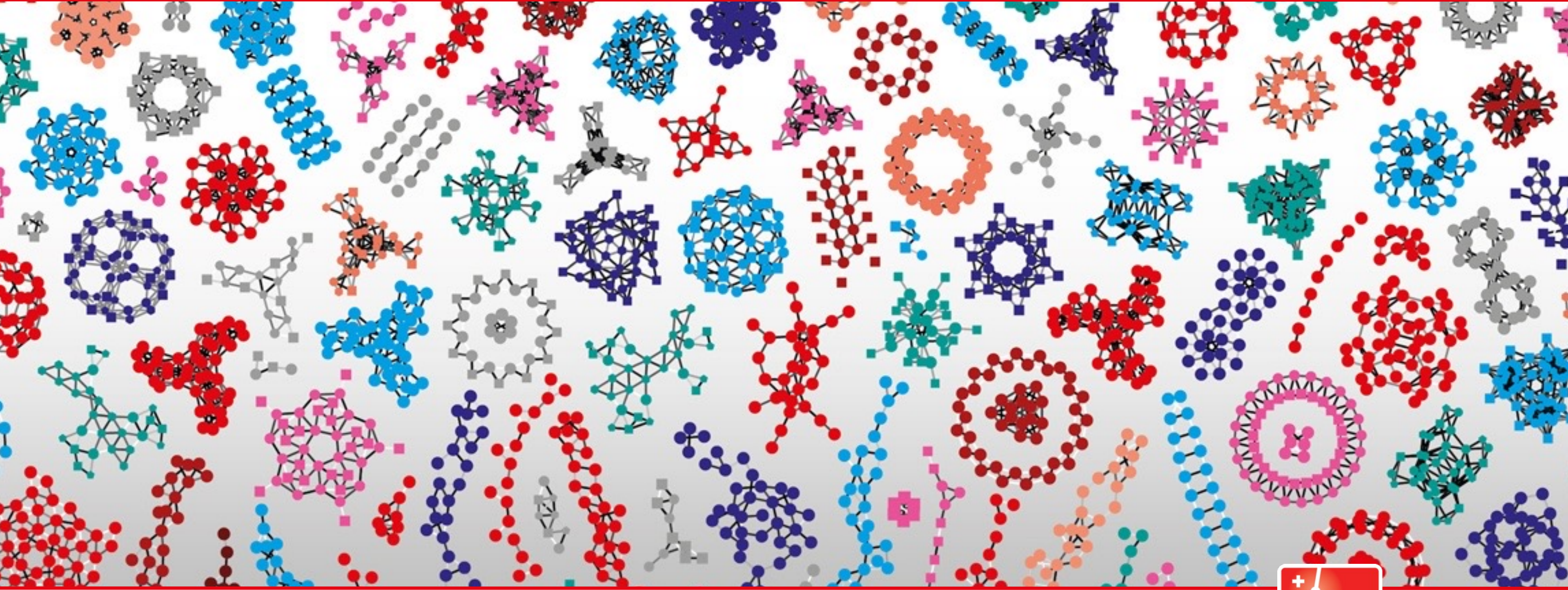


SIB Swiss Institute of Bioinformatics

CREMA tutorial

Erik van Nimwegen

Mikhail Pachkov

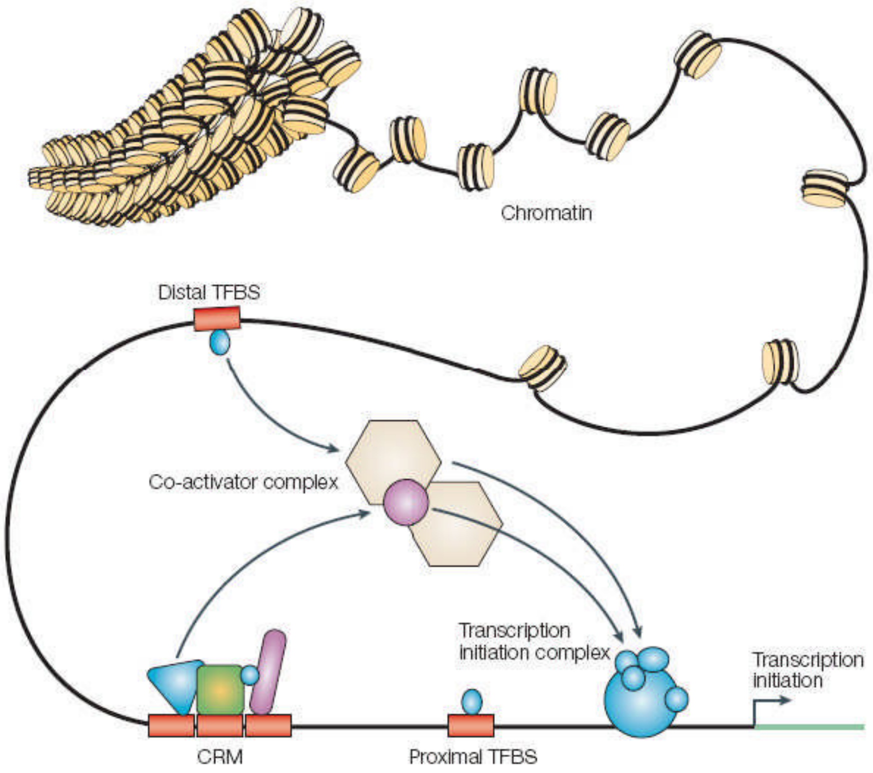
elixir SWITZERLAND

(BC)2 Basel Computational Biology Conference

@NimwegenLab

www.sib.swiss

What about distal regulation?



- ISMARA only considers regulatory elements near the transcription start site.
- But in higher eukaryotes, a lot (most?) of gene regulation is driven by distal cis-regulatory elements (enhancers).

Features of (distal) Cis-Regulatory Elements

- Activation requires local chromatin structure to become accessible.
- Each CRE is bound by different combinations of TFs.
- RNA polymerase is recruited to active CREs.
- Active CREs can produce short aborted transcripts.
- Chromatin is looped (actively) so that CREs contact target promoters.
- CRE state is associated with particular chromatin marks.

Why is including the effects of distal CREs challenging?

1. **There are too many!** A substantial fraction of the genome can act as a CRE *in particular tissues/conditions*.

2. **CREs are highly condition-dependent.** In contrast to elements like genes and promoters, the set of active CREs in the genome is highly condition-dependent.

3. **Disagreement between different methods for CRE identification** (e.g. DNA accessibility, H3K4me1, H3K27ac, p300, eRNAs).

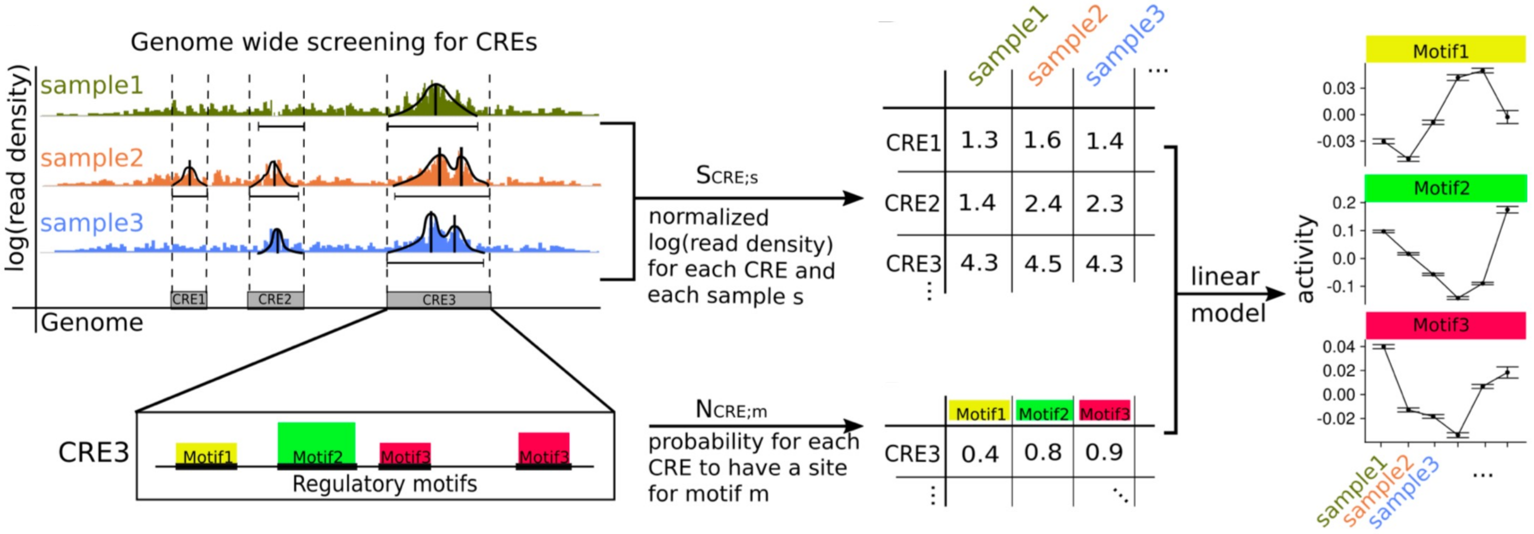
4. **Poor understanding of CRE-promoter interaction**
   - We typically do not know which CREs target which promoters.
   - Little understanding of how CRE activity affects target gene expression.

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB Swiss Institute of Bioinformatics

CREMA

Automated modeling of genome-wide chromatin state in terms of local constellations of regulatory sites



$S_{CRE;s}$
normalized log(read density) for each CRE and each sample s

	sample1	sample2	sample3	...
CRE1	1.3	1.6	1.4	
CRE2	1.4	2.4	2.3	
CRE3	4.3	4.5	4.3	
⋮				

$N_{CRE;m}$
probability for each CRE to have a site for motif m

	Motif1	Motif2	Motif3
CRE3	0.4	0.8	0.9
⋮			⋱

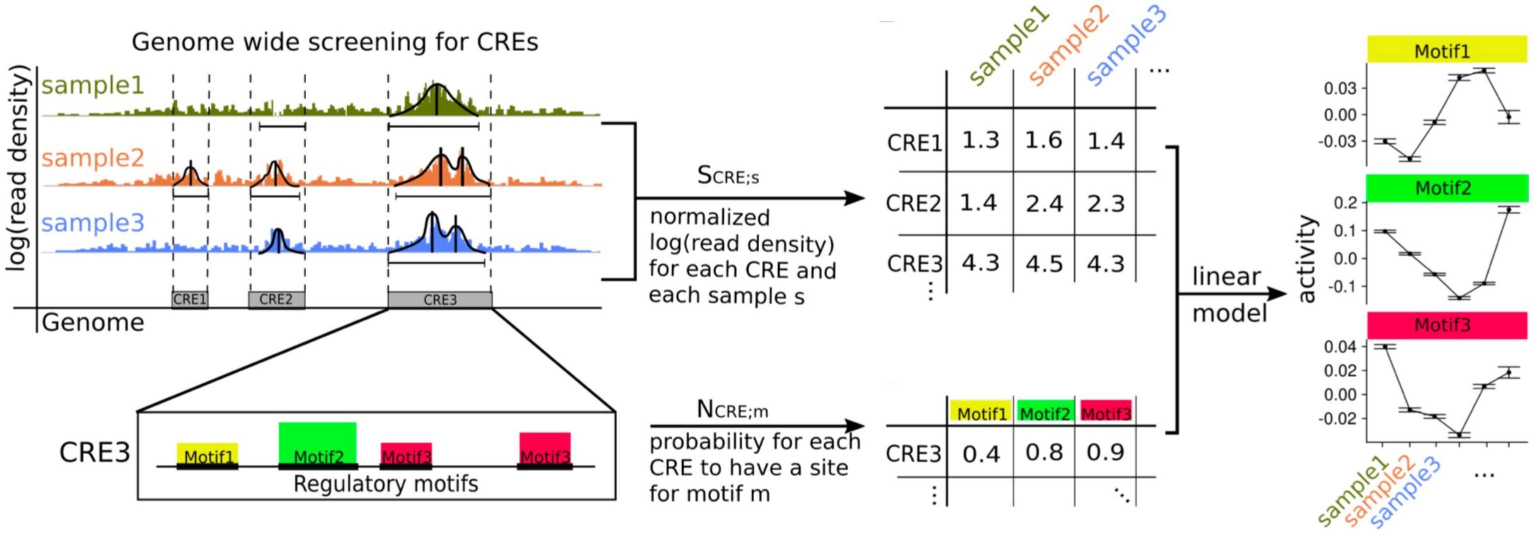
Anne Krämer

Summary of the approach

- **Input**: raw sequencing data of enhancer marks (Dnase-seq, ATAC-seq, ChIP-seq) across a set of samples.
- **CRE detection:** All genomic regions that show a significant enrichment in at least one sample.
- **CRE signal matrix:** Quantify the strength of each CRE's signal across conditions.
- **TFBS annotation:** Predict TFBSs in all CREs genome-wide.
- **Model CRE activity:** Model the CRE signal strength across samples in terms the the TFBSs in each CRE and activities of regulators.

CREMA

Automated modeling of genome-wide chromatin state in terms of local constellations of regulatory sites



$S_{CRE;s}$

normalized log(read density) for each CRE and each sample s

	sample1	sample2	sample3	...
CRE1	1.3	1.6	1.4	
CRE2	1.4	2.4	2.3	
CRE3	4.3	4.5	4.3	
⋮				

$N_{CRE;m}$

probability for each CRE to have a site for motif m

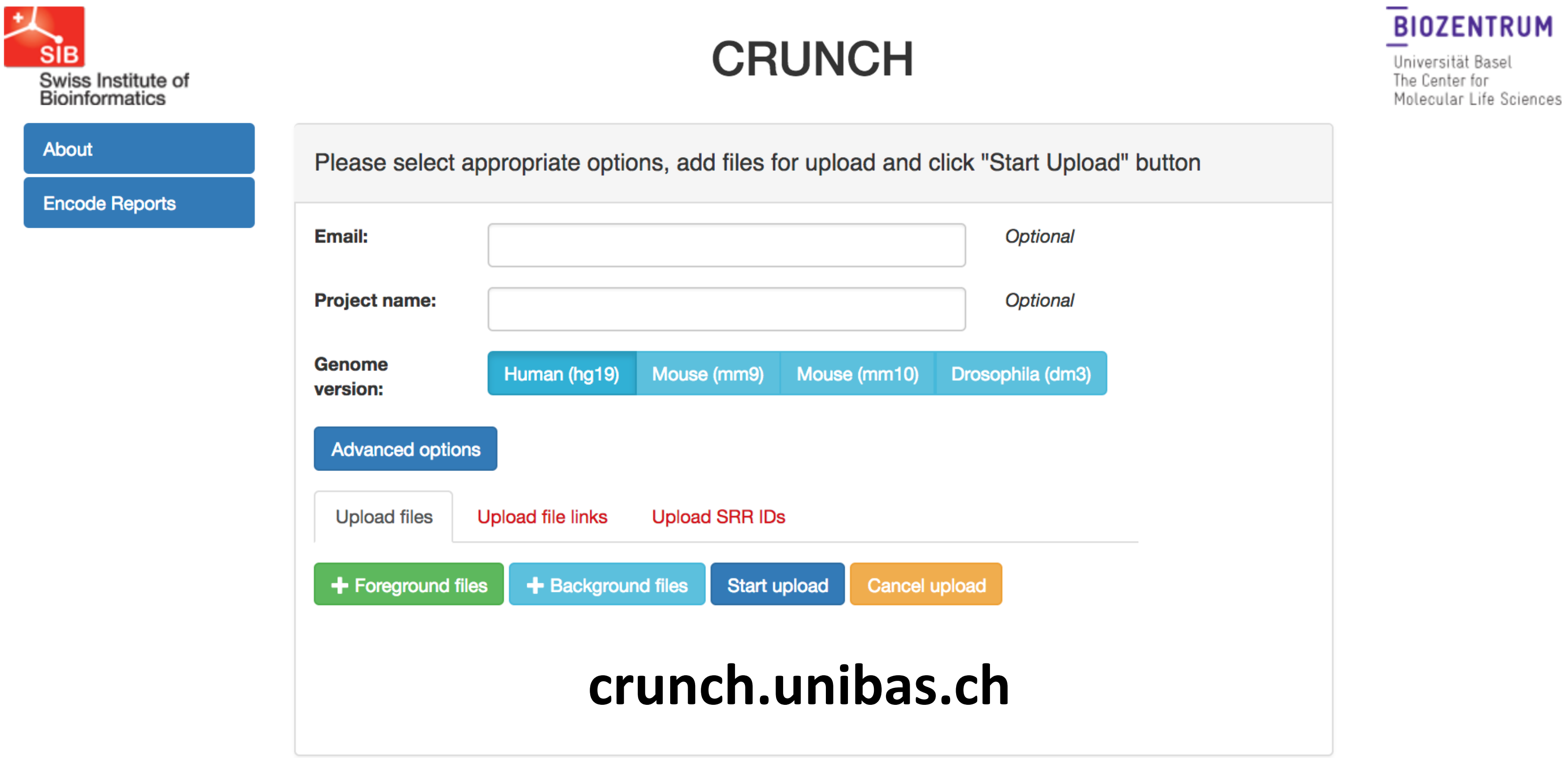
	Motif1	Motif2	Motif3
CRE3	0.4	0.8	0.9
⋮			⋱

Anne Krämer

Summary of the approach

- **Input**: raw sequencing data of enhancer marks (Dnase-seq, ATAC-seq, ChIP-seq) across a set of samples.
- **CRE detection:** All genomic regions that show a significant enrichment in at least one sample.
- **CRE signal matrix:** Quantify the strength of each CRE's signal across conditions.
- **TFBS annotation:** Predict TFBSs in all CREs genome-wide.
- **Model CRE activity:** Model the CRE signal strength across samples in terms the the TFBSs in each CRE and activities of regulators.

Completely automated analysis of ChIP-seq data



Please select appropriate options, add files for upload and click "Start Upload" button

Email: *Optional*

Project name: *Optional*

Genome version:

crunch.unibas.ch

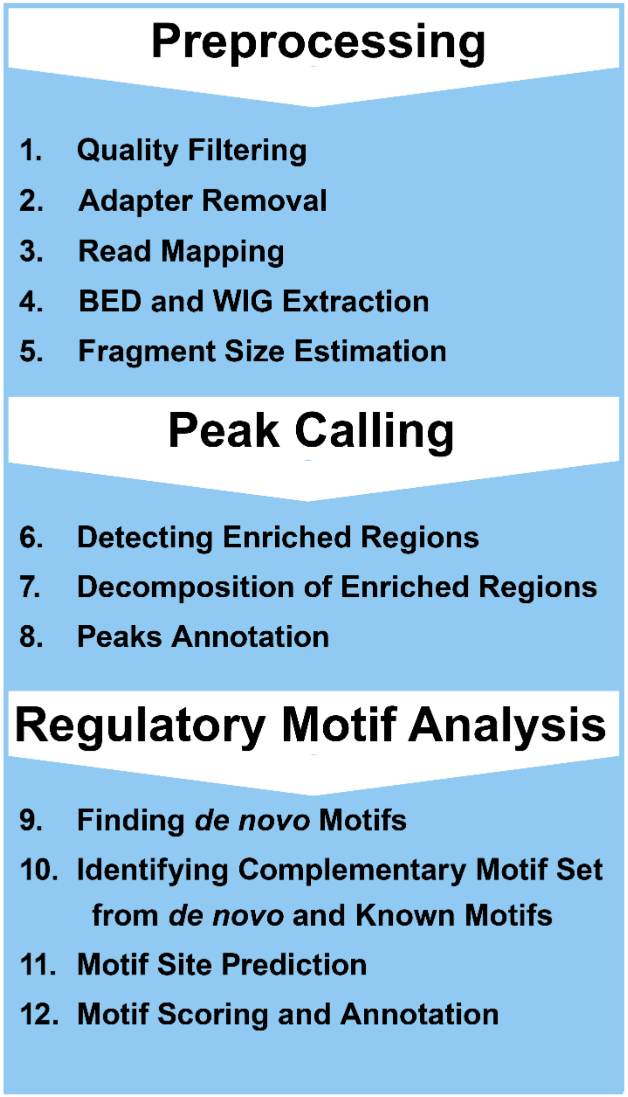
Citation:

Genome Res. 2019 Jul;29(7):1164-1177. doi: 10.1101/gr.239319.118. Epub 2019 May 28.

Crunch: integrated processing and modeling of ChIP-seq data in terms of regulatory motifs.

Berger S[1], Pachkov M[1], Arnold P[1], Omidi S[1], Kelley N[1], Salatino S[1], van Nimwegen E[1].

Overview of CRUNCH analysis steps



Preprocessing

1. Quality Filtering
2. Adapter Removal
3. Read Mapping
4. BED and WIG Extraction
5. Fragment Size Estimation

Peak Calling

6. Detecting Enriched Regions
7. Decomposition of Enriched Regions
8. Peaks Annotation

Regulatory Motif Analysis

9. Finding *de novo* Motifs
10. Identifying Complementary Motif Set from *de novo* and Known Motifs
11. Motif Site Prediction
12. Motif Scoring and Annotation

Overview of CRUNCH analysis steps

Preprocessing

1. Quality Filtering
2. Adapter Removal
3. Read Mapping
4. BED and WIG Extraction
5. Fragment Size Estimation

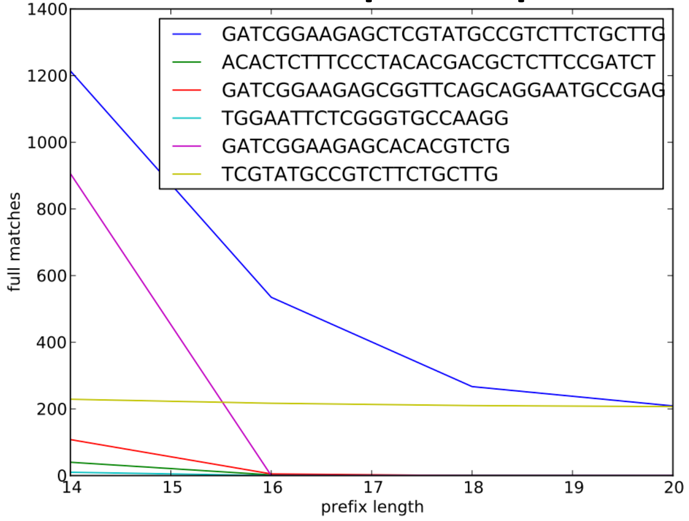
Peak Calling

6. Detecting Enriched Regions
7. Decomposition of Enriched Regions
8. Peaks Annotation

Regulatory Motif Analysis

9. Finding *de novo* Motifs
10. Identifying Complementary Motif Set from *de novo* and Known Motifs
11. Motif Site Prediction
12. Motif Scoring and Annotation

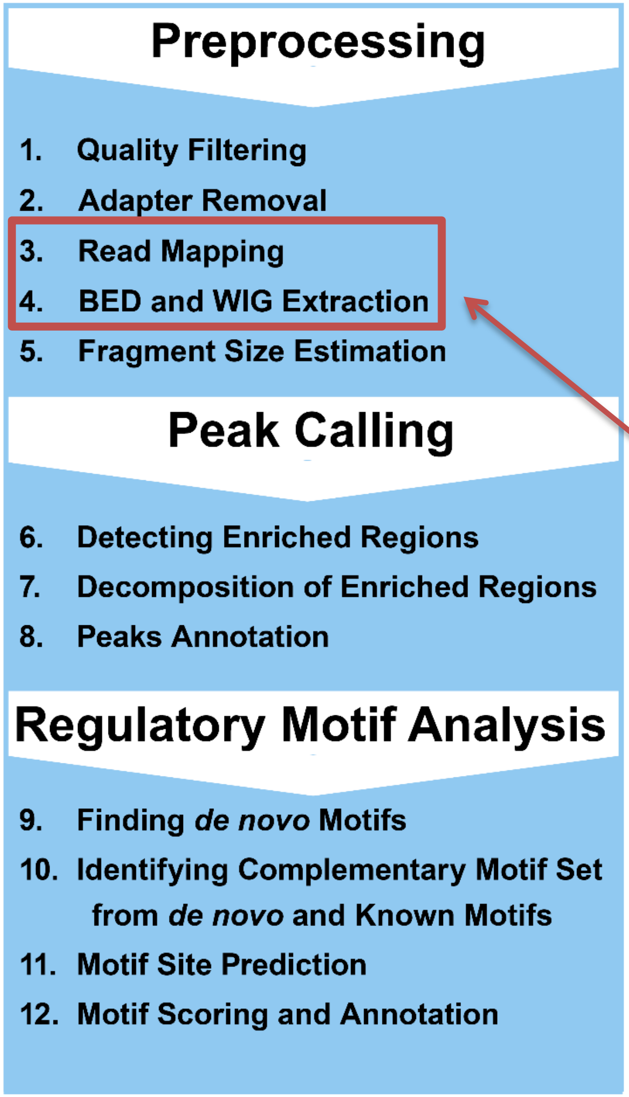
Matches to adaptor sequences



- Truncate low quality 3’ ends of reads.
- Remove reads that are:
  - too short
  - too low sequencing quality (phred score)
  - too many Ns
  - too low dinucleotide entropy.

- Identify which of a library of 3’ adapter sequences has most prefix matches to the reads.
- Remove adaptor matches.

Overview of CRUNCH analysis steps



Preprocessing

1. Quality Filtering
2. Adapter Removal
3. Read Mapping
4. BED and WIG Extraction
5. Fragment Size Estimation

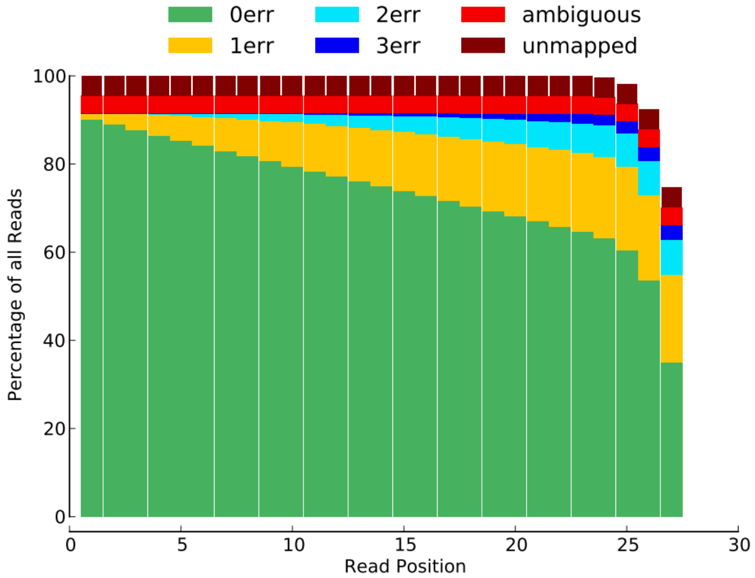
Peak Calling

6. Detecting Enriched Regions
7. Decomposition of Enriched Regions
8. Peaks Annotation

Regulatory Motif Analysis

9. Finding *de novo* Motifs
10. Identifying Complementary Motif Set from *de novo* and Known Motifs
11. Motif Site Prediction
12. Motif Scoring and Annotation

Mapping quality



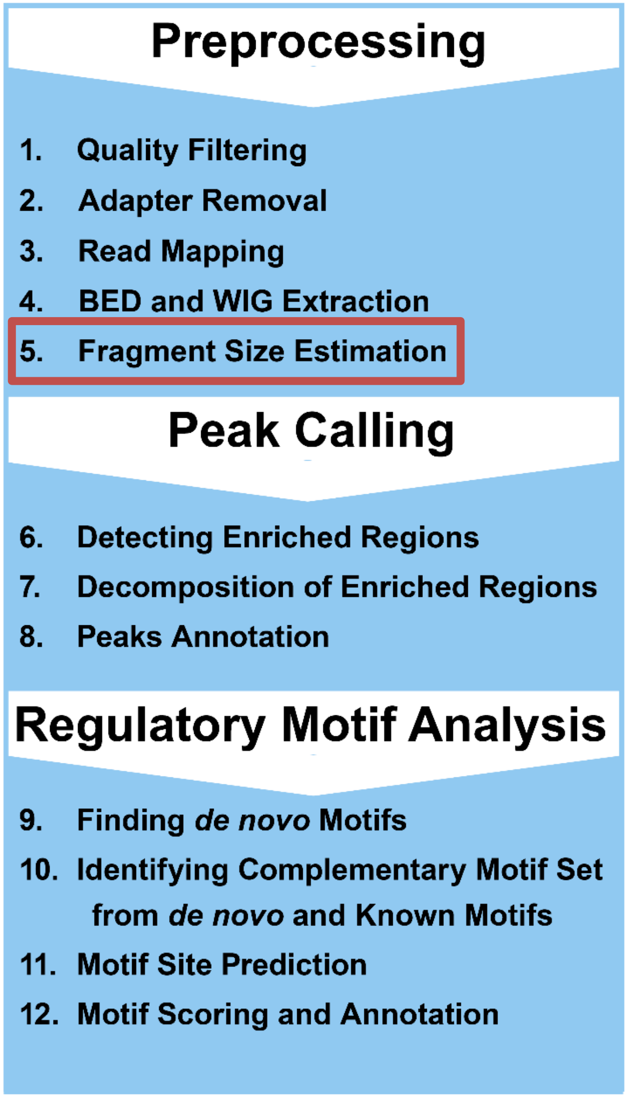
- Map reads to the genome (Bowtie).
- Use only ‘best’ mappings for each read.

