# Table of Contents

Introduction ........................................................................................................... 3
Introduction to Genomatix Genome Analyzer ......................................................... 4
Creating a project .................................................................................................... 6
Data background ...................................................................................................... 8
  Tbx20 transcription factor binding and effects on expression in the adult mouse heart .................................................................................................................. 8
RNA-Seq analysis .................................................................................................. 9
  Principle component analysis ................................................................................. 9
  Differential expression in Tbx-/- knockout compared to wild type adult mouse hearts .................................................................................................................. 20
    Comparative expression analysis ......................................................................... 20
    Biology of differentially expressed genes ............................................................ 28
Chip-sequencing analysis ....................................................................................... 30
  ChIP-Seq workflow: regions bound by Tbx20 in the adult mouse heart .............. 30
    Available peak finding algorithms ....................................................................... 30
    Peak finding ........................................................................................................ 35
    Read classification ................................................................................................ 36
    Peak classification ................................................................................................ 38
    Sequence extraction ............................................................................................... 38
    TFBS overrepresentation ....................................................................................... 39
    Definition of new TFBS ......................................................................................... 40
    TFBS module overrepresentation ......................................................................... 44
Integration of expression and ChIP-Seq data ......................................................... 47
  Positional correlation of Tbx20 peaks with differentially expressed transcripts 47
  Identification of direct regulatory targets based on correlation ............................. 49
In-depth transcription factor binding site analysis of correlated peaks ................. 56
  Trimming and conversion to sequence ................................................................... 58
  FrameWorker: common TFBS patterns ................................................................. 60
  ModelInspector: check for relevant biology .......................................................... 64
Annotation of Tbx20 binding regions – target prediction ........................................ 69
  Comparison of Tbx20-neighboring genes with regulated genes ............................ 74
Literature .................................................................................................................. 78
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, covering the analysis of RNA-Seq and ChIP-Seq data.
Introduction to 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.

**CNV analysis**
Pair-wise comparison of BAM files predicting copy number variations, including annotation, filter options, visualization, and links to downstream analysis tools.
Open the home page of the Genomatix Genome Analyzer in your web browser. 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
Creating a project

At the top of each page, you'll find a navigation menu bar which allows you to access the available programs. Select the Projects & Results item from the Projects & Account menu.

Press the New project button, enter a name for your project in the pop-up dialog, and click on OK.
Using the controls, set the new project as your default project.

Choose your default project:

workshop ▼ Set default project!

Current default project: MyProject

The project will be the default in the upper left hand corner project selection on the different program pages.
Data background

*Tbx20 transcription factor binding and effects on expression in the adult mouse heart*

The following examples are based on publicly available RNA-Seq and ChIP-Seq data from adult mouse heart (accession number GSE30943 on the NCBI Gene Expression Omnibus at http://www.ncbi.nlm.nih.gov/geo).

Tbx20, a transcription factor required for cardiac development, has key roles in early heart development. It has been associated with congenital heart diseases in humans, including defects in septation, chamber growth and valvulogenesis. Conditional ablation of Tbx20 in adult cardiomyocytes leads to a rapid onset and progression of heart failure, with prominent conduction and contractility phenotypes that lead to death. Tbx20 can act both as an activator and a repressor of transcription (Sakabe et al., 2012).

The available data comprise expression data from wild type and Tbx20 knockout adult mouse hearts in triplicates, as well as Tbx20 ChIP-Seq data and input DNA controls from wild type hearts. For this tutorial, sequence files were downloaded from GEO, transferred into fastq format, and mapped to the mouse genome (NCBI build 38) using the Genomatix Mining Station. The genomic positions of the uniquely mapping reads are available in bigBed (*.bb) format on the Genomatix Genome Analyzer server used during the workshop.
RNA-Sequencing analysis

**Principal component analysis**

Principal component analysis (PCA) is a statistical procedure that can be used for exploratory data analysis. PCA uses linear combinations of the original data (e.g. gene expression values) to define a new set of unrelated variables (principal components). These new variables are orthogonal to each other, avoiding redundant information.

PCA can be thought of as fitting an n-dimensional ellipsoid to the data, where each axis of the ellipsoid represents a principal component. If some axis of the ellipse is small, then the variance along that axis is also small, and by omitting that axis and its corresponding principal component from our representation of the dataset, we lose only a commensurately small amount of information.

Thus, PCA can be used to reduce the dimensions of a data set, allowing the description of data sets and their variance with a reduced number of variables. Since similarities between data sets are correlated to the distances in the projection of the space defined by the principal components, PCA can also be used to identify outliers with respect to the principal components.

It is often sufficient to look at the first two components, as these describe the largest variability.

A PCA tool can be found in the **NGS Analysis** menu in the navigation bar; please open this now.

This task can be used to get an impression of the similarity of RNA-sequencing samples, i.e. to identify subgroups or outliers.

Based on the read distribution in the input files, a normalized expression value (NE) will be calculated for each locus (or transcript) for each input file. The NE value is based on the number of reads located in the exons of the locus/transcript and is normalized to the length of the locus/transcript and the density of the data set. The resulting NE matrix is then used as input for the PCA, using the R package pcaMethods (Stacklies et al., 2007).
For this analysis, we'll need read position files in BED file format, or as bigBed, the corresponding binary format, or, alternatively, as BAM file.

Here is an example for a BED file:

```
chr1 3007329  3007356  4_112_715_245  0.962963  +
chr1 3007329  3007356  4_97_641_338  0.962963  +
chr1 3011584  3011611  4_74_929_759  1.000000  -
chr1 3014985  3015012  4_139_94_580  1.000000  +
chr1 3020759  3020786  4_99_752_96   1.000000  +
chr1 3020873  3020900  4_137_571_605  1.000000  -
chr1 3024593  3024620  4_197_207_931  0.925926  +
chr1 3025020  3025047  4_124_676_441  1.000000  +
chr1 3025020  3025047  4_54_459_727  0.925926  +
chr1 3025914  3025941  4_110_349_304  1.000000  +
chr1 3026179  3026206  4_95_762_768  0.925926  -
chr1 3038718  3038745  4_182_675_953  0.962963  -
```

The first three columns are mandatory:

Col 1: chromosome (starting with chr)
Col 2: start position of the read (counting starts from 0)
Col 3: end position of the read (start < end, represents the last nucleotide of the sequence + 1)

Additional optional information can be provided in the next columns; it is important that the order of the columns is maintained, i.e. if the file contains strand information, it must be placed in column 6, and both columns 4 and 5 cannot be empty.

