ChIP-Seq Data Analysis with Genomatix® Software

Prepared especially for:
The National Cancer Institute
National Institutes of Health
November 18-19, 2014

For more information please contact:

Genomatix Software, Inc.
3025 Boardwalk Drive
Suite 160
Ann Arbor, MI 48108
USA

Phone:	1-877-436-6628
Fax:	1-734-622-0477
Email:	support-us@genomatix.com
WWW:	http://www.genomatix.com

Table of Contents

Introduction

Next Generation Sequencing (NGS) offers a sensitive and unbiased method for high-throughput genomic studies. NGS is complementing, and to a considerable extent supplanting longer established methods, such as microarrays, in the analysis of *e.g.* gene expression, protein-DNA binding, or chromatin modification on a genome-wide scale.

A number of suppliers offer platforms for massive parallel sequencing. Throughput grows with each new sequencer generation, and with increasing numbers of reads per experiment, the scalability of the mapping algorithm is becoming an important performance factor.

The major challenge, though, is faced following the mapping of the reads: data must be turned into biological information. Pivotal for this is the availability of efficient software and strategies for downstream analysis.

In this tutorial you will learn how you can analyze NGS data with the Genomatix system, specifically covering the analysis of ChIP-Seq reads.

This will include ChIP-Seq peak finding and annotation, TFBS analysis, distance correlations with the publicly-available ENCODE project data and pathway analysis of downstream target genes.

Introduction to the Genomatix Mining Station

The Genomatix Mining Station (GMS) is Genomatix' integrated software/hardware solution for first level analysis of Next Generation sequence reads.

- Mapping is based on indexing of the target sequences (Eland and other mapping software index the source sequences).
- The index is based on „shortest unique subwords".
- The complete index is stored in main memory.
  - In case of mapping to vertebrate genomes, hardware architecture with 64GB main memory is required.

Shortest unique subwords

- For every position in a target sequence: calculate the smallest downstream sequence which is unique in the target sequence.
- The minimum word length considered is 8 bps.
- Only words consisting of A, G, C, and T are accepted.

Example: human genome NCBI build 37:

- Number of positions with downstream sequences of at least 8bps consisting only of A,G,C,T: 2.976.839.776 (96%)
- Coverage by shortest unique words in range [8;25]: 2.495.605.837 (80%)

Indexing

- Shortest unique subwords are stored in a proprietary data structure which allows to search subwords with tolerances i.e. insertions/deletions/point mutations.
- Not only unique but also small words with low copy numbers (up to 50 times in genome) are stored.
- The overall memory requirement for the human genome index is about 30 GB.

Mapping

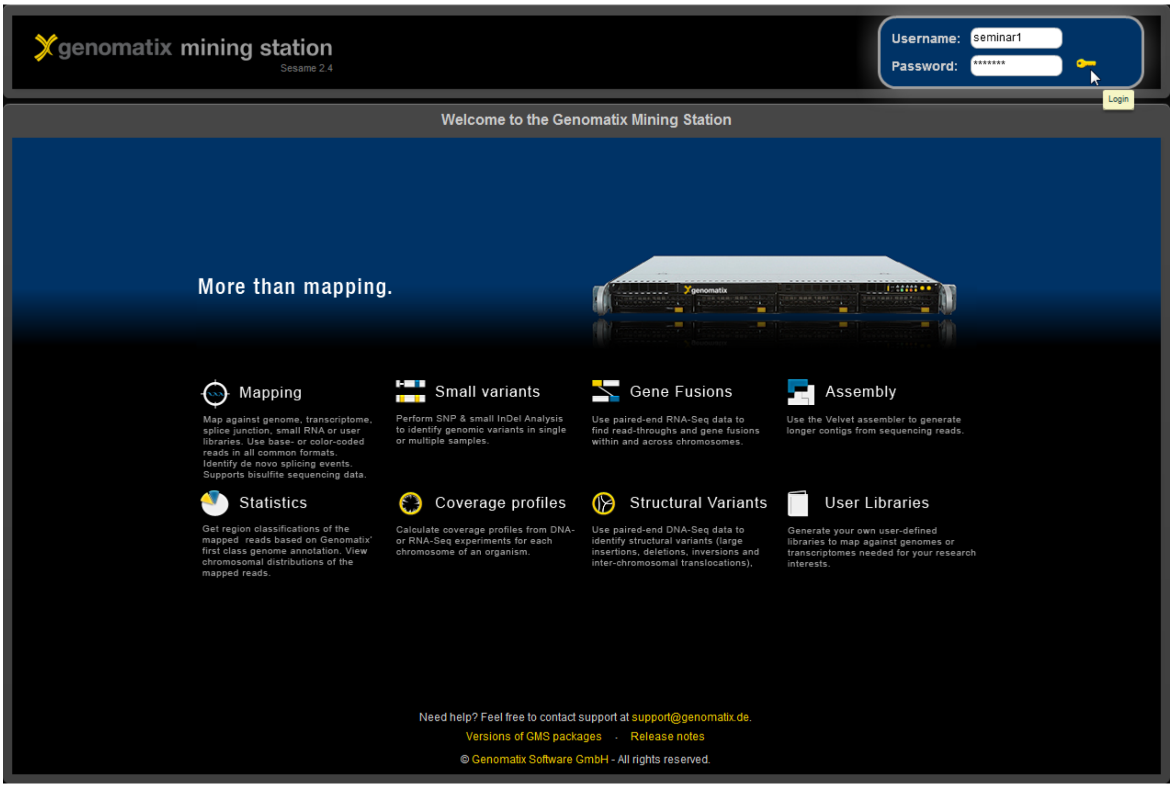
- Mapping is done in two steps:
    1. Find a seed word in a source sequence via the index
    2. Alignment of the complete source sequence
- Both steps can be done with different strictness
    1. Seed search
        - Fast: search only exact matches
        - Deep: allow max. one mismatch
    2. Complete alignment:
        - Needleman-Wunsch alignment (point mutations / indels)
        - Alignment allowing point mutations only
        - User-definable alignment quality thresholds
- Mapping time depends on selected strictness, number and quality of reads

Demo example: mapping NGS reads on the GMS

You can use the GMS web interface for most available analysis tasks on the GMS, including mapping, variant calling, and generation of statistics. The interface allows you to define and start analyses, and view and export your results. You can also view results in public projects.

Jobs with large memory footprints are automatically queued by the server's grid engine. Therefore, the following will be shown as a demonstration.

The system has a web browser interface for user access. Users log on with their user name and password, which must be provided by the system administrator:



Username: seminar1
Password: *******

Mapping
Map against genome, transcriptome, splice junction, small RNA or user libraries. Use base- or color-coded reads in all common formats. Identify de novo splicing events. Supports bisulfite sequencing data.

Small variants
Perform SNP & small InDel Analysis to identify genomic variants in single or multiple samples.

Gene Fusions
Use paired-end RNA-Seq data to find read-throughs and gene fusions within and across chromosomes.

Assembly
Use the Velvet assembler to generate longer contigs from sequencing reads.

Statistics
Get region classifications of the mapped reads based on Genomatix' first class genome annotation. View chromosomal distributions of the mapped reads.

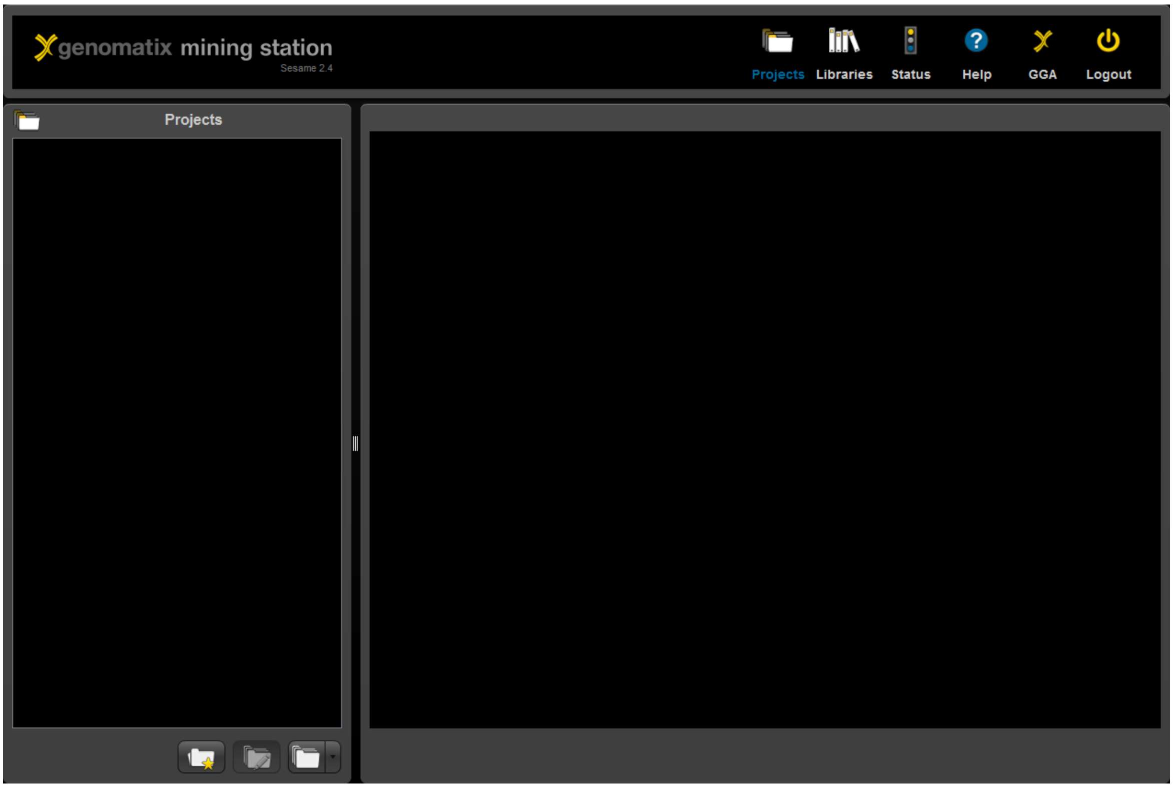
Coverage profiles
Calculate coverage profiles from DNA- or RNA-Seq experiments for each chromosome of an organism.

Structural Variants
Use paired-end DNA-Seq data to identify structural variants (large insertions, deletions, inversions and inter-chromosomal translocations),

User Libraries
Generate your own user-defined libraries to map against genomes or transcriptomes needed for your research interests.

Need help? Feel free to contact support at support@genomatix.de.
Versions of GMS packages · Release notes
© Genomatix Software GmbH - All rights reserved.

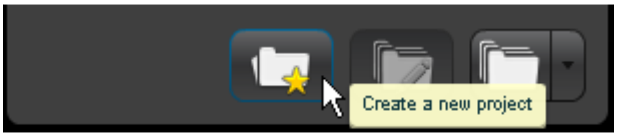
The first time you log on to the system, the interface will look like this, with an empty project panel on the left:



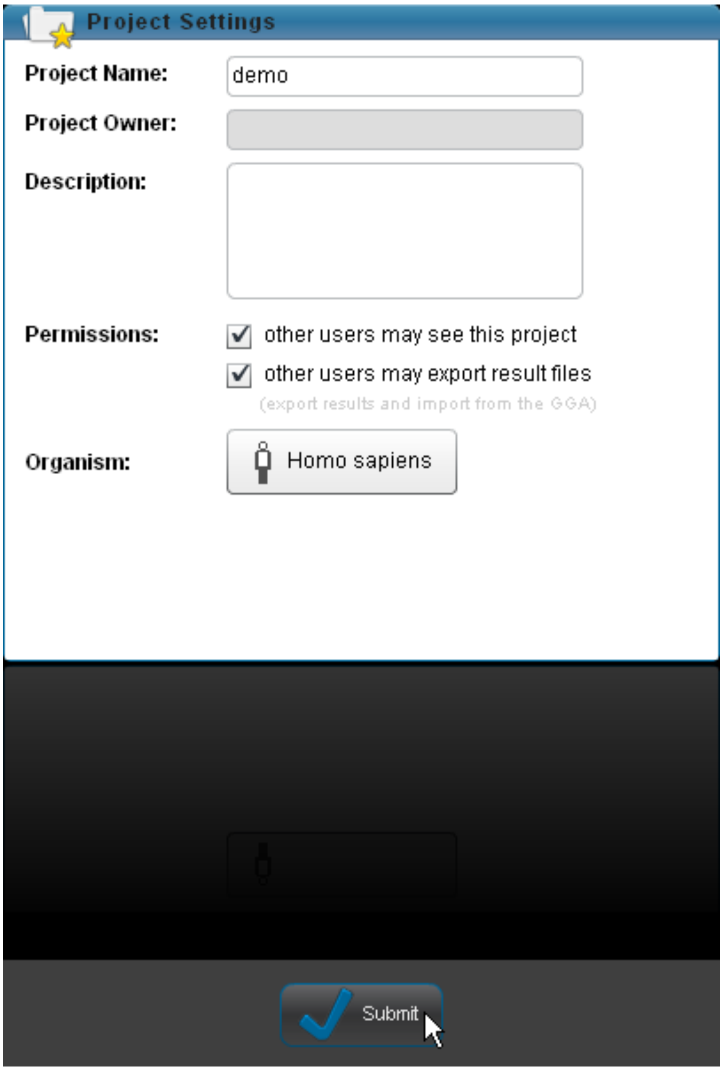
Before data can be uploaded and analyzed, a project that will contain the sequence files and analysis results has to be created. This will be shown in the next step.

Creating a project

You start by defining a project and importing sequence data to it. To do this you click the ‘Create a new project’ button in the lower left hand corner of the screen.



In the 'Project Settings' dialog, you provide a name, and optionally a description for your project. You can also allow other users access to the project and to export 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. In order to create the project, press Submit.



Project Settings

Project Name: demo

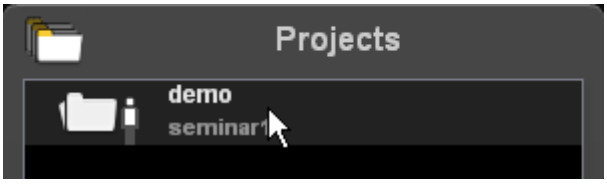
Project Owner:

Description:

Permissions:
☑ other users may see this project
☑ other users may export result files
(export results and import from the GGA)

Organism: Homo sapiens

An entry for your new project is automatically added to the project list on the left. To open the project for importing data, click on the project name.

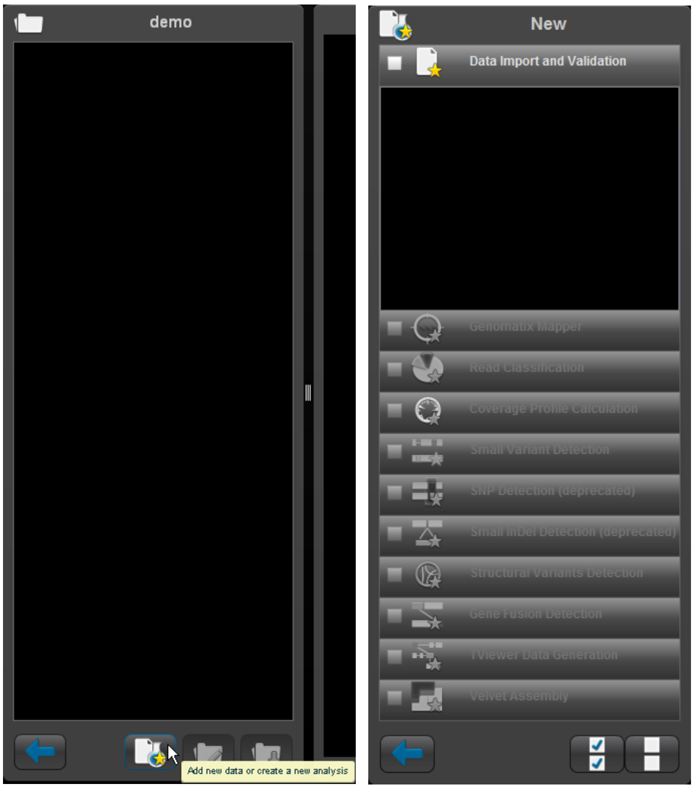


Importing sequence data to a project

The panel on the left now shows the empty project folder. 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. Here you can import data and start analyses.

Some analysis types depend on output from other analyses; as long as these results are not present, the dependent analysis types are grayed out and can’t be selected. As long as no data have been uploaded to the project, only the data import and validation option is active.

To import data, tick the checkbox in the ‘Data Import and Validation’ section; this will open a file upload dialog.


demo
Add new data or create a new analysis
New
□ Data Import and Validation
□ Genomatix Mapper
□ Read Classification
□ Coverage Profile Calculation
□ Small Variant Detection
□ SNP Detection (deprecated)
□ Small InDel Detection (deprecated)
□ Structural Variants Detection
□ Gene Fusion Detection
□ TViewer Data Generation
□ Velvet Assembly

By default, the dialog window shows the directory /home/gx_sesame/import on the GMS. Depending on the setup of your server, sequence data files will be found here or in a subdirectory, which could also be a mounted and linked file server directory, or in your home directory on the GMS (/home/<username>). Select the sequence file(s) you want to import, and click 'OK'.

The demo data set that we will use were downloaded from the Canada's Michael Smith Genome Sciences Center

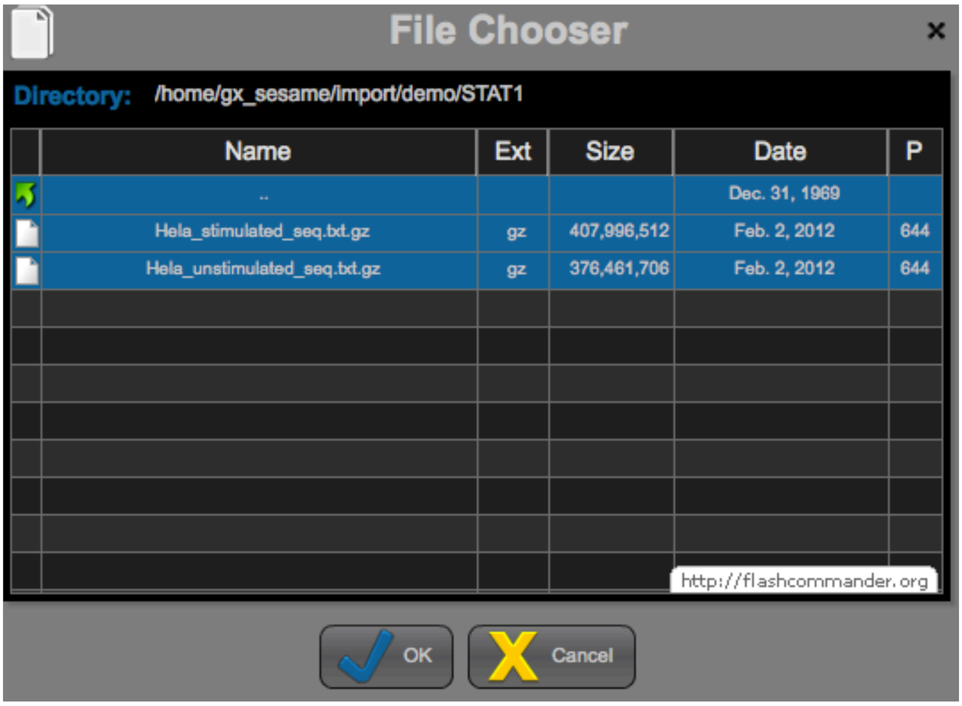
http://www.bcgsc.ca/downloads/chiptf/human/STAT1/

The raw sequence tags from each experiment can be found in

"…/stimulated/July_23_2008/*_seq.txt.gz"
"…/unstimulated/July_23_2008//*_seq.txt.gz"

These data represent a ChIP-Seq experiment containing STAT1 DNA binding in IFN-gamma stimulated and unstimulated human HeLa S3 cells (Robertson *et al*., 2007). Libraries were generated for 3 biological replicates for each condition. All data are single-end reads that were generated on Illumina 1G sequencer.

For the demonstration lanes (8) were combined from each flow cell and the treatment and control groups will be uploaded. Several files can be selected and uploaded at a time.

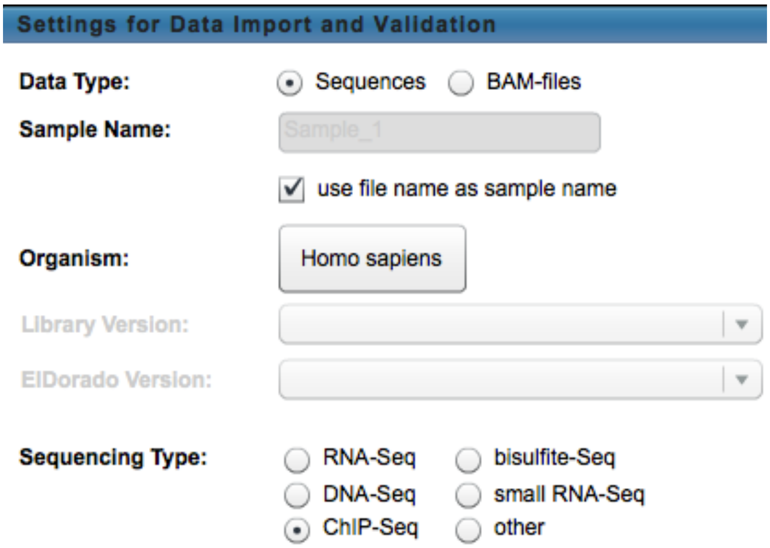


Directory: /home/gx_sesame/import/demo/STAT1

Name	Ext	Size	Date	P
..			Dec. 31, 1969	
Hela_stimulated_seq.txt.gz	gz	407,996,512	Feb. 2, 2012	644
Hela_unstimulated_seq.txt.gz	gz	376,461,706	Feb. 2, 2012	644

Clicking the OK button will open a settings dialog.

Here, the data type of the files can be defined. In this case, we have human DNA sequences (BAM files can also be uploaded), so the appropriate options are selected. Also, the file names are used as sample names (alternatively, you can provide your own sample name). Pressing Submit starts the data upload and validation.



Settings for Data Import and Validation

Data Type: ● Sequences ○ BAM-files

Sample Name: Sample_1

☑ use file name as sample name

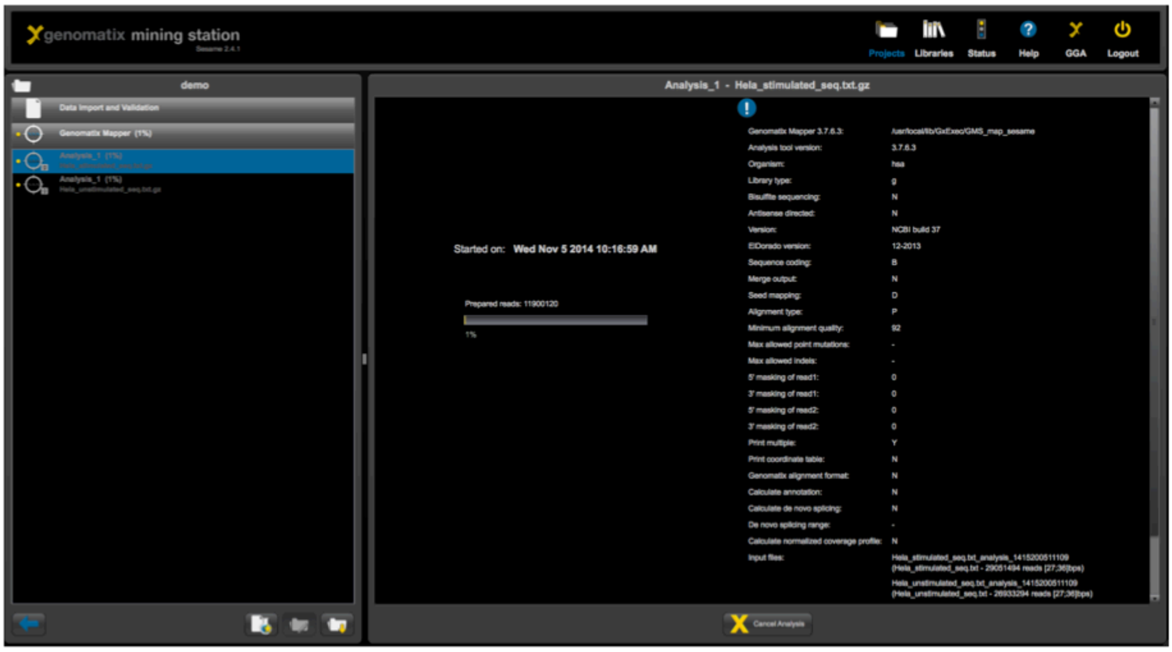
Organism: Homo sapiens

Library Version:

ElDorado Version:

Sequencing Type: ○ RNA-Seq ○ bisulfite-Seq ○ DNA-Seq ○ small RNA-Seq ● ChIP-Seq ○ other

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



Analysis_1 - Hela_stimulated_seq.txt.gz

Started on: Wed Nov 5 2014 10:16:59 AM

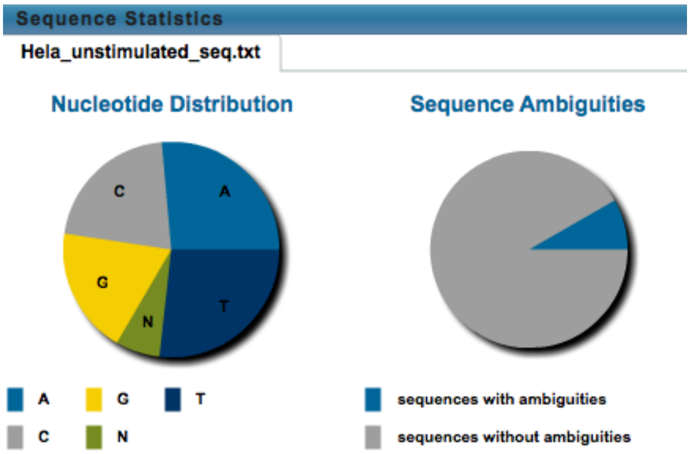
Prepared reads: 11900120

1%

Genomatix Mapper 3.7.6.3:	/usr/local/lib/GxExec/GMS_map_sesame
Analysis tool version:	3.7.6.3
Organism:	hsa
Library type:	g
Bisulfite sequencing:	N
Antisense directed:	N
Version:	NCBI build 37
ElDorado version:	12-2013
Sequence coding:	B
Merge output:	N
Seed mapping:	D
Alignment type:	P
Minimum alignment quality:	92
Max allowed point mutations:	-
Max allowed indels:	-
5' masking of read1:	0
3' masking of read1:	0
5' masking of read2:	0
3' masking of read2:	0
Print multiple:	Y
Print coordinate table:	N
Genomatix alignment format:	N
Calculate annotation:	N
Calculate de novo splicing:	N
De novo splicing range:	-
Calculate normalized coverage profile:	N
Input files:	Hela_stimulated_seq.txt_analysis_1415200511109 (Hela_stimulated_seq.txt - 29051494 reads [27;36]bps) Hela_unstimulated_seq.txt_analysis_1415200511109 (Hela_unstimulated_seq.txt - 26933294 reads [27;36]bps)

Looking at sequence statistics

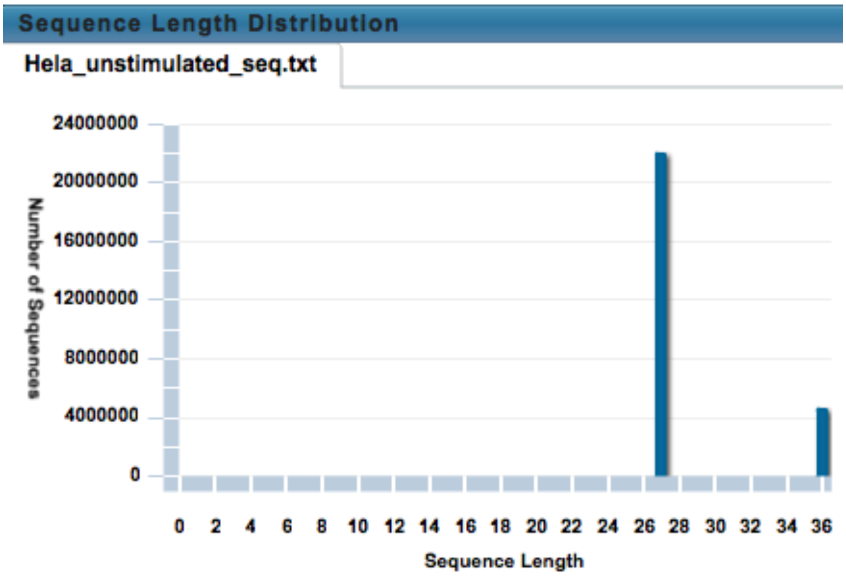
Click on a data set name to display the corresponding sequence statistics:


Sequence Statistics
Hela_unstimulated_seq.txt
Nucleotide Distribution
Sequence Ambiguities
C
A
G
N
T
A
G
T
C
N
sequences with ambiguities
sequences without ambiguities

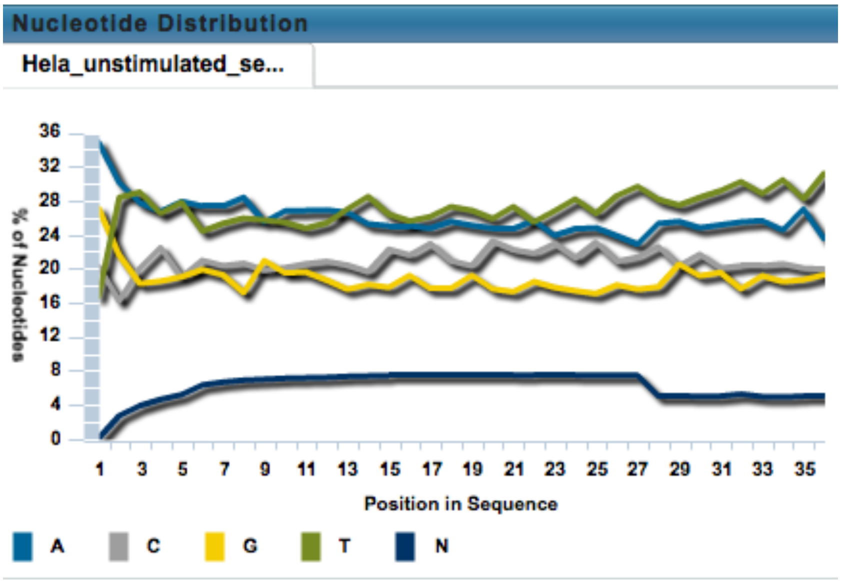
The left pie chart shows the nucleotide distribution in the reads. Positioning the mouse pointer over a part of a chart will display the corresponding numbers in a tool tip. Some numbers are also provided to the right of the graph panel. The average GC content here is 43.1%.

