(Negative) controls

Morgane Thomas-Chollier

Computational systems biology - IBENS

mthomas@biologie.ens.fr

M2 – Computational analysis of cis-regulatory sequences 2015/20165

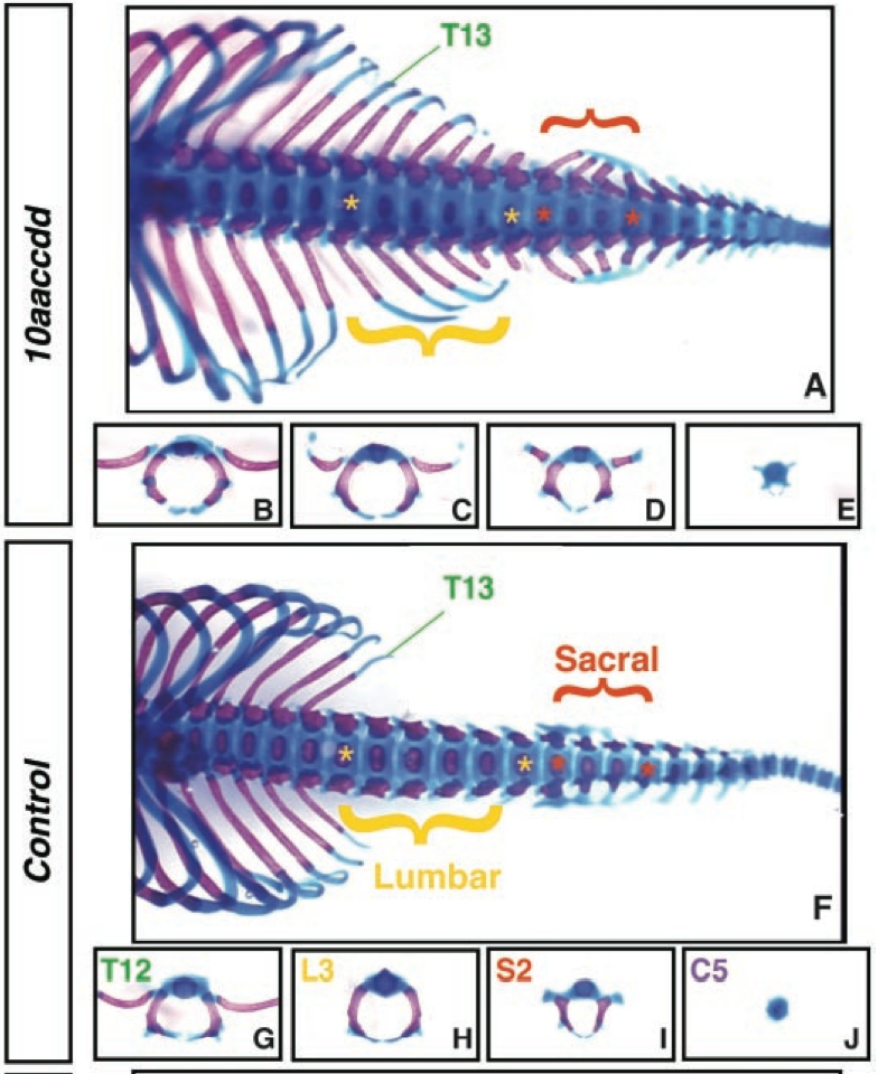
Denis Thieffry, Jacques van Helden and Carl Herrmann kindly shared some of their slides.

Aim of the course

1 – Understand the need for controls in bioinformatics

2 – Some strategies to build controls

Controls in biology



Wellik and Mario R Capecchi, Science, 2003

Evaluate predictions with controls

- Quantify the capability of the program to
  - » detect known features
  - » = Return a positive answer for a positive feature
  - » Not detect false features
  - » = Return a negative answer for a negative feature

