SNPtrack manual and quick start guide v1.0

Method description and required data

SNPtrack software method relies on the fact that all individuals with monogenic recessive traits (including those caused by mutations), derived from the same allele, lack diversity in the genomic region harboring the causal variation. In DNA sequence terms it means that such a homozygous genomic region will have identical sequence in any number of affected individuals (Fig. 1). The size and position of these regions will be defined by recombination rate and specific recombination points in each affected chromosome.

By pooling DNA from many affected individuals into a single pool one can identify such a genomic area defined by an identical haplotype. Single Nucleotide Polymorphisms (SNPs) can play a role of versatile markers to track such behavior. On the other hand Next Generation Sequencing (NGS) is a fast and cost-effective way to assess such SNPs in a single experiment.

In contrast with other methods we recommend using a control pool from phenotypically WT/unaffected individuals (Fig. 2) as it allows to account for non-informative (same in all parental chromosomes) SNPs and greatly decreases false positive results. Additionally it allows for not performing outcrosses into other genetic backgrounds and elevates the requirement to know/track strain history and variation. For more detailed method description please refer to: *Leshchiner I, Alexa K... Beier D, Goessling W, Sunyaev S. Mutation mapping and identification by whole genome sequencing. Genome Res. 2012* (http://genome.cshlp.org/content/early/2012/05/03/gr.135541.111).

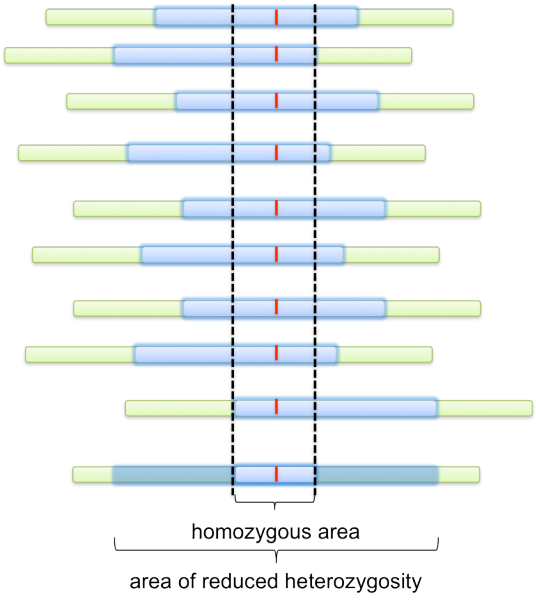

Fig.1 Structure of chromosomes in the affecteds pool.

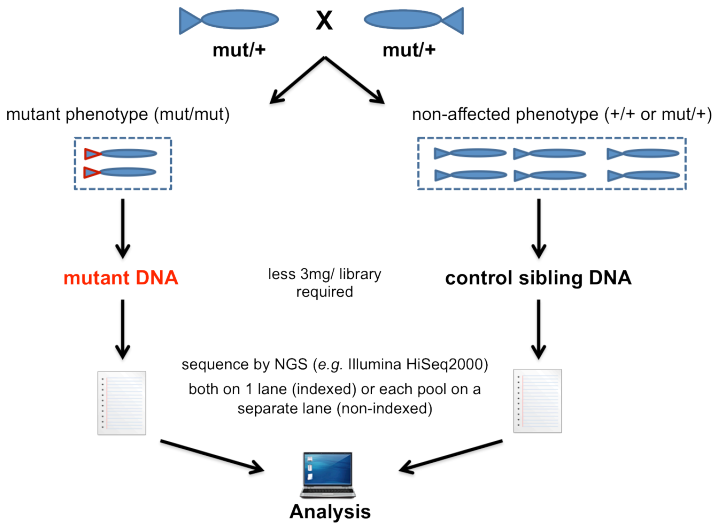

Fig.2 Experiment setup scheme.

Before submitting data to SNPtrack online tool you have to obtain raw sequencing files for both pools in .fastq format. Only gziped (.gz) or bzip2 (.bz2, preferred) files are accepted to save time and network traffic.

Quick Start Guide

Goto http://genetics.bwh.harvard.edu/snptrack/index.html.

Google Chrome and session id

When you land on the start page make sure you are using Google Chrome (if you don't – follow the link on the screen and install; it takes a few moments); some issues are reported with other browsers, please do follow this requirement as it is usually the fastest and the most advanced browser available.