Col 4: SeqId (alpha-numerical value, <=50 characters)
Col 5: Score (usually the quality score of the mapping)
Col 6: strand information
   + : plus strand
   - : minus strand
   0 : no strand information available
As you will work with mouse data, use the controls in the upper right hand corner of the input page to change the current genome selection to *Mus musculus*.

Press the *Add BED files* button to open a dialog for adding BED or bigBed files to your project.

Then select *Import from the GGA*, and press the *Browse GGA* button.
Open the directory structure until you come to the subdirectory at the path `workbench_home/Demo/NGS_Seminar/mmu_heart`. There you'll find the files with the expression data, and also the ChIP-Seq files which we will use later. For now, select the first 6 files starting with `mmu_heart_expression` by ticking the check boxes.

Press the Submit button at the bottom of the file selection dialog to close it.
In the upload dialog, press Submit.

The upload will start; when it is finished, press the Close this window button in the dialogue.

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

- mmu_heart_expression_tbx20ko_1.bb (8708085 regions)
- mmu_heart_expression_tbx20ko_2.bb (9105462 regions)
- mmu_heart_expression_tbx20ko_3.bb (8980364 regions)
- mmu_heart_expression_wt_1.bb (8028478 regions)
- mmu_heart_expression_wt_2.bb (8591698 regions)
- mmu_heart_expression_wt_3.bb (7845462 regions)

To delete, rename or protect the uploaded file(s) from automatic deletion please use the Project Management.
The uploaded files will be listed as below.

<table>
<thead>
<tr>
<th>Available files</th>
</tr>
</thead>
<tbody>
<tr>
<td>menu heart expression tub90ko 1 bb (670668 regions)</td>
</tr>
<tr>
<td>menu heart expression tub90ko 2 bb (695462 regions)</td>
</tr>
<tr>
<td>menu heart expression tub90ko 3 bb (688054 regions)</td>
</tr>
<tr>
<td>menu heart expression wt 1 bb (659160 regions)</td>
</tr>
<tr>
<td>menu heart expression wt 2 bb (745462 regions)</td>
</tr>
</tbody>
</table>

Parameters for PCA:

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Files: 0</td>
<td>Files: 0</td>
</tr>
<tr>
<td>Color: black</td>
<td>Color: green</td>
</tr>
</tbody>
</table>

Options:

- Do log transformation

Transcript/Locus:

- Locus-based expression analysis (union of exons for all loci, e.g. gene bodies)
- Transcript-based expression analysis (all transcripts separately)

Select the transcript-based analysis (for consistency with the comparative expression analysis that we'll run later) and submit the job, which will run in the background.

Rename the groups, e.g. Group 1 to *Tbx20 KO*, Group 2 to *WT*. Drag & drop the files into the corresponding group fields.
Check the *Project Management* page to see running jobs. The PCA analysis will be listed as *RUNNING* or *PENDING* (in case it's waiting for a free processor core). Please note that the list does not automatically update; if you wish to see the current state, reload the page.

When the job is finished, the result will appear in the current project under *Principal Component Analysis*. Click on the result name to display the result.

The Overview page displays the overview table and a number of analytic plots.

- **Samples**
  Number of samples submitted to analysis
- **PCs**
  Number of principal components calculated (max 10)
- **Variables**
  Number of loci or transcripts considered for analysis
- **Method**
  svd = singular value decomposition
- **R2**
  The proportion of variance explained by each PC calculated (eigenvalue)
- **R2cum**
  The cumulative proportion of the variance explained by the current and all preceding principal components.
Score plot

The score plot displays each sample in the data set with respect to the first two principal components and can therefore be used to interpret the relations among the samples. This information can be used to identify outliers.

In this data set, replicates from the WT group show high similarity with respect to the first two principal components. Replicates in the Tbx20 KO group show a greater variation, mainly due to the values for replicate 3. However, the two groups separate from each other.
Scree plot

The scree plot visualizes which principal components account for which fraction of total variance in the data. The principal components are listed by decreasing order of contribution to the total variance. The bars show the proportion of variance represented by each component (R^2) and the points shows the cumulative variance (R^{2cum}). In this case, the first component explains almost 80% of the total variance, the first three components together over 90% of it.
Loadings plot

The loadings plot is a plot of the relationship between original variables (genes) and subspace dimensions. It summarizes correlation and anti-correlation of genes/transcripts with the first two principal components.
Details for principal components

For the top principal components that are needed to account for 90% of the variance in the data (or up to a maximum of 10 PCs) the 40 transcripts/loci with the highest absolute loadings are shown in a table and a plot.

In the current example, the first 3 PCs account for >90% of the variance; below you see part of the results for the first component. Please note that a gene name can be listed several times for transcript-based analyses.
3D score plot

This score plot displays each sample in the data set with respect to the first three principal components.

Differential expression in Tbx20/- knockout compared to wild type adult mouse hearts

Comparative expression analysis

In this example, we'll carry out a differential expression analysis with the files that were subjected to a PCA in the previous step. Open the input page for Expression Analysis for RNA-Seq from the NGS Analysis menu:
Select the *tbx20ko* files to use them as the treatment group. Then tick the *Use second set...* checkbox and select the *wt* files in the second list as controls. You can choose from a number of methods and parameter settings for differential expression analysis. For this example, please leave the default settings for analyses with replicates: DESeq2 using the Wald test with parametric dispersion fitting.
Note that you have the option to run the analysis locus-based or transcript based. For this example, please take the transcript-based option. In this case, you can then choose from different transcript annotations. Please leave the latter at the default, activate the read classification, provide a result name, and run the analysis in the background, which should take about 10 minutes.