Predictions

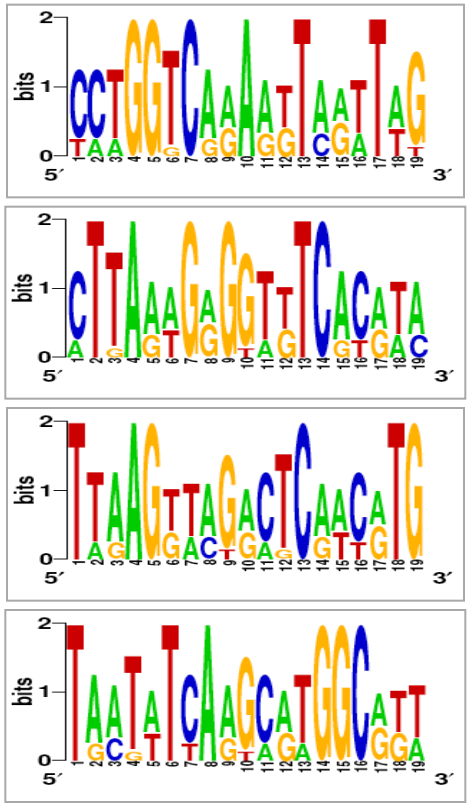
Annotation	**Positive**	**Negative**
Positive	**True Positive**	False negative
Negative	False Positive	**True Negative**

In the context of cis-regulation

Use different set of *sequences*

```
5′- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT ...HIS7
5′- ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG ...ARO4
5′- CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT ...ILV6
5′- TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC ...THR4
5′- ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA ...ARO1
5′- ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA ...HOM2
5′- GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA ...PRO3
```

