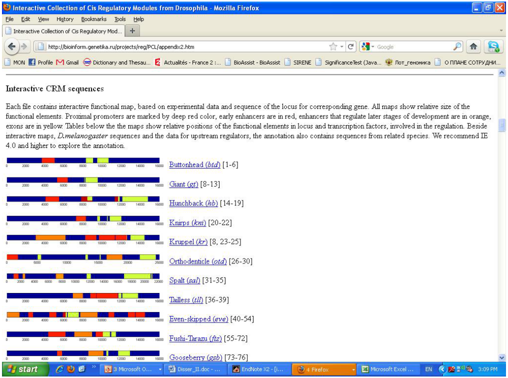
From Transfac to HOCOMOCO: using cross-validation and human curation to take most from the high throughput data compiling a complete collection of transcription factor binding motifs

Vsevolod J. Makeev

Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow

March 8, 2018

D.melanogaster enhancers


Interactive Collection of Cis Regulatory Modules from Drosophila - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Interactive Collection of Cis Regulatory Mod...
http://bioinform.genetika.ru/projects/reg/PCL/appendix2.htm
Google
MON Profile Gmail Dictionary and Thesau... Actualités - France 2 :... BioAssist - BioAssist SIRENE SignificanceTest (Java... Лот_реномика О ПЛАНЕ СОТРУДНИ...
start
3 Microsoft O...
Disser_II.doc - ...
EndNote X2 - [i...
4 Firefox
Microsoft Excel ...
EN
3:09 PM

Interactive CRM sequences

Each file contains interactive functional map, based on experimental data and sequence of the locus for corresponding gene. All maps show relative size of the functional elements. Proximal promoters are marked by deep red color, early enhancers are in red, enhancers that regulate later stages of development are in orange, exons are in yellow. Tables below the the maps show relative positions of the functional elements in locus and transcription factors, involved in the regulation. Beside interactive maps, *D.melanogaster* sequences and the data for upstream regulators, the annotation also contains sequences from related species. We recommend IE 4.0 and higher to explore the annotation.

Buttonhead (*btd*) [1-6]

Giant (*gt*) [8-13]

Hunchback (*hb*) [14-19]

Knirps (*kni*) [20-22]

Kruppel (*kr*) [8, 23-25]

Orthodenticle (*otd*) [26-30]

Spalt (*sal*) [31-35]

Tailless (*tll*) [36-39]

Even-skipped (*eve*) [40-54]

Fushi-Tarazu (*ftz*) [55-72]

Gooseberry (*gsb*) [73-76]

- We started to work with regulatory genomics in 1998
- Dima Papatsenko studied *Drosophila* enhancers
- he was interested in TF binding sites

Our first collection of TFBS

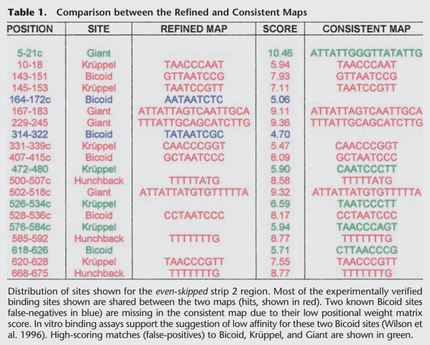


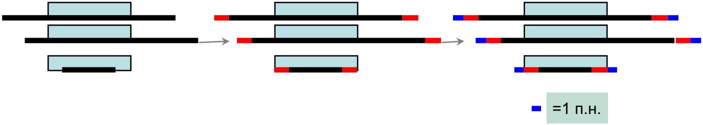
Table 1. Comparison between the Refined and Consistent Maps

POSITION	SITE	REFINED MAP	SCORE	CONSISTENT MAP
5-21c	Giant		10.46	ATTATTGGGTTATATTG
10-18	Krüppel	TAACCCAAT	5.94	TAACCCAAT
143-151	Bicoid	GTTAATCCG	7.93	GTTAATCCG
145-153	Krüppel	TAATCCGTT	7.11	TAATCCGTT
164-172c	Bicoid	AATAATCTC	5.06	
167-183	Giant	ATTATTAGTCAATTGCA	9.11	ATTATTAGTCAATTGCA
229-245	Giant	TTTATTGCAGCATCTTG	9.36	TTTATTGCAGCATCTTG
314-322	Bicoid	TATAATCGC	4.70	
331-339c	Krüppel	CAACCCGGT	5.47	CAACCCGGT
407-415c	Bicoid	GCTAATCCC	8.09	GCTAATCCC
472-480	Krüppel		5.90	CAATCCCTT
500-507c	Hunchback	TTTTTATG	8.58	TTTTTATG
502-518c	Giant	ATTATTATGTGTTTTTA	9.32	ATTATTATGTGTTTTTA
526-534c	Krüppel		6.59	TAATCCCTT
528-536c	Bicoid	CCTAATCCC	8.17	CCTAATCCC
576-584c	Krüppel		5.94	TAACCCAGT
585-592	Hunchback	TTTTTTTG	8.77	TTTTTTTG
618-626	Bicoid		5.71	CTTAACCCG
620-628	Krüppel	TAACCCGTT	7.55	TAACCCGTT
668-675	Hunchback	TTTTTTTG	8.77	TTTTTTTG

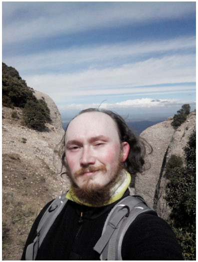
Distribution of sites shown for the *even-skipped* strip 2 region. Most of the experimentally verified binding sites shown are shared between the two maps (hits, shown in red). Two known Bicoid sites false-negatives in blue) are missing in the consistent map due to their low positional weight matrix score. In vitro binding assays support the suggestion of low affinity for these two Bicoid sites (Wilson et al. 1996). High-scoring matches (false-positives) to Bicoid, Krüppel, and Giant are shown in green.