The right chart displays the portions of sequences with and without ambiguities (Ns). 8.3% of the reads in the Hela_unstimulated_seq.txt data contain Ns.

The next chart shows the distribution of sequence lengths in the data set. In this case, the read length is either 27 or 36 nt.



The last chart in this panel displays the nucleotide distribution at each position in the sequence reads.



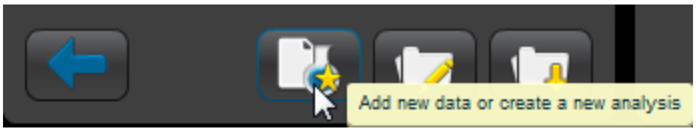
The nucleotide distribution is fairly variable over most of the sequence length, with slightly disparate percentages for A and T (blue and green curves), and G and C, respectively. N content is between 1-7%.

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

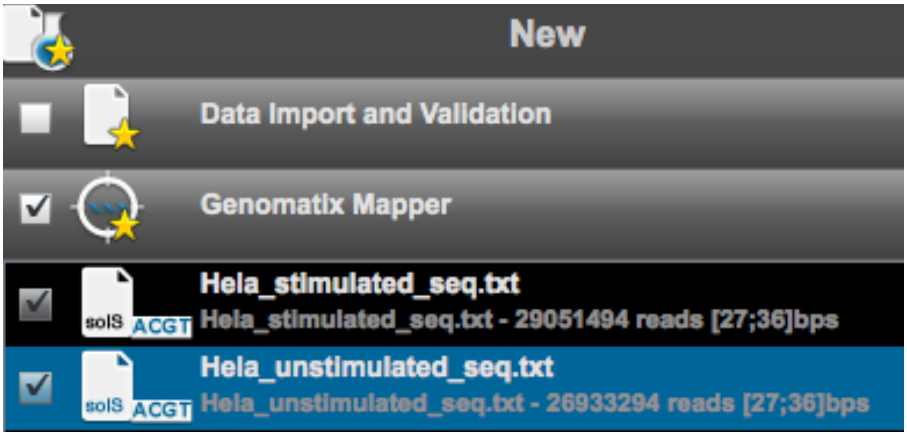
Starting a mapping

On the Genomatix Mining Station two mapping jobs can be run in parallel. For the purposes of training, mapping and analysis will be shown as a demonstration. Please note that while you can view and export results in another user's project if the owner allows it, you can map and analyze data only in projects you own. The following describes the process of analysis from the presenter's view.

Clicking the 'Analyze Data' button in the lower left hand corner gives you once more access to the analysis menu, where, after data have been uploaded, additional analysis options are selectable.



Ticking the checkbox in the ‘Genomatix Mapper’ section will display the list of available sequence files and a settings dialog. To select files, tick the checkbox next to the name.



New

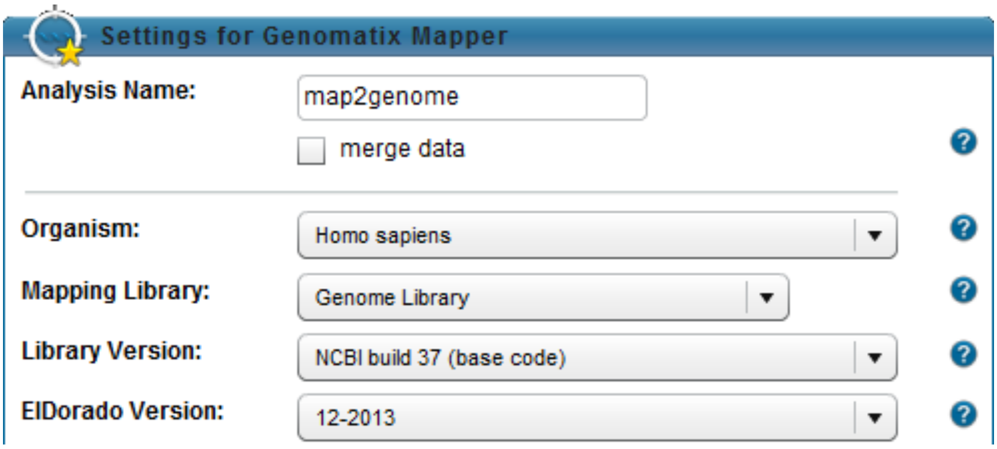
□ Data Import and Validation

☑ Genomatix Mapper

☑ **Hela_stimulated_seq.txt**
Hela_stimulated_seq.txt - 29051494 reads [27;36]bps

☑ **Hela_unstimulated_seq.txt**
Hela_unstimulated_seq.txt - 26933294 reads [27;36]bps

In the settings dialog an analysis name can be provided. To obtain a separate result set for each selected sequence file, the 'merge data' option is left empty. We'll select the genome library of *Homo sapiens*, using the newest genome library and Genomatix genome annotation (ElDorado) versions.



Settings for Genomatix Mapper

Analysis Name: map2genome
□ merge data

Organism: Homo sapiens

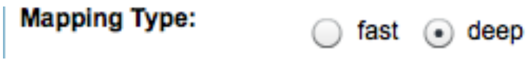
Mapping Library: Genome Library

Library Version: NCBI build 37 (base code)

ElDorado Version: 12-2013

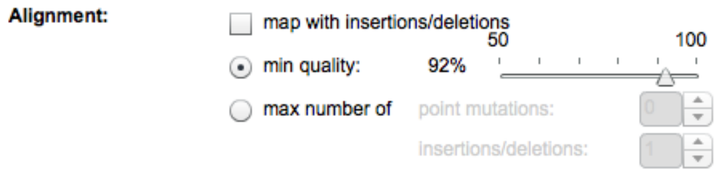
For strand-specific RNA-Seq protocols which generate antisense sequences, the 'antisense directed' option must be used to obtain the correct mapping result. Here, this is not needed.

For 'Mapping Type', ‘deep’ will be used for this example. The first mapping step – the seed search in the index – will allow up to one mismatch in the seed search, which can give you more mapping hits at the expense of speed. The latter option is most useful for very short sequences (like miRNAs) and for sequence files with high error rates, in which too many reads lack perfect seed sequences to maintain good mapping efficiency.



Mapping Type: ○ fast ● deep

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, a minimum alignment quality of 85% will be used. The ‘map with insertions/deletions’ checkbox is left empty. Mapping with indels would be necessary for pyrosequencing data (454, IonTorrent), where over- and undercalling in homopolymer runs is an issue. As the reads in our data set have a mixed lenth of 27 or 36nt, this will be equivalent to up to 2-3 point mutations per read.



Alignment:

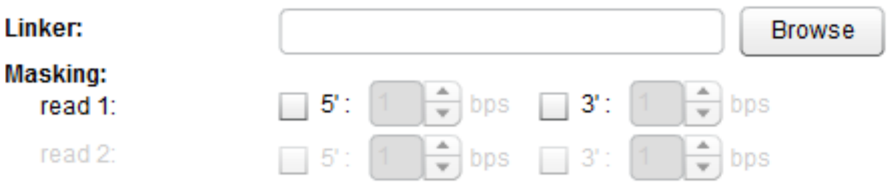
□ map with insertions/deletions

● min quality: 92%

○ max number of point mutations: 0

insertions/deletions: 1

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. Linker sequences can also be removed if a file with the linker sequences is uploaded here. The nucleotide distribution statistics did not indicate the need for masking or linker removal, so this is left empty here.



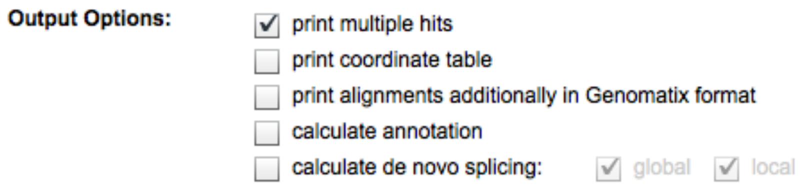
Linker:

Masking:

read 1: □ 5' : 1 bps □ 3' : 1 bps

read 2: □ 5' : 1 bps □ 3' : 1 bps

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 formatted text file with the positions of uniquely mapping reads. The 'Output Options' allow you to generate additional result files.



Output Options:

☑ print multiple hits

□ print coordinate table

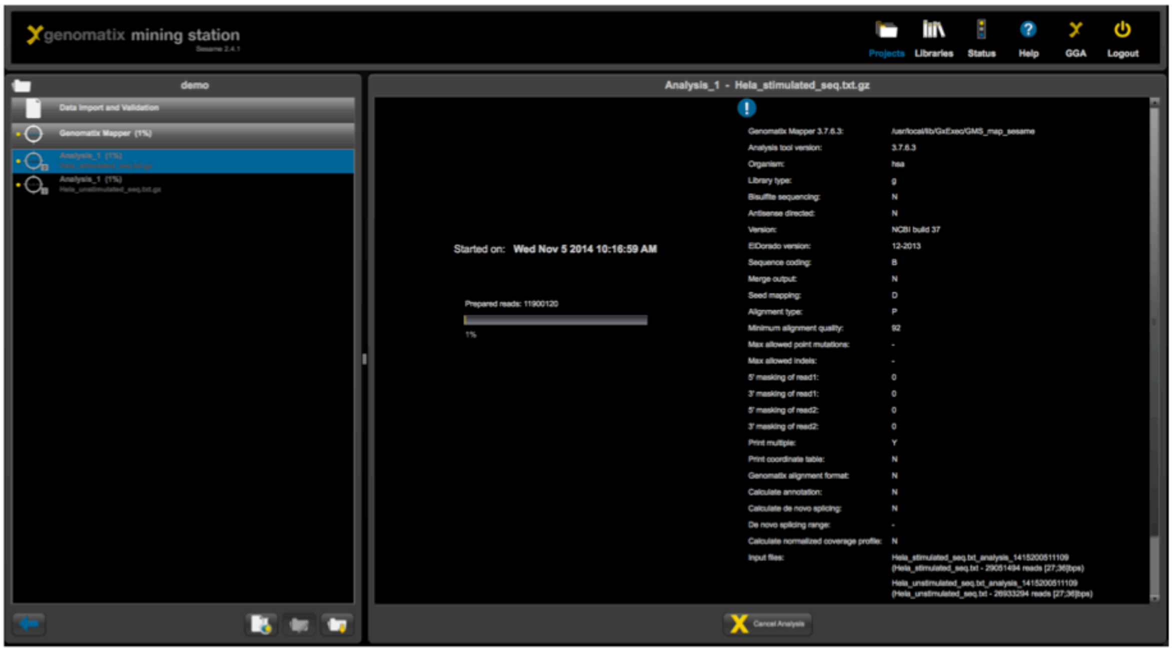
□ print alignments additionally in Genomatix format

□ calculate annotation

□ calculate de novo splicing: ☑ global ☑ local

Pressing ‘Submit’ starts the mapping.

The progress and the parameters of any running analysis are shown as below. In the demo, a pre-mapped dataset will be used for the next steps.



Started on: **Wed Nov 5 2014 10:16:59 AM**

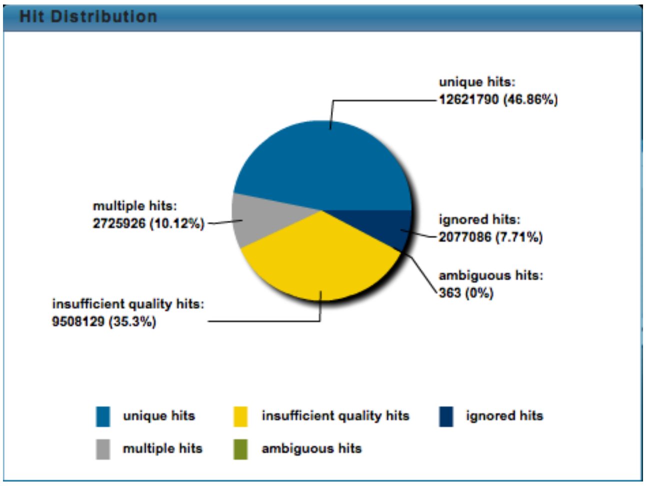
Prepared reads: 11900120

1%

Genomatix Mapper 3.7.6.3:	/usr/local/lib/GxExec/GMS_map_sesame
Analysis tool version:	3.7.6.3
Organism:	hsa
Library type:	g
Bisulfite sequencing:	N
Antisense directed:	N
Version:	NCBI build 37
ElDorado version:	12-2013
Sequence coding:	B
Merge output:	N
Seed mapping:	D
Alignment type:	P
Minimum alignment quality:	92
Max allowed point mutations:	-
Max allowed indels:	-
5' masking of read1:	0
3' masking of read1:	0
5' masking of read2:	0
3' masking of read2:	0
Print multiple:	Y
Print coordinate table:	N
Genomatix alignment format:	N
Calculate annotation:	N
Calculate de novo splicing:	N
De novo splicing range:	-
Calculate normalized coverage profile:	N
Input files:	Hela_stimulated_seq.txt_analysis_1415200511109 (Hela_stimulated_seq.txt - 29051494 reads [27;36]bps) Hela_unstimulated_seq.txt_analysis_1415200511109 (Hela_unstimulated_seq.txt - 26933294 reads [27;36]bps)

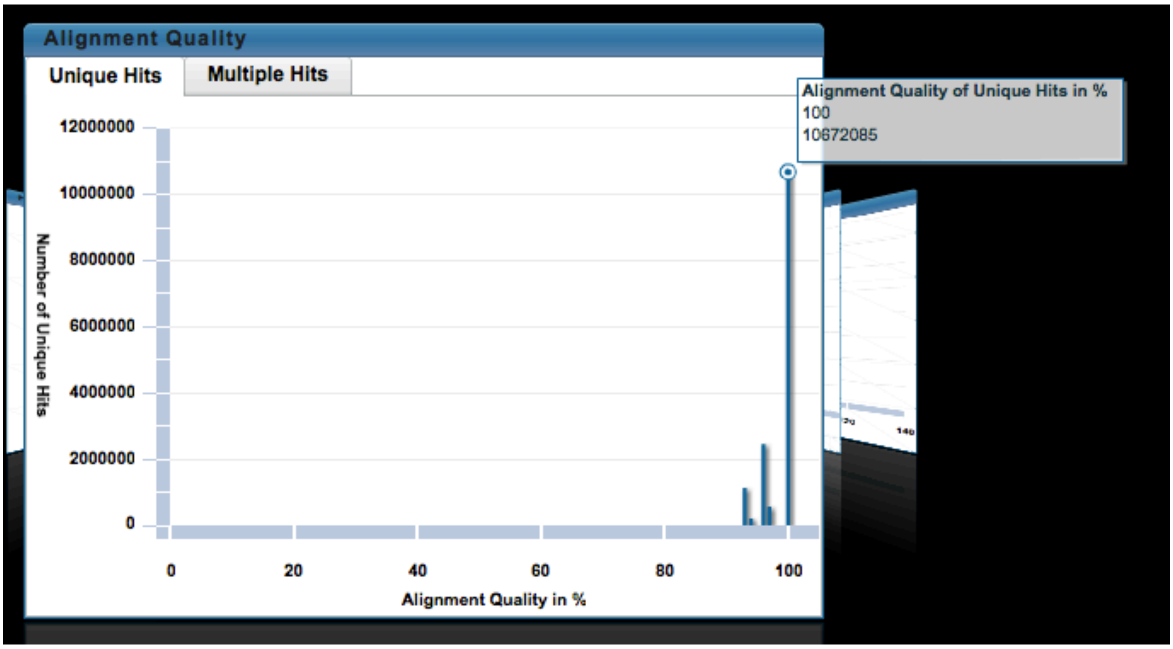
Mapping statistics

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; multiple hits have 2-50 equally best matches; unique hits have exactly one best match. The unique hit percentage of about 47% in the unstimulated sample and 51% in the stimulated sample This is slightly less than the results reported by Robertson *et al*. (2007) who reports that approximately 60% of their reads will map to unique locations in the genome.

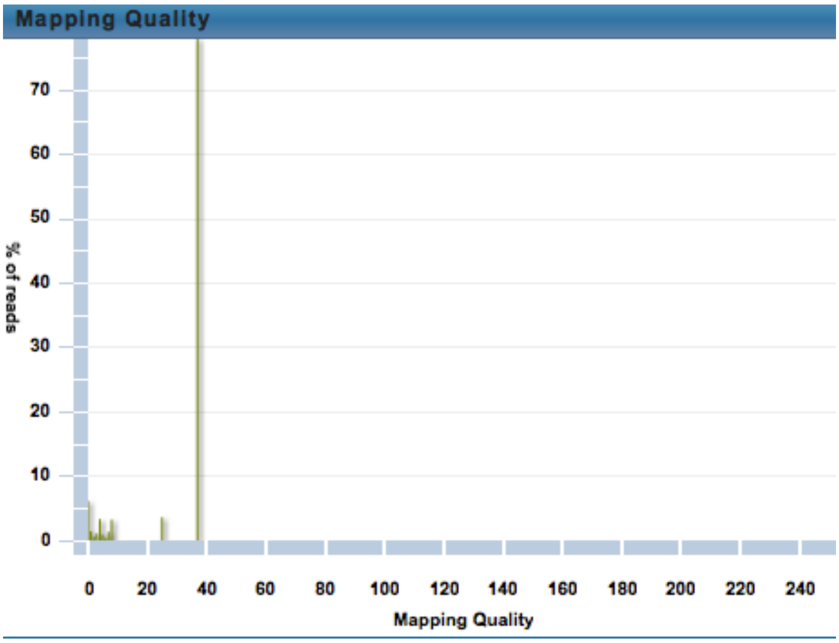


We also see that with a 92% alignment quality, more than one-third of the reads map below this threshold (insufficient quality hits). Adjusting the alignment threshold to below 92% (*e.g.*, 85%) will result in more uniquely mapped reads.

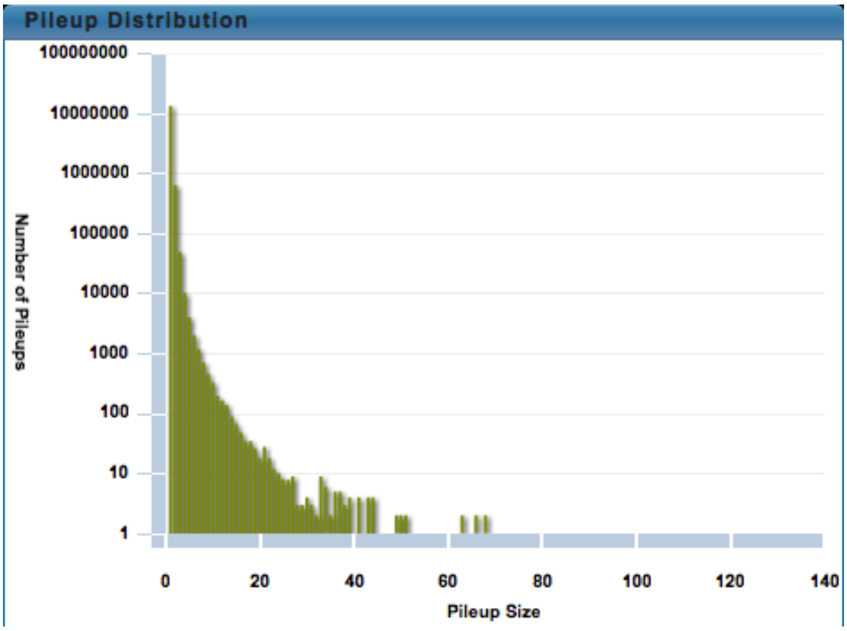
Move the slider below the graph to the right to view the alignment quality profile for the unique hits. The majority of reads map perfectly (rightmost column); additionally, we have a smaller percentage of the reads mapping with lower thresholds (4-5 differences).



In the following diagram, you see the distribution of mapping qualities. Mapping quality scores are a measure for the confidence that the read is correctly placed. For example, a mapping quality of 20 that there is at least a 1 in 100 chance that the read truly originated elsewhere. A value 255 indicates that the mapping quality is not available. For paired-end alignment, the pairing information (distance and strand orientation of the mates) will also be included.

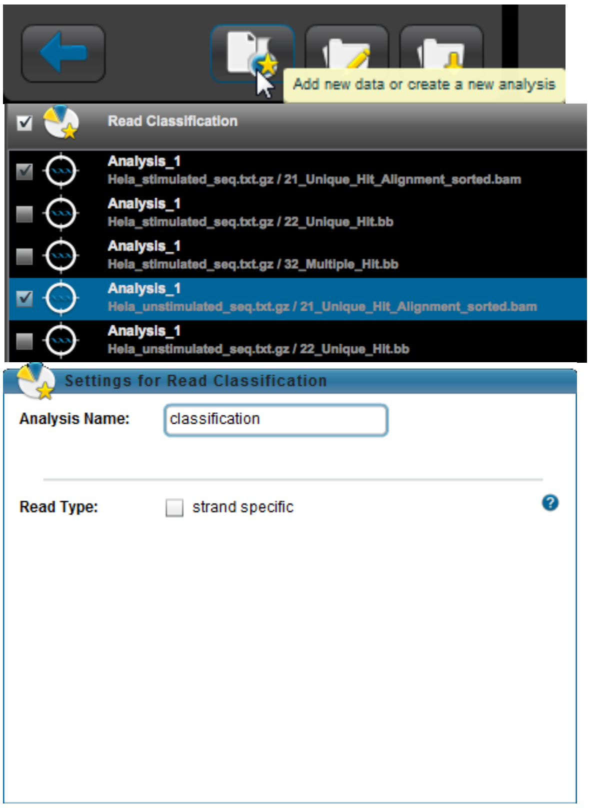


The next graph tab shows the pileup size distribution. Pileups are isolated stacks of reads with identical sequence mapping at identical positions, and are normally discarded as artifacts. The 0.95 quantile for the pileup size is generally used as a threshold for determining the maximum allowed pileup size in some downstream analyses.



Read classification

Mapped reads can be classified according to the annotation of the region they map in. The analysis is set up after clicking the ‘Analyze Data’ button. Select ‘Read Classification’ and the .bb or .bam files containing the unique hits from the previous step as shown below. The settings dialog takes an analysis name. Use the 'strand specific' option only if a strand-specific sequencing protocol was used. ‘Submit’ starts the analysis.



☑ Read Classification

☑ Analysis_1
Hela_stimulated_seq.txt.gz / 21_Unique_Hit_Alignment_sorted.bam

□ Analysis_1
Hela_stimulated_seq.txt.gz / 22_Unique_Hit.bb

□ Analysis_1
Hela_stimulated_seq.txt.gz / 32_Multiple_Hit.bb

☑ Analysis_1
Hela_unstimulated_seq.txt.gz / 21_Unique_Hit_Alignment_sorted.bam

□ Analysis_1
Hela_unstimulated_seq.txt.gz / 22_Unique_Hit.bb

Settings for Read Classification

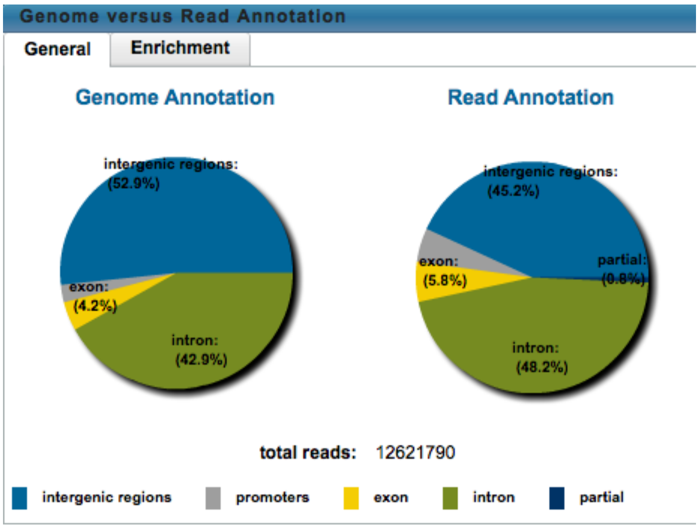
Analysis Name: classification

Read Type: □ strand specific

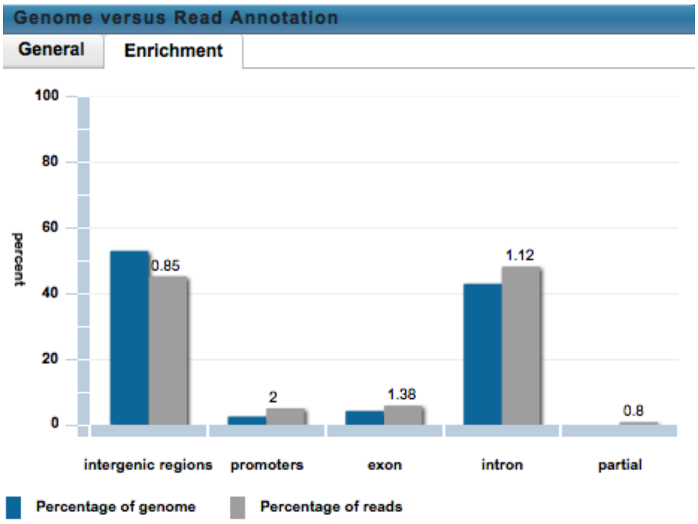
The analysis will take only a few minutes.

The output includes a collection of statistics graphs.

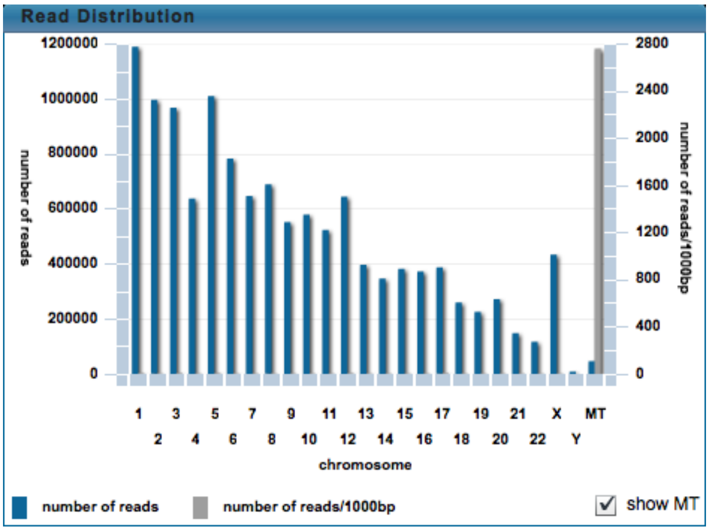
The first tab of the first graph contains two pie charts: one shows the portions of the human genome annotated as intergenic, exon, intron, and promoter in ElDorado genome annotation. The second chart represents the corresponding distribution of the analyzed reads. ‘Partial’ denotes reads that partially overlap with an annotated exon. As can be expected for ChIP-Seq data, promoters and exons are strongly overrepresented in the reads because promoter annotation overlaps with first exons of transcripts. 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.



