) The odds ratios (ORs) for these 49 GDs in DDs were inversely correlated with the combined CNV frequencies in the gnomAD-SV and UKBB datasets (R^2=0.28; P=1.18x10^{-3}; Pearson correlation test).[32] (**d**) Using the larger combined sample size of gnomAD-SV and the UKBB, we re-estimated ORs for each of the 49 GDs by comparing to the 29,085 DD cases from (c).[32]*

pectations from genetic architecture studies.

We next considered previously defined recurrent CNVs associated with syndromic phenotypes, or genomic disorders (GD), which are often mediated by recombination of long flanking segments of homologous sequences.[35] These GDs are among the most prevalent genetic causes of DDs,[36] and accordingly CMA remains the recommended first-tier genetic diagnostic screen for DDs of unknown etiology.[31] Thus, it is critical that these GD CNVs are able to be reliably captured from WGS for both routine clinical screening and studies of developmental and neuropsychiatric disease. Here, we calculated sequence-resolution CNV carrier frequencies in gnomAD for 49 GDs recently reported in the UK BioBank, and found consistent carrier frequency estimates between WGS in gnomAD-SV and those reported by CMA in the UK BioBank (UKBB; R^2=0.69; P=1.22x10^{-13}; Pearson correlation test; **Figure 5a**),[37] further confirming the accuracy of read depth-based discovery of large repeat-mediated CNVs from WGS. GD carrier frequencies did not vary dramatically between populations in gnomAD-SV, with the exception of a single GD (duplications of *NPHP1* at 2q13), where carrier frequencies in East Asian samples were 2.5-to-4.9-fold higher than other populations (**Figure 5b**). This finding underscores the value of characterizing putatively disease-associated SVs across diverse populations. Finally, given the correlation of CNV frequencies between gnomAD-SV and the UKBB, we calculated the combined CNV frequencies from these resources, which were inversely correlated with previously reported odds ratios (ORs) (**Figure 5c**).[32] These data estimate that roughly 0.05% of the population (~1:2,000 individuals) is a carrier of a GD-associated CNV with an estimated OR > 14.0 (*e.g.*, the top quartile of the 49 GDs), compared to 1.86% (~1:54 individuals) for GDs with an OR < 1.7 that represent relatively common polymorphic variants in the population (*e.g.*, the bottom quartile) (**Figure 5d**).

As genomic medicine advances toward diagnostic screening at sequence resolution, publicly accessible WGS references will be indispensable for variant interpretation. The current gnomAD-SV dataset will permit a screening threshold of AF < 0.1% when matching on ancestry to the African, East Asian, European, or Latino populations sampled here, and AF < 0.004% when compared against all samples collectively. For instance, filtering all SVs found in an individual genome versus gnomAD-SV dramatically reduced the number of singleton SVs in that genome to a median of 13, just one of which was pLoF (**Figure 6a** and **Extended Data Figure 3**). This reference dataset also aids in gene-level interpretation, as we catalogued at least one SV resulting in pLoF or CG for 40.4% and 23.5% of all autosomal genes, respectively, and 586 genes with at least one homozygous pLoF SV (**Figure 6b**, **Extended Data Figure 5e**, and **Supplementary Figure 12**). However, these data are still extremely sparse as compared to SNVs and indels, where analyses of the 120,000 gnomAD exomes have documented at least one pLoF SNV for 95.8% of all genes.[20] When further restricted to clinically relevant SVs using American College of Medical Genetics criteria,[38] we find that 0.4% of samples carry a very rare (AF<0.1%) SV resulting in pLoF of a gene for which incidental findings are clinically reportable, roughly half of which (*i.e.*, 0.26% of all samples) likely meet ACMG diagnostic criteria as pathogenic or likely pathogenic (**Figure 6c**). We also observed that 9.5% of individuals were heterozygous carriers of rare pLoF SVs in known recessive DD genes. Finally, we used the gnomAD-SV dataset to catalog rare chromosomal abnormalities (SVs ≥ 1Mb). We estimate that 3.1% of the general population (95% CI: 2.5-3.9%) carries at least one rare autosomal SV ≥1Mb in size, roughly half of which are balanced or complex (**Figure 6d**). Among these events was an example of highly complex localized chromosome shattering involving at least 49 breakpoints yet resulting in largely balanced products, reminiscent of chromothripsis, which was identified in an adult individual from the general population with no indication of severe developmental or pediatric disease (**Figure 6e** and **Extended Data Figure 8**).[8,14,15]

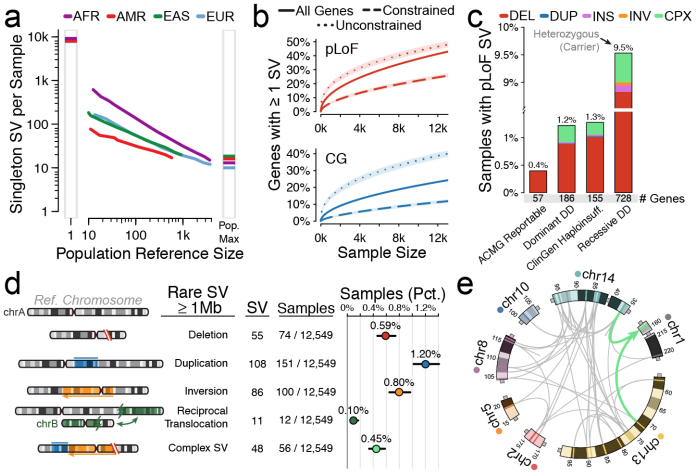


Figure 6 | gnomAD-SV as a resource for clinical WGS interpretation
*(**a**) Filtering SVs against gnomAD-SV reduces individual genomes to ~13 singleton variants at current sample sizes. (**b**) At least one pLoF or CG SV was detected in 40.4% and 23.5% of all autosomal genes, respectively. "Constrained" and "unconstrained" includes the least and most constrained 15% of all genes based on pLoF SNV observed:expected ratios, respectively.[20] (**c**) Up to 1.3% of genomes in gnomAD-SV harbored a very rare (AF<0.1%) pLoF SV in a medically relevant gene across several gene lists.[38-40] Manual review of all very rare pLoF SVs indicated that 0.24% of genomes carry a pathogenic or likely pathogenic variant in a clinically reportable gene for incidental findings.[38] We also found that 9.5% of genomes carried pLoF SVs of recessive DD genes in the heterozygous state.[39] (**d**) We found 308 rare autosomal SVs ≥ 1Mb, revealing that ~3.1% of genomes carry a large, rare chromosomal abnormality. Bars represent binomial 95% confidence intervals. (**e**) An extremely complex SV involving at least 49 breakpoints that localized in clusters across seven chromosomes in a single individual, yielding largely balanced derivatives, reminiscent of chromothripsis (see also **Extended Data Figure 8**). Chromosome coordinates provided as Mb.*

An online, interactive SV reference browser

A key aspect of ExAC and gnomAD for analyses of coding SNVs and indels was the open release of variant information via an user-friendly online interface.[19] Now in its second generation, the gnomAD browser (https://gnomad.broadinstitute.org) has been augmented to incorporate the gnomAD-SV callset described here. Users can query genes and regions to view all SVs, including their mutational class, frequency across populations, predicted genic effects, genotype quality, and other variant metadata (**Extended Data Figure 9**). These features are directly integrated into the existing interface as the gnomAD SNV and indel callsets, where users can toggle between viewing SVs and smaller point mutations within the same window. Finally, all SVs described in this study are provided for download in two common file formats via the gnomAD browser, with no use restrictions on the reanalysis of these data.

