Using the UCSC Genome Browser to evaluate putative genetic variants

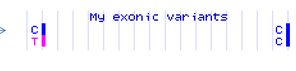
Angie S. Hinrichs, Ann S. Zweig, Brian J. Raney, Robert M. Kuhn, Fan Hsu, Belinda Giardine[2], Donna Karolchik, UCSC Genome Bioinformatics Group, David Haussler[1], W. Jim Kent
Center for Biomolecular Science and Engineering and [1]Howard Hughes Medical Institute, University of California Santa Cruz (UCSC); Center for Comparative Genomics and Bioinformatics, Pennsylvania State University

View your variant calls

The Genome Browser can display uploaded variant calls in several file formats, including the Variant Call Format (VCF) and Personal Genome SNP (pgSnp). A "track" line that specifies the data type and optional settings is required for VCF and pgSnp.

pgSnp specifies alleles, per-allele counts and optional quality scores, and is rendered as stacked bars with colors and relative heights determined by allele bases and counts.

```
track type=pgSnp visibility=3 name=exonVs description="My exonic variants" color=255,200,0
#chrom     start          end     alleles  #als  alCounts noQuals
chr17   12899901        12899902        C/T     2       1,1     0,0
chr17   12899962        12899963        C/C     2       1,1     0,0
```



VCF is a rich and flexible format that may include phased genotypes. The Genome Browser can display VCF that has been compressed and indexed by tabix (samtools.sourceforge.net), and is available by HTTP, FTP or HTTPS. In addition to viewing your own variants as VCF, you can view 1000 Genomes VCF files by browsing ftp://ftp.ncbi.nih.gov/1000genomes/ftp/release/ for the latest genotype VCF files, and constructing a track line such as this (no line breaks):

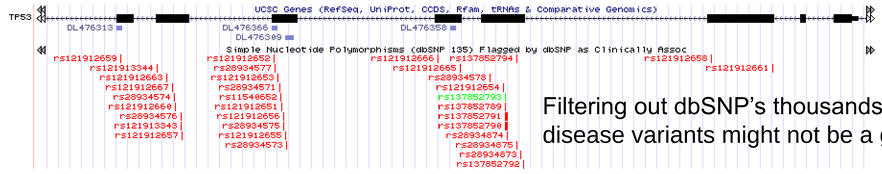
```
track type=vcfTabix name=phase1v3chr1 description="1000 Genomes Project, Phase 1 Release v3, chr1 SNPs, indels, SVs" visibility=pack
bigDataUrl=ftp://ftp.ncbi.nih.gov/1000genomes/ftp/release/20110521/ALL.chr1.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
```

More information about custom track formats can be found in the Genome Browser's Help and FAQ sections.

dbSNP: Not just polymorphisms

or, please don't throw out the baby with the bathwater!

It has come to our attention that some research groups are using the entire dbSNP as a filter to remove "boring" polymorphisms. However, dbSNP contains many rare variants, and in fact has incorporated variants from many Locus-Specific Databases (LSDBs) that catalog known disease variants for specific genes.



Filtering out dbSNP's thousands of known disease variants might not be a good idea.

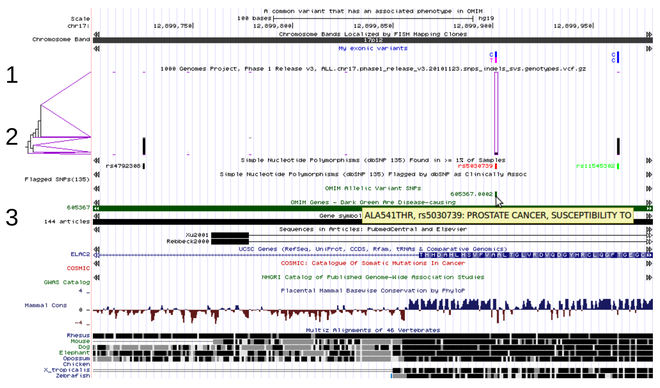
We have generated two subsets of dbSNP 135 that we hope will be useful for filtering variants:

- "Common SNPs", i.e. uniquely mapped variants for which dbSNP has allele frequency data and whose minor allele frequency is at least 1%
- "Mult. SNPs": variants which dbSNP has mapped to multiple locations in the reference genome, calling into question whether they truly vary in the population or are simply duplicated and diverged sequences

Is there an associated phenotype?

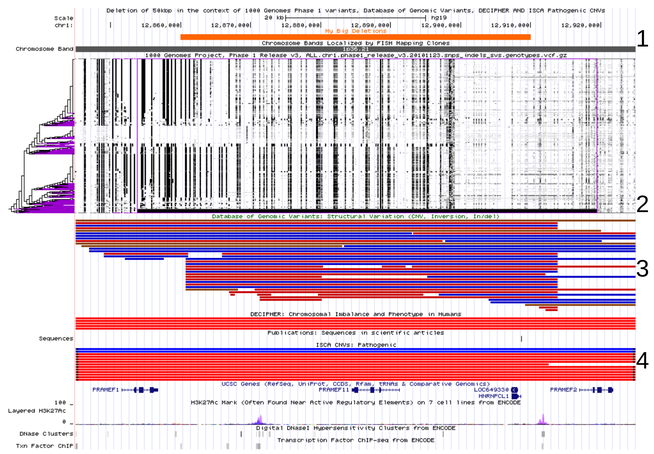
The browser image shows example exonic variants (1) with several informative tracks. 1000 Genomes Phase 1 integrated variant calls are displayed with haplotypes sorted by similarity to the selected variant (2), indicating the relative frequency of alternate alleles (black vertical bars) and linkage with neighboring variants. Three of the 1000 Genomes variants, including our two example variants, also appear in Common SNPs from dbSNP. However, one of the SNPs happens to be associated with a disease phenotype per OMIM (3); see mouse-over text box for the peptide change and phenotype.