- **Note:** Multi-mappers are divided with equal weight over the loci that they map to.

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

Overview of CRUNCH analysis steps



Preprocessing

1. Quality Filtering
2. Adapter Removal
3. Read Mapping
4. BED and WIG Extraction
5. Fragment Size Estimation

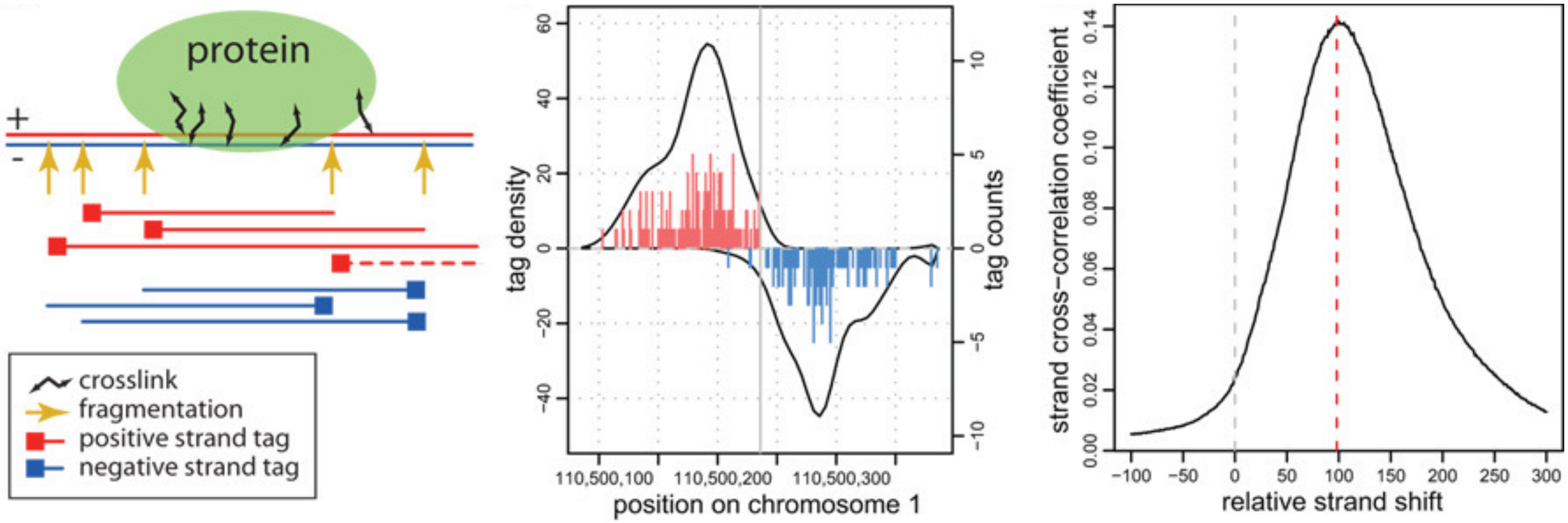
Peak Calling

6. Detecting Enriched Regions
7. Decomposition of Enriched Regions
8. Peaks Annotation

Regulatory Motif Analysis

9. Finding *de novo* Motifs
10. Identifying Complementary Motif Set from *de novo* and Known Motifs
11. Motif Site Prediction
12. Motif Scoring and Annotation

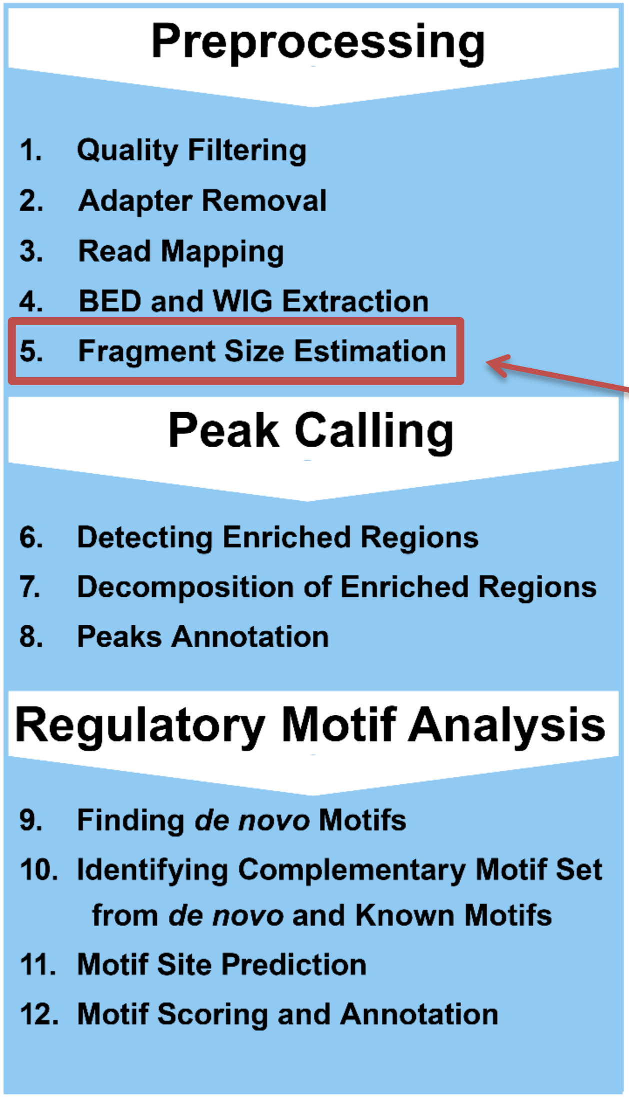
Fragment length can be estimated from cross-correlation of reads on opposite strands



From: Kharchenko et al *Nat Biotech* (2008), after Schmid and Bucher *Cell* (2007)

- DNA fragments are either sequenced from the left end on the plus strand.
- Or from their right end on the negative strand.
- The mapping position on pos/neg strand corresponds to the start/end of the fragment.
- One binding peak leads to *two* peaks of mapped reads: one on plus strand, and one shifted by fragment length on the negative strand.
- The cross-correlation between starts/ends of reads on pos/neg strand captures the fragment length.

Overview of CRUNCH analysis steps



Preprocessing

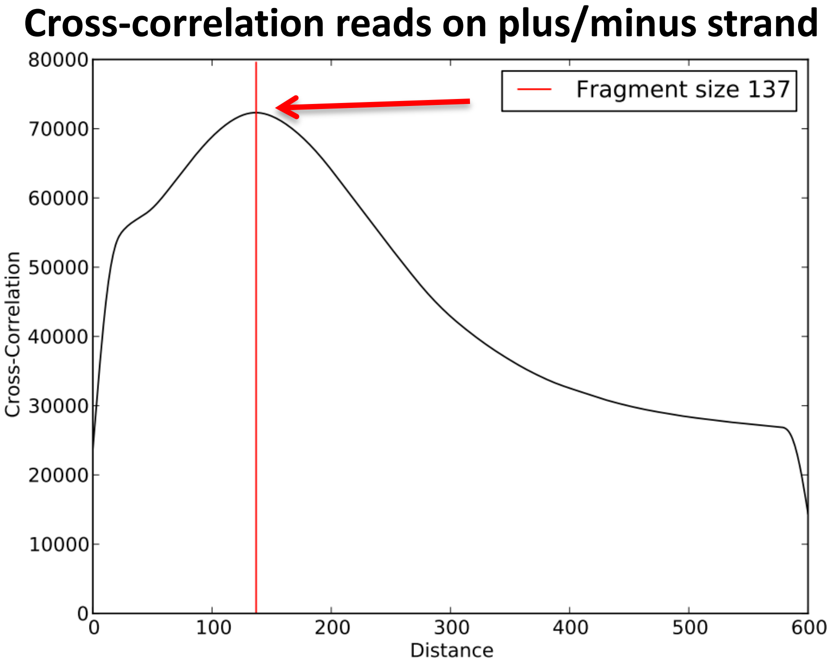
1. Quality Filtering
2. Adapter Removal
3. Read Mapping
4. BED and WIG Extraction
5. Fragment Size Estimation

Peak Calling

6. Detecting Enriched Regions
7. Decomposition of Enriched Regions
8. Peaks Annotation

Regulatory Motif Analysis

9. Finding *de novo* Motifs
10. Identifying Complementary Motif Set from *de novo* and Known Motifs
11. Motif Site Prediction
12. Motif Scoring and Annotation



Cross-correlation reads on plus/minus strand

- Cross-correlation *C*(*d*) between reads starting on plus strand and ending *d* nucleotides downstream on minus strand:

$$C(d) = \sum_i r_+(i) r_-(i+d)$$

- Using this, we estimate the (strand independent) central position for each read.

Overview of CRUNCH analysis steps

Preprocessing

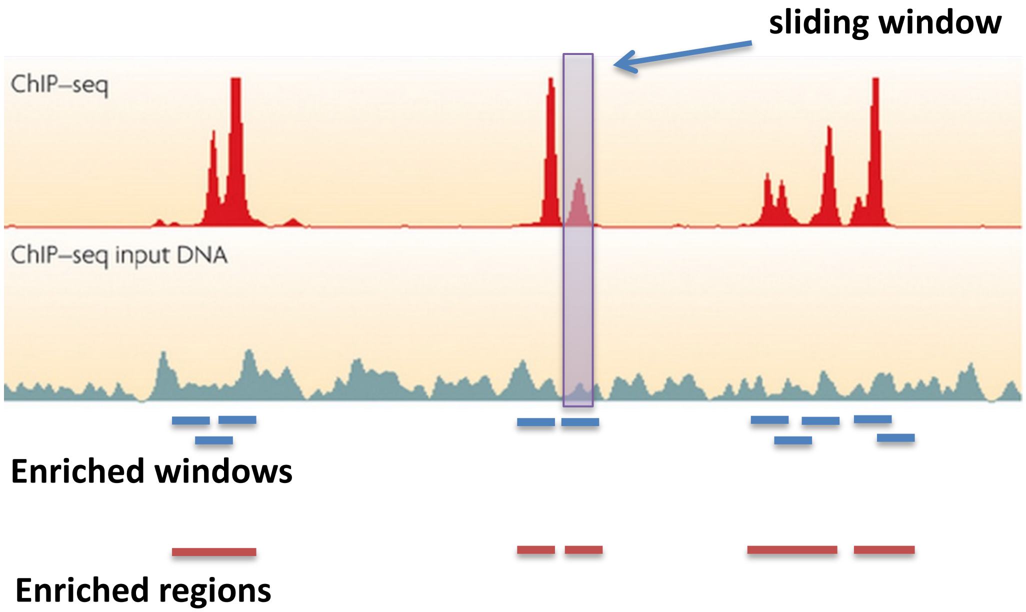
1. Quality Filtering
2. Adapter Removal
3. Read Mapping
4. BED and WIG Extraction
5. Fragment Size Estimation

Peak Calling

6. Detecting Enriched Regions
7. Decomposition of Enriched Regions
8. Peaks Annotation

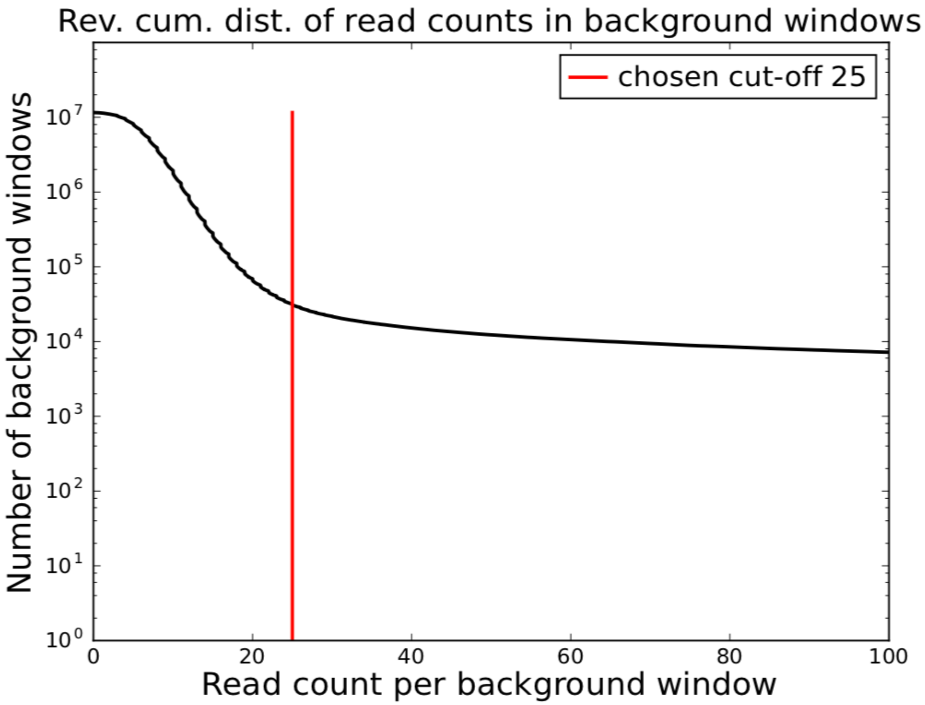
Regulatory Motif Analysis

9. Finding *de novo* Motifs
10. Identifying Complementary Motif Set from *de novo* and Known Motifs
11. Motif Site Prediction
12. Motif Scoring and Annotation



- Slide 500 bp window across the genome.
- Quantify significance of the enrichment of ChIP-seq over input DNA in each window.
- Collect all windows over a significance threshold.
- Fuse consecutive windows into enriched regions.

Removing regions with abnormally high coverage in background samples



- Reverse cumulative distribution of background reads per window.
- About 1 in 1000 windows has abnormally large coverage.
- These regions are often associated with repetitive elements and map poorly to other species.
- These are likely an artefact, e.g. the assembly may underestimate the size of these repeats.
- The statistics of the peak finding model breaks down in these regions.
- CRUNCH thus removes these regions from consideration.

Bayesian model for identifying enriched regions

Noise model for read-counts in un-enriched windows

- *Multiplicative* noise plus *Poisson* sampling, i.e. as previously developed in:

  Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, **van Nimwegen** E.
  Genome Biol. 2009;10(7):R79. doi: 10.1186/gb-2009-10-7-r79. Epub 2009 Jul 22.

Variables:

- *n*,*m* = reads in ChIP/input sample.
- *N*,*M* = total reads in ChIP/input sample.
- *σ* = standard-deviation of the multiplicative noise.
- *μ* = Shift in average log read-density.

Enrichment ***x*****:**

$$x = \log\left[\frac{n}{N}\right] - \log\left[\frac{m}{M}\right]$$

Probability of observing ***x*** **if there is no true enrichment**: $P(x\,|\,\mu,\sigma) \propto \exp\left[-\frac{(x-\mu)^2}{2\left(2\sigma^2 + \frac{1}{n} + \frac{1}{m}\right)}\right]$

Mixture model

- The enrichment x_i for each window *i* derives from either the noise model or a uniform distribution (= 'something else'):

$$P(D\,|\,\mu,\sigma,\rho) = \prod_i \left[P(x_i\,|\,\mu,\sigma)\rho + \frac{1-\rho}{x_{\max} - x_{\min}}\right]$$

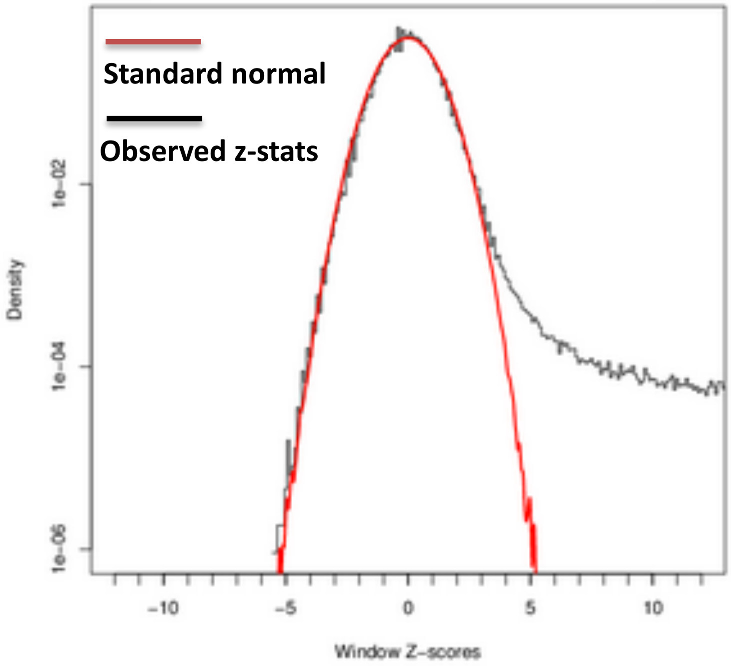
- We fit *μ, σ,* and *ρ* to *maximize* $P(D\,|\,\mu,\sigma,\rho)$, and calculate an enrichment z-score for each window.

The noise model accurately captures the observed genome-wide enrichment statistics

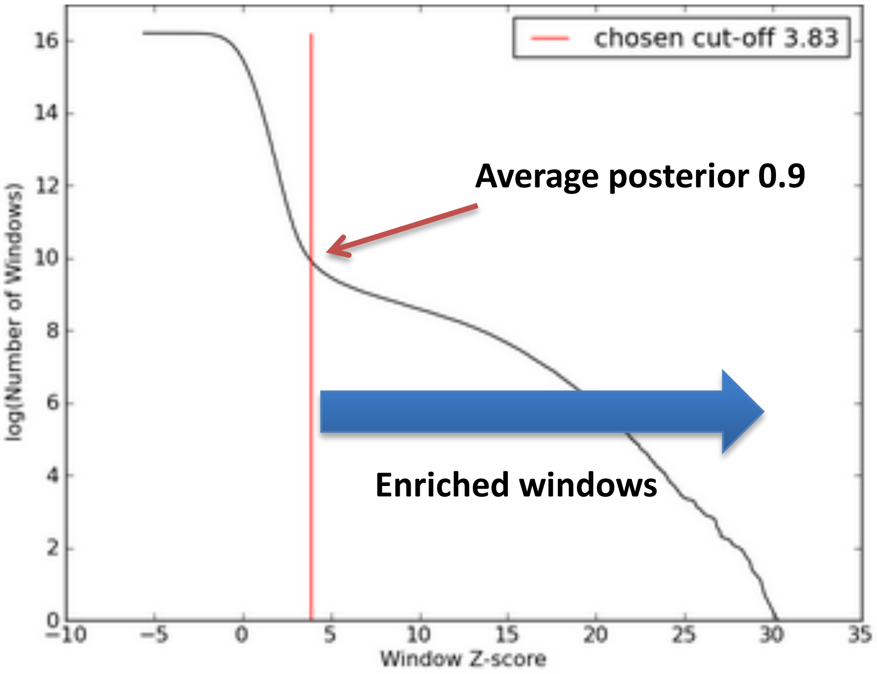
Z-score for each window:

$$z_i = \frac{\log\left[\frac{n_i}{N}\right] - \log\left[\frac{m_i}{M}\right] - \mu}{\sqrt{2\sigma^2 + \frac{1}{n_i} + \frac{1}{m_i}}}$$

Distribution of z-scores



Reverse cumulative distribution of z-scores

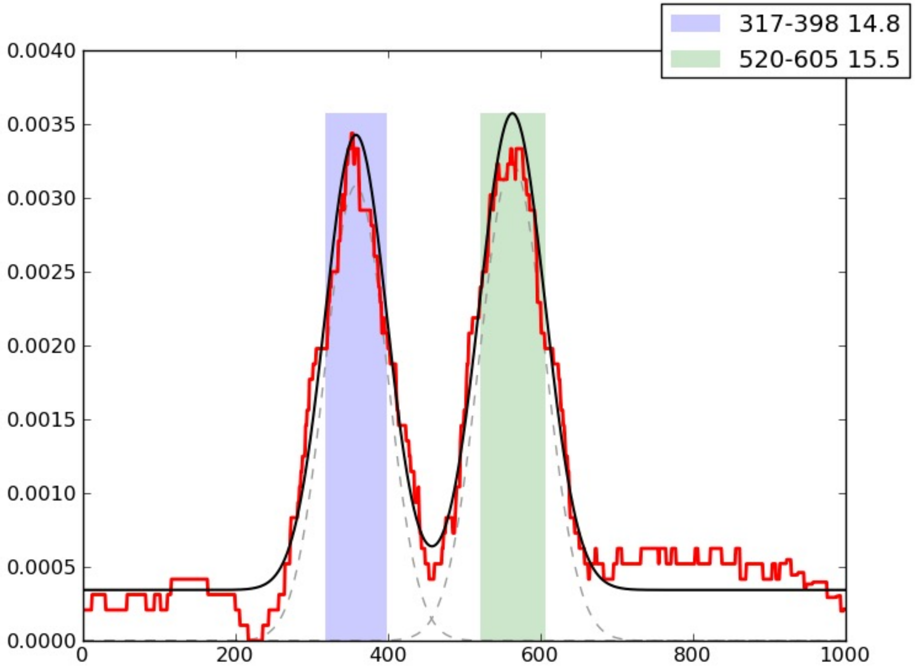


Average posterior 0.9

Enriched windows

As far as we are aware, **CRUNCH has the only peak-finder that demonstrably matches the data's statistics**.

Automated decomposition of each enriched region into individual binding peaks using a Gaussian mixture

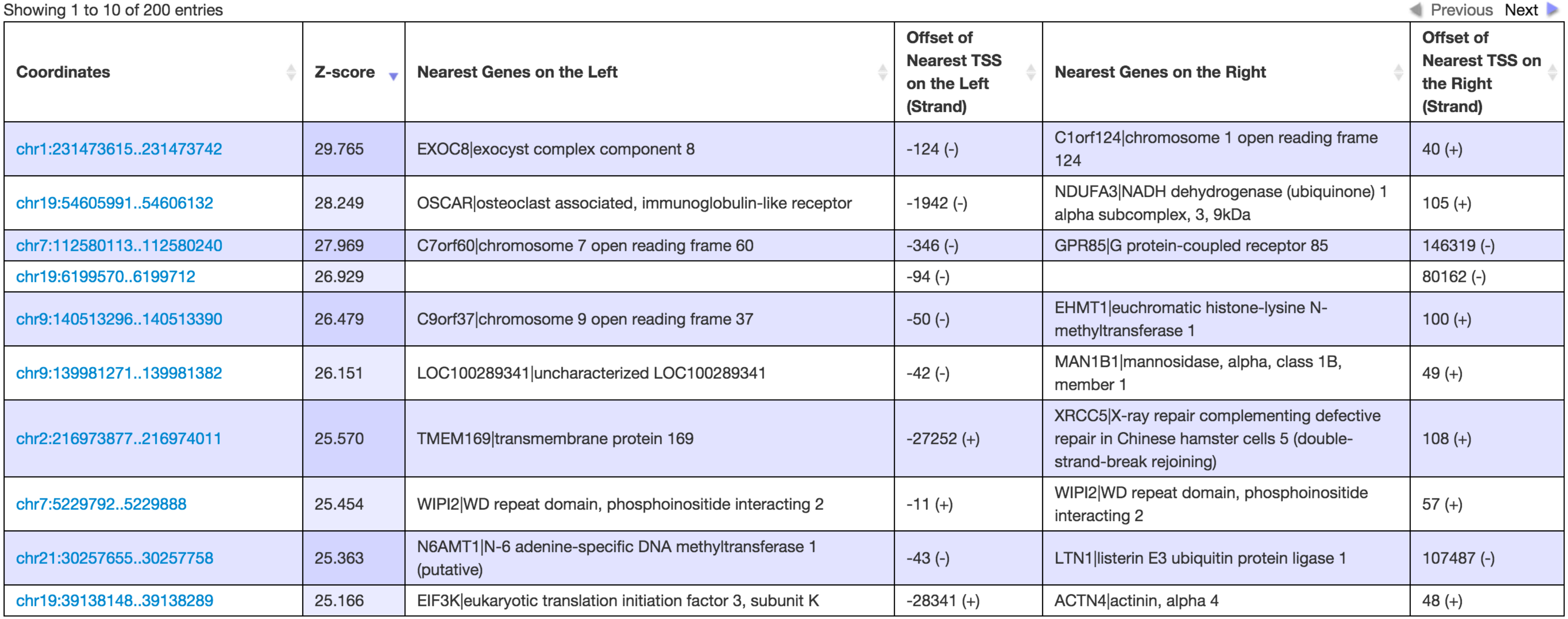


- Read-density profile modeled as a *Gaussian mixture* plus background read-density.
- Informative prior on peak-width from fragment sizes.
- Each individual peak assigned a final significance.
- Final individual peaks sorted by their significance.
- **Peak annotation**: Identify nearest neighboring genes.

