
JASPAR: a comprehensive database of transcription factor binding profiles

Dr. Oriol Forner

Deputy Group Leader @ Wasserman Lab
Centre for Molecular Medicine and Therapeutics
BC Children's Hospital Research Institute
University of British Columbia

✉ oriol@cmmt.ubc.ca
🐦 @OForner

GREEKC meeting – February 13, 2018



Good morning to everyone. First of all, I want to thank the organization for giving me the opportunity to come to Ljubljana. Today, I will be talking about the latest release of JASPAR for 2018, which has been recently published in the NAR database issue, and describe its new and improved features and tools.

Outline

- **Overview**
- **Manual Curation**
- **Tools:**
 - **Profile Inference**
 - **Matrix Clustering**
 - **Genome Tracks**
- **New Web**
- **RESTful API**
- **Perspectives**
- **MANTA2**

2

Here is an outline of my presentation. I will start with a brief overview of JASPAR, followed by the curation process that we perform before adding/updating any new binding profiles to the database. Then I will describe you some of the most recent tools that we have developed. I will also describe some of the improved new web features, the RESTful API, which grants universal programmatic access to JASPAR, and finally some future perspectives that we are exploring, and, if I have time, our recently updated MANTA database, which we believe will be a very useful tool for the community. By the way, interruptions are of course welcome.

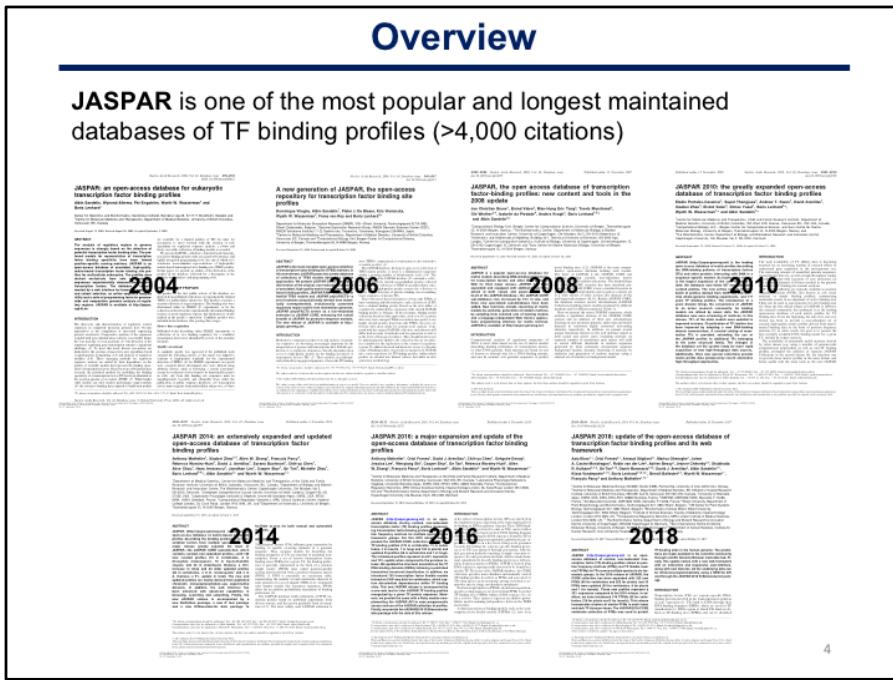
Outline

- **Overview**
 - Manual Curation
 - Tools:
 - Profile Inference
 - Matrix Clustering
 - Genome Tracks
 - New Web
 - RESTful API
 - Perspectives
 - MANTA2

3

Overview

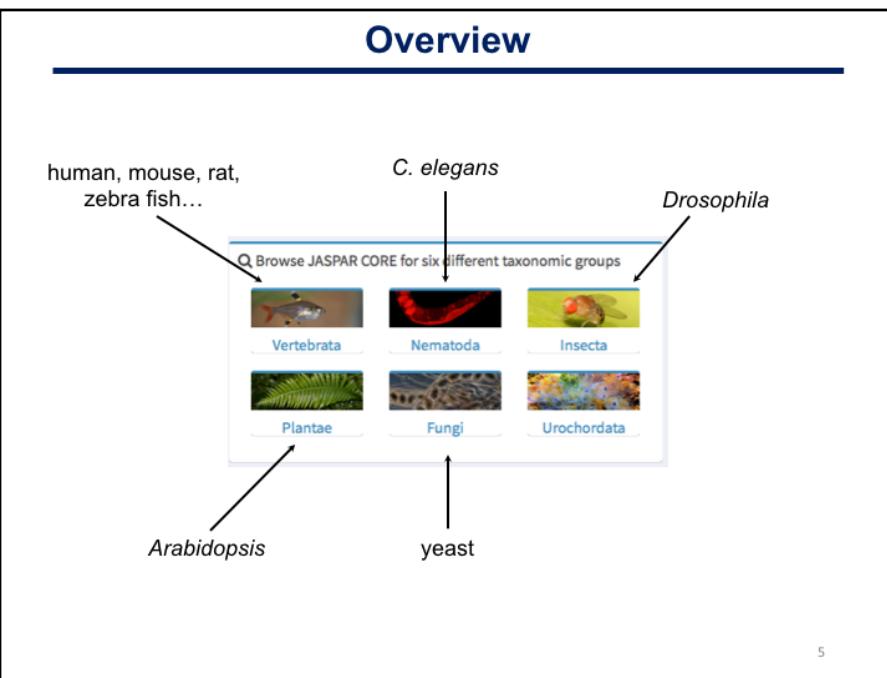
JASPAR is one of the most popular and longest maintained databases of TF binding profiles (>4,000 citations)



4

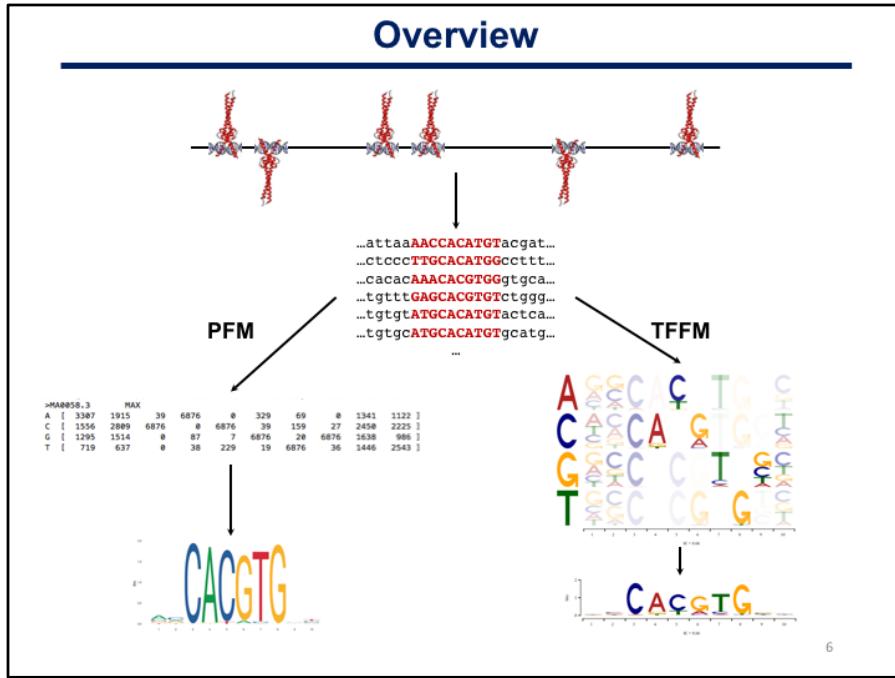
So JASPAR is one of the most popular and longest maintained databases of TF binding profiles. It was initially published in 2004 and has undergone 7 updates.

Overview



5

It comprises TF binding profiles from species from 6 different taxonomic groups. For example, for vertebrates, we include human, mouse, rat, and zebra fish profiles, among others, while for nematodes, insects, plants and fungi, profiles are mainly from *C. elegans*, *Drosophila*, *Arabidopsis* and yeast, respectively. We only have one profile for urocordates.



6

JASPAR includes two different kinds of profiles. For all TFs, it includes position frequency matrices, PFM, while only for some, binding sites are also described as TF flexible models. These more advanced models, are based on hidden-Markov models and account for both dependencies between nucleotide positions and binding sites of flexible length.

www.cisreg.ca/TFFM/doc/

TFFM documentation »

Table Of Contents

- Welcome to TFFM's documentation!
- System requirements
- Contents
- Tutorial
- Download
- Reference
- Licence
- Indices and tables

Next topic

[tffm Module](#)

This Page

[Show Source](#)

Quick search

Go

Enter search terms or a module, class or function name.

Welcome to TFFM's documentation!

We provide here the documentation of the TFFM-framework developed in Python. The **Transcription Factor Flexible Models (TFFMs)** represent TFBSs and are based on hidden Markov models (HMM). They are flexible and are able to model both position interdependence within TFBSs and variable length motifs within a single dedicated framework.

The framework also implements methods to generate a new graphical representation of the modeled motifs that convey properties of position interdependences.

