GTRD - a database on gene transcription regulation

Ivan Yevshin, Ruslan Sharipov,
Yuriy Kondrakhin,Semyon Kolmykov,
Fedor Kolpakov

Biosoft.RU LLC
Institute of Computational technologies SB RAS

http://gtrd.biouml.org

Nucleic Acids Res. 2017 Jan 4;45(D1):D61-D67.

30 Years TRANSFAC

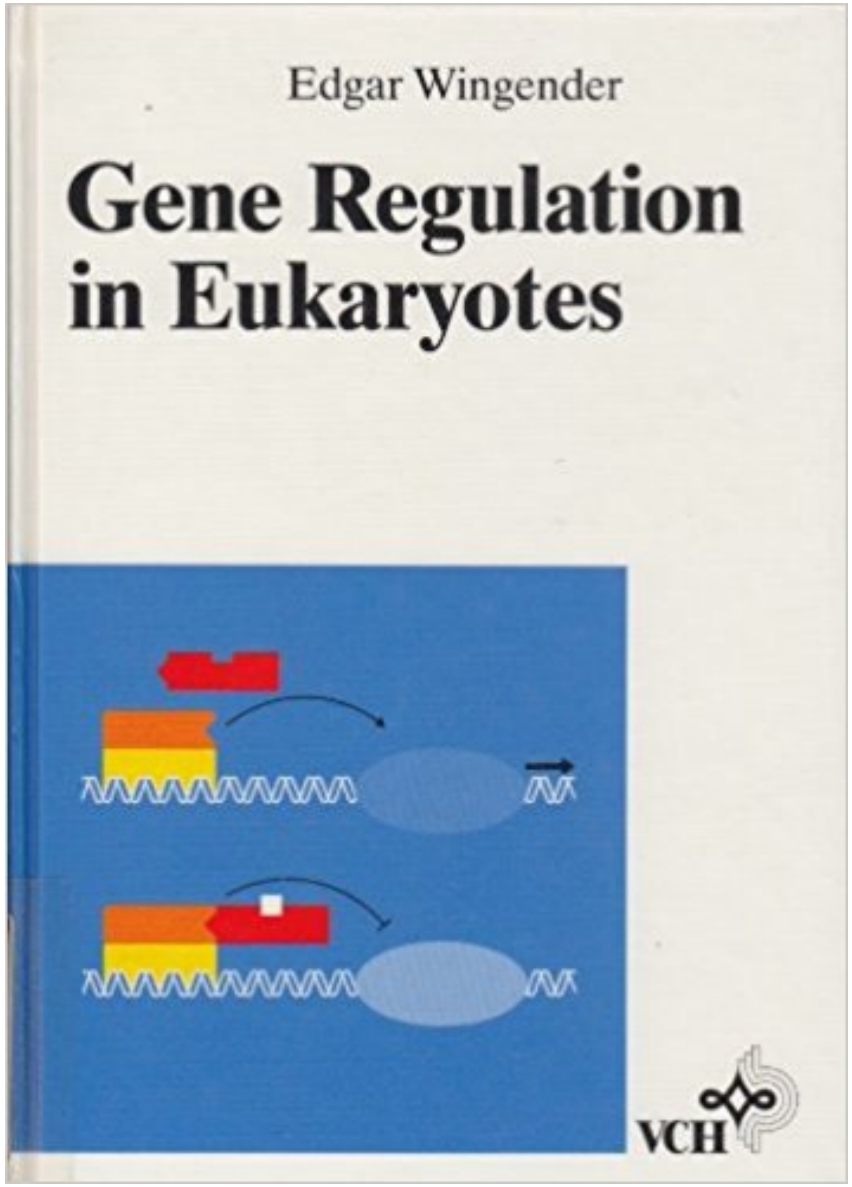
24 years of my life

1993, October

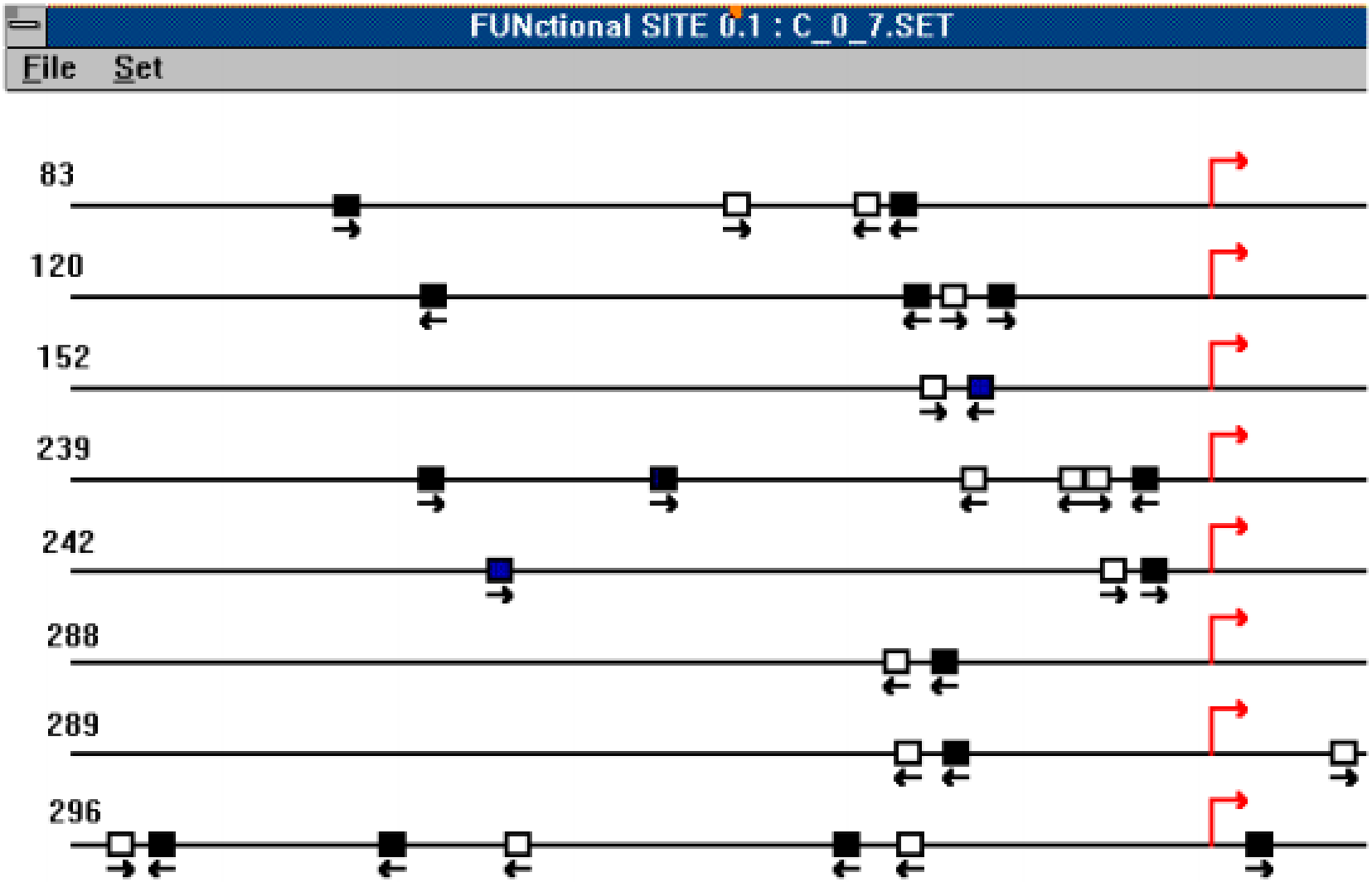
- laboratory of Theoretical Genetics, Institute of Cytology & Genetics, Novosibirsk, head of lab Prof. Nikolay Kolchanov

- start work on diploma “Computer analyses of promoter regions”
  supervisers: Alexander Kel
  Yuriy Kondrakhin

1994, July (summer holidays)



1994 – my first genome browser (C++, Windows 3.1)



Localization of the potential composite elements formed by AP-1(white) and RAR (black) binding sites in promoter sequences (from -500 to +100).
Genes are: 83 (number from EPD) – human islet amiloid polipeptide gene; 120 - …

1995

COMPUTER TOOL **FUNSITE** FOR ANALYSIS OF EUKARYOTIC REGULATORY GENOMIC SEQUENCES

Kel A.E., Kondrakhin Y.V., Kolpakov Ph., Kel O.V., Romashenko A.G., Wingender E.[a], Milanesi L.[b], Kolchanov N.A.

Institute of Cytology and Genetics, Siberian Branch, Russian Academy of Sciences, 630090 Novosibirsk, Russia, e.mail: kol@cgi.nsk.su, fax: (3832) 356558

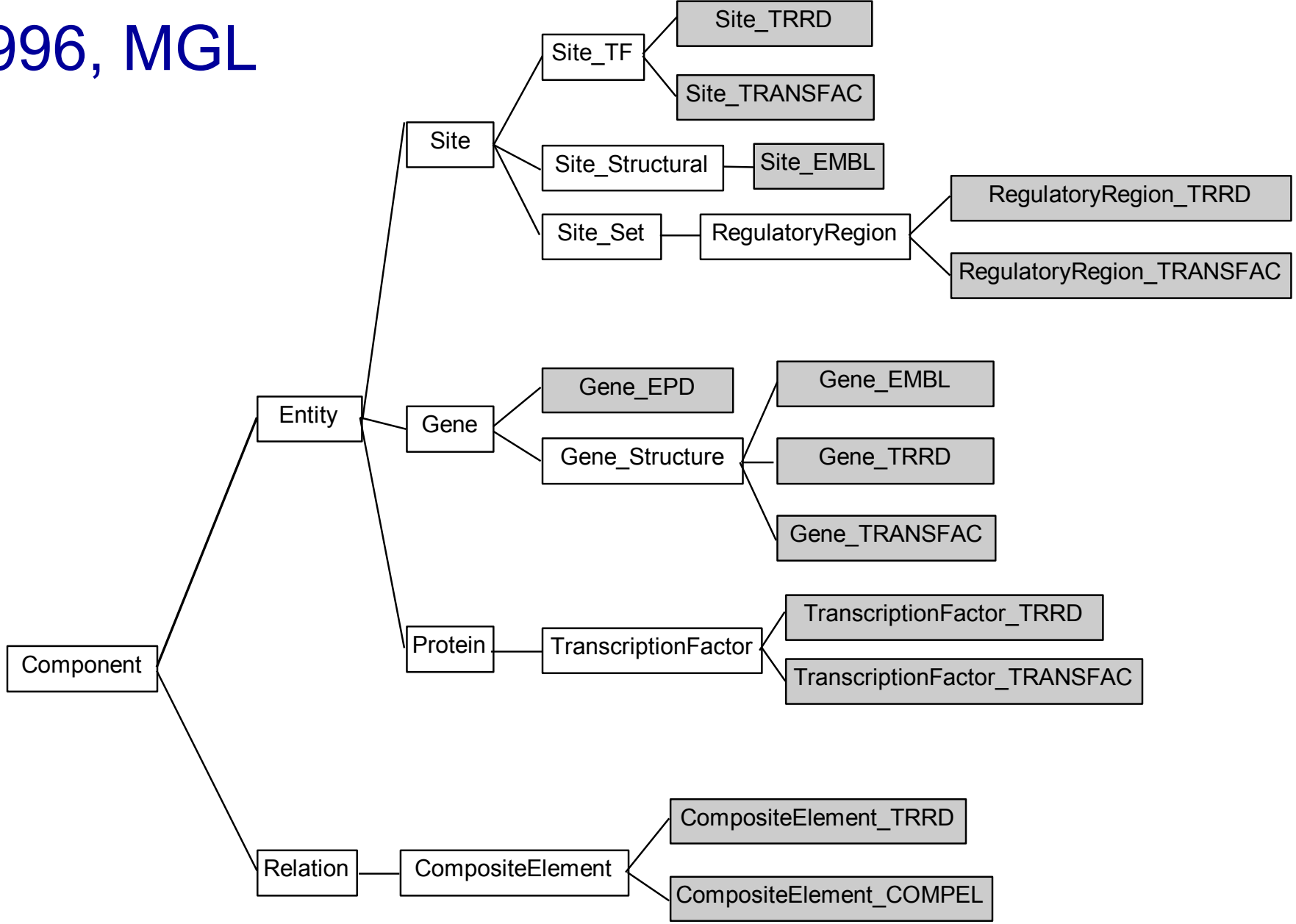
a) Gesellschaft fur Biotechnologische Forschung mbH, Maschroder Weg 1., D-38124 Braunschweig, Germany, E.mail: ewi@venus.gbf-braunschweig.d400.de

b) Instituto di Tecnologie Biomediche Avanzate, CNR, via Ampere n.56, 20131 Milano, Italy

Abstract

We present the computer tool FunSite for description and analysis of regulatory sequences of eukaryotic genomes. The tool consists of the following main parts: 1) An integrated database for genomic regulatory sequences. The integrated database was designed on the basis of the databases TRANSFAC [1] and TRRD [2] that are currently under development. The following functions are performed: i) linkage to the EMBL database; ii) preparing samples of definite types of functional sites with their flanking sequences; iii) preparing samples of promoter sequences; iv) preparing samples of transcription factors classified with regard to structural and functional features of DNA binding and activating domains, functional families of the factors, their tissue specificity and other functional features; v) access to data on mutual disposition of cis-elements within the regulatory regions. 2) The second component of FunSite tool is the set of programs for analysis of the structural organization of regulatory sequences: i) Program for revealing of potential transcription factors binding sites based on their consensi; ii) program for

1996, MGL



MGL computer system – class hierarchy for integration information from different databases on gene expression regulation

1996 – MGL (Molecular Genetics Language) (C++, Windows)

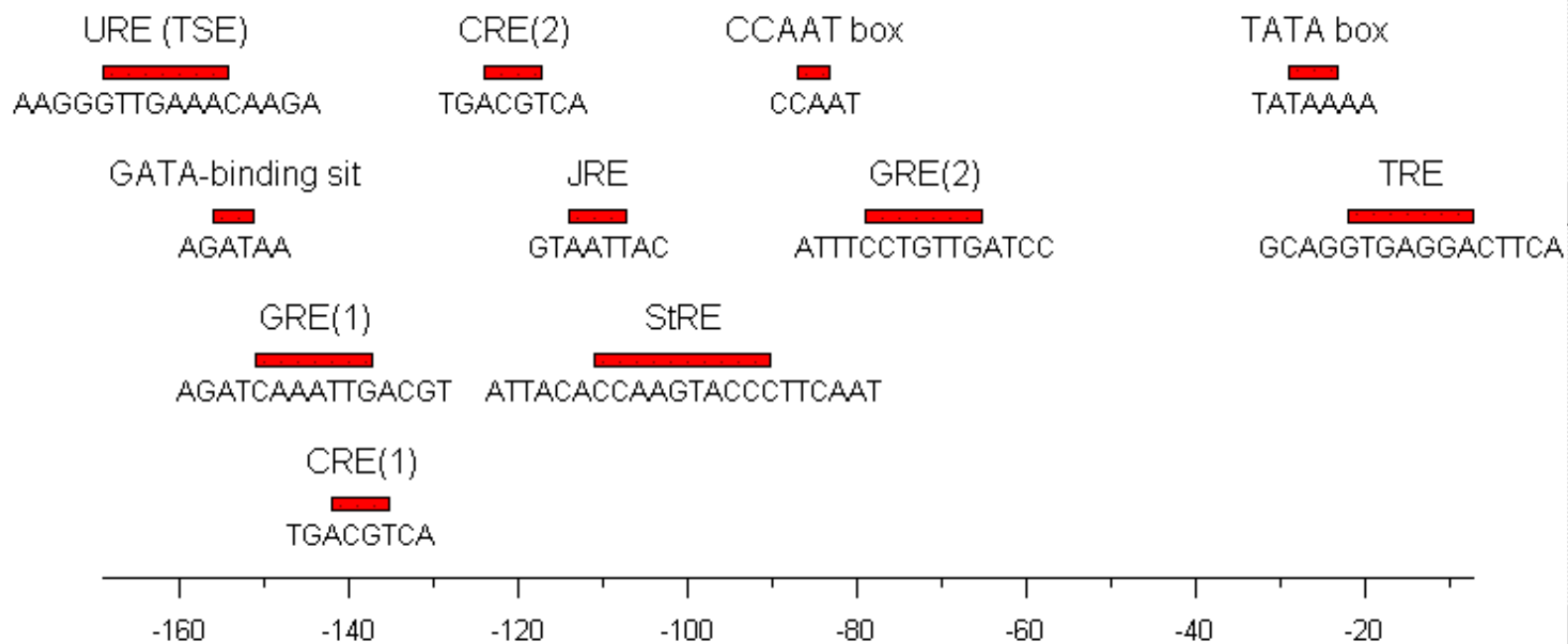
MGL (Molecular Genetic Language) alpha version 1.0

