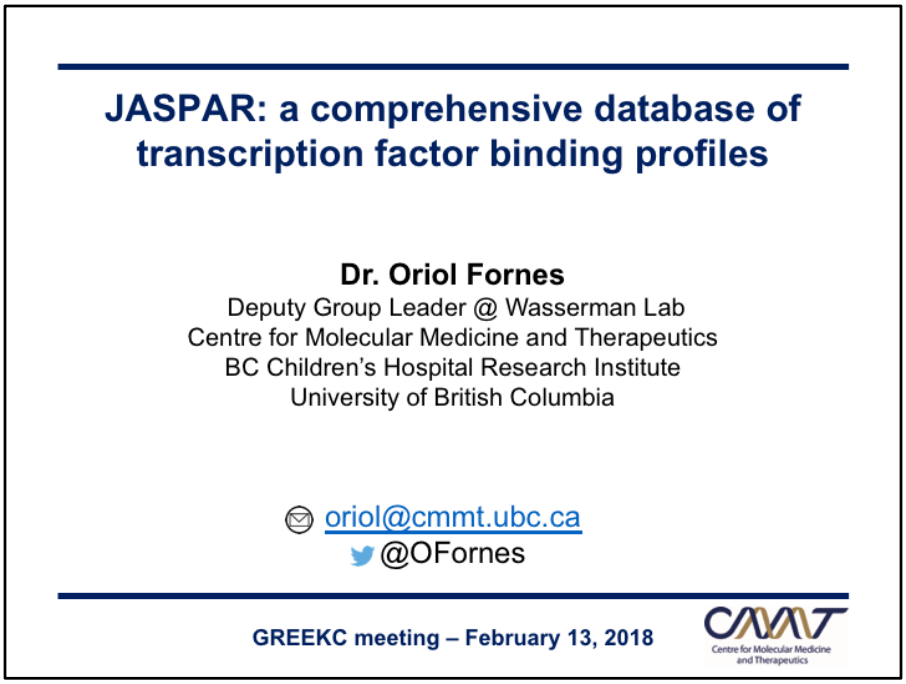
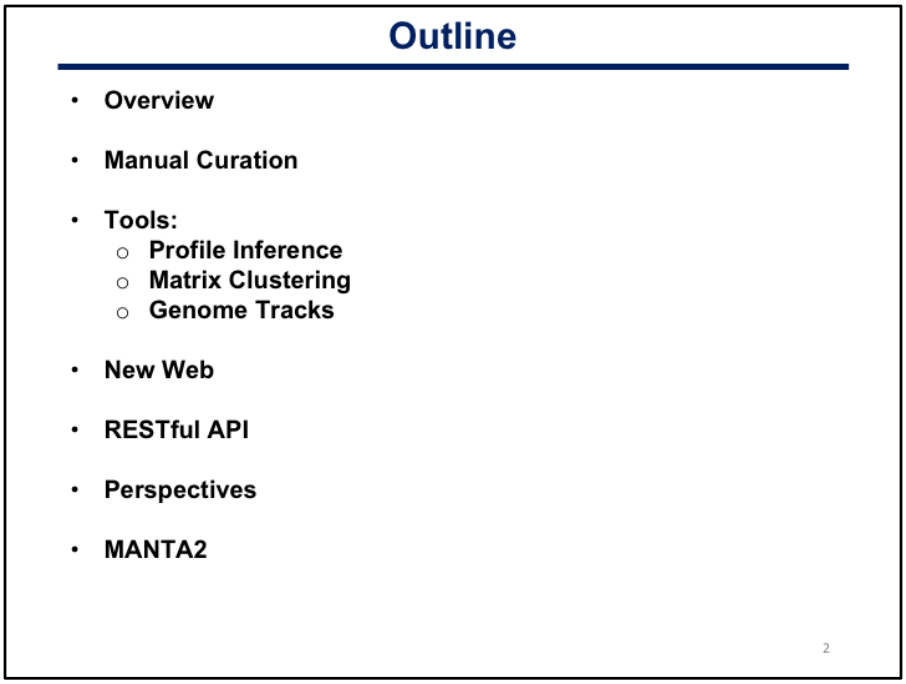


Good morning to everyone. First of all, I want to thank the organization for giving me the opportunity to come to Ljubljana. Today, I will be talking about the latest release of JASPAR for 2018, which has been recently published in the NAR database issue, and describe its new and improved features and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

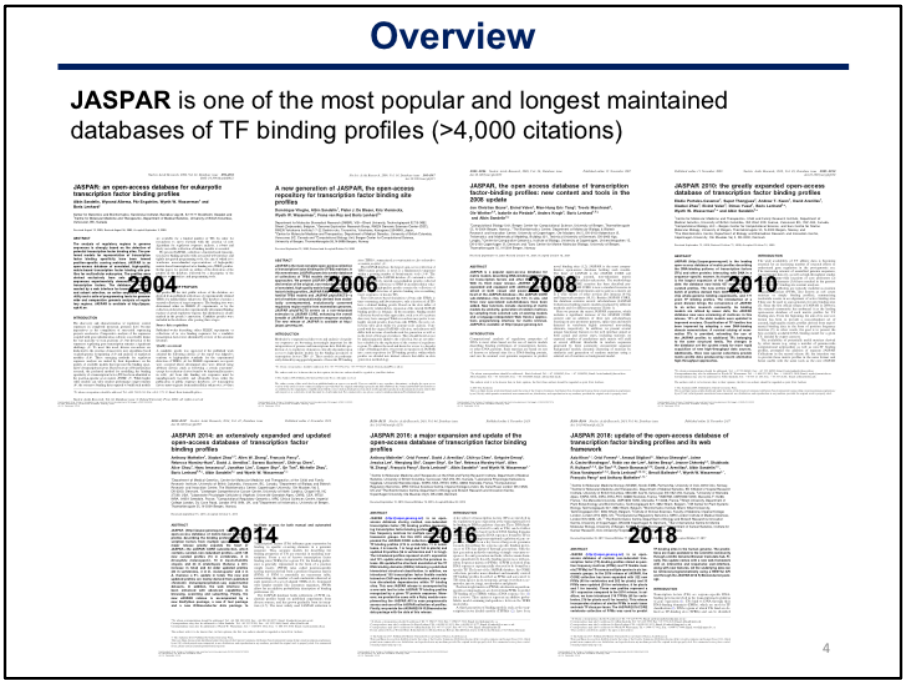


Here is an outline of my presentation. I will start with a brief overview of JASPAR, followed by the curation process that we perform before adding/updating any new binding profiles to the database. Then I will describe you some of the most recent tools that we have developed. I will also describe some 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.

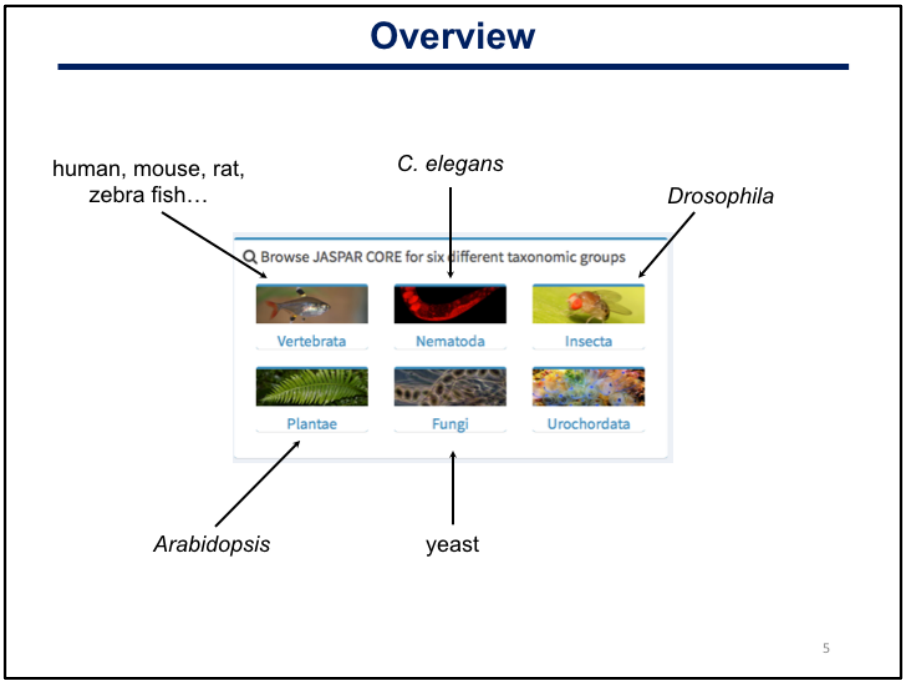
Outline

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

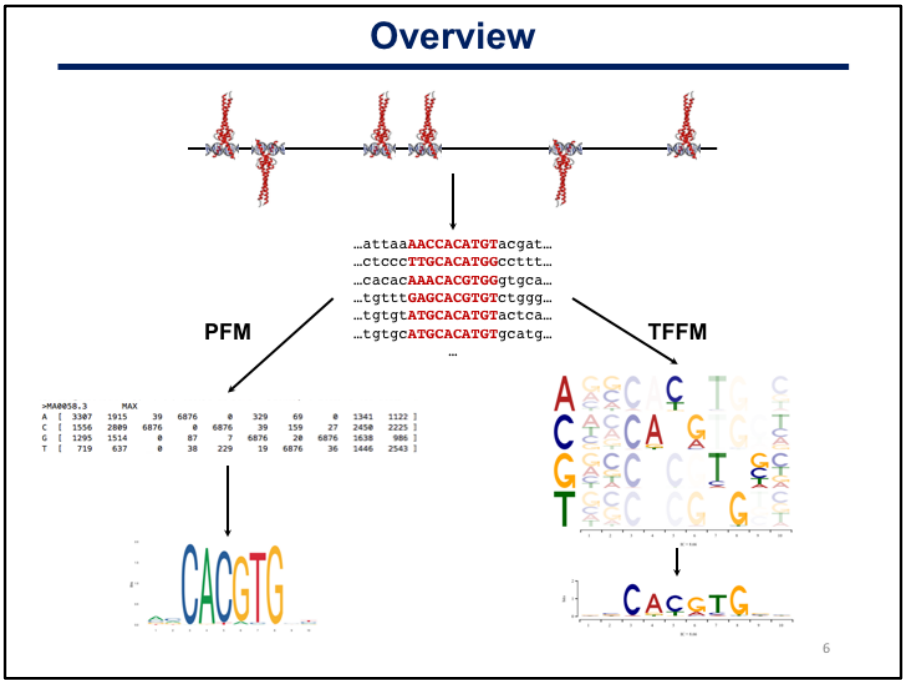
3



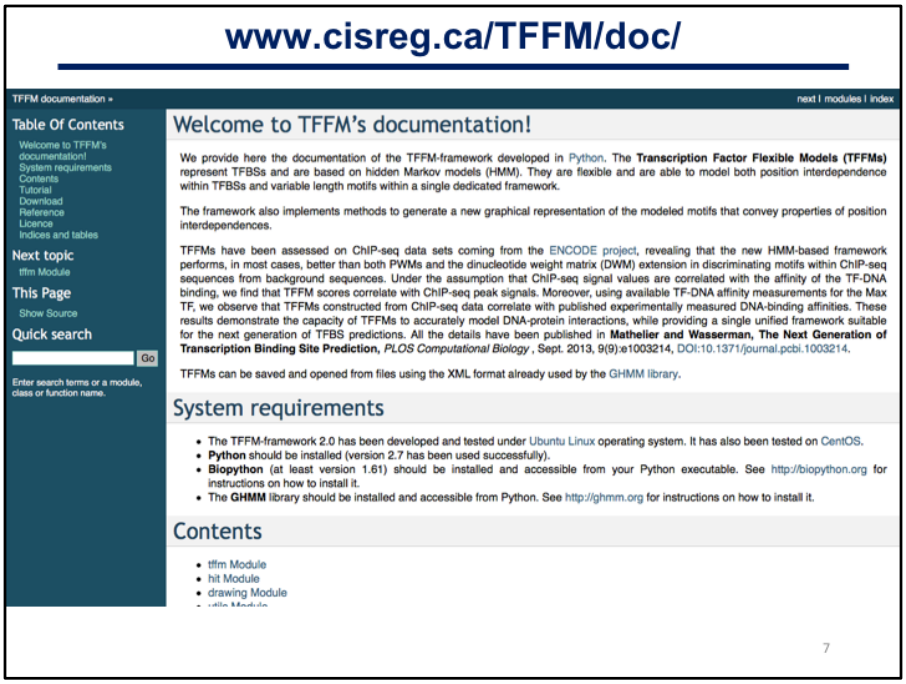
So JASPAR is one of the most popular and longest maintained databases of TF binding profiles. It was initially published in 2004 and has undergone 7 updates.