DISCUSSION

The fields of human genetics research and clinical diagnostics are becoming increasingly invested in defining the complete spectrum of variation in individual genomes. Ambitious international initiatives to generate short-read WGS in hundreds of thousands of individuals from complex disease cohorts have underwritten this goal,[41-44] and millions of genomes from unselected individuals will be sequenced in the coming years from national biobanks.[45,46] A central challenge to these efforts will be the uniform analysis and interpretation of all variation accessible to short-read WGS, particularly SVs, which are frequently cited as a source of added value offered by WGS over conventional technologies.[47] Indeed, early efforts to deploy WGS in cardiovascular disease, ASD, and type 2 diabetes were largely consistent in their analyses of SNVs using GATK, but all studies have differed in their analyses of SVs.[23,36,42-44,48,49] Thus, while ExAC and gnomAD have catalyzed remarkable advances in medical and population genetics, the same opportunities for new discovery and translational impact have not yet been realized for SVs. Although gnomAD-SV is by no means comprehensive, the half-million SVs it contains will begin to address the dearth of population SV datasets. Given that gnomAD-SV was constructed with contemporary WGS technologies and a reference genome that match those currently used in clinical settings, we anticipate that these data will augment disease

association studies and provide a useful screening tool for clinical interpretation of rare variation.

Most foundational assumptions of human genetic variation were consistent between SNVs/indels from the gnomAD exome study and SVs reported here,[20] most notably that SVs experience selection commensurate with their predicted biological consequences. This study also spotlights unique aspects of SVs, such as their remarkable mutational diversity, their varied functional impact on coding sequence, and the strong selection against large and complex SVs in the genome. We provide resolved structures for nearly six thousand such complex SVs, and predict that SVs comprise up to 25% of all rare pLoF variation in each genome. These analyses also demonstrate that gene-altering effects of SVs beyond pLoF parallel measures of mutational constraint derived from analyses of SNVs. Despite the strong correlation between SNV and SV constraint in this study, we made several assumptions that likely underestimate the true diversity of possible functional outcomes. For instance, we assigned any deletion of an exon from a canonical gene transcript as pLoF. There are technical and biological explanations for why that assumption will not universally hold,[3] yet the proportion of singleton SVs was nearly identical for partial or single exon deletions as for loss of a full copy of a gene (**Extended Data Figure 5d**). More sophisticated models of SV annotation will continue to refine future predictions of their biological impact. The patterns we observed for whole-gene copy gains (CG) and intragenic exonic duplications (IEDs) against pLoF constraint imply that existing SNV constraint metrics are not specific to depletion of pLoF variation, but rather underlie a more generalizable intolerance to alterations of both gene dosage and structure. Indeed, similar patterns of selection were observed for CG and pLoF SVs among the most constrained genes in the genome. Like complex SVs, IEDs are also an intriguing class of SVs that may operate in a context-dependent manner. Analogous to missense variation, IEDs can result in pLoF, neutral variation, or perhaps other effects, and thus represent an exciting area for future investigation. Finally, the strong selection against canonical and complex inversions despite no clear correspondence with existing gene constraint metrics is intriguing, and our analyses suggest that this may be related to large inversions blocking recombination through meiotic interference.

Technical barriers associated with short-read WGS preclude the establishment of a complete catalogue of SVs in gnomAD-SV. A recent study incorporating most extant genomics technologies demonstrated that short-read WGS is limited in low-complexity and repetitive sequence contexts.[17] The technology and methods relied upon here are thus blind to a disproportionate fraction of repeat-mediated SVs, and underestimate the true mutation rates within these hypermutable regions. Similarly, high copy state MCNVs often require specialized algorithms and manual curation to fully delineate their numerous haplotypes,[12,50,51] suggesting that the 1,053 MCNVs reported here comprise an incomplete portrait of extreme copy-number polymorphisms. We expect that emerging technologies, *de novo* assemblies, and graph-based genome representations are likely to expand our knowledge of SVs in repetitive sequences.[51,52] Nevertheless, based on current estimates, 92.7% of known autosomal protein-coding nucleotides are not localized to simple- and low-copy repeats. This suggests that catalogues of SVs accessible to short-read WGS across large populations, like gnomAD-SV, will likely capture a majority of the most interpretable gene-disrupting SVs in humans.

The oncoming deluge of short-read WGS datasets has magnified the need for publicly available large-scale resources of SVs. In this study, we aimed to begin to bridge the gap between the existence of such references for SNVs/indels and those for SVs. While the dataset provided here significantly exceeds current references in terms of sample size and sensitivity, these data remain insufficient to derive accurate estimates of gene-level constraint, and are dramatically underpowered to explore sequence-specific mutation rates and intolerance to noncoding SVs. Nonetheless, these data provide an initial step toward these goals, and demonstrate the value of a commitment to open data sharing and joint analyses of aggregated datasets by the field. The gnomAD-SV resource has been made available without restrictions on reuse, and has been integrated directly into the widely adopted gnomAD Browser (https://gnomad.broadinstitute.org), where users can freely view, sort, and filter the SV dataset described here. This resource will catalyze new discoveries in basic research and provide immediate clinical utility for the interpretation of rare structural rearrangements in the human genome.

METHODS & SUPPLEMENTARY INFO

There is supplementary information associated with this study, which includes detailed methods. These materials have been provided in a separate document, which will be linked directly from *bioRxiv*.

ACKNOWLEDGEMENTS

This work was supported by resources from the Broad Institute, the National Institutes of Health (R01MH115957, R01HD081256, P01GM061354, HD091797 to MET; U01MH105669 to MJD, BN, and MET) and the Simons Foundation for Autism Research Initiative (SFARI #573206 to MET). We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the SSC principal investigators. RLC was supported by NHGRI T32HG002295 and NSF GRFP #2017240332. HB was supported by NIDCR K99DE026824. MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420. MESA Family is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support is provided by grants and contracts R01HL071051, R01HL071205, R01HL071250, R01HL071251, R01HL071258, R01HL071259, by the National Center for Research Resources, Grant UL1RR033176, and the National Center for Advancing Translational Sciences ULTR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center. We thank Loyal Goff and the Goff Laboratory for providing the Adobe InDesign typesetting template adapted for this document.

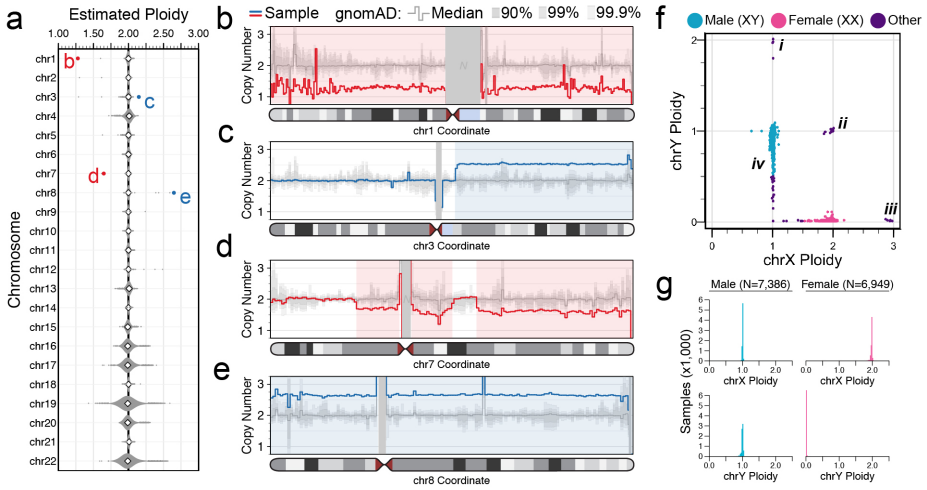
REFERENCES

1. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81, doi:10.1038/nature15394 (2015).
2. Perry, G. H. *et al.* Copy number variation and evolution in humans and chimpanzees. *Genome research* **18**, 1698-1710, doi:10.1101/gr.082016.108 (2008).
3. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature reviews. Genetics* **14**, 125-138, doi:10.1038/nrg3373 (2013).
4. Lucito, R. *et al.* Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. *Genome research* **10**, 1726-1736 (2000).
5. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science (New York, N.Y.)* **316**, 445-449, doi:10.1126/science.1138659 (2007).
6. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899-905, doi:10.1038/nature08822 (2010).
7. Talkowski, M. E. *et al.* Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell* **149**, 525-537, doi:10.1016/j.cell.2012.03.028