Use different set of *matrices*



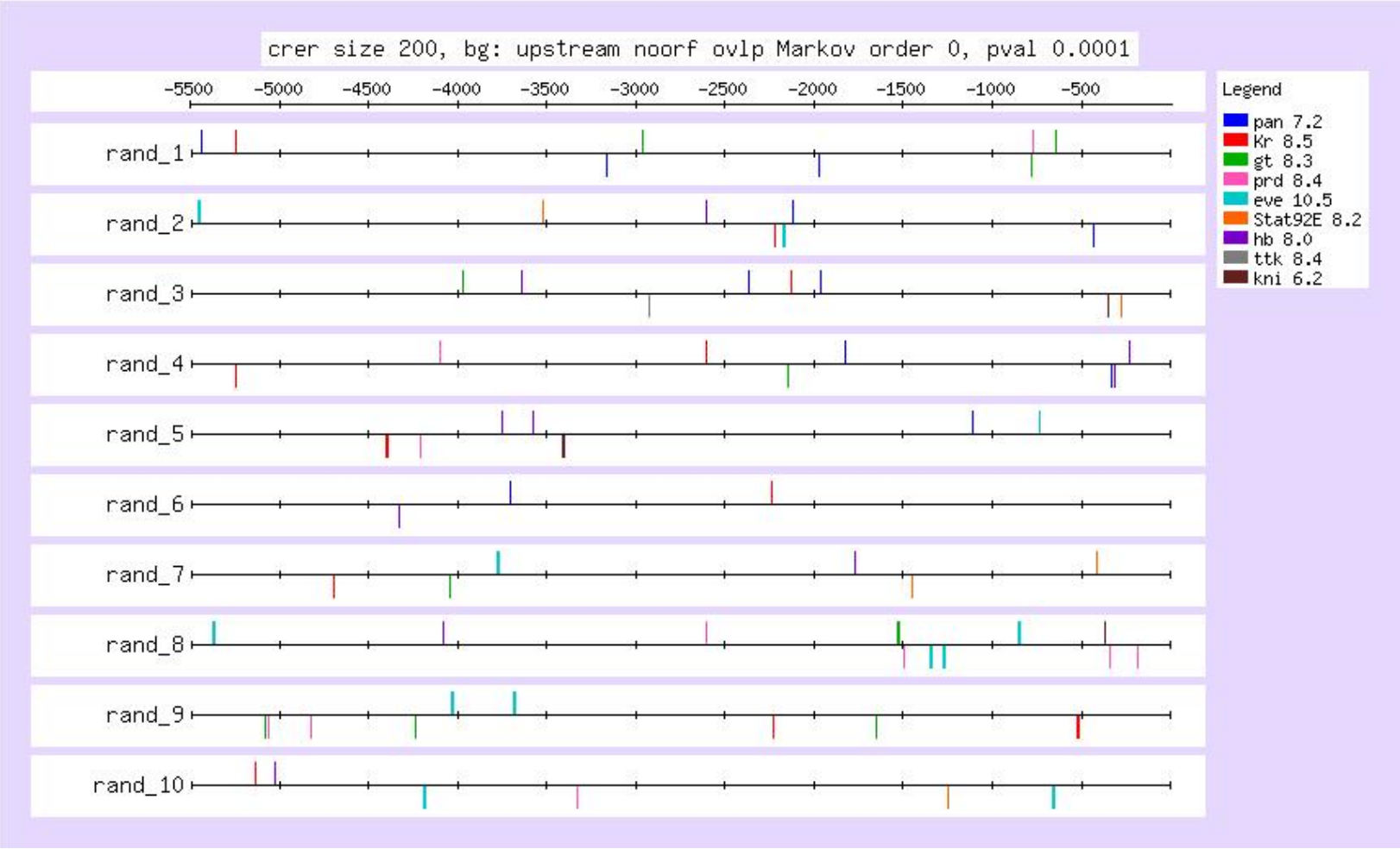
Sequences

- **Positive control**: quantify the capability of the program to detect known regulatory elements
  - » Annotated sites (e.g. sites from TRANSFAC) in their original context (the promoter sequences).
  - » Annotated sites implanted in other context
    - Biological sequences (random selection).
    - Artificial sequences.
  - » Artificial sites implanted in artificial sequences.

- **Negative control**: quantify the capability of the program to return a negative answer when there are no regulatory elements.
  - » Artificial sequences
    (generated according to a Bernoulli or a Markov model to mimic an organism of interest)
  - » Biological sequences without common regulation
    (random selection of genes)

Artificial sequences

- **Random-seq** in RSAT
  - » Generate artificial sequences (mimicking real biological sequences)
  - » Re-run the exact same analysis



Randomized (shuffling) sequences

- **Randomized sequences**
  - » Maintain composition (=nb of A,C,G,T)
  - » Conservation of higher-order dependencies ?
  - » Is it likely that the signal is still there ?

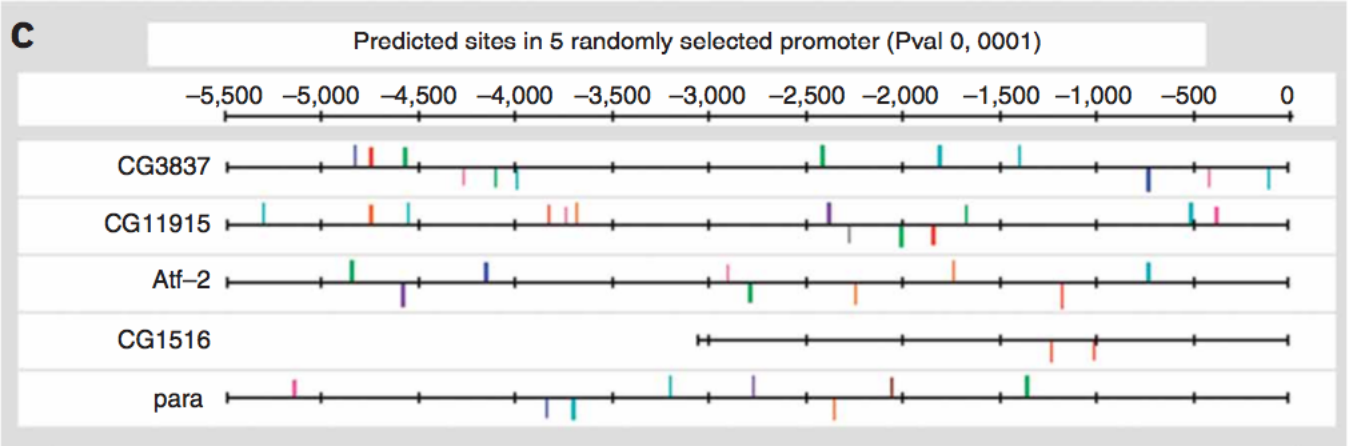
Sequences

- **Positive control**: quantify the capability of the program to detect known regulatory elements
  - » Annotated sites (e.g. sites from TRANSFAC) in their original context (the promoter sequences).
  - » Annotated sites implanted in other context
    - Biological sequences (random selection).
    - Artificial sequences.
  - » Artificial sites implanted in artificial sequences.