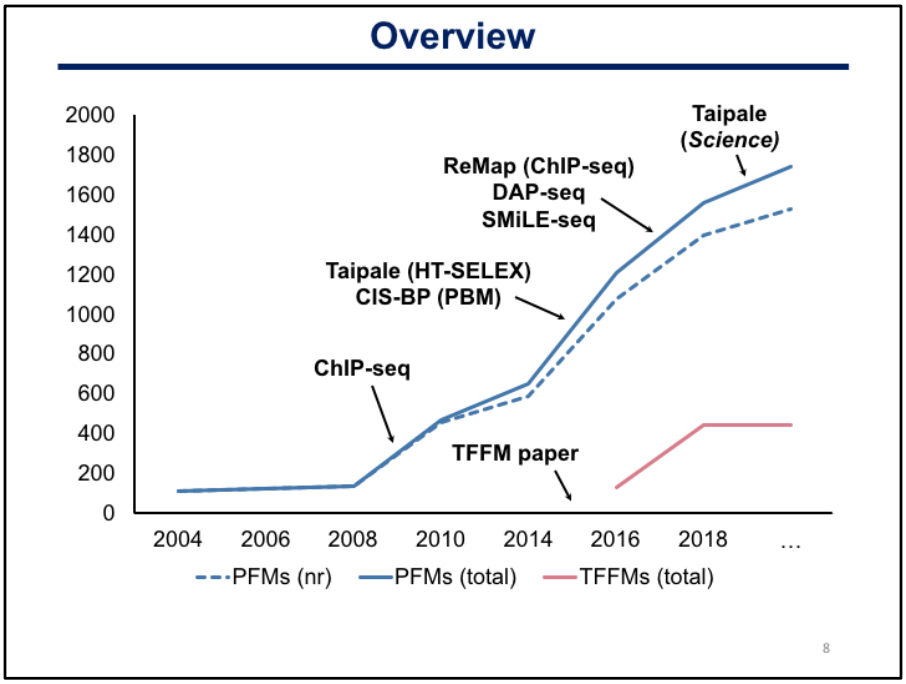
It comprises TF binding profiles from species from 6 different taxonomic groups. For example, for vertebrates, we include human, mouse, rat, and zebra fish profiles, among others, while for nematodes, insects, plants and fungi, profiles are mainly from C. elegans, Drosophila, Arabidopsis and yeast, respectively.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



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



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



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

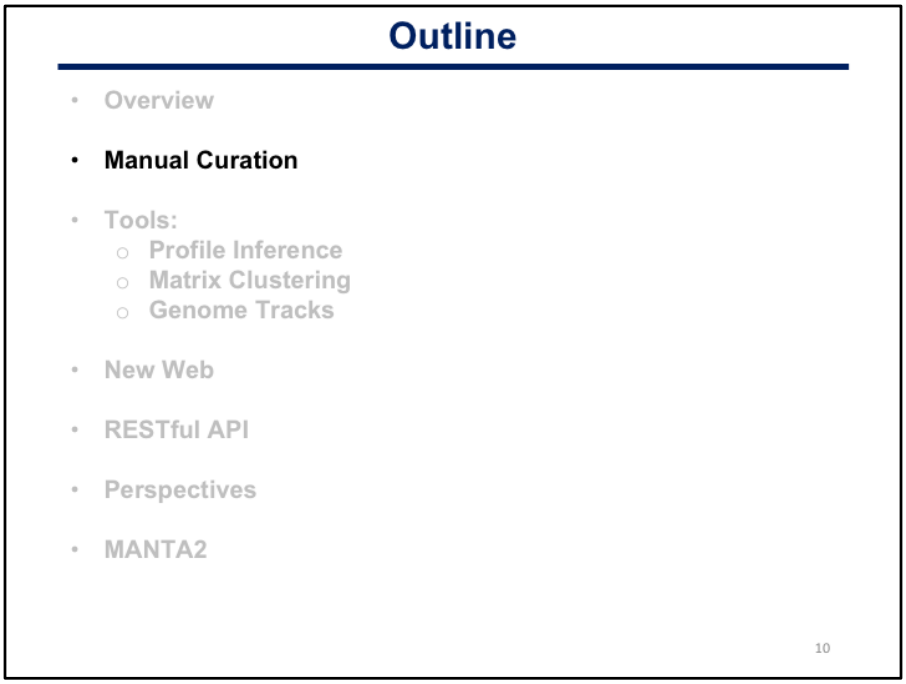
The JASPAR database as of 2018:

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**

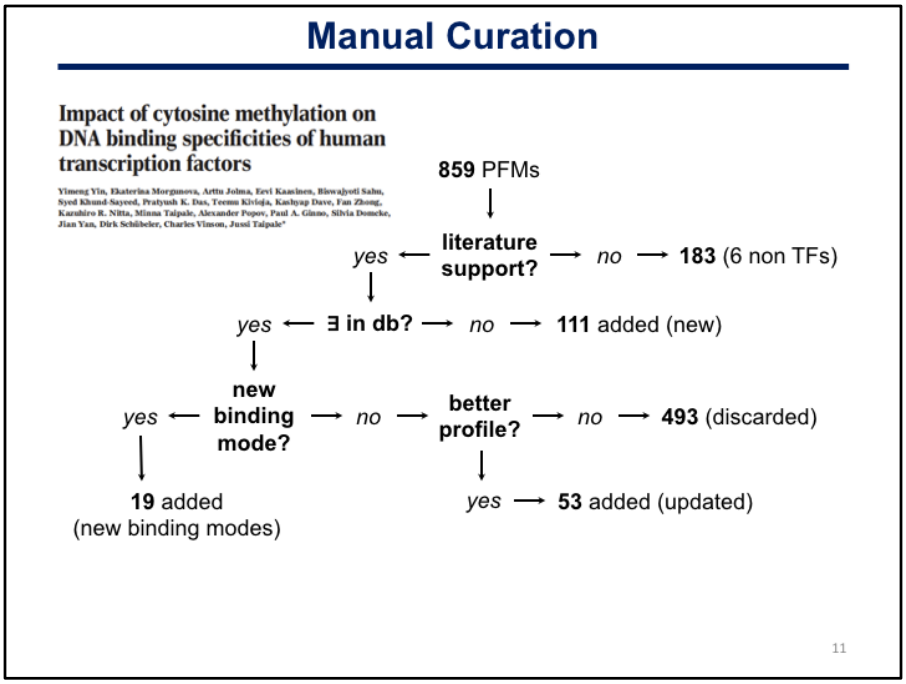
Experimental data sources: ChIP-chip/seq, DAP-seq, EMSA, PBM, (HT-)SELEX, SMiLE-seq…

9

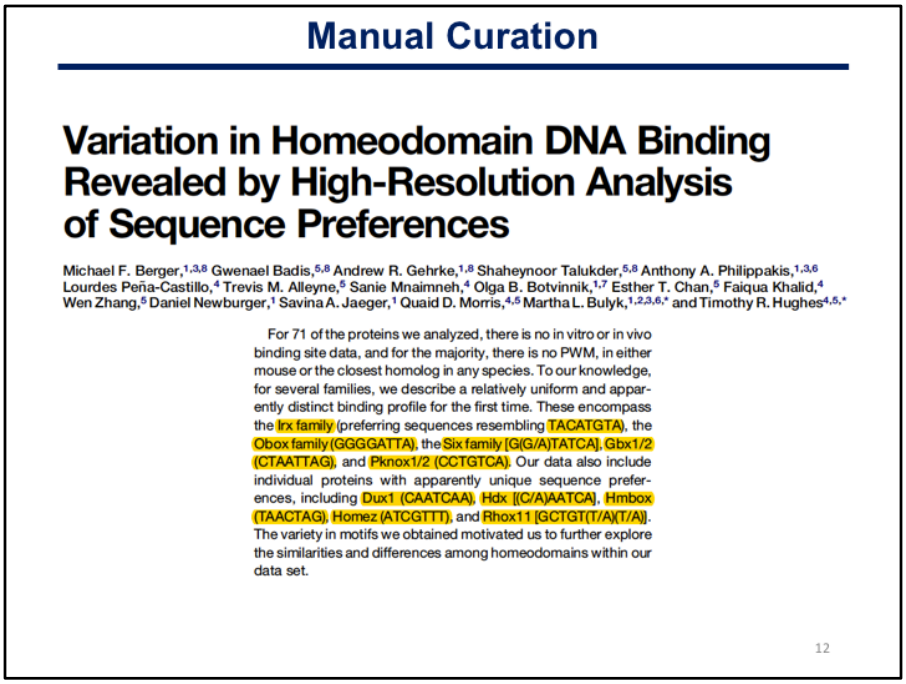
All in all, as of 2018, JASPAR contains over 14 hundred non redundant PFMs, and almost 500 TFFMs, with TFs belonging to 64 different structural classes and 115 families.



Now, I will describe the manual curation of profiles.



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

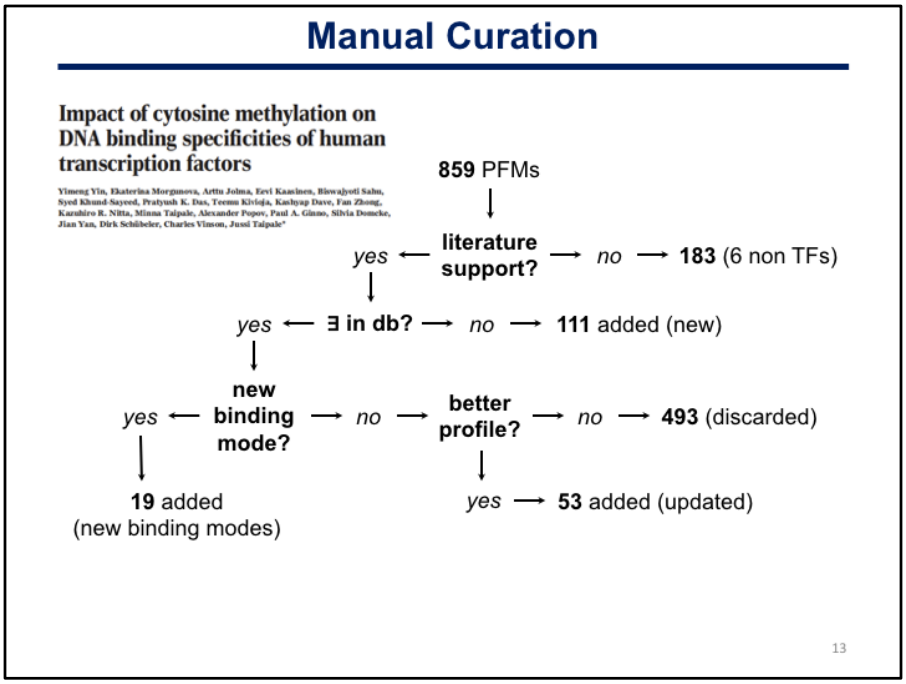
Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences

Michael F. Berger,[1,3,8] Gwenael Badis,[5,8] Andrew R. Gehrke,[1,8] Shaheynoor 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 (GGGGATTA), the Six family [G(G/A)TATCA], Gbx1/2 (CTAATTAG), and Pknox1/2 (CCTGTCA). 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.

