# Combining transcriptional and post-transcriptional regulation to predict somatic mutations altering the gene regulatory program in cancer cells

Anthony Mathelier

Centre for Molecular Medicine Norway (NCMM),
Nordic EMBL Partnership for Molecular Medicine
and
Department of Cancer Genetics, Institute for Cancer Research,
Oslo University Hospital

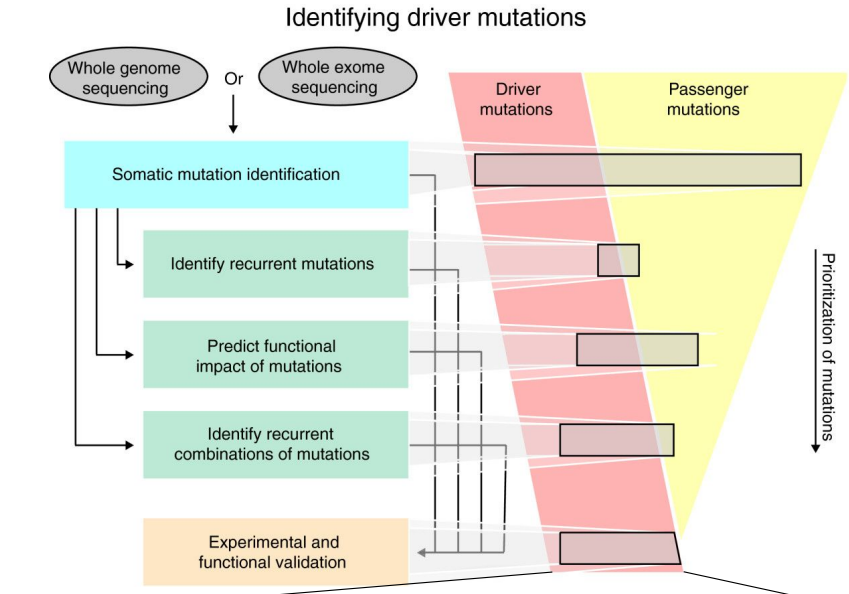
anthony.mathelier@ncmm.uio.no

@AMathelier

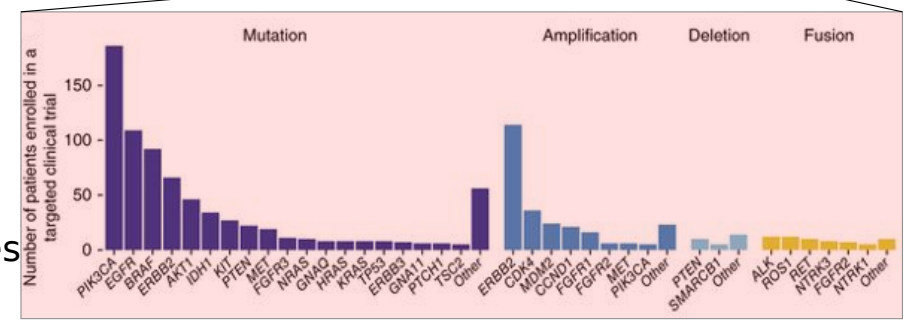
Barcelona Supercomputing Centre - 2019 Sept. 23rd

