Database of Genomic Variants

DGV Newsletter November 2012

Hello!

The *New Database of Genomic Variants (Beta version)* has recently been updated. In this newsletter, we will give an overview of the data added, and highlight the changes that have been made to the website. The latest updates include four new datasets, and we have incorporated an additional 12 original studies into the new database. We have updated and included new annotations, and implemented a number of modifications and corrections to the existing data to improve the overall functionality of the database. The database is accessible at the following location; http://dgvbeta.tcag.ca/dgv/app/home

This will be the last update of the DGV Beta site, and this version will transition to become the new Database of Genomic Variants following a brief period of review. With this final update, we have incorporated all of the fully curated and accessioned versions of the original studies in addition to 10 new datasets. Once the transition is complete, our new home page will be available at http://dgv.tcag.ca, and links to the original database will be automatically redirected here. There are a number of changes to the content and format, and over the next few weeks, please test the database and send feedback to us so that we can make the required changes before launching the new site. For those researchers or collaborators that have links to the old database, we will keep the underlying database active so that users will have the opportunity to develop new, stable links to the new database, while still maintaining a fully functional system. In addition we will continue to provide a track of the original DGV data in the genome browser (gbrowse) with fully functional links to the variant details page. This will ensure that all data (new and old) will be fully available to users.

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

1. Pinto et al. 2011. Study Accession = estd188

Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants.
Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, MacDonald JR, Mills R, Prasad A, Noonan K, Gribble S, Prigmore E, Donahoe PK, Smith RS,

Park JH, Hurles ME, Carter NP, Lee C, Scherer SW, Feuk L. Nat Biotechnol. 2011 May 8;29(6):512-20. doi: 10.1038/nbt.1852.
The authors systematically compared copy number variant (CNV) detection on eleven microarrays to evaluate data quality and CNV calling, reproducibility, concordance across array platforms and laboratory sites, breakpoint accuracy and analysis tool variability. Different analytic tools applied to the same raw data typically yield CNV calls with <50% concordance. Moreover, reproducibility in replicate experiments is <70% for most platforms. Nevertheless, these findings should not preclude detection of large CNVs for clinical diagnostic purposes because large CNVs with poor reproducibility are found primarily in complex genomic regions and would typically be removed by standard clinical data curation. The striking differences between CNV calls from different platforms and analytic tools highlight the importance of careful assessment of experimental design in discovery and association studies and of strict data curation and filtering in diagnostics. The CNV resource presented here allows independent data evaluation and provides a means to benchmark new algorithms.

2. Campbell et al. 2011. Study Accession = nstd46

Population-genetic properties of differentiated human copy-number polymorphisms.
Campbell CD, Sampas N, Tsalenko A, Sudmant PH, Kidd JM, Malig M, Vu TH, Vives L, Tsang P, Bruhn L, Eichler EE. Am J Hum Genet. 2011 Mar 11;88(3):317-32.

Copy number variants (CNVs) can reach appreciable frequencies in the human population, and several of these copy number polymorphisms (CNPs) have been recently associated with human diseases including lupus, psoriasis, Crohn disease, and obesity. Despite new advances, significant biases remain in terms of CNP discovery and genotyping. We developed a novel method based on single channel intensity data and benchmarked against copy numbers determined from sequencing read-depth to successfully obtain CNP genotypes for 1489 CNPs from 487 human DNA samples from diverse ethnic backgrounds. This customized microarray was enriched for segmental duplication-rich regions and novel insertions of sequences not represented in the reference genome assembly or on standard single nucleotide polymorphism (SNP) microarray platforms. We observe that CNPs in segmental duplications are more likely to be population differentiated than CNPs in unique regions ($p = 0.015$) and that bi-allelic CNPs show greater stratification when compared to frequency-matched SNPs ($p = 0.0026$). Although bi-allelic CNPs show a strong correlation of copy number with flanking SNP genotypes, the majority of multi-copy CNPs do not (40% with $r > 0.8$). We selected a subset of CNPs for further characterization in 1873 additional samples from 62 populations; this revealed striking population-differentiated structural variants in genes of clinical significance such as the OCLN gene, a tight junction protein involved in hepatitis C viral entry. Our new microarray design allows these variants to be rapidly tested for disease association and our results suggest that CNPs (especially those not in linkage disequilibrium with SNPs) may have contributed disproportionately to human diversity and selection.