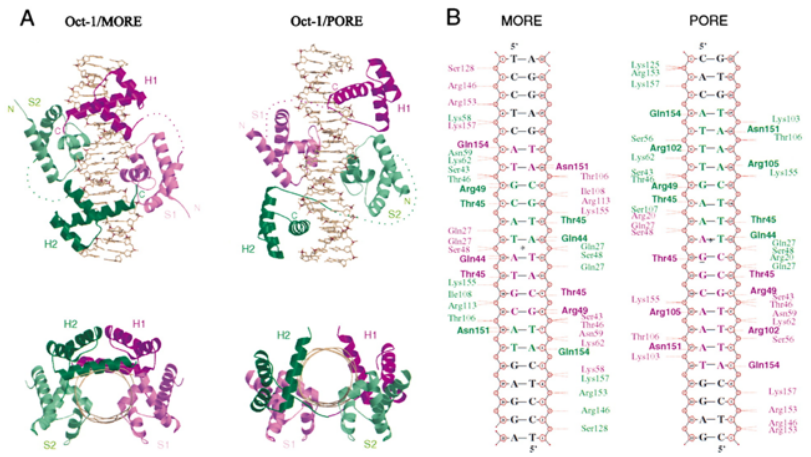
12

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.



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.

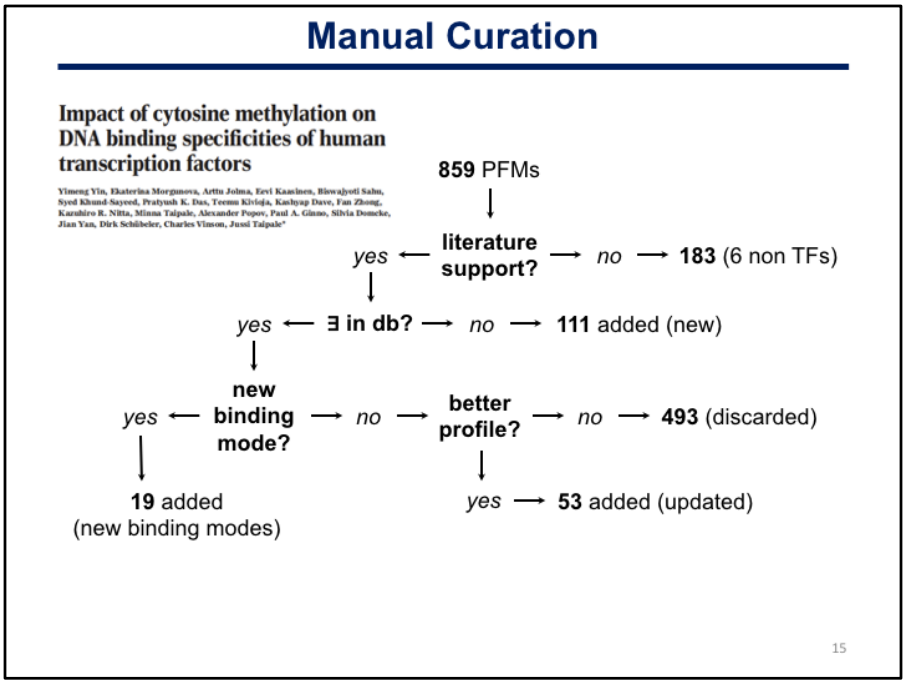
Manual Curation



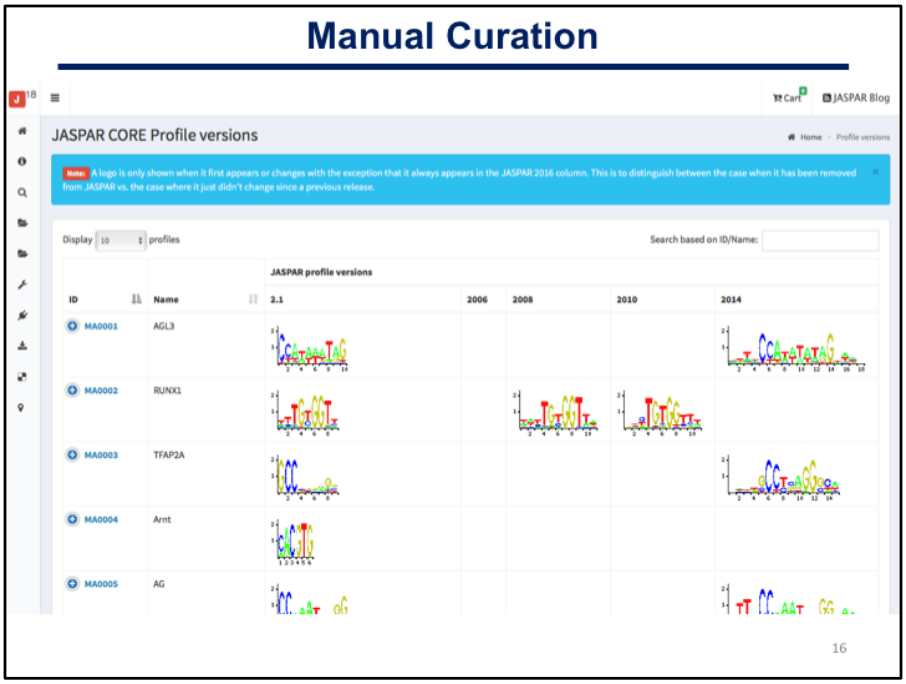
Reményi A, *et al.* (2001)

14

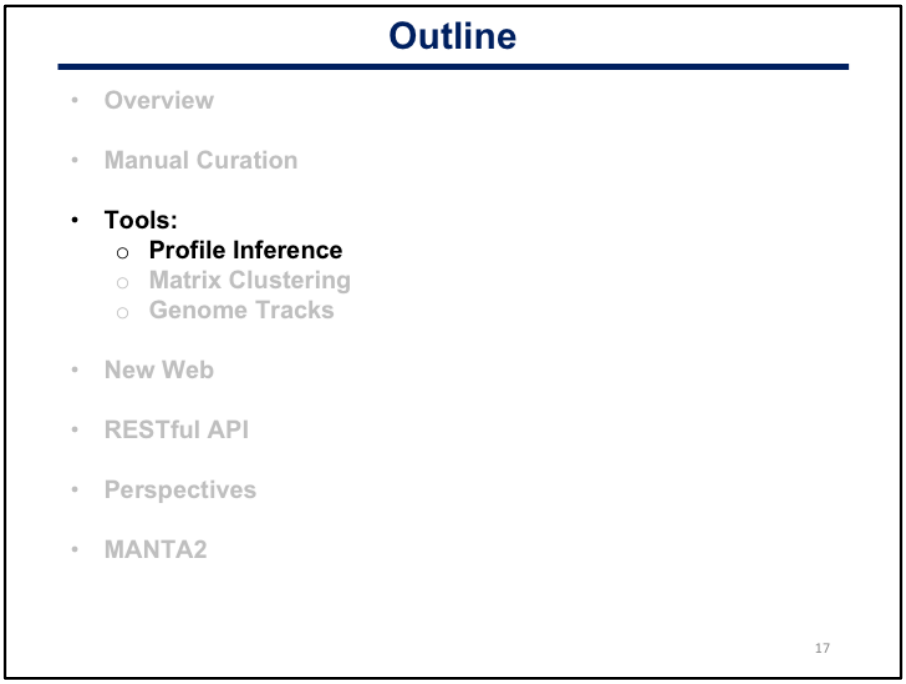
By binding mode we refer to the ability of some TFs, such as Oct-1, to bind DNA in different conformations, resulting in different DNA binding preferences for the MORE and PORE conformations.



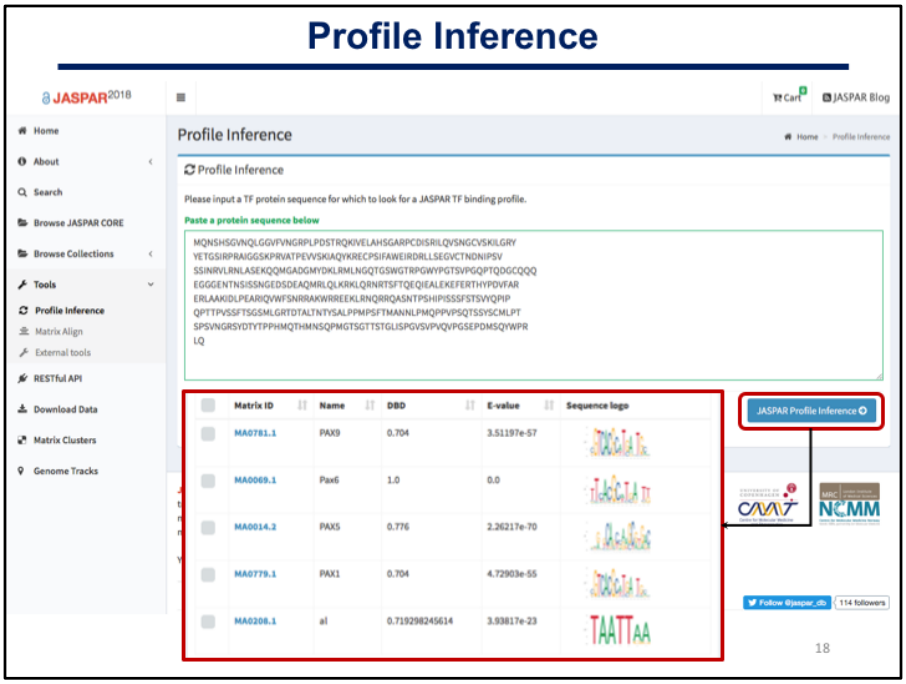
If it's a new binding mode, it is added to JASPAR. If not, we check if the profile is better than the previous profile. If it is, that profile is updated with the new version, otherwise it is discarded.



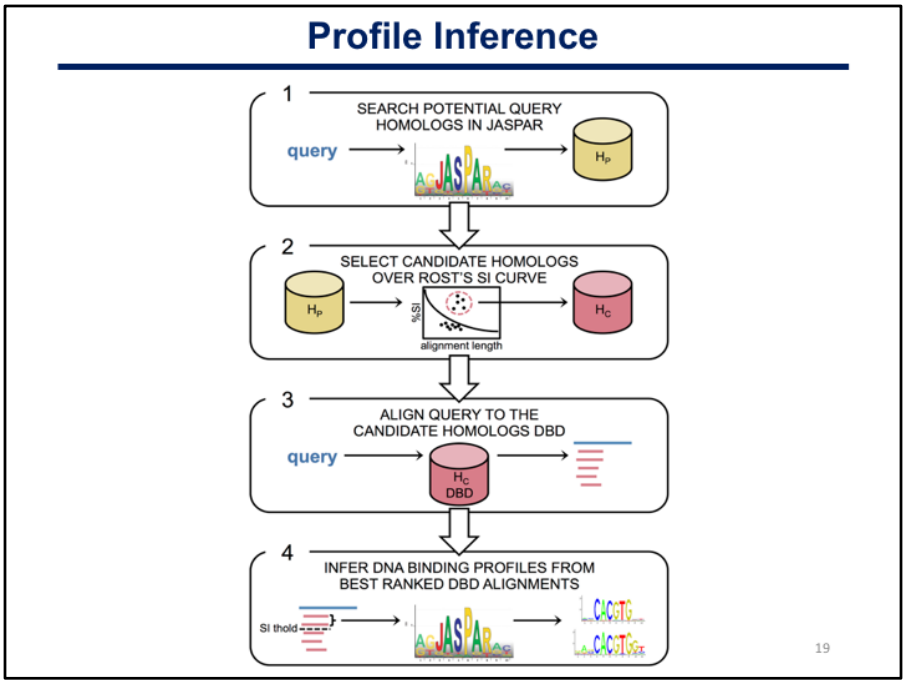
We have recently added to the JASPAR web a functionality that allows users to navigate across all versions of TF binding profiles.