# Variant prioritization to identify cancer drivers

Driver variants identification




#patients in clinical trials identified based on actionable variants in genes




adapted from Lectures.GersteinLab.org

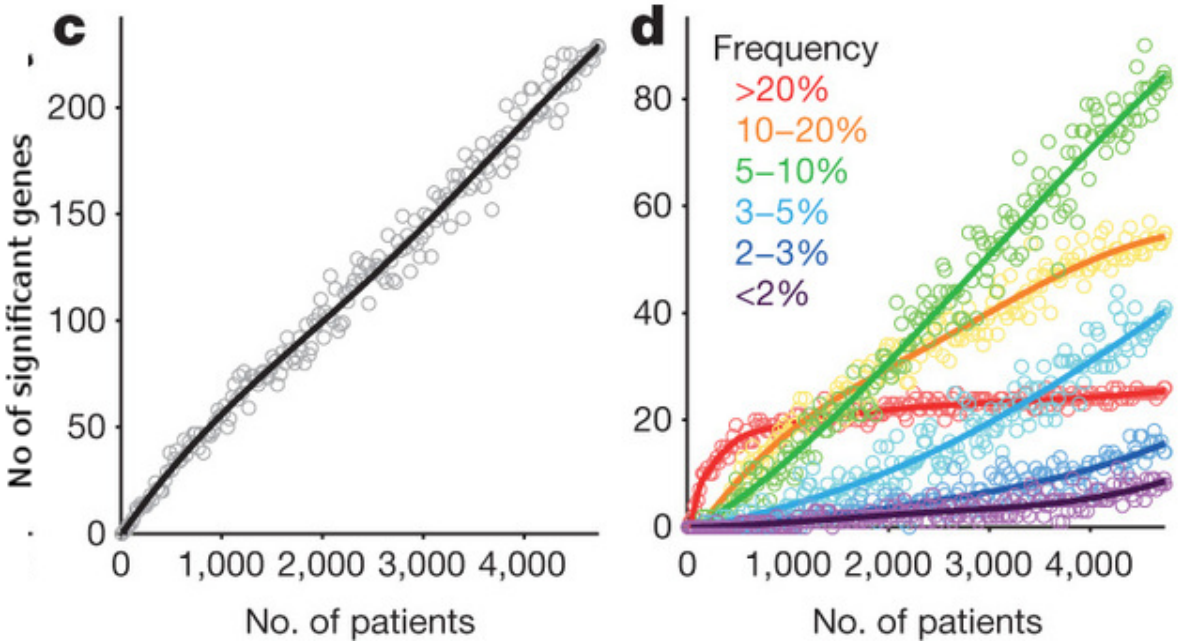
Raphael *et al.*, 2014.

Zehir *et al.*, 2017.

## Goal

Identify driver somatic events to shed light into molecular mechanisms and enable more precise diagnostics and targeted therapies.

