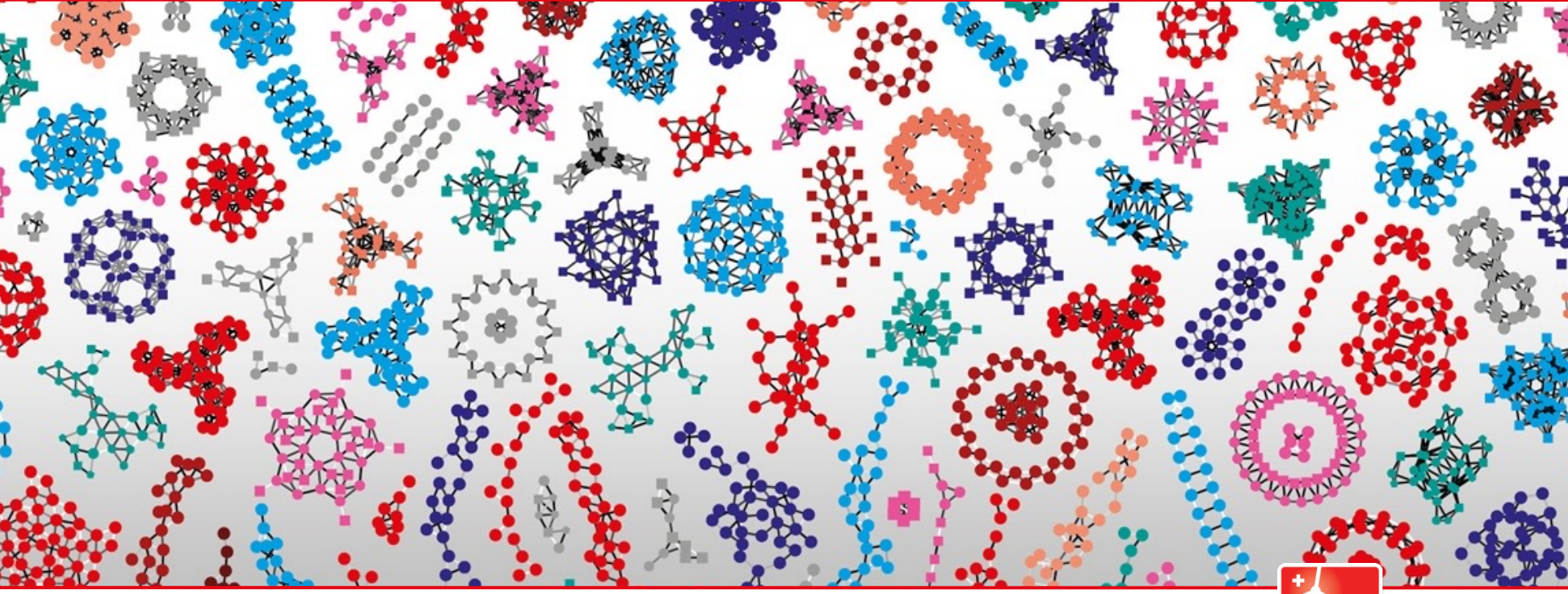


ISMARA & CREMA tutorials

SIB Swiss Institute of Bioinformatics


Erik van Nimwegen


Mikhail Pachkov


Daan de Groot


Anurag Ranjak

elixir SWITZERLAND

(BC)2 Basel Computational Biology Conference

@NimwegenLab

www.sib.swiss

Agenda

- 9:00 – 10:30 **ISMARA**: Introduction to Motif Activity Response Analysis (MARA) modeling gene expression in terms of regulatory sites. Theory and overview of the results.
- 10:30 – 11:00 Coffee break
- 11:00 – 12:30 **CREMA**: Cis-regulatory Element Motif Activities. Modeling chromatin state genome-wide in terms of regulatory sites. Theory and overview of the results.
- 12:30 – 13:30 Lunch break
- 13:30 – 15:00 **Using the web interface**: Supported species, data types and formats, uploading data, downloading result, and advanced interactive features.
- 15:00 – 15:30 Coffee break
- 15:30 – 17:00 Hands-on exercises. Users explore results using their own datasets.

SwissRegulon Portal

Resources for regulatory genomics

NimwegenLab

- Web-based tools & services
- Software
- Genome annotations
- Documentation & Help
- Contact us

swissregulon.unibas.ch

Web-based tools & services

ISMARA

ISMARA: The Integrated System for Motif Actitivity Response Analysis.

Input DATA: RNA-Seq, ChIP-Seq, microarray.
Analysis: Infers key regulators (TFs/microRNAs) and gene regulatory interactions from expression data.

CREMA

CREMA: Cis-Regulatory Element Motif Activities.

Input DATA: ATAC-Seq, DNase-Seq, ChIP-Seq.
Analysis: Infers CREs genome-wide and the key TFs that regulate their chromatin state (i.e. accessibility or epigenetic marks) across a set of samples.

CRUNCH

CRUNCH: A completely automated pipe-line for TF ChIP-Seq analysis.

Input DATA: TF ChIP-Seq.
Analysis: Peak identification and comprehensive annotation of regulatory motifs and sites in peaks.

REALPHY

REALPHY: Reference sequence Alignment based Phylogeny.

Input DATA: Genome assemblies or raw genomic sequencing reads.
Analysis: Core genome alignment and phylogenetic tree.

RECOPHY

RECOPHY: A recombination pattern analysis in related prokaryotic species.

Input DATA: Genome assemblies or raw genomic sequencing reads.
Analysis: A range of summary statistics, such as the fraction of SNPs supporting each branch of a tree and the fraction of clonal and recombined genome between each pair of strains.

PhyloGibbs

Phylogibbs: A Gibbs sampling motif finder for multiple alignments.

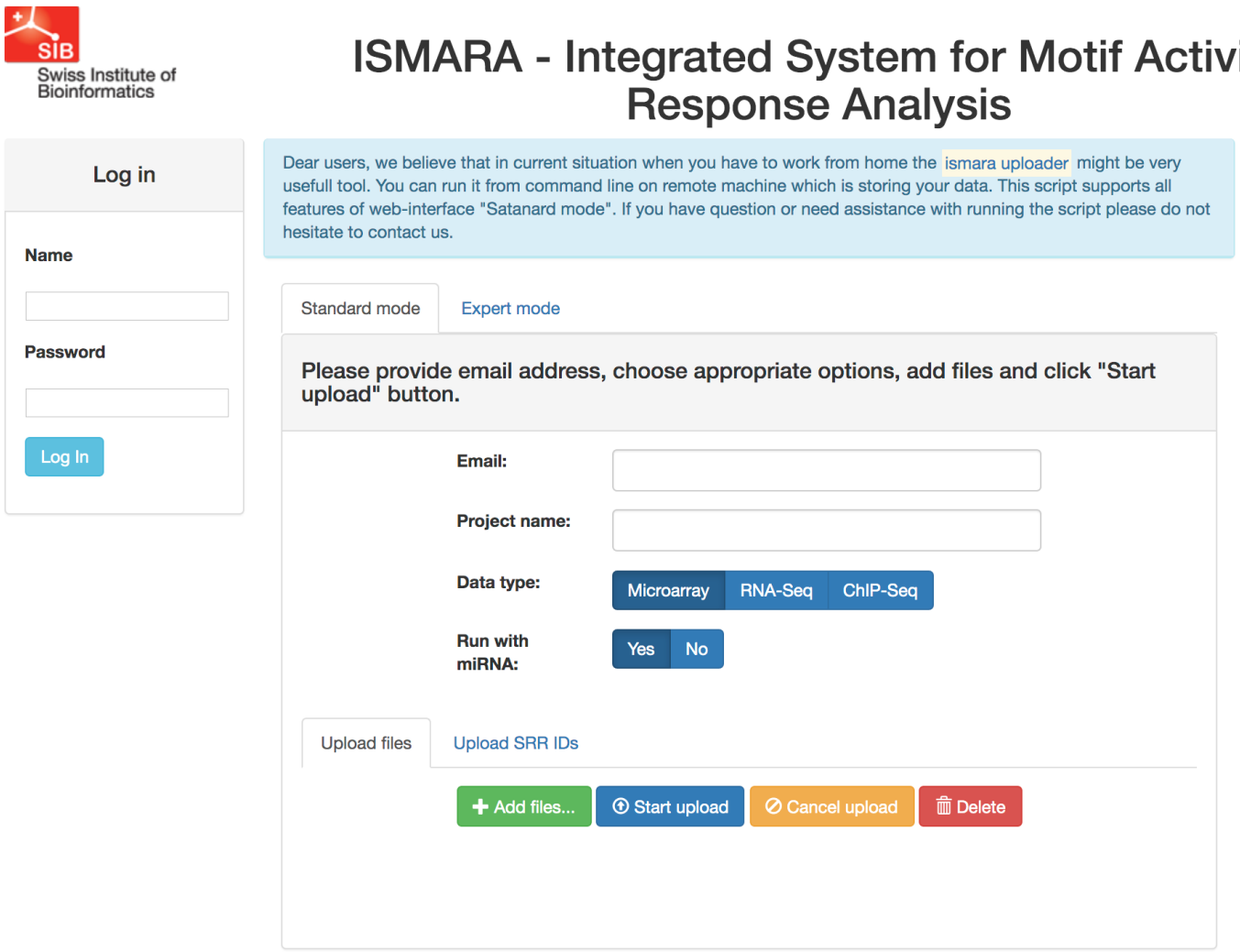
Input DATA: Multiple alignments of DNA sequences.
Analysis: *De novo* identification of regulatory motifs with rigorous incorporation of conservation information.

ISMARA

Automatically inferring key gene regulatory circuitry from gene expression data

Swiss Institute of Bioinformatics

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

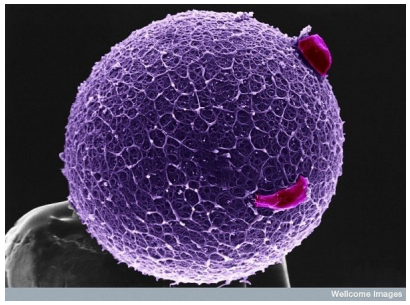


ismara.unibas.ch

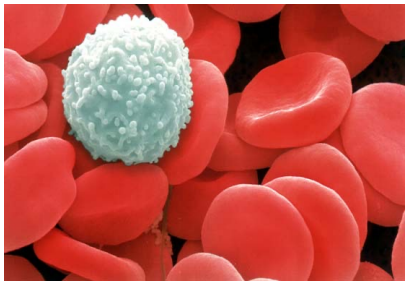
BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

Swiss Institute of Bioinformatics

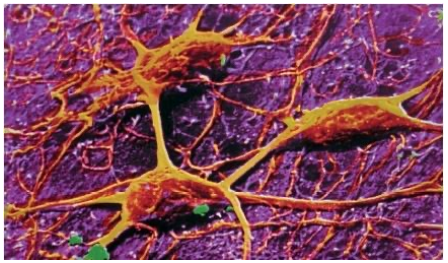
How is the regulatory code in the DNA `read out' to control cell fate and identity?



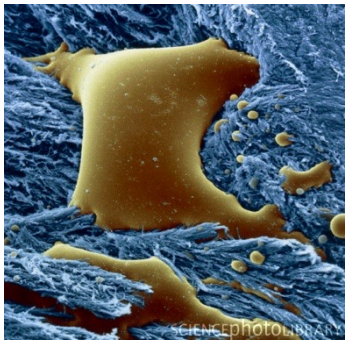
egg cell with 2 coronal cells



white and red blood cells



three neurons

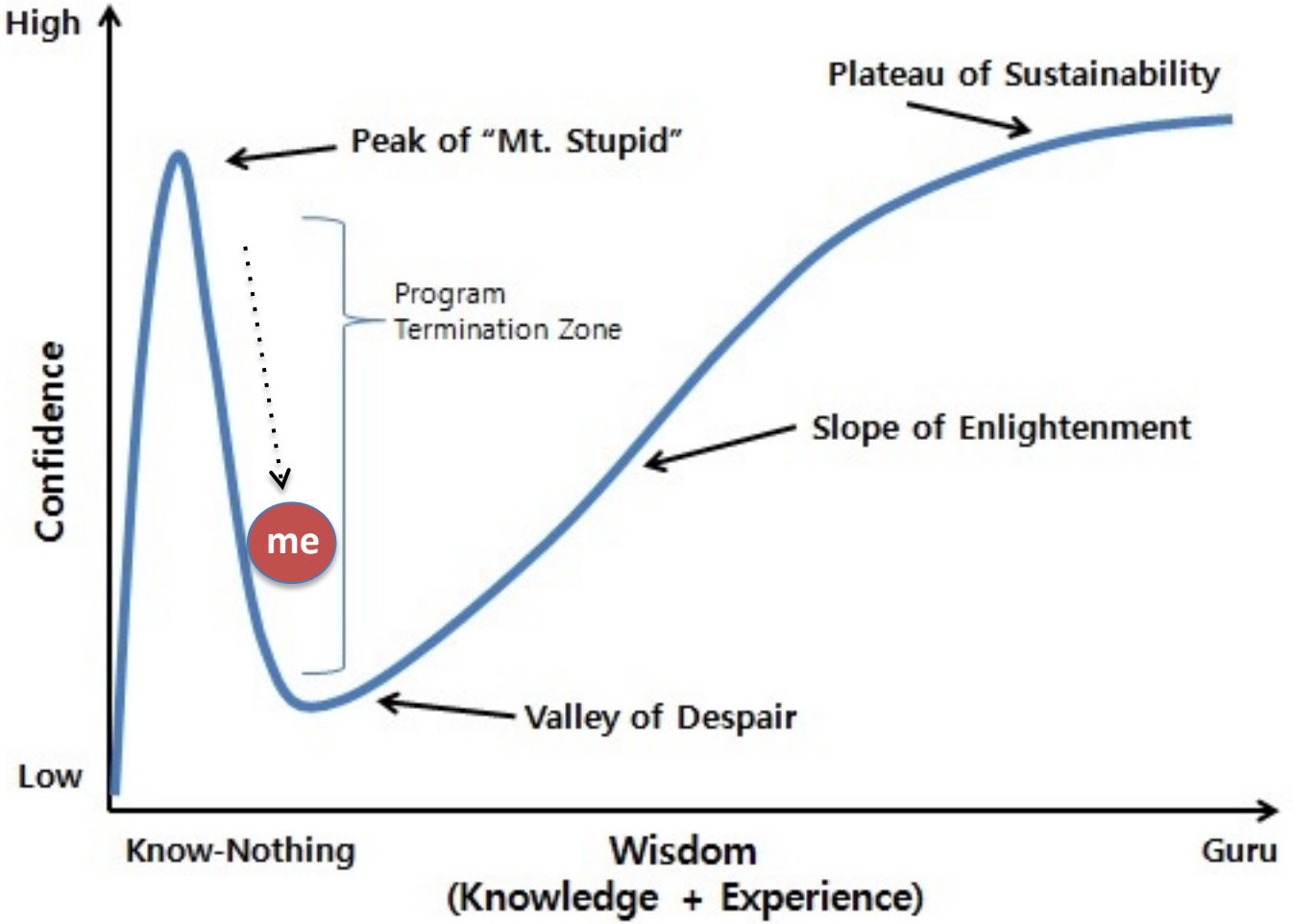


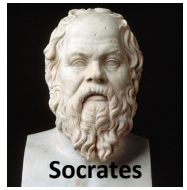
osteoclasts

How do gene regulatory networks function as *systems*.

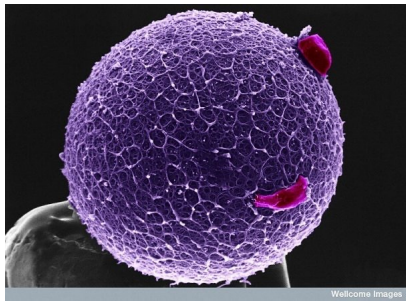
- What is a cell type?
- How is cell identity stabilized?
- Where is the key information? What does not matter?

Dunning-Kruger Effect

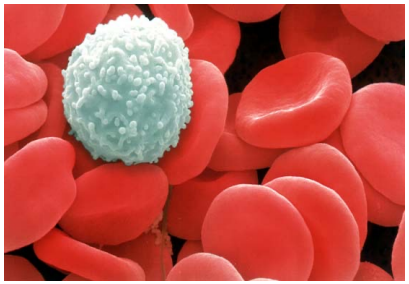




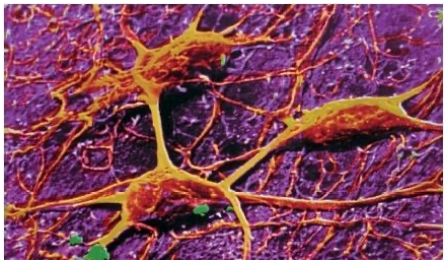
How is the regulatory code in the DNA `read out' to control cell fate and identity?



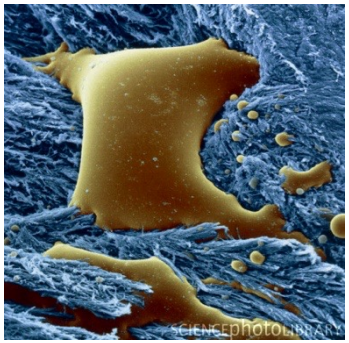
egg cell with 2 coronal cells



white and red blood cells



three neurons



osteoclasts

How do gene regulatory networks function as *systems*.

- What is a cell type?
- How is cell identity stabilized?
- Where is the key information? What does not matter?

My worries

- We think we know/measure a lot, but there is orders of magnitude more we do not know.
- Nowhere near the ability to meaningfully model what is going on.
- High-throughput measurements full of artifacts and biases that we poorly understand.
- Data analysis typically involves dizzying arrays of normalizations, filters, and transformations.

What useful things can computational analysis offer?

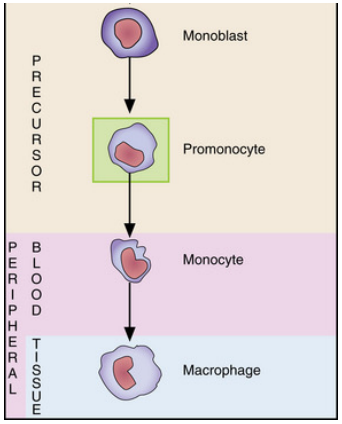
Robust and transparent methods that help guide experimental efforts.

What does my transcriptome/epigenome data say about regulation in my system?

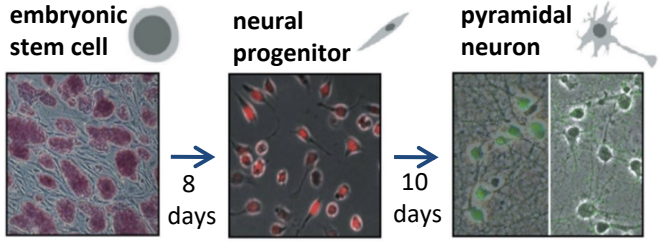
Typical questions:

What are the key regulators? What are their roles? Which pathways do they target?

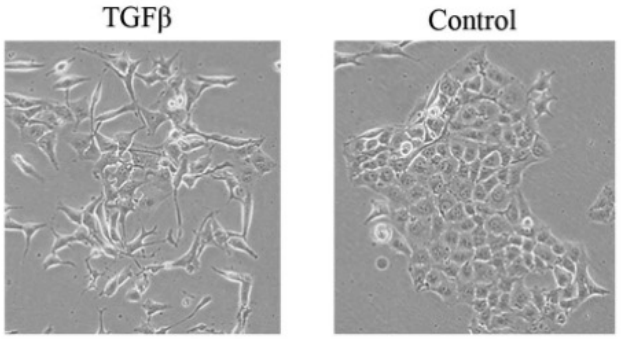
monoblast to macrophage differentiation



Mouse ES cells differentiating into pyramidal neurons



TGF-β induced EMT



Challenges

- Cannot do saturating genetic screens (too many candidate TF/miRNA regulators).
- Easy to do high-throughput measurements (microarray, RNA-seq, ChIP-seq, ATAC-seq).
- *Experimental labs often do not have the expertise to infer regulation from such data.*
- Collaborations with dedicated computational labs on a *per case* basis are big investment of time and effort.