File Run Window Help

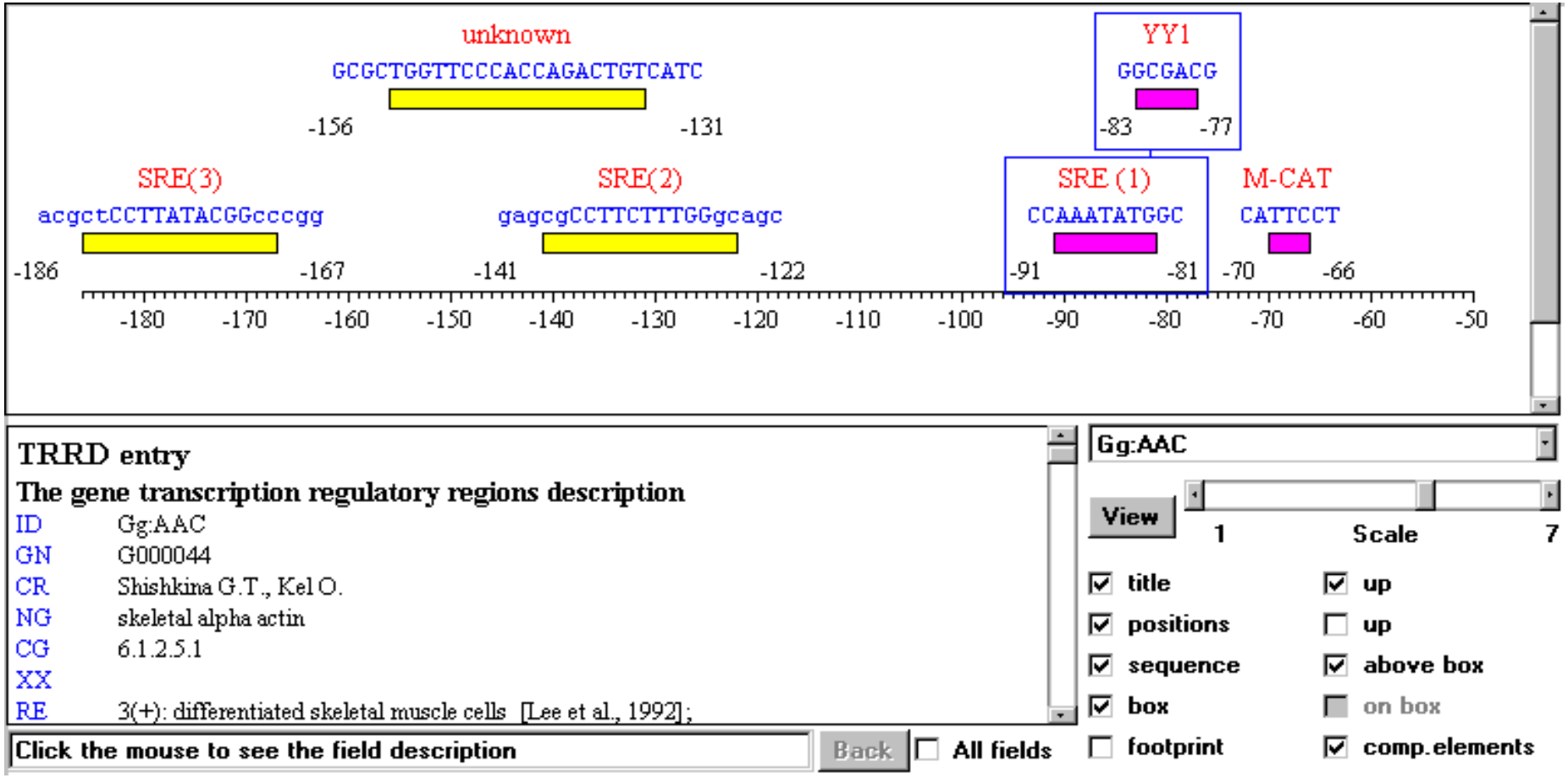
e:\MGL\example8.mgl

```
// Пример 8. MGL программа для графического представления карты строения
// транскрипционного регуляторного района гена из БД TRRD
DATABASE  trrd = DB_Connect ("TRRD");              // установление связи системы MGL с БД
CARD_SET  c = DB_Search ( trrd, " ID='Hs:GHA' " ); // поиск заданного гена
          TRRDViewGene ( c,                        // представление в графическом виде это
                         307 );                    // режим графического представления
```

Gene: glycoprotein hormone alpha subunit gene (56:271:Hs:GHA)



1996 – TRRD/TRANSFAC viewer (Java 1.0)



199
8

362–367 *Nucleic Acids Research, 1998, Vol. 26, No. 1*

Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL

T. Heinemeyer, E. Wingender*, I. Reuter, H. Hermjakob, A. E. Kel[1], O. V. Kel[1], E. V. Ignatieva[1], E. A. Ananko[1], O. A. Podkolodnaya[1], F. A. Kolpakov[1], N. L. Podkolodny[1] and N. A. Kolchanov[1]

Gesellschaft für Biotechnologische Forschung mbH, Mascheroder Weg 1, D-38124 Braunschweig, Germany and [1]Institute of Cytology and Genetics SB RAS, pr. Lavrentyeva-10, 630090 Novosibirsk, Russia

Received September 30, 1997; Accepted October 3, 1997

ABSTRACT

TRANSFAC, TRRD (Transcription Regulatory Region Database) and COMPEL are databases which store information about transcriptional regulation in eukaryotic cells. The three databases provide distinct views on the components involved in transcription: transcription factors and their binding sites and binding profiles (TRANSFAC), the regulatory hierarchy of

(Transcription Regulatory Region Database, developed at the Institute of Cytology and Genetics SB RAS since 1993; 4,5) and COMPEL (about Composite Elements, a common effort of both groups; 6) try to match these requirements. Their specific aims and present status as well as their linkages will be described subsequently.

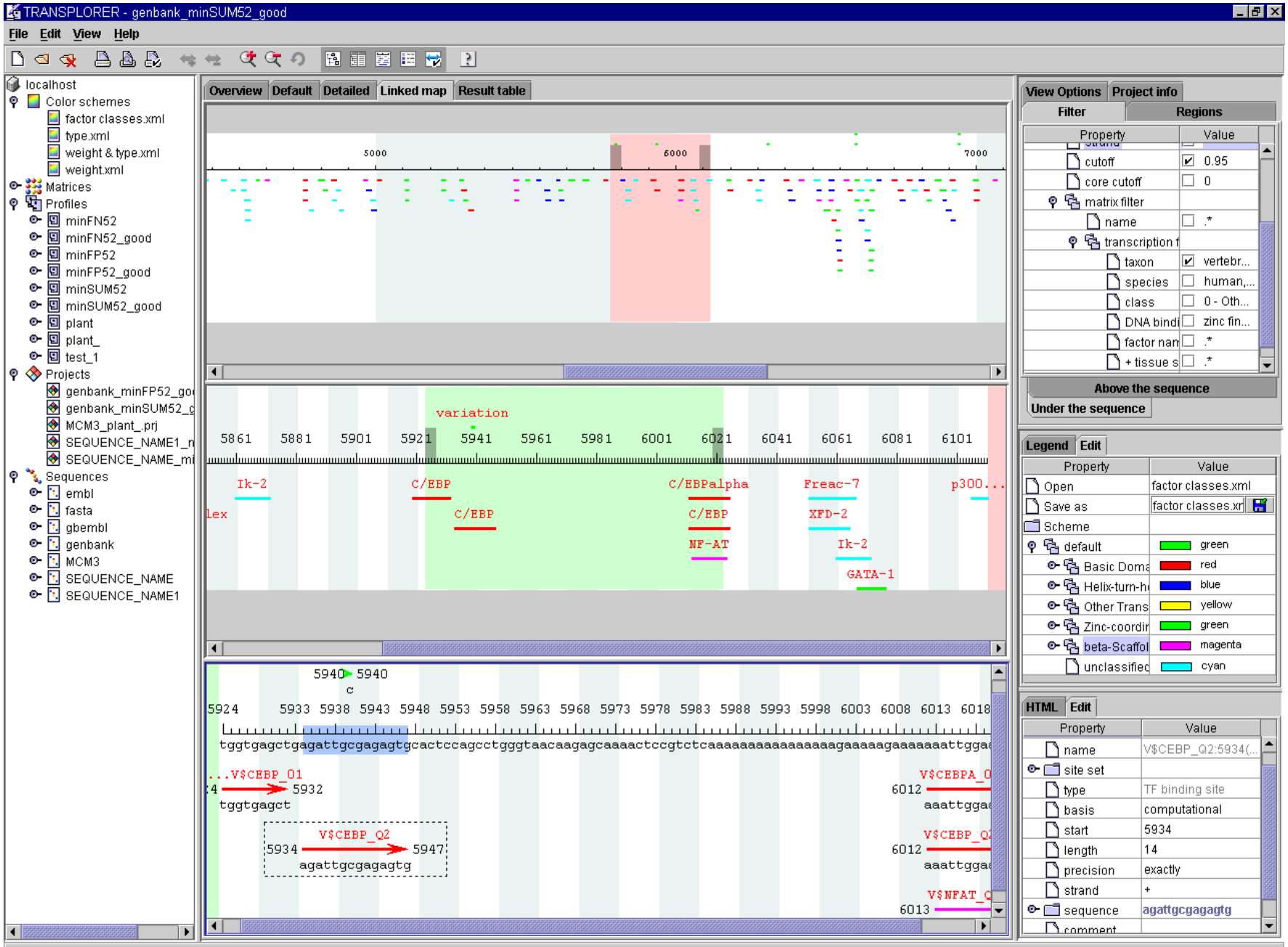
Users are asked to cite this article when publishing results which have been obtained with the database tools described here.

2000 – new company – DevelopmentOnTheEdge.com
- first contract – development of TRANSPLORER software for BIOBASE GmbH

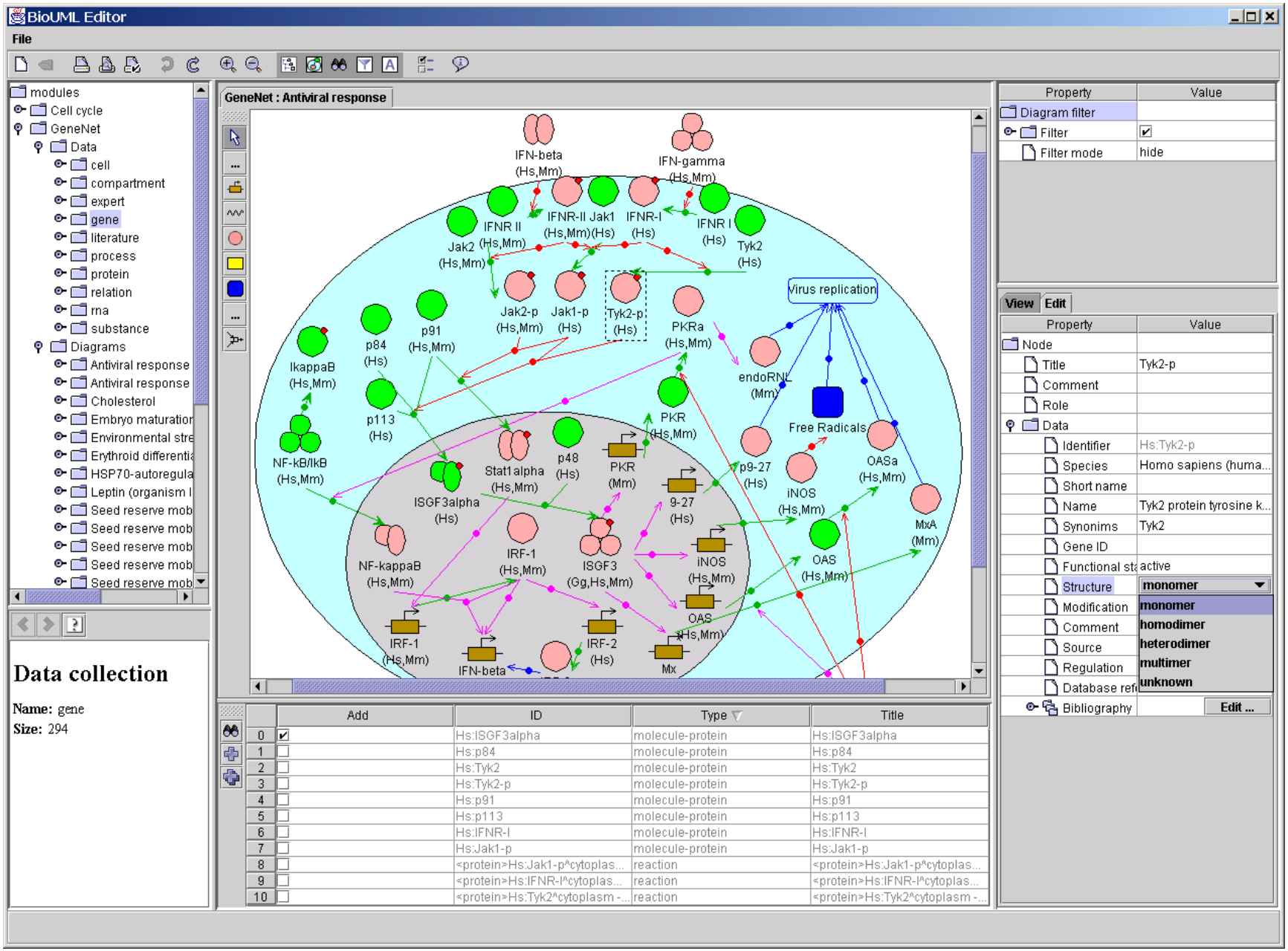
From TRANSPLORER user guide:

TRANSPLORER (TRANScription exPLORER) is a software package for the analysis of transcription regulatory sequences. It includes a tool for prediction of potential binding sites for transcription factors in any sequence that may be of interest. Currently, TRANSPLORER site prediction tool uses position weight matrices (PWM) collections. It is able to use several matrix sources: the largest and most up-to-date library of matrices derived from TRANSFAC® Professional database, other matrix libraries as well as any user-developed matrix libraries. This means that it provides an opportunity to search for a great variety of different transcription factor binding sites. A search can be made using all or subsets of matrices from the libraries.

2001 – TRANSPLORER was released



2002 – BioUML project is started



2006-2007 – development of web interface for BIOBASE Knowledge Library

Locus Report - Human p53 (general)

BIOBASE
BIOLOGICAL DATABASES

BIOBASE Knowledge Library

SEARCH

▼ Genome
FACTOR
GENE
SITE
ChIP-chip
MATRIX
CLASS

PROMOTER
COMPOSITE ELEMENT
S/MART

▼ Proteome
PROTEIN
FAMILY
DOMAIN

▼ Pathways
MOLECULE
REACTION
GENE
PATHWAY
MAPS

▼ ???
DISEASE
MUTATED FACTOR
MUTATED SITE
GENOTYPE
PHENOTYPE
DIAGNOSIS METHODS

▼ References

Reaction Search

Help

Select a previous search you want to process NONE VIEW it DELETE it

and choose an operator AND OR NOT
or select
NEW
to create a new search.

Search Terms | Search Fields | Output Fields

Entire words
Case sensitive
Fuzzy

Search Terms 1
Paste a text or upload a file with queries
(e.g. accession numbers and expression levels).
Обзор...

Accession number, Sec. acc. number
All fields

Accession number
Classification
Comments
Decomposed reactions

BROWSE the selected search field

AND OR NOT
Reaction Name
Search Term 2

AND OR NOT
Reaction Name
Search Term 3

Accession number
Accession numbers, secondary
Classification
Comments
Decomposed reactions
Decompositions
Effect
Enzymes
Evidence level
External database hyperlinks
Inhibitors
Location negative and experiment(s)
Location positive and experiment(s)
Medline database hyperlink
Molecule/gene downstream
Molecule/gene upstream
Pathway level

You can select multiple output fields by using the CTRL key while clicking

Submit Query | Reset Query

2009 – BioUML – web edition

Databases | Data | Analyses

- Sequences
  - chromosomes NCBI35
  - chromosomes NCBI36
    - 1
    - 10
    - 11
    - 12
    - 13
    - 14
    - 15
    - 16
    - 17
    - 18
    - 19
    - 2
    - 20
    - 21
    - 22
    - 3
    - 4
    - 5
    - 6
    - 7
    - 8
    - 9
    - c22_H2
    - c5_H2
    - c6_COX
    - c6_QBL
    - MT

About | JavaScript log | SQL query result | 21

Position: 1 Set Find: Go

0Mb 10.00Mb 20.00Mb 30.00Mb 40.00Mb 46.94Mb

Karyotype.. p13 p11.2 q21.1 q21.2 q21.3 q22.11 q22.2 q22.3

GeneTrack

RepeatTrack 98912 sites

Variation.. 219897 sites

Default

ID: KaryotypeTrack

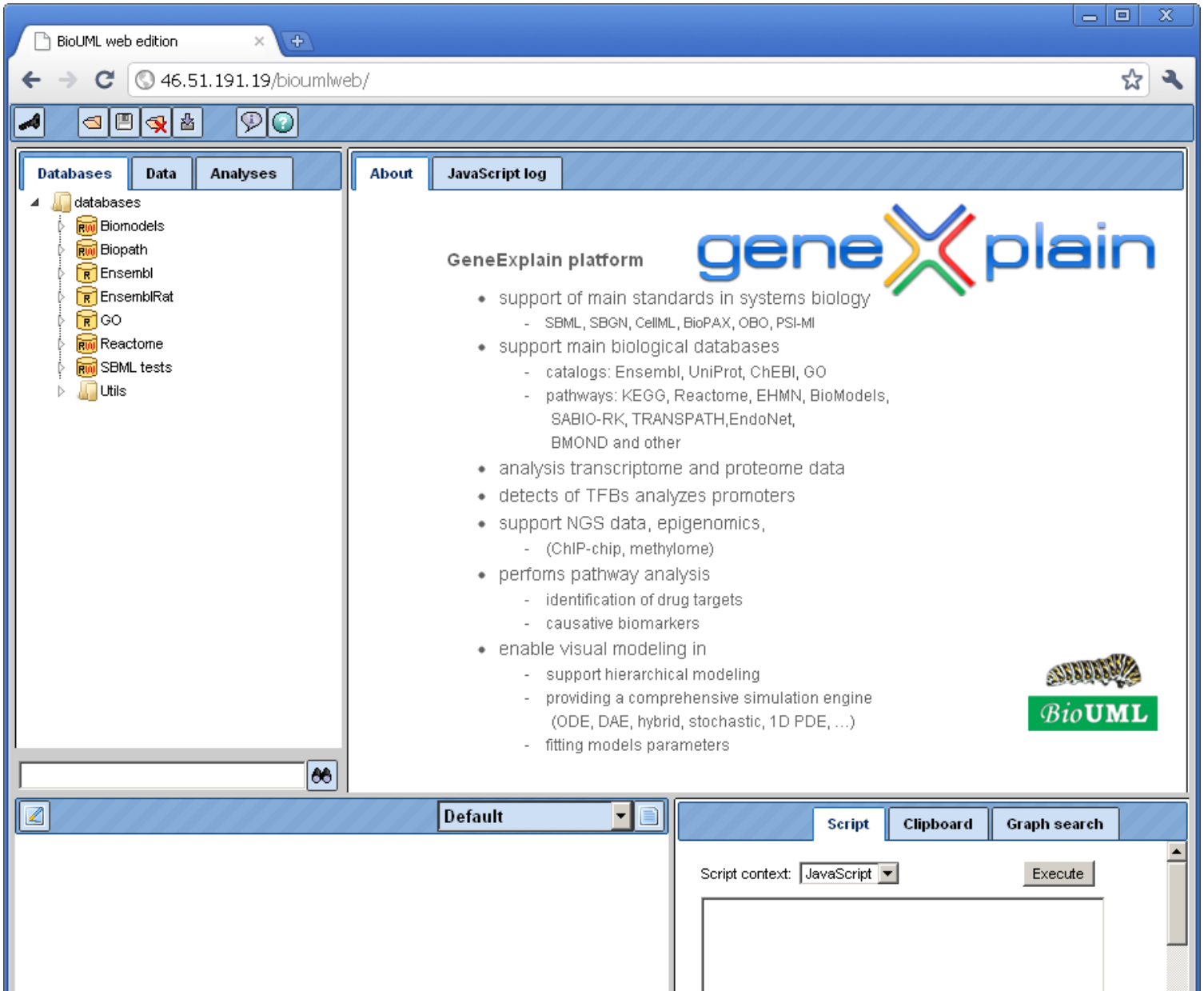
Complete name: databases/Ensembl/Tracks/KaryotypeTrack

