Database of Genomic Variants

DGV Newsletter June 2013

Hello!

We would like to announce the official launch of the new *Database of Genomic Variants*! The latest release includes a number of updates and corrections to the current data, and this completes the transition from the original (http://projects.tcag.ca/variation) and the *Beta* version (http://dgvbeta.tcag.ca/dgv/app/home) to the new *Database of Genomic Variants* (http://dgv.tcag.ca/dgv/app/home).

We will now host only one version of the database, and the original site will be retired. We will continue to provide a track of the original DGV data in the genome browser (gbrowse) which will be searchable and include details from the original variant details page. Any links from third party sites and software which use the "VariationID" to point to the original DGV genome browser or variant details page will be automatically redirected to the corresponding entry in our new database. This will ensure that all data (new and old) will be fully available to users.

We will work with the various partners and websites that provide links to the original data to update the content to reflect the information available in the new site.

With this final update, we have included a total of 53 studies, representing all of the fully curated and accessioned versions of the original studies, in addition to 10 new datasets. There are a number of changes to the content and format, and we have summarized some of these in the newsletter, with additional details in our FAQ and Training Resources pages.

Updates, Modifications and Improvements

We have replaced all of the original internal DGV identifiers with the updated accessions generated at NCBI (dbVAR) and EBI (DGVa). . These are stable, universal accessions that will allow for standardized comparisons of structural variation data throughout the community. The new 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.

Details on the new DGV Accessions:

Each study from DGV has been archived and accessioned by one of the two groups; dbVAR has assigned nsv/nssv accessions, while DGVa has assigned esv/essv accessions. An nsv is an NCBI structural variant, and an nssv is an NCBI supporting structural variant. An esv is an EBI structural variant, and an essv is an EBI supporting structural variant.

Supporting structural variants ("ssv") are typically sample level variants, where each ssv represents the variant called in a single sample/individual. In a few studies the ssv represents the variant called by a single algorithm. If multiple algorithms were used, overlapping ssv's from the same individual would be combined to generate a sample level sv. If there are many samples analysed in a study and if there are many samples which have the same variant, there will be multiple ssv's with the same start and end coordinates. These sample level variants are then merged and combined to form a representative variant that highlights the common variant found in that study. This is called a structural variant ("sv") record.

DGV has always provided this type of summary/merged variant and we have continued to do so in cases where there are a number of overlapping variants that are almost identical, but may be slightly different due to the inherent variability between experiments. If there are clusters of variants within a single study that share at least 70% reciprocal overlap in size/location, we will merge these together and provide an accession record that has our internal "dgv" prefixed identifier. The dgv merged identifier has been updated to help improve the consistency and stability across updates of the database. The format of the dgv accession is now "dgv + variant number + study accession. As an example, the first merged variant from the Shaikh et al 2009 study (study accession=nstd21) would be dgv1n21. The second merged variant would be dgv2n21 and so forth.

Updates to the DGV Downloads File:

We have updated our download files, and have added a number of new columns to address issues and requests made by DGV users. The sample identifiers have been added to each variant record (where available). We have also updated our mergeid column to contain the accessions which link the dgv, sv and ssv records. The number of gains and number of losses have been added to this file as well.

We have updated the content that we provide for download and have generated two distinct files for each version of the reference assembly (hg17, hg18 and hg19). The first file contains the variants that are represented in our primary "DGV Structural Variants" track. The second file contains the underlying supporting variants from our "Supporting Variants" track.
The first file containing the DGV Structural Variants data will be analogous to the data that were made available in the original DGV download file. The supporting variants data are available for users who wish to have a complete set of related data for their studies. The number of entries in this file is quite large, and we hope that by splitting the data into these two distinct files, it will be much clearer and more manageable for the majority of users.

Updates to the content and display of structural variants:

The primary structural variation data in DGV will be displayed in two distinct tracks, which is a change from the original site.

1. The DGV Structural Variants track is synonymous with the original DGV track, where the top level/merged variant regions are displayed.
2. The "Supporting Variants" track contains the underlying sample level and supporting variant calls.
3. The DGV Version 1 track contains a copy of the data from our original site which has been included in the event that users need to go back to compare to analyses that were done on the previous dataset.

Change in Representation of Structural Variant Boundaries

To accurately reflect the inherent differences in the resolution of different approaches, the assignment of boundaries for the structural variants has been updated and the display has been updated.

1. Variant boundaries may be assigned a start and stop position. This will be common for sequencing based studies where the actual breakpoints are known.
2. An outer start and outer stop coordinate will be assigned for studies that use a mapping based strategy (paired-end, optical mapping) or BAC clone approach where the variant boundaries are likely overestimated, and the maximum or outer boundaries are known, but the actual variant likely resides somewhere within this region.
3. The inner start and inner stop coordinates are used for studies where the boundaries are likely underestimated and may include oligo (probe) based CGH experiments. The actual boundary of the variant would likely reside somewhere between the last positive probe on the array and the next neighbouring negative probe.
4. For some studies, a combination of outer start-outer stop and inner start-inner stop coordinates are described when information is available on the boundary regions.