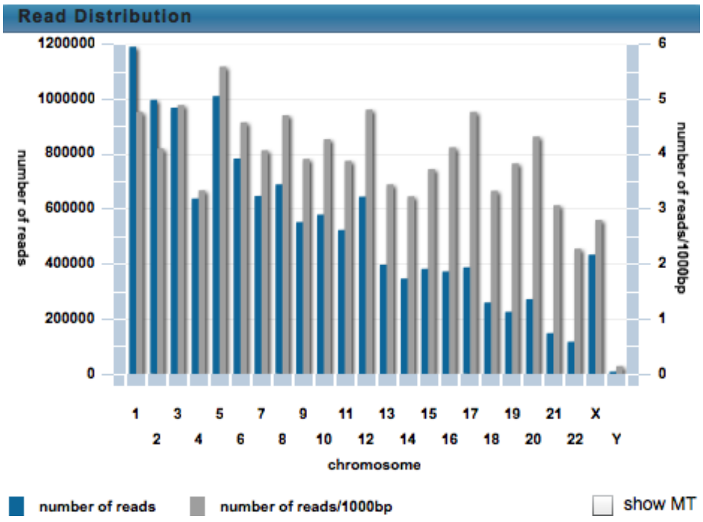
The second tab is a side-by-side comparison of the percentages of each annotation in genome and reads, with fold over/underrepresentation numbers:



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.



Un-tick the 'show MT' checkbox to hide the MT values and thus rescale the other read density columns.

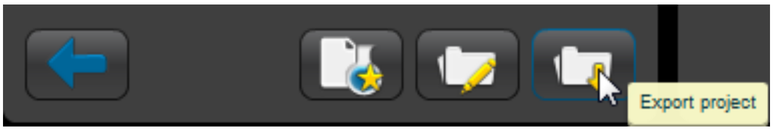


ea

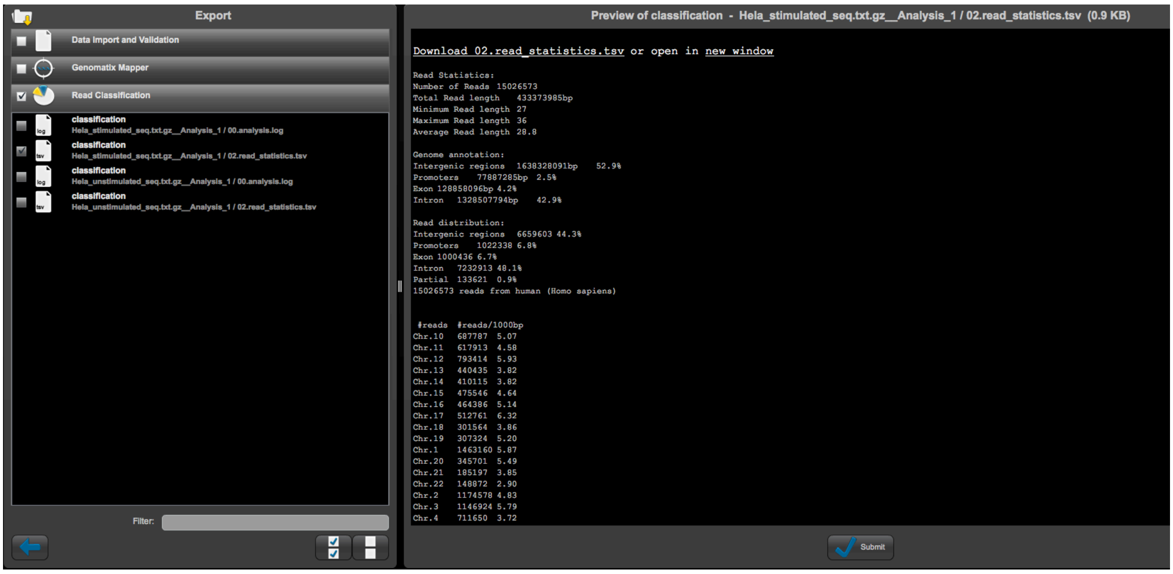
Preview, download, and export of result files

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. Depending on your setup, they might be available on the GGA directly, but in this case no export is needed.

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

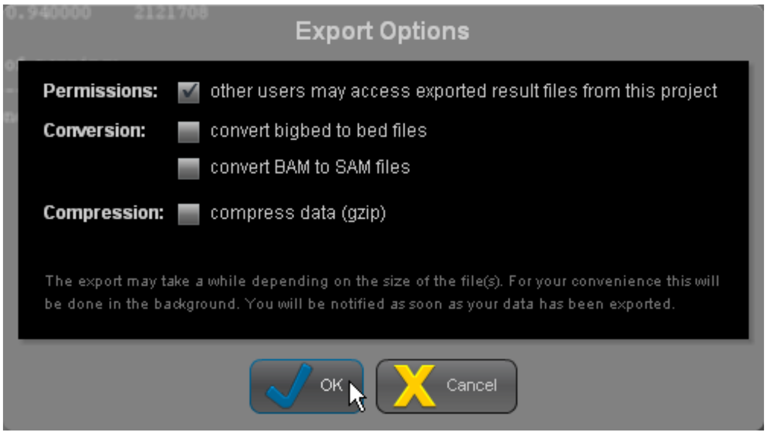


Results in the ‘Export’ menu are grouped just as in the ‘Analyze Data’ menu. For a preview of a file, click on the header of the according section and then click on the file name in the list. To select files for export to the GMS file system, tick the checkbox in the header (*e.g.* ‘Read Classification’ 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. Small files can also be downloaded individually to your local computer using the 'Download <filename>' link in the preview window.


Export
□ Data Import and Validation
□ Genomatix Mapper
☑ Read Classification
□ classification
Hela_stimulated_seq.txt.gz__Analysis_1 / 00.analysis.log
☑ classification
Hela_stimulated_seq.txt.gz__Analysis_1 / 02.read_statistics.tsv
□ classification
Hela_unstimulated_seq.txt.gz__Analysis_1 / 00.analysis.log
□ classification
Hela_unstimulated_seq.txt.gz__Analysis_1 / 02.read_statistics.tsv
Filter:
Preview of classification - Hela_stimulated_seq.txt.gz__Analysis_1 / 02.read_statistics.tsv (0.9 KB)
Submit

```
Download 02.read_statistics.tsv or open in new window

Read Statistics:
Number of Reads 15026573
Total Read length    433373985bp
Minimum Read length 27
Maximum Read length 36
Average Read length 28.8

Genome annotation:
Intergenic regions  1638328091bp    52.9%
Promoters    77887285bp  2.5%
Exon 128858096bp 4.2%
Intron   1328507794bp    42.9%

Read distribution:
Intergenic regions   6659603 44.3%
Promoters    1022338 6.8%
Exon 1000436 6.7%
Intron   7232913 48.1%
Partial  133621  0.9%
15026573 reads from human (Homo sapiens)


 #reads  #reads/1000bp
Chr.10   687787  5.07
Chr.11   617913  4.58
Chr.12   793414  5.93
Chr.13   440435  3.82
Chr.14   410115  3.82
Chr.15   475546  4.64
Chr.16   464386  5.14
Chr.17   512761  6.32
Chr.18   301564  3.86
Chr.19   307324  5.20
Chr.1    1463160 5.87
Chr.20   345701  5.49
Chr.21   185197  3.85
Chr.22   148872  2.90
Chr.2    1174578 4.83
Chr.3    1146924 5.79
Chr.4    711650  3.72
```

For exporting to the GMS, click the ‘Submit’ button 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.



Export Options

Permissions: ☑ other users may access exported result files from this project

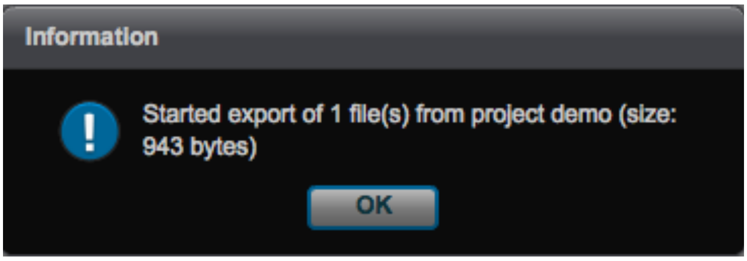
Conversion: □ convert bigbed to bed files

□ convert BAM to SAM files

Compression: □ compress data (gzip)

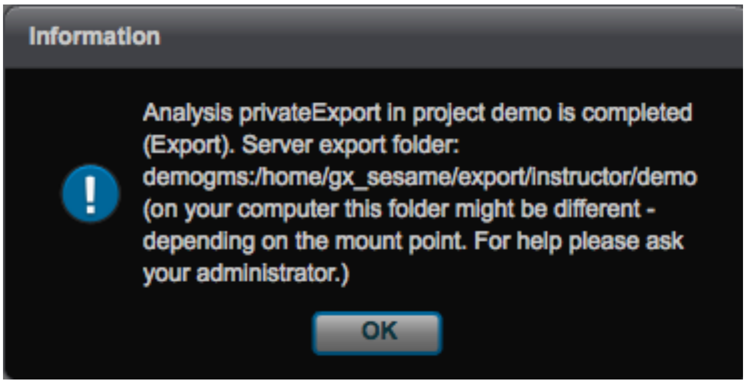
The export may take a while depending on the size of the file(s). For your convenience this will be done in the background. You will be notified as soon as your data has been exported.

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



Information

Started export of 1 file(s) from project demo (size: 943 bytes)



Information

Analysis privateExport in project demo is completed (Export). Server export folder: demogms:/home/gx_sesame/export/instructor/demo (on your computer this folder might be different - depending on the mount point. For help please ask your administrator.)

Exported files can then be accessed in the file system of your GMS. 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.

Introduction to the Genomatix Genome Analyzer

The Genomatix Genome Analyzer (GGA) is an integrated software/hardware solution for second level analysis of NGS data, after reads have been mapped to the respective genomic target sequences. An easy to use web interface gives access to a broad range of analysis applications for Chip-Seq, RNA-Seq, and DNA-Seq data, among them:

Peak finding
Position data of mapped single reads can be clustered to detect peaks and separate signal from background.

Genome annotation
NGS data can be integrated, correlated, and visualized within the extensive genome annotation in ElDorado. Comparative genomics allows cross-species analysis for phylogenetically conserved regions and regulatory structures.

Expression analysis
The GGA generates normalized transcript expression values from your NGS data and genomic annotation. Compare data sets for differential expression and upload the results into Genomatix Pathway System to generate and analyze gene networks.

Transcription factor analysis
Genome-wide transcription factor (TF) analysis identifies overrepresented TF binding sites and phylogenetically conserved functional elements. Correlation with genomic annotation finds potential regulatory targets of TF binding. Use CoreSearch for de novo binding site definition from your ChIP-Seq data.

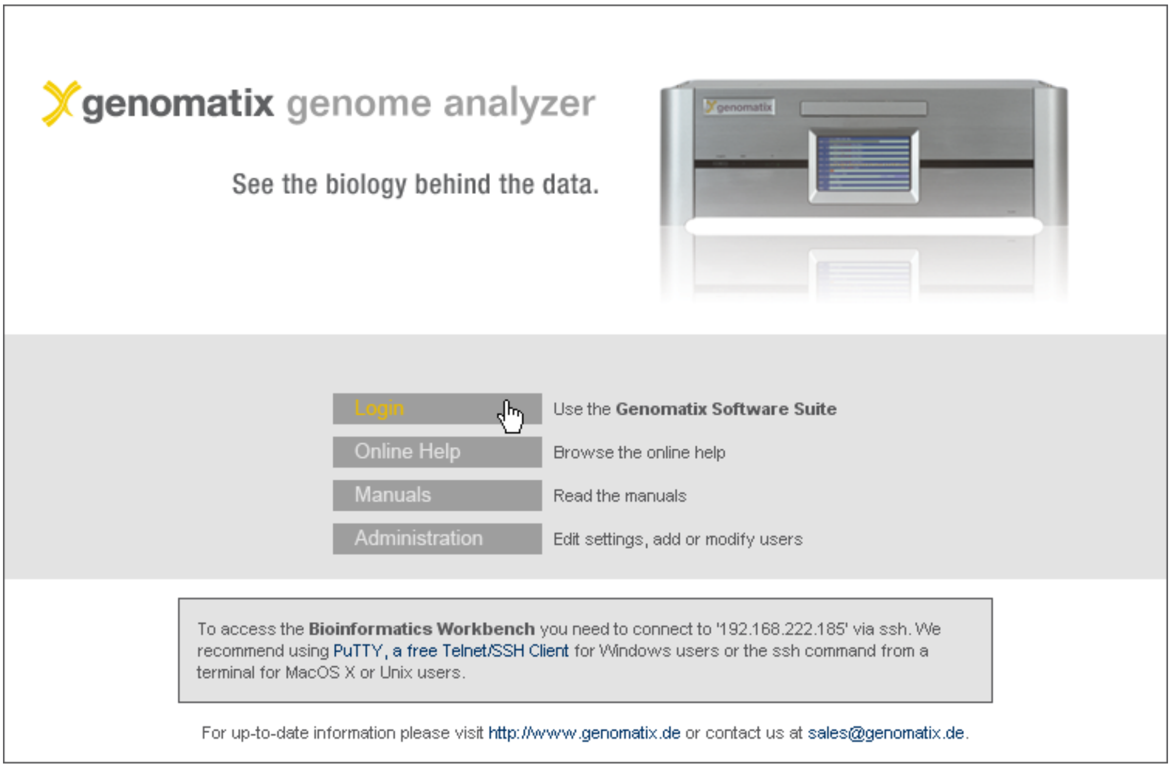
Data meta analysis
Compare several data sets in position correlation graphs, *e.g.* for the genome wide elucidation of TF interaction, and retrieve regions based on correlation.

Variant analysis
Genome wide small variant analysis identifies effects on protein sequences and TF binding sites, using the genome and TF binding site annotation in ElDorado and MatBase.

Hands-on examples

The first examples will show you how to analyze mapped sequence reads of RNA-Seq studies and what information can be found in the output files. You'll learn how to use downstream analysis tools, and how to view NGS data in ElDorado.

Start your browser and open the home page of your Genomatix Genome Analyzer. You should see a page like this:



Use the **Genomatix Software Suite**

Browse the online help

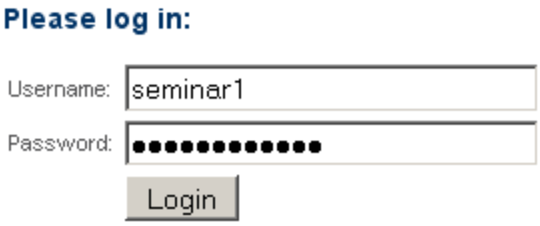
Read the manuals

Edit settings, add or modify users

To access the **Bioinformatics Workbench** you need to connect to '192.168.222.185' via ssh. We recommend using PuTTY, a free Telnet/SSH Client for Windows users or the ssh command from a terminal for MacOS X or Unix users.

For up-to-date information please visit http://www.genomatix.de or contact us at sales@genomatix.de.

Click the 'Login' button and enter your user name and password:

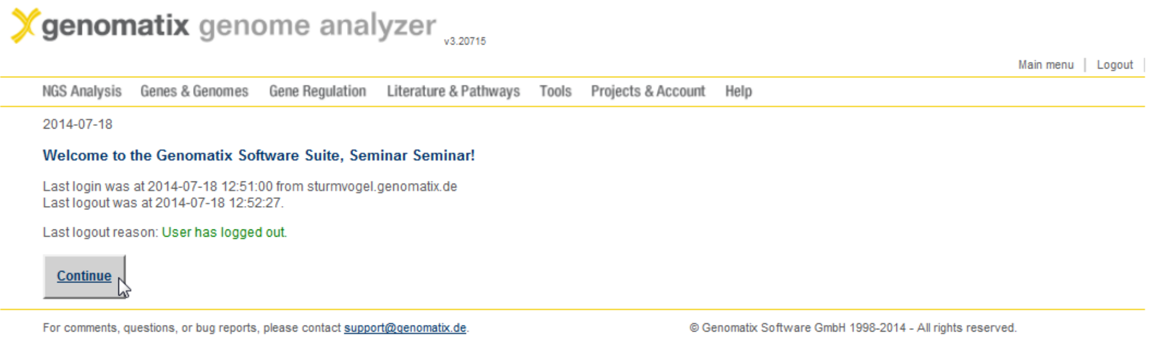


Please log in:

Username: seminar1

Password: ●●●●●●●●●●●●●

A welcome page with news will be shown. Programs can be started from the navigation bar, which always stays visible. Pressing the Continue button will open the main menu page.



2014-07-18

Welcome to the Genomatix Software Suite, Seminar Seminar!

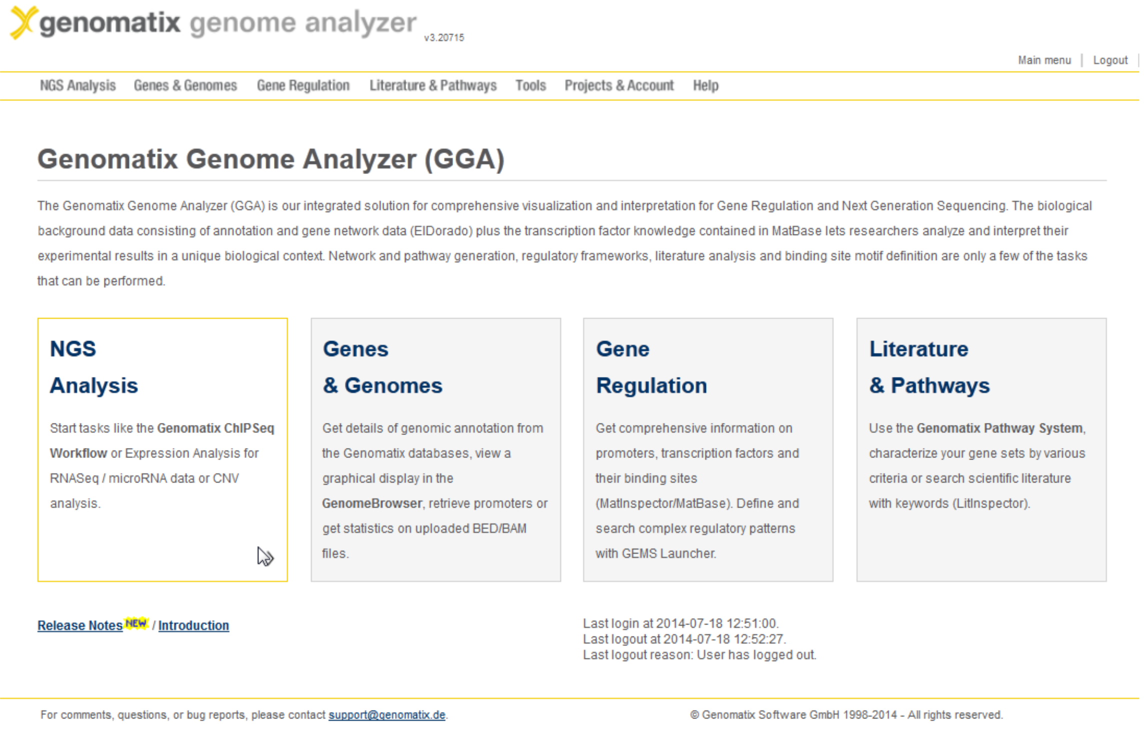
Last login was at 2014-07-18 12:51:00 from sturmvogel.genomatix.de
Last logout was at 2014-07-18 12:52:27.

Last logout reason: User has logged out.

For comments, questions, or bug reports, please contact support@genomatix.de.

© Genomatix Software GmbH 1998-2014 - All rights reserved.

From the main menu, you can also access the programs in the four main packages, as well as the release notes.



Genomatix Genome Analyzer (GGA)

The Genomatix Genome Analyzer (GGA) is our integrated solution for comprehensive visualization and interpretation for Gene Regulation and Next Generation Sequencing. The biological background data consisting of annotation and gene network data (ElDorado) plus the transcription factor knowledge contained in MatBase lets researchers analyze and interpret their experimental results in a unique biological context. Network and pathway generation, regulatory frameworks, literature analysis and binding site motif definition are only a few of the tasks that can be performed.

NGS Analysis

Start tasks like the **Genomatix ChIPSeq Workflow** or Expression Analysis for RNASeq / microRNA data or CNV analysis.

Genes & Genomes

Get details of genomic annotation from the Genomatix databases, view a graphical display in the **GenomeBrowser**, retrieve promoters or get statistics on uploaded BED/BAM files.

Gene Regulation

Get comprehensive information on promoters, transcription factors and their binding sites (MatInspector/MatBase). Define and search complex regulatory patterns with GEMS Launcher.

Literature & Pathways

Use the **Genomatix Pathway System**, characterize your gene sets by various criteria or search scientific literature with keywords (LitInspector).

Release Notes NEW / Introduction

Last login at 2014-07-18 12:51:00.
Last logout at 2014-07-18 12:52:27.
Last logout reason: User has logged out.

For comments, questions, or bug reports, please contact support@genomatix.de.

© Genomatix Software GmbH 1998-2014 - All rights reserved.

ChIP-Seq workflow: STAT1 binding in IFN-γ stimulated HeLa cells

In the next example, you will learn how to analyze ChIP-Seq data, including peak finding, TFBS analysis, and target prediction.

Available peak finding algorithms

As ChIP-Seq data are inherently noisy, clustering of mapped ChIP-Seq reads is a prerequisite step for their analysis. Clustering algorithms use a distribution model of the reads for separating signal from noise.

Three different algorithms are available in RegionMiner for cluster detection in ChIP-Seq data: NGS Analyzer, and the public algorithms MACS (Model based Analysis for ChIP-Seq) and SICER (Spatial clustering for Identification of ChIP-Enriched Regions).

NGS Analyzer was developed by Genomatix; it identifies local enrichments (clusters) representing genomic regions bound by protein (ChIP-Seq) or being expressed (RNA-Seq). By default, the threshold applied by the clustering algorithm takes the density of the data set into account, assuming a Poisson distribution. A control data file can be provided.

Two alternative ways of background subtraction are possible:

Either, clusters in the experimental data sets that overlap with unspecific enrichments detected in the control data are completely removed from the ChIP experiment.

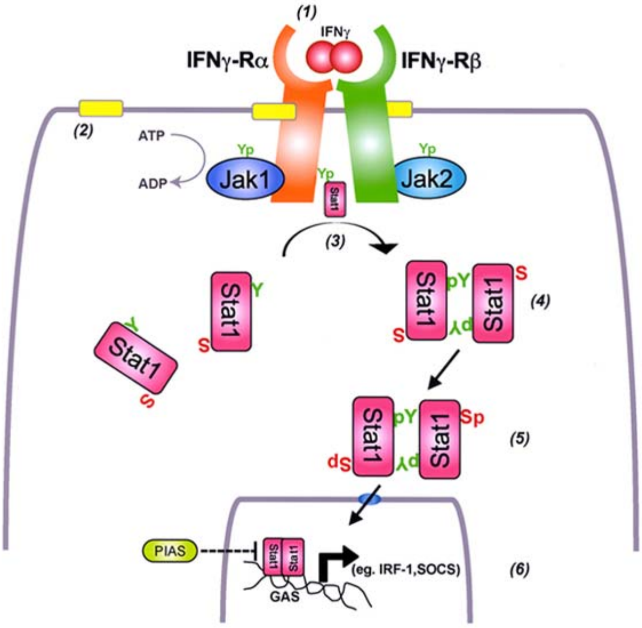
Alternatively, a quantitative comparison of the clustered reads in the experimental data file to the reads in corresponding regions in the control file using the Audic-Claverie algorithm (Audic & Claverie, 1997) can be applied.

MACS is specifically designed for clustering of ChIP-Seq data with narrow peaks as you typically get from transcription factor binding. It uses a sliding window approach and assumes a Poisson distribution of the reads just as NGS Analyzer does. However, it uses a peak model generated from high confidence read cluster regions in the data to shift the reads to the assumed center of a protein binding region. It also uses the local read density background for peak calling, which NGS Analyzer does not do. MACS comes with its own quantitative background subtraction method against a control file.

MACS has been developed at the Dana-Farber Cancer Institute (Zhang *et al*, 2008). The GGA uses the original MACS implementation.

SICER (Zang *et al*., 2009) is particularly recommended for the analysis of histone modifications, which form broad peaks. It scores non-overlapping windows (typically of nucleosome length) based on the read count, assuming a Poisson distribution. Windows are flagged eligible based on a read count significance threshold, and adjacent eligible windows are grouped as islands (peaks). Small gaps of ineligible windows can be allowed within islands. The island score is the sum of the scores of the eligible windows in the island.

We will look at some data from a ChIP-Seq experiment comparing STAT1 DNA binding in IFN-gamma stimulated and unstimulated human HeLa S3 cells (Robertson *et al*., 2007).



Graph from Ceponis *et al*., 2005

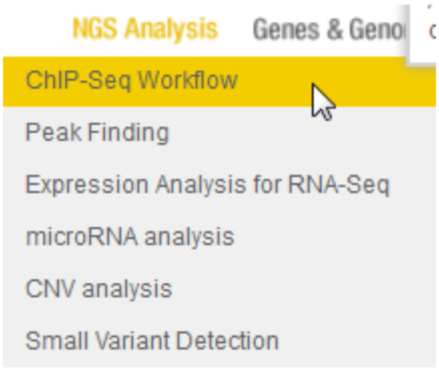
IFN-gamma regulates transcription via the JAK-STAT pathway. Binding of IFN-gamma to its cognate receptor stimulates phosphorylation of STAT1 by Janus kinase, followed by dimerization and translocation of the STAT1 homodimers into the nucleus, where they bind GAS (gamma activated sequence) motifs on the DNA.

A comparison of IFN-gamma stimulated and untreated cells reveals genomic regions of IFN-gamma dependent STAT1 binding as well as potential regulatory targets of IFN-gamma.

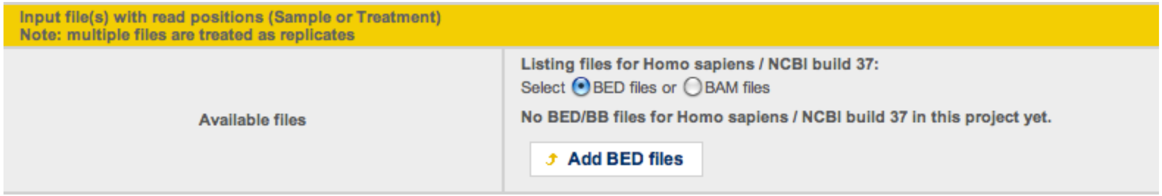
The raw sequence tags from the experiment have been mapped to the human genome using the GMS. For this example, a random sample containing 1000000 read positions was generated from the output BED files for each condition (stimulated and unstimulated). You find the files in the folder `HeLa_STAT1` in your working directory.

The Chip-Seq workflow is an automated process that includes a number of analyses: clustering including read and cluster classification, creation of a cluster sequence file, and TFBS overrepresentation analysis. Additionally, a *de novo* definition of TF binding sites from the ChIP cluster sequences is possible. This uses the program CoreSearch, which can, of course, also be run separately.

Please select “ChipSeq Workflow” in the NGS Analysis menu.



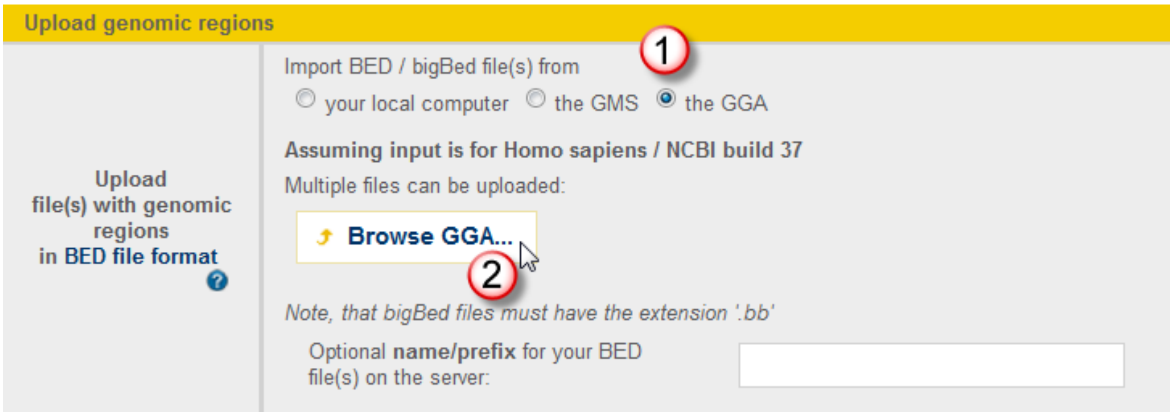
On the input page, press the Add BED files button.


Input file(s) with read positions (Sample or Treatment)
Note: multiple files are treated as replicates

Available files

Listing files for Homo sapiens / NCBI build 37:
Select ● BED files or ○ BAM files
No BED/BB files for Homo sapiens / NCBI build 37 in this project yet.

In the upload dialog, select the GGA for the file import and press the Browse GGA button.