The example variants are not found in Catalog of Somatic Mutations in Cancer (COSMIC) and GWAS Catalog tracks.



Zooming in to base-level view shows that the mutated peptide is conserved across mammals.

How rare is this deletion?



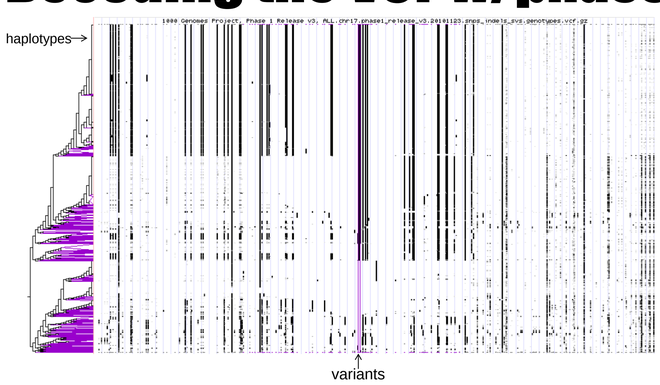
A hypothetical deletion of ~50,000 bases is shown in orange (1). In the same region, 1000 Genomes has identified a slightly larger deletion in 53 out of 2184 phased haplotypes (2). The Database of Genomic Variants (DGV) contains several reported deletions as well as duplications at the same approximate location.

Zooming out to view a larger region shows that the gains and losses in the Database of Chromosomal Imbalance and Phenotype (DECIPHER) and the International Standards for Cytogenetic Arrays (ISCA) Pathogenic Copy Number Variant tracks (4) are for much larger regions.

Novel variant: what's in the neighborhood?

If a variant is not found in dbSNP, COSMIC, or other sets of annotated variants, other data types might still be informative. This hypothetical variant (1) does not fall within any UCSC gene, but it does happen to fall in an exon of a pseudogene annotated by Ensembl (2), a small peak of histone methylation with a marker of possible regulatory activity in at least one cell line (3), regions with moderate DNase hypersensitivity and transcription factor binding (4), and some evidence for transcription in several ENCODE expression datasets (4).

Decoding the VCF w/phased genotypes display

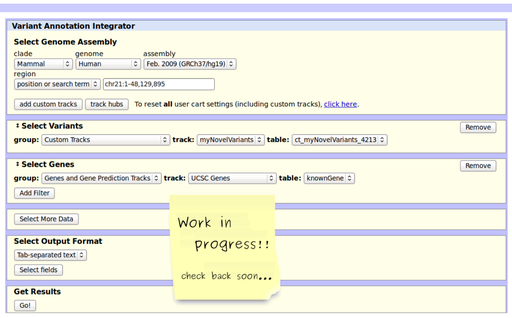


When a VCF file contains phased genotypes, i.e. an attempt has been made to separate the maternal and paternal haplotypes, the Genome Browser sorts the haplotypes by similarity so that strong linkage patterns can be identified visually. Each variant is represented as a vertical column; haplotypes run horizontally. By default, reference alleles are invisible and alternate alleles are drawn in black; when multiple variants and/or haplotypes are squeezed into the same pixel, grayscale is used to indicate the portion of alternate alleles.

To the left of the main image, a clustering tree shows the hierarchical pairings of similar haplotypes. Similarity is measured by weighted distance from a central variant outlined by a tall purple box. Due to computational complexity, the number of variants used to measure distance is limited; purple marks appear above and below variants used. Beyond those, no sorting is performed; haplotypes might appear more fragmented than they are. Purple angles in the tree indicate sets of identical haplotypes (leaves of the tree).

Work in Progress: Integrated Variant Annotation

Before a variant can be scrutinized in the Genome Browser, usually it must survive the weeding out of millions of variant calls. Several existing tools use gene annotations to predict the functional impact of variants. We are developing an interactive tool that can add not only gene-based functional predictions, but also data from almost any data track. There will also be a command-line version for offline processing when data sizes are too large for web queries. Our existing Table Browser tool can identify variants that overlap data in other tracks, but lacks the ability to add those data to the variant annotations. In addition to producing combined output, the new tool will support filtering of variants using data from other tracks, and enhanced display of results in the Genome Browser.



More Information

Click "**Help**" link (upper right) to get tool-specific help pages.

Search for answers to questions:
http://genome.ucsc.edu/contacts.html

Or **email** your question to the actively monitored public list:
genome@soe.ucsc.edu

OpenHelix provides free training material:
http://www.openhelix.com/downloads/ucsc/ucsc_home.shtml
and also offers training seminars (some free or discounted).

http://genomewiki.ucsc.edu/index.php/BoG2012VariationPoster

Acknowledgements

This work was funded by National Human Genome Research Institute award 5P41 HG002371-12 to UCSC Center for Genomic Science and 5 U41 HG004598-05 to UCSC ENCODE Data Coordination Center. Thanks to the 1000 Genomes Project for releasing Phase 1 variant calls in advance of publication.
We would like to acknowledge our many collaborators and data providers, and our users for their feedback and support.

Reference

The UCSC Genome Browser database: extensions and updates 2011. Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Pohl A, Malladi VS, Li CH, Learned K, Kirkup V, Hsu F, Harte RA, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, Kent WJ. Nucleic Acids Res. 2012 Jan;40(Database issue):D918-23..