nature reviews genetics

https://doi.org/10.1038/s41576-024-00808-9

Supplementary information

Diversity and consequences of structural variation in the human genome

In the format provided by the authors and unedited

Supplementary Information

Diversity and consequences of structural variation in the human genome
Ryan L. Collins & Michael Talkowski

Data availability: at several points in this manuscript, we performed reanalysis of publicly available datasets to quantify certain topics or themes. All data are available from their original sources as described below, which are ordered alphabetically:

- **Cooper *et al.* (2011)**: we downloaded microarray-based control CNV data from Cooper *et al., Nat. Genet.* (2011) via NCBI dbVaR accession #nstd54 (https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd54/).

- **Exome Aggregation Consortium (ExAC)**: we downloaded exome-based CNV data from ExAC as described by Ruderfer *et al., Nat. Genet.* (2017) via the ExAC (now "gnomAD") website (https://gnomad.broadinstitute.org/data#exac-other).

- **NIH Center for Common Disease Genomics (CCDG)**: the CCDG SV callset described in Abel *et al., Nature* (2020) is freely available to download from GitHub (https://github.com/hall-lab/sv_paper_042020).

- **Gencode:** we used protein-coding genes defined by Gencode for all gene-related analyses. For analyses in hg19 coordinates (gnomAD v2.1, 1000G ph.III), we used Gencode v19. For analyses in hg38 coordinates (CCDG, 1000G 30X), we used Gencode v44. All Gencode versions are freely available from the Gencode website (https://www.gencodegenes.org/human/).

- **The Genome Aggregation Database (gnomAD)**: the gnomAD v2.1 SV callset described in Collins *et al., Nature* (2020) is freely available to download from the gnomAD website (https://gnomad.broadinstitute.org/data#v2-structural-variants).

- **Genomic disorder CNVs**: for genomic disorder analyses presented in Figure 3, we relied on a recent analysis of CNVs in ~one million individuals described in Collins *et al., Cell* (2022). The annotated list of disease-associated CNV loci from this study is available in Table S4 of Collins *et al.* (DOI: 10.1016/j.cell.2022.06.036).

- **Ontario PGP**: we downloaded microarray-based CNV data for the Ontario PGP as described in Uddin *et al., Genet. Med.* (2014) via NCI dbVar accession # estd212 (https://www.ncbi.nlm.nih.gov/dbvar/studies/estd212/).

- **Spit for Science**: we downloaded microarray-based CNV data for the Spit for Science project as described in Zarrei *et al., Hum. Mol. Genet.* (2023) via NCI dbVar accession # nstd224 (https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd224/).

- **1000 Genomes Project (1000G)**: the 1000G phase III SV callset based on low-coverage (7.5X) WGS of 2,504 samples as described in Sudmant *et al., Nature* (2015) is freely available from the International Genome Sample Resource (IGSR) FTP server (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/). The more recent 1000G SV dataset generated on high-coverage (30X) resequencing of these

same samples as described in Byrska-Bishop *et al., Cell* (2022) is also available from the IGSR FTP server (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20210124.SV_Illumina_Integration/).

For datasets that were referenced in this Review but not explicitly mentioned above (such as DiscovEHR, UK Biobank, and deCODE), we extracted the most comparable metrics as reported in their original publications (for example, SV per genome). All summary metrics for Figure 4 are available from GitHub (https://github.com/RCollins13/collins_talkowski_nrg_sv).

Code availability: the code necessary to reproduce the new analyses presented in this Review is available on GitHub (https://github.com/RCollins13/collins_talkowski_nrg_sv).