Database of Genomic Variants

DGV Newsletter July 2013

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 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. 1000 Genomes Phase 1. Study Accession = estd199

An integrated map of genetic variation from 1,092 human genomes.
1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. Collaborators (692)
Nature. 2012 Nov 1;491(7422):56-65 PMID:23128226

Through characterising the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help understand the genetic contribution to disease. Following from the Pilot Phase, in which we established the key principles underlying the project design, we now report on the genomes of 1,092 individuals drawn from 14 populations, constructed using a combination of low-coverage whole- genome and exome targeted sequencing. By developing methodologies to combine information across multiple algorithms and diverse data sources we provide an integrated and validated haplotype map of 38 million SNPs, 1.4 million indels and over 14 thousand larger deletions. We show how individuals from different populations have different profiles of rare and common variants and that low-frequency variants show elevated geographic differentiation, which is further increased by the action of purifying selection. By measuring the excess of rare alleles, we show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, and that rare- variant load varies substantially across biological pathways. We show that each individual harbours hundreds of rare, non-coding variants, such as

transcription-factor-motif disrupting changes at conserved sites. This resource, which captures up to 98% of variants at 1% frequency in populations of medical genetics focus, enables imputation of common and low-frequency variants in individuals from diverse, including admixed, populations.

2. Wong2012b. Study Accession = estd201

Deep whole-genome sequencing of 100 southeast Asian Malays.
Wong LP, Ong RT, Poh WT, Liu X, Chen P, Li R, Lam KK, Pillai NE, Sim KS, Xu H, Sim NL, Teo SM, Foo JN, Tan LW, Lim Y, Koo SH, Gan LS, Cheng CY, Wee S, Yap EP, Ng PC, Lim WY, Soong R, Wenk MR, Aung T, Wong TY, Khor CC, Little P, Chia KS, Teo YY.
Am J Hum Genet. 2013 Jan 10;92(1):52-66 PMID:23290073

Whole-genome sequencing across multiple samples in a population provides an unprecedented opportunity to comprehensively characterize the polymorphic variants in the population. While the 1000 Genomes Project (1KGP) has offered brief insights into the value of population-level sequencing, the low coverage inadvertently compromised the ability to confidently detect rare and low-frequency variants. In addition, the composition of populations in the 1KGP is not complete, despite the extension of the study design to more than 2,500 samples from more than 20 population groups. The Malays are one of the Austronesian groups predominantly present in Southeast Asia and Oceania, and the Singapore Sequencing Malay Project (SSM) aims to perform deep whole-genome sequencing of 100 healthy Malays. Sequencing at an average of 30-fold coverage, we illustrate the higher sensitivity at detecting low-frequency and rare variants, and the ability to investigate the presence of hotspots of functional mutations. The deeper coverage allows more functional variants to be identified for each person when compared to the low-pass sequencing in 1KGP. This set of whole-genome sequence data is expected to be the benchmark for evaluating the value of deep population-level sequencing versus low-pass sequencing, especially in populations that are poorly represented in population genetic studies. We also expect the high coverage will enable methodological and technological assessments of current strategies in sequence data analysis.

3. Xu2011. Study Accession = nstd71

SgD-CNV, a database for common and rare copy number variants in three Asian populations.
Xu H, Poh WT, Sim X, Ong RT, Suo C, Tay WT, Khor CC, Seielstad M, Liu J, Aung T, Tai ES, Wong TY, Chia KS, Teo YY. *Hum Mutat*. 2011 Dec;32(12):1341-9. PMID: 21882294

Copy number variants (CNVs) extend our understanding of the genetic diversity in humans. However, the distribution and characteristics of CNVs in Asian populations remain largely unexplored, especially for rare CNVs that have emerged as important genetic factors for complex traits. In the present study, we performed an in-depth investigation of common and rare CNVs across 8,148 individuals from the three major Asian ethnic groups: Chinese (n = 1,945), Malays (n = 2,399), and Indians (n = 2,217) in Singapore, making this investigation the

most comprehensive genome-wide survey of CNVs outside the European-ancestry populations to date. We detected about 16 CNVs per individual and the ratio of loss to gain events is ~2:1. The majority of the CNVs are of low frequency (<10%), and 40% are rare (<1%). In each population, ~20% of the CNVs are not previously catalogued in the Database of Genomic Variants (DGV). Contrary to findings from European studies, the common CNVs (>5%) in our populations are not well tagged by SNPs in Illumina 1M and 610K arrays, and most disease-associated common CNVs previously reported in Caucasians are rare in our populations. We also report noticeable population differentiation in the CNV landscape of these Asian populations, with the greatest diversity seen between the Indians and the Chinese.