- A site verified by at least two methods from footprints, mutant, or highly conserved blocks
- Bicoid (34 sites), Caudal (15), Ftz (25), Hunchback (43), Knirps (47), Kruppel (21), and Tramtrak (7)
- Aligned with CLUSTALW and manually and cut the flanks

Aligning footprints with genome mapping

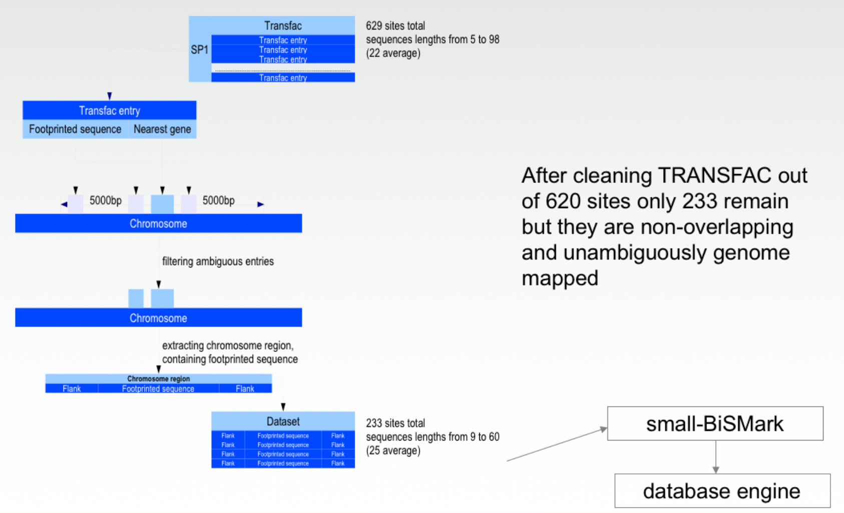


- 2008
- Mapping footprints on the genome allows recovering up to 40
- Usually it is enough to add only two letters
- Genome data may be very useful for interpretation *in vitro* results
- http://autosome.ru/dmmpmm/ DMMPMM collection



Ivan Kulakovskiy

TRANSFAC appears!

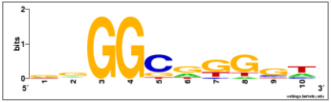


629 sites total
sequences lengths from 5 to 98
(22 average)

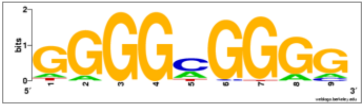
233 sites total
sequences lengths from 9 to 60
(25 average)

After cleaning TRANSFAC out of 620 sites only 233 remain but they are non-overlapping and unambiguously genome mapped

Nice Sp1 model for studying CpG islands



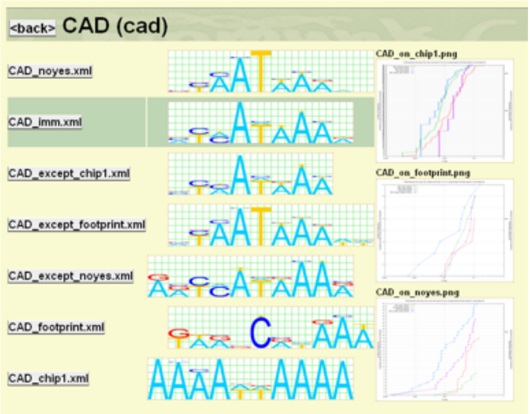
Sp1 JASPAR 2007
(SELEX data)



Sp1 Remapped and realigned
TRANSFAC 2008

Chip-on-chip data

- Chip-on-chip yielded long regions (up to 20K)
- Wasn't suitable for motif discovery
- But perhaps could be helped with *in vitro* data



<back> CAD (cad)

CAD_noyes.xml

CAD_imm.xml

CAD_except_chip1.xml

CAD_except_footprint.xml

CAD_except_noyes.xml

CAD_footprint.xml

CAD_chip1.xml

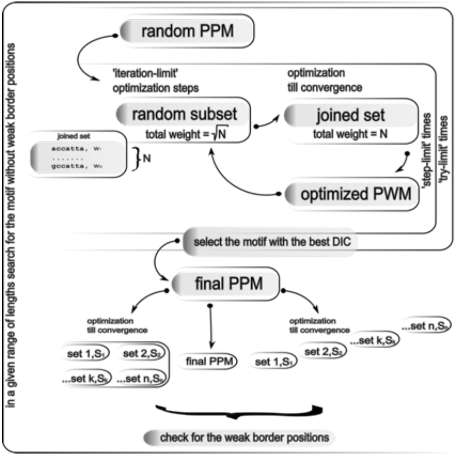
CAD_on_chip1.png

CAD_on_footprint.png

CAD_on_noyes.png

Integrative motif discovery: early ChIPmunk

Subsampling on many sets of sequences then optimization on total set of weighted sequencies



in a given range of lengths search for the motif without weak border positions

random PPM

'iteration-limit' optimization steps

optimization till convergence

random subset

total weight = $\sqrt{N}$

joined set

total weight = N

joined set

accatta, w_1

....

