

NGS Data Analysis Workshop

Course Tutorial

© 2016 Genomatix Software GmbH

For more information please contact:

Genomatix Software GmbH
Bayerstr. 85a
80335 Munich
Germany

Phone: +49 89 599766 0
Fax: +49 89 599766 55
Email: info@genomatix.de
WWW: <http://www.genomatix.de>

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-Sequencing 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
ModellInspector: 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 EIDorado. 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.

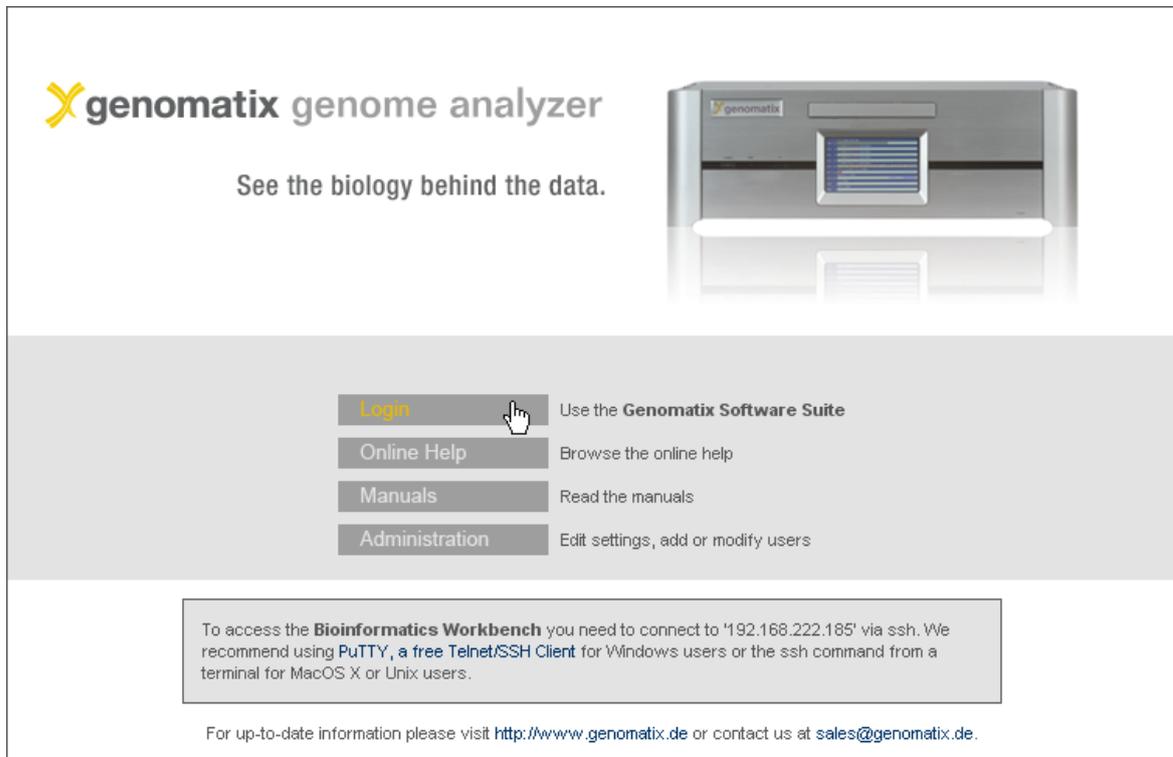
Variant analysis

Genome wide small variant analysis identifies effects on protein sequences and TF binding sites, using the genome and TF binding site annotation in EIDorado 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:

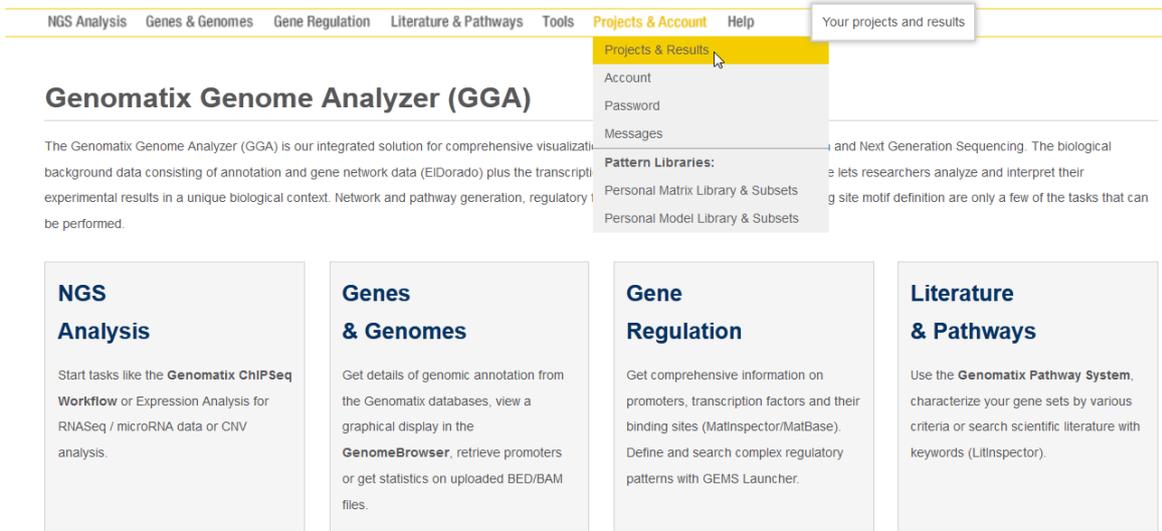
Please log in:

Username:

Password:

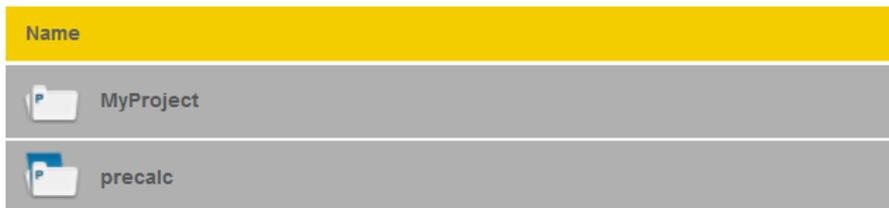
Creating a project

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.

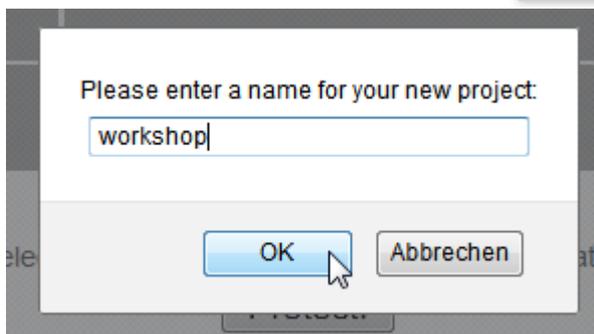
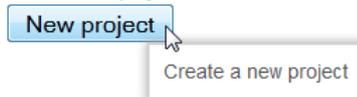


The screenshot shows the Genomatix Genome Analyzer (GGA) interface. At the top, there is a navigation menu bar with items: NGS Analysis, Genes & Genomes, Gene Regulation, Literature & Pathways, Tools, Projects & Account, and Help. A dropdown menu is open under 'Projects & Account', showing options: Projects & Results (highlighted), Account, Password, Messages, and Pattern Libraries (with sub-items: Personal Matrix Library & Subsets, Personal Model Library & Subsets). Below the menu, the main content area is titled 'Genomatix Genome Analyzer (GGA)' and contains a brief description of the tool. Below the description are four main sections: NGS Analysis, Genes & Genomes, Gene Regulation, and Literature & Pathways, each with a short description of its capabilities.

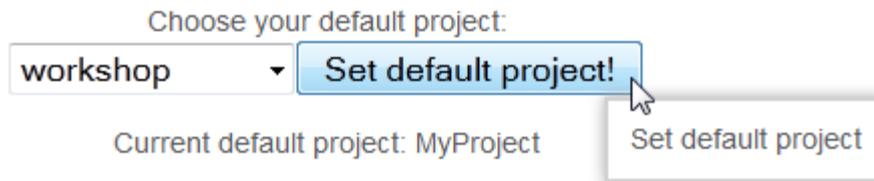
Press the New project button, enter a name for your project in the pop-up dialog, and click on OK.



Create a new project:



Using the controls, set the new project as your default project.



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

Principle component analysis

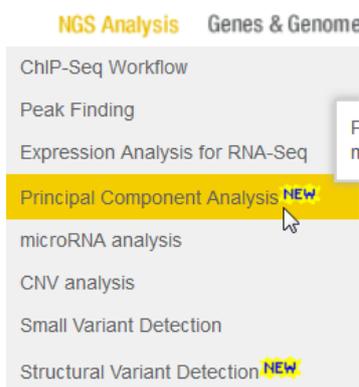
Principal component analysis (PCA) is a statistical procedure that can be used for exploratory data analysis. PCA uses linear combinations of the original data (e.g. gene expression values) to define a new set of unrelated variables (principal components). These new variables are orthogonal to each other,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

Current project: Current Genome: GRCh38

Principal Component Analysis

Principal Component Analysis for RNASeq data

Input	
Available files	Listing files for Homo sapiens / GRCh38: Select <input checked="" type="radio"/> BED files or <input type="radio"/> BAM files No BED/BB files for Homo sapiens / GRCh38 in this project yet. <input type="button" value="Add BED files"/>
	Use drag & drop to fill the groups below with available files from above list: Number of Groups: <input type="text" value="1"/>

Anopheles gambiae
 Apis mellifera
 Arabidopsis thaliana
 Bos taurus
 Caenorhabditis elegans
 Camponotus floridanus
 Canis familiaris
 Danio rerio
 Drosophila melanogaster
 Equus caballus
 Gallus gallus
 Glycine max
 Harpegnathos saltator
 Homo sapiens
 Macaca mulatta
 Monodelphis domestica
Mus musculus
 Neurospora crassa
 Ornithorhynchus anatinus
 Oryctolagus cuniculus

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

Current project: Current Genome: NCBI build 38

Principal Component Analysis for NGS Data

Principal Component Analysis for RNASeq data. See help for more.

Input	
Available files	Listing files for Mus musculus / NCBI build 38: Select <input checked="" type="radio"/> BED files or <input type="radio"/> BAM files No BED/BB files for Mus musculus / NCBI build 38 in this project yet. <input type="button" value="Add BED files"/>
	Use drag & drop to fill the groups below with available files from above list: <input type="button" value="Upload more files to your project"/>

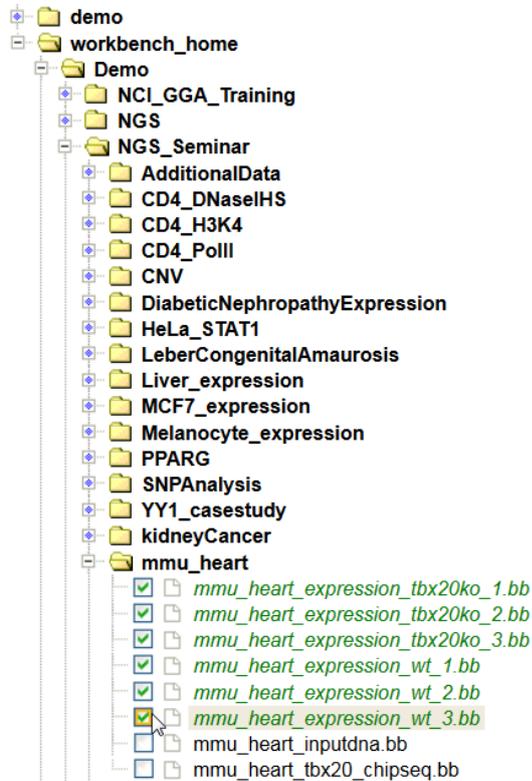
Then select *Import from the GGA*, and press the *Browse GGA* button.

BED File Upload

Current Project: "workshop"

Upload genomic regions	
Upload file(s) with genomic regions in BED file format	Import BED / bigBed file(s) from 1 <input type="radio"/> your local computer <input type="radio"/> the GMS <input checked="" type="radio"/> the GGA
	Assuming input is for <i>Mus musculus</i> / NCBI build 38 Multiple files can be uploaded: <input type="button" value="Browse GGA..."/> 2
	Note, that bigBed files must have the extension '.bb' Optional name/prefix for your BED file(s) on the server: <input type="text"/>

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*.

Upload genomic regions

<p>Upload file(s) with genomic regions in BED file format ?</p>	<p>Import BED / bigBed file(s) from</p> <p> <input type="radio"/> your local computer <input type="radio"/> the GMS <input checked="" type="radio"/> the GGA </p> <p>Assuming input is for Mus musculus / NCBI build 38</p> <p>Multiple files can be uploaded:</p> <p style="text-align: center;">↗ Browse GGA...</p> <p> <input checked="" type="checkbox"/> mmu_heart_expression_tbx20ko_1.bb <input checked="" type="checkbox"/> mmu_heart_expression_tbx20ko_2.bb <input checked="" type="checkbox"/> mmu_heart_expression_tbx20ko_3.bb <input checked="" type="checkbox"/> mmu_heart_expression_wt_1.bb <input checked="" type="checkbox"/> mmu_heart_expression_wt_2.bb <input checked="" type="checkbox"/> mmu_heart_expression_wt_3.bb </p> <p><i>Note, that bigBed files must have the extension '.bb'</i></p> <p>Optional name/prefix for your BED file(s) on the server: <input style="width: 100%;" type="text"/></p>
Email option (for very large, zipped files)	
<p>Your email address ?</p>	<p> <input checked="" type="radio"/> Show result directly in browser window <input type="radio"/> Send the URL of the result to <input style="width: 100%;" type="text" value="courses@genomatix.de"/> </p> <p><i>Use the email option for long-running jobs, to avoid server-timeout messages</i></p> <p>You may set a default email address by filling or modifying the 'email address' field on your personal account page</p>

Submit Reset Form

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 (8980354 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](#)

Close this window or add more BED files...

The uploaded files will be listed as below.

Input	
Available files	<p>Listing files for <i>Mus musculus</i> / NCBI build 38:</p> <p>Select <input checked="" type="radio"/> BED files or <input type="radio"/> BAM files</p> <ul style="list-style-type: none"> mmu_heart_expression_tbx20ko_1.bb (8708085 regions) mmu_heart_expression_tbx20ko_2.bb (9105462 regions) mmu_heart_expression_tbx20ko_3.bb (8980354 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) <p>Add BED files</p>
Parameters for PCA	<p>Use drag & drop to fill the groups below with available files from above list:</p> <p>Number of Groups: 2</p> <div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid #ccc; padding: 5px; width: 45%;"> <p>Group 1</p> <hr/> <p>Files: 0</p> <p>black</p> </div> <div style="border: 1px solid #ccc; padding: 5px; width: 45%;"> <p>Group 2</p> <hr/> <p>Files: 0</p> <p>green</p> </div> </div>
Options	<p><input checked="" type="checkbox"/> Do rlog transformation</p>
Transcript/Locus	<p><input type="radio"/> Locus-based expression analysis (union of exons for all loci, i.e. gene bodies) NEW</p> <p><input type="radio"/> Transcript-based expression analysis (all transcripts separately)</p>

Rename the groups, e.g. Group 1 to *Tbx20 KO*, Group 2 to *WT*. Drag & drop the files into the corresponding group fields. 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.

Input	
Available files	<p>Listing files for <i>Mus musculus</i> / NCBI build 38:</p> <p>Select <input checked="" type="radio"/> BED files or <input type="radio"/> BAM files</p> <p>Add BED files</p>
Parameters for PCA	<p>Use drag & drop to fill the groups below with available files from above list:</p> <p>Number of Groups: 2</p> <div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid #ccc; padding: 5px; width: 45%;"> <p>Tbx20 KO</p> <ul style="list-style-type: none"> mmu_heart_expression_tbx20ko_1.bb (8708085 regions) mmu_heart_expression_tbx20ko_2.bb (9105462 regions) mmu_heart_expression_tbx20ko_3.bb (8980354 regions) <p>Files: 3</p> <p>black</p> </div> <div style="border: 1px solid #ccc; padding: 5px; width: 45%;"> <p>WT</p> <ul style="list-style-type: none"> mmu_heart_expression_wt_1.bb (8028478 regions) mmu_heart_expression_wt_2.bb (8591698 regions) mmu_heart_expression_wt_3.bb (7845462 regions) <p>Files: 3</p> <p>green</p> </div> </div>
Options	<p><input checked="" type="checkbox"/> Do rlog transformation</p>
Transcript/Locus	<p><input type="radio"/> Locus-based expression analysis (union of exons for all loci, i.e. gene bodies) NEW</p> <p><input checked="" type="radio"/> Transcript-based expression analysis (all transcripts separately)</p> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 5px;"> <p>Source of transcripts</p> <ul style="list-style-type: none"> <input checked="" type="radio"/> All sources (non-redundant transcripts) <input type="radio"/> NCBI RefSeq <input type="radio"/> Ensembl <input type="radio"/> NCBI GenBank </div>
Output	
Result name	<p>Result name: <input type="text" value="result_pca"/></p> <p>(special characters except +, ., * are not allowed and will be replaced by _)</p>
Your email address	<p><input type="text" value="courses@genomatix.de"/></p> <p>You may set a default email address by filling or modifying the 'email address' field on your personal account page</p>
<p><input type="button" value="Submit"/> <input type="button" value="Reset Form"/></p>	

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.

Your submitted jobs				
Job-ID	Task	State	Submitted at	Remove job
3459	Principal Component Analysis	RUNNING	2015-07-10T10:44:57	Remove job

Project Management

Name	Comment	Created	Automatic deletion in	Action
 MyProject		2006-09-18 19:42:29		

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.

 workshop	
 <input type="checkbox"/> BED files	containing 6 BED files
 <input type="checkbox"/> Principal Component Analysis	containing 1 result
<input type="checkbox"/> result_pca 	

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.

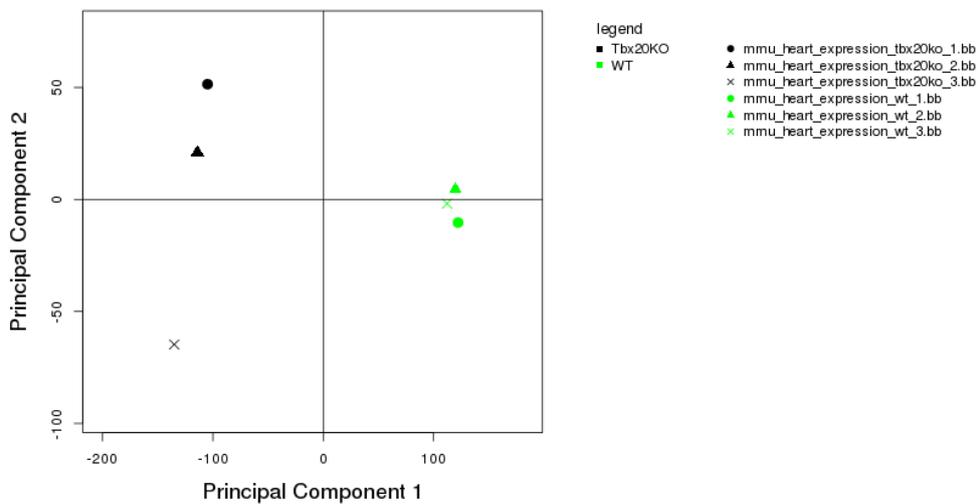
Overview	PC1	PC2	PC3	3D	Download of Results
----------	-----	-----	-----	----	---------------------

PCA Info		Coefficient of Determination						
		1	2	3	4	5	6	
samples	6							
PCs	6							
variables	217159	R2	0.793	0.070	0.057	0.042	0.038	0.000
method	svd	R2cum	0.793	0.863	0.920	0.962	1.000	1.000

Score plot

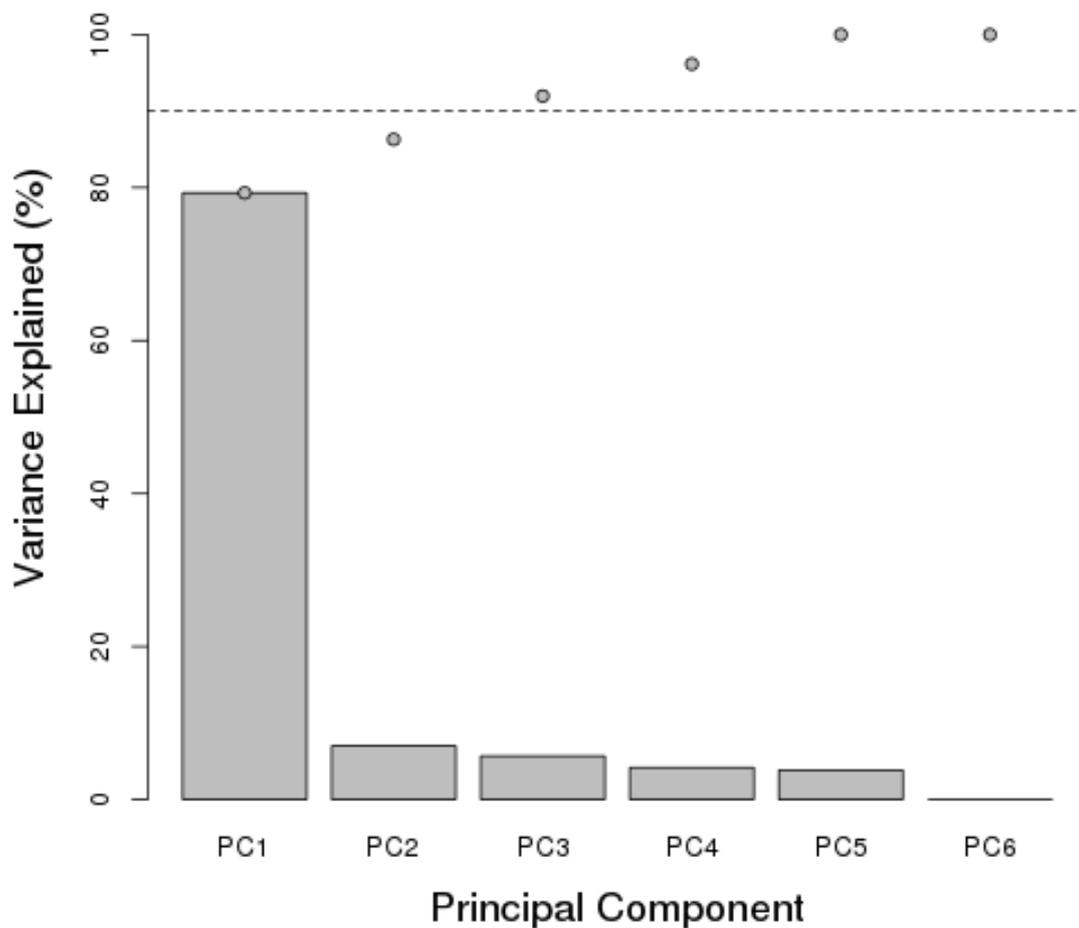
The score plot displays each sample in the data set with respect to the first two principal components and can therefore be used to interpret the relations among the samples. This information can be used to identify outliers.