Script | Clipboard | Graph search | SQL Editor | Tasks | Sites | Tracks

Script context: JavaScript Execute

2010 – geneXplain GmbH

geneXplain platform (branch of BioUML), v. 1.0 is released



GTRD - a database on gene transcription regulation

The Ensembl gene annotation process

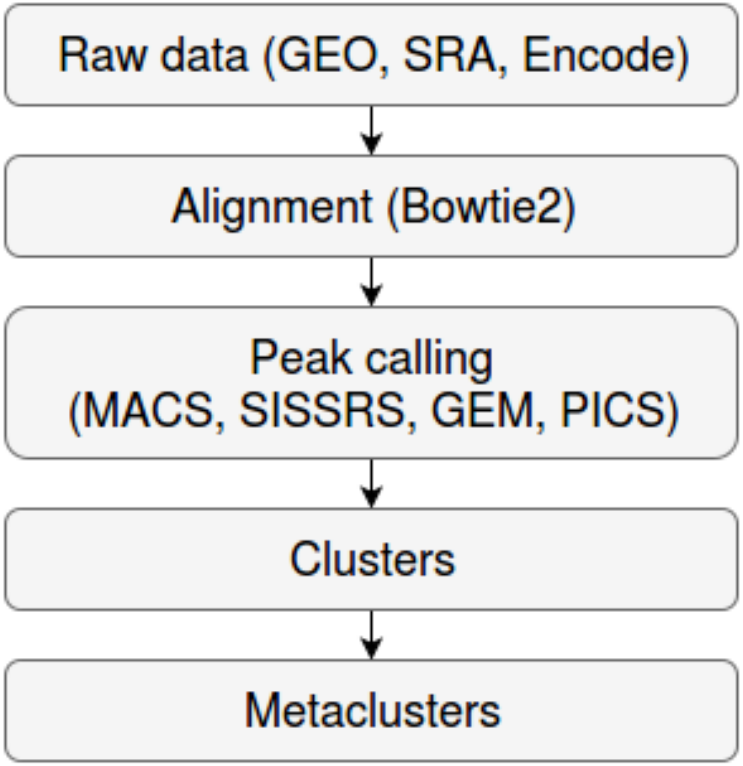
gene structure

The Ensembl gene annotation process

gene structure

The GTRD annotation workflow

gene regulation

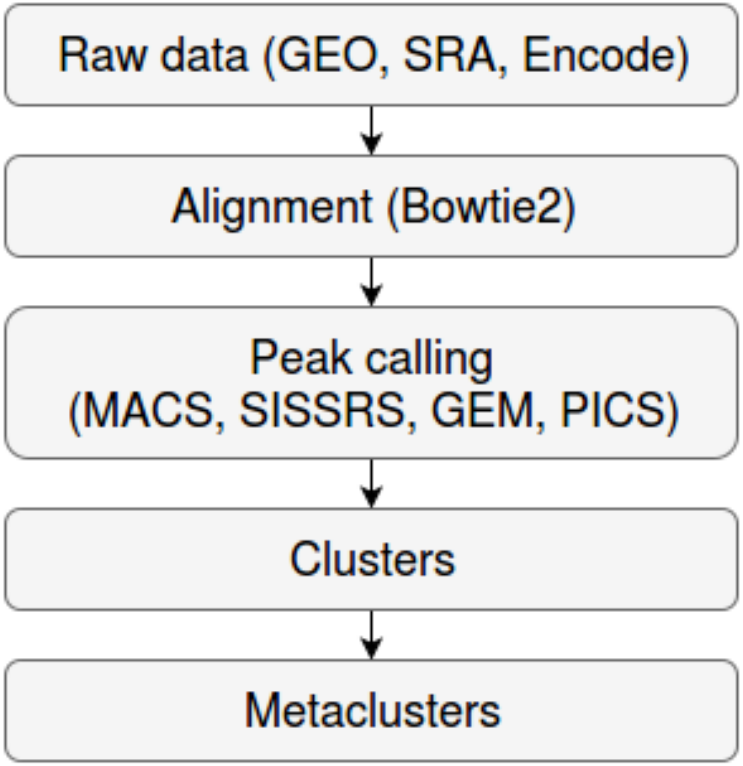


The Ensembl gene annotation process

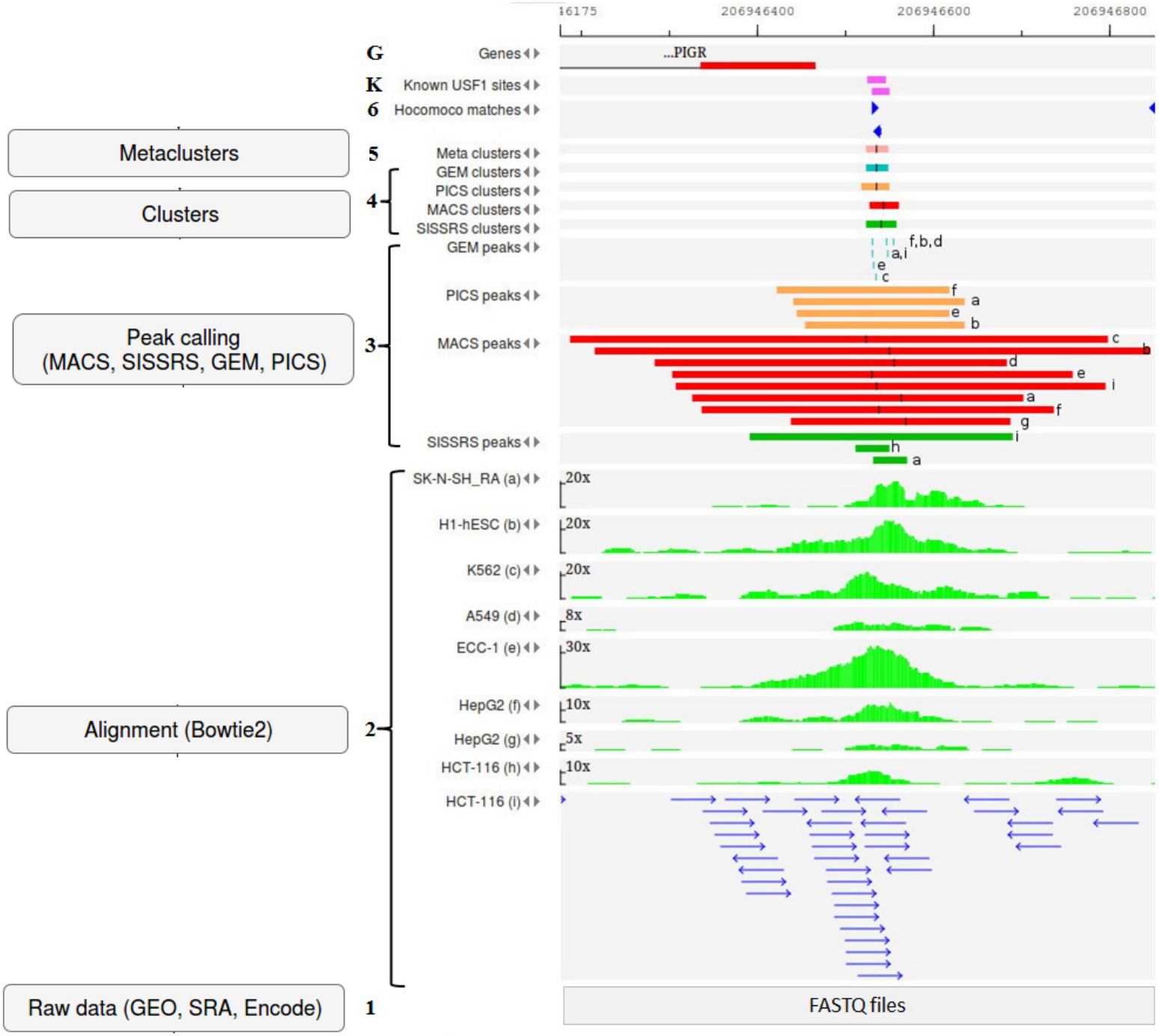
gene structure

The GTRD annotation workflow

gene regulation



The goal of GTRD is to be Ensembl for gene regulation


Metaclusters
Clusters
Peak calling
(MACS, SISSRS, GEM, PICS)
Alignment (Bowtie2)
Raw data (GEO, SRA, Encode)
G
K
6
5
4
3
2
1
Genes
Known USF1 sites
Hocomoco matches
Meta clusters
GEM clusters
PICS clusters
MACS clusters
SISSRS clusters
GEM peaks
PICS peaks
MACS peaks
SISSRS peaks
SK-N-SH_RA (a)
H1-hESC (b)
K562 (c)
A549 (d)
ECC-1 (e)
HepG2 (f)
HepG2 (g)
HCT-116 (h)
HCT-116 (i)
206946400
206946600
206946800
...PIGR
20x
8x
30x
10x
5x
FASTQ files

BioUML platform - main features

- **Systems biology**
  - Supports main standards used in systems biology:
    SBML, SBGN, CellML, BioPAX, OBO, PSI-MI…
  - visual modeling:
    - simulation engine supports (ODE, DAE, hybrid, stochastic, 1D PDE)
    - composite models
    - agent based modeling, rule based modeling
  - parameters fitting

- **Omics data analyses**
  - powerful workflow engine and scripts (R, Javascript)
  - integration with Galaxy, R/Bioconductor
  - workflows and methods for omics data analysis
  - integrated genome browser

- **Collaborative reproducible research**
  - web interface for collaborative work
  - user's data are organized as projects
  - project administrator grants access rights
  - all user actions are tracked in project journal
  - collaborative work on diagrams, models, workflows (like Google documents)

Merge peaks - BioUML 2

gtrd.biouml.org/bioumlweb/#de=analyses/Methods/GTRD/Merge%20peaks

Default

Research

Databases | Data | Analyses | Users

- GTRD
  - Compare Experiments
  - Create flat files
  - Diff Encode
  - Gene features
  - Import Encode
  - Join GTRD Tracks
  - Join GTRD clusters
  - Make meta tracks
  - Match Encode antibodies
  - Merge peak callers
  - **Merge peaks**
  - Open per TF view
  - Open tracks for all TF
  - Predict and merge
  - Prepare Search by regulatio
  - Prepare cluster to exp table
  - Prepare finished tables
  - Search binding sites
  - Search regulated genes
  - Symmetry points
  - Upgrade
  - Validate Experiments

Start page | Join GTRD Tracks X | Merge peaks X

inputTrack	(select element)
maxDistance	50
useDensityClusterization	
useGlobalSD	
bindingSiteWidth	0
maxPropWidth	100
outputTrack	(select element)
sdTable	(select element)
clusterToPeakTable	(select element)

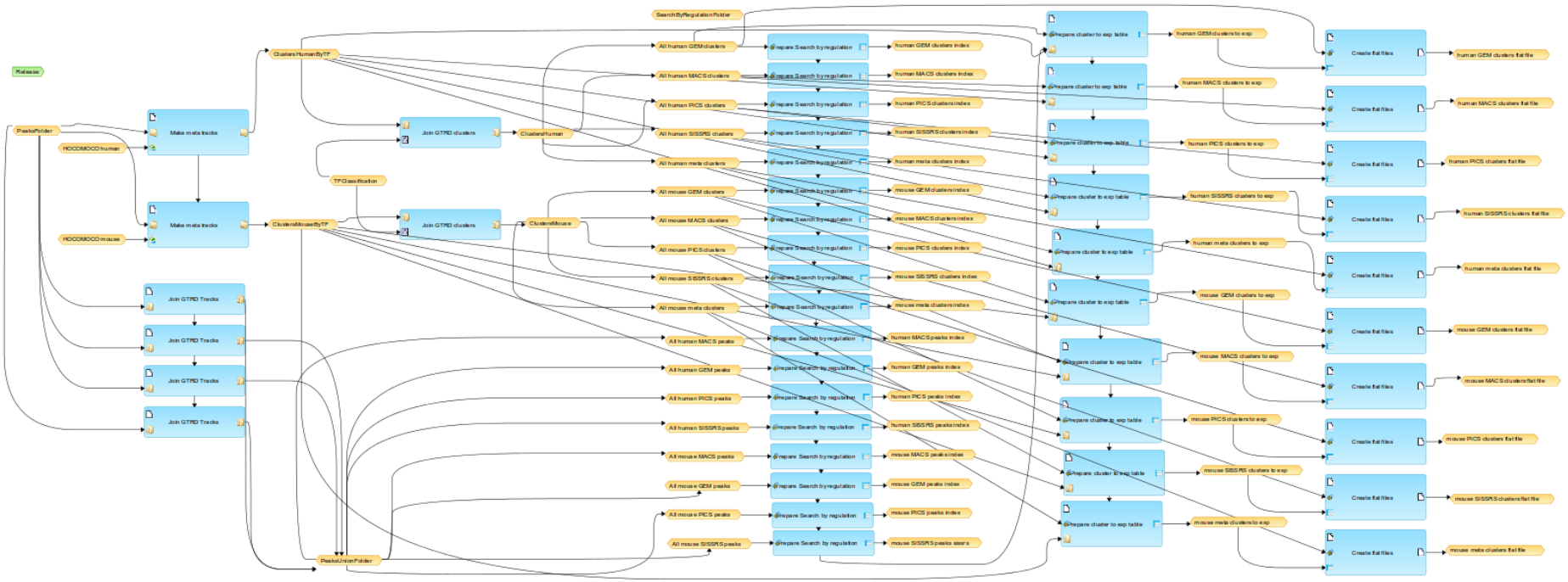
Run

Search | Info

Default

My description | Graph sea

BioUML workflow (fragment) for chip-seq data processing for GTRD database



GTRD

http://gtrd.biouml.org

v18.01

Gene Transcription Regulation Database

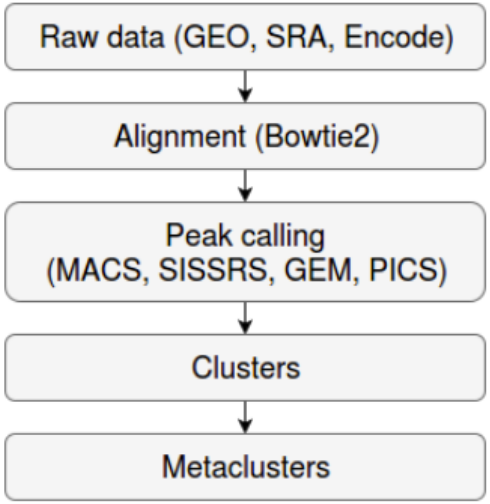
The most complete collection of uniformly processed ChIP-seq data to identify transcription factor binding sites for human and mouse. Convenient web interface with advanced search, browsing and genome browser based on the BioUML platform. For support or any questions contact *ivan@dote.ru*

Start » | Documentation » | Download »

Previous release »

Workflow

How was it constructed?



ChIP-seq experiment information and raw data were collected from publically available sources. Sequenced reads were aligned using Bowtie2 and ChIP-seq peaks were called using 4 different methods. Peaks were merged into clusters and metaclusters to produce non-redundant set of transcription factor binding sites.