gccatta, w_N

}N

'step-limit' times

'try-limit' times

optimized PWM

select the motif with the best DIC

final PPM

optimization till convergence

set 1,S_1

set 2,S_2

...set k,S_k

...set n,S_n

final PPM

optimization till convergence

set 1,S_1

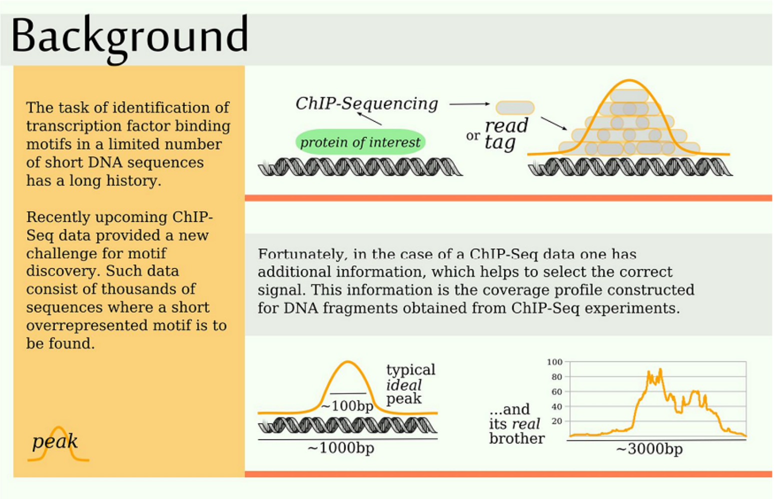
set 2,S_2

...set k,S_k

...set n,S_n

check for the weak border positions

Chip-seq data



Background

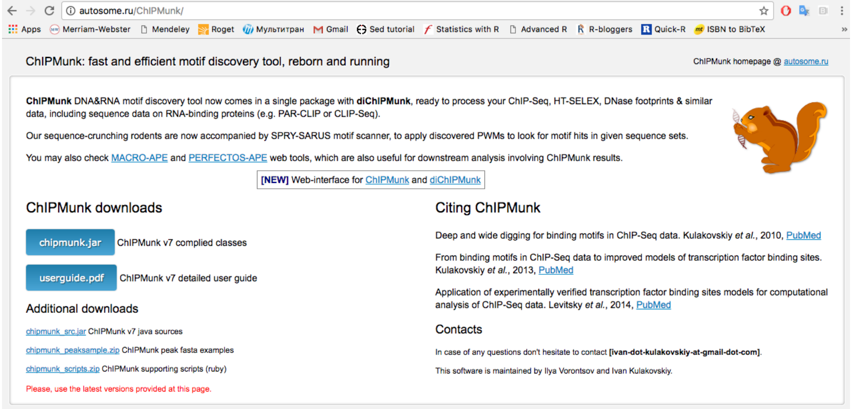
The task of identification of transcription factor binding motifs in a limited number of short DNA sequences has a long history.

Recently upcoming ChIP-Seq data provided a new challenge for motif discovery. Such data consist of thousands of sequences where a short overrepresented motif is to be found.

Fortunately, in the case of a ChIP-Seq data one has additional information, which helps to select the correct signal. This information is the coverage profile constructed for DNA fragments obtained from ChIP-Seq experiments.

ChIPmunk page

Peak shape and motif shape prior (like double box)
available at http://autosome.ru/ChIPMunk/



ChIPMunk: fast and efficient motif discovery tool, reborn and running

ChIPMunk homepage @ autosome.ru

ChIPMunk DNA&RNA motif discovery tool now comes in a single package with **diChIPMunk**, ready to process your ChIP-Seq, HT-SELEX, DNase footprints & similar data, including sequence data on RNA-binding proteins (e.g. PAR-CLIP or CLIP-Seq).

Our sequence-crunching rodents are now accompanied by SPRY-SARUS motif scanner, to apply discovered PWMs to look for motif hits in given sequence sets.

You may also check MACRO-APE and PERFECTOS-APE web tools, which are also useful for downstream analysis involving ChIPMunk results.

[NEW] Web-interface for ChIPMunk and diChIPMunk

ChIPMunk downloads

chipmunk.jar ChIPMunk v7 complied classes

userguide.pdf ChIPMunk v7 detailed user guide

Additional downloads

chipmunk_src.jar ChIPMunk v7 java sources

chipmunk_peaksample.zip ChIPMunk peak fasta examples

chipmunk_scripts.zip ChIPMunk supporting scripts (ruby)

Please, use the latest versions provided at this page.