Upload genomic regions

Upload file(s) with genomic regions in BED file format

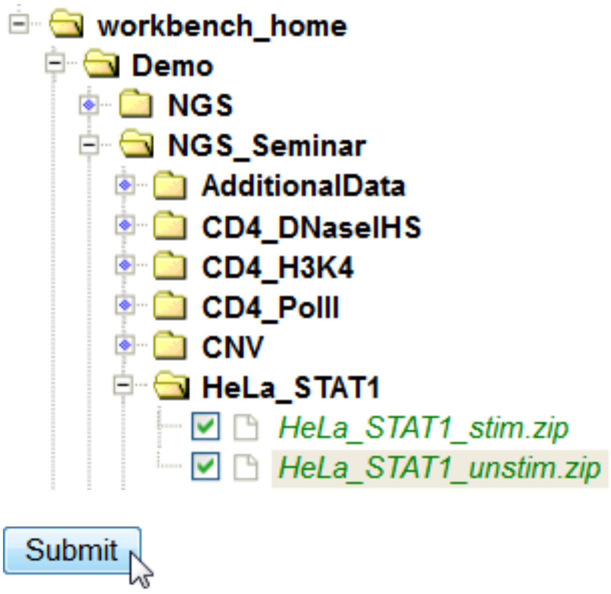
Import BED / bigBed file(s) from
○ your local computer ○ the GMS ● the GGA

Assuming input is for Homo sapiens / NCBI build 37
Multiple files can be uploaded:

Note, that bigBed files must have the extension '.bb'

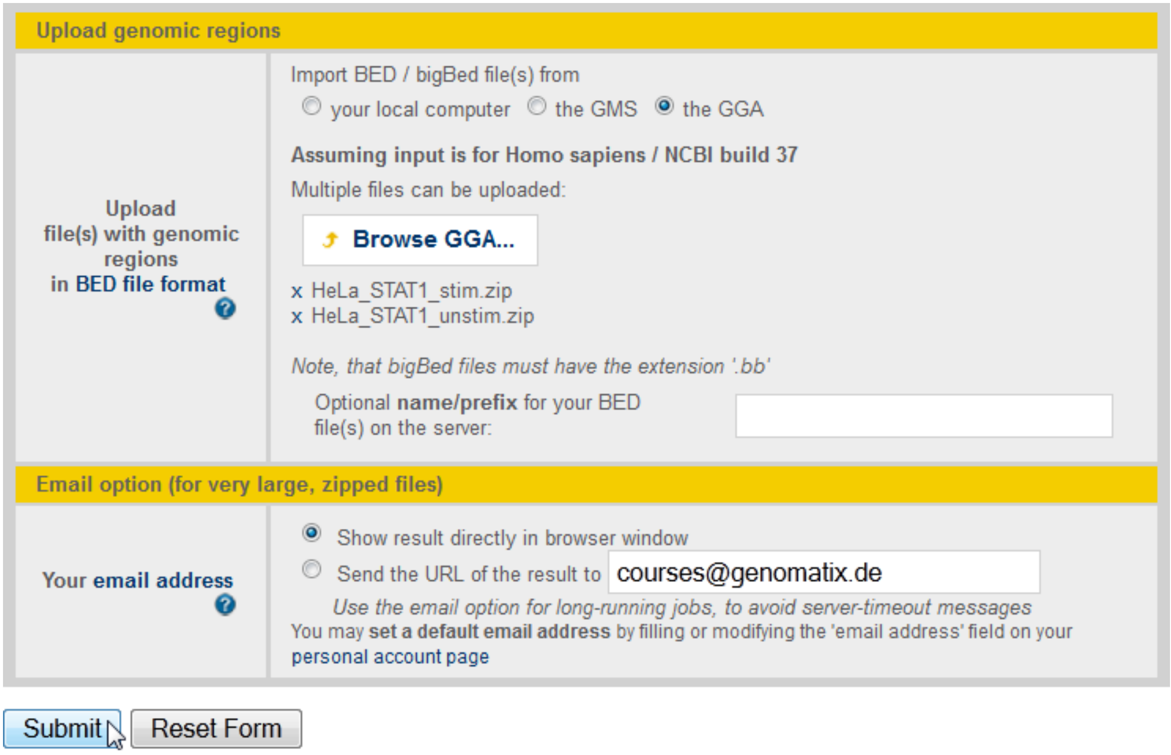
Optional **name/prefix** for your BED file(s) on the server:

You will find the files `HeLa_STAT1_stim.zip` and `HeLa_STAT1_unstim.zip` in the directory `/workbench_home/Demo/NGS_Seminar/HeLa_STAT1`.



- workbench_home
  - Demo
    - NGS
    - NGS_Seminar
      - AdditionalData
      - CD4_DNaseIHS
      - CD4_H3K4
      - CD4_PolII
      - CNV
      - HeLa_STAT1
        - ☑ *HeLa_STAT1_stim.zip*
        - ☑ *HeLa_STAT1_unstim.zip*

Press Submit in the upload dialog to start the import process.



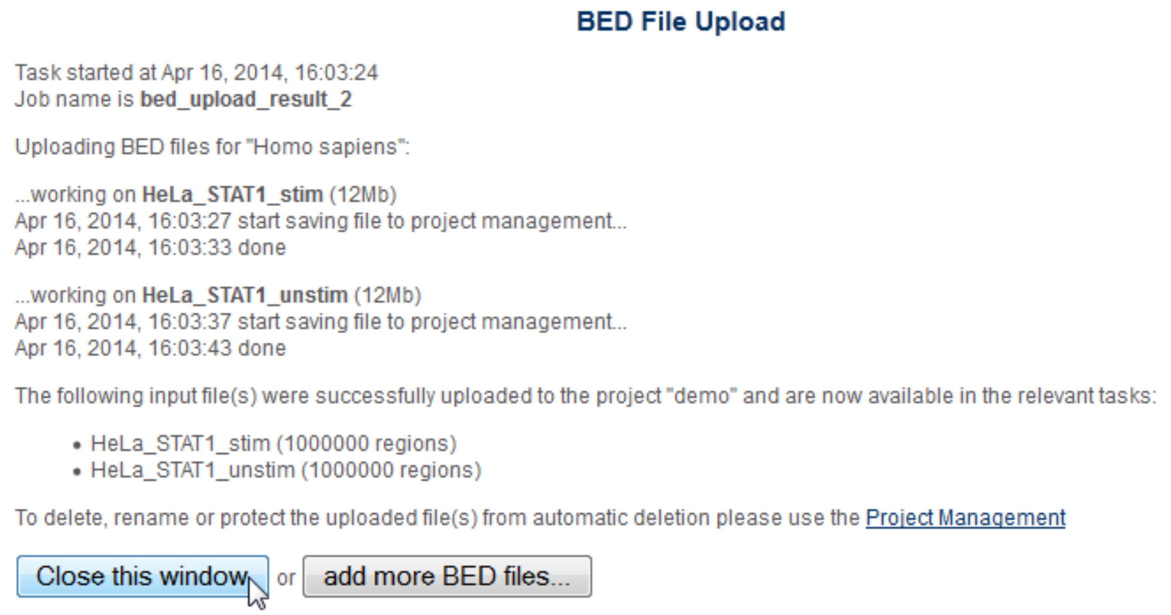
Upload genomic regions

Upload file(s) with genomic regions in BED file format	Import BED / bigBed file(s) from ○ your local computer ○ the GMS ● the GGA **Assuming input is for Homo sapiens / NCBI build 37** Multiple files can be uploaded: Browse GGA... x HeLa_STAT1_stim.zip x HeLa_STAT1_unstim.zip *Note, that bigBed files must have the extension '.bb'* Optional **name/prefix** for your BED file(s) on the server:

Email option (for very large, zipped files)

Your email address	● Show result directly in browser window ○ Send the URL of the result to courses@genomatix.de *Use the email option for long-running jobs, to avoid server-timeout messages* You may **set a default email address** by filling or modifying the 'email address' field on your personal account page

When the upload has finished, press the Close this window button.



BED File Upload

Task started at Apr 16, 2014, 16:03:24
Job name is **bed_upload_result_2**

Uploading BED files for "Homo sapiens":

...working on **HeLa_STAT1_stim** (12Mb)
Apr 16, 2014, 16:03:27 start saving file to project management...
Apr 16, 2014, 16:03:33 done

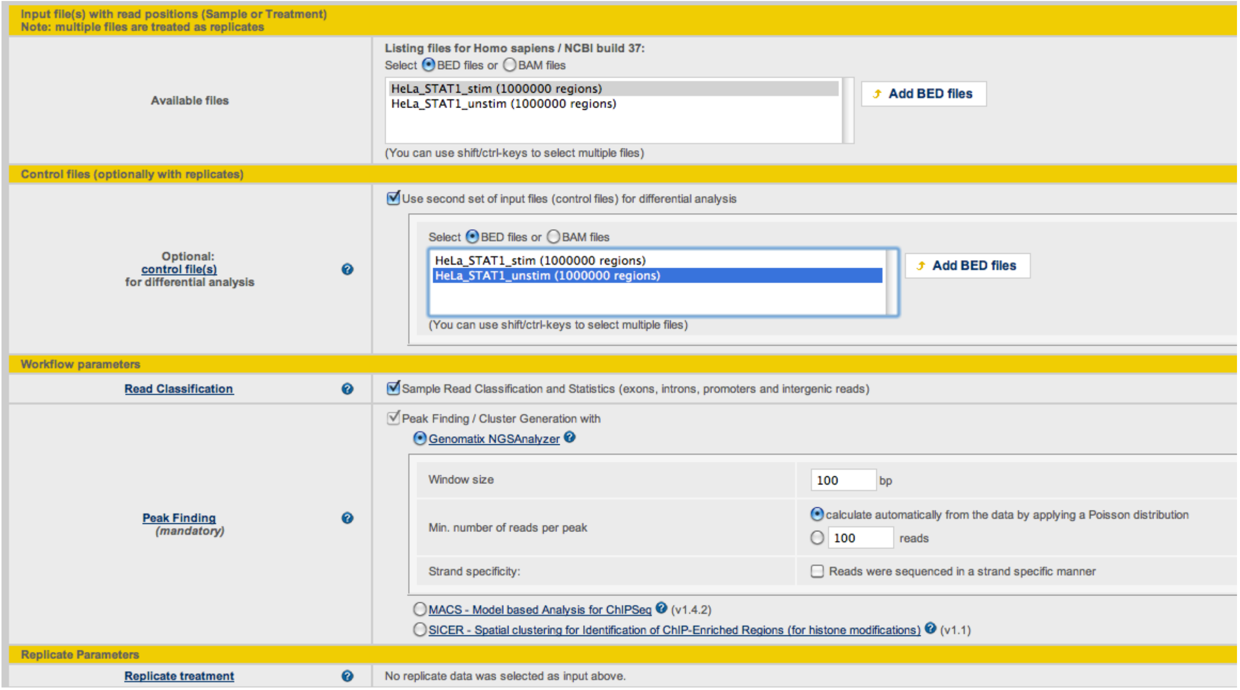
...working on **HeLa_STAT1_unstim** (12Mb)
Apr 16, 2014, 16:03:37 start saving file to project management...
Apr 16, 2014, 16:03:43 done

The following input file(s) were successfully uploaded to the project "demo" and are now available in the relevant tasks:

- HeLa_STAT1_stim (1000000 regions)
- HeLa_STAT1_unstim (1000000 regions)

To delete, rename or protect the uploaded file(s) from automatic deletion please use the Project Management

In the BED file lists, choose `HeLa_STAT1_stim` as sample and `HeLa_STAT1_unstim` as control file.



Input file(s) with read positions (Sample or Treatment)
Note: multiple files are treated as replicates

Available files

Listing files for Homo sapiens / NCBI build 37:
Select ● BED files or ○ BAM files
HeLa_STAT1_stim (1000000 regions)
HeLa_STAT1_unstim (1000000 regions)
(You can use shift/ctrl-keys to select multiple files)

Control files (optionally with replicates)

Optional:
control file(s)
for differential analysis

☑ Use second set of input files (control files) for differential analysis
Select ● BED files or ○ BAM files
HeLa_STAT1_stim (1000000 regions)
HeLa_STAT1_unstim (1000000 regions)
(You can use shift/ctrl-keys to select multiple files)

Workflow parameters

Read Classification: ☑ Sample Read Classification and Statistics (exons, introns, promoters and intergenic reads)

Peak Finding (mandatory): ☑ Peak Finding / Cluster Generation with
● Genomatix NGSAnalyzer

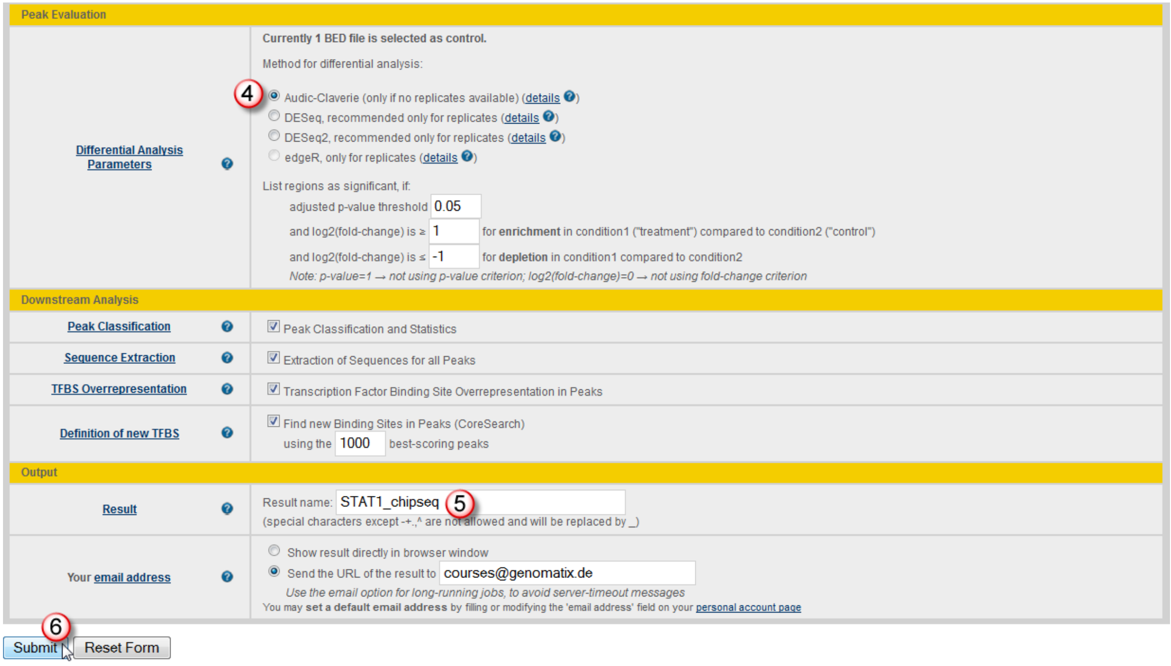
Window size	100 bp
Min. number of reads per peak	● calculate automatically from the data by applying a Poisson distribution ○ 100 reads
Strand specificity:	□ Reads were sequenced in a strand specific manner

○ MACS - Model based Analysis for ChIPSeq (v1.4.2)
○ SICER - Spatial clustering for Identification of ChIP-Enriched Regions (for histone modifications) (v1.1)

Replicate Parameters

Replicate treatment: No replicate data was selected as input above.

Make sure “Audic-Claverie” is selected as differential analysis method. Provide a result name, and start the analysis with the default e-mail option.



Peak Evaluation

Differential Analysis Parameters

Currently 1 BED file is selected as control.

Method for differential analysis:

4

● Audic-Claverie (only if no replicates available) (details)

○ DESeq, recommended only for replicates (details)

○ DESeq2, recommended only for replicates (details)

○ edgeR, only for replicates (details)

List regions as significant, if:

adjusted p-value threshold 0.05

and log2(fold-change) is ≥ 1 for **enrichment** in condition1 ("treatment") compared to condition2 ("control")

and log2(fold-change) is ≤ -1 for **depletion** in condition1 compared to condition2

Note: p-value=1 → not using p-value criterion; log2(fold-change)=0 → not using fold-change criterion

Downstream Analysis

Peak Classification	☑ Peak Classification and Statistics
Sequence Extraction	☑ Extraction of Sequences for all Peaks
TFBS Overrepresentation	☑ Transcription Factor Binding Site Overrepresentation in Peaks
Definition of new TFBS	☑ Find new Binding Sites in Peaks (CoreSearch) using the 1000 best-scoring peaks

Output

Result

Result name: STAT1_chipseq 5

(special characters except -+.,^ are not allowed and will be replaced by _)

Your email address

○ Show result directly in browser window

● Send the URL of the result to courses@genomatix.de

Use the email option for long-running jobs, to avoid server-timeout messages

You may **set a default email address** by filling or modifying the 'email address' field on your personal account page

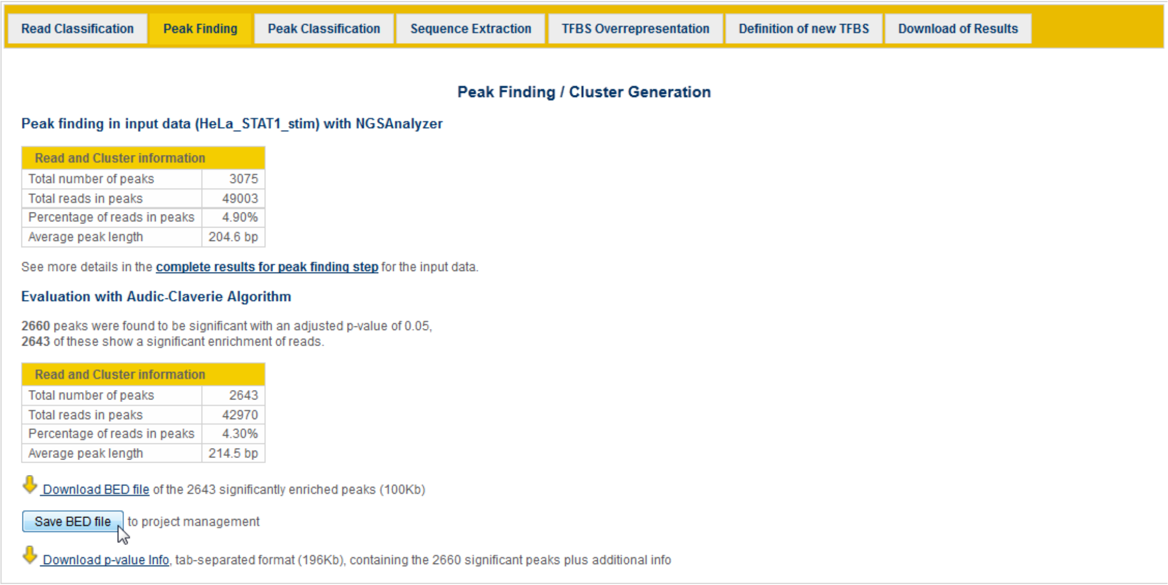
6

When the analysis is done, open the result from the project management page.

Peak finding

The output page has its own navigation bar, which is used to access each workflow result. The peak finding result is shown by default.

In the experimental sample, 3075 peaks were found originally, of which 2643 enriched peaks remain after Audic-Claverie evaluation. 4.3% of the reads are in these clusters, which is a typical value.



Peak Finding / Cluster Generation

Peak finding in input data (HeLa_STAT1_stim) with NGSAnalyzer

Read and Cluster information	
Total number of peaks	3075
Total reads in peaks	49003
Percentage of reads in peaks	4.90%
Average peak length	204.6 bp

See more details in the complete results for peak finding step for the input data.

Evaluation with Audic-Claverie Algorithm

2660 peaks were found to be significant with an adjusted p-value of 0.05,
2643 of these show a significant enrichment of reads.

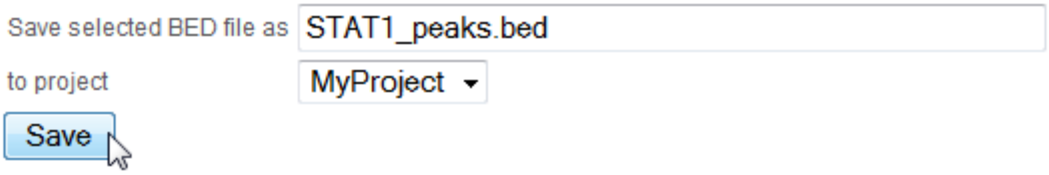
Read and Cluster information	
Total number of peaks	2643
Total reads in peaks	42970
Percentage of reads in peaks	4.30%
Average peak length	214.5 bp

Download BED file of the 2643 significantly enriched peaks (100Kb)

Save BED file to project management

Download p-value Info, tab-separated format (196Kb), containing the 2660 significant peaks plus additional info

Please save the BED file with significantly enriched clusters to the project management; we will need it at a later step.

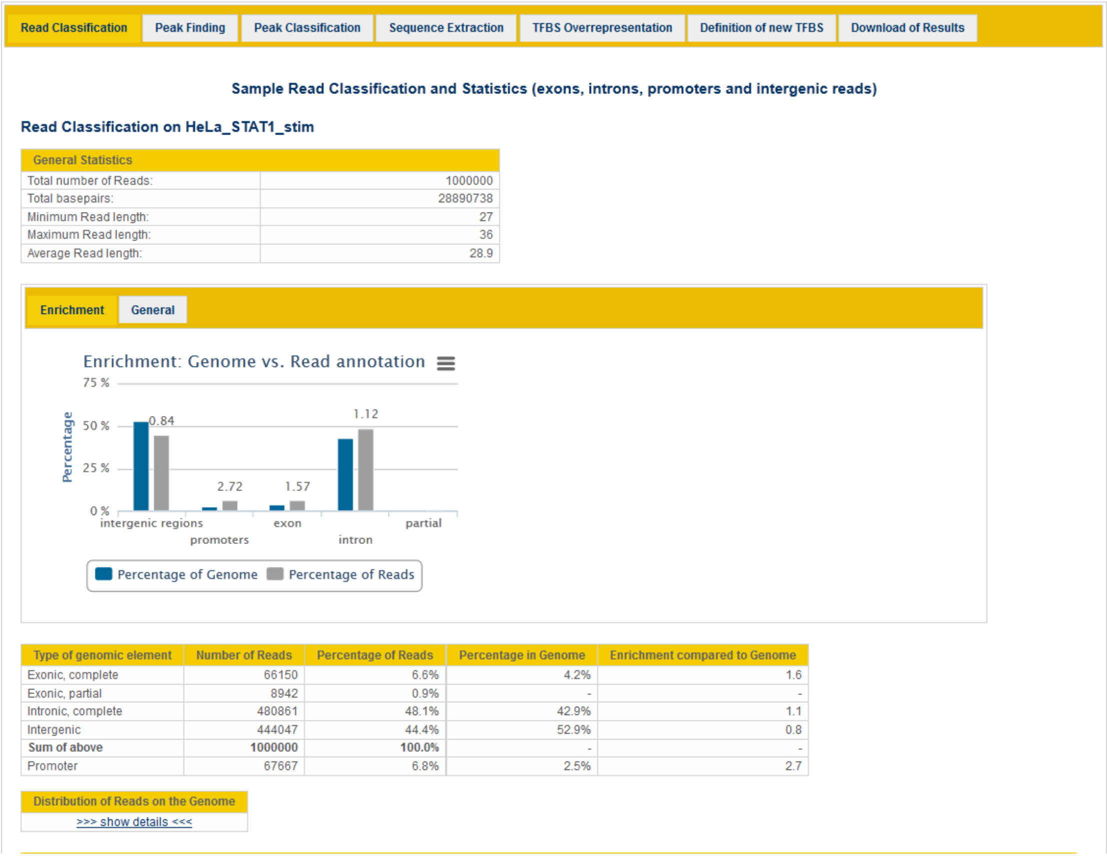


Save selected BED file as: STAT1_peaks.bed

to project: MyProject

Read classification

The read classification shows some enrichment in promoters, a little more pronounced in IFN-gamma stimulated compared to unstimulated cells:



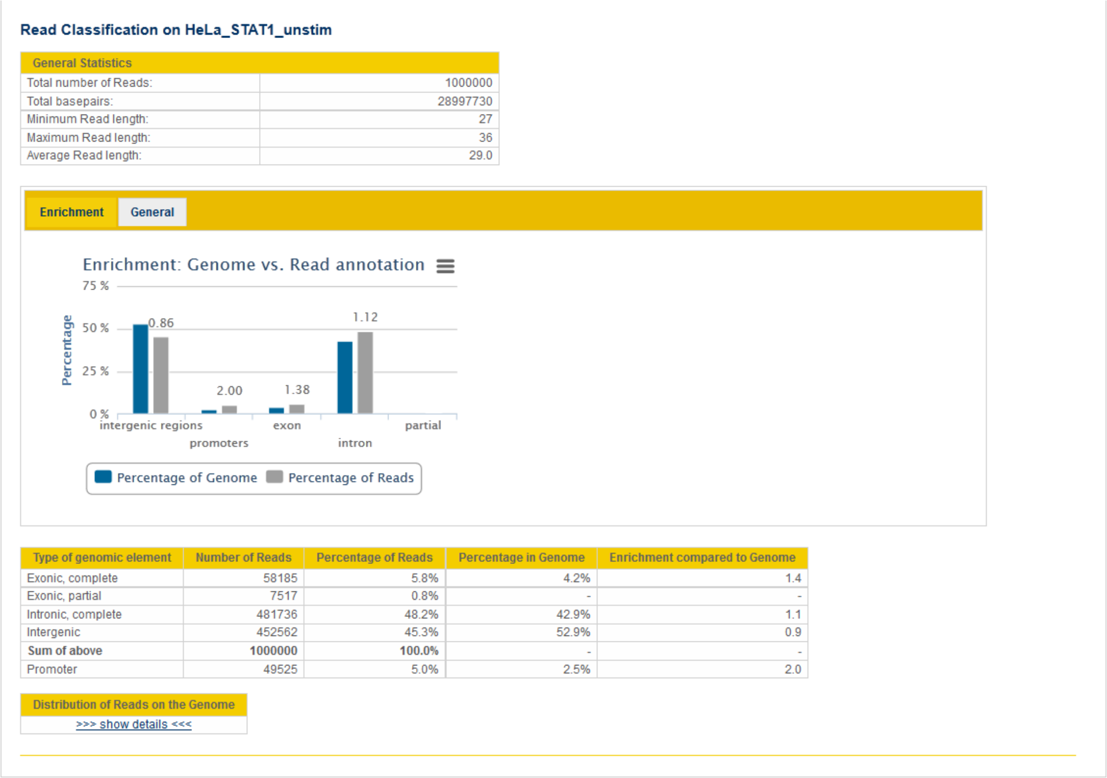
Sample Read Classification and Statistics (exons, introns, promoters and intergenic reads)

Read Classification on HeLa_STAT1_stim

General Statistics	
Total number of Reads:	1000000
Total basepairs:	28890738
Minimum Read length:	27
Maximum Read length:	36
Average Read length:	28.9

Type of genomic element	Number of Reads	Percentage of Reads	Percentage in Genome	Enrichment compared to Genome
Exonic, complete	66150	6.6%	4.2%	1.6
Exonic, partial	8942	0.9%	-	-
Intronic, complete	480861	48.1%	42.9%	1.1
Intergenic	444047	44.4%	52.9%	0.8
Sum of above	**1000000**	**100.0%**	-	-
Promoter	67667	6.8%	2.5%	2.7

Distribution of Reads on the Genome
>>> show details <<<



Read Classification on HeLa_STAT1_unstim

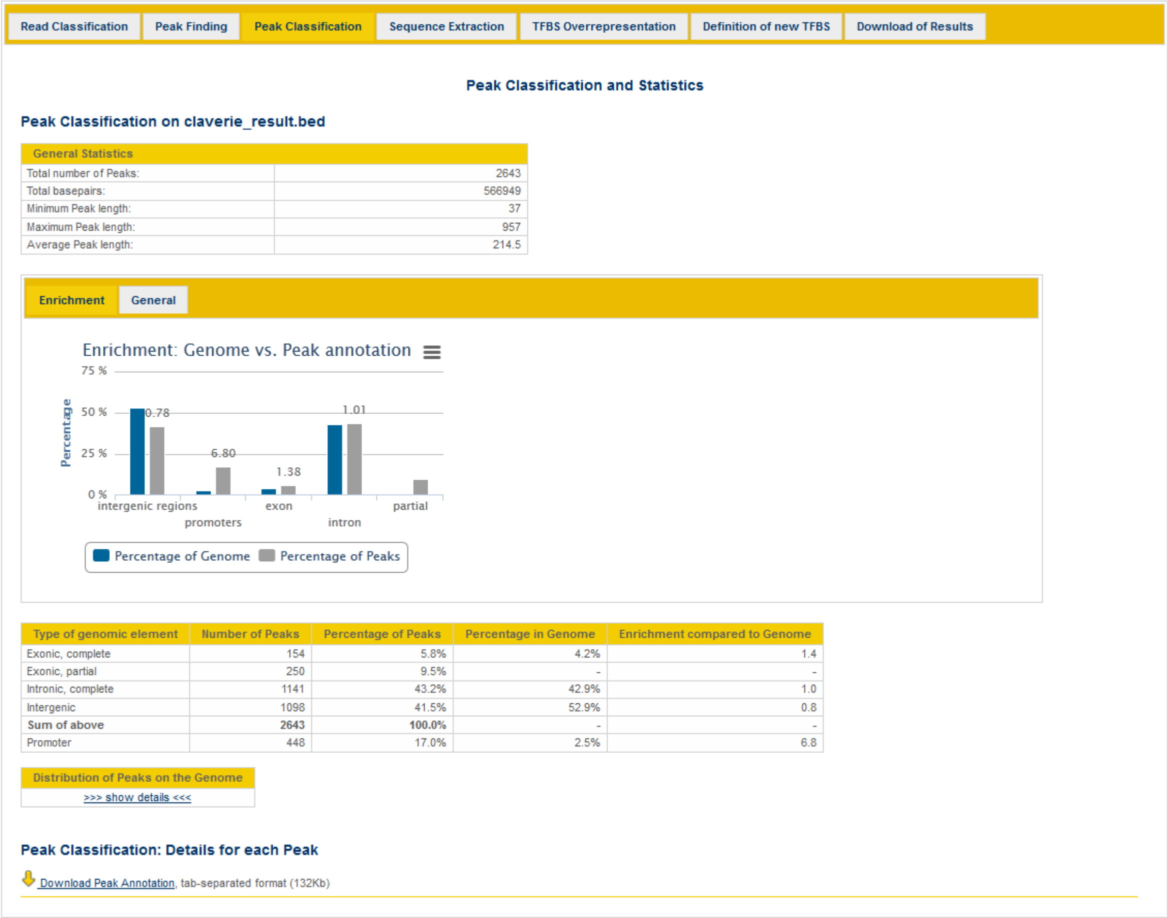
General Statistics	
Total number of Reads:	1000000
Total basepairs:	28997730
Minimum Read length:	27
Maximum Read length:	36
Average Read length:	29.0

Type of genomic element	Number of Reads	Percentage of Reads	Percentage in Genome	Enrichment compared to Genome
Exonic, complete	58185	5.8%	4.2%	1.4
Exonic, partial	7517	0.8%	-	-
Intronic, complete	481736	48.2%	42.9%	1.1
Intergenic	452562	45.3%	52.9%	0.9
Sum of above	**1000000**	**100.0%**	-	-
Promoter	49525	5.0%	2.5%	2.0

Distribution of Reads on the Genome
>>> show details <<<

Peak classification

The enrichment in promoters is 6.6 fold for peaks (reads: 2.7 fold for the stimulated data set).



