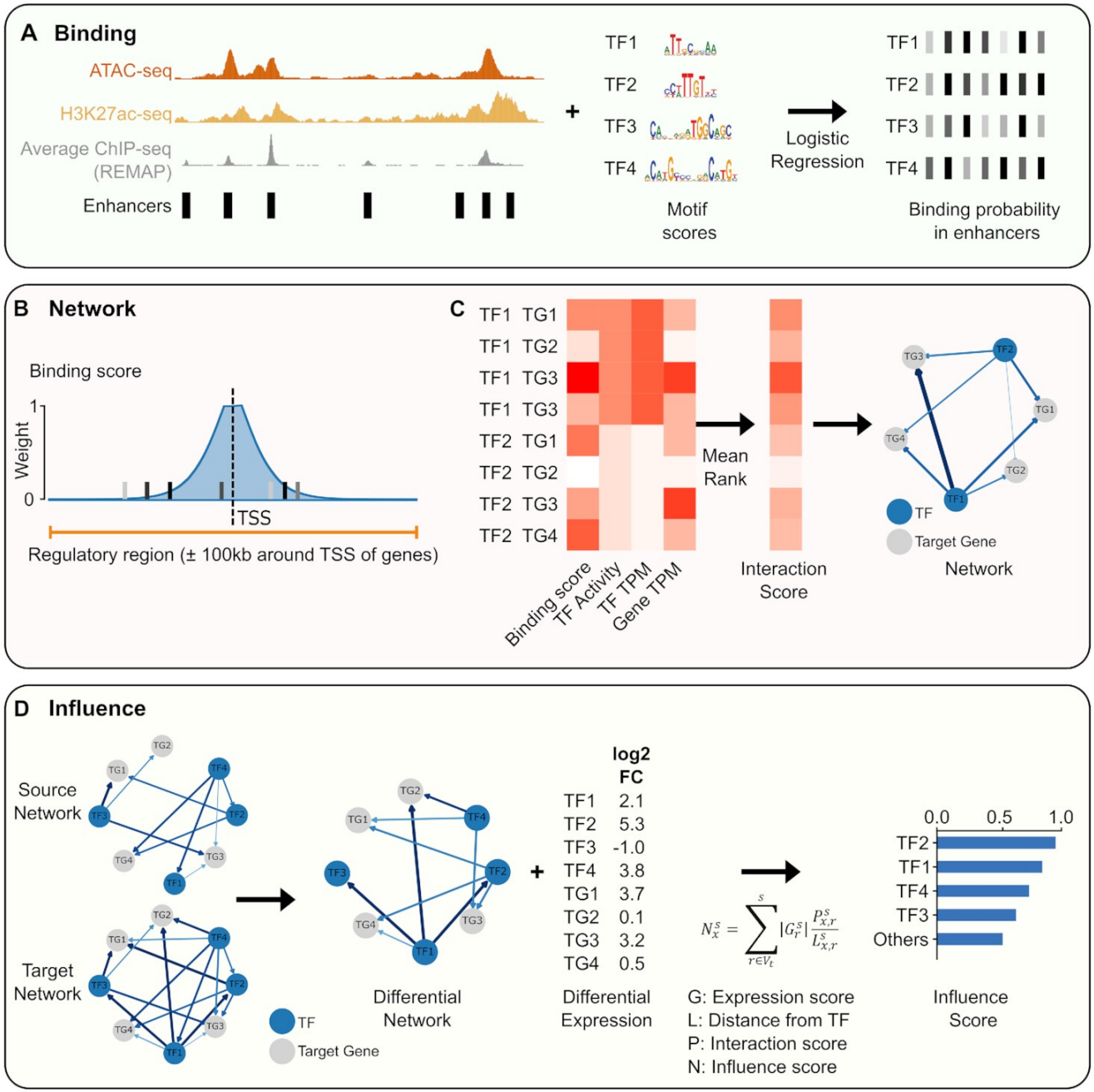
ANANSE and Cavefish: A love/hate history