(2012).
8. Chiang, C. *et al.* Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nature genetics* **44**, 390-397, S391, doi:10.1038/ng.2202 (2012).
9. Spielmann, M., Lupianez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nature reviews. Genetics* **19**, 453-467, doi:10.1038/s41576-018-0007-0 (2018).
10. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nature reviews. Genetics* **7**, 85-97, doi:10.1038/nrg1767 (2006).
11. Stewart, C. *et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS genetics* **7**, e1002236, doi:10.1371/journal.pgen.1002236 (2011).
12. Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nature genetics* **47**, 296-303, doi:10.1038/ng.3200 (2015).
13. Carvalho, C. M. *et al.* Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nature genetics* **43**, 1074-1081, doi:10.1038/ng.944 (2011).
14. Collins, R. L. *et al.* Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome biology* **18**, 36, doi:10.1186/s13059-017-1158-6 (2017).
15. Kloosterman, W. P. *et al.* Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Human molecular genetics* **20**, 1916-1924, doi:10.1093/hmg/ddr073 (2011).
16. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).
17. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv* (2017).
18. Hehir-Kwa, J. Y. *et al.* A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nature communications* **7**, 12989, doi:10.1038/ncomms12989 (2016).
19. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).
20. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* (2019).
21. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nature genetics* **46**, 944-950, doi:10.1038/ng.3050 (2014).
22. Walsh, R. *et al.* Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genetics in medicine : official journal of the American College of Medical Genetics* **19**, 192-203, doi:10.1038/gim.2016.90 (2017).
23. Werling, D. M. *et al.* An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nature genetics* **50**, 727-736, doi:10.1038/s41588-018-0107-y (2018).
24. Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nature genetics* **49**, 692-699, doi:10.1038/ng.3834 (2017).
25. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215-1233, doi:10.1016/j.neuron.2015.09.016 (2015).
26. Barbujani, G. & Colonna, V. Human genome diversity: frequently asked questions. *Trends in genetics : TIG* **26**, 285-295, doi:10.1016/j.tig.2010.04.002 (2010).
27. Brand, H. *et al.* Paired-Duplication Signatures Mark Cryptic Inversions and Other Complex Structural Variation. *American journal of human genetics* **97**, 170-176, doi:10.1016/j.ajhg.2015.05.012 (2015).
28. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theoretical population biology* **7**, 256-276 (1975).
29. Carvalho, C. M. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nature reviews. Genetics* **17**, 224-238, doi:10.1038/nrg.2015.25 (2016).
30. Hurles, M. E., Dermitzakis, E. T. & Tyler-Smith, C. The functional impact of structural variation in humans. *Trends in genetics : TIG* **24**, 238-245, doi:10.1016/j.tig.2008.03.001 (2008).
31. Miller, D. T. *et al.* Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *American journal of human genetics* **86**, 749-764, doi:10.1016/j.ajhg.2010.04.006 (2010).
32. Coe, B. P. *et al.* Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nature genetics* **46**, 1063-1071, doi:10.1038/ng.3092 (2014).
33. Psychiatric Genetics Consortium, T. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature genetics*, doi:10.1038/ng.3725 (2016).
34. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature genetics* **45**, 1134-1140, doi:10.1038/ng.2760 (2013).
35. Dittwald, P. *et al.* NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. *Genome research* **23**, 1395-1409, doi:10.1101/gr.152454.112 (2013).
36. Yuen, R. K. *et al.* Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nature neuroscience* **20**, 602-611, doi:10.1038/nn.4524 (2017).
37. Owen, D. *et al.* Effects of pathogenic CNVs on physical traits in participants of the UK Biobank. *BMC genomics* **19**, 867, doi:10.1186/s12864-018-5292-7 (2018).
38. Green, R. C. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in medicine : official journal of the American College of Medical Genetics* **15**, 565-574, doi:10.1038/gim.2013.73 (2013).
39. Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet (London, England)* **385**, 1305-1314, doi:10.1016/s0140-6736(14)61705-0 (2015).
40. Rehm, H. L. *et al.* ClinGen--the Clinical Genome Resource. *The New England journal of medicine* **372**, 2235-2242, doi:10.1056/NEJMsr1406261 (2015).
41. Sanders, S. J. *et al.* Whole genome sequencing in psychiatric disorders: the WGSPD consortium. *Nature neuroscience* **20**, 1661-1668, doi:10.1038/s41593-017-0017-9 (2017).
42. Natarajan, P. *et al.* Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nature communications* **9**, 3391, doi:10.1038/s41467-018-05747-8 (2018).
43. Choi, S. H. *et al.* Association Between Titin Loss-of-Function Variants and Early-Onset Atrial Fibrillation. *Jama* **320**, 2354-2364, doi:10.1001/jama.2018.18179 (2018).
44. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41-47, doi:10.1038/nature18642 (2016).
45. Collins, F. S. & Varmus, H. A new initiative on precision medicine. *The New England journal of medicine* **372**, 793-795, doi:10.1056/NEJMp1500523 (2015).
46. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **12**, e1001779, doi:10.1371/journal.pmed.1001779 (2015).
47. Meienberg, J., Bruggmann, R., Oexle, K. & Matyas, G. Clinical sequencing: is WGS the better WES? *Human genetics* **135**, 359-362, doi:10.1007/s00439-015-1631-9 (2016).
48. An, J. Y. *et al.* Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science (New York, N.Y.)* **362**, doi:10.1126/science.aat6576 (2018).
49. Turner, T. N. *et al.* Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171**, 710-722.e712, doi:10.1016/j.cell.2017.08.047 (2017).
50. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177-183, doi:10.1038/nature16549 (2016).
51. Chaisson, M. J., Wilson, R. K. & Eichler, E. E. Genetic variation and the de novo assembly of human genomes. *Nature reviews. Genetics*

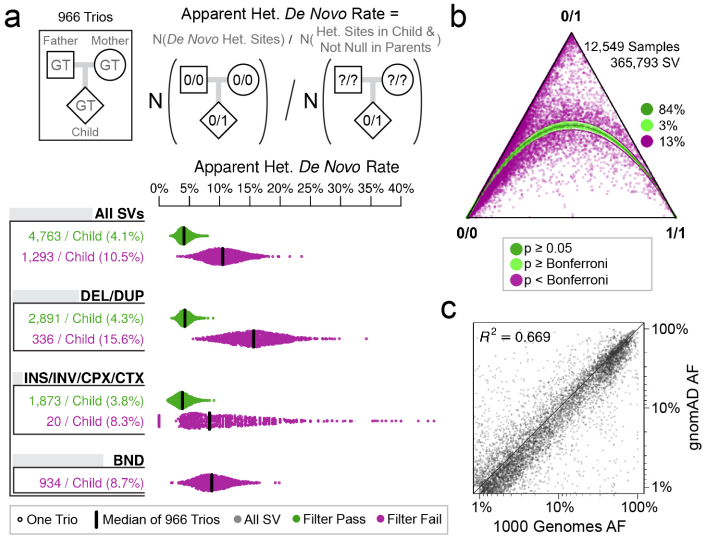
16, 627-640, doi:10.1038/nrg3933 (2015).

52. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature biotechnology* **36**, 875-879, doi:10.1038/nbt.4227 (2018).



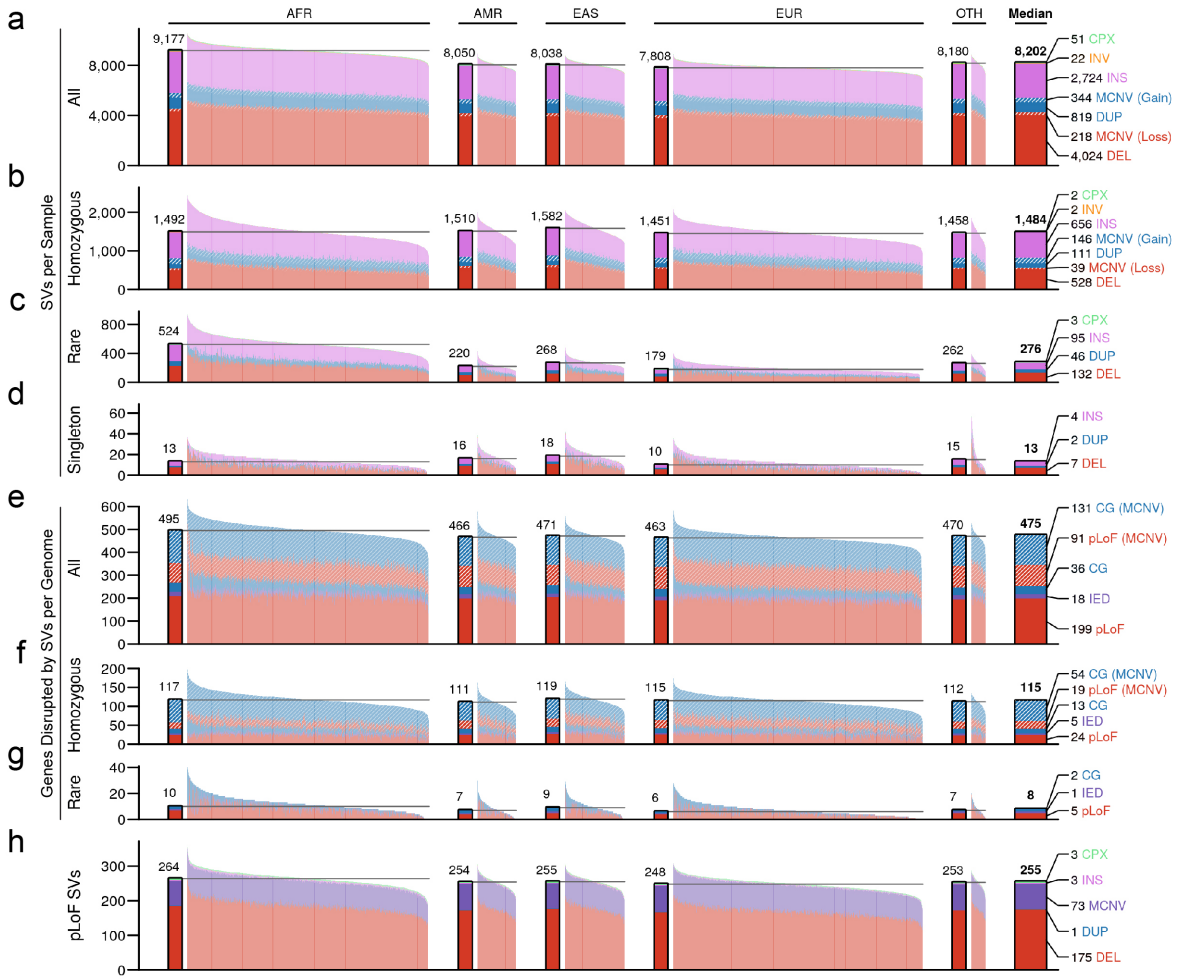
Extended Data Figure 1 | Detection of chromosome-scale dosage alterations

*We estimated ploidy (i.e. whole-chromosome copy number) for all 24 chromosomes per sample. (**a**) Distribution of autosome ploidy estimates across 14,378 samples passing initial data quality thresholds. The outlier points marked in red and blue correspond to the samples highlighted in panels (b-e). (**b-e**) Samples with outlier autosome ploidy estimates typically harbored somatic or mosaic chromosomal abnormalities, such as somatic aneuploidy of chr1 (b) or chr8 (e), or large focal somatic or mosaic CNVs on chr3 (c) and chr7 (d). Each panel depicts copy-number estimates in 1Mb bins for each rearranged sample in red or blue. Dark, medium, and light grey background shading indicates the range of copy number estimates for 90%, 99%, and 99.9% of all gnomAD-SV samples, respectively, and the medium grey line indicates the median copy number estimate across all samples. Regions of unalignable N-masked bases >1Mb in the reference genome are masked with grey rectangles. (**f**) Sex chromosome ploidy estimates for all samples from (a). We inferred karyotypic sex by clustering samples to their nearest integer ploidy for sex chromosomes. Several abnormal sex chromosome ploidies are marked, including XYY (i), XXY (ii), XXX (iii), and mosaic loss-of-Y (iv). (**g**) The overwhelming majority of samples conformed to canonical sex chromosome ploidies.*



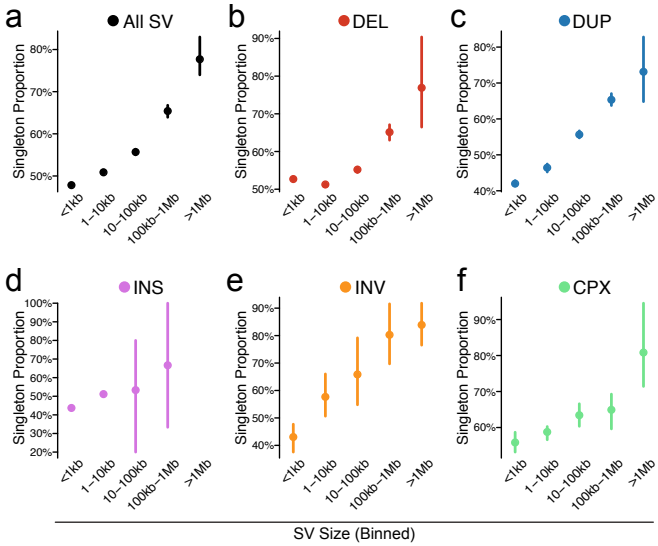
Extended Data Figure 2 | Benchmarking the technical qualities of the gnomAD-SV callset

*We evaluated the quality of gnomAD-SV with five orthogonal analyses detailed in **Supplementary Table 4** and **Supplementary Figures 5-7**. Three core analyses are presented here. (**a**) We assessed Mendelian transmission for heterozygous SVs in 966 parent-child trios. Given the expected mutation rate of SVs accessible to short-read WGS (<1 true* de novo *SV per trio; see also **Figure 3a**),[1,23] most* de novo *SVs represented a combination of false-positive genotypes in children and/or false-negative genotypes in parents. These analyses revealed an apparent* de novo *rate of 4.1% in gnomAD-SV final variants ("Filter Pass"; green). For comparison, the apparent* de novo *rate is provided for variants that did not pass* post hoc *site-level filters ("Filter Fail"; purple). Notably, these failed variants with lower quality metrics are predominantly comprised of unresolved BNDs (72.2% of all failing SVs per trio). (**b**) We assessed Hardy-Weinberg Equilibrium (HWE) for all biallelic SVs localized to autosomes. Vertex labels reflect genotypes: 0/0=homozygous reference; 0/1=heterozygous; 1/1=homozygous alternate, with all sites shaded by HWE p-value. (**c**) AFs were strongly correlated between common (AF>1%) SVs captured by both the 1000 Genomes Project and gnomAD-SV.[1] A Pearson correlation coefficient is provided here.*