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.



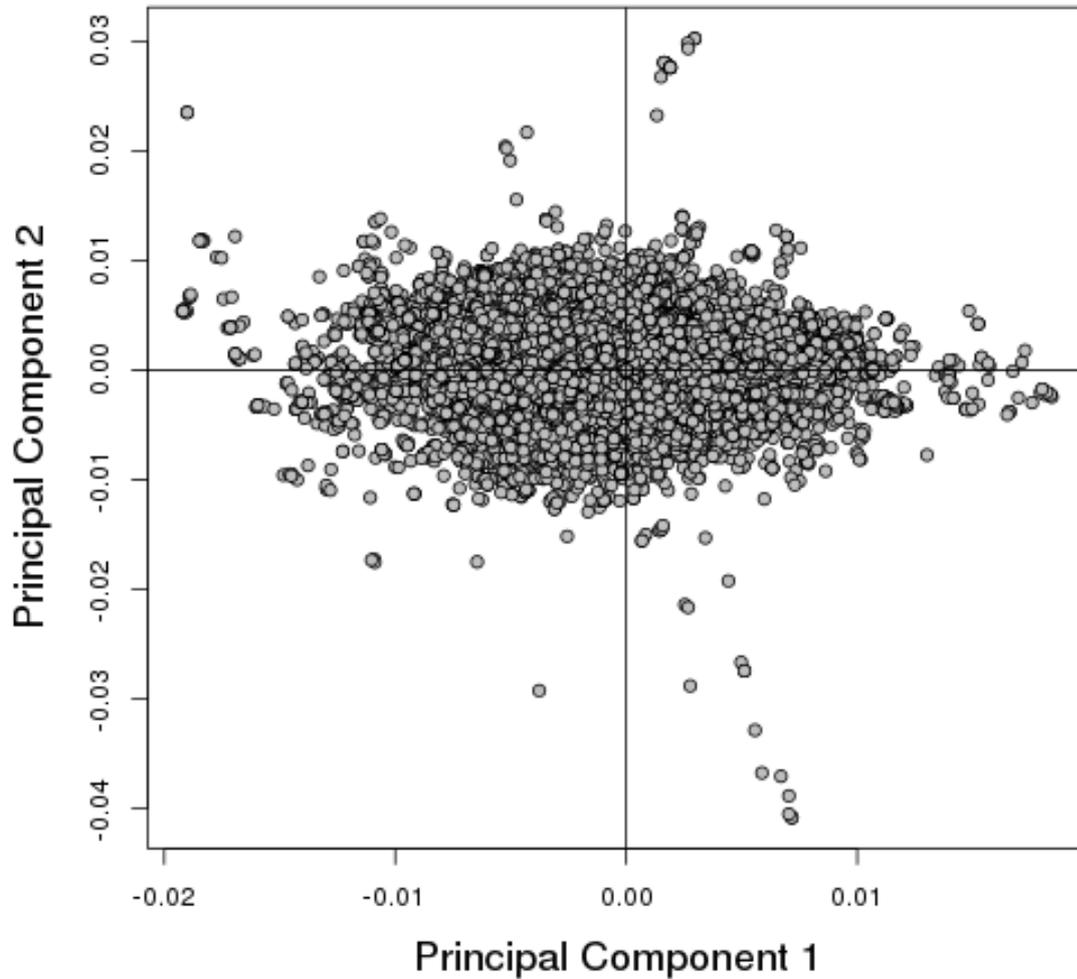
Scree plot

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^2_{cum}). 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

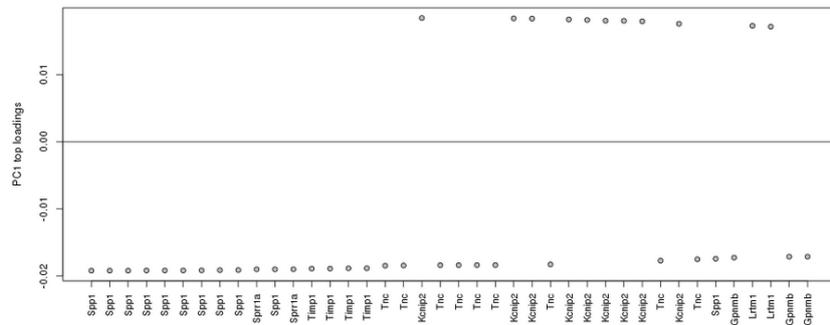
For the top principal components that are needed to account for 90% of the variance in the data (or up to a maximum of 10 PCs) the 40 transcripts/loci with the highest absolute loadings are shown in a table and a plot.

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.

[Overview](#)
[PC1](#)
[PC2](#)
[PC3](#)
[3D](#)
[Download of Results](#)

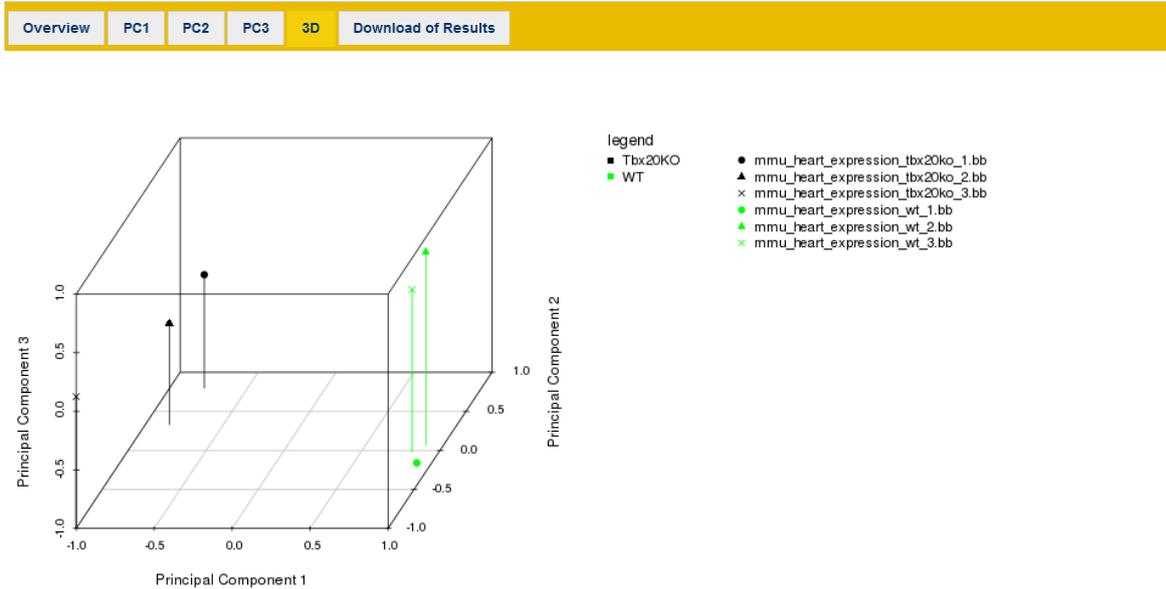
Top 40 Loadings for PC1

Rank	GeneID	Symbol	Loading
1	20750	Spp1	-0.0192
2	20750	Spp1	-0.0192
3	20750	Spp1	-0.0192
4	20750	Spp1	-0.0192
5	20750	Spp1	-0.0192
6	20750	Spp1	-0.0192
7	20750	Spp1	-0.0192
8	20750	Spp1	-0.0191
9	20750	Spp1	-0.0191
10	20753	Sprr1a	-0.0190
11	20750	Spp1	-0.0190
12	20753	Sprr1a	-0.0190
13	21857	Timp1	-0.0189
14	21857	Timp1	-0.0189
15	21857	Timp1	-0.0189
16	21857	Timp1	-0.0189
17	21923	Tnc	-0.0185
18	21923	Tnc	-0.0184



3D score plot

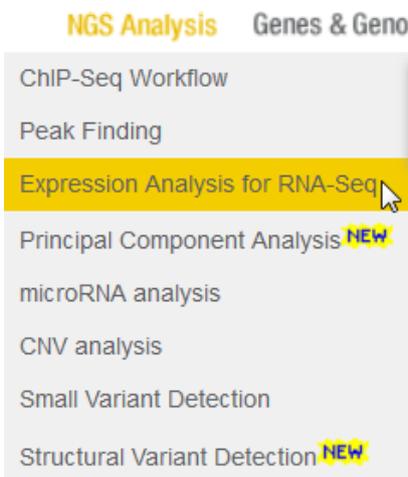
This score plot displays each sample in the data set with respect to the first three principal components.



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.

Input file(s) with read positions from RNA-Seq ("Treatment")
 Note: multiple files are treated as replicates

Available files Add BED files

Listing files for Mus musculus / NCBI build 38:
 Select BED files or BAM files

- mmu_heart_expression_tbx20ko_1.bb (8708085 regions)
- mmu_heart_expression_tbx20ko_2.bb (9105462 regions)
- mmu_heart_expression_tbx20ko_3.bb (8980354 regions)
- mmu_heart_expression_wt_1.bb (8028478 regions)

(You can use shift/ctrl-keys to select multiple files)

Control files / Different condition (optionally with replicates)

Optional: control file(s) for differential analysis Add BED files

Use second set of input files (different condition / control files) for differential gene expression analysis

Select BED files or BAM files

- mmu_heart_expression_tbx20ko_2.bb (9105462 regions)
- mmu_heart_expression_tbx20ko_3.bb (8980354 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)

(You can use shift/ctrl-keys to select multiple files)

Currently 3 BED files are selected as control.

Method for differential analysis:

Audic-Clavien (only if no replicates available) [\(details\)](#)

DESeq, recommended only for replicates [\(details\)](#)

DESeq2, recommended only for replicates [\(details\)](#)

Statistical testing method: Wald test Likelihood ratio test

Dispersion fitting method: parametric local mean

edgeR, only for replicates [\(details\)](#)

List transcripts as significant, if:

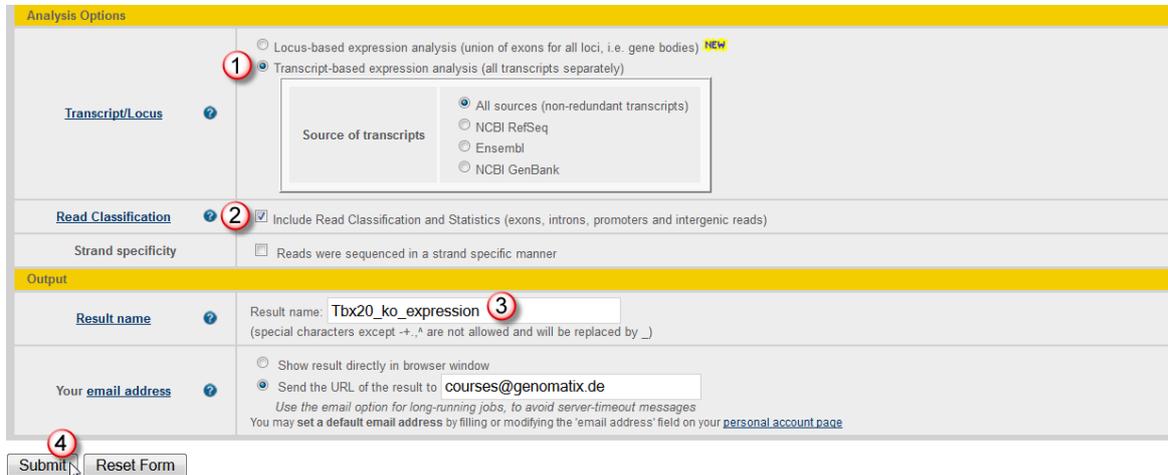
adjusted p-value threshold

and $\log_2(\text{fold-change}) \geq 1$ for up-regulation in condition1 ("treatment") compared to condition2 ("control")

and $\log_2(\text{fold-change}) \leq -1$ for down-regulation in condition1 compared to condition2

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

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.



Analysis Options

Locus-based expression analysis (union of exons for all loci, i.e. gene bodies) **NEW**

Transcript-based expression analysis (all transcripts separately)

Transcript/Locus

Source of transcripts

- All sources (non-redundant transcripts)
- NCBI RefSeq
- Ensembl
- NCBI GenBank

Read Classification

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

Strand specificity Reads were sequenced in a strand specific manner

Output

Result name

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

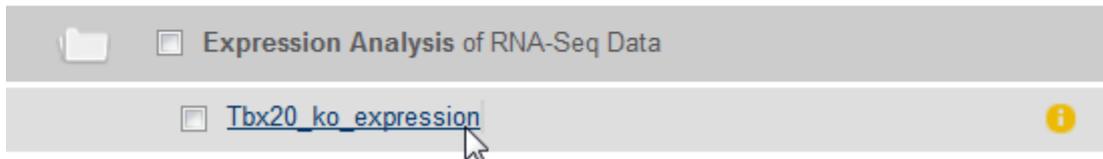
Your email address

Show result directly in browser window

Send the URL of the result to

Use the email option for long-running jobs, to avoid server-timeout messages. You may set a default email address by filling or modifying the 'email address' field on your [personal account page](#)

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

	Transcripts	Genes (known GeneID)
Total number analyzed	217159 ↓ download details (tab-separated) (62Mb)	29812 ↓ download details (tab-separated) (2.2Mb)
Differential expression	31049 ↓ download details (tab-separated) (9.0Mb)	4927
Up-regulation	17021 ↓ Download BED file of Transcripts (1.1Mb) ✏ Save BED file to project management	2729 ↓ download details (tab-separated) (288Kb) ↓ download gene list (52Kb)
Down-regulation	14028 ↓ Download BED file of Transcripts (881Kb) ✏ Save BED file to project management	2214 ↓ download details (tab-separated) (244Kb) ↓ download gene list (48Kb)
Up- and down-regulated genes (with different transcripts)	-	16 ↓ download details (tab-separated) (4.0Kb)

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.

TranscriptId	Accn	LocusId	Symbol	GeneId	ContigAccn	Chromosome	Strand	Start	End	Length	#exons	p-value	adj. p-value	log2(fold change)	Regulation
GXT_12942264	AK090041	GXL_1787596	Slamf9	98365 NC_000067	chr1	+	172475374	172478575	1297	4	2.55E-04	1.06E-03		1.05 up	
GXT_12942270	AK089400	GXL_1196246	Kif21b	16565 NC_000067	chr1	+	136131454	136149993	2606	6	2.24E-05	1.14E-04		2.16 up	
GXT_12942315	AK088077	GXL_87684	Trmt1l	98685 NC_000067	chr1	+	151428666	151436707	1528	3	1.25E-03	4.42E-03		-1.17 down	
GXT_12942316	AK088027	GXL_742666	Cd48	12506 NC_000067	chr1	+	171682009	171705256	920	3	1.26E-02	3.36E-02		1.62 up	
GXT_12942320	AK087631	GXL_20287	Irf6	54139 NC_000067	chr1	+	193153154	193166868	1760	5	2.18E-03	7.21E-03		1.85 up	
GXT_12942323	AK087427	GXL_87676	Arpc5	67771 NC_000067	chr1	+	152766676	152775503	1687	4	8.91E-14	1.33E-12		1.41 up	
GXT_12942325	AK086974	GXL_110155	Stradb	227154 NC_000067	chr1	+	58973641	58991512	1229	7	4.23E-07	2.88E-06		-1.15 down	
GXT_12942344	AK085015	GXL_6599	Arid5a	214855 NC_000067	chr1	+	36307760	36322975	4444	5	1.67E-13	2.42E-12		1.6 up	
GXT_12942346	AK084971	GXL_110144	Fastkd2	75619 NC_000067	chr1	+	63730651	63753385	3179	12	3.20E-06	1.89E-05		-1.02 down	
GXT_12942350	AK084836	GXL_110247	AK084836	0 NC_000067	chr1	+	74295592	74297905	1745	2	7.56E-03	2.15E-02		1.35 up	

...

#reads treat1	#reads treat2	#reads treat3	#reads ctrl1	#reads ctrl2	#reads ctrl3	NE treat1	NE treat2	NE treat3	NE ctrl1	NE ctrl2	NE ctrl3	mean NE(treat)	stddev NE(treat)	mean NE(ctrl)	stddev NE(ctrl)
102	79	79	24	31	33	0.09079	0.06781	0.0689	0.02337	0.02824	0.03292	0.07583	0.01059	0.02818	0.0039
29	32	33	5	5	3	0.01253	0.01282	0.01259	0.00242	0.00208	0.00149	0.01265	0.00012	0.002	0.00038
39	28	12	45	50	38	0.02899	0.02002	0.00888	0.0364	0.03834	0.03185	0.0193	0.00823	0.03553	0.00272
35	12	10	3	8	0	0.04435	0.01452	0.0123	0.00412	0.01027	0	0.02372	0.01461	0.0048	0.00422
20	19	16	2	5	2	0.01259	0.01075	0.01028	0.00072	0.00336	0.00147	0.01121	0.001	0.00185	0.00111
414	395	577	144	116	109	0.28403	0.26067	0.38556	0.10707	0.08124	0.0813	0.31009	0.05421	0.08987	0.01216
76	73	59	99	120	116	0.0645	0.05979	0.0497	0.0925	0.09902	0.10633	0.058	0.00617	0.09928	0.00565
286	447	503	111	81	94	0.07477	0.11098	0.12669	0.03127	0.02154	0.02737	0.10415	0.02174	0.02673	0.004
97	88	68	127	131	114	0.03447	0.03082	0.02313	0.05007	0.04683	0.04558	0.02947	0.00473	0.04749	0.00189
13	30	43	7	6	9	0.00869	0.01914	0.02787	0.00507	0.00406	0.00667	0.01857	0.00784	0.00527	0.00107

...

RPKM treat1	RPKM treat2	RPKM treat3	RPKM ctrl1	RPKM ctrl2	RPKM ctrl3	mean RPKM(treat)	stddev RPKM(treat)	mean RPKM(ctrl)	stddev RPKM(ctrl)
9.03104	6.68937	6.78256	2.30483	2.78191	3.24306	7.50099	1.08258	2.7766	0.38305
1.27791	1.34857	1.41009	0.23898	0.22331	0.14673	1.34552	0.05401	0.20301	0.0403
2.93102	2.01249	0.87451	3.66822	3.80862	3.16987	1.93934	0.84116	3.5489	0.27408
4.36875	1.43249	1.21037	0.40616	1.0121	0	2.3372	1.43938	0.47275	0.41586
1.30495	1.1856	1.01231	0.14154	0.33066	0.14484	1.16762	0.12014	0.20568	0.08838
28.1814	25.71461	38.08617	10.63198	8.0032	8.23556	30.66073	5.34629	8.95691	1.18824
7.10132	6.52332	5.34573	10.03345	11.3645	12.03061	6.32346	0.73052	11.14285	0.83026
7.39042	11.04667	12.60377	3.11111	2.12145	2.6961	10.34695	2.18509	2.64289	0.40578
3.50936	3.04012	2.38191	4.976	4.79625	4.57084	2.97533	0.46036	4.78103	0.16576
0.85551	1.8881	2.74397	0.49965	0.4002	0.6574	1.82919	0.77208	0.51908	0.1059

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:

GeneId	Symbol	#transcripts regulated	total #transcripts for gene	mean log2(fold change) of reg. trans.	min fold change of reg. trans.	max fold change of reg. trans.	fc stddev	min p_value
80906	Kcnp12	13	13	-6.141	-6.409	-5.799	0.19	2.05E-230
68052	Rps13	3	7	-6.139	-6.139	-6.139	0	6.16E-14
13643	Efnb3	3	3	-5.773	-5.796	-5.76	0.016	1.16E-152
319476	Lrtm1	5	5	-5.391	-5.516	-5.276	0.091	5.28E-123
142687	Asb14	8	8	-5.305	-5.737	-3.437	0.73	7.60E-74
319942	A530016L24Rik	6	6	-5.299	-5.457	-5.125	0.109	7.50E-65
30952	Cngb3	3	3	-5.081	-5.165	-4.946	0.097	2.03E-10
213402	Armc2	10	12	-4.801	-5.495	-2.308	0.877	8.12E-42
78910	Asb15	7	7	-4.791	-5.115	-4.349	0.288	8.58E-57
238564	Mylk4	10	11	-4.64	-4.902	-3.6	0.352	4.09E-114

...

mean NE(treat.reg.)	stddev NE(treat.reg.)	mean NE(ctrl.reg.)	stddev NE(ctrl.reg.)	mean RPKM(treat.reg.)	stddev RPKM(treat.reg.)	mean RPKM(ctrl.reg.)	stddev RPKM(ctrl.reg.)
0.01348	0.005	0.81806	0.229	1.32801	0.506	81.58881	22.09
0	0	0.28783	0.165	0	0	28.36193	16.262
0.00722	0.001	0.32802	0.043	0.71178	0.106	32.45018	4.281
0.01947	0.01	0.72033	0.236	1.93015	0.94	71.11432	23.38
0.00445	0.002	0.18749	0.06	0.46146	0.266	19.15481	5.898
0.00459	0.001	0.16635	0.039	0.46903	0.13	16.40701	3.813
0	0	0.00876	0.003	0	0	0.88762	0.325
0.00179	0.002	0.0375	0.008	0.17655	0.219	3.77305	0.785
0.00683	0.004	0.19036	0.042	0.67324	0.381	18.99107	4.015
0.00964	0.004	0.21467	0.083	0.96041	0.442	21.88284	8.548

You can download a simple list of regulated genes with Gene IDs, log₂ fold changes, and gene symbols.

80906	-6.141	Kcnp2
68052	-6.139	Rps13
13643	-5.773	Efnb3
319476	-5.391	Lrtm1
142687	-5.305	Asb14
319942	-5.299	A530016L24Rik
30952	-5.081	Cngb3
213402	-4.801	Armc2
78910	-4.791	Asb15
238564	-4.64	Mylk4

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.

Up-regulation	17021 Download BED file of Transcripts (1.1Mb) Save BED file to project management	2729 download details (tab-separated) (288Kb) download gene list (52Kb)
Down-regulation	14028 Download BED file of Transcripts (881Kb) Save BED file to project management	2214 download details (tab-separated) (244Kb) download gene list (48Kb)
Up- and down-regulated genes (with different transcripts)	-	16 download details (tab-separated) (4.0Kb)

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

Save selected BED file as

to project

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.

