Database of Genomic Variants

DGV Newsletter July 2015

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 five new datasets and we have included a newly developed DGV curated set of variants that were reported in a recent issue of *Nature Reviews Genetics*.

Updates, Modifications and Improvements

1. We have updated the database to include the newly developed DGV curated set of variants. This is currently posted as a separate track in the Genome Browser. Two independent datasets are included for the GRCh37 version of the assembly. The first dataset "Nat. Rev. Gen. CNV Map – Inclusive" includes a set of variants found in one or more samples in at least 2 different studies. The second dataset "Nat. Rev. Gen. CNV Map – Stringent" contains variants that are found in 2 or more unique samples in at least 2 or more unique studies. Please see the article below for additional details.

A copy number variation map of the human genome. Zarrei M, MacDonald JR, Merico D, Scherer SW. *Nat Rev Genet*. 2015 Mar;16(3):172-83. doi: 10.1038/nrg3871. Epub 2015 Feb 3

A major contribution to the genome variability among individuals comes from deletions and duplications - collectively termed copy number variations (CNVs) - which alter the diploid status of DNA. These alterations may have no phenotypic effect, account for adaptive traits or can underlie disease. We have compiled published high-quality data on healthy individuals of various ethnicities to construct an updated CNV map of the human genome. Depending on the level of stringency of the map, we estimated that 4.8-9.5% of the genome contributes to CNV and found approximately 100 genes that can be completely deleted without producing apparent phenotypic consequences. This map will aid the interpretation of new CNV findings for both clinical and research applications.

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

1. Uddin2014. Study Accession = estd212

A high-resolution copy-number variation resource for clinical and population genetics. Uddin M, Thiruvahindrapuram B, Walker S, Wang Z, Hu P, Lamoureux S, Wei J, MacDonald JR, Pellecchia G, Lu C, Lionel AC, Gazzellone MJ, McLaughlin JR, Brown C, Andrulis IL, Knight JA, Herbrick JA, Wintle RF, Ray P, Stavropoulos DJ, Marshall CR, Scherer SW. Genet Med. 2014 Dec 11. doi: 10.1038/gim.2014.178.

Purpose: Chromosomal microarray analysis to assess copy-number variation has become a first-tier genetic diagnostic test for individuals with unexplained neurodevelopmental disorders or multiple congenital anomalies. More than 100 cytogenetic laboratories worldwide use the new ultra-high resolution Affymetrix CytoScan-HD array to genotype hundreds of thousands of samples per year. Our aim was to develop a copy-number variation resource from a new population sample that would enable more accurate interpretation of clinical genetics data on this microarray platform and others.Methods:Genotyping of 1,000 adult volunteers who are broadly representative of the Ontario population (as obtained from the Ontario Population Genomics Platform) was performed with the CytoScan-HD microarray system, which has 2.7 million probes. Four independent algorithms were applied to detect copy-number variations. Reproducibility and validation metrics were quantified using sample replicates and quantitative-polymerase chain reaction, respectively.Results:DNA from 873 individuals passed quality control and we identified 71,178 copy-number variations (81 copy-number variations/individual); 9.8% (6,984) of these copy-number variations were previously unreported. After applying three layers of filtering criteria, from our highest confidence copy-number variation data set we obtained >95% reproducibility and >90% validation rates (73% of these copy-number variations overlapped at least one gene).Conclusion:The genotype data and annotated copy-number variations for this largely Caucasian population will represent a valuable public resource enabling clinical genetics research and diagnostics.

2. Boomsma2014. Study Accession = estd215

The Genome of the Netherlands: design, and project goals.
Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, Ye K, Guryev V, Vermaat M, van Dijk F, Francioli LC, Hottenga JJ, Laros JF, Li Q, Li Y, Cao H, Chen R, Du Y, Li N, Cao S, van Setten J, Menelaou A, Pulit SL, Hehir-Kwa JY, Beekman M, Elbers CC, Byelas H, de Craen AJ, Deelen P, Dijkstra M, den Dunnen JT, de Knijff P, Houwing-Duistermaat J, Koval V, Estrada K, Hofman A, Kanterakis A, Enckevort Dv, Mai H, Kattenberg M, van Leeuwen EM, Neerincx PB, Oostra B, Rivadeneira F, Suchiman EH, Uitterlinden AG, Willemsen , Wolffenbuttel

BH, Wang J, de Bakker P, van Ommen GJ, van Duijn CM.Eur J Hum Genet. 2014 Feb;22(2):221-7. doi: 10.1038/ejhg.2013.118. Epub 2013 May 29.

