Database of Genomic Variants

DGV Newsletter October 2014

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 eight new datasets and we have updated the genome browser to display the newest reference assembly, GRCh38. A number of modifications and corrections to the existing data and overall improved functionality have also been completed.

Updates, Modifications and Improvements

1. We have updated the database to include the newest reference genome assembly, GRCh38/hg38. Variants are included in our query tool, and are present in our genome browser. The number of annotation tracks is limited at the moment, but we will continually update these as they become available.

Corrections and Re-evaluation of Studies in DGV

Study nstd71, Xu2011:
We have temporarily removed this study based on a reanalysis of the data and feedback from numerous DGV users, indicating a likely high false positive rate in this dataset.

Study estd59, 1000 Genomes Pilot Phase:
With the release of the Phase 1 data from the 1,000 Genomes Project, the deletions from the Pilot phase were removed in the last update, as the underlying data was reanalysed and improved detection methods were employed. We have now processed all variants (including deletions) through our pipeline and the deletions are now available in our Downloads Page in the Filtered Variants section.

Study nstd54, Cooper2011:
The dataset has been reanalysed and we are now able to identify variants found only in control samples and these have now been added into DGV.

Study *estd20 Conrad2009:*
The number of samples reported in our Query Tool has been changed to 40 (from 451) to reflect the actual number of samples tested, vs. the number of samples used in the validation experiments.

Study nstd41 *Iafrate2004:*
The number of samples reported in our Query Tool and variant details pages have been corrected to reflect the actual number of samples tested (n=39).

Study nstd46 *Campbell2011:*
The underlying sample level calls (with sample identifiers) are now available for this study.

Study nstd22 *McCarroll2008:*
The underlying sample level calls (with sample identifiers) are now available for this study.

Study estd1 *Redon2006:*
The underlying data has been reanalysed, and the relationship between the variant calls and variant regions have been corrected. As a result, the proper variant subtype is annotated in the database.

Study estd49 *Gusev2009:*
The underlying data has been reanalysed, and the relationship between the variant calls and variant regions have been corrected. As a result, the proper variant subtype is annotated in the database.

Study nstd29 *Locke2006:*
The underlying sample level calls were updated with corrected versions now available for this study.

Study nstd23 *Young2008:*
The underlying data has been reanalysed, and the relationship between the variant calls and variant regions have been corrected. As a result, the complete set of variant calls is available with the sample level information annotated to each region.

Study nstd35 *Kidd2010:*

Miscellaneous: Variant region nsv519796 and its supporting variants nssv689256 and nssv658757 were removed, as these could not be validated experimentally.

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

1. Vogler 2010. Study Accession = estd203

Microarray-based maps of copy-number variant regions in European and sub-Saharan populations. Vogler C, Gschwind L, Röthlisberger B, Huber A, Filges I, Miny P, Auschra B, Stetak A, Demougin P, Vukojevic V, Kolassa IT, Elbert T, de Quervain DJ, Papassotiropoulos A. PLoS One. 2010 Dec 16;5(12):e15246. PubMed PMID: 21179565.

The genetic basis of phenotypic variation can be partially explained by the presence of copy-number variations (CNVs). Currently available methods for CNV assessment include high-density single-nucleotide polymorphism (SNP) microarrays that have become an indispensable tool in genome-wide association studies (GWAS). However, insufficient concordance rates between different CNV assessment methods call for cautious interpretation of results from CNV-based genetic association studies. Here we provide a cross-population, microarray-based map of copy-number variant regions (CNVRs) to enable reliable interpretation of CNV association findings. We used the Affymetrix Genome-Wide Human SNP Array 6.0 to scan the genomes of 1167 individuals from two ethnically distinct populations (Europe, N=717; Rwanda, N=450). Three different CNV-finding algorithms were tested and compared for sensitivity, specificity, and feasibility. Two algorithms were subsequently used to construct CNVR maps, which were also validated by processing subsamples with additional microarray platforms (Illumina 1M-Duo BeadChip, Nimblegen 385K aCGH array) and by comparing our data with publicly available information. Both algorithms detected a total of 42669 CNVs, 74% of which clustered in 385 CNVRs of a cross-population map. These CNVRs overlap with 862 annotated genes and account for approximately 3.3% of the haploid human genome.We created comprehensive cross-populational CNVR-maps. They represent an extendable framework that can leverage the detection of common CNVs and additionally assist in interpreting CNV-based association studies.

