Database of Genomic Variants

DGV Newsletter October 2011

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 the launch of an entirely new database, genome browser and query tool. The new database now contains the fully accessioned data which has been synchronized with our partner databases at EBI (DGVa), and NCBI (dbVar), and is available at this link http://dgvbeta.tcag.ca/dgv/app/home. Throughout the process we followed the consistent rule of maintaining simplicity in presenting the data.

DGV2 Beta Version released

The Database of Genomic Variants (DGV) has been working in partnership with the new database archives (DGVa and dbVar). Moving forward, DGV will use them as the primary source of structural variation information, ensuring that the data is synchronized with the accessioned structural variations contained in these archives. The main role of DGV going forward will be to curate and visualize selected studies to facilitate interpretation of structural variation data, including implementing the highest-level quality standards required by the clinical and diagnostic communities. The complex nature of structural variation often makes it less amenable than other genomic data (for example, SNPs) to single-step electronic mapping to reference genomes, necessitating significant manual curation efforts, a critical contribution which DGV will continue to perform. For more information about this collaboration, please refer to the publication in Nature Genetics (PubMedID: 20877315).

The Database of Genomic Variants has launched a Beta (test) version of the data currently contained in the DGVa and dbVar archives. The database is being launched as a Beta version, while we continue to refine and improve the data and the presentation to ensure that a complete and highly accurate dataset is available for our users. In the meantime, the current, full version of DGV will still be active (http://projects.tcag.ca/variation). This will ensure continuity and complete availability of the resources users require for their studies. We encourage everyone to try, test and utilize some of the new and interactive features of the database and provide feedback to the DGV team so we can continue to improve the new version which we hope to launch in production mode within the next few months. In total there are 37 studies represented in the new database, eight of which are new. In addition, we have included the original DGV data as a separate track in the new database for comparison, and to provide users the option to view all available data in a single location.

In addition to the new data, a new version of the genome browser has been made available which has many new and useful tools to help users navigate the data. We have also redeveloped the underlying database schema, allowing us to capture many additional data types and descriptions, and have integrated this to allow more complex queries to be submitted by users. A new Query Tool interface has been developed that will allow users to select and filter the database across studies to extract out features, regions or items of interest. For example, if a user was interested in obtaining all studies that sampled the common reference NA15510, they could filter the studies by this name. The query tool has also been integrated with the genome browser, so users may filter and view only variants that meet specific requirements. We have spent some time developing methods to integrate and resolve sample duplicates and have preloaded all common sample cohorts such as the HapMap and Human Genome Diversity Panel samples, which will allow users to obtain non-redundant records for commonly used samples.

We have prepared a tutorial that is available from the home page that will help users navigate the new features and components of the database, and as always we are available for assistance by email at: dgv-contact@sickkids.ca

Genome Browser Update

A new version of Gbrowse (version 2.39) has been implemented. Some of the new functions associated with this update include the ability to click and drag tracks, rearrange the order, click/drag and zoom or re-centre the view. The underlying data (DGV or Annotations) can be extracted for the region that you are viewing or for the entire chromosome/genome and downloaded as a text file. A filter has been placed, linking our new query tool to the DGV variants data, allowing users to filter which variants are displayed in the browser. For example, if a user only wants to display variants from a particular study or a particular sample (across studies), this function now exists.

Query Tool

A new database was created, and many additional fields have been captured and stored. A query tool has been developed that will allow users to query, extract, filter and search for variants that fit a set of criteria they are interested in. This functionality will allow for a much more flexible and interactive interface for the data contained in DGV.

New Studies Added to the Database of Genomic Variants

Eight new studies have been added to DGV. These studies have been collected and archived by either staff at dbVar (NCBI) or DGVa (EBI). This includes data from pilot 1 and 2 of the 1,000 genomes project (Durbin et al, 2010). A summary of the new datasets is included below.