Ale Gil

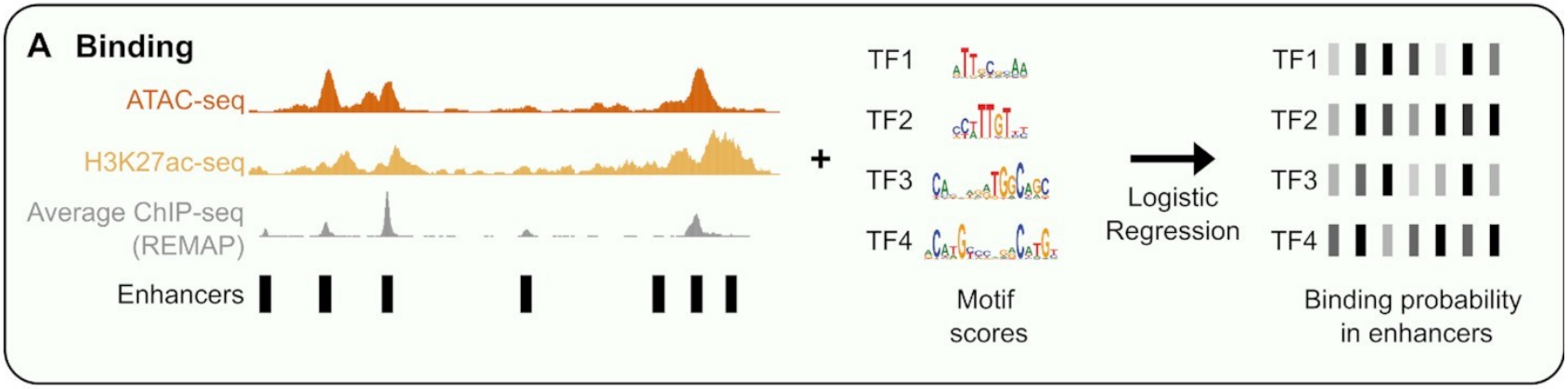
Fishmeeting @ CABD

11/11/2022

How does ANANSE work? An outline



ANANSE outline: Binding

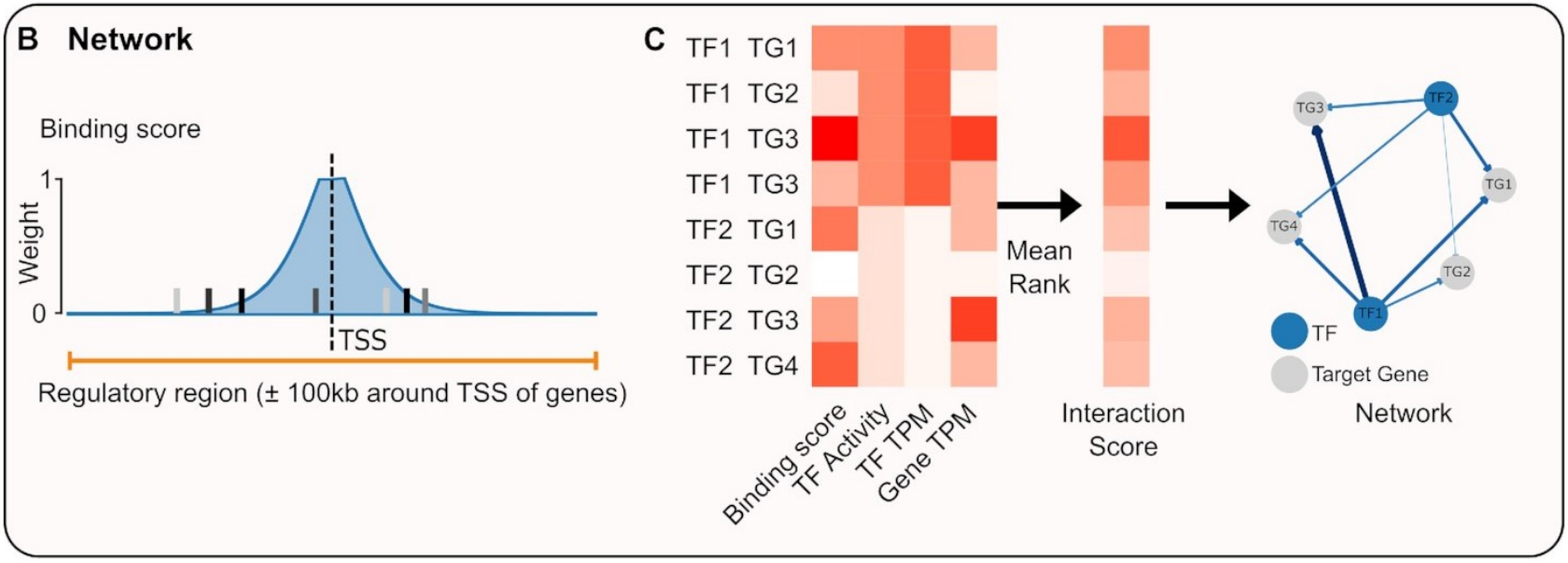


How does `binding` works?

The `binding` command of ANANSE is the one that generates probabilities of TF binding in all the enhancers. It does this integrating different information:

- ATAC-seq and H3K27ac ChIP signal (this last optional). This is the enhancer activity.
- Enhancer positions. Either you provide it or ANANSE compute them from the data above.
- TF motif scores. ANANSE has a predetermined set of motifs that uses to scan the genome