3. Forsberg et al. 2012. Study Accession = nstd58

Age-related somatic structural changes in the nuclear genome of human blood cells.
Forsberg LA, Rasi C, Razzaghian HR, Pakalapati G, Waite L, Thilbeault KS, Ronowicz A, Wineinger NE, Tiwari HK, Boomsma D, Westerman MP, Harris JR, Lyle R, Essand M, Eriksson F, Assimes TL, Iribarren C, Strachan E, O'Hanlon TP, Rider LG, Miller FW, Giedraitis V, Lannfelt L, Ingelsson M, Piotrowski A, Pedersen NL, Absher D, Dumanski JP.
Am J Hum Genet. 2012 Feb 10;90(2):217-28. Epub 2012 Feb 2

Structural variations are among the most frequent inter-individual genetic differences in the human genome. The frequency and distribution of de novo somatic structural variants in normal cells is, however, poorly explored. Using age-stratified cohorts of 318 monozygotic (MZ) twins and 296 single-born subjects, we describe age-related accumulation of copy-number variation in the nuclear genomes in vivo and frequency changes for both megabase- and kilobase-range variants. Megabase-range aberrations were found in 3.4% (9 of 264) of subjects =60 years old; these subjects included 78 MZ twin pairs and 108 single-born individuals. No such findings were observed in 81 MZ pairs or 180 single-born subjects who were =55 years old. Recurrent region- and gene-specific mutations, mostly deletions, were observed. Longitudinal analyses of 43 subjects whose data were collected 7-19 years apart suggest considerable variation in the rate of accumulation of clones carrying structural changes. Furthermore, the longitudinal analysis of individuals with structural aberrations suggests that there is a natural self-removal of aberrant cell clones from peripheral blood. In three healthy subjects, we detected somatic aberrations characteristic of patients with myelodysplastic syndrome. The recurrent rearrangements uncovered here are candidates for common age-related defects in human blood cells. We anticipate that extension of these results will allow determination of the genetic age of different somatic-cell lineages and estimation of possible individual differences between genetic and chronological age. Our work might also help to explain the cause of an age-related reduction in the number of cell clones in the blood; such a reduction is one of the hallmarks of immuno-senescence.

4. Cooper et al. 2011. Study Accession = nstd54

A copy number variation morbidity map of developmental delay.
Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, Abdel-Hamid H, Bader P, McCracken E, Niyazov D, Leppig K, Thiese H, Hummel M, Alexander N, Gorski J, Kussmann J, Shashi V, Johnson K, Rehder C, Ballif BC, Shaffer LG, Eichler EE. Nat Genet. 2011 Aug 14;43(9):838-46. doi: 10.1038/ng.909.

To understand the genetic heterogeneity underlying developmental delay, we compared copy number variants (CNVs) in 15,767 children with intellectual disability and various congenital defects (cases) to CNVs in 8,329 unaffected adult controls. We estimate that ~14.2% of disease in these children is caused by CNVs >400 kb. We observed a greater enrichment of CNVs in individuals with craniofacial anomalies and cardiovascular defects compared to those with epilepsy or autism. We identified 59 pathogenic CNVs, including 14 new or previously weakly supported candidates, refined the critical interval for several genomic disorders, such as the 17q21.31 microdeletion syndrome, and identified 940 candidate dosage-sensitive genes. We also

developed methods to opportunistically discover small, disruptive CNVs within the large and growing diagnostic array datasets. This evolving CNV morbidity map, combined with exome and genome sequencing, will be critical for deciphering the genetic basis of developmental delay, intellectual disability and autism spectrum disorders.

Personal Genome Variants:

To avoid assigning accessions to the small InDels from this and future studies which have already been submitted (and accessioned) to dbSNP, we have excluded these from the DGV Structural Variants datasets. To ensure that the data is still available and easily accessible, we have provided this information in our newly developed "Personal Genome Variants" track.

1. Kim et al. 2009 Genome Variants (InDels).
   Variants were submitted to dbSNP. The dataset contains over 170,000 indels from 1bp up to 29bp.

Clinically Relevant Genomic Variation

1. ISCA: International Standards for Cytogenomic Arrays Clinical Cytogenetic Testing.

This dataset (nstd37) has been curated and accessioned by staff at dbVar and has recently been updated to include an additional 1,777 variants.

Updates, Modifications and Improvements

We have replaced all of the original internal DGV identifiers with the corresponding accession from the archives. These are stable, universal accessions that will allow for standardized comparisons of structural variation data throughout the community. The accession will be designated as the unique ID in our Downloads Page, and will be fully searchable using the genome browser or the query tool.

For a relatively large number of variants, we did not have the proper variant subtype assigned, and this has been corrected for studies where this information is available. Previously these were designated as unknown (or simply as a CNV) and shown as a black bar in the genome browser. We have gone back to the original studies and identified the original variant subtype and appended this to the record.

The database hardware has been upgraded to improve the speed and stability of the servers.

Data posted on the statistics page has been updated and corrected.

The original DGV variation identifiers are searchable from the new DGV home page. Results will be highlighted and displayed in the DGV1 track in the genome browser.