Differential Expression Overview

	Transcripts	Genes (known GenelD)
Total number analyzed	217159 download details (tab-separated) (62Mb)	29812 download details (tab-separated) (2.2Mb)
Differential expression	31049 download details (tab-separated) (9.0Mb)	4927
Up-regulation	17021 Download BED file of Transcripts (1.1Mb) Save BED file to project management	2729 download details (tab-separated) (288Kb) 1 download gene list (52Kb)
Down-regulation	14028 Download BED file of Transcripts (881Kb) Save BED file to project management	2214 download details (tab-separated) (244Kb) 2 download gene list (48Kb)
Up- and down-regulated genes (with different transcripts)	-	16 download details (tab-separated) (4.0Kb)

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, ...):

Symbol	GeneId	mean log ₂ (fold change) of up-reg. transcripts
Spp1	20750	7.01
Timp1	21857	6.54
Sprr1a	20753	5.91
Bglap3	12095	5.90
Tnc	21923	5.85
>>> 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, ...):

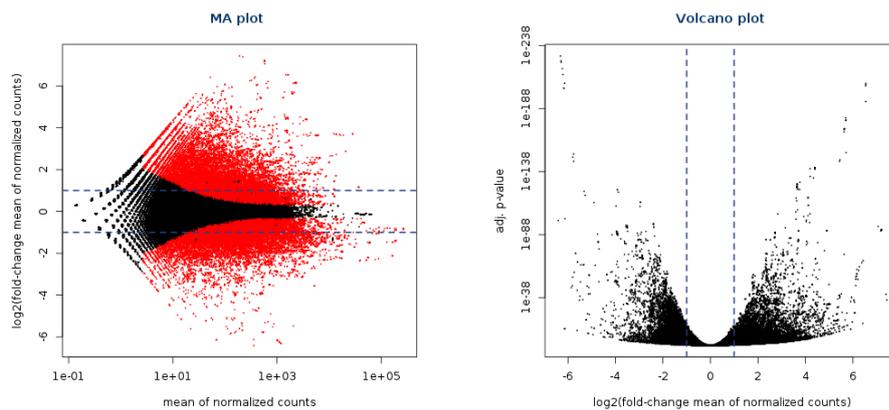
Symbol	GeneId	mean log ₂ (fold change) of down-reg. transcripts
Kcnip2	80906	-6.14
Rps13	68052	-6.14
Efnb3	13643	-5.77
Lrtm1	319476	-5.39
Asb14	142687	-5.30
>>> show more genes <<< (top 50)		

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

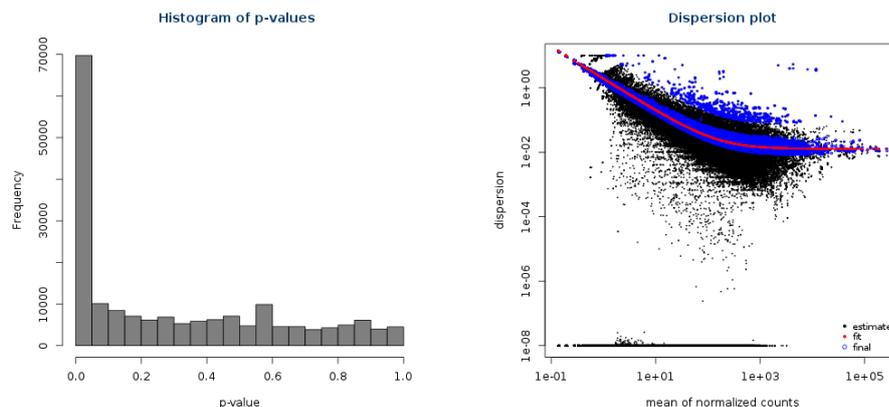
Right: volcano plot of adjusted p-value (y-axis, inverted scale) vs. log₂ fold-change mean of normalized counts (x-axis). The volcano plot shows statistical significance (p-value) and biological significance (effect size as log₂ fold change) in one graph.



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:

Differential Expression Analysis | **Read Classification for all files** | Expression Analysis for sample(s) | Expression Analysis for control(s) | Download of Results

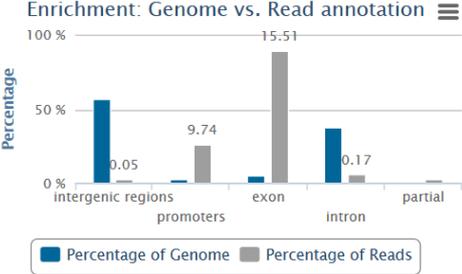
Read Classification on mmu_heart_expression_tbx20ko_1.bb

General Statistics

Total number of Reads:	8708085
Total basepairs:	308793936
Minimum Read length:	9
Maximum Read length:	36
Average Read length:	35.5

Enrichment | **General**

Enrichment: Genome vs. Read annotation



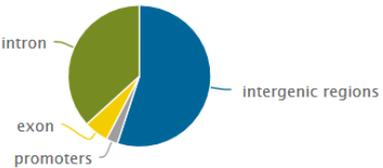
Type of genomic element	Number of Reads	Percentage of Reads	Percentage in Genome	Enrichment compared to Genome
Exon	7698046	88.4%	5.7%	15.5
Partial	218211	2.5%	-	-
Intron	566581	6.5%	37.8%	0.2
Intergenic regions	225247	2.6%	56.5%	0.0
Sum of above	8708085	100.0%	-	-
Promoters	2292235	26.3%	2.7%	9.7

Distribution of Reads on the Genome
[>>> show details <<<](#)

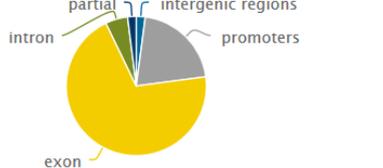
The read classification results can also be shown as pie charts; the left graph shows the fractions of the different annotations in the genome; the right diagram shows the percentages of the corresponding read annotations:

Enrichment | **General**

Genome Annotation



Read Annotation



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.

Pathway and Network analysis

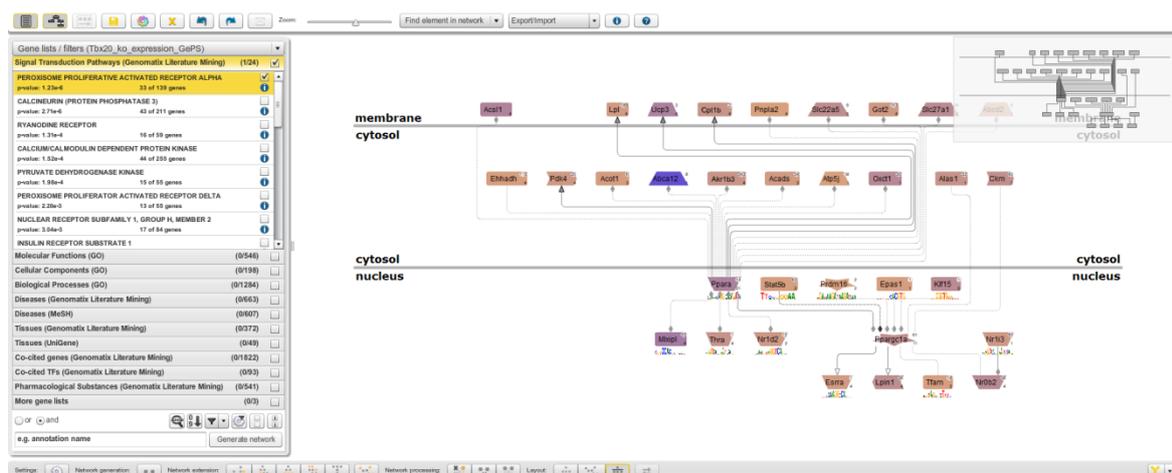
Start [Genomatix Pathway System](#) 

with the top up-regulated genes
 and with the top down-regulated genes
 and name result (opens new window)
 Use [orthologous genes in human](#)  for the pathway analysis

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.

Cellular Components (GO) (0/198)	Diseases (Genomatix Literature Mining) (0/663)
mitochondrion p-value: 2.02e-164 515 of 1664 genes	CARDIOMYOPATHIES p-value: 1.22e-35 105 of 291 genes
mitochondrial part p-value: 1.27e-94 244 of 647 genes	PEARSON'S MARROW PANCREAS SYNDROME p-value: 3.30e-24 49 of 100 genes
cytoplasmic part p-value: 7.09e-92 944 of 6158 genes	NICOTINAMIDE ADENINE DINUCLEOTIDE COE... p-value: 7.85e-23 31 of 43 genes
cytoplasm p-value: 3.46e-90 1235 of 9227 genes	LEIGH DISEASE p-value: 1.38e-18 32 of 57 genes
mitochondrial inner membrane p-value: 4.72e-83 170 of 363 genes	MITOCHONDRIAL DISEASES p-value: 2.79e-18 47 of 120 genes
mitochondrial membrane p-value: 1.08e-80 197 of 496 genes	SUDDEN CARDIAC DEATH p-value: 1.13e-17 59 of 184 genes
mitochondrial envelope p-value: 1.73e-79 202 of 527 genes	HEART FAILURE p-value: 1.23e-16 82 of 328 genes
organelle inner membrane	DILATED CARDIOMYOPATHY
Tissues (Genomatix Literature Mining) (0/372)	Tissues (UniGene) (0/49)
ENTIRE HEART p-value: 2.50e-64 303 of 1181 genes	heart p-value: 7.19e-86 1301 of 9787 genes
HEART p-value: 2.23e-59 293 of 1172 genes	cardiovascular system p-value: 2.16e-85 1316 of 9967 genes
SKELETAL MUSCLE STRUCTURE p-value: 3.52e-49 246 of 986 genes	central nervous system p-value: 9.39e-41 1428 of 12752 genes
MUSCLE p-value: 2.55e-47 258 of 1087 genes	nervous system p-value: 9.39e-41 1428 of 12752 genes
MYOCARDIUM p-value: 7.60e-27 132 of 522 genes	brain p-value: 9.39e-41 1428 of 12752 genes
MUSCLE CELLS p-value: 5.20e-22 140 of 638 genes	integumental system p-value: 1.94e-36 1558 of 14509 genes
CARDIAC MYOCYTE p-value: 4.95e-21 164 of 823 genes	tongue p-value: 1.22e-35 551 of 3764 genes
STRIATED MUSCLE	pharynx

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.

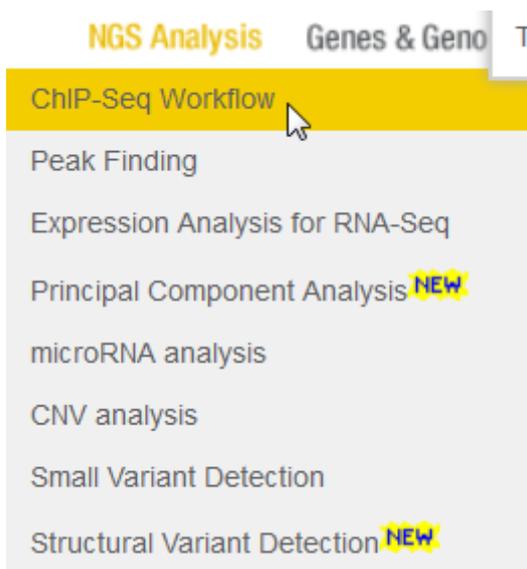
In the first step of the analysis, we will identify genomic regions bound by Tbx20 in wild type adult mouse heart, and run some downstream analyses on these ChIP peak regions.

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.

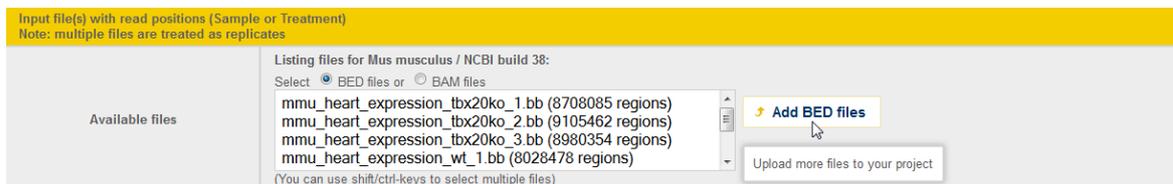
Additionally, a de novo definition of TF binding sites from the ChIP cluster sequences is possible. This uses the program CoreSearch, which can, of course, also be run separately.

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.

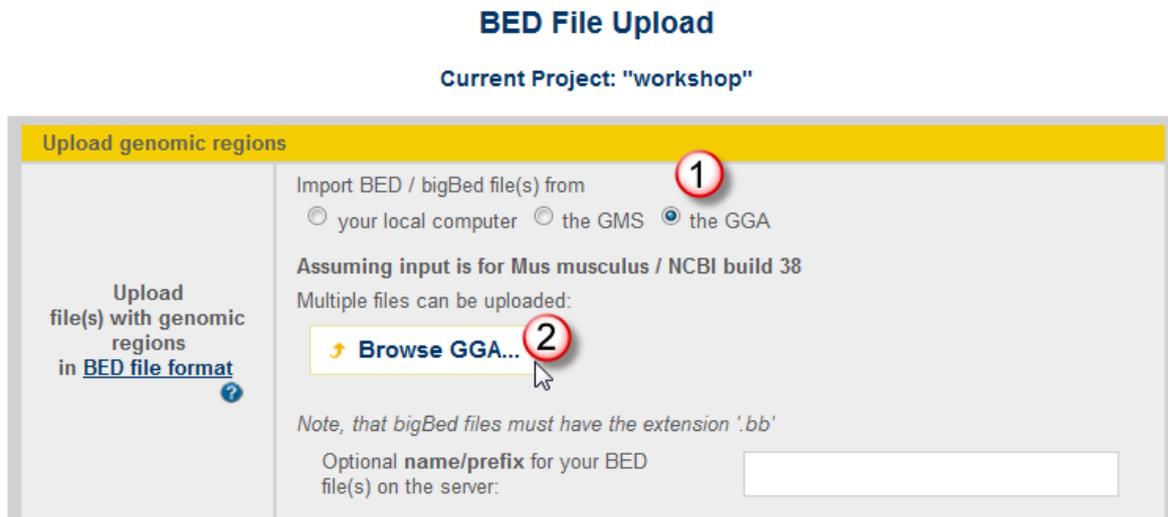
Please open the Genomatix Genome Analyzer in your browser, and select “ChipSeq Workflow” in the NGS Analysis menu.



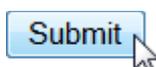
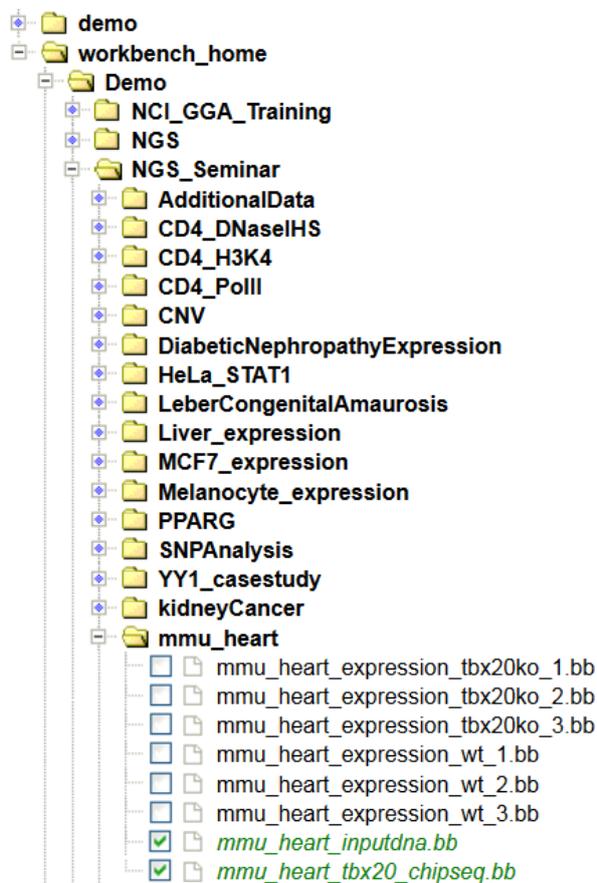
On the input page, press the Add BED files button.



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



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.

Upload genomic regions

Upload file(s) with genomic regions in [BED file format](#) 

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

Assuming input is for **Mus musculus / NCBI build 38**

Multiple files can be uploaded:

 **Browse GGA...**

mmu_heart_inputdna.bb
 mmu_heart_tbx20_chipseq.bb

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

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

Email option (for very large, zipped files)

Your [email address](#) 

Show result directly in browser window
 Send the URL of the result to

Use the email option for long-running jobs, to avoid server-timeout messages
You may set a **default email address** by filling or modifying the 'email address' field on your [personal account page](#)

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

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

- mmu_heart_inputdna.bb (41091391 regions)
- mmu_heart_tbx20_chipseq.bb (5963202 regions)

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

or

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.

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

Available files

Listing files for Mus musculus / NCBI build 38:
Select BED files or BAM files

- mmu_heart_expression_wt_3.bb (8881888 regions)
- mmu_heart_expression_wt_3.bb (7845462 regions)
- mmu_heart_inputdna.bb (41091391 regions)
- mmu_heart_tbx20_chipseq.bb (5963202 regions)**
- Tbx20_ko_expression_transcripts_down.bed (10338 regions)

(You can use shift/ctrl-keys to select multiple files)

Control files (optionally with replicates)

Use second set of input files (control files) for differential analysis

Optional: control file(s) for differential analysis

Select BED files or BAM files

- mmu_heart_expression_wt_3.bb (8881888 regions)
- mmu_heart_expression_wt_3.bb (7845462 regions)
- mmu_heart_inputdna.bb (41091391 regions)**
- mmu_heart_tbx20_chipseq.bb (5963202 regions)
- Tbx20_ko_expression_transcripts_down.bed (10338 regions)

(You can use shift/ctrl-keys to select multiple files)

Workflow parameters

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

Peak Finding (mandatory) Peak Finding / Cluster Generation with

- Genomatix NGSAnalyzer
 - Window size: 100 bp
 - Min. number of reads per peak:
 - calculate automatically from the data by applying a Poisson distribution
 - 100 reads
 - Strand specificity: Reads were sequenced in a strand specific manner
- MACS2/MACS - Model based Analysis for ChIPSeq
- SICER - Spatial clustering for Identification of ChIP-Enriched Regions (for histone modifications) (v1.1)

Replicate Parameters

Replicate treatment No replicate data was selected as input above.

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.

Peak Evaluation

Currently 1 BED file is selected as control.

Method for differential analysis:

- Audic-Claverie (only if no replicates available) (details)
- DESeq, recommended only for replicates (details)
- DESeq2, recommended only for replicates (details)
- edgeR, only for replicates (details)

List regions as significant, if:

adjusted p-value threshold: 0.05

and $\log_2(\text{fold-change}) \geq 1$ for enrichment in condition1 ("treatment") compared to condition2 ("control")

and $\log_2(\text{fold-change}) \leq -1$ for depletion in condition1 compared to condition2

Note: p-value=1 → not using p-value criterion; $\log_2(\text{fold-change})=0$ → not using fold-change criterion

Downstream Analysis

Peak Classification Peak Classification and Statistics

Sequence Extraction Extraction of Sequences for all Peaks

TFBS Overrepresentation Transcription Factor Binding Site Overrepresentation in Peaks

Definition of new TFBS Find new Binding Sites in Peaks (CoreSearch) using the 1000 best-scoring peaks

Output

Result Result name: Tbx20_chipseq (special characters except -,.,* are not allowed and will be replaced by _)

Your email address Send the URL of the result to: courses@genomatix.de

Show result directly in browser window

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

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

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.

Read Classification
Peak Finding
Peak Classification
Sequence Extraction
TFBS Overrepresentation
Definition of new TFBS
Download of Results

Peak Finding / Cluster Generation

Peak finding in input data (mmu_heart_tbx20_chipseq.bb) with NGSAnalyzer

Read and Cluster information	
Total number of peaks	3374
Total reads in peaks	205538
Percentage of reads in peaks	3.45%
Average peak length	144.5 bp

See more details in the [complete results for peak finding step](#) for the input data.

Evaluation with Audic-Claverie Algorithm

2814 peaks were found to be significant with an adjusted p-value of 0.05, 2698 of these show a significant enrichment of reads.

Read and Cluster information	
Total number of peaks	2698
Total reads in peaks	61950
Percentage of reads in peaks	1.04%
Average peak length	152.5 bp

- [Download BED file](#) of the 2698 significantly enriched peaks (104Kb)
- [Save BED file](#) to project management
- [Download p-value Info](#), tab-separated format (212Kb), containing the 2814 significant peaks plus additional info

Please save the BED file with significantly enriched clusters to the project management.

