

More than mapping.



Genomatix Mining Station

Quickstart Guide

Welcome...

... to the Genomatix Mining Station manual. The Genomatix Mining Station (GMS) is Genomatix' ultra high speed solution for genomic positioning, genotyping and splice analysis.

It provides read mapping, advanced de-novo SNP detection, splice junction identification, low frequency copies, correlation with known SNPs, haplotype analysis, general statistics and quality reporting.

The features of Genomatix Mining Station include:

- Unlimited read length
- Alignment with point mutations, insertions and deletions
- De novo detection of splice events from single sequence reads (spliced alignment)
- Genome wide gapped alignment
- Advanced handling of repetitive reads
- Mapping of bisulfite sequencing data
- Support of sequence space as well as color space
- Mapping speed of up to 2 million bases per minute to the human genome
- Dedicated special purpose built hardware with UNIX operating system

Proprietary Genome Library genomic pattern database technology delivers regional information content for each genomic / transcriptome position and gives comprehensive information for each sequence read.

The Genomatix Mining Station supports all established NGS sequencing platforms including:

- SOLiD™ System by Applied Biosystems™
- 454 Life Sciences™ platforms (a Roche Company)
- Genome Analyzer, HiSeq and MiSeq platforms by Illumina™
- IonTorrent platforms by LifeTech

If there are any open questions that are not covered in this manual you are always welcome to contact us

via email: support@genomatix.de

via phone: +49 89 599766 0

US or Canadian customers are welcome to contact: support-us@genomatix.com

We're confident that the Mining Station will help you reach your scientific goals quicker and more easily.

Thank you for choosing a Genomatix product.

The Genomatix Team

Introduction	5
Prerequisites (aka "to-dos before you start!")	7
Sequence data analysis	9
Creating a project and importing sequence data	10
Looking at sequence statistics	13
Starting a mapping	14
Mapping statistics	16
Read classification	19
SNP detection	22
Exporting your results	25
Literature	27
Mapping library creation	29
Creating a new library	29
Conclusion	33

GMS QuickStart Guide

Introduction

This QuickStart guide is intended to introduce you to the graphical user interface of the Genomatix Mining Station (GMS). It will walk you through two of the most common tasks by showing you how to analyze data from a sequencing experiment and how to build a user defined library that can be used for mapping.

Prerequisites (aka "to-dos before you start!")

To access the graphical user interface of the GMS you will need a username and password.

Please note that this username and password need to be set up in addition to any command-line account you may already have on the system. Setting up an account is described in the Sysadmin Guide for the GMS.

You need to know the URL for accessing the user interface. This should usually be the name of your GMS followed by your domain name, e.g. 'http://gms.company.com'.

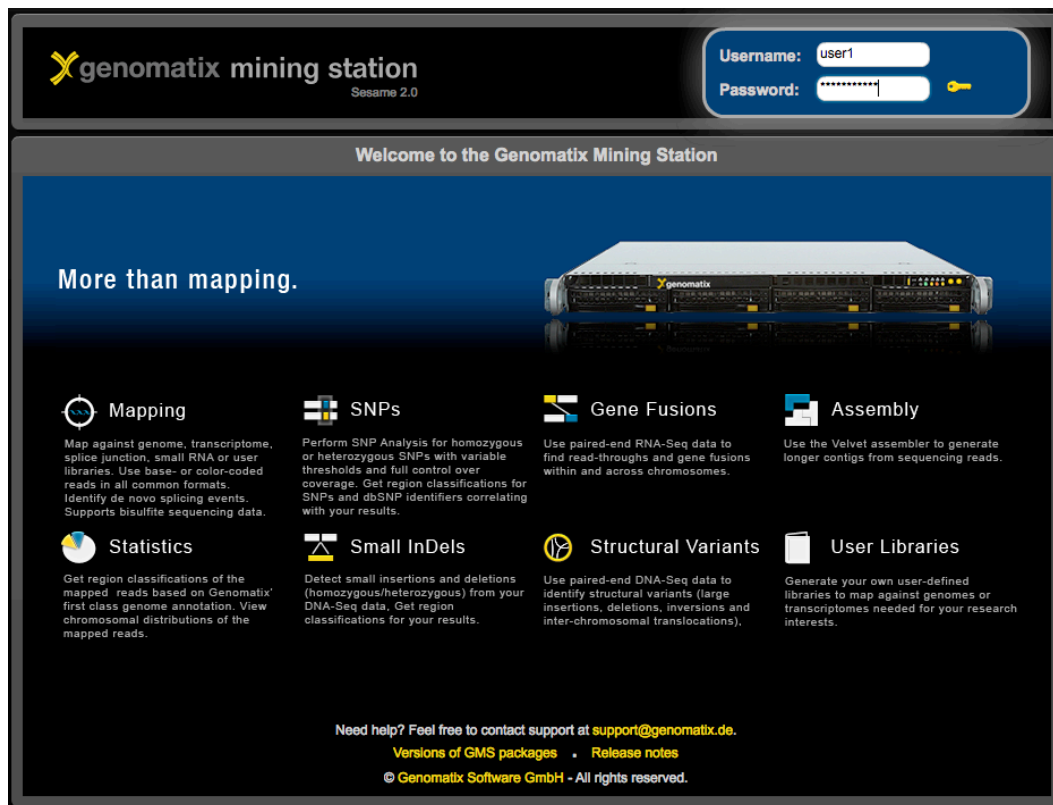
In order for you to be able to run the example, your administrator needs to install the package `gx-gmsdemo` (please consult the Installation guide for instructions). This package will copy the sequence data file `SRR018005all.fastq` into the import directory. It will also generate example results in the public project 'demo' of user 'demo' on your GMS.

Sequence data analysis

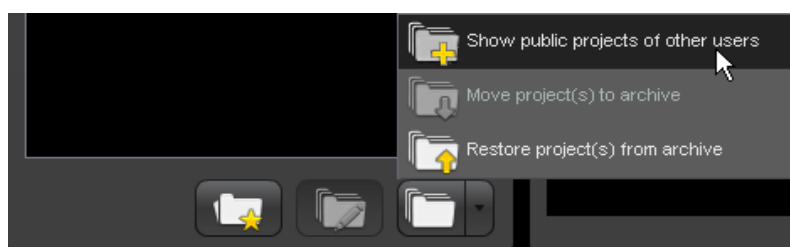
In this analysis example you will create a project, import sequence data, map the imported data, annotate mapped reads, call SNPs based on the mapping, and export some of your results. Please make sure that the gx-gmsdemo package has been installed as described in the previous section before you continue.

The example sequence data were downloaded from the NCBI Sequence Read Archive, experiment SRX005936 (Summerer et al., 2009). The data has been generated by sequencing the HapMap human reference sample NA18558, obtained from a man of Han Chinese origin, on an Illumina GA II sequencer. The sample was enriched for different genomic target sequences by HybSelect, a microarray-based sequence capture method.

Please connect to the GMS user interface in your browser (e.g. by going to the URL <http://gms.company.com>) and log in with your user name and password ('user1' is used as an example here):



The following description shows the course of action for creating the results yourself. If you just want to view the pre-generated results, you can activate the 'Show public projects of other users' option, and then select the 'demo' project of user 'demo':

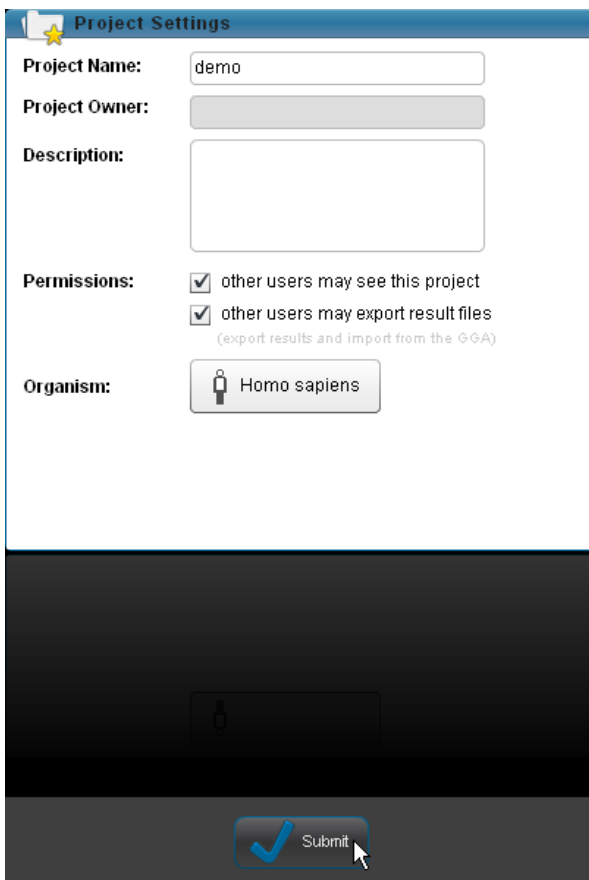


Creating a project and importing sequence data

To follow this guide, please start by defining a project and importing sequence data to it; click the 'Create a new project' button in the lower left hand corner of the screen.



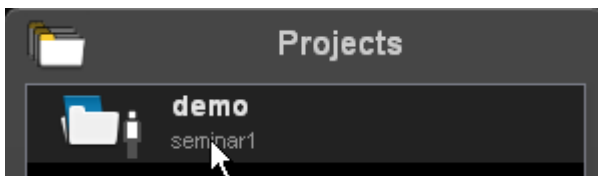
Provide a name for your project in the 'Project Settings' dialog. You can allow other users access to the project and its results by ticking the appropriate checkboxes. The organism is used for pre-setting parameters in your analyses, but you can use sequences from different organisms in any project. Press Submit to generate the project.

A screenshot of the "Project Settings" dialog box. It has a blue header with a star icon and the text "Project Settings". The form contains the following fields:

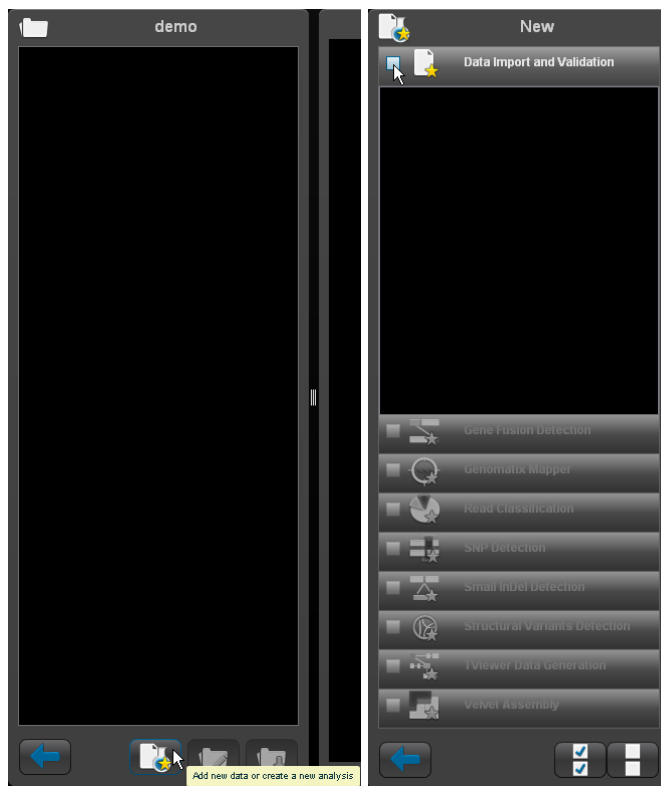
- Project Name:** A text input field containing "demo".
- Project Owner:** A greyed-out text input field.
- Description:** A large empty text area.
- Permissions:** Two checked checkboxes:
 - other users may see this project
 - other users may export result files
(export results and import from the GGA)
- Organism:** A dropdown menu showing "Homo sapiens" with a person icon.