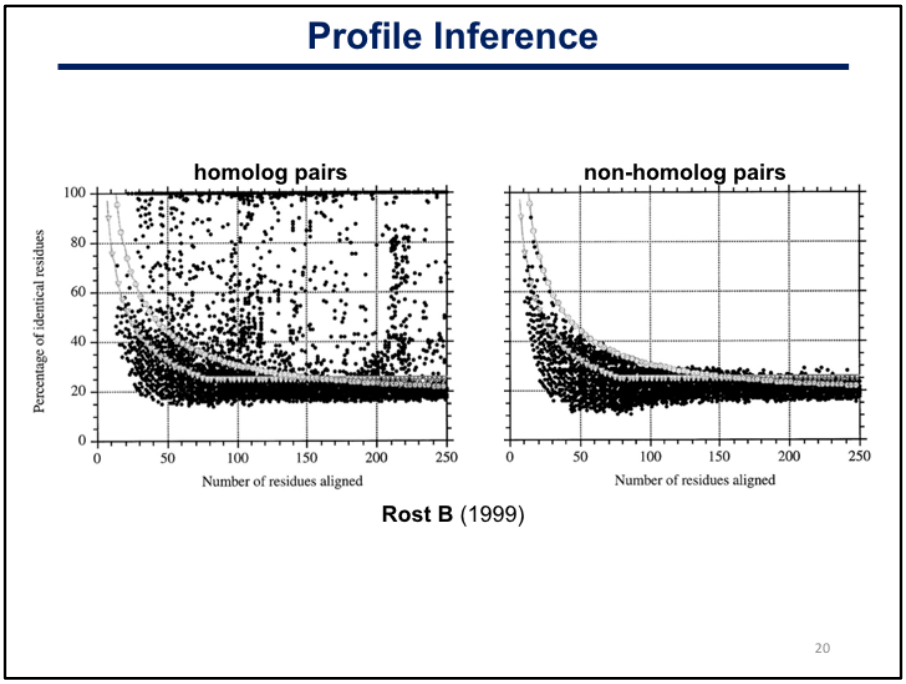
Now I will introduce some of the recent tools that we have developed. First, the profile inference tool was motivated to help JASPAR users who would recurrently ask for the right profile to use in their specie of interest.



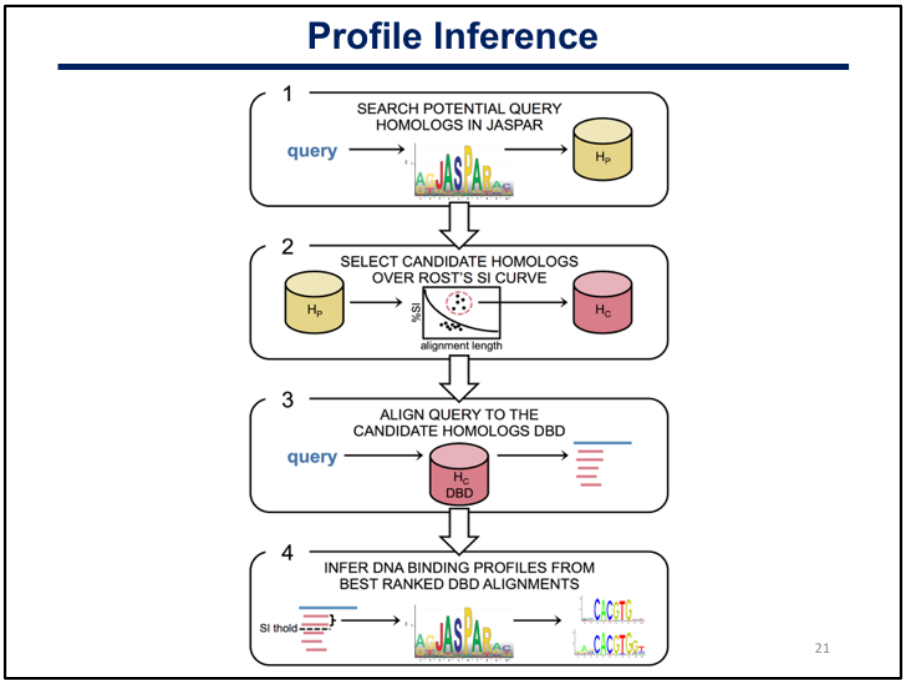
The tools offers a textbox to be filled with your sequence of interest, and upon clicking the blue button, it provides a list of profiles that potentially represent the binding affinities of you TF of interest.



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

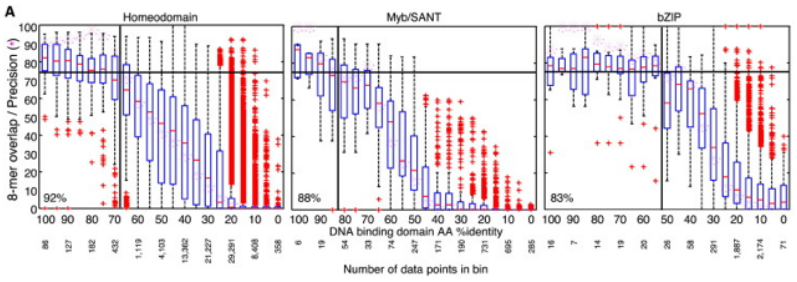


The Rost's curve sets a flexible threshold on the sequence identity for pairwise alignments. The threshold is very stringent for short alignments, two infer homology from short alignments one requires a very high percentage of identical residues, and becomes more lax with the number of aligned residues.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

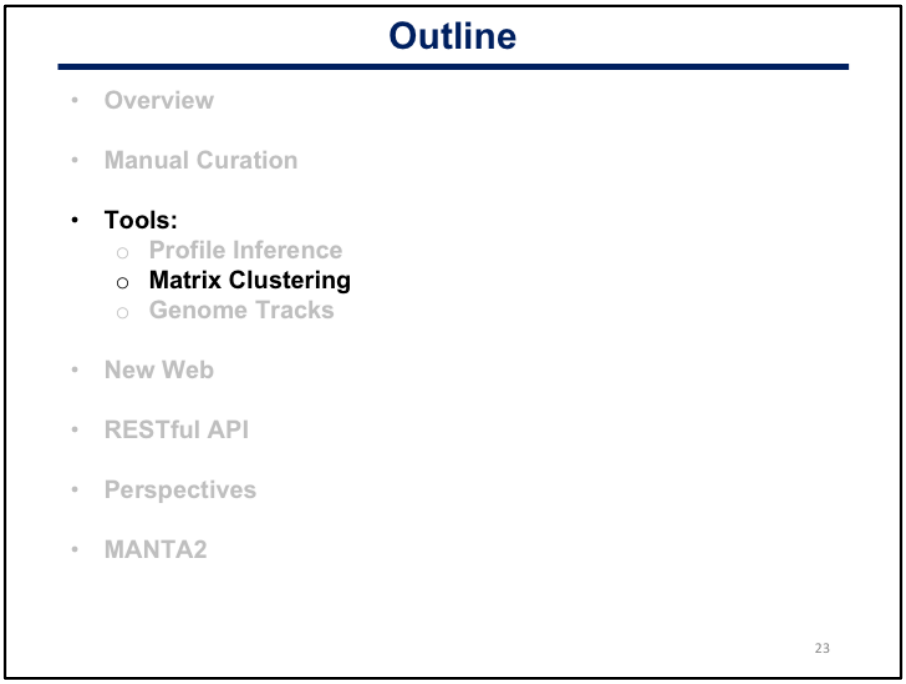
Profile Inference



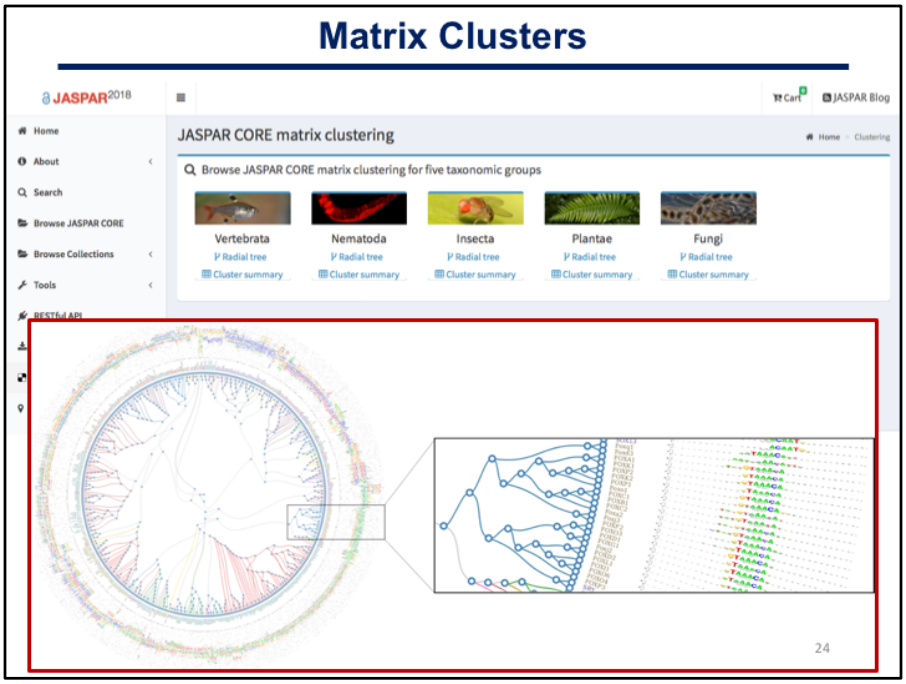
Weirauch M, *et al.* (2014)

22

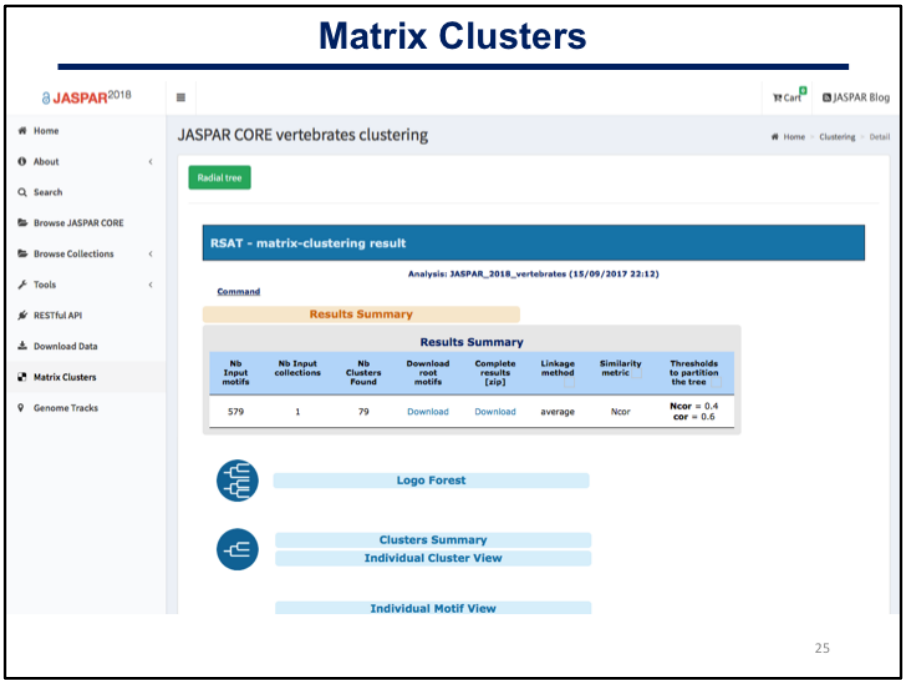
The thresholds on the sequence identity for DNA binding domain are extracted from the CiS-BP paper, in which it was observed that for certain families, the percentage of shared identical bound DNA sequences was very high even when the sequence similarity between the binding domains of the members of 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.



