Database of Genomic Variants

DGV Newsletter March 2009

Hello!

The *Database of Genomic Variants* has just 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 update of the database includes three new datasets adding a total of 6,800 new variants to the database. Two of the datasets are from recent individual genome sequencing efforts. The details of the changes in this update are described below.

New datasets in March 2009 release (hg18.v7)

Accurate whole human genome sequencing using reversible terminator chemistry
Bentley DR *et al*. 2008. *Nature* 456(7218):53-9. PMID: 18987734

This study used Illumina sequencing to characterize the genome of a male Yoruba from Ibadan, Nigeria (NA18507). The authors constructed a consensus sequence based on >30x average depth of paired 35bp reads, based on insert sizes of 200bp and 2kb, respectively. In the paper, the authors reported that they discovered 5,704 structural variants larger than 50bp in size. Here we have included a subset of the deletions detected in the study, as per the authors' recommendations. In total, we are adding 4,116 deletions from this study as part of the new update. Of these, 693 are >1kb in size and the rest are in the 100bp-1kb size range.

The diploid genome sequence of an Asian individual
Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Li D, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J. 2008. *Nature* 456(7218):60-5. PMID: 18987735

In this study, the authors used Illumina sequencing to characterize the genome of a Chinese individual. Sequence was produced using both single and paired read strategies, with 36-fold average coverage. The majority of structural variants identified in this project were deletions. In total, we have added 2,469 variants from this study as part of the new DGV update, including deletions, tandem duplications and inversions.
.

Database of Genomic Variants

Whole population, genome-wide mapping of hidden relatedness
Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I. 2009. *Genome Res.* 19(2):318-26. PMID: 18971310

This study describes the development of an algorithm called GERMLINE, which identifies segmental sharing indicative of recent common ancestry between pairs of individuals. The approach can also be used to detect polymorphic deletions. Here, the authors used GERMLINE on a sample of 3,000 individuals from the Micronesian Kosrae population, as well as the HapMap samples. The 215 deletion regions identified in the study have been added to DGV.

Browser updates
We have now added color to the variants to indicate whether they were reported as gain, loss or a combination of both. Variants reported as “gains” are now shown in red color and “losses” are shown as blue bars. For regions where both gains and losses are reported in a study, or where gains and losses were merged, we show the variant in green color. The same color scheme is used for both the CNV and InDel tracks. When interpreting this data, it is important to remember that gains and losses are relative and may be reported differently depending on the approach used for detection (e.g. in CGH, the direction of the call may be relative to a specific reference sample).

Gene information added to download file
We have received many requests to provide information about which genes overlap the CNVs. In this update we have therefore added the RefSeq genes overlapping the variants as part of the download file. This is simply based on any overlap of a variant region and a gene. As boundaries of CNVs are often inaccurate or overestimated, these overlaps should be interpreted with caution.

Regions removed
It was brought to our attention that a small number of regions that have been linked to specific disorders were reported in the database as found in healthy controls. After considering this, we have now decided to remove entries that represent variants that are known to cause microdeletion/microduplication syndromes, in order to facilitate correct interpretation of the data. The 9 regions that were removed are either false positive calls, or regions identified in “population controls” for which individuals that had specific genetic disorders were not excluded. Only calls entirely matching causative loci in well-established genomic disorders were removed.