If your browser cookies are not disabled you'll be given an automatically generated session id (**sid**). This is your experiment identification name please keep it written down (you can also request it to be sent to you by email upon data submission). It will be automatically saved on the computer you first use to enter the SNPtrack site, but you can also use it to retrieve your experiment and results from any other computer. Keep it safe, as if anybody from the outside will gain access to your data if he knows your **sid**.

Upload your files into two "boxes" seen on screen (Fig 3.)

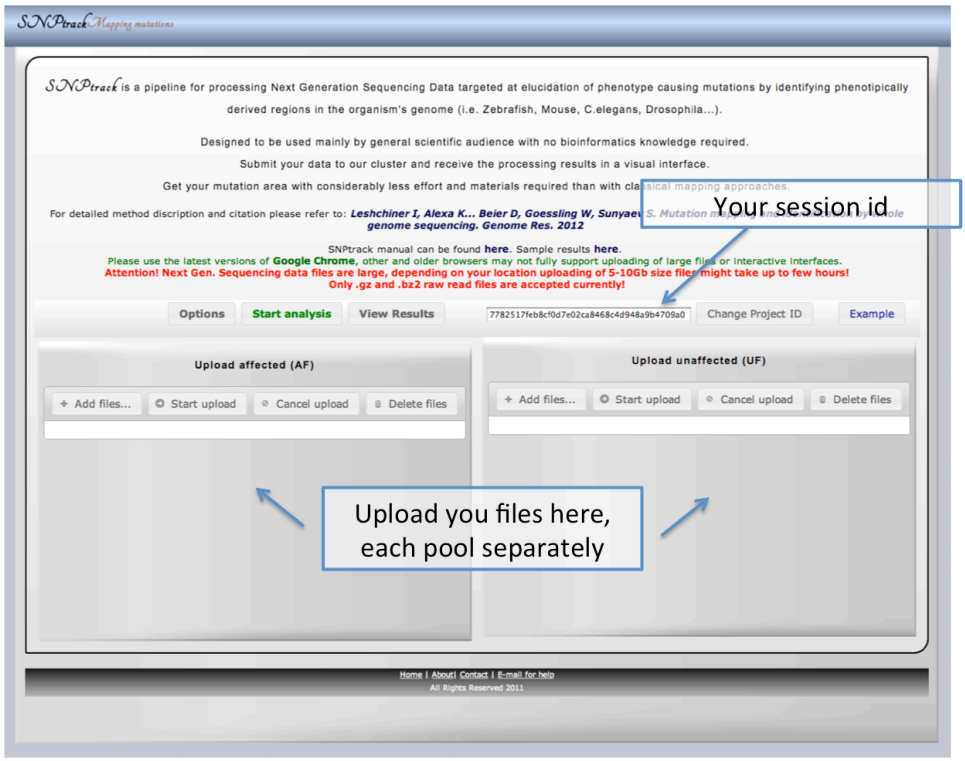

Fig. 3 The SNPtrack starting page.

Uploading will likely take several hours due to very large file size. When the uploading is complete

you'll see files in both containers as depicted in Fig. 4.