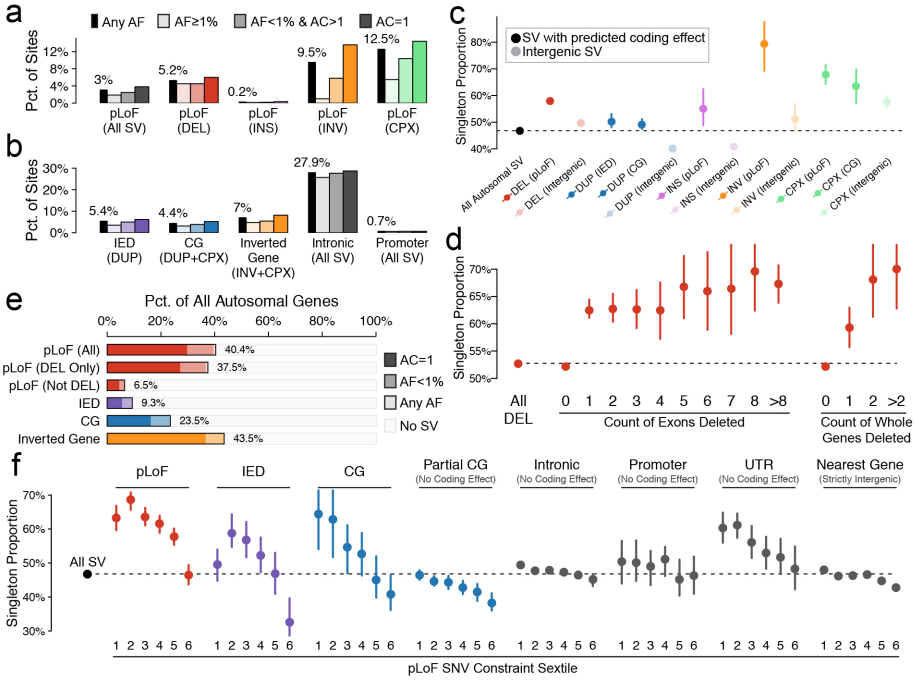
Extended Data Figure 3 | SVs contribute a substantial burden of rare, homozygous, and coding mutations per genome

*(**a**-**d**) We tabulated the count of SVs per genome across a variety of parameters, finding a median of (a) 8,202 total SVs, (b) 1,484 homozygous SVs, (c) 276 rare SVs, and (d) 13 singleton SVs. These counts varied by population, with African/African-American (AFR) samples having the greatest genetic diversity and East Asian (EAS) samples having the most homozygosity. Colors correspond to SV types as indicated to the right of each panel, and the solid bar to the left of each population indicates the population median. (**e**-**g**) We evaluated the contribution of SVs to disruptions of protein-coding genes per genome, and found median counts of genes disrupted by SVs of (e) 467 due to all SVs (including MCNVs; 253 genes altered by biallelic SVs), (f) 115due to homozygous SVs (including MCNVs), and (g) eight genes due to rare SVs. Colors correspond to predicted functional consequence as indicated to the right of each panel. (**h**) We found that a diverse spectrum of SV classes contribute pLoF alleles to every genome. The average genome harbored 255 pLoF SVs. For certain categories, such as genes disrupted by rare SVs per genome, a subset of samples (<5%) were enriched above the population average, as expected for individuals carrying large, rare CNVs predicted to cause the disruption of dozens or hundreds of genes (see **Extended Data Figure 1**); thus, for the purposes of visualization, the y-axis for all panels presented here has been restricted to a maximum of three interquartile ranges above the third quartile across all samples for each category.*



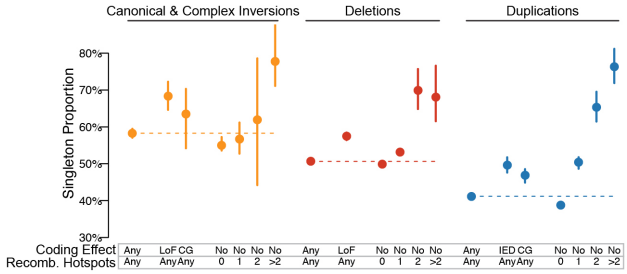
Extended Data Figure 4 | Rearrangement size is a primary determinant of negative selection for most classes of SVs

*We calculated the proportion of singletons in five SV size bins for (**a**) all SVs, (**b**) biallelic deletions, (**c**) biallelic duplications, (**d**) insertions, (**e**) inversions, and (**f**) complex SVs. For this analysis, we excluded SVs that were in highly repetitive or low-complexity sequence (≥30% coverage by annotated segmental duplications or simple repeats). The proportion of singletons for all SV classes exhibited a clear dependency on SV size. Bars reflect 95% confidence intervals from 100-fold bootstrapping.*



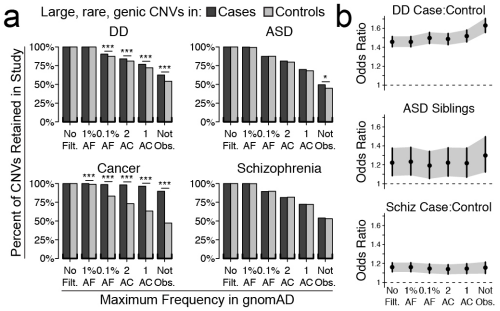
Extended Data Figure 5 | Most SVs within genes appear under negative selection

*These data suggest that negative selection acts against SVs overlapping genes across a broad spectrum of mutational classes and predicted functional consequences. (**a**) Rare and singleton variants were enriched for SVs with pLoF consequences on at least one gene, with singleton variants being enriched to a greater extent than rare variants. (**b**) SVs with predicted genic effects beyond pLoF, such as IED, CG, and whole-gene inversions, displayed similar enrichments as pLoF among rare and singleton variants. The same trend was also observed for intronic SVs and variants overlapping gene promoters, suggesting that these classes of variation are also likely subjected to negative selection. (**c**) Using the proportion of singletons as a proxy for the strength of negative selection, we found that SVs predicted to directly alter coding sequence (pLoF, IED, or CG) were enriched for singletons above the baseline of all autosomal SV, or also when compared to intergenic SVs of the same SV class (lighter dots), suggesting the overall observation of negative selection on pLoF SVs was not specific to a single SV class or context. For panels (c), (d), and (f), bars represent 95% confidence intervals from 100-fold bootstrapping. (**d**) The proportion of singletons for pLoF deletions was correlated with the number of exons and number of whole genes deleted, although this is also closely linked to SV size (see **Extended Data Figure 4**). Notably, even single-exon deletions exhibited a clear enrichment of singletons above the baselines of both all deletions and noncoding deletions, implying that our pLoF annotations were capturing likely disruptive SVs down to single-exon resolution. For this analysis, we excluded SVs that were in highly repetitive or low-complexity sequence (≥30% coverage by annotated segmental duplications or simple repeats). (**e**) In total, we found one pLoF SV for 40.4% of all autosomal protein-coding genes, and similar or lower fractions for all other SV effects considered in this study. (**f**) Most categories of SVs, including intronic, promoter, and UTR SVs, exhibited an inverse relationship between proportion of singleton SVs and pLoF SNV constraint.*