Statistics

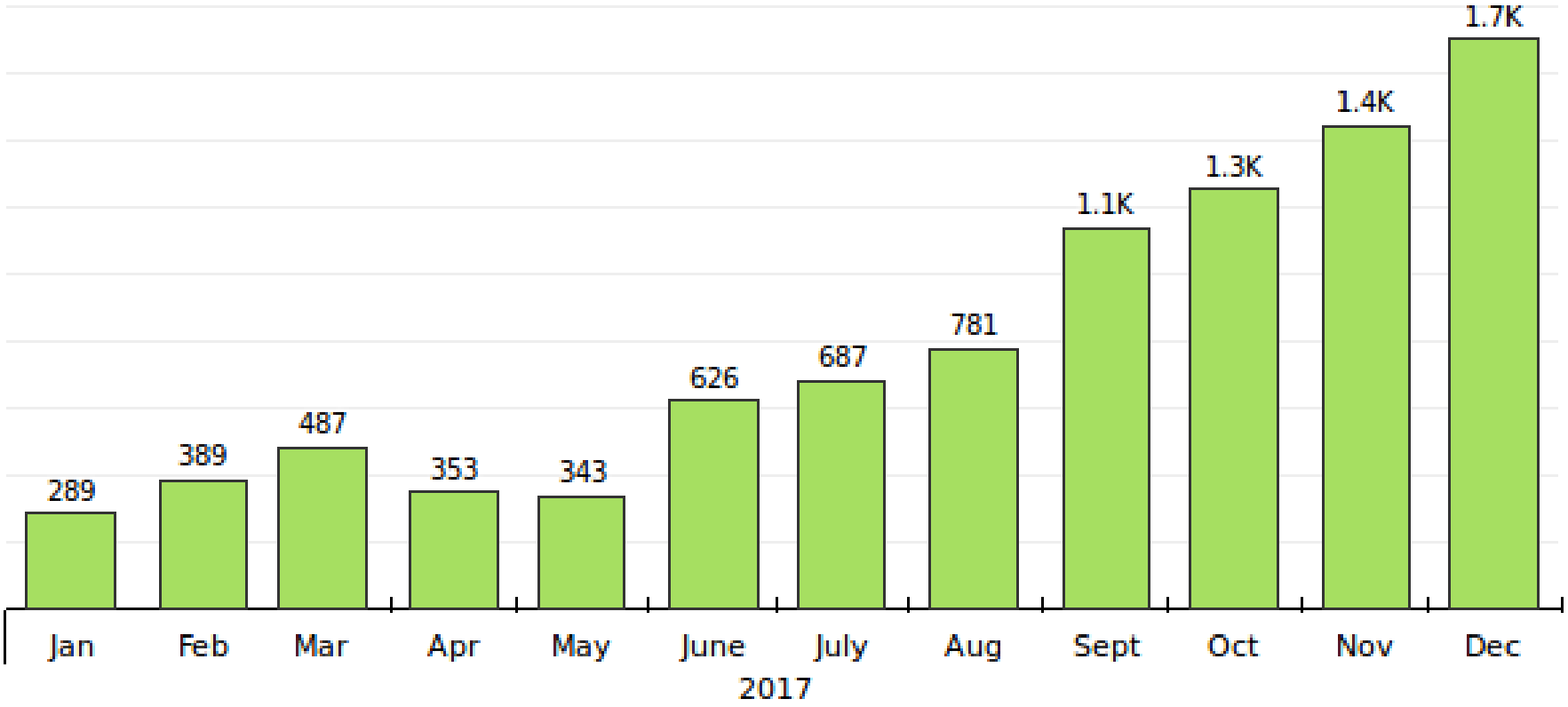
version 18.01

ChIP-seq experiments	12168	3340 new
Transcription factors	766	53 new
ChIP-seq reads	508 023 543 763	27% new
Reads aligned	335 707 127 761	29% new
ChIP-seq peaks	961 923 622	31% new
Clusters	527 498 674	20% new
Metaclusters	70 338 813	14% new

Learn more »

Database	Number of samples * - and other	Number of TFs * - and other	ChIP-seq peak callers	Metacluster approach
GTRD v18.01	10 418 – total 5 603 – human 4 815 – mouse	766 – TFClass classes 682 – human 384 – mouse	4 (MACS, SISSRs, GEM, PICS)	Yes
ChIP-Atlas	10 774 – total 5 914 – human 4 860 – mouse	699* – human 502* – mouse	1(MACS2)	No
Cistrome DB	10 276* – total 5 774* – human 4 502* – mouse	260*	1 (MACS2)	No
ReMap 2018	3 549 – human	486* – human	MACS2	Yes (CRMs)
ENCODE	1 448 – total 1 254 – human 194 – mouse	295* – human 52* – mouse	5 (SPP, GEM, PeakSeq, MACS, Hotspot/Hotspot2)	No
ChIPBase	3 549 – total 2 498 – human 1 036 – mouse	252* – for 10 species	each ChIP-seq is processed by different peak caller	No
Factorbook	1 007 – total 837 – human 170 – mouse	167* – human 51* – mouse	None	No
GeneProf	1 692 – total 693 – human 999 – mouse	133 – human 131 – mouse	1(MACS)	No
NGS-QC	6 672 – total 4 234 – human 2 438 – mouse	unknown	None	No

Number of GTRD users, 2017



Location of GTRD users (last 200)

Yevshin, I., Sharipov, R., Valeev, T., Kel, A., Kolpakov, F.
GTRD: A database of transcription factor binding sites identified by ChIP-seq experiments // Nucleic Acids Res. – 2017. – V. 45(D1). – P. D61–D67.

Citations during 2017

1. Eukaryotic and prokaryotic promoter databases as valuable tools in exploring the regulation of gene transcription: a comprehensive overview. **Gene**. 2017 Nov 2. **IF - 2.4**
2. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. **Nucleic Acids Res**. 2017 Nov 11. **IF - 10.162**
3. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments **Nucleic Acids Res**. 2017 Nov 8. **IF - 10.162**
4. DMS-Seq for In Vivo Genome-wide Mapping of Protein-DNA Interactions and Nucleosome Centers. **Cell Reports** 2017 Oct 3;21(1):289-300. **IF - 8.3**
5. EpiDenovo: a platform for linking regulatory de novo mutations to developmental epigenetics and diseases. **Nucleic Acids Research** 2017, gkx918 **IF - 10.162**
6. Genetic variants in ADAMTS13 as well as smoking are major determinants of plasma ADAMTS13 levels. **Blood Advances** 2017 1:1037-1046;
7. Master-regulators driving resistance of non-small cell lung cancer cells to p53 reactivator Nutlin-3. **Virtual Biology** 2017, 0(4), 1-31.
8. Discovering relationships between nuclear receptor signaling pathways, genes, and tissues in Transcriptomine. **Sci Signal.** 2017 Apr 25;10(476). **IF - 7.4**
9. A comprehensive review of web-based non-coding RNA resources for cancer research. **Cancer Lett**. 2017 Aug 18;407:1-8. **IF - 6.3**
10. RUNX1 promote invasiveness in pancreatic ductal adenocarcinoma through regulating miR-93. **Oncotarget** 2017 Aug 24 **IF - 5.17**

Use cases

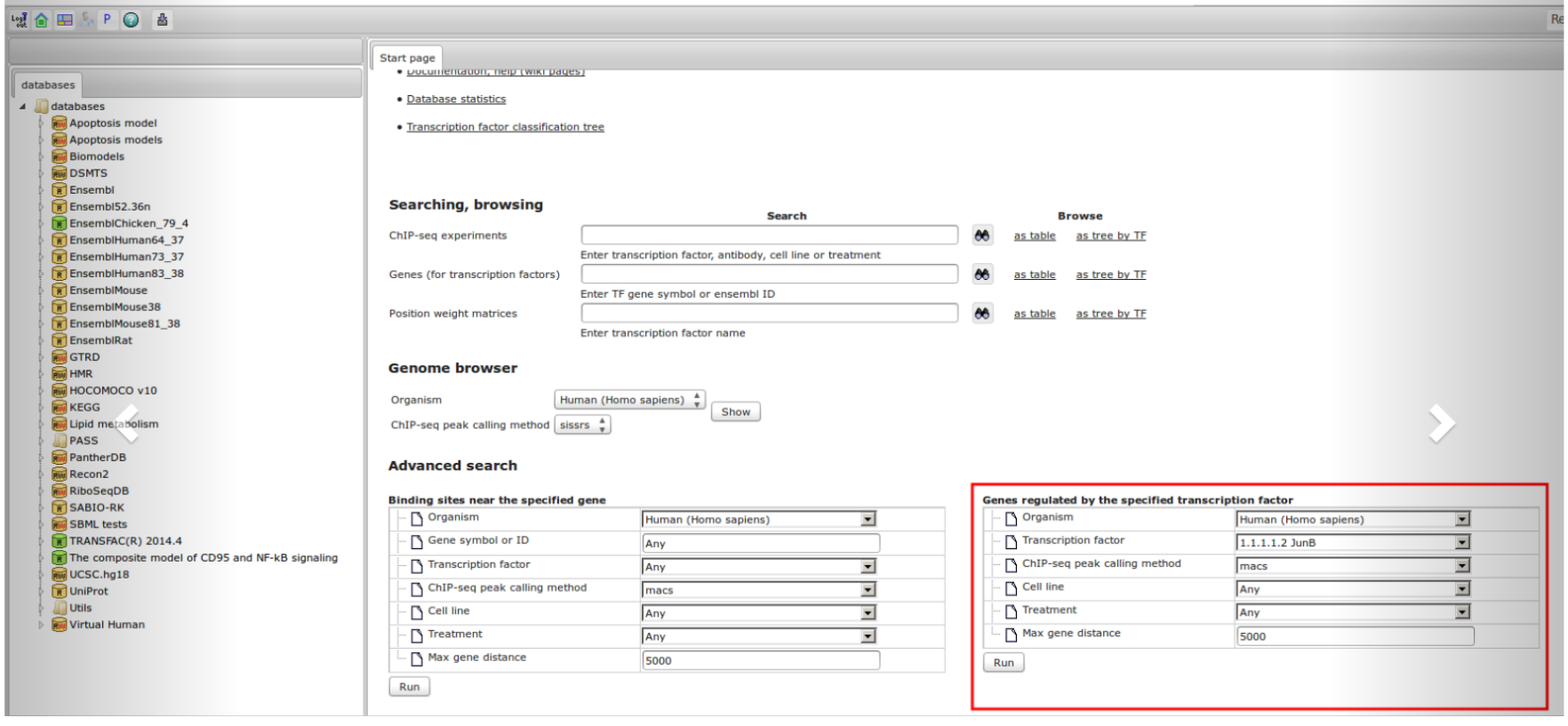
- Search ChIP-seq experiments by transcription factor
- Browse ChIP-seq peaks in genome browser
- Find transcription factor binding sites on gene
- Find genes regulated by transcription factor

Step 1

Select JunB in 'Transcription factor' parameter to search for genes potentially regulated by JunB.

You also can restrict search to specific cell line and treatment.

Press Run button.



BioUML 2016.4 web edit

gtrd.biouml.org/bioumlweb/#

Research: Demo | GTRD

databases

- databases
  - EnsemblHuman85_38
  - EnsemblMouse81_38
  - GTRD
    - Data
      - clusters
      - experiments
      - generic
      - matrices
      - peaks
      - site models
      - views
    - Dictionaries
  - HOCOMOCO v10
  - Utils

Start page

Gene Transcription Regulation Database

- Documentation, help (wiki pages)
- Transcription factor classification tree

Searching, browsing

	Search		Browse	
ChIP-seq experiments	gr		as table	as tree by TF
	Enter transcription factor, antibody, cell line or treatment			
Genes (for transcription factors)			as table	as tree by TF
	Enter TF gene symbol or Ensembl ID			

Genome browser

Display tracks for all TF

Organism	Human (Homo sapiens)
Data type	meta clusters

Show

Display per TF workflow results

Organism	Human (Homo sapiens)
Transcription factor	1.1.1.1.1 c-Jun

Show

Search | Info | Default

ID: Data
Size: 7
Complete name: databases/GTRD/Data
Description: Data folders in GTRD database:

- alignments - alignments created via Bowtie2 on raw ChIP-seq data
- experiments - information about ChIP-seq experiments.

Search result

First | Previous | 1 | 2 | 3 | Next | Last

Showing 1 to 50 of 132 entries

Show 50 entries

ID	Title	Name	TF class	TF title	Cell line	Treatm
				Glucocorticoid		

1. Search, browsing
2. View in genome browser
3. Advanced search

Searching, browsing

	Search		Browse	
ChIP-seq experiments			as table	as tree by TF
	Enter transcription factor, antibody, cell line or treatment			
Genes (for transcription factors)			as table	as tree by TF
	Enter TF gene symbol or ensembl ID			
Position weight matrices			as table	as tree by TF
	Enter transcription factor name			

1

Genome browser

Organism: Human (Homo sapiens)

ChIP-seq peak calling method: sissrs

Show

2

Advanced search

Binding sites near the specified gene

Organism	Human (Homo sapiens)
Gene symbol or ID	Any
Transcription factor	Any
ChIP-seq peak calling method	macs
Cell line	Any
Treatment	Any
Max gene distance	5000

Run

3

Genes regulated by the specified transcription fa

Organism	Huma
Transcription factor	Any
ChIP-seq peak calling method	macs
Cell line	Any
Treatment	Any
Max gene distance	5000

Run

Search ChIP-seq experiments by transcription factor

Research: Demo ▾ GTRD ▾

databases

- databases
  - EnsemblHuman85_38
  - EnsemblMouse81_38
  - GTRD
    - Data
      - clusters
      - experiments
      - generic
      - matrices
      - peaks
      - site models
      - views
    - Dictionaries
  - HOCOMOCO v10
  - Utils

Start page

Gene Transcription Regulation Database

- Documentation, help (wiki pages)
- Transcription factor classification tree

Searching, browsing

	Search		Browse	
ChIP-seq experiments	stat*		as table	as tree by TF
	Enter transcription factor, antibody, cell line or treatment			
Genes (for transcription factors)			as table	as tree by TF
	Enter TF gene symbol or Ensembl ID			

Search | Info | GTRD Experiment ▾

EXP031359

Transcription factor class: 6.2.1.0.1
Transcription factor: STAT1
Antibody: sc-346, Bethyl A302-753A
Cell line: bone marrow-derived dendritic cells
Specie: Mouse (Mus musculus)
Treatment: LPS, 0 min
Control: EXP031267
Reads:
READS031825.fastq.gz (fetched from URL)
READS033999.fastq.gz (fetched from URL)
READS034000.fastq.gz (fetched from URL)
READS034001.fastq.gz (fetched from URL)
READS034002.fastq.gz (fetched from URL)

Search result

First | Previous | 1 | 2 | 3 | 4 | 5 | Next | Last — Showing 1 to 50 of 286 entries — Show 50 entries

ID	Title	Name	TF class	TF title	Cell line	Treatment	Sp
0	EXP031359	EXP031359	6.2.1.0.1	**STAT1**	bone marrow-derived dendritic cells	LPS, 0 min	Mou (Mu mus
1	EXP031360	EXP031360	6.2.1.0.1	**STAT1**	bone marrow-derived dendritic cells	LPS, 120 min	Mou (Mu mus
2	EXP031326	EXP031326	6.2.1.0.3	**STAT3**	bone marrow-derived dendritic cells	LPS, 0 min	Mou (Mu mus
3	EXP031327	EXP031327	6.2.1.0.3	**STAT3**	bone marrow-derived dendritic cells	LPS, 120 min	Mou (Mu

Browse ChIP-seq experiments as table

Searching, browsing

ChIP-seq experiments

Search

Enter transcription factor, antibody, cell line or treatment

Browse

as table

as tree by TF

First | Previous | Page 1 of 49 | Next | Last

Showing 1 to 50 of 2443 entries

Name	Antibody	TF class	TF title	Cell line	Treatment	Specie	External References	Is control experiment	Control	Peak	Alignment
EXP000001	Input			HeLa cells		Human (Homo sapiens)	GSM357350, GSE14283, 19171782				ALIGNS00082
EXP000002	Oct1	3.1.10.2.1	POU2F1 (Oct-1, OTF-1)	HeLa cells	H2O2 1hr	Human (Homo sapiens)	GSM357351, GSE14283, 19171782		EXP000001	PEAKS010302	ALIGNS00082
EXP000004	anti-GFP(ab290-050)			HeLa Kyoto cell line		Human (Homo sapiens)	GSM566168, GSE20303, 20850016				ALIGNS00093
EXP000005	anti-GFP(ab290-050) GATAD1 tagged	2.2.1.2.1	GATAD1 (ODAG)	HeLa Kyoto cell line		Human (Homo sapiens)	GSM566155, GSM566161, 20850016, GSE20303		EXP000004	PEAKS000471	ALIGNS00093
EXP000011	None			Mouse embryonic fibroblasts		Mouse (Mus musculus)	GSM560357, GSE22562, 20720539				ALIGNS00094
EXP000012	CTCF	2.3.3.50.1	CTCF [11]	Mouse embryonic fibroblasts		Mouse (Mus musculus)	GSM560351, GSM560352, 20720539, GSE22562		EXP000011	PEAKS000472	ALIGNS00093
EXP000013	IgG (Millipore)			ES-derived neurons		Mouse (Mus musculus)	GSM818945, GSE33059, 22085726				ALIGNS00094
EXP000014	Sox3 (T. Edlund)	4.1.1.2.3	SOX-3	Sox3-transfected C2C12 cells		Mouse (Mus musculus)	GSM818950, GSE33059, 22085726		EXP000013	PEAKS000473	ALIGNS00094
EXP000015	Sox3 (T. Edlund)	4.1.1.2.3	SOX-3	ES-derived neural progenitor cells		Mouse (Mus musculus)	22085726, GSE33059, GSM818938, GSM818937, GSM818936		EXP000013	PEAKS000474	ALIGNS00094
EXP000016	Sox2 (Millipore)	4.1.1.2.2	SOX-2	ES-derived neural progenitor		Mouse (Mus	GSM818941, 22085726, GSE33059,		EXP000013	PEAKS000475	ALIGNS00094