Typical analysis of transcriptomic data

METHOD | OPEN ACCESS

Differential expression analysis for sequence count data

Simon Anders and Wolfgang Huber

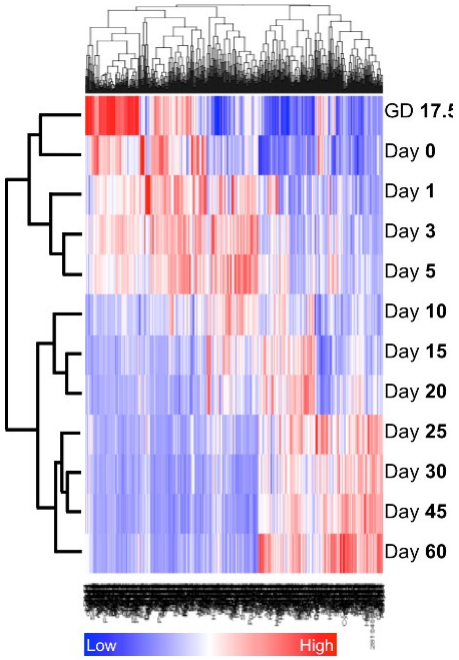
Genome Biology 2010 11:R106 | https://doi.org/10.1186/gb-2010-11-10-r106 | © Anders et al 2010

Received: 20 April 2010 | Accepted: 27 October 2010 | Published: 27 October 2010

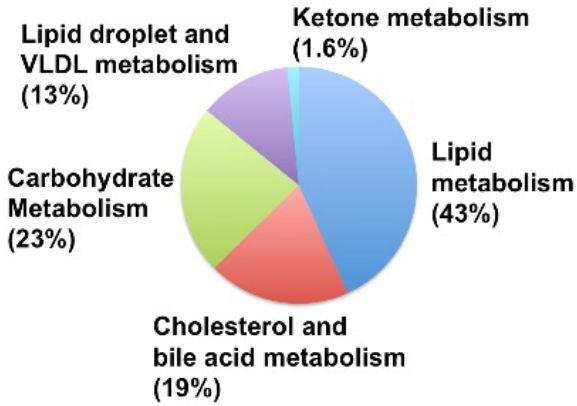
Basic processing

- Map raw reads to transcripts.
- Find all genes that are expressed.
- Find genes that are *differently expressed* across conditions, e.g. using *DESeq*.

Clustering genes with similar expression



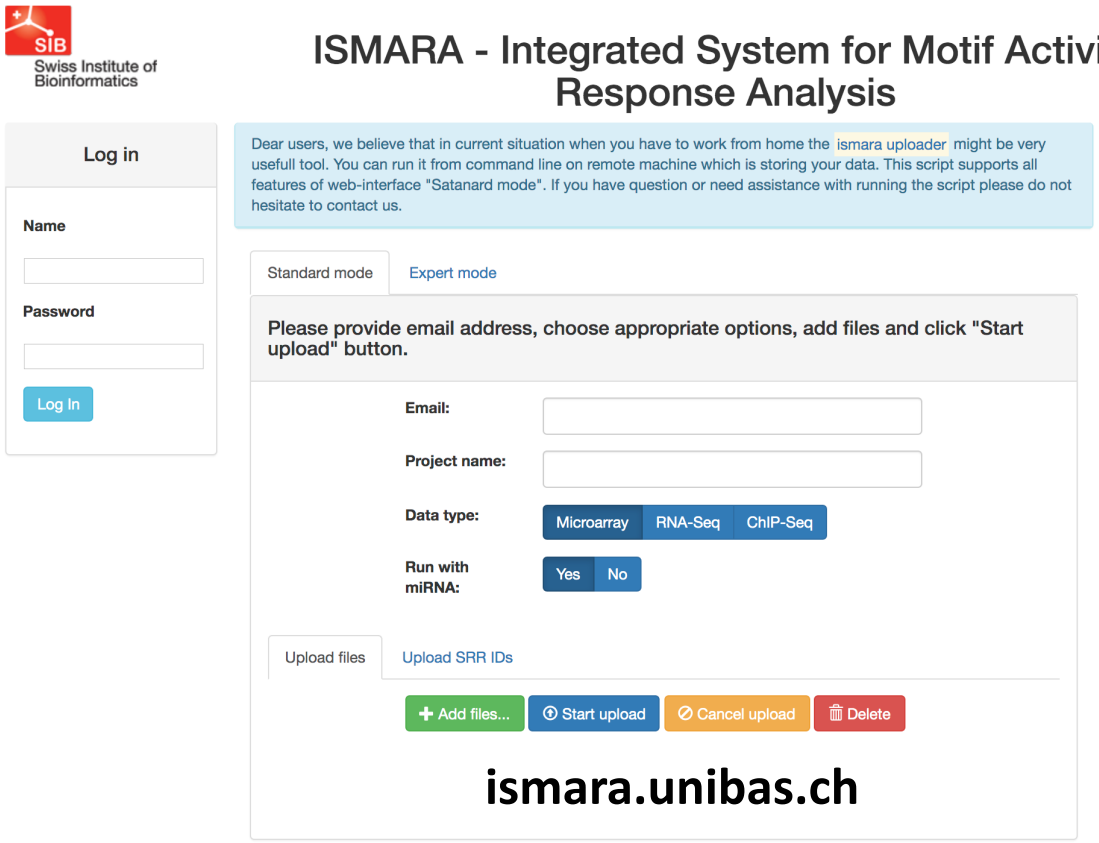
Enriched categories among gene sets



Limitations of these traditional approaches

- Does not infer anything about gene *regulation*.
- Often unclear how to experimentally follow-up.

Completely automated prediction of regulatory interactions from high-throughput data



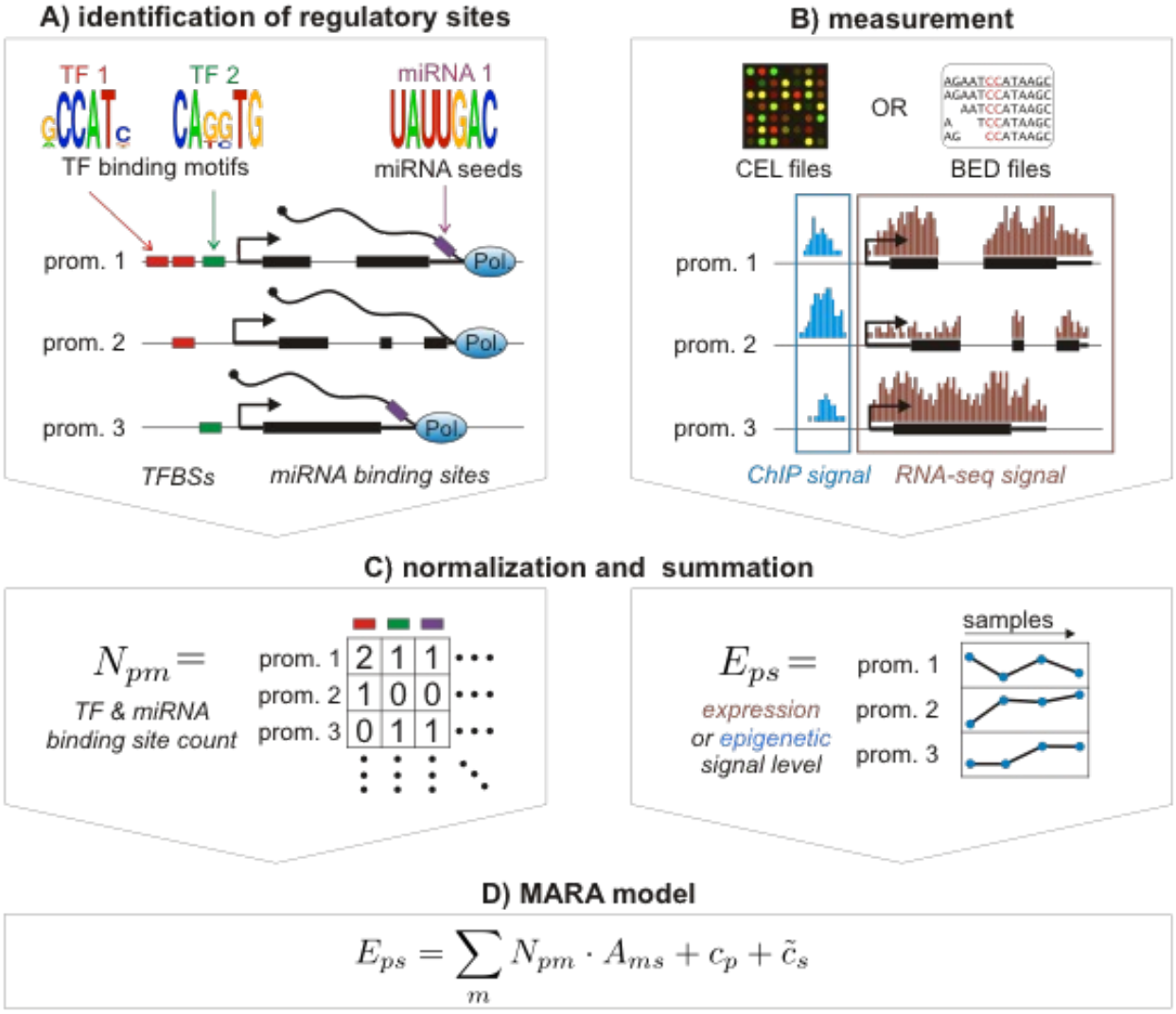
Suzuki et al.
***Nat Genet* 2009**

Balwierz et al.
***Genome Res* 2014**

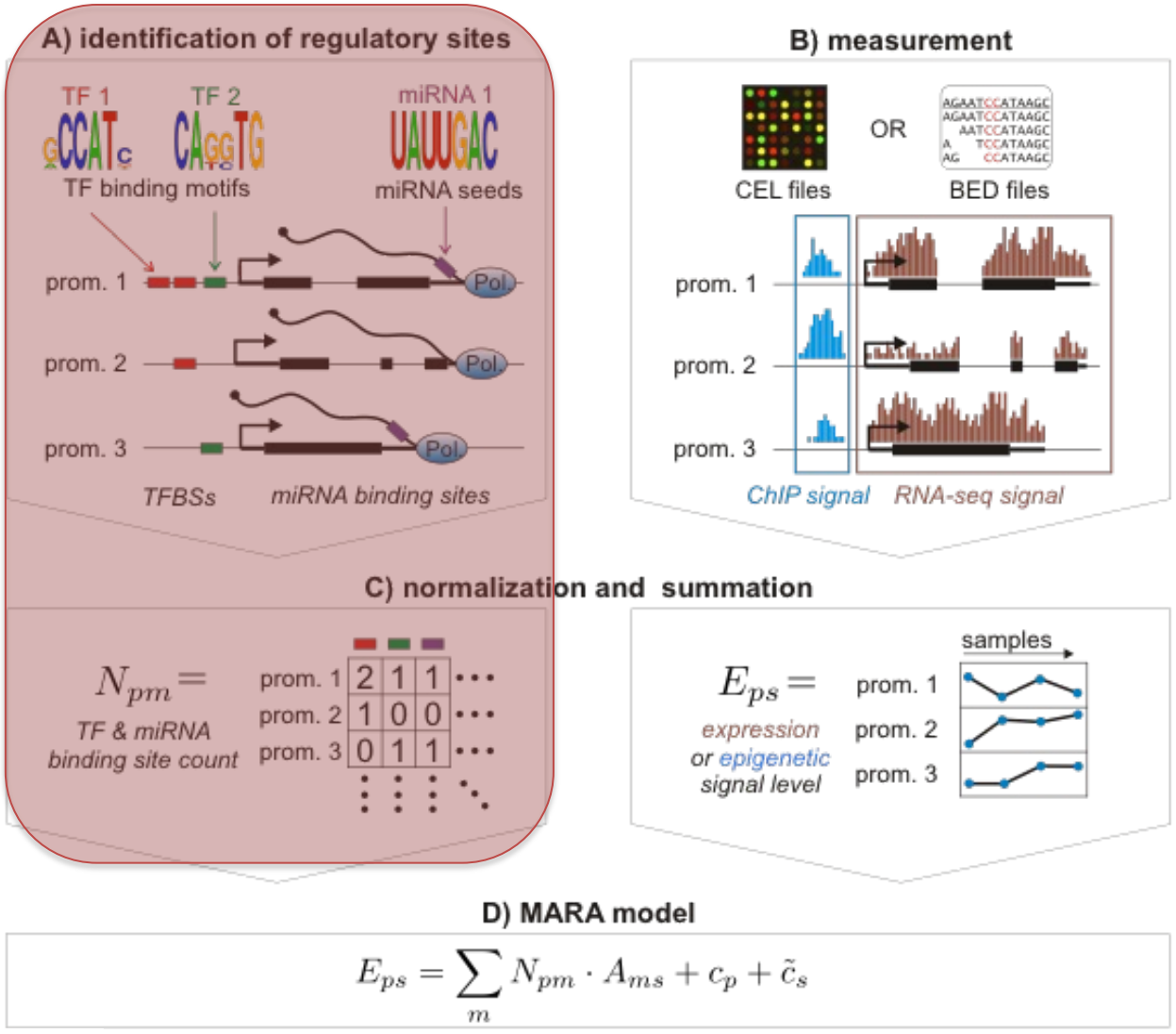
Upload raw micro-array oe RNA-seq data and predict:

- Key regulators (TFs/miRNAs) in the system.
- Regulator activities across the input samples.
- Sets of target genes and pathways for each regulator.
- The regulatory sites on the genome through which each regulator acts.
- Interactions between the regulators.

Modelling gene expression and chromatin state in terms of TFBS using a linear model



Modelling gene expression and chromatin state in terms of TFBS using a linear model



Constructing reference promoteromes and transcriptomes

Input data

- Collections of experimentally measured transcription start sites (e.g. CAGE).
- Collections of know full-length mRNAs (e.g. Genbank, Gencode, or Ensembl).

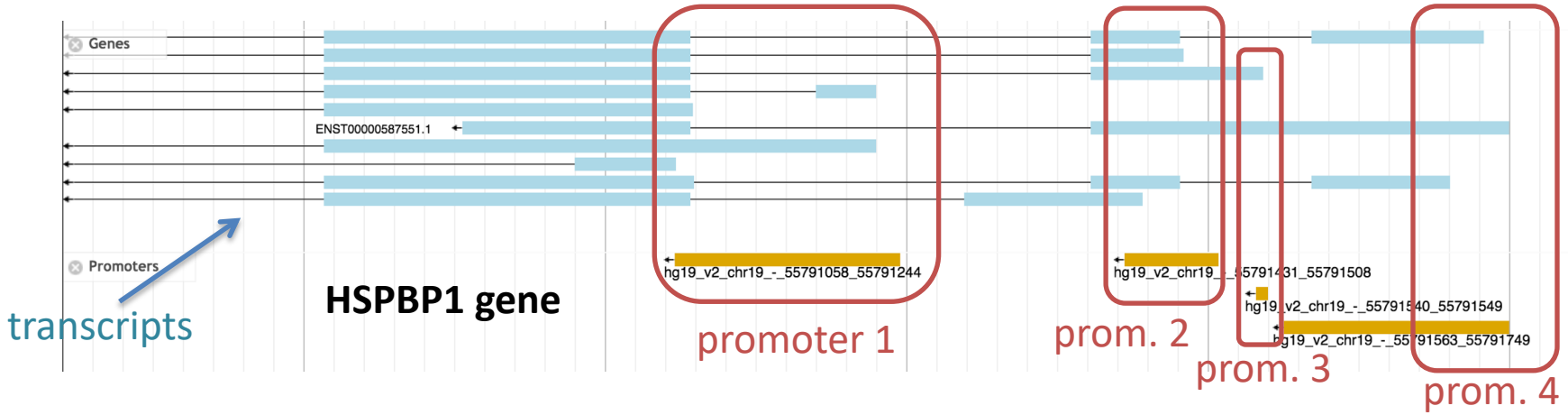
Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data

Piotr J Balwierz, Piero Carninci, Carsten O Daub, Jun Kawai, Yoshihide Hayashizaki, Werner Van Belle, Christian Beisel and Erik van Nimwegen

Genome Biology 2009 10:R79 | https://doi.org/10.1186/gb-2009-10-7-r79

Received: 23 October 2008 | Accepted: 22 July 2009 | Published: 22 July 2009

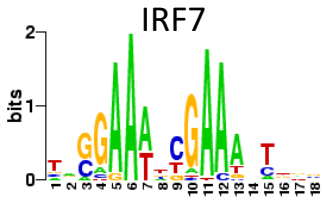
Procedure: Cluster nearby TSSs with mRNA starts

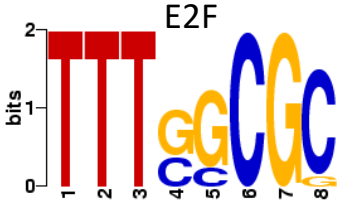


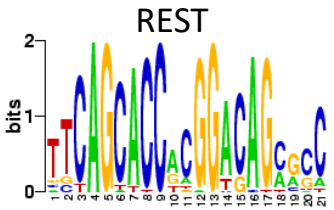
- mRNA starts are clustered with TSSs within 150bps (one nucleosome) of each other.
- Each cluster corresponds to a promoter.
- Only clusters with associated transcripts are retained.

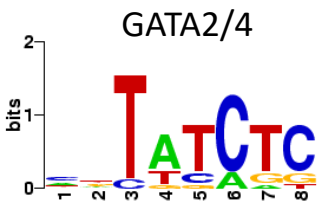
Regulatory motifs represent sequence binding preferences of transcription factors

Position specific weight matrix representation (sequence logos):



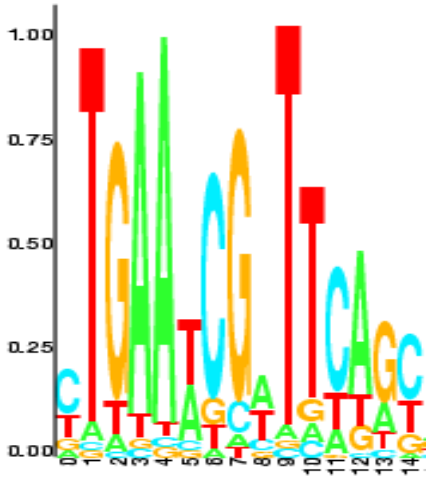






Example: E. coli's **fruR** binding sites and weight matrix

```
AAGCTGAATCGATTTTATGATTTGGT
AGGCTGAATCGTTTCAATTCAGCAAG
CTGCTGAATTGATTCAGGTCAGGCCA
GTGCTGAAACCATTCAAGAGTCAATT
GTGGTGAATCGATACTTTACCGGTTG
CGACTGAAACGCTTCAGCTAGGATAA
TGACTGAAACGTTTTTGCCCTATGAG
TTCTTGAAACGTTTCAGCGCGATCTT
ACGGTGAATCGTTCAAGCAAATATAT
GCACTGAATCGGTTAACTGTCCAGTC
ATCGTTAAGCGATTCAGCACCTTACC
**gcTGAAtCG*TTcAg**c******
```



w_α^i = Probability of finding base α at position i.