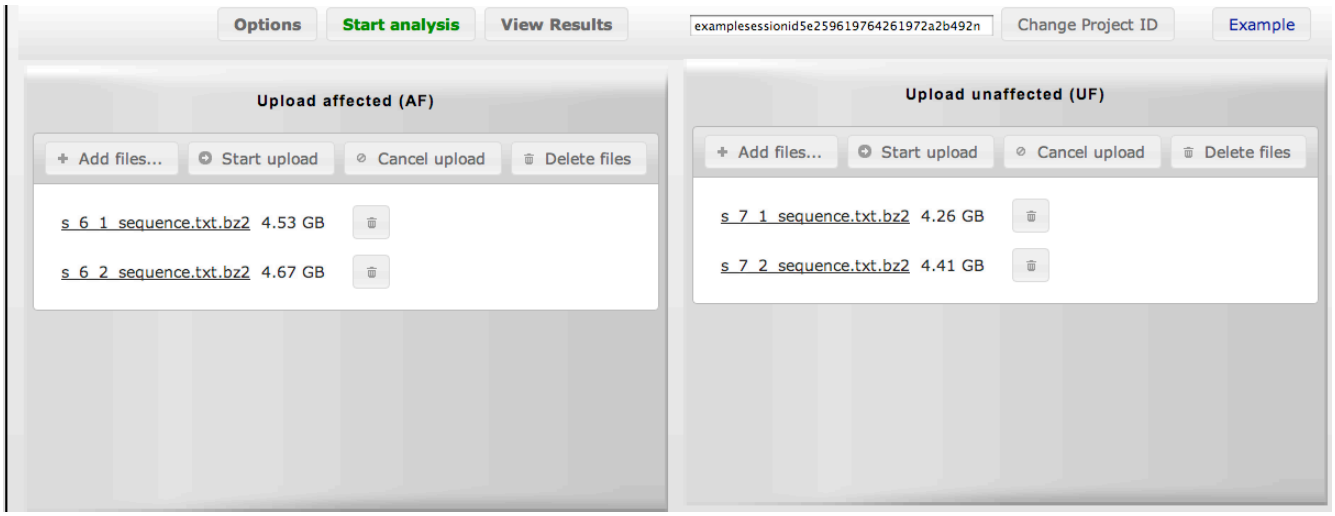

Fig. 4 Files uploaded in the start screen.

When you make sure that all files are uploaded and the sizes are similar to those seen on your local system (although they might differ somewhat) you can start submitting your files for analysis.

Submitting uploaded sequence files for analysis

Click on the **"Start analysis"** button. A dialog with experiment options will appear (Fig. 5).

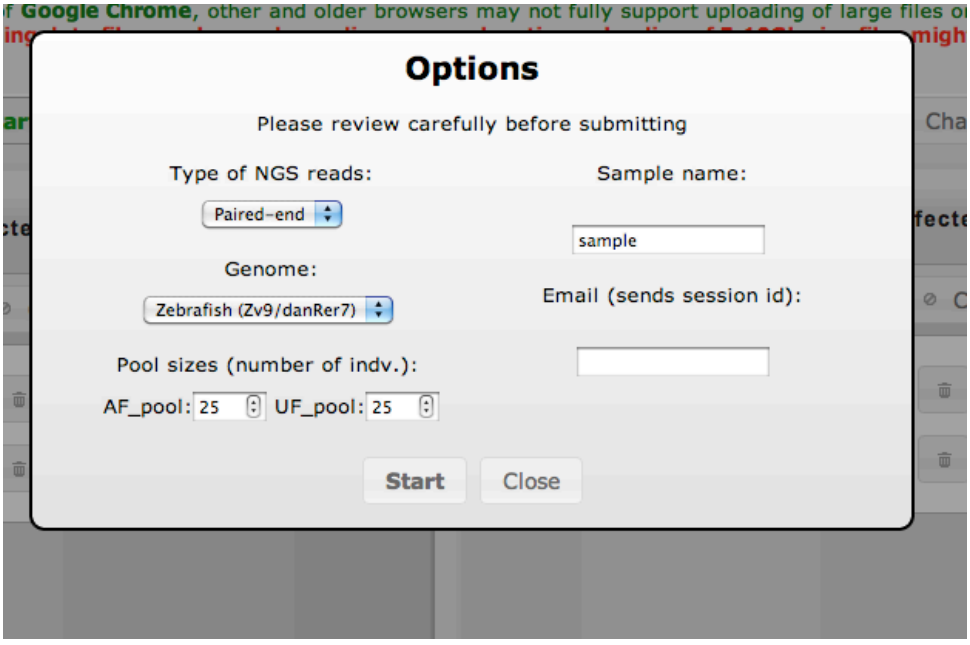

Fig. 5 Options dialog before submission.

On this dialog window please make sure all the settings are correct:

- Please specify the correct organism's genome. (If your organism is not in the list – please email to snptrack@gmail.com and the team will try to assist you and add the organism's genome to the list.)
- Please specify your pool sizes for the affecteds and unaffecteds pools. **These values are used in the analysis, take care to specify them as close as possible to real numbers!** If you're not completely sure of the exact number – specify a number that is in +/- 10 region of the supposed value. More inaccuracy in specifying these values will lead to less precise prediction of the mutation position/region, but a rough mapping position should still be produced.
- Make sure you choose the correct sequencing data (paired or single-end).
- You can specify a name to go with this submission (to help differentiate separate runs under same **sid**).
- Specify your email address if you want your **sid** emailed to you. This address will also be used by the SNPtrack team as the means of contacting you if some problems arise during the analysis.