After completion, load the result from the project management page.

Different files with analysis results on transcript and gene level can be downloaded. Of 217159 annotated transcripts, 31049 are differentially expressed (17021 up-regulated, 14028 down-regulated), corresponding to 4927 genes (2729 up-regulated, 2214 down-regulated).

### Differential Expression Overview

<table>
<thead>
<tr>
<th></th>
<th>Transcripts</th>
<th>Genes (known Geneid)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number analyzed</td>
<td>217159</td>
<td>28812</td>
</tr>
<tr>
<td>Differential expression</td>
<td>31049</td>
<td>4927</td>
</tr>
<tr>
<td>Up-regulation</td>
<td>17021</td>
<td>2729</td>
</tr>
<tr>
<td>Down-regulation</td>
<td>14028</td>
<td>2214</td>
</tr>
<tr>
<td>Up- and down-regulated genes (with different transcripts)</td>
<td>-</td>
<td>16</td>
</tr>
</tbody>
</table>
Click the **download details** link for the differentially expressed transcripts, and open the file in a spreadsheet program; this will show you the list of the transcripts which are regulated according to the selected analysis method and thresholds (adjusted p-value ≤ 0.05; log2 fold change ≥ 1 or ≤ -1) including detailed information. NE (normalized expression) and RPKM (reads per thousand base pairs per million mapped reads) values are used as measures for expression. The output below is broken down into three blocks.

An unfiltered file with the same structure listing all analyzed transcripts is also available.

For detailed result lists on gene level, click on the corresponding links in the rightmost column of the differential expression overview. For example, the top of the list of down-regulated genes looks like this:
You can download a simple list of regulated genes with Gene IDs, log2 fold changes, and gene symbols.

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Log2 Fold Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>80906</td>
<td>-6.141</td>
</tr>
<tr>
<td>08052</td>
<td>-6.139</td>
</tr>
<tr>
<td>13643</td>
<td>-5.773</td>
</tr>
<tr>
<td>315476</td>
<td>-5.591</td>
</tr>
<tr>
<td>142587</td>
<td>-5.305</td>
</tr>
<tr>
<td>319942</td>
<td>-5.299</td>
</tr>
<tr>
<td>30952</td>
<td>-5.081</td>
</tr>
<tr>
<td>212402</td>
<td>-4.801</td>
</tr>
<tr>
<td>76910</td>
<td>-4.791</td>
</tr>
<tr>
<td>228564</td>
<td>-4.644</td>
</tr>
</tbody>
</table>

For later comparison with the Tbx ChIP-Seq data, we'll use the BED file with the positions of the down-regulated transcript. Please save this now to your project management. Click the Save BED file link for the down-regulated transcripts.

On the next page, provide a name for the BED file and press the Save button.

Next, please download the gene lists of the up-regulated and of the down-regulated genes to your local computer; we will use them later.
After you've saved the files, please go back to the output page. The top 5 and top 50 up- and down-regulated genes are also available on the HTML page:

**Up-Regulation:**

Genes with the highest log2(fold change) for up-regulated Transcripts in input file(s) (mmu_heart_expression_tbx20ko_1.bb, ...) compared to control file(s) (mmu_heart_expression_wt_1.bb, ...):

<table>
<thead>
<tr>
<th>Symbol</th>
<th>GenId</th>
<th>mean log2(fold change) of up-reg transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spp1</td>
<td>20750</td>
<td>7.01</td>
</tr>
<tr>
<td>Tmp1</td>
<td>21857</td>
<td>6.54</td>
</tr>
<tr>
<td>Spprta</td>
<td>20753</td>
<td>5.91</td>
</tr>
<tr>
<td>Belap3</td>
<td>12095</td>
<td>5.90</td>
</tr>
<tr>
<td>Tnc</td>
<td>21923</td>
<td>5.65</td>
</tr>
</tbody>
</table>

[>> show more genes] (top 50)

**Down-Regulation:**

Genes with the smallest log2(fold change) for down-regulated Transcripts in input file(s) (mmu_heart_expression_tbx20ko_1.bb, ...) compared to control file(s) (mmu_heart_expression_wt_1.bb, ...):

<table>
<thead>
<tr>
<th>Symbol</th>
<th>GenId</th>
<th>mean log2(fold change) of down-reg transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knplc2</td>
<td>80906</td>
<td>-6.14</td>
</tr>
<tr>
<td>Rps13</td>
<td>68052</td>
<td>-6.14</td>
</tr>
<tr>
<td>Efb3</td>
<td>13643</td>
<td>-5.77</td>
</tr>
<tr>
<td>Lrmt</td>
<td>319476</td>
<td>-5.39</td>
</tr>
<tr>
<td>Axb14</td>
<td>142687</td>
<td>-5.30</td>
</tr>
</tbody>
</table>

[>> show more genes] (top 50)

The top up- and down-regulated genes can directly be used as input for the Genomatix Pathway System from the result page (see next step).

Four different diagnostic plots can be viewed and downloaded. The first two are an MA plot and a volcano plot. Points represent transcripts, dashed lines are fold change thresholds.
Left: MA plot (log2 fold-change mean of normalized counts (y-axis) vs. mean of normalized counts (x-axis)). Red dots represent values for significantly regulated transcripts (according to the adjusted p-value, but not taking the log2 fold-change into account). Note that no transcripts with a mean below ~10 normalized counts are considered regulated.

Right: volcano plot of adjusted p-value (y-axis, inverted scale) vs. log2 fold-change mean of normalized counts (x-axis). The volcano plot shows statistical significance (p-value) and biological significance (effect size as log2 fold change) in one graph.

The next two are p-value histogram and a dispersion plot.

Left: p-value histogram showing the distribution of observed p-values in bins of 0.05. As expected for a comparison with significant differences, there is an enrichment of small p-values.