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

Sorted list of annotated peaks



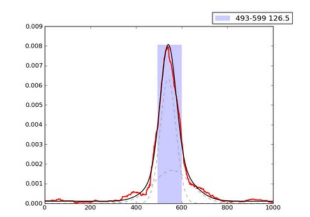
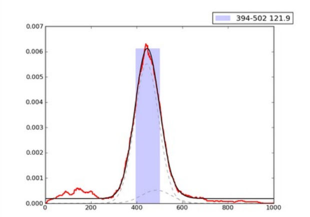
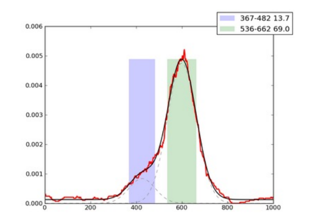
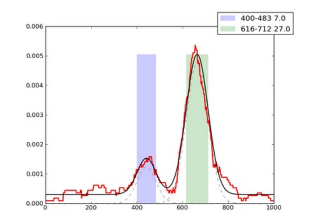
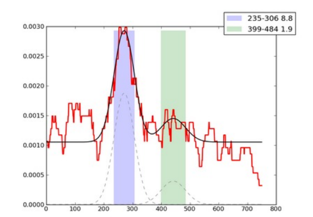
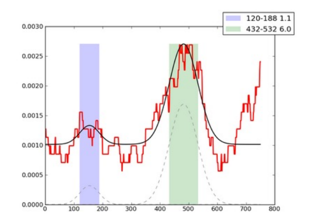
Showing 1 to 10 of 200 entries

◀ Previous Next ▶

Coordinates	Z-score ▼	Nearest Genes on the Left	Offset of Nearest TSS on the Left (Strand)	Nearest Genes on the Right	Offset of Nearest TSS on the Right (Strand)
chr1:231473615..231473742	29.765	EXOC8|exocyst complex component 8	-124 (-)	C1orf124|chromosome 1 open reading frame 124	40 (+)
chr19:54605991..54606132	28.249	OSCAR|osteoclast associated, immunoglobulin-like receptor	-1942 (-)	NDUFA3|NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 3, 9kDa	105 (+)
chr7:112580113..112580240	27.969	C7orf60|chromosome 7 open reading frame 60	-346 (-)	GPR85|G protein-coupled receptor 85	146319 (-)
chr19:6199570..6199712	26.929		-94 (-)		80162 (-)
chr9:140513296..140513390	26.479	C9orf37|chromosome 9 open reading frame 37	-50 (-)	EHMT1|euchromatic histone-lysine N-methyltransferase 1	100 (+)
chr9:139981271..139981382	26.151	LOC100289341|uncharacterized LOC100289341	-42 (-)	MAN1B1|mannosidase, alpha, class 1B, member 1	49 (+)
chr2:216973877..216974011	25.570	TMEM169|transmembrane protein 169	-27252 (+)	XRCC5|X-ray repair complementing defective repair in Chinese hamster cells 5 (double-strand-break rejoining)	108 (+)
chr7:5229792..5229888	25.454	WIPI2|WD repeat domain, phosphoinositide interacting 2	-11 (+)	WIPI2|WD repeat domain, phosphoinositide interacting 2	57 (+)
chr21:30257655..30257758	25.363	N6AMT1|N-6 adenine-specific DNA methyltransferase 1 (putative)	-43 (-)	LTN1|listerin E3 ubiquitin protein ligase 1	107487 (-)
chr19:39138148..39138289	25.166	EIF3K|eukaryotic translation initiation factor 3, subunit K	-28341 (+)	ACTN4|actinin, alpha 4	48 (+)

Examples of peaks fitted within regions

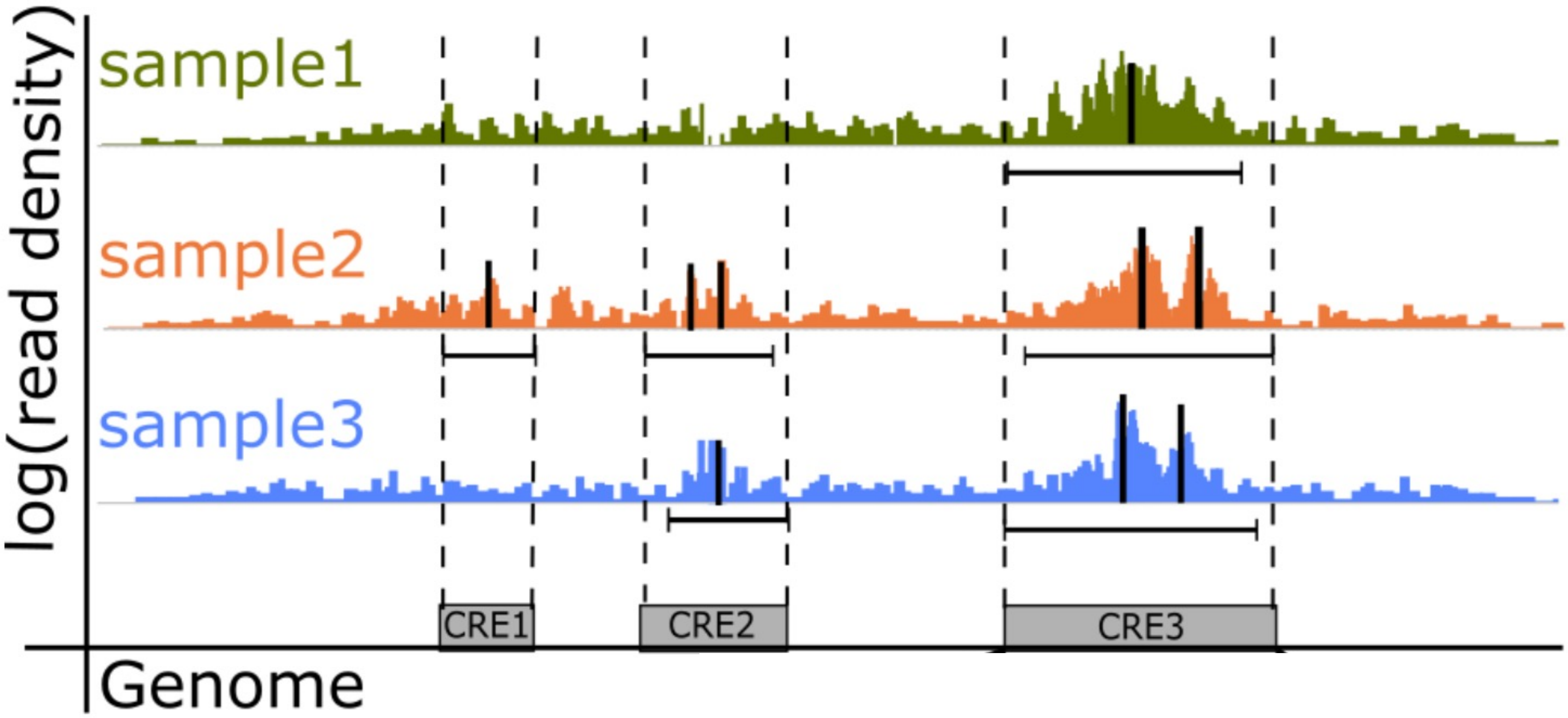
hide

Example from top 5% of regions	Example from top 10% of regions	Example from top 20% of regions	Example from top 40% of regions	Example from top 60% of regions	Example from top 90% of regions
					

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

Defining a set of CREs and their signal across samples



- The CRUNCH pipe-line is used to identify peaks within each sample.
- All peaks from all samples with centers within 75bp are fused into CREs.
- 90% of all CREs are less than 500bp in datasets processed so far.
- Typically on the order of 100'000 CREs genome-wide in a given dataset.

CRE signal strength across samples

Signal strength is defined as the log-ratio of the read-density in the foreground sample relative to a `background' sample:

$$S_{cs} = \log\left(\frac{f_{cs}}{F_s} \cdot \tilde{F} + 1\right) - \log\left(\frac{b_{cs}}{B_s} \cdot \tilde{F} + 1\right)$$

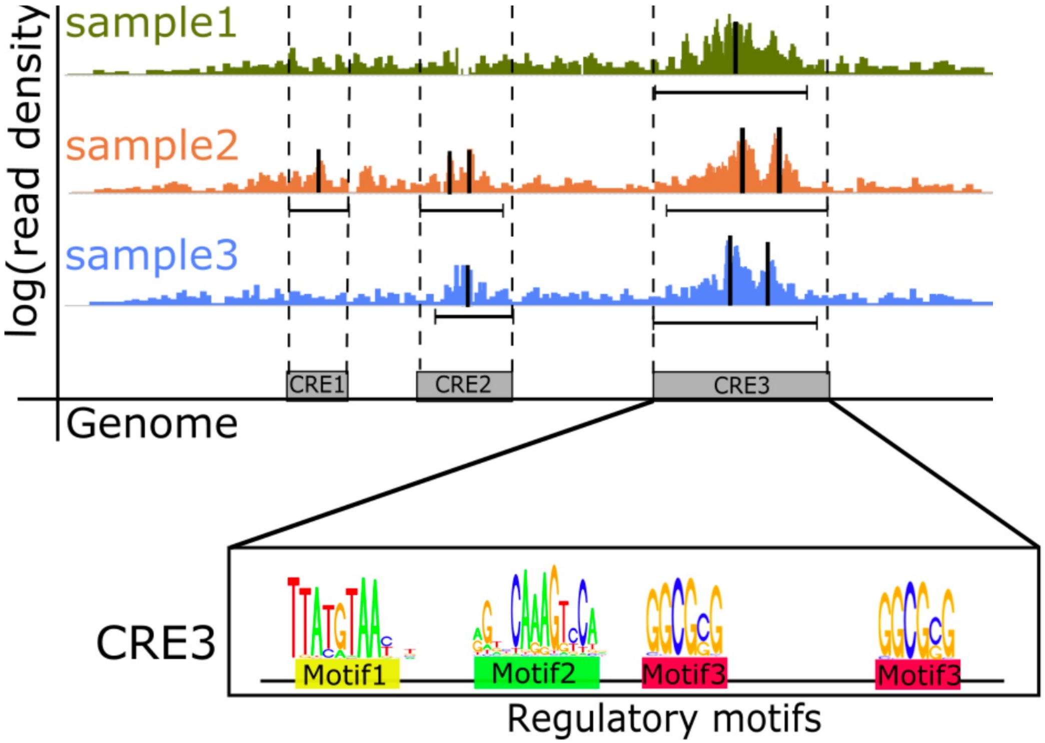
- S_{cs} = Signal of CRE *c* in sample *s.*
- f_{cs} = Number of reads from sample *s* falling in CRE *c.*
- F_s = Total number of reads in sample *s*.
- $\tilde{F}$ = Median number of total reads across samples.
- b_{cs} = Number of *background* reads from sample *s* falling in CRE *c.*
- B_s = Total number of reads in sample *s*.

Background

- For ChIP-seq: Provided background samples of input DNA (or reference ChIP-seq background sample that we have precalculated).

- For ATAC-seq/DNase-seq: A simple *uniform distribution* of background read counts.

TFBS annotation in CREs



- We use our curated collection of ~500 motif groups representing ~600 mammalian TFs.
- We use MotEvo to predict TFBSs for each motif *m* in each CRE *c*.
- The TFBS predictions are summarized in the sitecount matrix:

$$N_{cm} = \text{Sum of the posteriors of sites for motif } m \text{ in CRE } c.$$

MARA model for CREs

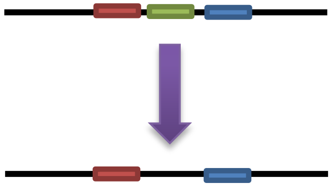
- We employ the MARA model *exactly* as it is performed for gene expression data, i.e. we fit the model:

$$S_{cs} = \sum_{m} N_{cm} \cdot A_{ms} + \tilde{c}_c + c_s + noise$$

- A_{ms} = Average effect on CRE signal in sample *s* from removing 1 binding site for motif *m*.
- We again use a *Gausian prior* on the motif activities (ridge regression) and optimize its parameter using 80/20 cross-validation.
- Motif significances are: $$z_m = \sqrt{\frac{1}{S}\sum_s \left(\frac{A'_{ms}}{\delta A'_{ms}}\right)^2}$$
- Target scores are (changes in chi-squared of the fit): $$\zeta_{cm} = \frac{\sum_s \chi^2_{csm} - \chi^2_{cs}}{\langle \chi^2 \rangle}$$

Predicting targets of each motif (conceptual)

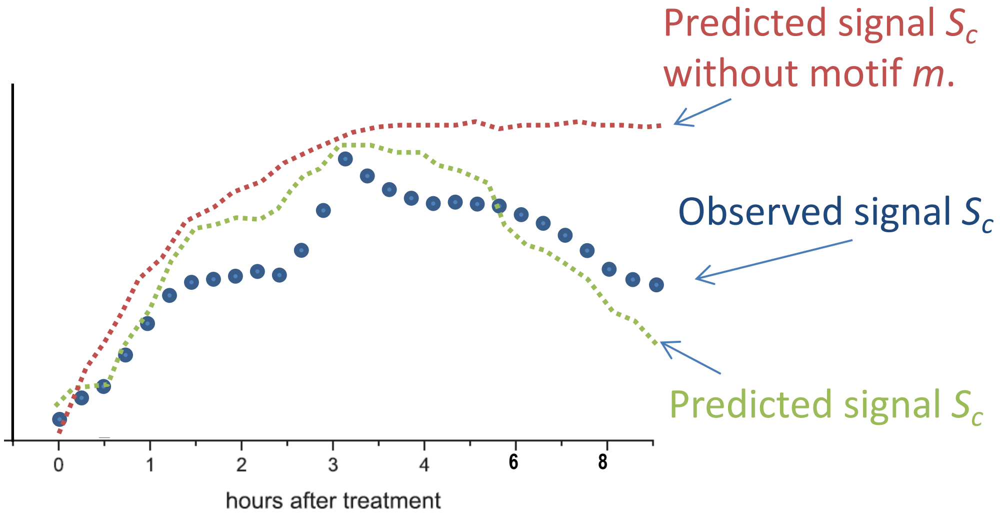
- For each motif, select promoters with predicted sites, i.e with $N_{cm} > 0$



- *Mutate* CRE *c* to *remove* the binding site(s) for motif *m*: $N_{cm} \to 0$
- Updated site-count matrix: $N \to \tilde{N}$
- Log-likelihood ratio of fitting *all data* with N versus the mutated $\tilde{N}$:

$$\zeta_{cm} = \log\left[\frac{\int dAP(S|N,A)}{\int dAP(S|\tilde{N},A)}\right]$$

Quantifies the contribution of motif *m* to explaining the signal across samples of CRE *c*.

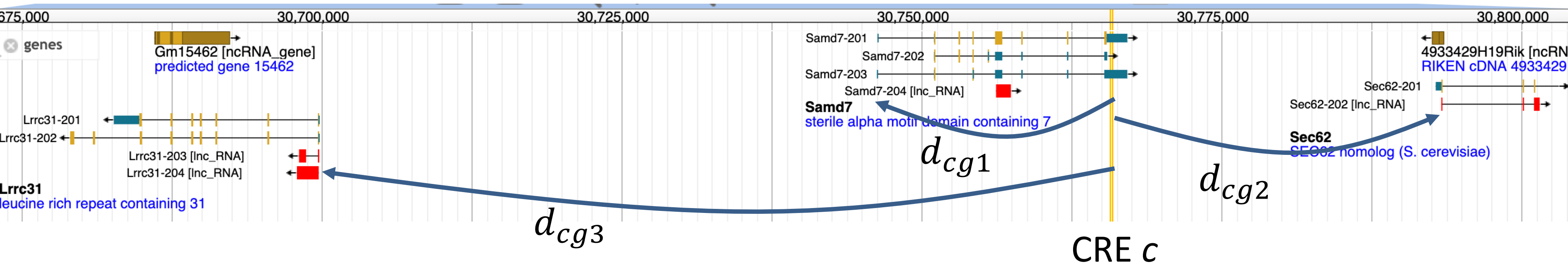


The log-likelihood ratio ζ_{cm} quantifies how much the quality of the fit is reduced when the sites for motif *m* in CRE *c* are removed.

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

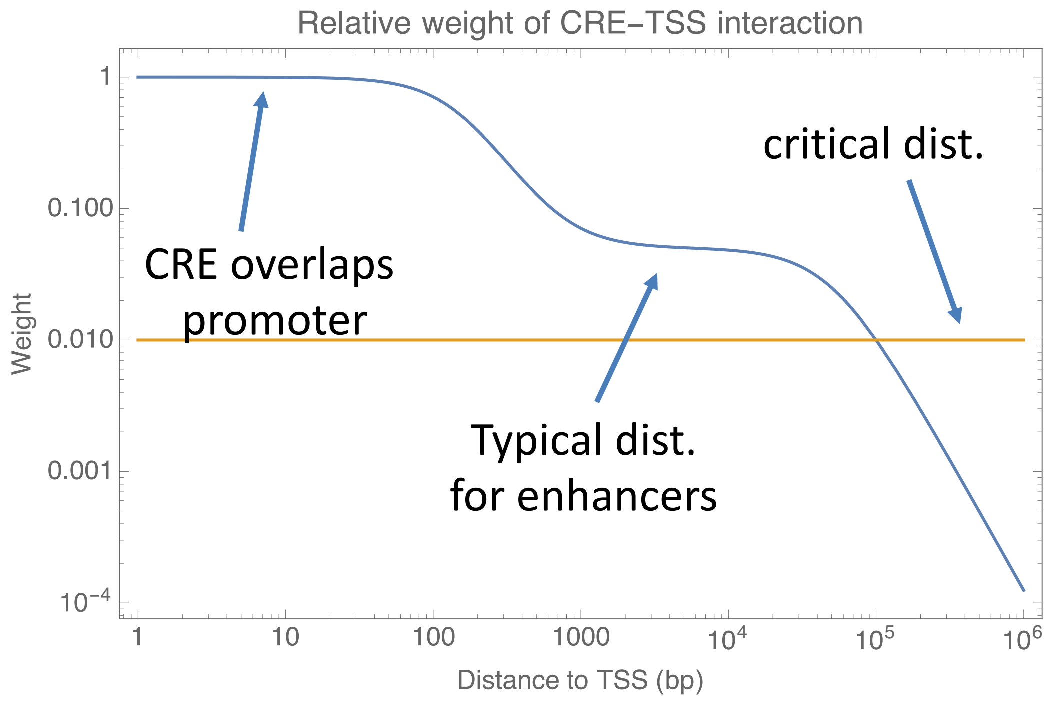
Associating CREs with genes

SIB
Swiss Institute of Bioinformatics



Distance based weights between CRE and TSS of nearby genes:

$$w_c(G) = \frac{0.95}{1+(\frac{d_{CG}}{d_p})^2} + \frac{0.05}{1+(\frac{d_{CG}}{d_d})^2}$$



CRE overlaps promoter

Typical dist. for enhancers

critical dist.

Probability of associating CRE *c* with gene G based on relative weights:

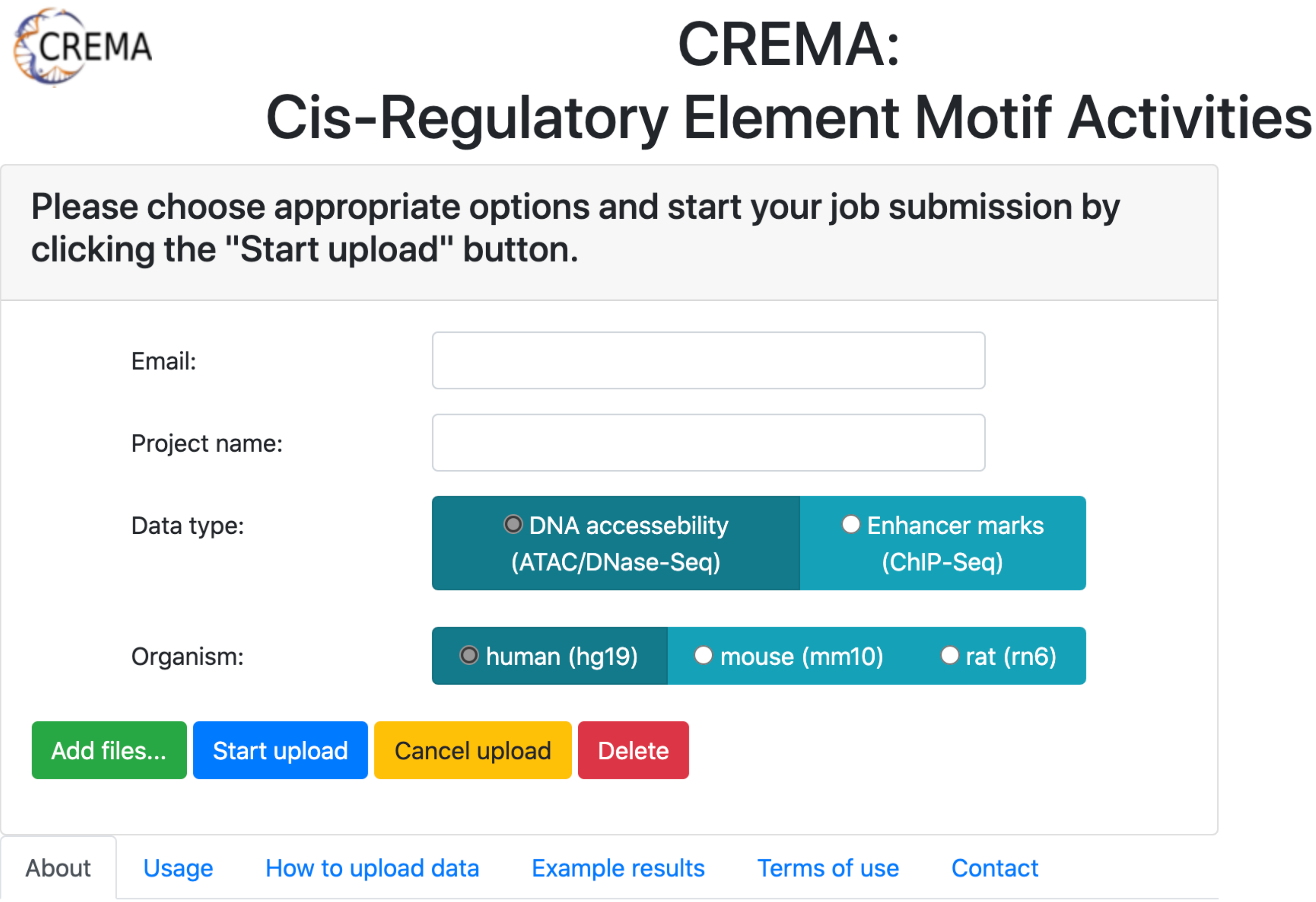
$$P_c(G) = \frac{w_c(G)}{w_0 + \sum_g w_c(g)}$$

$$w_0 = 0.01$$

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

crema.unibas.ch

SIB
Swiss Institute of
Bioinformatics



CREMA

CREMA:
Cis-Regulatory Element Motif Activities

Please choose appropriate options and start your job submission by clicking the "Start upload" button.

Email:

Project name:

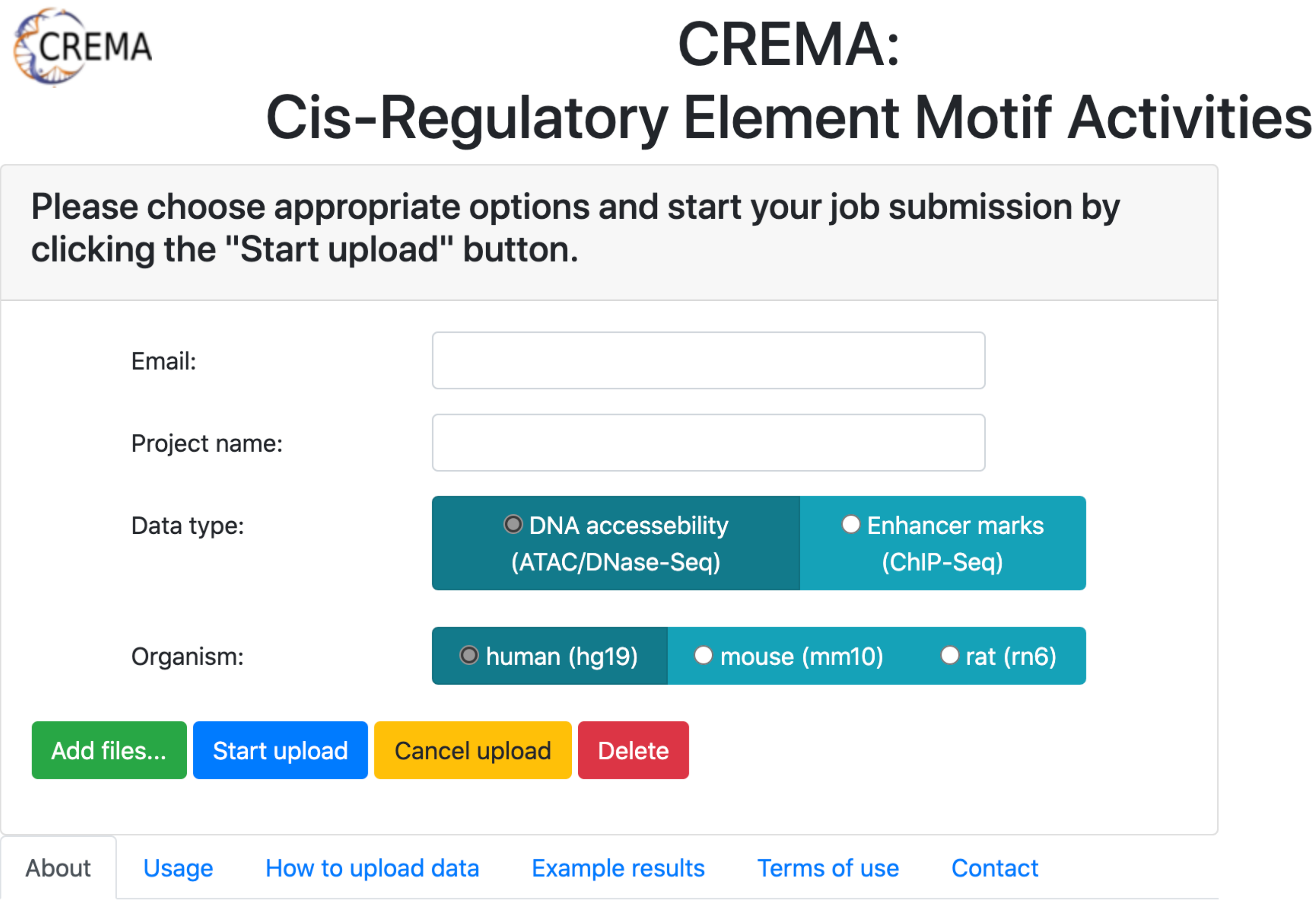
Data type: ● DNA accessibility (ATAC/DNase-Seq) ○ Enhancer marks (ChIP-Seq)

Organism: ● human (hg19) ○ mouse (mm10) ○ rat (rn6)

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

crema.unibas.ch

SIB
Swiss Institute of
Bioinformatics



CREMA:
Cis-Regulatory Element Motif Activities

Please choose appropriate options and start your job submission by clicking the "Start upload" button.