Peak Classification and Statistics

Peak Classification on claverie_result.bed

General Statistics	
Total number of Peaks:	2643
Total basepairs:	566949
Minimum Peak length:	37
Maximum Peak length:	957
Average Peak length:	214.5

Type of genomic element	Number of Peaks	Percentage of Peaks	Percentage in Genome	Enrichment compared to Genome
Exonic, complete	154	5.8%	4.2%	1.4
Exonic, partial	250	9.5%	-	-
Intronic, complete	1141	43.2%	42.9%	1.0
Intergenic	1098	41.5%	52.9%	0.8
Sum of above	**2643**	**100.0%**	-	-
Promoter	448	17.0%	2.5%	6.8

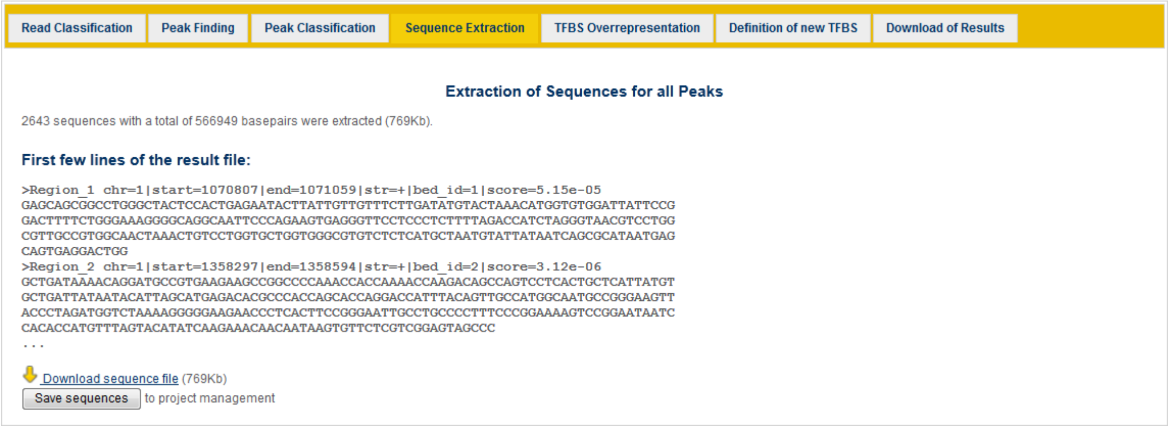
Distribution of Peaks on the Genome
>>> show details <<<

Peak Classification: Details for each Peak

Download Peak Annotation, tab-separated format (132Kb)

Sequence extraction

The peak sequences can be saved in the next section:



Extraction of Sequences for all Peaks

2643 sequences with a total of 566949 basepairs were extracted (769Kb).

First few lines of the result file:

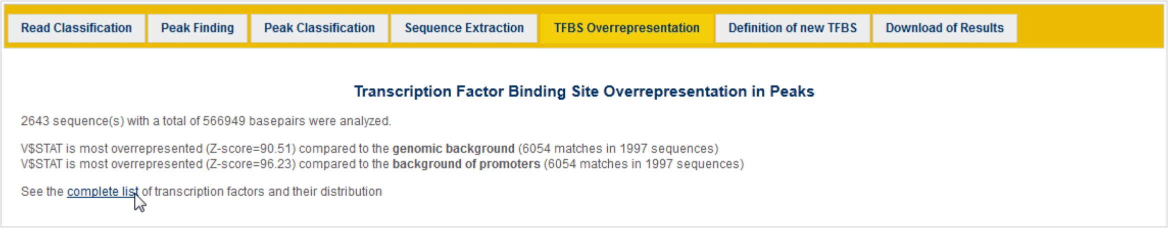
```
>Region_1 chr=1|start=1070807|end=1071059|str=+|bed_id=1|score=5.15e-05
GAGCAGCGGCCTGGGCTACTCCACTGAGAATACTTATTGTTGTTTCTTGATATGTACTAAACATGGTGTGGATTATTCCG
GACTTTTCTGGGAAAGGGGCAGGCAATTCCCAGAAGTGAGGGTTCCTCCCTCTTTTAGACCATCTAGGGTAACGTCCTGG
CGTTGCCGTGGCAACTAAACTGTCCTGGTGCTGGTGGGCGTGTCTCTCATGCTAATGTATTATAATCAGCGCATAATGAG
CAGTGAGGACTGG
>Region_2 chr=1|start=1358297|end=1358594|str=+|bed_id=2|score=3.12e-06
GCTGATAAAACAGGATGCCGTGAAGAAGCCGGCCCCAAACCACCAAAACCAAGACAGCCAGTCCTCACTGCTCATTATGT
GCTGATTATAATACATTAGCATGAGACACGCCCACCAGCACCAGGACCATTTACAGTTGCCATGGCAATGCCGGGAAGTT
ACCCTAGATGGTCTAAAAGGGGGAAGAACCCTCACTTCCGGGAATTGCCTGCCCCTTTCCCGGAAAAGTCCGGAATAATC
CACACCATGTTTAGTACATATCAAGAAACAACAATAAGTGTTCTCGTCGGAGTAGCCC
...
```

Download sequence file (769Kb)

to project management

TFBS overrepresentation

Next, we'll have a look which transcription factor binding sites can be found in the clusters. A short summary of the TFBS analysis is given in the overview: V$STAT, the binding site family for STAT1, is most overrepresented, both against a genomic and a promoter background.



Transcription Factor Binding Site Overrepresentation in Peaks

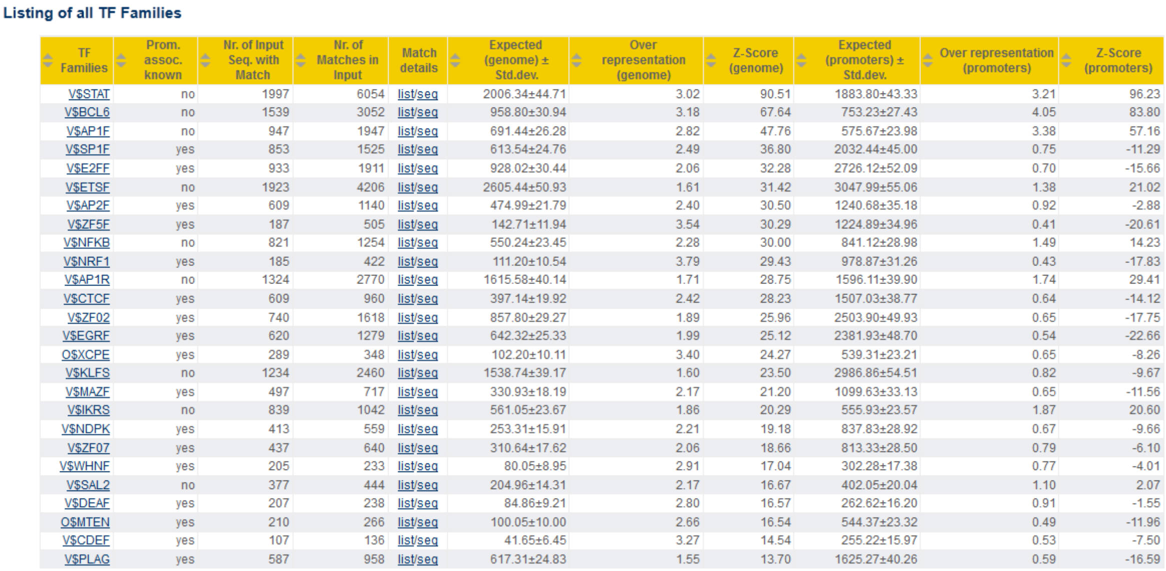
2643 sequence(s) with a total of 566949 basepairs were analyzed.

V$STAT is most overrepresented (Z-score=90.51) compared to the **genomic background** (6054 matches in 1997 sequences)
V$STAT is most overrepresented (Z-score=96.23) compared to the **background of promoters** (6054 matches in 1997 sequences)

See the complete list of transcription factors and their distribution

Click the “complete list” link to open the detailed result page.

You'll see some statistics on top and then a table containing all transcription factor binding site matches together with overrepresentation values and Z-scores.



Listing of all TF Families

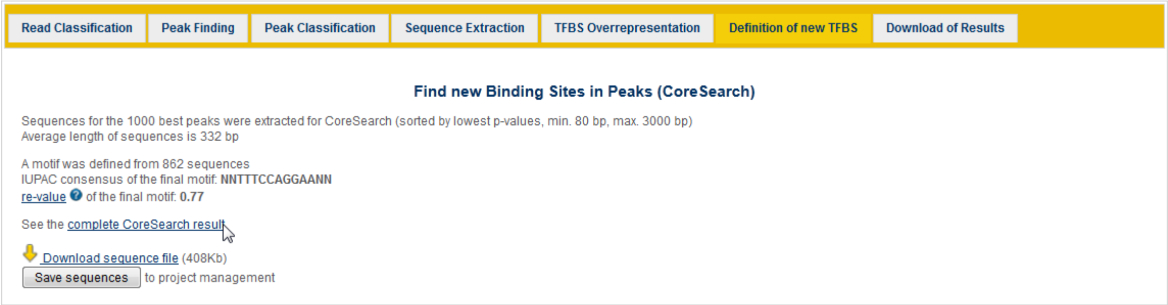
TF Families	Prom. assoc. known	Nr. of Input Seq. with Match	Nr. of Matches in Input	Match details	Expected (genome) ± Std.dev.	Over representation (genome)	Z-Score (genome)	Expected (promoters) ± Std.dev.	Over representation (promoters)	Z-Score (promoters)
V$STAT	no	1997	6054	list/seq	2006.34±44.71	3.02	90.51	1883.80±43.33	3.21	96.23
V$BCL6	no	1539	3052	list/seq	958.80±30.94	3.18	67.64	753.23±27.43	4.05	83.80
V$AP1F	no	947	1947	list/seq	691.44±26.28	2.82	47.76	575.67±23.98	3.38	57.16
V$SP1F	yes	853	1525	list/seq	613.54±24.76	2.49	36.80	2032.44±45.00	0.75	-11.29
V$E2FF	yes	933	1911	list/seq	928.02±30.44	2.06	32.28	2726.12±52.09	0.70	-15.66
V$ETSF	no	1923	4206	list/seq	2605.44±50.93	1.61	31.42	3047.99±55.06	1.38	21.02
V$AP2F	yes	609	1140	list/seq	474.99±21.79	2.40	30.50	1240.68±35.18	0.92	-2.88
V$ZF5F	yes	187	505	list/seq	142.71±11.94	3.54	30.29	1224.89±34.96	0.41	-20.61
V$NFKB	no	821	1254	list/seq	550.24±23.45	2.28	30.00	841.12±28.98	1.49	14.23
V$NRF1	yes	185	422	list/seq	111.20±10.54	3.79	29.43	978.87±31.26	0.43	-17.83
V$AP1R	no	1324	2770	list/seq	1615.58±40.14	1.71	28.75	1596.11±39.90	1.74	29.41
V$CTCF	yes	609	960	list/seq	397.14±19.92	2.42	28.23	1507.03±38.77	0.64	-14.12
V$ZF02	yes	740	1618	list/seq	857.80±29.27	1.89	25.96	2503.90±49.93	0.65	-17.75
V$EGRF	yes	620	1279	list/seq	642.32±25.33	1.99	25.12	2381.93±48.70	0.54	-22.66
O$XCPE	yes	289	348	list/seq	102.20±10.11	3.40	24.27	539.31±23.21	0.65	-8.26
V$KLFS	no	1234	2460	list/seq	1538.74±39.17	1.60	23.50	2986.86±54.51	0.82	-9.67
V$MAZF	yes	497	717	list/seq	330.93±18.19	2.17	21.20	1099.63±33.13	0.65	-11.56
V$IKRS	no	839	1042	list/seq	561.05±23.67	1.86	20.29	555.93±23.57	1.87	20.60
V$NDPK	yes	413	559	list/seq	253.31±15.91	2.21	19.18	837.83±28.92	0.67	-9.66
V$ZF07	yes	437	640	list/seq	310.64±17.62	2.06	18.66	813.33±28.50	0.79	-6.10
V$WHNF	yes	205	233	list/seq	80.05±8.95	2.91	17.04	302.28±17.38	0.77	-4.01
V$SAL2	no	377	444	list/seq	204.96±14.31	2.17	16.67	402.05±20.04	1.10	2.07
V$DEAF	yes	207	238	list/seq	84.86±9.21	2.80	16.57	262.62±16.20	0.91	-1.55
O$MTEN	yes	210	266	list/seq	100.05±10.00	2.66	16.54	544.37±23.32	0.49	-11.96
V$CDEF	yes	107	136	list/seq	41.65±6.45	3.27	14.54	255.22±15.97	0.53	-7.50
V$PLAG	yes	587	958	list/seq	617.31±24.83	1.55	13.70	1625.27±40.26	0.59	-16.59

The list is sorted by the Z-score of the overrepresentation over the genome. The overrepresentation for V$STAT is about 3 fold over the genome background and 3.2 over the promoter background, and the Z-scores are quite high, indicating that it is statistically highly unlikely to find such an overrepresentation. You can click any column header to sort by that column; repeated clicking inverts the sort order.

Definition of new TFBS

The TFBS overrepresentation analysis uses pre-defined binding site matrices from the MatBase/MatInspector library provided with the Genomatix Genome Analyzer. It is, however, also possible to define your own matrices from the data generated by the ChIP-Seq experiment. In the workflow, the STAT1 cluster sequences were submitted to CoreSearch to generate a new STAT1 binding site matrix.

The next item in the workflow output overview is the CoreSearch result. The sequences of all clusters were used to generate a new matrix. The IUPAC consensus of the defined motif is very similar to the palindromic GAS motif (TTTCCNGGAAA) that binds STAT1 homodimers (described *e.g.* by Schindler *et al.*, 2007). For details, please click the “complete CoreSearch result” link.



Find new Binding Sites in Peaks (CoreSearch)

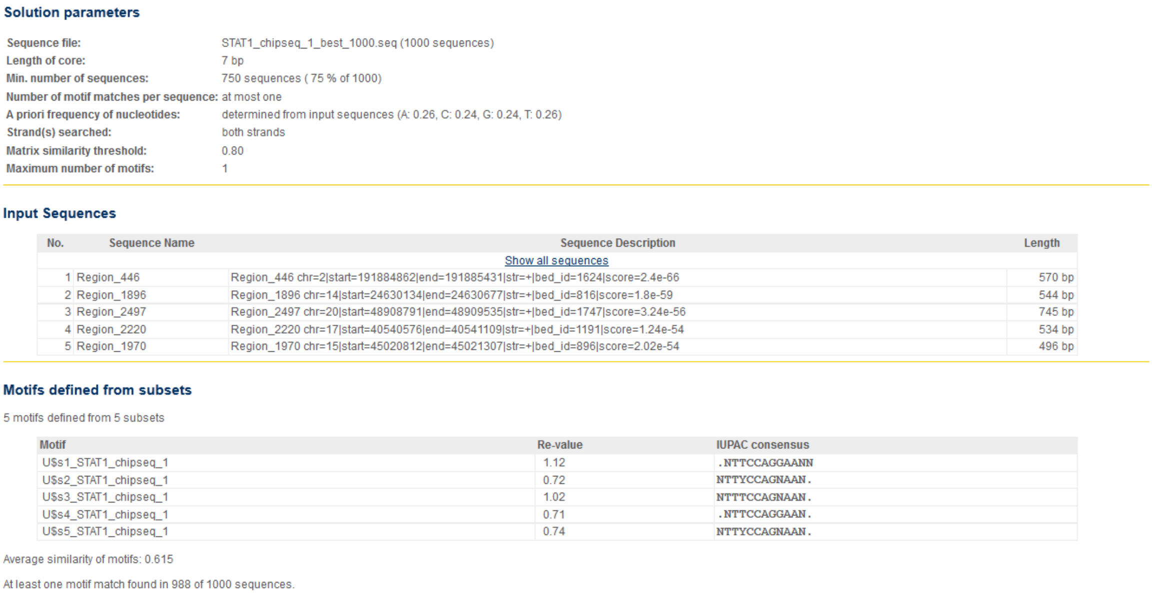
Sequences for the 1000 best peaks were extracted for CoreSearch (sorted by lowest p-values, min. 80 bp, max. 3000 bp)
Average length of sequences is 332 bp

A motif was defined from 862 sequences
IUPAC consensus of the final motif: **NNTTTCCAGGAANN**
re-value of the final motif: **0.77**

See the complete CoreSearch result

Download sequence file (408Kb)
Save sequences to project management

Here is an outline of the CoreSearch algorithm: as a first step, CoreSearch randomly picks sets of 100 input sequences to generate 5 matrices, which are grouped into a family. The IUPAC sequences of the matrices are displayed in the output below the list of input sequences:



Solution parameters

Sequence file:	STAT1_chipseq_1_best_1000.seq (1000 sequences)
Length of core:	7 bp
Min. number of sequences:	750 sequences (75 % of 1000)
Number of motif matches per sequence:	at most one
A priori frequency of nucleotides:	determined from input sequences (A: 0.26, C: 0.24, G: 0.24, T: 0.26)
Strand(s) searched:	both strands
Matrix similarity threshold:	0.80
Maximum number of motifs:	1

Input Sequences

No.	Sequence Name	Sequence Description	Length
		Show all sequences	
1	Region_446	Region_446 chr=2|start=191884862|end=191885431|str=+|bed_id=1624|score=2.4e-66	570 bp
2	Region_1896	Region_1896 chr=14|start=24630134|end=24630677|str=+|bed_id=816|score=1.8e-59	544 bp
3	Region_2497	Region_2497 chr=20|start=48908791|end=48909535|str=+|bed_id=1747|score=3.24e-56	745 bp
4	Region_2220	Region_2220 chr=17|start=40540576|end=40541109|str=+|bed_id=1191|score=1.24e-54	534 bp
5	Region_1970	Region_1970 chr=15|start=45020812|end=45021307|str=+|bed_id=896|score=2.02e-54	496 bp

Motifs defined from subsets

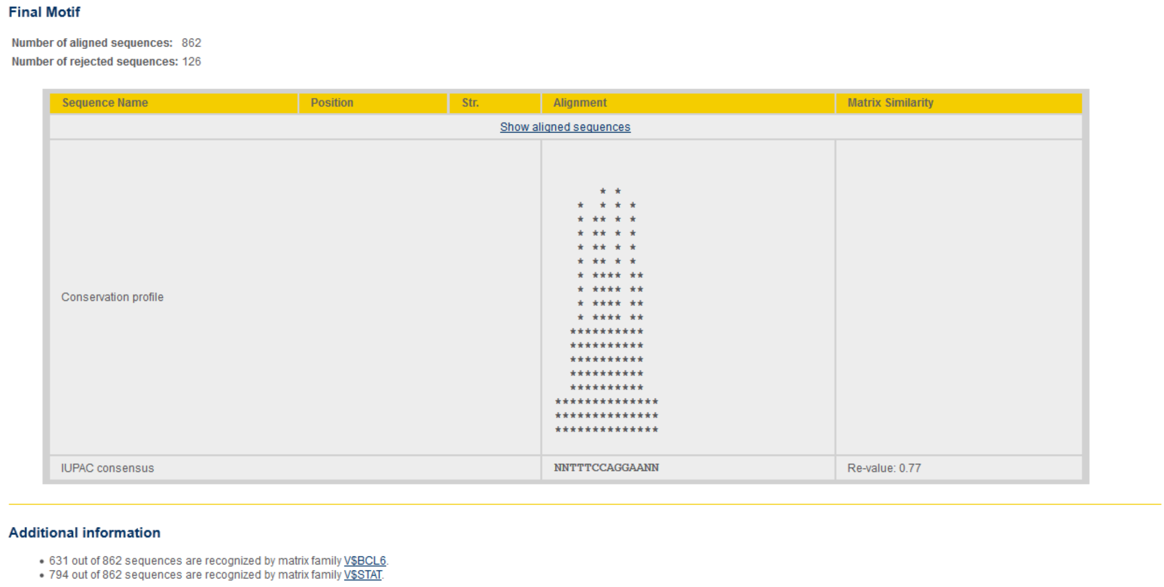
5 motifs defined from 5 subsets

Motif	Re-value	IUPAC consensus
U$s1_STAT1_chipseq_1	1.12	.NTTCCAGGAANN
U$s2_STAT1_chipseq_1	0.72	NTTYCCAGNAAN.
U$s3_STAT1_chipseq_1	1.02	NTTTCCAGNAAN.
U$s4_STAT1_chipseq_1	0.71	.NTTCCAGGAAN.
U$s5_STAT1_chipseq_1	0.74	NTTYCCAGNAAN.

Average similarity of motifs: 0.615

At least one motif match found in 988 of 1000 sequences.

All input sequences are then scanned for matches to the new matrix family, and the best match of each sequence is used to generate the final matrix. Its conservation profile is displayed at the end of the output page.



Final Motif

Number of aligned sequences: 862
Number of rejected sequences: 126

Sequence Name	Position	Str.	Alignment	Matrix Similarity
Show aligned sequences				
Conservation profile			(profile below)	
IUPAC consensus			NNTTTCCAGGAANN	Re-value: 0.77

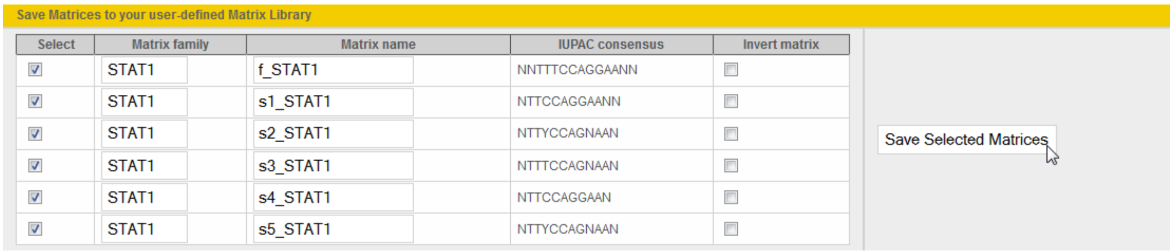
```
      * *
   *  * * *
   * ** * *
   * ** * *
   * ** * *
   * ** * *
   * **** **
   * **** **
   * **** **
   * **** **
  **********
  **********
  **********
  **********
  **********
**************
**************
**************
```

Additional information

- 631 out of 862 sequences are recognized by matrix family V$BCL6.
- 794 out of 862 sequences are recognized by matrix family V$STAT.

Most of the sequences used for generation of the matrix are also recognized by the existing STAT matrix family.

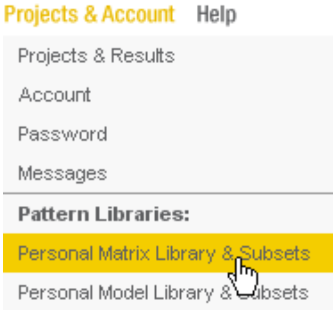
You can save any of the new matrices (the final one as well as the five matrices generated in the first step) in the 'Save Matrices to your user-defined Matrix Library' section at the bottom of the page. They are then available in tools applying matrix searches, such as MatInspector or RegionMiner.



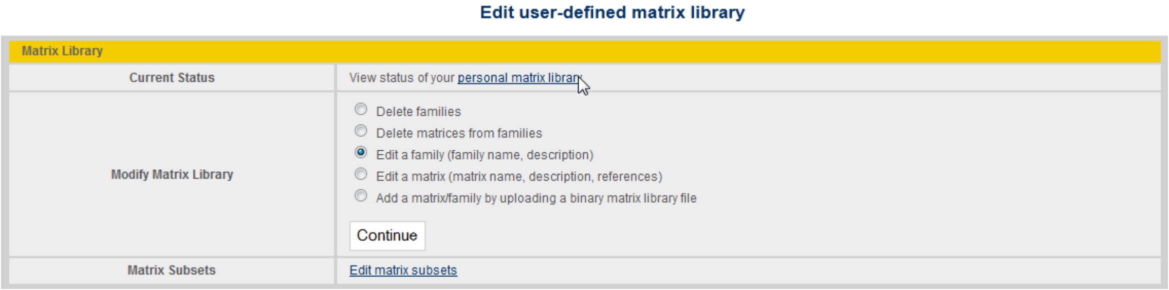
Save Matrices to your user-defined Matrix Library

Select	Matrix family	Matrix name	IUPAC consensus	Invert matrix
☑	STAT1	f_STAT1	NNTTTCCAGGAANN	☐
☑	STAT1	s1_STAT1	NTTCCAGGAANN	☐
☑	STAT1	s2_STAT1	NTTYCCAGNAAN	☐
☑	STAT1	s3_STAT1	NTTTCCAGNAAN	☐
☑	STAT1	s4_STAT1	NTTCCAGGAAN	☐
☑	STAT1	s5_STAT1	NTTYCCAGNAAN	☐

You can view your new matrices if you click the ‘Personal Matrix Library’ link in the menu:



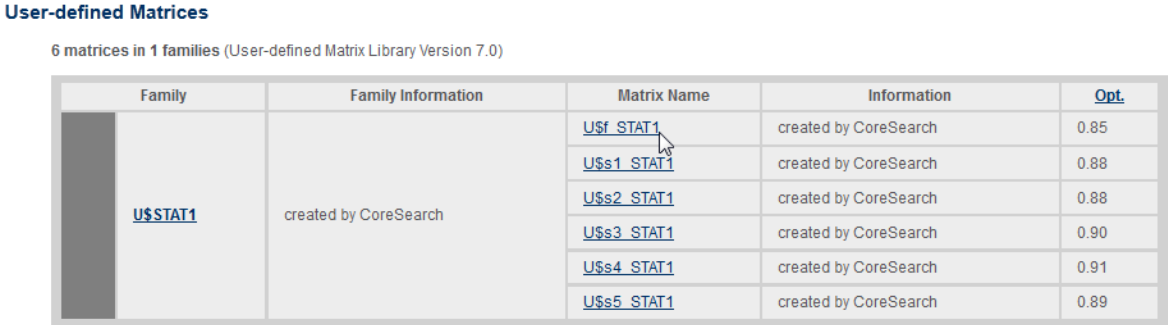
Select the “personal matrix library” link as shown below:



Edit user-defined matrix library

Matrix Library	
Current Status	View status of your personal matrix library
Modify Matrix Library	○ Delete families ○ Delete matrices from families ● Edit a family (family name, description) ○ Edit a matrix (matrix name, description, references) ○ Add a matrix/family by uploading a binary matrix library file Continue
Matrix Subsets	Edit matrix subsets

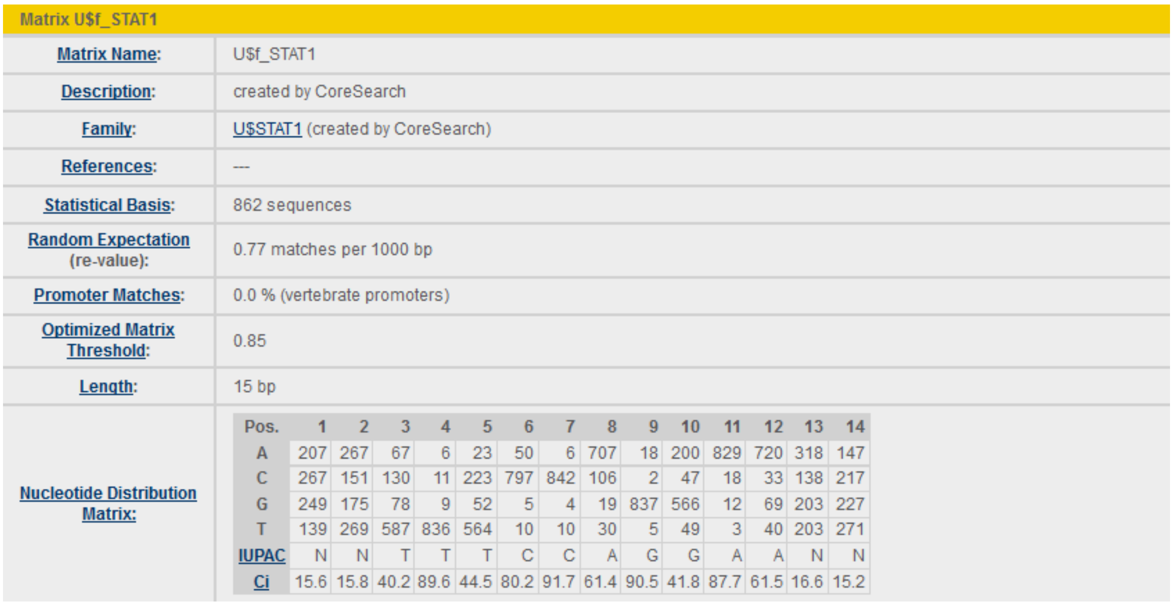
Click the first matrix name to display detailed information for this matrix.