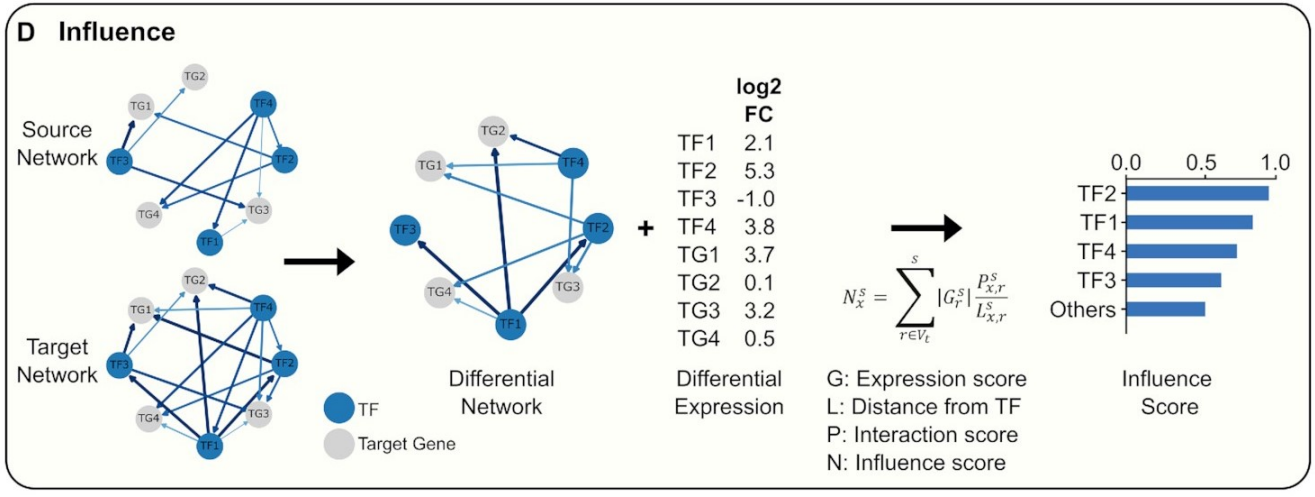
ANANSE outline: network



What is network doing?

The `network` command integrates the previously computed binding information and computes the interaction score between a TF factor and target genes. It integrates information from RNA-seq, the distance to the target, and interactions with other genes.

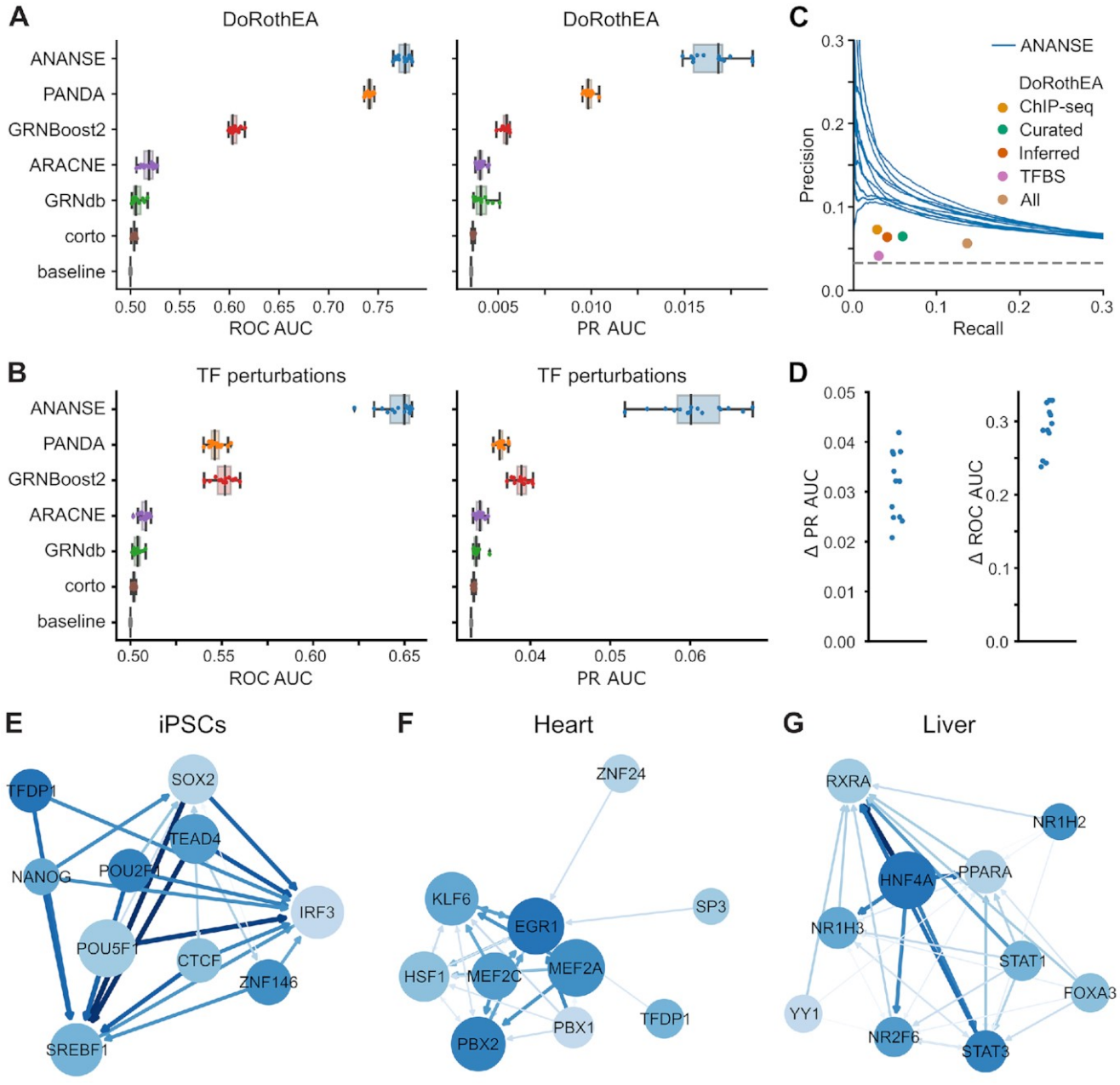
ANANSE outline: Influence (the cool thing)



Why the cool thing?