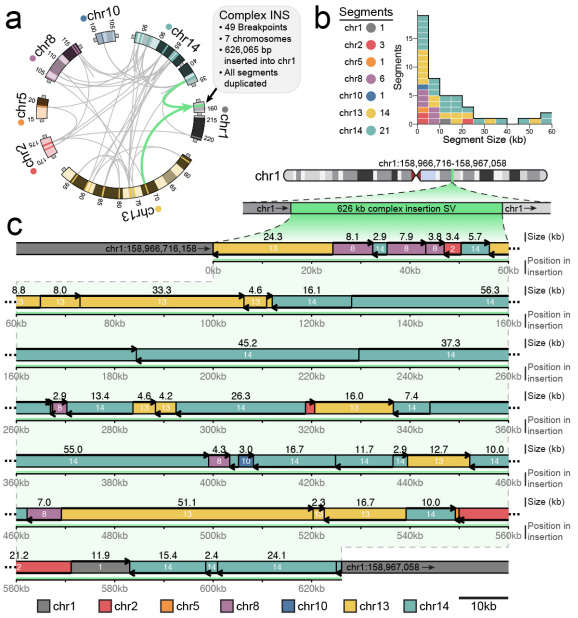
Extended Data Figure 6 | Evidence of selection against possible meiotic interference caused by large noncoding inversions and CNVs

We evaluated the hypothesis that large SVs might cause meiotic interference by blocking recombination by evaluating a proxy for strength of selection (singleton proportion) between various categories of SVs. We compared SVs in this dataset against recombination hotspots, conditioned by whether or not each SV had any predicted direct effects on coding sequence. For inversions, deletions, and duplications, we found that rearrangements with no predicted genic effect that also were predicted to alter more than two recombination hotspots were under particularly strong selection, surpassing the degree of selection against SVs from the same class with direct coding effects. Although sample sizes were small, these analyses suggest that noncoding SVs may be selected against when predicted to disrupt multiple recombination hotspots, potentially due to mechanisms of meiotic interference.



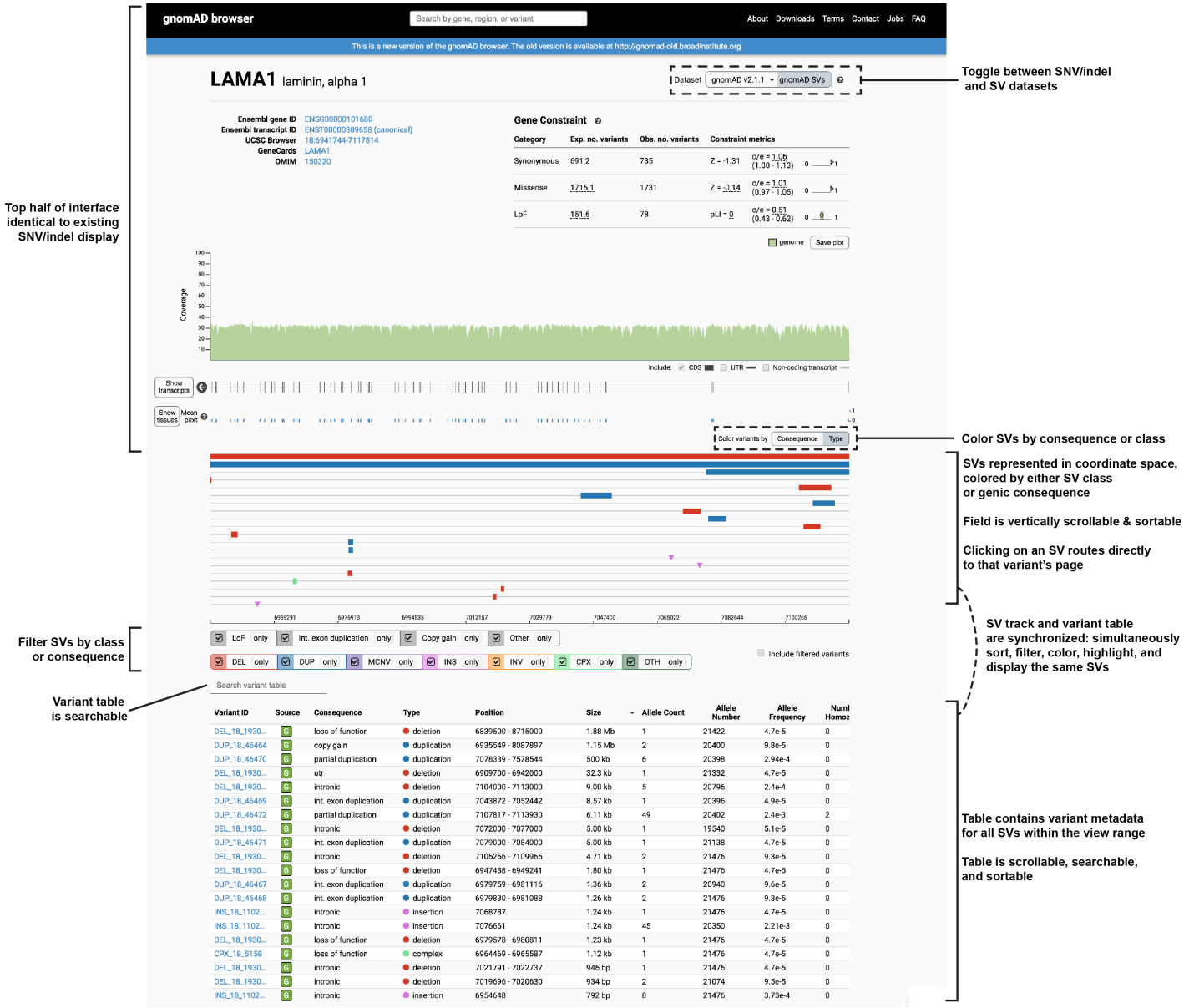
Extended Data Figure 7 | gnomAD-SV can augment disease association studies of SVs

*(**a**) Filtering CNV calls from microarray disease association studies against gnomAD-SV can magnify reported signals of association between ultra-rare genic CNVs and various diseases, including DDs,[32] ASD,[25] and cancer.[34] Bars represent the total number of large (≥100kb), rare (frequency<0.1% in the original study) CNVs overlapping at least one protein-coding exon across all cases or controls after filtering versus gnomAD-SV. Asterisks correspond to P-value thresholds of 0.05, 0.005, and 0.0005, respectively. AF=max allele frequency in gnomAD-SV; AC=max allele count in gnomAD-SV; "Not. Obs"=not observed in gnomAD-SV. (**b**) Odds ratios and 95% confidence intervals corresponding to the filtering procedures used in (a).*



Extended Data Figure 8 | An extremely complex SV involving 49 breakpoints and seven chromosomes

*In the gnomAD-SV cohort, we identified one highly complex insertion rearrangement where 47 segments from six different chromosomes were duplicated and inserted into a single locus on chromosome 1, forming a 626,065 bp stretch of contiguous inserted sequence composed of shattered fragments. Given the involvement of multiple chromosomes, the signature of localized shattering, and the clustered breakpoints, we note that this rearrangement has several hallmarks of germline chromothripsis.[14] However, unlike previous reports of germline chromothripsis, there are no apparent whole-chromosome translocations, and all segments were duplicated before being inserted in a compound manner into chromosome 1, potentially suggesting a replication-based repair mechanism. The exact origin of this rearrangement is unclear. (**a**) Circos representation of all 49 breakpoints and seven chromosomes involved in this SV, reproduced from **Figure 6** for clarity. (**b**) The median segment size was 8.4kb. (**c**) Linear representation of the rearranged inserted sequence. Colors correspond to chromosome of origin, and arrows indicate strandedness of inserted sequence, relative to the GRCh37 reference.*



Extended Data Figure 9 | An online, interactive platform to query, filter, and download the gnomAD-SV resource