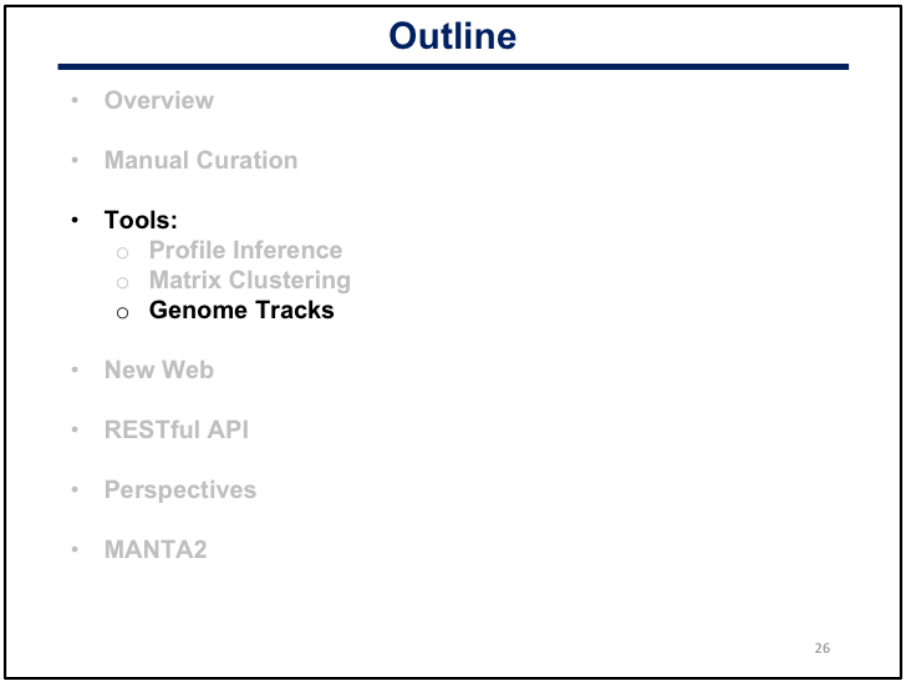
While the non-redundancy is one of the guiding principles of JASPAR, TFs with similar DBDs often have similar binding preferences. To facilitate the exploration of similar profiles in the JASPAR CORE collection, we performed hierarchical clustering of PFMs using
the RSAT matrix-clustering tool developed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



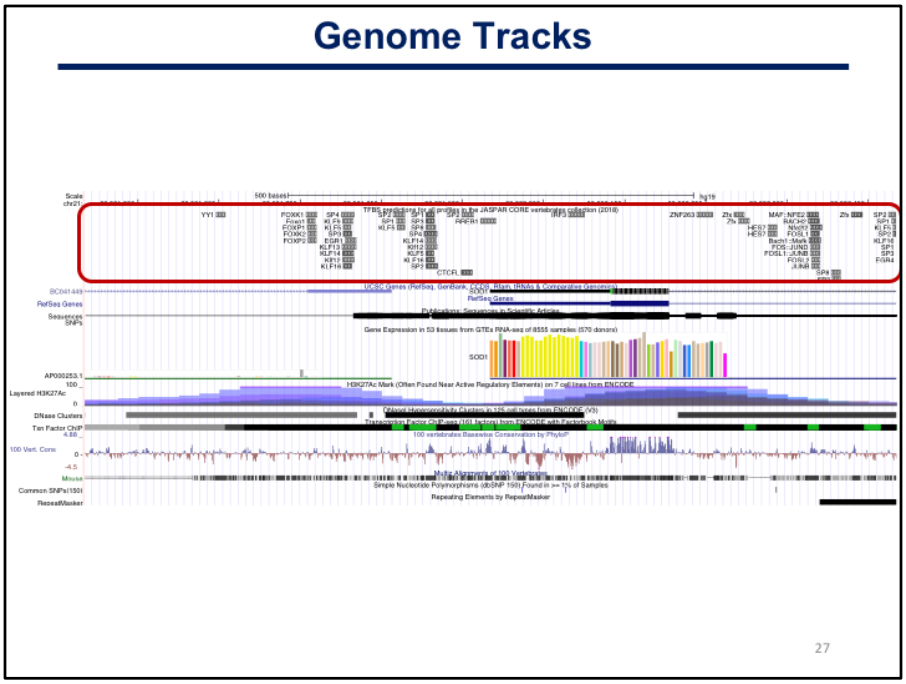
Specifically, the tool was applied to PFMs in each taxon independently as well as in each TF class per taxon. The clustering results are provided as radial trees (below).



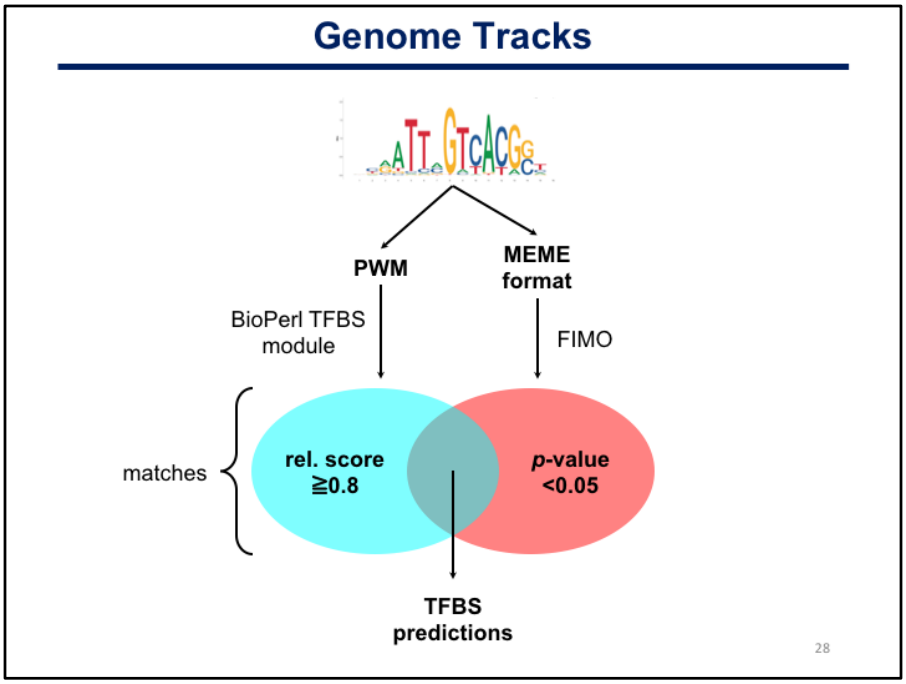
Clusters can further be explored through dedicated web pages, and downloaded for custom analysis.



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 convertd 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 ≤10-10), thus allowing for comparison of prediction confidence between different profiles.

Genome Tracks

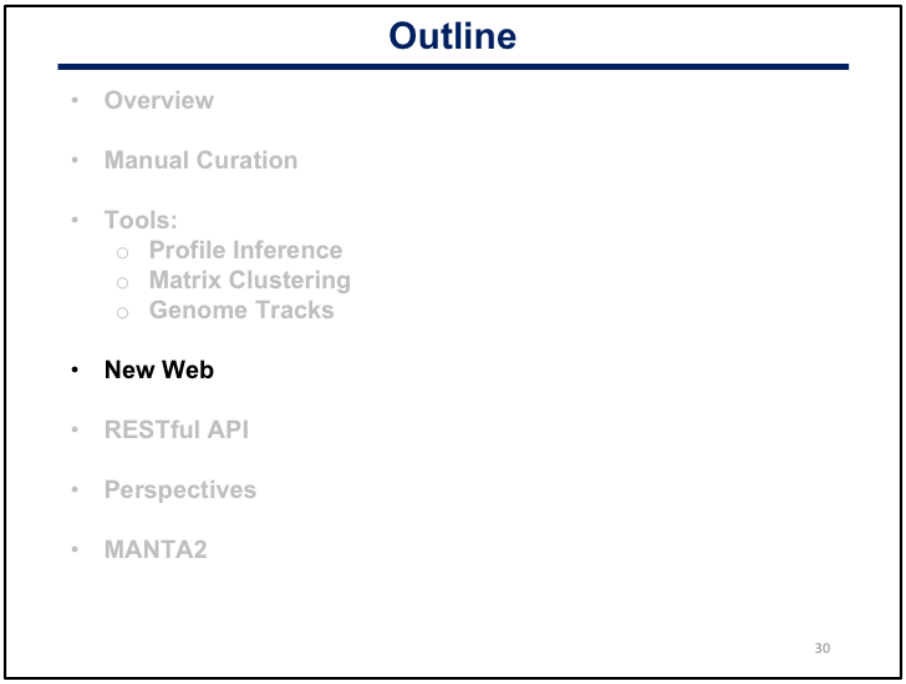
Organism	Genome Assembly	JASPAR CORE
Arabidopsis	araTha1	Plants
C. Elegans	ce10	Nematodes
Drosophila	dm6	Insects
Human	hg19, hg38	Vertebrates
Mouse	mm10	Vertebrates
Yeast	sacCer3	Fungi
Zebrafish	danRer10, danRer11 (soon)	Vertebrates

Availability: UCSC and Ensembl browsers, Track Hub Registry

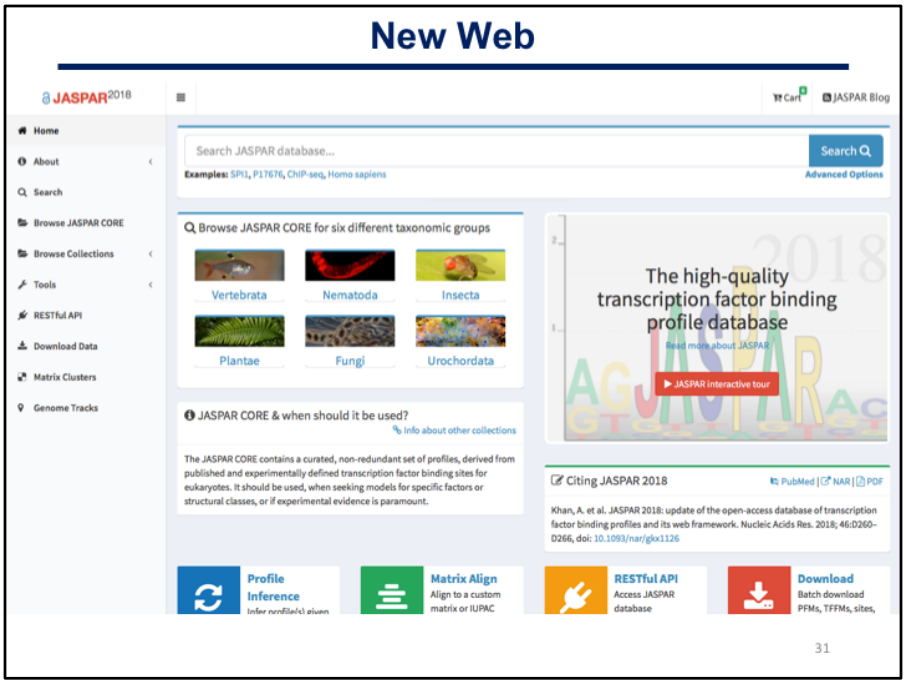
http://expdata.cmmt.ubc.ca/JASPAR/downloads/UCSC_tracks/2018/