High-resolution human genome structure by single-molecule analysis.
Teague B, Waterman MS, Goldstein S, Potamousis K, Zhou S, Reslewic S, Sarkar D, Valouev A, Churas C, Kidd JM, Kohn S, Runnheim R, Lamers C, Forrest D, Newton MA, Eichler EE, Kent-First M, Surti U, Livny M, Schwartz DC. Proc Natl Acad Sci U S A. 2010 Jun 15;107(24):10848-53. Epub 2010 Jun 1. PubMed PMID: 20534489

The authors used optical mapping to identify thousands of structural variants from kb to Mb in size. A genome-wide restriction map was generated for a complete hydatidiform mole and three lymphoblast-derived cell lines (GM15510, GM10860, GM18994). The process was validated by demonstrating a strong concordance with existing methods. The authors describe the process, "Optical Mapping is a high-throughput system that combines single-molecule measurements with dedicated computational analysis to produce ordered restriction maps from individual molecules of genomic DNA: essentially, a single-molecule realization of traditional restriction fragment length polymorphism mapping. Each single-molecule restriction map is a direct measurement of the source genome, free from biases introduced by cloning, amplification, or hybridization."

A human genome structural variation sequencing resource reveals insights into mutational mechanisms
Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, Eichler EE. Cell. 2010 Nov 24;143(5):837-47. PubMed PMID: 21111241

In order to understand the mutational mechanisms responsible for structural variation, the authors analysed the genomes of 17 individuals by developing a resource based on capillary end sequencing of 13.8 million fosmid clones. The complete sequence of 1054 large structural variants was ascertained, and the breakpoint junctions were analysed to infer the potential mechanism of origin.

Characterization of missing human genome sequences and copy-number polymorphic insertions.
Kidd JM, Sampas N, Antonacci F, Graves T, Fulton R, Hayden HS, Alkan C, Malig M, Ventura M, Giannuzzi G, Kallicki J, Anderson P, Tsalenko A, Yamada NA, Tsang P, Kaul R, Wilson RK, Bruhn L, Eichler EE. Nat Methods. 2010 May;7(5):365-71. PubMed PMID: 20440878

The goal of this study was to identify and characterize novel sequences that were not represented in the human genome assembly. Using 9.7 million fosmid end-sequence pairs from 9 individuals (4 YRI, 2 CEU, 2 CHB and NA15510), the authors identified clones where only one end mapped to the reference, while the other end did not. They also searched for clones where neither end mapped even though high quality fosmid end sequence was available. This method resulted in the identification of 2,363 novel insertion sequences which correspond to 720 distinct genomic loci. They also determined that approximately 18-37% of these are copy number polymorphic. Complete sequencing of 156 insertions identified novel exons and conserved non-coding sequence in the human genome.

Complete Khoisan and Bantu genomes from southern Africa
Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, Harris RS, Petersen DC, Zhao F, Qi J, Alkan C, Kidd JM, Sun Y, Drautz DI, Bouffard P, Muzny DM, Reid JG, Nazareth LV, Wang Q, Burhans R, Riemer C, Wittekindt NE, Moorjani P, Tindall EA, Danko CG, Teo WS, Buboltz AM, Zhang Z, Ma Q, Oosthuysen A, Steenkamp AW, Oostuisen H, Venter P, Gajewski J, Zhang Y, Pugh BF, Makova KD, Nekrutenko A, Mardis ER, Patterson N, Pringle TH, Chiaromonte F, Mullikin JC, Eichler EE, Hardison RC, Gibbs RA, Harkins TT, Hayes VM. Nature. 2010 Feb 18;463(7283):943-7. PubMed PMID: 20164927