TFFMs have been assessed on ChIP-seq data sets coming from the ENCODE project, revealing that the new HMM-based framework performs, in most cases, better than both PWMs and the dinucleotide weight matrix (DWM) extension in discriminating motifs within ChIP-seq sequences from background sequences. Under the assumption that ChIP-seq signal values are correlated with the affinity of the TF-DNA binding, we find that TFFM scores correlate with ChIP-seq peak signals. Moreover, using available TF-DNA affinity measurements for the Max TF, we observe that TFFMs constructed from ChIP-seq data correlate with published experimentally measured DNA-binding affinities. These results demonstrate the capacity of TFFMs to accurately model DNA-protein interactions, while providing a single unified framework suitable for the next generation of TFBS predictions. All the details have been published in [Mathelier and Wasserman, The Next Generation of Transcription Binding Site Prediction, PLOS Computational Biology](#), Sept. 2013, 9(9):e1003214, DOI:10.1371/journal.pcbi.1003214.

TFFMs can be saved and opened from files using the XML format already used by the GHMM library.

System requirements

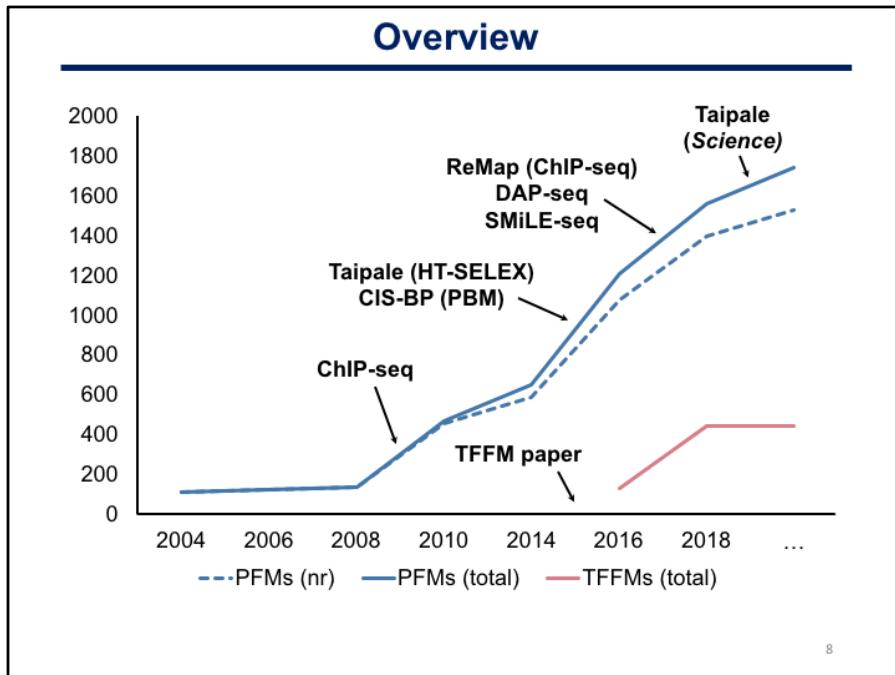
- The TFFM-framework 2.0 has been developed and tested under Ubuntu Linux operating system. It has also been tested on CentOS.
- Python should be installed (version 2.7 has been used successfully).
- Biopython (at least version 1.61) should be installed and accessible from your Python executable. See <http://biopython.org> for instructions on how to install it.
- The GHMM library should be installed and accessible from Python. See <http://ghmm.org> for instructions on how to install it.

Contents

- [tffm Module](#)
- [hit Module](#)
- [drawing Module](#)
- [util Module](#)

7

I am not going to describe the TFFMs in detail, but for more information you can check the documentation provided at this website.



8

JASPAR owes its growth to different technological advances that have occurred for the years. For example, the growth between the versions 2008 and 2010 is owed to the emergence of the ChIP-seq technology, while the expansion of the last release is mainly thanks to the ReMap of ChIP-seq data from ENCODE, GEO and ArrayExpress, and the emergence of DAP-seq (for plants) and SMiLE-seq. As you can also see, TFFMs were only included to JASPAR during the previous release, as it's a recent method that was only published in 2014. Finally, we have already started processing new data from the methylation paper recently published in Science by the Taipale group. So far, we are only including profiles on non-methylated DNA.

Overview

The JASPAR database as of 2018:

Taxon	PFMs (nr)	PFMs (total)	TFFMs (total)	TF classes	TF families
Vertebrates	579	719	225	38	95
Plants	489	501	218	22	24
Insects	133	140	3	18	25
Nematodes	26	26	0	10	14
Fungi	176	177	0	19	11
Urochordata	1	1	0	1	0
Total	1,404	1,564	446	64	115

Experimental data sources: ChIP-chip/seq, DAP-seq, EMSA, PBM, (HT-)SELEX, SMiLE-seq...

9

All in all, as of 2018, JASPAR contains over 14 hundred non redundant PFMs, and almost 500 TFFMs, with TFs belonging to 64 different structural classes and 115 families.

Outline

- Overview
- **Manual Curation**
- Tools:
 - Profile Inference
 - Matrix Clustering
 - Genome Tracks
- New Web
- RESTful API
- Perspectives
- MANTA2

10

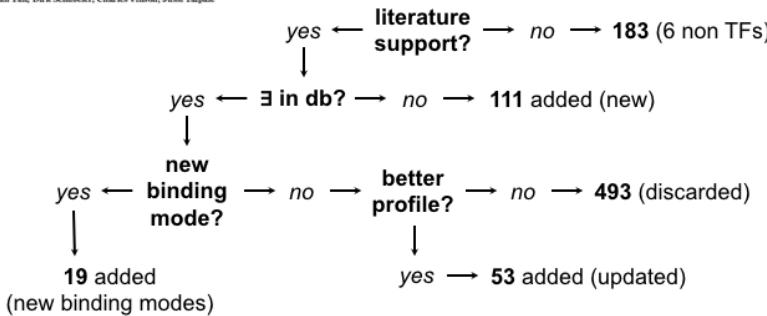
Now, I will describe the manual curation of profiles.

Manual Curation

Impact of cytosine methylation on DNA binding specificities of human transcription factors

Vincent Yin, Ekaterina Moryanova, Arttu Jolma, Eerik Kaasinen, Rishabhayoti Saha, Syed Khidir-Sayed, Pratyush K. Das, Teremu Khirolja, Kashyap Dave, Fan Zhong, Kazuhiko R. Niita, Mianna Taipale, Alexander Popov, Paul A. Glino, Silvia Domcke, Jian Yan, Dirk Schibeler, Charles Vinson, Jussi Taipale*

859 PFMs



11

As an example, I will provide statistics from my curation process on the Taipale's Lab data. We started with 859 profiles with sufficient information content (there were a few whose quality was not good). The first step in the curation process is to check whether a profile has literature support. For 183 profiles, we could not find any. In addition, 6 of the profiles were from non TF proteins.

Manual Curation

Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences

Michael F. Berger,^{1,3,8} Gwenael Badis,^{5,8} Andrew R. Gehrke,^{1,8} Shaheyenor Talukder,^{5,8} Anthony A. Philippakis,^{1,3,6} Lourdes Peña-Castillo,⁴ Trevis M. Alleyne,⁵ Sanie Mnaimneh,⁴ Olga B. Botvinnik,^{1,7} Esther T. Chan,⁵ Faiqua Khalid,⁴ Wen Zhang,⁵ Daniel Newburger,¹ Savina A. Jaeger,¹ Quaid D. Morris,^{4,5} Martha L. Bulyk,^{1,2,3,6,*} and Timothy R. Hughes^{4,5,*}

For 71 of the proteins we analyzed, there is no *in vitro* or *in vivo* binding site data, and for the majority, there is no PWM, in either mouse or the closest homolog in any species. To our knowledge, for several families, we describe a relatively uniform and apparently distinct binding profile for the first time. These encompass the **Irx family** (preferring sequences resembling **TACATGTA**), the **Obox family** (**GGGGATTAA**), the **Six family** (**G(G/A)TATCA**), **Gbx1/2 (CTAATTAG)**, and **Pknox1/2 (CCGTGTC)**. Our data also include individual proteins with apparently unique sequence preferences, including **Dux1 (CAATCAA)**, **Hdx [(C/A)AATCA]**, **Hmbox (TAACTAG)**, **Homez (ATCGTTT)**, and **Rhox11 (GCTGT(T/A)(T/A))**. The variety in motifs we obtained motivated us to further explore the similarities and differences among homeodomains within our data set.

12

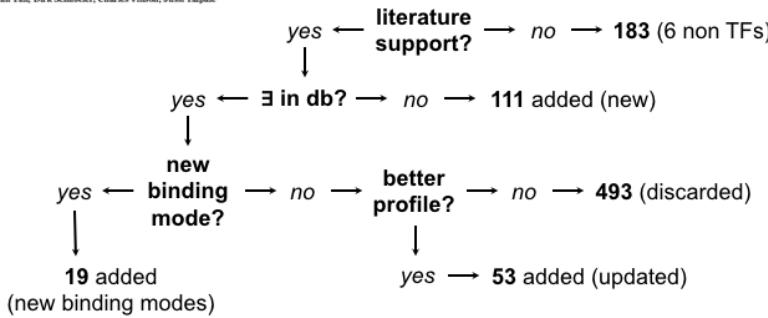
By literature support we mean finding papers that say something like this highlighted text. For instance, the Irx family of TFs prefers this kind of sequence. I am showing you this paper because it has been key to add most homeodomain TFs to JASPAR.