Save selected BED file as

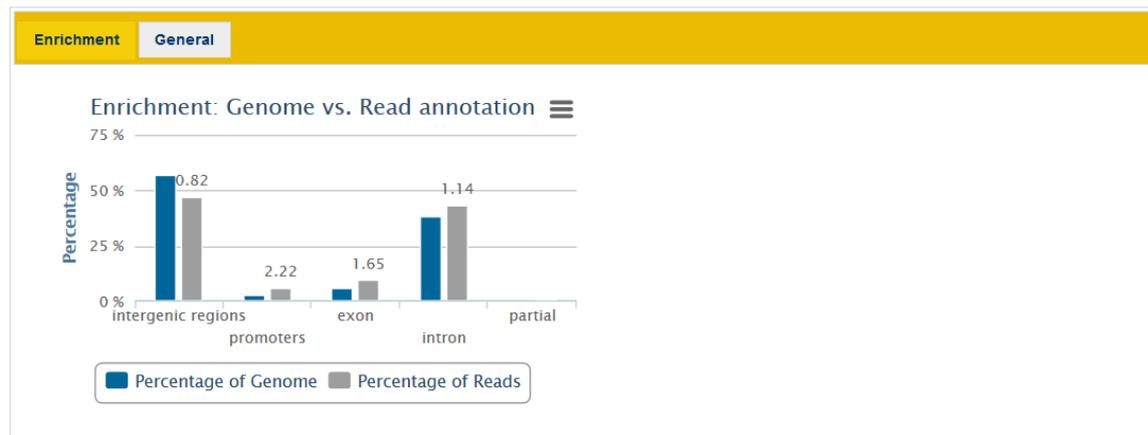
to project

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

General Statistics	
Total number of Reads:	5963202
Total basepairs:	214675272
Minimum Read length:	36
Maximum Read length:	36
Average Read length:	36.0

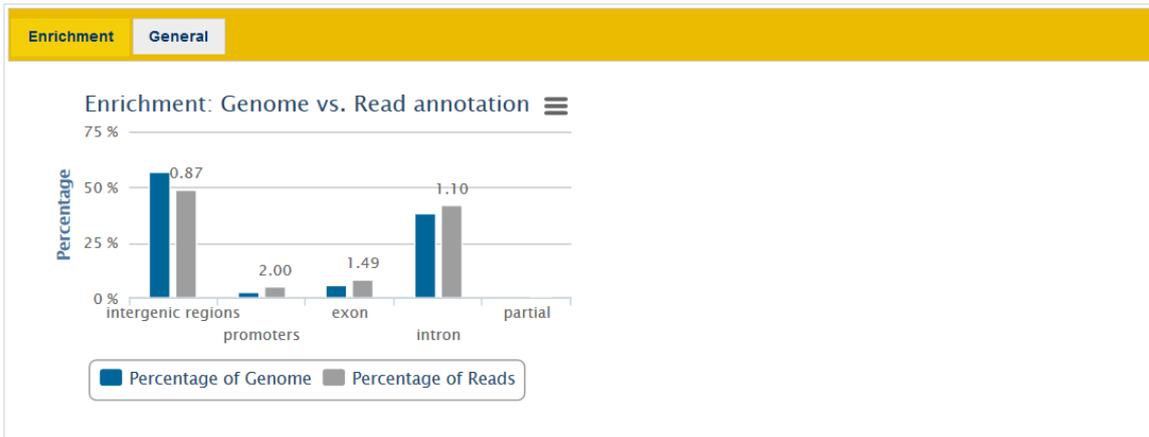


Type of genomic element	Number of Reads	Percentage of Reads	Percentage in Genome	Enrichment compared to Genome
Exon	560195	9.4%	5.7%	1.6
Partial	62348	1.0%	-	-
Intron	2569635	43.1%	37.8%	1.1
Intergenic regions	2771024	46.5%	56.5%	0.8
Sum of above	5963202	100.0%	-	-
Promoters	360351	6.0%	2.7%	2.2

Distribution of Reads on the Genome
[>>> show details <<<](#)

Read Classification on mmu_heart_inputdna.bb

General Statistics	
Total number of Reads:	41091391
Total basepairs:	1479290076
Minimum Read length:	36
Maximum Read length:	36
Average Read length:	36.0



Type of genomic element	Number of Reads	Percentage of Reads	Percentage in Genome	Enrichment compared to Genome
Exon	3489333	8.5%	5.7%	1.5
Partial	379444	0.9%	-	-
Intron	17134331	41.7%	37.8%	1.1
Intergenic regions	20088283	48.9%	56.5%	0.9
Sum of above	41091391	100.0%	-	-
Promoters	2206277	5.4%	2.7%	2.0

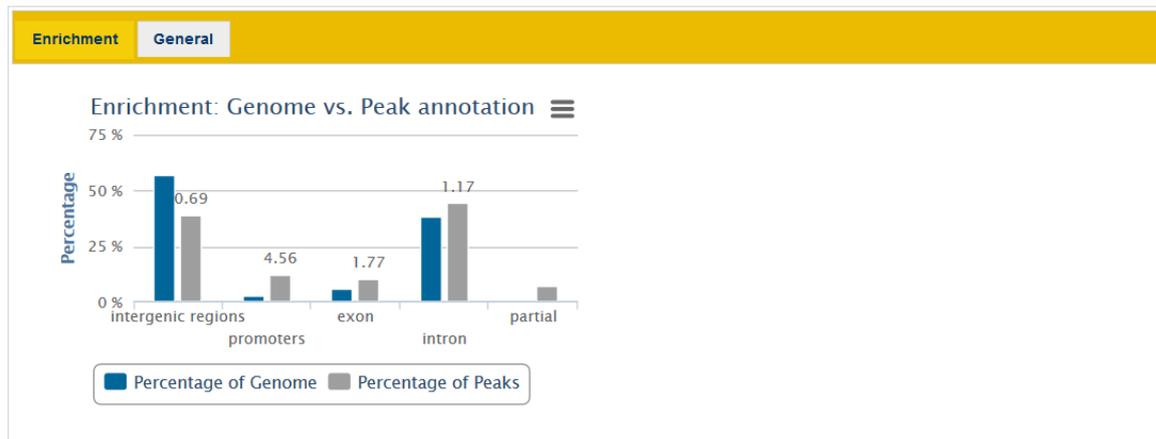
Distribution of Reads on the Genome
[>>> show details <<<](#)

Peak classification

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

Peak Classification on claverie_result.bed

General Statistics	
Total number of Peaks:	2698
Total basepairs:	411423
Minimum Peak length:	36
Maximum Peak length:	4808
Average Peak length:	152.5



Type of genomic element	Number of Peaks	Percentage of Peaks	Percentage in Genome	Enrichment compared to Genome
Exon	272	10.1%	5.7%	1.8
Partial	183	6.8%	-	-
Intron	1190	44.1%	37.8%	1.2
Intergenic regions	1053	39.0%	56.5%	0.7
Sum of above	2698	100.0%	-	-
Promoters	332	12.3%	2.7%	4.6

Distribution of Peaks on the Genome
[>>> show details <<<](#)

Sequence extraction

The peak sequences can be saved in the next section:

Read Classification | **Peak Finding** | Peak Classification | **Sequence Extraction** | TFBS Overrepresentation | Definition of new TFBS | Download of Results

Extraction of Sequences for all Peaks

2698 sequences with a total of 411423 basepairs were extracted (621Kb).

First few lines of the result file:

```
>Region_1 chr=1|start=4412567|end=4412753|str=+|bed_id=1|score=9.12e-13
AGCGGGCAGGAACCGGAGCTTTCCACAGGGCTGAGCCTGGCCCTCCACTGAGCAGTGTCTGCATTCCAAGGCTCCAG
CCTGTACCACCCCTTCCAATCCCTTTGAAGCTGGGCAAAAGGCCTGCCAACAGCACCAAACTTGAGAGCTCCTCTGCCA
GCCCTGGGAGGGGCTGTTCTCCTGCTG
>Region_2 chr=1|start=7819512|end=7819581|str=+|bed_id=2|score=0.0141
TCCCGTGACGACAACTCGCCGATGGGCTGCAGCCAAACAGGGAGTGACACGTCCTAGGCGGAGGATAA
>Region_3 chr=1|start=11003565|end=11003662|str=+|bed_id=3|score=2.63e-08
AAGTGCCTCCTGTCCCTCAGGGTTCTGTGTTTCAAACCTTAGCTCAACAGGATGAGCCCTTTCAGGTTCCACAT
TATCTGATAACTGGTATG
>Region_4 chr=1|start=14227616|end=14227812|str=+|bed_id=4|score=4.42e-25
...
```

[Download sequence file \(621Kb\)](#)
 to project management

TFBS overrepresentation

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

Read Classification	Peak Finding	Peak Classification	Sequence Extraction	TFBS Overrepresentation	Definition of new TFBS	Download of Results
<p>Transcription Factor Binding Site Overrepresentation in Peaks</p> <p>2698 sequence(s) with a total of 411423 basepairs were analyzed.</p> <p>V\$TALE is most overrepresented (Z-score=42.89) compared to the genomic background (1980 matches in 1293 sequences) V\$TALE is most overrepresented (Z-score=45.17) compared to the background of promoters (1980 matches in 1293 sequences)</p> <p>See the complete list of transcription factors and their distribution</p>						

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.

Listing of all TF Families

TF Families	Prom. assoc. known	Nr. of Input Seq. with Match	Nr. of Matches in Input	Match details	Expected (genome) ± Std.dev.	Over representation (genome)	Z-Score (genome)	Expected (promoters) ± Std.dev.	Over representation (promoters)	Z-Score (promoters)
V\$TALE	no	1293	1980	list/seq	781.52±27.93	2.53	42.89	746.54±27.30	2.65	45.17
V\$BRAC	no	1444	2198	list/seq	944.36±30.70	2.33	40.83	873.14±29.52	2.52	44.87
V\$MYOD	no	913	1782	list/seq	834.02±28.85	2.14	32.84	1028.98±32.04	1.73	23.49
V\$NF1F	no	671	1000	list/seq	414.95±20.36	2.41	28.71	463.37±21.51	2.16	24.92
V\$ZF5F	yes	137	318	list/seq	84.52±9.19	3.76	25.34	697.82±26.39	0.46	-14.41
V\$ZF11	no	398	475	list/seq	157.44±12.55	3.02	25.27	240.11±15.49	1.98	15.13
V\$CTCF	yes	524	725	list/seq	294.94±17.17	2.46	25.02	836.10±28.89	0.87	-3.86
V\$AP4R	no	352	451	list/seq	151.13±12.29	2.98	24.36	228.73±15.12	1.97	14.67
V\$AP1R	no	1197	2099	list/seq	1263.12±35.49	1.66	23.54	1271.20±35.60	1.65	23.24
V\$AP2F	yes	439	708	list/seq	309.95±17.60	2.28	22.59	676.47±25.99	1.05	1.19
V\$NRF1	yes	157	313	list/seq	99.57±9.98	3.14	21.34	589.03±24.25	0.53	-11.40
V\$E2FF	yes	725	1239	list/seq	695.03±26.34	1.78	20.63	1732.38±41.53	0.72	-11.89
V\$MYRF	no	267	285	list/seq	89.46±9.46	3.19	20.62	117.76±10.85	2.42	15.37
OSMTEN	yes	174	226	list/seq	67.62±8.22	3.34	19.20	328.87±18.13	0.69	-5.70
V\$HDBP	yes	60	91	list/seq	15.57±3.95	5.84	18.99	186.21±13.64	0.49	-7.02

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.

Read Classification
Peak Finding
Peak Classification
Sequence Extraction
TFBS Overrepresentation
Definition of new TFBS
Download of Results

Find new Binding Sites in Peaks (CoreSearch)

Sequences for the 1000 best peaks were extracted for CoreSearch (sorted by lowest p-values, min. 80 bp, max. 3000 bp)
Average length of sequences is 194 bp

A motif was defined from 910 sequences
IUPAC consensus of the final motif: **NNSTGNTGACAGSN**
[re-value](#) of the final motif: 1.69

See the [complete CoreSearch result](#)

[Download sequence file](#) (272Kb)

[Save sequences](#) to project management

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

Solution parameters

Sequence file: Tbx20_chipseq_best_1000.seq (1000 sequences)
 Length of core: 6 bp
 Min. number of sequences: 750 sequences (75 % of 1000)
 Number of motif matches per sequence: at most one
 A priori frequency of nucleotides: determined from input sequences (A: 0.23, C: 0.27, G: 0.26, T: 0.23)
 Strand(s) searched: both strands
 Matrix similarity threshold: 0.80
 Maximum number of motifs: 1

Input Sequences

No.	Sequence Name	Sequence Description	Length
Show all sequences			
1	Region_2390	Region_2390 chr=17 start=39846450 end=39846796 str=+ bed_id=1288	339 bp
2	Region_1766	Region_1766 chr=11 start=109011644 end=109012100 str=+ bed_id=570	457 bp
3	Region_2393	Region_2393 chr=17 start=39847175 end=39848831 str=+ bed_id=1291	1657 bp
4	Region_889	Region_889 chr=5 start=146260991 end=146261359 str=+ bed_id=2364	369 bp
5	Region_484	Region_484 chr=3 start=5860624 end=5860823 str=+ bed_id=1886	200 bp

Motifs defined from subsets

5 motifs defined from 5 subsets

Motif	Re-value	IUPAC consensus
USs1_Tbx20_chipseq_c	0.84	.NSTGNTGACAGN.
USs2_Tbx20_chipseq_c	1.51	.NNTGNTGACAGSN
USs3_Tbx20_chipseq_c	1.31	.NNTGNTGACAGS.
USs4_Tbx20_chipseq_c	1.26	.NNTGNTGACAGN.
USs5_Tbx20_chipseq_c	2.20	.NNNTGACAGN.

Average similarity of motifs: 0.610

At least one motif match found in 975 of 1000 sequences.

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

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

GEMS Launcher: Edit user-defined matrix library

Matrix Library	
Current Status	View status of your personal matrix library
Modify Matrix Library	<input type="radio"/> Delete families <input type="radio"/> Delete matrices from families <input checked="" type="radio"/> Edit a family (family name, description) <input type="radio"/> Edit a matrix (matrix name, description, references) <input type="radio"/> Add a matrix/family by uploading a binary matrix library file <input type="button" value="Continue"/>
Matrix Subsets	Edit matrix subsets

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

User-defined Matrices

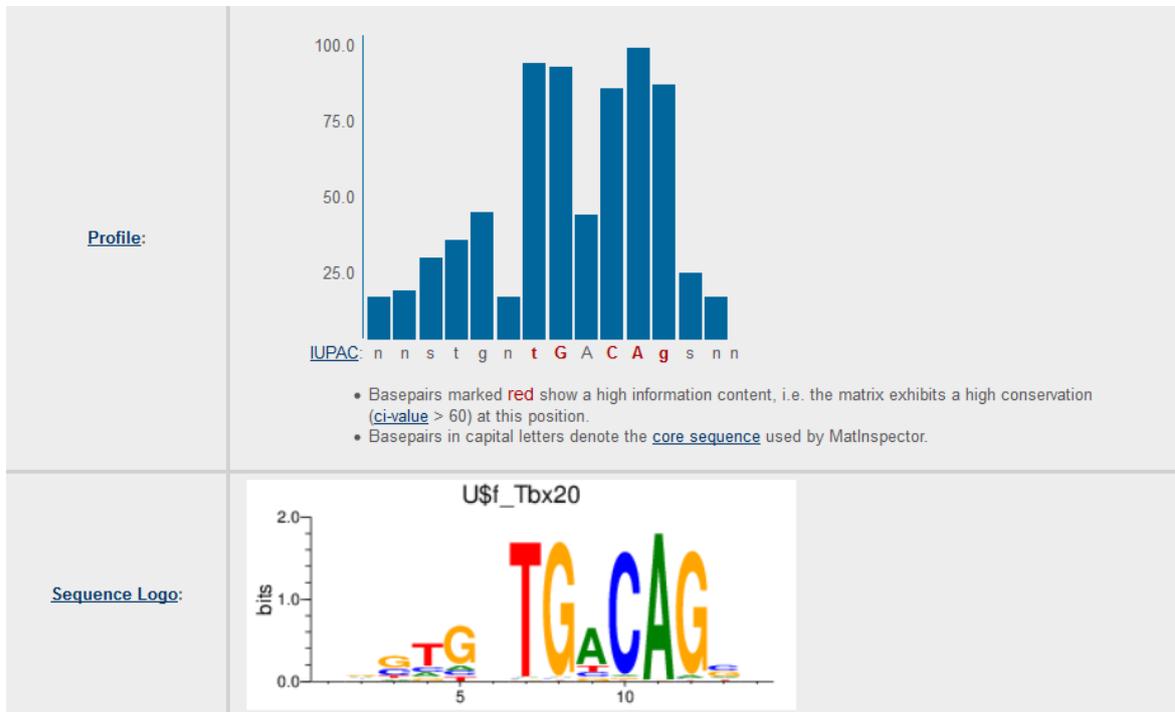
1 matrices in 1 families (User-defined Matrix Library Version 7.0)

Family	Family Information	Matrix Name	Information	Opt.
U\$Tbx20	created by CoreSearch	U\$f_Tbx20	created by CoreSearch	0.94

You'll see some statistics and the nucleotide distribution including IUPAC translation and consensus index for each position, which is a measure for conservation.

Matrix U\$f_Tbx20																																																																																																										
Matrix Name:	U\$f_Tbx20																																																																																																									
Description:	created by CoreSearch																																																																																																									
Family:	U\$Tbx20 (created by CoreSearch)																																																																																																									
References:	---																																																																																																									
Statistical Basis:	910 sequences																																																																																																									
Random Expectation (re-value):	1.69 matches per 1000 bp																																																																																																									
Promoter Matches:	0.0 % (vertebrate promoters)																																																																																																									
Optimized Matrix Threshold:	0.94																																																																																																									
Length:	15 bp																																																																																																									
Nucleotide Distribution Matrix:	<table border="1"> <thead> <tr> <th>Pos.</th> <th>1</th> <th>2</th> <th>3</th> <th>4</th> <th>5</th> <th>6</th> <th>7</th> <th>8</th> <th>9</th> <th>10</th> <th>11</th> <th>12</th> <th>13</th> <th>14</th> </tr> </thead> <tbody> <tr> <td>A</td> <td>253</td> <td>221</td> <td>71</td> <td>130</td> <td>92</td> <td>208</td> <td>11</td> <td>8</td> <td>632</td> <td>18</td> <td>902</td> <td>25</td> <td>123</td> <td>245</td> </tr> <tr> <td>C</td> <td>204</td> <td>180</td> <td>256</td> <td>135</td> <td>90</td> <td>261</td> <td>5</td> <td>8</td> <td>88</td> <td>863</td> <td>0</td> <td>9</td> <td>351</td> <td>233</td> </tr> <tr> <td>G</td> <td>272</td> <td>340</td> <td>459</td> <td>82</td> <td>645</td> <td>214</td> <td>6</td> <td>887</td> <td>68</td> <td>19</td> <td>7</td> <td>865</td> <td>333</td> <td>162</td> </tr> <tr> <td>T</td> <td>181</td> <td>169</td> <td>124</td> <td>563</td> <td>83</td> <td>227</td> <td>888</td> <td>7</td> <td>122</td> <td>10</td> <td>1</td> <td>11</td> <td>103</td> <td>270</td> </tr> <tr> <td>IUPAC</td> <td>N</td> <td>N</td> <td>S</td> <td>T</td> <td>G</td> <td>N</td> <td>T</td> <td>G</td> <td>A</td> <td>C</td> <td>A</td> <td>G</td> <td>S</td> <td>N</td> </tr> <tr> <td>Ci</td> <td>14.7</td> <td>16.4</td> <td>27.1</td> <td>33.2</td> <td>42.7</td> <td>14.1</td> <td>91.4</td> <td>91.0</td> <td>41.5</td> <td>84.0</td> <td>96.7</td> <td>84.7</td> <td>22.2</td> <td>14.9</td> </tr> </tbody> </table>	Pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	A	253	221	71	130	92	208	11	8	632	18	902	25	123	245	C	204	180	256	135	90	261	5	8	88	863	0	9	351	233	G	272	340	459	82	645	214	6	887	68	19	7	865	333	162	T	181	169	124	563	83	227	888	7	122	10	1	11	103	270	IUPAC	N	N	S	T	G	N	T	G	A	C	A	G	S	N	Ci	14.7	16.4	27.1	33.2	42.7	14.1	91.4	91.0	41.5	84.0	96.7	84.7	22.2	14.9
Pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14																																																																																												
A	253	221	71	130	92	208	11	8	632	18	902	25	123	245																																																																																												
C	204	180	256	135	90	261	5	8	88	863	0	9	351	233																																																																																												
G	272	340	459	82	645	214	6	887	68	19	7	865	333	162																																																																																												
T	181	169	124	563	83	227	888	7	122	10	1	11	103	270																																																																																												
IUPAC	N	N	S	T	G	N	T	G	A	C	A	G	S	N																																																																																												
Ci	14.7	16.4	27.1	33.2	42.7	14.1	91.4	91.0	41.5	84.0	96.7	84.7	22.2	14.9																																																																																												

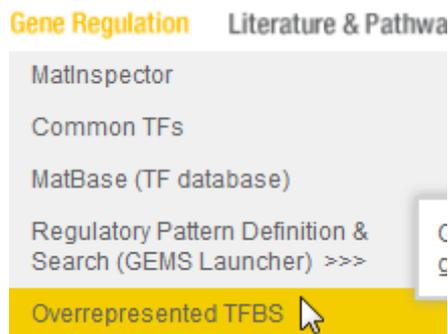
The conservation profile of the binding site definition is also shown in a column chart and as a sequence logo.



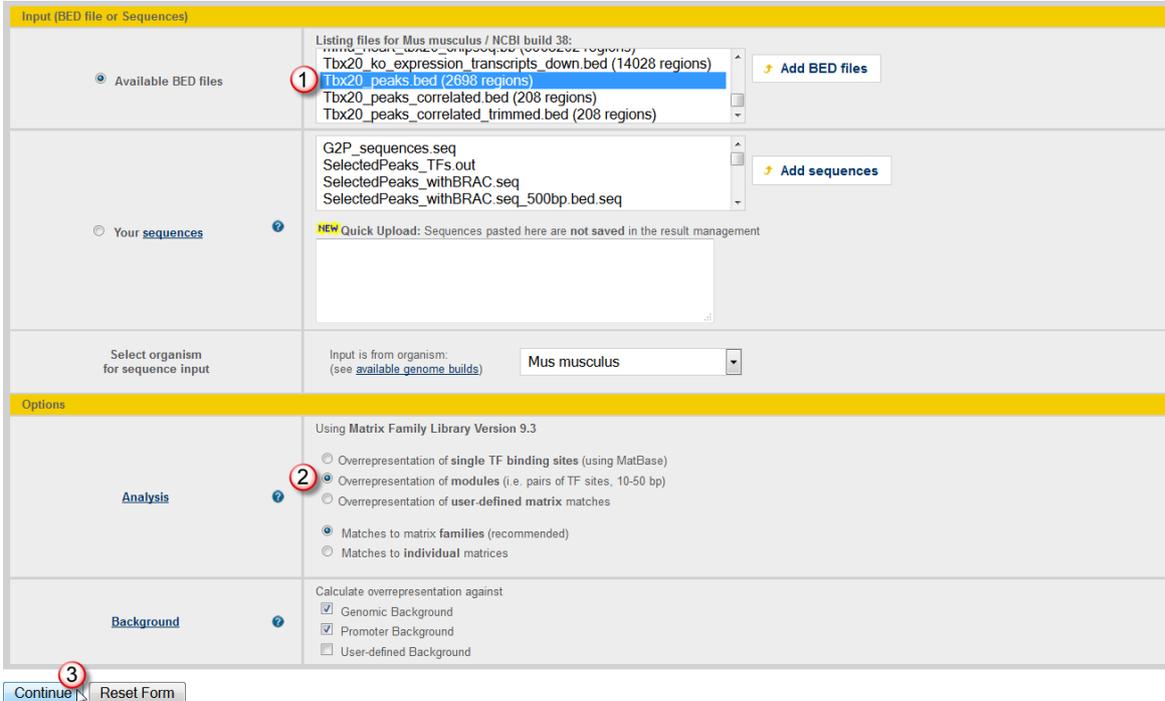
TFBS module overrepresentation

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.



The image shows a web interface for the TFBS overrepresentation analysis. It is divided into two main sections: 'Input (BED file or Sequences)' and 'Options'.

Input (BED file or Sequences):

- Available BED files:** A list of files for *Mus musculus* / NCBI build 38. The file 'Tbx20_peaks.bed (2698 regions)' is selected and highlighted. A red circle with the number '1' is next to it. There is an 'Add BED files' button.
- Your sequences:** A list of sequence files including 'G2P_sequences.seq', 'SelectedPeaks_TFs.out', 'SelectedPeaks_withBRAC.seq', and 'SelectedPeaks_withBRAC.seq_500bp.bed.seq'. There is an 'Add sequences' button and a 'NEW Quick Upload' note.
- Select organism for sequence input:** A dropdown menu set to 'Mus musculus'.

Options:

- Analysis:** Radio buttons for 'Overrepresentation of single TF binding sites (using MatBase)', 'Overrepresentation of modules (i.e. pairs of TF sites, 10-50 bp)', and 'Overrepresentation of user-defined matrix matches'. The 'Overrepresentation of modules' option is selected. A red circle with the number '2' is next to it.
- Background:** Checkboxes for 'Genomic Background', 'Promoter Background', and 'User-defined Background'. Both 'Genomic Background' and 'Promoter Background' are checked.

At the bottom, there are 'Continue' and 'Reset Form' buttons. A red circle with the number '3' is next to the 'Continue' button.

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.