4. Zhu2011. Study Accession = nstd55

X-linked congenital hypertrichosis syndrome is associated with interchromosomal insertions mediated by a human-specific palindrome near SOX3. Zhu H, Shang D, Sun M, Choi S, Liu Q, Hao J, Figuera LE, Zhang F, Choy KW, Ao Y, Liu Y, Zhang XL, Yue F, Wang MR, Jin L, Patel PI, Jing T, Zhang X. *Am J Hum Genet*. 2011 Jun 10;88(6):819-26 PMID:21636067

X-linked congenital generalized hypertrichosis (CGH), an extremely rare condition characterized by universal overgrowth of terminal hair, was first mapped to chromosome Xq24-q27.1 in a Mexican family. However, the underlying genetic defect remains unknown. We ascertained a large Chinese family with an X-linked congenital hypertrichosis syndrome combining CGH, scoliosis, and spina bifida and mapped the disease locus to a 5.6 Mb critical region within the interval defined by the previously reported Mexican family. Through the combination of a high-resolution copy-number variation (CNV) scan and targeted genomic sequencing, we identified an interchromosomal insertion at Xq27.1 of a 125,577 bp intragenic fragment of COL23A1 on 5q35.3, with one X breakpoint within and the other very close to a human-specific short palindromic sequence located 82 kb downstream of SOX3. In the Mexican family, we found an interchromosomal insertion at the same Xq27.1 site of a 300,036 bp genomic fragment on 4q31.2, encompassing PRMT10 and TMEM184C and involving parts of ARHGAP10 and EDNRA. Notably, both of the two X breakpoints were within the short palindrome. The two palindrome-mediated insertions fully segregate with the CGH phenotype in each of the families, and the CNV gains of the respective autosomal genomic segments are not present in the public database and were not found in 1274 control individuals. We used a PCR-based sequencing method to detect deletions mediated by a human-specific palindromic sequence in 740 individuals of different ethnic origins.Analysis of control individuals revealed deletions ranging from 173 bp to 9104 bp at the site of the insertions with no phenotypic consequence. Taken together, our results strongly support the pathogenicity of the identified insertions and establish X-linked congenital hypertrichosis syndrome as a genomic disorder.

Updates, Modifications and Improvements

1. We have updated the method by which the number of gains and number of losses are calculated. Previously the number of variants were counted and reported for each region within a study. In some cases, overlapping variants detected in the same sample inflated the numbers we presented. We have now modified the counts to reflect the non-redundant number of samples which have been identified to harbour a gain or loss for that particular region (when this information is available). Not all studies have included the sample information, and in these cases, we have reported the number of variants within the region.
2. The platform that was utilized by the authors has been added to the variant details page. This will help to provide additional details on the specific approach used.

Corrections and Re-evaluation of Studies in DGV

Study estd188, Pinto2011:
We have temporarily removed this study based on feedback from the authors. The underlying data will be re-evaluated and a refined, high confidence subset of the data will be presented in a future update.

Study estd50, Durbin2010:
With the release of the Phase 1 data from the 1,000 Genomes Project, the deletions from the Pilot phase have been removed. The underlying data was reanalysed and improved detection methods were employed, providing a high-quality set of calls in a larger set of samples. The insertions and tandem duplications were kept, as these were unique to the Pilot Phase.

Study nstd54, Cooper2011:
All of the variants in this study were associated to sets of samples containing both cases and controls and thus we are unable to present the data from only the control individuals.

Changes to Study Names:
We have updated the study names for estd59, and nstd50 to “1000 Genomes Pilot Phase” and “Arlt et al 2010” respectively. These are now consistent with the entries at DGVa and dbVAR.

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