Invalid values will be highlighted in red.
When you're ready – click **"Start"**.
A usage agreement will appear (Fig. 6), reminding you of that this is a Research Tool and that we anticipate to include non-causal SNPs into the databases for known variation for the benefit of other users.
Please read carefully and click **"I agree"** if you choose to proceed.

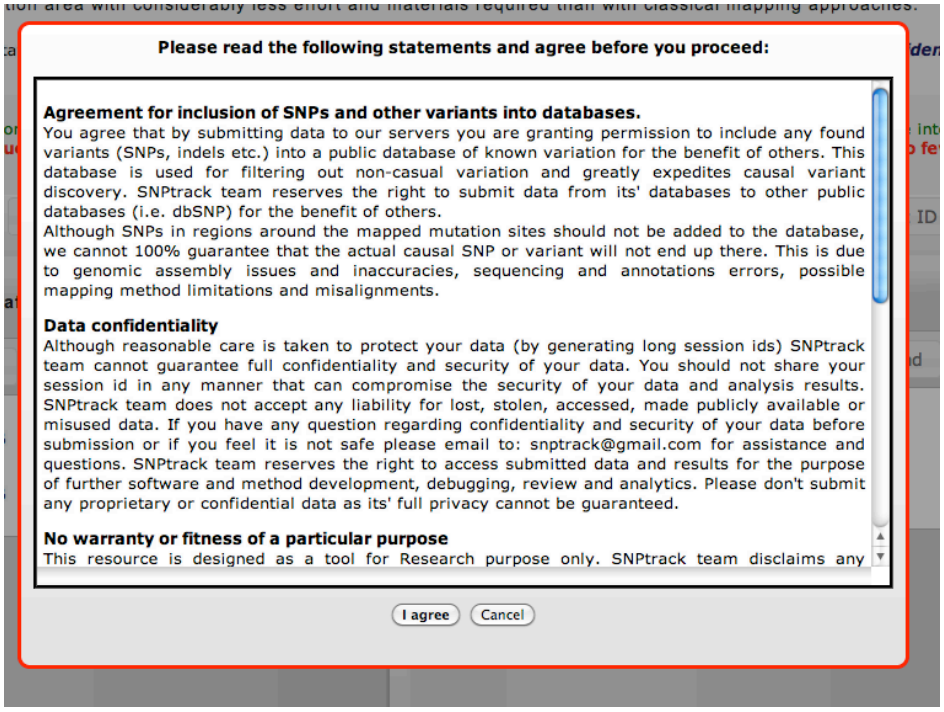

Fig. 6 Usage agreement before submission.