Citing ChIPMunk

Deep and wide digging for binding motifs in ChIP-Seq data. Kulakovskiy *et al.*, 2010, PubMed

From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. Kulakovskiy *et al.*, 2013, PubMed

Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. Levitsky *et al.*, 2014, PubMed

Contacts

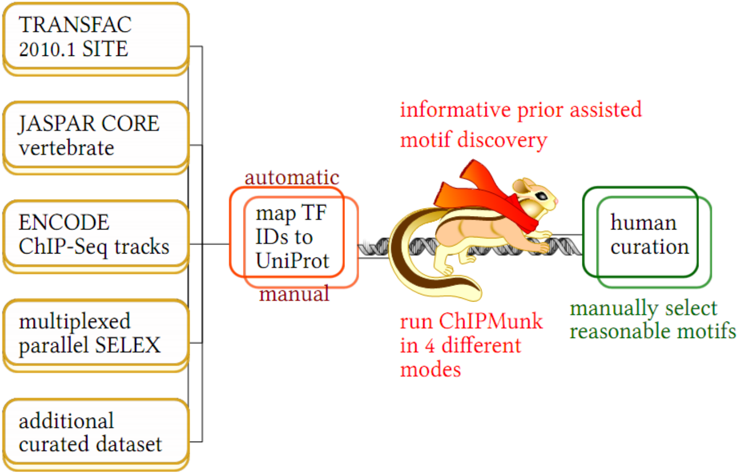
In case of any questions don't hesitate to contact **[ivan-dot-kulakovskiy-at-gmail-dot-com]**.

This software is maintained by Ilya Vorontsov and Ivan Kulakovskiy.

TRANSFAC comes into view again

... and supplies us with a new version of SITE database (for free)

Core workflow (2011), with Vlad Bajic from KAUST



TRANSFAC 2010.1 SITE

JASPAR CORE vertebrate

ENCODE ChIP-Seq tracks

multiplexed parallel SELEX

additional curated dataset

automatic

map TF IDs to UniProt

manual

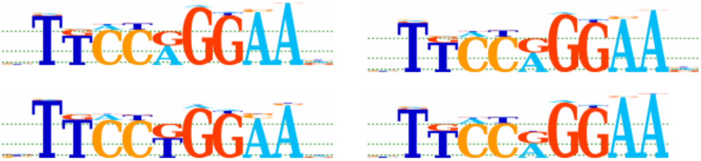
informative prior assisted motif discovery

run ChIPMunk in 4 different modes

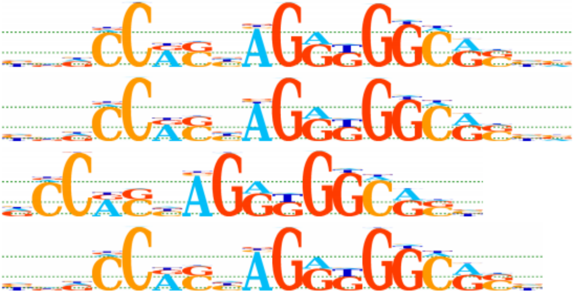
human curation

manually select reasonable motifs

Discovery strategies usually agree!



STAT3

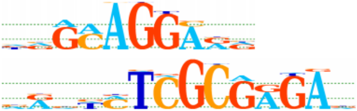


CTCF

Human curation

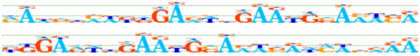
From a set of (**f1**,**f2**,**si**,**do**) motifs we **manually** select reasonable ones according to the following criteria:

- select similar motifs for the TFs from a particular family;
- select motifs having higher weight / number of aligned sequences;
- for huge sequence sets: trust flanking regions;
- for small sequence sets: take motif cores;
- take >1 motifs for one TF when the motifs have completely different consensi;
- use information from other sources (compare to known existing motifs).

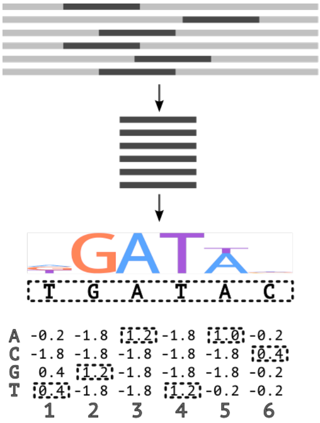


KAISO - both motifs are significant
(known to have two distinct binding motifs)

XRCC4 - no significant motif
(long and unstructured)



Some notes on PWMs



A	-0.2	-1.8	1.2	-1.8	1.0	-0.2
C	-1.8	-1.8	-1.8	-1.8	-1.8	0.4
G	0.4	1.2	-1.8	-1.8	-1.8	-0.2
T	0.4	-1.8	-1.8	1.2	-0.2	-0.2
	1	2	3	4	5	6

- PWM can be used to calculate a score for any sequence
- $Score[j] = \sum_{j}^{j+L-1} PWM[j, s(j)]$
- $s(j)$ is the letter in the position j of the alignment of PWM with the sequence
- L is the PWM length

PWM and the scoring threshold as a binary classifier

Each pair **(PWM , threshold)** classifies any word as a motif hit (YES/NO)

	1	2	3	4	5	6
A	-1.6	-1.6	0.96	-1.6	-1.6	0.96
C	-1.6	-1.6	0.00	-1.6	-1.6	-1.6
G	1.22	1.22	-1.6	-1.6	-1.6	-1.6
T	-1.6	-1.6	-1.6	1.22	1.22	0.00