Email:

Project name:

Data type: ● DNA accessibility (ATAC/DNase-Seq) ○ Enhancer marks (ChIP-Seq)

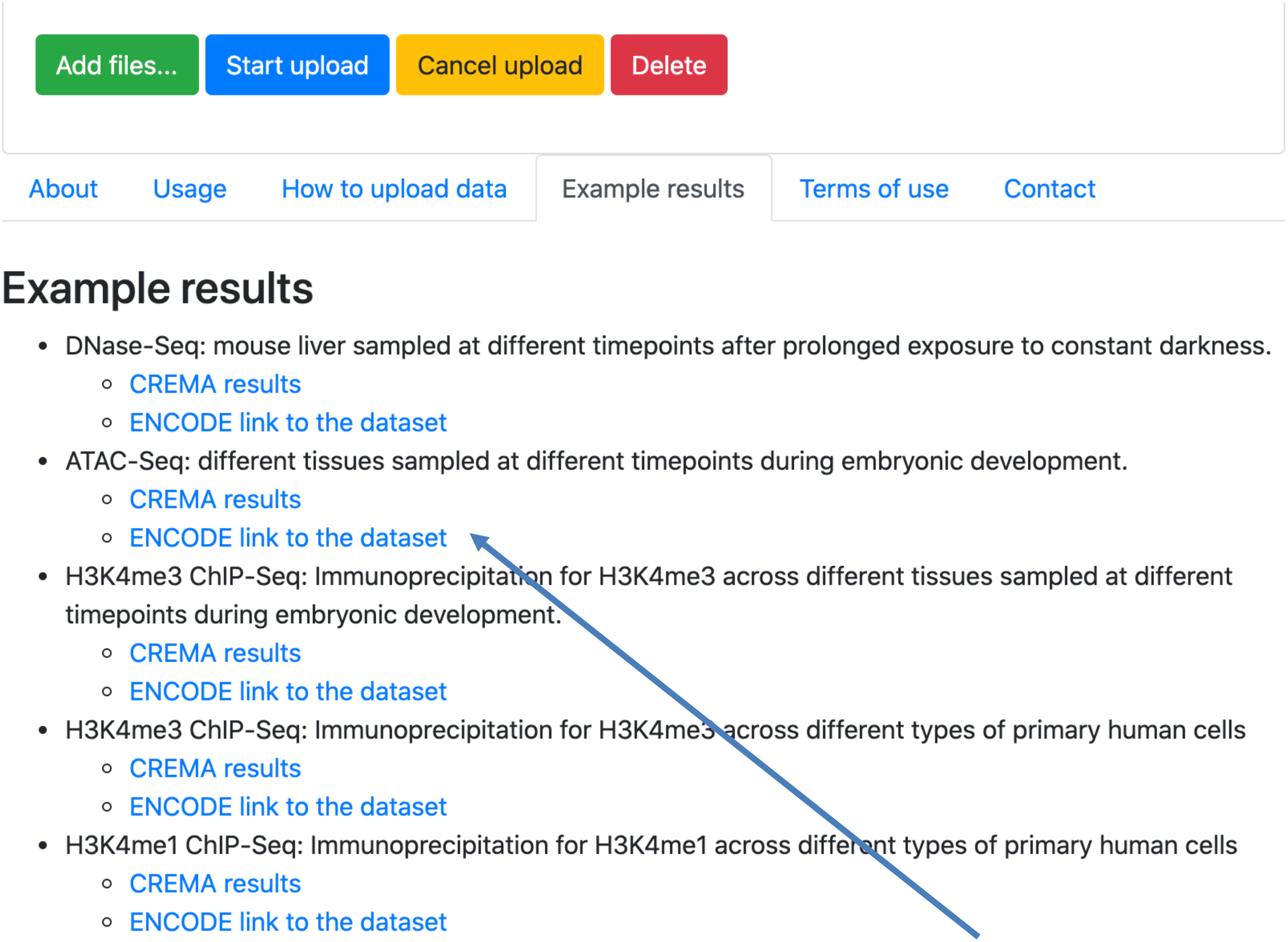
Organism: ● human (hg19) ○ mouse (mm10) ○ rat (rn6)

Click on example results

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

crema.unibas.ch

SIB
Swiss Institute of
Bioinformatics

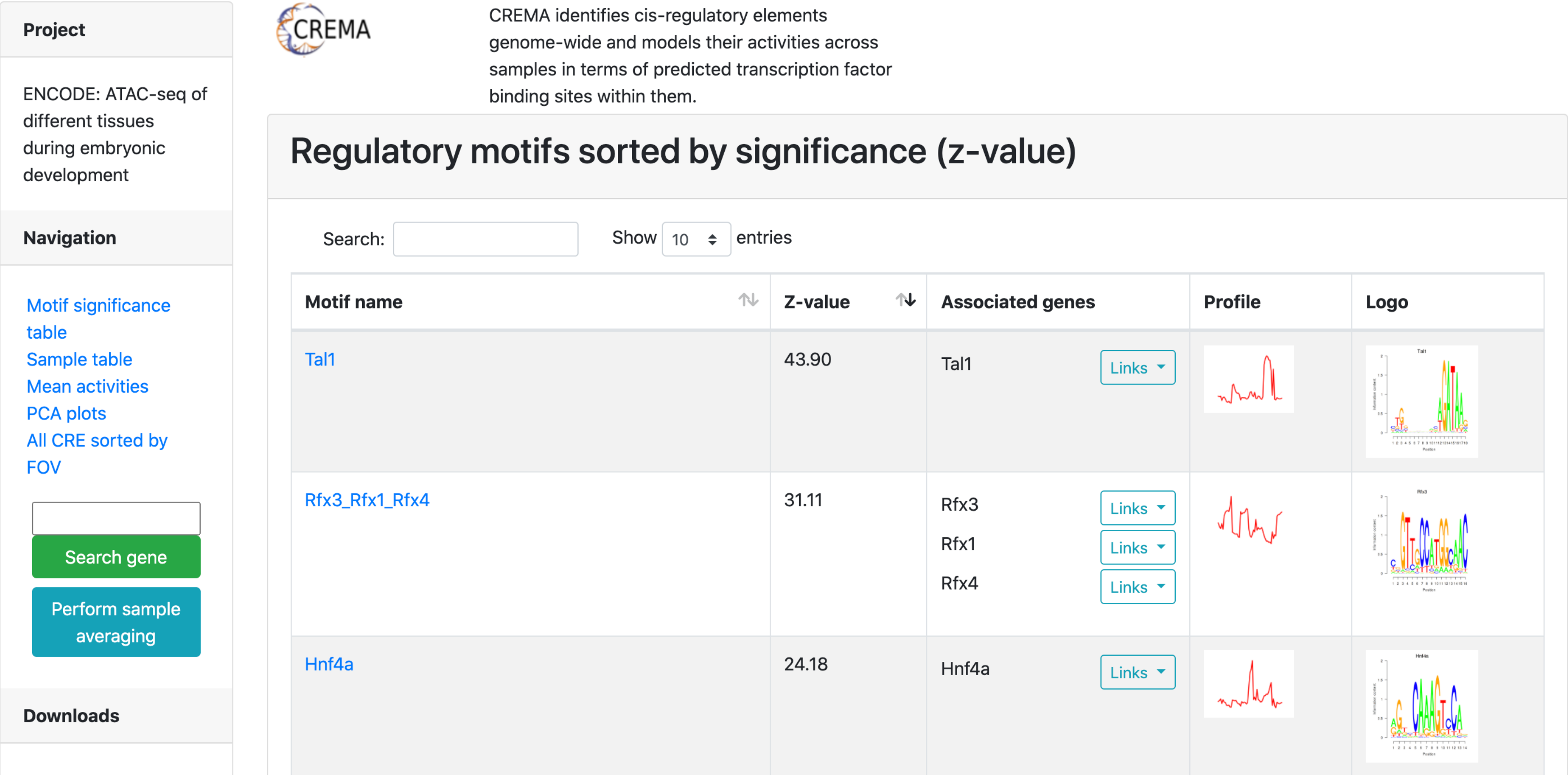


Example results

- DNase-Seq: mouse liver sampled at different timepoints after prolonged exposure to constant darkness.
  - CREMA results
  - ENCODE link to the dataset
- ATAC-Seq: different tissues sampled at different timepoints during embryonic development.
  - CREMA results
  - ENCODE link to the dataset
- H3K4me3 ChIP-Seq: Immunoprecipitation for H3K4me3 across different tissues sampled at different timepoints during embryonic development.
  - CREMA results
  - ENCODE link to the dataset
- H3K4me3 ChIP-Seq: Immunoprecipitation for H3K4me3 across different types of primary human cells
  - CREMA results
  - ENCODE link to the dataset
- H3K4me1 ChIP-Seq: Immunoprecipitation for H3K4me1 across different types of primary human cells
  - CREMA results
  - ENCODE link to the dataset

- Chromatin accessibility in the developing mouse embryo.
- ATAC-seq from the Bing Ren lab (ENCODE).
- 10 tissues, multiple time points in each.

Results chromatin accessibility in mouse development



Project

ENCODE: ATAC-seq of different tissues during embryonic development

Navigation

- Motif significance table
- Sample table
- Mean activities
- PCA plots
- All CRE sorted by FOV

Search gene

Perform sample averaging

Downloads

CREMA identifies cis-regulatory elements genome-wide and models their activities across samples in terms of predicted transcription factor binding sites within them.

Regulatory motifs sorted by significance (z-value)

Search:

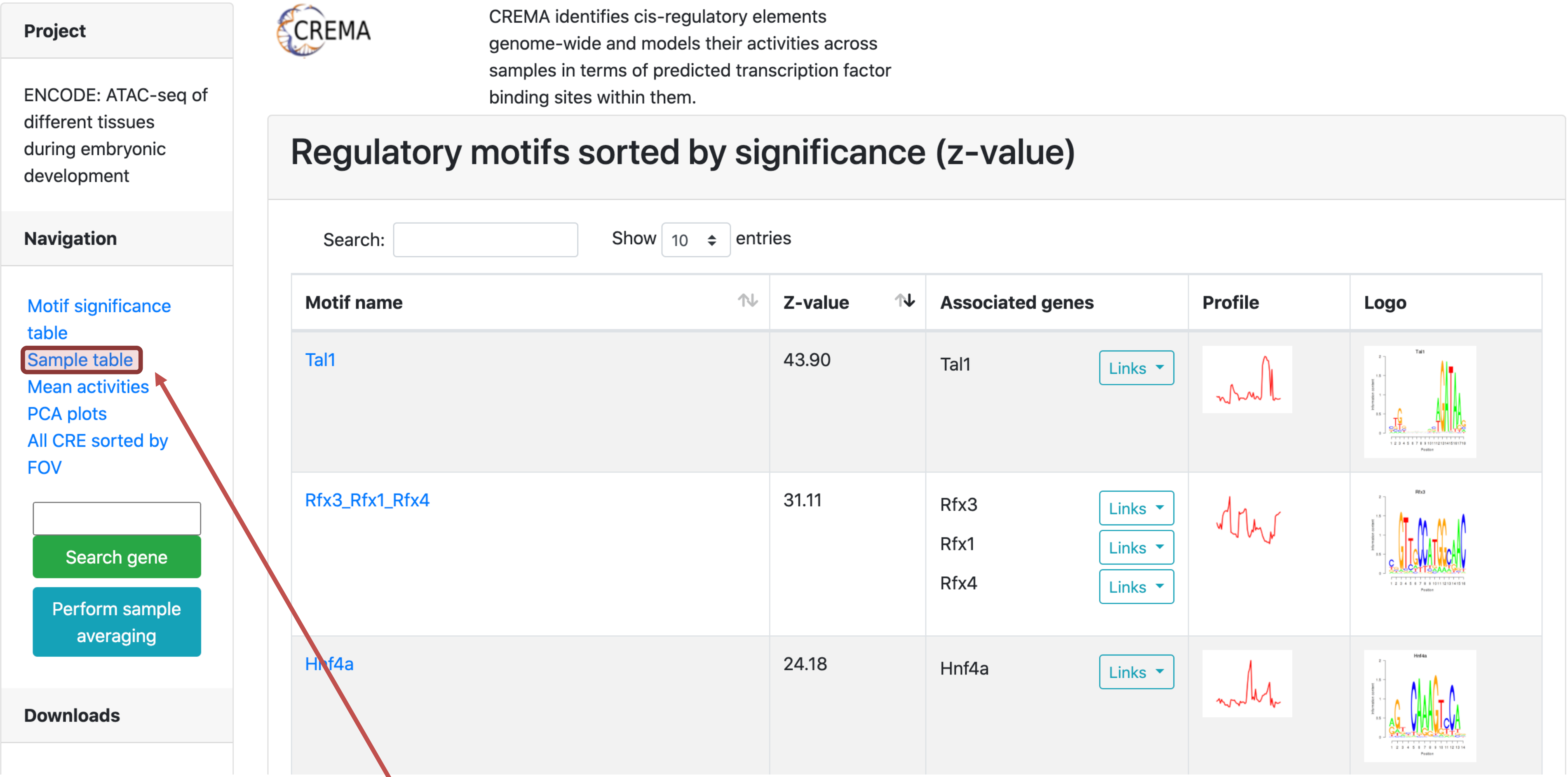
Show 10 entries

Motif name	Z-value	Associated genes	Profile	Logo
Tal1	43.90	Tal1 Links		
Rfx3_Rfx1_Rfx4	31.11	Rfx3 Links Rfx1 Links Rfx4 Links		
Hnf4a	24.18	Hnf4a Links		

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

Results chromatin accessibility in mouse development



Project

ENCODE: ATAC-seq of different tissues during embryonic development

Navigation

- Motif significance table
- Sample table
- Mean activities
- PCA plots
- All CRE sorted by FOV

Downloads

CREMA

CREMA identifies cis-regulatory elements genome-wide and models their activities across samples in terms of predicted transcription factor binding sites within them.

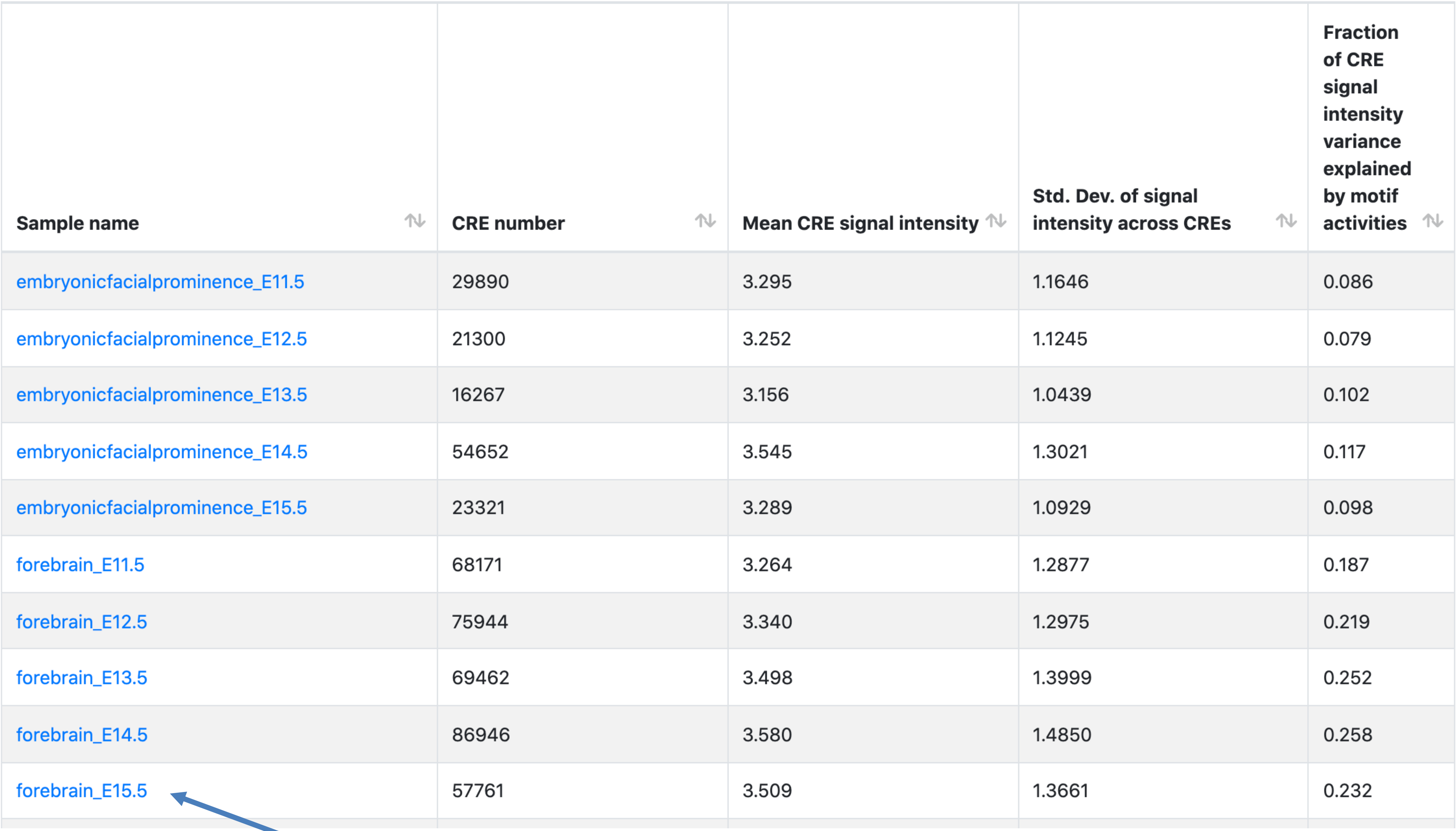
Regulatory motifs sorted by significance (z-value)

Search: Show 10 entries

Motif name	Z-value	Associated genes	Profile	Logo
Tal1	43.90	Tal1 Links		
Rfx3_Rfx1_Rfx4	31.11	Rfx3 Links Rfx1 Links Rfx4 Links		
Hnf4a	24.18	Hnf4a Links		

Information per sample.

List of samples with CRE summary statistics



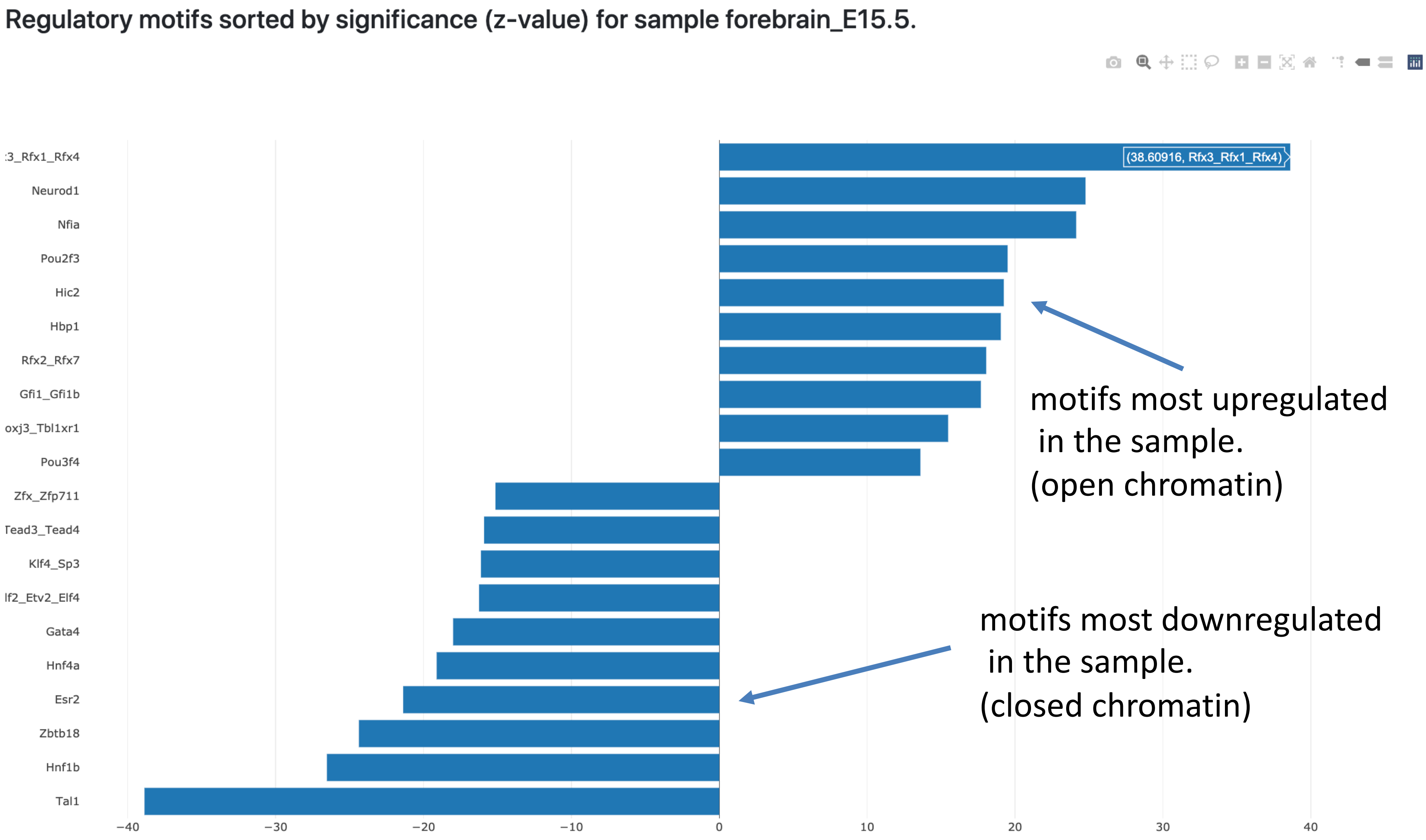
Sample name	CRE number	Mean CRE signal intensity	Std. Dev. of signal intensity across CREs	Fraction of CRE signal intensity variance explained by motif activities
embryonicfacialprominence_E11.5	29890	3.295	1.1646	0.086
embryonicfacialprominence_E12.5	21300	3.252	1.1245	0.079
embryonicfacialprominence_E13.5	16267	3.156	1.0439	0.102
embryonicfacialprominence_E14.5	54652	3.545	1.3021	0.117
embryonicfacialprominence_E15.5	23321	3.289	1.0929	0.098
forebrain_E11.5	68171	3.264	1.2877	0.187
forebrain_E12.5	75944	3.340	1.2975	0.219
forebrain_E13.5	69462	3.498	1.3999	0.252
forebrain_E14.5	86946	3.580	1.4850	0.258
forebrain_E15.5	57761	3.509	1.3661	0.232

Links with more information about each sample.

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

Most significant motifs for forebrain_E15.5



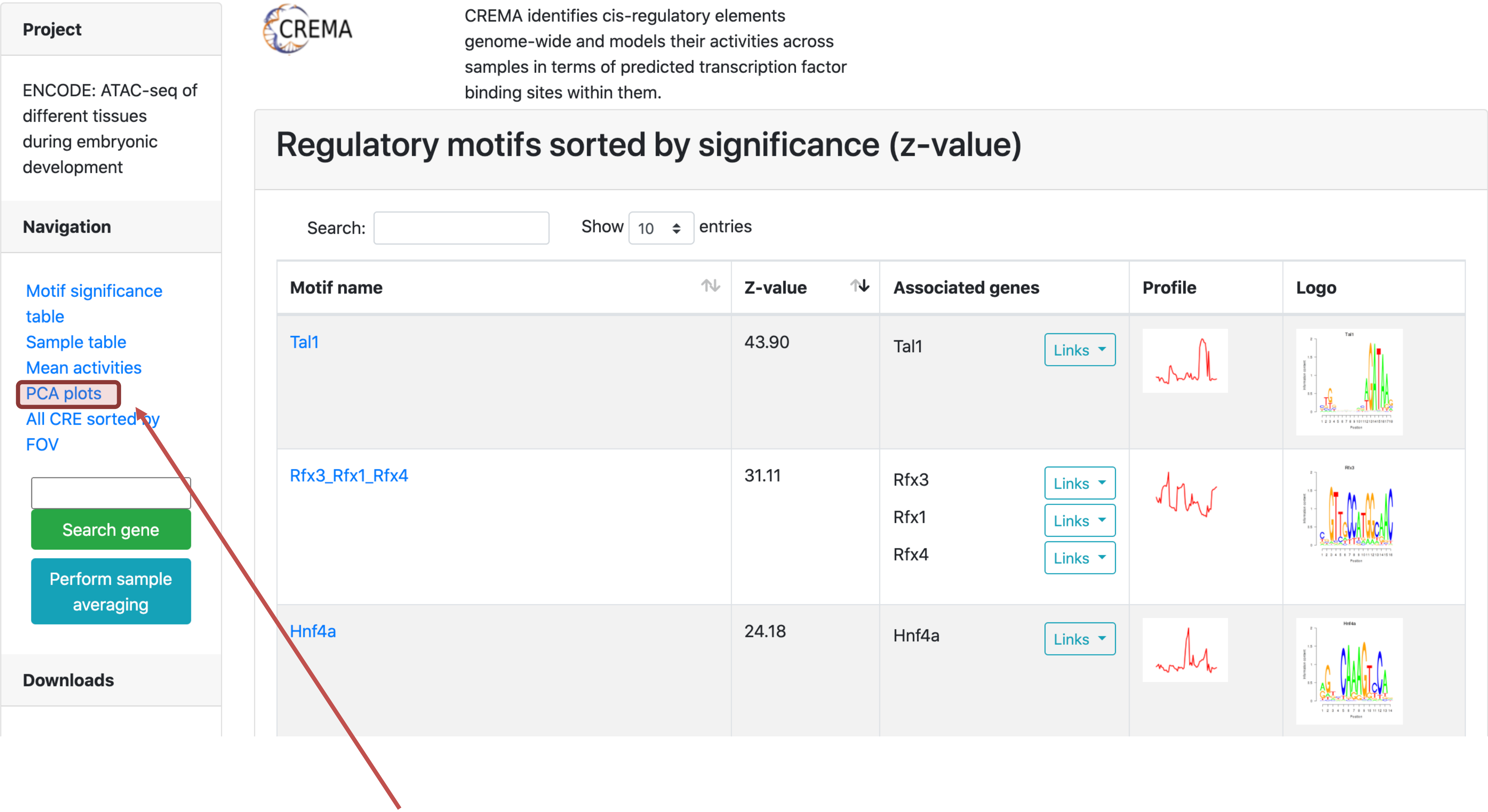
Regulatory motifs sorted by significance (z-value) for sample forebrain_E15.5.

motifs most upregulated
in the sample.
(open chromatin)

motifs most downregulated
in the sample.
(closed chromatin)

z-value of motif activity

Results chromatin accessibility in mouse development



Project

ENCODE: ATAC-seq of different tissues during embryonic development

Navigation

- Motif significance table
- Sample table
- Mean activities
- PCA plots
- All CRE sorted by FOV

Downloads

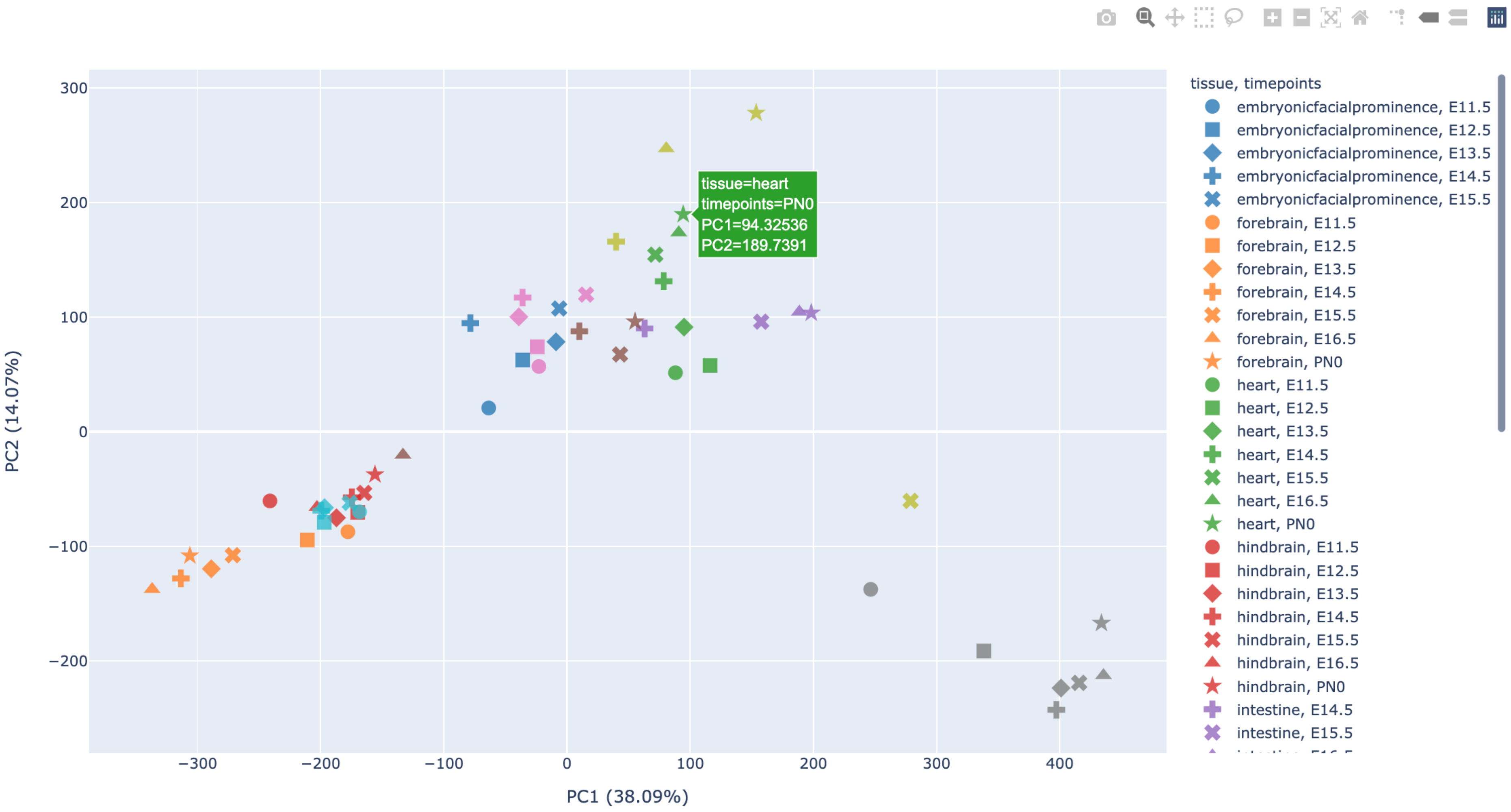
CREMA identifies cis-regulatory elements genome-wide and models their activities across samples in terms of predicted transcription factor binding sites within them.