At the bottom of the dialog is a "Submit" button with a blue checkmark icon.

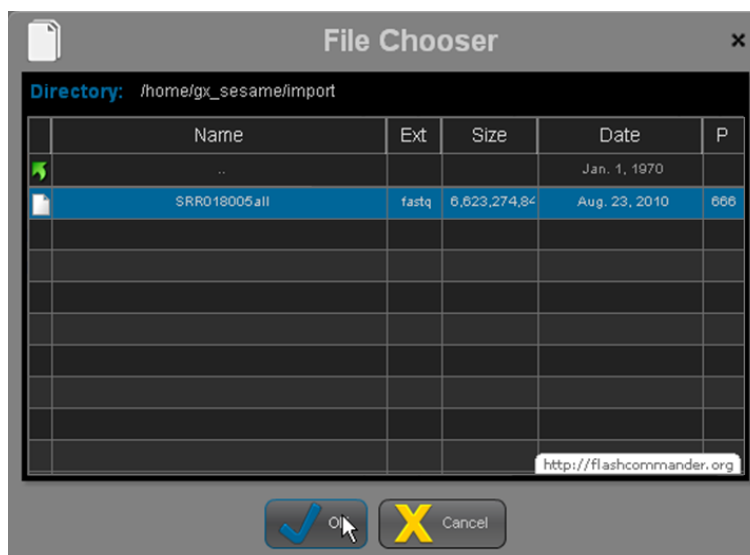
Click on the name of the project in the project list to open it.



Clicking the 'Add new data or create a new analysis' button in the lower left hand corner (see left panel below) gives you access to the analysis menu. Some analysis types depend on output of other analyses; as long as these results are not present, the dependent analysis types are blocked and can't be selected.

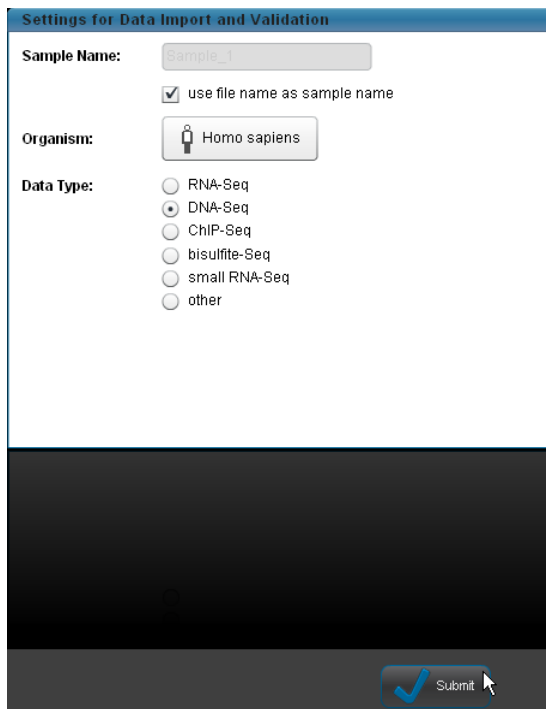


Tick the checkbox in the 'Data import and Validation' section; this will open a file chooser dialog. Select the file 'SRR018005all.fastq' in the directory /home/gx_sesame_import, and click 'Ok':



Sequence data analysis

Select the option „use file name as sample name“ (alternatively, you can provide your own sample name for the file), and select the data type “DNA-seq”; then press Submit to start the data validation.

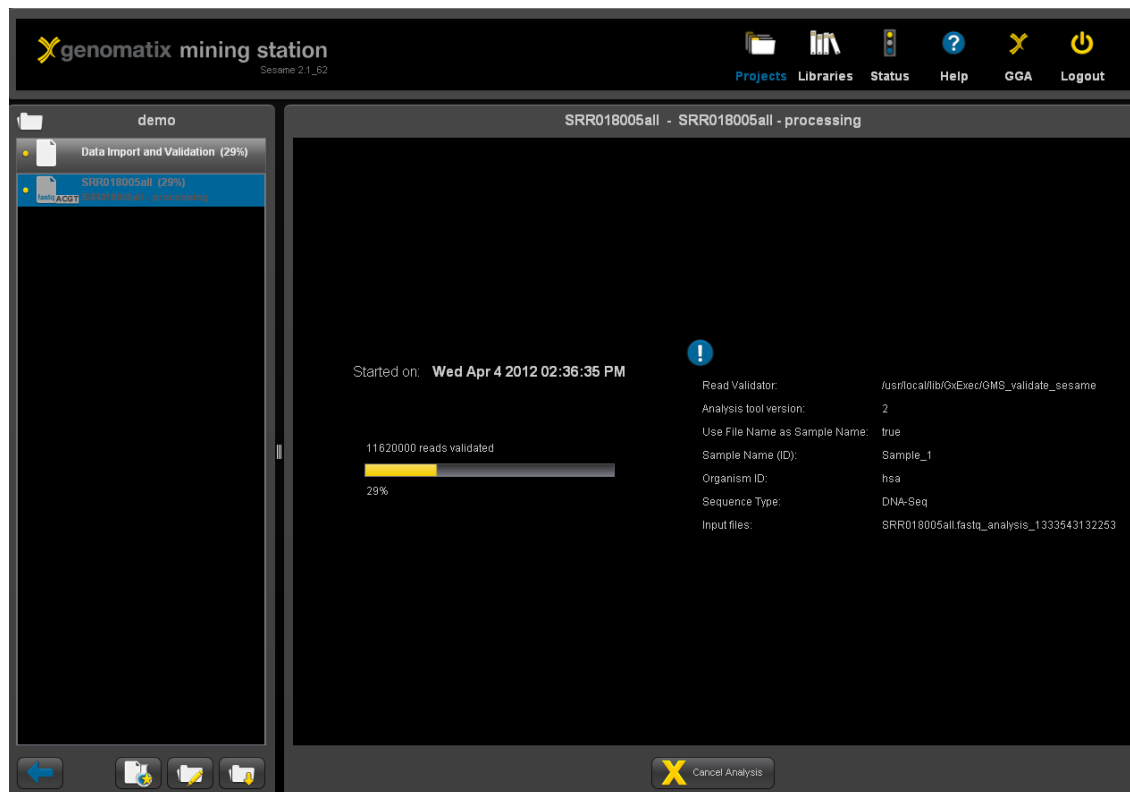


The screenshot shows a dialog box titled "Settings for Data Import and Validation". It contains the following fields and options:

- Sample Name:** A text input field containing "Sample_1".
- use file name as sample name
- Organism:** A dropdown menu showing "Homo sapiens".
- Data Type:** A list of radio buttons:
 - RNA-Seq
 - DNA-Seq
 - ChIP-Seq
 - bisulfite-Seq
 - small RNA-Seq
 - other

At the bottom right of the dialog is a "Submit" button with a checkmark icon.

A progress bar will show the status of the validation. After validation has completed, sequence statistics are displayed.



The screenshot shows the Genomatix Mining Station interface. The top bar includes the logo "genomatix mining station" and navigation icons for Projects, Libraries, Status, Help, GGA, and Logout. The main window is titled "demo" and "SRR018005all - SRR018005all - processing".

The left sidebar shows a file tree with "Data Import and Validation (29%)" and "SRR018005all (29%)".

The main area displays the validation progress:

- Started on: **Wed Apr 4 2012 02:36:35 PM**
- 11620000 reads validated
- 29% progress bar

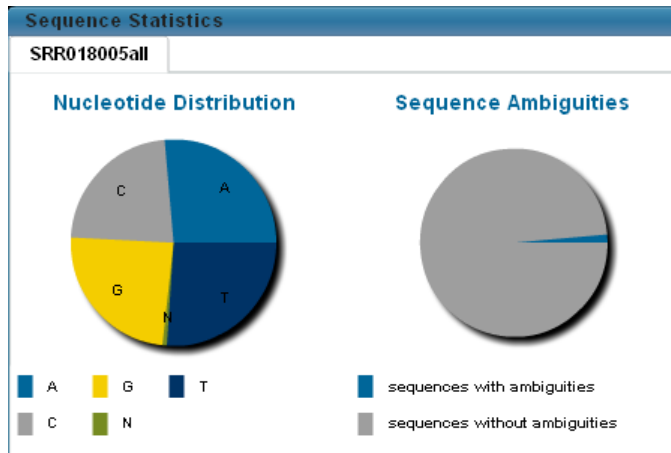
On the right, a table shows the validation parameters:

Read Validator:	/usr/local/lib/QtExec/GMS_validate_sesame
Analysis tool version:	2
Use File Name as Sample Name:	true
Sample Name (ID):	Sample_1
Organism ID:	hsa
Sequence Type:	DNA-Seq
Input files:	SRR018005all.fastq_analysis_133543132253

At the bottom right, there is a "Cancel Analysis" button with a red X icon.

Looking at sequence statistics

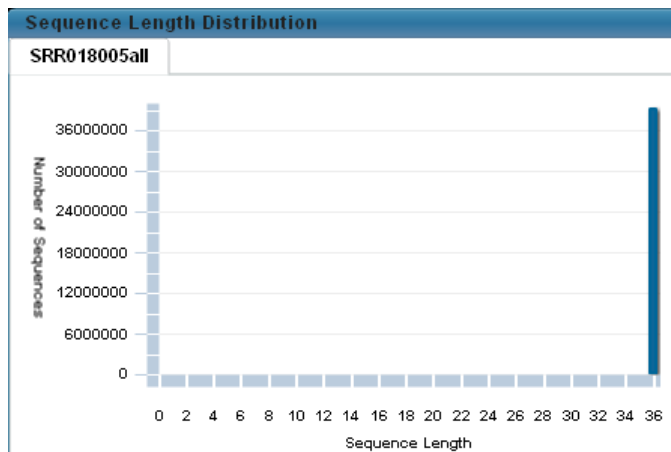
Click on the data set name to display the sequence statistics:



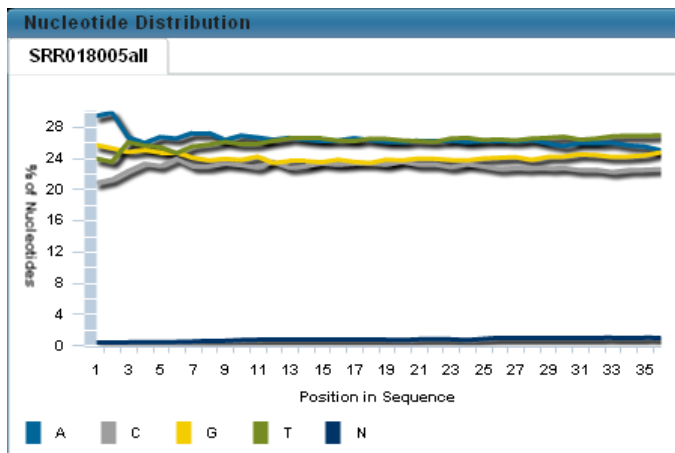
The left pie chart shows the nucleotide distribution in the reads. Positioning the mouse pointer over a part of a chart will display the corresponding numbers in a tool tip. Some numbers are also provided to the right of the graph panel. The average GC content is 47%.