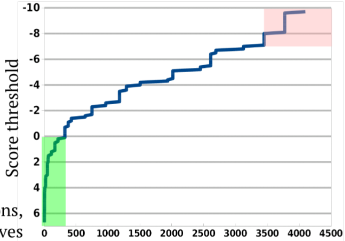
PWM

GGATTA → $S_{\text{GGATTA}}=1.22+1.22+0.96+1.22+1.22+0.96=\mathbf{6.8}$ the best score

$S_{\text{GGGGGG}}=2.44\text{-}6.4=\mathbf{-3.96}$

$S=\mathbf{-9.6}$ the worst score

more predicted TFBS,
more false positive predictions

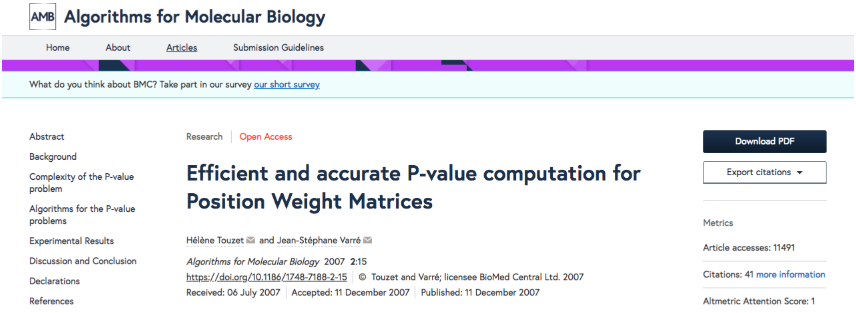


less TFBS predictions,
less true positives

Number of words passing the threshold
(i.e. scoring not less than the threshold)

Fast exact calculation of motif P-vlaue

- Suppose there is a probability distribution upon the *l*-words
- Motif *P*-value is the sum of probabilities of all words scoring above the threshold
- In 2007 Hélèn Touzet and Jean Stéfan Varré designed nice precise algorithm



AMB
Algorithms for Molecular Biology
Home
About
Articles
Submission Guidelines

What do you think about BMC? Take part in our survey our short survey

Abstract
Background
Complexity of the P-value problem
Algorithms for the P-value problems
Experimental Results
Discussion and Conclusion
Declarations
References

Efficient and accurate P-value computation for Position Weight Matrices

Hélène Touzet and Jean-Stéphane Varré

Algorithms for Molecular Biology 2007 2:15

https://doi.org/10.1186/1748-7188-2-15 | © Touzet and Varré; licensee BioMed Central Ltd. 2007

Received: 06 July 2007 | Accepted: 11 December 2007 | Published: 11 December 2007

Metrics

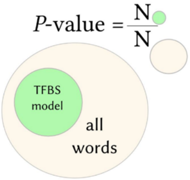
Article accesses: 11491

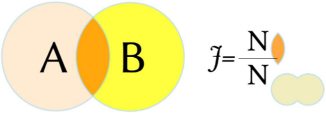
Citations: 41 more information

Altmetric Attention Score: 1

Motfs can be compared as clussifiers i.e. pairs (**PWM, threshold**)

- One needs to set both thresholds
- ... but after that it is possible to calculate the percentage of common words recognized by both motfs
- and compare it with a larger set of words recognized by any of them
- Matrices of different origine (or even PWM and PCM) can be compared without additional normalization



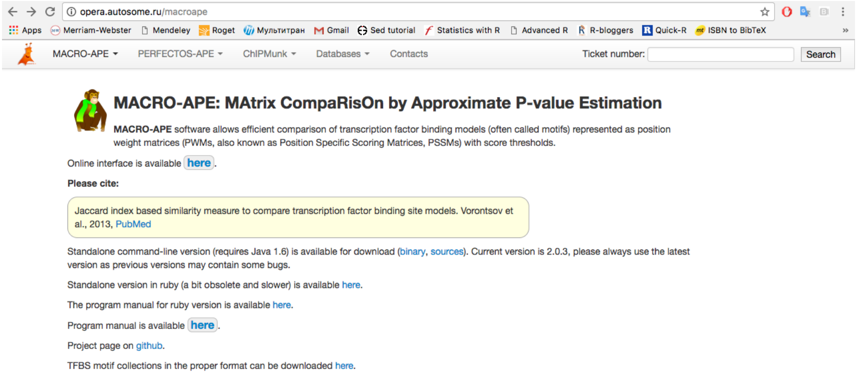


...or in case of probabilities:

$$\mathcal{J}1(\Omega_1,\Omega_2)=\frac{P(\{\omega\}:\omega\in\Omega_1\cap\Omega_2)}{P(\{\omega\}:\omega\in\Omega_1\cup\Omega_2)}$$

MacroApe to compare motifs

We modified Touzet - Varré algorithm to compare PMWs
Available at http://opera.autosome.ru/macroape
Can be used to extract motifs from various motif databases



MACRO-APE: MAtrix CompaRisOn by Approximate P-value Estimation

MACRO-APE software allows efficient comparison of transcription factor binding models (often called motifs) represented as position weight matrices (PWMs, also known as Position Specific Scoring Matrices, PSSMs) with score thresholds.

Online interface is available **here**.