Updates to Variation Classification and Annotations

For a relatively large number of variants, we did not have the proper variant subtype assigned (gain, loss, inversion etc.), 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.

We have updated our variant details pages to capture a number of annotations that were not available in the past.

1. Allele length

a. In many studies, insertions into the reference have been identified, and the size of the variant has been captured and reported. This information is available for a large number of variants in the following studies.
    i. Teague2010 (nstd49)
    ii. Tuzun2005 (nstd1)
    iii. Levy2007 (estd22)
    iv. McKernan2009 (estd197)
    v. Pang2010 (estd180)
    vi. deSmith2007 (estd24)
    vii. Kim2009 (nstd43)
    viii. Kidd2008 (nstd2)
    ix. Mills2006 (nstd6)

2. Allele state
    a. There are a few studies that report if a variant is homozygous or heterozygous, and we have included this where available. Examples of studies where this is available include,
        i. Cooper2008 (nstd14)
        ii. Conrad2006 (nstd17)
        iii. Levy2007 (estd22)
        iv. Pang2010 (estd180)
        v. Shaikh2009 (nstd21)
        vi. Jakobsson (nstd30)

3. Allele origin
    a. A couple of studies have indicated if a specific variant is inherited or *de novo*, and we have included this for the following studies.
        i. Conrad2006 (nstd17)
        ii. Forsberg2012 (nstd58)

4. Copy number
    a. Currently two studies have copy number values for sample level variants and we have reported this information on the variant details pages.
        i. Kim2009 (nstd43)
        ii. Altshuler2010 (estd195)

We have also updated the variant details pages to include information on the frequency of the variant within each study.

1. Number of gains
2. Number of losses
3. Number of complex
4. Number of samples

New and Updated Datasets

OPGP Affymetrix CytoHD Variants

The Ontario Population Genomics Platform (OPGP) was conceived as an opportunity to establish a collection of fully consented, Ontario-based population control DNA samples and corresponding immortalized cell lines, as a resource for researchers in Ontario and elsewhere. Blood-derived DNA from 873 samples was genotyped using the Affymetrix CytoScanHD array and analyzed for copy number variation (CNV). A primary set of 71,178 calls were made by ChAS (Chromosome Analysis Suite) using minimum size and probe cut-offs of 1kb and 8 markers, respectively. Additionally we generated CNV calls using IPN (iPattern), NXS (Nexus) and PTK (Partek) that we used to annotate our primary dataset. Three groups of variants were generated depending on the following criteria. The AFFY_FILTER set of variants represent those defined by Affymetrix as clinically relevant variants. These are losses which are a minimum of 25kb in size and contain a minimum of 25 probes and gains which are a minimum of 50kb and contain a minimum of 50 probes. The MULTI_ALGO set of variants represent those which have been identified using the ChAS algorithm and at least 1 additional algorithm to call the variant (min 1kb, 8 probes). The ONLY_CHAS set of variants represent those which have been identified by the ChAS algorithm only (min 1kb, 8 probes).

DECIPHER: Chromosomal Imbalance and Phenotype in Humans:
The content for the DECIPHER dataset has been updated to reflect the most up-to-date content available.

ISCA Clinical cytogenetic testing
The content for the ISCA clinical cytogenetic testing dataset has also been updated.

Corrections and Re-evaluation of Studies in DGV

Study estd55, Pinto2007:
We have gone back to the authors of this study and retrieved the underlying raw data to accurately represent the content in this study. The authors used multiple algorithms to call structural variants and presented a high-confidence set of calls that were called by 2 or more different algorithms. The ssv records in this study represent algorithm level calls. If two or more essv calls support a variant in a single individual, an esv variant is generated and is annotated as "S" in the query table and "S" in the Merge Status category in the variant details pages. If overlapping esv variants are found in more than one sample, the variant will be tagged as "M" in both the query tool and variant details pages. If an esv variant is supported by a single essv algorithm level call, they are assigned an "is_low_quality" tag and have been filtered. These can be found in our filtered dataset posted on the download page of the website.

Corrections to Study estd24 (deSmith2007):
It was brought to our attention that there were several variants mapped to NCBI35 (hg17) in the download file that extended past the end of the chromosome. This error was a result of the assignment of the end positions for variants as the midpoint between the last positive probe and the neighbouring probe. When the last positive probe was the last probe on the chromosome, the

next (neighbouring) probe position was arbitrarily set at 1MB downstream, causing an error in correctly assigning the end position of the variant.

Summary

We hope that the transition to the new site is easy and straightforward. If you have any questions or comments, please feel free to contact us by email at dgv-contact@sickkids.ca

Sincerely,

The DGV team