Manual Curation

Impact of cytosine methylation on DNA binding specificities of human transcription factors

Vincent Yin, Ekaterina Moryanova, Arttu Jolma, Eerik Kaasinen, Rishabhayoti Saha, Syed Khidir-Sayed, Pratyush K. Das, Teremu Khirolja, Kashyap Dave, Fan Zhong, Kazuhiko R. Niita, Minala Talpale, Alexander Popov, Paul A. Glino, Silvia Domcke, Jian Yan, Dirk Schibeler, Charles Vinson, Jussi Taipale*

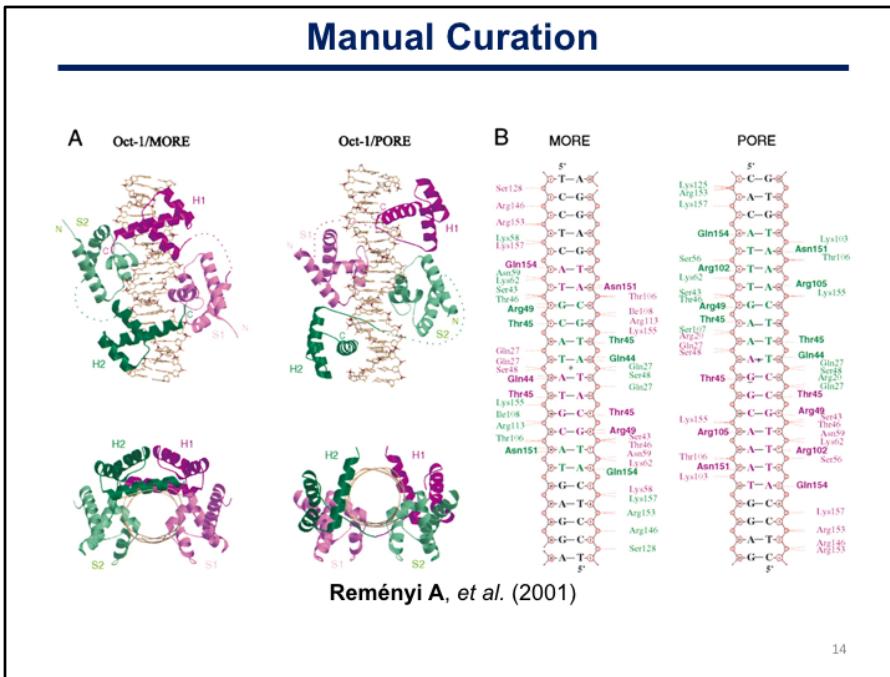
859 PFM



13

The next step is to check if the profile already exists in JASPAR. If not, it is added. If exists, we check if the profile is a new binding mode for the TF.

Manual Curation



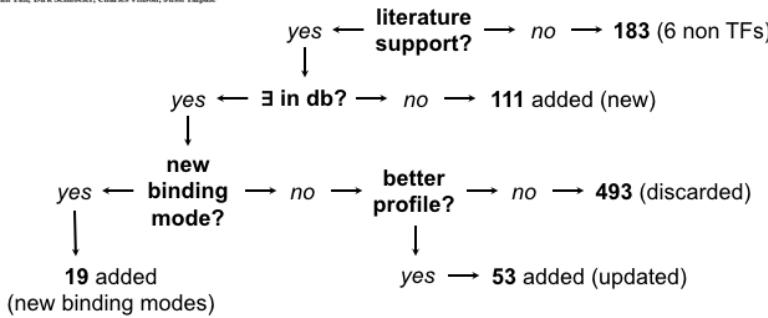
By binding mode we refer to the ability of some TFs, such as Oct-1, to bind DNA in different conformations, resulting in different DNA binding preferences for the MORE and PORE conformations.

Manual Curation

Impact of cytosine methylation on DNA binding specificities of human transcription factors

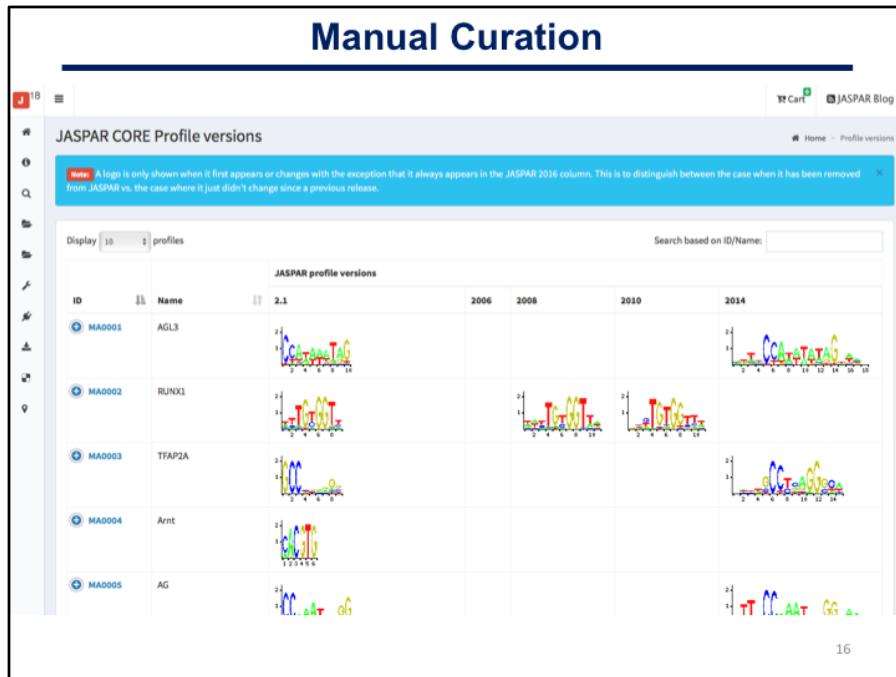
Vincent Yin, Ekaterina Moryanova, Arttu Jolma, Eerik Kaasinen, Rishabh Iyer, Saha, Syed Khadem-Sayeed, Pratyush K. Das, Teremu Khiogja, Kashyap Dave, Fan Zhong, Kazuhiko R. Niita, Mianan Talpale, Alexander Popov, Paul A. Glino, Silvia Domcke, Jian Yan, Dirk Schibeler, Charles Vinson, Jussi Talpale*

859 PFM



15

If it's a new binding mode, it is added to JASPAR. If not, we check if the profile is better than the previous profile. If it is, that profile is updated with the new version, otherwise it is discarded.



We have recently added to the JASPAR web a functionality that allows users to navigate across all versions of TF binding profiles.

Outline

- Overview
- Manual Curation
- **Tools:**
 - Profile Inference
 - Matrix Clustering
 - Genome Tracks
- New Web
- RESTful API
- Perspectives
- MANTA2

17

Now I will introduce some of the recent tools that we have developed. First, the profile inference tool was motivated to help JASPAR users who would recurrently ask for the right profile to use in their specie of interest.

Profile Inference

Please input a TF protein sequence for which to look for a JASPAR TF binding profile.

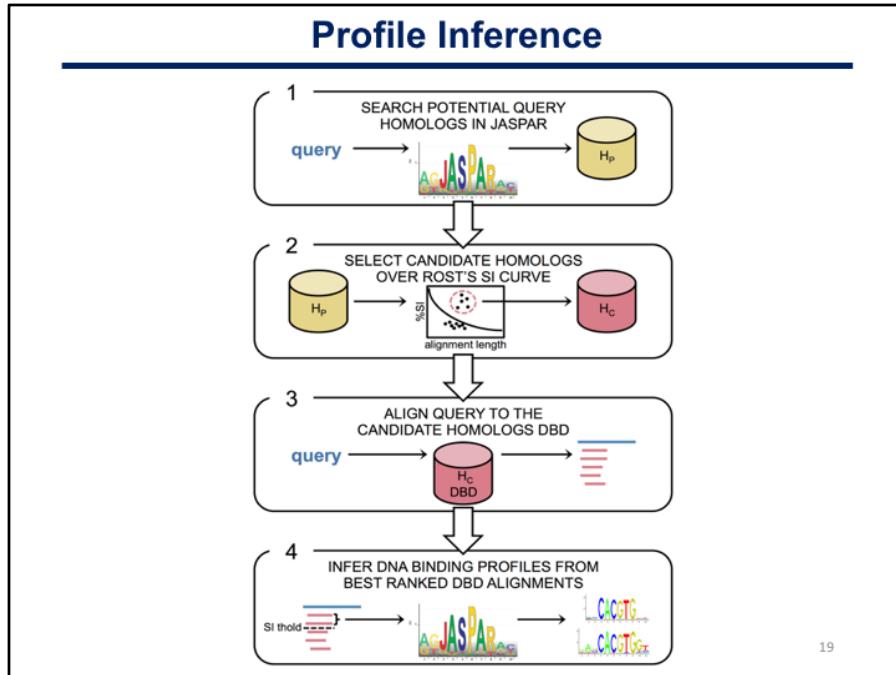
Paste a protein sequence below

Matrix ID	Name	DBD	E-value	Sequence logo
MA0781_1	PAX9	0.704	3.51197e-57	
MA0069_1	Pax6	1.0	0.0	
MA0014_2	PAX5	0.776	2.26217e-70	
MA0779_1	PAX1	0.704	4.72903e-55	
MA0208_1	al	0.719298245614	3.93817e-23	