Example, position 4:

$w_A^4 = 0.06,\ w_C^4 = 0.53\ ,\ w_G^4 = 0.27\ ,\ w_T^4 = 0.13$

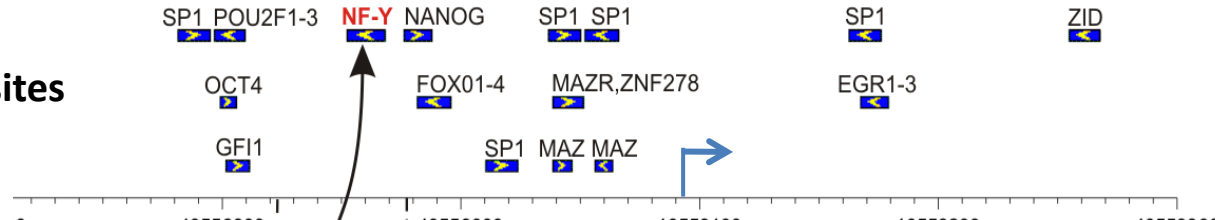
Probability that a site for the TF has sequence s: $P(s \mid w) = \prod_{i=1}^{l} w_{s_i}^i$

Predicting transcription factor binding sites using comparative genomics: **MotEvo**

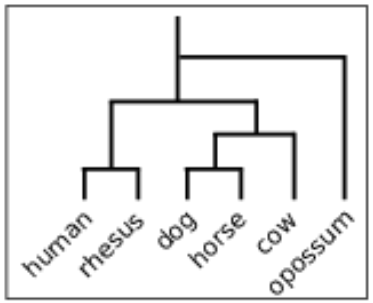
TF binding site prediction procedure

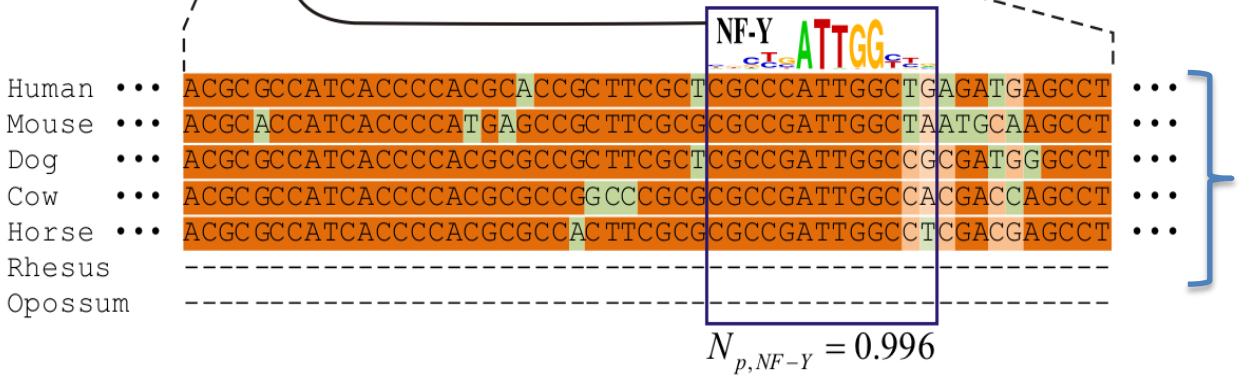
- For each promoter, collect promoter sequence plus 500bp upstream and 500bp downstream.
- Align each promoter region with orthologous regions from other species.
- For each motif in the motif collection, predict binding sites using the MotEvo algorithm.

Predicted sites



Phylogeny:





Alignment of orthologous promoter sequences

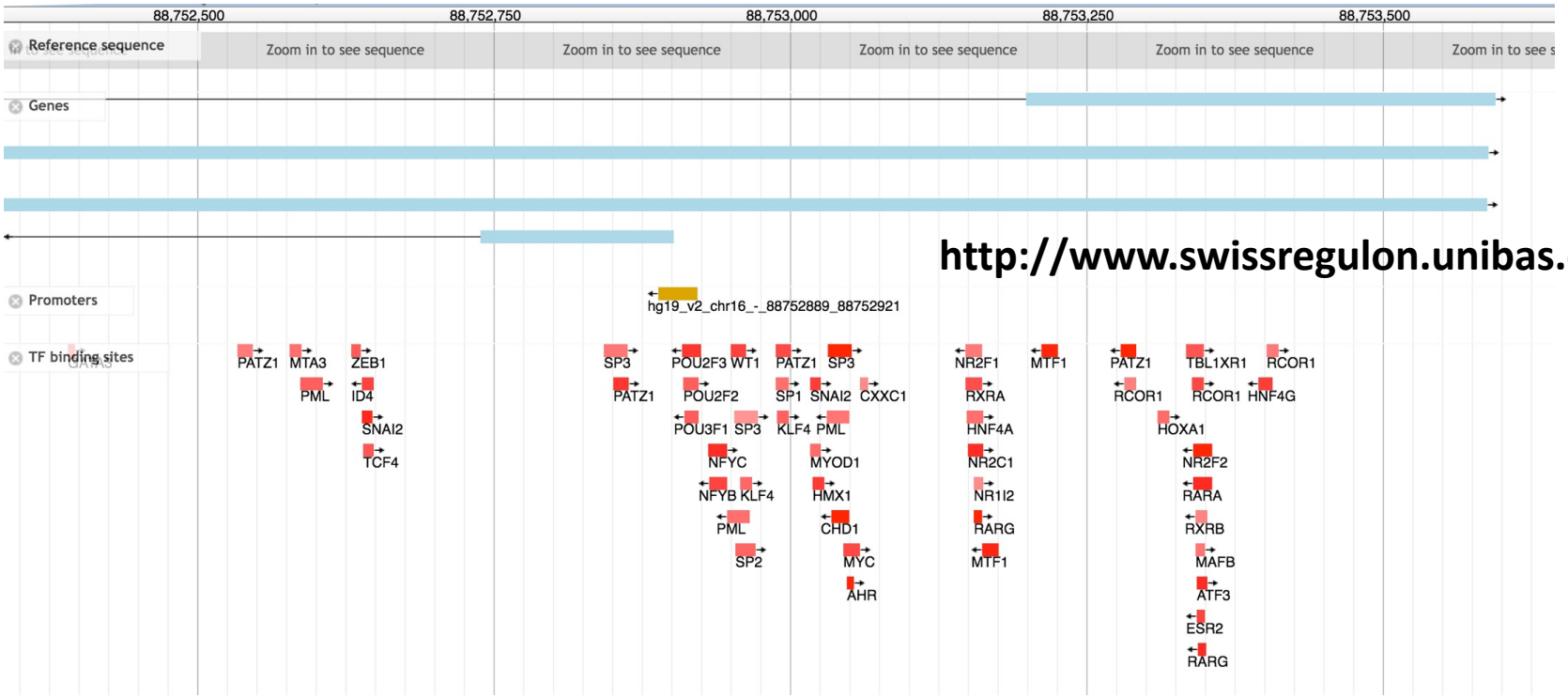
Arnold et al, *Bioinformatics.* 2012 Feb 15;28(4):487-94.

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

Genome-wide annotation of regulatory sites in promoters

Example: Predicted TFBSs in the proximal promoter of the SNAI3 TF.



http://www.swissregulon.unibas.ch

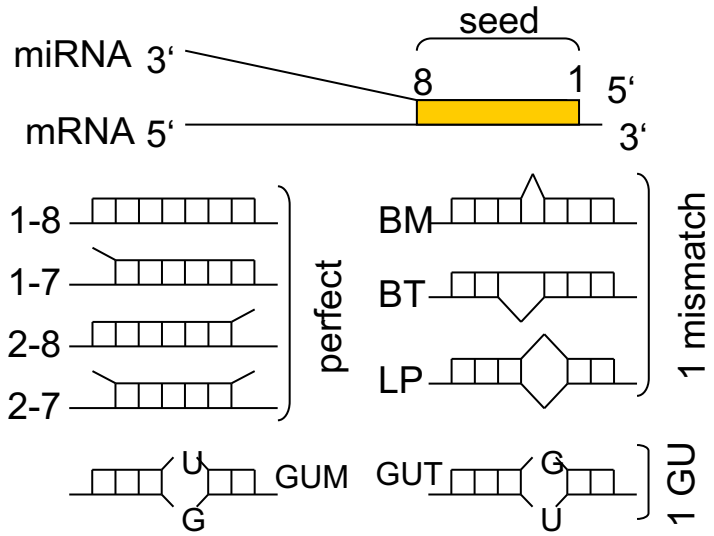
Summarizing the TFBS predictions

Sum the posteriors of the predicted sites for each motif to obtain a **matrix of site-counts**:

$$N_{pm} = \text{Total number of sites for motif } m \text{ in promoter } p.$$

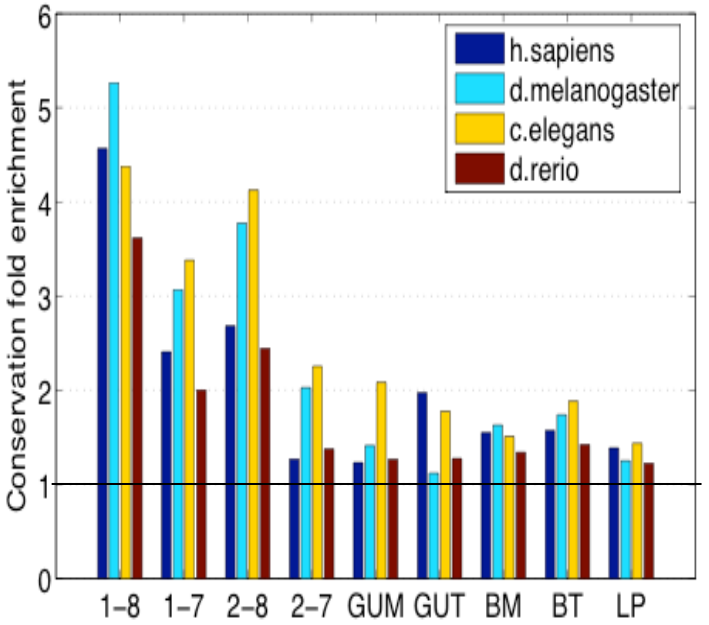
Including regulation by miRNAs

miRNAs destabilize mRNAs by hybridizing to sites in their 3' UTRs.



miRNAs bind mainly through their 5' *seed* region

Gaidatzis et al. BMC Bioinformatics 2007



Sites with a perfect seed match show strong conservation

We include predicted sites for 86 seed families (based on seed conservation analysis from TargetScan version 7). **Site counts**:

$$N_{p\mu} = \text{Average number of sites for seed motif } \mu \text{ in transcripts associated with promoter } p.$$

Curating a set of mammalian motifs and sites

1. Large motif collection from: SwissRegulon, CRUNCH, JASPAR, Hocomoco, Homer, Uniprobe, Encode, HT-SELEX.

2. Multiple candidate motifs for each transcription factor (TF).

3. Selecting an optimal set of motifs: Run ISMARA on the **FANTOM5 expression atlas** (889 human/388 mouse samples) selecting one motif per TF (simulated annealing).

4. **Redundancy removal**: Collapse similar motifs with statistically indistinguishable activities in the FANTOM5 atlas (Bayesian model selection).

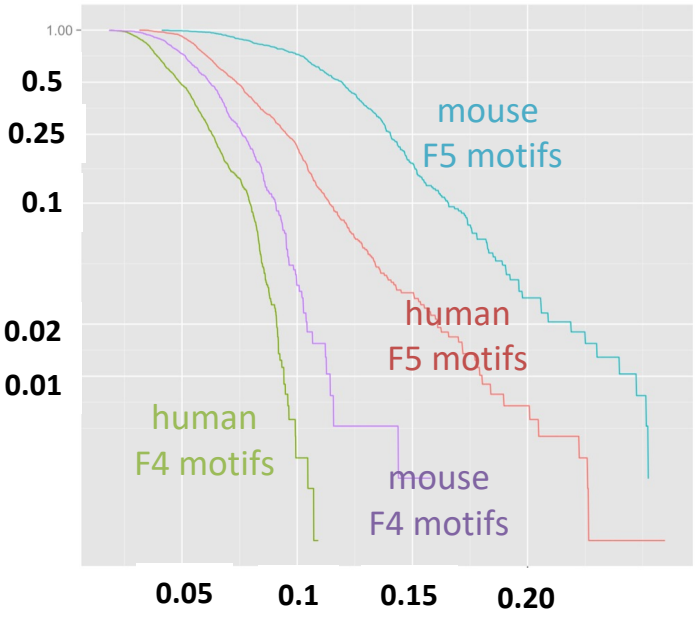
FANTOM
FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

Nature. 2014 Mar 27;507(7493):462-70. doi: 10.1038/nature13182.

A promoter-level mammalian expression atlas.

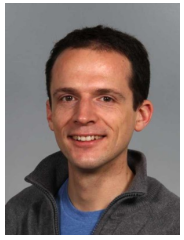
FANTOM Consortium and the RIKEN PMI and CLST (DGT).

Reverse cumulative: explained variance per sample



	human	mouse
Initial motifs	2181	2035
Associated TFs	682	679
miRNAs	106	99
Motif groups	499	503

Daniel Schmocker

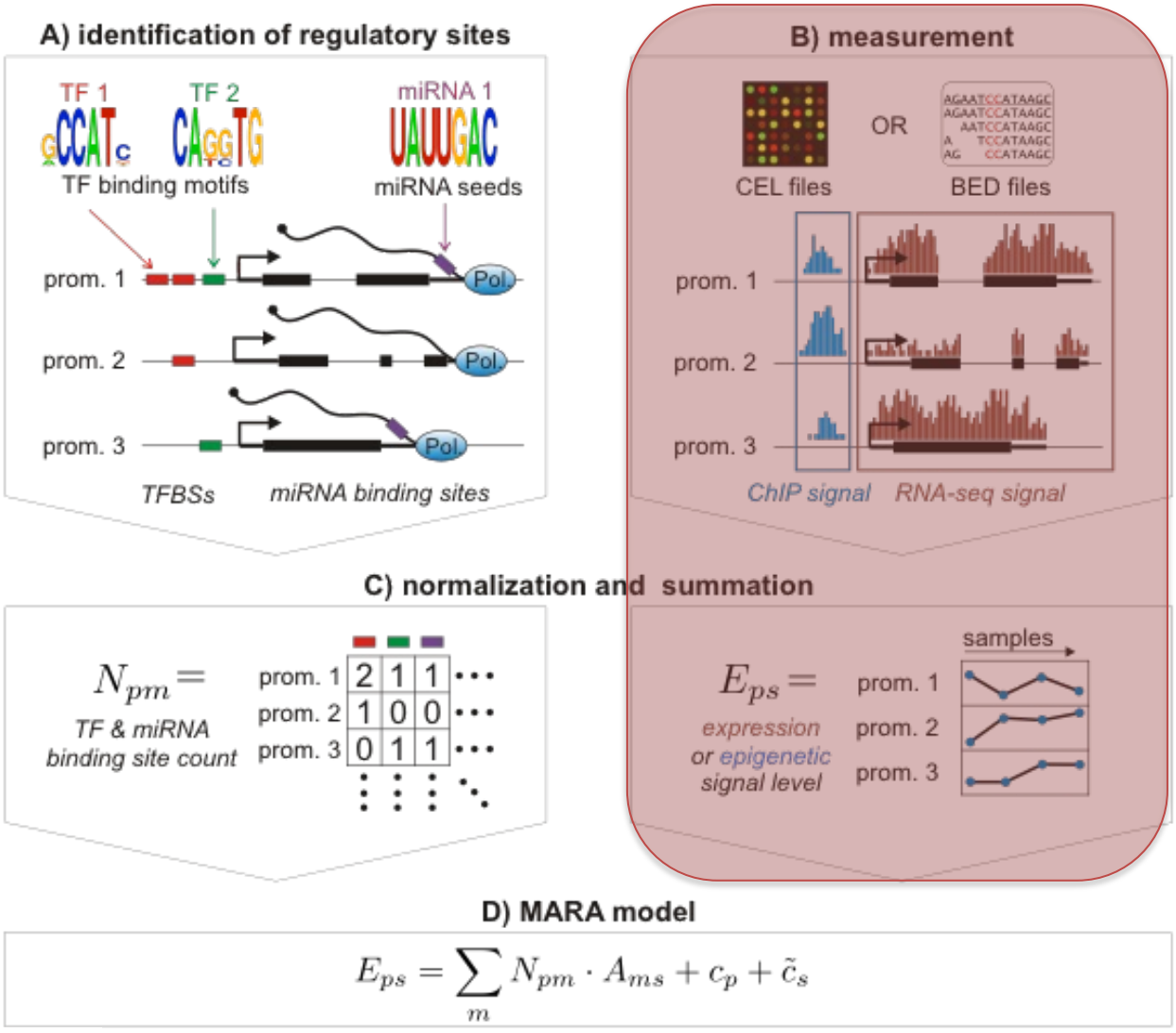


Florian Geier

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

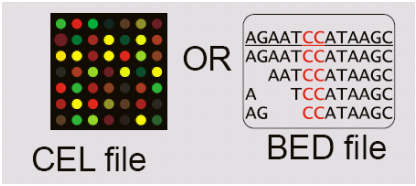
SIB
Swiss Institute of Bioinformatics

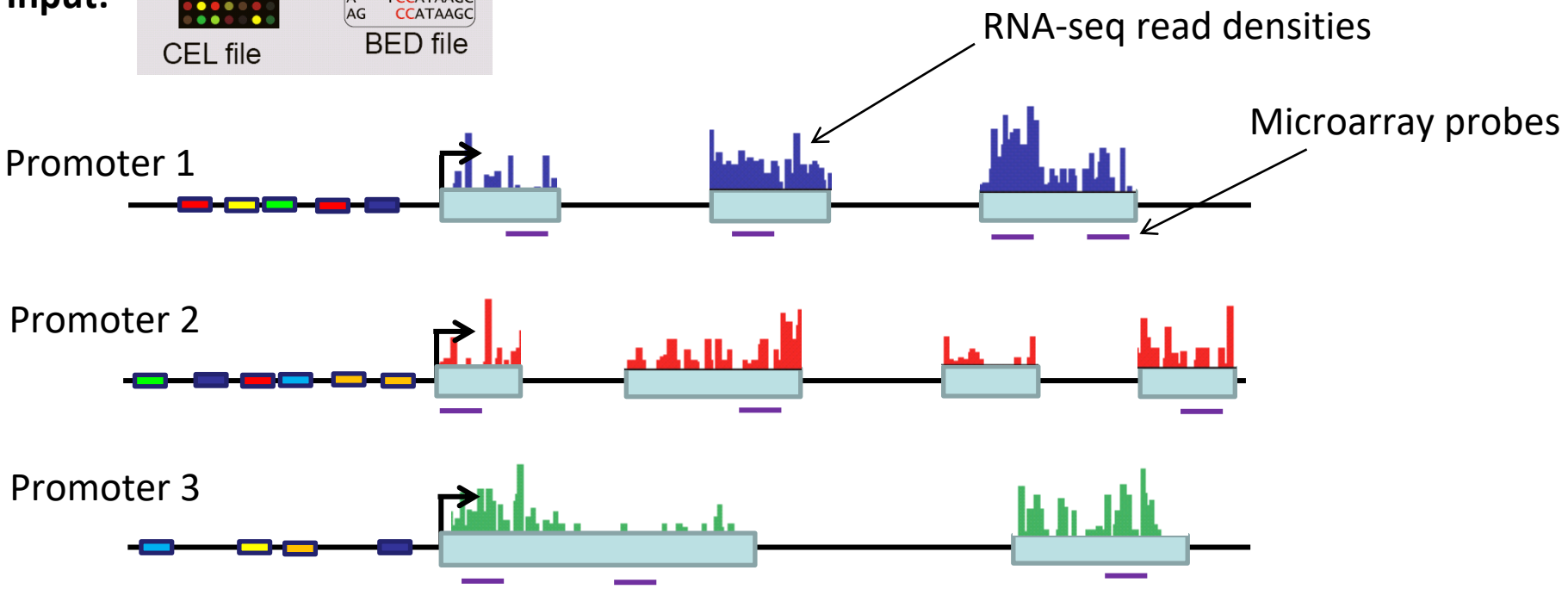
Modeling gene expression and chromatin state in terms of TFBS using a linear model