User-defined Matrices

6 matrices in 1 families (User-defined Matrix Library Version 7.0)

Family	Family Information	Matrix Name	Information	Opt.
U$STAT1	created by CoreSearch	U$f_STAT1	created by CoreSearch	0.85
		U$s1_STAT1	created by CoreSearch	0.88
		U$s2_STAT1	created by CoreSearch	0.88
		U$s3_STAT1	created by CoreSearch	0.90
		U$s4_STAT1	created by CoreSearch	0.91
		U$s5_STAT1	created by CoreSearch	0.89

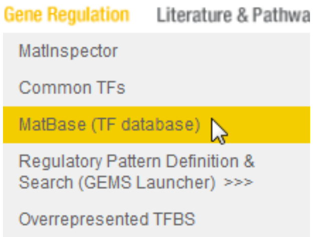


Matrix U$f_STAT1	
Matrix Name:	U$f_STAT1
Description:	created by CoreSearch
Family:	U$STAT1 (created by CoreSearch)
References:	---
Statistical Basis:	862 sequences
Random Expectation (re-value):	0.77 matches per 1000 bp
Promoter Matches:	0.0 % (vertebrate promoters)
Optimized Matrix Threshold:	0.85
Length:	15 bp

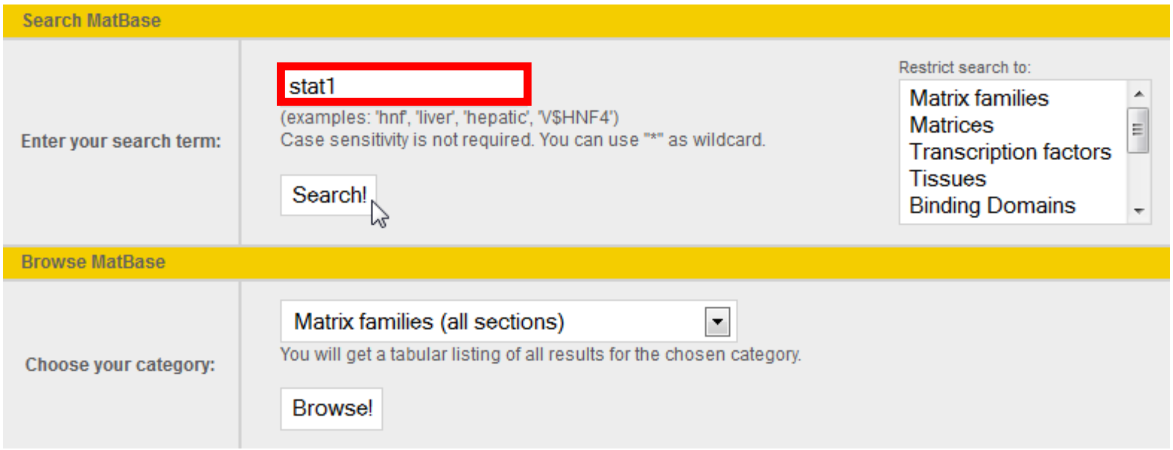
Nucleotide Distribution Matrix:

Pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	207	267	67	6	23	50	6	707	18	200	829	720	318	147
C	267	151	130	11	223	797	842	106	2	47	18	33	138	217
G	249	175	78	9	52	5	4	19	837	566	12	69	203	227
T	139	269	587	836	564	10	10	30	5	49	3	40	203	271
IUPAC	N	N	T	T	T	C	C	A	G	G	A	A	N	N
Ci	15.6	15.8	40.2	89.6	44.5	80.2	91.7	61.4	90.5	41.8	87.7	61.5	16.6	15.2

To compare the sequence logo to existing STAT matrices, select MatBase from the ‘Gene Regulation’ menu in the navigation bar:



Enter *e.g.* ‘stat1’ in the search field, and start the search.



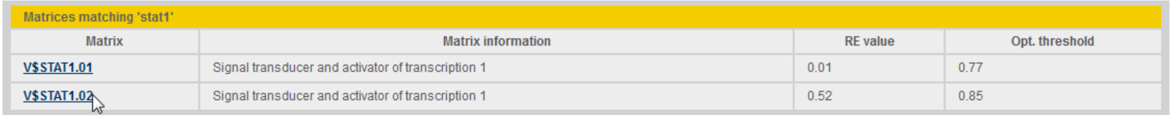
Search MatBase

Enter your search term: stat1

(examples: 'hnf', 'liver', 'hepatic', 'V$HNF4')
Case sensitivity is not required. You can use "*" as wildcard.

You will get a tabular listing of all results for the chosen category.

In the output, select one of the matching matrices, *e.g.* V$STAT1.02, as below:



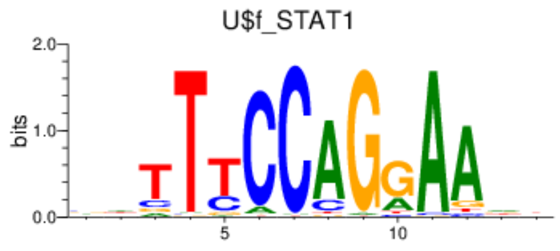
Matrices matching 'stat1'

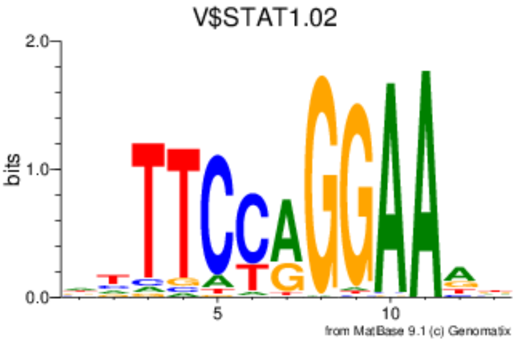
Matrix	Matrix information	RE value	Opt. threshold
V$STAT1.01	Signal transducer and activator of transcription 1	0.01	0.77
V$STAT1.02	Signal transducer and activator of transcription 1	0.52	0.85

Here is a side-by-side comparison of the new STAT site and the STAT1.02 site from MatBase:

New STAT site

STAT1 site from MatBase



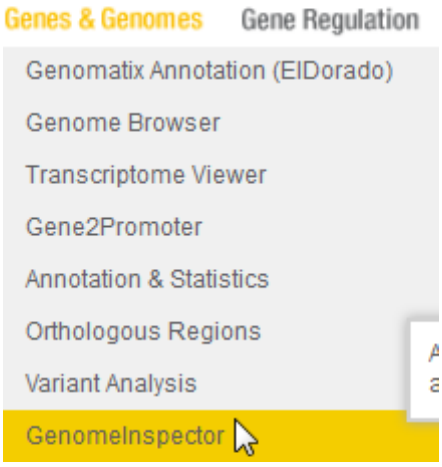


Positional correlation of ChIP-Seq data sets

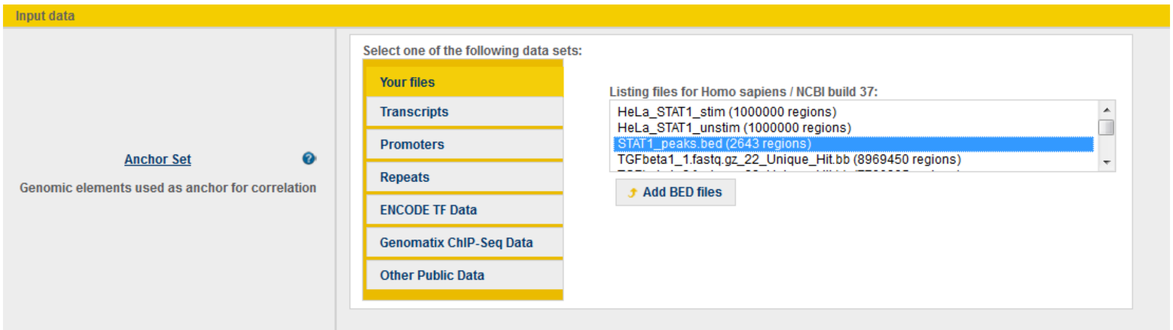
In order to characterize the identified STAT1 ChIP-Seq peak regions further, we will correlate the genomic positions in the STAT1 peak BED file with different data sets originating from the ENCODE project (ENCODE Project Consortium *et al.*, 2012).

The program GenomeInspector uses one BED file (the anchor set) and draws a correlation graph for up to 6 additional BED files (the partner sets). The graph shows the summarized coverage with regions from the partner sets in the vicinity of the regions in the anchor set.

Please start GenomeInspector from the Gene & Genomes menu.



Select the STAT1 peak BED file from your files in the list for the anchor set.



Anchor Set

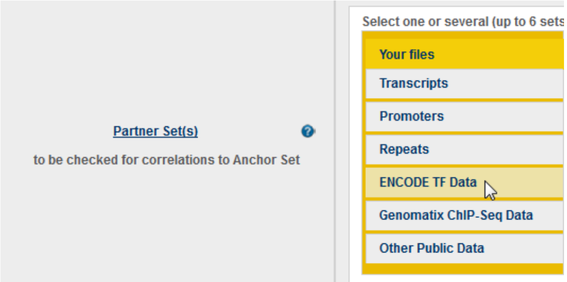
Genomic elements used as anchor for correlation

Select one of the following data sets:

Listing files for Homo sapiens / NCBI build 37:

```
HeLa_STAT1_stim (1000000 regions)
HeLa_STAT1_unstim (1000000 regions)
STAT1_peaks.bed (2643 regions)
TGFbeta1_1.fastq.gz_22_Unique_Hit.bb (8969450 regions)
```

For the partner set, select the ENCODE TF Data group.

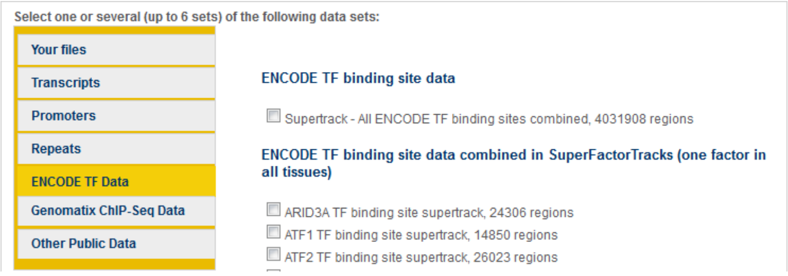


Partner Set(s)

to be checked for correlations to Anchor Set

Select one or several (up to 6 sets

Then, select the available STAT family TFBS supertracks. They contain all STAT1. STAT2, and STAT3 peak regions from the ENCODE project, merged from different cell lines.



Select one or several (up to 6 sets) of the following data sets:

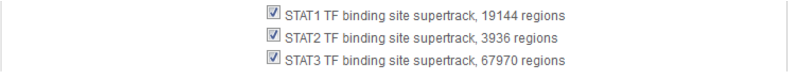
ENCODE TF binding site data

□ Supertrack - All ENCODE TF binding sites combined, 4031908 regions

ENCODE TF binding site data combined in SuperFactorTracks (one factor in all tissues)

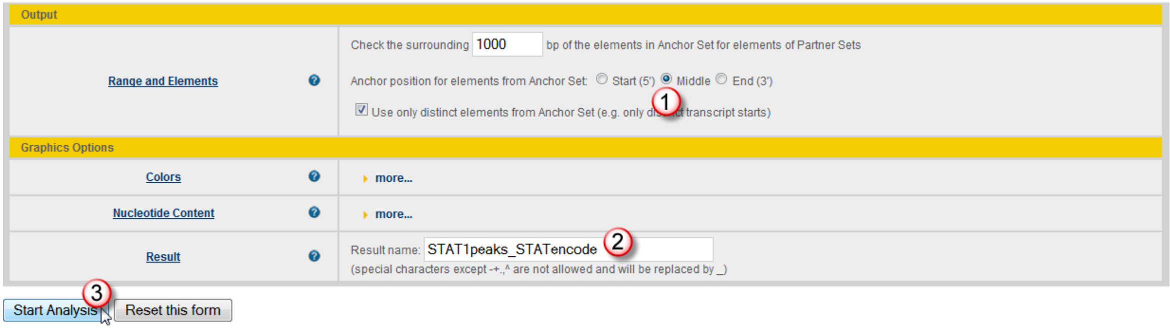
□ ARID3A TF binding site supertrack, 24306 regions
□ ATF1 TF binding site supertrack, 14850 regions
□ ATF2 TF binding site supertrack, 26023 regions

...



☑ STAT1 TF binding site supertrack, 19144 regions
☑ STAT2 TF binding site supertrack, 3936 regions
☑ STAT3 TF binding site supertrack, 67970 regions

Set the anchor position to the middle of the anchor set, provide a result name, and start the analysis.



Output

Range and Elements

Check the surrounding 1000 bp of the elements in Anchor Set for elements of Partner Sets

Anchor position for elements from Anchor Set: ○ Start (5') ● Middle ○ End (3')

1

☑ Use only distinct elements from Anchor Set (e.g. only di ct transcript starts)

Graphics Options

Colors: more...

Nucleotide Content: more...

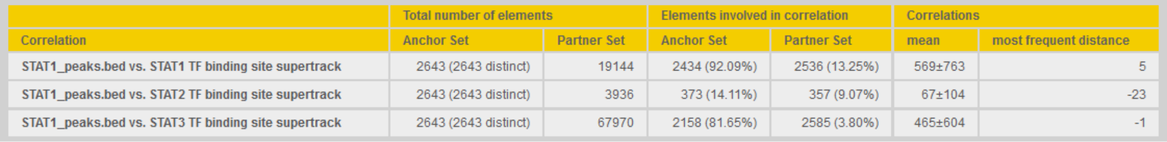
Result: Result name: STAT1peaks_STATencode

2

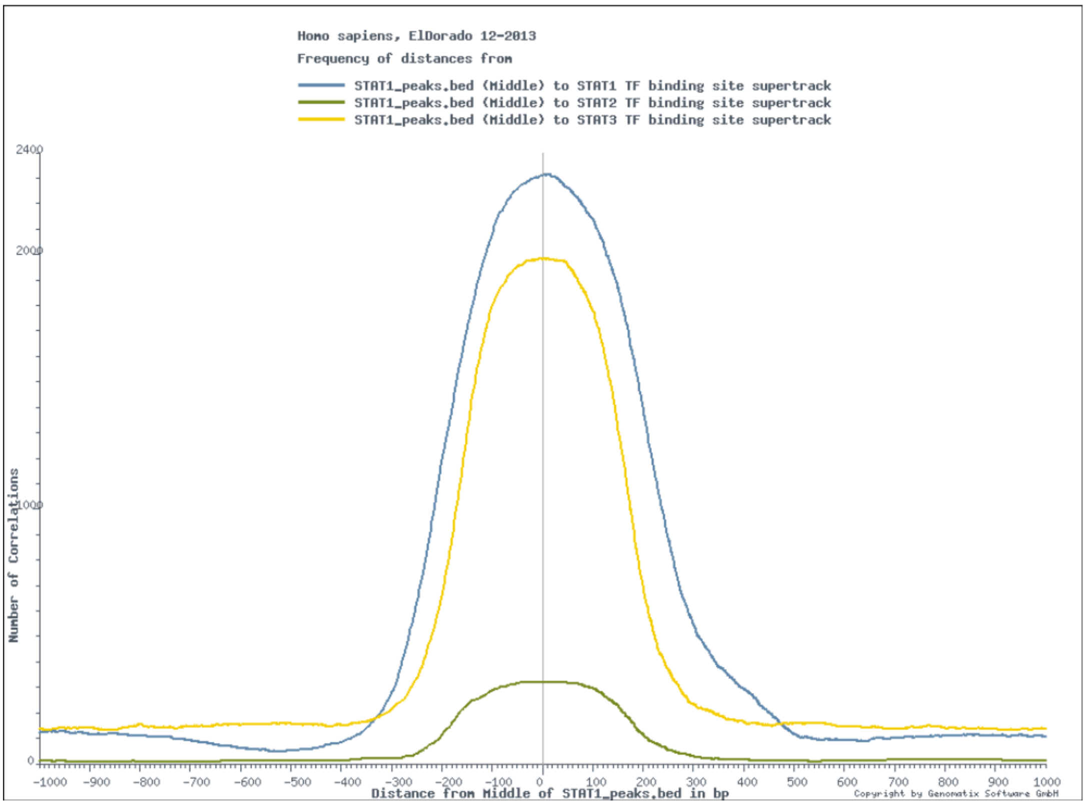
(special characters except -+.,^ are not allowed and will be replaced by _)

3

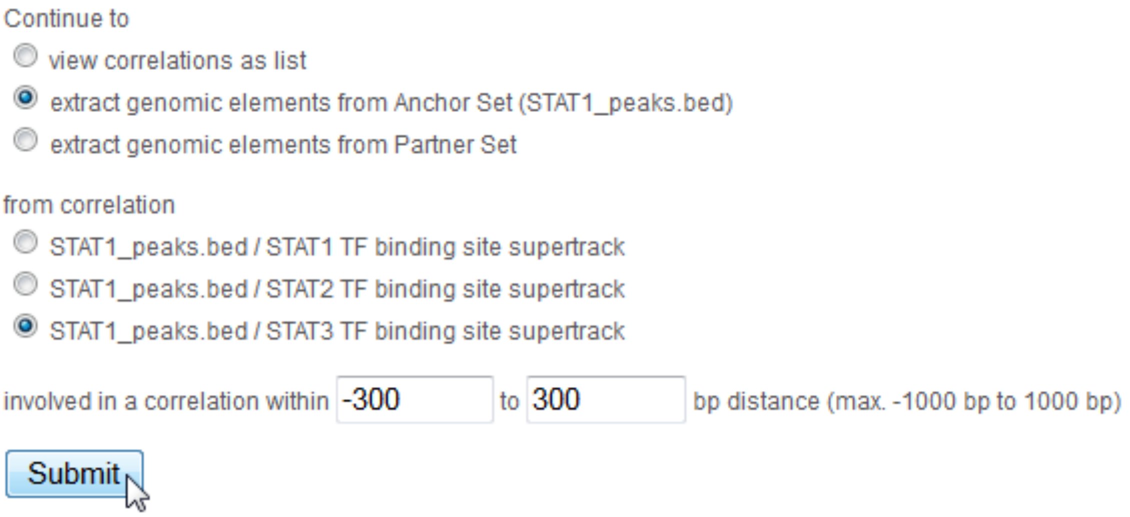
The graph shows a strong correlation of the STAT1 peaks from the HeLa cells with the ENCODE STAT1 data with the majority of correlated peaks covering a region of +/- 300 bp around the anchor point. 92.09% of the peaks in the anchor set have a positional correlation with a STAT1 partner region within the selected window of +/+ 1000 bp. The maximum coverage is >2300. The correlation with STAT3 peaks is similarly strong, reflecting the fact that STAT1 and STAT3 can bind the same sequence motifs. The STAT2 partner set is markedly smaller than the other two (3936 peaks versus 19144 (STAT1) and 67970 (STAT3)); only 14.11% of the anchor set peaks have an overlapping or neighboring STAT2 peak region within +/- 1000 bp.



	Total number of elements		Elements involved in correlation		Correlations	
Correlation	Anchor Set	Partner Set	Anchor Set	Partner Set	mean	most frequent distance
STAT1_peaks.bed vs. STAT1 TF binding site supertrack	2643 (2643 distinct)	19144	2434 (92.09%)	2536 (13.25%)	569±763	5
STAT1_peaks.bed vs. STAT2 TF binding site supertrack	2643 (2643 distinct)	3936	373 (14.11%)	357 (9.07%)	67±104	-23
STAT1_peaks.bed vs. STAT3 TF binding site supertrack	2643 (2643 distinct)	67970	2158 (81.65%)	2585 (3.80%)	465±604	-1



Correlations, as well as the peak regions from the anchor or partner set, can be retrieved based on a correlation distance range. The settings below show how to get the STAT1 peaks from the anchor set that have a correlation with at least one peak region in the STAT3 partner set in a window of +/- 300 bp around the anchor point.



Continue to

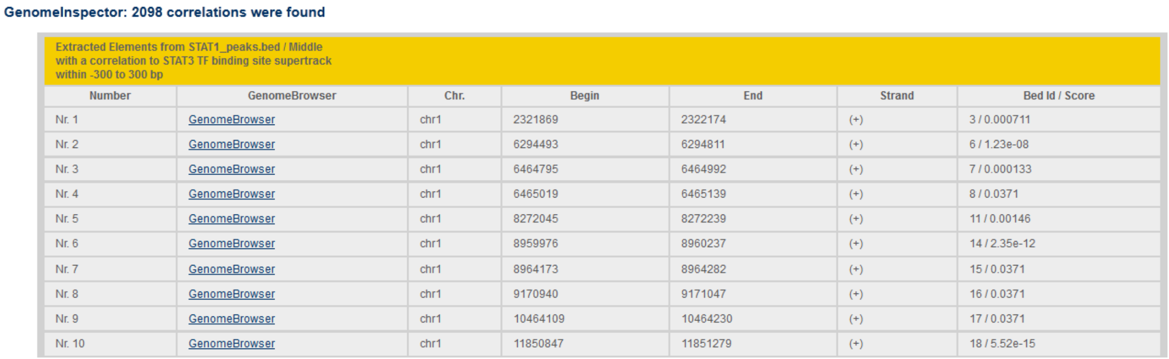
○ view correlations as list
● extract genomic elements from Anchor Set (STAT1_peaks.bed)
○ extract genomic elements from Partner Set

from correlation

○ STAT1_peaks.bed / STAT1 TF binding site supertrack
○ STAT1_peaks.bed / STAT2 TF binding site supertrack
● STAT1_peaks.bed / STAT3 TF binding site supertrack

involved in a correlation within -300 to 300 bp distance (max. -1000 bp to 1000 bp)

2098 of 2643 peak regions are found in this way; i.e. about 80% of the STAT1 peaks from the HeLa cells are overlapping or very close to STAT3 peaks in the ENCODE set.



GenomeInspector: 2098 correlations were found

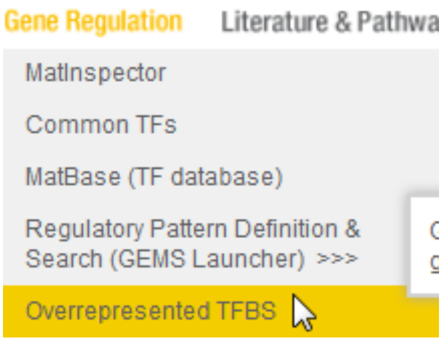
Extracted Elements from STAT1_peaks.bed / Middle
with a correlation to STAT3 TF binding site supertrack
within -300 to 300 bp

Number	GenomeBrowser	Chr.	Begin	End	Strand	Bed Id / Score
Nr. 1	GenomeBrowser	chr1	2321869	2322174	(+)	3 / 0.000711
Nr. 2	GenomeBrowser	chr1	6294493	6294811	(+)	6 / 1.23e-08
Nr. 3	GenomeBrowser	chr1	6464795	6464992	(+)	7 / 0.000133
Nr. 4	GenomeBrowser	chr1	6465019	6465139	(+)	8 / 0.0371
Nr. 5	GenomeBrowser	chr1	8272045	8272239	(+)	11 / 0.00146
Nr. 6	GenomeBrowser	chr1	8959976	8960237	(+)	14 / 2.35e-12
Nr. 7	GenomeBrowser	chr1	8964173	8964282	(+)	15 / 0.0371
Nr. 8	GenomeBrowser	chr1	9170940	9171047	(+)	16 / 0.0371
Nr. 9	GenomeBrowser	chr1	10464109	10464230	(+)	17 / 0.0371
Nr. 10	GenomeBrowser	chr1	11850847	11851279	(+)	18 / 5.52e-15

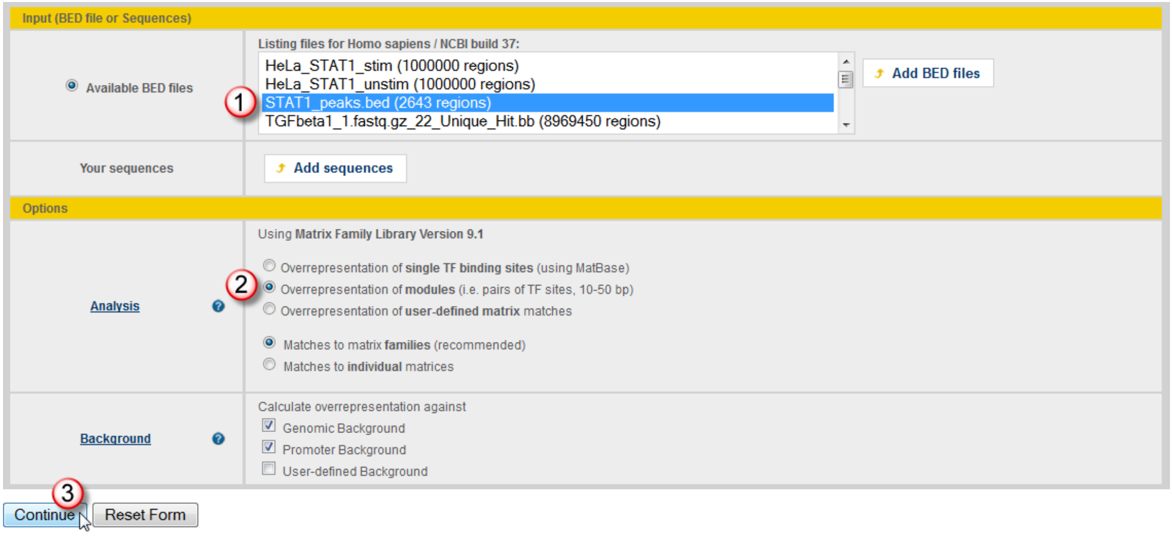
TFBS module overrepresentation

The TFBS overrepresentation analysis in the ChIP-Seq workflow considers only single binding site matches. As TFs often work in concert, it makes sense to analyze the ChIP regions for combinations of binding sites that could represent transcriptional modules, or parts thereof. Let's see if there are any combinations with other binding sites that can be found more often than others in our STAT1 peaks.

Please select "Overrepresented TFBS" from the Gene Regulation menu



On the input page, select the STAT1 peak file you saved on the ChIP-Seq workflow output in the list of previously uploaded BED files.