Please cite:

Jaccard index based similarity measure to compare transcription factor binding site models. Vorontsov et al., 2013, PubMed

Standalone command-line version (requires Java 1.6) is available for download (binary, sources). Current version is 2.0.3, please always use the latest version as previous versions may contain some bugs.

Standalone version in ruby (a bit obsolete and slower) is available here.

The program manual for ruby version is available here.

Program manual is available **here**.

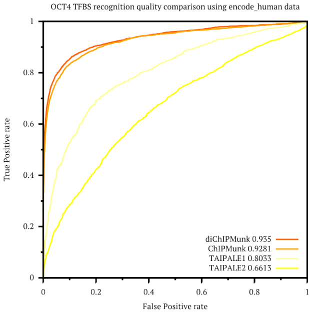
Project page on github.

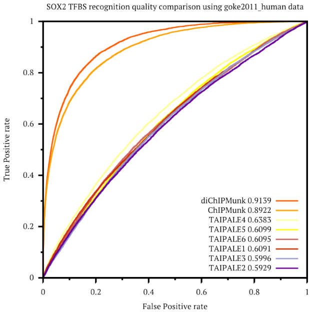
TFBS motif collections in the proper format can be downloaded here.

Measuring performance with AU ROC

We can use theoretically calculated P-values for a false-positive rate

This allows us to compare performance of different motifs on the same benchmark datasets

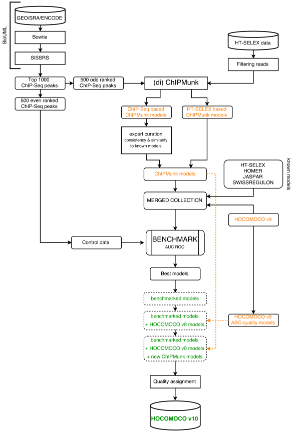




Hocomoco database log

- 2011 first website published
- 2012, first publication, v.9, *Nucleic acids research, database 2013*
- 2015, second publication, v.10, *Nucleic acids research, database, 2016*
- 2017, third publication, v.11, *Nucleic acids research, database 2018*
- http://hocomoco11.autosome.ru/
- http://www.cbrc.kaust.edu.sa/hocomoco11

Extension from HT-SELEX data (v.10)


BioUML
GEO/SRA/ENCODE
Bowtie
SISSRS
HT-SELEX data
Filtering reads
Top 1000 ChIP-Seq peaks
500 odd ranked ChIP-Seq peaks
(di) ChIPMunk
500 even ranked ChIP-Seq peaks
ChIP-Seq based ChIPMunk models
HT-SELEX based ChIPMunk models
expert curation
consistency & similarity to known models
ChIPMunk models
HT-SELEX
HOMER
JASPAR
SWISSREGULON
known models
MERGED COLLECTION
HOCOMOCO v9
Control data
BENCHMARK
AUC ROC
Best models
benchmarked models
benchmarked models + HOCOMOCO v9 models
HOCOMOCO v9 ABC-quality models
benchmarked models + HOCOMOCO v9 models + new ChIPMunk models
Quality assignment
HOCOMOCO v10

- large number of HT-SELEX data and new ChIP-seq data allowed us to extend the core base only by benchmarking and curation

Curation of extantion v.10

- similar to known models (0.05 Jaccard similarity)
- consistent within a TF family, TFclass families are taken
- or at least with a clearly exhibited consensus (based on LOGO representation, manually assessed).

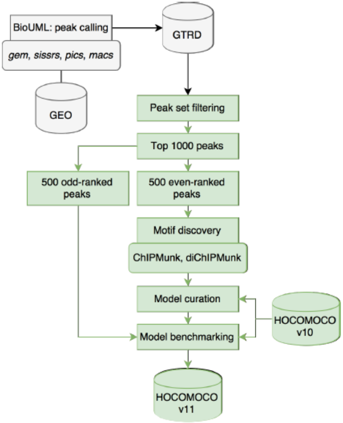
Extension from GTRD ChIP-seq database

Gather as many datasets as possible

Motif discovery in all datasets

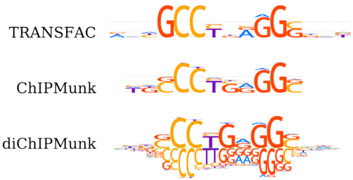
Benchmarking and conservative filtering

Machine dataset filtering v.11



- Cross-validation based dataset filtering
- If known motif performs better than the genuine dataset motif the entire dataset is discarded

Dinucleotide models

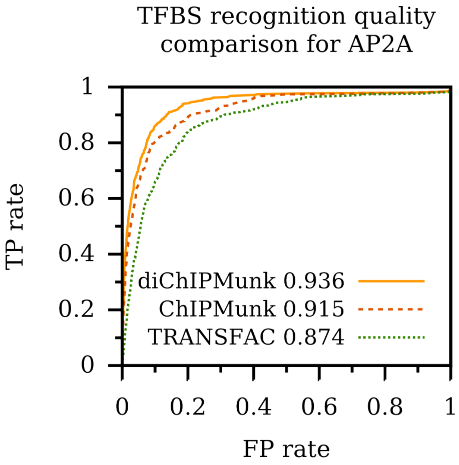


GATA:

G'A'T'A or GA'AT'TA

mononucleotide alphabet {A,C,G,T}

dinucleotide superalphabet {AA,AC,...TT}


TFBS recognition quality comparison for AP2A

Many motifs are very similar

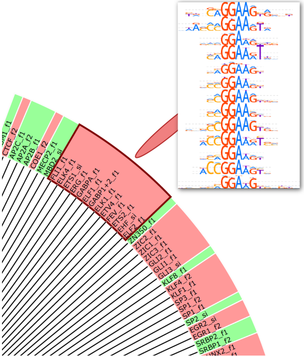
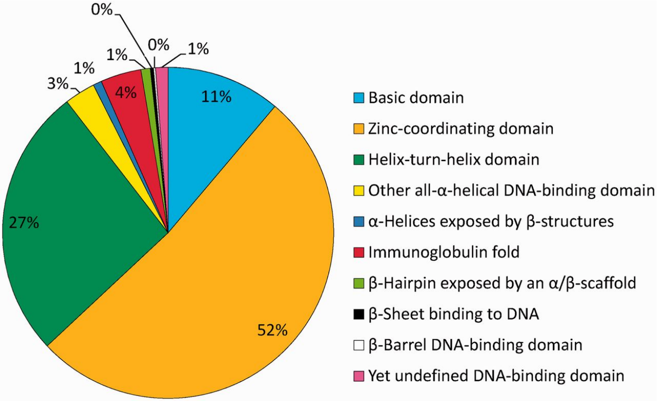


Figure: ETC family

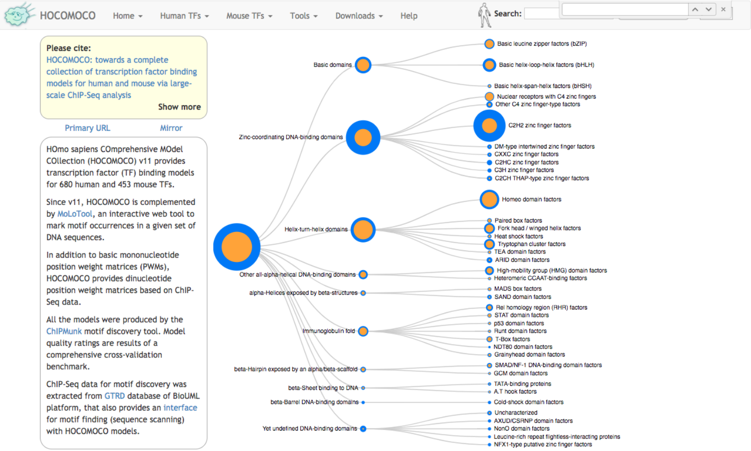
Difficulties for MARA style analysis. SwissRegulon contains small number of "isolated" motifs

Motif classes correspond to structural classes of TFs



Adapted from TFclass database, Wingender et al., 2015

http://www.cbrc.kaust.edu.sa/hocomoco11
http://hocomoco.autosome.ru



Please cite:
HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis
Show more

Primary URL   Mirror

HOmo sapiens COmprehensive MOdel COllection (HOCOMOCO) v11 provides transcription factor (TF) binding models for 680 human and 453 mouse TFs.

Since v11, HOCOMOCO is complemented by MoLoTool, an interactive web tool to mark motif occurrences in a given set of DNA sequences.

In addition to basic mononucleotide position weight matrices (PWMs), HOCOMOCO provides dinucleotide position weight matrices based on ChIP-Seq data.

All the models were produced by the ChIPMunk motif discovery tool. Model quality ratings are results of a comprehensive cross-validation benchmark.

ChIP-Seq data for motif discovery was extracted from GTRD database of BioUML platform, that also provides an interface for motif finding (sequence scanning) with HOCOMOCO models.

Basic domains
Zinc-coordinating DNA-binding domains
Helix-turn-helix domains
Other all-alpha-helical DNA-binding domains
alpha-Helices exposed by beta-structures
Immunoglobulin fold
beta-Hairpin exposed by an alpha/beta-scaffold
beta-Sheet binding to DNA
beta-Barrel DNA-binding domains
Yet undefined DNA-binding domains
Basic leucine zipper factors (bZIP)
Basic helix-loop-helix factors (bHLH)
Basic helix-span-helix factors (bHSH)
Nuclear receptors with C4 zinc fingers
Other C4 zinc finger-type factors
C2H2 zinc finger factors
DM-type intertwined zinc finger factors
CXXC zinc finger factors
C2HC zinc finger factors
C3H zinc finger factors
C2CH THAP-type zinc finger factors
Homeo domain factors
Paired box factors
Fork head / winged helix factors
Heat shock factors
Tryptophan cluster factors
TEA domain factors
ARID domain factors
High-mobility group (HMG) domain factors
Heteromeric CCAAT-binding factors
MADS box factors
SAND domain factors
Rel homology region (RHR) factors
STAT domain factors
p53 domain factors
Runt domain factors
T-Box factors
NDT80 domain factors
Grainyhead domain factors
SMAD/NF-1 DNA-binding domain factors
GCM domain factors
TATA-binding proteins
A.T hook factors
Cold-shock domain factors
Uncharacterized
AXUD/CSRNP domain factors
NonO domain factors
Leucine-rich repeat flightless-interacting proteins
NFX1-type putative zinc finger factors

v.11 Hocomoco statistics

- models for 453 mouse and 680 human transcription factors
- contains 1302 mononucleotide and 576 dinucleotide PWMs
- build from more than 3000 ChIP-seq tracks and four peak callers

What one needs motifs for ?