Browse ChIP-seq experiments in TFClass tree

Searching, browsing

ChIP-seq experiments

Search

Enter transcription factor, antibody, cell line or treatment

Browse

as table | as tree by TF

ID	Title	Experiments	Peaks
1	Basic domains	440	
2	Zinc-coordinating DNA-binding domains	734	
3	Helix-turn-helix domains	435	
4	Other all-α-helical DNA-binding domains	79	
5	α-Helices exposed by β-structures	26	
5.1	MADS box factors	24	
5.1.1	Regulators of differentiation	8	
5.1.1.1	MEF-2	8	
5.1.1.1.1	MEF-2A	EXP010153, EXP010683, EXP010983, EXP011067, EXP030003	PEAKS020153, PEA
5.1.1.1.2	MEF-2B (RSRFR2, xMEF2)	0	
5.1.1.1.3	MEF-2C	EXP010469, EXP030002	PEAKS020469, PEA
5.1.1.1.4	MEF-2D	EXP030081	PEAKS030079
5.1.2	Responders to external signals (SRF/RLM1)	16	
5.2	E2-related factors	0	
5.3	SAND domain factors	2	
6	Immunoglobulin fold	155	

Browse ChIP-seq peaks in genome browser

Genome browser

Display tracks for all TF

Organism: Human (Homo sapiens)

Data type: meta clusters

- meta clusters
- sissrs clusters
- macs clusters
- gem clusters
- pics clusters
- sissrs peaks
- macs peaks
- gem peaks
- pics peaks

Show

Display per TF workflow results

Organism: Human (Homo sapiens)

Transcription factor: (not selected)

- (not selected)
- 0.0.6.0.1 NRF-1 (α-pal)
- 0.2.1.0.2 SFPQ (PSF)
- 1.1.1.1.1 c-Jun
- 1.1.1.1.2 JunB
- 1.1.1.1.3 JunD
- 1.1.1.2.1 NF-E2 p45
- 1.1.1.2.2 NF-E2L1 (NRF1)
- 1.1.1.2.3 NF-E2L2 (NRF2)
- 1.1.1.2.4 NF-E2L3 (NRF3)
- 1.1.1.2.5 BACH1
- 1.1.1.2.6 BACH2
- 1.1.1.3.1 ATF-2

Show

Start page | chromosomes GRCh38 X

Sequence (chromosome): 1 Position: 1:963816-965543 Set

816 964000 964500 965000 96550

Genes ...KLHL17...

all meta clusters c-Myc Estrogen receptor α (ERα) (NR3A1) Runx1 (PEBP2αB, CBF-α2, AML-1) NF-E2 p4

E2A (TCF-3, ITF-1) ERG Estrogen receptor α (ERα) (NR3A1) CTCF [11]

ERG p63 NF-E2 p45 ERG p53 Fli-1 Elf-1

Sp1 CLOCK STAT5B

Fli-1 FOXA2 (HNF-3β)

SCRT1

E2F-6

Sp1

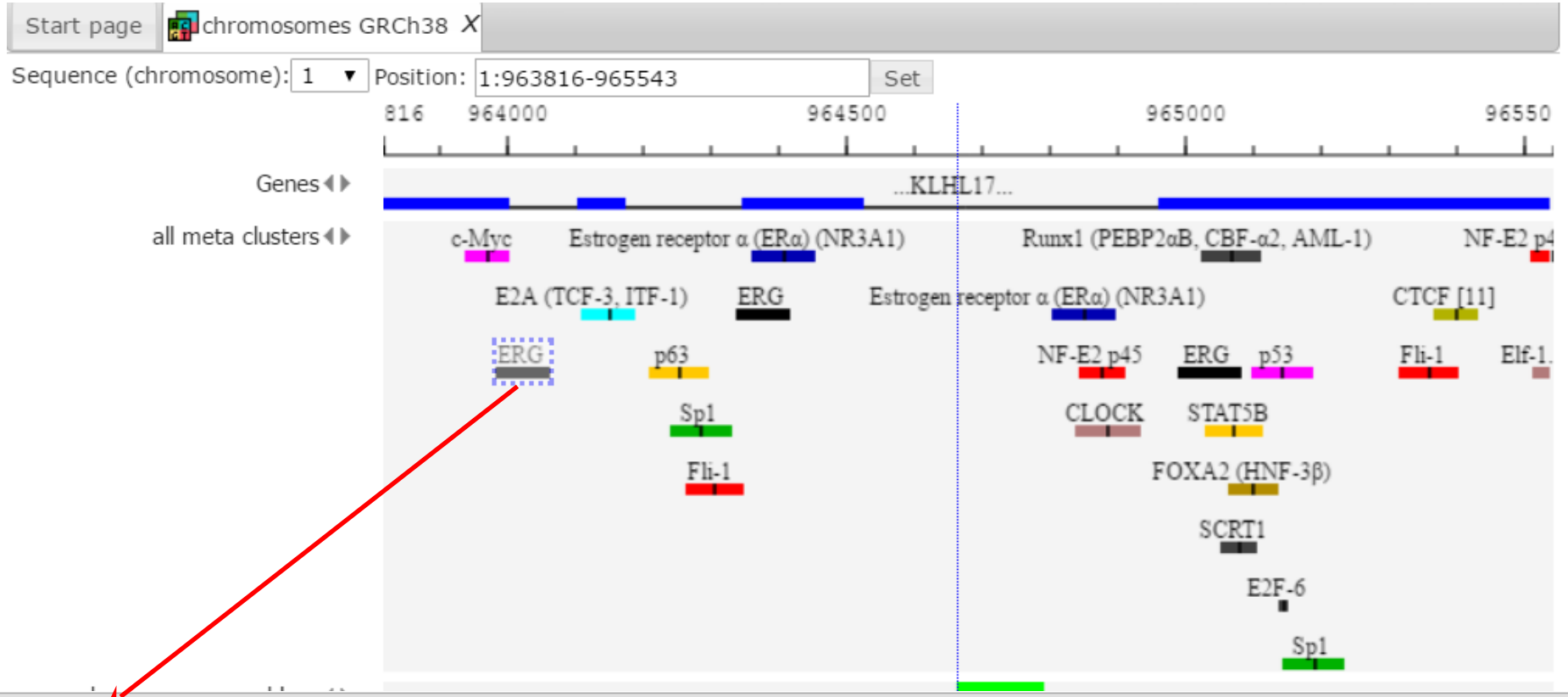
human_unmappable...

Repeats trf (TTTA)n

dust dust dust

964 660 trf

trf



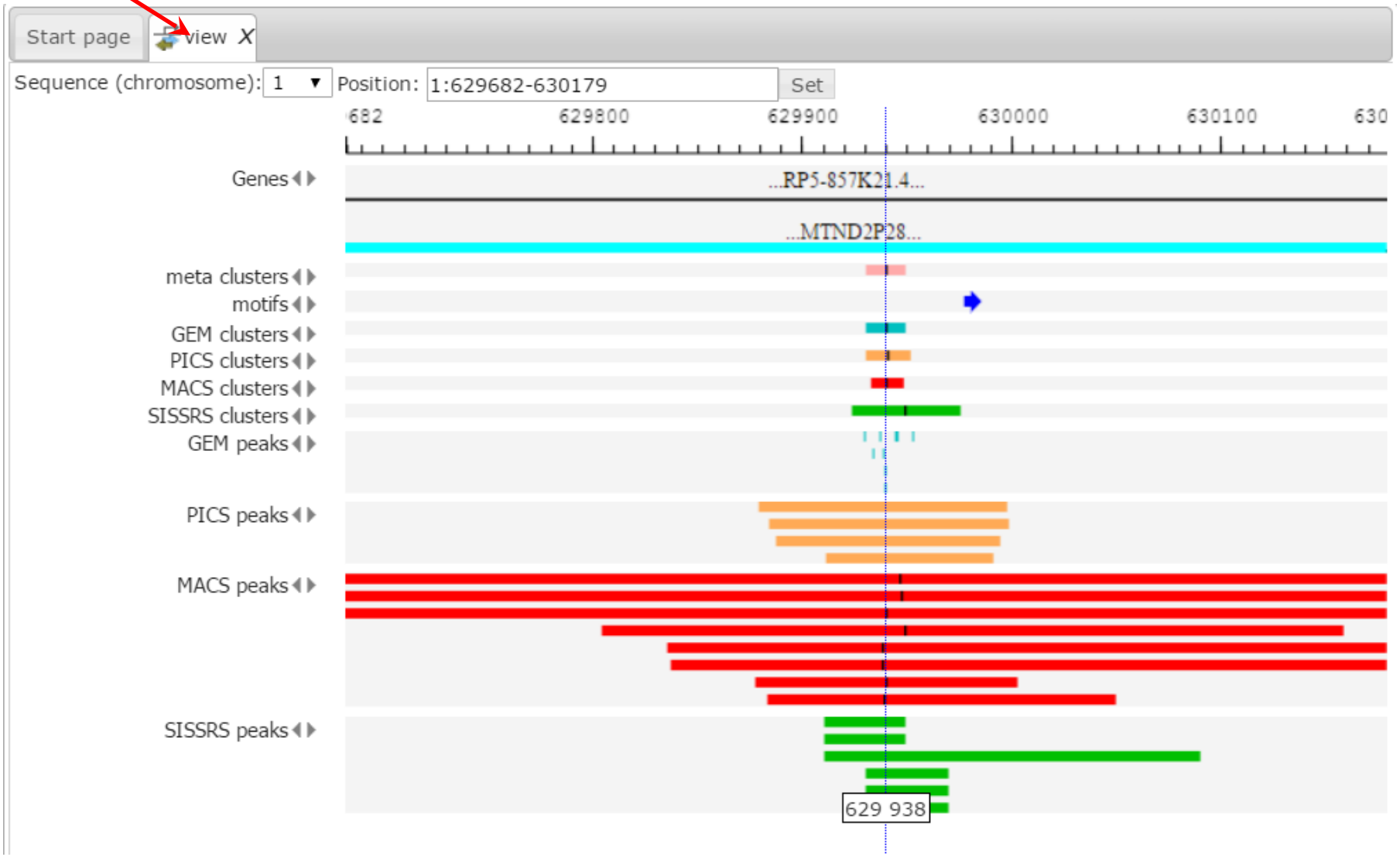
Site ID: 28228446
Type: ERG
Sequence name: 1
Sequence: GCTATGTGCGAGTGGCCACGCTTGGTGGGTGATGGGGCCTGCCTGGGGGGCATCCCCACCTTCCCCCACCGTGGAGACCCCAC
Position: 963985 - 964067 (83)
Properties:

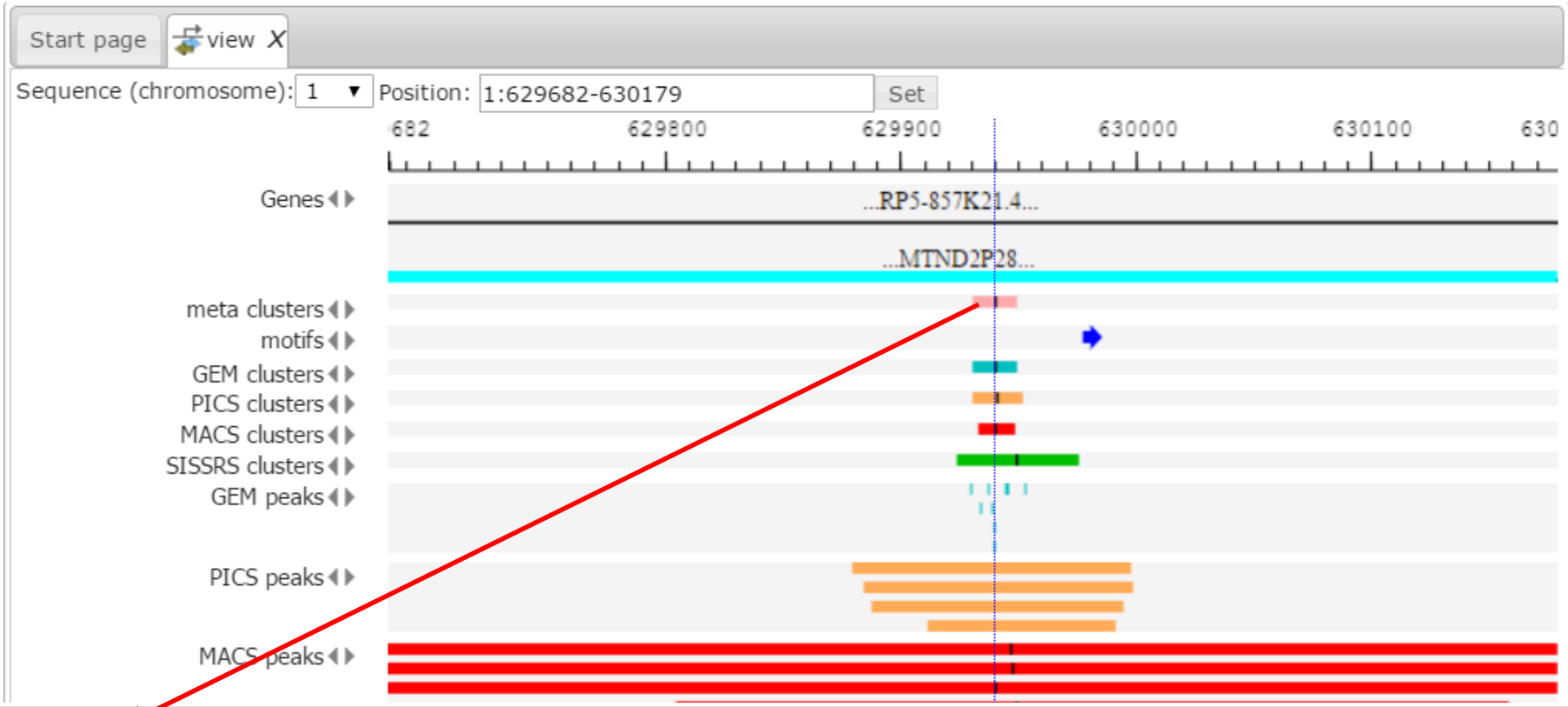
- **antibody.set**: ERG
- **cell.set**: CD34+ nr29 (normal);AML pz12 (leukemic)
- **exp.set**: EXP030902;EXP030903
- **id**: 372
- **peak-caller.count**: 2
- **peak-caller.list**: SISSRS;GEM
- **peak-caller.set**: GEM;SISSRS
- **peak.count**: 3
- **summit**: 41

Display per TF workflow results

Organism	Human (Homo sapiens)
Transcription factor	1.1.1.1.1 c-Jun

Show





Search | Info | Default

Site ID: 3
Type: TF binding site
Sequence name: 1
Sequence: ATCATAATGGCTATAGCAAT
Position: 629931 - 629950 (20)
Properties:

- **antibody.set**: c-Jun;Jun;c-Jun Antibody (H-79) from Santa Cruz Biotechnology;JUN;c-Jun (H-79, Santa Cruz);...
- **cell.set**: BT549;K562;MCF-7;LoVo;MDA-MB-231
- **exp.set**: EXP010452;EXP033442;EXP000309;EXP035406;EXP033444;EXP030024;EXP033443;EXP000622;EXP000621
- **id**: 18
- **peak-caller.count**: 4
- **peak-caller.list**: GEM;MACS;PICS;SISSRS
- **peak-caller.set**: GEM;MACS;PICS;SISSRS
- **peak.count**: 26
- **summit**: 10
- **treatment.set**: ;dexametasone;E2;compaund A;None

Find transcription factor binding sites on gene

Advanced search

Binding sites near the specified gene

Organism	Human (Homo sapiens)
Gene symbol or ID	Any
Transcription factor	Any
Data set	meta clusters
Max gene distance	5000
Output type	Open in genome browser

Run

Genes regulated by the specified transcription factor

Organism	Human (Homo sapiens)
Transcription factor	1.1.1.1.1 c-Jun
Data set	meta clusters
Max gene distance	5000

Run

Advanced search

Binding sites near the specified gene

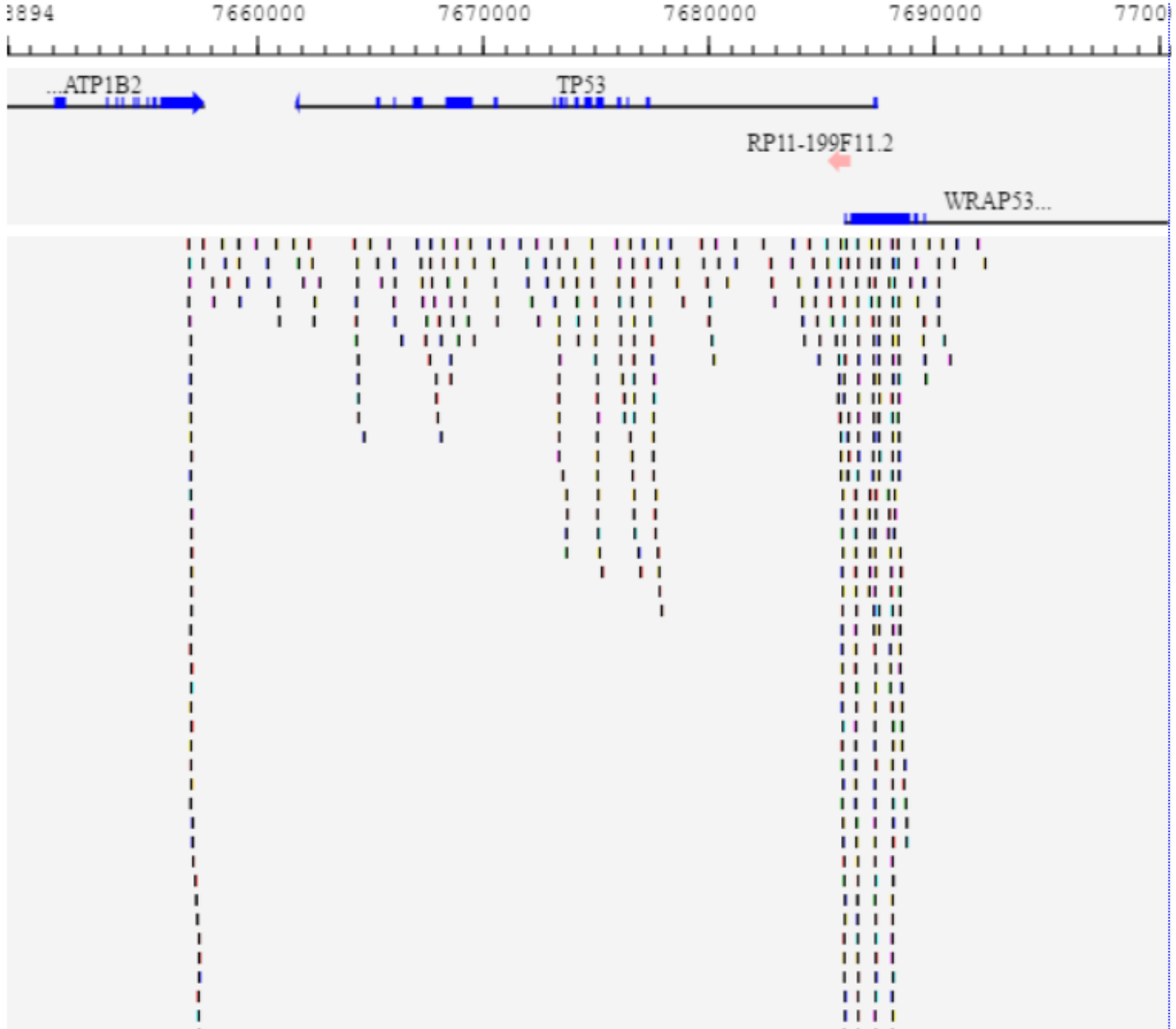
Organism	Human (Homo sapiens)
Gene symbol or ID	tp53
Transcription factor	Any
Data set	meta clusters
Max gene distance	5000
Output type	Open in genome browser