29

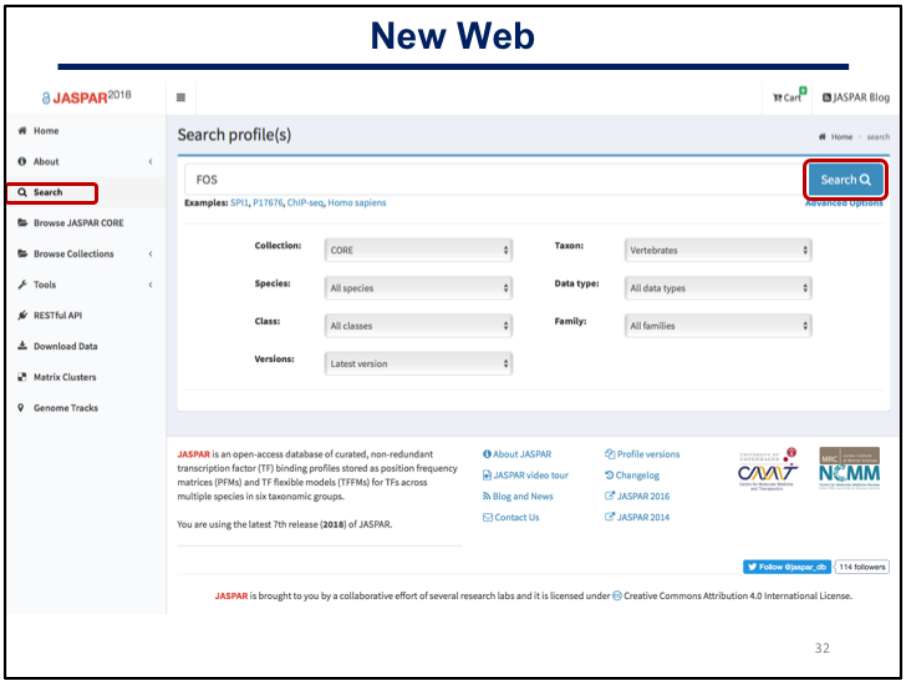
To date, we offer genome tracks for all these species and genome assemblies via UCSC, Ensembl and the Track Hub Registry. All TFBS predictions and underlying BED files are available in our web servers.



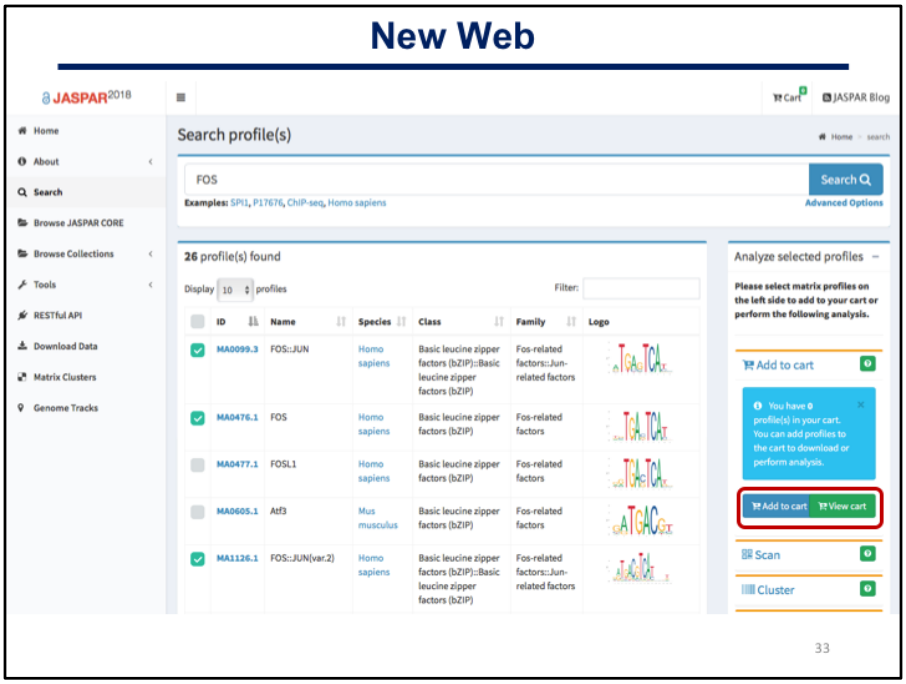
One of the most significant updates of this release is the new, more user friendly web interface of JASPAR.



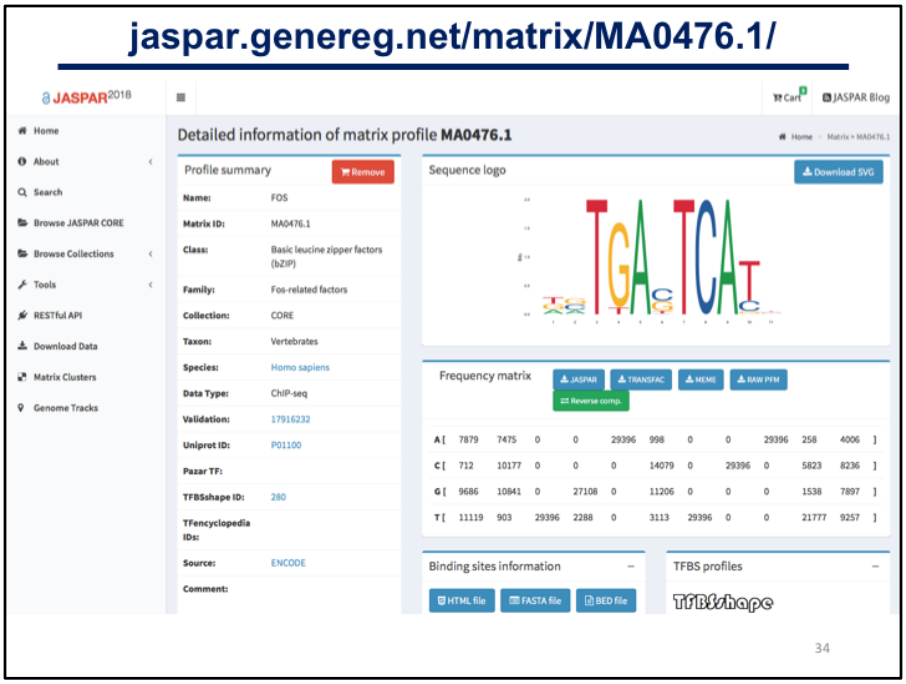
Here is an overview of the new web site. I remember when Anthony first interviewed me, he asked me why I hadn't used JASPAR during my PhD research. I told him that the main reason was that it was very complicated to navigate. Since then, we had been talking about updating the web, making it more 2.0.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



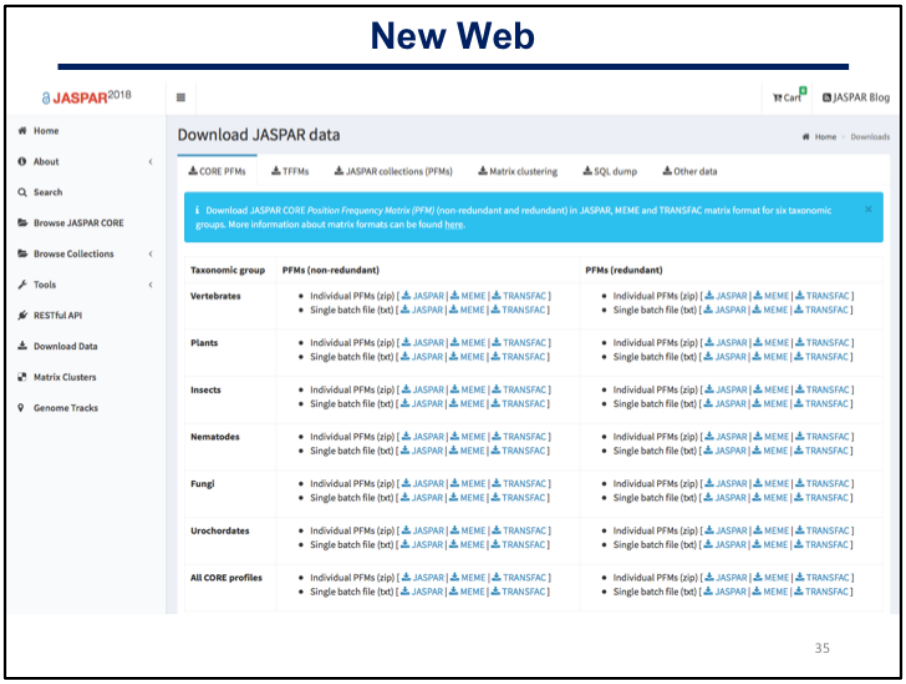
The web comes with an improved search engine. It allows, for example, search for a given TF name in a specific collection, version or taxon.



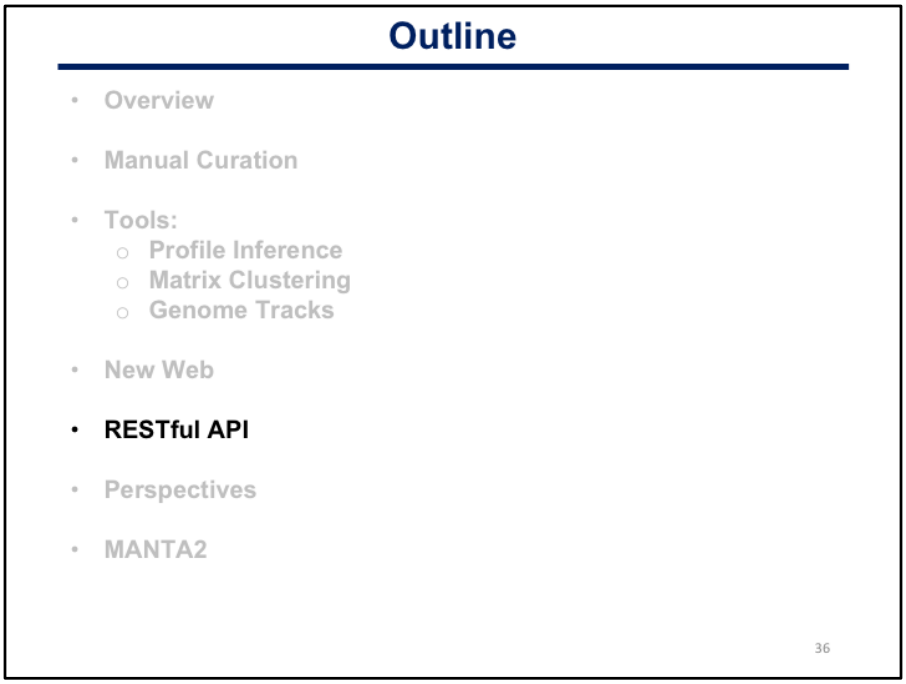
And here are the resulting profiles matching the criteria of our search, which can be, for example, selected and added to cart for custom analysis.