The right chart displays the portions of sequences with and without ambiguities (Ns). Only 1.4% of reads contain Ns, which is fairly low.

The next chart shows the distribution of sequence lengths in the data set. In this case, all sequences are of the length 36.



Click the 'Nucleotide Statistics' button to switch to a graph with the nucleotide distribution at each position in the sequence reads.



The nucleotide distribution is fairly uniform over most of the sequence length, with similar percentages for A and T (blue and green curves), and G and C, respectively, as would be expected. At the first two positions, there is a sequence bias, but this is not an indication of base calling errors (Hansen et al., 2010). Towards the end, G content increases slightly over C; to a lesser extent, also T over A. N is below 1% at each position.

Next, we will map the reads to the human genome.

Starting a mapping

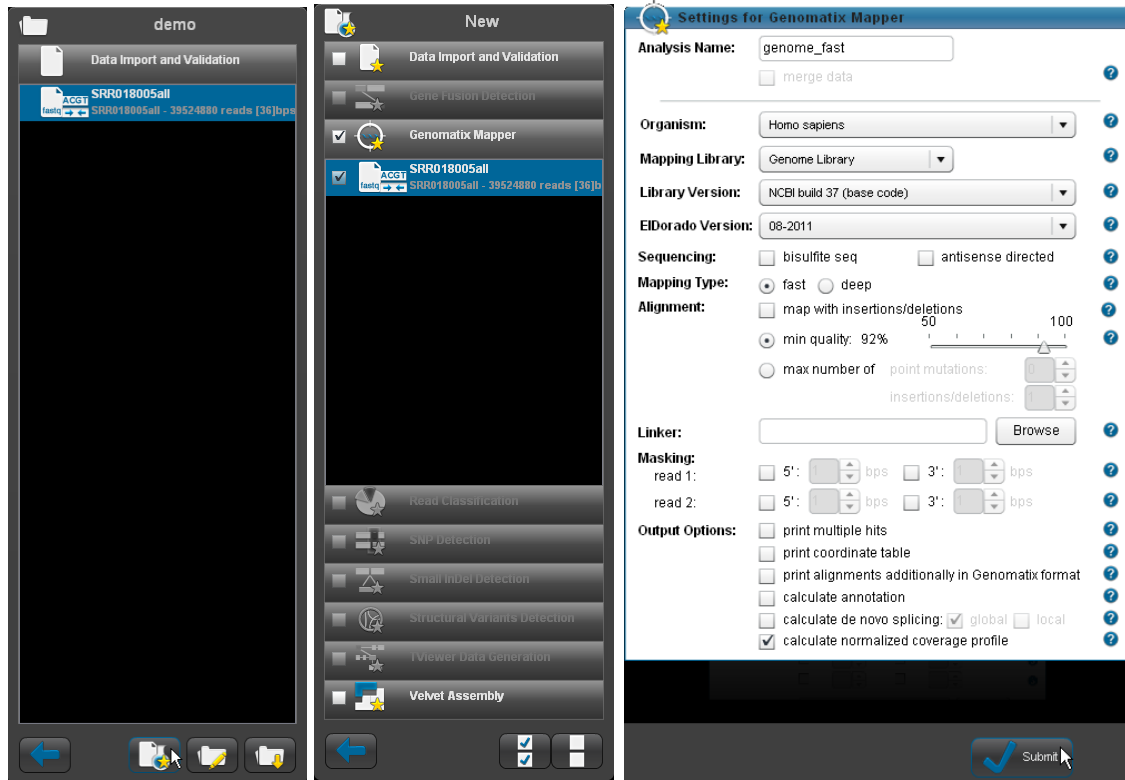
Press the 'Add new data or create a new analysis' button once more, and tick the checkbox in the 'Genomatix Mapper' section; this will display the list of available sequence files and a settings dialog. In the file list, activate the checkbox next to the name of the file you just imported (see center panel below).

In the settings dialog (see right panel on the next page), enter a name for your analysis. The reads will be mapped to the current genomic library of *Homo sapiens*, so you can leave the parameters 'Organism', 'Mapping Library', 'Library Version', and 'Eidorado Version' at their defaults. For 'Mapping Type', please select 'fast' for this example. The first mapping step – the seed search in the mapping library – will then only use perfect matches (for a detailed description of the mapping procedure, please consult the *GMS* manual).

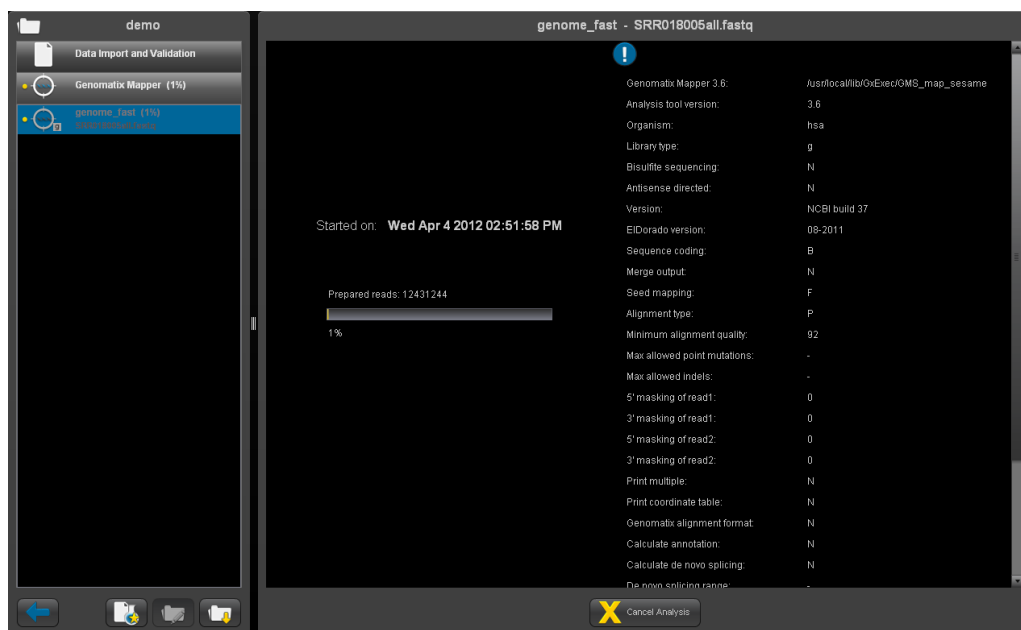
The quality threshold for the second mapping step – the alignment of the complete read – can be set by the 'Alignment' parameter in two alternative ways: you can either set a minimum quality threshold or specify a fixed number of allowed mismatches. For this example, please use a minimum alignment quality of 92%, and leave the 'map with insertions/deletions' checkbox empty. As the reads in our data set have a uniform length of 36bp, this will be equivalent to allowing up to 2 point mutations per read.

Masking can be used to cut off a number of nucleotides from either end of the reads, e.g. linker sequences or low sequence quality regions, which would strongly decrease mapping efficiency. In the nucleotide distribution statistics, there was no evidence for any of this, so we don't need to mask anything here.

The standard output generated by a mapping run depends on the type of library that is used: mapping to a Genome Library will, for example, always include a BigBED file with the positions of uniquely mapped reads. The 'Output Options' allow you to generate additional result files; please deselect every additional output with the exception of 'calculate normalized coverage profile'. You'll find descriptions of the output options in the GMS manual.

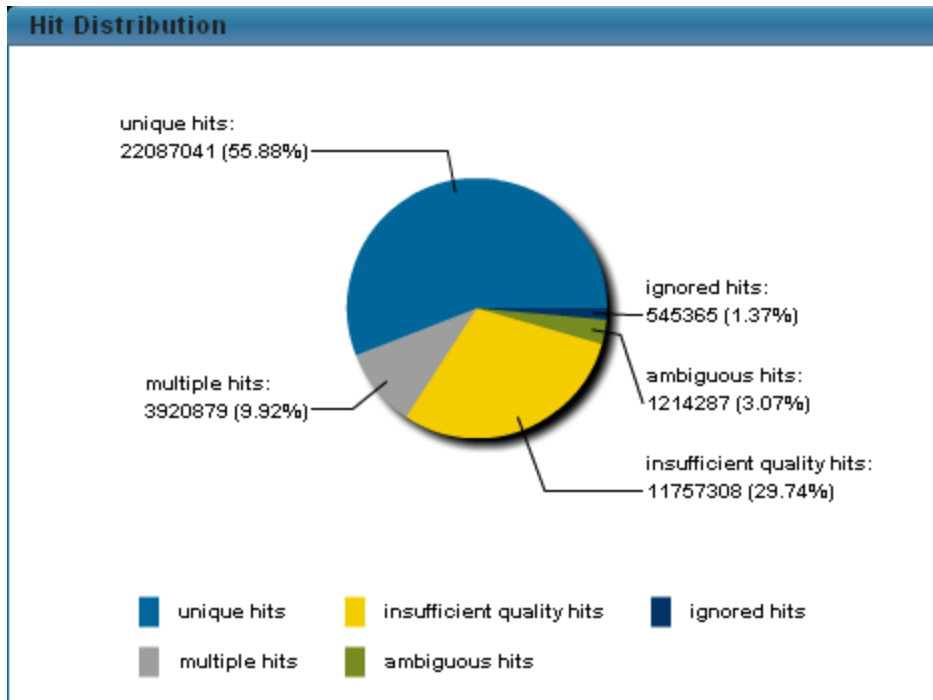


When you're done setting the parameters, click on 'Submit' to start the mapping. The progress and the parameters of any running analysis are shown as below. On a 12 core GMS, this mapping should take about one hour.