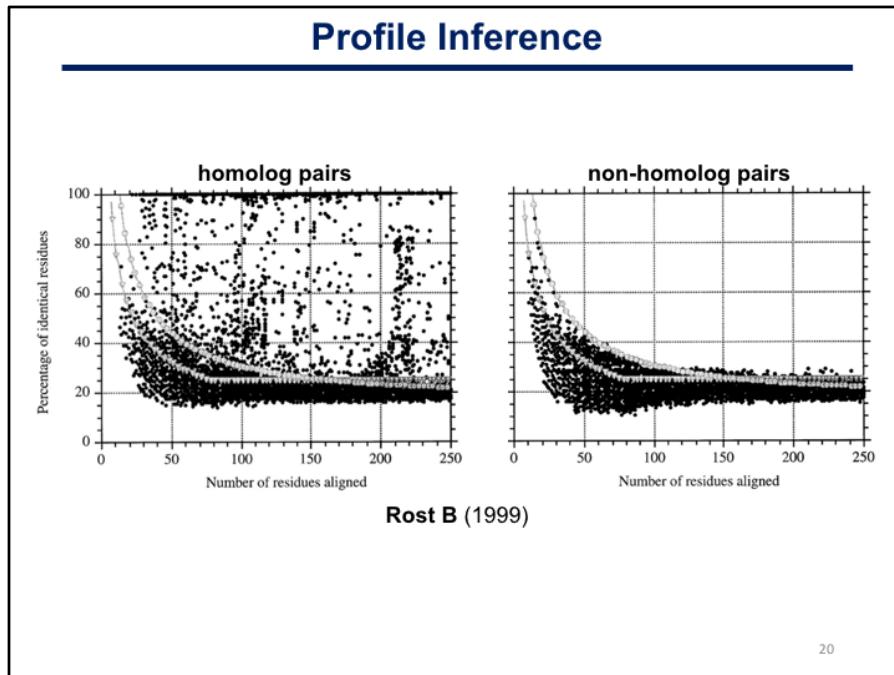
JASPAR Profile Inference

Follow @jaspar_db 114 followers

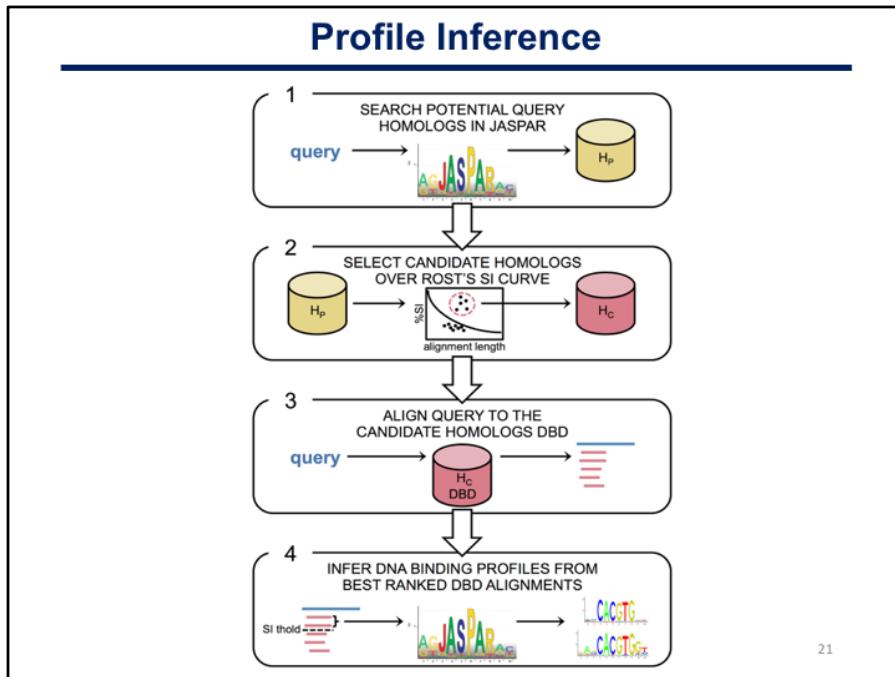
The tools offers a textbox to be filled with your sequence of interest, and upon clicking the blue button, it provides a list of profiles that potentially represent the binding affinities of you TF of interest.



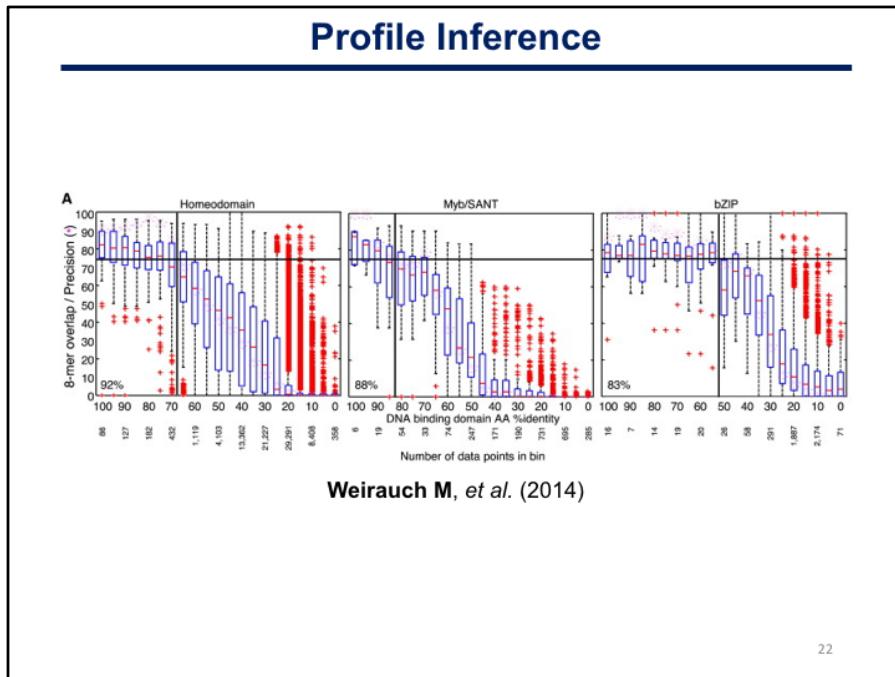
Internally, the tool searches JASPAR for potential homologs of the query. Only those above the Rost's sequence identity curve are kept.



The Rost's curve sets a flexible threshold on the sequence identity for pairwise alignments. The threshold is very stringent for short alignments, two infer homology from short alignments one requires a very high percentage of identical residues, and becomes more lax with the number of aligned residues. In principle, the curve separates correct from incorrect alignments between pairs of homologs.



Then, the query protein is aligned against the DNA binding domains of its candidate homologs, and if the alignments are above a certain threshold, then the binding profiles from the candidate homologs are assigned to the query.



22

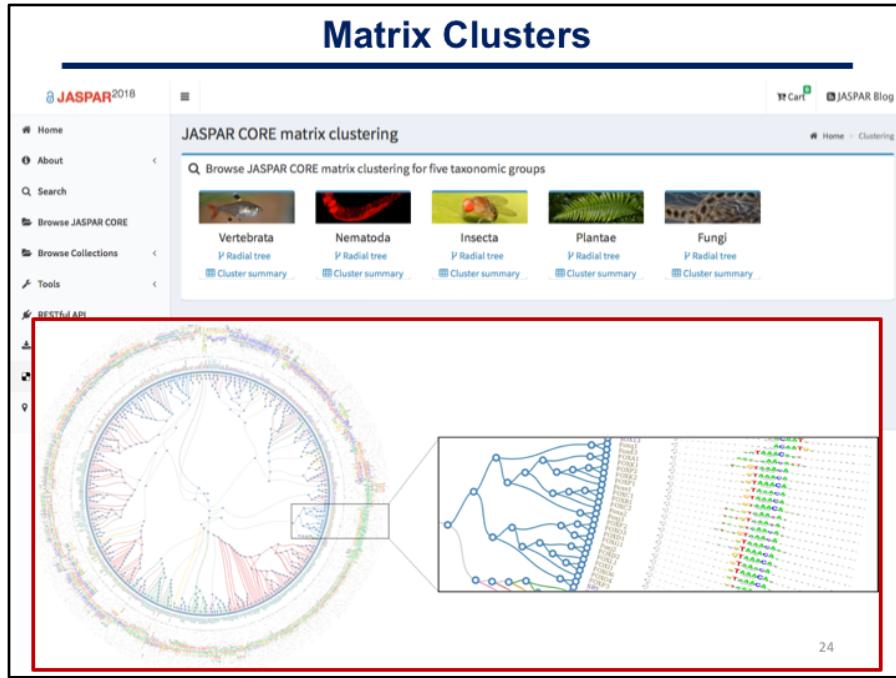
The thresholds on the sequence identity for DNA binding domain are extracted from the CiS-BP paper, in which it was observed that for certain families, the percentage of shared identical bound DNA sequences was very high even when the sequence similarity between the binding domains of the members of the family was low. Yet for other TF families, a few amino acid changes were enough as to make members of the family bind to different DNA sequences.