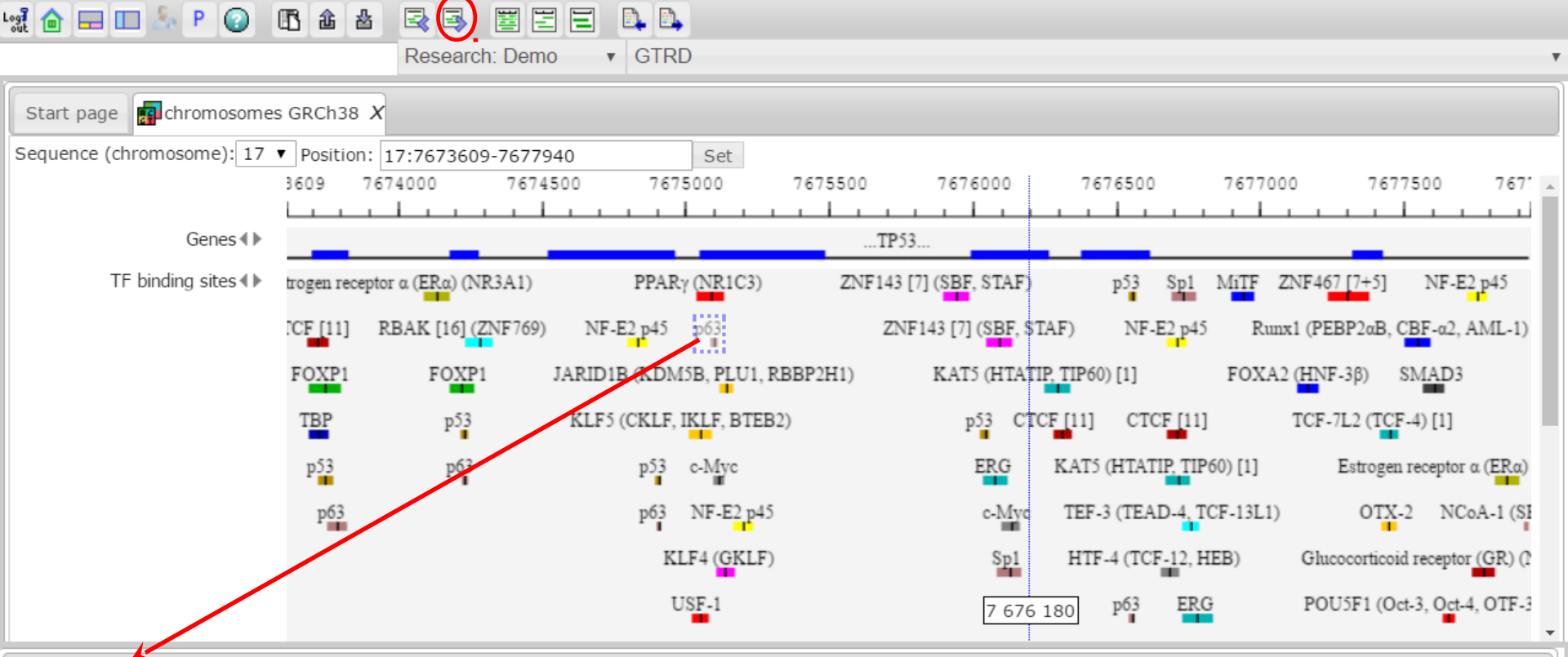
Run

48%

Start page | chromosomes GRCh38 X

Sequence (chromosome): 17 Position: 17:7648894-7700436 Set





Site ID: 26358133
Type: p63
Sequence name: 17
Sequence: gcgcctcacaacctccgtcatgtgctgtgactgct
Position: 7675087 - 7675121 (35)
Properties:

- **antibody.set**: p63 (4A4);anti-p63 4A4;anti-p63
- **cell.set**: neonatal keratinocytes;Human neonatal foreskin keratinocytes
- **exp.set**: EXP032995;EXP031101;EXP030433;EXP032996;EXP032997
- **id**: 166426
- **peak-caller.count**: 4
- **peak-caller.list**: GEM;PICS;SISSRS;MACS;GEM;SISSRS
- **peak-caller.set**: GEM;MACS;PICS;SISSRS
- **peak.count**: 14
- **summit**: 17
- **tfClassId**: 6.3.1.0.2
- **tfTitle**: p63
- **treatment.set**: ;25uM Cisplatin;progenitor;none;350nM Adrimycin

Please cite:
HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models

Show more

Primary URL Mirror

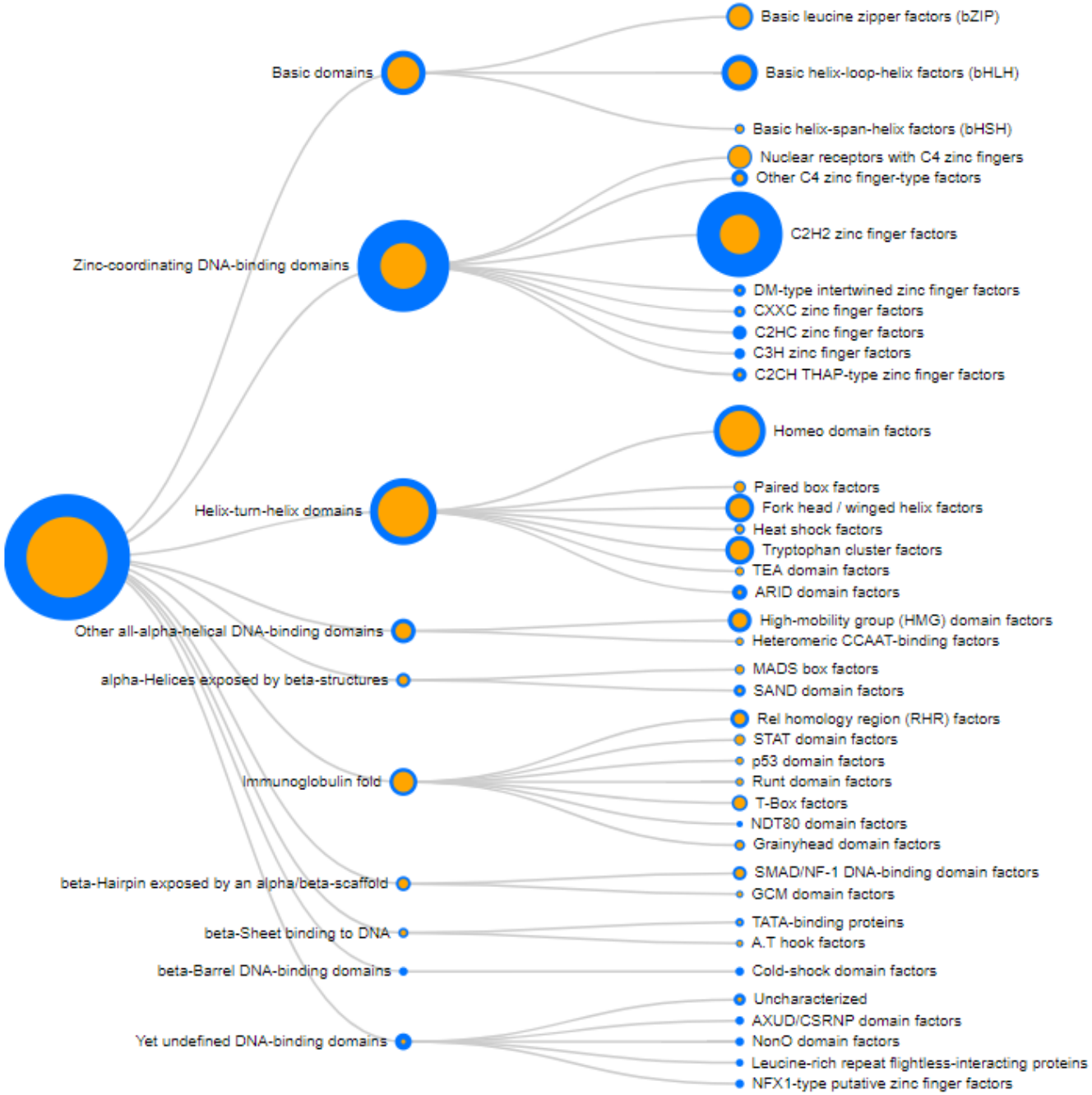
HOmo sapiens COmprehensive MOdel COllection (HOCOMOCO) v11 provides transcription factor (TF) binding models for 680 human and 453 mouse TFs.

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.



BioUML 2016.4 web editi

gtrd.biouml.org/bioumlweb/#

Research: Demo | HOCOMOCO

HOCOMOCO

- HOCOMOCO
  - Matrices and profiles
    - PCM_HUMAN_mono
    - PCM_MOUSE_mono
    - PWM_HUMAN_mono_
    - PWM_HUMAN_mono_
    - PWM_HUMAN_mono_
    - PWM_MOUSE_mono_
    - PWM_MOUSE_mono_
    - PWM_MOUSE_mono_
    - factors
  - Public datasets
  - User data

Start page

HOCOMOCO

HOmo sapiens **CO**mprehensive **MO**del **CO**llection

HOCOMOCO in BioUML

- **HO**mo sapiens **CO**mprehensive **MO**del **CO**llection (HOCOMOCO) is a collection of transcription factor (TF) binding models.
  From this page you can use BioUML tools and HOCOMOCO collection to search for binding sites in DNA sequences.

- Import data
  Load your bed or fasta file

- Site search
  Search for binding sites in your track or sequences

Track	...TP53_gene_promoter_-1000+100
Sequences source	Ensembl 85.38 Human
Profile	...iles/PWM_HUMAN_mono_pval=0.0001 Auto
Output name	HOCOMOCO/User data/Site search result Auto

Run

- View/Export results
  View and export site search results

Genome track	HOCOMOCO/User data/Site search result

View in genome browser | View as table | Export

GTRD further development

1) processing ChIP-seq data for new species

current

in process

Top Organisms [Tree] GEO

- Homo sapiens *(5231)*
- Mus musculus *(3935)*
- Drosophila melanogaster *(694)*
- Caenorhabditis elegans *(450)*
- Saccharomyces cerevisiae *(262)*
- Arabidopsis thaliana *(204)*
- Schizosaccharomyces pombe *(92)*
- Rattus norvegicus *(74)*
- Danio rerio *(58)*
- Escherichia coli *(46)*
- Gallus gallus *(34)*
- Oryza sativa *(31)*
- Zea mays *(23)*
- Macaca mulatta *(16)*
- Caulobacter vibrioides *(16)*
- Xenopus laevis *(15)*
- Xenopus tropicalis *(15)*
- Drosophila simulans *(15)*
- Plasmodium falciparum *(15)*
- Drosophila pseudoobscura *(13)*

Less...

GTRD further development

1) processing ChIP-seq data on TFBS for new species

2) processing ChIP-seq data for:
   - co-repressors and co-activators
   - histone modifications

3) processing data on chromatin availability
   - DNase-seq
   - ATAC-seq

Gene transcription regulation
grand challenges

Volume 16 Number 5 1988 Nucleic Acids Research

Compilation of transcription regulating proteins

Edgar Wingender

Gesellschaft für Biotechnologische Forschung mbH, Mascheroder Weg 1, D-3300 Braunschweig, FRG

Received November 28, 1987; Revised and Accepted January 28, 1988

Introduction

As gene regulation is one of the central topics of molecular biology efforts have been made to define the regulating elements. On DNA level, cis-acting sequences have successfully been defined for a large number of genes. Promoter, enhancer, and regulating (or responsive) elements have been determined which govern constitutive gene expression on a basal and often low level, which enhance or repress transcription of the respective gene in a cell- or stage-specific manner or which make the gene responsive to external trigger signals, e.g. to hormones or metal ions.

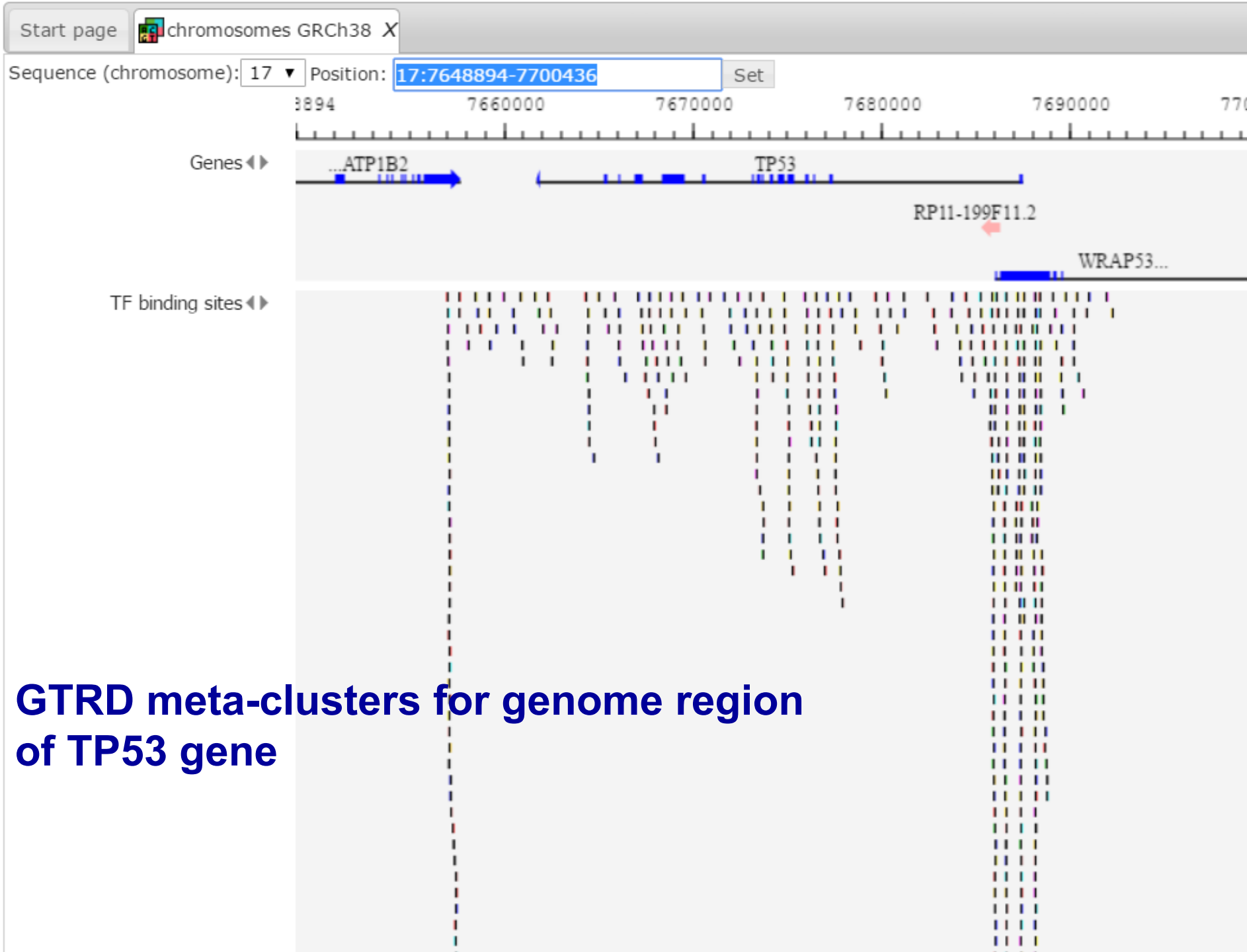
In the last years, it became evident that these sequences exert their influences by interaction with specific proteins, e.g. general or specific transcription factors or steroid hormone receptors. Accordingly, the number of reports increased dramatically showing the occupation of "regulating" DNA sequences by these trans-acting factors.

The aim of the following compilation is (i) to give an survey of the genes for which regulatory and protein-interacting elements are known and to localize these regions (Tab. 1); (ii) to assign to these elements the factors by which they are recognized (Tab. 1); (iii) to list the regulating factors, their target genes, some of their molecular properties and corresponding proteins of (presumably) similar function (Tab. 2); (iv) to compare the DNA-binding domains of those regulating proteins, which hypothetically possess a finger structure (Tab. 3).

