Database of Genomic Variants

DGV Newsletter February 2020

Hello!

The *Database of Genomic Variants* has recently been updated. In this newsletter, we will give an overview of the data added, and the changes that have been made to the website. The latest updates include three new datasets and we have provided a version of the DGV Gold Standard curated set of variants mapped to the GRCh38 assembly.

New Studies and New Datasets Added to the Database of Genomic Variants

1. Besenbacher_et_al_2015 (estd217)

Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. Besenbacher S, Liu S, Izarzugaza JM, Grove J, Belling K, Bork-Jensen J, Huang S, Als TD, Li S, Yadav R, Rubio-García A, Lescai F, Demontis D, Rao J, Ye W, Mailund T, Friborg RM, Pedersen CN, Xu R, Sun J, Liu H, Wang O, Cheng X, Flores D, Rydza E, Rapacki K, Damm Sørensen J, Chmura P, Westergaard D, Dworzynski P, Sørensen TI, Lund O, Hansen T, Xu X, Li N, Bolund L, Pedersen O, Eiberg H, Krogh A, Børglum AD, Brunak S, Kristiansen K, Schierup MH, Wang J, Gupta R, Villesen P, Rasmussen S. Nat Commun. 2015 Jan 19;6:5969. doi: 10.1038/ncomms6969. PubMed PMID: 25597990; PubMed Central PMCID: PMC4309431.

Building a population-specific catalogue of single nucleotide variants (SNVs), indels and structural variants (SVs) with frequencies, termed a national pan-genome, is critical for further advancing clinical and public health genetics in large cohorts. Here we report a Danish pan-genome obtained from sequencing 10 trios to high depth (50 ×). We report 536k novel SNVs and 283k novel short indels from mapping approaches and develop a population-wide de novo assembly approach to identify 132k novel indels larger than 10 nucleotides with low false discovery rates. We identify a higher proportion of indels and SVs than previous efforts showing the merits of high coverage and de novo assembly approaches. In addition, we use trio information to identify de novo mutations and use a probabilistic method to provide direct estimates of 1.27e-8 and 1.5e-9 per nucleotide per generation for SNVs and indels, respectively.

2. Audano_et_al_2019 (nstd162)

Characterizing the Major Structural Variant Alleles of the Human Genome. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, Warren WC, Magrini V, McGrath SD, Li YI, Wilson RK, Eichler EE. . Cell. 2019 Jan 24;176(3):663-675.e19. doi:10.1016/j.cell.2018.12.019. Epub 2019 Jan 17. PubMed PMID: 30661756; PubMed Central PMCID: PMC6438697.

In order to provide a comprehensive resource for human structural variants (SVs), we generated long-read sequence data and analyzed SVs for fifteen human genomes. We sequence resolved 99,604 insertions, deletions, and inversions including 2,238 (1.6 Mbp) that are shared among all discovery genomes with an additional 13,053 (6.9 Mbp) present in the majority, indicating minor alleles or errors in the reference. Genotyping in 440 additional genomes confirms the most common SVs in unique euchromatin are now sequence resolved. We report a ninefold SV bias toward the last 5 Mbp of human chromosomes with nearly 55% of all VNTRs (variable number of tandem repeats) mapping to this portion of the genome. We identify SVs affecting coding and noncoding regulatory loci improving annotation and interpretation of functional variation. These data provide the framework to construct a canonical human reference and a resource for developing advanced representations capable of capturing allelic diversity.

3. gnomAD SV (nstd166)

An open resource of structural variation for medical and population genetics. Ryan L. Collins, Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C. Francioli, Amit V. Khera, Chelsea Lowther, Laura D. Gauthier, Harold Wang, Nicholas A. Watts, Matthew Solomonson, Anne O'Donnell-Luria, Alexander Baumann, Ruchi Munshi, Mark Walker, Christopher Whelan, Yongqing Huang, Ted Brookings, Ted Sharpe, Matthew R. Stone, Elise Valkanas, Jack Fu, Grace Tiao, Kristen M. Laricchia, Valentin Ruano-Rubio, Christine Stevens, Namrata Gupta, Lauren Margolin, Genome Aggregation Database Production Team, Genome Aggregation Database Consortium, Kent D. Taylor, Henry J. Lin, Stephen S. Rich, Wendy Post, Yii-Der Ida Chen, Jerome I. Rotter, Chad Nusbaum, Anthony Philippakis, Eric Lander, Stacey Gabriel, Benjamin M. Neale, Sekar Kathiresan, Mark J. Daly, Eric Banks, Daniel G. MacArthur, Michael E. Talkowski. https://www.biorxiv.org/content/10.1101/578674v2

Structural variants (SVs) rearrange large segments of the genome and can have profound consequences for evolution and human diseases. As national biobanks, disease association studies, and clinical genetic testing grow increasingly reliant on genome sequencing, population references such as the Genome Aggregation Database (gnomAD) have become integral for interpreting genetic variation. To date, no large-scale reference maps of SVs exist from high-coverage sequencing comparable to those available for point mutations in protein-coding genes. Here, we constructed a reference atlas of SVs across 14,891 genomes from diverse global populations (54% non-European) as a component of gnomAD. We discovered a rich landscape of 433,371 distinct SVs, including 5,295 multi-breakpoint complex SVs across 11

mutational subclasses, and examples of localized chromosome shattering, as in chromothripsis. The average individual harbored 7,439 SVs, which accounted for 25-29% of all rare protein-truncating events per genome. We found strong correlations between constraint against damaging point mutations and rare SVs that both disrupt and duplicate protein-coding sequence, suggesting intolerance to reciprocal dosage alterations for a subset of tightly regulated genes. We also uncovered modest selection against noncoding SVs in cis-regulatory elements, although selection against protein-truncating SVs was stronger than any effect on noncoding SVs. Finally, we benchmarked carrier rates for medically relevant SVs, finding very large (≥1Mb) rare SVs in 3.8% of genomes (~1:26 individuals) and clinically reportable incidental SVs in 0.18% of genomes (~1:556 individuals). These data have been integrated directly into the gnomAD browser (https://gnomad.broadinstitute.org) and will have broad utility for population genetics, disease association, and diagnostic screening.

Summary

If you have any questions or comments, please feel free to contact us by email at dgv-contact@sickkids.ca

Sincerely,

The DGV team