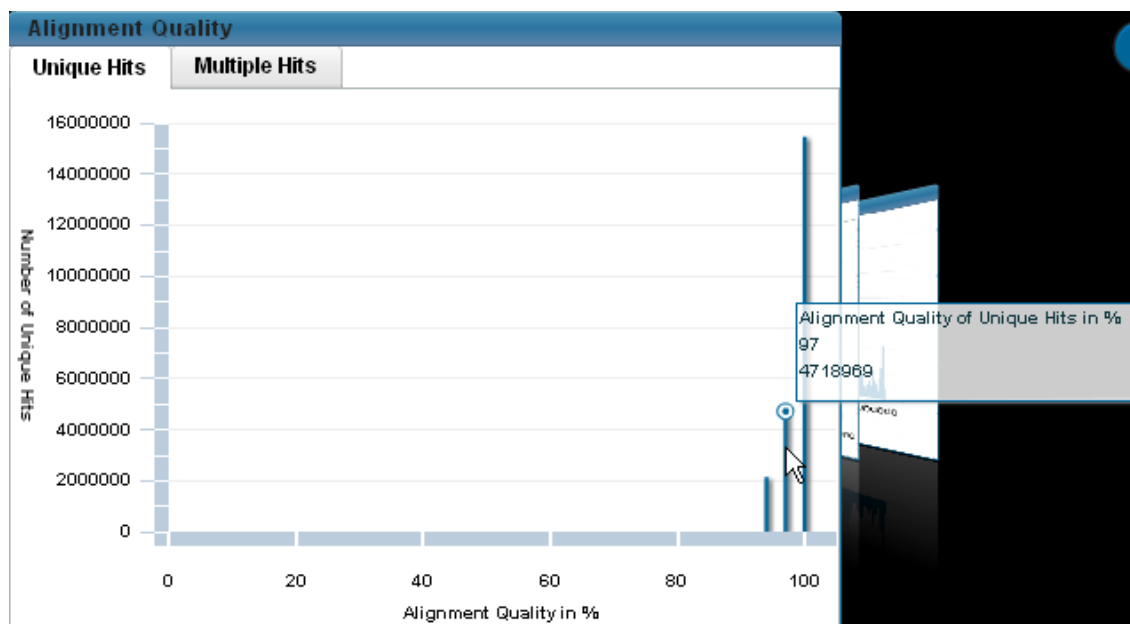


Mapping statistics

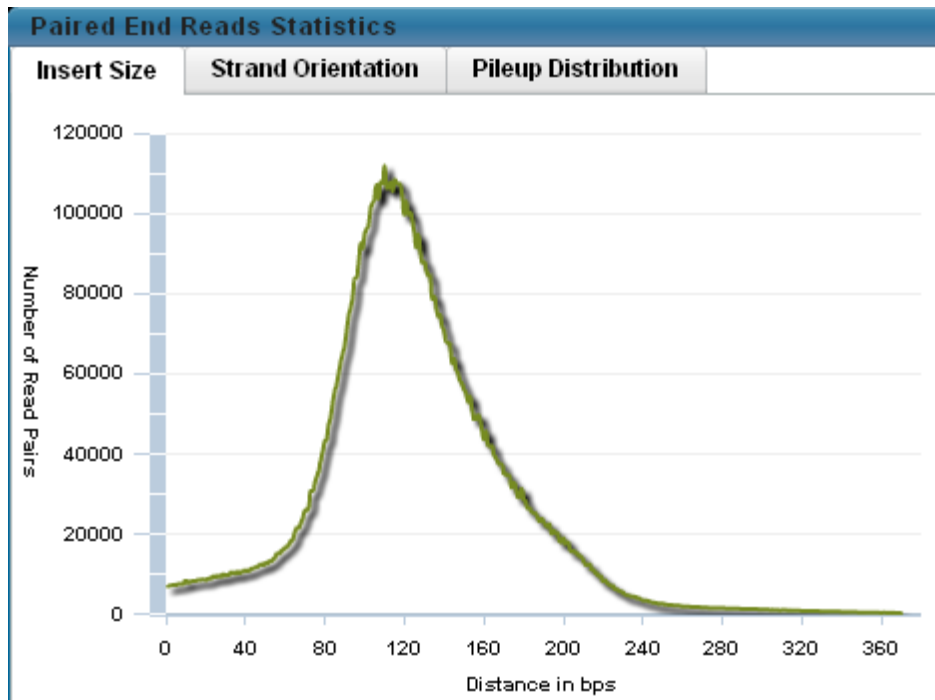
After completion of the mapping, numbers of mapped and non-mapped reads are shown in a pie chart. For ignored hits, no seed could be found in the index; ambiguous hits match more than 50 times with equal best quality in the genome; insufficient quality hits have too many mismatches to pass the specified alignment quality threshold; copy hits have 2-50 equally best matches; unique hits have exactly one best match. The unique hit percentage of over 50% is an indicator for good data quality.



Move the slider below the graph one step to the right to view the alignment quality profile for the unique hits. The majority of reads maps perfectly (rightmost column); additionally, we have reads mapping with one (alignment quality 97%) or two mismatches. Moving the mouse pointer over one of the columns shows the numbers.



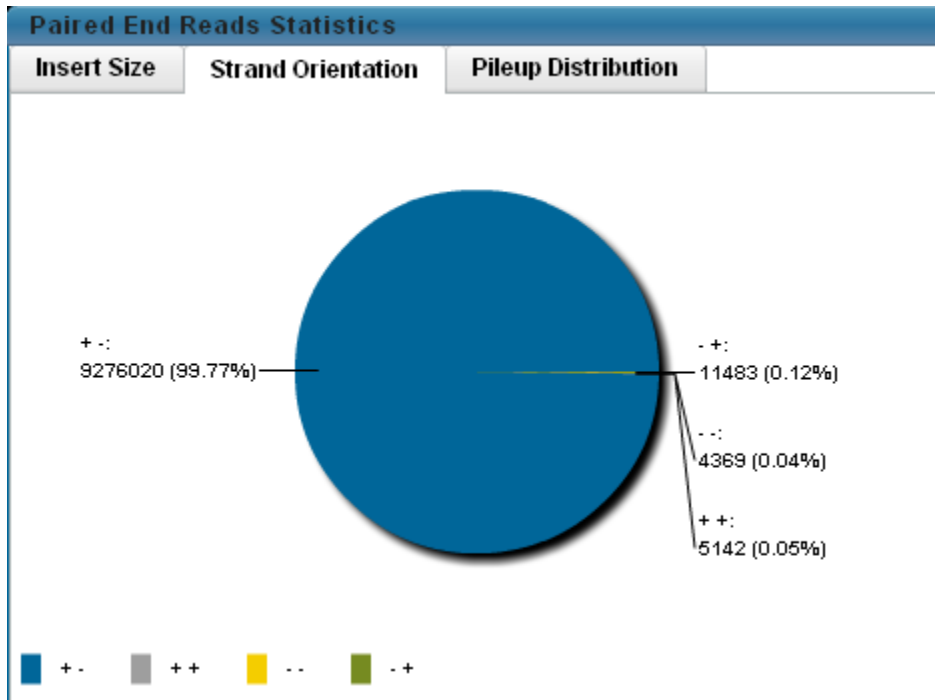
For paired-end data, an extra set of statistics is provided. The first tab shows the insert size distribution:



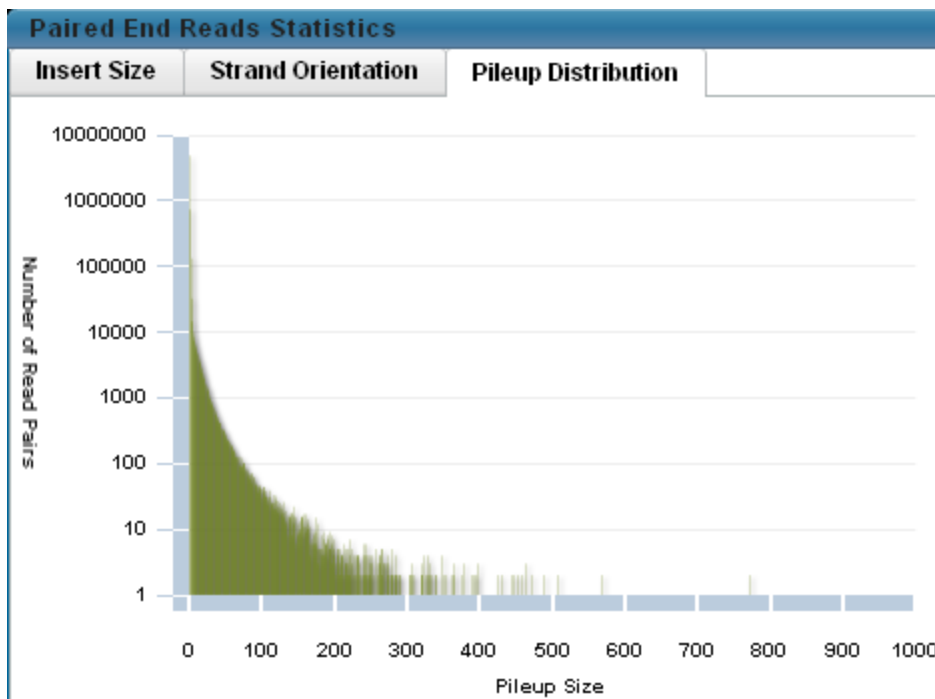
Numbers are given separately:

Min Insert Size:	1 bp
Median Insert Size:	120 bp
Max Insert Size:	242201978 bp
Mean Insert Size:	125 bp
1 Sigma Insert Size:	50 bp (76.64%)
2 Sigma Insert Size:	100 bp (94.87%)
3 Sigma Insert Size:	150 bp (98.74%)
0.05-Quantile Insert Size:	49
0.10-Quantile Insert Size:	74
0.90-Quantile Insert Size:	185
0.95-Quantile Insert Size:	209
0.95-Quantile Pileup Size:	85

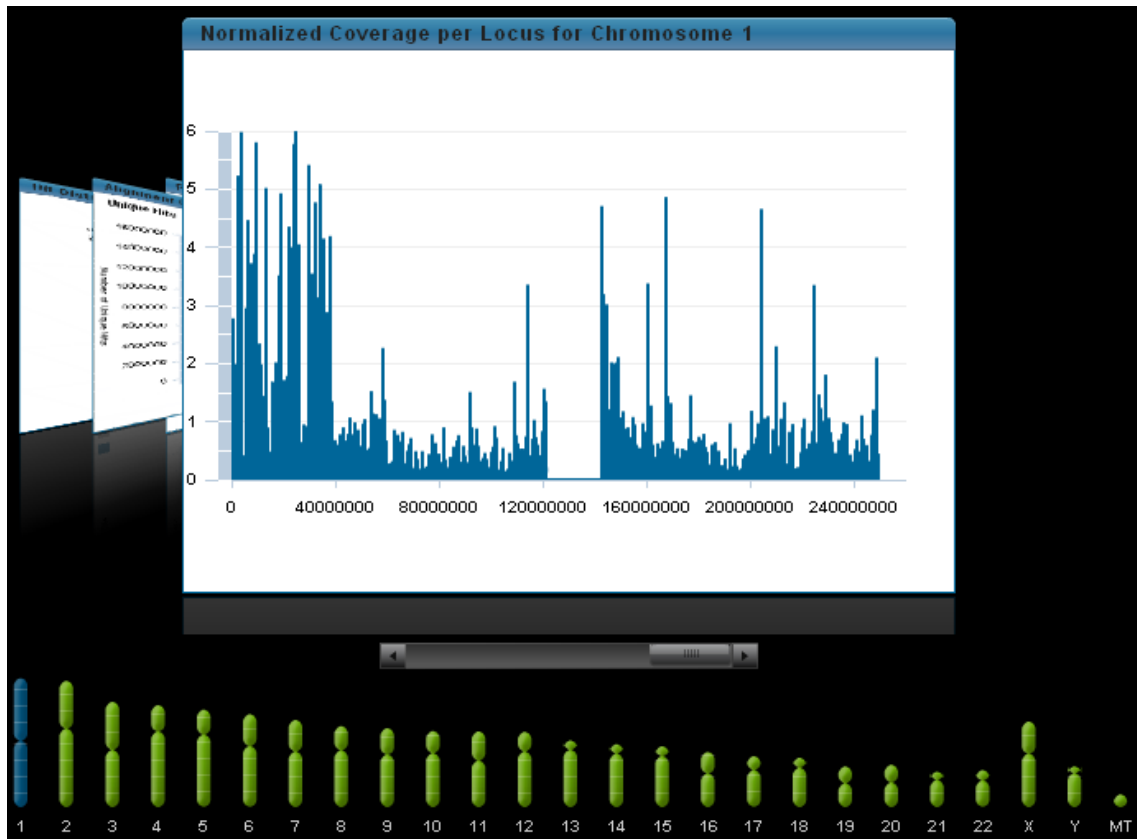
Next is a pie chart with strand orientation numbers for the reads in each pair. It should be identical for almost all reads in your data set; the orientation will depend on the protocol used.