Quantifying genome-wide expression

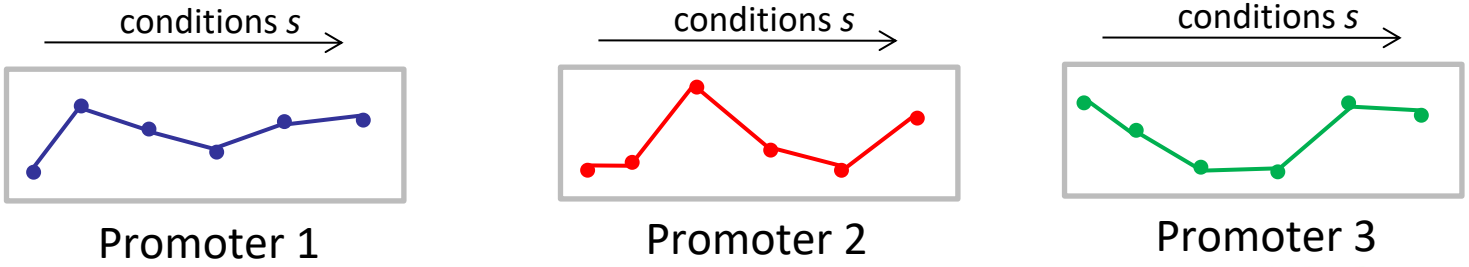
Input:





Output:
expression profiles

$$E_{ps}$$



BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

Mapping reads to transcripts

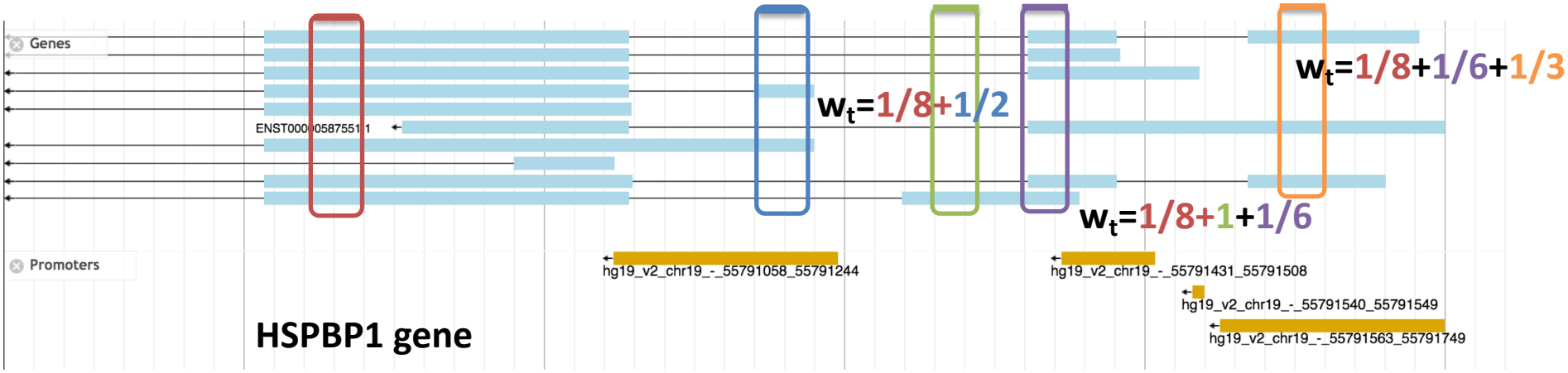
- Each RNA-seq read is mapped to the *transcriptome* using Kallisto.
- The weight of each read is distributed *uniformly* over all transcripts consistent with it.

Near-optimal probabilistic RNA-seq quantification

Nicolas L Bray, Harold Pimentel, Páll Melsted & Lior Pachter

Affiliations | Contributions | Corresponding author

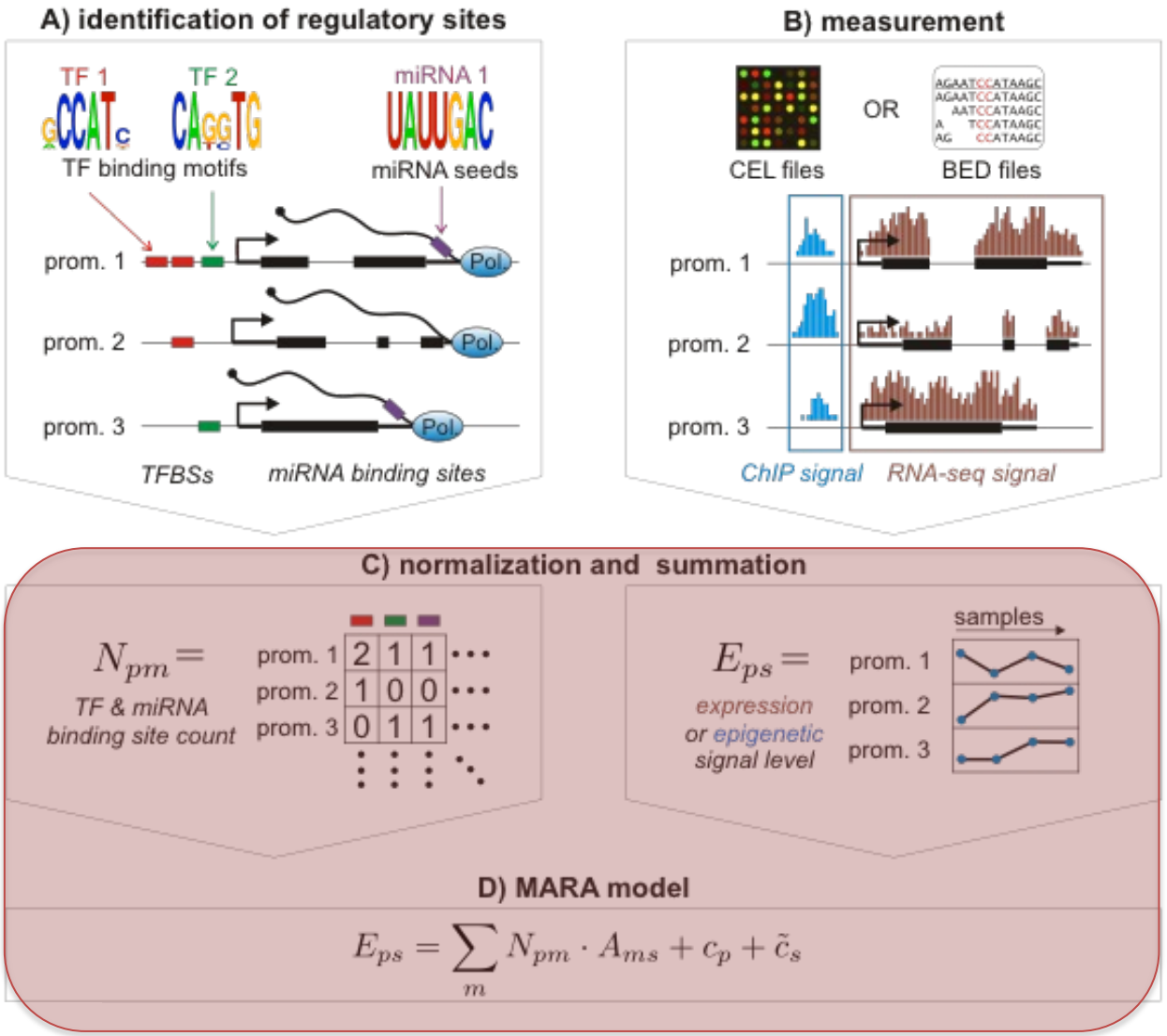
Nature Biotechnology **34**, 525–527 (2016) | doi:10.1038/nbt.3519
Received 15 October 2015 | Accepted 25 February 2016 | Published online 04 April 2016



- Each transcript's total weight w_t is the sum of the weights of all reads mapping to the transcript. Then weight is divided by transcript length $W_t = w_t/L_t$.
- A promoter's total weight W_p is the sum of the weights of its transcripts: $W_p = \sum_{t \in p} W_t$
- A pseudo-count is added (constant corresponding to 0.5 tpm): $W_p \rightarrow W_p + \lambda$
- The weights are rescaled to represent transcripts-per-million (tpm), and log-transformed:

$$E_p = \log_2 \left[10^6 \frac{W_p}{\sum_{\tilde{p}} W_{\tilde{p}}} \right]$$

Modeling gene expression and chromatin state in terms of TFBS using a linear model



MARA's linear model

- Measurements are represented as a matrix E_{ps} of expression across all promoters and samples.
- Each sample (column of the matrix) is normalized by subtracting the mean expression, and site counts are normalized to sum to zero across promoters.

$$E_s = \frac{1}{P}\sum_p E_{ps} \qquad E_{ps} \rightarrow E'_{ps} = E_{ps} - E_s \qquad N_m = \frac{1}{P}\sum_p N_{pm} \qquad N_{pm} \rightarrow \tilde{N}_{pm} = N_{pm} - N_m$$

- We model the expression in terms of the site counts and *motif activities*

$$E'_{ps} = noise + \sum_m \tilde{N}_{pm} A_{ms}$$

- We separate the fitting into a fit of the **average expression**:

$$\left\langle E'_p \right\rangle = \frac{1}{S}\sum_s E'_{ps} \qquad \left\langle A_m \right\rangle = \frac{1}{S}\sum_s A_{ms} \qquad \text{model: } \left\langle E'_p \right\rangle = noise + \sum_m \tilde{N}_{pm} \left\langle A_m \right\rangle$$

- And fitting of **expression changes** across the conditions:

$$\tilde{E}_{ps} = E'_{ps} - \left\langle E'_p \right\rangle \qquad \tilde{A}_{ms} = A_{ms} - \left\langle A_m \right\rangle \qquad \tilde{E}_{ps} = noise + \sum_m N_{pm} \tilde{A}_{ms}$$

Note: $\sum_s \tilde{E}_{ps} = \sum_s \tilde{A}_{ms} = 0$

Fitting MARA's linear model (technical)

$$\tilde{E}_{ps} = noise + \sum_m \tilde{N}_{pm}\tilde{A}_{ms}$$

Assume the noise is Gaussian gives likelihood:

$$P(\tilde{E} \mid \tilde{A}) \propto \exp\left[-\frac{\sum_{p,s}\left(\tilde{E}_{ps} - \sum_m \tilde{N}_{pm}\tilde{A}_{ms}\right)^2}{2\sigma^2} \right]$$

To avoid overfitting, we include a Gaussian prior (with average zero) over motif activities (ridge regression):

$$P(\tilde{A} \mid \lambda) \propto \exp\left[-\frac{\lambda^2}{2\sigma^2}\sum_{m,s}\tilde{A}_{ms}^2 \right]$$

The optimal posterior activities $\tilde{A}^*_{ms}$ and the posterior distribution over the activities can be easily determined through Singular Value Decomposition:

$$P(\tilde{A}_s \mid \tilde{E}_s, \lambda) \propto \exp\left[-\frac{P}{2\chi_s^2}\sum_{m,\tilde{m}}(A_{ms} - A^*_{ms})W_{m\tilde{m}}(A_{\tilde{m}s} - A^*_{\tilde{m}s}) \right]$$

Here we have defined:

$$W_{m\tilde{m}} = \sum_p \tilde{N}_{pm}\tilde{N}_{p\tilde{m}} + \lambda^2\delta_{m\tilde{m}} \qquad A^*_{ms} = \sum_{m,p} W^{-1}_{m\tilde{m}}\tilde{N}_{p\tilde{m}}\tilde{E}_{ps} \qquad \chi_s^2 = \sum_p \left(\tilde{E}_{ps} - \sum_m \tilde{N}_{pm}A^*_{ms} \right)^2$$

The parameter λ of the prior is optimized by maximizing the likelihood of the data, marginalizing over all motif activities.

Fitting MARA's linear model (conceptual)

Linear model: $E_{ps} = \text{noise} + \sum_m N_{pm} A_{ms}$

- E_{ps}: Expression from promoter *p* in sample *s*
- N_{pm}: Number of functional sites in promoter *p* for regulator motif *m*
- A_{ms}: Activity of motif *m* in sample *s*

Bayesian inference of the motif activities

Obtain both best-fit activities and error-bars on the activities:

A^*_{ms} = Fitted activity of motif *m* in sample *s*.

δA_{ms} = Error-bar on the activity.

Significance of motif *m*:

$$z_m = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\frac{A^*_{ms}}{\delta A_{ms}}\right)^2}$$

Notes

- Motif activities capture the expression *changes* across the input samples.
- Activity meaning: A^*_{ms} is the amount by which log-expression of a transcript is predicted to go up in sample *s* when a site for motif *m* is added to its promoter.
- Significance meaning: z_m is the typical number of standard-deviations that the activity of motif *m* is away from its average of zero.

Example dataset: Mouse liver development

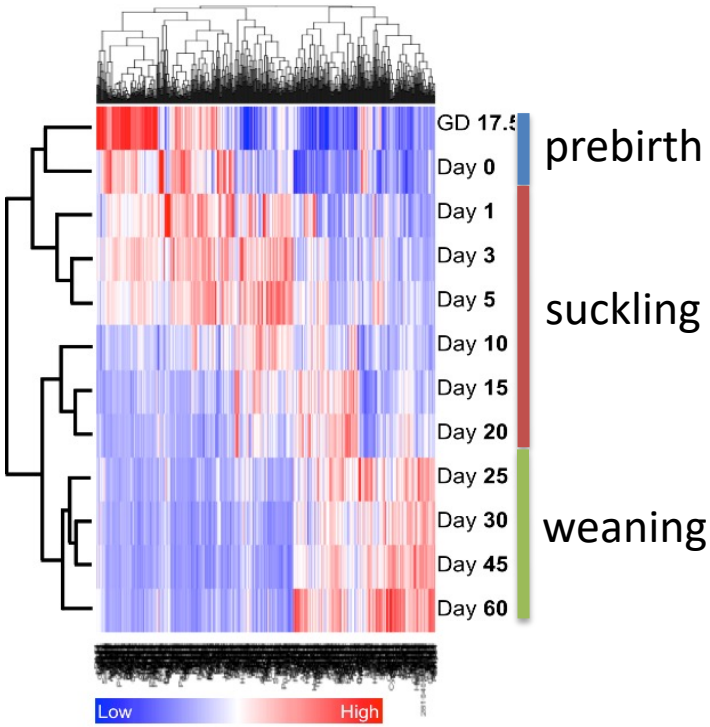
Ontogeny of Hepatic Energy Metabolism Genes in Mice as Revealed by RNA-Sequencing

Helen J. Renaud, Yue Julia Cui, Hong Lu, Xiao-bo Zhong, Curtis D. Klaassen

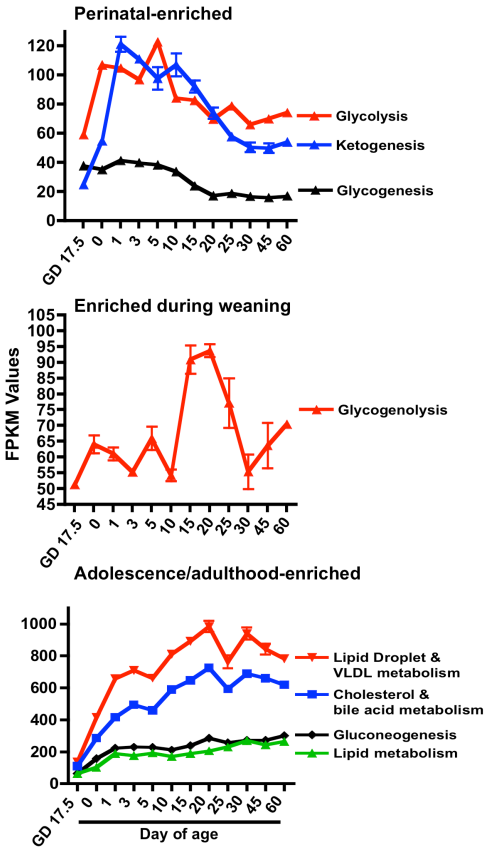
Published: August 7, 2014 • https://doi.org/10.1371/journal.pone.0104560

RNA-seq at 12 time points (in triplicate). Starting 2 days before birth, until 60 days after birth.

Clustering of expression profiles



Time-dependent expression of genes in different metabolic categories

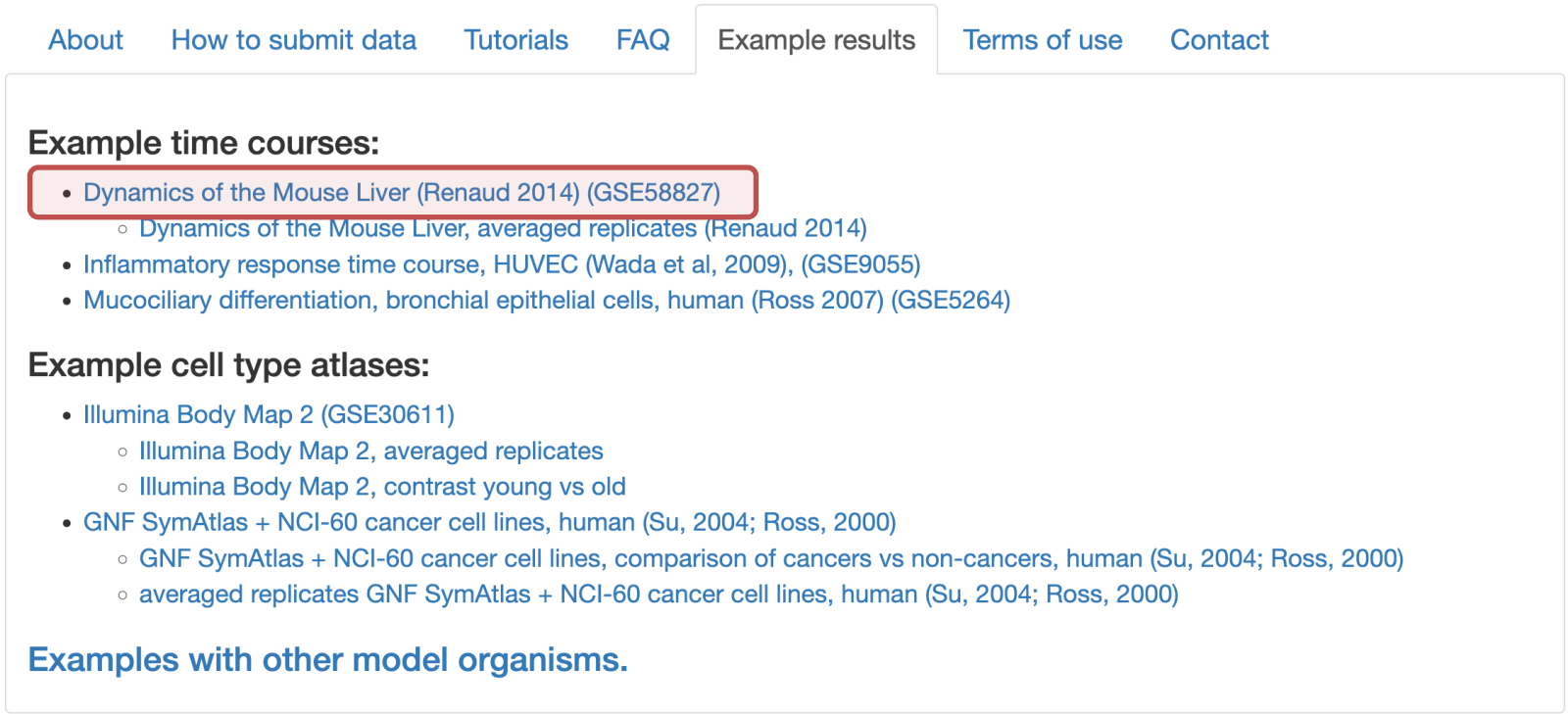


BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

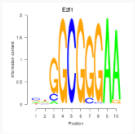
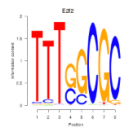
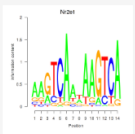
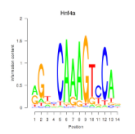
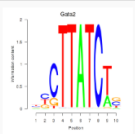
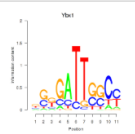
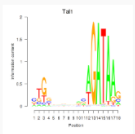
SIB
Swiss Institute of Bioinformatics