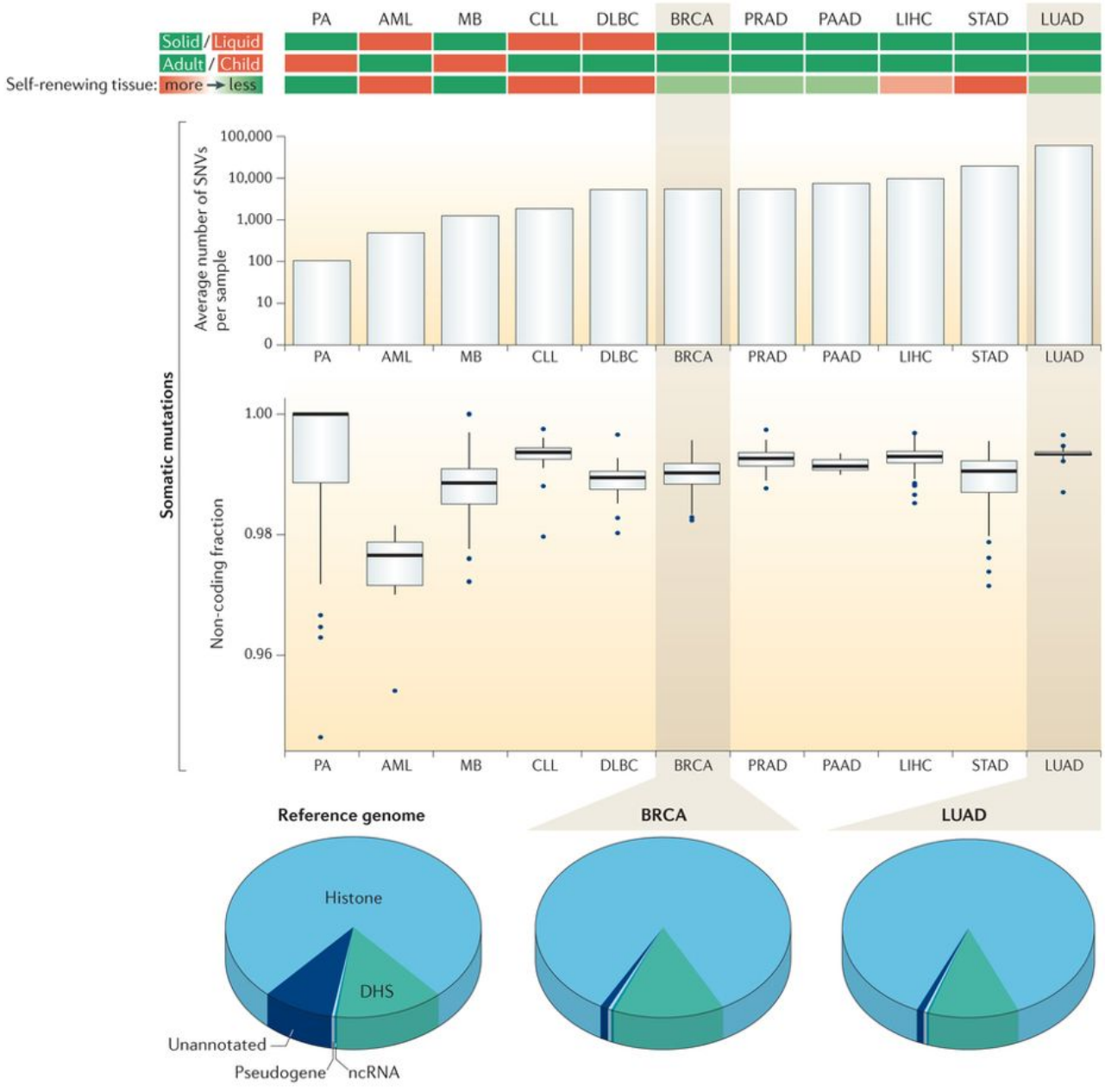
# A cancer gene discovery gap




M.S. Lawrence *et al.*, 2013.

- Highly mutated cancer genes revealed through The Cancer Genome Atlas project.
- Still a discovery gap in the search of new cancer genes.
- We assert this gap can be partially filled through the analysis of the non-coding genome.

# The vast majority of somatic mutations are non-coding



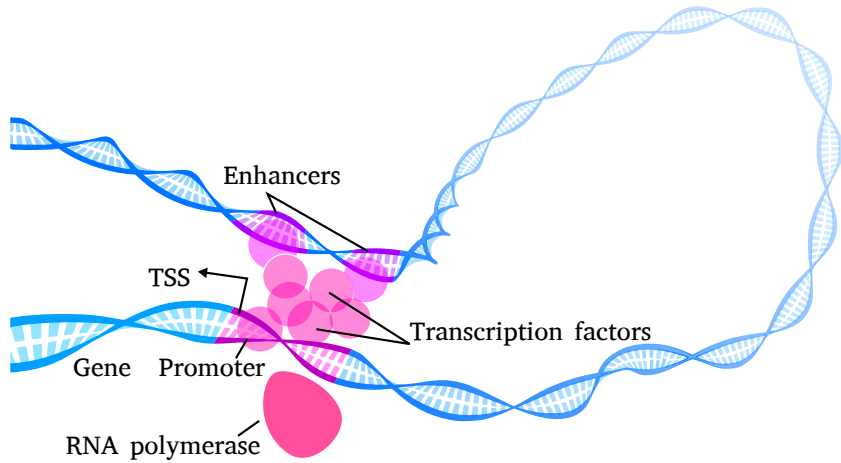

Khurana *et al.*, 2016.