Right: dispersion plot. The dispersion quantifies the within-group variability of each transcript. Black dots: transcript-wise dispersion estimates. Red line: trend line showing the dispersions' dependence on the mean; its shape is influenced by the selected dispersion fitting method. Blue dots near the trend line: final (shrunk towards the trend line) dispersion estimates. Blue dots above main cloud: dispersion outliers, which are not shrunk towards the trend line. Values represented by blue dots are used for significance testing.
The next part shows the read classification for all input files. It also provides enrichment graphs; below are the numbers for one of the knockout samples:

<table>
<thead>
<tr>
<th>General Statistics</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of Reads</td>
<td>8709056</td>
<td></td>
</tr>
<tr>
<td>Total transcripts</td>
<td>39979352</td>
<td></td>
</tr>
<tr>
<td>Minimum Read length</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>Maximum Read length</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>Average Read length</td>
<td>38.5</td>
<td></td>
</tr>
</tbody>
</table>

**Enrichment: Genome vs. Read annotation**

- **Percentage**
  - 55.2%
  - 9.74%
  - 11.7%

**Distribution of Reads on the Genome**

<table>
<thead>
<tr>
<th>Type of genomic element</th>
<th>Number of Reads</th>
<th>Percentage of Reads</th>
<th>Percentage in Genome</th>
<th>Enrichment compared to Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon</td>
<td>7940606</td>
<td>88.4%</td>
<td>5.1%</td>
<td>35.5</td>
</tr>
<tr>
<td>Partial</td>
<td>24511</td>
<td>2.5%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Intronic</td>
<td>56081</td>
<td>6.5%</td>
<td>37.8%</td>
<td>0.2</td>
</tr>
<tr>
<td>Intergenic regions</td>
<td>22537</td>
<td>2.6%</td>
<td>56.5%</td>
<td>0.6</td>
</tr>
<tr>
<td>Sum of above</td>
<td>8709056</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Promoters</td>
<td>228235</td>
<td>26.3%</td>
<td>2.7%</td>
<td>8.7</td>
</tr>
</tbody>
</table>

The read classification results can also be shown as pie charts; the left graph shows the fractions of the different annotations in the genome; the right diagram shows the percentages of the corresponding read annotations:
Biology of differentially expressed genes

With the Genomatix Pathway System (GePS), you can generate gene networks and identify the biology that is overrepresented in a set of genes. Depending on the organism, there is a selection of biological categories, e.g. signal transduction pathway associations, GeneOntology (GO), diseases, and tissues.

From the *Differential Expression Analysis* section, run the Genomatix Pathway System for the down-regulated genes. To do this, remove the number from the field for the up-regulated genes, and change the entry for the down-regulated genes to 2300 to include all of them; then press the *Go* button.

In the output, you’ll find lists of overrepresented terms in the different categories based on the Gene ID list you uploaded.

The top enriched literature mining based pathway is PPAR alpha, which plays an important role in heart physiology.

Click on the first entry to display the corresponding literature-based gene network.

The input genes are shown with a orange (weak down-regulation) to blue (strong down-regulation) colored background. Details will be shown during the workshop.
Other overrepresented biological annotations include *mitochondrion* in the GO Cellular Components category, cardiomyopathies among the literature-mining based diseases, and heart tissue based both on literature mining and UniGene tissue annotation.
Chip-sequencing analysis

**ChIP-Seq workflow: regions bound by Tbx20 in the adult mouse heart**

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

### 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. A quantitative comparison of the clustered reads in the experimental data file to the reads in corresponding regions in the control file uses the Audic-Claverie algorithm (Audic & Claverie, 1997).

**MACS** was originally designed specifically 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 provides both versions 1.4 and 2 of the MACS implementation; the latter can also be used for broader peaks.

**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.
In the first step of the analysis, we will identify genomic regions bound by Tbx20 in wild type adult mouse heart, and run some downstream analyses on these ChIP peak regions.

For this we will use the Chip-Seq workflow, which is an automated process that includes a number of analyses: peak finding, read and peak classification, creation of a peak 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.

The raw sequence tags from the experiment have been mapped to the human genome using the GMS. You find the files once more in the folder workbench_home/Demo/NGS_Seminar/mmu_heart on the GGA.

Please open the Genomatix Genome Analyzer in your browser, and 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.

**BED File Upload**

**Current Project: "workshop"**

Select the last two files (input DNA and Tbx20 ChIP-Seq) in the folder `workbench_home/Demo/NGS_Seminar/mmu_heart`, and click on *Submit*. 
Press Submit in the upload dialog to start the import process.

When the upload has finished, press the Close this window button.
In the Available files list, choose the Tbx20 ChIPSeq data set as treatment file. Activate the option Use second set of input files... and select the input DNA data set as control file. Please leave the workflow parameters at the default values.

In this example, we'll also use the defaults of the peak evaluation and downstream analysis parameters. Please provide a result name, and start the analysis with the standard e-mail option.
You'll see a message informing you that the job has been started.

The task "Complete Workflow for ChIP-Seq Analysis" has been started!

As soon as the result/data is available on the server, a mail with a link to the output will be sent to courses@genomatix.de

You can stop this job via the project management.

When the job has finished, open your project folder and the result group "ChIP-Seq Workflow" and click on the entry to open the result.