Parameters

Partner in module search Search for modules where one of the partners is

V\$DECF
V\$BHLH
V\$BNCF
V\$BPTF
V\$BRAC
V\$BRN5

1 Note: you can check [MatBase \(Matrix Library 9.3\)](#) for specific search terms or see the [list of available weight matrices](#)

Strand-sepcificity check for strand-specific modules
i.e. same-strand modules (+/+ and -/-) from different-strand modules (+/- and -/+)

Output

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

Your email address
 Show result directly in browser window
 Send the URL of the result to:
Use the email option for long-running jobs, to avoid server-timeout messages
You may set a default email address by filling or modifying the 'email address' field on your [personal account page](#)

3

This is the start of the output list:

Listing of all Modules with V\$BRAC

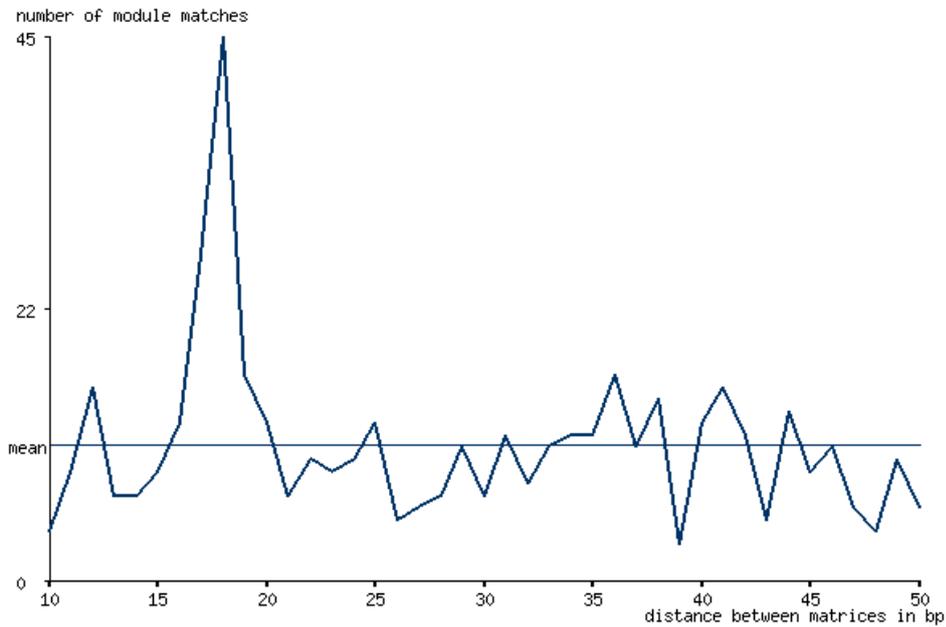
Modules with V\$BRAC	Distance Score	Prom. assoc. known	Nr. of Input Seq. with Match	Nr. of Matches in Input	Match details	Expected (genome) ± Std.dev.	Over representation (genome)	Z-Score (genome)	Expected (promoters) ± Std.dev.	Over representation (promoters)	Z-Score (promoters)
V\$BRAC-V\$TALE	2.161	no	490	879	list	184.63±13.58	4.76	51.08	150.68±12.27	5.83	59.30
V\$BRAC-V\$MYOD	2.802	no	355	843	list	188.02±13.71	4.48	47.74	201.72±14.20	4.18	45.13
V\$AP1R-V\$BRAC	2.024	no	472	908	list	249.56±15.79	3.64	41.66	230.86±15.19	3.93	44.55
V\$BRAC-V\$NF1F	4.847	no	286	459	list	84.76±9.21	5.42	40.60	84.39±9.19	5.44	40.73
V\$AP4R-V\$BRAC	2.142	no	142	235	list	31.77±5.64	7.40	35.97	41.23±6.42	5.70	30.10
V\$BRAC-V\$ZF11	1.858	no	156	216	list	27.51±5.24	7.85	35.84	40.40±6.36	5.35	27.55
V\$BRAC-V\$CTCF	3.152	yes	205	306	list	55.85±7.47	5.48	33.41	112.92±10.63	2.71	18.13
V\$AP2F-V\$BRAC	2.738	no	172	319	list	62.51±7.91	5.10	32.38	105.76±10.28	3.02	20.69
V\$BRAC-V\$EGRF	1.595	no	202	442	list	115.03±10.72	3.84	30.44	217.41±14.74	2.03	15.20
V\$BRAC-V\$P33F	3.862	no	237	476	list	129.60±11.38	3.67	30.39	121.81±11.03	3.91	32.05
V\$BRAC-V\$SP1F	2.244	no	236	391	list	99.45±9.97	3.93	29.19	186.04±13.64	2.10	14.99
V\$BRAC-V\$HAND	1.980	no	412	679	list	233.31±15.27	2.91	29.15	226.40±15.04	3.00	30.06
V\$BRAC-V\$NELR	3.513	no	283	433	list	119.61±10.93	3.62	28.61	116.49±10.79	3.72	29.28
V\$BRAC-V\$EBOX	2.193	no	255	445	list	126.43±11.24	3.52	28.29	145.64±12.07	3.06	24.77

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.

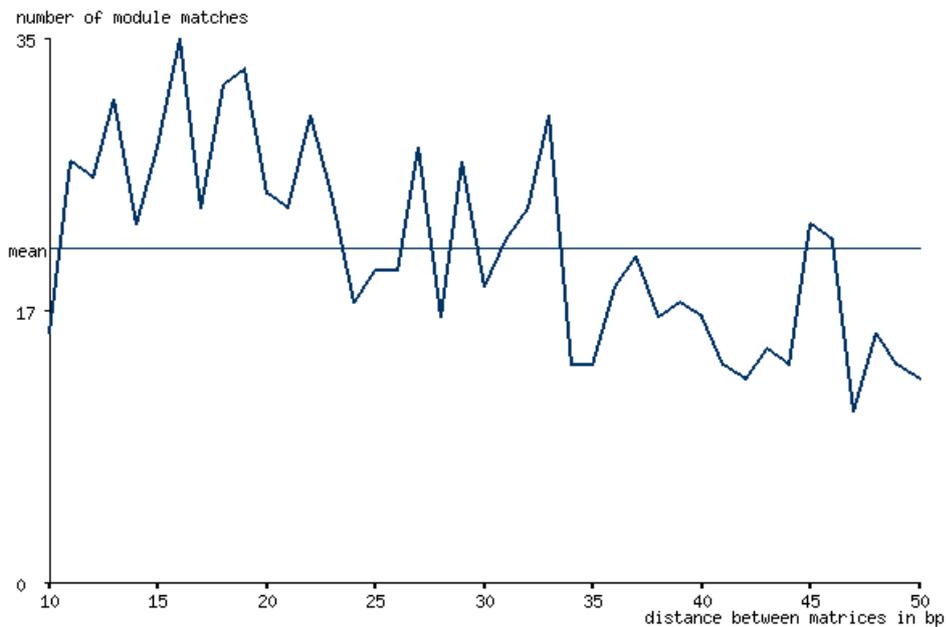
The distance score can be used for sorting module matches with one or a few preferred distances between the sites in the input sequences. A high score would indicate a strong distance preference.

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.

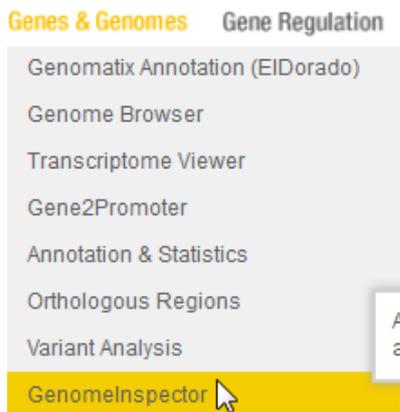
Integration of expression and ChIP-Seq data

Positional correlation of Tbx20 peaks with differentially expressed transcripts

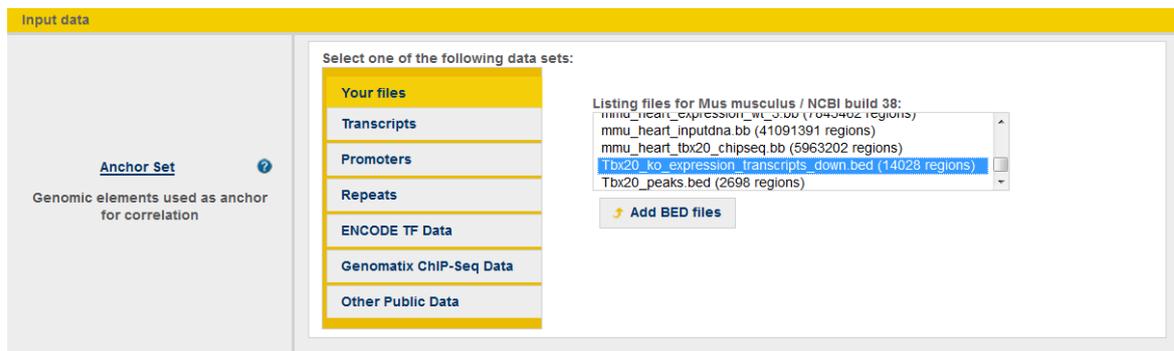
In the next step, we will predict which genes that are differentially expressed in Tbx20 knock-out mouse hearts are direct targets of Tbx20. For this, we will use the program GenomInspector and visualize the positional correlation of the starts of the down-regulated transcripts with the Tbx20 ChIP peaks.

GenomInspector 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 GenomInspector 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.

Partner Set(s) ?

to be checked for correlations to Anchor Set

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

Your files
Transcripts
Promoters
Repeats
ENCODE TF Data
Genomatix ChIP-Seq Data
Other Public Data

mmu_heart_expression_enr_06 (147472 regions)

mmu_heart_inputdna.bb (41091391 regions)

mmu_heart_tbx20_chipseq.bb (5963202 regions)

Tbx20_ko_expression_transcripts_down.bed (14028 regions)

Tbx20_peaks.bed (2698 regions)

(You can use shift/ctrl-keys to select multiple files)

[Add BED files](#)

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.

Output

Range and Elements ?

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

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

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

Graphics Options

Colors ?

Nucleotide Content ?

Result ?

[more...](#)

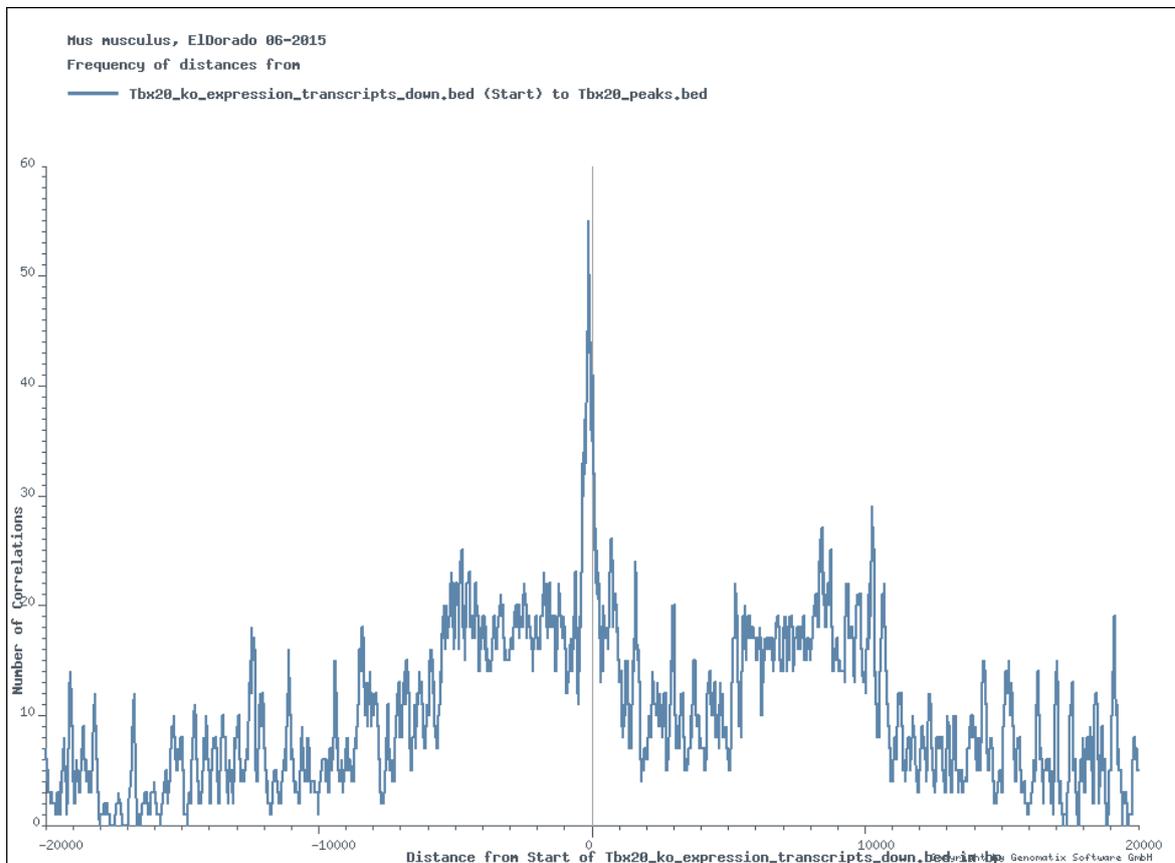
[more...](#)

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

[Start Analysis](#)

[Reset this form](#)

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*.

Continue to

view correlations as list

1 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 to bp distance (max. -20000 bp to 20000 bp)

4 **2** **3**

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

GenomeInspector: 635 correlations were found

Extracted Elements from Tbx20_ko_expression_transcripts_down.bed / Start with a correlation to Tbx20_peaks.bed within -6000 to 11000 bp						
Number	GenomeBrowser	Chr.	Begin	End	Strand	Bed Id / Score
Nr. 1	GenomeBrowser	chr1	3999403	4409266	(-)	GXT_26095811/XM_006495473/Rp1 / -2.04
Nr. 2	GenomeBrowser	chr1	23995939	24005640	(-)	GXT_13127351/NM_026503/1110058L19Rik / -1.34
Nr. 3	GenomeBrowser	chr1	23995968	24005598	(-)	GXT_13007139/AK003789/1110058L19Rik / -1.48
Nr. 4	GenomeBrowser	chr1	24002966	24005630	(-)	GXT_24302361/ENSMUST00000155767/1110058L19Rik / -1.79
Nr. 5	GenomeBrowser	chr1	52845044	52885337	(+)	GXT_24324887/ENSMUST00000161125/Hibch / -1.69
Nr. 6	GenomeBrowser	chr1	52845046	52920860	(+)	GXT_13033472/NM_146108/Hibch / -1.56
Nr. 7	GenomeBrowser	chr1	52845048	52920860	(+)	GXT_12942462/AK076038/Hibch / -1.56
Nr. 8	GenomeBrowser	chr1	75383556	75384975	(+)	GXT_24322105/ENSMUST00000146705/Speg / -1.51
Nr. 9	GenomeBrowser	chr1	75384700	75387948	(+)	GXT_24322106/ENSMUST00000125118/Speg / -1.39
Nr. 10	GenomeBrowser	chr1	75384828	75391923	(+)	GXT_23717585/ENSMUST00000132228/Speg / -1.32

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

Nr. 98	GenomeBrowser	chr2	132690283	132751055	(+)	GXT_23381210/NM_028637/1110034G24Rik / -2.18
Nr. 99	GenomeBrowser	chr2	146239879	146512004	(-)	GXT_25620894/ENSMUST00000109986/Ralgapa2 / -1.54
Nr. 100	GenomeBrowser	chr2	146239879	146512321	(-)	GXT_26122546/XR_374469/Ralgapa2 / -1.54

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.

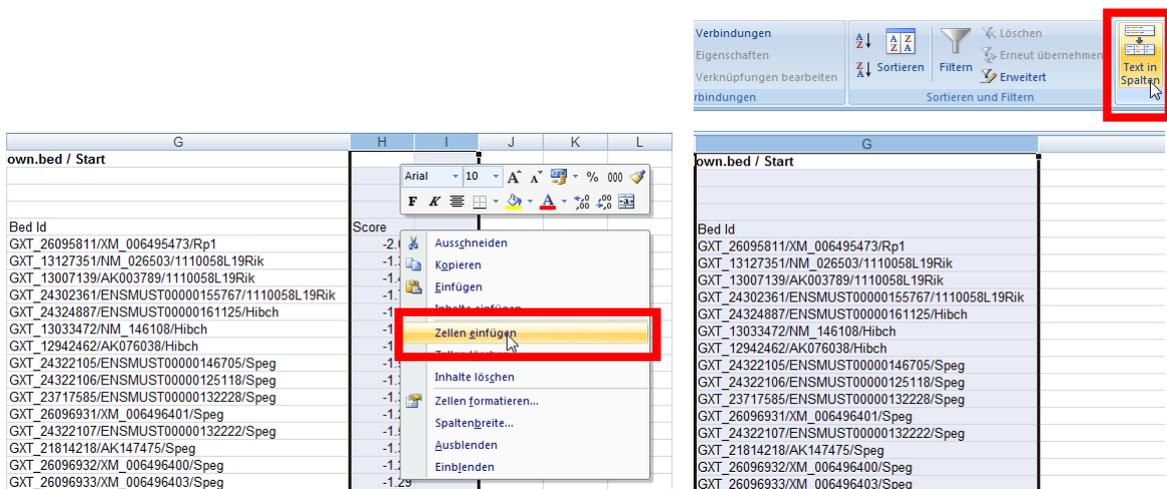
as

Extract table as

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.

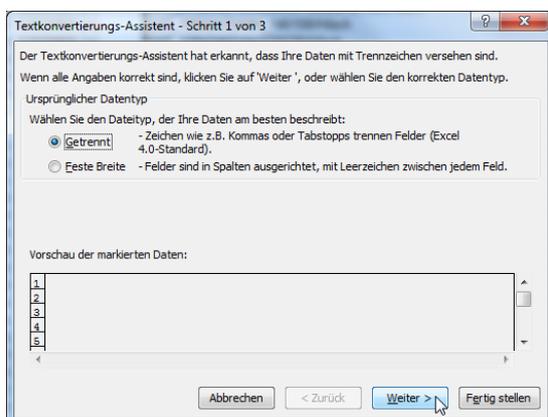
Extracted Elements from Tbx20_ko_expression_transcripts_down.bed / Start							
with a correlation to Tbx20_peaks.bed							
within -6000 to 11000 bp							
Number	GenomeBr	Chr.	Begin	End	Strand	Bed Id	Score
1	/cgi-bin//el	chr1	3999403	4409266	(-)	GXT_26095811/XM_006495473/Rp1	-2.04
2	/cgi-bin//el	chr1	23995939	24005640	(-)	GXT_13127351/NM_026503/1110058L19Rik	-1.34
3	/cgi-bin//el	chr1	23995968	24005598	(-)	GXT_13007139/AK003789/1110058L19Rik	-1.48
4	/cgi-bin//el	chr1	24002966	24005630	(-)	GXT_24302361/ENSMUST00000155767/1110058L19Rik	-1.79
5	/cgi-bin//el	chr1	52845044	52885337	(+)	GXT_24324887/ENSMUST00000161125/Hibch	-1.69
6	/cgi-bin//el	chr1	52845046	52920860	(+)	GXT_13033472/NM_146108/Hibch	-1.56
7	/cgi-bin//el	chr1	52845048	52920860	(+)	GXT_12942462/AK076038/Hibch	-1.56
8	/cgi-bin//el	chr1	75383556	75384975	(+)	GXT_24322105/ENSMUST00000146705/Spieg	-1.51
9	/cgi-bin//el	chr1	75384700	75387948	(+)	GXT_24322106/ENSMUST00000125118/Spieg	-1.39
10	/cgi-bin//el	chr1	75384828	75391923	(+)	GXT_23717585/ENSMUST00000132228/Spieg	-1.32
11	/cgi-bin//el	chr1	75385158	75432320	(+)	GXT_26096931/XM_006496401/Spieg	-1.29
12	/cgi-bin//el	chr1	75385512	75389104	(+)	GXT_24322107/ENSMUST00000132222/Spieg	-1.51
13	/cgi-bin//el	chr1	75385610	75432304	(+)	GXT_21814218/AK147475/Spieg	-1.31
14	/cgi-bin//el	chr1	75385676	75432320	(+)	GXT_26096932/XM_006496400/Spieg	-1.29
15	/cgi-bin//el	chr1	75398588	75432320	(+)	GXT_26096933/XM_006496403/Spieg	-1.29

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.

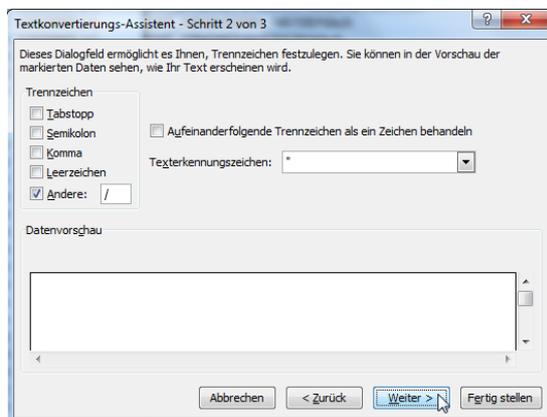


The screenshot shows the 'Text in Spalten' (Text to Columns) wizard in Microsoft Excel. The wizard is set to 'Text' and 'Delimited' options, with the 'Separator' set to a slash (/). The 'Advanced' options are also visible. Below the wizard, the resulting data table is shown, where the 'Bed Id' column has been split into multiple columns based on the slash separator.

own.bed / Start	Score	Bed Id				
GXT_26095811/XM_006495473/Rp1	-2.04	GXT_26095811	XM_006495473	Rp1		
GXT_13127351/NM_026503/1110058L19Rik	-1.34	GXT_13127351	NM_026503	1110058L19Rik		
GXT_13007139/AK003789/1110058L19Rik	-1.48	GXT_13007139	AK003789	1110058L19Rik		
GXT_24302361/ENSMUST00000155767/1110058L19Rik	-1.79	GXT_24302361	ENSMUST00000155767	1110058L19Rik		
GXT_24324887/ENSMUST00000161125/Hibch	-1.69	GXT_24324887	ENSMUST00000161125	Hibch		
GXT_13033472/NM_146108/Hibch	-1.56	GXT_13033472	NM_146108	Hibch		
GXT_12942462/AK076038/Hibch	-1.56	GXT_12942462	AK076038	Hibch		
GXT_24322105/ENSMUST00000146705/Spieg	-1.51	GXT_24322105	ENSMUST00000146705	Spieg		
GXT_24322106/ENSMUST00000125118/Spieg	-1.39	GXT_24322106	ENSMUST00000125118	Spieg		
GXT_23717585/ENSMUST00000132228/Spieg	-1.32	GXT_23717585	ENSMUST00000132228	Spieg		
GXT_26096931/XM_006496401/Spieg	-1.29	GXT_26096931	XM_006496401	Spieg		
GXT_24322107/ENSMUST00000132222/Spieg	-1.51	GXT_24322107	ENSMUST00000132222	Spieg		
GXT_21814218/AK147475/Spieg	-1.31	GXT_21814218	AK147475	Spieg		
GXT_26096932/XM_006496400/Spieg	-1.29	GXT_26096932	XM_006496400	Spieg		
GXT_26096933/XM_006496403/Spieg	-1.29	GXT_26096933	XM_006496403	Spieg		



The screenshot shows the 'Textkonvertierungs-Assistent - Schritt 1 von 3' dialog box. The 'Ursprünglicher Datentyp' is 'Getrennt'. The 'Vorschau der markierten Daten' shows the first five rows of the data table.