- **Negative control**: quantify the capability of the program to return a negative answer when there are no regulatory elements.
  - » Artificial sequences
    (generated according to a Bernoulli or a Markov model)
  - » Biological sequences without common regulation
    (random selection of genes)

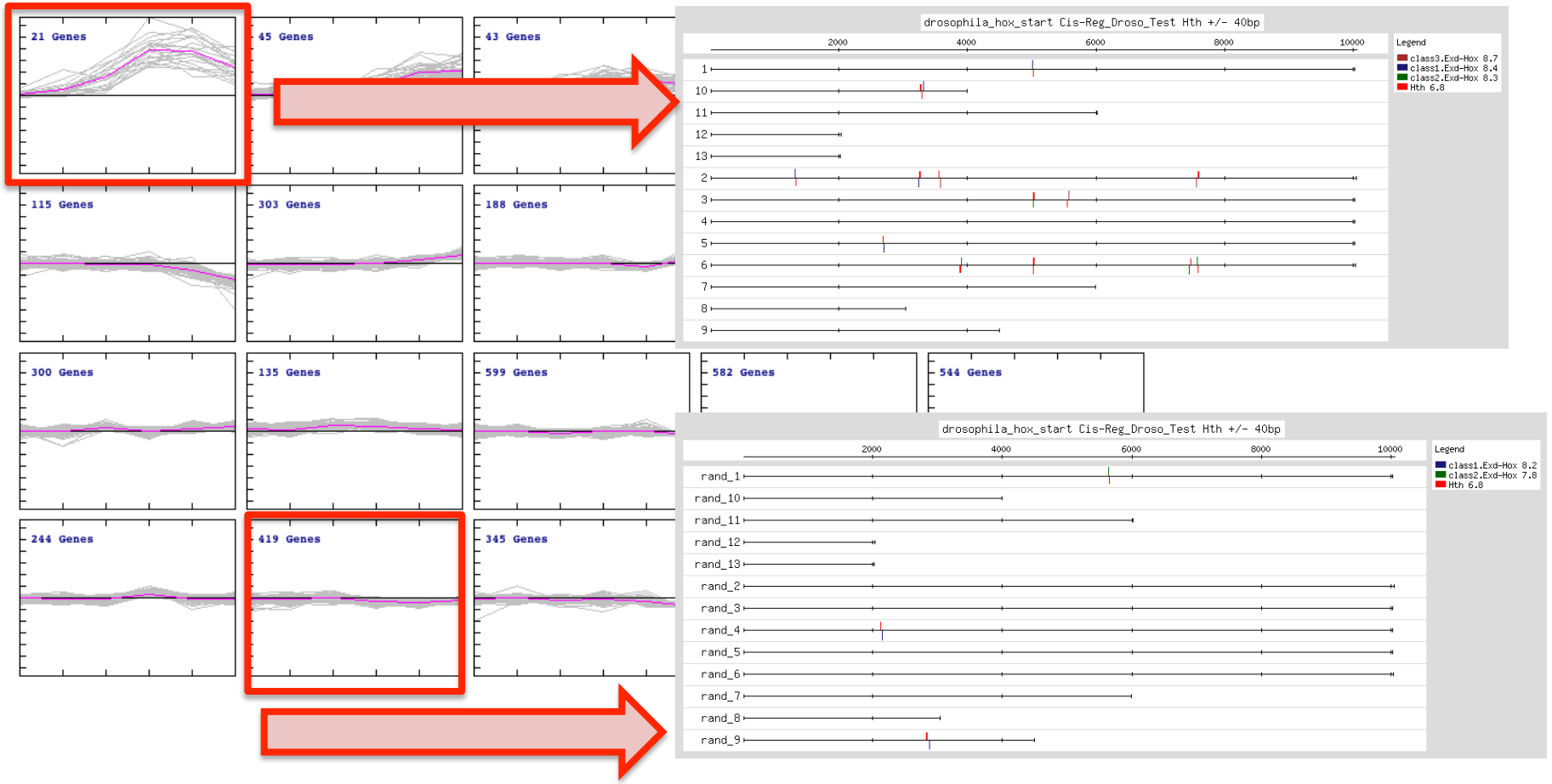
Biological sequences

- **Random-genes** in RSAT
  - » Select X genes randomly within a given genomes
  - » Obtain the upstream sequences
  - » Re-run the exact same analysis