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, 3374 peaks were found originally, of which 2698 enriched peaks remain after Audic-Claverie evaluation. 1.04% of the reads are in peaks, which is relatively low.
Please save the BED file with significantly enriched clusters to the project management.

```
Save selected BED file as Tbx20_peaks.bed

to project: workshop

Save
```

**Read classification**

The read classification in shows that enrichment in promoters is only slightly higher for the Tbx20 ChIP-Seq reads than for the input control:

```
Read Classification on mmu_heart_tbx20_chipseq.bb

<table>
<thead>
<tr>
<th>Enrichment</th>
<th>General</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enrichment: Genome vs. Read annotation</td>
<td></td>
</tr>
<tr>
<td>Percentage of Genome</td>
<td>Percentage of Reads</td>
</tr>
<tr>
<td>Intergenic regions</td>
<td>promoters</td>
</tr>
<tr>
<td>0%</td>
<td>2.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of genomic element</th>
<th>Number of Reads</th>
<th>Percentage of Reads</th>
<th>Percentage in Genome</th>
<th>Enrichment compared to Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon</td>
<td>560195</td>
<td>9.4%</td>
<td>5.7%</td>
<td>1.6</td>
</tr>
<tr>
<td>Partial</td>
<td>62348</td>
<td>1.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Intron</td>
<td>2569635</td>
<td>43.1%</td>
<td>37.8%</td>
<td>1.1</td>
</tr>
<tr>
<td>Intergenic regions</td>
<td>2771034</td>
<td>46.5%</td>
<td>66.5%</td>
<td>0.8</td>
</tr>
<tr>
<td>Sum of above</td>
<td>6563202</td>
<td>100.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Promoters</td>
<td>360331</td>
<td>6.0%</td>
<td>2.7%</td>
<td>2.2</td>
</tr>
</tbody>
</table>
```