Within the Netherlands a national network of biobanks has been established (Biobanking and Biomolecular Research Infrastructure-Netherlands (BBMRI-NL)) as a national node of the European BBMRI. One of the aims of BBMRI-NL is to enrich biobanks with different types of molecular and phenotype data. Here, we describe the Genome of the Netherlands (GoNL), one of the projects within BBMRI-NL. GoNL is a whole-genome-sequencing project in a representative sample consisting of 250 trio-families from all provinces in the Netherlands, which aims to characterize DNA sequence variation in the Dutch population. The parent-offspring trios include adult individuals ranging in age from 19 to 87 years (mean=53 years; SD=16 years) from birth cohorts 1910-1994. Sequencing was done on blood-derived DNA from uncultured cells and accomplished coverage was 14-15x. The family-based design represents a unique resource to assess the frequency of regional variants, accurately reconstruct haplotypes by family-based phasing, characterize short indels and complex structural variants, and establish the rate of de novo mutational events. GoNL will also serve as a reference panel for imputation in the available genome-wide association studies in Dutch and other cohorts to refine association signals and uncover population-specific variants. GoNL will create a catalog of human genetic variation in this sample that is uniquely characterized with respect to micro-geographic location and a wide range of phenotypes. The resource will be made available to the research and medical community to guide the interpretation of sequencing projects. The present paper summarizes the global characteristics of the project.

3. 1000 Genomes Phase 3. Study Accession = estd214

Submitter: Laura Clark; Project: PRJEB6930; Submitter URL: http://www.1000genomes.org/

This study contains the structural variants from the combined release set which contains more than 79 million variant sites and includes not just biallelic snps but also indels, deletions, complex short substitutions and other structural variant classes. It is based on data from 2504 unrelated individuals from 26 different populations around the world.

4. Coe2014. Study Accession = nstd100

Refining analyses of copy number variation identifies specific genes associated with developmental delay. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BW, Vulto-van Silfhout AT, Bosco P, Friend KL, Baker C, Buono S, Vissers LE, Schuurs-Hoeijmakers JH, Hoischen A, Pfundt R, Krumm N, Carvill GL, Li D, Amaral D, Brown N, Lockhart PJ, Scheffer IE, Alberti A, Shaw M, Pettinato R, Tervo R, de Leeuw N, Reijnders MR, Torchia BS, Peeters H, O'Roak BJ, Fichera M, Hehir-Kwa JY, Shendure J, Mefford HC, Haan E, Gécz J, de Vries BB, Romano C, Eichler EE. Nat Genet. 2014 Oct;46(10):1063-71. doi: 10.1038/ng.3092. Epub 2014 Sep 14

Copy number variants (CNVs) are associated with many neurocognitive disorders; however, these events are typically large, and the underlying causative genes are unclear. We created an

expanded CNV morbidity map from 29,085 children with developmental delay in comparison to 19,584 healthy controls, identifying 70 significant CNVs. We resequenced 26 candidate genes in 4,716 additional cases with developmental delay or autism and 2,193 controls. An integrated analysis of CNV and single-nucleotide variant (SNV) data pinpointed 10 genes enriched for putative loss of function. Follow-up of a subset of affected individuals identified new clinical subtypes of pediatric disease and the genes responsible for disease-associated CNVs. These genetic changes include haploinsufficiency of SETBP1 associated with intellectual disability and loss of expressive language and truncations of ZMYND11 in individuals with autism, aggression and complex neuropsychiatric features. This combined CNV and SNV approach facilitates the rapid discovery of new syndromes and genes involved in neuropsychiatric disease despite extensive genetic heterogeneity.

5. Mokhtar2014. Study Accession = estd213

Novel population specific autosomal copy number variation and its functional analysis amongst Negritos from Peninsular Malaysia. Mokhtar SS, Marshall CR, Phipps ME, Thiruvahindrapuram B, Lionel AC, Scherer SW, Peng HB. PLoS One. 2014 Jun 23;9(6):e100371. doi: 10.1371/journal.pone.0100371. eCollection 2014.

Copy number variation (CNV) has been recognized as a major contributor to human genome diversity. It plays an important role in determining phenotypes and has been associated with a number of common and complex diseases. However CNV data from diverse populations is still limited. Here we report the first investigation of CNV in the indigenous populations from Peninsular Malaysia. We genotyped 34 Negrito genomes from Peninsular Malaysia using the Affymetrix SNP 6.0 microarray and identified 48 putative novel CNVs, consisting of 24 gains and 24 losses, of which 5 were identified in at least 2 unrelated samples. These CNVs appear unique to the Negrito population and were absent in the DGV, HapMap3 and Singapore Genome Variation Project (SGVP) datasets. Analysis of gene ontology revealed that genes within these CNVs were enriched in the immune system (GO:0002376), response to stimulus mechanisms (GO:0050896), the metabolic pathways (GO:0001852), as well as regulation of transcription (GO:0006355). Copy number gains in CNV regions (CNVRs) enriched with genes were significantly higher than the losses (P value <0.001). In view of the small population size, relative isolation and semi-nomadic lifestyles of this community, we speculate that these CNVs may be attributed to recent local adaptation of Negritos from Peninsular Malaysia.

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