Example dataset: Mouse liver development

1. Go to: ismara.unibas.ch.
2. Click on the 'Example results' tab.
3. The mouse liver datasets are at the top.



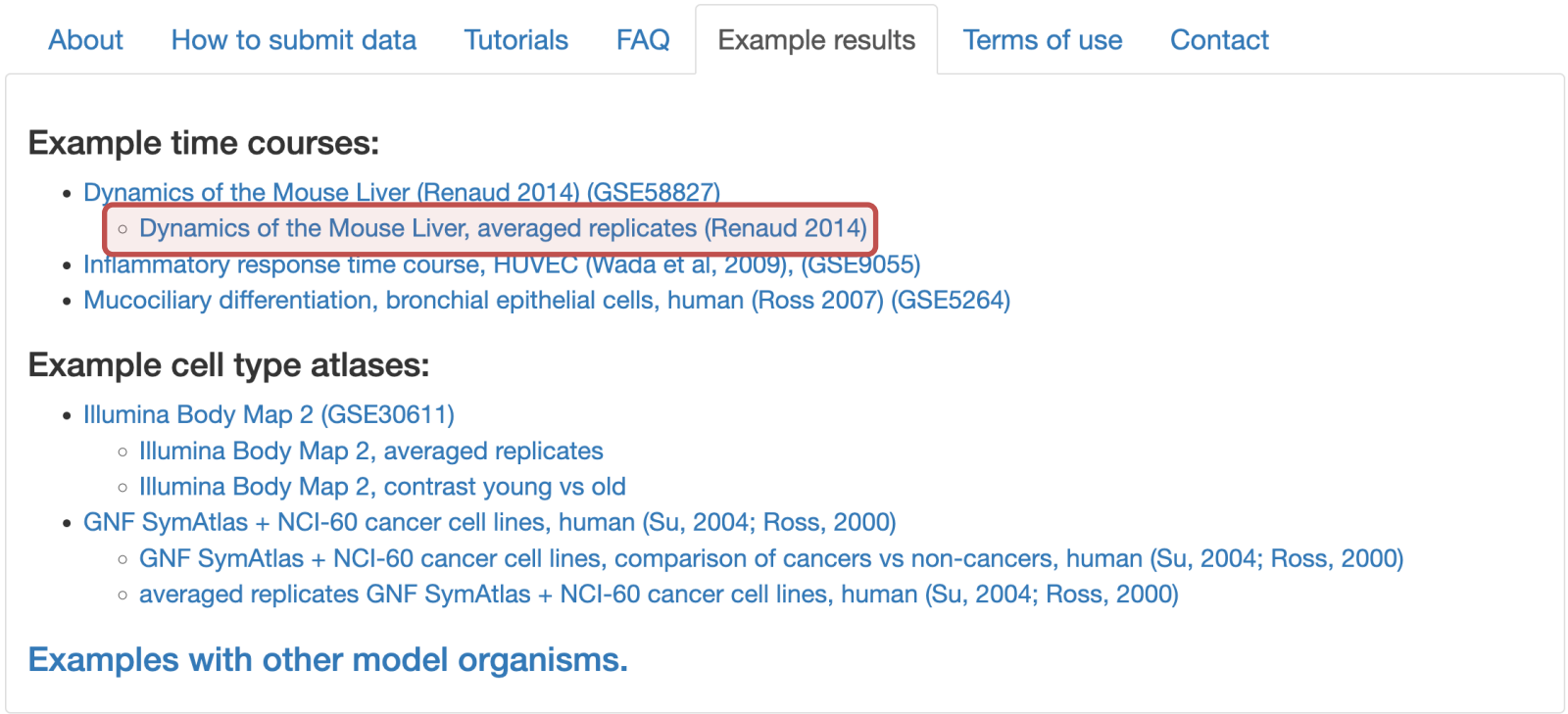
Most significant motifs (all samples)

Motif name	Z-value	Associated genes	Profile	Logo
E2f1	5.25	E2f1 Links		
E2f2_E2f5	5.21	E2f2 Links E2f5 Links		
Nr2e1	4.87	Nr2e1 Links		
Hnf4a	4.78	Hnf4a Links		
Gata2_Gata1	4.26	Gata2 Links Gata1 Links		
Ybx1_Nfya_Nfyb_Nfyc_Cebpz	4.22	Ybx1 Links Nfya Links Nfyb Links Nfyc Links Cebpz Links		
Tal1	4.01	Tal1 Links		

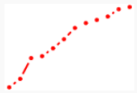
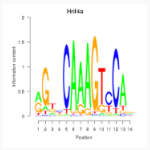
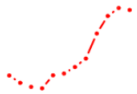
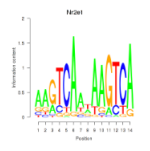
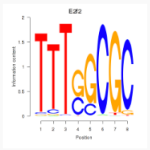
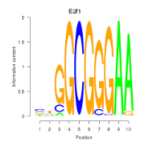
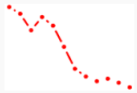
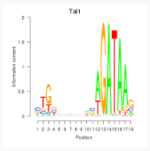
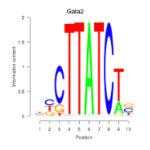
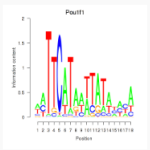
- Top motifs sorted by significance.
- Z-values.
- Names of the associated TF genes.
- Thumbnails of the motif activity across the time course.
- Sequence logos of the binding patterns of these motifs.

Example dataset: Mouse liver development

1. Go to: ismara.unibas.ch.
2. Click on the 'Example results' tab.
3. The mouse liver datasets are at the top.



Most significant motifs (replicate averaged)

Motif name	Z-value	Associated genes	Profile	Logo
Hnf4a	7.61	Hnf4a Links		
Nr2e1	7.31	Nr2e1 Links		
E2f2_E2f5	7.12	E2f2 Links E2f5 Links		
E2f1	6.52	E2f1 Links		
Tal1	6.09	Tal1 Links		
Gata2_Gata1	5.58	Gata2 Links Gata1 Links		
Pou1f1	5.28	Pou1f1 Links		

- Reorders motif significance.
- Top Z-values increase.

Most significant motifs (replicate averaged)

Motif name	Z-value	Associated genes	Profile	Logo
Hnf4a	7.61	Hnf4a Links		
		Nr2e1 Links		
E2f2_E2f5	7.12	E2f2 Links E2f5 Links		
E2f1	6.52	E2f1 Links		
Tal1	6.09	Tal1 Links		
Gata2_Gata1				
Pou1f1				

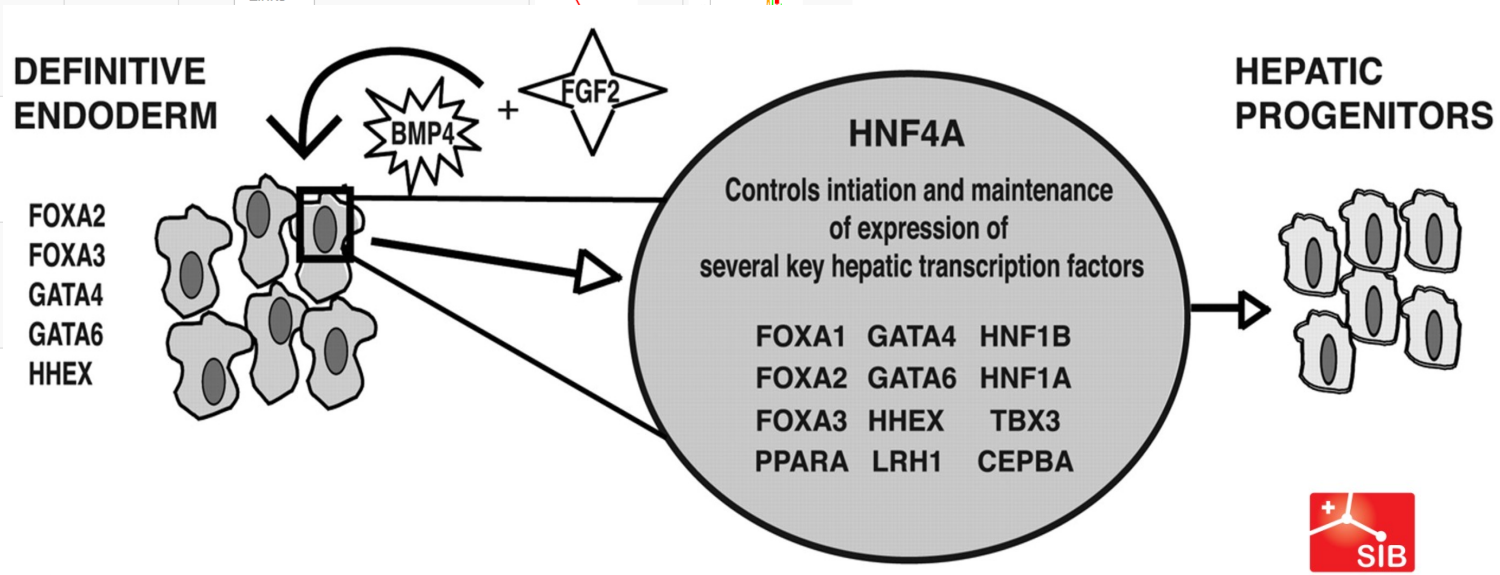
The most significant motif in the liver maturation data is HNF4a. HNF4a is a key TF in hepatocyte development, e.g.

DEVELOPMENT AND STEM CELLS

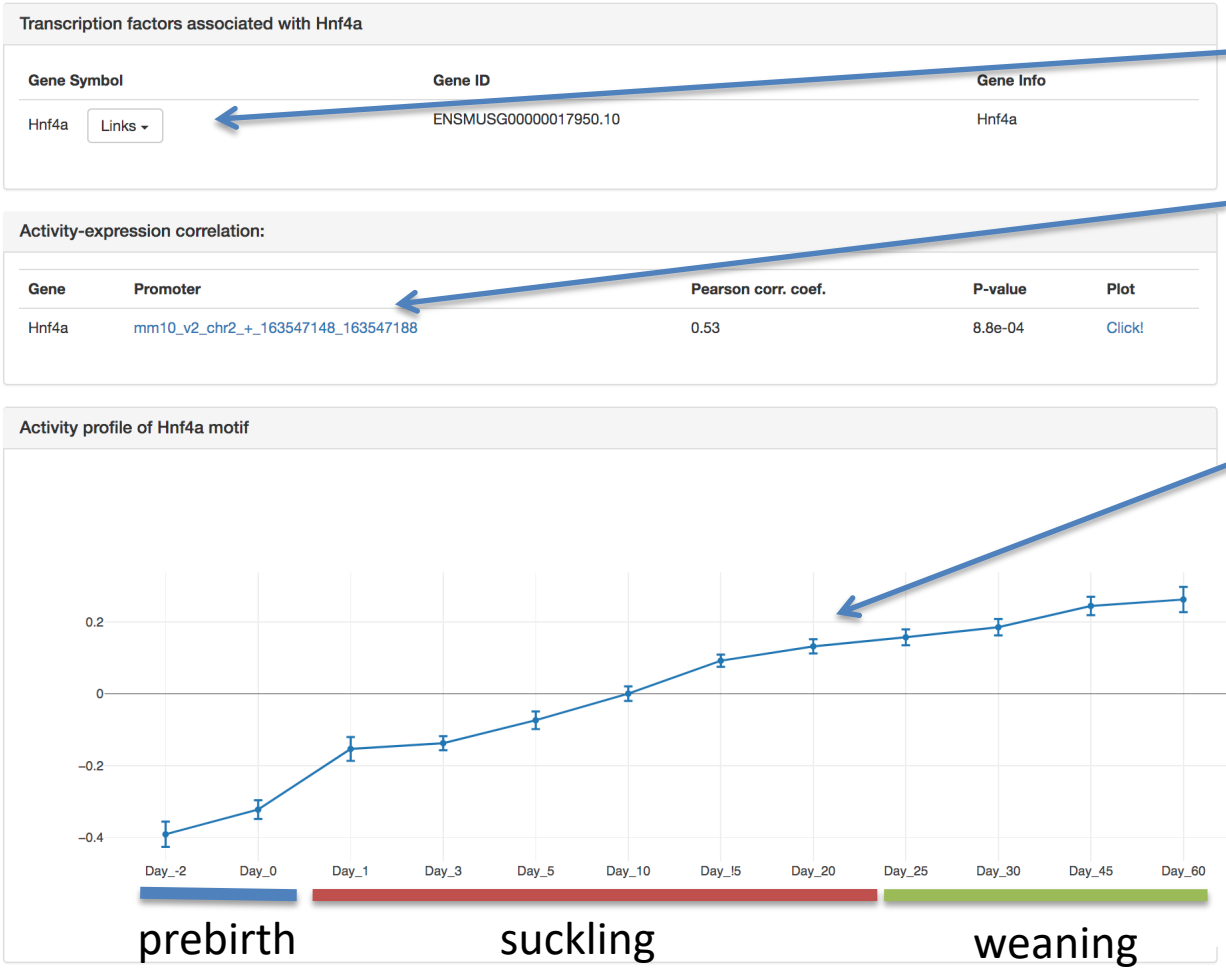
HNF4A is essential for specification of hepatic progenitors from human pluripotent stem cells

Ann DeLaForest, Masato Nagaoka, Karim Si-Tayeb , Fallon K. Noto, Genevieve Konopka, Michele A. Battle, Stephen A. Duncan

Development 2011 138: 4143-4153; doi: 10.1242/dev.062547



Information regarding the HNF4a motif



Links with information about the TF.

TF's promoter on the chromosome.

Activity of the motif at each time-point, and error bars on those activities.

Notes

- Motif activity increases with time.
- This means the *targets* of HNF4a (on average) increase expression with time.
- From -0.4 to 0.3 means the average effect of a single HNF4a site goes from 40% reduction of expression to 30% increase in expression relative to average expression.

Information regarding the HNF4a motif

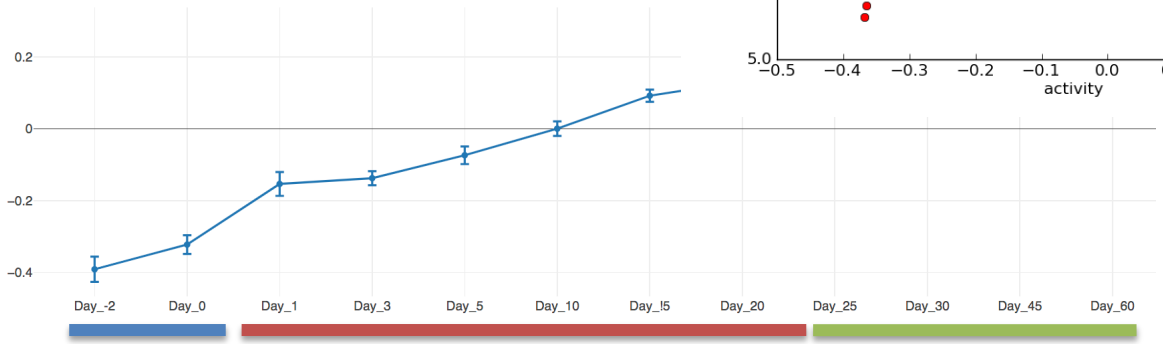
Transcription factors associated with Hnf4a

Gene Symbol	Gene ID	Gene Info
Hnf4a Links	ENSMUSG00000017950.10	

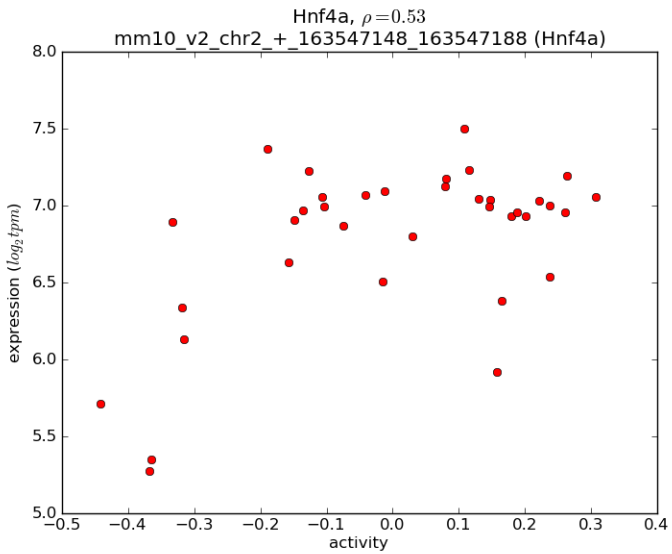
Activity-expression correlation:

Gene	Promoter	Pearson
Hnf4a	mm10_v2_chr2_+_163547148_163547188	0.53

Activity profile of Hnf4a motif



prebirth suckling weaning



HNF4a activity vs. mRNA expression of the HNF4a TF itself.

Notes

- Positive correlation indicates HNF4a acts as an *activator*.
- Transcripts per million go from $2^5 = 32$ to $2^7 = 128$.

Example with two TFs for one motif: E2F2_E2F5

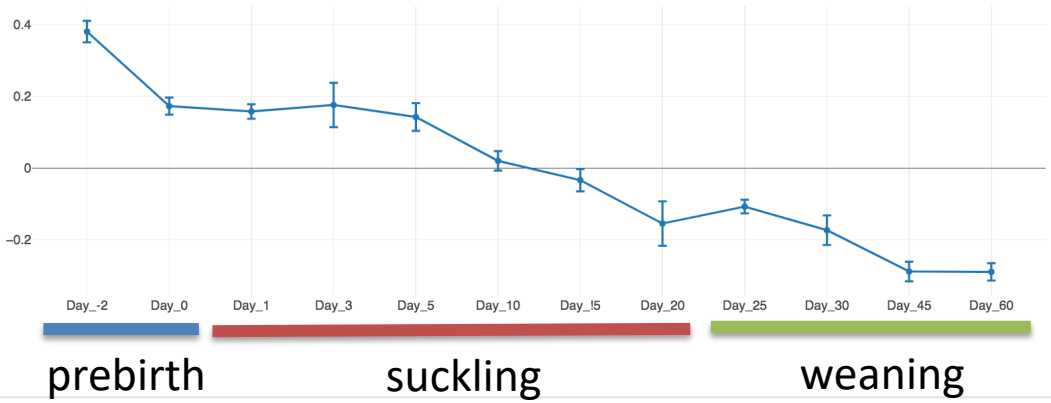
Transcription factors associated with E2f2_E2f5

Gene Symbol		Gene ID	Gene Info
E2f2	Links	ENSMUSG00000018983.9	E2f2
E2f5	Links	ENSMUSG00000027552.8	E2f5

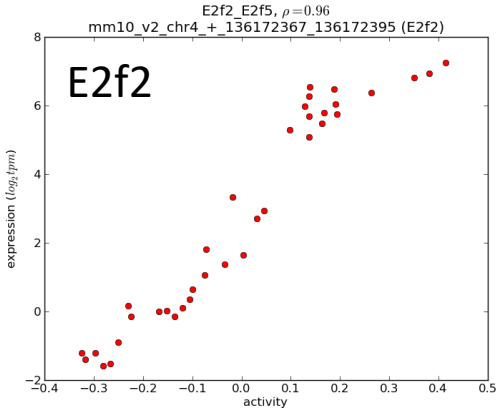
Activity-expression correlation:

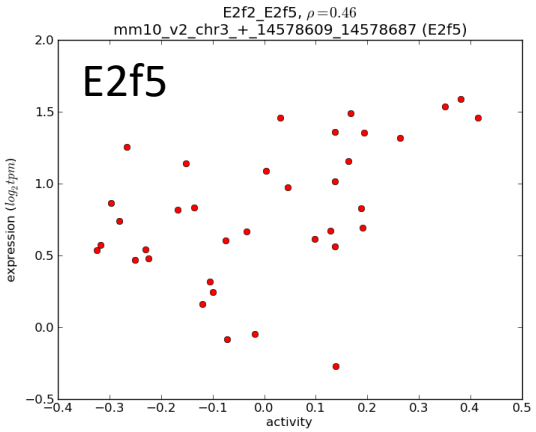
Gene	Promoter	Pearson corr. coef.	P-value	Plot
E2f2	mm10_v2_chr4_+_136172367_136172395	0.96	1.0e-20	Click!
E2f5	mm10_v2_chr3_+_14578609_14578687	0.46	4.3e-03	Click!