Distribution of Reads on the Genome

```
Read Classification on mmu_heart_inputdna.bb

General Statistics

| Total number of Reads | 41091391 |
| Total basepairs       | 1476089576 |
| Minimum Read length   | 36 |
| Maximum Read length   | 36 |
| Average Read length   | 36.0 |

Enrichment

<table>
<thead>
<tr>
<th>Percentage of Genome</th>
<th>Percentage of Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>promoters</td>
<td>2.00</td>
</tr>
<tr>
<td>exon</td>
<td>1.49</td>
</tr>
<tr>
<td>intron</td>
<td>1.10</td>
</tr>
<tr>
<td>intergenic regions</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Type of genomic element | Number of Reads | Percentage of Reads | Percentage in Genome | Enrichment compared to Genome |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon</td>
<td>3409335</td>
<td>8.5%</td>
<td>5.7%</td>
<td>1.5</td>
</tr>
<tr>
<td>Partial</td>
<td>370444</td>
<td>0.9%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Intron</td>
<td>17134331</td>
<td>41.7%</td>
<td>37.0%</td>
<td>1.1</td>
</tr>
<tr>
<td>Intergenic regions</td>
<td>20088283</td>
<td>48.9%</td>
<td>56.0%</td>
<td>0.9</td>
</tr>
<tr>
<td>Sum of above</td>
<td>41091391</td>
<td>100.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Promoters</td>
<td>2206227</td>
<td>5.4%</td>
<td>2.7%</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Distribution of Reads on the Genome

>>> show details <<<
Peak classification

The enrichment in promoters is 4.56 fold for peaks, approximately double of that for reads.

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$TALE is most overrepresented, both against a genomic and a promoter background.

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. V$BRAC, the binding site family for Tbx20, ranks second after V$TALE in the overrepresentation.

The list is sorted by the Z-score of the overrepresentation over the genome. The overrepresentation for V$BRAC is about 2.3 - 2.5 fold over genome and promoter background, respectively, 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 Tbx20 cluster sequences were submitted to CoreSearch to generate a new Tbx20 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 shown. 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.
A bit more than half of the sequences used for generation of the matrix are also recognized by the existing V$BRAC 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.

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

You'll see some statistics and the nucleotide distribution including IUPAC translation and consensus index for each position, which is a measure for conservation.
The conservation profile of the binding site definition is also shown in a column chart and as a sequence logo.

- Basepairs marked red show a high information content, i.e. the matrix exhibits a high conservation (consValue > 60) at this position.
- Basepairs in capital letters denote the core sequence used by MatInspector.
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 Tbx20 peaks.

Please select “Overrepresented TFBS” from the Gene Regulation menu.

On the input page, select the Tbx20 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$BRAC’ family (containing transcription factor binding sites for Tbx20 matrices). Provide a result name, and press the Submit button.

This is the start of the output list:

<table>
<thead>
<tr>
<th>Module Description</th>
<th>Distance Score</th>
<th>Number of Modules</th>
<th>Number of Matches</th>
<th>Matches Score</th>
<th>Expected Modules</th>
<th>Expected Matches</th>
<th>Z Score</th>
<th>Expected Modules</th>
<th>Expected Matches</th>
<th>Z Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>V$TALE</td>
<td>2.181</td>
<td>no</td>
<td>490</td>
<td>879</td>
<td>211</td>
<td>55.81</td>
<td>4.71</td>
<td>51.99</td>
<td>23.27</td>
<td>5.63</td>
</tr>
<tr>
<td>V$MYOD</td>
<td>2.082</td>
<td>no</td>
<td>155</td>
<td>483</td>
<td>280</td>
<td>45.12</td>
<td>4.1</td>
<td>41.97</td>
<td>23.27</td>
<td>4.1</td>
</tr>
<tr>
<td>V$AP1R</td>
<td>2.024</td>
<td>no</td>
<td>472</td>
<td>908</td>
<td>240</td>
<td>45.12</td>
<td>3.64</td>
<td>41.97</td>
<td>23.27</td>
<td>3.64</td>
</tr>
<tr>
<td>V$NF1F</td>
<td>4.047</td>
<td>no</td>
<td>290</td>
<td>409</td>
<td>64</td>
<td>40.6</td>
<td>5.42</td>
<td>40.6</td>
<td>23.27</td>
<td>5.42</td>
</tr>
<tr>
<td>V$TALE</td>
<td>2.142</td>
<td>no</td>
<td>142</td>
<td>235</td>
<td>211</td>
<td>45.12</td>
<td>7.40</td>
<td>35.97</td>
<td>41.21</td>
<td>7.40</td>
</tr>
<tr>
<td>V$MYOD</td>
<td>1.090</td>
<td>no</td>
<td>115</td>
<td>216</td>
<td>280</td>
<td>45.12</td>
<td>7.85</td>
<td>45.12</td>
<td>44.60</td>
<td>7.85</td>
</tr>
<tr>
<td>V$AP1R</td>
<td>3.152</td>
<td>yes</td>
<td>285</td>
<td>306</td>
<td>55</td>
<td>33.41</td>
<td>5.43</td>
<td>33.41</td>
<td>21.27</td>
<td>5.43</td>
</tr>
<tr>
<td>V$NF1F</td>
<td>2.728</td>
<td>no</td>
<td>172</td>
<td>319</td>
<td>322</td>
<td>32.38</td>
<td>5.10</td>
<td>32.38</td>
<td>20.69</td>
<td>5.10</td>
</tr>
<tr>
<td>V$TALE</td>
<td>1.085</td>
<td>no</td>
<td>282</td>
<td>442</td>
<td>103</td>
<td>45.12</td>
<td>3.04</td>
<td>40.44</td>
<td>27.14</td>
<td>3.04</td>
</tr>
<tr>
<td>V$AP1R</td>
<td>2.261</td>
<td>no</td>
<td>335</td>
<td>391</td>
<td>89</td>
<td>29.15</td>
<td>3.90</td>
<td>29.15</td>
<td>16.95</td>
<td>3.90</td>
</tr>
<tr>
<td>V$NF1F</td>
<td>1.080</td>
<td>no</td>
<td>410</td>
<td>679</td>
<td>223</td>
<td>20.16</td>
<td>2.91</td>
<td>20.16</td>
<td>10.86</td>
<td>2.91</td>
</tr>
<tr>
<td>V$TALE</td>
<td>3.510</td>
<td>no</td>
<td>283</td>
<td>453</td>
<td>155</td>
<td>20.16</td>
<td>3.62</td>
<td>20.16</td>
<td>9.72</td>
<td>3.62</td>
</tr>
<tr>
<td>V$MYOD</td>
<td>2.193</td>
<td>no</td>
<td>255</td>
<td>446</td>
<td>108</td>
<td>20.16</td>
<td>3.52</td>
<td>20.16</td>
<td>4.54</td>
<td>3.52</td>
</tr>
</tbody>
</table>

V$TALE, V$MYOD, V$AP1R, and V$NF1F are the most overrepresented partners of Tbx20 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 BRAC-NF1F combinations, with a distance score of 4.847, clearly shows a peak at 183 bp over a low background.

In contrast, the top overrepresented combination of BRAC with TALE has lower distance score (2.161), and doesn’t show a clear peak:

In summary, regions of Tbx20 binding sometimes show specific distance-conserved patterns of BRAC sites with other TF binding sites. The fraction of matches with preferred distances can be up to 20% of the total matches in the regions.
Integration of expression and ChIP-Seq data

Positional correlation of Tbx20 peaks with differentially expressed transcripts

In the next step, we will predict which genes that are differentially expressed in Tbx20 knock-out mouse hearts are direct targets of Tbx20. For this, we will use the program Genomelnspector and visualize the positional correlation of the starts of the down-regulated transcripts with the Tbx20 ChIP peaks.

Genomelnspector uses one BED file as anchor set and, based on the genomic positions of the regions in the file, draws a correlation graph for up to 6 other 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 Genomelnspector from the Gene & Genomes menu.

Select the BED file of down-regulated transcripts from your files in the anchor set list.
Select the Tbx20 peaks as partner set.

Set the range to the surrounding 20000 bps; in this way, also more distal regulatory regions will be included. Make sure the anchor position is at the start of the anchor set (i.e. the transcript starts), provide a result name, and start the analysis.
The graph shows a narrow peak around the transcript start sites, representing the region with the highest density of Tbx20 peaks. There is also a slightly elevated plateau ranging from about 6 kbp upstream to 11 kbp downstream of the TSS.

**Identification of direct regulatory targets based on correlation**

Next, we will identify the potential Tbx20 target genes whose down-regulated transcripts have a positional correlation with a Tbx20 peak in the range defined by the -6kbp/+11kbp plateau above. Correlations between the elements in the two data sets, as well as regions from the anchor (down-regulated transcripts) and partner (Tbx20 peaks) set, can be extracted based on the distances.
To retrieve the list of correlated transcripts and genes for the Tbx20 peaks in the -6kbp/+11kbp plateau, select the extraction of elements from the anchor set (the down-regulated transcripts), enter the range, and click on Submit.

1. view correlations as list
2. extract genomic elements from Anchor Set (Tbx20_ko_expression_transcripts_down.bed)
3. extract genomic elements from Partner Set
4. Submit

All list of correlations will be shown, including distances and gene names (only the first 100 entries).

Genominspector: 635 correlations were found

Press the EXCEL file download button at the end of the list, and open the file in Excel.

Note: 635 correlations were found. The list is too long to be displayed. Only the first 100 matches are listed, the complete list can be downloaded.

Download regions as BED file  Save BED file to project management  Extract table as EXCEL file  tab-separated file
The **Bed Id** column for the anchor set contains the internal transcript identifiers (GXT-...), the transcript accession numbers, and the corresponding gene symbols. Note that you may need to adjust the column width to see the complete content.

Use Excel functions to write the contents into separate columns: add two empty columns left of the Score column; then separate the text in the **Bed Id** column into different columns, using the slash (/) as separator.
Open the Genomatin Pathway System from the navigation bar, and start a gene set characterization.
Copy the transcript accession numbers from the Excel list to the gene keyword input field (duplicates will be removed by the system). Select Transcript Accession Numbers as the keyword type. Select Mus musculus as organism, and start the search.

Some accession numbers will not be mapped to a gene ID; ignore the warning the program gives you, and proceed with the analysis.
A total of 190 genes are found in this way. Overrepresented terms include ion binding in GO: Molecular Function, cardiac muscle contraction in GO: Biological Processes, and cardiomyopathies in Diseases. The pathway graphs below use the hierarchical layout, which you can activate with the leftmost Layout button in the lower control bar:

**Ion binding gene network:**

**Cardiac muscle contraction gene network:**

**Cardiomyopathies gene network:**
Many of the ion binding genes are also associated with cardiomyopathies, as can be seen by selecting the ion binding network, and then ticking the checkbox for the cardiomyopathies associated genes:

Only ion binding network genes fulfilling both criteria are shown with a colored background.
In-depth transcription factor binding site analysis of correlated peaks

The next analysis step will take a closer look at the peak regions which form the correlation plateau in the GenomeInspector graph. You will retrieve a BED file of the correlated peaks, prepare it in the BED file toolbox for downstream analysis, and find common transcription factor binding site patterns that include Tbx20 binding sites, which will then be assessed further.

Go back to the GenomeInspector output or open the result from the project management, and select the extraction of elements from the partner set (the Tbx20 peaks), again setting the distance range to -6kbp/+11kbp.

Continue to
- view correlations as list
- extract genomic elements from Anchor Set (Tbx20_ko_expression_transcripts_down.bed)
- extract genomic elements from Partner Set

from correlation
- Tbx20_ko_expression_transcripts_down.bed / Tbx20_peaks.bed

involved in a correlation within **-6000** to **11000** bp distance (max. -20000 bp to 20000 bp)

Submit
208 correlated peaks are found.

*Genomelnspector: 208 correlations were found*

Scroll down to the end of the list and save the regions as BED file in your project management.

Note: 208 correlations were found. The list is too long to be displayed. Only the first 100 matches are listed, the complete list can be downloaded.
Trimming and conversion to sequence

For downstream analysis, the peak data need to be modified in two ways: one, some of the regions are too long to be accepted as input for the FrameWorker program, which we will use for detection of common transcription factor binding site patterns in the peaks. Therefore, we will give the peaks a uniform length. Secondly, FrameWorker needs sequences as input; therefore we will generate a sequence file from the modified BED file.

Open the BED file tools in the navigation bar.

Select the file with the correlated peaks in the list. Then select the option *Trim regions to the same size*, set the size to 500bp, and select the option *symmetrical around the center of the regions*; then press *Submit*.
Save the file to the project management.

First few lines of the result file:

#BED file created with Genomatix BED file toolbox
#extension/trimming of input regions to 500 bp
#Eldorado: E30R1410
#TaxonID: 10090

chr1  4412410  4412910  1  9.12e-13  +
chr1  24010788  24011288  12  1.01e-06  +
chr1  52844732  52845322  40  1.66e-07  +
chr1  75393171  75393671  70  1.31e-10  +
chr1  75549201  75549701  71  1.47e-14  +
chr1  79775828  79776328  77  2.95e-05  +

Download BED file of trimmed regions (12Kb)
Save BED file to project management

Back to BED File Toolbox

Save selected BED file as Tbx20_peaks_correlated_trimmed.bed

to project workshop

Save

Open the BED file tools once more to convert the trimmed file to a sequence file. Select the trimmed file in the list, and activate the Convert BED file to DNA sequence file function. Start the conversion, and save the result in your project management.
FrameWorker: common TFBS patterns

The next step in the analysis will employ FrameWorker, which searches for common patterns of TF binding sites in a set of input sequences – here, the Tbx20 peaks which are correlated with genes that are down-regulated in Tbx20 knock-out mouse hearts.

You'll find the program in the navigation bar under Gene Regulation - Regulatory Pattern Definition & Search - FrameWorker:
Select the saved sequence file in the list and continue by clicking the *Load Sequence* button.

The next step lets you choose the elements the frameworks can be made up of. For this analysis, please leave the settings at the defaults.

Ignore the warning on the next page, and press *Continue*.

*WARNING:* No pairwise similarity check was performed, because of too many sequences!
In the next step, parameters defining the stringency of the pattern search are set. Specifically the quorum constraint, and sometimes also the distance constraints, are usually changed in an iterative process, checking the result and adapting the stringency so that a handful of patterns of the desired complexity are found, which are then further evaluated.

For this example, please set the parameters as follows:
- Quorum constraint = 7 of 208
- Distance constraints: maximum distance variance = 20
- Element constraints: V$BRAC (the binding site for Tbx20) mandatory element.

Then start the analysis.

<table>
<thead>
<tr>
<th>Frame/Worker Parameters</th>
<th>Minimum number of input sequences to contain a framework</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quorum constraint</strong></td>
<td>7 of 208 (3 %)</td>
</tr>
<tr>
<td></td>
<td>- Region_1</td>
</tr>
<tr>
<td></td>
<td>- Region_2</td>
</tr>
<tr>
<td></td>
<td>- Region_3</td>
</tr>
<tr>
<td></td>
<td>- Region_4</td>
</tr>
<tr>
<td></td>
<td>- Region_5</td>
</tr>
<tr>
<td><strong>Distance constraints</strong></td>
<td>Maximum distance variance between two elements: 20 (max. 100)</td>
</tr>
<tr>
<td></td>
<td>Distance between two elements: min. 5, max. 200 (max. 500)</td>
</tr>
<tr>
<td><strong>Element constraints</strong></td>
<td>V$BRAC (the binding site for Tbx20) mandatory element.</td>
</tr>
</tbody>
</table>
One model with 4 elements is found. Click the link to jump to the description.

The model consists of two SORY sites, a TALE site, and the mandatory BRAC site. You can click on the links for each binding site symbol to find more information about it. The V$SORY binding site family binds, among others, the Sox6 protein, which has a role in cardiomyocyte differentiation. V$TALE can bind Meis1, which is a regulator of the cardiomyocyte cell cycle.

From the associated biology, the pattern could be of interest. Please tick its checkbox, marking it for saving, give it a name, e.g. SORY_SORY_TALE_BRAC, and press the Save selected models button.
ModellInspector: check for relevant biology

For further evaluation of the model, we will run a ModellInspector search, and try to find patterns matches in a mouse promoter database. The output will include an overrepresentation analysis for GO terms in the categories biological process, molecular function, and cellular component, which allows us to assess whether the binding site pattern is associated with relevant biology.

ModellInspector uses model definitions as are generated by FrameWorker to scan DNA sequences for matches. A model is defined as a set of various individual elements (here: transcription factor binding sites), their strand orientation, their sequential order, and their distance ranges.

Click on the ModellInspector link on the notification page you see after the model has been saved.

We will scan all mouse promoters of annotated genes. Click the more... option in the Database Input section to display the available parameters, then check Mouse Promoters in the section Promoters of annotated genes and proceed with Load Sequence.
On the following parameter screen, select *User defined models* and *continue with subset selection* and continue.

Select the newly saved model in the list and start.

**Please select a number of models for ModelInspector to check your sequence:**

- [ ] Check all Models
- [ ] Uncheck all Models
- [ ] Invert Selection

- [ ] defined: SORY_SORY_TALES_BRAC
- [ ] Save this model subset with the following name

- [ ] Start Task

Please make sure you selected at least one checkbox!

The analysis will run in the background; repeatedly check in your project management if the job is still running:

<table>
<thead>
<tr>
<th>Job ID</th>
<th>Task</th>
<th>State</th>
<th>Submitted at</th>
<th>Remove job</th>
</tr>
</thead>
</table>

When it is done, click the link in the results directory to open the result page.
There are 376 matches in the promoter database. Click on Evaluation in the table header to display the GO statistics for the matching genes.

One of the overrepresented terms in the Molecular Functions category is ion binding, with a model match in the promoters of 105 genes. The same term was also overrepresented in the list of genes that were downregulated in Tbx20 knockout hearts and in the subset of down-regulated transcripts with a correlated Tbx20 ChIP-Seq peak in a +/- 10kb window around the TSS. This last analysis shows that the model finds the term also in all promoters, independent of expression or ChIP-Seq analysis results.
In order to find more associated biology for the genes with a model match in the promoter, we'll use the program GeneRanker. Go back to the ModelInspector match list output, and scroll down to the end of the page. Here, press the 'Extract GeneIDs' button to open a page showing the GeneIDs.

Select the Gene IDs on this page and copy them to the clipboard.

Open GeneRanker from the Literature & Pathways menu in the navigation bar.
Paste the Gene IDs from the clipboard into the keyword field, select the mouse as organism, and start the analysis.

The result shows the cardiomyopathy hypertrophy, left ventricular in the top ten overrepresented MeSH Disease terms, and several heart associated terms in the category Tissues (Genomatix Literature Mining).
Annotation of Tbx20 binding regions – target prediction

An alternative way for finding potential regulatory targets of a transcription factor based on ChIP-Seq peaks, which can also be applied in the absence of expression data, is to analyze the genomic annotation in the vicinity of the TF peak positions and look for overlapping and neighboring promoters and gene loci.

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 with the Tbx20 peaks from the BED file list, and activate the Next Neighbor Analysis, Exons/Introns, and Promoters checkboxes. This is necessary for identification of neighboring and overlapping promoters and loci. To include the information which peaks have a match for a Tbx20 binding site, click on the TF analysis more... option, tick the TFBS search checkbox, and select V$BRAC from the binding site list. 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.

Overlap details can be viewed in the Overlap Statistics section.
Based on this annotation, different data sets can be generated. Select *Detailed Annotation and Download*.

Select the option *Browse table with details*..., and start the task.

The output shows the neighboring gene loci for each region in both directions and on both strands, as well as overlaps with promoters, exons, and introns, and the number of V$BRAC binding site matches in each peak.
Next, please go back to the overview page, and select the option *Extract GenelIDs of neighboring genes*. For this example, set the maximum distance to 10,000 bp. To include the identifiers of the corresponding peaks, activate the *keep region assignment* option. Provide a file name, and save the file with the GenelIDs on your local computer.

The file contains the GenelIDs and associated peak identifiers based on the peak IDs in the BED file.