The complete genome sequences of an indigenous hunter-gatherer (Khoisan) from the Kalahari Desert and a Bantu from southern Africa was generated using next generation sequencing approaches. The Khoisan individual (KB1), was sequenced to 10.2 fold coverage using the Roche/454 GS FLX Titanium chemistry. An additional long insert library (up to 17kb) was constructed for KBI, and sequenced using the same platform to 12.3 fold non-redundant clone coverage. An additional Khoisan individual, NH1 was sequenced to 2 fold coverage using the same platform. The Bantu individual (Archbishop Desmond Tutu, ABT) was sequenced to approximately 30 fold non-redundant clone coverage using Applied Biosystems SOLiD 3 platform. The protein coding regions of samples ABT, KB1 plus an additional 3 individuals (NB1, TK1, MD8; Khoisan) were targeted using a NimbleGen 2.1M capture array, with subsequent sequencing by Roche/454 GS FLX Titanium to at least 16 fold coverage.

Comparison of constitutional and replication stress-induced genome structural variation by SNP array and mate-pair sequencing.
Arlt MF, Ozdemir AC, Birkeland SR, Lyons RH Jr, Glover TW, Wilson TE. Genetics. 2011 Mar;187(3):675-83. Epub 2011 Jan 6. PubMed PMID: 21212237

To address the impact that replication stress may have as a causative factor in CNV formation, the authors utilized two high resolution approaches to compare CNVs that occur constitutionally to those that arise following aphidicolin-induced DNA replication stress in the same human cells. Both single nucleotide polymorphism (SNP) arrays (Illumina HumanOmni1-Quad BeadChip) and mate-pair sequencing (Illumina) was used to call CNVs. The results were compared and the authors determined that the majority of constitutional and all of the aphidicolin-induced CNVs appear to have been formed by homology-independent mechanisms. Aphidicolin-induced CNVs were larger, and would seem to resemble human pathogenic CNVs and the subset of larger nonhomologous constitutional CNVs.

Copy number variation and evolution in humans and chimpanzees.
Perry GH, Yang F, Marques-Bonet T, Murphy C, Fitzgerald T, Lee AS, Hyland C, Stone AC, Hurles ME, Tyler-Smith C, Eichler EE, Carter NP, Lee C, Redon R. Genome Res. 2008 Nov;18(11):1698-710. Epub 2008 Sep 4. PubMed PMID: 18775914

The aims of this study were to improve the understanding of the evolutionary significance of CNVs and how they impact human phenotypic diversity by providing the raw material for gene duplication and gene family expansion. An updated version of the WGTP CGH array from Redon et al 2006 was performed on 30 human and 30 chimpanzee samples. The updated array included more than 2000 new clones added to target gaps in the previous version. The authors detected over 350 discrete autosomal CNV regions in the human samples and the number of common CNVs detected (found in 2 or more samples) was comparable between human and chimpanzee groups (only human CNVs were added to DGV). The authors revealed that comparing patterns of variation within and between species provided important insights into the selective forces and mutational mechanisms which contribute to this class of genetic diversity.

XueZhang et al, 2001
Article in Press: 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).

A map of human genome variation from population-scale sequencing.
1000 Genomes Project Consortium. Durbin et. al. Nature. 2010 Oct 28;467(7319):1061-73. PubMed PMID: 20981092

This study from the 1000 Genomes Project Consortium contains all the structural variations detected in the first phase of the project. The types of variants identified include deletions, tandem duplications, novel sequences and mobile element insertions. Data from both Pilot 1 and Pilot 2 are included and calls from all chromosomes are represented. The structural variations for Pilot 1 were identified using low coverage re-sequencing of the HapMap samples, while variations in Pilot 2 were identified using high coverage re-sequencings of two HapMap trios (CEU and YRI).

As the field of structural variation has grown and developed we wanted to ensure that the Database of Genomic Variants has evolved and grown as well. The quality and quantity of data has increased exponentially over the past few years and as a result there is a significant increase in the need to have a complete, accurate and integrated database. By generating a new interface (query tool), we hope to make the data more easily accessible and interactive for all users. We will continue to work on curating and assessing the data to provide a high quality, comprehensive dataset for all users.