The existing gnomAD browser (https://gnomad.broadinstitute.org) has been modified to incorporate the gnomAD-SV data described in this study. These data can be queried on a per-gene, per-locus, or per-variant basis, and toggled between SNV/indel SV datasets within the same view range. Shown here is an example of the SV mode for a gene-level query, for the gene LAMA1. *This screenshot is annotated with several new SV features, and highlights some of the functionality of the new gnomAD browser mode, such as coloring by SV class or genic consequence, a synchronized display of SVs in the variant track and metadata table below, and the ability to sort, filter, and scroll freely among SVs in the view range.*

GROUP AUTHORS

Genome Aggregation Database Production Team

Jessica Alföldi[1,2], Irina M. Armean[3,1,2], Eric Banks[4], Louis Bergelson[4], Kristian Cibulskis[4], Ryan L Collins[1,5,6], Kristen M. Connolly[7], Miguel Covarrubias[4], Beryl Cummings[1,2,8], Mark J. Daly[1,2,9], Stacey Donnelly[1], Yossi Farjoun[4], Steven Ferriera[10], Laurent Francioli[1,2], Stacey Gabriel[10], Laura D. Gauthier[4], Jeff Gentry[4], Namrata Gupta[10,1], Thibault Jeandet[4], Diane Kaplan[4], Konrad J. Karczewski[1,2], Kristen M. Laricchia[1,2], Christopher Llanwarne[4], Eric V. Minikel[1], Ruchi Munshi[4], Benjamin M Neale[1,2], Sam Novod[4], Anne H. O'Donnell-Luria[1,11,12], Nikelle Petrillo[4], Timothy Poterba[9,2,1], David Roazen[4], Valentin Ruano-Rubio[4], Andrea Saltzman[1], Kaitlin E. Samocha[13], Molly Schleicher[1], Cotton Seed[9,2], Matthew Solomonson[1,2], Jose Soto[4], Grace Tiao[1,2], Kathleen Tibbetts[4], Charlotte Tolonen[4], Christopher Vittal[9,2], Gordon Wade[4], Arcturus Wang[9,2,1], Qingbo Wang[1,2,6], James S Ware[14,15,1], Nicholas A Watts[1,2], Ben Weisburd[4], Nicola Whiffin[14,15,1]

1. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
2. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
3. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom
4. Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
5. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
6. Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA 02115, USA
7. Genomics Platform, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
8. Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, 02115, USA
9. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
10. Broad Genomics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
11. Division of Genetics and Genomics, Boston Children's Hospital, Boston, Massachusetts 02115, USA
12. Department of Pediatrics, Harvard Medical School, Boston, Massachusetts 02115, USA
13. Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK
14. National Heart & Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London UK
15. Cardiovascular Research Centre, Royal Brompton & Harefield Hospitals NHS Trust, London UK

Genome Aggregation Database Consortium

Carlos A Aguilar Salinas[1], Tariq Ahmad[2], Christine M. Albert[3,4], Diego Ardissino[5], Gil Atzmon[6,7], John Barnard[8], Laurent Beaugerie[9], Emelia J. Benjamin[10,11,12], Michael Boehnke[13], Lori L. Bonnycastle[14], Erwin P. Bottinger[15], Donald W Bowden[16,17,18], Matthew J Bown[19,20], John C Chambers[21,22,23], Juliana C. Chan[24], Daniel Chasman[3,25], Judy Cho[15], Mina K. Chung[26], Bruce Cohen[27,25], Adolfo Correa[28], Dana Dabelea[29], Mark J. Daly[30,31,32], Dawood Darbar[33], Ravindranath Duggirala[34], Josée Dupuis[35,36], Patrick T. Ellinor[30,37], Roberto Elosua[38,39,40], Jeanette Erdmann[41,42,43], Tõnu Esko[30,44], Martti Färkkilä[45], Jose Florez[46], Andre Franke[47], Gad Getz[48,49,25], Benjamin Glaser[50], Stephen J. Glatt[51], David Goldstein[52,53], Clicerio Gonzalez[54], Leif Groop[55,56], Christopher Haiman[57], Craig Hanis[58], Matthew Harms[59,60], Mikko Hiltunen[61], Matti M. Holi[62], Christina M. Hultman[63,64], Mikko Kallela[65], Jaakko Kaprio[56,66], Sekar Kathiresan[67,68,25], Bong-Jo Kim[69], Young Jin Kim[69], George Kirov[70], Jaspal Kooner[23,22,71], Seppo Koskinen[72], Harlan M. Krumholz[73], Subra Kugathasan[74], Soo Heon Kwak[75], Markku Laakso[76,77], Terho Lehtimäki[78], Ruth J.F. Loos[15,79], Steven A. Lubitz[30,37], Ronald C.W. Ma[24,80,81], Daniel G. MacArthur[31,30], Jaume Marrugat[82,39], Kari M. Mattila[78], Steven McCarroll[32,83], Mark I McCarthy[84,85,86], Dermot McGovern[87], Ruth McPherson[88], James B. Meigs[89,25,90], Olle Melander[91], Andres Metspalu[44], Benjamin M Neale[30,31], Peter M Nilsson[92], Michael C O'Donovan[70], Dost Ongur[27,25], Lorena Orozco[93], Michael J Owen[70], Colin N.A. Palmer[94], Aarno Palotie[56,32,31], Kyong Soo Park[75,95], Carlos Pato[96], Ann E. Pulver[97], Nazneen Rahman[98], Anne M. Remes[99], John D. Rioux[100,101], Samuli Ripatti[56,66,102], Dan M. Roden[103,104], Danish Saleheen[105,106,107], Veikko Salomaa[108], Nilesh J. Samani[19,20], Jeremiah Scharf[30,32,67], Heribert Schunkert[109,110], Moore B. Shoemaker[111], Pamela Sklar*[112,113,114], Hilkka Soininen[115], Harry Sokol[9], Tim Spector[116], Patrick F. Sullivan[63,117], Jaana Suvisaari[108], E Shyong Tai[118,119,120], Yik Ying Teo[118,121,122], Tuomi Tiinamaija[56,123,124], Ming Tsuang[125,126], Dan Turner[127], Teresa Tusie-Luna[128,129], Erkki Vartiainen[66], James S Ware[130,131,30], Hugh Watkins[132], Rinse K Weersma[133], Maija Wessman[123,56], James G. Wilson[134], Ramnik J. Xavier[135,136]