Outline

- Overview
- Manual Curation
- **Tools:**
 - Profile Inference
 - **Matrix Clustering**
 - Genome Tracks
- New Web
- RESTful API
- Perspectives
- MANTA2

23

While the non-redundancy is one of the guiding principles of JASPAR, TFs with similar DBDs often have similar binding preferences. To facilitate the exploration of similar profiles in the JASPAR CORE collection, we performed hierarchical clustering of PFM^s using the RSAT matrix-clustering tool developed by Jaime who is sitting there (point to him).



Specifically, the tool was applied to PFM_s in each taxon independently as well as in each TF class per taxon. The clustering results are provided as radial trees (below).

Matrix Clusters

Nb Input motifs	Nb Input collections	Nb Clusters Found	Download motif motifs	Download complete results [zip]	Linkage method	Similarity metric	Thresholds to partition the tree
579	1	79	Download	Download	average	Ncor	Ncor = 0.4 cor = 0.6

25

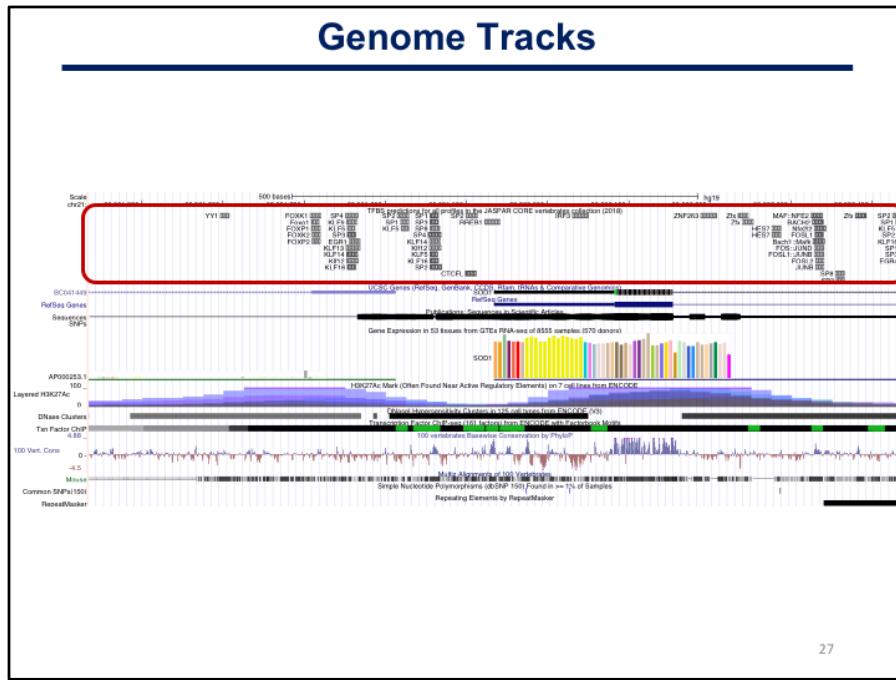
Clusters can further be explored through dedicated web pages, and downloaded for custom analysis.

Outline

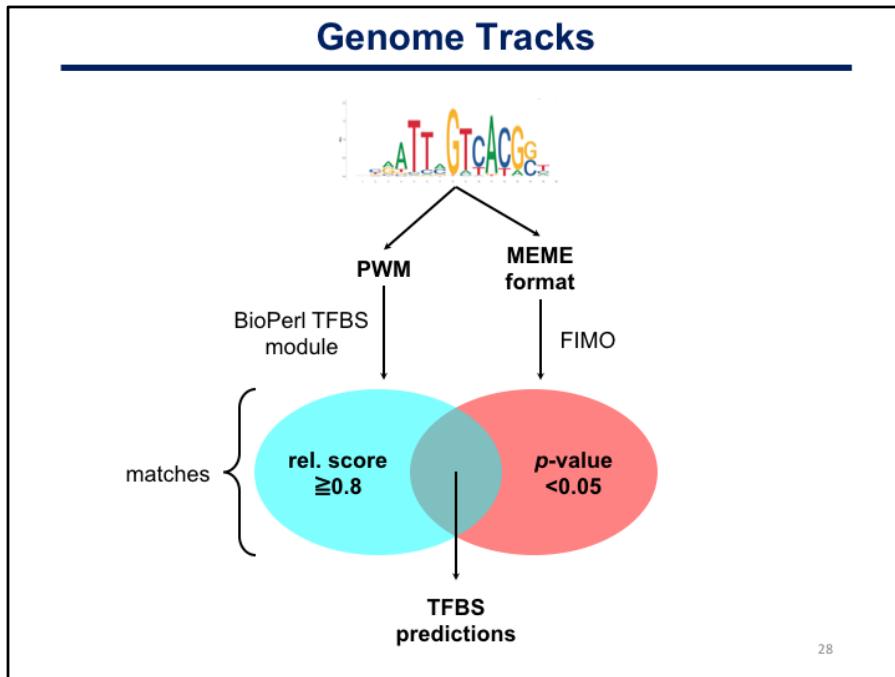
- Overview
- Manual Curation
- **Tools:**
 - Profile Inference
 - Matrix Clustering
 - **Genome Tracks**
- New Web
- RESTful API
- Perspectives
- MANTA2

26

Finally, I will introduce the Genome Tracks.



The tracks contain genome-wide TFBS predictions using the JASPAR CORE profiles.



Specifically, for each profile, a given genome is scanned in parallel using the TFBS Perl module and FIMO from the MEME suite. For scanning the genome with the BioPerl TFBS module, profiles are converted to PWMs and matches with a relative score ≥ 0.8 are kept. For the FIMO scan, profiles are reformatted to MEME motifs and matches with a p-value < 0.05 are kept. We only keep consistent TFBS predictions between the two methods. Finally, TFBS predictions are converted to genome tracks and colored according to their FIMO p-value (scaled between 0-1000, where 0 corresponds to a p-value of 1 and 1000 to a p-value $\leq 10^{-10}$), thus allowing for comparison of prediction confidence between different profiles.

Genome Tracks

Organism	Genome Assembly	JASPAR CORE
<i>Arabidopsis</i>	araTha1	Plants
<i>C. Elegans</i>	ce10	Nematodes
<i>Drosophila</i>	dm6	Insects
Human	hg19, hg38	Vertebrates
Mouse	mm10	Vertebrates
Yeast	sacCer3	Fungi
Zebrafish	danRer10, danRer11 (soon)	Vertebrates

Availability: UCSC and Ensembl browsers, Track Hub Registry

http://expdata.cmmt.ubc.ca/JASPAR/downloads/UCSC_tracks/2018/

29

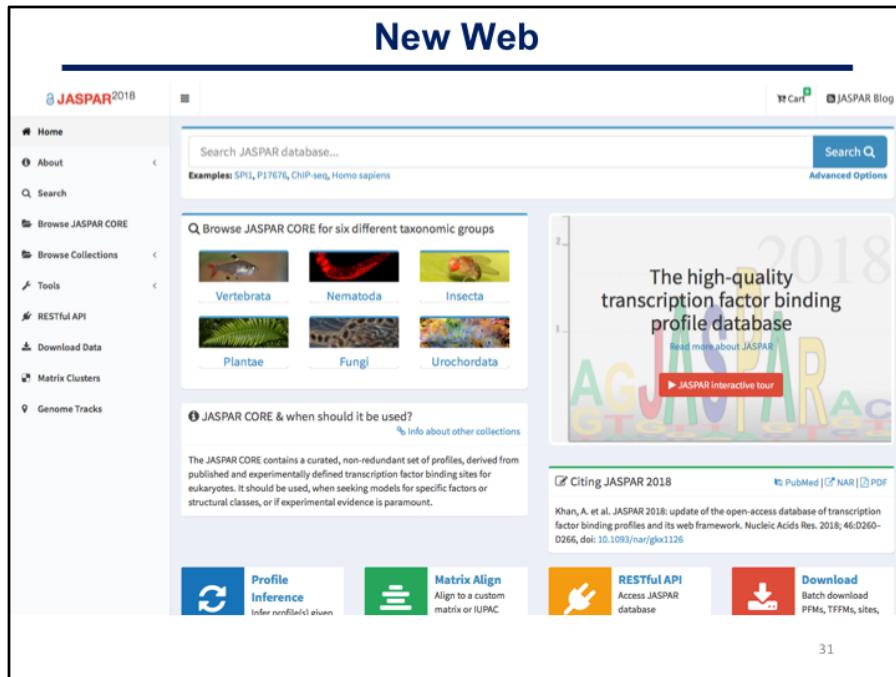
To date, we offer genome tracks for all these species and genome assemblies via UCSC, Ensembl and the Track Hub Registry. All TFBS predictions and underlying BED files are available in our web servers.