2. Chia2012. Study Accession = estd198

High-resolution SNP microarray investigation of copy number variations on chromosome 18 in a control cohort. Chia NL, Bryce M, Hickman PE, Potter JM, Glasgow N, Koerbin G, Danoy P, Brown MA, Cavanaugh J. Cytogenet Genome Res. 2013;141(1):16-25. doi: 10.1159/000350767. Epub 2013 Apr 26. PubMed PMID: 23635498.

Copy number variations (CNVs) as described in the healthy population are purported to contribute significantly to genetic heterogeneity. Recent studies have described CNVs using

lymphoblastoid cell lines or by application of specifically developed algorithms to interrogate previously described data. However, the full extent of CNVs remains unclear. Using high-density SNP array, we have undertaken a comprehensive investigation of chromosome 18 for CNV discovery and characterisation of distribution and association with chromosome architecture. We identified 399 CNVs, of which loss represents 98%, 58% are less than 2.5 kb in size and 71% are intergenic. Intronic deletions account for the majority of copy number changes with gene involvement. Furthermore, one-third of CNVs do not have putative breakpoints within repetitive sequences. We conclude that replicative processes, mediated either by repetitive elements or microhomology, account for the majority of CNVs in the healthy population. Genomic instability involving the formation of a non-B structure is demonstrated
in one region.

3. Pang2013b. Study Accession = estd209

Performance of high-throughput sequencing for the discovery of genetic variation across the complete size spectrum. Pang AW, Macdonald JR, Yuen RK, Hayes VM, Scherer SW. G3 (Bethesda). 2014 Jan 10;4(1):63-5. PubMed PMID: 24192839.

We observed that current high-throughput sequencing approaches only detected a fraction of the full size-spectrum of insertions, deletions, and copy number variants compared with a previously published, Sanger-sequenced human genome. The sensitivity for detection was the lowest in the 100- to 10,000-bp size range, and at DNA repeats, with copy number gains harder to delineate than losses. We discuss strategies for discovering the full spectrum of genetic variation necessary for disease association studies.

4. Schrider2013. Study Accession = nstd78

Gene copy-number polymorphism caused by retrotransposition in humans. Schrider DR, Navarro FC, Galante PA, Parmigiani RB, Camargo AA, Hahn MW, de Souza SJ. PLoS Genet. 2013;9(1):e1003242. Epub 2013 Jan 24. PubMed PMID: 23359205.

The era of whole-genome sequencing has revealed that gene copy-number changes caused by duplication and deletion events have important evolutionary,functional, and phenotypic consequences. Recent studies have therefore focused on revealing the extent of variation in copy-number within natural populations of humans and other species. These studies have found a large number of copy-number variants (CNVs) in humans, many of which have been shown to have clinical or evolutionary importance. For the most part, these studies have failed to detect an important class of gene copy-number polymorphism: gene duplications caused by retrotransposition, which result in a new intron-less copy of the parental gene being inserted into a random location in the genome. Here we describe a computational approach leveraging next-generation sequence data to detect gene copy-number variants caused by retrotransposition (retroCNVs), and we report the first genome-wide analysis of these variants in humans. We find that retroCNVs account for a substantial fraction of gene copy-number differences between any two individuals. Moreover, we show that these variants may often result in expressed chimeric transcripts, underscoring their potential for the evolution of novel gene functions. By locating the insertion sites of these duplicates, we are able to show that retroCNVs have had an important role in recent human adaptation, and we also uncover evidence that positive selection may currently be driving multiple retroCNVs toward fixation. Together these findings imply that retroCNVs are an especially important class of polymorphism, and that future studies of copy-number variation should search for these variants in order to illuminate their potential evolutionary and functional relevance.

5. Cooper2011. 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. PubMed PMID: 21841781;

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.

6. Sudmant2013. Study Accession = nstd82

Evolution and diversity of copy number variation in the great ape lineage. Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, Antonacci F, Ventura M, Prado-Martinez J; Great Ape Genome Project, Marques-Bonet T, Eichler EE. Genome Res. 2013 Sep;23(9):1373-82. PubMed PMID:23825009.