The screenshot shows the 'Textkonvertierungs-Assistent - Schritt 2 von 3' dialog box. The 'Trennzeichen' (Separator) is set to a slash (/). The 'Texterkennungszeichen' (Text marker) is set to an asterisk (*). The 'Vorschau' (Preview) shows the resulting data table.

You should end up with a structure like this:

Extracted Elements from Tbx20_ko_expression_transcripts_down.bed / Start
with a correlation to Tbx20_peaks.bed
within -6000 to 11000 bp

Number	GenomeBr	Chr.	Begin	End	Strand	Bed Id			Score
1	/cgi-bin//el	chr1	3999403	4409266	(-)	GXT_26095811	XM_006495473	Rp1	-2.04
2	/cgi-bin//el	chr1	23995939	24005640	(-)	GXT_13127351	NM_026503	1110058L19Rik	-1.34
3	/cgi-bin//el	chr1	23995968	24005598	(-)	GXT_13007139	AK003789	1110058L19Rik	-1.48
4	/cgi-bin//el	chr1	24002966	24005630	(-)	GXT_24302361	ENSMUST00000155767	1110058L19Rik	-1.79
5	/cgi-bin//el	chr1	52845044	52885337	(+)	GXT_24324887	ENSMUST00000161125	Hibch	-1.69
6	/cgi-bin//el	chr1	52845046	52920860	(+)	GXT_13033472	NM_146108	Hibch	-1.56
7	/cgi-bin//el	chr1	52845048	52920860	(+)	GXT_12942462	AK076038	Hibch	-1.56
8	/cgi-bin//el	chr1	75383566	75384975	(+)	GXT_24322105	ENSMUST00000146705	Speg	-1.51
9	/cgi-bin//el	chr1	75384700	75387948	(+)	GXT_24322106	ENSMUST00000125118	Speg	-1.39
10	/cgi-bin//el	chr1	75384828	75391923	(+)	GXT_23717585	ENSMUST00000132228	Speg	-1.32
11	/cgi-bin//el	chr1	75385158	75432320	(+)	GXT_26096931	XM_006496401	Speg	-1.29
12	/cgi-bin//el	chr1	75385512	75389104	(+)	GXT_24322107	ENSMUST00000132222	Speg	-1.51
13	/cgi-bin//el	chr1	75385610	75432304	(+)	GXT_21814218	AK147475	Speg	-1.31
14	/cgi-bin//el	chr1	75385676	75432320	(+)	GXT_26096932	XM_006496400	Speg	-1.29
15	/cgi-bin//el	chr1	75398588	75432320	(+)	GXT_26096933	XM_006496403	Speg	-1.29

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



Genomatix Pathway System (GePS)

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

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

Characterization of gene sets

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

Co-cited genes for one gene

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

Co-cited genes for one term

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

Pathways for one gene

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

Browse human pathways

Browse, search and load canonical human pathways.

Build networks from scratch

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

Extracted Elements from Tbx20_ko_expression_transcripts_down.bed / Start
with a correlation to Tbx20_peaks.bed
within -6000 to 11000 bp

Number	GenomeBr Chr.	Begin	End	Strand	Bed Id		Score
1	/cgi-bin//el chr1	3999403	4409266	(-)	GXT_26095811	XM_006495473	Rp1 -2.04
2	/cgi-bin//el chr1	23995939	24005640	(-)	GXT_13127351	NM_026503	1110058L19Rik -1.34
3	/cgi-bin//el chr1	23995968	24005598	(-)	GXT_13007139	AK003789	1110058L19Rik -1.48
4	/cgi-bin//el chr1	24002966	24005630	(-)	GXT_24302361	ENSMUST00000155767	1110058L19Rik -1.79
5	/cgi-bin//el chr1	52845044	52885337	(+)	GXT_24324887	ENSMUST00000161125	Hibch -1.69
6	/cgi-bin//el chr1	52845046	52920860	(+)	GXT_13033472	NM_146108	Hibch -1.56
7	/cgi-bin//el chr1	52845048	52920860	(+)	GXT_12942462	AK076038	Hibch -1.56
8	/cgi-bin//el chr1	75383556	75384975	(+)	GXT_24322105	ENSMUST00000146705	Speg -1.51
9	/cgi-bin//el chr1	75384700	75387948	(+)	GXT_24322106	ENSMUST00000125118	Speg -1.39
10	/cgi-bin//el chr1	75384828	75391923	(+)	GXT_23717585	ENSMUST00000132228	Speg -1.32
11	/cgi-bin//el chr1	75385158	75432320	(+)	GXT_26096931	XM_006496401	Speg -1.29
12	/cgi-bin//el chr1	75385512	75389104	(+)	GXT_24322107	ENSMUST00000132222	Speg -1.51
13	/cgi-bin//el chr1	75385610	75432304	(+)	GXT_21814218	AK147475	Speg -1.31
14	/cgi-bin//el chr1	75385676	75432320	(+)	GXT_26096932	XM_006496400	Speg -1.29
15	/cgi-bin//el chr1	75398588	75432320	(+)	GXT_26096933	XM_006496403	Speg -1.29

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.

Parameters

Specify what kind of gene keywords you will provide:

Entrez and/or Ensembl Gene IDs 1
 Transcript Accession Numbers

Gene Symbols/Names
 Affymetrix Probe Set IDs

Paste a list of gene keywords...

AK162500
 AK165865
 AK140152
 ENSMUST00000082405 2
 AK131579
 AK141672
 AK131599

or upload a [text file](#) containing gene keywords, optionally with corresponding expression values.

Keine Datei ausgewählt.

OR

[Use example gene set](#)

"Inflammation in H.sapiens"

The example data set is from a microarray analysis of Systemic Inflammation in Humans (Calvano et al (2005) Nature 437,1032-7; PMID: 16136080).

Gene expression changes relative to t=0 are displayed at 5 timepoints (2,4,6,9 and 24 hours) after inoculation with bacterial endotoxin.

[Organism](#)

3

Some accession numbers will not be mapped to a gene ID; ignore the warning the program gives you, and proceed with the analysis.

GeneRanker couldn't map the following input keywords to a gene of the selected organism and therefore they won't be used in the analysis!

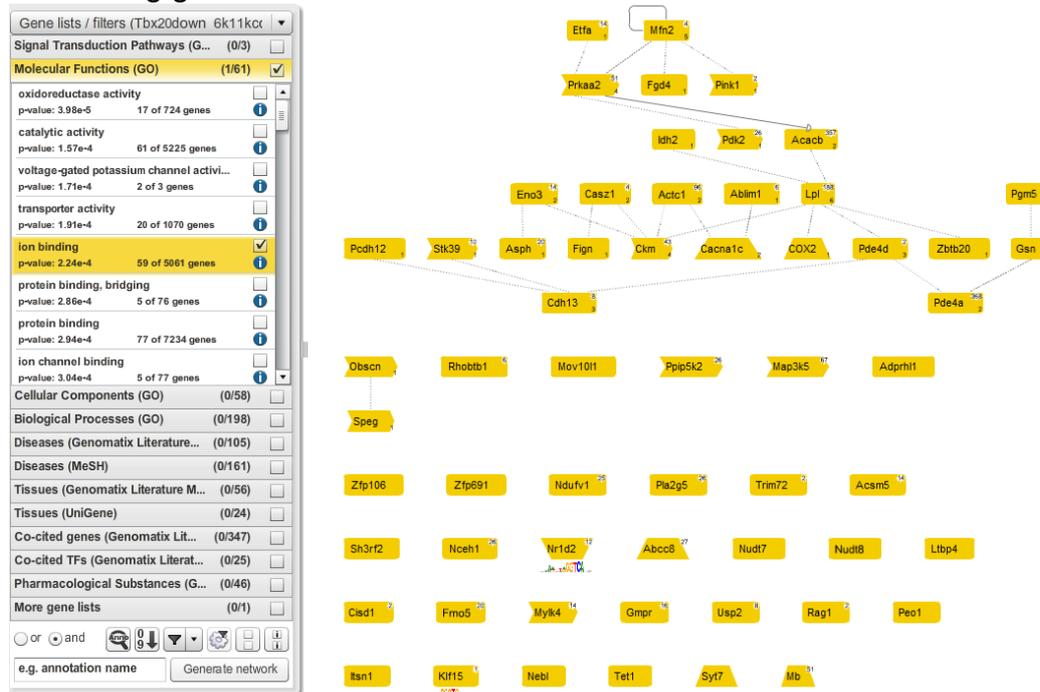
AK155508, AK084991, AK140265, AK139000, AK137643, AK156495, AK136371, AK084726, AK131599, ENSMUST00000082396, AK142161, AK164731, AK140152, AK040421, AK153841, AK141672, AK076583, AK156840

[Proceed with GeneRanker analysis](#)

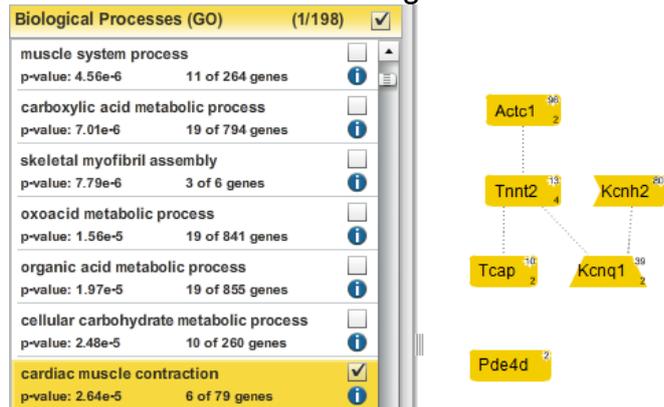
A total of 190 genes are found in this way. Overrepresented terms include *ion binding* in GO: Molecular Function, *cardiac muscle contraction* in GO: Biological Processes, and *cardiomyopathies* in Diseases. The pathway graphs below use the hierarchical layout, which you can activate with the leftmost *Layout* button in the lower control bar:



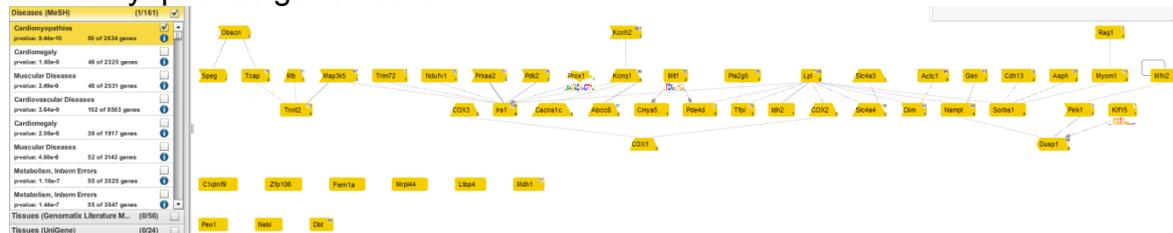
Ion binding gene network:



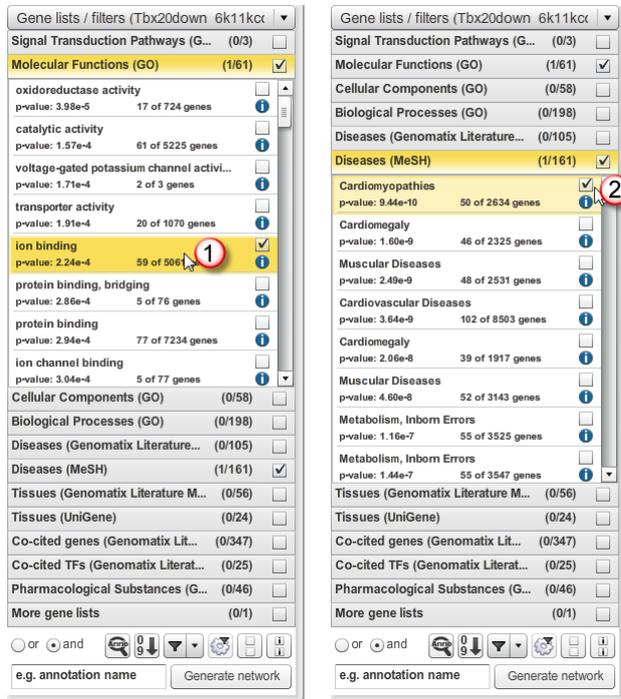
Cardiac muscle contraction gene network:



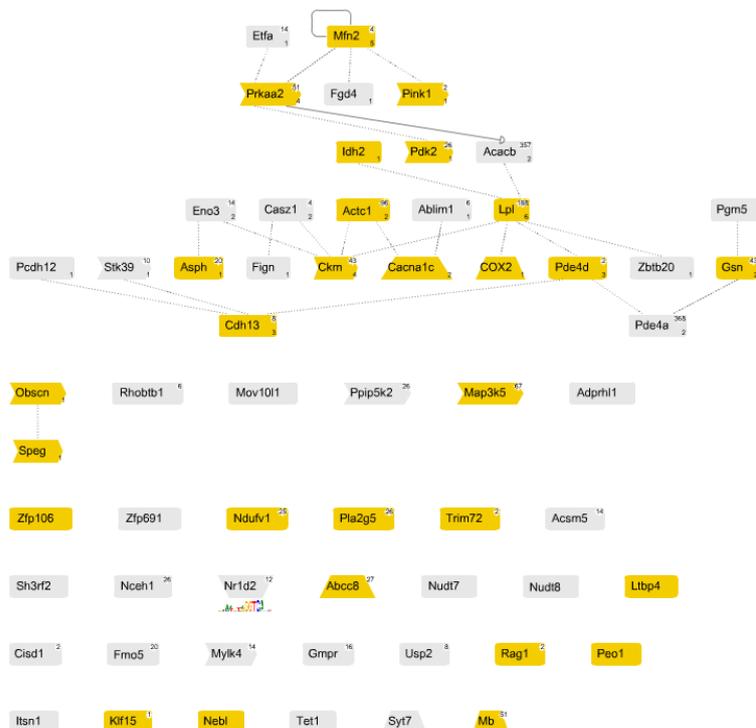
Cardiomyopathies gene network:



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 GenomInspector 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 GenomInspector 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 to bp distance (max. -20000 bp to 20000 bp)

208 correlated peaks are found.

GenomInspector: 208 correlations were found

Extracted Elements from Tbx20_peaks.bed with a correlation to Tbx20_ko_expression_transcripts_down.bed / Start within -6000 to 11000 bp						
Number	GenomeBrowser	Chr.	Begin	End	Strand	Bed Id / Score
Nr. 1	GenomeBrowser	chr1	4412567	4412753	(+)	1 / 9.12e-13
Nr. 2	GenomeBrowser	chr1	24010974	24011102	(+)	12 / 1.01e-06
Nr. 3	GenomeBrowser	chr1	52844928	52845037	(+)	40 / 1.66e-07
Nr. 4	GenomeBrowser	chr1	75393329	75393513	(+)	70 / 1.31e-10
Nr. 5	GenomeBrowser	chr1	75549351	75549552	(+)	71 / 1.47e-14
Nr. 6	GenomeBrowser	chr1	79776017	79776139	(+)	77 / 2.95e-05
Nr. 7	GenomeBrowser	chr1	82291491	82291596	(+)	78 / 3.87e-06
Nr. 8	GenomeBrowser	chr1	97761621	97761790	(+)	93 / 8.83e-09
Nr. 9	GenomeBrowser	chr1	118479426	118479585	(+)	108 / 2.31e-10
Nr. 10	GenomeBrowser	chr1	118481924	118482088	(+)	109 / 1.1e-08
Nr. 11	GenomeBrowser	chr1	135850907	135851029	(+)	137 / 1.16e-05
Nr. 12	GenomeBrowser	chr1	155234365	155234481	(+)	157 / 7.66e-07

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

Nr. 100	GenomeBrowser	chr8	68908384	68908503	(+)	2793 / 2.95e-05
---------	-------------------------------	------	----------	----------	-----	-----------------

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.

as

Extract table as

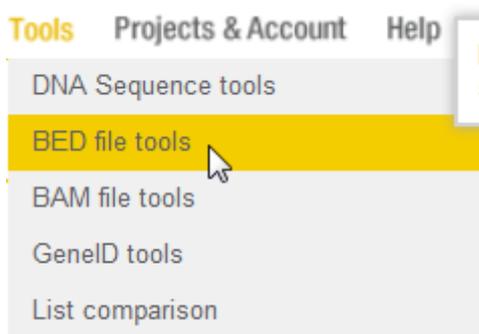
Save selected BED file as

to project

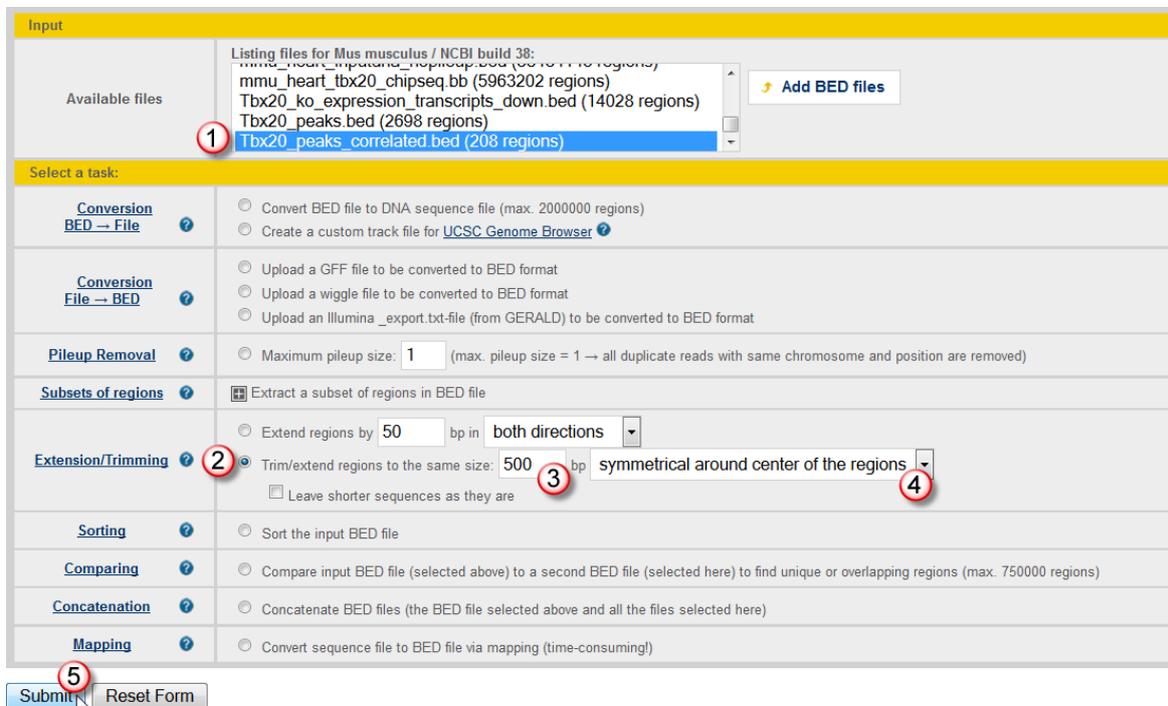
Trimming and conversion to sequence

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 52845232 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

to project

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.

Input

Available files

Listing files for Mus musculus / NCBI build 38:

- trmm_peaks_tbx20_trimseq.bed (200202 regions)
- Tbx20_ko_expression_transcripts_down.bed (14028 regions)
- Tbx20_peaks.bed (2698 regions)
- Tbx20_peaks_correlated.bed (208 regions)
- Tbx20_peaks_correlated_trimmed.bed (208 regions)**

Select a task:

Conversion
BED → File

Create a custom track file for [UCSC Genome Browser](#)

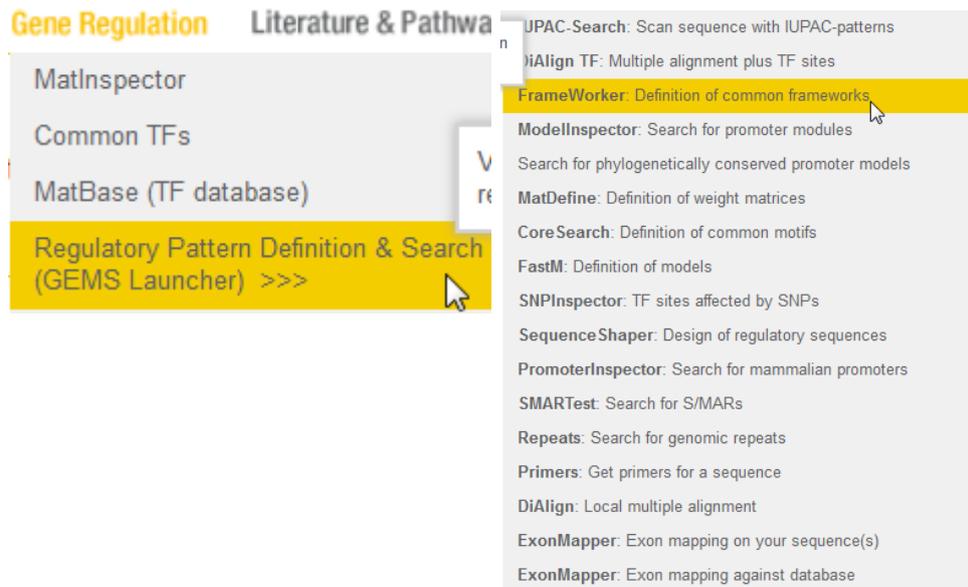
Save selected sequence file as

to project

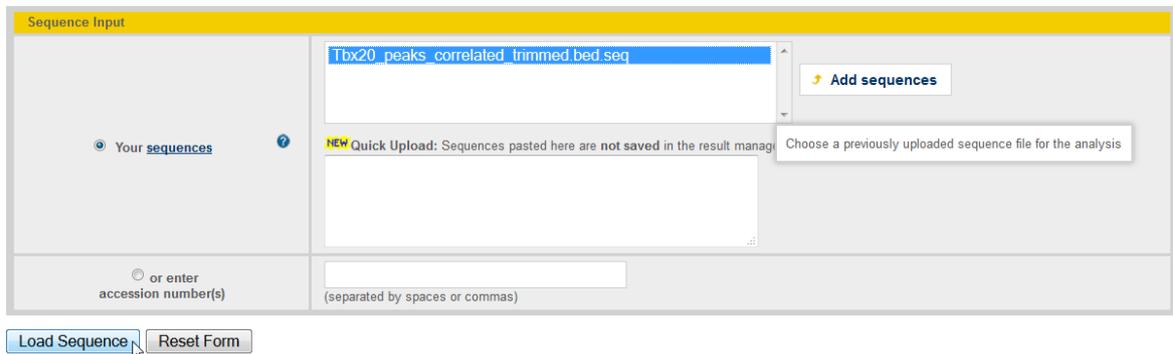
FrameWorker: common TFBS patterns

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*:

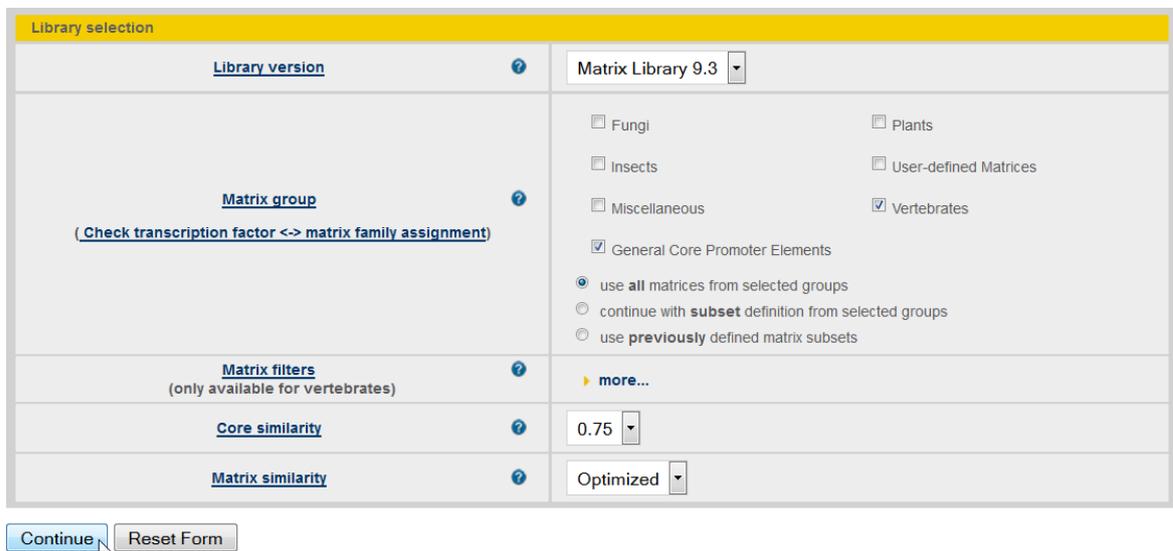


Select the saved sequence file in the list and continue by clicking the *Load Sequence* button.