Nature Reviews | Genetics

# Transcriptional regulation

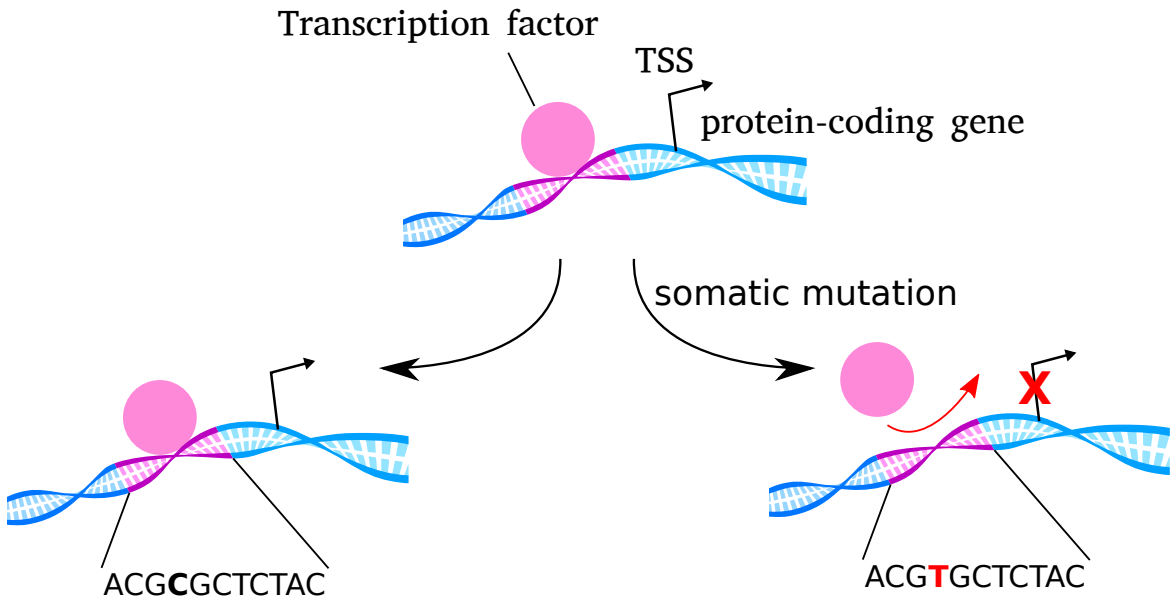
## Transcriptional regulation

Transcription factors, epigenetics, open chromatin, close chromatin, etc.