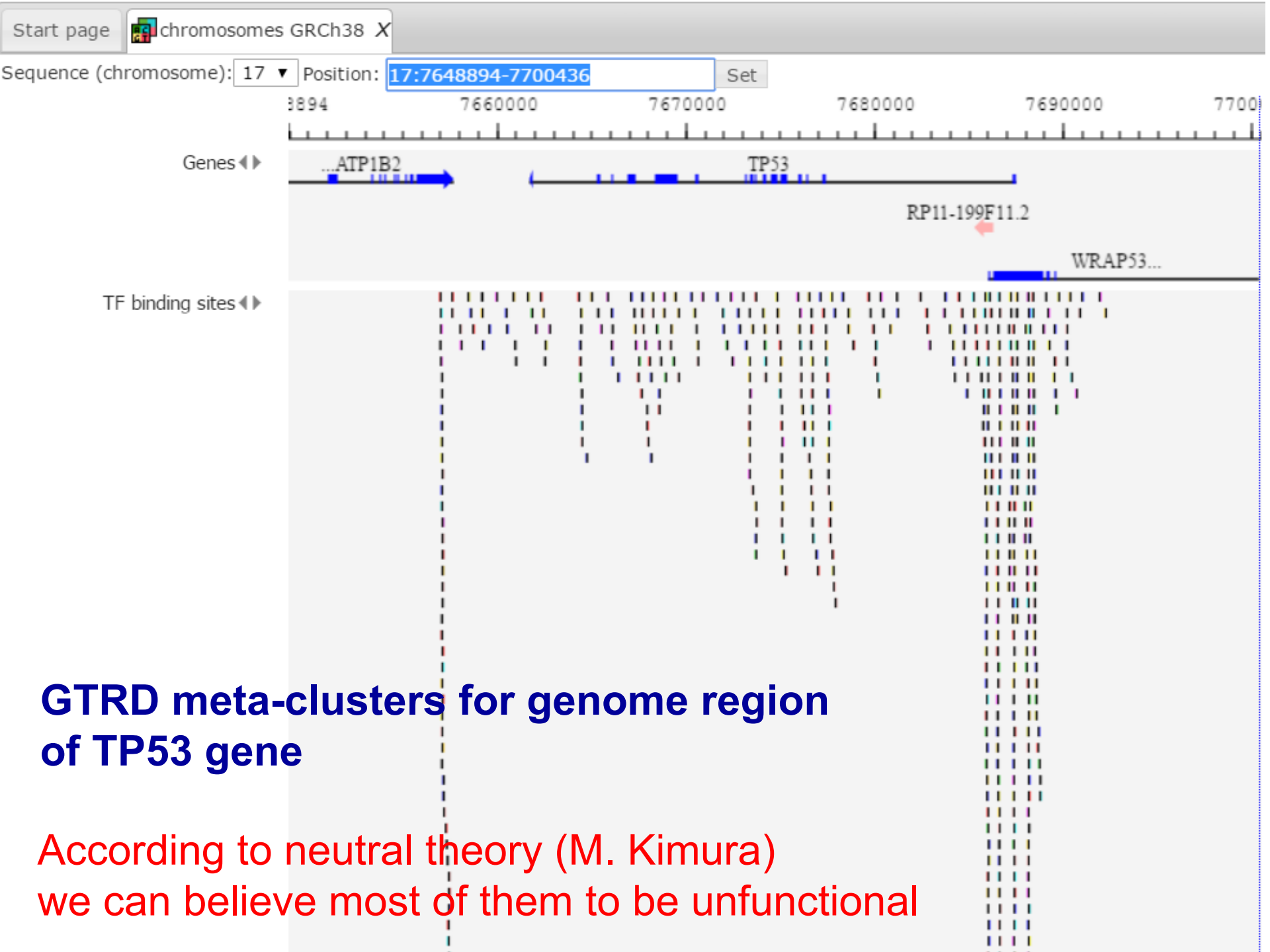
This listing might provide a basis to systematize the puzzle of transcrip-

The aim of the following compilation is (i) to give an survey of the genes for which regulatory and protein-interacting elements are known and to localize these regions (Tab. 1); (ii) to assign to these elements the factors by which they are recognized (Tab. 1); (iii) to list the regulating factors, their target genes, some of their molecular properties and corresponding proteins of (presumably) similar function (Tab. 2); (iv) to compare the DNA-binding domains of those regulating proteins, which hypothetically possess a finger structure (Tab. 3).

This listing might provide a basis to systematize the puzzle of transcription factors, particularly those for the genes which are transcribed by eukaryotic RNA polymerase II. Accordingly, only genes transcribed by this enzyme are included in Table 1. Depending on which term is more commonly used, either the genes or the gene products are listed in alphabetical order.



GTRD meta-clusters for genome region of TP53 gene

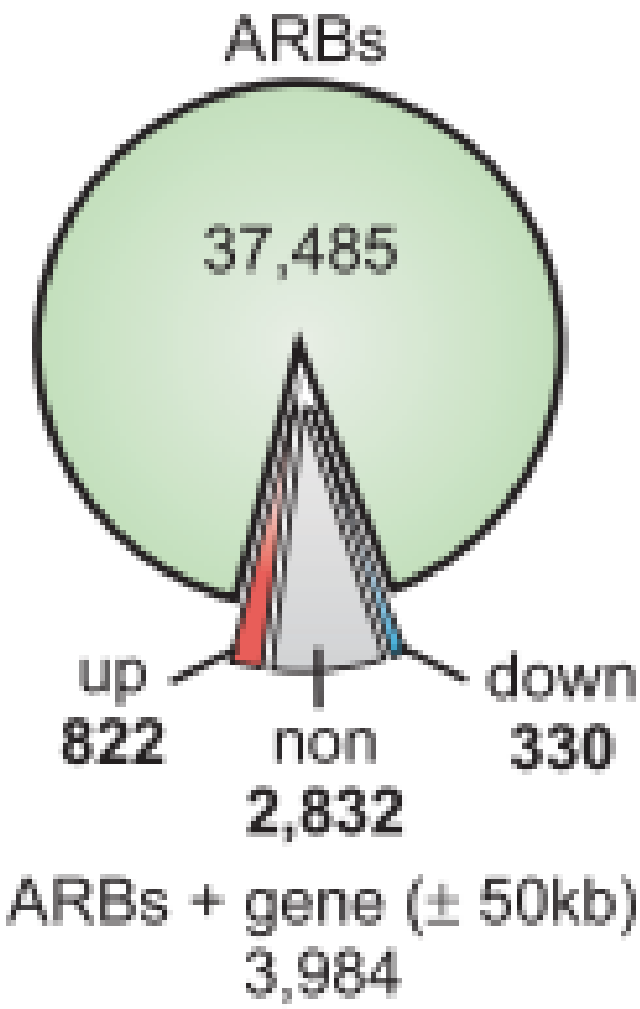


GTRD meta-clusters for genome region of TP53 gene

According to neutral theory (M. Kimura)
we can believe most of them to be unfunctional

Global analysis of transcription in castration-resistant prostate cancer cells uncovers active enhancers and direct androgen receptor targets

Toropainen et al., Scientific Reports, 2016, 6:33510



AR plays an important role in the development of prostate cancer (PC), and changes in androgen signaling are thought to critically contribute to the development of castrate resistant prostate cancer (CRPC).

However, our understanding of the gene programs that are directly targeted by the AR in CRPC cells is still limited. One of the challenges in deciphering these gene programs is to identify functionally active AR-binding sites from the vast number of AR-binding sites on chromatin.

…

According to our stringent analyses, only <3% of the intergenic and intragenic ARBs qualifed as androgen-regulated eRNA-producing enhancers.


SUDHA
TATA
RJ15 P0251


SUDHA
TATA
RJ15 P0251
HDAC3
CoR
PIAS1
FOXA1
AR
AR
ERG
eRNA

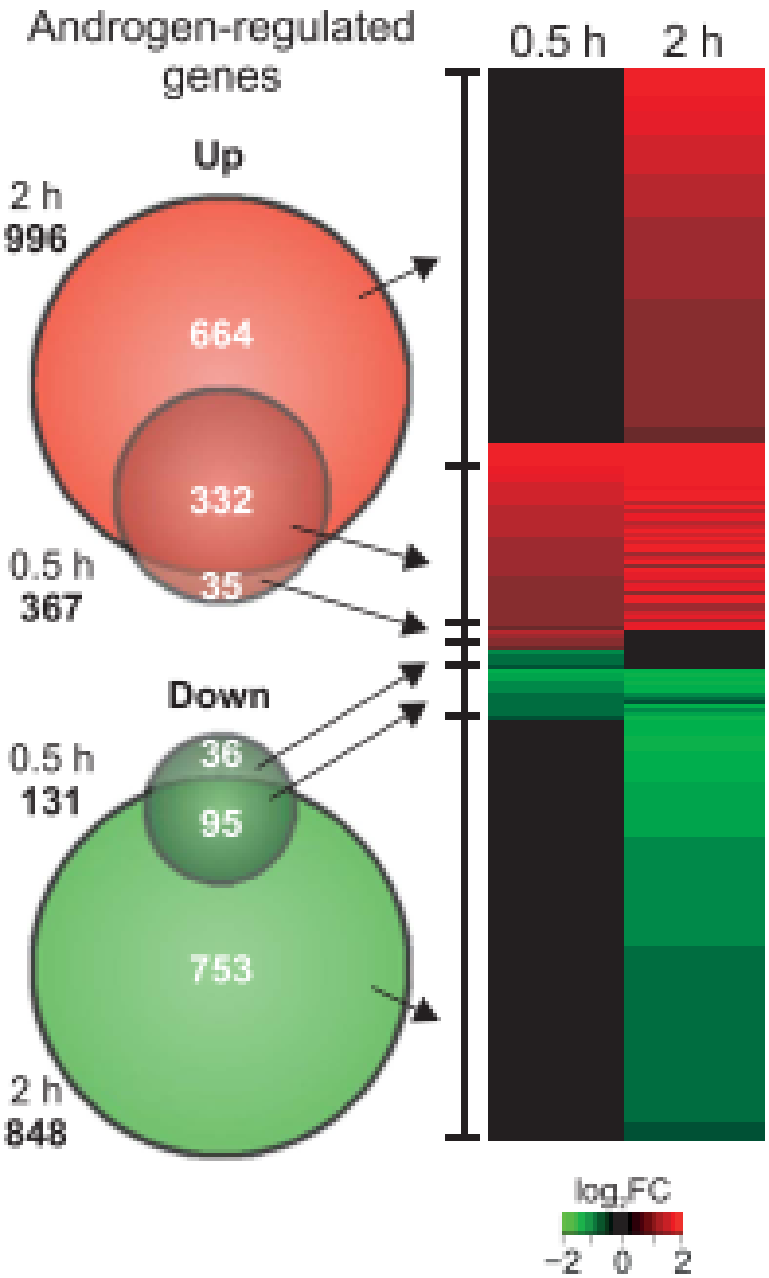

SUDHA
TATA
RJ15 P0251
HDAC3
CoR
PIAS1
FOXA1
AR
AR
ERG
eRNA

Grand challenge 1

What transcription factor binding sites are **functional?**

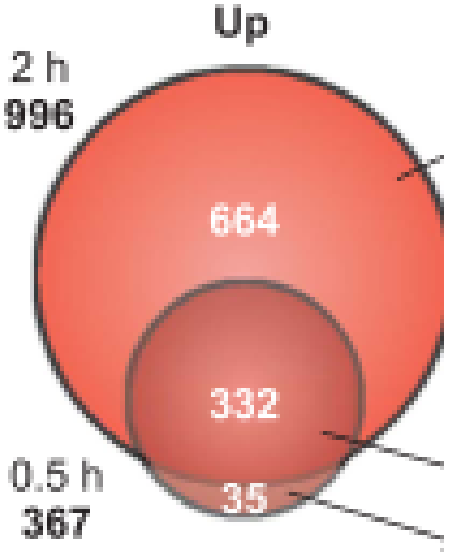
How to predict effect of variation (SNV, deletion) in functional transcription factor binding site on expression of the specified gene?

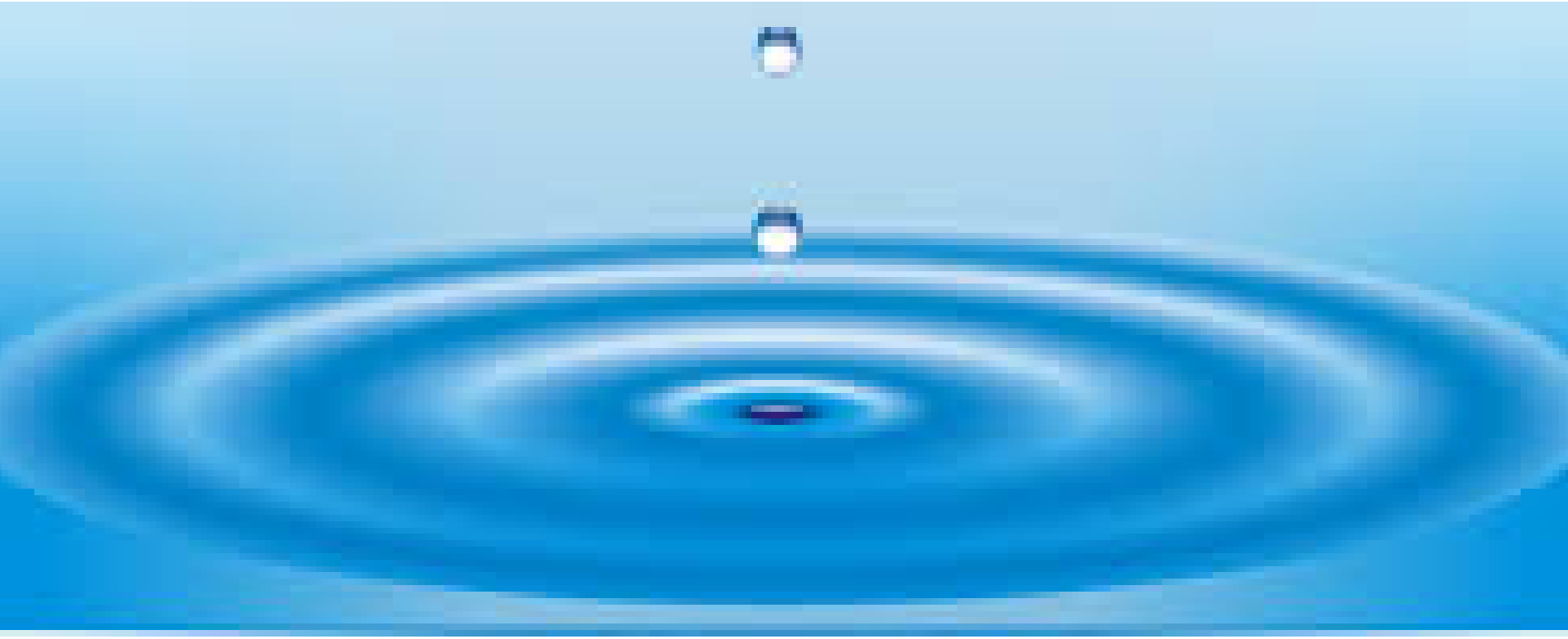
In which conditions (cell line or tissue, development stage, treatment, etc.) specified transcription factor binding site is functional?



Global analysis of transcription in castration-resistant prostate cancer cells uncovers active enhancers and direct androgen receptor targets

Toropainen et al., Scientific Reports, 2016, 6:33510

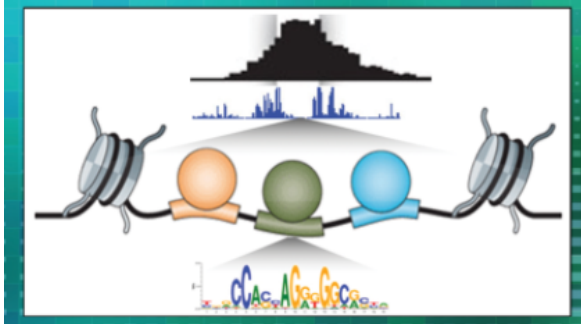

Up
2 h
996
664
332
0.5 h
367
35



Grand challenge 2

When and which genes will be up/down regulated by action of specified TF?

ENCODE-DREAM *in vivo* Transcription Factor Binding Site Prediction Challenge



DREAM CHALLENGES powered by Sage Bionetworks · Sage BIONETWORKS · Stanford University · OHSU · IBM Research · HelmholtzZentrum münchen Deutsches Forschungszentrum für Gesundheit und Umwelt · School of Medicine Department of Pharmacology University of Colorado Anschutz Medical Campus

2 - Challenge Overview

Scientific Rationale

Transcription factors (TFs) are regulatory proteins that bind specific DNA sequence patterns (motifs) in the genome and affect transcription rates of target genes. Binding sites of TFs differ across cell types and experimental conditions. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is an experimental method that is commonly used to obtain the genome-wide binding profile of a TF of interest in a specific cell type/condition. However, profiling the binding landscape of every TF in every cell type/condition is infeasible due to constraints on cost, material and effort. Hence, accurate computational prediction of *in vivo* TF binding sites is critical to complement experimental results.