Regulatory motifs sorted by significance (z-value)

Motif name	Z-value	Associated genes	Profile	Logo
Tal1	43.90	Tal1 Links		
Rfx3_Rfx1_Rfx4	31.11	Rfx3 Links Rfx1 Links Rfx4 Links		
Hnf4a	24.18	Hnf4a Links		

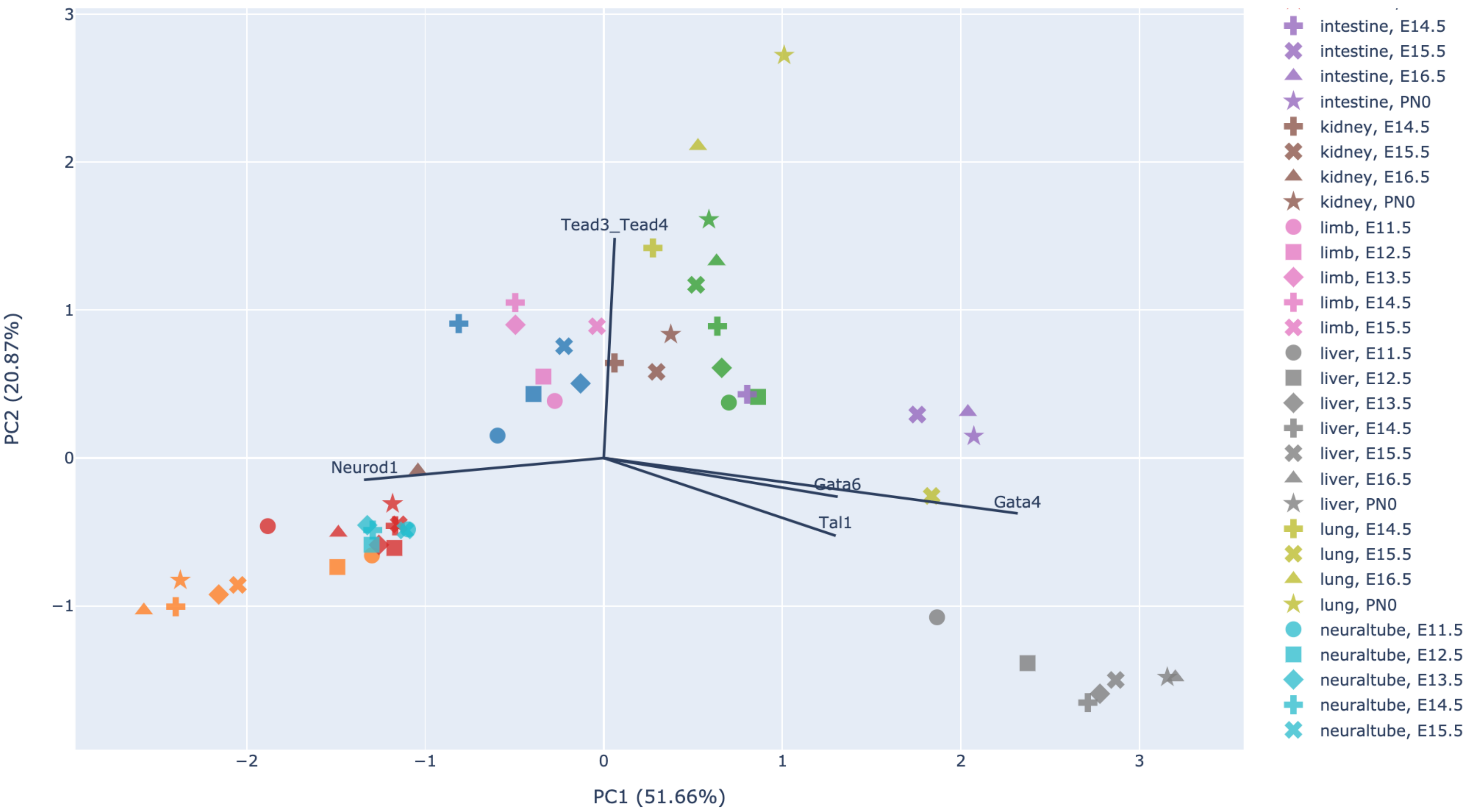
PCA plots summarize the overall structure in the data

PCA of the CRE signal vectors across samples



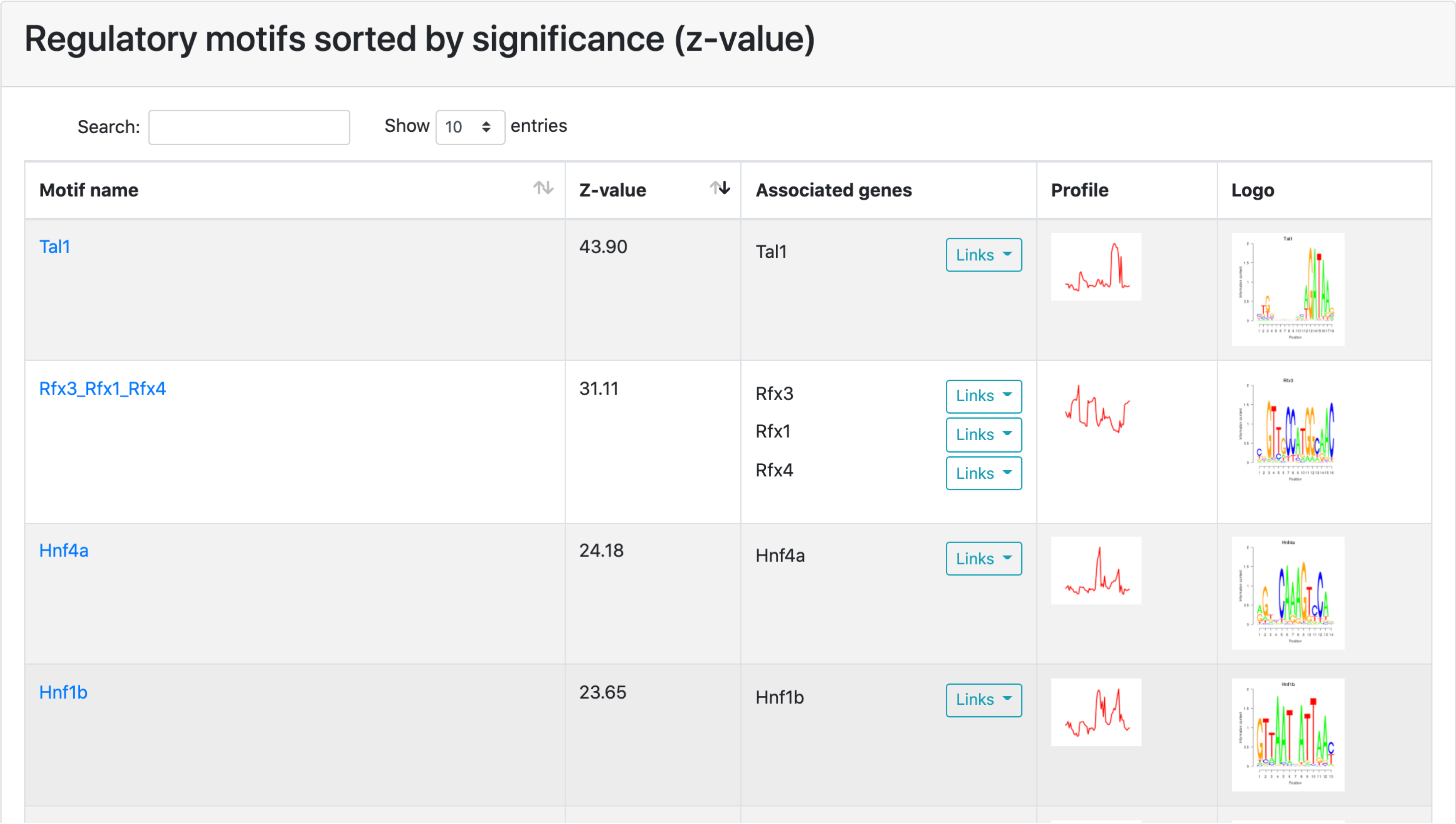
- Interactive figure (mouse over, zoom, screen shot, etc.)
- Colors correspond to tissues.
- Symbols correspond to developmental time.

PCA of the motif activities across samples



- More than 70% of the variance is captured by the first two PCA components.
- Samples tend to move radially outward with developmental time.
- Projections of top motifs onto these two PCA components are indicated.

Motifs sorted by significance
(explaining changes in accessibility across samples)

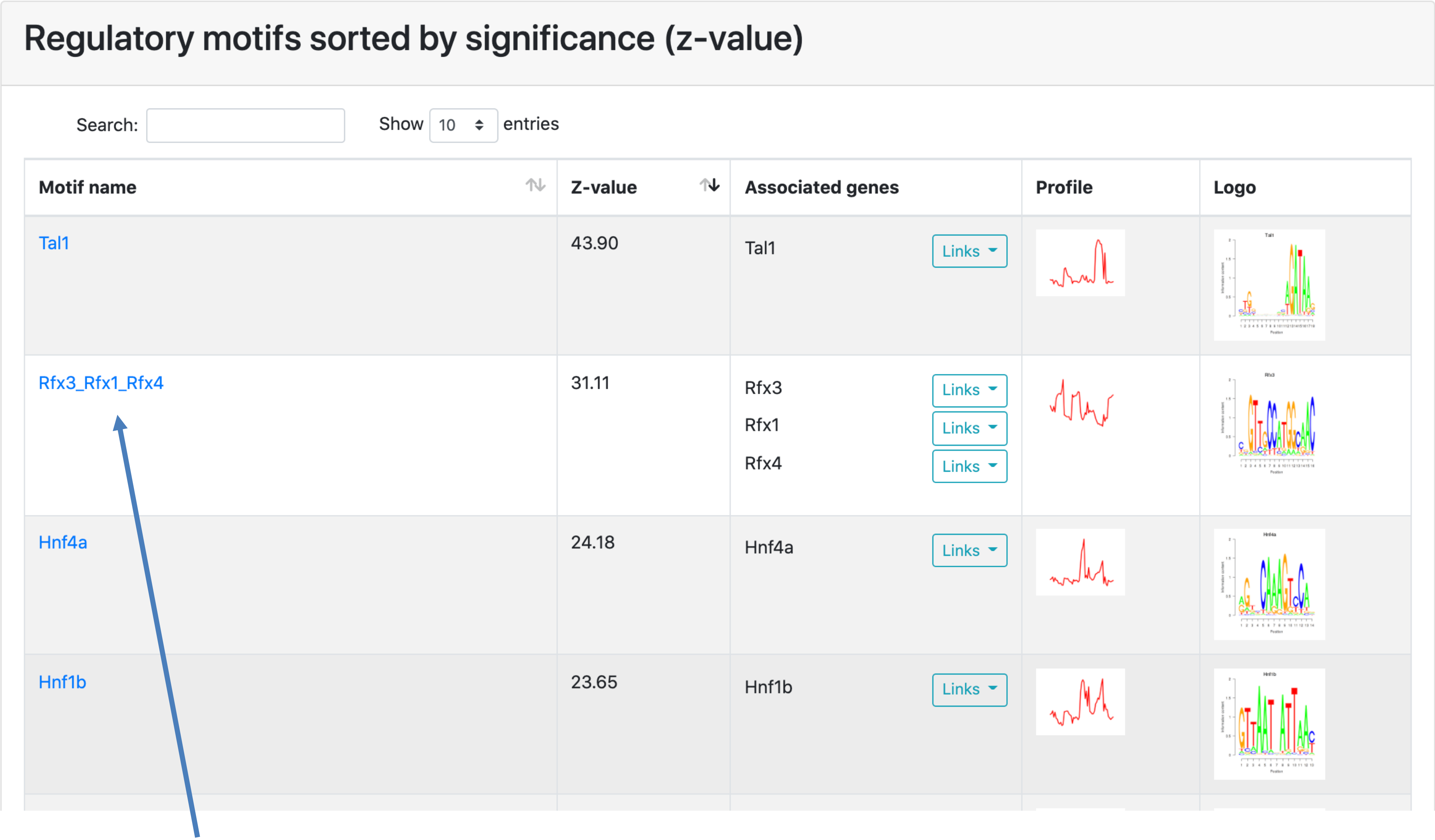


Regulatory motifs sorted by significance (z-value)

Search: Show 10 entries

Motif name	Z-value	Associated genes	Profile	Logo
Tal1	43.90	Tal1 Links ▾		
Rfx3_Rfx1_Rfx4	31.11	Rfx3 Links ▾ Rfx1 Links ▾ Rfx4 Links ▾		
Hnf4a	24.18	Hnf4a Links ▾		
Hnf1b	23.65	Hnf1b Links ▾		

Motifs sorted by significance
(explaining changes in accessibility across samples)



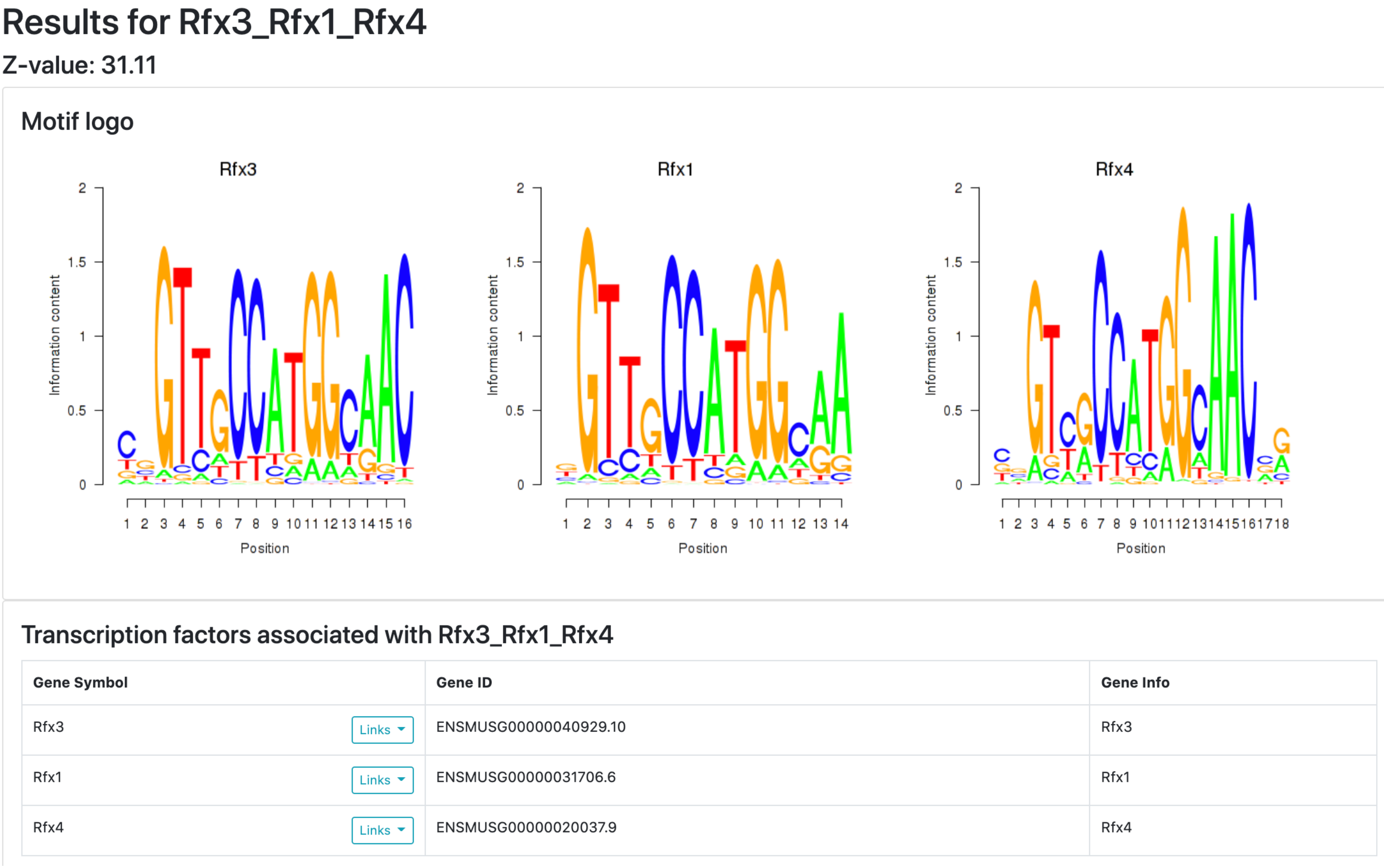
Regulatory motifs sorted by significance (z-value)

Search: Show 10 entries

Motif name	Z-value	Associated genes	Profile	Logo
Tal1	43.90	Tal1 Links		
Rfx3_Rfx1_Rfx4	31.11	Rfx3 Links Rfx1 Links Rfx4 Links		
Hnf4a	24.18	Hnf4a Links		
Hnf1b	23.65	Hnf1b Links		

Rfx motif is second in the list.

Three Rfx TFs bind this motif



Results for Rfx3_Rfx1_Rfx4

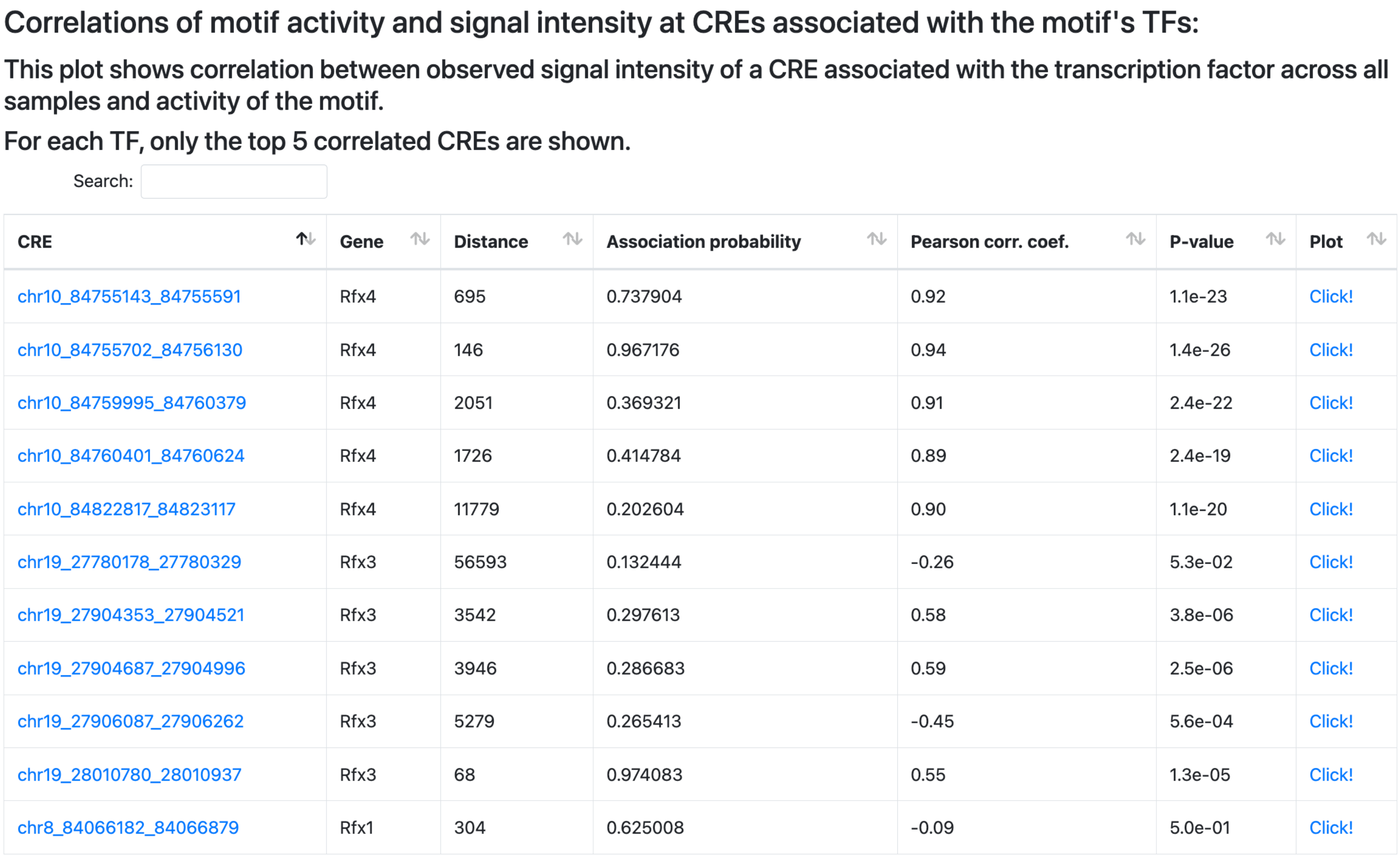
Z-value: 31.11

Motif logo

Transcription factors associated with Rfx3_Rfx1_Rfx4

Gene Symbol	Gene ID	Gene Info
Rfx3 Links ▾	ENSMUSG00000040929.10	Rfx3
Rfx1 Links ▾	ENSMUSG00000031706.6	Rfx1
Rfx4 Links ▾	ENSMUSG00000020037.9	Rfx4

CREs near the TFs associated with the motif



Correlations of motif activity and signal intensity at CREs associated with the motif's TFs:

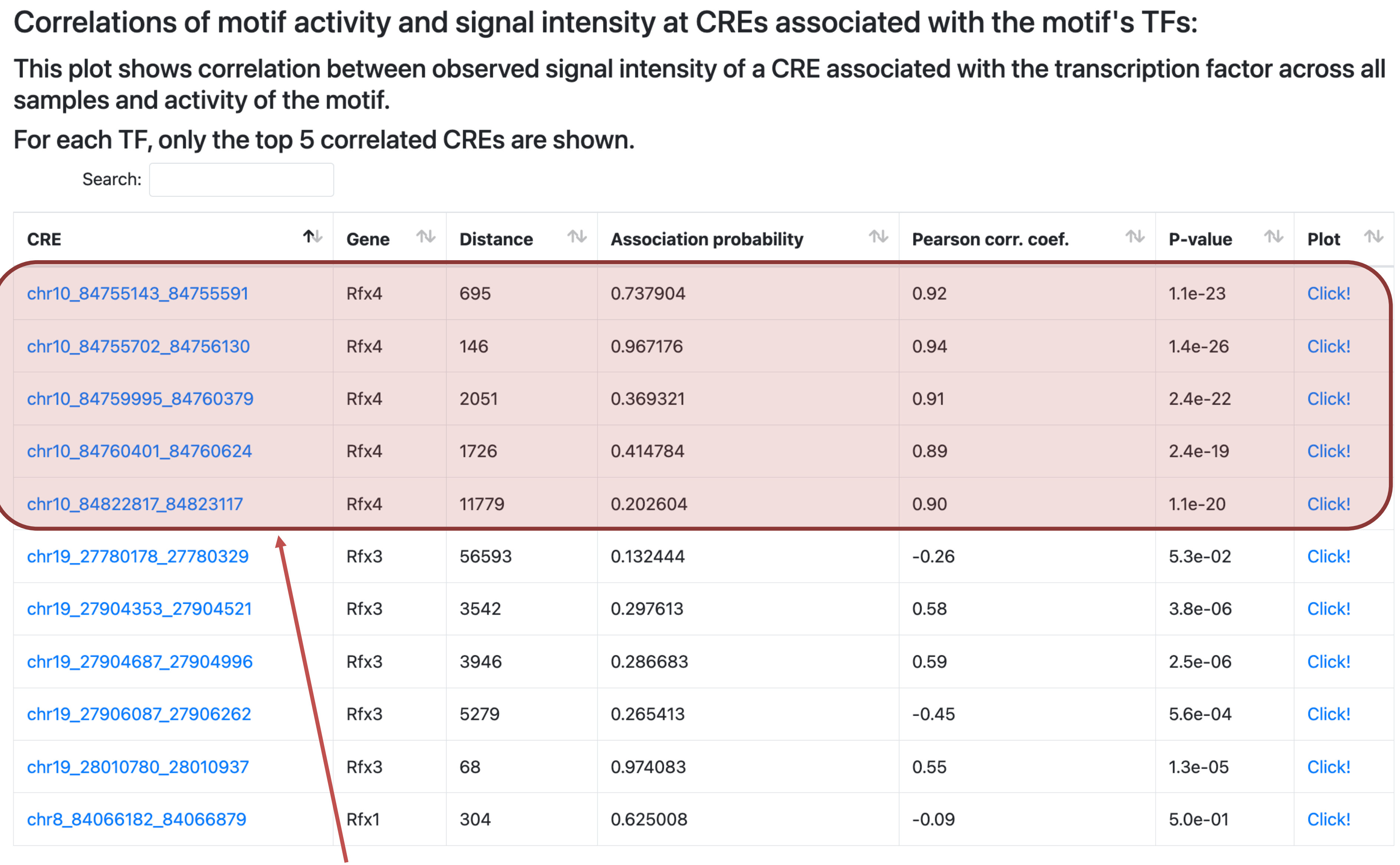
This plot shows correlation between observed signal intensity of a CRE associated with the transcription factor across all samples and activity of the motif.

For each TF, only the top 5 correlated CREs are shown.

Search:

CRE	Gene	Distance	Association probability	Pearson corr. coef.	P-value	Plot
chr10_84755143_84755591	Rfx4	695	0.737904	0.92	1.1e-23	Click!
chr10_84755702_84756130	Rfx4	146	0.967176	0.94	1.4e-26	Click!
chr10_84759995_84760379	Rfx4	2051	0.369321	0.91	2.4e-22	Click!
chr10_84760401_84760624	Rfx4	1726	0.414784	0.89	2.4e-19	Click!
chr10_84822817_84823117	Rfx4	11779	0.202604	0.90	1.1e-20	Click!
chr19_27780178_27780329	Rfx3	56593	0.132444	-0.26	5.3e-02	Click!
chr19_27904353_27904521	Rfx3	3542	0.297613	0.58	3.8e-06	Click!
chr19_27904687_27904996	Rfx3	3946	0.286683	0.59	2.5e-06	Click!
chr19_27906087_27906262	Rfx3	5279	0.265413	-0.45	5.6e-04	Click!
chr19_28010780_28010937	Rfx3	68	0.974083	0.55	1.3e-05	Click!
chr8_84066182_84066879	Rfx1	304	0.625008	-0.09	5.0e-01	Click!

CREs near the TFs associated with the motif



Correlations of motif activity and signal intensity at CREs associated with the motif's TFs:

This plot shows correlation between observed signal intensity of a CRE associated with the transcription factor across all samples and activity of the motif.