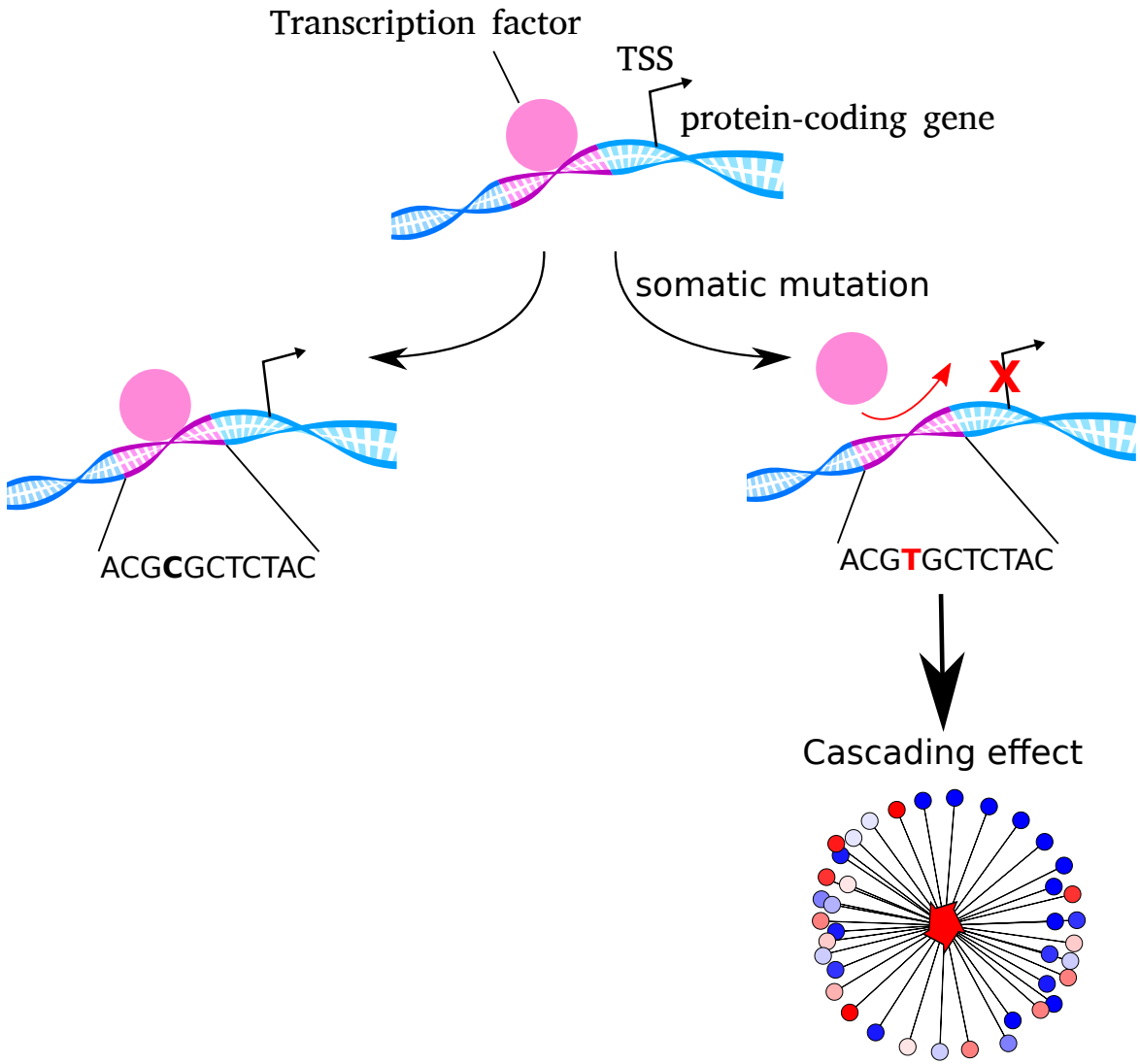



adapted from Kelvin Song's work on Wikimedia Commons

# Transcriptional deregulation and cascading effect



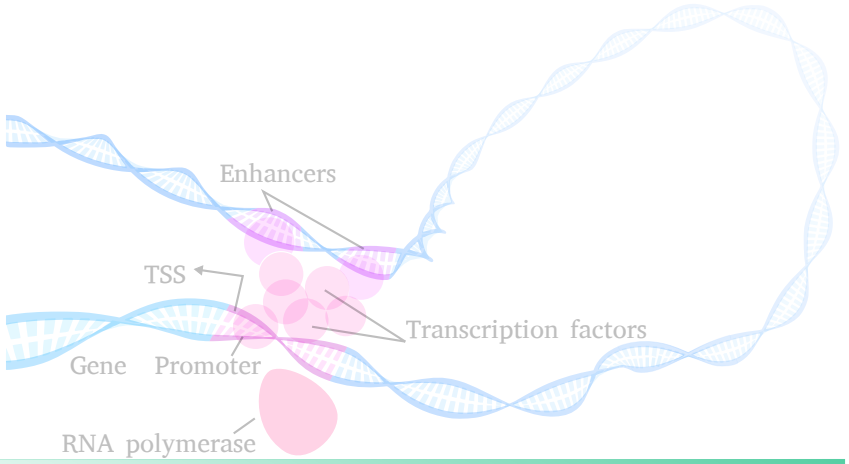
# Transcriptional deregulation and cascading effect