Activity profile of E2f2_E2f5 motif



Interpretation: both TFs bind to the same binding sites.





E2f2 is both higher expressed and correlates much better with the motif activity.

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

Example of a negatively correlated motif: Cebpe

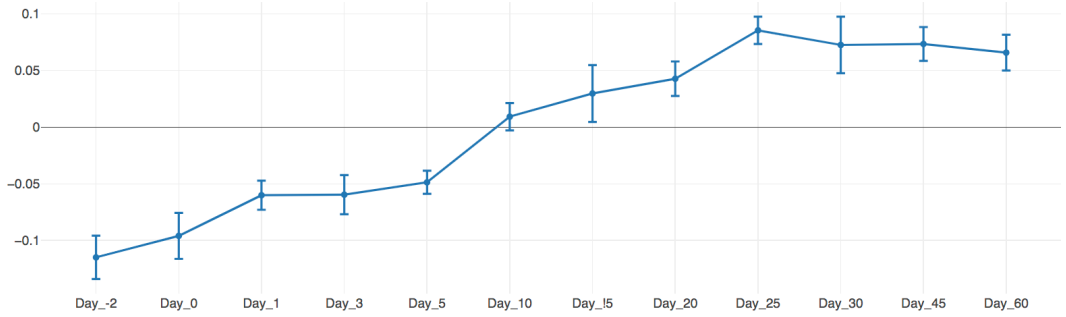
Transcription factors associated with Cebpe

Gene Symbol	Gene ID	Gene Info
Cebpe Links	ENSMUSG00000052435.6	Cebpe

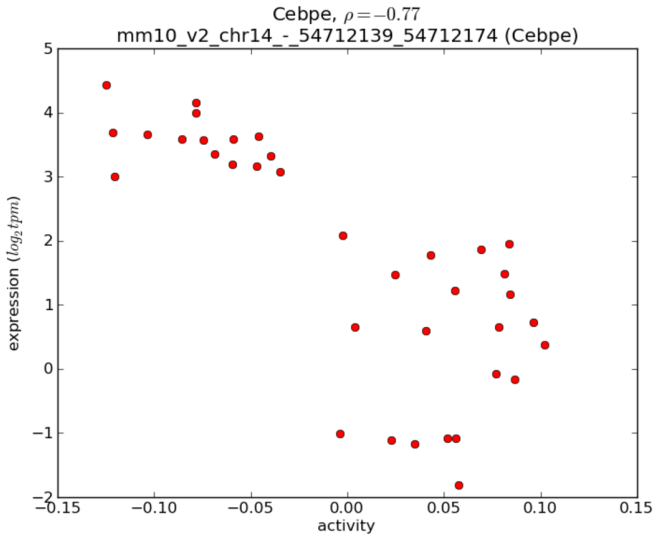
Activity-expression correlation:

Gene	Promoter	Pearson corr. coef.	P-value	Plot
Cebpe	mm10_v2_chr14_-_54712139_54712174	-0.77	5.0e-08	Click!

Activity profile of Cebpe motif



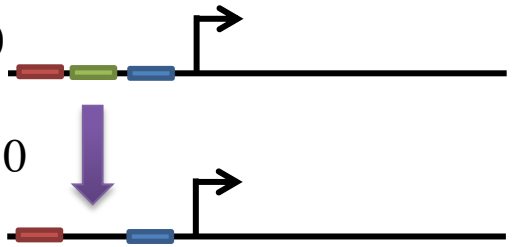
mRNA expression of the TF Cebpe is negatively correlated with its motif activity, as inferred from the expression of it targets.



Interpretation: Cebpe is acting as a *repressor* in this system.

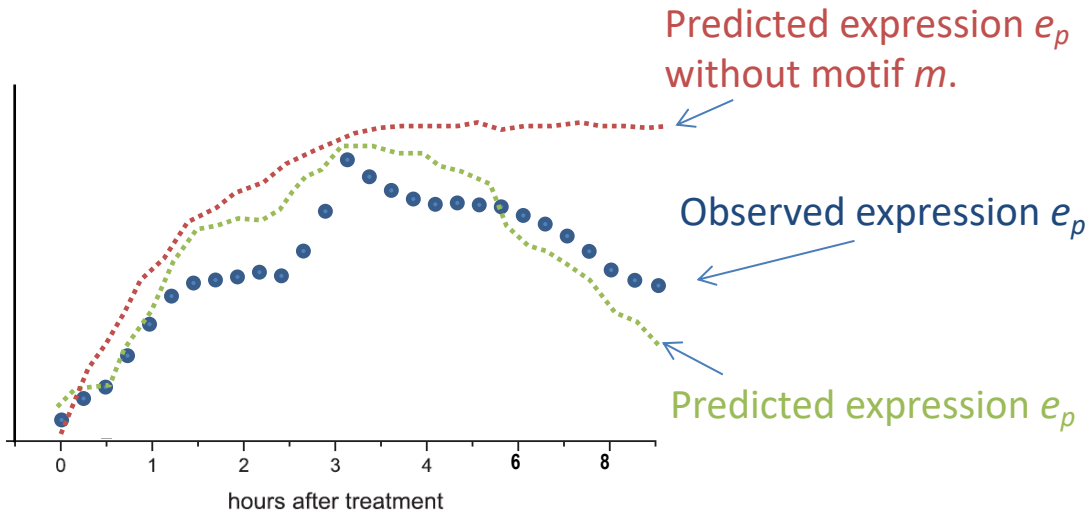
Predicting targets of each motif (conceptual)

- For each motif, select promoters with predicted sites, i.e with $N_{pm} > 0$
- *Mutate* promoter *p* to *remove* the binding site(s) for motif *m*: $N_{pm} \rightarrow 0$
- Updated site-count matrix: $N \rightarrow \tilde{N}$
- Log-likelihood ratio of fitting *all data* with N versus the mutated $\tilde{N}$:



$$S_{pm} = \log\left[\frac{\int dA P(E \mid N, A)}{\int dA P(E \mid \tilde{N}, A)}\right]$$

Quantifies the contribution of motif *m* to explaining the expression pattern of promoter *p*.



The log-likelihood ratio S_{pm} quantifies how much the quality of the fit is reduced when the sites for motif *m* in promoter *p* are removed.

Predicting targets of each motif (technical)

$$S_{pm} = \log\left[\frac{\int dAP(E \mid N, A)}{\int dAP(E \mid \tilde{N}, A)}\right]$$

The target score is the log-likelihood ratio of the fit of the model when the binding sites for motif *m* are removed from promoter p, i.e. when $N \rightarrow \tilde{N}$.

Chi-squared gives square deviation observed and predicted expression: $\chi^2_{ps} = \left(E'_{ps} - \sum_m N'_{pm} A'^{*}_{ms}\right)^2$

When sites for *m* in promoter *p* are removed, chi-squared becomes: $\chi^2_{psm} = \left(E'_{ps} - \sum_{m'} \tilde{N}'_{pm'} A'^{*}_{m's}\right)^2$

To a good approximation, the target score log-likelihood ratio is given by $S_{pm} = \frac{\sum_s \chi^2_{psm} - \chi^2_{ps}}{\langle \chi^2 \rangle}$

where $\langle \chi^2 \rangle = \frac{1}{PS} \sum_{p,s} \chi^2_{ps}$ is the average chi-squared across all promoters and samples.

Interpretation:

- The target-score measures how much the squared-deviation between fit and model increases when the sites for motif *m* in promoter *p* are removed, relative to the average squared-deviation across all promoters and samples.

Notes:

- Generally, the more samples, the higher the target scores are.
- Target scores can be negative as well (when the predictions are better without the site).

List of target promoter/genes of HNF4a

Top of the list of HNF4a target promoters, sorted by their significance:

Top targets:

Search:

Show 25 entries

Promoter	Score	Transcript	Gene	Gene Info
chr19_+_39287074	95.78	ENSMUST00000003137.8	Cyp2c29	cytochrome P450, family 2, subfamily c, polypeptide 29
chr17_-_46438471	84.62	ENSMUST00000087012.5	Slc22a7	solute carrier family 22 (organic anion transporter), member 7
chr4_-_62087261	81.03	ENSMUST00000107488.3 ENSMUST00000107472.1 ENSMUST00000084531.4	Mup3	major urinary protein 3
chr19_+_39007019	61.77	ENSMUST00000025966.4	Cyp2c55	cytochrome P450, family 2, subfamily c, polypeptide 55
chr4_-_60501903	60.52	ENSMUST00000084548.4 ENSMUST00000103012.3 ENSMUST00000107499.3	Mup1	major urinary protein 1
chr19_-_8405060	58.81	ENSMUST00000064507.5 ENSMUST00000120540.1 ENSMUST00000096269.4	Slc22a30	solute carrier family 22, member 30
chr19_-_40073731	58.08	ENSMUST00000048959.3	Cyp2c54	cytochrome P450, family 2, subfamily c, polypeptide 54
chr4_-_62054112	57.75	ENSMUST00000074018.3	Mup20	major urinary protein 20
chr19_-_8131982	56.33	ENSMUST00000065651.4	Slc22a28	solute carrier family 22, member 28
chr4_-_60741275	55.97	ENSMUST00000117932.1	Mup12	major urinary protein 12
chr19_-_39463067	55.17	ENSMUST00000035488.2	Cyp2c38	cytochrome P450, family 2, subfamily c, polypeptide 38
chr15_-_82764176	55.09	ENSMUST00000055721.4	Cyp2d40	cytochrome P450, family 2, subfamily d, polypeptide 40
chr10_-_128960965	54.11	ENSMUST00000026398.3	Mettl7b	methyltransferase like 7B

List of target promoter/genes of HNF4a

Top of the list of HNF4a target promoters, sorted by their significance:

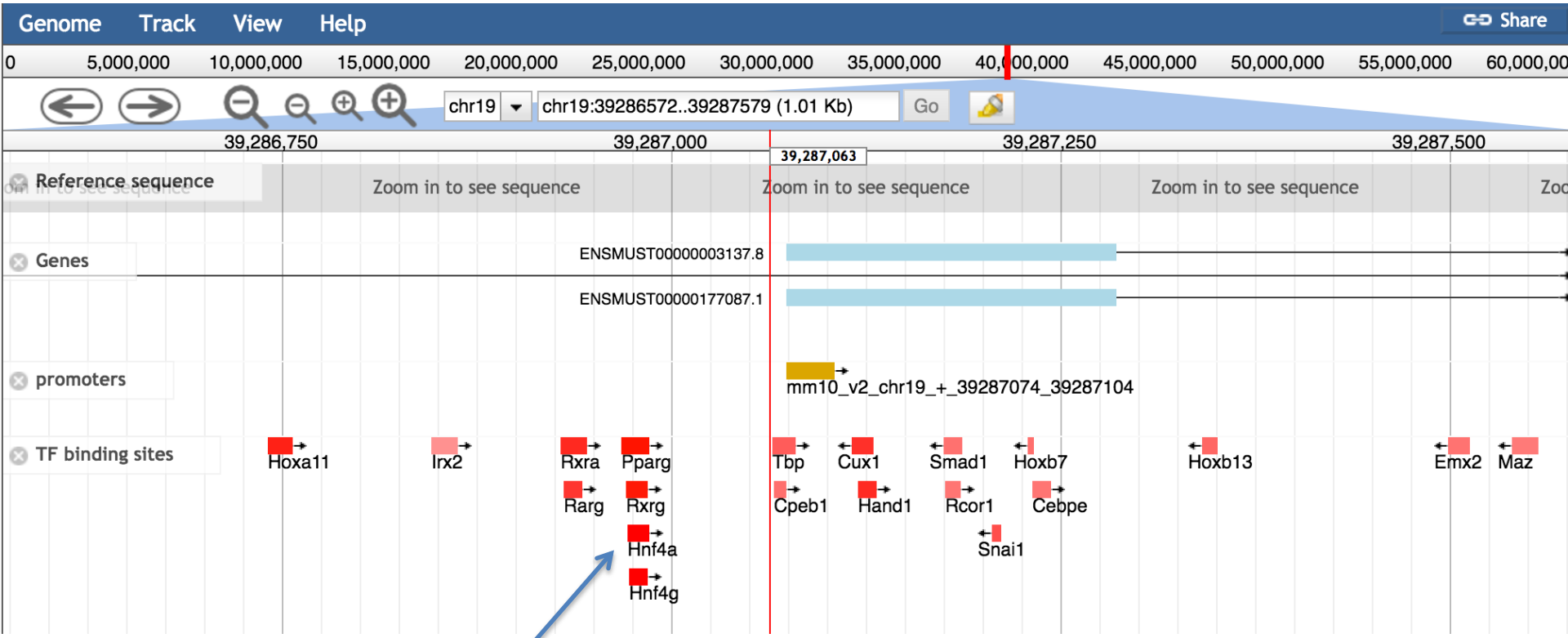
Top targets:

Search:

Show 25 entries

Promoter	Score	Transcript	Gene	Gene Info
chr19_+_39287074	95.78	ENSMUST00000003137.8	Cyp2c29	cytochrome P450, family 2, subfamily c, polypeptide 29
chr17_-_46438471	84.62	ENSMUST00000087012.5	Slc22a7	solute carrier family 22 (organic anion transporter), member 7
chr4_-_62087261	81.03	ENSMUST00000107488.3 ENSMUST00000107472.1 ENSMUST00000084531.4	Mup3	major urinary protein 3
chr19_+_39007019	61.77	ENSMUST00000025966.4	Cyp2c55	cytochrome P450, family 2, subfamily c, polypeptide 55
chr4_-_60501903	60.52	ENSMUST00000084548.4 ENSMUST00000103012.3 ENSMUST00000107499.3	Mup1	major urinary protein 1
chr19_-_8405060	58.81	ENSMUST00000064507.5 ENSMUST00000120540.1 ENSMUST00000096269.4	Slc22a30	solute carrier family 22, member 30
chr19_-_40073731	58.08	ENSMUST00000048959.3	Cyp2c54	cytochrome P450, family 2, subfamily c, polypeptide 54
chr4_-_62054112	57.75	ENSMUST00000074018.3	Mup20	major urinary protein 20
chr19_-_8131982	56.33	ENSMUST00000065651.4	Slc22a28	solute carrier family 22, member 28
chr4_-_60741275	55.97	ENSMUST00000117932.1	Mup12	major urinary protein 12
chr19_-_39463067	55.17	ENSMUST00000035488.2	Cyp2c38	cytochrome P450, family 2, subfamily c, polypeptide 38
chr15_-_82764176	55.09	ENSMUST00000055721.4	Cyp2d40	cytochrome P450, family 2, subfamily d, polypeptide 40
chr10_-_128960965	54.11	ENSMUST00000026398.3	Mettl7b	methyltransferase like 7B

SwissRegulon view of the Cyp2c29 promoter



Location of the Hnf4A binding site.

This predicts which bases in the promoter are crucial for the regulation by HNF4a.

What pathways does HNF4a target?
Enriched Gene Ontology categories

Gene overrepresentation in biological_process category:

Search:

Show 10 entries

Log-likelihood per target	Total log-likelihood	Term	Description
11.9	391.3	GO:0019373	epoxygenase P450 pathway(GO:0019373)
9.6	248.7	GO:0035634	response to stilbenoid(GO:0035634)
4.2	155.1	GO:0019369	arachidonic acid metabolic process(GO:0019369)
9.7	135.1	GO:0015747	urate transport(GO:0015747)
19.3	77.4	GO:0071718	sodium-independent icosanoid transport(GO:0071718)
9.6	57.8	GO:0008355	olfactory learning(GO:0008355)
18.6	55.9	GO:0042450	arginine biosynthetic process via ornithine(GO:0042450)
1.4	50.1	GO:0050892	intestinal absorption(GO:0050892)
12.4	49.6	GO:1903966	monounsaturated fatty acid metabolic process(GO:1903964) monounsaturated fatty acid biosynthetic process(GO:1903966)
3.1	49.5	GO:0006957	complement activation, alternative pathway(GO:0006957)

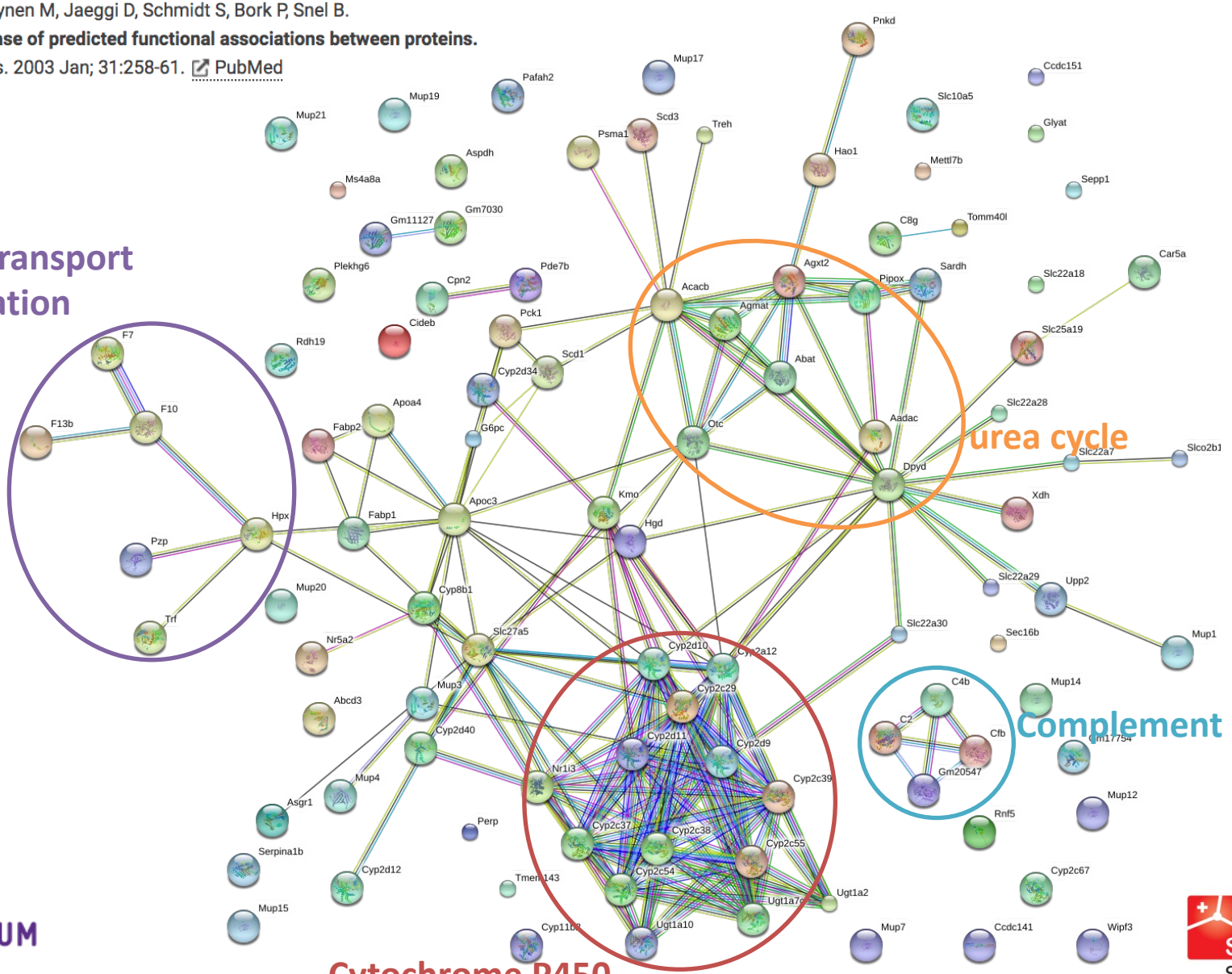
Showing 1 to 10 of 157 entries

Previous 1 2 3 4 5 … 16 Next

- For each Gene Ontology category (starting from the most specific), calculate the sum and average of target log-likelihood scores for the genes in the category.
- Sort all categories by average target score or summed log-likelihood of all genes.
- For each category, remove all genes in this category from other categories lower in the list.
- The table can be searched, expanded, and sorted in different ways.

