Database of Genomic Variants

DGV Newsletter April 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 the changes that have been made to the website. The latest updates include several new datasets and annotations, a number of modifications and corrections to the existing data and overall improved functionality.

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

1. Banerjee et al. 2011. Study Accession = estd176

A computational framework discovers new copy number variants with functional importance. Banerjee S, Oldridge D, Poptsova M, Hussain WM, Chakravarty D, Demichelis F. PLoS One. 2011 Mar 29;6(3):e17539.

The authors proposed a three step computational framework (Identification of germline Changes in Copy Number or IgC2N) to discover and genotype germline CNVs. First, they detect candidate CNV loci by combining information across multiple samples without imposing restrictions to the number of coverage markers or to the variant size. They then refine the detection of rare variants and inferred the putative copy number classes for each locus. Last, they combine the relative distance between consecutive copy number classes along with genetic information to estimate the reference model bias. This computational approach was applied to genome-wide data from 1250 HapMap individuals.

Novel variants were discovered and characterized in terms of size, minor allele frequency, type of polymorphism (gains, losses or both), and mechanism of formation. The authors validated the majority of calls, using data previously generated for a subset of these individuals on a 42 million marker platform. They reported that the highest validation rate (66.7%) was for variants of size larger than 1 kb. The results support the validity of the computational framework to detect novel variants relevant to disease susceptibility studies and provide evidence of the importance of genetic variants in regulatory network studies.

Unannotated Studies:
This track will include studies that have not yet been published or archived by our partners at EBI or NCBI. The data will be available in this track until it has been published or fully archived and accessioned.

1. Singapore Genome Variation Project (PubMed ID= 19700652)

This study aims to characterize the extent of common variation in the human genome across at least 1 million single nucleotide polymorphisms (SNPs) for DNA samples from each of the three ethnic groups in Singapore – Chinese, Malays and Indians. The Affymetrix Genome-Wide Human SNP Array 6.0 and the Illumina Human1M single BeadChip were used to assay 292 samples (99 Chinese, 98 Malays and 95 Indians). The inclusion criterion specifies that parents and both sets of grandparents have to belong to the same ethnic group.

2. XueZhang et al, 2011 (unpublished)

The authors used a PCR-based sequencing method to detect deletions mediated by a human-specific palindromic sequence in 740 individuals of different ethnic origins. The data for this study was obtained through a direct submission to dbVar (nstd55).

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.

3. Wheeler et al. Genome Variants (InDels).
   Variants were submitted to dbSNP under handle bcmhgsc_jdw. The dataset contains over 220,000 indels from 2bp up to 40Kb.

Clinically Relevant Genomic Variation

1. DECIPHER: Chromosomal Imbalance and Phenotype in Humans

The DECIPHER database of submicroscopic chromosomal imbalance collects clinical information about chromosomal microdeletions/duplications/insertions, translocations and inversions and displays this information on the human genome map. This track shows genomic regions of reported cases and their associated phenotype information. All links from this data have been deactivated, as access to the underlying records is protected.

2. ISCA: International Standards for Cytogenomic Arrays

Two new datasets have been curated and accessioned by staff at dbVar and made available as separate tracks in our browser.

a) ISCA Curated clinically relevant regions. ISCA annotated clinically relevant regions used in interpretation of cytogenetic testing. Phenotypes include developmental delay and additional significant developmental and morphological phenotypes referred for genetic testing.

b) ISCA Clinical cytogenetic testing. Patients referred for cytogenetic testing due to clinical phenotype. A total of 7605 samples with data available via dbGaP.

Updates, Modifications and Improvements

Training Resources Page

We have provided additional information and resources which will help users navigate and utilize the new database. In addition to the tutorial, we have posted a link to the DGV webinar video and slides as well as links to tips and hints from Open Helix.

Downloads Page

We have included gene annotations for each of the download files, and included the accessions for each record.

Correction of Filtering Steps

We have made some modifications to our filtering process, and this has resulted in the inclusion of many additional insertions that were previously removed. If the size of the inserted sequence is known, the standard size filters are applied. In many cases this information is not recorded by the authors, and if no reported size is available, the variants are kept in the database.

Variant Details Page

The summary page format has been updated and we have included functional links from the reported Gene entries to NCBI to facilitate interoperability.

Genome Browser Tracks

The ABI TaqMan assay track has been made available for hg18 and hg19. We have also reported the date and time when each track was last updated in the track description.