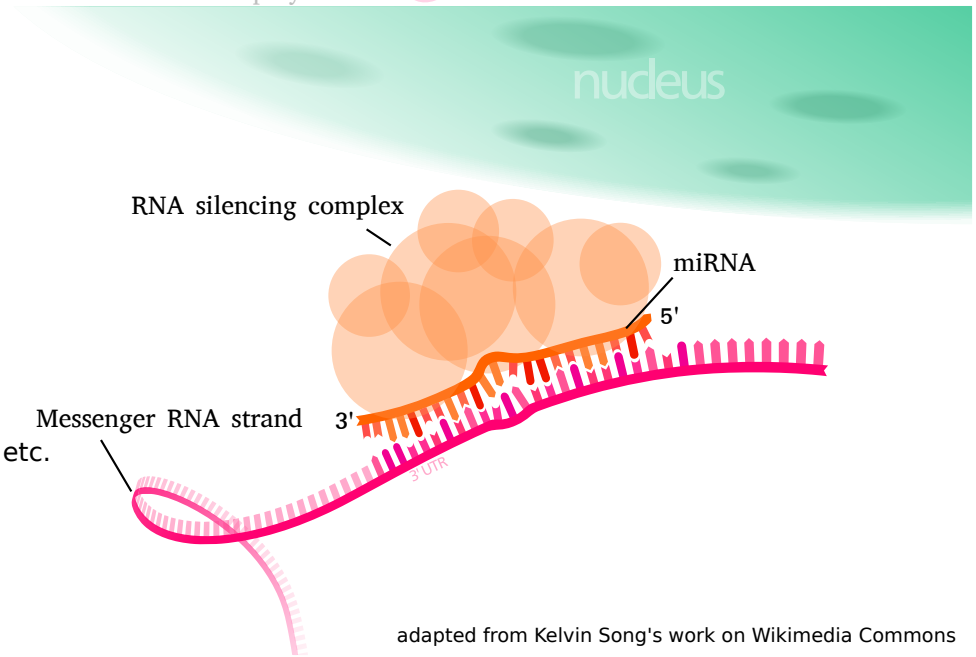


# Multiple layers of gene expression regulation

## Transcriptional regulation

Transcription factors, epigenetics,
open chromatin, close chromatin, etc.



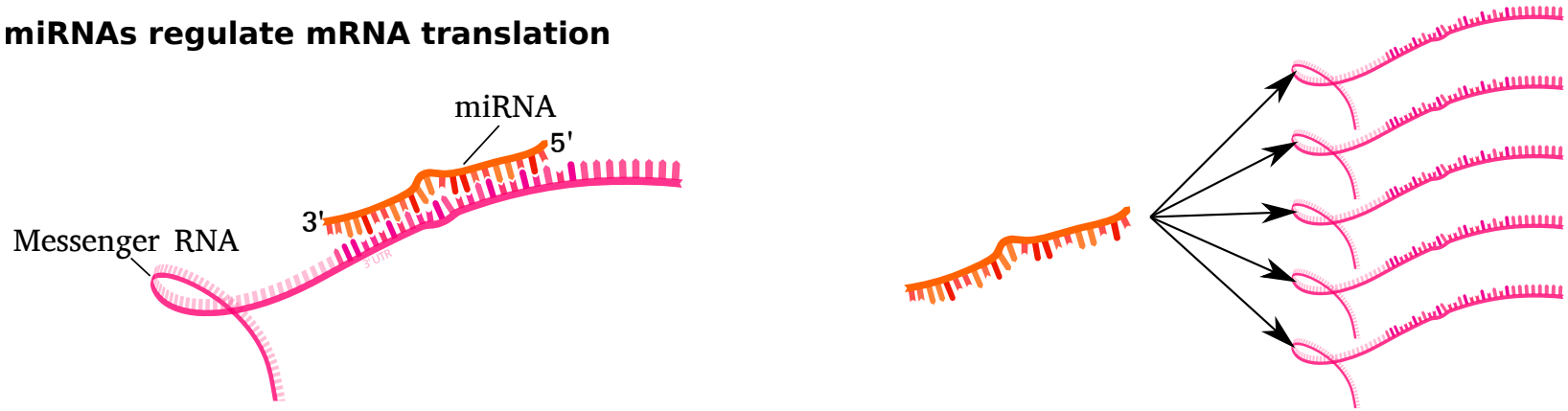




## Post-transcriptional regulation

miRNAs, mRNA localization,
RNA-binding proteins, splicing, etc.