What pathways does HNF4a target?
STRING-db picture of the network of HNF4a targets

von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B.
STRING: a database of predicted functional associations between proteins.
Nucleic Acids Res. 2003 Jan; 31:258-61. PubMed

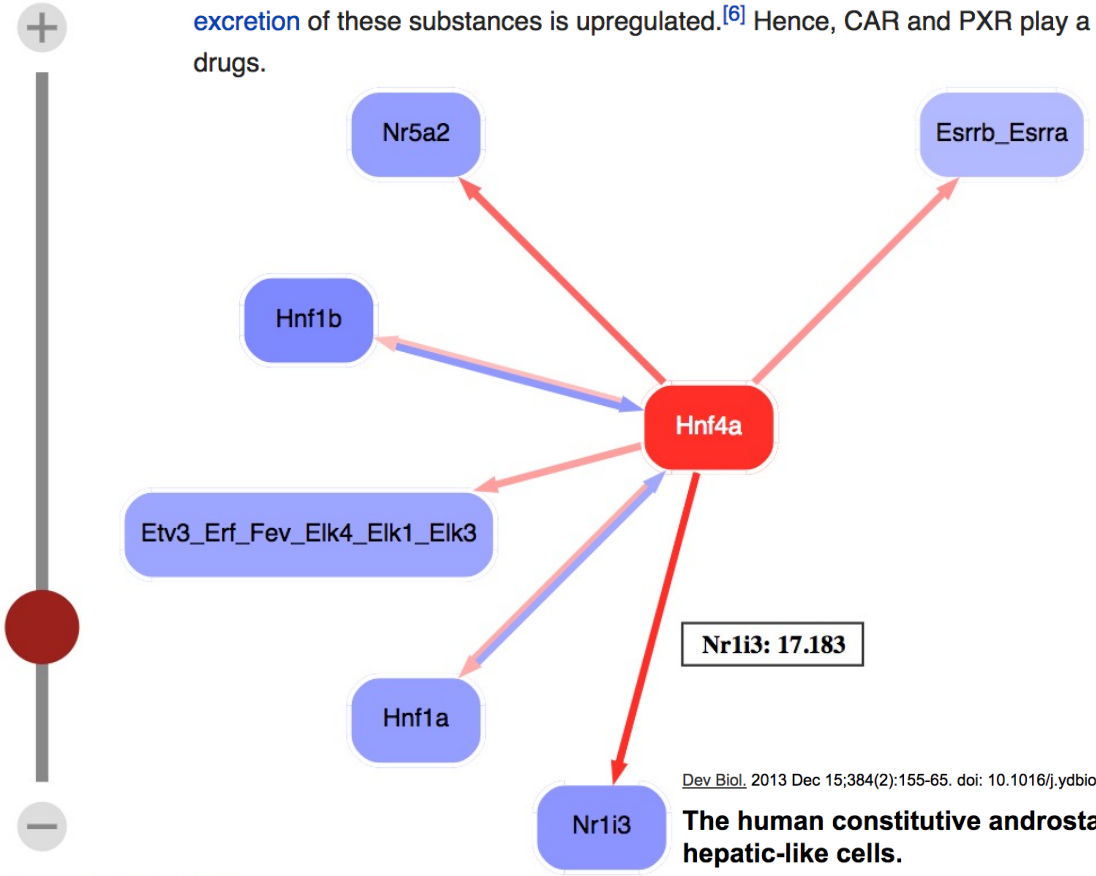


BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

Direct interactions between HNF4a and other regulators

The **constitutive androstane receptor** (**CAR**) also known as **nuclear receptor subfamily 1, group I, member 3** is a protein that in humans is encoded by the *NR1I3* gene.[5] CAR is a member of the nuclear receptor superfamily and along with pregnane X receptor (PXR) functions as a sensor of endobiotic and xenobiotic substances. In response, expression of proteins responsible for the metabolism and excretion of these substances is upregulated.[6] Hence, CAR and PXR play a major role in the detoxification of foreign substances such as drugs.



Score threshold: 4.675

Use slider or -/+ buttons to hide/show edges.

Dev Biol. 2013 Dec 15;384(2):155-65. doi: 10.1016/j.ydbio.2013.10.012. Epub 2013 Oct 18.

The human constitutive androstane receptor promotes the differentiation and maturation of hepatic-like cells.

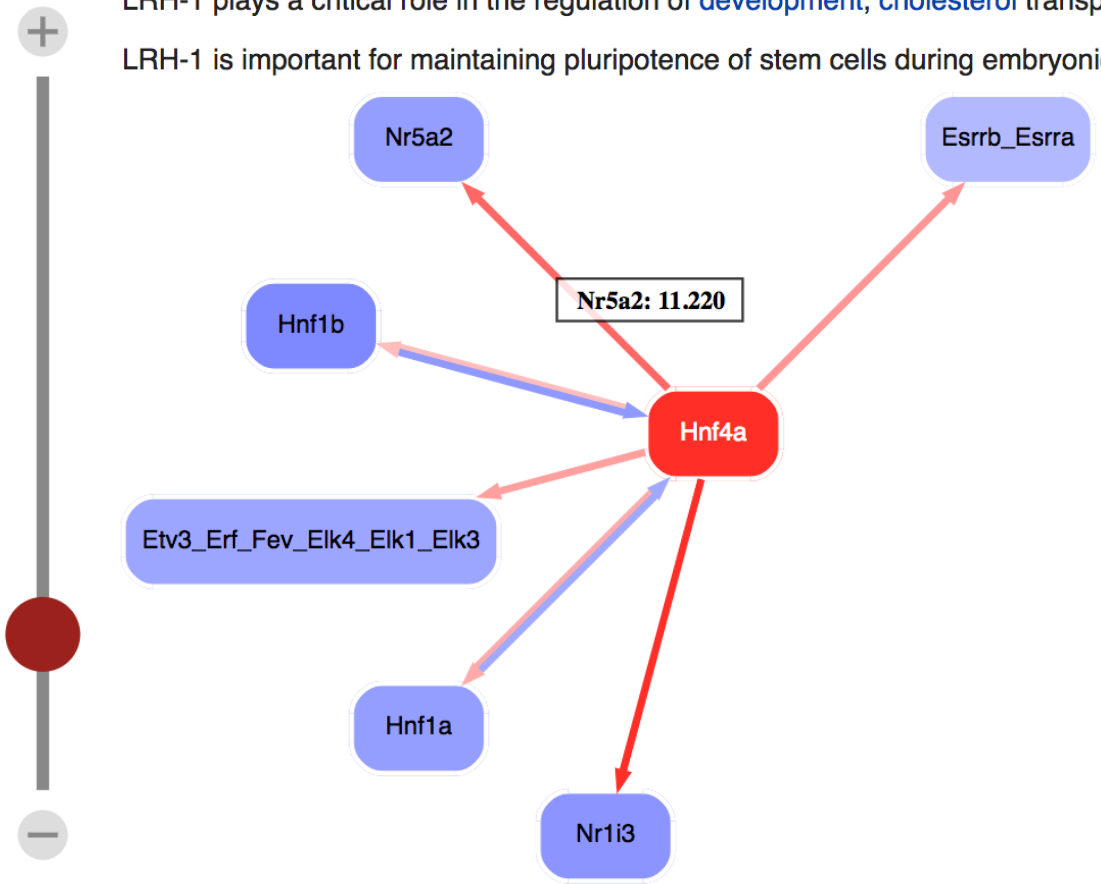
Chen F[1], Zamule SM, Coslo DM, Chen T, Omiecinski CJ.

Direct interactions between HNF4a and other regulators

The **liver receptor homolog-1** (**LRH-1**) also known as **NR5A2** (nuclear receptor subfamily 5, group A, member 2) is a protein that in humans is encoded by the *NR5A2* gene.[5][6] LRH-1 is a member of the nuclear receptor family of intracellular transcription factors.

LRH-1 plays a critical role in the regulation of development, cholesterol transport, bile acid homeostasis and steroidogenesis.[7][8][9]

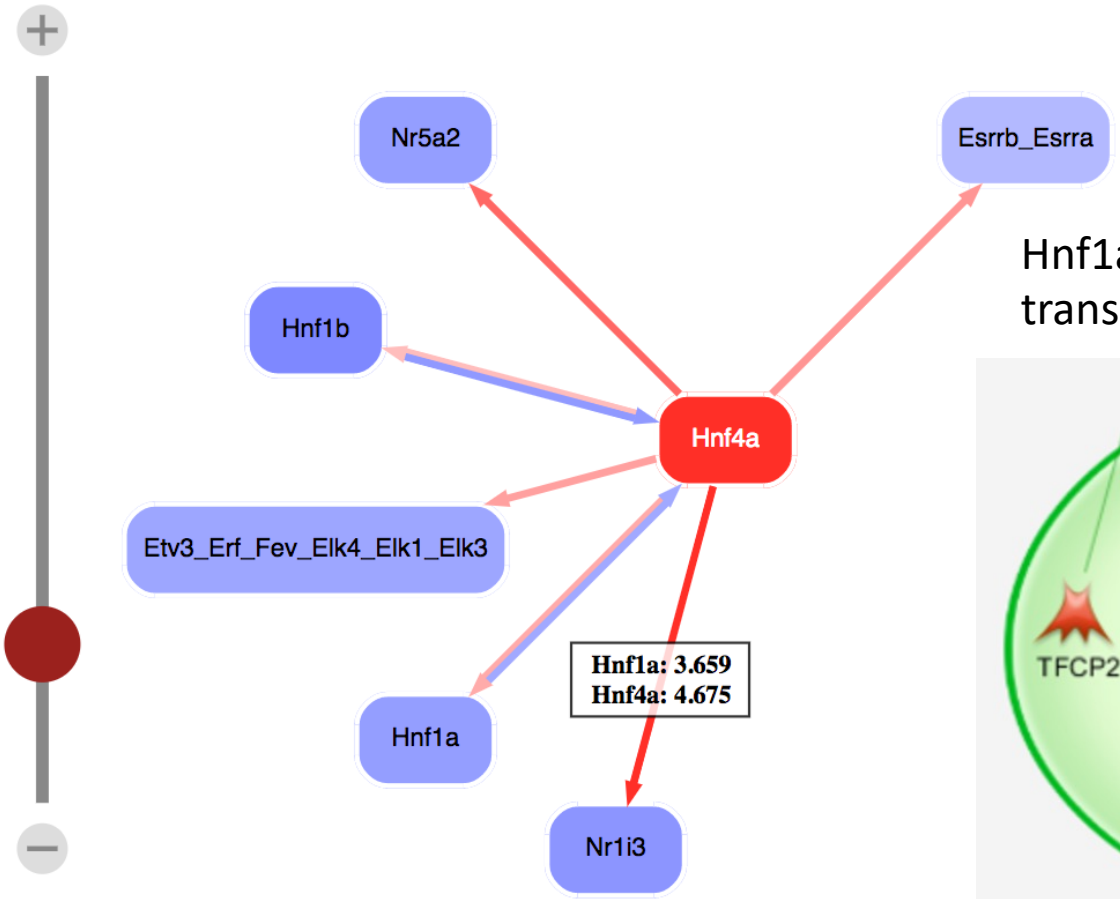
LRH-1 is important for maintaining pluripotence of stem cells during embryonic development.[10]



Score threshold: 4.675

Use slider or -/+ buttons to hide/show edges.

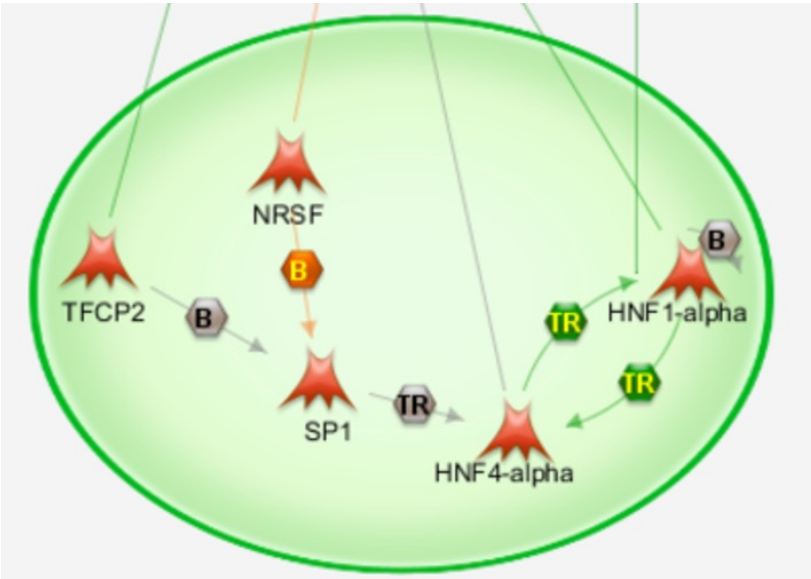
Direct interactions between HNF4a and other regulators



Score threshold: 4.675

Use slider or -/+ buttons to hide/show edges.

Hnf1a and Hnf4a are in fact known to transcriptionally regulate each other.



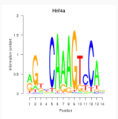
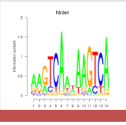
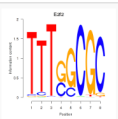
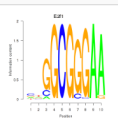
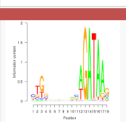
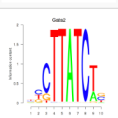
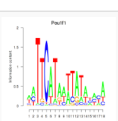
J Cell Sci. 1998 Aug;111 (Pt 16):2411-21.

Phenotypic effects of the forced expression of HNF4 and HNF1alpha are conditioned by properties of the recipient cell.

Bailly A[1], Späth G, Bender V, Weiss MC.

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

Most significant motifs

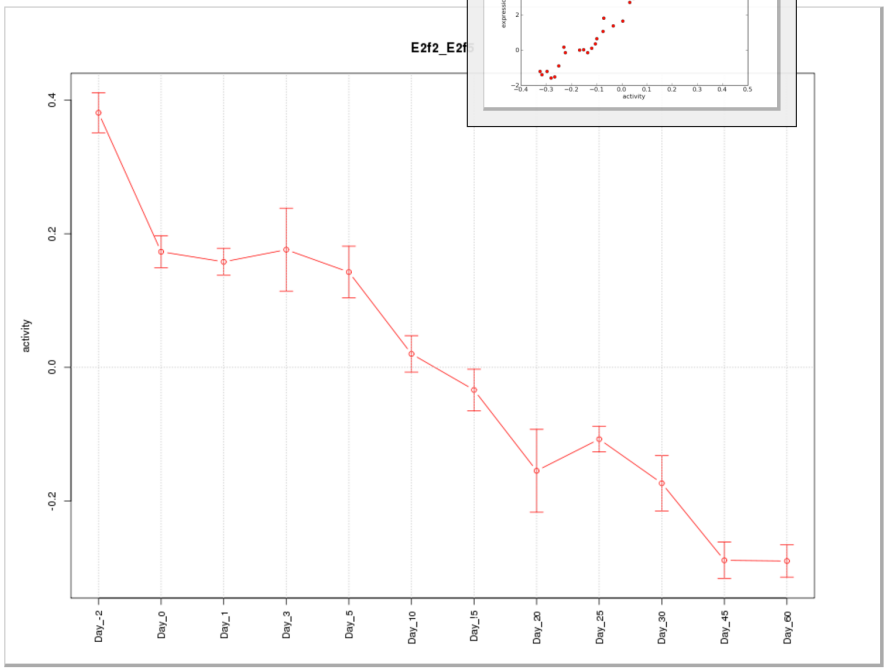
Motif name	Z-value	Associated genes	Profile	Logo
Hnf4a	7.61	Hnf4a Links		
Nr2e1	7.31	Nr2e1 Links		
E2f2_E2f5	7.12	E2f2 Links E2f5 Links		
E2f1	6.52	E2f1 Links		
Tal1	6.09	Tal1 Links		
Gata2_Gata1	5.58	Gata2 Links Gata1 Links		
Pou1f1	5.28	Pou1f1 Links		

What is the role of the E2F motifs?

E2f2 and E2f1 targets are down-regulated over time

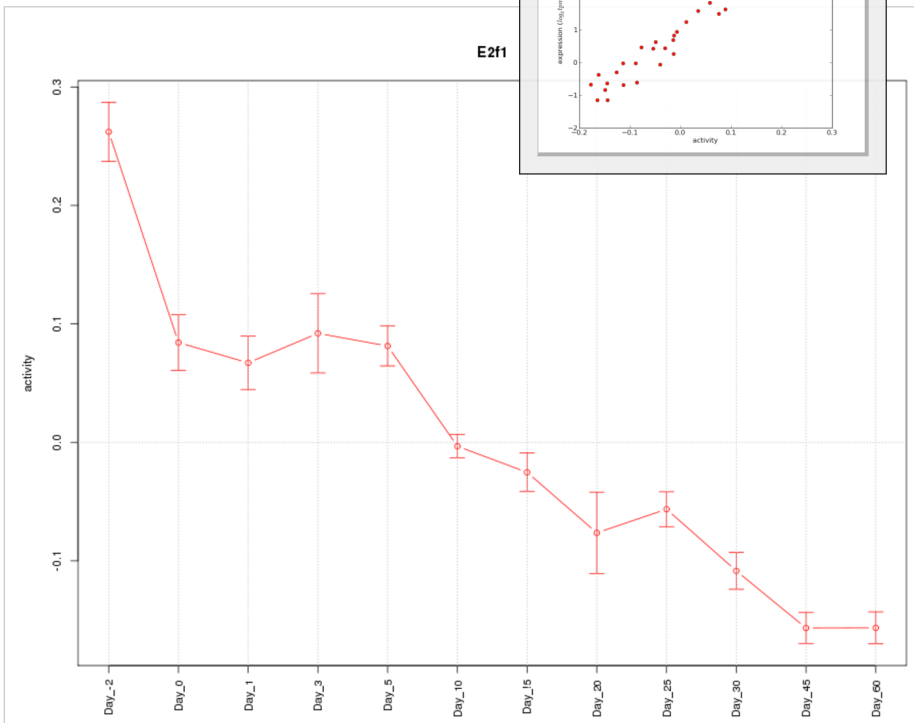
Gene Symbol	Promoter	Pearson corr. coef.	P-value	Plot
E2f2	mm10_v2_chr4_+_136172367_136172395	0.96	1.0e-20	Click!
E2f5	mm10_v2_chr3_+_14578609_1457868	0.46	4.3e-03	Click!

Activity profile for motif E2f2_E2f5.



Gene Symbol	Promoter	Pearson corr. coef.	P-value	Plot
E2f1	mm10_v2_chr2_-_154569845_154569892	0.95	2.2e-19	Click!

Activity profile for motif E2f1.