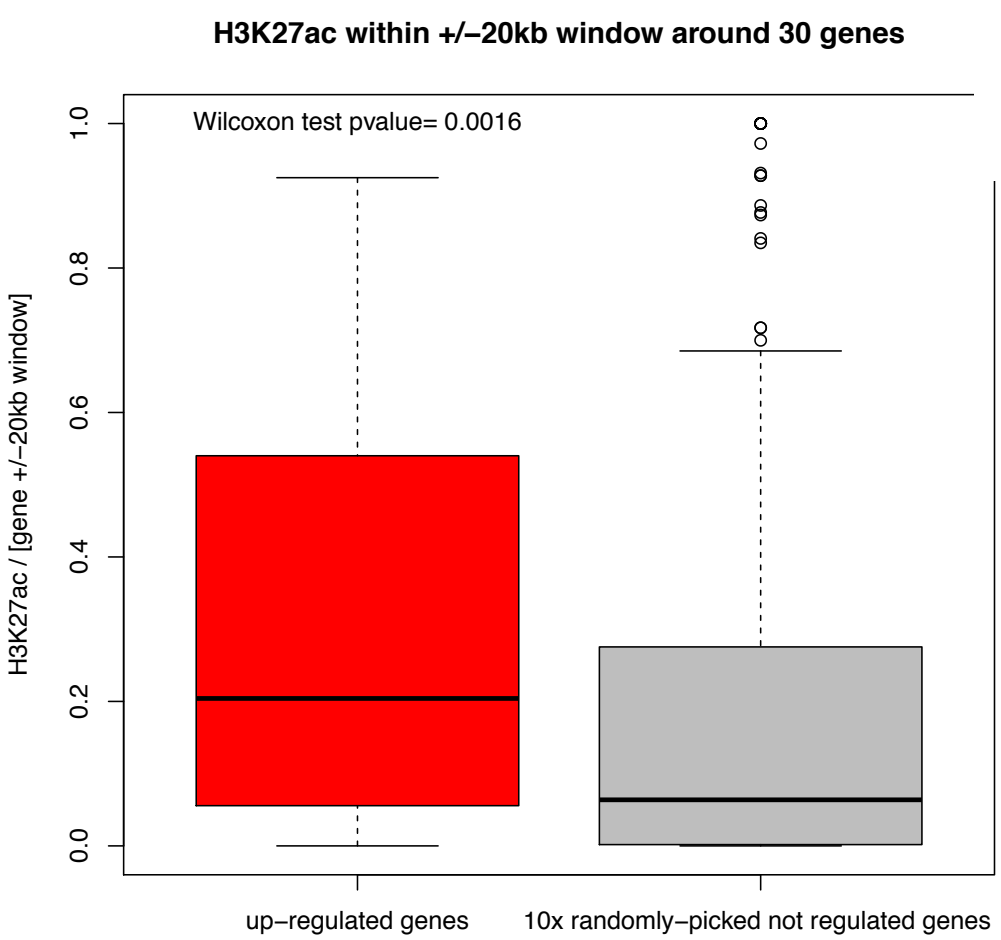
Biological sequences

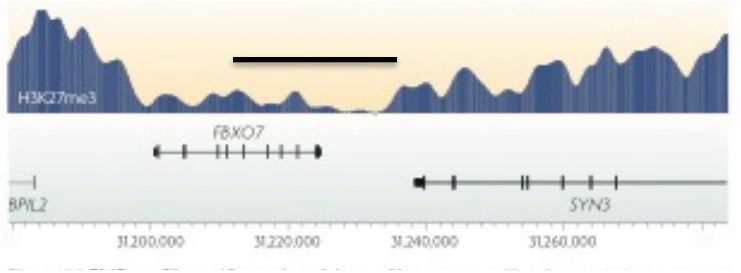
- **Genes not differentially regulated**
  - » Select X genes among genes that do not show changes in expression
  - » Obtain the upstream sequences
  - » Re-run the exact same analysis



Biological sequences

- **Genes not differentially regulated**
  - » Coverage in reads in windows aroung TSS (histone marks)





Biological sequences

- **Random genome fragments** in RSAT
  - » Select a set of fragments with random positions in a given genome, and return their coordinates and/or sequences
  - » Adapted to chip-seq ?
    - Yes: same number of peaks + same size
    - No: composition of the sequences (dinucleotides) not respected