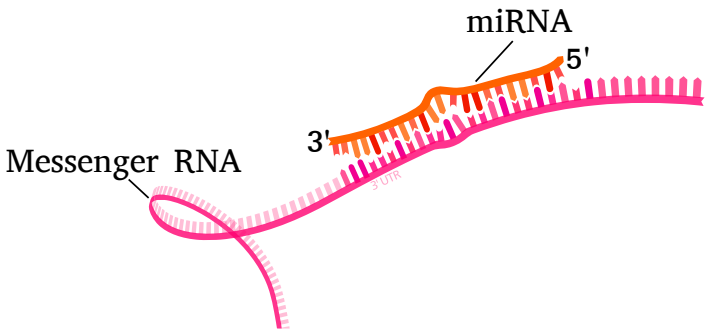
adapted from Kelvin Song's work on Wikimedia Commons

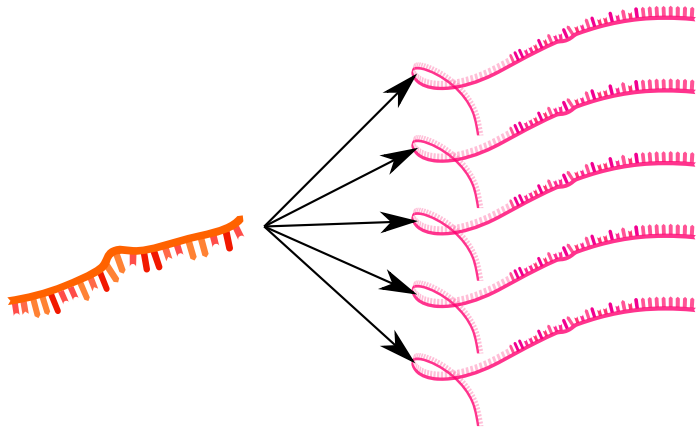
# Transcriptional and post-transcriptional deregulation



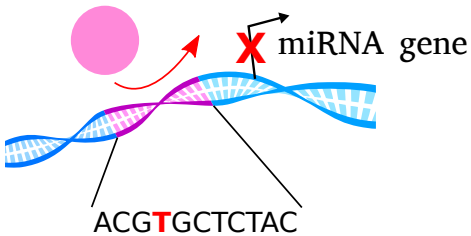
# Transcriptional and post-transcriptional deregulation

**miRNAs regulate mRNA translation**

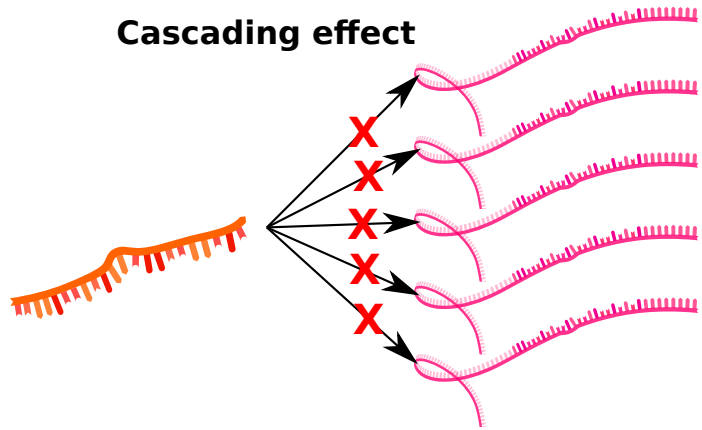


**miRNA transcription must be accurately controlled**

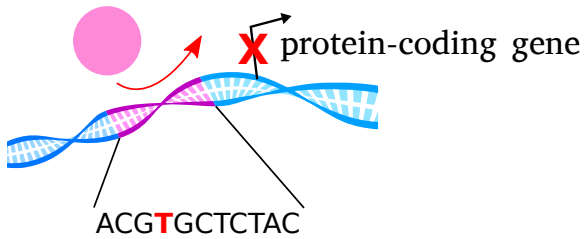


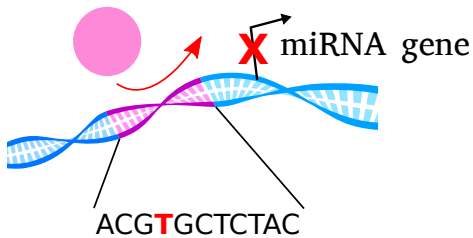

**Cascading effect**



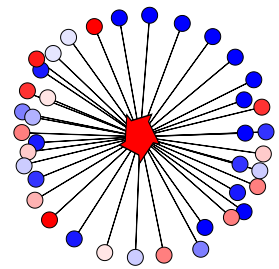
# Predicting cis-regulatory mutations altering the regulatory program in cancer cells