Running/Submitted and Finished jobs page

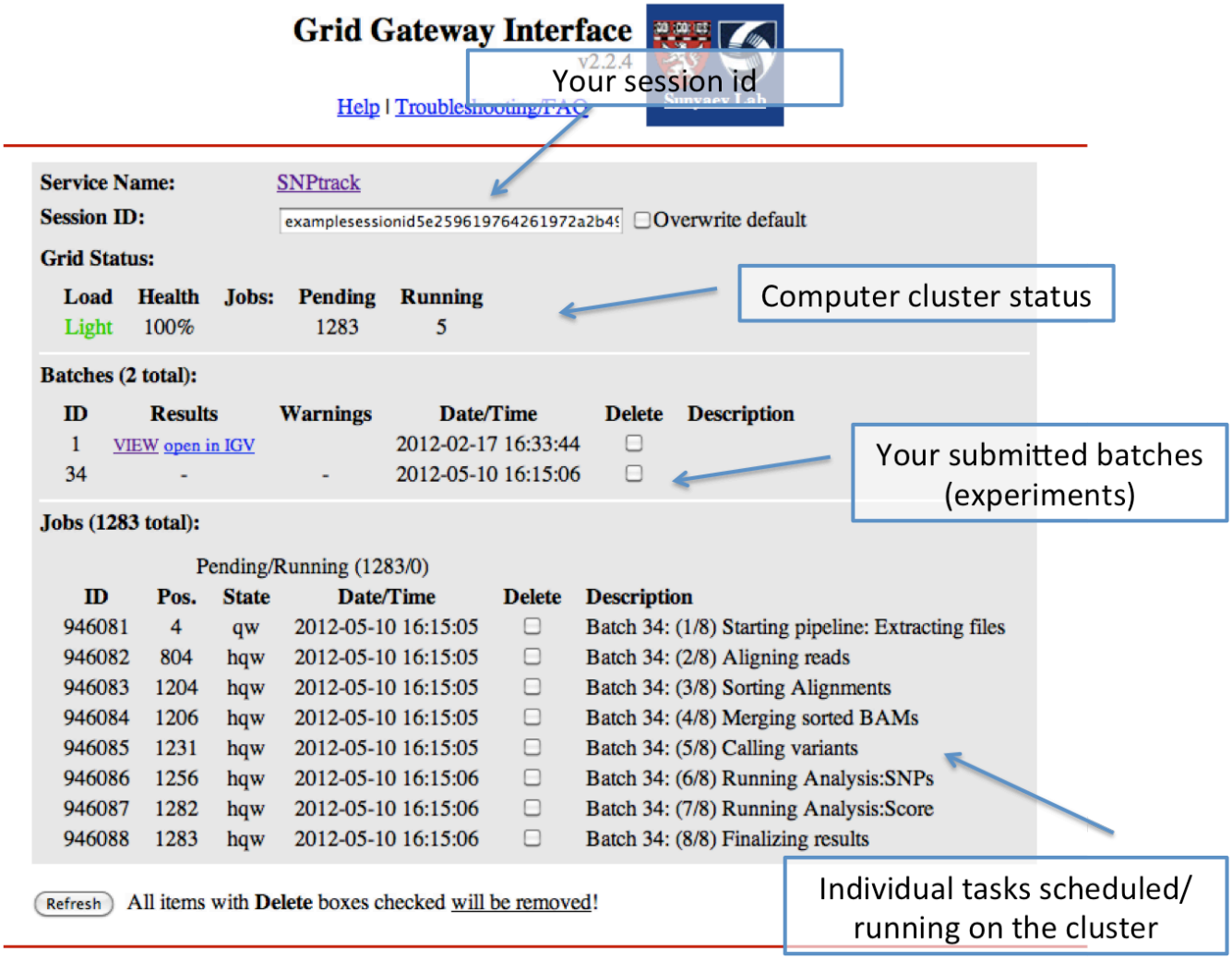

Fig. 7 Cluster status/control page.

After submission you should be carried over (Fig. 7) to the Running/Submitted and Finished jobs page (Cluster status/control). This page can be accessed anytime form the previous screen by clicking the **"View Results"** button (Fig. 3).

Here you can see your submitted experiments that either are in process of running or finished.

Individual tasks/SNPtrack pipeline steps in the job queue are shown on the bottom.

Click **"Refresh"** to update the displayed information. You can also change your session id to any other valid, existing session id (*e.g.* from a previous experiment).

The cluster **Load** will give you some information on how long will it take for your analysis to be ready. If the cluster is empty and the load is Light the average time of one sample analysis is close to several hours.

If your data is successfully processed and analyzed you will see blue "VIEW" and "open in IGV" links on the line corresponding to your sample (batch) id (Fig. 7). To access your results please click on one of these links.

Viewing your analysis results online in a webtool interface (preferred)

When you click on the "VIEW" link the corresponding results for your sample will be displayed (Fig.8).