```plaintext
11267 Region_1019
11304 Region_571
11426 Region_699
11430 Region_1774 Region_1774
11459 Region_1330
11461 Region_884 Region_884 Region_885 Region_885 Region_886 Region_885
11464 Region_392 Region_393
11465 Region_676
11472 Region_1902
11504 Region_2308
11512 Region_2214 Region_2214
11520 Region_659
11539 Region_107
11639 Region_674 Region_675
11652 Region_1053
11790 Region_54 Region_54
11804 Region_1379
11831 Region_2405
11818 Region_1765
11829 Region_2473
```
Comparison of Tbx20-neighboring genes with regulated genes

As we have expression data available, we can now compare the list of Tbx20-neighboring genes with the previously saved lists of up- and down-regulated genes in Tbx20 knock-out mouse hearts.

Start the List comparison tool from the Tools menu in the navigation bar.

As you will compare three lists to one another, namely the list of Tbx20 neighboring genes the list of up-regulated genes, and the list of down-regulated genes from the expression analysis, set the number of lists accordingly to 3 (marked with 1 in the screenshot below). Provide a name for each list (2,4,7), and upload the corresponding files from your computer (3,5,8).
The list comparison tool allows to include associated values in the output. For uploaded tab-separated text files, you can select how many columns should be evaluated for each file. The default is 2, i.e. the identifier column plus one column with associated values. Set this value to 3 for the files with up- and down-regulated genes (6,9). This will include fold change values and gene names in addition to the gene IDs.

To keep the case as it is in the uploaded files, activate the Case Sensitivity option (otherwise lower case will be converted to upper case in the output). This also makes the ID comparison case-sensitive. The start the comparison.

In the result, you'll find a Venn diagram with the overlap numbers. Of the 1107 neighboring genes, 134 are also found in the up-regulated list, and 207 in the down-regulated list.
To see the complete comparison, export the union of all lists to Excel.
Genes that were present in each input list have an associated value (Region ID for neighboring genes; log fold change and gene symbol for regulated genes); the others get only a dash in the value columns.

Thus you can use Excel functionality to filter e.g. for up-regulated Tbx20-neighboring genes (which would correspond to genes whose expression is probably directly repressed by Tbx20).

The identifiers can then, for example, be uploaded to the Genomatix Pathway System for further analysis.
Literature


List of resources available on the web:

Gene Expression Omnibus:

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

This tutorial was compiled for Genomatix Genome Analyzer v3.51106.

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

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.