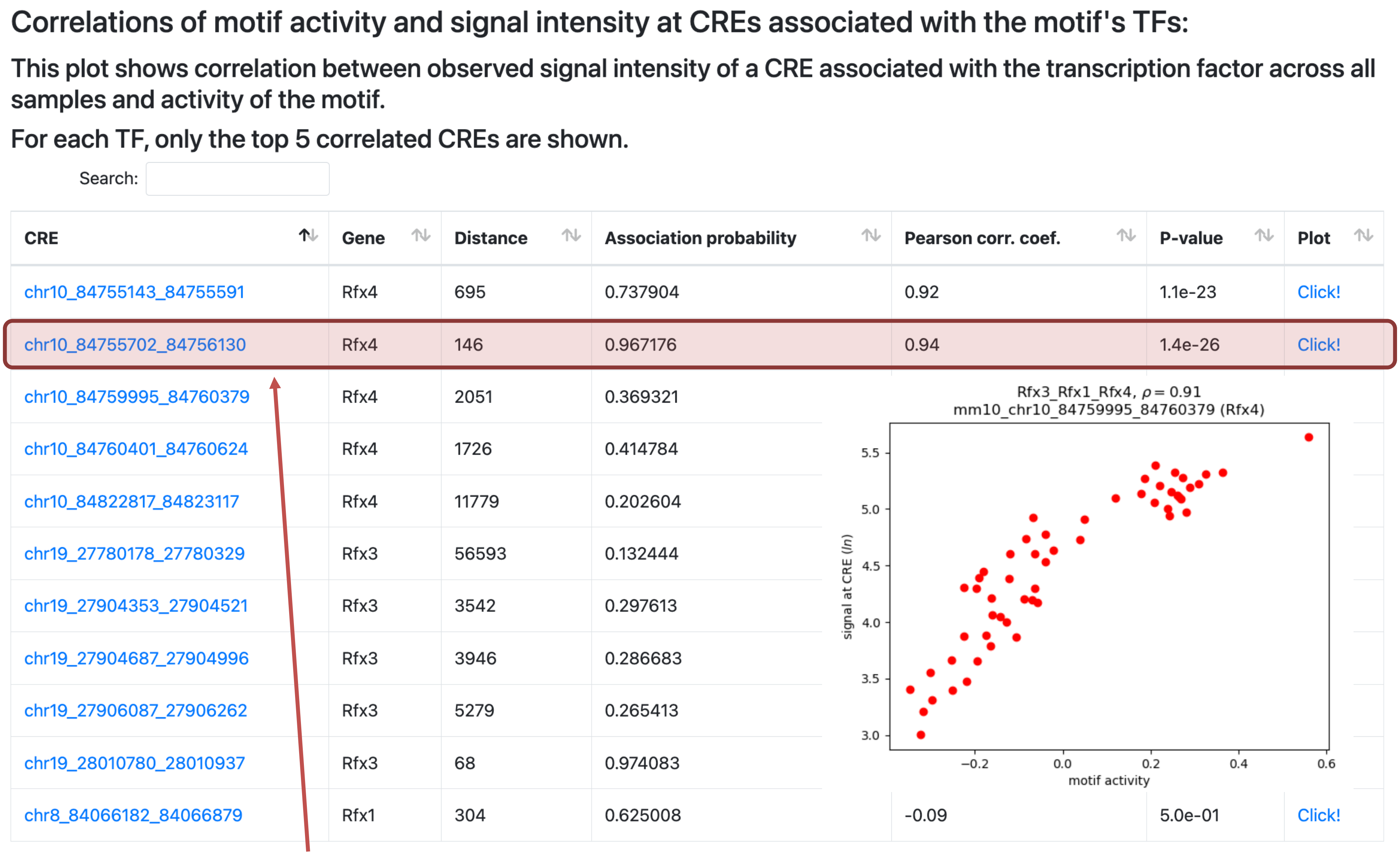
For each TF, only the top 5 correlated CREs are shown.

Search:

CRE	Gene	Distance	Association probability	Pearson corr. coef.	P-value	Plot
chr10_84755143_84755591	Rfx4	695	0.737904	0.92	1.1e-23	Click!
chr10_84755702_84756130	Rfx4	146	0.967176	0.94	1.4e-26	Click!
chr10_84759995_84760379	Rfx4	2051	0.369321	0.91	2.4e-22	Click!
chr10_84760401_84760624	Rfx4	1726	0.414784	0.89	2.4e-19	Click!
chr10_84822817_84823117	Rfx4	11779	0.202604	0.90	1.1e-20	Click!
chr19_27780178_27780329	Rfx3	56593	0.132444	-0.26	5.3e-02	Click!
chr19_27904353_27904521	Rfx3	3542	0.297613	0.58	3.8e-06	Click!
chr19_27904687_27904996	Rfx3	3946	0.286683	0.59	2.5e-06	Click!
chr19_27906087_27906262	Rfx3	5279	0.265413	-0.45	5.6e-04	Click!
chr19_28010780_28010937	Rfx3	68	0.974083	0.55	1.3e-05	Click!
chr8_84066182_84066879	Rfx1	304	0.625008	-0.09	5.0e-01	Click!

CREs near Rfx4 have CRE signal intensities that highly correlate with motif activity

CREs near the TFs associated with the motif



Correlations of motif activity and signal intensity at CREs associated with the motif's TFs:

This plot shows correlation between observed signal intensity of a CRE associated with the transcription factor across all samples and activity of the motif.

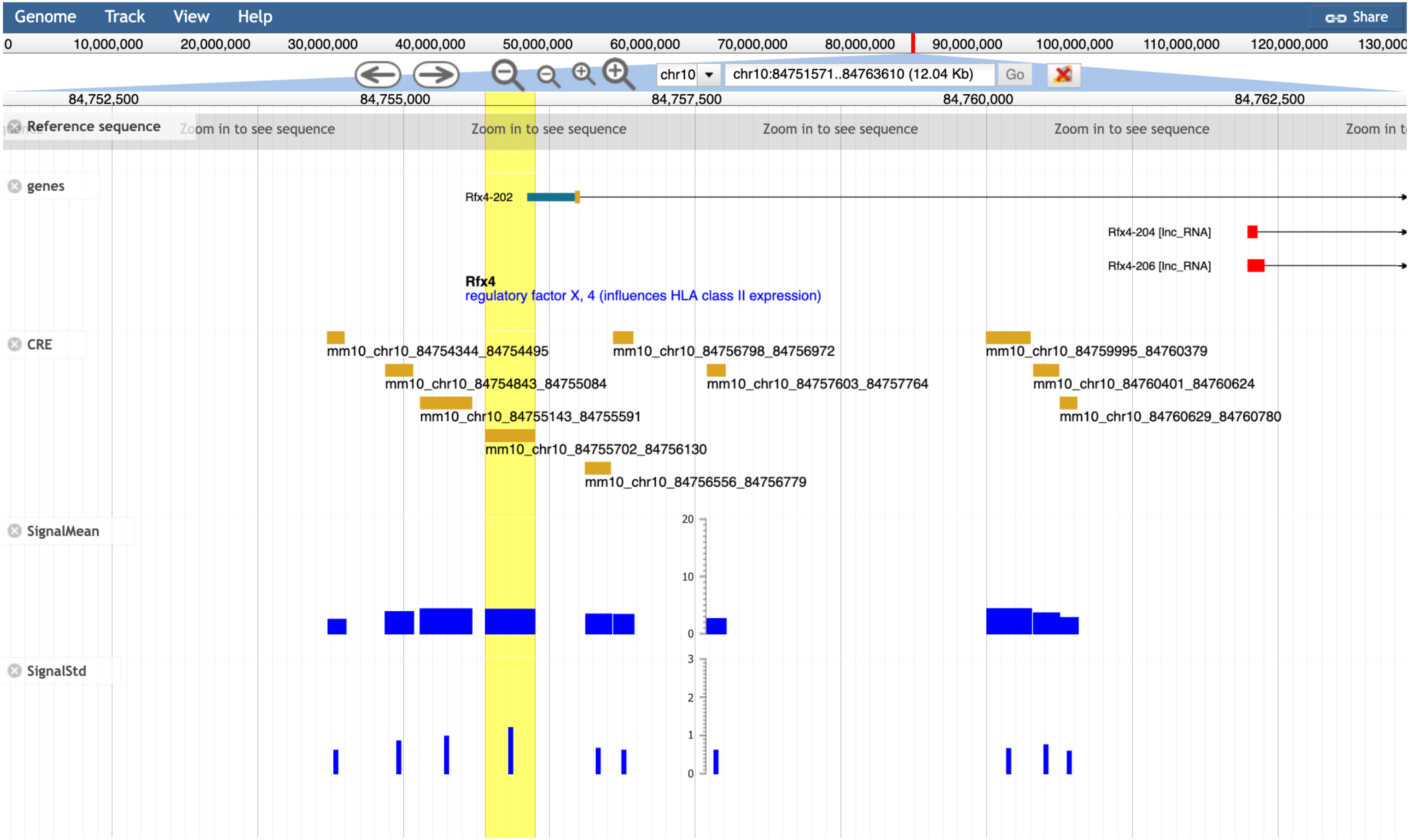
For each TF, only the top 5 correlated CREs are shown.

Search:

CRE	Gene	Distance	Association probability	Pearson corr. coef.	P-value	Plot
chr10_84755143_84755591	Rfx4	695	0.737904	0.92	1.1e-23	Click!
chr10_84755702_84756130	Rfx4	146	0.967176	0.94	1.4e-26	Click!
chr10_84759995_84760379	Rfx4	2051	0.369321			
chr10_84760401_84760624	Rfx4	1726	0.414784			
chr10_84822817_84823117	Rfx4	11779	0.202604			
chr19_27780178_27780329	Rfx3	56593	0.132444			
chr19_27904353_27904521	Rfx3	3542	0.297613			
chr19_27904687_27904996	Rfx3	3946	0.286683			
chr19_27906087_27906262	Rfx3	5279	0.265413			
chr19_28010780_28010937	Rfx3	68	0.974083			
chr8_84066182_84066879	Rfx1	304	0.625008	-0.09	5.0e-01	Click!

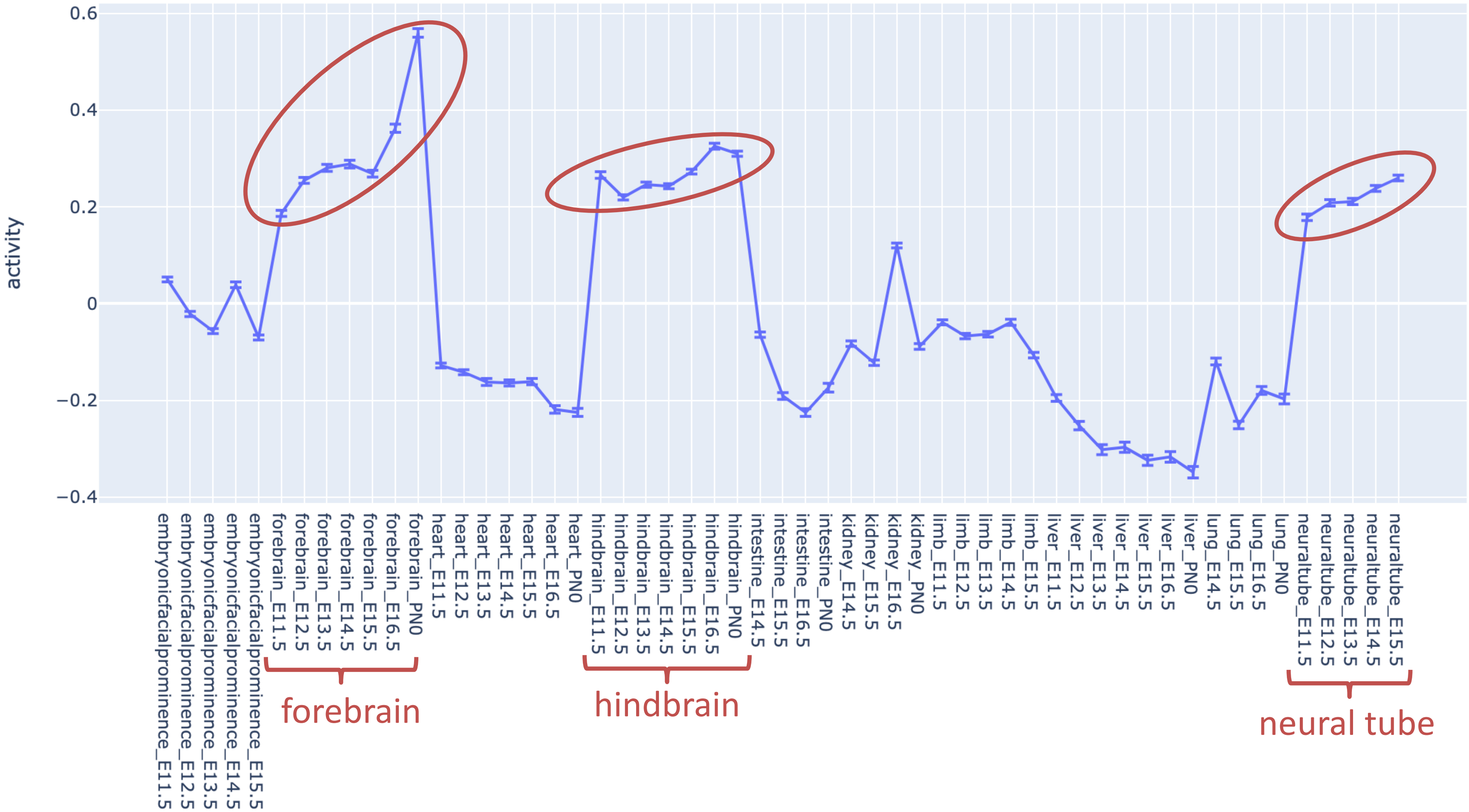
Rfx4 promoter accessibility matches the activity of the motif across samples.

CREs near the Rfx4 promoter


Genome
Track
View
Help
Share
0
10,000,000
20,000,000
30,000,000
40,000,000
50,000,000
60,000,000
70,000,000
80,000,000
90,000,000
100,000,000
110,000,000
120,000,000
130,000
chr10
chr10:84751571..84763610 (12.04 Kb)
Go
84,752,500
84,755,000
84,757,500
84,760,000
84,762,500
Reference sequence
Zoom in to see sequence
genes
Rfx4-202
Rfx4-204 [lnc_RNA]
Rfx4-206 [lnc_RNA]
Rfx4
regulatory factor X, 4 (influences HLA class II expression)
CRE
mm10_chr10_84754344_84754495
mm10_chr10_84754843_84755084
mm10_chr10_84755143_84755591
mm10_chr10_84755702_84756130
mm10_chr10_84756556_84756779
mm10_chr10_84756798_84756972
mm10_chr10_84757603_84757764
mm10_chr10_84759995_84760379
mm10_chr10_84760401_84760624
mm10_chr10_84760629_84760780
SignalMean
20
10
0
SignalStd
3
2
1
0

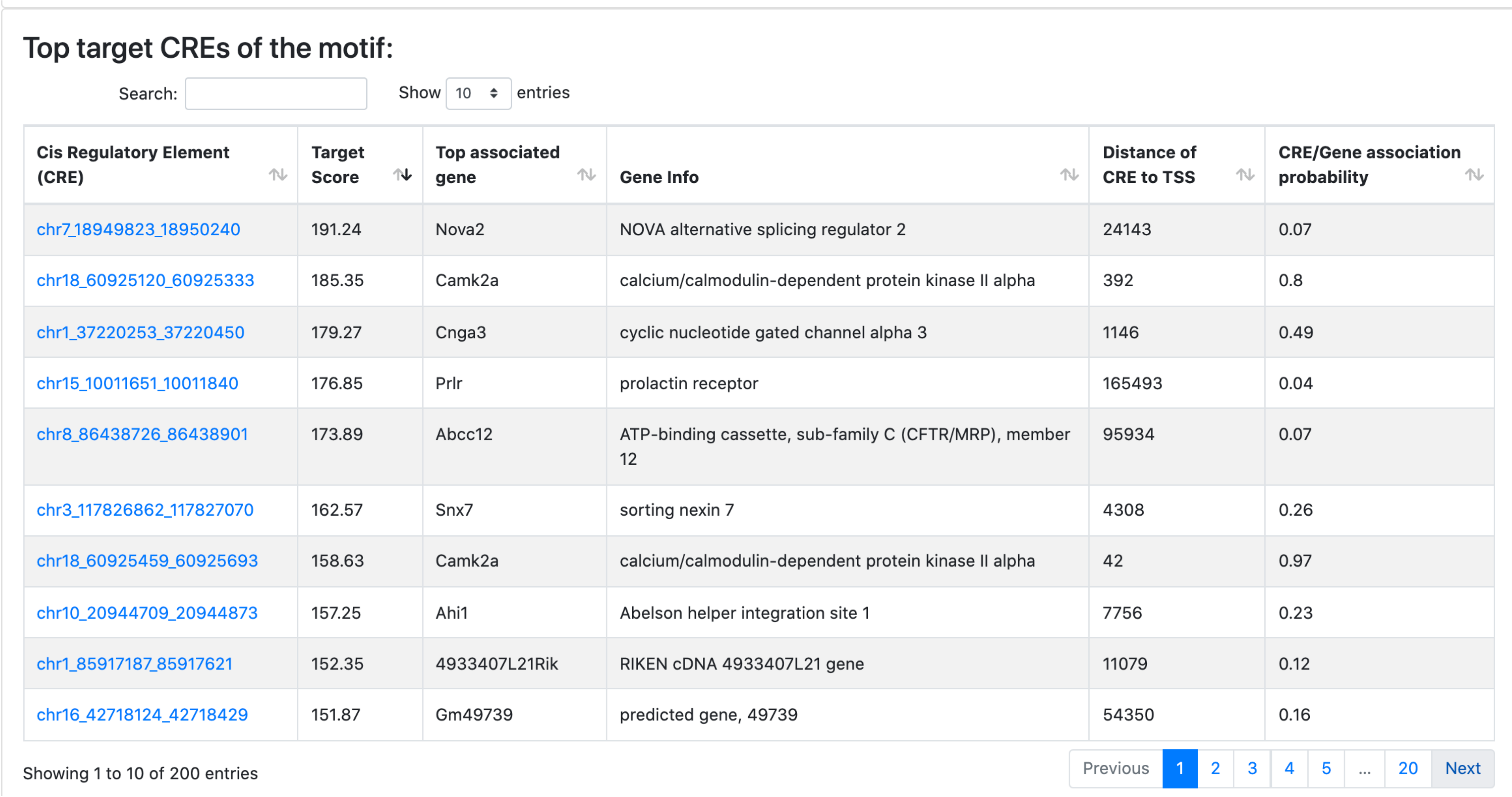
- 7 Separate CREs in a 10Kb region around the start of the Rfx4 gene.
- 3 more CREs downstream of the promoter and upstream of 2 lincRNAs.

Activity of the Rfx3_Rfx1_Rfx4 motif across the samples



- The motif is strongly upregulated in all neural tissues.
- The motif increases in activity across development.
- Especially in late development and postnatally in forebrain.

List of top target CREs of the Rfx motifs



Top target CREs of the motif:

Search:

Show 10 entries

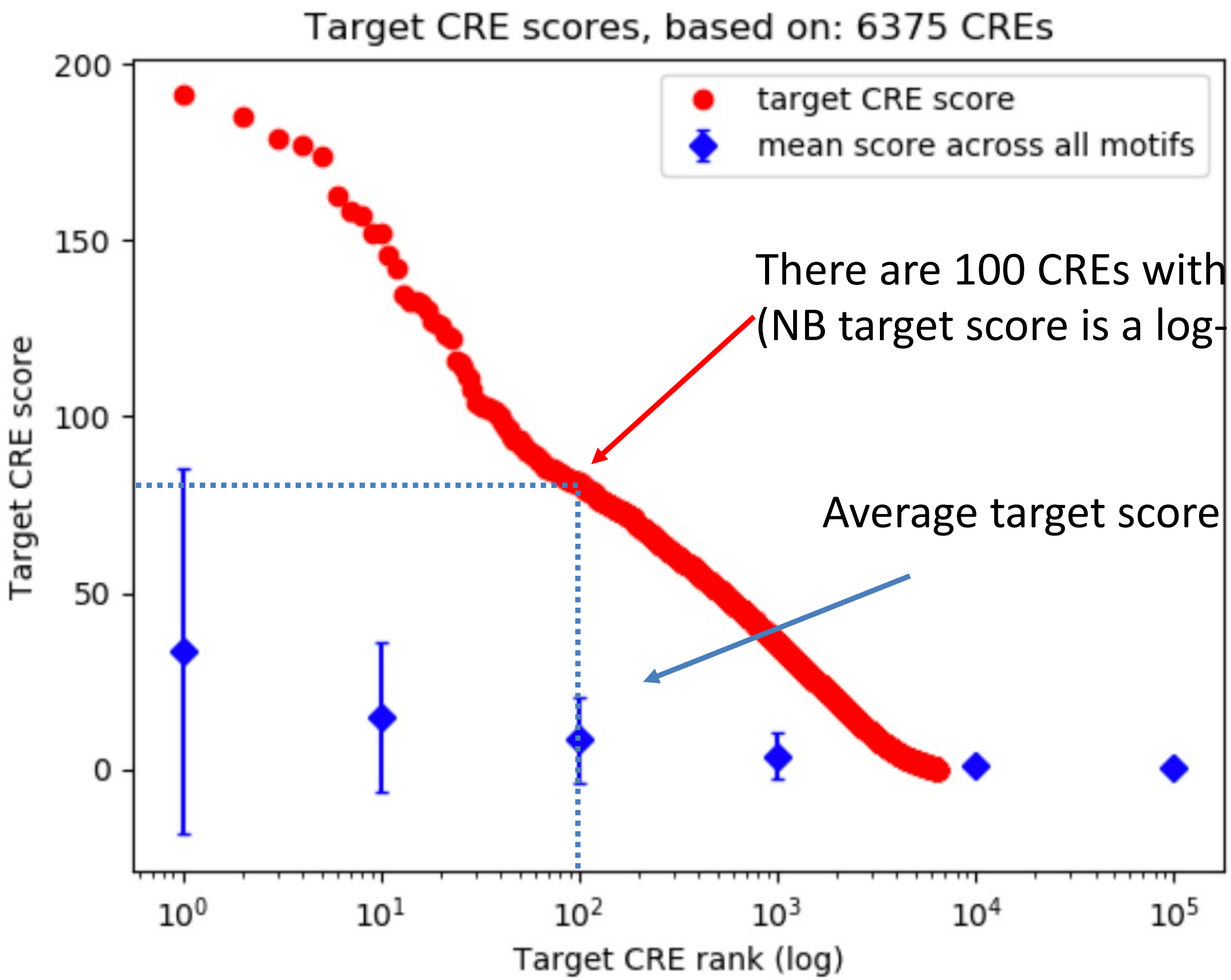
Cis Regulatory Element (CRE)	Target Score	Top associated gene	Gene Info	Distance of CRE to TSS	CRE/Gene association probability
chr7_18949823_18950240	191.24	Nova2	NOVA alternative splicing regulator 2	24143	0.07
chr18_60925120_60925333	185.35	Camk2a	calcium/calmodulin-dependent protein kinase II alpha	392	0.8
chr1_37220253_37220450	179.27	Cnga3	cyclic nucleotide gated channel alpha 3	1146	0.49
chr15_10011651_10011840	176.85	Prlr	prolactin receptor	165493	0.04
chr8_86438726_86438901	173.89	Abcc12	ATP-binding cassette, sub-family C (CFTR/MRP), member 12	95934	0.07
chr3_117826862_117827070	162.57	Snx7	sorting nexin 7	4308	0.26
chr18_60925459_60925693	158.63	Camk2a	calcium/calmodulin-dependent protein kinase II alpha	42	0.97
chr10_20944709_20944873	157.25	Ahi1	Abelson helper integration site 1	7756	0.23
chr1_85917187_85917621	152.35	4933407L21Rik	RIKEN cDNA 4933407L21 gene	11079	0.12
chr16_42718124_42718429	151.87	Gm49739	predicted gene, 49739	54350	0.16

Showing 1 to 10 of 200 entries

All tables like this are searchable and sortable by each of their columns.

How many CREs does the Rfx motif target?

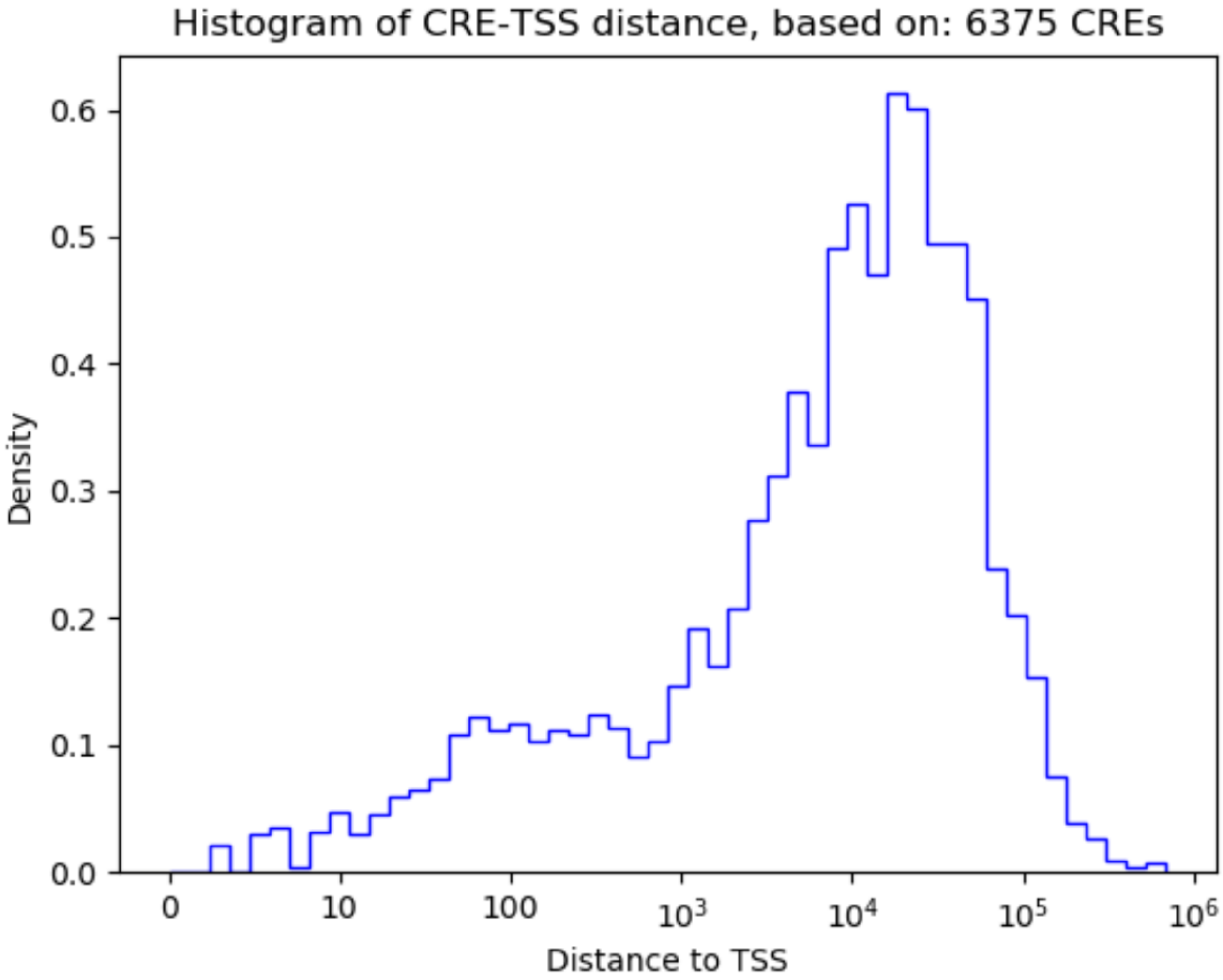
Rank distribution of CRE target scores:



There are 100 CREs with target score of 80 or more. (NB target score is a log-likelihood ratio).

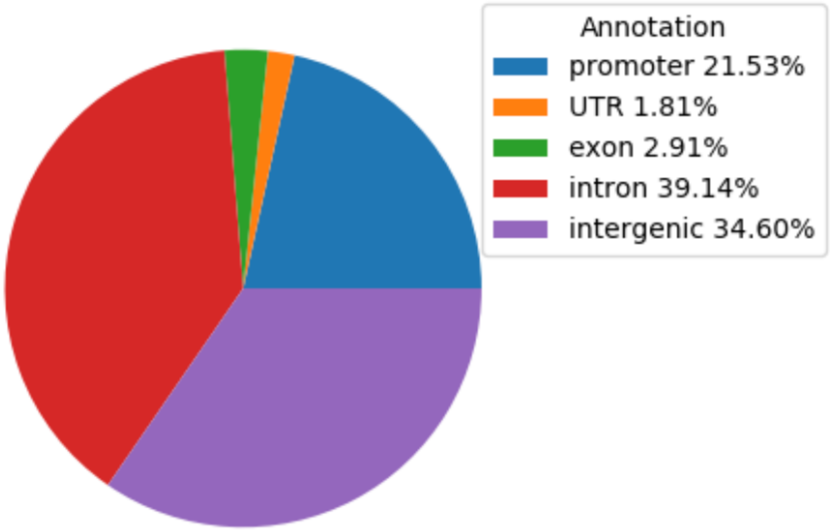
Average target score distribution (over all motifs).