The 'Sequence Input' form features a yellow header. On the left, there is a radio button labeled 'Your sequences'. The main area contains a text input field with the text 'Tbx20_peaks_correlated_trimmed_bed_seq'. To the right of this field is an 'Add sequences' button and a tooltip that says 'Choose a previously uploaded sequence file for the analysis'. Below the text input is a larger empty text area. At the bottom left, there is a radio button labeled 'or enter accession number(s)' and a text input field with the placeholder '(separated by spaces or commas)'. At the bottom of the form are two buttons: 'Load Sequence' and 'Reset Form'.

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.



The 'Library selection' form has a yellow header. It is divided into several sections. The top section is 'Library version' with a dropdown menu set to 'Matrix Library 9.3'. The 'Matrix group' section includes a link '(Check transcription factor <-> matrix family assignment)' and a list of checkboxes: 'Fungi', 'Insects', 'Miscellaneous', 'General Core Promoter Elements', 'Plants', 'User-defined Matrices', and 'Vertebrates'. The 'General Core Promoter Elements' checkbox is checked. Below this is a radio button selection: 'use all matrices from selected groups' (selected), 'continue with subset definition from selected groups', and 'use previously defined matrix subsets'. The 'Matrix filters' section is noted as '(only available for vertebrates)' and has a 'more...' link. The 'Core similarity' section has a dropdown menu set to '0.75'. The 'Matrix similarity' section has a dropdown menu set to 'Optimized'. At the bottom are 'Continue' and 'Reset Form' buttons.

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.

FrameWorker Parameters	
Quorum constraint for framework	Minimum number of input sequences to contain a framework: <input type="text" value="7 of 208 (3%)"/> 1 of input sequences
Sequence constraints <small>NEW</small>	Mandatory sequences (sequences that must contain framework, max. 10): <input type="text" value="Region_1"/> <input type="text" value="Region_2"/> <input type="text" value="Region_3"/> <input type="text" value="Region_4"/> <input type="text" value="Region_5"/>
Distance constraints for framework	Maximum distance VARIANCE between two elements: <input type="text" value="20"/> 2 (max: 100) Distance between two elements: min. <input type="text" value="5"/> max. <input type="text" value="200"/> (max: 500) <input type="checkbox"/> Do restricted model search <small>(FrameWorker lists more specific models where distance variations are as small as possible)</small>
Element constraints	Number of elements in models: min. <input type="text" value="2"/> max. <input type="text" value="6"/> <input type="checkbox"/> Show intermediate models (else only the longest models are shown) <hr/> Mandatory elements for models (max. 5): <input type="text" value="V\$DCDF"/> <input type="text" value="V\$BHLH"/> <input type="text" value="V\$BNCF"/> <input type="text" value="V\$BPTF"/> <input type="text" value="V\$BRAC"/> 3 <input type="text" value="V\$BRN5"/> Combination of mandatory elements: <input checked="" type="radio"/> ALL selected elements must be present in model <input type="radio"/> ONE of the selected elements must be present in model
Output Options	more...
p-values	<input type="checkbox"/> Determine p-values of models
Your email address	<input checked="" type="radio"/> Show result directly in browser window <input type="radio"/> Send the URL of the result to <input type="text" value="courses@genomatix.de"/> <small>Use the email option for long-running jobs, to avoid server-timeout messages</small> <small>You may set a default email address by filling or modifying the 'email address' field on your personal account page</small>
Result	
Result name (optional)	<input type="text"/> <small>(special characters like "#\$%&+./:;<=>?@ not allowed)</small>
<div style="display: flex; justify-content: space-between;"> 4 Start FrameWorker Reset Form </div>	

One model with 4 elements is found. Click the link to jump to the description.

Graphical View **Model Overview** Model Details Common Elements

Overview: Models common to at least 7 sequences (3%)

Models consisting of	# of different models containing mandatory element(s)	# of models checked
single element	417 common elements found	-
2 elements	2544 models found	40586 models checked
3 elements	128 models found	5293 models checked
4 elements	1 model found	194 models checked
5 elements	0 models found	3 models checked

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.

Graphical View **Model Overview** Model Details Common Elements

1 model with 4 elements:

Model "model_4el_1": (click opens graphics)

Save this model as 2

Element	Strand	Matrix sim.	Distance to next element	Common to
1 V\$SORY	-	Optimized (min. 0.71)	5 - 12 bp	8 matches in 7 seq. (3%), 7 non-overlapping
2 V\$SORY	+	Optimized (min. 0.73)	83 - 101 bp	
V\$TALE	+	Optimized (min. 0.89)	5 - 10 bp	
V\$BRAC	-	Optimized (min. 0.90)	---	

Check all Models Uncheck all Models Invert Selection

Save selected models 3 with the prefix as model subset

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.

FrameWorker: Definition of common frameworks

Find common transcription factor frameworks (common TFs in a specific order with conserved distances) working on Tbx20_peaks_correlated_trimmed.bed.seq (208 sequences, 103967 bp)

The model(s)

- User-defined/SORY_SORY_TALE_BRAC.model

were saved in your [Personal Model Library](#)

To search for matches to your new model(s) you can use the the following task, selecting user-defined models and subset selection:

[ModellInspector: Search for promoter modules](#)

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*.

or Database Input

1 more...

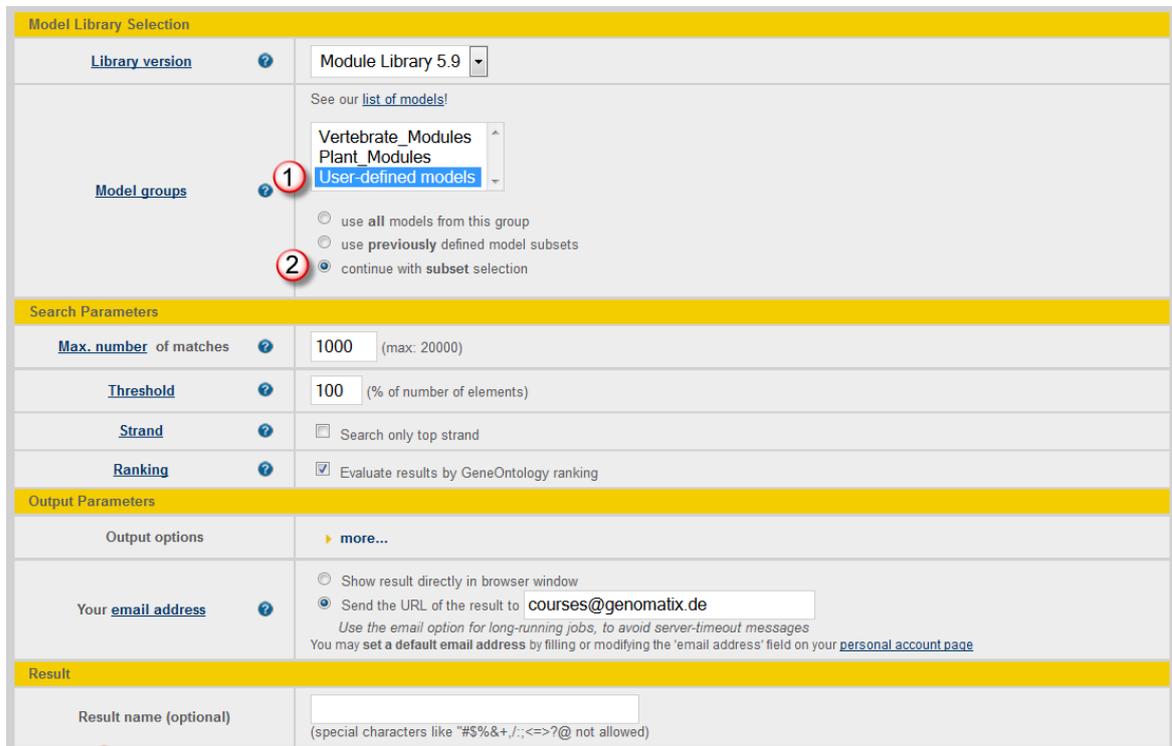
Genomatix Promoter Database: Promoters of annotated genes from EIDorado 06-2015:

Human Promoters Mouse Promoters Rat Promoters

2 Genomatix Promoter Database: Promoters of all genes from EIDorado 06-2015:

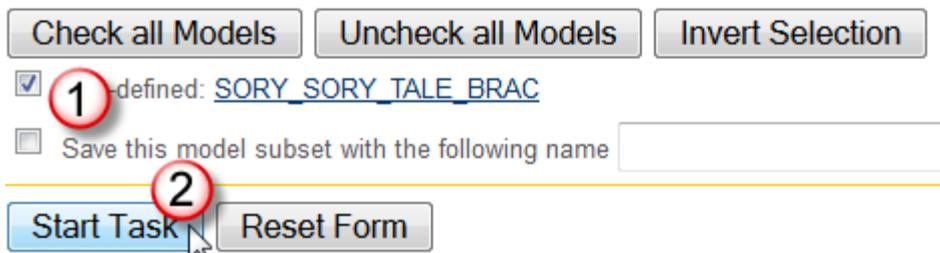
<input type="checkbox"/> Anopheles Promoters	<input type="checkbox"/> Dog Promoters	<input type="checkbox"/> Neurospora Crassa Promoters	<input type="checkbox"/> Rice Promoters
<input type="checkbox"/> Arabidopsis Promoters	<input type="checkbox"/> Drosophila Promoters	<input type="checkbox"/> Opossum Promoters	<input type="checkbox"/> Sorghum Promoters
<input type="checkbox"/> Baker's Yeast Promoters	<input type="checkbox"/> Fission Yeast Promoters	<input type="checkbox"/> Pig Promoters	<input type="checkbox"/> Soybean Promoters
<input type="checkbox"/> C.elegans Promoters	<input type="checkbox"/> Frog Promoters	<input type="checkbox"/> Plasmodium Promoters	<input type="checkbox"/> Wine Grape Promoters
<input type="checkbox"/> Carpenter Ant Promoters	<input type="checkbox"/> Honey Bee Promoters	<input type="checkbox"/> Platypus Promoters	<input type="checkbox"/> Zebra Finch Promoters
<input type="checkbox"/> Chicken Promoters	<input type="checkbox"/> Horse Promoters	<input type="checkbox"/> Poplar Promoters	<input type="checkbox"/> Zebrafish Promoters
<input type="checkbox"/> Chimpanzee Promoters	<input type="checkbox"/> Human Promoters (all)	<input type="checkbox"/> Rabbit Promoters	
<input type="checkbox"/> Corn Promoters	<input type="checkbox"/> Jumping Ant Promoters	<input type="checkbox"/> Rat Promoters (all)	
<input type="checkbox"/> Cow Promoters	<input type="checkbox"/> Mouse Promoters (all)	<input type="checkbox"/> Rhesus Macaque Promoters	

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:



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:

Your submitted jobs				
Job-ID	Task	State	Submitted at	Remove job
488	ModelInspector: Search for promoter modules	RUNNING	2015-07-16T14:41:58	Remove job

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.

Output overview of ModelInspector matches (376 matches)

go to: [[Output overview](#)] [[Detailed output](#)] [[Statistics](#)]

ModelInspector Release 5.6.8.7 Nov 2013

Thu Jul 16 14:41:58 2015

Solution parameters:

Sequence file: Mouse Promoters
 Models: User-defined/SORY_SORY_TALE_BRAC.model
 Matrix library: Matrix Family Library Version 9.3
 Strand(s) searched: both strands
 Threshold for number of elements: 100.0 % (4 of 4 elements)
 Output sorted by: match positions on the sequences
 Output filtered for: sequences with at least 1 different model matches
 Maximum number of matches: 1000

Match List	Evaluation	Further Analysis			
Match List:					
Sequence	Model Name	Position	Strand	Genomic Position	Select Match
GXP_5056712 [GXP_5056712] (1 - 601) Gpr143, GXL_49, GeneID: 18241, Mus musculus chr. X G protein-coupled receptor 143	SORY_SORY_TALE_BRAC	246 - 120	(-)	chrX: 152797998 - 152798124 (-)	<input checked="" type="checkbox"/>
GXP_5039589 [GXP_5039589] (1 - 856) Il1a, GXL_2277, GeneID: 16175, Mus musculus chr. 2 interleukin 1 alpha	SORY_SORY_TALE_BRAC	825 - 706	(-)	chr2: 129309031 - 129309150 (+)	<input checked="" type="checkbox"/>
GXP_5039672 [GXP_5039672] (1 - 601) Slc23a2, GXL_2326, GeneID: 54338, Mus musculus chr. 2 solute carrier family 23 (nucleobase transporters), member 2	SORY_SORY_TALE_BRAC	268 - 400	(+)	chr2: 132103666 - 132103798 (-)	<input checked="" type="checkbox"/>

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.

Annotation Type: Molecular Functions (GO)

Number of input genes mapped to GO-Terms: 322

Number of significant GO-Terms: 111

Show/Hide column GO-Term Show all columns Show default columns

Displayed rows: 1-10 / 111

Page 1 of 12

Results per page 10

GO-Term	GO-Term id	P-value	# Genes (observed)	# Genes (expected)	# Genes (total)
binding	GO:0005488	2.24e-06	202	160.84	11365
cAMP-dependent protein kinase regulator activity	GO:0008603	2.64e-06	4	0.11	8
ion binding	GO:0043167	1.05e-05	105	71.62	5061
small molecule binding	GO:0030244	2.09e-05	38	33.77	2300
transferase activity	GO:0016740	3.53e-05	51	28.57	2019
natural killer cell lectin-like receptor binding	GO:0046703	4.75e-05	4	0.21	15
adenyl ribonucleotide binding	GO:0032559	7.45e-05	38	19.61	1386
adenyl nucleotide binding	GO:0030554	8.28e-05	38	19.71	1393
nucleotide binding	GO:0000166	1.09e-04	51	29.86	2110
nucleoside phosphate binding	GO:1901265	1.09e-04	51	29.86	2110

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.

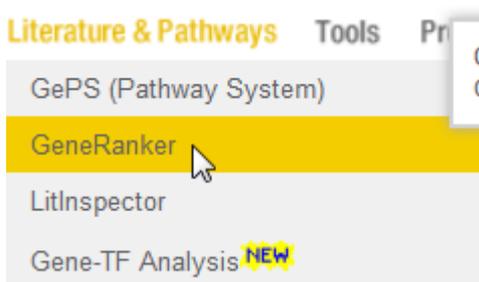
Extraction Options	
Sequence Extraction	<input checked="" type="radio"/> Selected elements with start/end \pm <input type="text" value="0"/> bp <input type="radio"/> Complete sequence <input type="checkbox"/> reverse complement in <input checked="" type="radio"/> FASTA format <input type="radio"/> GenBank format <input type="text" value="Extract Sequence(s)"/>
GeneID Extraction	<input type="button" value="Extract GeneIDs"/> <input type="button" value="Extract GeneIDs by Chromosome"/> e.g. for input into GenomatixPathwaySystem (GePS)
Match Extraction	Export Matches as <input type="button" value="Excel file"/> <input type="button" value="tab-separated file"/> <input type="button" value="BED file"/>

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

GeneIDs of matching sequences

SORY SORY TALE BRAC:18241,16175,54338,75823,14388,244562,170755,226861,252870,67980,19084,16416,69976,22371,27277,66611,56397,56191,225600,69190,243755,239188,18478,16622,60594,71729,109323,69940,66768,19346,269108,83560,230376,75691,102626,252837,67972,12974,71750,17938,21853,22685,17346,230623,16975,21843,69234,22787,77647,80859,106407,94212,75497,19877,170780,208258,66379,51791,28077,268670,218343,319822,74637,329679,242050,67812,117592,70999,11854,71682,68636,16468,216549,17449,66713,76441,268859,75619,70892,68151,56224,74776,328448,12554,12857,18117,69047,67528,329738,99512,14225,100415913,212073,17454,84605,59013,257886,13859,244431,64176,11796,76568,67781,69514,230103,12557,212377,66362,18260,108071,69354,257943,223864,67197,75459,19274,66609,170654,16559,230863,24131,12289,18300,320343,271786,228998,109054,70638,195522,230558,20979,17878,71817,114889,17064,81910,268482,68768,12491,108099,67661,56384,194744,104570,216650,790909,20641,66878,242286,212514,73379,63985,18130,75288,72565,74254,73998,16513,258715,75387,170756,170472,114679,71481,50781,18576,15931,102857,78920,271005,67759,208518,71738,216835,67863,83429,212627,259025,14302,224860,214523,20265,226747,383592,258786,80985,50765,17151,15356,66966,211673,226977,14088,212483,104601,21391,74468,258366,320225,54562,106068,77652,100043468,224796,81630,218100,83428,234407,226025,252967,20873,212127,21416,209737,13869,70729,665574,226751,270685,103268,22025,54712,59044,211914,66923,219140,20356,11429,545156,20084,241769,99952,66559,100040736,230824,231591,18439,231760,17425,16438,243813,106682,76130,65247,109272,18191,431706,383075,14651,13417,21873,258574,433653,16924,66116,319734,59035,319899,258742,93697,19087,20662,67988,209378,11049,66371,10043254,68526,211936,13175,21923,20260,13069,211383,78611,258435,100040724,14389,235567,110606,211151,67706,225895,19726,100034351,100177,100039968,68817,69594,231326,70809,80297,192192,234214,382421,380694,67201,68299,13682,14859,213059,209630,277463,241915,100039008,433801,243164,235441,18728,56292,112207,108912,432995,68897,100038948,624910,58554,379043,657281,20745,238393,27643,214601,100169864,432972,59036,50779,791340,108800,666955,239510,108013,277250,12894,80890,338349,100072,26561,76580,72587,320127,236266,12288,14802

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.

Parameters

Specify what kind of gene keywords you will provide:

Entrez and/or Ensembl Gene IDs Transcript Accession Numbers
 Gene Symbols/Names Affymetrix Probe Set IDs

Paste a list of gene keywords...

13682, 14859, 213056, 209630, 277463, 241915, 100039008, 433801, 243764, 235441, 16728, 56292, 171207, 108912, 432995, 68897, 100038948, 624910, 56554, 379043, 667281, 20715, 238393, 27643, 214601, 100169864, 432572, 59036, 50779, 791340, 108800, 666955, 239510, 108013, 277250, 12894, 80890, 338349, 100072, 26561, 76580, 72587, 320127, 236266, 12288, 14802

or upload a [text file](#) containing gene keywords, optionally with corresponding expression values.

Durchsuchen... Keine Datei ausgewählt.

OR

[Use example gene set](#)

"Inflammation in H. sapiens"

The example data set is from a microarray analysis of Systemic Inflammation in Humans (Calvano et al (2005) Nature 437,1032-7; PMID: 16136080).

Gene expression changes relative to t=0 are displayed at 5 timepoints (2,4,6,9 and 24 hours) after inoculation with bacterial endotoxin.

Organism

Mus musculus

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)*.