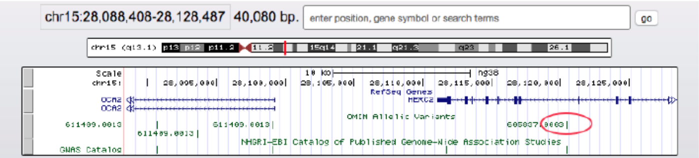
A:A brown eye colour, 80%

A:G brown eye colour

G:G blue eye colour, 99%

Found in the intron of HERC2, the non-pigment gene

21kb upstream of OCA2, the non-pigment gene



chr15:28,088,408-28,128,487 40,080 bp.

enter position, gene symbol or search terms

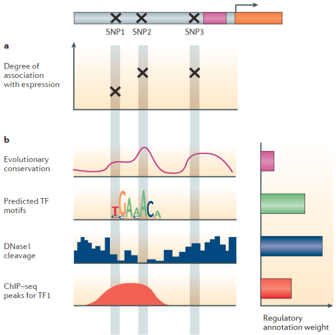
Mike Visser et al. Genome Research, 2012; 22:446-455

No experimental location of TFBS

method	*in vitro*	native or	segment	# segments	comment
	in vivo	synthetic	length		
ChIP	*in vivo*	native	40 (exo) 5000	150 - 50000	indirect binding
One-hybrid	*in vivo*	synthetic	∼30	20-50	in bacteria
SELEX, RSS	*in vitro*	synthetic	∼20	20-50	saturation
HT-SELEX	*in vitro*	synthetic	∼50	5000	saturation
PBA	*in vitro*	synthetic	∼50	10000	overlapping
Footprints	either	native	∼100	20 - 10000	indirect

Table: Experimental methods of TF binding identification

Limitations for using motifs to explain eQTLs


SNP1
SNP2
SNP3
a
Degree of association with expression
b
Evolutionary conservation
Predicted TF motifs
DNaseI cleavage
ChIP–seq peaks for TF1
Regulatory annotation weight

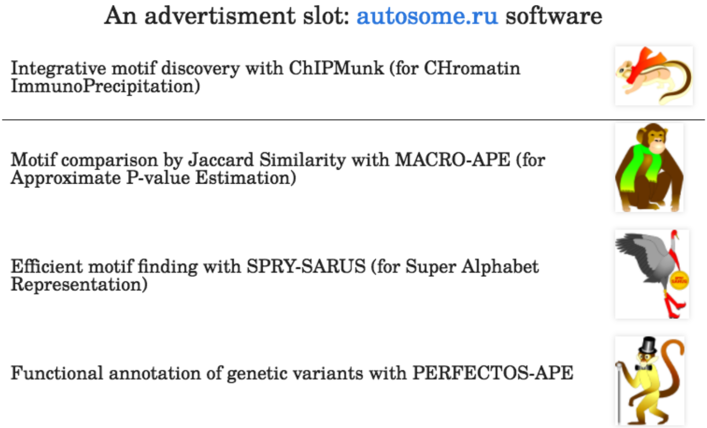
From Levo and Segal, 2014, Nat Rev Genet

Because many other processes (mostly chromatin related) contribute to the protein positioning at the genome

who cite HOCOMOCO (References on 2016 paper, 63 total for Jan. 2018)

Functional genomics (genome structure, annotation, etc)	15
Genetics: annotation of loci and rSNP	13
Systems biology (regulatory networks from DE data)	10
Algorithms and Machine learning assisted genome annotation	7
”Stories” about particular promoters etc	7
DNA - protein interaction studies	6
TF studies - databases, structure of DNA recognition motifs etc	4
Genetic engineering - prediction of genemics manipulation	2
General Molecular biology (transctiption initiation etc)	1

Autosome.RU software family + Hocomoco database



An advertisment slot: autosome.ru software

Integrative motif discovery with ChIPMunk (for CHromatin ImmunoPrecipitation)

Motif comparison by Jaccard Similarity with MACRO-APE (for Approximate P-value Estimation)

Efficient motif finding with SPRY-SARUS (for Super Alphabet Representation)

Functional annotation of genetic variants with PERFECTOS-APE

Who contributed this?

- VIGG RAS:
- Artem Kasianov
- Ivan Kulakovskiy
- Ilya Vorontsov
- Seva Makeev
- KAUST:
- Haitham Ashoor
- Wail Ba-alawi
- Arturo Magana-Mora
- Ulf Schaefer
- Vlad Bajic
- CB RAS:
- Julya Medvedeva
- ISB Ltd:
- Ruslan Shapirov
- Ivan Yevshin
- Fedor Kolpakov
- Skolkovo Tech:
- Dima Papatsenko
- students
- Alla Fedorova, MSU FBB
- Eugen Rumynskiy, MIPT
- Nastya Soboleva, MIPT

Thank you!

- Russian Fund of Basics Research
- Russian Scientific Fund
- Ministry of Science and Education of Russian Federation
- Biobase and personally Edgar Wingender and Alexander Kel
- RIKEN Fantom Project
- Ecole Polytechnique and personally Mireille Regnier