Input (BED file or Sequences)

● Available BED files

Listing files for Homo sapiens / NCBI build 37:

- HeLa_STAT1_stim (1000000 regions)
- HeLa_STAT1_unstim (1000000 regions)
- STAT1_peaks.bed (2643 regions)
- TGFbeta1_1.fastq.gz_22_Unique_Hit.bb (8969450 regions)

Add BED files

Your sequences — Add sequences

Options

Analysis

Using Matrix Family Library Version 9.1

○ Overrepresentation of **single TF binding sites** (using MatBase)
● Overrepresentation of **modules** (i.e. pairs of TF sites, 10-50 bp)
○ Overrepresentation of **user-defined matrix** matches

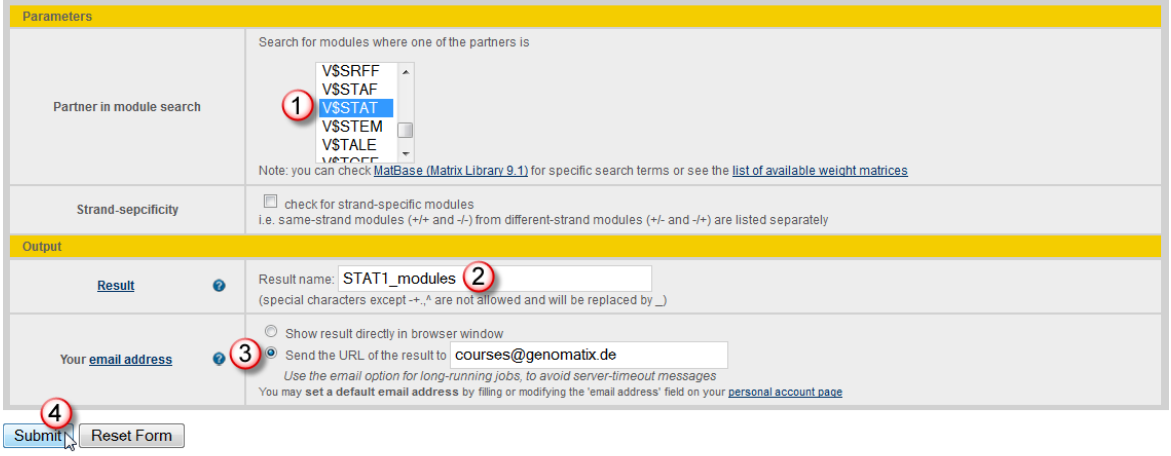
● Matches to matrix **families** (recommended)
○ Matches to **individual** matrices

Background

Calculate overrepresentation against
☑ Genomic Background
☑ Promoter Background
☐ User-defined Background

In the "options" section, click the radio button next to "Module overrepresentation (i.e. pairs of TF sites, 10-50 bp)", and continue.

On the next page, choose one TF binding site family as a partner for searching for modules. Otherwise the number of possible combinations would be too high to calculate meaningful results in appropriate time. Of course, we choose the 'V$STAT' family (containing transcription factor binding sites for STAT matrices). Provide a result name, select the e-mail option, and press the Submit button.



Parameters

Partner in module search

Search for modules where one of the partners is

1 V$STAT

Note: you can check MatBase (Matrix Library 9.1) for specific search terms or see the list of available weight matrices

Strand-sepcificity

□ check for strand-specific modules
i.e. same-strand modules (+/+ and -/-) from different-strand modules (+/- and -/+) are listed separately

Output

Result

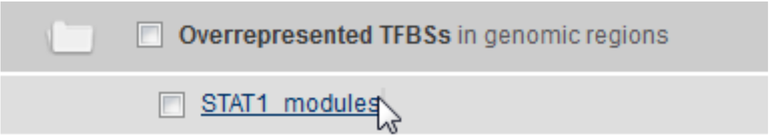
Result name: STAT1_modules 2
(special characters except -+.,^ are not allowed and will be replaced by _)

Your email address

○ Show result directly in browser window
3 ● Send the URL of the result to courses@genomatix.de
Use the email option for long-running jobs, to avoid server-timeout messages
You may **set a default email address** by filling or modifying the 'email address' field on your personal account page

4

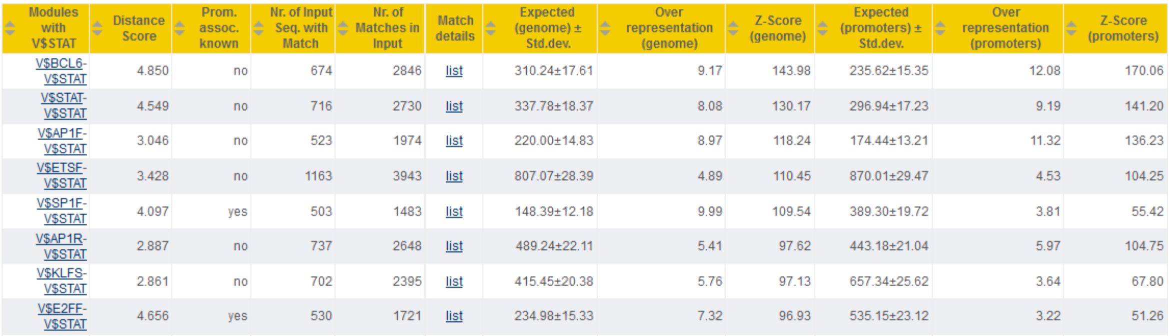
Now hit the 'Submit' button; when the result has arrived in your project management list, open it.



□ **Overrepresented TFBSs** in genomic regions

□ STAT1_modules

This is the start of the output list:

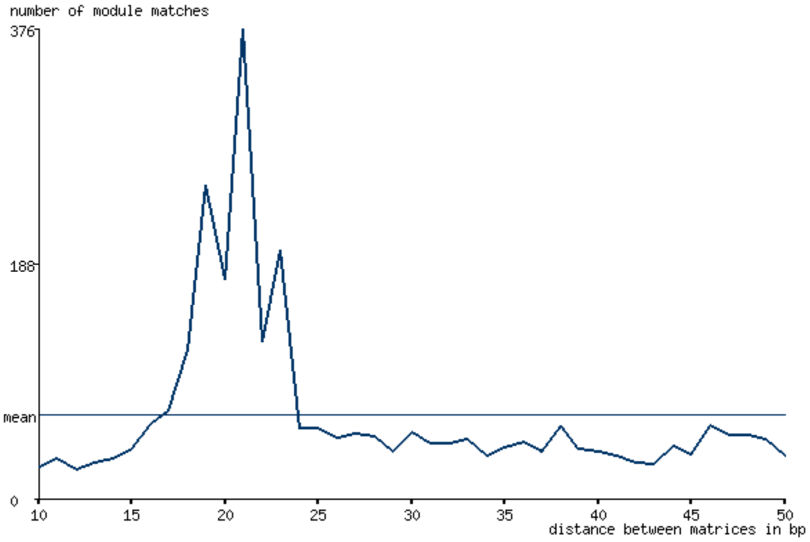


Modules with V$STAT	Distance Score	Prom. assoc. known	Nr. of Input Seq. with Match	Nr. of Matches in Input	Match details	Expected (genome) ± Std.dev.	Over representation (genome)	Z-Score (genome)	Expected (promoters) ± Std.dev.	Over representation (promoters)	Z-Score (promoters)
V$BCL6-V$STAT	4.850	no	674	2846	list	310.24±17.61	9.17	143.98	235.62±15.35	12.08	170.06
V$STAT-V$STAT	4.549	no	716	2730	list	337.78±18.37	8.08	130.17	296.94±17.23	9.19	141.20
V$AP1F-V$STAT	3.046	no	523	1974	list	220.00±14.83	8.97	118.24	174.44±13.21	11.32	136.23
V$ETSF-V$STAT	3.428	no	1163	3943	list	807.07±28.39	4.89	110.45	870.01±29.47	4.53	104.25
V$SP1F-V$STAT	4.097	yes	503	1483	list	148.39±12.18	9.99	109.54	389.30±19.72	3.81	55.42
V$AP1R-V$STAT	2.887	no	737	2648	list	489.24±22.11	5.41	97.62	443.18±21.04	5.97	104.75
V$KLFS-V$STAT	2.861	no	702	2395	list	415.45±20.38	5.76	97.13	657.34±25.62	3.64	67.80
V$E2FF-V$STAT	4.656	yes	530	1721	list	234.98±15.33	7.32	96.93	535.15±23.12	3.22	51.26

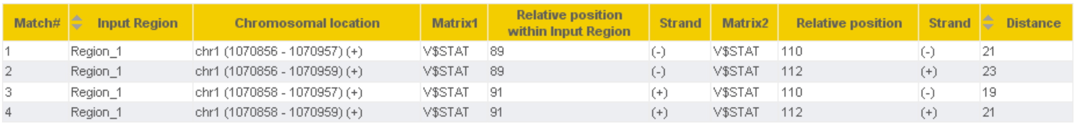
V$BCL6, V$STAT matrices themselves, V$AP1F, and V$ETSF are the most overrepresented partners of STAT sites in modules consisting of two sites with a distance of 10 to 50 bp in between.

The distance score can be used for sorting module matches with one or a few preferred distances between the sites in the input sequences. A high score would indicate a strong distance preference.

To see a profile of the distribution of distances between the binding sites in any model, please click the corresponding “list” link in the “match detail” column.
The distance profile of the pair of two STAT sites, with a distance score of 4.850, clearly shows a triple peak at 19, 21, and 23 bp over a low background. The triple peak is due to the nearly palindromic sequence of STAT sites. Because of this structure, one STAT site can give rise to two matches, one on the plus strand and one on the minus strand, with an offset of only 2 bp between them.



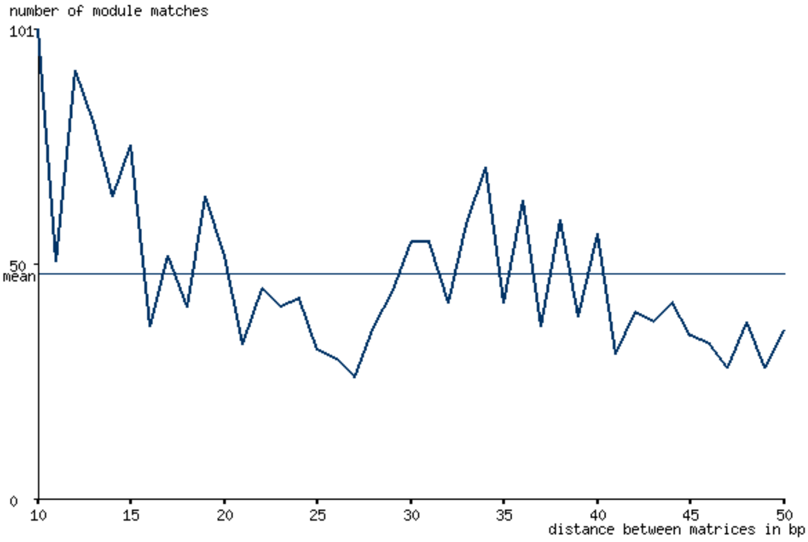
The first four entries in the match list below the distance profile exemplify this situation: two STAT sites resulting in together four matches at positions 89(-), 91(+), 110(-), and 112(+) combine to four STAT-STAT module matches with, 21, 23, 19, and 21 bp distance, respectively.



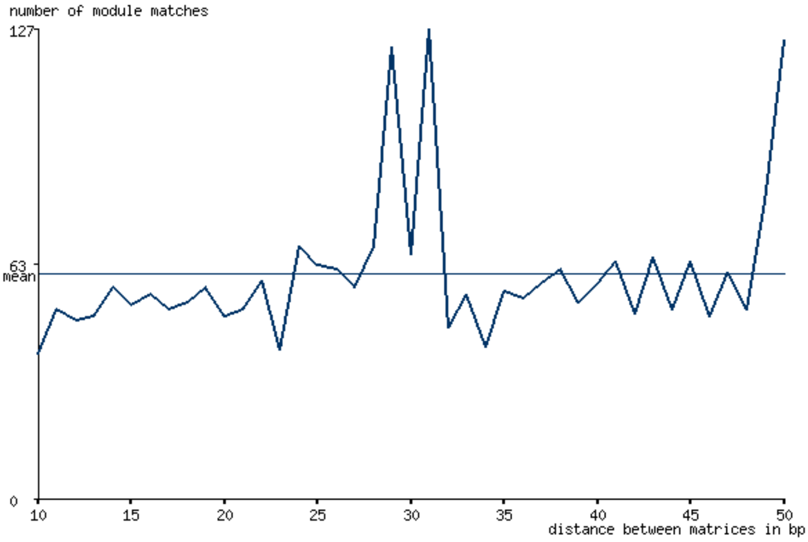
Match#	Input Region	Chromosomal location	Matrix1	Relative position within Input Region	Strand	Matrix2	Relative position	Strand	Distance
1	Region_1	chr1 (1070856 - 1070957) (+)	V$STAT	89	(-)	V$STAT	110	(-)	21
2	Region_1	chr1 (1070856 - 1070959) (+)	V$STAT	89	(-)	V$STAT	112	(+)	23
3	Region_1	chr1 (1070858 - 1070957) (+)	V$STAT	91	(+)	V$STAT	110	(-)	19
4	Region_1	chr1 (1070858 - 1070959) (+)	V$STAT	91	(+)	V$STAT	112	(+)	21

The highest (21 bp) peak results from the two possible same-strand match combinations (-/- and +/+). This distance corresponds to 2 turns of the DNA helix, suggesting a side-by-side position of the binding proteins on the DNA.

In contrast, the strongly overrepresented combination of STAT with AP1F has lower distance score (3.046), and doesn't show a clear peak:



FKHD-STAT modules show a double peak at 29/31, and another at 50. The distance between first peak in the pair to the peak at 50 is 21 bp, which is the preferred distance between two STAT sites as shown above. This suggests that FKHD sites are preferentially located near 21bp STAT site pairs in the regions identified by ChIP, thus forming a more complex pattern.

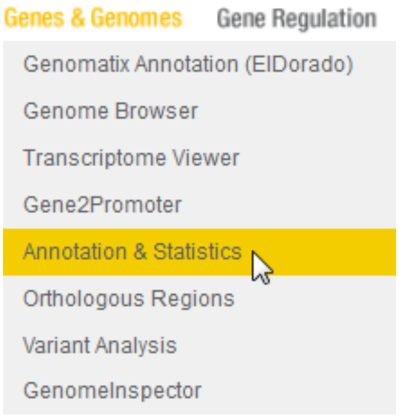


In summary, regions of STAT1 binding often show specific distance-conserved patterns of STAT sites with other TF binding sites. The fraction of matches with preferred distances can be up to 20-30% of the total matches in the regions.

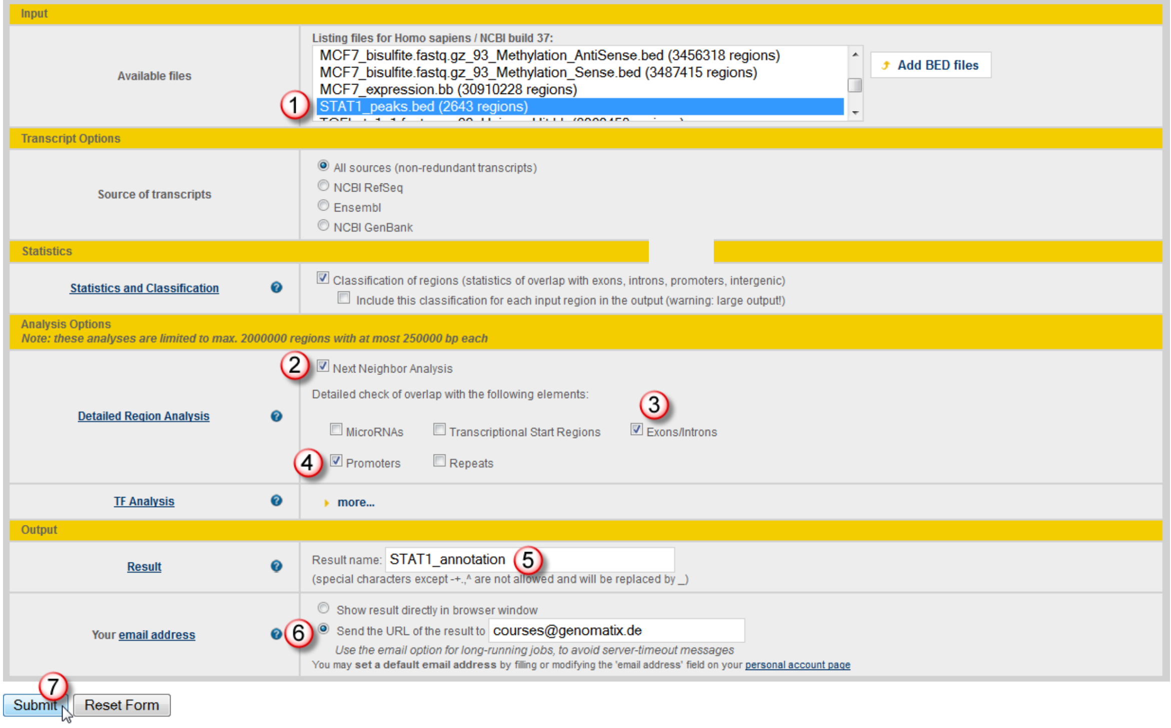
Annotation of STAT1 binding regions – target prediction

To find potential STAT1 targets, we need to look at the genomic annotation in regions where we find STAT1 binding.

The program “Annotation & Statistics” annotates your input regions for features such as promoter overlaps or neighboring loci. Please start this task from the Genes & Genomes menu in the navigation bar:



Please set the analysis parameters as below: select the BED file you saved in the GenomeInspector output from the BED file list, activate the ‘Next Neighbor Analysis’, ‘Exons/Introns’, and ‘Promoters’ checkboxes, provide a result name, make sure that you selected the e-mail option, and start the analysis. As we have more than 2000 regions to analyze in detail, the analysis will take about 10 minutes.



Input

Available files — Listing files for Homo sapiens / NCBI build 37:
MCF7_bisulfite.fastq.gz_93_Methylation_AntiSense.bed (3456318 regions)
MCF7_bisulfite.fastq.gz_93_Methylation_Sense.bed (3487415 regions)
MCF7_expression.bb (30910228 regions)
1 STAT1_peaks.bed (2643 regions)

Transcript Options

Source of transcripts:
● All sources (non-redundant transcripts)
○ NCBI RefSeq
○ Ensembl
○ NCBI GenBank

Statistics

Statistics and Classification:
☑ Classification of regions (statistics of overlap with exons, introns, promoters, intergenic)
□ Include this classification for each input region in the output (warning: large output!)

Analysis Options
Note: these analyses are limited to max. 2000000 regions with at most 250000 bp each

Detailed Region Analysis:
2 ☑ Next Neighbor Analysis
Detailed check of overlap with the following elements:
□ MicroRNAs □ Transcriptional Start Regions 3 ☑ Exons/Introns
4 ☑ Promoters □ Repeats

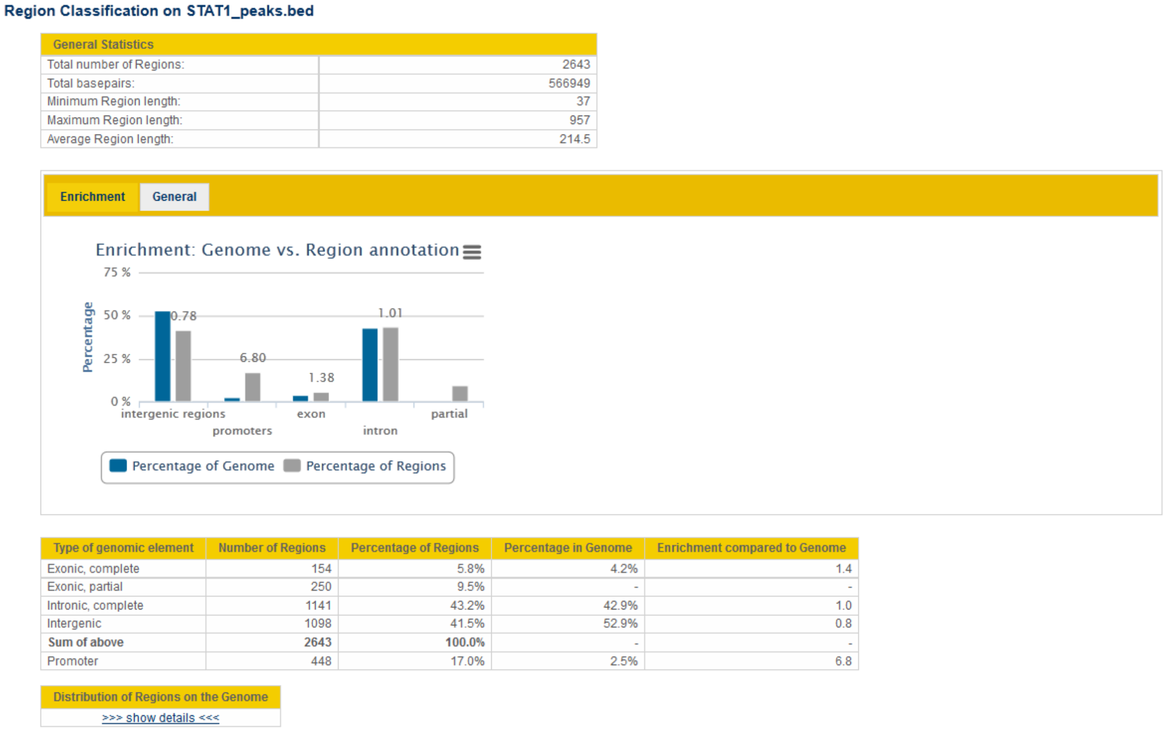
TF Analysis: more...

Output

Result — Result name: STAT1_annotation 5
(special characters except -+.,^ are not allowed and will be replaced by _)

Your email address:
○ Show result directly in browser window
6 ● Send the URL of the result to courses@genomatix.de
Use the email option for long-running jobs, to avoid server-timeout messages
You may **set a default email address** by filling or modifying the 'email address' field on your personal account page

When the analysis has completed, please open it in the project management. A classification table displays the numbers for the overlap of genome annotation with your input regions.



Region Classification on STAT1_peaks.bed

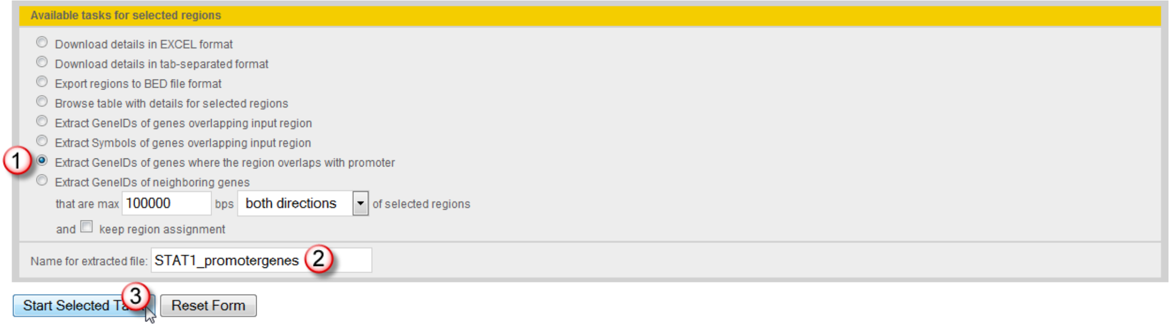
General Statistics	
Total number of Regions:	2643
Total basepairs:	566949
Minimum Region length:	37
Maximum Region length:	957
Average Region length:	214.5

Type of genomic element	Number of Regions	Percentage of Regions	Percentage in Genome	Enrichment compared to Genome
Exonic, complete	154	5.8%	4.2%	1.4
Exonic, partial	250	9.5%	-	-
Intronic, complete	1141	43.2%	42.9%	1.0
Intergenic	1098	41.5%	52.9%	0.8
Sum of above	**2643**	**100.0%**	-	-
Promoter	448	17.0%	2.5%	6.8

Distribution of Regions on the Genome

>>> show details <<<

Based on this annotation, different data sets can be generated. Please select the option ‘Extract GeneIDs of genes where the regions overlap with promoter’, and save the file with the GeneIDs on your local computer. We will use this later for further analysis.

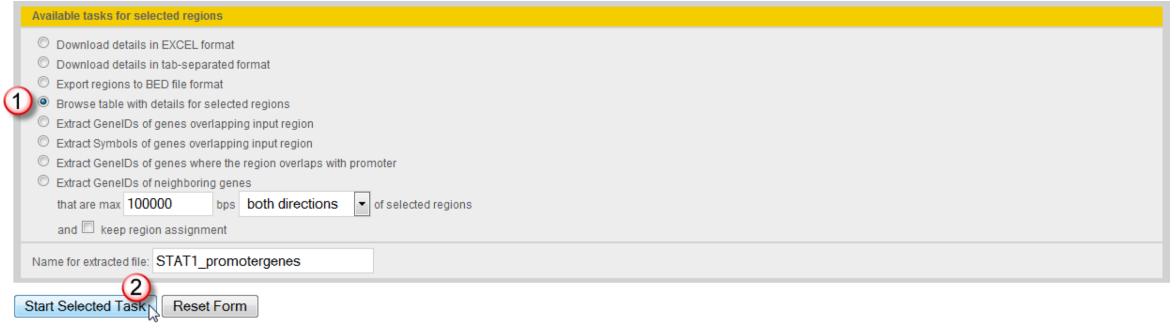


Available tasks for selected regions

○ Download details in EXCEL format
○ Download details in tab-separated format
○ Export regions to BED file format
○ Browse table with details for selected regions
○ Extract GeneIDs of genes overlapping input region
○ Extract Symbols of genes overlapping input region
● Extract GeneIDs of genes where the region overlaps with promoter
○ Extract GeneIDs of neighboring genes
that are max 100000 bps both directions of selected regions
and □ keep region assignment

Name for extracted file: STAT1_promotergenes

Back on the output page, select another option, ‘Browse table with details…’, and start.



Available tasks for selected regions

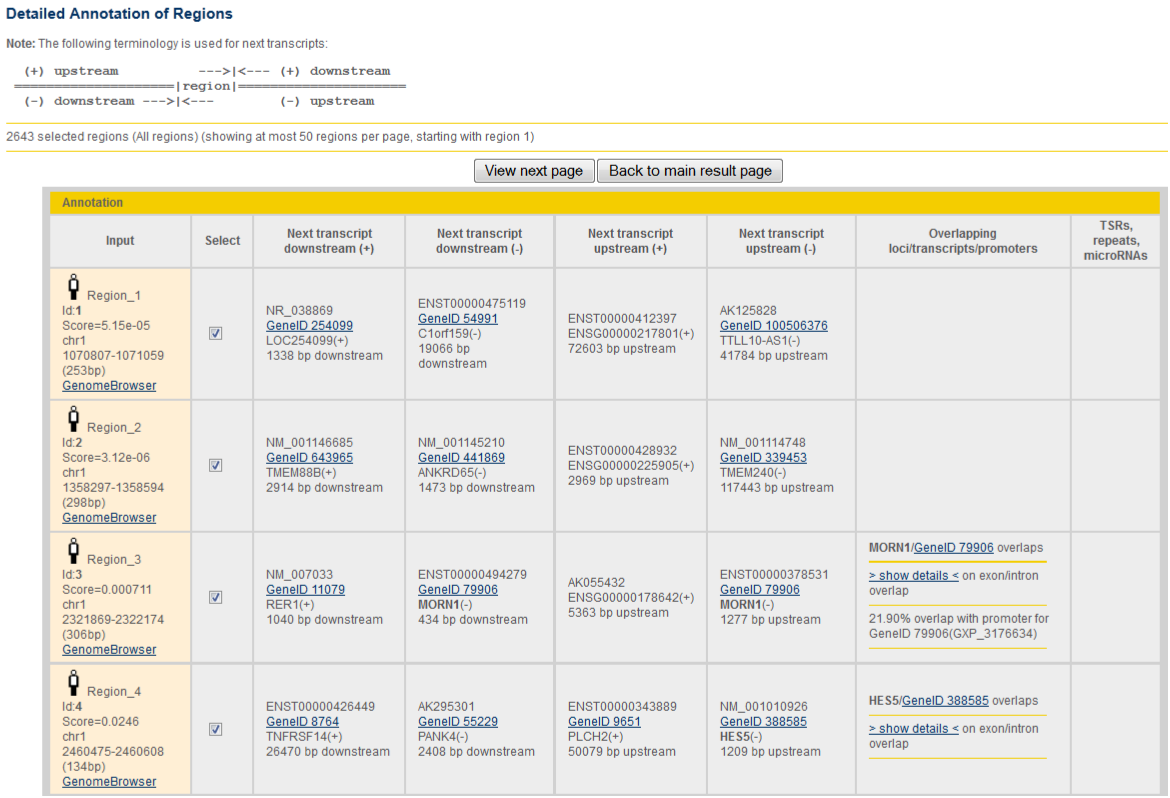
- ○ Download details in EXCEL format
- ○ Download details in tab-separated format
- ○ Export regions to BED file format
- ● Browse table with details for selected regions
- ○ Extract GeneIDs of genes overlapping input region
- ○ Extract Symbols of genes overlapping input region
- ○ Extract GeneIDs of genes where the region overlaps with promoter
- ○ Extract GeneIDs of neighboring genes

that are max 100000 bps both directions of selected regions

and □ keep region assignment

Name for extracted file: STAT1_promotergenes

The output shows the neighboring gene loci for each region, as well as overlaps with promoters, exons, and introns.