Signal Transduction Pathways (Genomatix Literature Mining)	Molecular Functions (GO)	Cellular Components (GO)	Biological Processes (GO)	Diseases (Genomatix Literature Mining)	Diseases (MeSH)			
Tissues (Genomatix Literature Mining)	Tissues (UniGene)	Co-cited genes (Genomatix Literature Mining)	Co-cited TFs (Genomatix Literature Mining)	Pharmacological Substances (Genomatix Literature Mining)				
MeSH-Term	MeSH-Term id(s)	P-value	Adjusted p-value	# Genes (observed)	# Genes (expected)	# Genes (total)	List of observed genes	Gene ids
Neoplasms	C04	1.73e-04	n/a	232	207.76	13767	Galk2, Abcc12, Hagh, Raet1e, Cacna	69976, 244562, 14651, 379043, 12288
Pathologic Processes	C23.550	1.81e-04	n/a	238	214.97	14245	Hagh, Raet1e, Cacna1d, Eps15l1, Slc	14651, 379043, 12289, 13859, 194744
Neoplasms by Site	C04.588	6.17e-04	n/a	210	185.51	12293	Galk2, Abcc12, Hagh, Raet1e, Cacna	69976, 244562, 14651, 379043, 12288
Digestive System Neoplasms	C04.588.274	2.67e-03	n/a	152	128.73	8530	Abcc12, Hagh, Raet1e, Eps15l1, Actr	244562, 14651, 379043, 13859, 66713
Digestive System Neoplasms	C06.301	2.91e-03	n/a	152	128.95	8545	Abcc12, Hagh, Raet1e, Eps15l1, Actr	244562, 14651, 379043, 13859, 66713
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	C16	4.33e-03	n/a	169	147.21	8755	Abcc12, Hagh, Cacna1d, Eps15l1, Slc	14651, 12289, 13859, 194744
Joint Instability	C05.550.521	5.00e-03	n/a	8	2.63	174	B3gal6, Serpina3f, Il1a, Il17, Cox4i1	117592, 238393, 16175, 67661, 12857
Hypersomnolence	C23.550.421	5.35e-03	n/a	8	2.66	176	Sirt4, Hmgcl, Il1a, Cpt1a, Npr2, Kcnj1	75387, 15356, 16175, 12894, 230103
Hypertrophy, Left Ventricular	C23.300.775.250.400, C14.280.195.400	5.41e-03	n/a	23	12.90	855	Fkbp1a, Rock1, Ldb3, Slc6a8, Il1a, Pi	14225, 19677, 24131, 102857, 16175
Consciousness Disorders	C23.888.592.604.359	5.70e-03	n/a	12	5.13	340	Hmgcl, Il1a, Pah, Naca, Nr2c1, Chrm3	15356, 16175, 18478, 17938, 22025, 2

Signal Transduction Pathways (Genomatix Literature Mining)	Molecular Functions (GO)	Cellular Components (GO)	Biological Processes (GO)	Diseases (Genomatix Literature Mining)	Diseases (MeSH)			
Tissues (Genomatix Literature Mining)	Tissues (UniGene)	Co-cited genes (Genomatix Literature Mining)	Co-cited TFs (Genomatix Literature Mining)	Pharmacological Substances (Genomatix Literature Mining)				
Tissue	Tissue id	P-value	Adjusted p-value	# Genes (observed)	# Genes (expected)	# Genes (total)	List of observed genes	Gene ids
ENTIRE HEART	C1281570	3.17e-04	n/a	35	18.93	1181	Uaca, Ucn3, Pde3b, Foxo1, Gpt2, Fkl	72565, 83428, 18576, 17425, 108682
HEART TISSUE	C1272575	1.13e-03	n/a	14	5.40	337	Fkbp1a, Serpina3f, Trpm3, Cpt1a, Cdi	14225, 238393, 226025, 12894, 12554
SPINAL CORD WHITE MATTER STRUCTURE	C0458457	1.70e-03	n/a	5	0.87	54	Sptbn4, Gria4, Itpr1, Tnc, Grm5	80297, 14802, 16438, 21923, 108071
HEART	C0018787	2.21e-03	n/a	32	18.78	1172	Ucn3, Pde3b, Gpt2, Fkbp1a, Rbfox1	83428, 18576, 108682, 14225, 268855
ENTIRE ANTERIOR CRURAL MUSCLE	C0448479	2.48e-03	n/a	2	0.08	5	Fkbp1a, Cs	14225, 12974
SKELETAL MYOCYTES	C1704336	3.37e-03	n/a	19	9.50	593	Sirt4, Foxo1, Clec5, Gpt2, Fkbp1a, Ldl	75387, 17425, 224796, 108682, 14225
PHOTORECEPTORS	C0031760	3.38e-03	n/a	16	7.42	463	Vsx1, Cacna2d4, Slc6a8, Kir7, Ankrk	114889, 319734, 102857, 16559, 2082
INTESTINAL WALL TISSUE	C1708548	4.21e-03	n/a	21	11.15	696	Abcc12, Eps15l1, Ahctf1, Shkbp1, Cd	244562, 13859, 226747, 192192, 1255
EMBRYONIC HEART	C1516821	4.43e-03	n/a	9	3.14	196	Cacna1d, Fkbp1a, Ldb3, Erbb4, Seml	12289, 14225, 24131, 13869, 20356
ENTIRE SPINAL CORD WHITE MATTER	C1281097	4.48e-03	n/a	4	0.67	42	Gria4, Itpr1, Tnc, Grm5	14802, 16438, 21923, 108071

Annotation of Tbx20 binding regions – target prediction

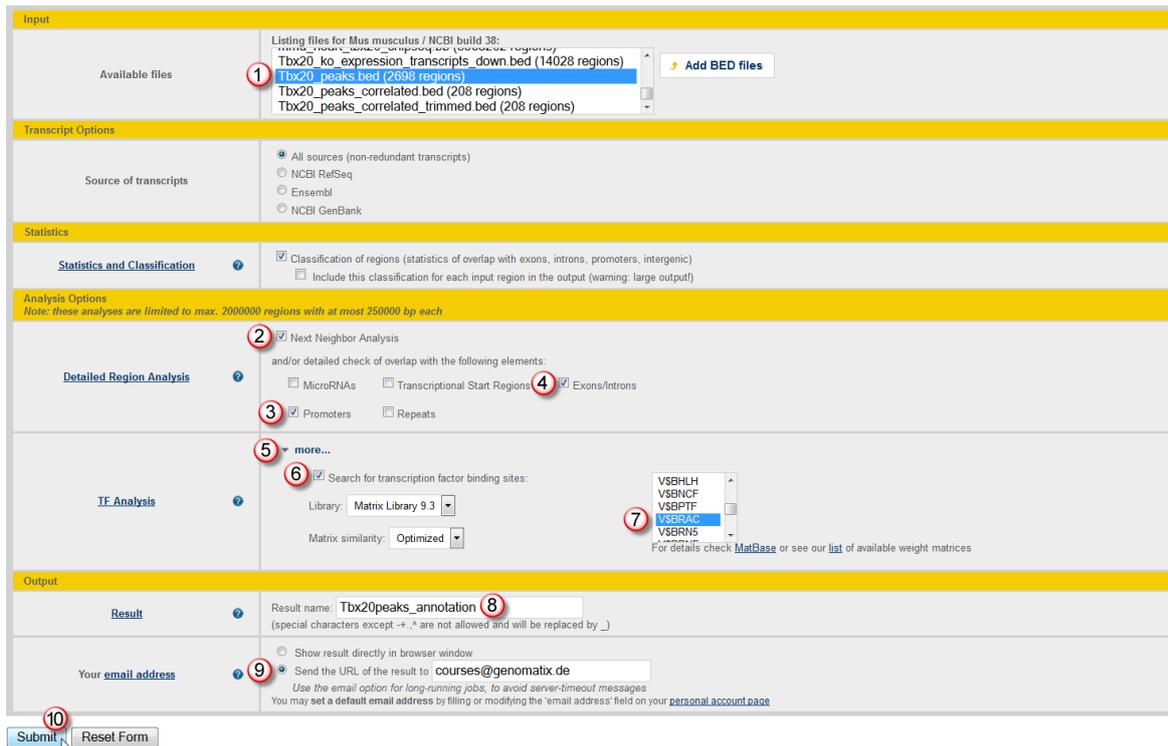
An alternative way for finding potential regulatory targets of a transcription factor based on ChIP-Seq peaks, which can also be applied in the absence of expression data, is to analyze the genomic annotation in the vicinity of the TF peak positions and look for overlapping and neighboring promoters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

Genes & Genomes Gene Regulation



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.



Input

Available files: Listing files for *Mus musculus* / NCBI build 38:
 Tbx20_ko_expression_transcripts_down.bed (14028 regions)
 Tbx20_peaks.bed (2636 regions) **1**
 Tbx20_peaks_correlated.bed (208 regions)
 Tbx20_peaks_correlated_trimmed.bed (208 regions)

Transcript Options

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

Statistics

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

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

Detailed Region Analysis
 Next Neighbor Analysis
 and/or detailed check of overlap with the following elements:
 MicroRNAs Transcriptional Start Regions Exons/Introns **4**
 Promoters Repeats **3**

TF Analysis
 Search for transcription factor binding sites: **6**
 Library: Matrix Library 9.3
 Matrix similarity: Optimized
 VSBHLH
 VSBNCF
 VSBPIF
 V\$BRAC **7**
 V\$BRN5
 For details check [MatBase](#) or see our [list](#) of available weight matrices

Output

Result
 Result name: Tbx20peaks_annotation **8**
 (special characters except +, ., ^ are not allowed and will be replaced by _)

Your email address **9**
 Show result directly in browser window
 Send the URL of the result to courses@genomatix.de
 Use the email option for long-running jobs, to avoid server-timeout messages
 You may set a default email address by filling or modifying the 'email address' field on your [personal account page](#)

10 Submit Reset Form

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

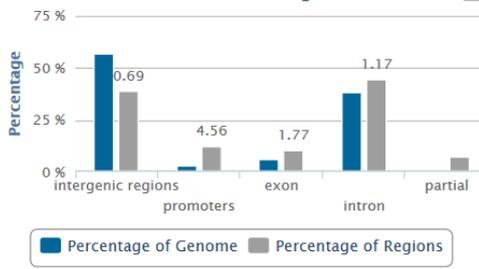
Region Classification
Overlap Statistics
Detailed Annotation and Download

Region Classification on Tbx20_peaks.bed

General Statistics	
Total number of Regions:	2698
Total basepairs:	411423
Minimum Region length:	36
Maximum Region length:	4808
Average Region length:	152.5

Enrichment
General

Enrichment: Genome vs. Region annotation



Type of genomic element	Number of Regions	Percentage of Regions	Percentage in Genome	Enrichment compared to Genome
Exon	272	10.1%	5.7%	1.8
Partial	183	6.8%	-	-
Intron	1190	44.1%	37.8%	1.2
Intergenic regions	1053	39.0%	56.5%	0.7
Sum of above	2698	100.0%	-	-
Promoters	332	12.3%	2.7%	4.6

Distribution of Regions on the Genome
[>>> show details <<<](#)

Overlap details can be viewed in the Overlap Statistics section.

Region Classification
Overlap Statistics
Detailed Annotation and Download

Overlap Statistics

A total of 2698 regions was checked (=100%)

Number of input regions	Percentage input regions	Description
1645	61.0%	overlap with at least one locus
1053	39.0%	overlap with intergenic regions
455	16.9%	overlap with at least one exon (of alternative transcripts)
1472	54.6%	overlap with at least one intron (of alternative transcripts)
332	12.3%	overlap with promoters
>>> show details <<< on exon and intron overlap		
1444	53.5%	regions have a match to the matrix family V\$BRAC (a total of 2198 matches)

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.

Region Classification
Overlap Statistics
Detailed Annotation and Download 1

Detailed annotation for all regions or subsets

For details of the next neighbor analysis please use the download-details or browse-table option below.

Select regions containing a match to V\$BRAC
 Select regions overlapping with at least one exon
 Select regions overlapping with at least one intron
 Select regions overlapping with intergenic regions
 Select regions overlapping with promoters

(Use shift/ctrl-keys to select combinations)

Invert Selection (i.e. not this type of element)

Available tasks for selected regions

Download details in EXCEL format

Download details in tab-separated format

Export regions to BED file format

2 Browse table with details for selected regions

Extract GeneIDs of genes overlapping input region

Extract Symbols of genes overlapping input region

Extract GeneIDs of genes where the region overlaps with promoter

Extract GeneIDs of neighboring genes

that are max bps of selected regions

and keep region assignment

Name for extracted file:

3

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.

Detailed Annotation of Regions

Note: The following terminology is used for next transcripts:



2698 selected regions (All regions)

showing at most 50 regions per page, starting with region 1

Annotation								
Input	Select	Next transcript downstream (+)	Next transcript downstream (-)	Next transcript upstream (+)	Next transcript upstream (-)	Overlapping loci/transcripts/promoters	TSRs, repeats, microRNAs	TF binding sites
Region_1 Id:1 Score=9.12e-13 chr1 4412567-4412753 (187bp) GenomeBrowser	<input checked="" type="checkbox"/>	AK051370 GeneID_20671 Sox17(+) 83798 bp downstream	XM_006495473 GeneID_19888 Rp1(-) 3301 bp downstream	AK140060 GeneID_110038431 Gm10568(+) 732598 bp upstream	NM_011441 GeneID_20671 Sox17(-) 83846 bp upstream			1 matches to V\$BRAC Start MainInspector on this region
Region_2 Id:2 Score=0.0141 chr1 7819512-7819581 (70bp) GenomeBrowser	<input checked="" type="checkbox"/>	AK142798 AK142798(+) 708199 bp downstream	AK142999 AK043789(-) 421643 bp downstream	AK036865 GeneID_102632303 LOC102632303(+) 321613 bp upstream	XM_006495574 GeneID_21096 Sntg1(-) 963371 bp upstream			Start MainInspector on this region
Region_3 Id:3 Score=2.63e-08 chr1 11003665-11003662 (90bp) GenomeBrowser	<input checked="" type="checkbox"/>	XM_006495442 GeneID_109294 Prex2(+) 82943 bp downstream	AK156457 AK156457(-) 219198 bp downstream	XM_006495441 GeneID_109294 Prex2(+) 10419 bp upstream	ENSMUST00000187404 ENSMUSG00000101627(-) 229609 bp upstream	Prex2/GeneID_109294 overlaps > show details < on exon/intron overlap		1 matches to V\$BRAC Start MainInspector on this region

Next, please go back to the overview page, and select the option *Extract GeneIDs 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 GeneIDs on your local computer.

Available tasks for selected regions

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

that are max bps of selected regions

and keep region assignment

Name for extracted file:

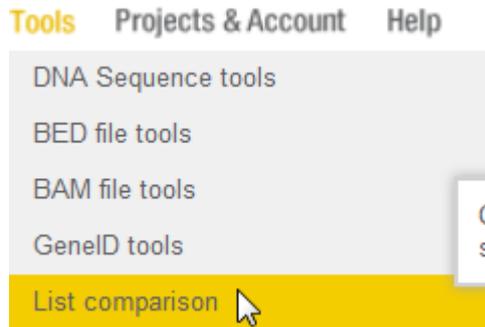
The file contains the GeneIDs and associated peak identifiers based on the peak IDs in the BED file.

11287	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_886
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					
11811	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).

List Input

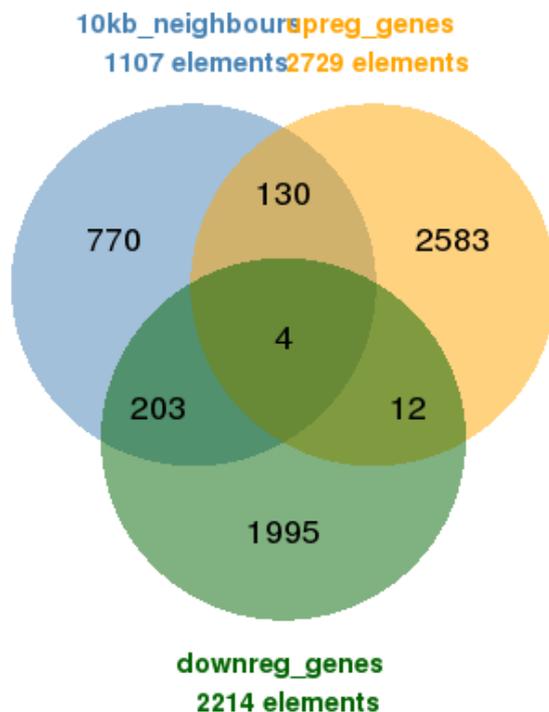
Number of Lists	How many lists do you want to compare? <input type="text" value="3"/> 1 <small>(Venn diagrams will be available for 2-, 3- and 4-list comparisons)</small>
List 1	Name for List 1: <input type="text" value="10kb neighbv"/> 2 Enter your list elements separated by blanks, newlines or commas: <input type="text"/> Optional: enter a list of associated values (must be same number and order as the list elements) <input type="text"/> or alternatively upload a text file containing List1: <input type="button" value="Durchsuchen..."/> <input type="text" value="Genes_within10kb_of_Tbx20peaks.txt"/> 3 and limit analysis to the first <input type="text" value="2"/> columns <small>Format: one element per line, first value is used for comparison, optionally tab-delimited associated value(s). Note: text files only, i.e. binary Excel files will not be accepted</small>
List 2	Name for List 2: <input type="text" value="upreg_genes"/> 4 Enter list elements: <input type="text"/> Optional: Associated values <input type="text"/> or alternatively upload a text file containing List2 (and opt. associated values): <input type="button" value="Durchsuchen..."/> <input type="text" value="Tbx20_ko_expression_diff_expressed_genes_up.list"/> 5 Limit analysis to the first <input type="text" value="3"/> 6 columns
List 3	Name for List 3: <input type="text" value="downreg_genes"/> 7 Enter list elements: <input type="text"/> Optional: Associated values <input type="text"/> or alternatively upload a text file containing List3 (and opt. associated values): <input type="button" value="Durchsuchen..."/> <input type="text" value="Tbx20_ko_expression_diff_expressed_genes_down.list"/> 8 Limit analysis to the first <input type="text" value="3"/> 9 columns

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 included 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.

Options	
Case Sensitivity	<input checked="" type="checkbox"/> 10 treat elements in lists CaSe-SeNsitiVe
Header Line	<input type="checkbox"/> remove the first line of all uploaded files
Compute Probability (only for 2 lists)	more...
<input type="button" value="Compare lists"/> <input type="button" value="Reset Form"/>	

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.

Union / Intersection		
Union		
<input checked="" type="radio"/> Union of 3 lists	5697 elements	100017, 100034251, 100036535, 100036537, 100037258, 100037282, 100038347, 100038353, 100038355, 100038356, ... (list truncated)
Intersection		
<input type="radio"/> Common to exactly 3 Lists	4 elements	108000, 212307, 58194, 64291
<input type="radio"/> Common to exactly 2 Lists	345 elements	100039027, 100040872, 100303644, 100379605, 100502602, 100503434, 100503471, 100503659, 100504518, 100705, ... (list truncated)
Single Lists		
10kb_neighbours (Input List1)	1107 elements	100037258, 100038347, 100038353, 100038381, 100038388, 100038412, 100038424, 100038512, 100038543, 100038570, ... (list truncated)
<input type="radio"/> Non-redundant in List1	1107 elements	100037258, 100038347, 100038353, 100038381, 100038388, 100038412, 100038424, 100038512, 100038543, 100038570, ... (list truncated)
<input type="radio"/> Duplicates within List1	0 elements	-
<input type="radio"/> Elements only in List1 (in no other list)	770 elements	100037258, 100038347, 100038353, 100038381, 100038388, 100038412, 100038424, 100038512, 100038543, 100038570, ... (list truncated)
upreg_genes (Input List2)	2729 elements	100017, 100034251, 100036535, 100036537, 100038355, 100038369, 100038405, 100038452, 100038468, 100038531, ... (list truncated)
<input type="radio"/> Non-redundant in List2	2729 elements	100017, 100034251, 100036535, 100036537, 100038355, 100038369, 100038405, 100038452, 100038468, 100038531, ... (list truncated)
<input type="radio"/> Duplicates within List2	0 elements	-
<input type="radio"/> Elements only in List2 (in no other list)	2583 elements	100017, 100034251, 100036535, 100036537, 100038355, 100038369, 100038405, 100038452, 100038468, 100038531, ... (list truncated)
downreg_genes (Input List3)	2214 elements	100037282, 100038356, 100038368, 100038395, 100038453, 100038564, 100038605, 100038712, 100038761, 100039027, ... (list truncated)
<input type="radio"/> Non-redundant in List3	2214 elements	100037282, 100038356, 100038368, 100038395, 100038453, 100038564, 100038605, 100038712, 100038761, 100039027, ... (list truncated)
<input type="radio"/> Duplicates within List3	0 elements	-
<input type="radio"/> Elements only in List3 (in no other list)	1995 elements	100037282, 100038356, 100038368, 100038395, 100038453, 100038564, 100038605, 100038712, 100038761, 100040293, ... (list truncated)

Export selected list as

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.

Element	value(s) from 10kb neighbours	value(s) from upreg_genes	value(s) from downreg_genes
100017	-	1.307 Ldlrap1	-
100034251	-	1.782 Wfdc17	-
100036535	-	2.338 Gm9913	-
100036537	-	2.263 Gm11149	-
100037258 Region_2099	-	-	-
100037282	-	-	-1.645 Rsph3b
100038347 Region_1097	-	-	-
100038353 Region_2543	-	-	-
100038355	-	2.291 Runx2os1	-
100038356	-	-	-1.313 Gm15612
100038368	-	-	-1.474 Gm10609
100038369	-	1.653 F630201L12Rik	-
100038381 Region_1416	-	-	-
100038388 Region_1318	-	-	-
100038395	-	-	-1.49 1700061E17Rik
100038405	-	3.872 Gm10827	-
100038412 Region_1195	-	-	-
100038424 Region_606	-	-	-
100038452	-	2.469 Gm13372	-
100038453	-	-	-2.452 Gm12522
100038468	-	2.654 Gm10684	-
100038512 Region_635	-	-	-
100038531	-	1.503 D030062O11Rik	-
100038543 Region_2235	-	-	-
100038548	-	2.36 Gm10521	-
100038564	-	-	-1.81 Gm10524
100038570 Region_1775	-	-	-
100038605	-	-	-1.512 E030047D23Rik
100038610 Region_1606	-	-	-

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).

Element	value(s) from 10kb neighbours	value(s) from upreg_genes	value(s) from downreg_genes
100379605 Region_1501	-	2.152 Gm15270	-
100503434 Region_2453	-	2.537 Gm19689	-
100503471 Region_176	-	1.943 Gm15867	-
100503659 Region_1562	-	1.453 Dos	-
102595 Region_1413	-	1.302 Plekho2	-
102631551 Region_2547	-	1.669 LOC102631551	-
105245 Region_1933	-	1.336 Txndc5	-
105988 Region_2224	-	2.486 Esp1	-
106205 Region_2236	-	1.071 Zc3h7a	-
107702 Region_1192	-	1.722 Rnh1	-
107765 Region_2582	-	2.879 Ankrd1	-
108000 Region_161	-	2.281 Cenpf	-1.399 Cenpf
108099 Region_773	-	1.074 Prkag2	-
108903 Region_1787	-	1.214 Tbcd	-
108912 Region_2059	-	2.517 Cdca2	-
11304 Region_571	-	1.87 Abca4	-
11459 Region_1330	-	1.127 Acta1	-
11461 Region_884	-	1.774 Actb	-
11465 Region_676	-	1.771 Actg1	-
11504 Region_2308	-	2.044 Adamts1	-
11520 Region_659	-	1.241 Plin2	-
12181 Region_2177	-	1.276 Bop1	-
12523 Region_147	-	3.191 Cd84	-
12606 Region_1057	-	1.915 Cebpa	-
12982 Region_2617	-	1.834 Csf2ra	-
14087 Region_1329	-	2.129 Fanca	-

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

Literature

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

Sakabe NJ, Aneas I, Shen T, Shokri L et al. Dual transcriptional activator and repressor roles of TBX20 regulate adult cardiac structure and function. *Hum Mol Genet* 21(10), 2194-2204 (2012).

Stacklies W, Redestig H, Scholz M, Walther D, Selbig J: *pcaMethods* - a Bioconductor package providing PCA methods for incomplete data. *Bioinformatics* 23, 1164-1167 (2007).

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

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

List of resources available on the web:

Gene Expression Omnibus:

<http://www.ncbi.nlm.nih.gov/geo/>

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.

EIDorado 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.