Outline

- Overview
- Manual Curation
- Tools:
 - Profile Inference
 - Matrix Clustering
 - Genome Tracks
- **New Web**
- RESTful API
- Perspectives
- MANTA2

30

One of the most significant updates of this release is the new, more user friendly web interface of JASPAR.



Here is an overview of the new web site. I remember when Anthony first interviewed me, he asked me why I hadn't used JASPAR during my PhD research. I told him that the main reason was that it was very complicated to navigate. Since then, we had been talking about updating the web, making it more 2.0. The web had been the same since the release of JASPAR in 2004. Here is the result.

New Web

The screenshot shows the JASPAR 2018 web interface. On the left is a sidebar with links: Home, About, Search (which is highlighted with a red box), Browse JASPAR CORE, Browse Collections, Tools, RESTful API, Download Data, Matrix Clusters, and Genome Tracks. The main content area has a header "Search profile(s)" and a sub-header "FOS". It includes a search bar with examples: SP1, P17676, ChIP-seq, Homo sapiens. Below the search bar are several dropdown filters: Collection (set to CORE), Taxon (set to Vertebrates), Species (All species), Data type (All data types), Class (All classes), Family (All families), and Versions (Latest version). At the bottom of the search form, there is a note about JASPAR being an open-access database of curated, non-redundant transcription factor (TF) binding profiles. To the right of the search form are links to About JASPAR, Profile versions, JASPAR video tour, Changelog, Blog and News, Contact Us, and links to JASPAR 2016 and JASPAR 2014. There are also logos for CAVAT and NCMM. At the very bottom, it says "You are using the latest 7th release (2018) of JASPAR." and "Follow @jaspar_db 114 followers".

The web comes with an improved search engine. It allows, for example, search for a given TF name in a specific collection, version or taxon.

New Web

The screenshot shows the JASPAR 2018 web interface. The left sidebar contains links for Home, About, Search, Browse JASPAR CORE, Browse Collections, Tools, RESTful API, Download Data, Matrix Clusters, and Genome Tracks. The main search bar at the top has 'FOS' entered, with examples like SP1, P17676, ChIP-seq, Homo sapiens. Below the search bar, it says '26 profile(s) found'. A table lists profiles with columns for ID, Name, Species, Class, Family, and Logo. The first few rows are:

ID	Name	Species	Class	Family	Logo
MA0099.3	FOS::JUN	Homo sapiens	Basic leucine zipper factors (bZIP)	Basic:Jun-related factors	
MA0476.1	FOS	Homo sapiens	Basic leucine zipper factors (bZIP)	Fos-related factors	
MA0477.1	FOSL1	Homo sapiens	Basic leucine zipper factors (bZIP)	Fos-related factors	
MA0605.1	Atf3	Mus musculus	Basic leucine zipper factors (bZIP)	Fos-related factors	
MA1126.1	FOS::JUN(var.2)	Homo sapiens	Basic leucine zipper factors (bZIP)	Basic:Jun-related factors	

On the right, there's a sidebar for 'Analyze selected profiles' with a note to select matrix profiles. It includes 'Add to cart' (0 items), a message about items in the cart, 'Scan' (0 items), and 'Cluster' (0 items). Buttons for 'Add to cart' and 'View cart' are highlighted with a red box.

33

And here are the resulting profiles matching the criteria of our search, which can be, for example, selected and added to cart for custom analysis.

jaspar.genereg.net/matrix/MA0476.1/

[JASPAR2018](#)

[Home](#) [About](#) [Search](#) [Browse JASPAR CORE](#) [Browse Collections](#) [Tools](#) [RESTful API](#) [Download Data](#) [Matrix Clusters](#) [Genome Tracks](#)

[Cart](#) [JASPAR Blog](#)

Detailed information of matrix profile MA0476.1

[Remove](#)

Profile summary

Name: FOS
Matrix ID: MA0476.1
Class: Basic leucine zipper factors (bZIP)
Family: Fos-related factors
Collection: CORE
Taxon: Vertebrates
Species: Homo sapiens
Data Type: ChIP-seq
Validation: 17916232
Uniprot ID: P01100
Pazara TF:
TFBSshape ID: 280
TFencyclopedia IDs:
Source: ENCODE
Comment:

[Download SVG](#)

Sequence logo

The sequence logo displays the consensus sequence TGAATTCGAT. The vertical axis represents the probability of each nucleotide (A, T, G, C) at each position. The sequence is composed of two half-sites: TGAATTC and CGAT.

Frequency matrix

JASPAR TRANSFAC MEME RAW PFM

Reverse comp.

A [7879	7475	0	0	29396	998	0	0	29396	258	4006	1
C [712	10177	0	0	0	14079	0	29396	0	5823	8236	1
G [9686	10841	0	27108	0	11206	0	0	0	1538	7897	1
T [11119	903	29396	2288	0	3113	29396	0	0	21777	9257	1

Binding sites information

HTML file FASTA file BED file

TFBS profiles

TFBSshape

We have also introduced semantic URLs to facilitate external linking to the detailed pages of individual profiles.

35

Finally, the downloads section allows for downloading profiles per taxonomic group in different formats, as a single batch or one file per profile, or the entire database.

Outline

- Overview
- Manual Curation
- Tools:
 - Profile Inference
 - Matrix Clustering
 - Genome Tracks
- New Web
- **RESTful API**
- Perspectives
- MANTA2

36

Probably, the second most significant update to JASPAR 2018 is its RESTful API.

The screenshot shows the homepage of the JASPAR RESTful API. At the top, there is a dark header bar with the JASPAR logo and several navigation links: "Home", "Browsable API", "API Overview", "Documentation", "Live API", "API Clients", and "Contact Us". Below the header, the main content area has a light gray background. It features a large, bold title "Welcome to {JASPAR RESTful API}" where the opening brace is red and the rest of the text is black. Underneath the title, a short paragraph explains the API's purpose: "This API provides easy-to-use REST web interface to query/retrieve matrix profile data from the latest version of JASPAR database. The API comes with a human browsable interface and also programmatic interface, which return the results in eight different formats, including `json`, `jsonp`, `jaspar`, `meme`, `transfac`, `pfm`, `yaml` and `bed`". Below this text are two buttons: a green "Try It Now" button and a blue "Live API" button. Further down, there is a section titled "Read more about JASPAR and JASPAR RESTful API" containing two bullet points with links to academic papers. At the very bottom of the page, a small note says "A RESTful API to programmatically access the latest version of JASPAR database." and the number "37" is visible on the right side.

The API provides an easy-to-use interface to query and retrieve data from JASPAR. The API comes with a browsable interface and a programmatic interface, which return results in several common formats, as for example JSON.

The screenshot shows the JASPAR REST API's "Live API" section. The top navigation bar includes links for Home, Browsable API, API Overview, Documentation, **Live API** (which is highlighted with a red box), API Clients, and Contact Us. Below this, a breadcrumb trail shows "Home / Live API". The main content area is titled "JASPAR Live API" and lists several categories with their corresponding API endpoints and descriptions:

- collections**:
 - `GET /api/v1/collections/` | Show/Hide | List Operations | Expand Operations | Description: List all the collections available in JASPAR.
 - `GET /api/v1/collections/{collection}/` | Show/Hide | List Operations | Expand Operations | Description: API endpoint that returns a list of all matrix profiles based on collection name.
- infer**:
 - `GET /api/v1/infer/{sequence}/` | Show/Hide | List Operations | Expand Operations | Description: Infer matrix profiles, given profile sequence.
- matrix**:
 - `GET /api/v1/matrix/` | Show/Hide | List Operations | Expand Operations | Description: REST API endpoint that returns a list of all matrix profiles.
 - `GET /api/v1/matrix/{base_id}/versions/` | Show/Hide | List Operations | Expand Operations | Description: List matrix profile versions based on base_id.
 - `GET /api/v1/matrix/{matrix_id}/` | Show/Hide | List Operations | Expand Operations | Description: Gets profile detail information.
- releases**:
 - `GET /api/v1/releases/` | Show/Hide | List Operations | Expand Operations | Description: REST API endpoint that returns all releases of JASPAR database.
 - `GET /api/v1/releases/{release_number}/` | Show/Hide | List Operations | Expand Operations | Description: Gets JASPAR release information based on release number.
- sites**:
 - `GET /api/v1/sites/{matrix_id}/` | Show/Hide | List Operations | Expand Operations | Description: List matrix profile sites based on matrix_id.

38

The Live API provides several recipes to query the JASPAR collections and to access different JASPAR tools.

The screenshot shows the JASPAR REST API's 'Live API' section. The URL is `/api/v1/infer/{sequence}/`. The 'Parameters' table shows a single parameter `sequence` with the value `IPMGTSQTTSTGLISPGVSVPVQVPGSEPMQYWWRLO`. The 'Response Messages' table shows a 200 status code with a curl command and a URL. The 'Request Headers' table shows a header `Accept: application/json`.