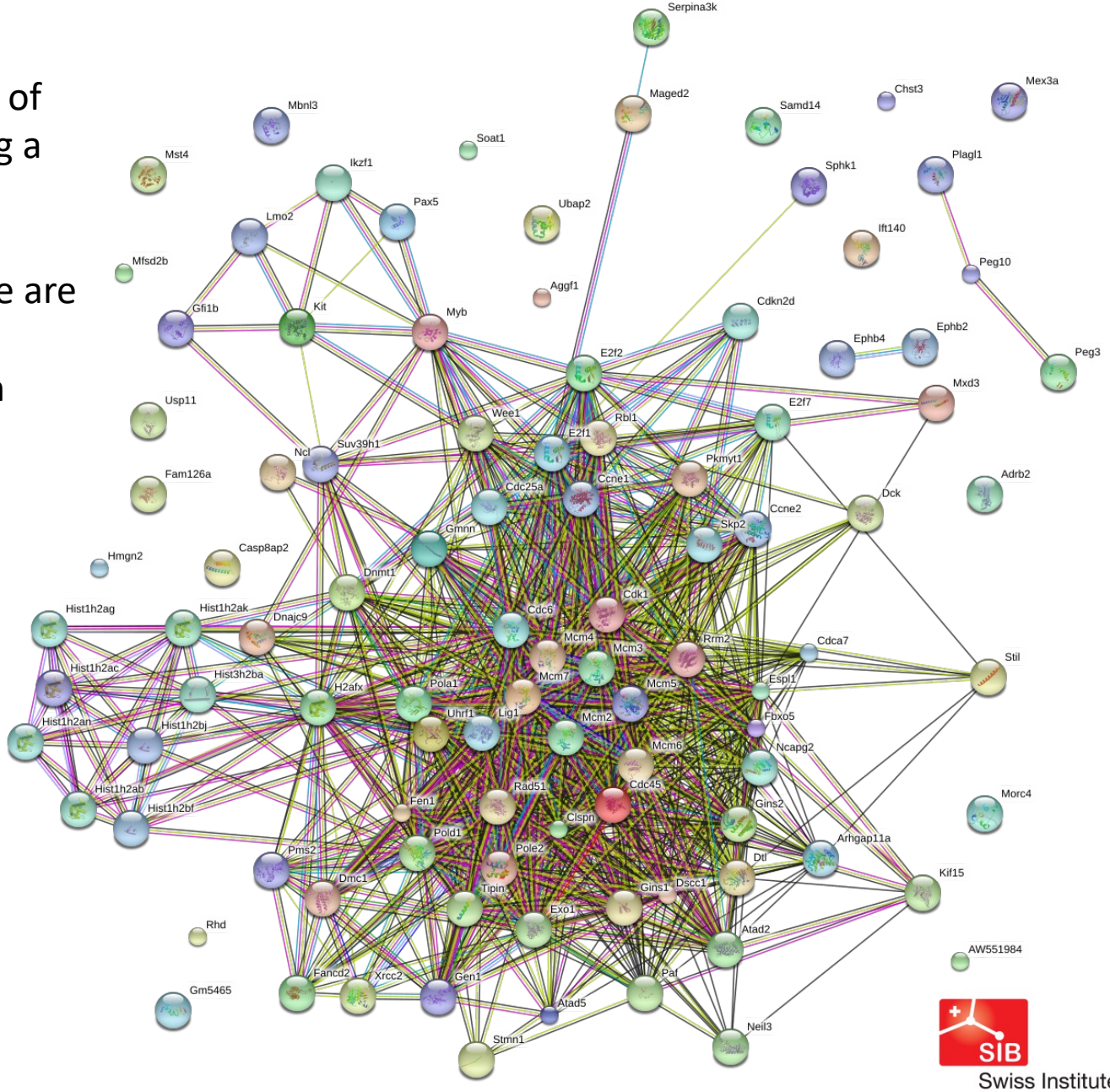


Note that the motif activities and expression of both factors are very similar.
Both are down-regulated across the time course.
This suggests we are looking at a single `pathway'.

Network of E2f2 target genes

The extremely high density of links shows E2f2 is targeting a very well-studied pathway.

Inspection shows that these are all cell cycle genes, and in particular genes involved in initiation of replication.



Pathways most targeted by E2f1 and E2f2
G1 – S transition of the cell cycle

This picture is confirmed when one looks at the top Gene Ontology categories and pathways among the E2f2/E2f1 targets:

Gene overrepresentation in biological_process category:

Log-likelihood per target	Total log-likelihood	Term	Description
2.8	95.9	GO:0006270	DNA replication initiation(GO:0006270)
9.2	92.2	GO:0006268	DNA unwinding involved in DNA replication(GO:0006268)

Gene overrepresentation in cellular_component category:

Log-likelihood per target	Total log-likelihood	Term	Description
3.0	158.1	GO:0042555	MCM complex(GO:0042555) MCM core complex(GO:0097373)
0.1	116.9	GO:0005694	chromosome(GO:0005694)
3.6	114.3	GO:0005657	replication fork(GO:0005657)
0.4	69.5	GO:0000775	chromosome, centromeric region(GO:0000775)

Gene overrepresentation in curated gene sets: REACTOME pathways category:

Log-likelihood per target	Total log-likelihood	Term
8.9	142.6	REACTOME_UNWINDING_OF_DNA
5.7	101.7	REACTOME_G1_S_SPECIFIC_TRANSCRIPTION

E2f1/E2f2 are regulating initiation of DNA replication, i.e. transition from G1 to S. The fact that their activity decreases with time likely indicates that the amount of cell division is steadily decreasing during liver maturation.

How is a given gene of interest regulated?

Project

avrg: GSE58827: Dynamics of the Mouse Liver

Navigation

Motif significance table

Sample table

Mean activities

All promoters sorted by FOV

Search gene

Downloads

Activity table
Activity delta table
Regulatory interactions
Motifs sorted by significance
Download the whole report

ISMARA results avrg: GSE58827: Dynamics of the Mouse Liver

ISMARA - Integrated System for Motif Actitivity Response Analysis is a free online tool that recognizes most important transcription factors that are changing their activity in a set of samples.

Original results

Show averaging configuration

All motifs sorted by activity significance

Search: Show 10 entries

Motif name	Z-value	Associated genes	Profile	Logo
Hnf4a	7.61	Hnf4a Links		
Nr2e1	7.31	Nr2e1 Links		
E2f2_E2f5	7.12	E2f2 Links E2f5 Links		
E2f1	6.52	E2f1 Links		

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

Sortable table of genes with expression statistics

Average log-tpm expression

Standard-deviation of log-tpm expression

Fraction of expression variance explained by MARA.

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

Show 100 entries

Search:

Promoter	Mean Expression	Std. deviation	FOV	Genes
mm10_v2_chr9_+_44334685_44334715	4.850	8.192	0.907	H2afx (H2A histone family, member X)
mm10_v2_chr6_-_47594967_47595047	1.422	7.068	0.882	Ezh2 (enhancer of zeste homolog 2 (Drosophila))
mm10_v2_chr10_+_110745433_110745572	-0.281	6.650	0.878	E2f7 (E2F transcription factor 7)
mm10_v2_chr15_+_55557399_55557436	-0.601	5.071	0.870	Mtbp (Mdm2, transformed 3T3 cell double minute p53 binding protein)
mm10_v2_chr6_-_8259098_8259173	2.621	3.798	0.863	Rpa3 (replication protein A3)
mm10_v2_chr2_+_85037448_85037530	4.696	3.178	0.856	Ssrp1 (structure specific recognition protein 1)
mm10_v2_chr12_+_55836365_55836408	0.034	3.343	0.854	Brms1l (breast cancer metastasis-suppressor 1-like)
mm10_v2_chr2_-_157204483_157204542	-0.216	7.126	0.848	Rbl1 (retinoblastoma-like 1 (p107))
mm10_v2_chr7_+_44816088_44816088	3.355	4.165	0.847	Nup62 (nucleoporin 62)
mm10_v2_chr9_-_70934808_70934847	6.527	4.890	0.843	Lipc (lipase, hepatic)
mm10_v2_chr10_-_13552838_13552870	0.808	3.197	0.838	Pex3 (peroxisomal biogenesis factor 3)
mm10_v2_chr10_+_127063599_127063655	5.237	4.263	0.837	Cdk4 (cyclin-dependent kinase 4)
mm10_v2_chr11_-_5099223_5099266	2.418	3.212	0.835	Ewsr1 (Ewing sarcoma breakpoint region 1)
mm10_v2_chr3_+_88621436_88621555	-0.542	5.529	0.835	Arhgef2 (rho/rac guanine nucleotide exchange factor (GEF) 2)
mm10_v2_chr11_+_88047693_88047727	2.027	4.360	0.833	Srsf1 (serine/arginine-rich splicing factor 1)
mm10_v2_chr19_+_46075842_46075863	2.458	4.348	0.830	Nolc1 (nucleolar and coiled-body phosphoprotein 1)
mm10_v2_chr5_-_115119277_115119346	6.414	4.051	0.829	Acads (acyl-Coenzyme A dehydrogenase, short chain)

Observed and predicted expression of H2afx

Promoter: mm10_v2_chr9_+_44334685_44334715

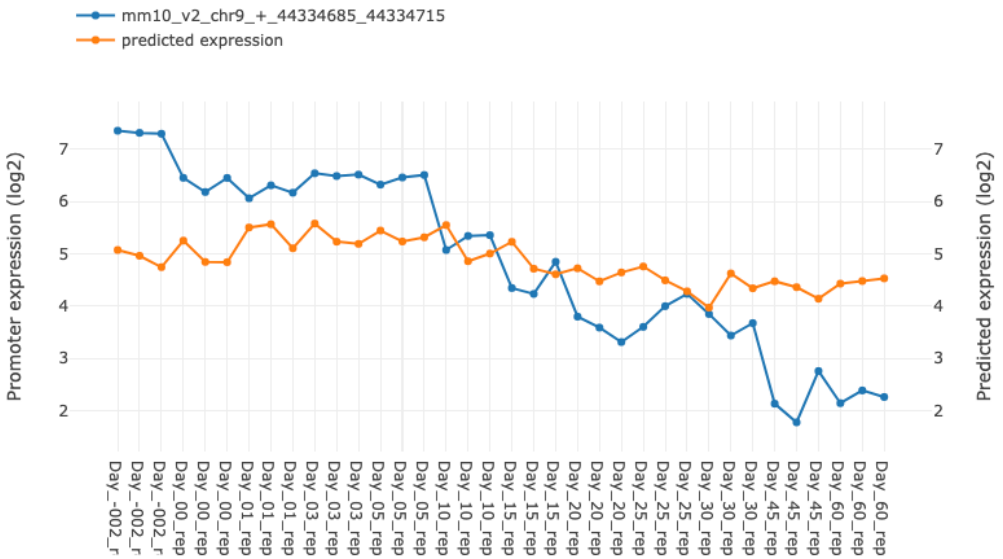
Fraction of explained variance: 0.907

SwissRegulon link: mm10_v2_chr9_+_44334685_44334715

Associated genes:

- **H2afx:** H2A histone family, member X Links
  Associated transcripts: ENSMUST00000052686.2

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



This plot shows expression and predicted expression of mm10_v2_chr9_+_44334685_44334715 promoter. Left vertical axis is a promoter expression on the log2 scale. Right vertical axis is a predicted promoter expression on the log2 scale. Horisontal axis indicates samples.

All motifs turned off.

Search: Show 10 entries

Motif	ChiSq	SiteCount	Z-val
E2f2_E2f5	16.34	2.05	5.21
Ybx1_Nfya_Nfyb_Nfyc_Cebpz	6.80	6.69	4.22
E2f4	4.52	1.00	2.93
E2f1	3.75	1.21	5.25
Tbp	2.76	0.85	3.29
Tfdp1_Wt1_Egr2	1.98	3.27	3.03
Wrnip1_Mta3_Rcor1	1.94	5.27	3.29
E2f7	1.41	0.60	1.40
Klf4_Sp3	1.34	1.81	2.56
Klf1	0.93	0.86	2.56

Showing 1 to 10 of 45 entries

Previous 1 2 3 4 5 Next

All On All Off

Observed and predicted expression of H2afx

Promoter: mm10_v2_chr9_+_44334685_44334715

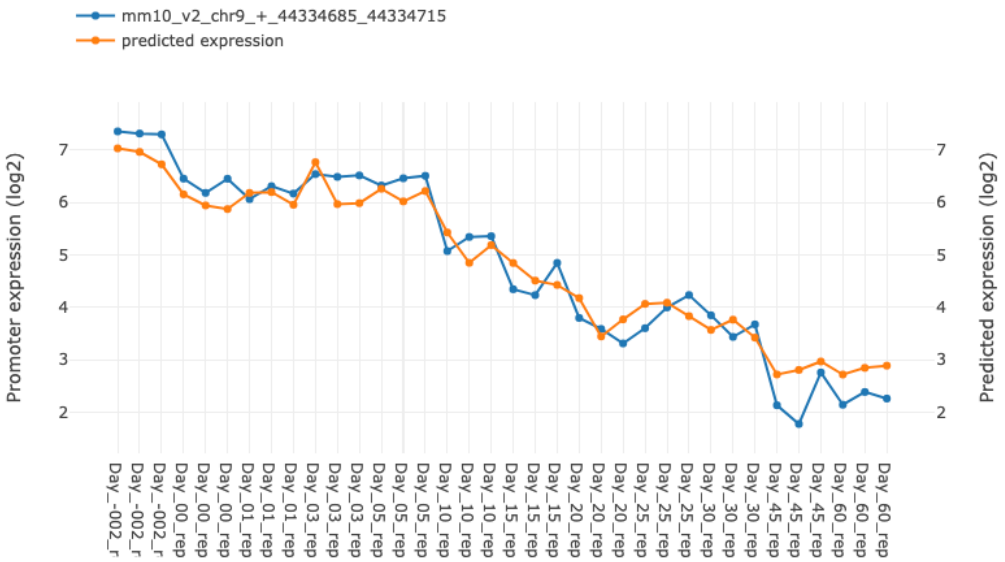
Fraction of explained variance: 0.907

SwissRegulon link: mm10_v2_chr9_+_44334685_44334715

Associated genes:

- **H2afx:** H2A histone family, member X Links
  Associated transcripts: ENSMUST00000052686.2

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



This plot shows expression and predicted expression of mm10_v2_chr9_+_44334685_44334715 promoter. Left vertical axis is a promoter expression on the log2 scale. Right vertical axis is a predicted promoter expression on the log2 scale. Horisontal axis indicates samples.

All motifs turned on.

Search: Show 10 entries

Motif	ChiSq	SiteCount	Z-val
E2f2_E2f5	16.34	2.05	5.21
Ybx1_Nfya_Nfyb_Nfyc_Cebpz	6.80	6.69	4.22
E2f4	4.52	1.00	2.93
E2f1	3.75	1.21	5.25
Tbp	2.76	0.85	3.29
Tfdp1_Wt1_Egr2	1.98	3.27	3.03
Wrnip1_Mta3_Rcor1	1.94	5.27	3.29
E2f7	1.41	0.60	1.40
Klf4_Sp3	1.34	1.81	2.56
Klf1	0.93	0.86	2.56

Showing 1 to 10 of 45 entries

Previous 1 2 3 4 5 Next

All On All Off

Observed and predicted expression of H2afx

Promoter: mm10_v2_chr9_+_44334685_44334715

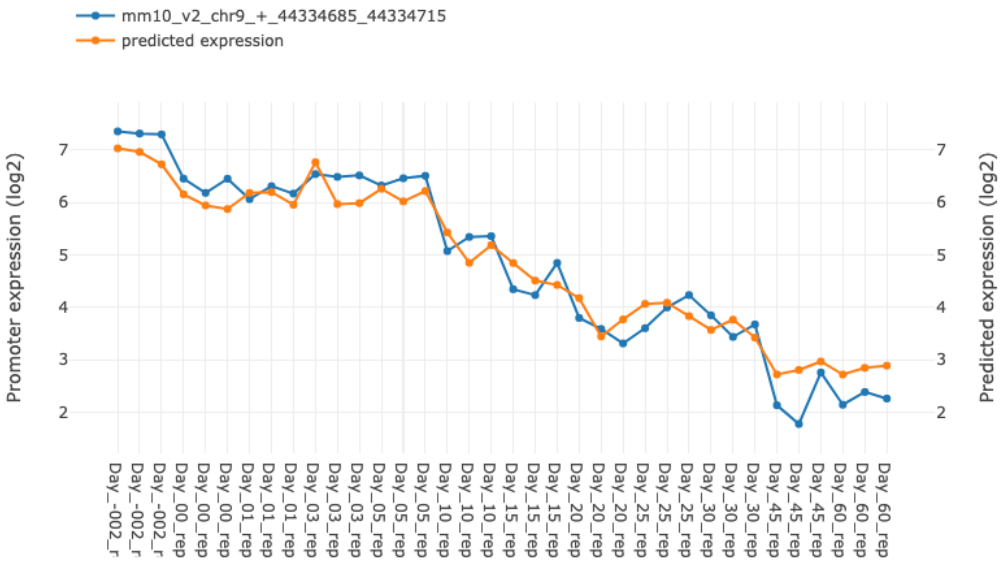
Fraction of explained variance: 0.907

SwissRegulon link: mm10_v2_chr9_+_44334685_44334715

Associated genes:

- **H2afx:** H2A histone family, member X Links
  Associated transcripts: ENSMUST00000052686.2

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



This plot shows expression and predicted expression of mm10_v2_chr9_+_44334685_44334715 promoter. Left vertical axis is a promoter expression on the log2 scale. Right vertical axis is a predicted promoter expression on the log2 scale. Horisontal axis indicates samples.

Z-score of the motif activity.

Target score for the motif.

Total sites for the motif.

Search: Show 10 entries

Motif	ChiSq	SiteCount	Z-val
E2f2_E2f5	16.34	2.05	5.21
Ybx1_Nfya_Nfyb_Nfyc_Cebpz	6.80	6.69	4.22
E2f4	4.52	1.00	2.93
E2f1	3.75	1.21	5.25
Tbp	2.76	0.85	3.29
Tfdp1_Wt1_Egr2	1.98	3.27	3.03
Wrnip1_Mta3_Rcor1	1.94	5.27	3.29
E2f7	1.41	0.60	1.40
Klf4_Sp3	1.34	1.81	2.56
Klf1	0.93	0.86	2.56

Showing 1 to 10 of 45 entries

Previous 1 2 3 4 5 Next

All On All Off

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

Downloadable results for downstream analysis

Project

avrg: GSE58827: Dynamics of the Mouse Liver

Navigation

Motif significance table

Sample table

Mean activities

All promoters sorted by FOV

Search gene

Downloads

Activity table
Activity delta table
Regulatory interactions
Motifs sorted by significance
Expression table
Download the whole report

ISMARA results avrg: GSE58827: Dynamics of the Mouse Liver

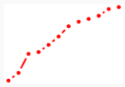
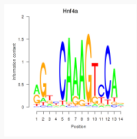
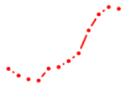
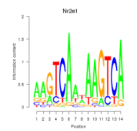
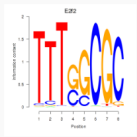
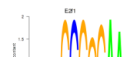
ISMARA - Integrated System for Motif Actitivity Response Analysis is a free online tool that recognizes most important transcription factors that are changing their activity in a set of samples.

Original results

Show averaging configuration

All motifs sorted by activity significance

Search: Show 10 entries

Motif name	Z-value	Associated genes	Profile	Logo
Hnf4a	7.61	Hnf4a Links		
Nr2e1	7.31	Nr2e1 Links		
E2f2_E2f5	7.12	E2f2 Links E2f5 Links		
E2f1	6.52	E2f1 Links		

These downloadable result files will be discussed in the afternoon session.

ISMARA Acknowledgments

People that helped develop the MARA tool

Piotr Balwierz
ISMARA development

Mikhail Pachkov
Web interface and support

Phil Arnold
MotEvo and epi-MARA

Jeremie Breda
single-cell MARA

Đorđe Relic
zebrafish ISMARA

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics