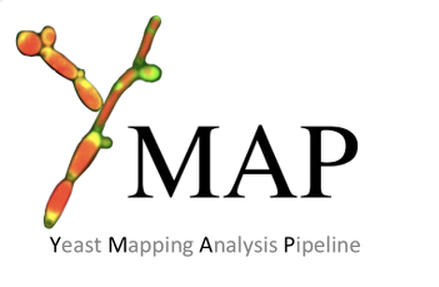
YMAP User Manual



<http://lovelace.cs.umn.edu/Ymap/>

**Table of content:**

Introduction

YMAP User interface

Creating a new project

Interpretation of results

**Introduction**

**Overview**

YMAP is a web-based tool for the analysis of genome alterations in single cell Eukaryotes. The purpose of this tool is to facilitate the analysis of three major genome changes: copy number variations (CNV), single nucleotide polymorphisms (SNPs) and loss of heterozygosity (LOH).

**The YMAP implementation**

The genome analysis pipeline is composed of three main components: a module that performs raw sequence alignment and processing, a module that performs custom copy number variation (CNV) and single nucleotide polymorphism (SNP)/loss of heterozygosity (LOH) analyses, and a module that constructs figures summarizing all completed analyses and then displays them on the webpage. The implementation details for each of these components are described in more detail in the following sections.

The first component of the central computation engine takes the user-input data and attempts to correct some typical file errors before outputting corrected FASTQ file(s) for use by later steps in the pipeline. Typical sequence data is input as one or two (for paired-end reads) FASTQ format files, either raw or compressed in the ZIP or GZ format. Uploading is facilitated by using compressed file format. Uploads of compressed files typically takes a few hours, depending on the file size. The large size of FASTQ files leaves them prone to file transfer errors that result in corruption because the file format doesn’t have an internal error correction/identification system. This corruption often results in the final read entry being incomplete, which can cause analysis programs to crash, and normally has to be dealt with on a case-by-case basis. Trimming the FASTQ file to remove incomplete read entries can then result in paired-end files having imbalanced numbers of reads. Raw paired-end read files also often contain a minority of single-end reads that are generated by some sequencing technologies. Both causes of imbalanced reads can interfere with the processing by some tools designed for paired-end reads exclusively. Trimming the longer of the paired-end FASTQ files to the length of the shorter file resolves this issue and completes the implemented automatic FASTQ file correction algorithm.

The second step in the central computation pipeline is to processes the corrected FASTQ file into a final Binary sequence Alignment/Mapping (BAM) file. The single- or paired-end reads are aligned to one of the installed reference genomes using Bowtie2 [27], resulting in a Sequence Alignment/Mapping (SAM) file. SAMtools [28] is used to compress this into a BAM file. PicardTools [29] is used to standardize the read-group headers in the BAM files, to resolve some formatting irregularities to the BAM file. SAMtools is then used to sort the BAM file, which is required for efficient later processing steps. FASTQC [30] is used to identify the quality coding system used in the input FASTQ files, as a prelude to defining the input parameters for processing by the Genome Analysis ToolKit (GATK [31]), which performs indel-realignment of the BAM files, removing spurious apparent SNPs around true indels in the primary alignment.

The third step in the sequence data processing component of the pipeline is to convert the BAM file into a simpler text file containing limited data for each coordinate across the genome, which simplifies later processing. The SAMtools function mpileup first processes the BAM file into a ‘pileup’ file, which contains information about all of the mapped reads at each chromosome coordinate in a simple format that facilitates subsequent processing by custom Python scripts. The Python scripts extract base call counts for each coordinate, discarding indel and read start/end information. The raw read-depth data per coordinate is saved to a text file that is input into the CNV analysis section of the pipeline. Any coordinates with more than one base call have that information saved to a separate text file that is input into the SNP and LOH analysis section of the pipeline.

**CNV Analysis**

CNV analysis of NGS data by the pipeline is based upon read depth across the genome. Several biases can impact read depth and thereby interfere with CNV analysis. Two separate biases, a chromosome-end bias and a GC-content bias, appear sporadically in all types of data examined (including microarray and whole genome sequence (WGseq) data). The mechanism that results in the chromosome end artifact is unclear, but the smooth change in the apparent copy number increase towards the chromosome ends (Fig. 2A) suggests that some DNA preparations may release more genomic DNA as a function of telomere proximity (Jane Usher, personal comm.). A GC-content bias is due to strong positional variations in GC content in the *C. albicans* genome. This, combined with the PCR amplification bias introduced during sequence library or array preparation, results in a strong positional effect in local copy number estimates (Fig. 3A). In datasets produced from the ddRADseq protocol, a third bias is associated with the length of restriction fragments. A fourth bias, seen consistently in all ddRADseq data sets, appears as a high frequency of short-range increases and decreases in read depth at specific genome positions across all strains analyzed, and thus can be removed by normalization to a control dataset from the reference genome. The YMAP pipeline includes filters, which can be deselected by the user, for each of these biases to correct the data before final presentation and to facilitate detection of *bona fide* CNVs. The final presentation of the corrected copy number data is in the form of a histogram drawn vertically from the figure centerline (Figs. 2-4, A-B).

The chromosome-end bias is normalized using locally weighted scatterplot smoothing (LOWESS) normalization

[32] of average read depth vs. distance to the nearest chromosome end, for 5000 bp windows tiled along each chromosome (Fig. 2C). The LOWESS fitting is performed with a smoothing window size determined for each dataset as that which produces the least error between the fit and the raw data, using 10-fold cross-validation [33]. Dividing the raw data by the fit curve normalizes the bias (Fig. 2D), allowing an unimpeded view of the mapped genome (Fig. 2B, a diploid with no significant CNVs). Because this bias is sporadically present, the correction is optional and is not performed by default.

The GC-content bias is normalized using LOWESS normalization of average read depth vs. GC content, for 5000 bp windows tiled along each chromosome] (Fig. 3C). The LOWESS fitting is performed with a smoothing window size determined for each dataset as that which produces the least error between the fit and the raw data using 10-fold cross-validation. Dividing the raw data by the fit curve normalizes this bias (Fig. 3D) allowing an unimpeded visual examination of CNVs across the genome. For example, it can distinguish chromosome number for a near-tetraploid strain with a small segmental duplication near the centromere of ChrR, three copies of chromosomes 4, 5R and 6, and with 7 copies of the left arm of chromosome 5R (due to the presence of 3 copies of whole Chr5 and 2 copies of an isochromosome (5L) with 2 copies of Chr5L per isochromosome) (Fig. 3B). Because this bias is always present to some degree in all data types examined, the correction is performed by default unless deselected by the user.

The ddRADseq protocol generates high read depths at a sub-sampling of genomic loci, resulting in a much-reduced total cost per strain sequenced. The protocol produces a library of restriction fragments digested with two different restriction enzymes (in this case *Mfo*I and *Mpe*I). A strong bias exists in the read depth vs. the length of each valid restriction fragment (obtained via a simulated digest of the reference genome, followed by selecting fragments that have the two restriction fragment ends), (Fig. 4C). The fragment-length-bias is filtered using LOWESS normalization of an average read depth vs. the simulated fragment frequency. The LOWESS fitting is performed with a smoothing window size determined for each dataset as that which produces the least error between the fit and the raw data. Restriction fragments less than 50 bp or greater than 1000 bp show average read depths that exhibit a too much noise and are considered unreliable. Where the LOWESS fit line drops below one read, the fragments are considered unreliable due to the reduced dynamic range in the data. These unreliable data are noted (red points in Fig. 4D) and not used in later steps of the analysis.

For ddRADseq analyses, first the chromosome-end and GC-content bias corrections are applied using data per valid restriction fragment instead of the standard-sized 5000 bp windows used in WGseq analysis. After these corrections are performed, the remaining strong position-effect bias in read depth that remains uncharacterized. This final bias is corrected by normalizing the corrected read depths for each usable restriction fragment by the corrected read depths from a euploid reference dataset. Because the earlier biases differ from dataset to dataset, the reference-normalization is performed as the final normalization step. The result of these corrections is a pronounced reduction in noise in the CNV data as seen by comparing the raw read depth (Fig. 4A) to the corrected read depth (Fig. 4B) for an example dataset.

After these corrections are applied to the raw sequence read data, the corrected copy number estimates are locally smoothed to reduce the impact of high-frequency noise. The estimates are then multiplied by the whole genome ploidy estimate that was determined by flow cytometry of DNA content and entered during setup of the project. The corrected estimates are plotted as a histogram along each chromosome, with the lines drawn vertically from the baseline ploidy entered during project setup. Copy number variations are then evident as regions with prominent black bars.

**SNP/LOH analysis**

SNPs are regions of a genome that have two different alleles at the same locus on different homologs. The allelic ratio (0 or 1 for homozygous regions and 0.5 for heterozygous regions in a diploid genome) is used to determine whether a region that had SNPs in the parent/reference strain has undergone LOH to become homozygous. An allelic ratio is calculated for each coordinate by dividing the number of reads with the more abundant base base call by the total number of reads at each coordinate (resulting in values ranging from 0.5 to 1.0).

Three styles of analysis are performed, depending on user input during the project setup. The first style is the default option, which is used when no reference strain or hapmap is available. In this case, the SNP distribution for the strain of interest is displayed as vertical grey bars in the background of each chromosome. Once analysis has completed, this strain can be used as the ‘parent’ for other related strains. In the second style of analysis, a parent strain is chosen and the SNPs in common between that parent and the test strain being analyzed are displayed as grey bars (as in the first style), while any SNPs in the parent that have different allelic ratios in the test strain are displayed in red, if allelic ratios approach 0 or 1, or in green, if ratios suggest unusual allele numbers (often due to CNVs or aneuploidy). The third style of analysis can be chosen if a hapmap for the parent strain background is available. SNPs that remain heterozygous are again displayed in grey, while those that have become homozygous are displayed in the color assigned to the homolog that is retained (e.g., cyan for the ‘a’ allele and magenta for the ‘b’ allele).

For the default option, any coordinates with an allelic ratio near 0.5 (0.50 to 0.75) are considered heterozygous. More extreme allelic ratios are considered to be homozygous, appearing in the dataset due to sequencing errors. The density of heterozygous SNPS is presented as vertical lines spanning the height of each chromosome cartoon, with the intensity of grey color representing the number of SNPs in each 5000 bp bin. If there are fewer than 100 SNPs in a bin, it is drawn with a lighter shade corresponding to the number of SNPs relative to the 100 SNPs threshold. This results in white backgrounds for homozygous regions and increasingly dark shades of grey for regions with higher numbers of SNPs (Fig. 5A).

When a parental type strain of unknown genotype (e.g., a clinical isolate) is selected for a project, the pipeline first calculates the distribution of SNPs across the parental genome in the manner described above. For comparison of the parental genotype to another related strain (e.g., another sample from the same patient), every heterozygous SNP locus in the parent is examined in the second dataset. If the allelic ratio changes from the 0.5 value observed in the reference strain, the SNP is assigned a red color and the final color of each 5000 bp display bin is calculated as the weighted average of all the SNPs within the bin (Fig. 5B). An alternate presentation assigns red color only to coordinates that have transitioned from heterozygous to homozygous (allelic ratio of 1.0) and assigns the green color to coordinates that have unusual allelic ratios (allelic ratios between 0.75 and 1.0, only excluding those with allelic ratios precisely at 1.0) (Fig. 5C). Low SNP counts are factored into the presented colors, as described above for the first style of analysis.

When a known hapmap is selected for a project, the pipeline loads SNP coordinates from the map and examines the allelic ratios of the dataset at those coordinates. For disomic regions of the genome, any SNP locus with an allelic ratio near 0.5 (0.50 to 0.75) is considered heterozygous and assigned the color grey. Any SNP locus with a more extreme allelic ratio is considered homozygous and assigned the color corresponding to the homolog with the matching allele in the map. For regions that are monosomic, trisomic, or larger, colors are assigned to SNPs based on the apparent ratio of homologs present. SNPs within each 5000 bp bin are gathered and the final presented color is determined as the weighted average of the colors assigned to the individual SNPs (Fig. 5D). Low SNP counts are factored into the presented colors as in the cases previously described.

The sparse datasets produced from the ddRADseq protocol introduce a high sampling error to allelic ratio calls, increasing the uncertainty of SNP calls and an increased incidence of coordinates that appear as a SNP in one dataset, but not another. This sampling error in allelic ratio calls interferes with the direct comparison of SNP loci between a dataset and a parental type dataset. If one dataset is examined without comparison to a reference – producing a very noisy CNV map – the allelic ratios are plotted as grey lines emanating from the top and bottom of each chromosome cartoon inwards to the ratio calculated for each coordinate (where the Y-axis ranges from 0.0 to 1.0 for the lines) (Fig. 6A). When a dataset is examined in comparison with a reference, the pipeline produces a figure with allelic ratios for the reference strain drawn as grey lines emanating from the bottom of the cartoon and allelic ratios for the test dataset plotted as red lines drawn from the top of each chromosome (Fig. 6B). Loci with a read-depth lower than 20 are ignored, because the corresponding high sampling error produces a high likelihood of spurious midrange allelic ratios that can appear as heterozygous.

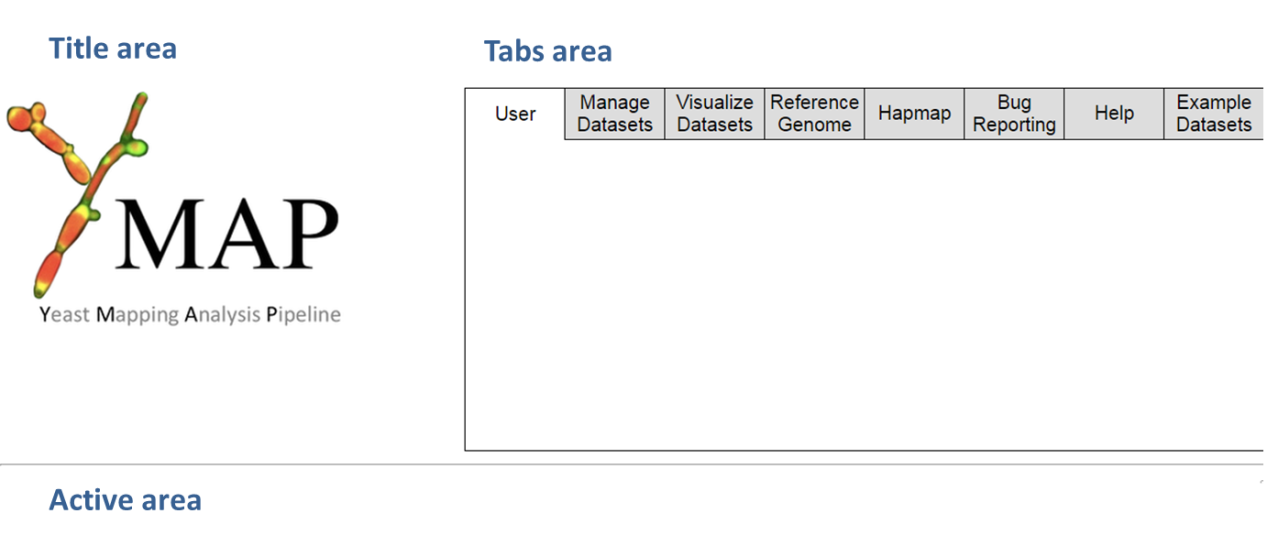
If the user selects a hapmap while setting up an analysis, the higher resolution data of the hapmap allows every SNP locus that appears in the dataset to be examined. The allelic ratios, coupled with the SNP homolog identity information from the hapmap [24, 25], allows coordinates to be assigned colors by how consistent they are with either homolog or with the heterozygous state. Lines are then drawn from the top to the bottom of each chromosome for coordinates with allelic ratios less than 1.0, in the color previously assigned (Fig. 6C). Allelic ratios of exactly 1.0 are not drawn because they often represent the sampling error found in low read depth areas of the sparse dataset. Visual comparison between the allelic ratio plots for related strains facilitates the identification of large regions of LOH (Fig. 6D: magenta at end of left arms of Chr1).

**The YMAP User interface:**

The YMAP user interface is divided into three parts: The title area, the tabs area and the active area.

The top section is the Tabs area, which shows the different options available to the user, while the bottom section, the active part, allows for the visualization of finished projects (Figure 1).

Figure 1: YMAP user interface



The tabs area contains all the different functions available to the user. Please see below for a detailed description of each tab.

Figure 2: The different tabs in the YMAP user interface and their functions

User Registration

Log in,

Log out.

Delete user.

Show analyzed

datasets.

Install new project datasets.

Delete projects.

Install\delete

user genome.

Generate new Hapmap

System status.

Bug reporting.

Descriptions of data types and input file format requirements.

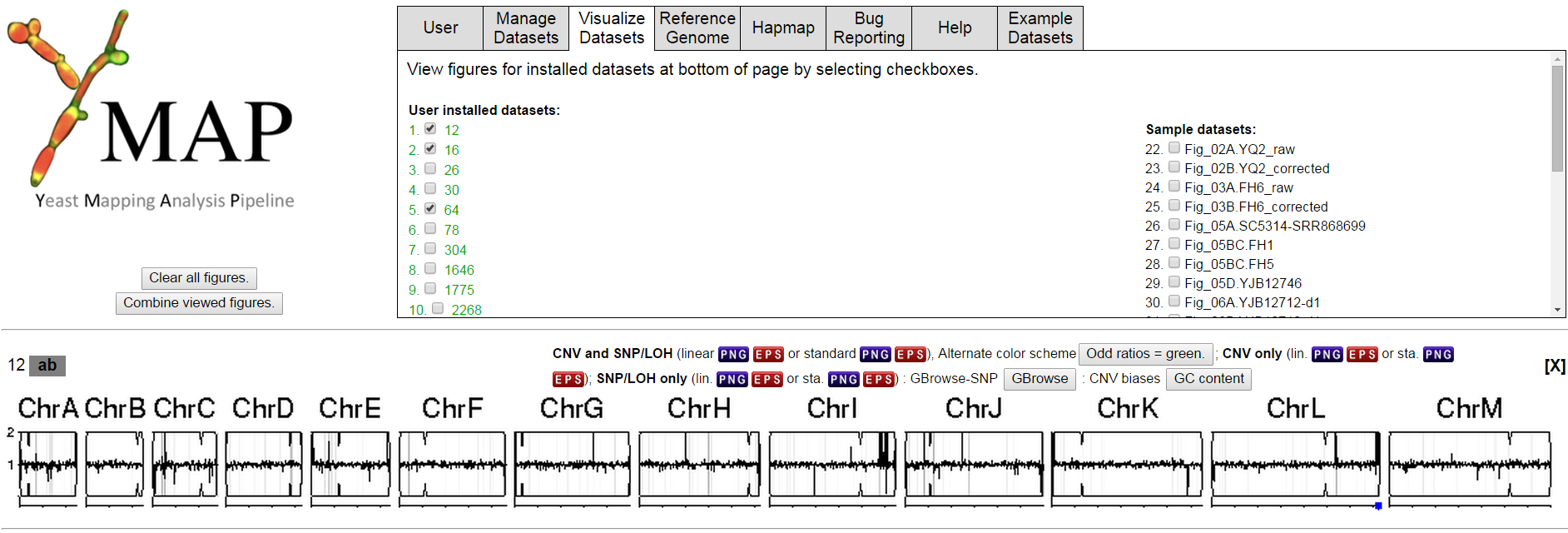
Reference and example datasets used in this paper



The “Manage Datasets” tab contains functions to install new projects, as well as functions to display or delete existing projects.

The "Visualize Datasets" tab allows for the visualization of finished projects in different formats. This window is separated into upper and lower sections. The upper section displays the list of all projects in the user’s account. The project data is displayed in lower section. The checkbox adjacent to the project name is ticked to display the data, which appears below in the order in which the data display was selected. When an additional project is chosen, an entry for the project is added to the bottom of the display section.

Figure 3: The "Visualize Datasets" tab. The check box which allows a dataset to be displayed is marked by a red circle.



The default format is a horizontal figure displaying CNVs and SNPs (as in Figure 3). Alternate formats (e.g., chromosomes displayed horizontally, one above the other) and options to display only CNVs or only SNPs are also available.

A displayed project can be removed from the viewing area by clicking the “X” at the top-right of the entry in the lower section of the window.

The “Reference Genome” tab contains functions to install or delete an installed reference genome. This is accomplished as described above. Importantly, reference genomes should be installed prior to addition of relevant project data as the uploading/analysis process will ask for the relevant reference genome for the analysis. During installation of a new genome, the loaded FASTA file is first processed to identify the names of included chromosomes. Locations of centromeres, rDNA, any other annotations, as well as any information about open reading frame (ORF) definitions are then loaded.

The “Hapmap” tab contains functions for constructing or deleting hapmap definitions. During construction of a new hapmap, the name for the new hapmap, the reference genome, and the first datasets are defined. If the hapmap is being constructed from two haploid/homozygous parents, the projects for those parents are selected in this step. The next step asks for user selection of colors to represent the two homologs. If the hapmap is being constructed from a diploid/heterozygous parent, the parent and a first partially homozygous progeny strain is chosen in this step. Next, the user is asked to select colors to represent the two homologs. This page also allows the user to define which regions of the first partially homozygous progeny strain represent an LOH event and which homologs remain. The system then processes the project data and user input to build a hapmap. A haploid hapmap will be automatically finalized at this stage, but a diploid hapmap can be improved with additional datasets until the user indicates that it is finished.

The “Bug reporting” tab contains system status indications and the option to report bugs to the developers.

The “Help” tab contains descriptions of the different input file requirements for the different data types.

The “Example Datasets” tab contains files or links to database accessions used to construct the figures in this paper.

**Creating a new project**

Figure 4: Location of the "Install New Dataset" button.

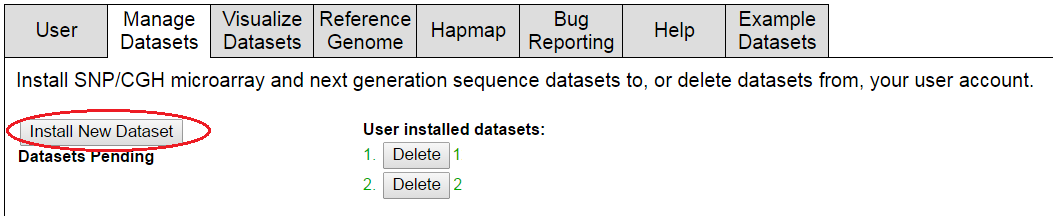
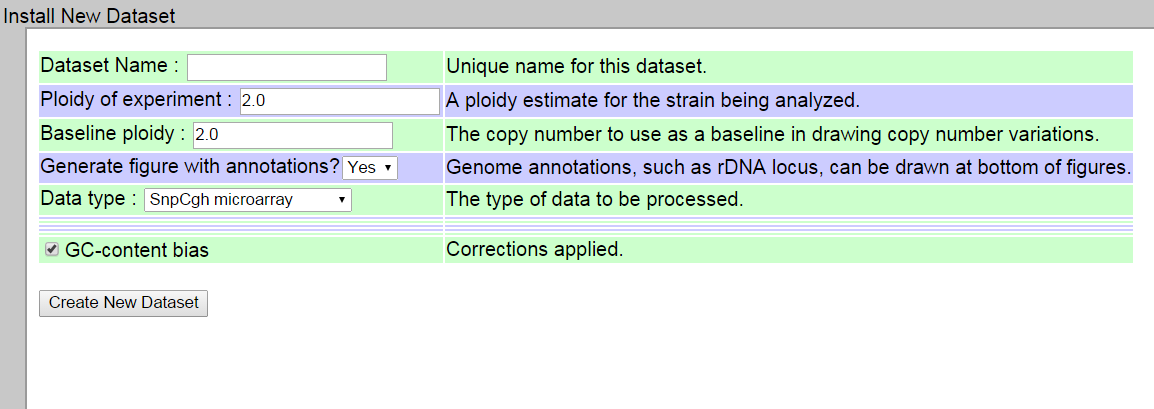


Figure 5: Appearance of the pop-up window which allows users to name their project, and enter all the parameters for its analysis.



The Dataset name is a user-defined name which allows user to recognize the project. It is best to use explicit project names, in order to avoid confusion.

The Ploidy of experiment means the expected ploidy of the dataset. For example, *C. albicans* is usually a diploid, therefore the ploidy number entered should be "2", while *C. glabrata* is usually a haploid, therefore this bar should be indicated "1" in its case. Naturally, in a case of a *C. albicans* haploid, the ploidy should be marked "1" as well.

There are several different data types:

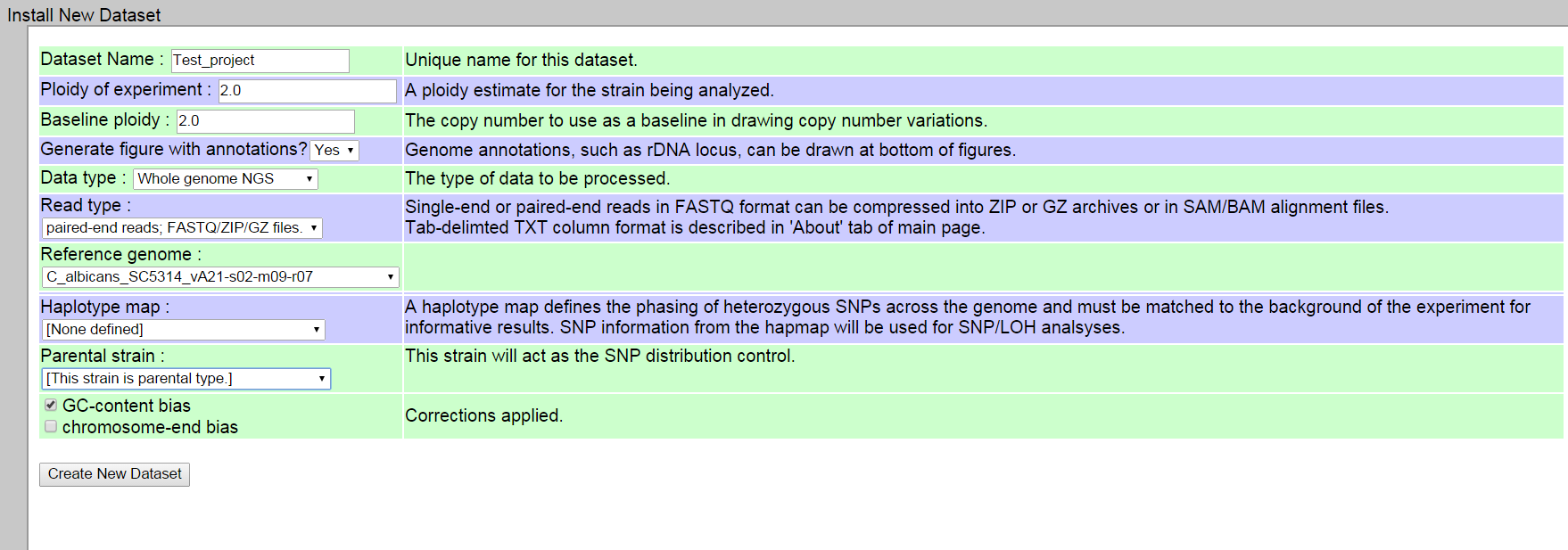
1. SNP/CGH microarray -
2. Whole genome NGS
3. ddRADSeq
4. RNAseq
5. IonExpress-seq

The first three options are currently available, while the remaining two are still in the process of being tested.

SNP/CGH microarray corresponds to the arrays defined in [25] and only has the option of correcting for the GC-bias. The other data types are all sequence-based and have additional input requirements.

Once the data type is chosen, new options would appear in this menu.

Figure 5: New options which appear in the pop-up window, once the data type is defined by the user. These options will allow user to define the desired characteristics of the project analysis.



Next, the format of the sequence read data should be defined: single or paired-ends in FASTQ format (compressed or uncompressed), BAM/SAM file or text file.

Though FASTQ files are acceptable for analysis, this file format is very large and therefore prone to file transfer errors and very long transfer times. Therefore, it is recommended to always use a compressed version of the FASTQ file (ZIP or GZ). Uploads of compressed files typically takes a few hours, depending on the file size.

Users are now requested to pick the reference genome, the hapmap information (if any) to be used, the parental strain for comparison, and a set of bias-correction filters depending on the type of sequence data. It is important to note that the choice of hapmap overrides the choice of parental strain.

After information about the specific project has been provided on the pop-up, the user must click the "Create New Dataset" button at the bottom of the page. This returns the user to the main page. It is then necessary for the user to reload/refresh the main page.

Once reloaded, a new entry will appear in the list of datasets (Figure 6). To upload the data into the project, the user then clicks on the "Add" button, which appears under the project name (Figure 6, circled in red). In the case of paired-end reads, after the first path to the first file is selected for upload, a second button will automatically appear beneath it, to allow for the second file path to be designated (Figure 7). Once the files are all designated, a green “upload” button appears (figure 8); clicking this button initiates data upload and analysis. Clicking the “Delete” button for a project removes it irreversibly from the site. To avoid inadvertent deletion of uploaded projects, a confirmation is requested from the user (Figure 6, Blue square).

Figure 6: Location of the "Add" and "delete" buttons in the "Manage Databases" tab.

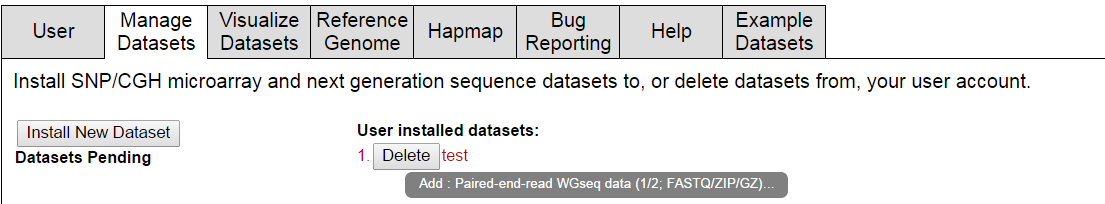


Figure 7: Appearance of a second designation buttons in the "Manage Databases" tab.

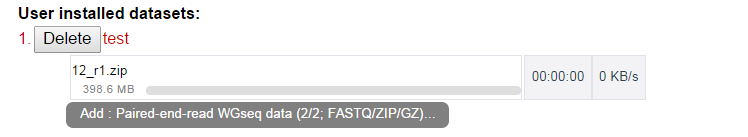


Figure 8: The appearance of the green "upload" button in the "Manage Databases" tab. This button appears only after all files are designated for upload.



As mentioned above, uploads of compressed files typically takes a few hours, depending on the file size. It is not possible to upload another project during this period. If the page is reloaded, the upload will stop and all data will be lost. The upload will then have to be restarted from the beginning.

Once uploaded, the files will be transferred directly to analysis whitout need of any further actions by the users.

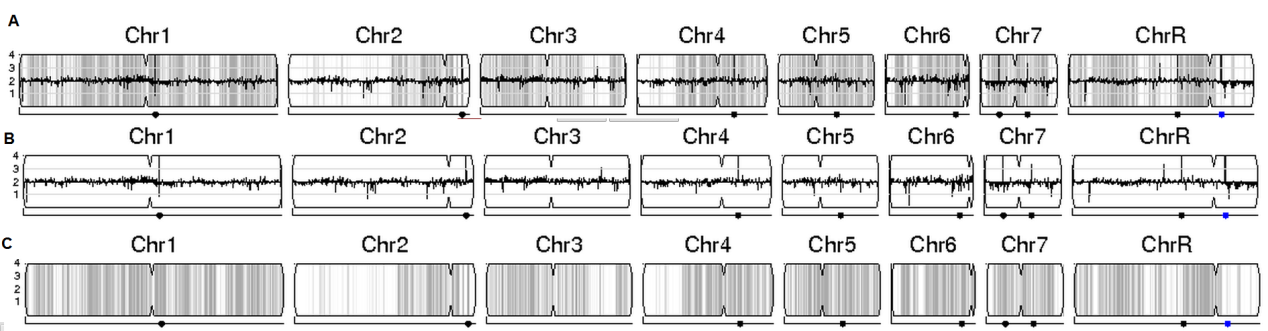
If during the analysis or upload of the data YMAP encounters a problem, particularly if the analysis process appears to take longer than expected, an error message will appear beneath the project name. If such message appears it is recommended to recheck the files and restart the project.

Project that are in the process of upload or analysis are marked in red on the "Manage Datasets" tab, when the analysis is over its color will change into green. Once a project has run to completion, it will automatically be displayed in the list of projects available under the "Visualize Datasets" tab. For more information regarding interpretation of results, please see the "Interpretation of results" section below.

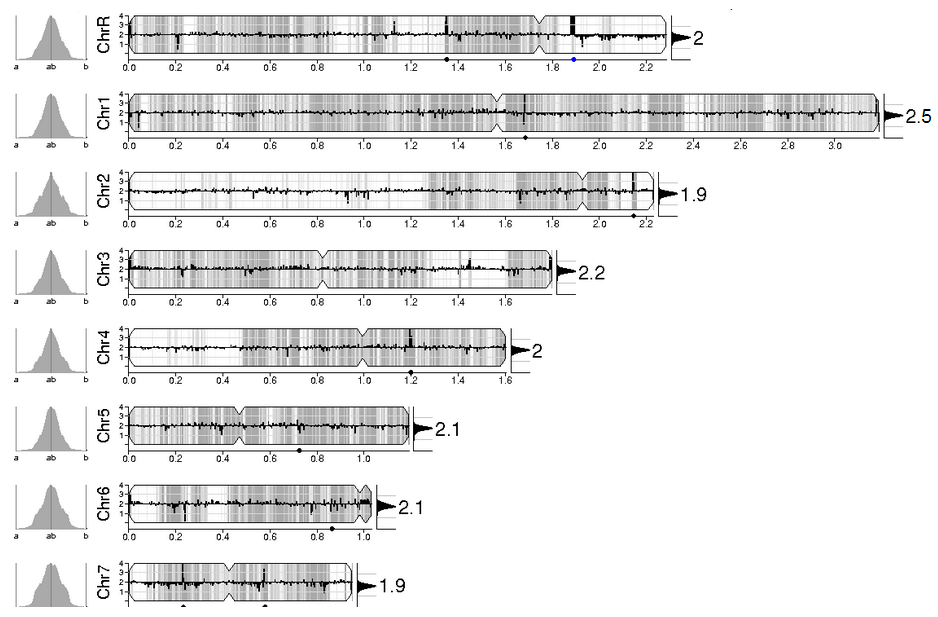
**Interpretation of results**

There are two basic visualization options in YMAP: A horizontal figure or a standard figure, where the chromosomes are displayed one above the other. In the horizontal figure each chromosome is illustrated to scale as in Abbey et al. 2011, with the position of its centromere indicated by an indentation in the appropriate place. CNVs per position are displayed as black histograms along the length of the chromosome. The numbers on the Y axis represent the relative chromosome copy numbers, based upon the whole genome ploidy, determined by flow cytometry. The dots represent the positions of the major repeat sequences (MRS).

Figure 9: A typical horizontal figure generated by YMAP. A. A figure which presents both CNVs and SNPs. B. Figure displaying only CNVs. C. Figure displaying only SNPs and LOH.

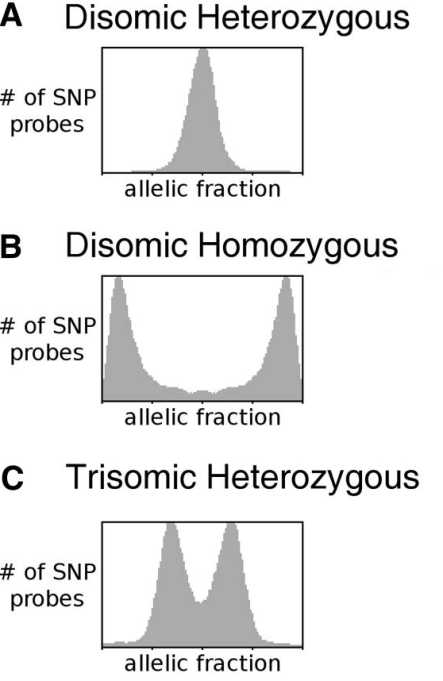


The standard figure is a more detailed version of the horizontal figure (Figure 10). In this figure too each chromosome is illustrated to scale as in Abbey et al. 2011, with the position of its centromere indicated by an indentation in the appropriate place. To make interpretation easier, the X axis represents the chromosomal distance, with 0.2 representing 200,000bp. There is an additional graph on the left of the chromosome representing allelic ratio. On the right hand side there is a single number, representing the ploidy of this specific chromosome, i.e. 2 for two alleles, 1 for a single allele etc. This number does not have to be a full number; ploidies such as 2.2 or 1.5 are also possible and represent cases of aneuploidy. Here too CNVs per position are displayed as black histograms along the length of the chromosome. The dots represent the positions of the major repeat sequences (MRS).

Figure 10: A typical standard figure generated by YMAP. The graph on the left of each chromosome represents this chromosome's allelic ratio.

The allelic ratio displayed to the left of the chromosome can be used to distinguish different statuses of allelic distribution and homozygocity versus heterozygosity. For example, figure 11 shows the different graphs that are generated for a disomic heterozygous (A) a disomic homozygous and a trisomic heterozygous. It is plain to see that each ratio is translated into a very different graph, allowing for a quick and easy indication of interesting characteristics of such data sets.

Figure 11: Allelic fraction indicating different allelic rations. Figure adapted from Abbey et al. 2011. A: Diosomic heterozygous; B: Disomic homozygous; trisomic heterozygous.

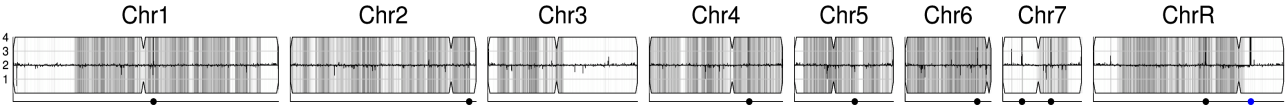
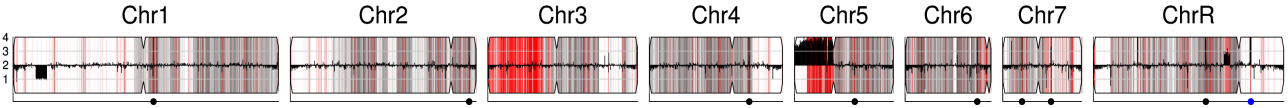
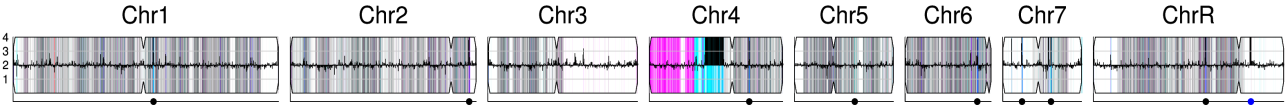
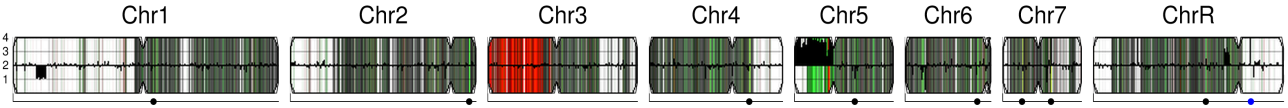


There are several display styles available in YMAP. These styles depend on the type of information chosen to be analyzed.

The first option is the default option, in which no reference strain or hapmap are defined. In this case, the SNP distribution for the strain of interest is displayed as vertical grey bars in the background of each chromosome (Figure 12A).

The second option is used to compare the SNPs found in a defined parent strain and the test strain. If a SNP is shared by both strains, the SNP will appear as a gray bar, while SNPs that have different allelic ratios in the test strain are displayed in red if allelic ratios approach 0 or 1 (Figure 12B). If allelic ratios suggest unusual allelic ratios, such as can be found in the case of aneuploidy, the ratios will be displayed in green (Figure 12C).

The third and last option can be used in cases where a hapmap of the parental strain is available. In this case, two new colors are in use, cyan and magenta, to help distinguish cases where the chromosome has become homozygous. If a SNPs remain homozygous, it is marked in gray, while any homozygous allele will be either colored cyan, representing the "a" allele, or magenta to indicate the "b" allele (Figure 12D).



A.

B.

C.

D.

For the default option, any coordinates with an allelic ratio near 0.5 (0.50 to 0.75) are considered heterozygous. More extreme allelic ratios are considered to be homozygous, appearing in the dataset due to sequencing errors. The density of heterozygous SNPS is presented as vertical lines spanning the height of each chromosome cartoon, with the intensity of grey color representing the number of SNPs in each 5000 bp bin.

**Exporting SNP data into the CGD website:**

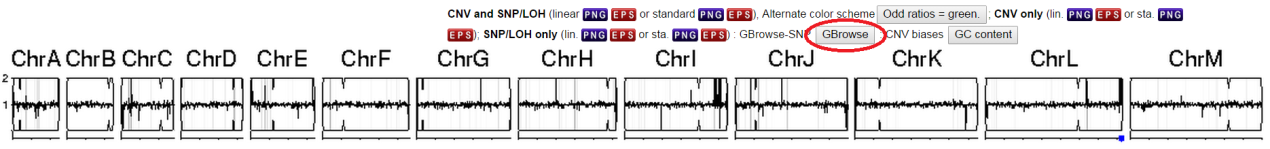
To better understand the SNPs found by YMAP it is sometimes necessary to take a closer look at the specific genes that these SNPs correspond to. It is therefore possible to visualize this data by uploading the SNPs data into the CGD website.

To do so, a user must first click the "GBrowse" button located right above the horizontal figure (Figure 13, circled in red). A new window will appear, containing the exact positions of each SNP located in the analysis, for example: "ChrA\_C\_glabrata\_CBS138 34 34 [\*/\*] 0 + 34 34 170,170,170".

This information should now be copied and saved into a TXT file. It is recommended to do so as follows:

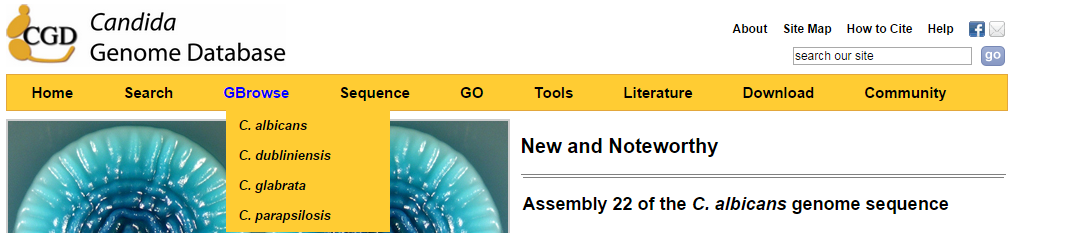
Click Ctrl+A, then Ctrl+C. This should copy the entire content of the new window. Now, to to "Notpad" and open a new TXT file. Click Ctrl+P to paste the entire content of the window into this new file. This file should now be saved in a comfortable location on the users computer.

Figure 13: Location of the GBrowse button.



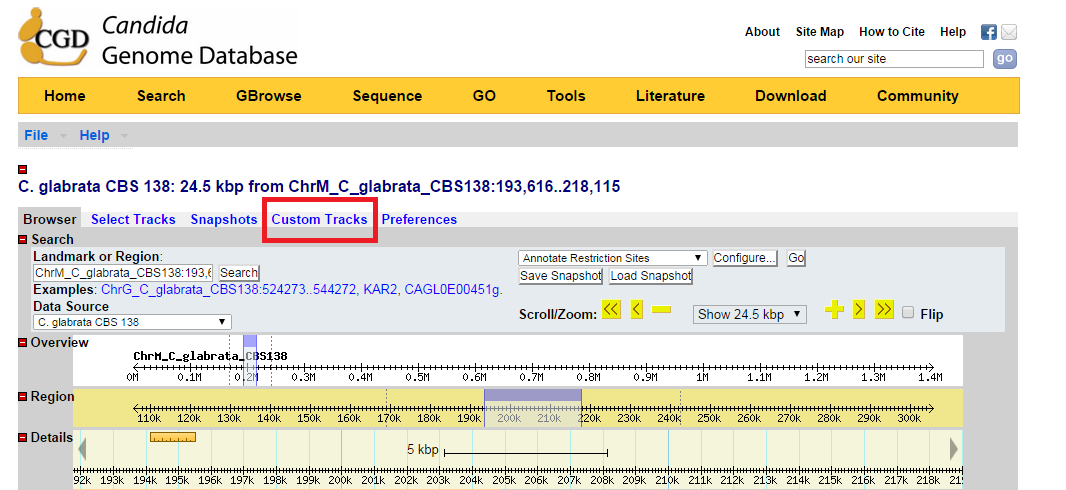
Now user should open the CGD website (<http://www.candidagenome.org/>) and go to the GBrowse tab. Please choose the relevant species, in this example it is *C. glabrata.*

Figure 14: Location of the GBrowse button in CGD.



On the top bar, above the figure, there are several options, one of them is the "Custom Tracks" option (Figure 15, red square). Clicking it will allow the user to add the SNP file to the genome figure shown below.

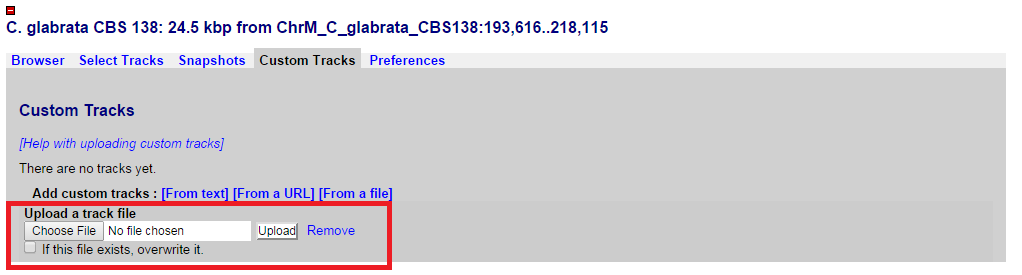
Figure 15: Location of the "Custom Tracks" button in CGD.



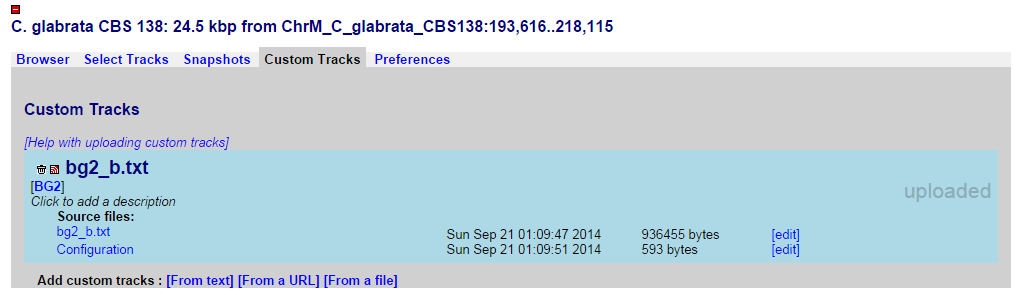
In the bottom of the gray box, go to "Add custom tracks" and choose the "From a file" option (Figure 16, red square).

Figure 16: Location of the button which allows choice of the type of file to be used in the "Custom Tracks" tab.

A new option would appear, allowing for the SNP file to be chosen and uploaded to the CGD website. Once clicking the "Upload" button, user should wait patiently until the file is uploaded into the website (Figure 17, red square).

Figure 17: Location of the button which allows choice of file in the "Custom Tracks" tab.

When the upload is done, the screen would indicate it by showing the file name inside a blue square (Figure 18).

Figure 18: A successfully uploaded file in the CGD "Custom Tracks" tab. 

Now, upon returning to the browser itself, the different SNPs should now be indicated by "\*" inside the genomic figure, indicating each SNPs position on the different chromosomes. It is therefore possible to utilize all of CGD's available tools to analyze these SNPs and their meaning.

Figure 19: Appearance of SNPs in the CGD browser, once they have been assigned as "Custom Tracks".