Where are the CREs that the Rfx motif targets?



- Targets = all CREs that have at least 1 binding site for the Rfx motif.
- The histogram is made by weighing each CRE with its target score for the Rfx motif.

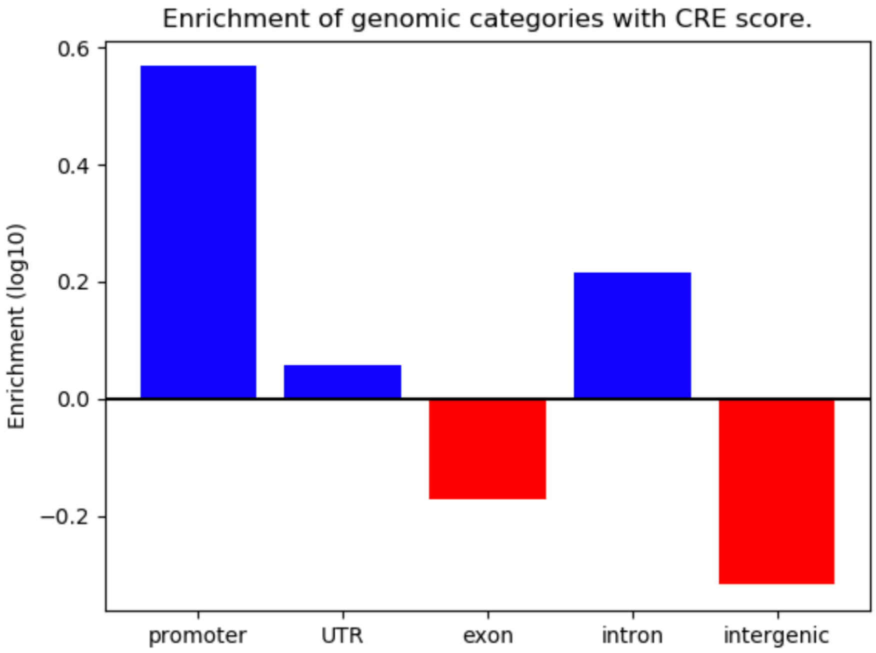
Where are the CREs that the Rfx motif targets?

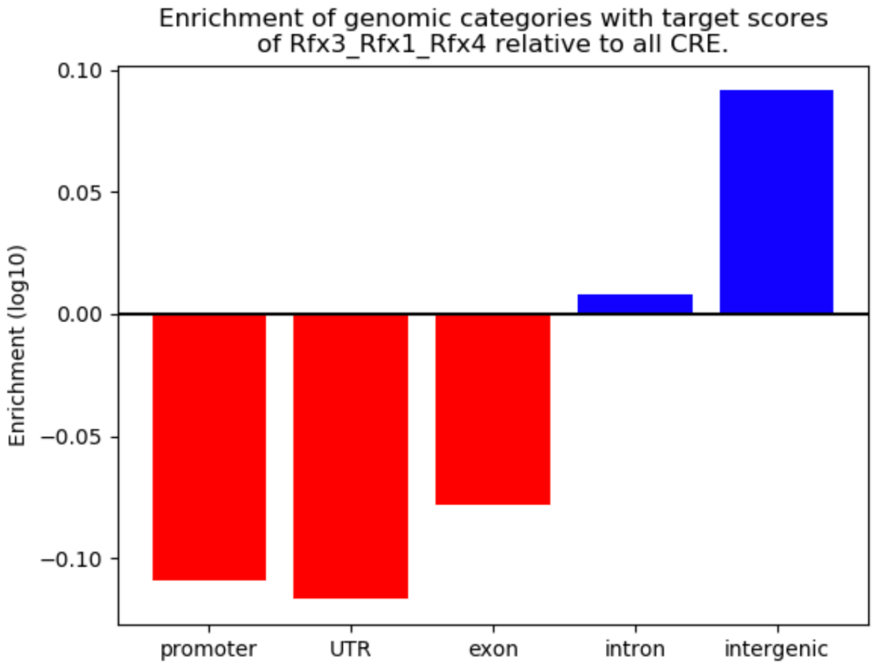


- Fractions of the CREs targeted by the Rfx motif that intersect different types of genomic regions.

- Enrichment of each region type relative to *random positions in the genome*.

- Enrichment of each region type relative to the set of *all CREs.*

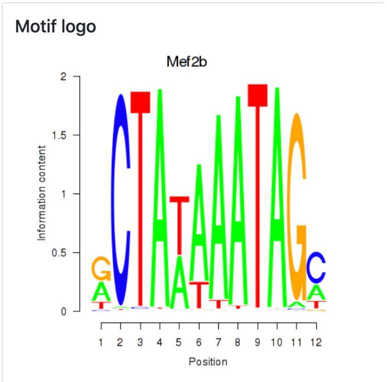


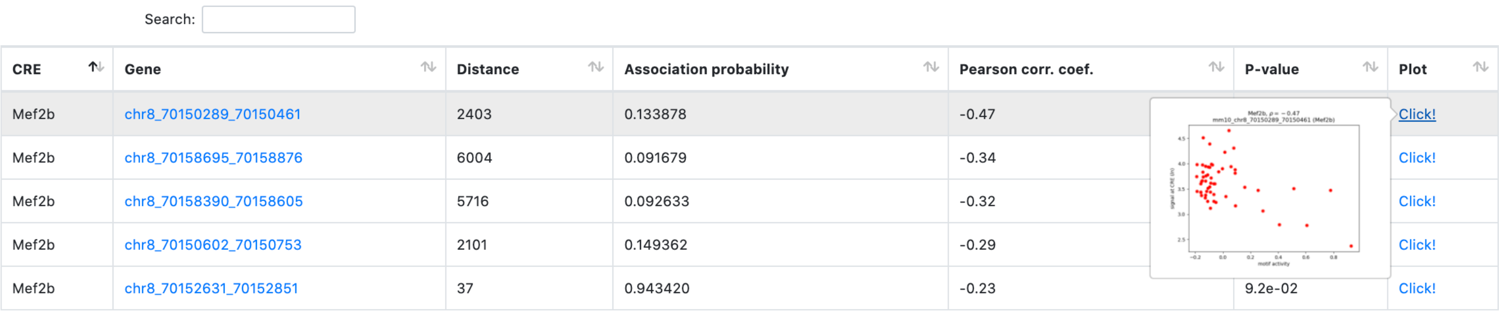


10th most significant motif is Mef2b

Results for Mef2b

Z-value: 14.50



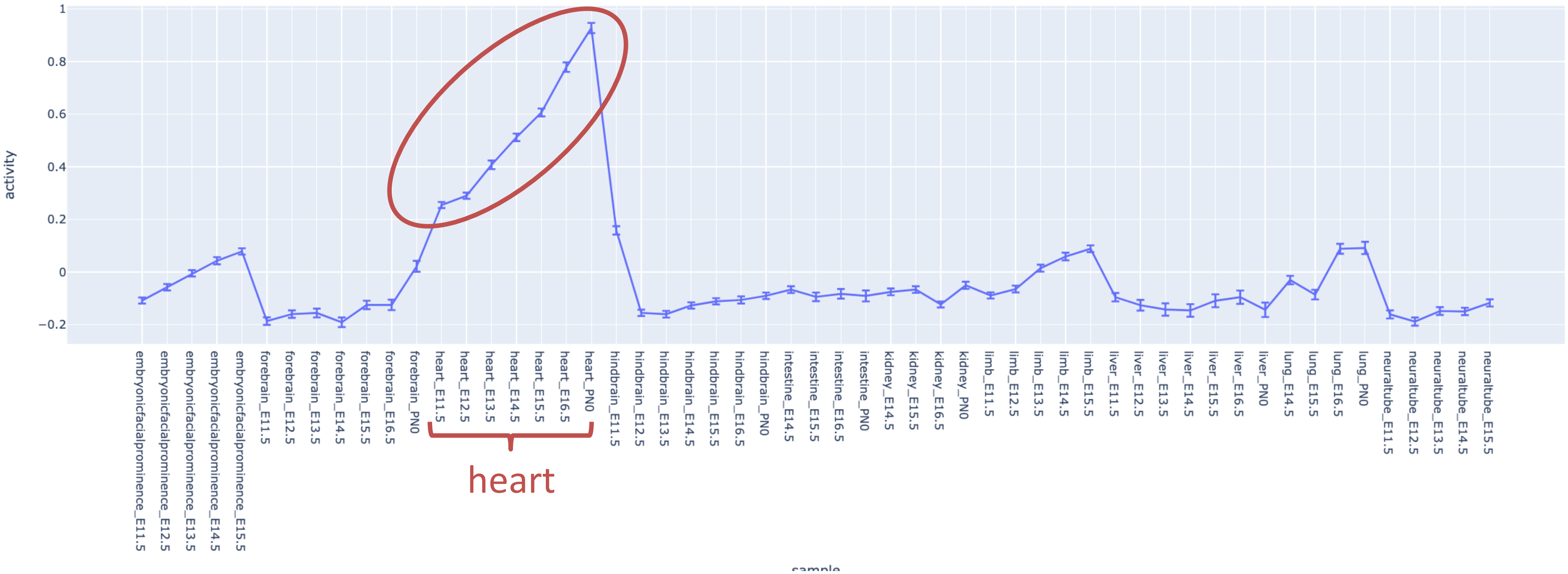


Search:

CRE	Gene	Distance	Association probability	Pearson corr. coef.	P-value	Plot
Mef2b	chr8_70150289_70150461	2403	0.133878	-0.47	[illegible]	Click!
Mef2b	chr8_70158695_70158876	6004	0.091679	-0.34	[illegible]	Click!
Mef2b	chr8_70158390_70158605	5716	0.092633	-0.32	[illegible]	Click!
Mef2b	chr8_70150602_70150753	2101	0.149362	-0.29	[illegible]	Click!
Mef2b	chr8_70152631_70152851	37	0.943420	-0.23	9.2e-02	Click!

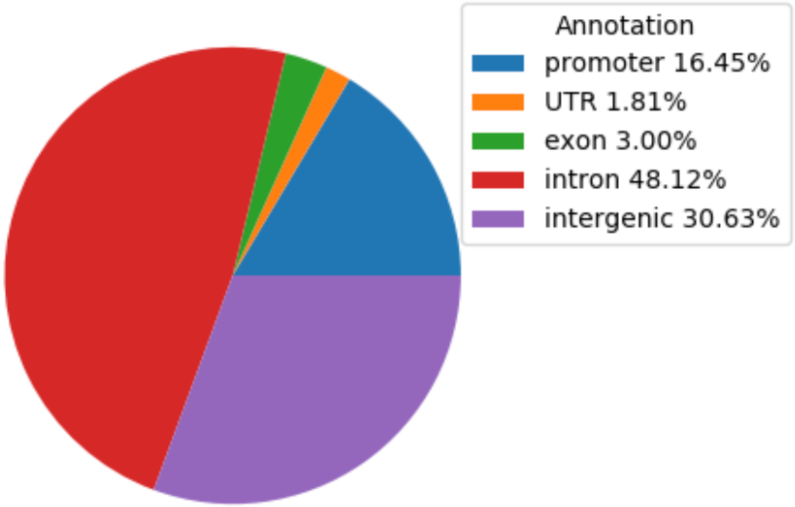
None of the CREs near Mef2b correlate strongly in accessibility with Mef2b motif activity.

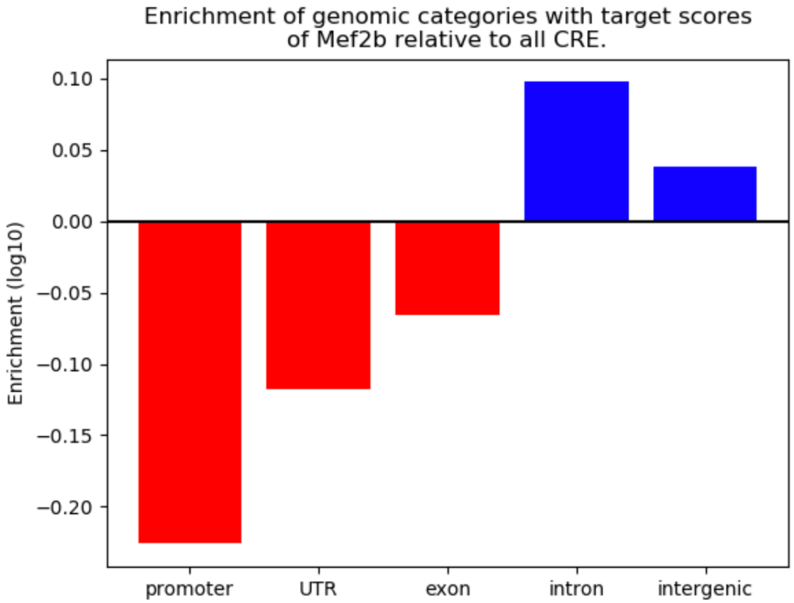
Mef2b motif activity is strongly up-regulated in the developing heart.

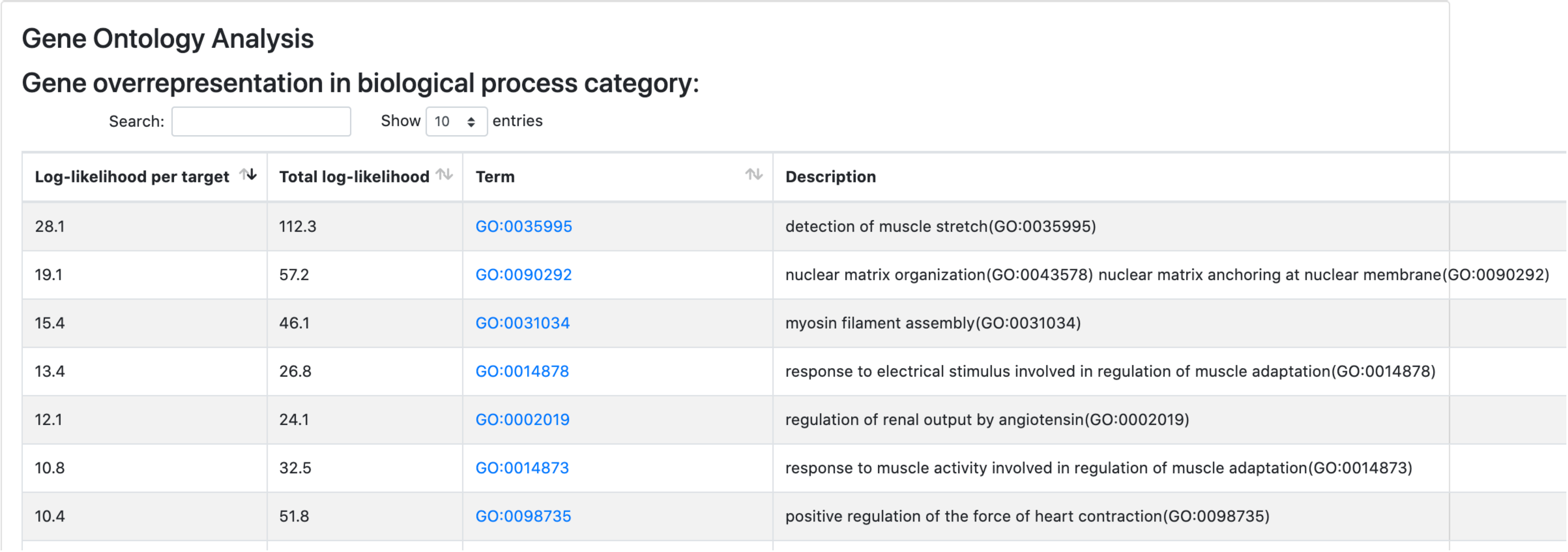


heart

Mef2b targets muscle genes, mostly in introns







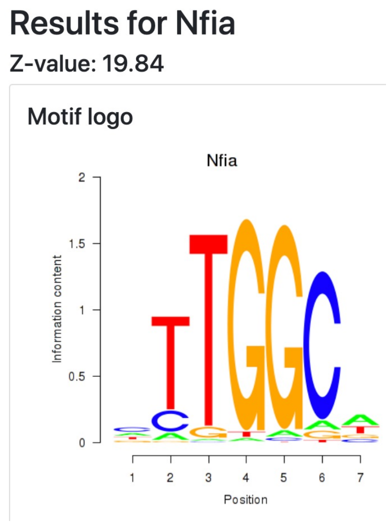
Gene Ontology Analysis

Gene overrepresentation in biological process category:

Search: Show 10 entries

Log-likelihood per target	Total log-likelihood	Term	Description
28.1	112.3	GO:0035995	detection of muscle stretch(GO:0035995)
19.1	57.2	GO:0090292	nuclear matrix organization(GO:0043578) nuclear matrix anchoring at nuclear membrane(GO:0090292)
15.4	46.1	GO:0031034	myosin filament assembly(GO:0031034)
13.4	26.8	GO:0014878	response to electrical stimulus involved in regulation of muscle adaptation(GO:0014878)
12.1	24.1	GO:0002019	regulation of renal output by angiotensin(GO:0002019)
10.8	32.5	GO:0014873	response to muscle activity involved in regulation of muscle adaptation(GO:0014873)
10.4	51.8	GO:0098735	positive regulation of the force of heart contraction(GO:0098735)

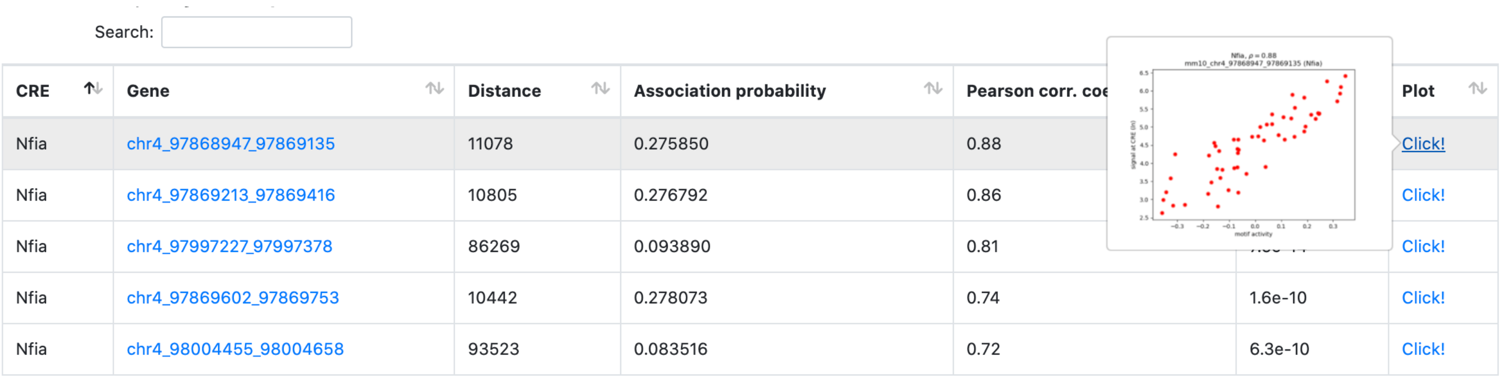
Nfia motif activity increases with time in many tissues



Results for Nfia

Z-value: 19.84

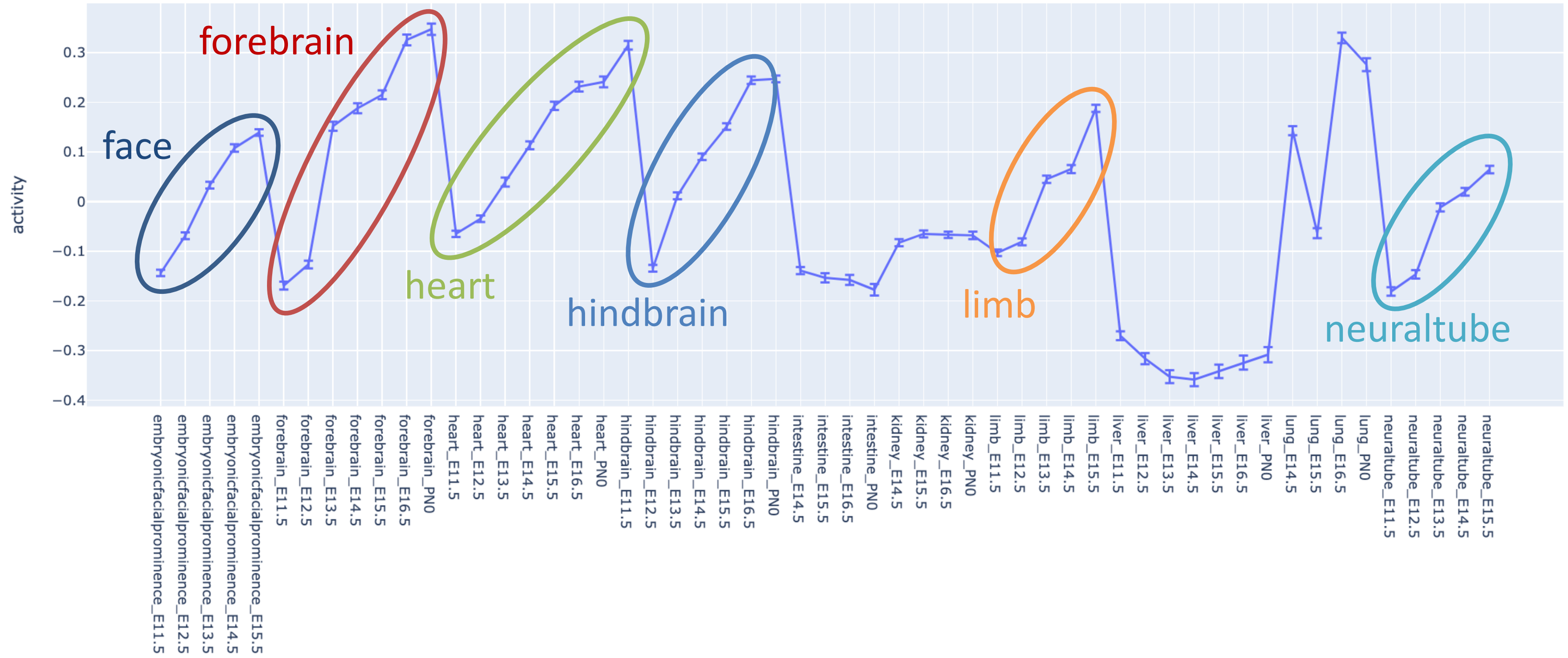
Motif logo



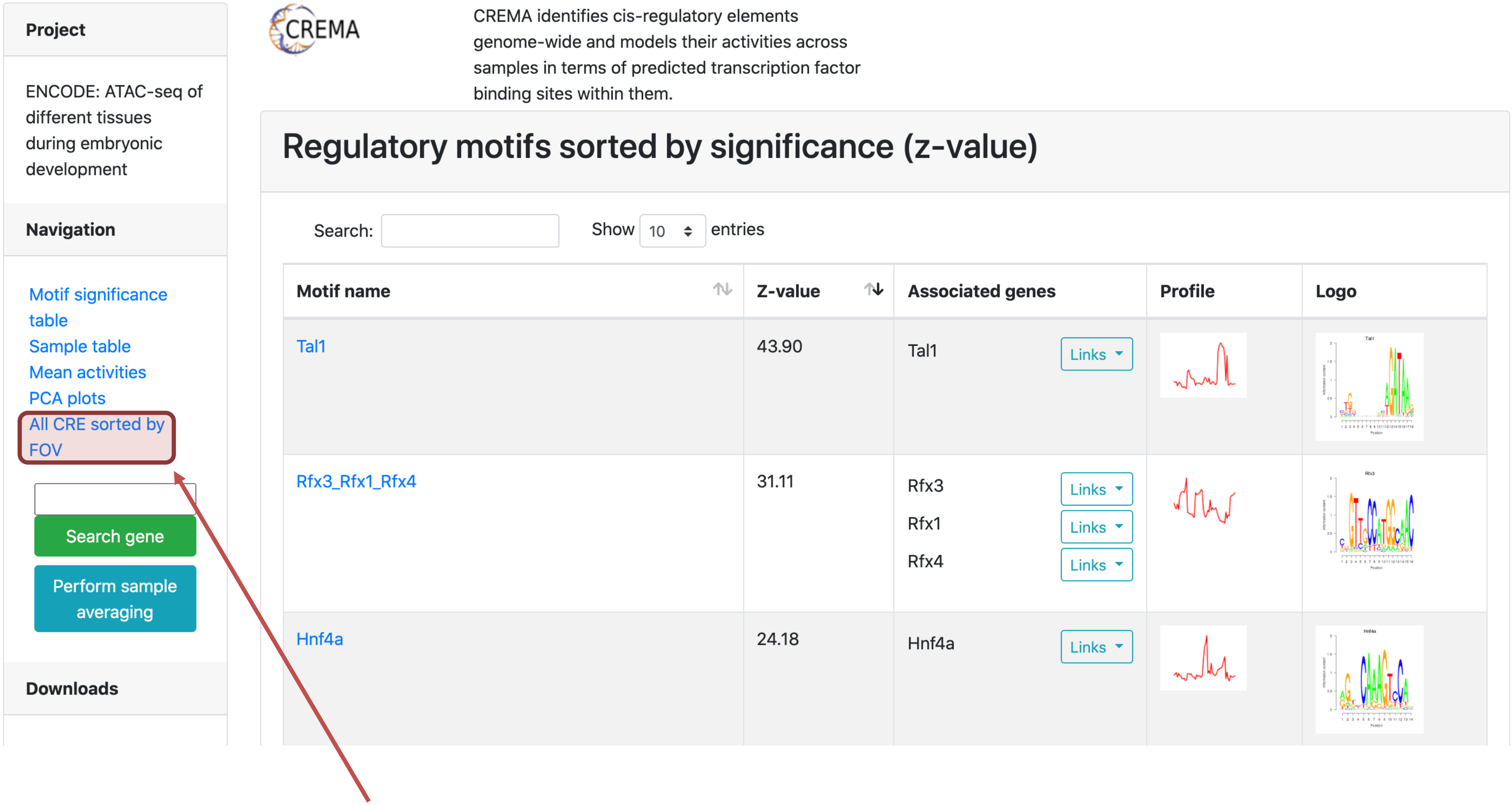
Search:

CRE	Gene	Distance	Association probability	Pearson corr. coe		Plot
Nfia	chr4_97868947_97869135	11078	0.275850	0.88		Click!
Nfia	chr4_97869213_97869416	10805	0.276792	0.86		Click!
Nfia	chr4_97997227_97997378	86269	0.093890	0.81	[illegible]	Click!
Nfia	chr4_97869602_97869753	10442	0.278073	0.74	1.6e-10	Click!
Nfia	chr4_98004455_98004658	93523	0.083516	0.72	6.3e-10	Click!

CREs near the Nfia TF have accessibility that correlate with Nfia motif activity.



Results chromatin accessibility in mouse development



Project

ENCODE: ATAC-seq of different tissues during embryonic development

Navigation

CREMA identifies cis-regulatory elements genome-wide and models their activities across samples in terms of predicted transcription factor binding sites within them.