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



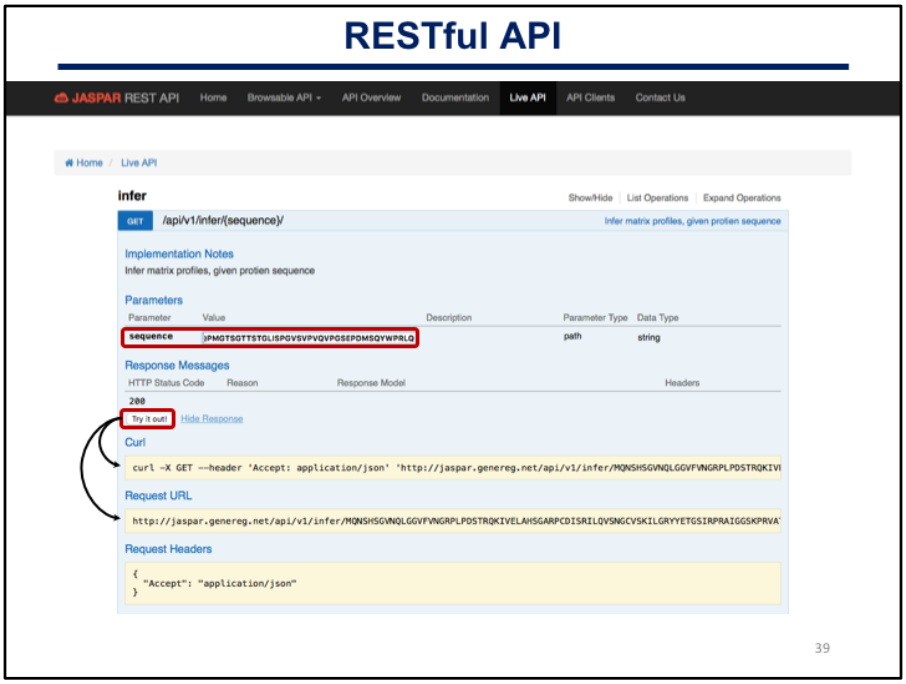
Finally, the downloads section allows for downloading profiles per taxonomic group in different formats, as a single batch or one file per profile, or the entire database.



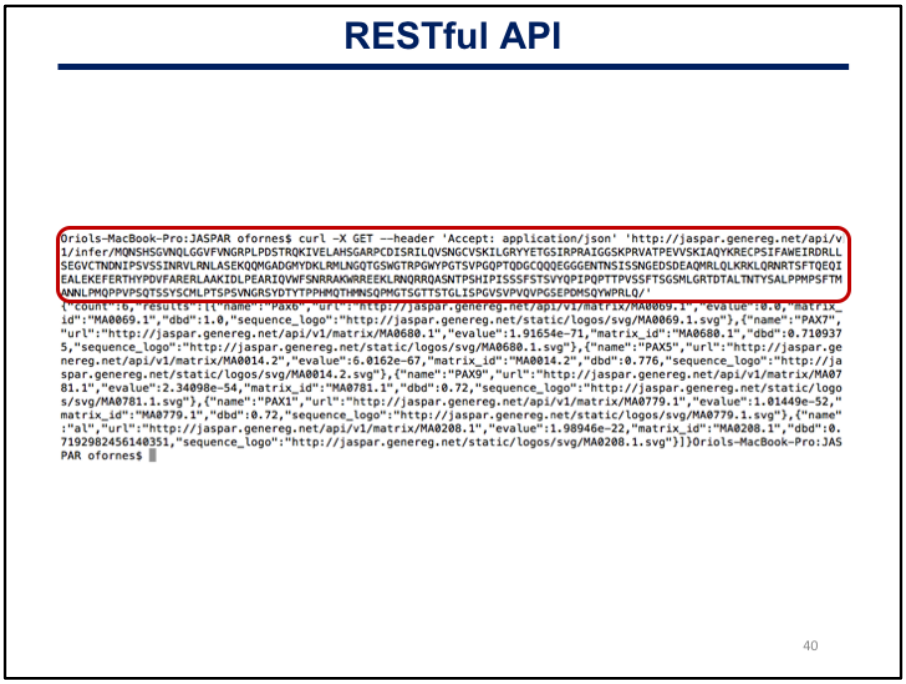
Probably, the second most significant update to JASPAR 2018 is its RESTful API.

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

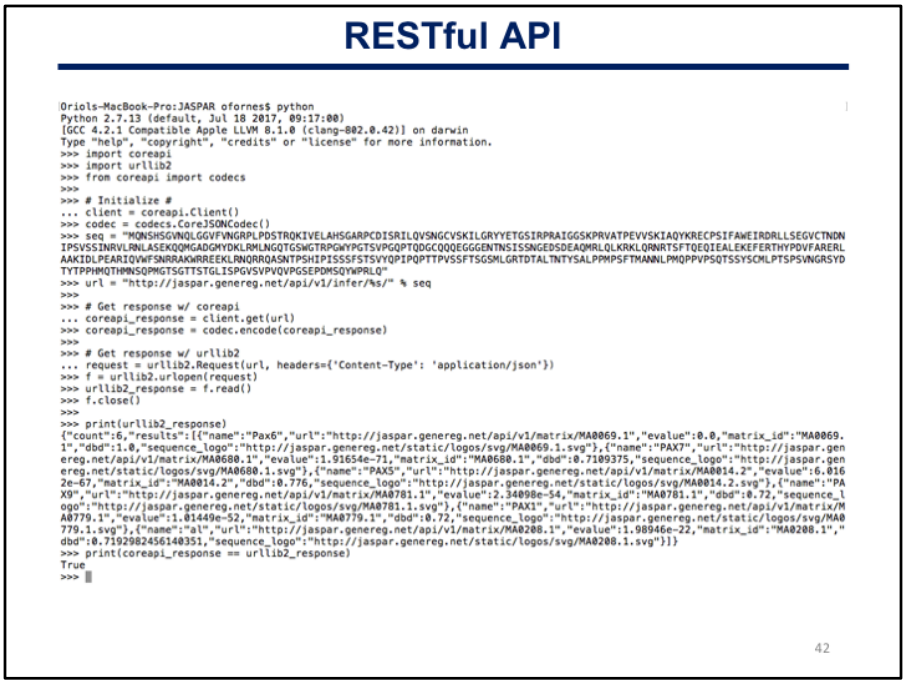


For example it provides access to the Profile Inference tool. By pasting a sequence and clicking on try it out, it provides two ways of fetching the inferred profiles: either using curl or through a URL request.



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

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.



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

Outline

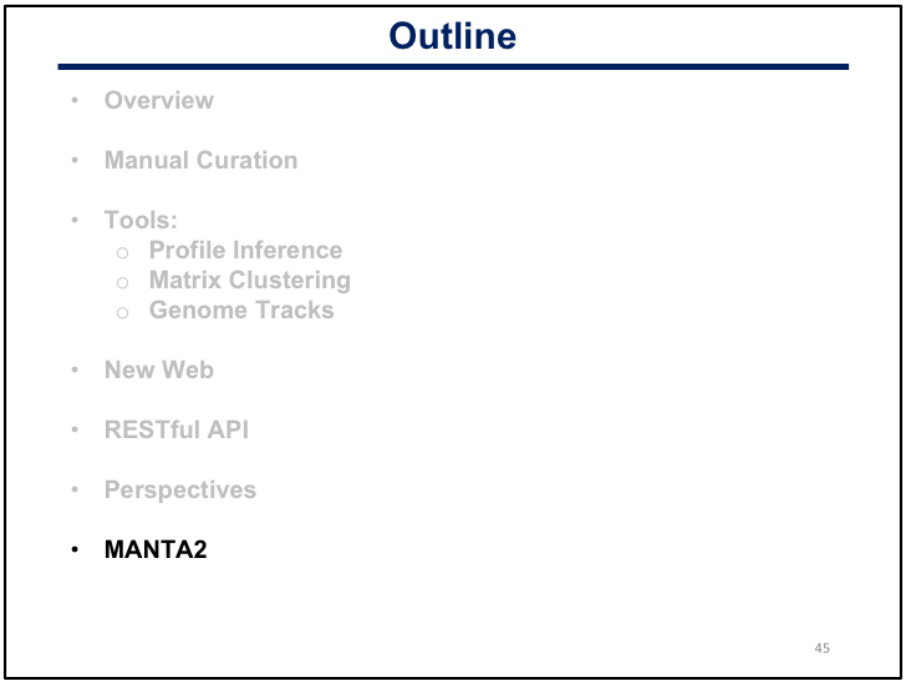
- Overview
- Manual Curation
- Tools:
  - Profile Inference
  - Matrix Clustering
  - Genome Tracks
- New Web
- RESTful API
- **Perspectives**
- MANTA2

43

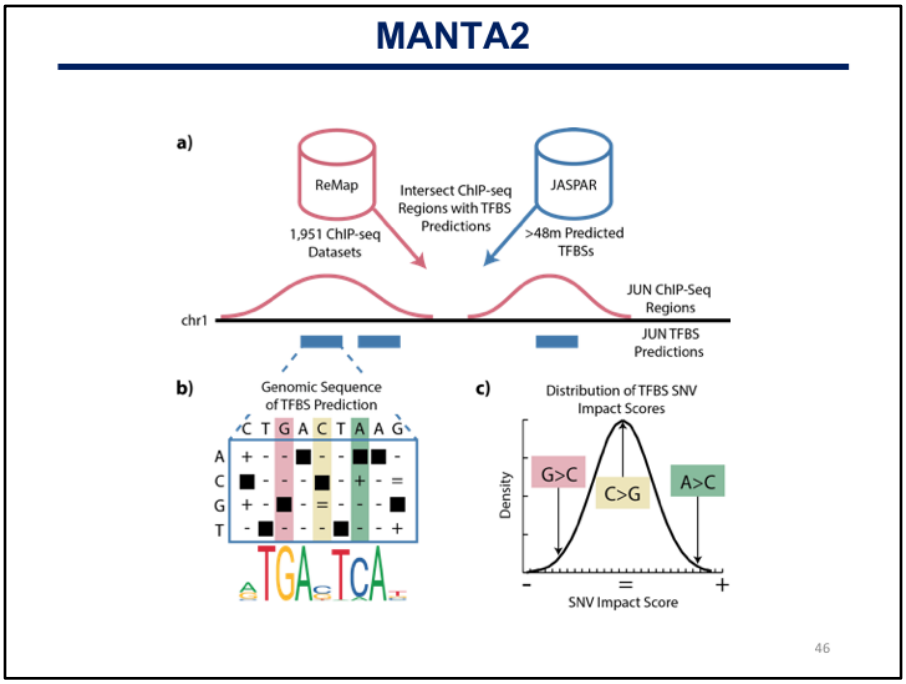
Perspectives

- 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.

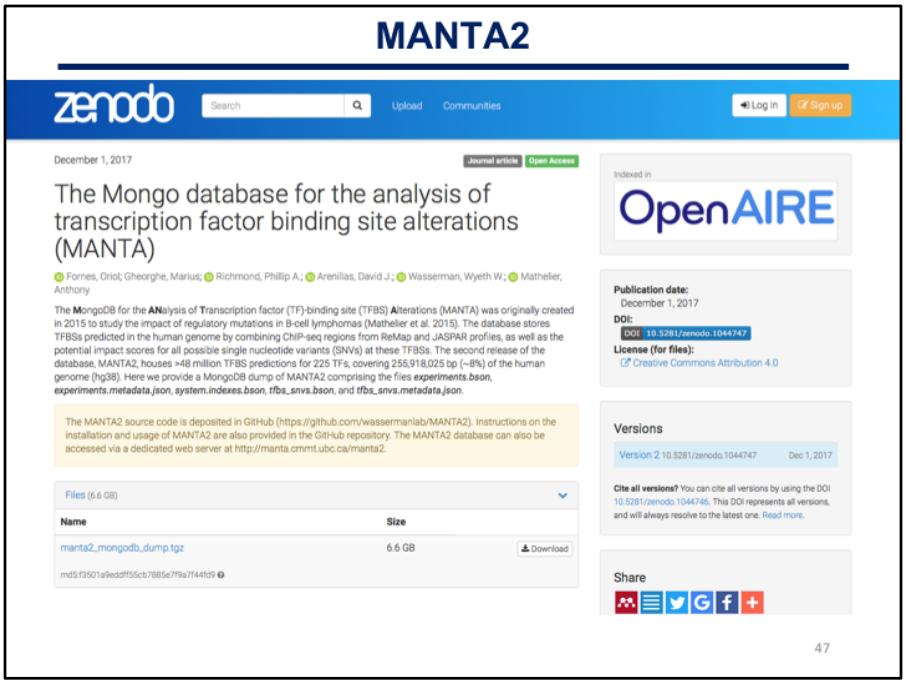
44



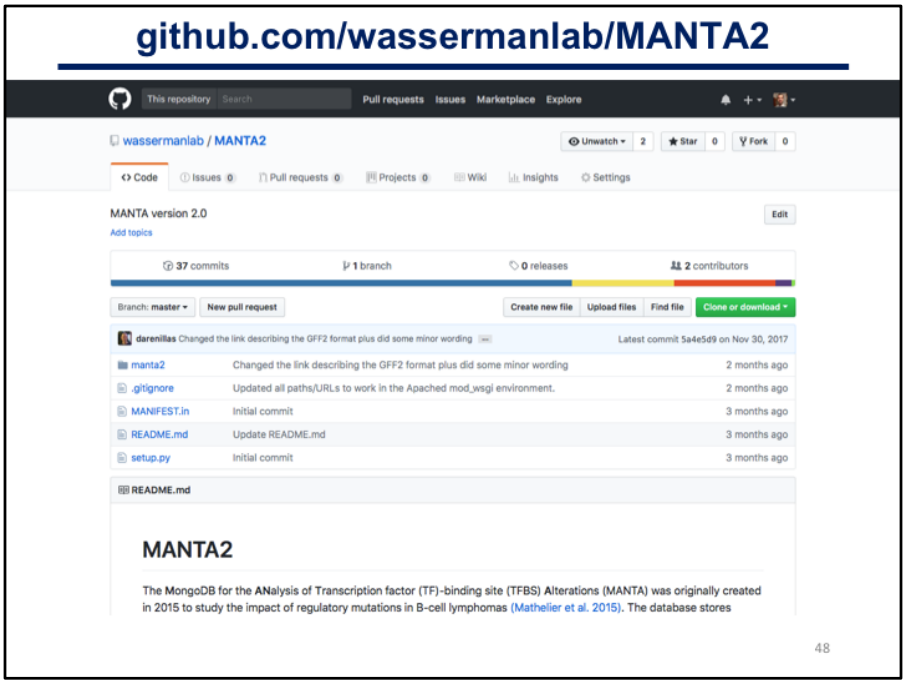
To finish my presentation, I would like to introduce you the 2018 update of our Mongo database for the analysis of TF binding site alterations MANTA. This work has been done in collaboration with the Mathelier Lab, specially with help from Marius here (point towards him).



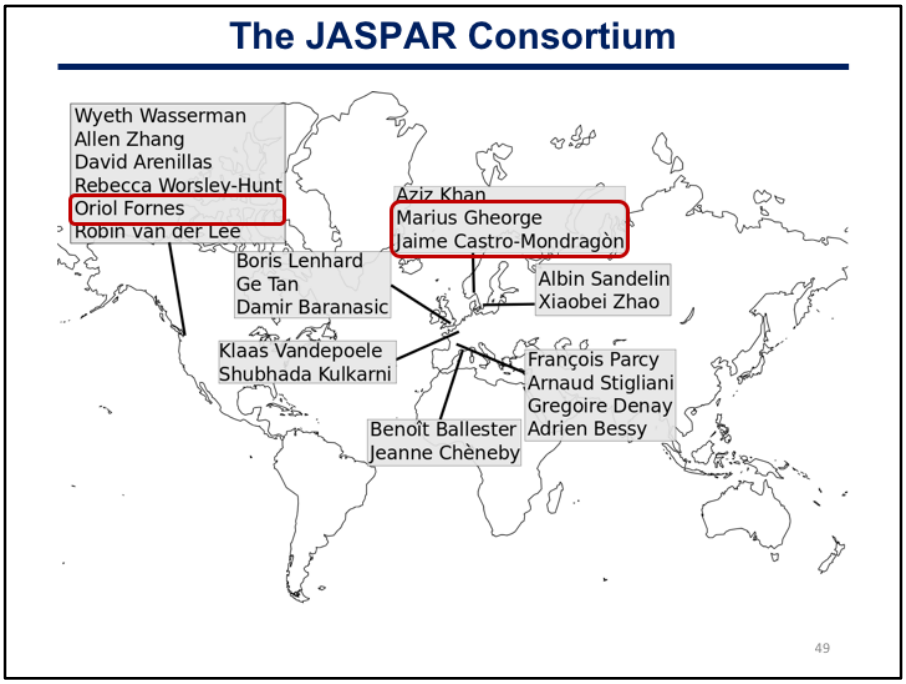
MANTA stores TF binding site predictions within ChIP-seq regions, as well as the potential impact of all single nucleotide variants within these binding sites on TF binding. Benefiting from the recent updates of ReMap and JASPAR, we intersected the ReMap ChIP-seq regions with the JASPAR genome-wide TFBS 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.



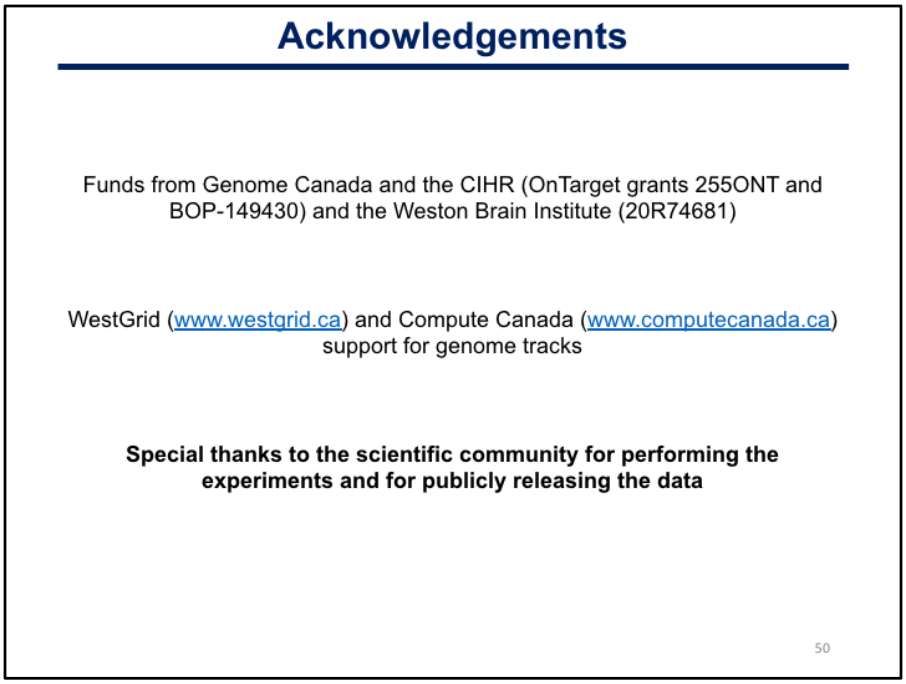
MANTA2 is available for download at Zenodo, with scripts granting programmatic access to the database in a GitHub repository.



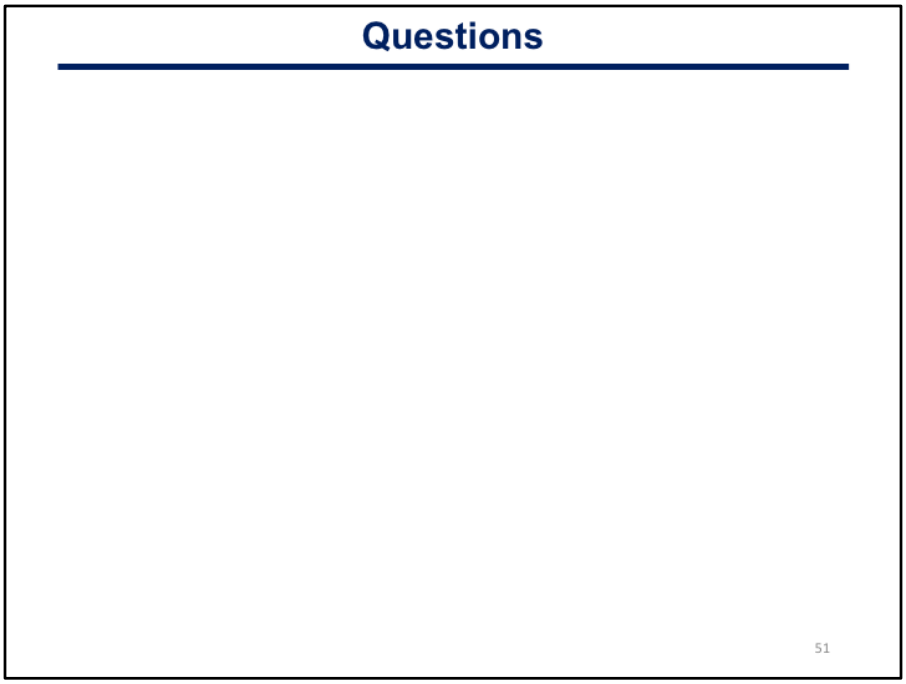
MANTA2 is available for download at Zenodo, with scripts granting programmatic access to the database in a GitHub repository.



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



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



And now I will be taking in questions.

TF Coverage for Human

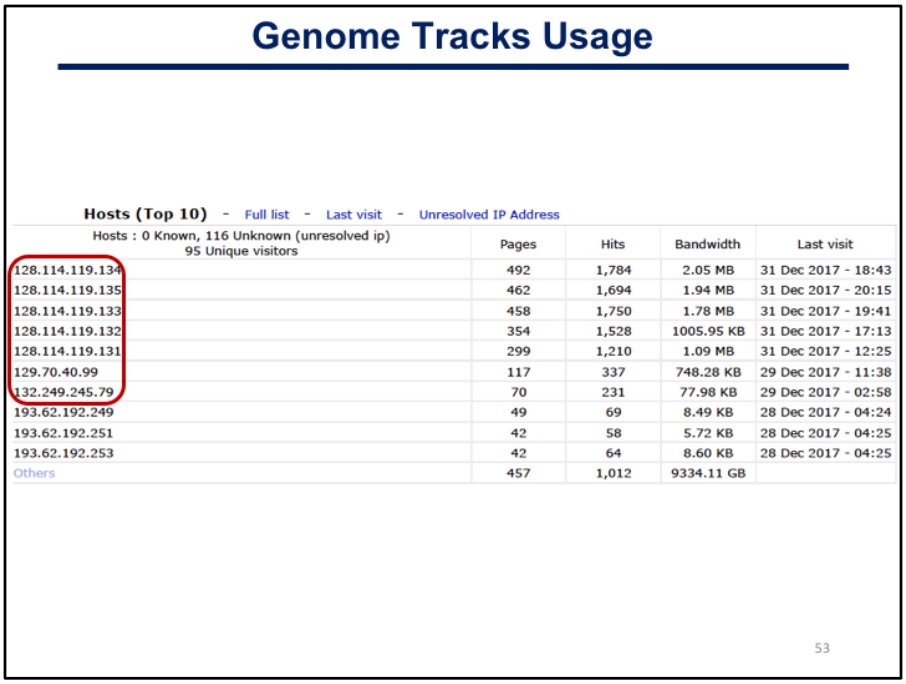
The Human Transcription Factors

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[5,7,8,*]
[1]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
[2]Donnelly Centre, University of Toronto, Toronto, ON, Canada
[3]Genome-Scale Biology Program, University of Helsinki, Helsinki, Finland
[4]Division of Functional Genomics and Systems Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Solna, Sweden
[5]Center for Autoimmune Genomics and Etiology (CAGE), Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[6]Department of Biochemistry, Cambridge University, Cambridge CB2 1GA, United Kingdom
[7]Divisions of Biomedical Informatics and Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[8]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA
[9]These authors contributed equally
*Correspondence: ajt208@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…



Describe this slide…