In the context of cis-regulation

Use different set of sequences

5′- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT …*HIS7*

5′- ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG …*ARO4*

5′- CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT …*ILV6*

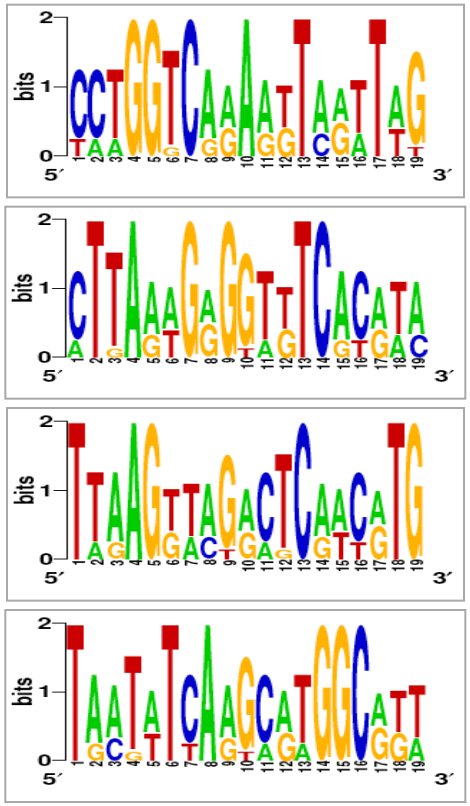
5′- TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC …*THR4*

5′- ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA …*ARO1*

5′- ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA …*HOM2*

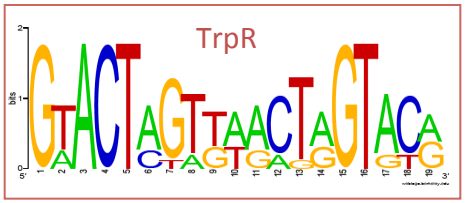
5′- GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA …*PRO3*