1. Unidad de Investigacion de Enfermedades Metabolicas. Instituto Nacional de Ciencias Medicas y Nutricion. Mexico City
2. Peninsula College of Medicine and Dentistry, Exeter, UK
3. Division of Preventive Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA.
4. Division of Cardiovascular Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.
5. Department of Cardiology, University Hospital, 43100 Parma, Italy
6. Department of Biology, Faculty of Natural Sciences, University of Haifa, Haifa, Israel
7. Departments of Medicine and Genetics, Albert Einstein College of Medicine, Bronx, NY, USA, 10461
8. Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44122, USA
9. Sorbonne Université, APHP, Gastroenterology Department, Saint Antoine Hospital, Paris, France
10. NHLBI and Boston University's Framingham Heart Study, Framingham, Massachusetts, USA.
11. Department of Medicine, Boston University School of Medicine, Boston, Massachusetts, USA.
12. Department of Epidemiology, Boston University School of Public Health, Boston, Massachusetts, USA.
13. Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan 48109
14. National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
15. The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY
16. Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA
17. Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA
18. Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, NC, USA
19. Department of Cardiovascular Sciences, University of Leicester, Leicester, UK
20. NIHR Leicester Biomedical Research Centre, Glenfield Hospital, Leicester, UK
21. Department of Epidemiology and Biostatistics, Imperial College London, London, UK
22. Department of Cardiology, Ealing Hospital NHS Trust, Southall, UK
23. Imperial College Healthcare NHS Trust, Imperial College London, London, UK
24. Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China.
25. Department of Medicine, Harvard Medical School, Boston, MA
26. Departments of Cardiovascular Medicine, Cellular and Molecular Medicine, Molecular Cardiology, and Quantitative Health Sciences, Cleveland Clinic, Cleveland, Ohio, USA.
27. McLean Hospital, Belmont, MA
28. Department of Medicine, University of Mississippi Medical Center, Jackson, Mississippi, USA
29. Department of Epidemiology, Colorado School of Public Health, Aurora, Colorado, USA.
30. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
31. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
32. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA
33. Department of Medicine and Pharmacology, University of Illinois at Chicago
34. Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA
35. Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA
36. National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA 01702, USA
37. Cardiac Arrhythmia Service and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA
38. Cardiovascular Epidemiology and Genetics, Hospital del Mar Medical Research Institute (IMIM). Barcelona, Catalonia, Spain
39. CIBER CV, Barcelona, Catalonia, Spain
40. Department of Medicine, Medical School, University of Vic-Central University of Catalonia. Vic, Catalonia, Spain
41. Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany
42. 1. DZHK (German Research Centre for Cardiovascular Research), partner site Hamburg/Lübeck/Kiel, 23562 Lübeck, Germany
43. University Heart Center Lübeck, 23562 Lübeck, Germany
44. Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia
45. Helsinki University and Helsinki University Hospital, Clinic of Gastroenterology, Helsinki, Finland.

46. Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital; Programs in Metabolism and Medical & Population Genetics, Broad Institute; Department of Medicine, Harvard Medical School
47. Institute of Clinical Molecular Biology (IKMB), Christian-Albrechts-University of Kiel, Kiel, Germany
48. Bioinformatics Program, MGH Cancer Center and Department of Pathology
49. Cancer Genome Computational Analysis, Broad Institute.
50. Endocrinology and Metabolism Department, Hadassah-Hebrew University Medical Center, Jerusalem, Israel
51. Department of Psychiatry and Behavioral Sciences; SUNY Upstate Medical University
52. Institute for Genomic Medicine, Columbia University Medical Center, Hammer Health Sciences, 1408, 701 West 168th Street, New York, New York 10032, USA.
53. Department of Genetics & Development, Columbia University Medical Center, Hammer Health Sciences, 1602, 701 West 168th Street, New York, New York 10032, USA.
54. Centro de Investigacion en Salud Poblacional. Instituto Nacional de Salud Publica MEXICO
55. Lund University, Sweden
56. Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland
57. Lund University Diabetes Centre
58. Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX 77030
59. Department of Neurology, Columbia University
60. Institute of Genomic Medicine, Columbia University
61. Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland
62. Department of Psychiatry, PL 320, Helsinki University Central Hospital, Lapinlahdentie, 00 180 Helsinki, Finland
63. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
64. Icahn School of Medicine at Mount Sinai, New York, NY, USA
65. Department of Neurology, Helsinki University Central Hospital, Helsinki, Finland.
66. Department of Public Health, Faculty of Medicine, University of Helsinki, Finland
67. Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
68. Cardiovascular Disease Initiative and Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
69. Center for Genome Science, Korea National Institute of Health, Chungcheongbuk-do, Republic of Korea.
70. MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University School of Medicine, Hadyn Ellis Building, Maindy Road, Cardiff CF24 4HQ
71. National Heart and Lung Institute, Cardiovascular Sciences, Hammersmith Campus, Imperial College London, London, UK.
72. Department of Health, THL-National Institute for Health and Welfare, 00271 Helsinki, Finland.
73. Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, Connecticut.
74. Division of Pediatric Gastroenterology, Emory University School of Medicine, Atlanta, Georgia, USA.
75. Department of Internal Medicine, Seoul National University Hospital, Seoul, Republic of Korea
76. The University of Eastern Finland, Institute of Clinical Medicine, Kuopio, Finland
77. Kuopio University Hospital, Kuopio, Finland
78. Department of Clinical Chemistry, Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Health Technology, Tampere University, Finland
79. The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY
80. Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China.
81. Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China.
82. Cardiovascular Research REGICOR Group, Hospital del Mar Medical Research Institute (IMIM). Barcelona, Catalonia.
83. Department of Genetics, Harvard Medical School, Boston, MA, USA
84. Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Old Road, Headington, Oxford, OX3 7LJ UK
85. Wellcome Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK
86. Oxford NIHR Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford OX3 9DU, UK
87. F Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA.
88. Atherogenomics Laboratory, University of Ottawa Heart Institute, Ottawa, Canada
89. Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, 02114
90. Program in Population and Medical Genetics, Broad Institute, Cambridge, MA
91. Department of Clinical Sciences, University Hospital Malmo Clinical Research Center, Lund University, Malmo, Sweden.
92. Lund University, Dept. Clinical Sciences, Skane University Hospital, Malmo, Sweden
93. Instituto Nacional de Medicina Genómica (INMEGEN), Mexico City, 14610, Mexico
94. Medical Research Institute, Ninewells Hospital and Medical School, University of Dundee, Dundee, UK.
95. Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Republic of Korea
96. Department of Psychiatry, Keck School of Medicine at the University of Southern California, Los Angeles, California, USA.
97. Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
98. Division of Genetics and Epidemiology, Institute of Cancer Research, London SM2 5NG
99. Medical Research Center, Oulu University Hospital, Oulu, Finland and Research Unit of Clinical Neuroscience, Neurology, University of Oulu, Oulu, Finland.
100. Research Center, Montreal Heart Institute, Montreal, Quebec, Canada, H1T 1C8
101. Department of Medicine, Faculty of Medicine, Université de Montréal, Québec, Canada
102. Broad Institute of MIT and Harvard, Cambridge MA, USA
103. Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA.
104. Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA.
105. Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA
106. Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA
107. Center for Non-Communicable Diseases, Karachi, Pakistan
108. National Institute for Health and Welfare, Helsinki, Finland
109. Deutsches Herzzentrum München, Germany
110. Technische Universität München
111. Division of Cardiovascular Medicine, Nashville VA Medical Center and Vanderbilt University, School of Medicine, Nashville, TN 37232-8802, USA.
112. Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA
113. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
114. Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
115. Institute of Clinical Medicine, neurology, University of Eastern Finland, Kuopio, Finland
116. Department of Twin Research and Genetic Epidemiology, King's College London, London UK
117. Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, USA
118. Saw Swee Hock School of Public Health, National University of Singapore, National University Health System, Singapore
119. Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
120. Duke-NUS Graduate Medical School, Singapore
121. Life Sciences Institute, National University of Singapore, Singapore.
122. Department of Statistics and Applied Probability, National University of Singapore, Singapore.
123. Folkhälsan Institute of Genetics, Folkhälsan Research Center, Helsinki, Finland
124. HUCH Abdominal Center, Helsinki University Hospital, Helsinki, Finland
125. Center for Behavioral Genomics, Department of Psychiatry, University of California, San Diego
126. Institute of Genomic Medicine, University of California, San Diego
127. Juliet Keidan Institute of Pediatric Gastroenterology, Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Israel
128. Instituto de Investigaciones Biomédicas UNAM Mexico City
129. Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán Mexico City
130. National Heart & Lung Institute & MRC London Institute of Medical Sciences, Imperial College London, London UK
131. Cardiovascular Research Centre, Royal Brompton & Harefield Hospitals NHS Trust, London UK
132. Radcliffe Department of Medicine, University of Oxford, Oxford UK
133. Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands
134. Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 39216, USA
135. Program in Infectious Disease and Microbiome, Broad Institute of MIT and Harvard, Cambridge, MA, USA
136. Center for Computational and Integrative Biology, Massachusetts General Hospital