This is the most novel feature of ANANSE. It can not only compute GRN, but also compare them. You can know which TF are essential for moving from one condition (network) to another. It integrates the GRN structure with differences in the expression of each target gene.

ANANSE outline: Final thoughts



But that's fantastic, where is the hate part of the history?

It's made for human & mouse data

Cavefish data outline

Data that we have

- **ATAC-seq**: We have ATAC-seq data for 80% epiboly, 5 somites, 24hpf and 48hpf
- **RNA-seq**: We have transcriptomics for 10hpf(5ss), 24hpf, 36hpf and 72hpf

We can explore how the GRNs have change from Cavefish to Surfacefish using ANANSE. Also we can know the TFs that have driven the main changes in GRN.

Cavefish step by step: Generating our own motif database

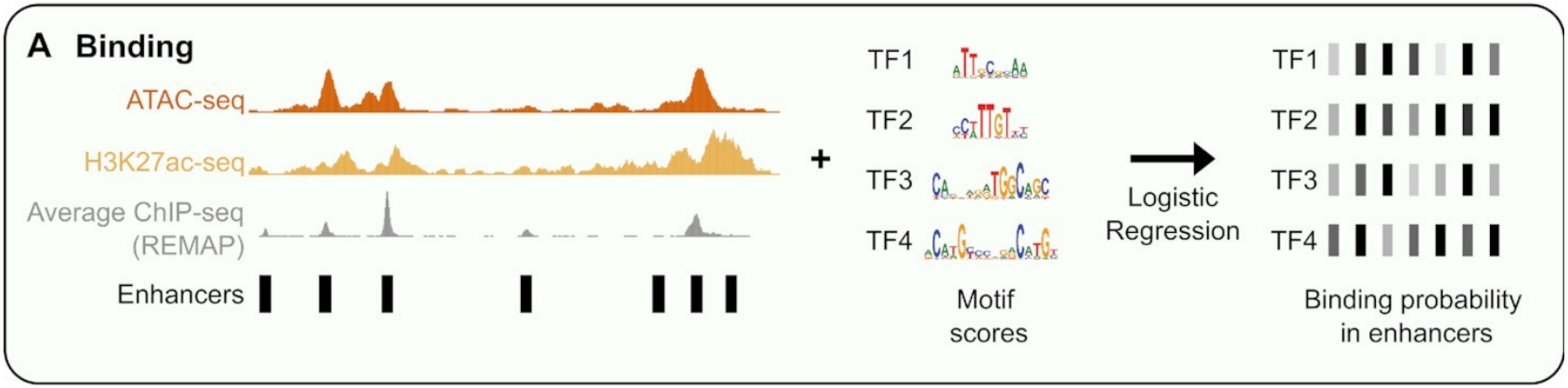
We can use `motif2factors` from the toolset `gimme-motif` in order to generate our own motif database based on orthology.

```
gimme motif2factors --new-reference AstMex2 \
--outdir results_motif2factors \
--threads 8
```

Considerations

If your genome is in UCSC/ENSEMBL/NCBI use that assembly name after `-new-reference`.
For example: danRer10 for zebrafish; xenTro10 for Xenopus; sScyCan1.2 for catshark; `Nfu_20140520` for killi.
If that's not the case we will have to use a FASTA with GeneIDs and peptide sequences. (see here)

Cavefish step by step: ANANSE binding



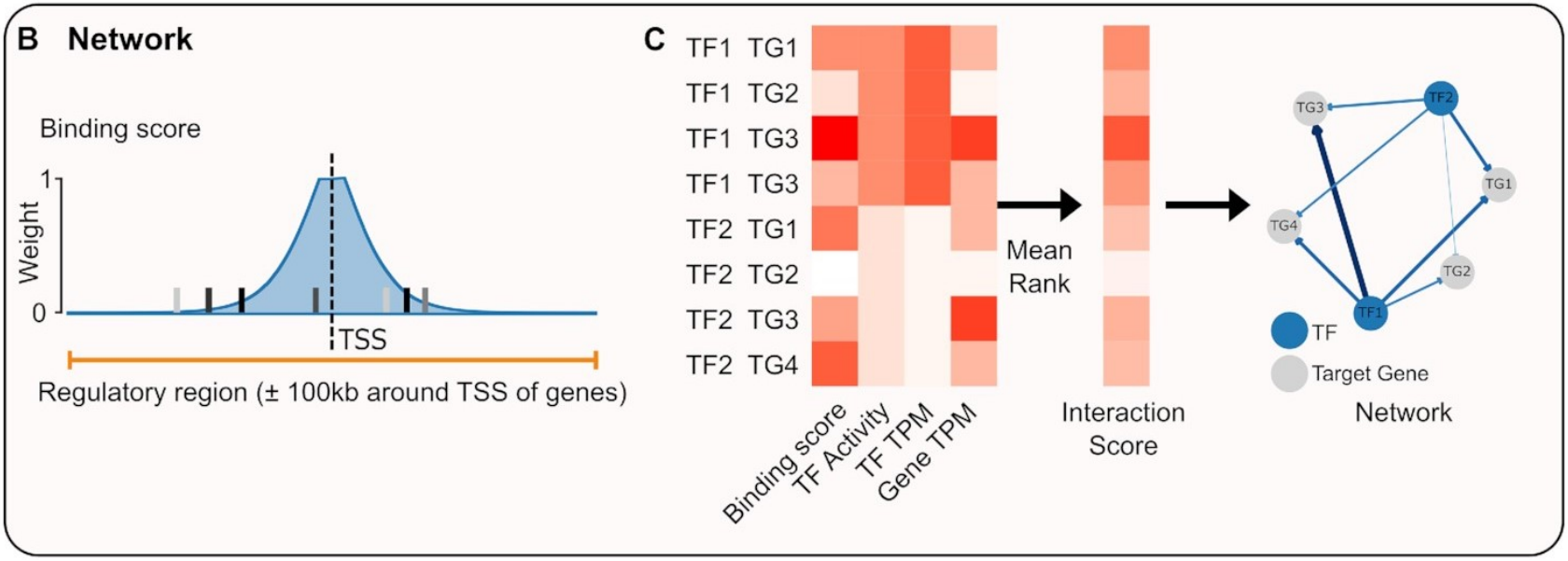
Once we have our motif database with our genes, we can compute the first step of ANANSE, the binding.

```
ananse binding -A Cave5ss_rep1.bam Cave5ss_rep2.bam \
-g AstMex2 -p Astyanax_mexicanus-2.0.pfm \
-o cave_binding_5ss \
-r Cave_5ss_IDR_peaks.bed -n 12
```

Considerations

We can use a custom fasta after -g option. Bam files **must** be indexed. You need `your_sample.bam.bai` in the same folder than `your_sample.bam`

Cavefish step by step: ANANSE network



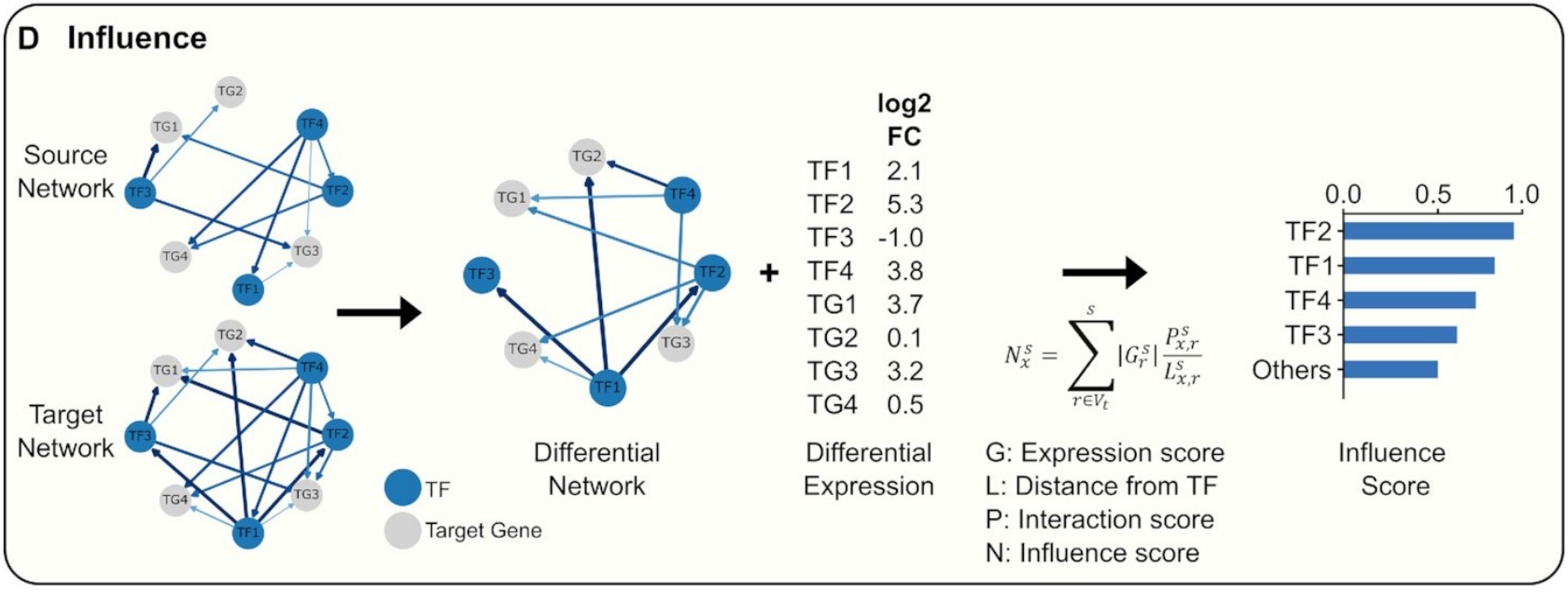
For generating the network we need to provide the binding output and the RNAseq data maped with Salmon or Kallisto

```
ananse network -n 4 -e RNASeq_cave_5ss.tpm \
-g AstMex2 -o cave_5ss.network \
cave_binding_5ss/binding.h5
```

Considerations

Do not increase -n above of 4, the program is EXTREMELY memory hungry.

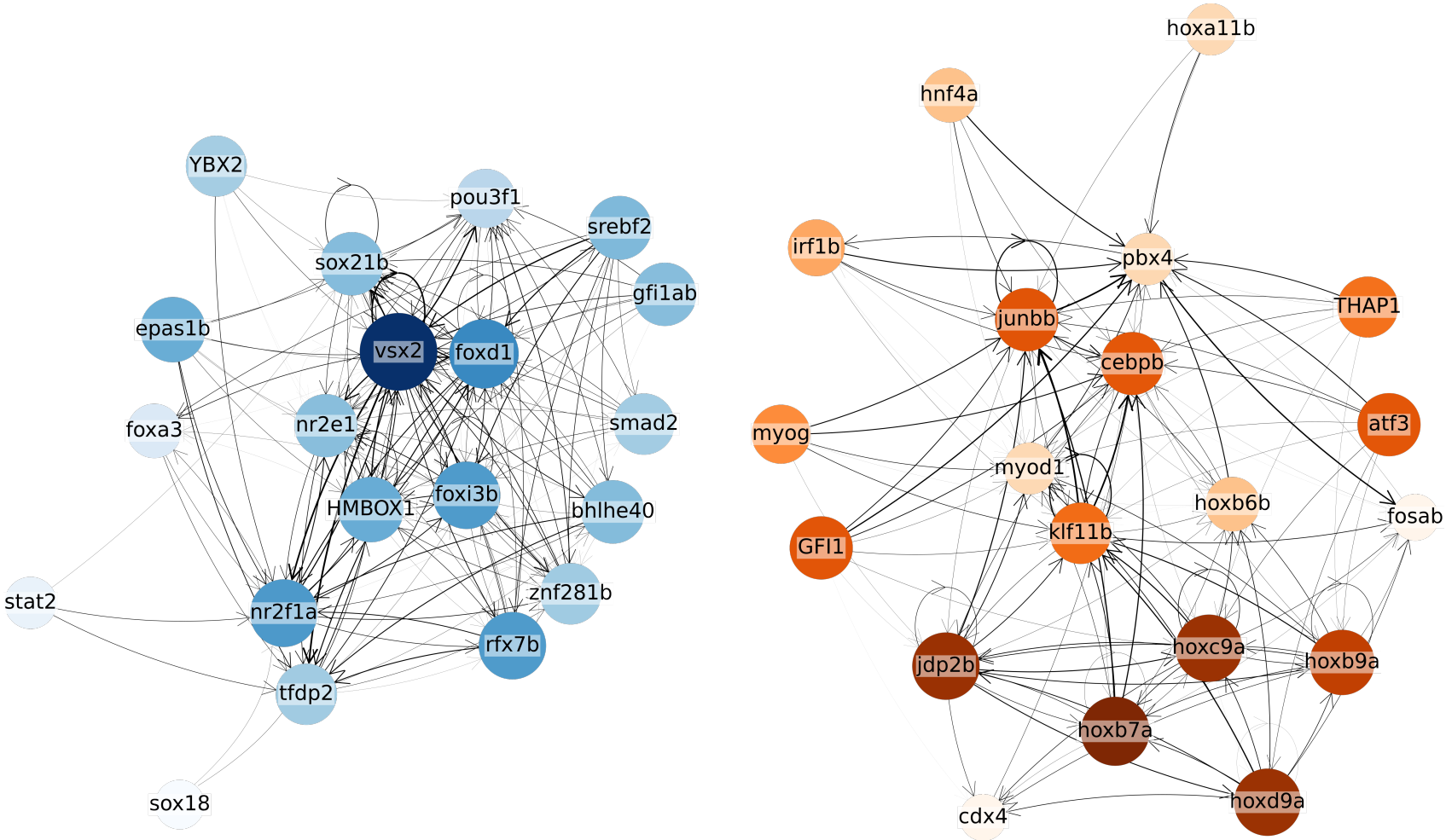
Cavefish step by step: ANANSE influence



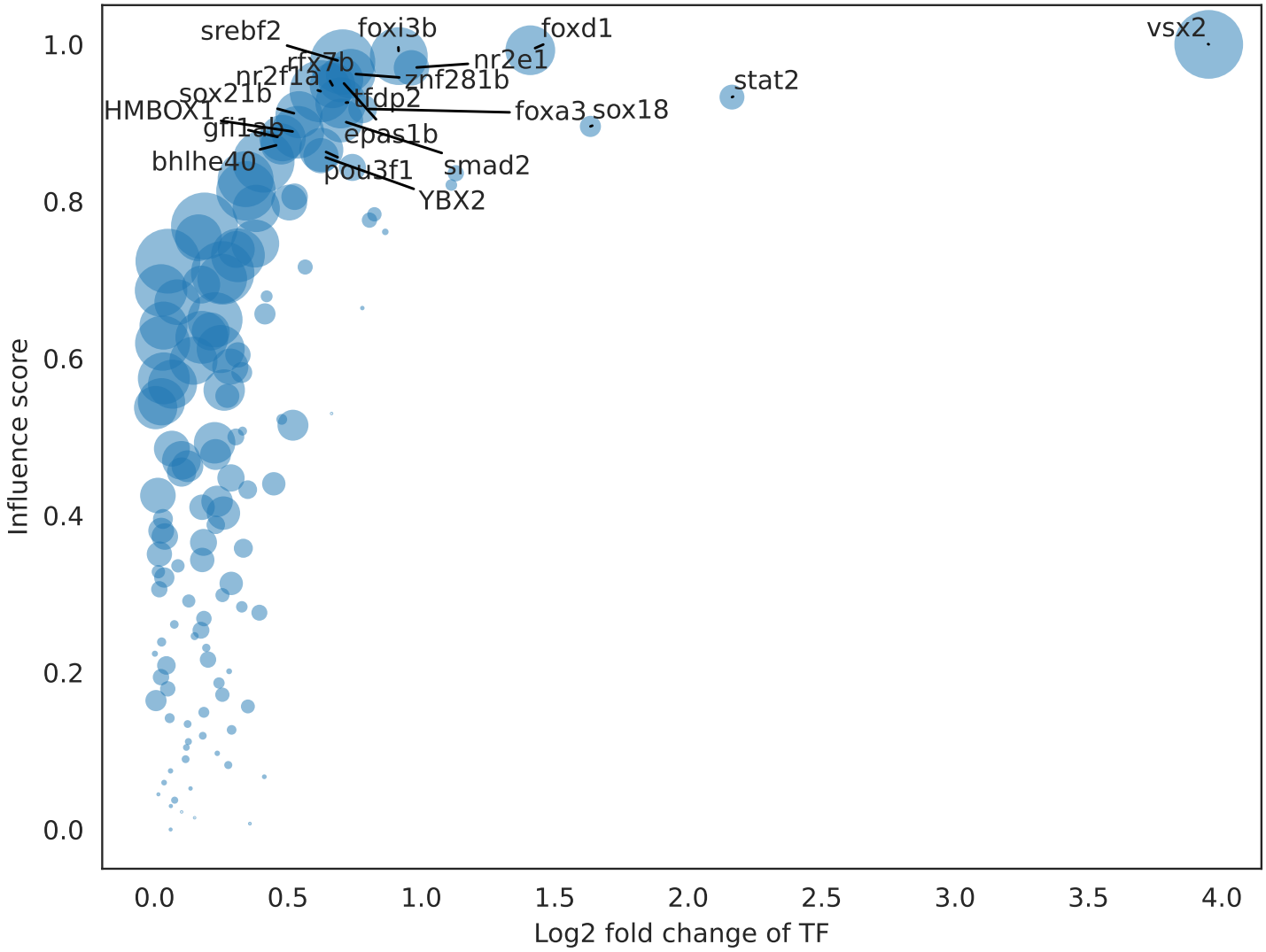

```
ananse influence -s results/cave_5ss.network \
-t results/surface_5ss.network \
-d cave_vs_SF_DE_genes.txt \
-o results/influence_cave_2_surface.txt \
-n 8
```

Okey, now what?

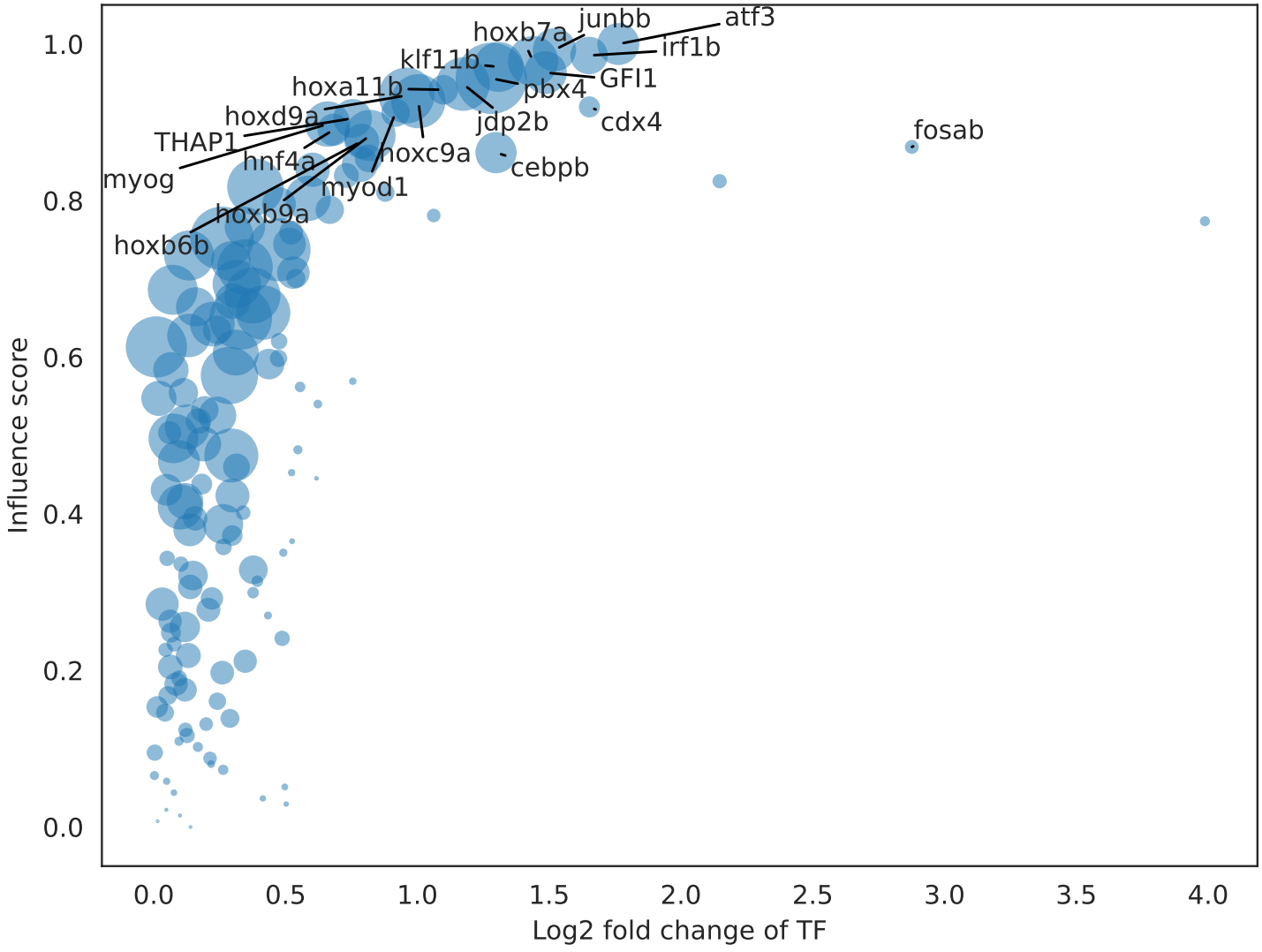
Cavefish results



We can see which are the TF that are driving differences in GRN between the two populations.



Cave to surface most influent TFs at 24hpf



Surface to cave most influent TFs 24hpf

Possibilities of analysis

- We can study which genes have more input or output (regulators or efectors)
- We can study the centrality of the TF in their GRN (their importance)
- We can annotate genes in the GRN according to the literature

Bonus track: Salmon to generate RNAseq TPM

Steps

- Generate the index for salmon. You will need a .fna file with 1 transcript per gene (DNA)
- Quantify the samples. You don't even need to map the FASTQs!
- Merge the TPM files in one matrix for ANANSE. I use a custom R script.

```
salmon index -t transcripts.fa -i transcripts_index -k 31

salmon quant -i transcripts_index -l U -r reads.fq \
--validateMappings -o transcripts_quant
```

If you have any doubts they have done a great job with the documentation.

Thanks a lot!

Links of interest

- Gimme motifs toolset
- ANANSE documentation
- Conda cheat sheet. For installing everything correctly.
- Salmon guide
- Seq2science pipeline (Advanced bioinformatics). This pipeline deals greatly with ATAC and RNA-seq, specially if the genome is in ENSEMBL or UCSC.

Generalized guide of Ananse

Ananse has a new version that can be easily installed on your computer without much problem. You can install it with the conda commands below.

```
conda update -n base conda
conda config --add channels defaults
conda config --add channels bioconda
conda config --add channels conda-forge

conda create -n ananse -c bioconda ananse
```

Considerations

These commands can take time to run. Run them in a screen if you don't want to have the connection dropped.

Generalized guide of Ananse

This new version integrates genomepy. So if you are lucky and your genome is in UCSC you can save a lot of problems. To install a genome of genomepy you do it as follows (example with danrer10).

- -p Is the provider name: for example UCSC
- -a Tells genomepy to also download gene annotation (you want to do this)

```
genomepy install -p UCSC -a danRer10
```

Considerations

You have to activate the environment each time you want to work with Ananse. You do that with the next command. Remember to do it to work with ananse!!!

```
conda activate ananse
```

Generalized guide of Ananse

The first step when working with non-human/mouse organisms in ananse is to generate the database of TF motifs. This can be done easily thanks to the gimmemotif toolset (already installed in the enviroment along ananse).

```
gimme motif2factors --new-reference danRer10 --outdir motifI
```

Considerations

You have to activate the environment each time you want to work with Ananse. You do that with the next command. Remember to do it to work with ananse!!!

```
conda activate ananse
```

Generalized guide of Ananse

Now you can run the binding with the new database. Go to 11 and follow the examples. You only need to change the -g argument and -p arguments (example below). The -p argument is the pfm file that you have generated before.

```
ananse binding -A danrer10_rep1.bam danrer10_rep2.bam \
-g danRer10 -p motifDB_danRer10/danRer10.pfm \
-o cave_binding_5ss \
-r Cave_5ss_IDR_peaks.bed -n 12
```

Considerations

You have to activate the environment each time you want to work with Ananse. You do that with the next command. Remember to do it to work with ananse!!!

```
conda activate ananse
```