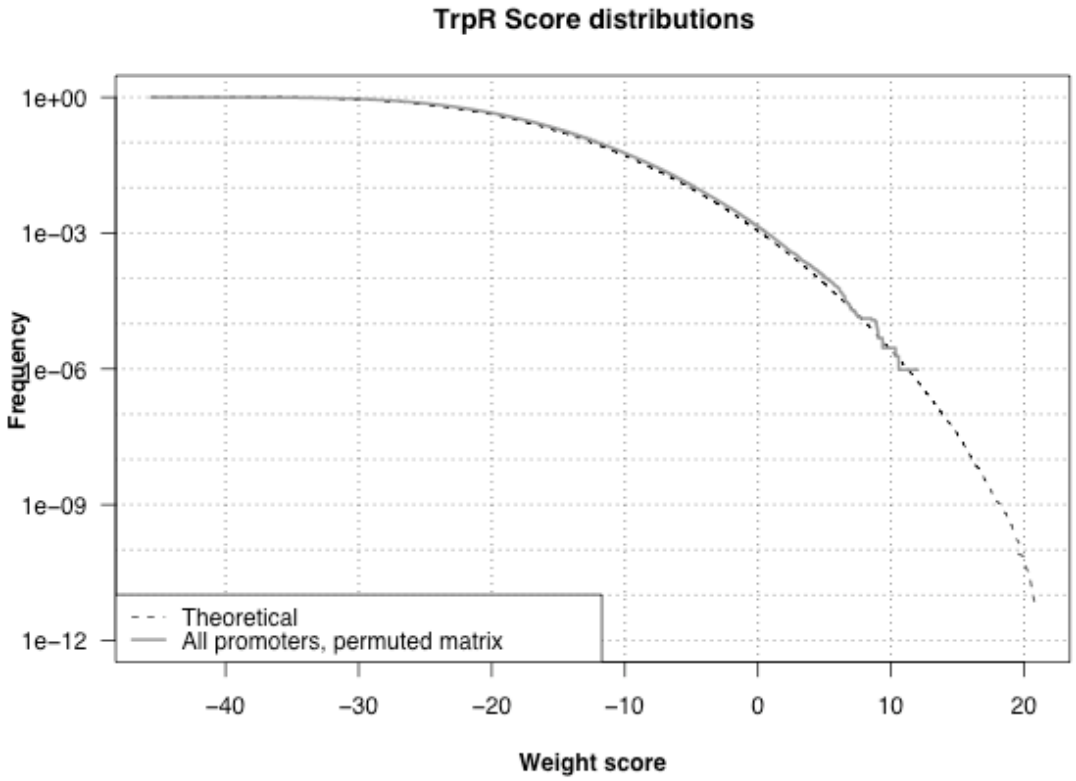
Use different set of matrices



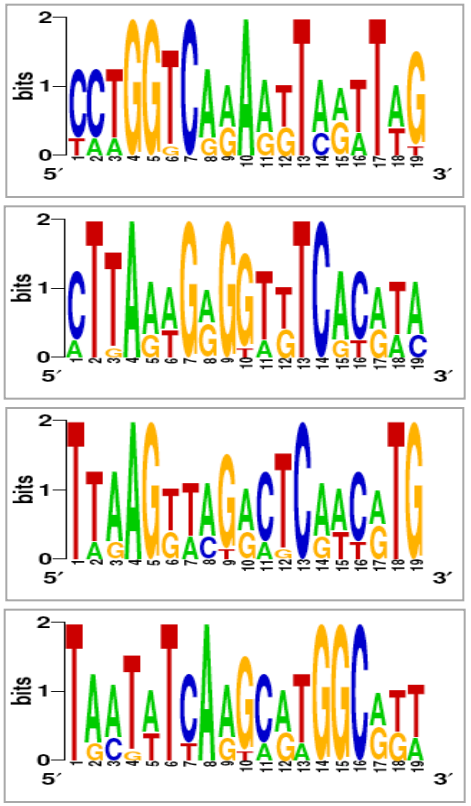
Matrix permutations

- **Matrix-quality** in RSAT
  - » Compare distributions of scores for PSSMs





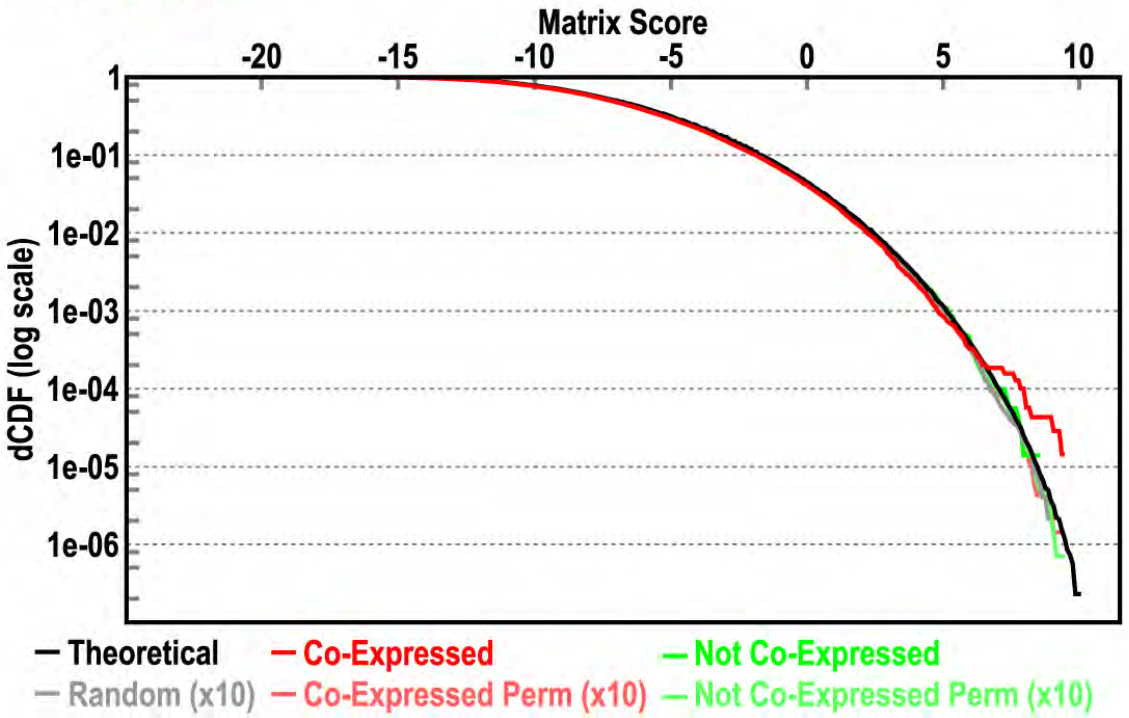
TrpR permutations



...

Matrix quality with negative datasets

- **Matrix-quality** in RSAT
  - » Not for randomly-generated sequences (random-seq) as it will ALWAYS follow the theoretical curve (= background = markov model used to generate the sequences !)
  - » OK for random selection of genes



Building controls in RSAT