Detailed Annotation of Regions

Note: The following terminology is used for next transcripts:

```
 (+) upstream           --->|<---  (+) downstream
=====================|region|=====================
 (-) downstream --->|<---          (-) upstream
```

2643 selected regions (All regions) (showing at most 50 regions per page, starting with region 1)

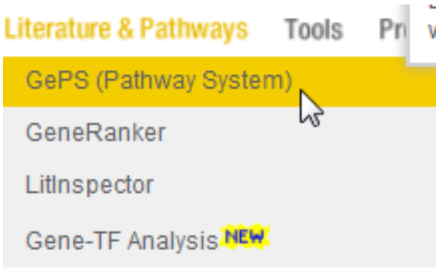
Annotation

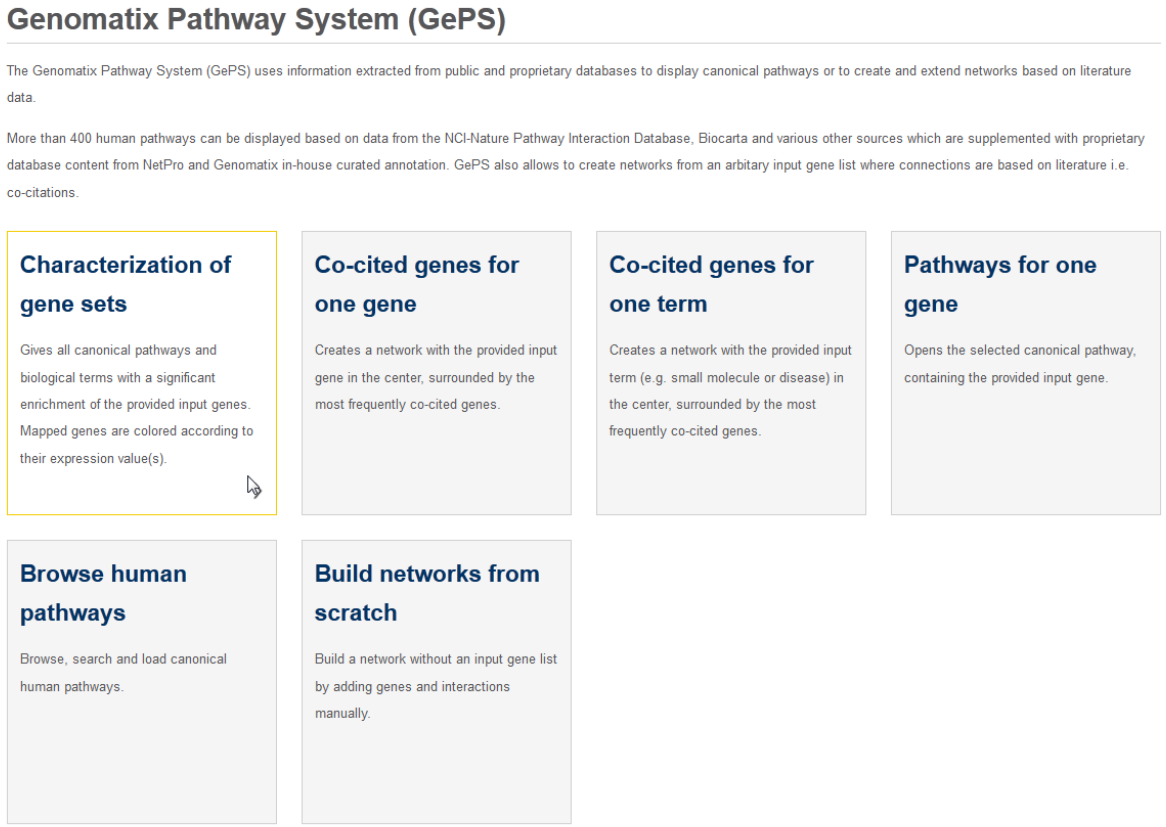
Input	Select	Next transcript downstream (+)	Next transcript downstream (-)	Next transcript upstream (+)	Next transcript upstream (-)	Overlapping loci/transcripts/promoters	TSRs, repeats, microRNAs
Region_1 Id:**1** Score=5.15e-05 chr1 1070807-1071059 (253bp) GenomeBrowser	☑	NR_038869 GeneID 254099 LOC254099(+) 1338 bp downstream	ENST00000475119 GeneID 54991 C1orf159(-) 19066 bp downstream	ENST00000412397 ENSG00000217801(+) 72603 bp upstream	AK125828 GeneID 100506376 TTLL10-AS1(-) 41784 bp upstream		
Region_2 Id:**2** Score=3.12e-06 chr1 1358297-1358594 (298bp) GenomeBrowser	☑	NM_001146685 GeneID 643965 TMEM88B(+) 2914 bp downstream	NM_001145210 GeneID 441869 ANKRD65(-) 1473 bp downstream	ENST00000428932 ENSG00000225905(+) 2969 bp upstream	NM_001114748 GeneID 339453 TMEM240(-) 117443 bp upstream		
Region_3 Id:**3** Score=0.000711 chr1 2321869-2322174 (306bp) GenomeBrowser	☑	NM_007033 GeneID 11079 RER1(+) 1040 bp downstream	ENST00000494279 GeneID 79906 **MORN1**(-) 434 bp downstream	AK055432 ENSG00000178642(+) 5363 bp upstream	ENST00000378531 GeneID 79906 **MORN1**(-) 1277 bp upstream	**MORN1**/GeneID 79906 overlaps > show details < on exon/intron overlap 21.90% overlap with promoter for GeneID 79906(GXP_3176634)	
Region_4 Id:**4** Score=0.0246 chr1 2460475-2460608 (134bp) GenomeBrowser	☑	ENST00000426449 GeneID 8764 TNFRSF14(+) 26470 bp downstream	AK295301 GeneID 55229 PANK4(-) 2408 bp downstream	ENST00000343889 GeneID 9651 PLCH2(+) 50079 bp upstream	NM_001010926 GeneID 388585 **HES5**(-) 1209 bp upstream	**HES5**/GeneID 388585 overlaps > show details < on exon/intron overlap	

Biology of potential STAT1 targets

Using the file with the GeneIDs that we saved in the previous step, we can now identify the biology represented by genes with STAT1 binding in their promoter region.

Please start the Genomatix Pathway System from the navigation bar, and start a gene set characterization.





Genomatix Pathway System (GePS)

The Genomatix Pathway System (GePS) uses information extracted from public and proprietary databases to display canonical pathways or to create and extend networks based on literature data.

More than 400 human pathways can be displayed based on data from the NCI-Nature Pathway Interaction Database, Biocarta and various other sources which are supplemented with proprietary database content from NetPro and Genomatix in-house curated annotation. GePS also allows to create networks from an arbitary input gene list where connections are based on literature i.e. co-citations.

Characterization of gene sets

Gives all canonical pathways and biological terms with a significant enrichment of the provided input genes. Mapped genes are colored according to their expression value(s).

Co-cited genes for one gene

Creates a network with the provided input gene in the center, surrounded by the most frequently co-cited genes.

Co-cited genes for one term

Creates a network with the provided input term (e.g. small molecule or disease) in the center, surrounded by the most frequently co-cited genes.

Pathways for one gene

Opens the selected canonical pathway, containing the provided input gene.

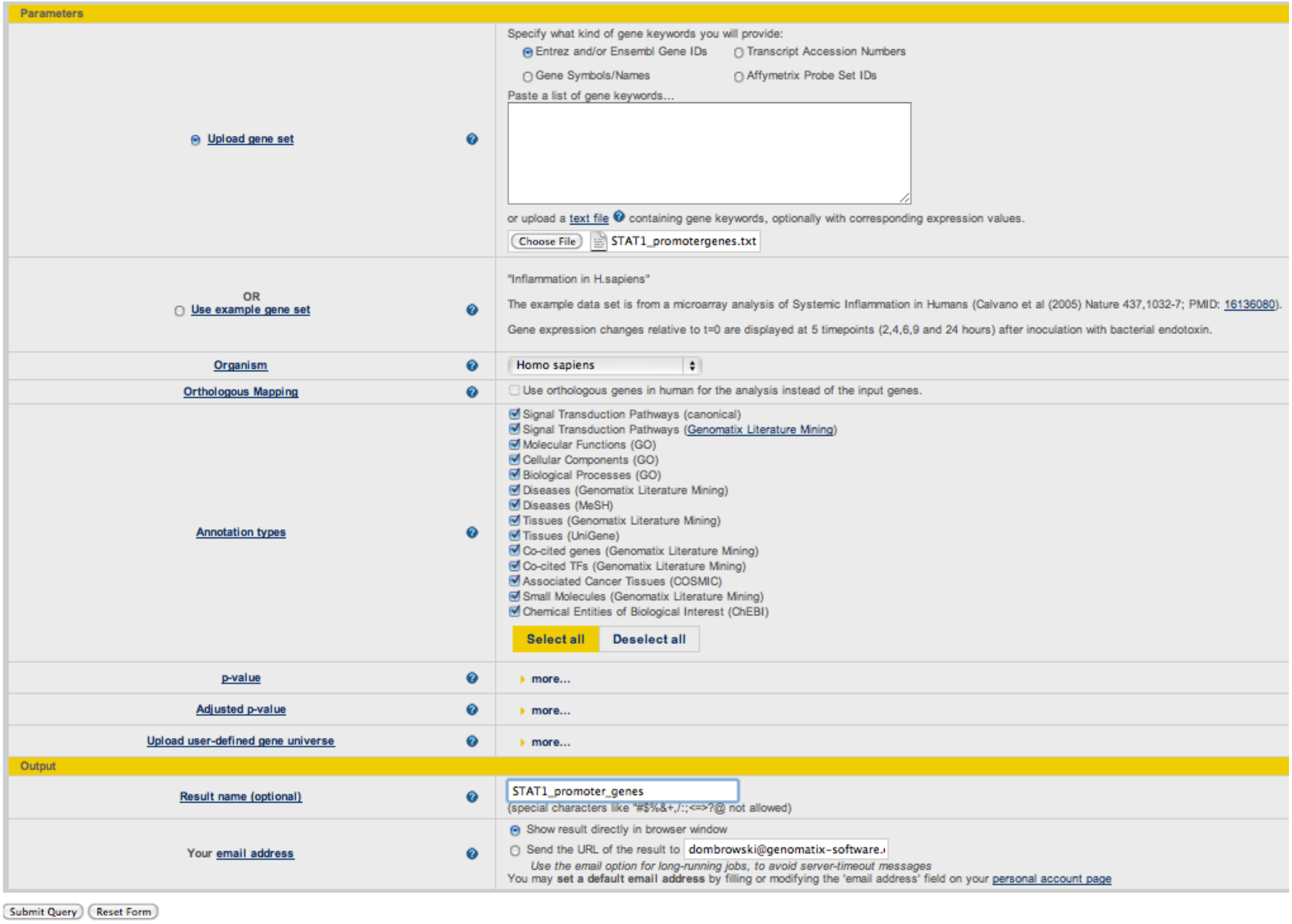
Browse human pathways

Browse, search and load canonical human pathways.

Build networks from scratch

Build a network without an input gene list by adding genes and interactions manually.

Upload the saved file with the GeneIDs of genes whose promoter overlaps with the STAT1 regions. Tick the checkboxes for all annotation types. Provide a result name and start the query.



Parameters

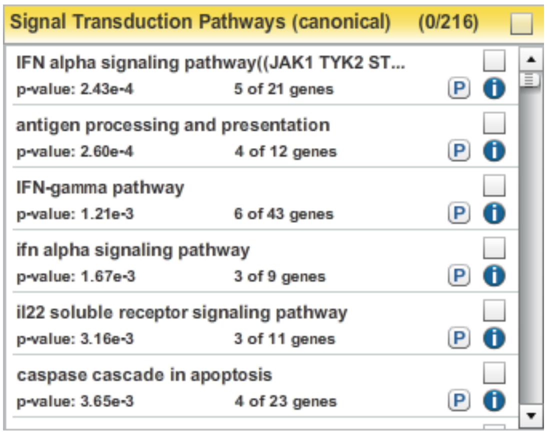
● Upload gene set	Specify what kind of gene keywords you will provide: ● Entrez and/or Ensembl Gene IDs ○ Transcript Accession Numbers ○ Gene Symbols/Names ○ Affymetrix Probe Set IDs Paste a list of gene keywords... or upload a text file containing gene keywords, optionally with corresponding expression values. Choose File STAT1_promotergenes.txt
OR ○ Use example gene set	"Inflammation in H.sapiens" The example data set is from a microarray analysis of Systemic Inflammation in Humans (Calvano et al (2005) Nature 437,1032-7; PMID: 16136080). Gene expression changes relative to t=0 are displayed at 5 timepoints (2,4,6,9 and 24 hours) after inoculation with bacterial endotoxin.
Organism	Homo sapiens
Orthologous Mapping	□ Use orthologous genes in human for the analysis instead of the input genes.
Annotation types	☑ Signal Transduction Pathways (canonical) ☑ Signal Transduction Pathways (Genomatix Literature Mining) ☑ Molecular Functions (GO) ☑ Cellular Components (GO) ☑ Biological Processes (GO) ☑ Diseases (Genomatix Literature Mining) ☑ Diseases (MeSH) ☑ Tissues (Genomatix Literature Mining) ☑ Tissues (UniGene) ☑ Co-cited genes (Genomatix Literature Mining) ☑ Co-cited TFs (Genomatix Literature Mining) ☑ Associated Cancer Tissues (COSMIC) ☑ Small Molecules (Genomatix Literature Mining) ☑ Chemical Entities of Biological Interest (ChEBI) Select all Deselect all
p-value	more...
Adjusted p-value	more...
Upload user-defined gene universe	more...

Output

Result name (optional)	STAT1_promoter_genes (special characters like '"#$%&+,/:;<=>?@ not allowed)
Your email address	● Show result directly in browser window ○ Send the URL of the result to dombrowski@genomatix-software. *Use the email option for long-running jobs, to avoid server-timeout messages* You may **set a default email address** by filling or modifying the 'email address' field on your personal account page

Submit Query Reset Form

In the overrepresented canonical pathways, we find IFN alpha and IFN gamma. The co-citation based pathway list is headed by interferon and STAT.



Signal Transduction Pathways (canonical) (0/216) □

IFN alpha signaling pathway((JAK1 TYK2 ST...
p-value: 2.43e-4 5 of 21 genes

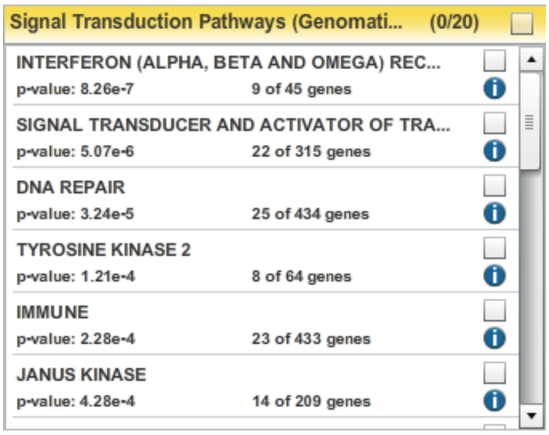
antigen processing and presentation
p-value: 2.60e-4 4 of 12 genes

IFN-gamma pathway
p-value: 1.21e-3 6 of 43 genes

ifn alpha signaling pathway
p-value: 1.67e-3 3 of 9 genes

il22 soluble receptor signaling pathway
p-value: 3.16e-3 3 of 11 genes

caspase cascade in apoptosis
p-value: 3.65e-3 4 of 23 genes



Signal Transduction Pathways (Genomati... (0/20) □

INTERFERON (ALPHA, BETA AND OMEGA) REC...
p-value: 8.26e-7 9 of 45 genes

SIGNAL TRANSDUCER AND ACTIVATOR OF TRA...
p-value: 5.07e-6 22 of 315 genes

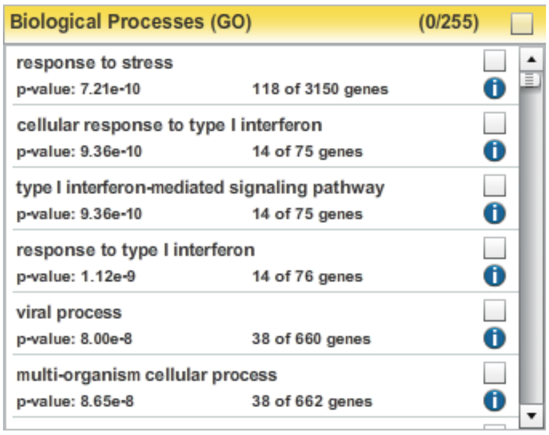
DNA REPAIR
p-value: 3.24e-5 25 of 434 genes

TYROSINE KINASE 2
p-value: 1.21e-4 8 of 64 genes

IMMUNE
p-value: 2.28e-4 23 of 433 genes

JANUS KINASE
p-value: 4.28e-4 14 of 209 genes

Among the top-ranking biological processes are interferon response and signaling. STAT and interferon regulatory factors are highly co-cited with the input genes.



Biological Processes (GO) (0/255) □

response to stress
p-value: 7.21e-10 118 of 3150 genes

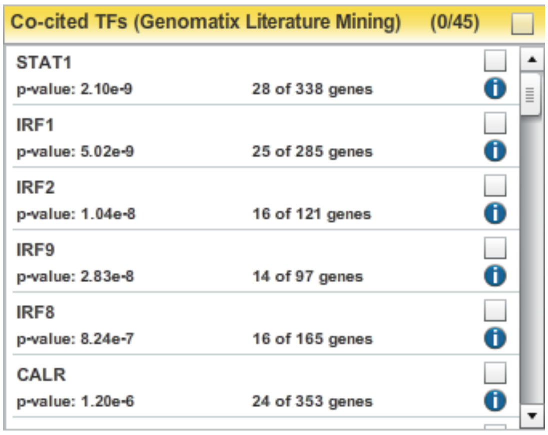
cellular response to type I interferon
p-value: 9.36e-10 14 of 75 genes

type I interferon-mediated signaling pathway
p-value: 9.36e-10 14 of 75 genes

response to type I interferon
p-value: 1.12e-9 14 of 76 genes

viral process
p-value: 8.00e-8 38 of 660 genes

multi-organism cellular process
p-value: 8.65e-8 38 of 662 genes



Co-cited TFs (Genomatix Literature Mining) (0/45) □

STAT1
p-value: 2.10e-9 28 of 338 genes

IRF1
p-value: 5.02e-9 25 of 285 genes

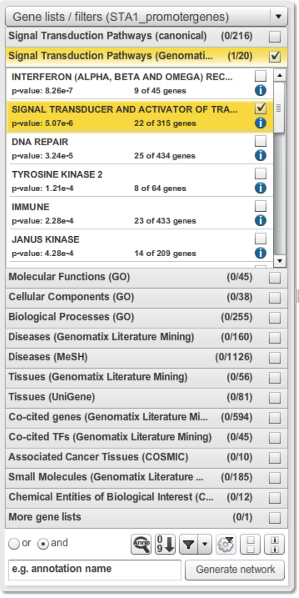
IRF2
p-value: 1.04e-8 16 of 121 genes

IRF9
p-value: 2.83e-8 14 of 97 genes

IRF8
p-value: 8.24e-7 16 of 165 genes

CALR
p-value: 1.20e-6 24 of 353 genes

The 22 genes binding STAT1 in their promoter that are co-cited with the STAT pathway include a number of transcription factors, among them STAT1 itself, which suggests a direct auto-regulatory loop. STAT1 also binds to promoters of other STAT factors (STAT2 and 3). STAT-inhibiting factors, such as SOCS3 and CISH, are also in this group. NMI interacts with STATs and augments IFN-gamma responsive transcription mediated by STATs.



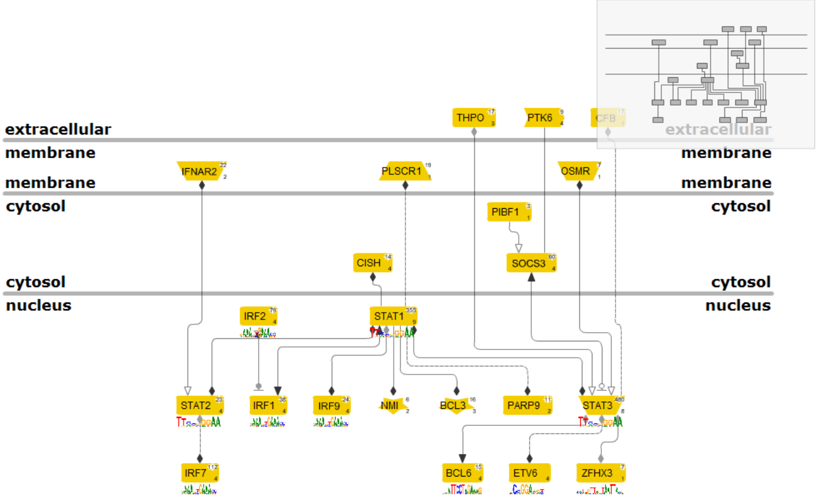
Gene lists / filters (STA1_promotergenes)

Signal Transduction Pathways (canonical)	(0/216)	□
Signal Transduction Pathways (Genomati...	(1/20)	☑
INTERFERON (ALPHA, BETA AND OMEGA) REC... p-value: 8.26e-7	9 of 45 genes	□
SIGNAL TRANSDUCER AND ACTIVATOR OF TRA... p-value: 5.07e-6	22 of 315 genes	☑
DNA REPAIR p-value: 3.24e-5	25 of 434 genes	□
TYROSINE KINASE 2 p-value: 1.21e-4	8 of 64 genes	□
IMMUNE p-value: 2.28e-4	23 of 433 genes	□
JANUS KINASE p-value: 4.28e-4	14 of 209 genes	□
Molecular Functions (GO)	(0/45)	□
Cellular Components (GO)	(0/38)	□
Biological Processes (GO)	(0/255)	□
Diseases (Genomatix Literature Mining)	(0/160)	□
Diseases (MeSH)	(0/1126)	□
Tissues (Genomatix Literature Mining)	(0/56)	□
Tissues (UniGene)	(0/81)	□
Co-cited genes (Genomatix Literature Mi...	(0/594)	□
Co-cited TFs (Genomatix Literature Mining)	(0/45)	□
Associated Cancer Tissues (COSMIC)	(0/10)	□
Small Molecules (Genomatix Literature ...	(0/185)	□
Chemical Entities of Biological Interest (C...	(0/12)	□
More gene lists	(0/1)	□

○ or ● and

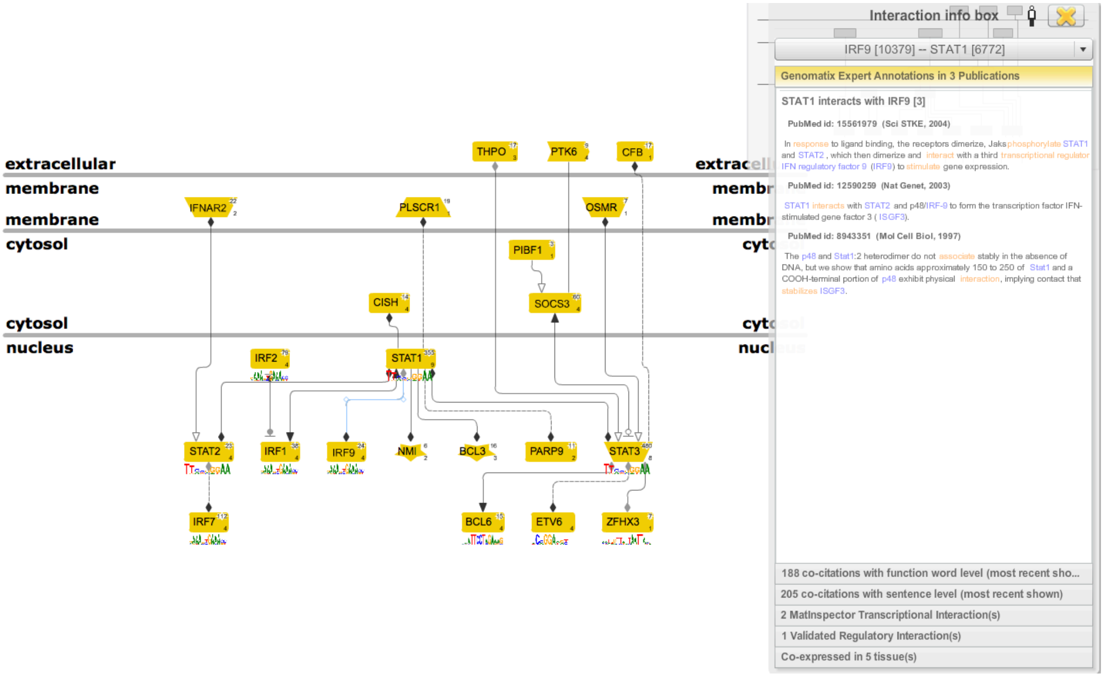
e.g. annotation name

Generate network



Based on this data set, the transcriptional repressor BCL6, which is transcriptionally-regulated by STAT3, is also a potential STAT1 target. STAT1 is known to interact with the IRF9 gene product, but obviously is also a transcriptional regulator of several IRF genes.

To view these potential regulatory interactions further, double-click on the node connecting STAT1 with IRF9. This will generate a pop-up window with more detailed information about the observed interaction.



IRF9 [10379] -- STAT1 [6772]

Genomatix Expert Annotations in 3 Publications

STAT1 interacts with IRF9 [3]

PubMed id: 15561979 (Sci STKE, 2004)

In response to ligand binding, the receptors dimerize, Jaks phosphorylate STAT1 and STAT2 , which then dimerize and interact with a third transcriptional regulator IFN regulatory factor 9 (IRF9) to stimulate gene expression.

PubMed id: 12590259 (Nat Genet, 2003)

STAT1 interacts with STAT2 and p48/IRF-9 to form the transcription factor IFN-stimulated gene factor 3 (ISGF3).

PubMed id: 8943351 (Mol Cell Biol, 1997)

The p48 and Stat1:2 heterodimer do not associate stably in the absence of DNA, but we show that amino acids approximately 150 to 250 of Stat1 and a COOH-terminal portion of p48 exhibit physical interaction, implying contact that stabilizes ISGF3.

188 co-citations with function word level (most recent sho...

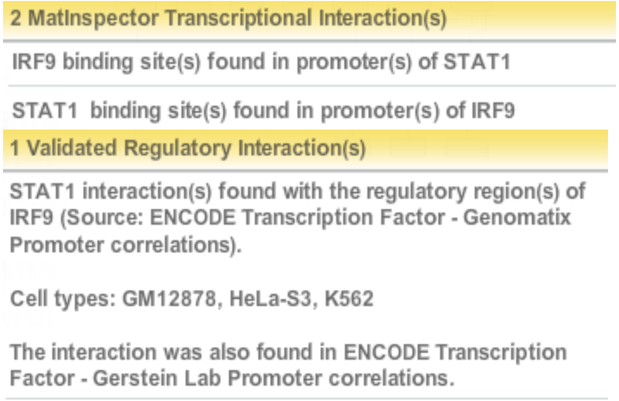
205 co-citations with sentence level (most recent shown)

2 MatInspector Transcriptional Interaction(s)

1 Validated Regulatory Interaction(s)

Co-expressed in 5 tissue(s)

The tabulated results contain both *in silico* transcription factor binding site information, as determined by MatInspector (Cartharius *et al*., 2005), as well as validated regulatory information generated from the ChIP-Seq studies that are part of the ENCODE project.



2 MatInspector Transcriptional Interaction(s)

IRF9 binding site(s) found in promoter(s) of STAT1

STAT1 binding site(s) found in promoter(s) of IRF9

1 Validated Regulatory Interaction(s)

STAT1 interaction(s) found with the regulatory region(s) of IRF9 (Source: ENCODE Transcription Factor - Genomatix Promoter correlations).

Cell types: GM12878, HeLa-S3, K562

The interaction was also found in ENCODE Transcription Factor - Gerstein Lab Promoter correlations.

Here we see that MatInspector has predicted a transcription factor binding site in the promoter of the IRF9 gene, and vice versa, suggesting a very intimate feedback loop of transcriptional control. Additionally, we learn that a validated regulatory interaction between STAT1 and IRF9 has been observed in HeLa-S3 cell line.

Literature

Audic S, Claverie JM. The significance of digital gene expression profiles. Genome Res 10, 986-995 (1997).

Cartharius, K, Frech K, Grote ., Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein, Werner T. MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics 21, 2933-42 (2005).

ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414), 57-74 (2012).

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S: Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4(8), 651-657 (2007).

Schindler C, Levy DE, Decker T: JAK-STAT signaling: from interferons to cytokines. J Biol Cem 282, 20059-20063 (2007).

Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W: A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25, 1952-1958 (2009)

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS: Model-based analysis of ChIP-Seq (MACS). Genome Biol 9(9), R137 (2008).

List of resources available on the web:

Gene Expression Omnibus:
http://www.ncbi.nlm.nih.gov/geo/

Canada's Michael Smith Genome Sciences Centre:
http://www.bcgsc.ca/

Further reading:
http://www.genomatix.de/expertise/publications.html

This tutorial was compiled for Genomatix Genome Analyzer v3.20715.

Please note that depending on the program versions and database releases used slight variations in results (*e.g.* gene numbers) may occur.