Top-performing Teams

J-TEAM

- Jens Keilwagen, Julius Kühn-Institut (JKI) - Federal Research Centre for Cultivated Plants, Quedlinburg, Germany
- Stefan Posch, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany
- Jan Grau, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

Yuanfang Guan

- Yuanfang Guan, University of Michigan, Ann Arbor, MI, USA

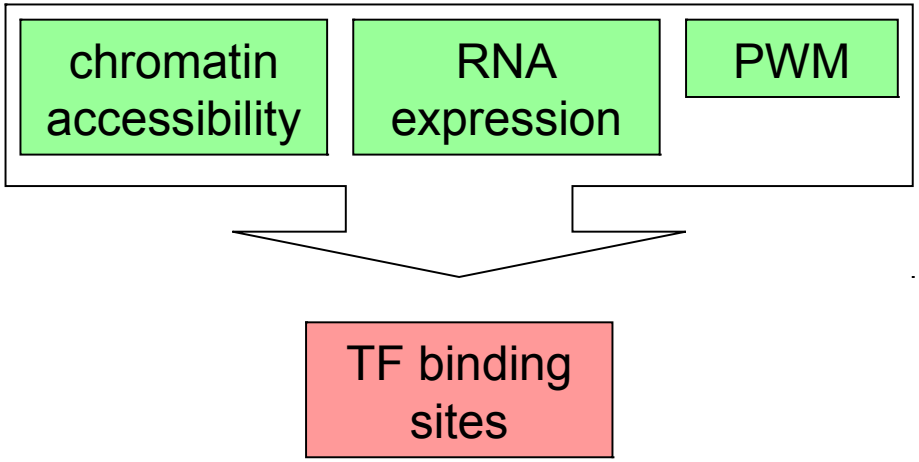
dxquang

- Daniel Quang, University of California, Irvine, CA, USA
- Xiaohui Xie, University of California, Irvine, CA, USA

automosome.ru

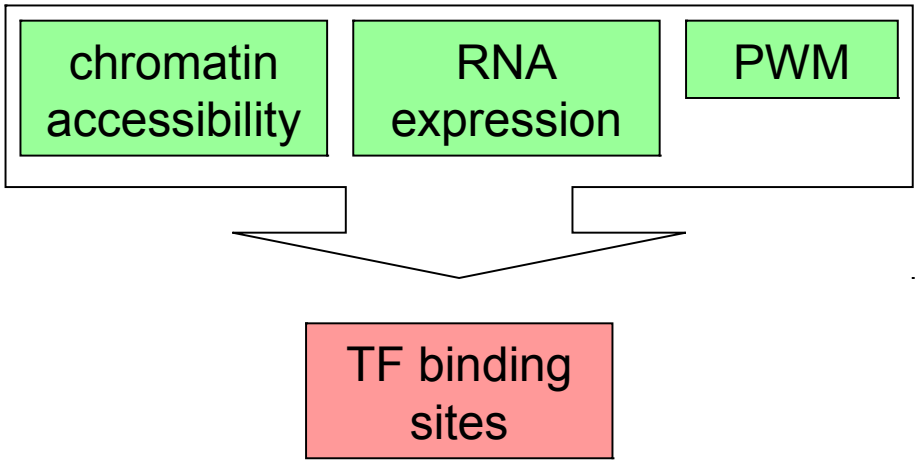
- Andrey Lando, Moscow Institute of Physics and Technology, Dolgoprudny, Russia
- Ilya Vorontsov, Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia
- Valentina Boeva, Institut Cochin, Paris, France
- Grigory Sapunov, Intento, https://inten.to
- Irina Eliseeva, Institute of Protein Research, Russian Academy of Sciences, Pushchino, Russia
- Vsevolod Makeev, Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia
- Ivan Kulakovskiy, Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia

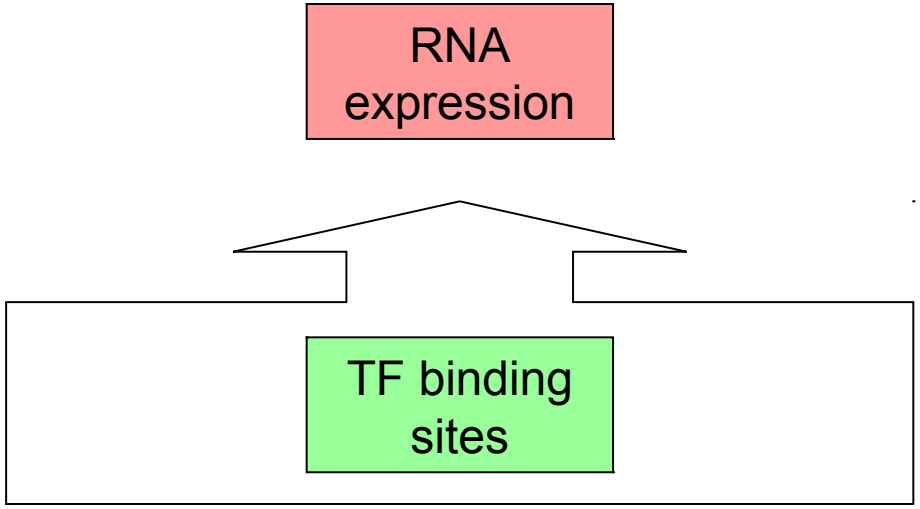
ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge



Grand challenge 3

ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge





Linking GTRD with FANTOM5 via special cell lines dictionary

Cell Line	Number of tracks	Number of distinct TF-classes	Number of FANTOM5 samples
Lovo	390	381	-----
HepG2	209	144	3
K562	229	117	54
HEK293	129	115	2
GM12878	123	73	3
MCF7	357	63	93
H1	66	40	9
A549	160	39	1
HeLa S3	50	36	3
HCT-116	76	25	------

Predicting gene expression (CAGE) for HepG2 cell line using GTRD tracks

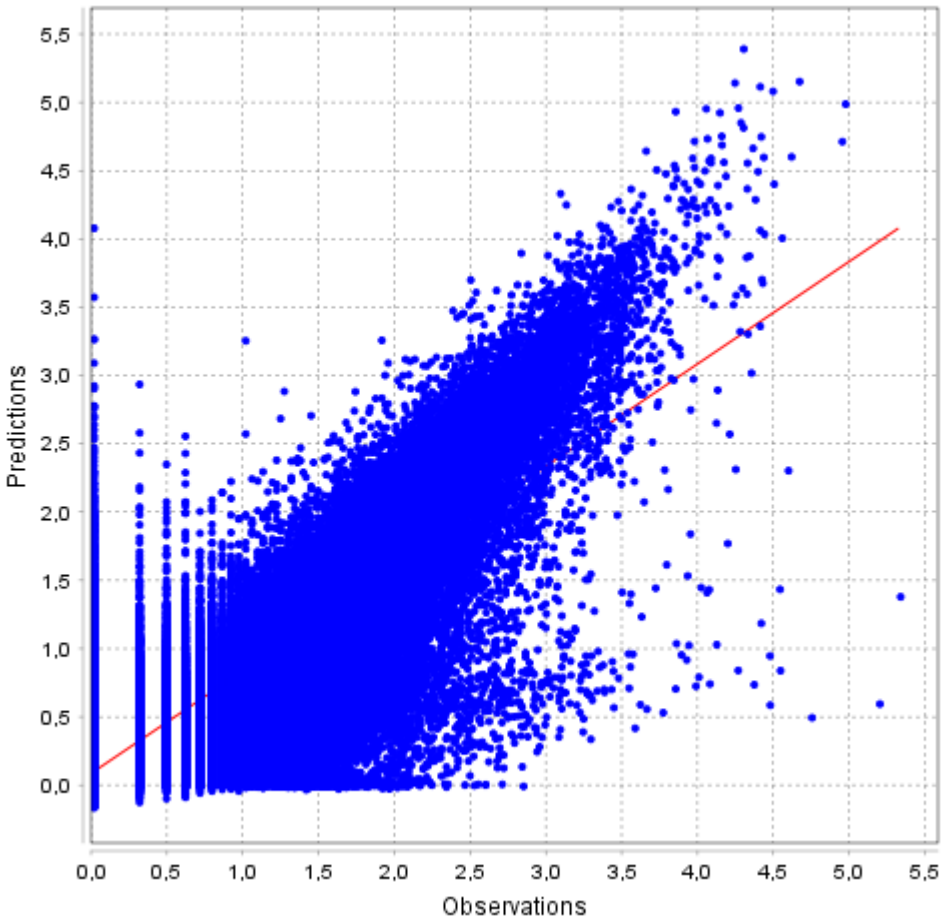
Cross-validation of Regression Models

Regression	Correlation (training set)	Correlation (test set)	Explained variance (training set)	Explained variance (test)
Least Square Regression (32 selected features)	0.866	0.862	74.9%	74.3%
Regression On Principle Components (all features)	0.829	0.712	68.7%	50.5%
Random Forest (all features)	0.934	0.707	85.3%	49.8%

Predicting gene expression (CAGE) for HepG2 cell line using GTRD tracks

Best model : Least Square Regression, 23 selected features

Pearson correlation = 0.866; Explained variance = 74.945



Predicting gene expression (CAGE) for HepG2 cell line using GTRD tracks

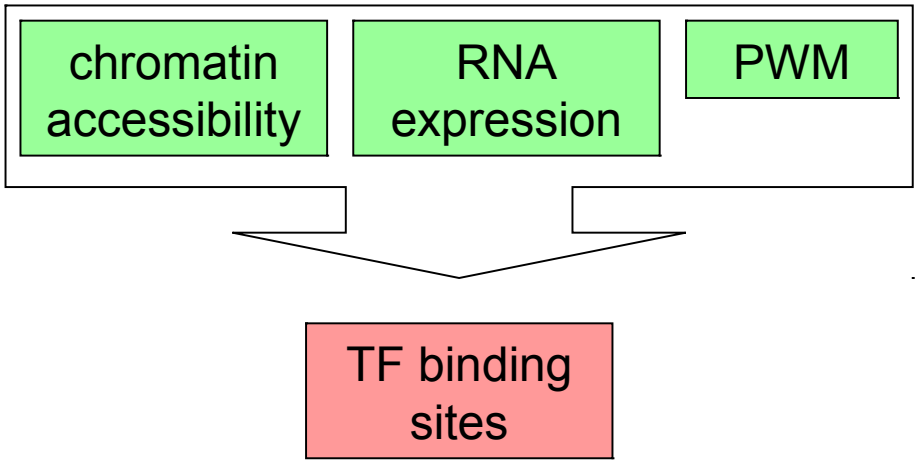
Least Square Regression: the most significant features

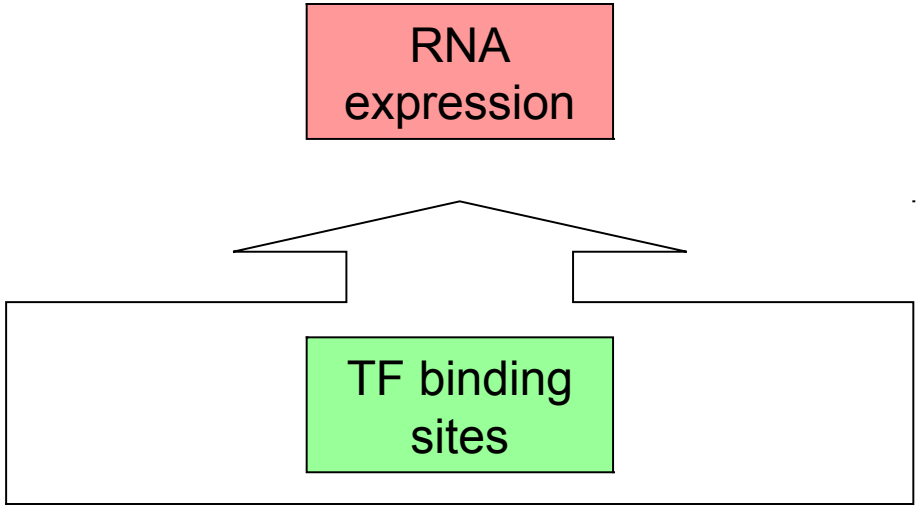
Pearson correlation = 0.866; Explained variance = 74.945;

TF	Promoter region	Coefficient	p-value
HNF-4a (2.1.3.2.1)	[-1, +1]	0.368	< 1.2E-255
TAF-1 (4.1.3.0.5)			

Grand challenge 3

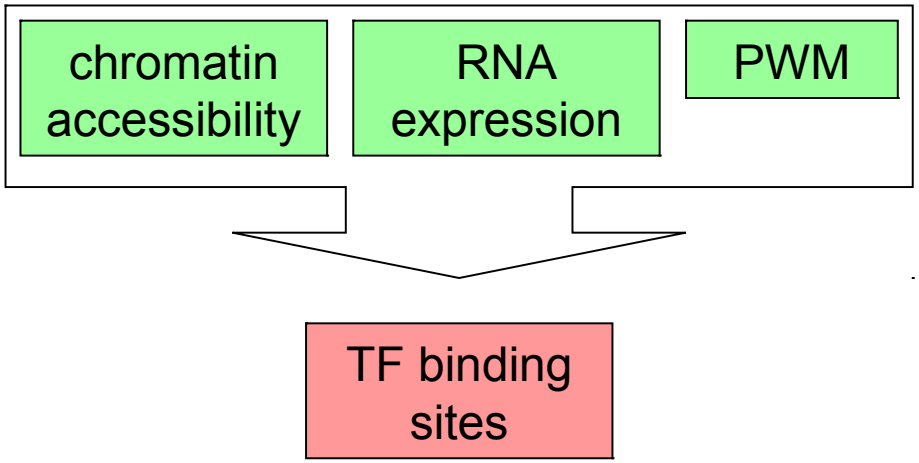
ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge

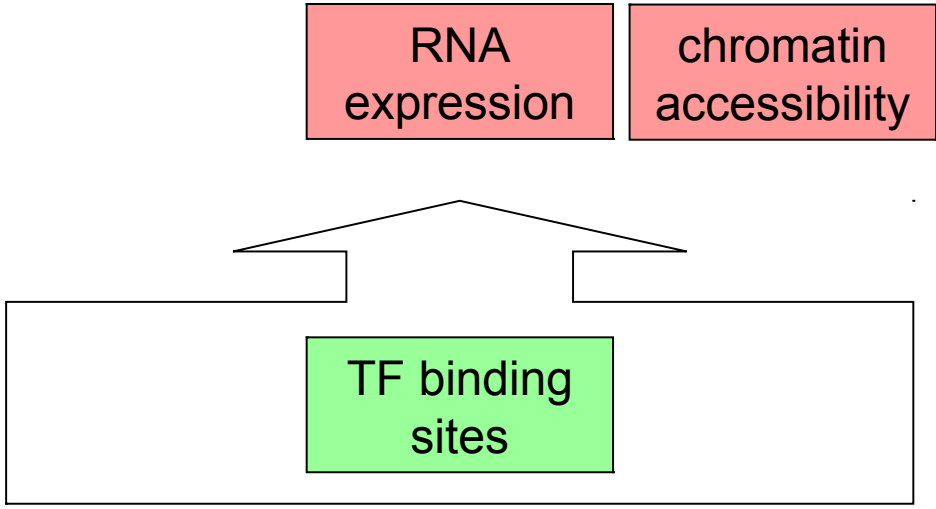




ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge

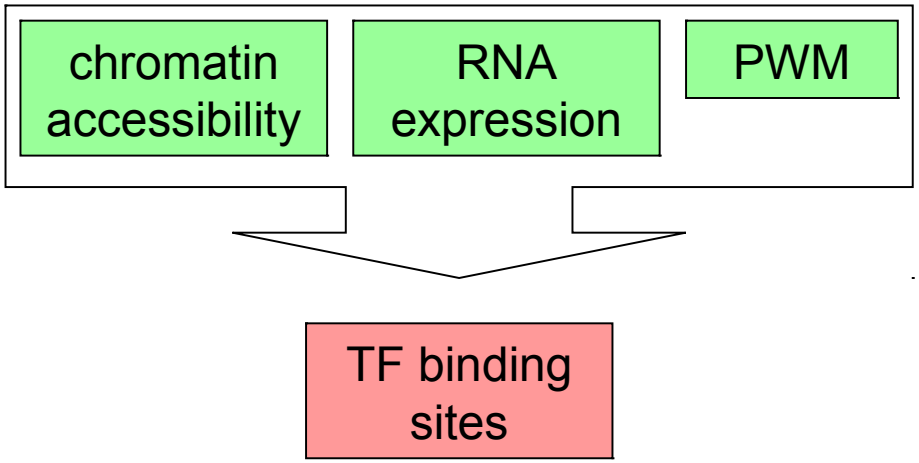
Grand challenge 3

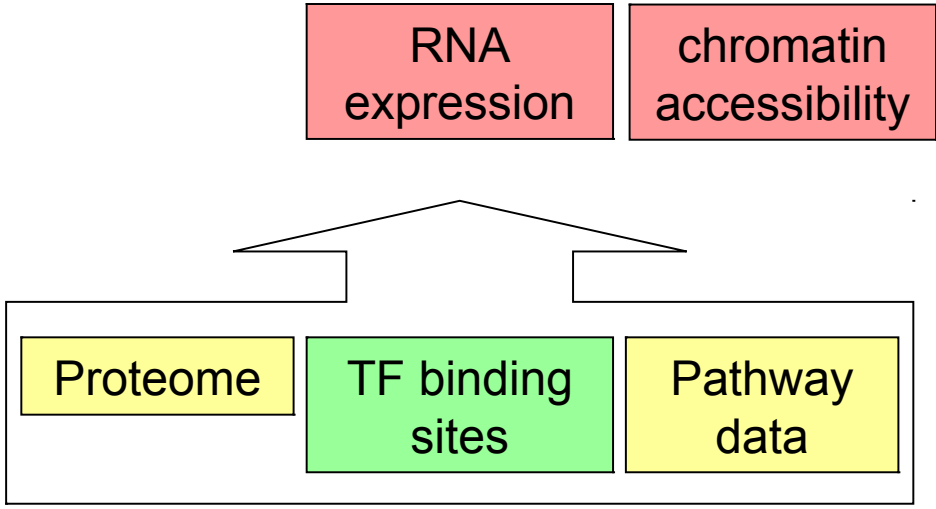




ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge

Grand challenge 3





in dynamics

Gene transcription regulation grand challenges

1. What transcription factor binding sites are functional?

2. When and which genes will be up/down regulated by action of specified TF?

3. How to predict gene expression on the base of TFBS and related pathway data in dynamics?