# Table of Contents

Introduction........................................................................................................................................... 3  
Introduction to the Genomatix Mining Station ................................................................................ 4  
  Shortest unique subwords.................................................................................................................. 4  
  Indexing........................................................................................................................................... 4  
  Mapping ........................................................................................................................................... 5  
Demo example: mapping NGS reads on the GMS ...................................................................... 6  
  Creating a project ............................................................................................................................. 7  
Importing sequence data to a project ............................................................................................... 9  
Looking at sequence statistics ......................................................................................................... 12  
Starting a mapping ............................................................................................................................ 13  
Mapping statistics .............................................................................................................................. 16  
Read classification ............................................................................................................................. 19  
Preview, download, and export of result files ............................................................................. 22  
Introduction to the Genomatix Genome Analyzer ....................................................................... 24  
Hands-on examples ............................................................................................................................ 25  
  ChIP-Seq workflow: STAT1 binding in IFN-γ stimulated HeLa cells ........................................ 27  
    Available peak finding algorithms............................................................................................... 27  
    Peak finding .................................................................................................................................. 34  
    Read classification ....................................................................................................................... 35  
    Peak classification .......................................................................................................................... 36  
    Sequence extraction ...................................................................................................................... 37  
    TFBS overrepresentation .............................................................................................................. 37  
    Definition of new TFBS ............................................................................................................... 39  
  Positional correlation of ChIP-Seq data sets ............................................................................... 43  
  TFBS module overrepresentation ................................................................................................. 47  
  Annotation of STAT1 binding regions – target prediction ......................................................... 51  
  Biology of potential STAT1 targets .............................................................................................. 54  
Literature ........................................................................................................................................ 59
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:
The first time you log on to the system, the interface will look like this, with an empty project panel on the left:

![Project Panel](image)

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 presetting parameters in your analyses, but you can use sequences from different organisms in any project. In order to create the project, press Submit.

<table>
<thead>
<tr>
<th>Project Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project Name:</strong> demo</td>
</tr>
<tr>
<td><strong>Project Owner:</strong></td>
</tr>
<tr>
<td><strong>Description:</strong></td>
</tr>
</tbody>
</table>
| **Permissions:** | ✓ other users may see this project  
| | ✓ other users may export result files  
| | (export results and import them via  
| | the GUI) |
| **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.
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.

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.

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

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

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

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.

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.
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 length of 27 or 36 nt, this will be equivalent to up to 2-3 point mutations per read.

Masking can be used to cut off a number of nucleotides from either end of the reads, e.g. linker sequences or low sequence quality regions, which would strongly decrease mapping efficiency. 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.

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.

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.

**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.

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.
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 (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.

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

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:

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

Please log in:

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

Login
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.

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

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.

In the upload dialog, select the GGA for the file import and press the Browse GGA button.
You will find the files `HeLa_STAT1_stim.zip` and `HeLa_STAT1_unstim.zip` in the directory `/workbench_home/Demo/NGS_Seminar/HeLa_STAT1`.

Press Submit in the upload dialog to start the import process.
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 10, 2014, 10:03:27 start saving file to project management...
Apr 16, 2014, 16:03:33 done

...working on HeLa_STAT1_unstim (12Mb)
Apr 10, 2014, 10:03:37 start saving file to project management...
Apr 10, 2014, 10:04:43 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.

[Close this window] or [add more BED files...]

In the BED file lists, choose HeLa_STAT1_stim as sample and HeLa_STAT1_unstim as control file.
Make sure “Audic-Claverie” is selected as differential analysis method. Provide a result name, and start the analysis with the default e-mail option.

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.

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

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

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

The peak sequences can be saved in the next section:

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.

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.

The list is sorted by the Z-score of the overrepresentation over the genome. The overrepresentation for VSSTAT 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 (TTTCNGGAAA) that binds STAT1 homodimers (described e.g. by Schindler et al., 2007). For details, please click the “complete CoreSearch result” link.

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:
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.

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.

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:

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

User-defined Matrices

<table>
<thead>
<tr>
<th>Family</th>
<th>Family Information</th>
<th>Matrix Name</th>
<th>Information</th>
<th>Opt.</th>
</tr>
</thead>
<tbody>
<tr>
<td>USSTAT</td>
<td>created by CoreSearch</td>
<td>USM_STAT1</td>
<td>created by CoreSearch</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Matrix USM_STAT1

- **Matrix Name:** USM_STAT1
- **Description:** created by CoreSearch
- **Family:** USM_STAT1 (created by CoreSearch)
- **References:** —
- **Statistical Basis:** 852 sequences
- **Random Expectation:** 0.77 matches per 1000 bp
- **Promoter Matches:** 0.0 % (vertebrate promoters)
- **Optimized Matrix Threshold:** 0.05
- **Length:** 15 bp

<table>
<thead>
<tr>
<th>Nucleotide Distribution Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pos. 1 2 3 4 5 6 7 8 9 10 11 12 13 14</td>
</tr>
<tr>
<td>A 207 267 67 8 23 50 7 7 108 2 18 200 829 720 318 217</td>
</tr>
<tr>
<td>C 267 151 130 11 223 797 842 106 2 47 18 33 138 217</td>
</tr>
<tr>
<td>G 249 175 79 5 9 52 4 19 837 566 12 69 203 227</td>
</tr>
<tr>
<td>T 120 269 567 636 584 10 10 50 49 3 30 203 271</td>
</tr>
<tr>
<td>IUPAC</td>
</tr>
<tr>
<td>CI</td>
</tr>
</tbody>
</table>

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.

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

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.

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

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

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.
**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.

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.

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

This is the start of the output list:

<table>
<thead>
<tr>
<th>Distance Score</th>
<th>V$BCL6</th>
<th>V$STAT</th>
<th>V$AP1F</th>
<th>V$ETSF</th>
<th>V$BCL6 in V$STAT</th>
<th>V$STAT in V$STAT</th>
<th>V$AP1F in V$ETSF</th>
<th>V$ETSF in V$ETSF</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.85</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>4.54</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3.06</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2.80</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2.68</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2.61</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>4.05</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

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.

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 GenomelInspector 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.
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.

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

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 transcript:

- ** upstream**
- ** downstream**
- ** exon**
- ** intron**

2643 selected regions (414 regions) (showing almost 50 regions per page, starting with region 1)
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.
Upload the saved file with the GenelIDs of genes whose promoter overlaps with the STAT1 regions. Tick the checkboxes for all annotation types. Provide a result name and start the query.
In the overrepresented canonical pathways, we find IFN alpha and IFN gamma. The co-citation based pathway list is headed by interferon and STAT.

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

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.
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.

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


List of resources available on the web:

Gene Expression Omnibus:  

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.

BiblioSphere, ElDorado and GEMS Launcher are registered trademarks of Genomatix Software GmbH in the USA and other countries. All other trademarks, service marks and trade names appearing in this publication are the property of their respective owners.