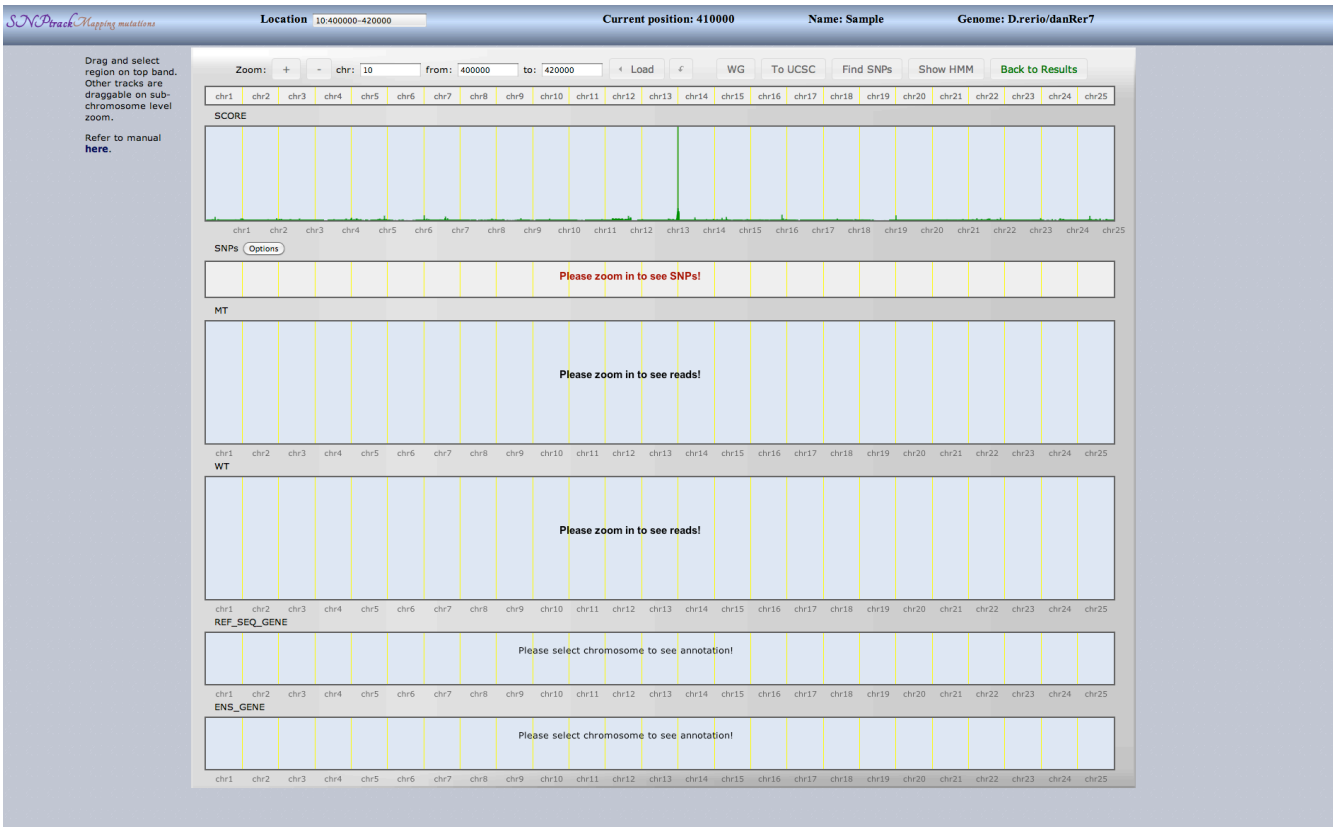

Fig. 8 View results in SNPtrack webtool.

Here you will see the traditional mapping score in the top. The upper band with genomic positions/contigs is draggable and selectable. You can select the area of interest by dragging the mouse with the left-button pressed. When a specific chromosome is selected the Hidden Markov Model (HMM) score should appear (in red).

The following tracks should be displayed:

- The score track which shows the traditional binned score in green and the HMM score in red (on chr zoom level). The HMM score can be toggled on/off by the button on the top toolbar.
- The SNP track with annotation. Options button on the left will allow to filter variants on different bases: Coverage, Effect, Existence in SNPtrack database, Repetitive regions

- Then two tracks with corresponding reads aligned to the reference genome are displayed. The MT (affected) and WT (unaffected) pool alignments. You'll have to zoom in to <50kb to be able to see reads and even more to see individual SNPs and Genome sequence.
- The last two tracks show annotation of genes from RefSeq and Ensembl databases.

All tracks can be dragged around to reach a specific genomic area, can be zoomed +/- with the buttons on the toolbar, reads can be scrolled by the mouse wheel. Undo button is provided for convenience. The **"WG"** button will bring you back to the whole genome view.

By pressing "to UCSC" button you can open up any set genomic interval in the UCSC Genome Browser (Fig. 9).

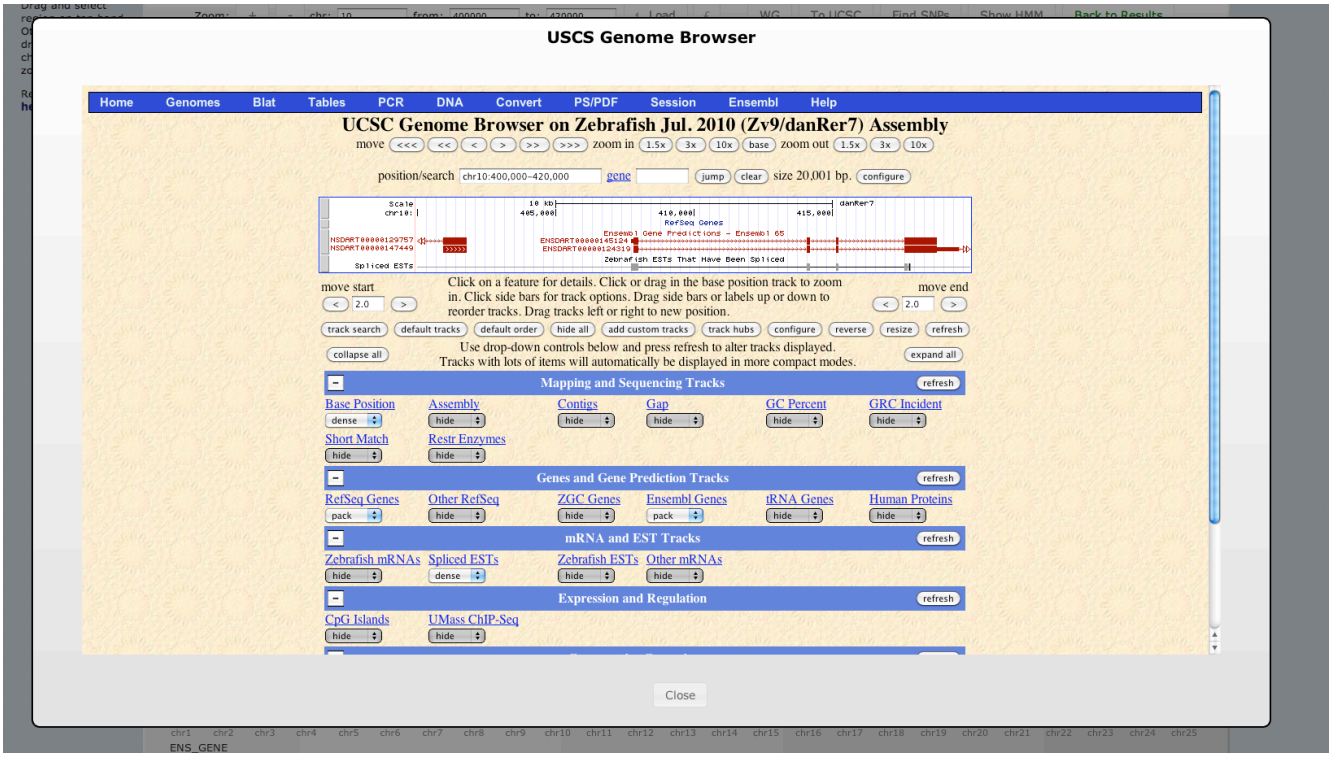

Fig. 9 View of UCSC Genome Browser Opened at specific position.

Searching for SNPs/ causal variants in the mapped area

By pressing **"Find SNPs"** button you should be able to search for SNPs in a large (up to 20Mb) area anywhere on the chromosome. A special window opens up with the list of SNPs (Fig. 10)

Various sophisticated filtering options are provided. Including filtering on Effect and homozygousity of SNPs in the MT or WT pool. You can click on genes/transcripts to open the corresponding Ensembl page; on positions - to open specific locations in the genome. Excel formatted files of shown SNPs are available for download. Results from two types of annotation programs are presented: Variant Effect Predictor by Ensembl (default, preferred) and snpEff.

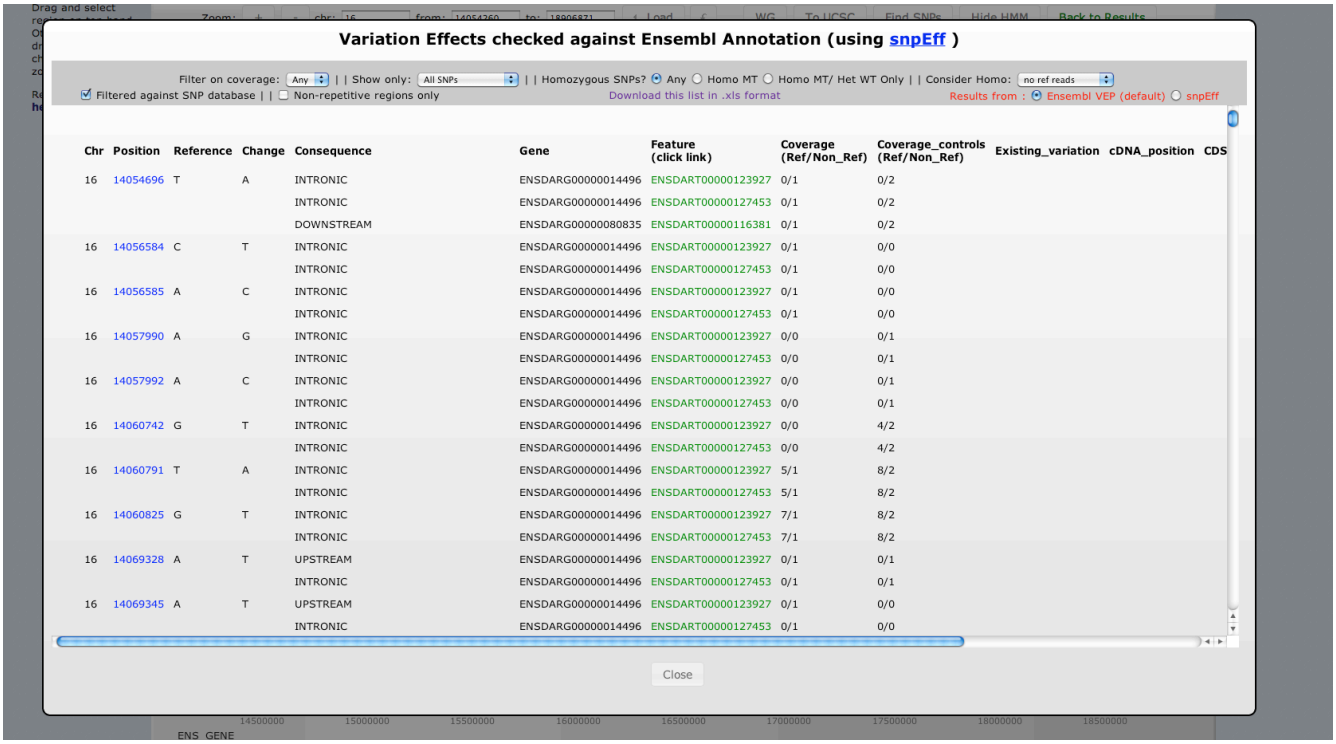

Fig. 10 Search for SNPs in the mapped region.

Viewing your analysis results remotely in Integrative Genomics Viewer (IGV)

As an alternative method for accessing your results the "open in IGV" link on Cluster Control page can be clicked. You'll be directed to the screen depicted in Fig. 11.

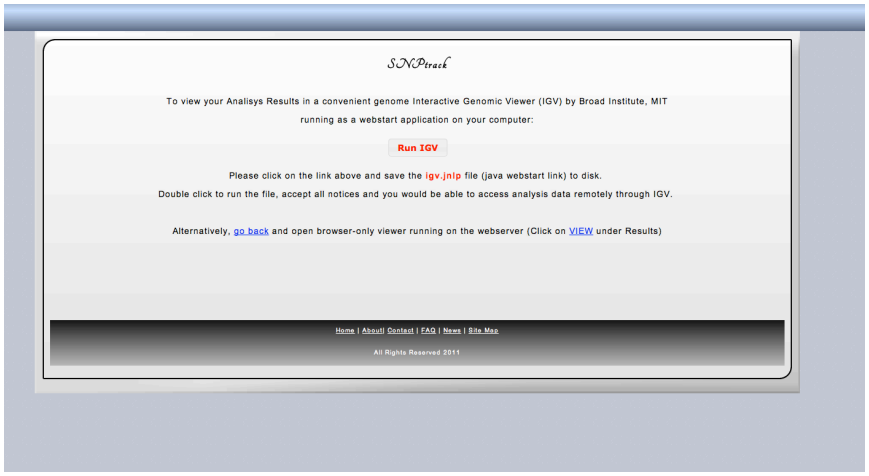

Fig. 11 IGV results link download page.

You should follow the instructions on screen, *i.e.* click the button, save the igv.jnlp file to disk and run it. This should automatically download IGV by Broad Institute MIT (http://www.broadinstitute.org/igv/) and load your analysis results. Detailed manual for IGV is available at their webpage.

IGV tracks include:

- All SNPs with annotation track for two (MT and WT) pools
- SNPs that are not found in SNPtrack database track
- Found SNP effects track
- Traditional score track
- HMM score track
- Two alignments for MT and WT pools (.bam files)
- Gene annotation track

Sometimes issues are seen with IGV if very strict firewall settings are used at your institution. Please email to snptrack@gmail.com if you encounter such issues.

Please address all questions, comments and concerns to snptrack@gmail.com.