Regulatory motifs sorted by significance (z-value)

Search: Show 10 entries

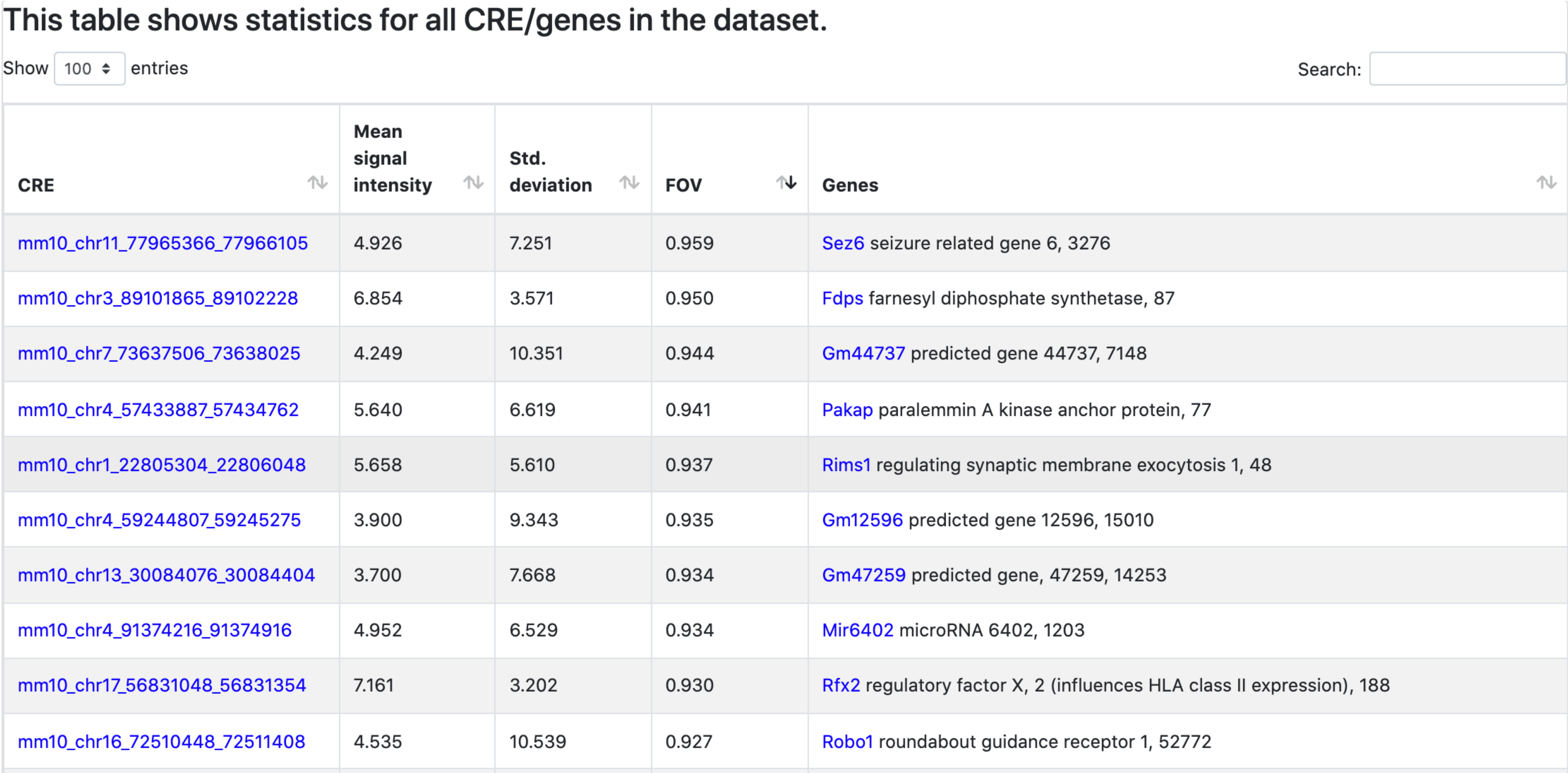
Motif name	Z-value	Associated genes	Profile	Logo
Tal1	43.90	Tal1 Links		
Rfx3_Rfx1_Rfx4	31.11	Rfx3 Links Rfx1 Links Rfx4 Links		
Hnf4a	24.18	Hnf4a Links		

Searchable list with all CREs.

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

List of CREs with summary statistics



This table shows statistics for all CRE/genes in the dataset.

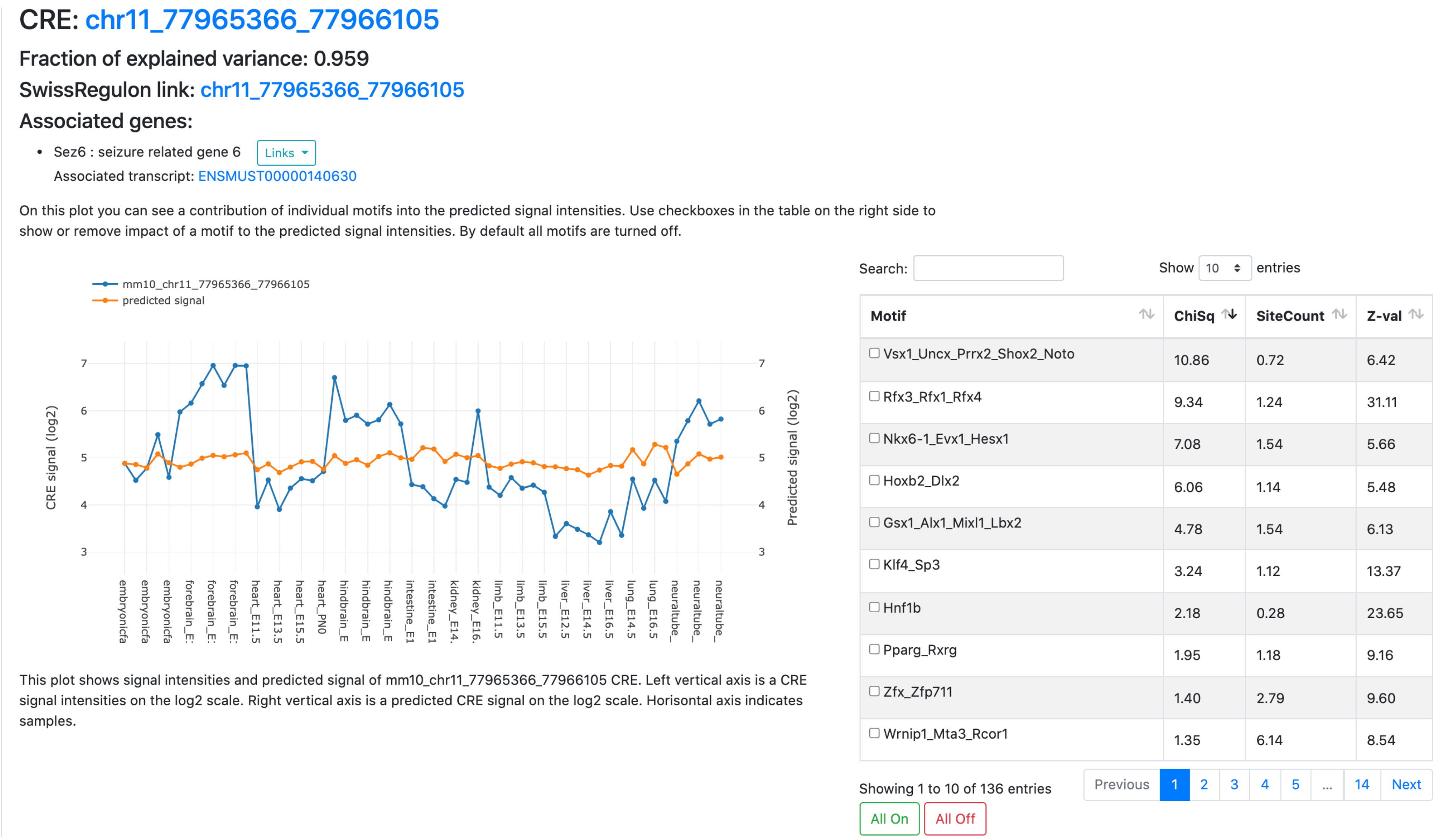
Show 100 entries

Search:

CRE	Mean signal intensity	Std. deviation	FOV	Genes
mm10_chr11_77965366_77966105	4.926	7.251	0.959	Sez6 seizure related gene 6, 3276
mm10_chr3_89101865_89102228	6.854	3.571	0.950	Fdps farnesyl diphosphate synthetase, 87
mm10_chr7_73637506_73638025	4.249	10.351	0.944	Gm44737 predicted gene 44737, 7148
mm10_chr4_57433887_57434762	5.640	6.619	0.941	Pakap paralemmin A kinase anchor protein, 77
mm10_chr1_22805304_22806048	5.658	5.610	0.937	Rims1 regulating synaptic membrane exocytosis 1, 48
mm10_chr4_59244807_59245275	3.900	9.343	0.935	Gm12596 predicted gene 12596, 15010
mm10_chr13_30084076_30084404	3.700	7.668	0.934	Gm47259 predicted gene, 47259, 14253
mm10_chr4_91374216_91374916	4.952	6.529	0.934	Mir6402 microRNA 6402, 1203
mm10_chr17_56831048_56831354	7.161	3.202	0.930	Rfx2 regulatory factor X, 2 (influences HLA class II expression), 188
mm10_chr16_72510448_72511408	4.535	10.539	0.927	Robo1 roundabout guidance receptor 1, 52772

- The table can be sorted by any of its columns (default by FoV).
- One can search for particular CREs or genes.
- Note a gene can have many CREs associated with it.
- The table is LARGE and typically takes ~1 minute to load.

Example of a CRE with very high FoV



CRE: chr11_77965366_77966105

Fraction of explained variance: 0.959

SwissRegulon link: chr11_77965366_77966105

Associated genes:

- Sez6 : seizure related gene 6 Links ▾
  Associated transcript: ENSMUST00000140630

On this plot you can see a contribution of individual motifs into the predicted signal intensities. Use checkboxes in the table on the right side to show or remove impact of a motif to the predicted signal intensities. By default all motifs are turned off.

This plot shows signal intensities and predicted signal of mm10_chr11_77965366_77966105 CRE. Left vertical axis is a CRE signal intensities on the log2 scale. Right vertical axis is a predicted CRE signal on the log2 scale. Horisontal axis indicates samples.

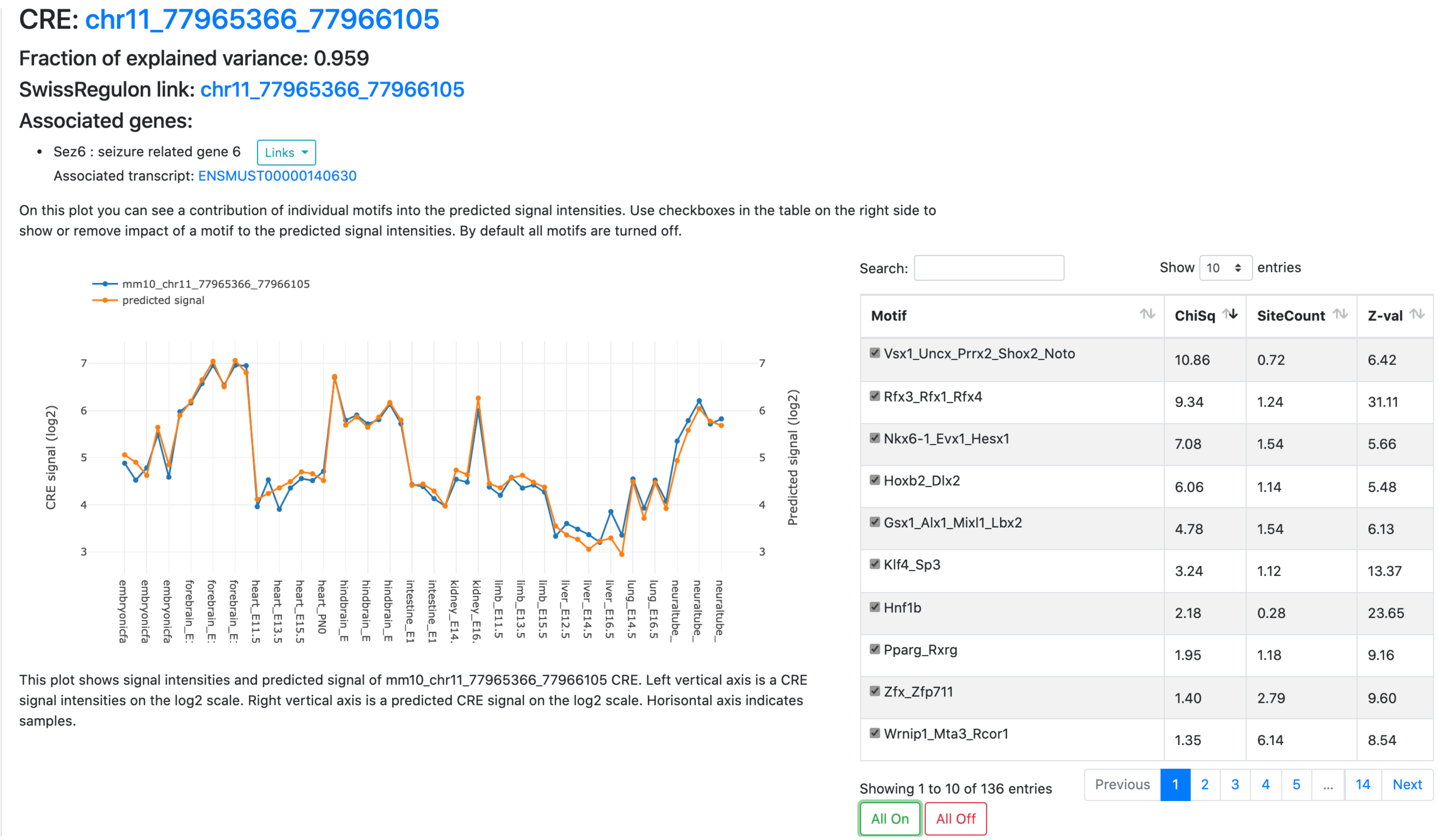
Search:

Show 10 entries

Motif	ChiSq	SiteCount	Z-val
□ Vsx1_Uncx_Prrx2_Shox2_Noto	10.86	0.72	6.42
□ Rfx3_Rfx1_Rfx4	9.34	1.24	31.11
□ Nkx6-1_Evx1_Hesx1	7.08	1.54	5.66
□ Hoxb2_Dlx2	6.06	1.14	5.48
□ Gsx1_Alx1_Mixl1_Lbx2	4.78	1.54	6.13
□ Klf4_Sp3	3.24	1.12	13.37
□ Hnf1b	2.18	0.28	23.65
□ Pparg_Rxrg	1.95	1.18	9.16
□ Zfx_Zfp711	1.40	2.79	9.60
□ Wrnip1_Mta3_Rcor1	1.35	6.14	8.54

Showing 1 to 10 of 136 entries

Example CRE with very high FoV



CRE: chr11_77965366_77966105

Fraction of explained variance: 0.959

SwissRegulon link: chr11_77965366_77966105

Associated genes:

- Sez6 : seizure related gene 6 Links ▾
  Associated transcript: ENSMUST00000140630

On this plot you can see a contribution of individual motifs into the predicted signal intensities. Use checkboxes in the table on the right side to show or remove impact of a motif to the predicted signal intensities. By default all motifs are turned off.

This plot shows signal intensities and predicted signal of mm10_chr11_77965366_77966105 CRE. Left vertical axis is a CRE signal intensities on the log2 scale. Right vertical axis is a predicted CRE signal on the log2 scale. Horisontal axis indicates samples.

Search:

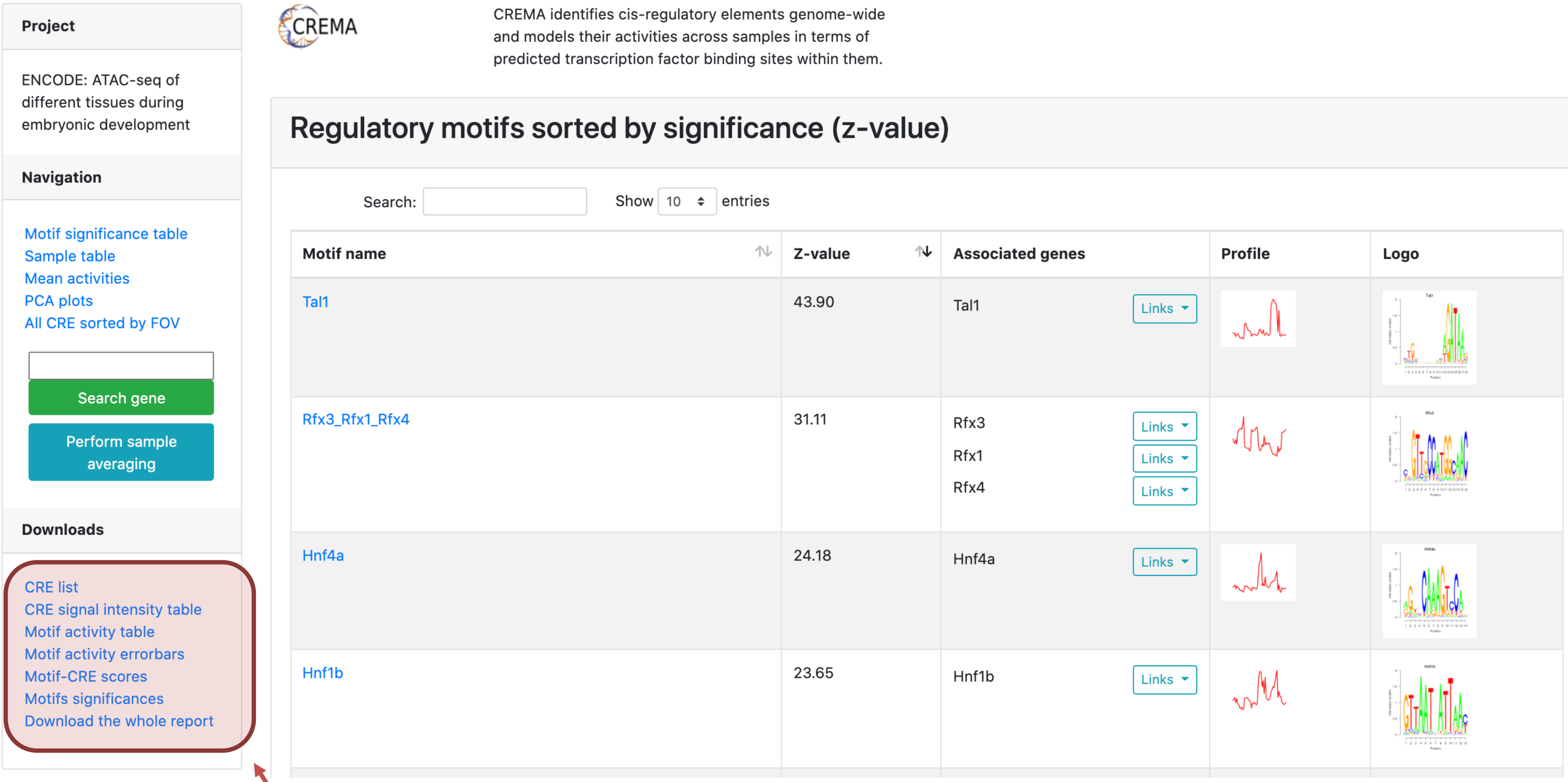
Show 10 entries

Motif	ChiSq	SiteCount	Z-val
☑ Vsx1_Uncx_Prrx2_Shox2_Noto	10.86	0.72	6.42
☑ Rfx3_Rfx1_Rfx4	9.34	1.24	31.11
☑ Nkx6-1_Evx1_Hesx1	7.08	1.54	5.66
☑ Hoxb2_Dlx2	6.06	1.14	5.48
☑ Gsx1_Alx1_Mixl1_Lbx2	4.78	1.54	6.13
☑ Klf4_Sp3	3.24	1.12	13.37
☑ Hnf1b	2.18	0.28	23.65
☑ Pparg_Rxrg	1.95	1.18	9.16
☑ Zfx_Zfp711	1.40	2.79	9.60
☑ Wrnip1_Mta3_Rcor1	1.35	6.14	8.54

Showing 1 to 10 of 136 entries

Of course, it is extremely rare for the model to fit accessibility across tissues so well.

All results are downloadable in flat file formats



Project

ENCODE: ATAC-seq of different tissues during embryonic development

Navigation

- Motif significance table
- Sample table
- Mean activities
- PCA plots
- All CRE sorted by FOV

Downloads

- CRE list
- CRE signal intensity table
- Motif activity table
- Motif activity errorbars
- Motif-CRE scores
- Motifs significances
- Download the whole report

CREMA identifies cis-regulatory elements genome-wide and models their activities across samples in terms of predicted transcription factor binding sites within them.

Regulatory motifs sorted by significance (z-value)

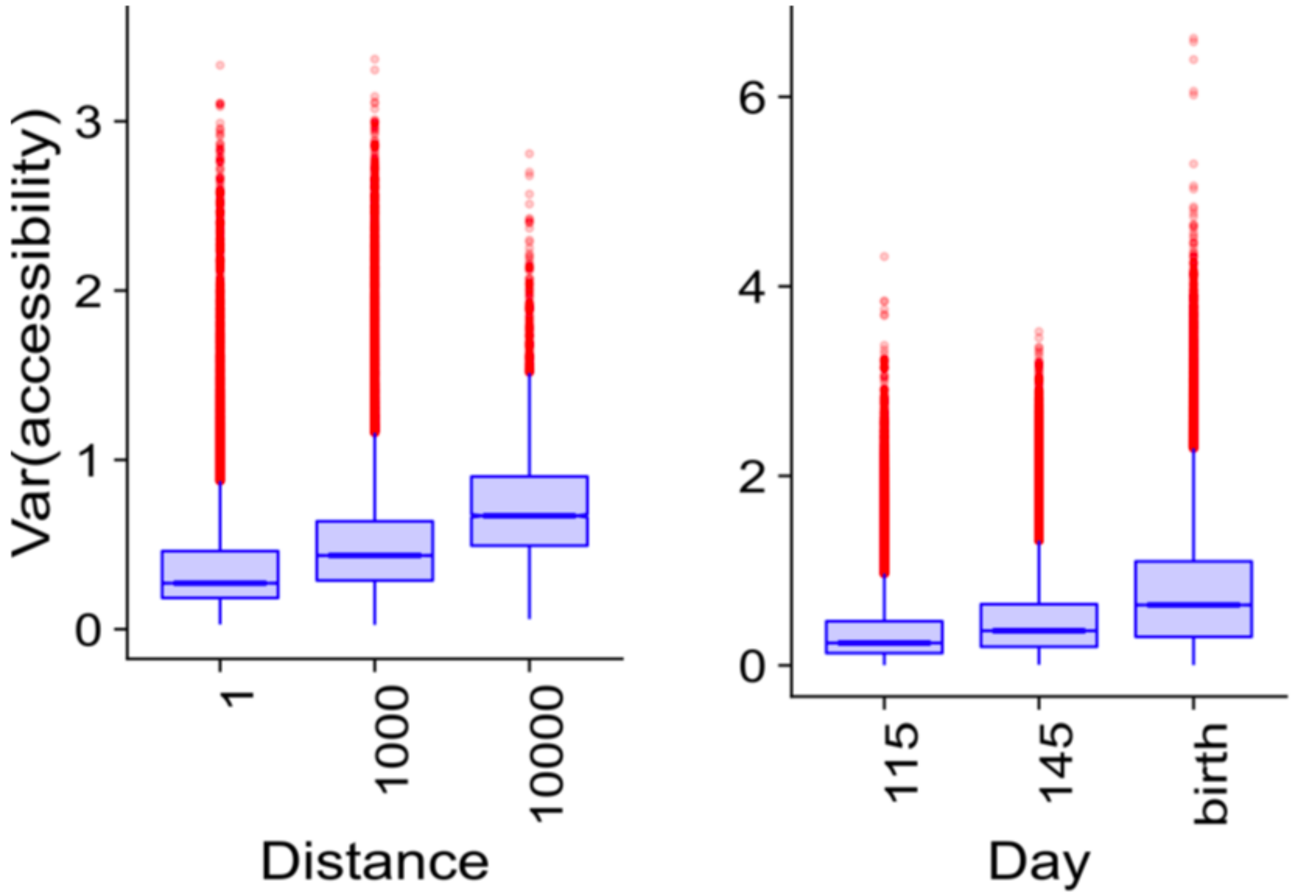
Search: Show 10 entries

Motif name	Z-value	Associated genes	Profile	Logo
Tal1	43.90	Tal1 Links		
Rfx3_Rfx1_Rfx4	31.11	Rfx3 Links Rfx1 Links Rfx4 Links		
Hnf4a	24.18	Hnf4a Links		
Hnf1b	23.65	Hnf1b Links		

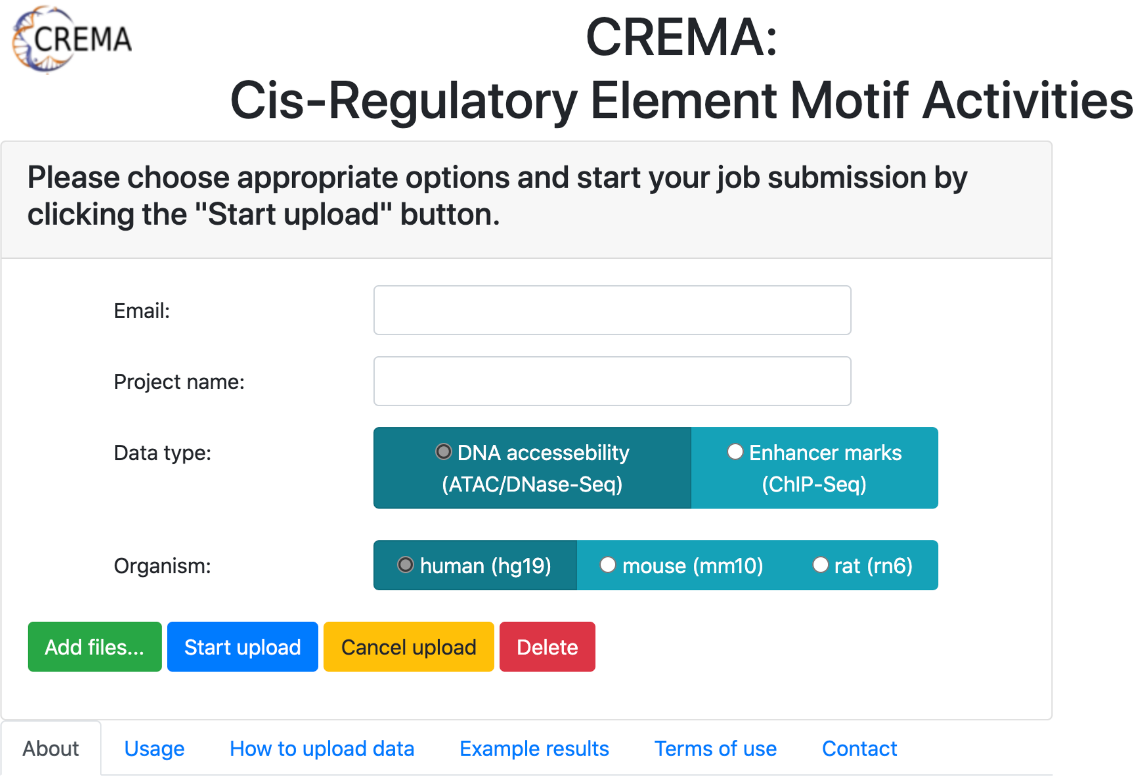
These results allow all kinds of downstream analyses of your own design.

Example

Variability in accessibility is larger for distal regions and larger at later developmental time points



CREMA: acknowledgments



CREMA:

Cis-Regulatory Element Motif Activities

Please choose appropriate options and start your job submission by clicking the "Start upload" button.

Email:	
Project name:	
Data type:	● DNA accessebility (ATAC/DNase-Seq) ○ Enhancer marks (ChIP-Seq)
Organism:	● human (hg19) ○ mouse (mm10) ○ rat (rn6)

Anne Krämer
CREMA developer

Mikhail Pachkov
web-interface developer

Severin Berger
CRUNCH developer

Phil Arnold
MotEvo

Saeed Omidi
CRUNCH pipeline

Nick Kelley
pre-processing

Silvia Salatino
pre-processing