Parameter	Value	Description	Parameter Type	Data Type
<code>sequence</code>	<code>IPMGTSQTTSTGLISPGVSVPVQVPGSEPMQYWWRLO</code>		path	string

HTTP Status Code	Reason	Response Model	Headers
200			

Request Headers	
{ "Accept": "application/json" }	

39

For example it provides access to the Profile Inference tool. By pasting a sequence and clicking on try it out, it provides two ways of fetching the inferred profiles: either using curl or through a URL request.

RESTful API

40

Here you can see that the CURL command returns a JSON document with the inferred profiles.

The screenshot shows the JASPAR REST API browsable interface. The title bar says "RESTful API". Below it is a navigation bar with links: "Home", "Browsable API", "API Overview", "Documentation", "Live API", "API Clients", and "Contact Us". The main content area has a heading "Matrix Inference". A tooltip message states: "This is JASPAR browsable API, which provides easy-to-use REST web interface to query/retrieve matrix profile data from JASPAR database. The API comes with a human browsable interface and also programmatic interface, which return the results in eight different formats, including json,jsonp,jaspar,transfac,pfm,eme and yaml, bed and also api for browsable interface." Below this is a green box stating "API endpoint that infer matrix models based on protein sequence.". The URL in the address bar is "GET /api/v1/infer/MQNSHGVNQOLGGVFNNGRLLPDSRKIVELAHSGARPCDISRLQVSNGCVSKILGRYYETGSIRPAZGGSKPRVATPEVVSKIAQYKRECP5IFAWEIRDRLLSEGV". The "GET" button is highlighted with a red box. The JSON response is shown in a code block:

```
{
  "count": 16,
  "results": [
    {
      "name": "Pax6",
      "url": "http://jaspar.genereg.net/api/v1/matrix/M00069.1",
      "value": 0.0,
      "matrix_id": "M00069.1",
      "dbd": "1.0",
      "sequence_logo": "http://jaspar.genereg.net/static/logos/svg/M00069.1.svg"
    },
    {
      "name": "PAX1",
      "url": "http://jaspar.genereg.net/api/v1/matrix/M00480.1",
      "value": 1.91654e-71,
      "matrix_id": "M00480.1",
      "dbd": "1.0",
      "sequence_logo": "http://jaspar.genereg.net/static/logos/svg/M00480.1.svg"
    },
    {
      "name": "Pax6",
      "url": "http://jaspar.genereg.net/api/v1/matrix/M00069.2",
      "value": 0.0,
      "matrix_id": "M00069.2",
      "dbd": "1.0",
      "sequence_logo": "http://jaspar.genereg.net/static/logos/svg/M00069.2.svg"
    },
    {
      "name": "PAX1",
      "url": "http://jaspar.genereg.net/api/v1/matrix/M00778.1",
      "value": 2.34098e-72,
      "matrix_id": "M00778.1",
      "dbd": "1.0",
      "sequence_logo": "http://jaspar.genereg.net/static/logos/svg/M00778.1.svg"
    },
    {
      "name": "Pax6",
      "url": "http://jaspar.genereg.net/api/v1/matrix/M00069.1",
      "value": 0.0,
      "matrix_id": "M00069.1",
      "dbd": "1.0",
      "sequence_logo": "http://jaspar.genereg.net/static/logos/svg/M00069.1.svg"
    },
    {
      "name": "PAX1",
      "url": "http://jaspar.genereg.net/api/v1/matrix/M00779.1",
      "value": 1.01449e-52,
      "matrix_id": "M00779.1",
      "dbd": "1.0",
      "sequence_logo": "http://jaspar.genereg.net/static/logos/svg/M00779.1.svg"
    },
    {
      "name": "Pax6",
      "url": "http://jaspar.genereg.net/api/v1/matrix/M00208.1",
      "value": 0.0,
      "matrix_id": "M00208.1",
      "dbd": "1.0",
      "sequence_logo": "http://jaspar.genereg.net/static/logos/svg/M00208.1.svg"
    },
    {
      "name": "PAX1",
      "url": "http://jaspar.genereg.net/api/v1/matrix/M00688.1",
      "value": 1.90934e-71,
      "matrix_id": "M00688.1",
      "dbd": "8.7109337e-05",
      "sequence_logo": "http://jaspar.genereg.net/static/logos/svg/M00688.1.svg"
    }
  ]
}
```

41

Alternatively, one can just copy and paste the provided URL in the internet browser, and the API will output the inferred profiles. By clicking on GET, one can download results as a JSON document.

RESTful API

```
[Oriols-MacBook-Pro:JASPAR orforness$ python
Python 2.7.13 (default, Jul 18 2017, 09:17:00)
[GCC 4.2.1 Compatible Apple LLVM 8.1.0 (clang-802.0.42)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import coreapi
>>> import urllib2
>>> from coreapi import codecs
>>> # Initialize #
... client = coreapi.Client()
>>> codec = codecs.CoreJSONCodec()
>>> seq = coreapi.SequentialCodec(codec)
>>> url = "http://jaspar.generereg.net/api/v1/infer/%s/" % seq
>>>
>>> # Get response w/ coreapi
... coreapi_response = client.get(url)
... coreapi_response = codec.encode(coreapi_response)
>>>
>>> # Get response w/ urllib2
... request = urllib2.Request(url, headers={'Content-Type': 'application/json'})
... f = urllib2.urlopen(request)
... urllib2_response = f.read()
... f.close()
...
>>> print(urllib2_response)
{"count":16,"results":[{"name":"Pax6","url":"http://jaspar.generereg.net/api/v1/matrix/MA0069.1","evalue":0.0,"matrix_id":"MA0069.1","dbd":10.76,"sequence_logo":"http://jaspar.generereg.net/static/logos/svg/MA0069.1.svg"}, {"name":"PAKX","url":"http://jaspar.generereg.net/api/v1/matrix/MA0069.1","evalue":0.0,"matrix_id":"MA0069.1","dbd":10.76,"sequence_logo":"http://jaspar.generereg.net/static/logos/svg/MA0069.1.svg"}, {"name":"PAK5","url":"http://jaspar.generereg.net/api/v1/matrix/MA0014.2","evalue":6.162e-67,"matrix_id":"MA0014.2","dbd":10.776,"sequence_logo":"http://jaspar.generereg.net/static/logos/svg/MA0014.2.svg"}, {"name":"PA X9","url":"http://jaspar.generereg.net/api/v1/matrix/MA0781.1","evalue":2.34098e-54,"matrix_id":"MA0781.1","dbd":0.72,"sequence_logo":"http://jaspar.generereg.net/static/logos/svg/MA0781.1.svg"}, {"name":"PAK1","url":"http://jaspar.generereg.net/api/v1/matrix/MA0779.1","evalue":1.01449e-52,"matrix_id":"MA0779.1","dbd":10.72,"sequence_logo":"http://jaspar.generereg.net/api/v1/matrix/MA0779.1.svg"}, {"name":"PAK2","url":"http://jaspar.generereg.net/api/v1/matrix/MA0208.1","evalue":1.98946e-22,"matrix_id":"MA0208.1","dbd":0.7192982456140351,"sequence_logo":"http://jaspar.generereg.net/static/logos/svg/MA0208.1.svg"}]}
>>> print(coreapi_response == urllib2_response)
True
>>> ]]
```

42

Finally, the RESTful API can also be accessed using the coreapi or the urllib2 python modules. Here is an example of the code.

Outline

- Overview
- Manual Curation
- Tools:
 - Profile Inference
 - Matrix Clustering
 - Genome Tracks
- New Web
- RESTful API
- **Perspectives**
- MANTA2

43

Perspectives

- JASPAR beta (contains profiles/functionalities to be added to the next release)
- Validation section (contains profiles that we failed to validate through literature support)
- Genome tracks at user's request?
- While we are close to profile every human TF (except for zinc fingers), we don't have an estimate on the % of covered binding modes, heterodimers, etc.