The third tab shows the pileup size distribution. Pileups are reads of identical sequence mapping at identical positions and are normally discarded as artifacts. The 0.95 quantile for the pileup size is generally used as a threshold for determining the maximum allowed pileup size.

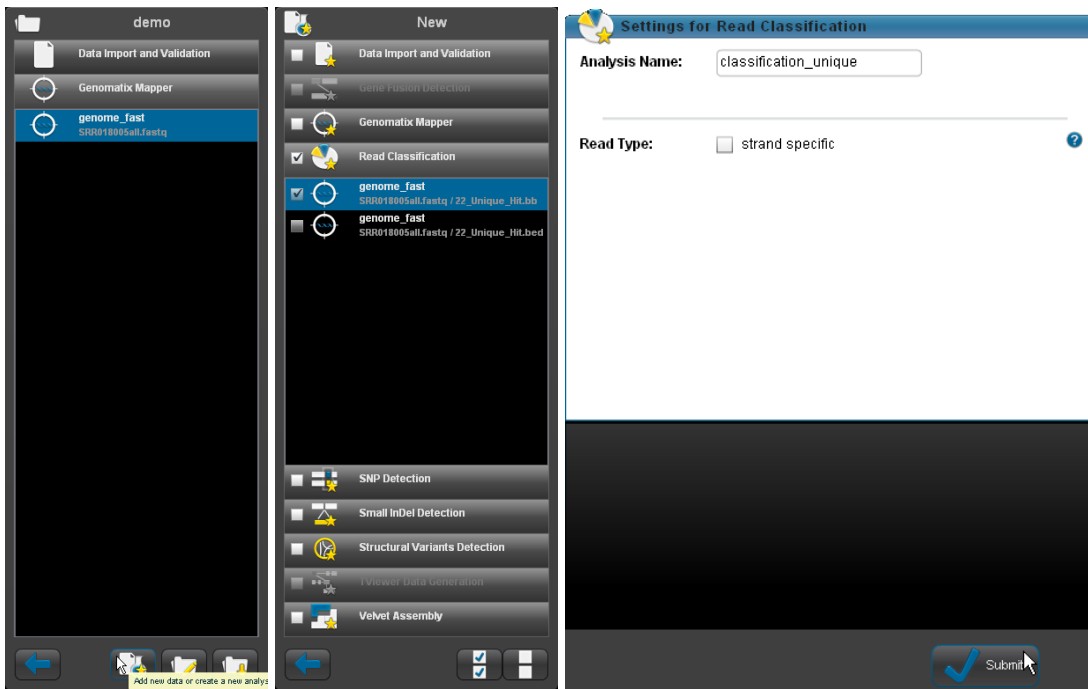


The last graph shows the normalized coverage (NE) per locus for chromosome 1. To see the same for a different chromosome, click its symbol below the slider. NE values are cut off at 6. The start of chr 1 has markedly higher coverage than the rest, which may be due to the targeted enrichment.



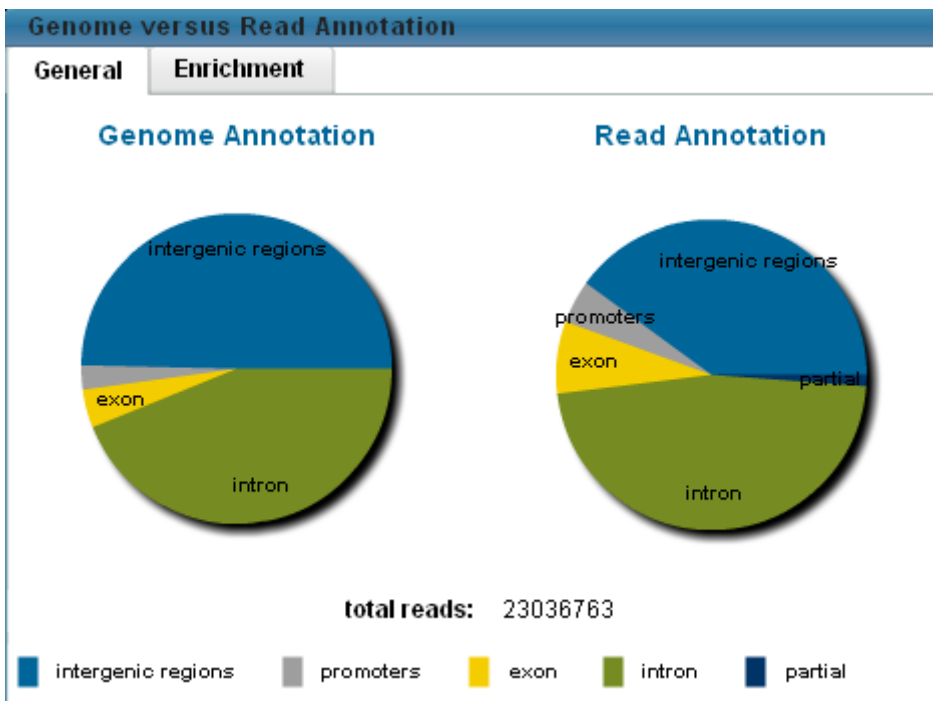
Read classification

Mapped reads can be classified according to the annotation of the region they map to. To run this analysis for your data set, click the new analysis button once more, then select 'Read Classification' and the file containing the unique hits from the previous step as shown on the next page. In the settings dialog, enter a name for the analysis, and click 'Submit'.

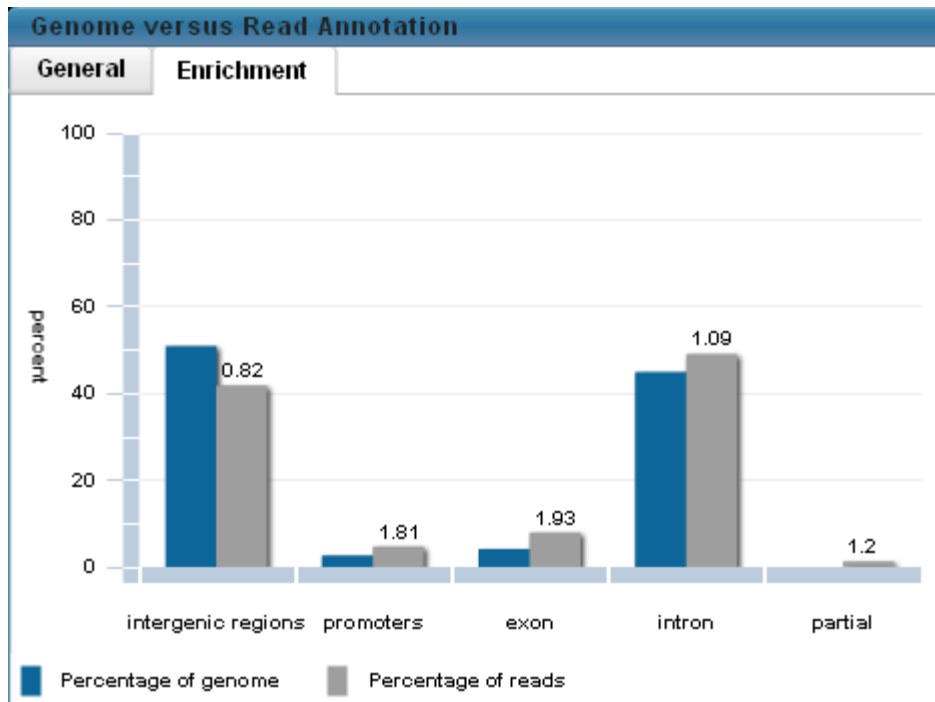


The analysis will take only a few minutes. The output includes a collection of statistics graphs.

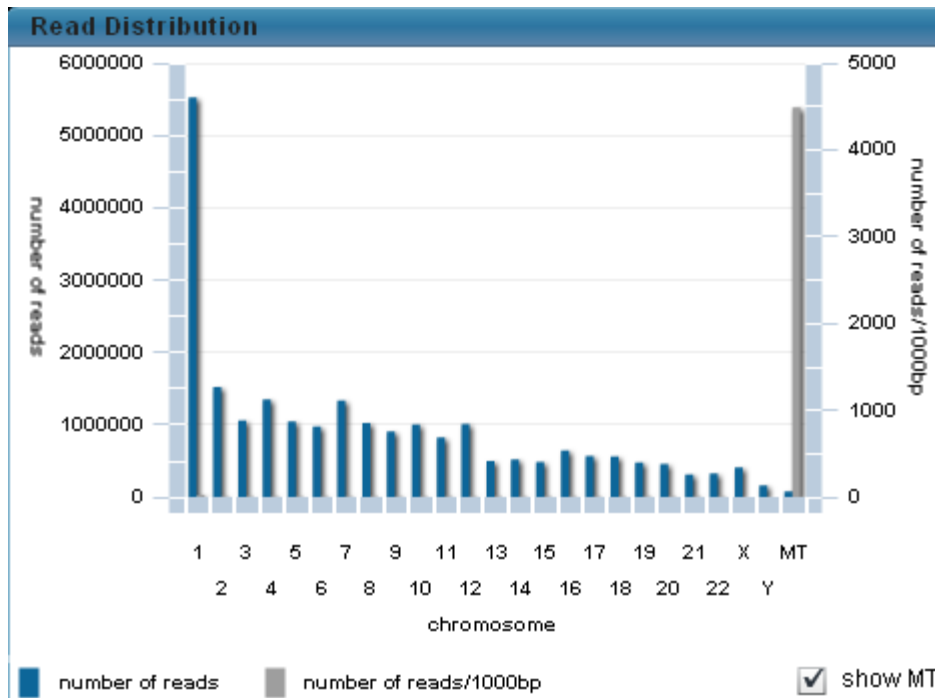
The first tab of the first graph contains two pie charts: one shows the portions of the human genome annotated as intergenic, exon, intron, and promoter in EIDorado. The second chart represents the corresponding distribution of the analyzed reads. 'Partial' denotes reads that partially overlap with an annotated exon. Exons are, for example, slightly overrepresented in the reads. Again, a mouse over shows you the relevant numbers. Percentages for intergenic, exon, intron, and partial add up to 100; promoters come on top of that, as annotated promoters always overlap with other annotation.