Copy number variation (CNV) contributes to disease and has restructured the genomes of great apes. The diversity and rate of this process, however, have not been extensively explored among great ape lineages. We analyzed 97 deeply sequenced great ape and human genomes and estimate 16% (469 Mb) of the hominid genome has been affected by recent CNV. We identify a comprehensive set of fixed gene deletions (n = 340) and duplications (n = 405) as well as >13.5 Mb of sequence that has been specifically lost on the human lineage. We compared the diversity and rates of copy number and single nucleotide variation across the hominid phylogeny. We find that CNV diversity partially correlates with single nucleotide diversity (r(2) = 0.5) and recapitulates the phylogeny of apes with few exceptions. Duplications significantly outpace deletions (2.8-fold). The load of segregating duplications remains significantly higher in bonobos, Western chimpanzees, and Sumatran orangutans-populations that have experienced recent genetic bottlenecks (P = 0.0014, 0.02, and 0.0088, respectively). The rate of fixed deletion has been more clocklike with the exception of the chimpanzee lineage, where we observe a twofold increase in the chimpanzee-bonobo ancestor (P = 4.79 × 10(-9)) and increased deletion load among Western chimpanzees (P = 0.002). The latter includes the first genomic disorder in a chimpanzee with features resembling Smith-Magenis syndrome mediated by a chimpanzee-specific increase in segmental duplication complexity. We hypothesize that demographic effects, such as bottlenecks, have contributed to larger and more gene-rich segments being deleted in the chimpanzee lineage and that this effect, more generally, may account for episodic bursts in CNV during hominid evolution.

7. Dogan2014. Study Accession = nstd73

Whole genome sequence of a Turkish individual. Dogan H, Can H, Otu HH. PLoS One. 2014 Jan 9;9(1):e85233. eCollection 2014. PubMed PMID: 24416366.

Although whole human genome sequencing can be done with readily available technical and financial resources, the need for detailed analyses of genomes of certain populations still exists. Here we present, for the first time, sequencing and analysis of a Turkish human genome. We have performed 35x coverage using paired-end sequencing, where over 95% of sequencing reads are mapped to the reference genome covering more than 99% of the bases. The assembly of unmapped reads rendered 11,654 contigs, 2,168 of which did not reveal any homology to known sequences, resulting in ~1 Mbp of unmapped sequence. Single nucleotide polymorphism (SNP) discovery resulted in 3,537,794 SNP calls with 29,184 SNPs identified in coding regions, where 106 were nonsense and 259 were categorized as having a high-impact effect. The homo/hetero zygosity (1,415,123:2,122,671 or 1:1.5) and transition/transversion ratios (2,383,204:1,154,590 or 2.06:1) were within expected limits. Of the identified SNPs, 480,396 were potentially novel with 2,925 in coding regions, including 48 nonsense and 95 high-impact SNPs. Functional analysis of novel high-impact SNPs revealed various interaction networks, notably involving hereditary and neurological disorders or diseases. Assembly results indicated 713,640 indels (1:1.09 insertion/deletion ratio), ranging from -52 bp to 34 bp in length and causing about 180 codon insertion/deletions and 246 frame shifts. Using paired-end- and read-depth-based methods, we discovered 9,109 structural variants and compared our variant findings with other populations. Our results suggest that whole genome sequencing is a valuable tool for understanding variations in the human genome across different populations. Detailed analyses of genomes of diverse origins greatly benefits research in genetics and medicine and should be conducted on a larger scale.

8. Watson2013. Study Accession = nstd76

Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, Willsey AJ, Joy JB, Scott JK, Graves TA, Wilson RK, Holt RA, Eichler EE, Breden F. Am J Hum Genet. 2013 Apr 4;92(4):530-46. Epub 2013 Mar 28. PubMed PMID: 23541343.

The immunoglobulin heavy-chain locus (IGH) encodes variable (IGHV), diversity (IGHD), joining (IGHJ), and constant (IGHC) genes and is responsible for antibody heavy-chain biosynthesis, which is vital to the adaptive immune response. Programmed V-(D)-J somatic rearrangement and the complex duplicated nature of the locus have impeded attempts to reconcile its genomic organization based on traditional B-lymphocyte derived genetic material. As a result, sequence descriptions of germline variation within IGHV are lacking, haplotype inference using traditional linkage disequilibrium methods has been difficult, and the human genome reference assembly is missing several expressed IGHV genes. By using a hydatidiform mole BAC clone resource, we present the most complete haplotype of IGHV, IGHD, and IGHJ gene regions derived from a single chromosome, representing an alternate assembly of ~1 Mbp of high-quality finished sequence. From this we add 101 kbp of previously uncharacterized sequence, including functional IGHV genes, and characterize four large germline copy-number variants (CNVs). In addition to this germline reference, we identify and characterize eight CNV-containing haplotypes from a panel of nine diploid genomes of diverse ethnic origin, discovering previously unmapped IGHV genes and an additional 121 kbp of insertion sequence. We genotype four of these CNVs by using PCR in 425 individuals from nine human populations. We find that all four are highly polymorphic and show considerable evidence of stratification (Fst = 0.3-0.5), with the greatest differences observed between African and Asian populations. These CNVs exhibit weak linkage disequilibrium with SNPs from two commercial arrays in most of the populations tested.

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