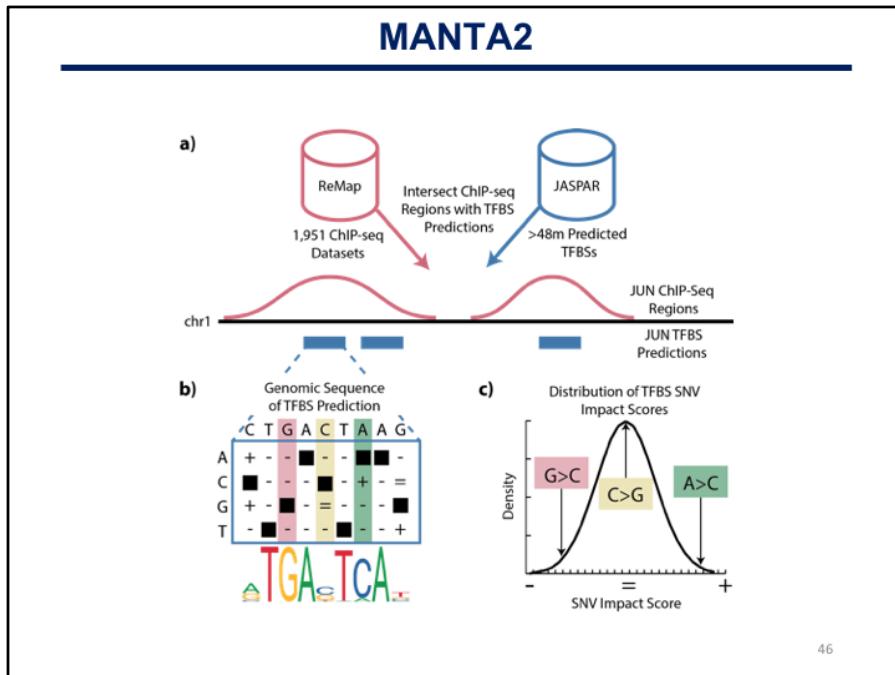
44

Outline

- Overview
- Manual Curation
- Tools:
 - Profile Inference
 - Matrix Clustering
 - Genome Tracks
- New Web
- RESTful API
- Perspectives
- **MANTA2**

45

To finish my presentation, I would like to introduce you the 2018 update of our Mongo database for the analysis of TF binding site alterations MANTA. This work has been done in collaboration with the Mathelier Lab, specially with help from Marius here (point towards him).



46

MANTA stores TF binding site predictions within ChIP-seq regions, as well as the potential impact of all single nucleotide variants within these binding sites on TF binding. Benefiting from the recent updates of ReMap and JASPAR, we intersected the ReMap ChIP-seq regions with the JASPAR genome-wide TFBS predictions, increasing the size of the database largely. SNV impact on TF binding was predicted by means of Z-scores obtained from the distribution of relative scores for that TFBS.

The Mongo database for the analysis of transcription factor binding site alterations (MANTA)

Fornes, Oriol; Gheorghe, Marius; Richmond, Phillip A.; Arenillas, David J.; Wasserman, Wyeth W.; Mathelier, Anthony

The MongoDB for the ANalysis of Transcription factor (TF)-binding site (TFBS) Alterations (MANTA) was originally created in 2015 to study the impact of regulatory mutations in B-cell lymphomas (Mathelier et al. 2015). The database stores TFBSs predicted in the human genome by combining ChIP-seq regions from ReMap and JASPAR profiles, as well as the potential impact scores for all possible single nucleotide variants (SNVs) at these TFBSs. The second release of the database, MANTA2, houses >48 million TFBS predictions for 225 TFA, covering 255,918,025 bp (~8%) of the human genome (hg38). Here we provide a MongoDB dump of MANTA2 comprising the files experiments.bson, experiments.metadata.json, system.indexes.bson, tfbs_snvs.bson, and tfbs_snvs.metadata.json.

Files (6.6 GB)

Name	Size	Download
manta2_mongodb.dump.tgz	6.6 GB	Download
md5:f3501a9eddf155cb7885e79a744fd9		

Versions

Version 2 10.5281/zenodo.1044747 Dec 1, 2017

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.1044747. This DOI represents all versions, and will always resolve to the latest one. [Read more](#).

Share

47

MANTA2 is available for download at Zenodo, with scripts granting programmatic access to the database in a GitHub repository.

The screenshot shows the GitHub repository page for `wassermanlab/MANTA2`. The repository has 37 commits, 1 branch, 0 releases, and 2 contributors. The commit history includes changes by `darenillas`, `manta2`, and others. The README section contains the following text:

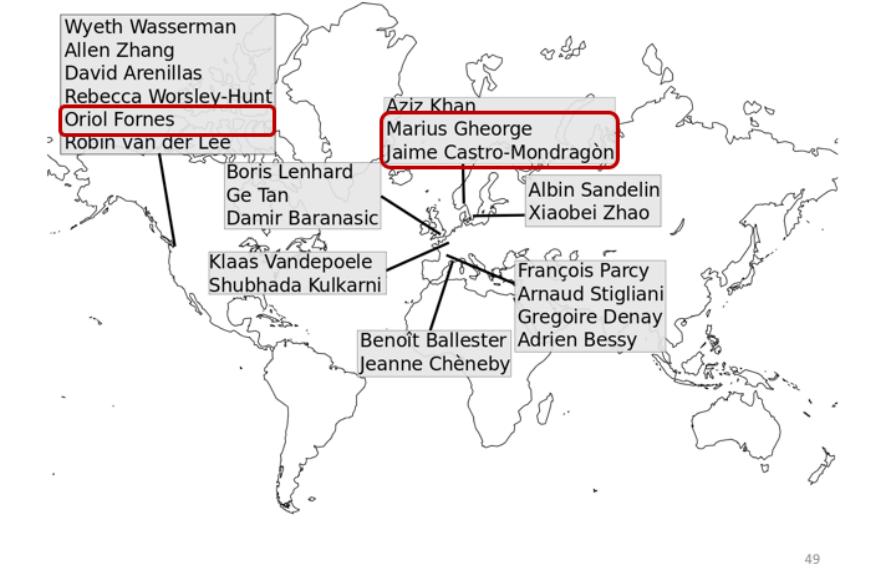
MANTA2

The MongoDB for the ANalysis of Transcription factor (TF)-binding site (TFBS) Alterations (MANTA) was originally created in 2015 to study the impact of regulatory mutations in B-cell lymphomas ([Mathelier et al. 2015](#)). The database stores

48

MANTA2 is available for download at Zenodo, with scripts granting programmatic access to the database in a GitHub repository.

The JASPAR Consortium



And that's it. Here is a distribution of members from the JASPAR consortium.

Acknowledgements

Funds from Genome Canada and the CIHR (OnTarget grants 255ONT and BOP-149430) and the Weston Brain Institute (20R74681)

WestGrid (www.westgrid.ca) and Compute Canada (www.computecanada.ca) support for genome tracks

Special thanks to the scientific community for performing the experiments and for publicly releasing the data

50

And finally I want to thank my funding agencies, Westgrid and Compute Canada for compute support, and specially the scientific community for performing the experiments and publicly releasing the data.

Questions

51

And now I will be taking in questions.

TF Coverage for Human

The Human Transcription Factors

Samuel A. Lambert,^{1,9} Arttu Jolma,^{2,9} Laura F. Campitelli,^{1,9} Pratyush K. Das,³ Yimeng Yin,⁴ Mihai Albu,² Xiaoting Chen,⁵ Jussi Taipale,^{3,4,6,*} Timothy R. Hughes,^{1,2,*} and Matthew T. Weirauch^{6,7,8,*}

¹Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada

²Donnelly Centre, University of Toronto, Toronto, ON, Canada

³Genome-Scale Biology Program, University of Helsinki, Helsinki, Finland

⁴Division of Functional Genomics and Systems Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Solna, Sweden

⁵Center for Autoimmune Genomics and Etiology (CAGE), Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

⁶Department of Biochemistry, Cambridge University, Cambridge CB2 1GA, United Kingdom

⁷Divisions of Biomedical Informatics and Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

⁸Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA

*These authors contributed equally

^{*}Correspondence: aj208@cam.ac.uk (J.T.), t.hughes@utoronto.ca (T.R.H.), Matthew.Weirauch@cchmc.org (M.T.W.)

<https://doi.org/10.1016/j.cell.2018.01.029>

Transcription factors (TFs) recognize specific DNA sequences to control chromatin and transcription, forming a complex system that guides expression of the genome. Despite keen interest in understanding how TFs control gene expression, it remains challenging to determine how the precise genomic binding sites of TFs are specified and how TF binding ultimately relates to regulation of transcription. This review considers how TFs are identified and functionally characterized, principally through the lens of a catalog of over 1,600 likely human TFs and binding motifs for two-thirds of them. Major classes of human TFs differ markedly in their evolutionary trajectories and expression patterns, underscoring distinct functions. TFs likewise underlie many different aspects of human physiology, disease, and variation, highlighting the importance of continued effort to understand TF-mediated gene regulation.

52

Describe this slide...

Genome Tracks Usage

[Hosts \(Top 10\)](#) - [Full list](#) - [Last visit](#) - [Unresolved IP Address](#)

Hosts : 0 Known, 116 Unknown (unresolved ip) 95 Unique visitors	Pages	Hits	Bandwidth	Last visit
128.114.119.134	492	1,784	2.05 MB	31 Dec 2017 - 18:43
128.114.119.135	462	1,694	1.94 MB	31 Dec 2017 - 20:15
128.114.119.133	458	1,750	1.78 MB	31 Dec 2017 - 19:41
128.114.119.132	354	1,528	1005.95 KB	31 Dec 2017 - 17:13
128.114.119.131	299	1,210	1.09 MB	31 Dec 2017 - 12:25
129.70.40.99	117	337	748.28 KB	29 Dec 2017 - 11:38
132.249.245.79	70	231	77.98 KB	29 Dec 2017 - 02:58
193.62.192.249	49	69	8.49 KB	28 Dec 2017 - 04:24
193.62.192.251	42	58	5.72 KB	28 Dec 2017 - 04:25
193.62.192.253	42	64	8.60 KB	28 Dec 2017 - 04:25
Others	457	1,012	9334.11 GB	

53

Describe this slide...