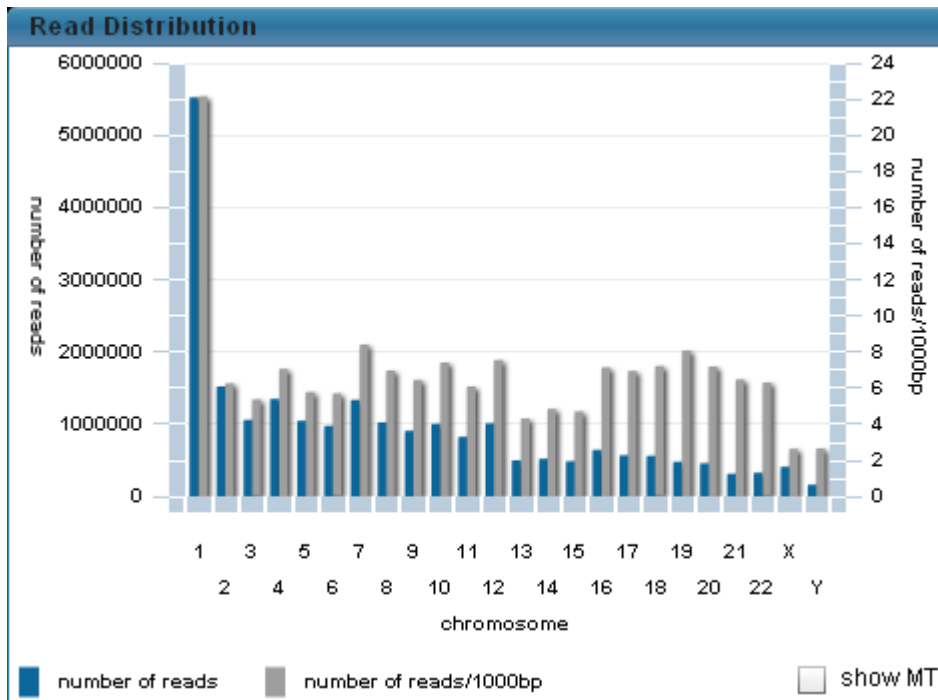
The second tab is a side-by-side comparison of the percentages of each annotation in genome and reads, with fold over-/underrepresentation numbers:



In the last panel, you see the numbers of reads (blue columns), and read densities (grey) for each chromosome. High read densities in the mitochondrial (MT) chromosome result in very small density columns for the other chromosomes.

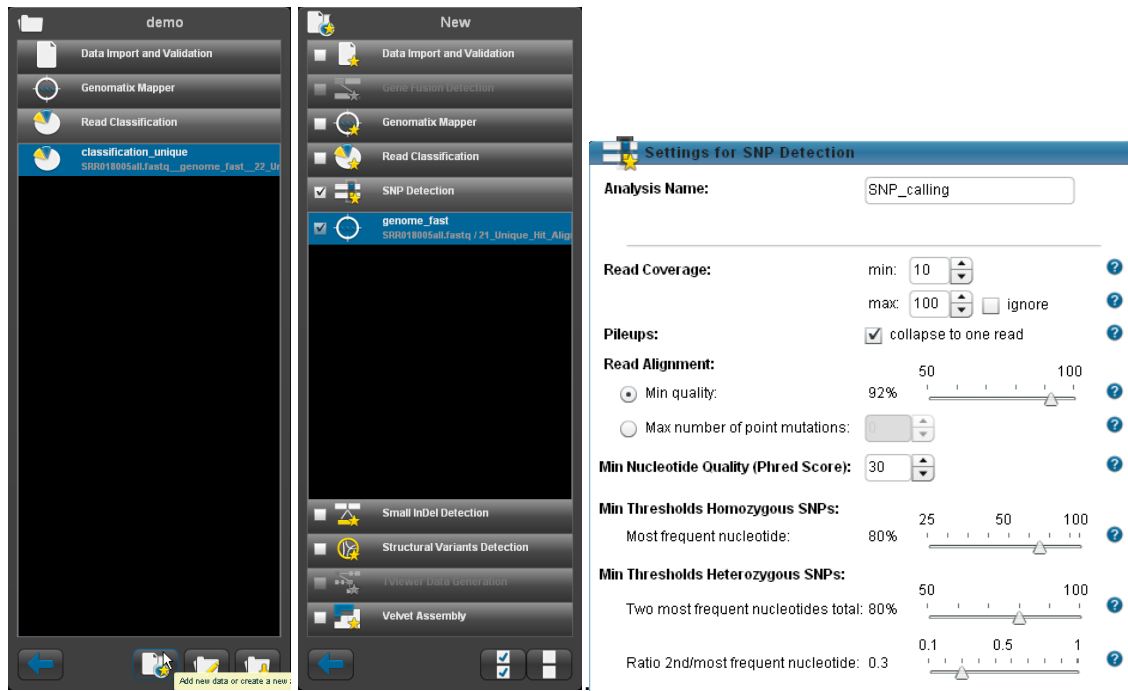


Untick the 'show MT' checkbox to hide the MT values and thus rescale the other density columns. Note that average read density in chromosome 1, which contains regions enriched by HybSelect, is about 4 times higher than in the other chromosomes.

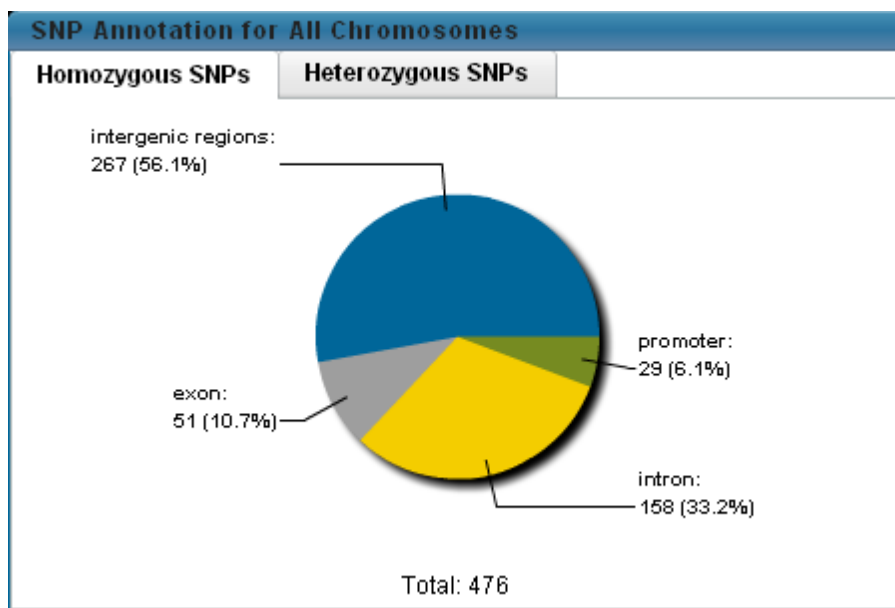


SNP detection

Next, we will use the alignment file from the mapping for detection of single nucleotide polymorphisms (SNPs) in our data. Open the New Analysis menu by clicking the 'New analysis button'. tick the checkboxes for 'SNP Detection' and then for the BAM file in the SNP Detection file list. In the Settings dialog, enter a name for the analysis. For SNP classification, set the minimum frequency ratio of the second to the most frequent allele for heterozygous SNPs to 0.3. Run the analysis with the other parameters at their default values, which are as follows: the minimum and maximum read coverage at a SNP are 10 and 100. Pileups of identical reads at identical positions are collapsed to one read. The minimum alignment quality of reads used for SNP calling is the same as for the mapping, 92%. For calling of homozygous SNPs, the most frequent allele (differing from the reference sequence) must represent at least 80% of the total read coverage at the SNP; for heterozygous SNPs, the two most frequent alleles together must make up at least 80% of the coverage (see panels on the next page).

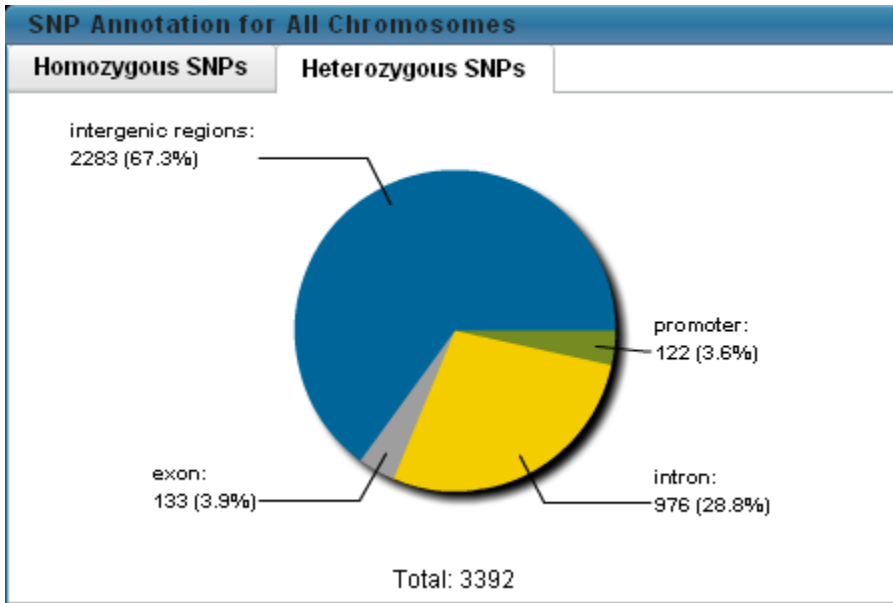


The result includes a number of statistics graphs that you can browse using the small scroll bar below the panel. The first tab in the first chart shows you the annotation of the homozygous SNPs for all chromosomes:

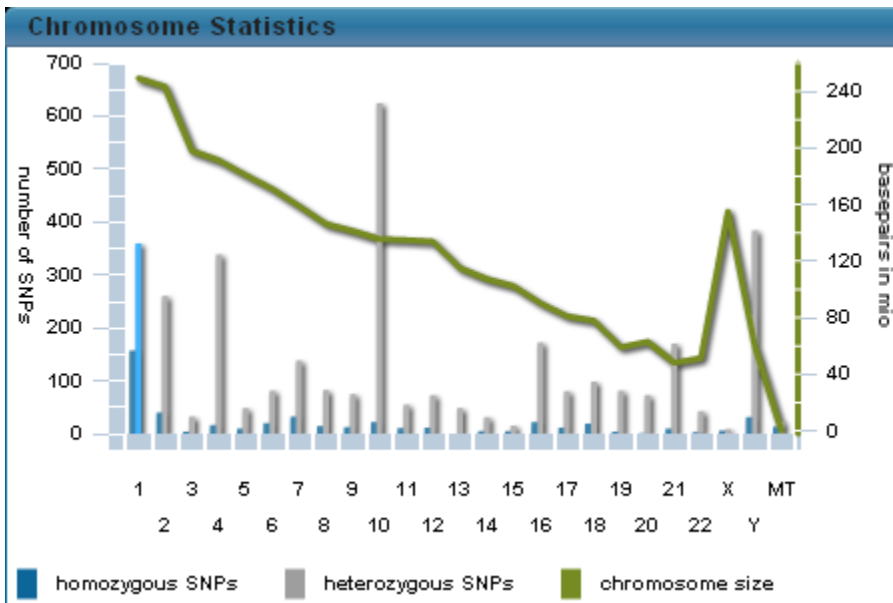


About two thirds of the homozygous SNPs have been detected in intergenic regions, a little over a quarter in exons.

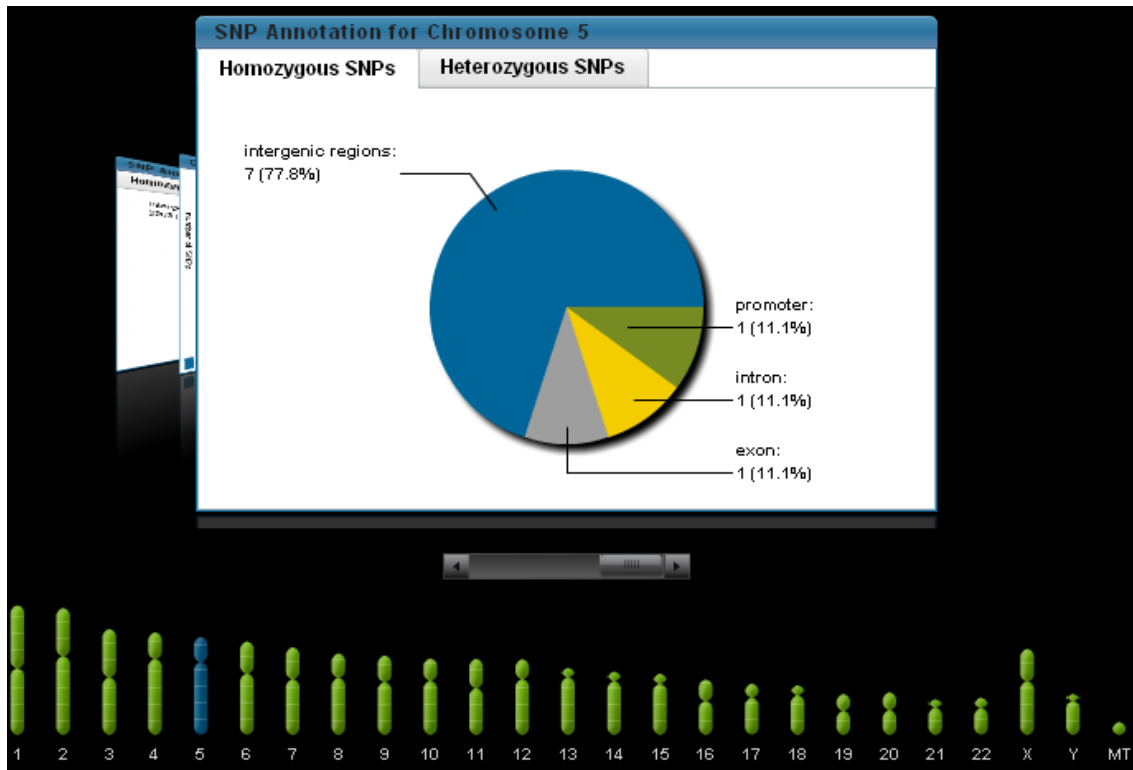
Click the 'Heterozygous SNPs' tab to display the same for the heterozygous SNPs; the distribution is very similar:



The next chart is an overview of the SNP numbers (blue: homozygous; grey: heterozygous) for each chromosome; the green line shows the chromosome sizes. Note that the ratio of detected homozygous to heterozygous SNPs is elevated in the enriched chromosome 1.



Pie charts for the annotation of homozygous and heterozygous SNPs are also available for each chromosome. To display one, browse one or two steps further, and click one of the green chromosome symbols (which will turn blue). For homozygous SNPs on chromosome 5, it should look like this:



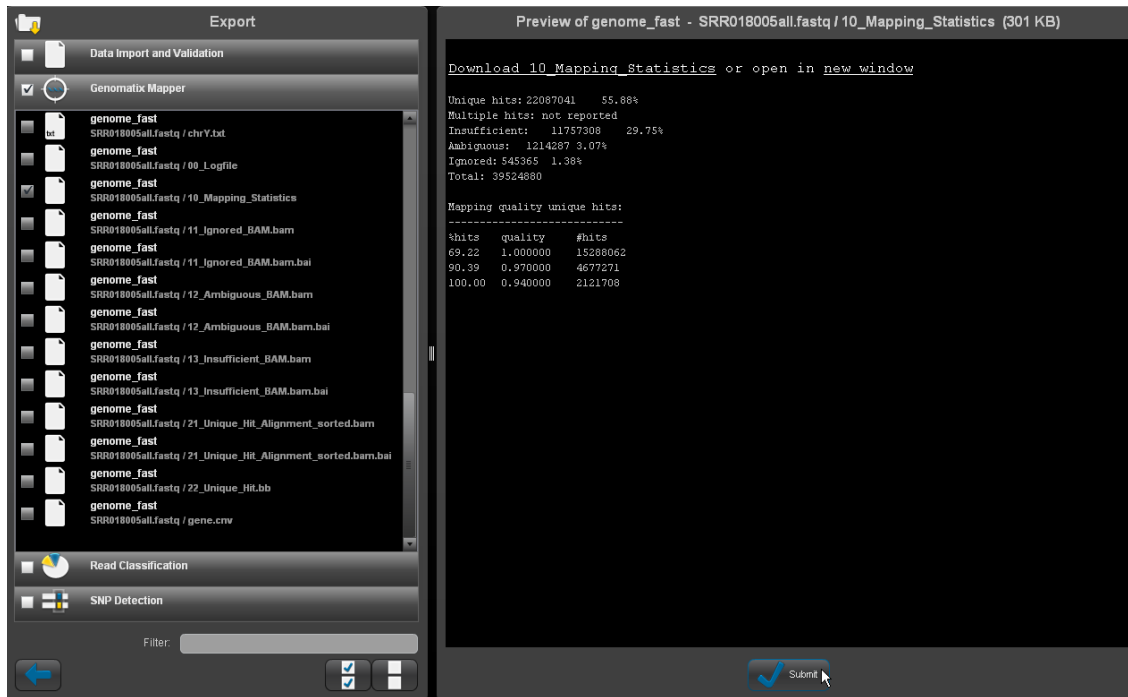
Exporting your results

The GMS GUI shows you mostly statistics graphs for your analysis results. The generated detailed data files, such as those containing the positions of mapped reads, can be previewed and exported for further downstream analysis, e.g. with the Genomatix Genome Analyzer (GGA). Depending on your setup they might be available on the GGA directly, in this case no export is needed.

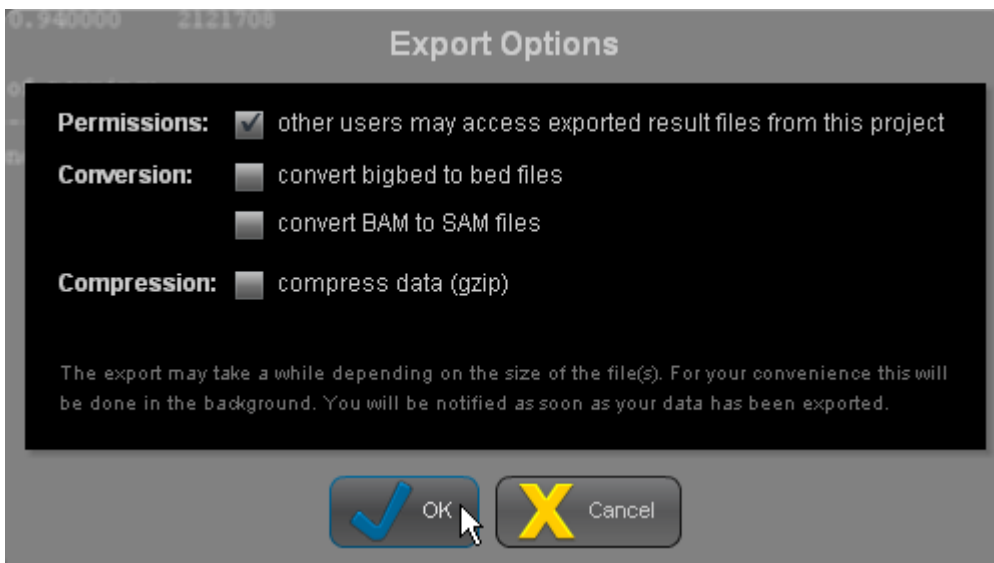
To preview and export result files in the current project, click the 'Export project' button in the lower left hand corner.



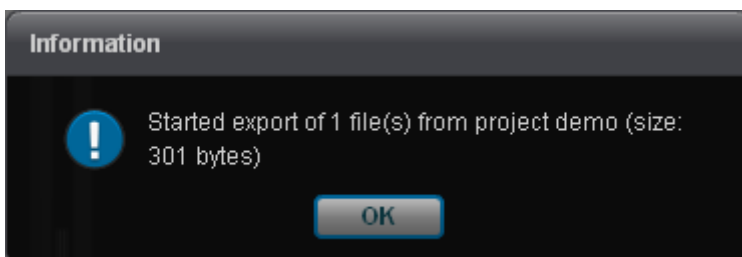
Results in the 'Export' menu are grouped just as in the 'Analyze Data' menu. For a preview, click the appropriate header and then a file name in the list. To select files for exporting, tick the checkbox in the header (e.g. 'Genomatix Mapper' as shown below), then make your selection using the checkboxes in the file list. You can select files from different groups and export them in one go. Some smaller files can also be downloaded directly.

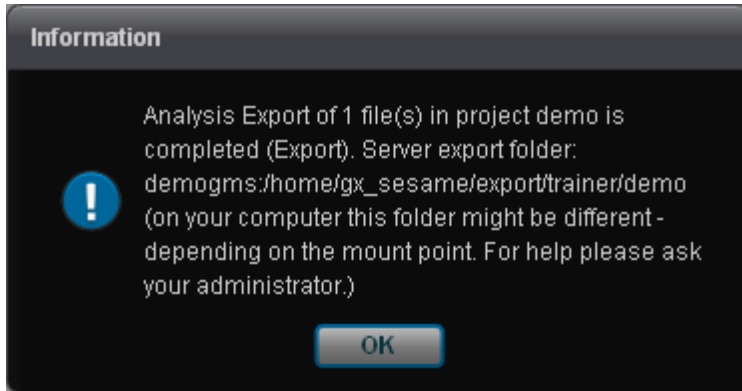


Click on 'Submit' to open an export dialog, where you can set a number of export options, including granting other users access to exported files, file format conversions, and compression of exported data.



The system notifies you when it starts and completes the export:





Exported files can then be accessed in the file system. By default, the results are in the base directory `/home/gx_sesame/export` in a subdirectory structure generated in this pattern: `/<username>/<project_name>/<analysis_type>/<analysis_name>`. Depending on the analysis type, the analysis directory may contain further subdirectories. To view a mapping statistics file using 'less' for example, type:

```
testgms:/home/gx_sesame/export/seminar1/demo# cd genomatiXMapper/genome_fast/SRR018005all/genome/
testgms:/home/gx_sesame/export/seminar1/demo/genomatiXMapper/genome_fast/SRR018005all/genome# less 10_Mapping_Statistics
```

which should give you this output:

```
Unique hits:      22087041      55.88%
Multiple hits:   not reported
Insufficient:    11757308      29.75%
Ambiguous:       1214287  3.07%
Ignored:         545365  1.38%
Total: 39524880

Mapping quality unique hits:
-----
%hits  quality  #hits
69.22  1.000000 15288062
90.39  0.970000 4677271
100.00 0.940000 2121708
```

Literature

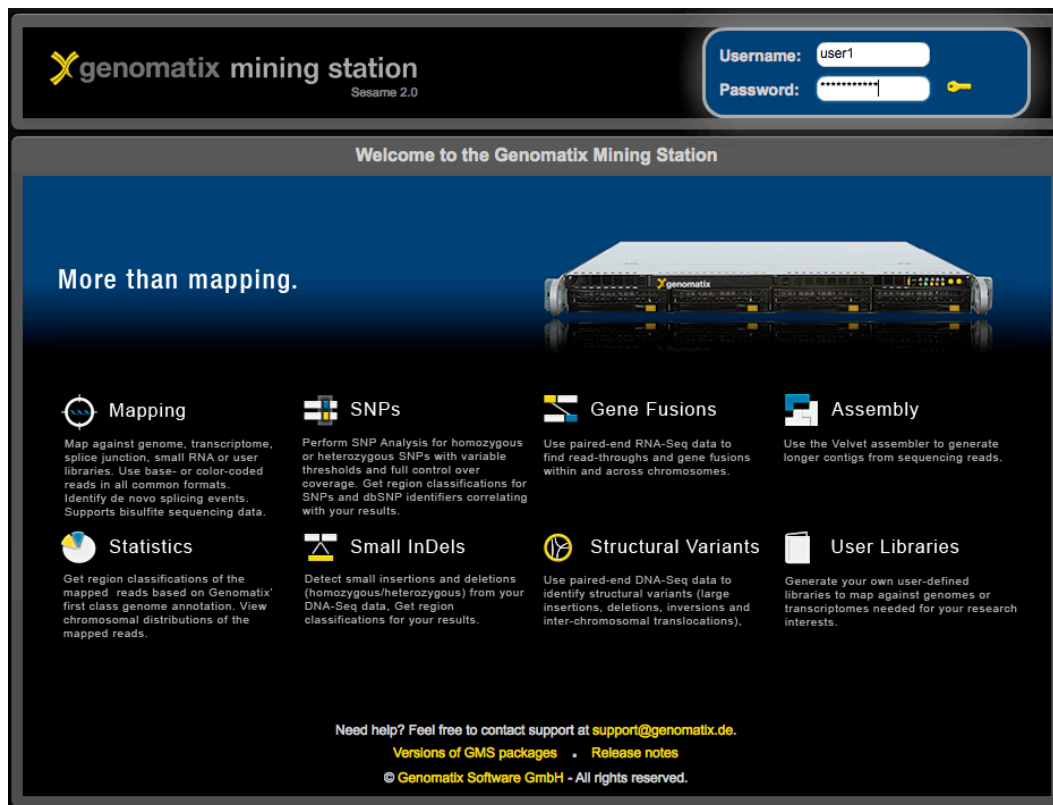
These are the references for the publications cited in this chapter:

- Hansen KD, Brenner SE, Dudoit S (2010)
Biases in Illumina transcriptome sequencing caused by random hexamer priming.
Nucleic Acids Res 38(1), e131
- Summerer D, Wu H, Haase B, Cheng Y, Schracke N, Stähler CF, Chee MS, Stähler PF, Beier M (2009)
Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing.
Genome Res. 19(9), 1616–21

Mapping library creation

The mapping software on the GMS depends on mapping indexes of the target sequences, e.g. genomes. A number of different libraries are available within the system, but you can also create your own index. This example walks you through the generation of an index for the human chromosome 21.

Please connect to the GMS user interface in your browser and log in with your user name and password.

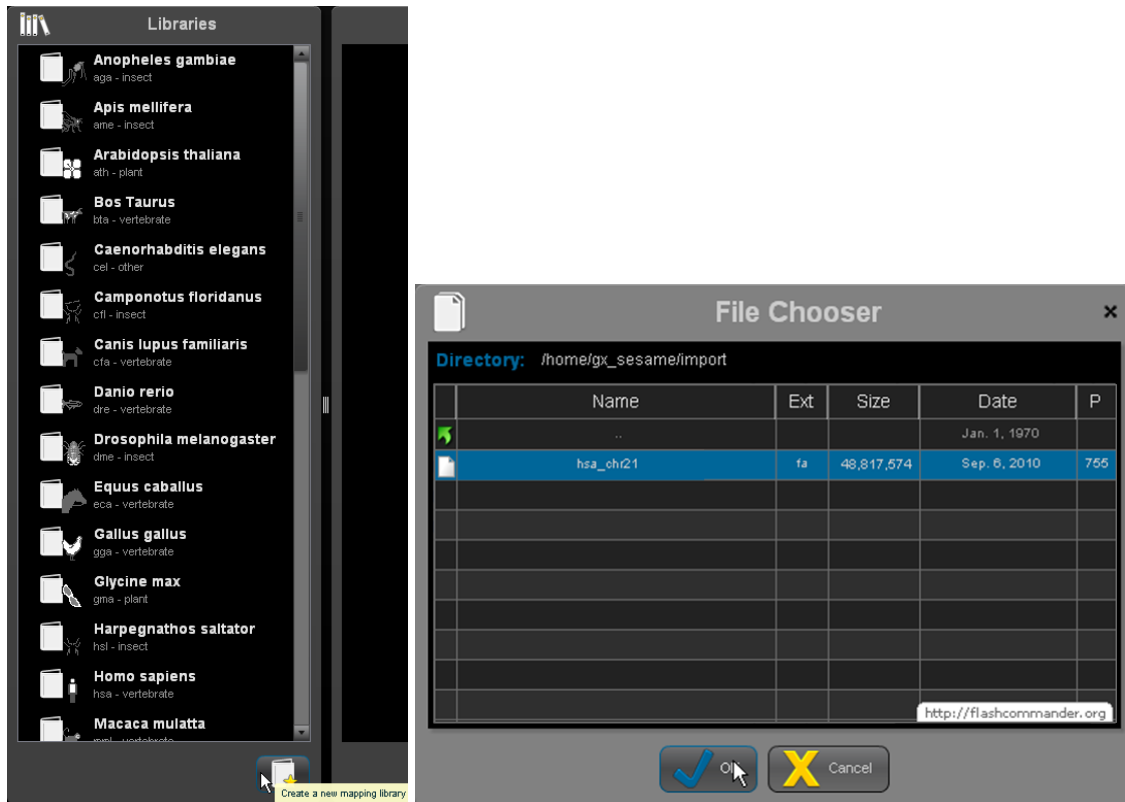


Creating a new library

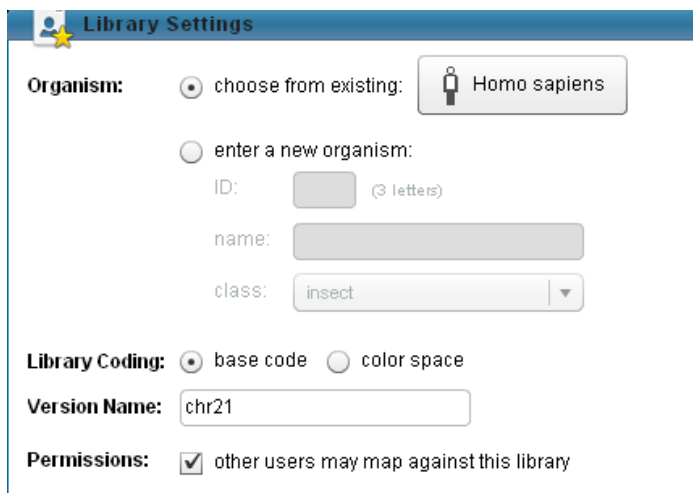
Click on 'Libraries' in the main menu bar



The list of available libraries is displayed. To generate your own library, please click the 'Create a new library' button below the list. In the file chooser dialog, select the file hsa_chr21.fa in the import directory, which contains the base code sequence of human chromosome 21 (NCBI build 37).



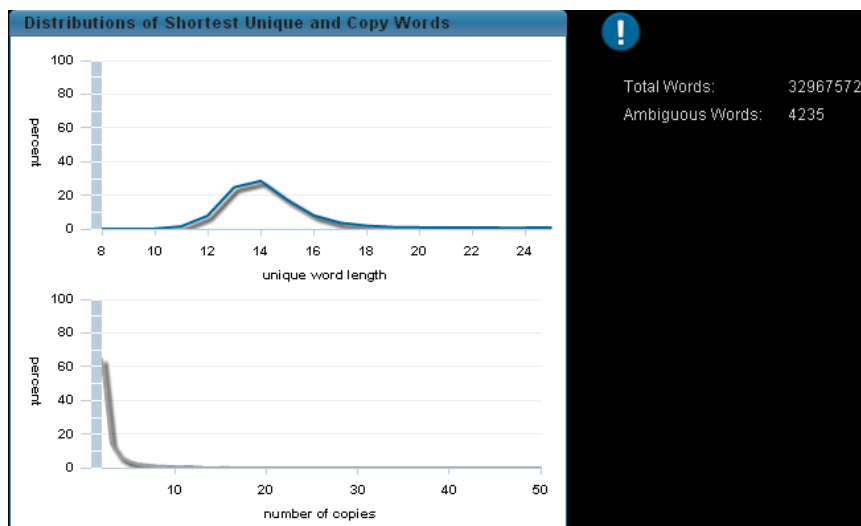
Click 'OK' to open the Library Settings dialog. The organism ID should be 'hsa', 'base code' should be selected for the library coding. Please change the version name to 'chr21'; this name will be displayed in the library list. If you want to allow other users to map against the new library, tick the corresponding checkbox.



When done, click the 'Start Library Indexing' button.

The progress is shown as below:

The generated index can contain both shortest unique words and words with up to 50 copies in the source sequence. Upon completion of the index generation you are shown graphs of the distribution of unique word lengths (top) and of the number of copies for the copy words (bottom). Here, the majority of unique words are between 13 and 15 bp long. Most of the copy words are present in a number of less than 5:



The new library is automatically added to the user library list for Homo sapiens, and will be available as a user library for mapping as shown below:

Conclusion

That's it! You're done with the GMS Quickstart Guide. Thank you for following along all the way. Hopefully you feel familiar enough now to run your own analyses on the system. For a more detailed description of parameters or the underlying methods please refer to the GMS Manual. For immediate help you can also refer to the online help which is accessible via the 'Help' button on top of every page.

For any questions or comments, you're most welcome to send us an email to either

support@genomatix.de

or

support-us@genomatix.com.