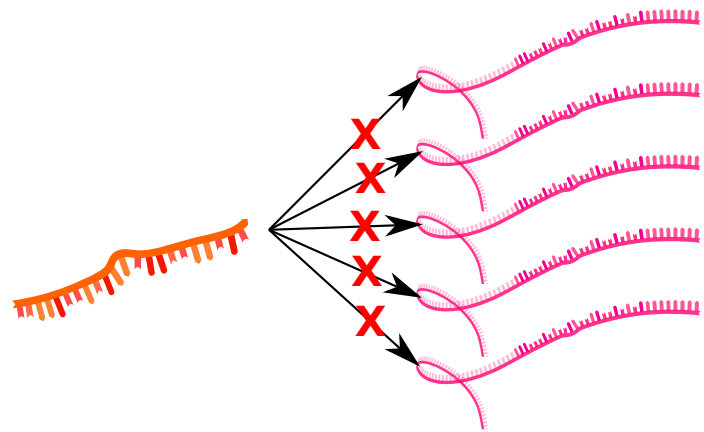
**Transcriptional dysregulation**



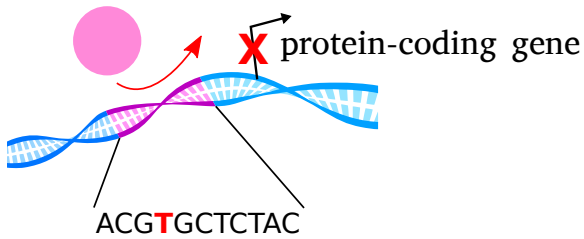

**Cascading effect**

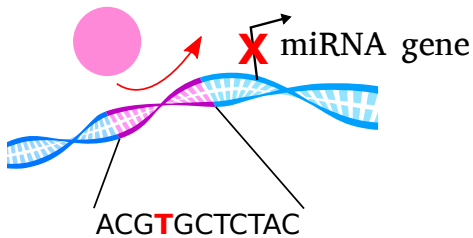




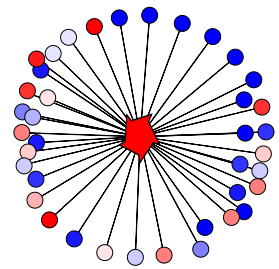
# Predicting cis-regulatory mutations altering the regulatory program in cancer cells

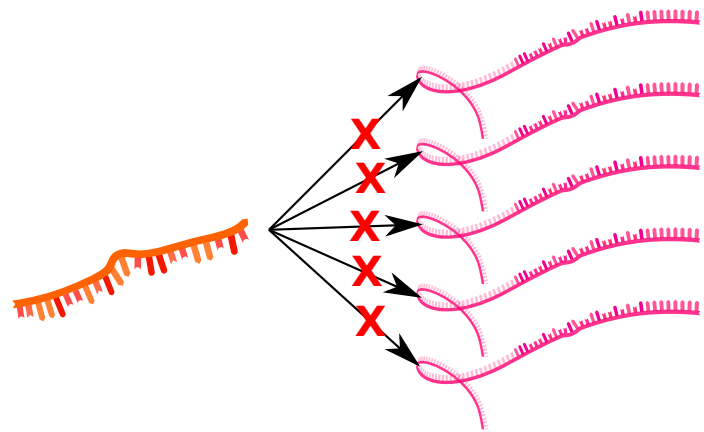
**Transcriptional dysregulation**




**Cascading effect**





One needs to accurately locate TFBSs to identify and characterize the regulatory sequences controlling specific genes transcription.

# Outline

1. Improving our capacity to predict transcription factor binding events

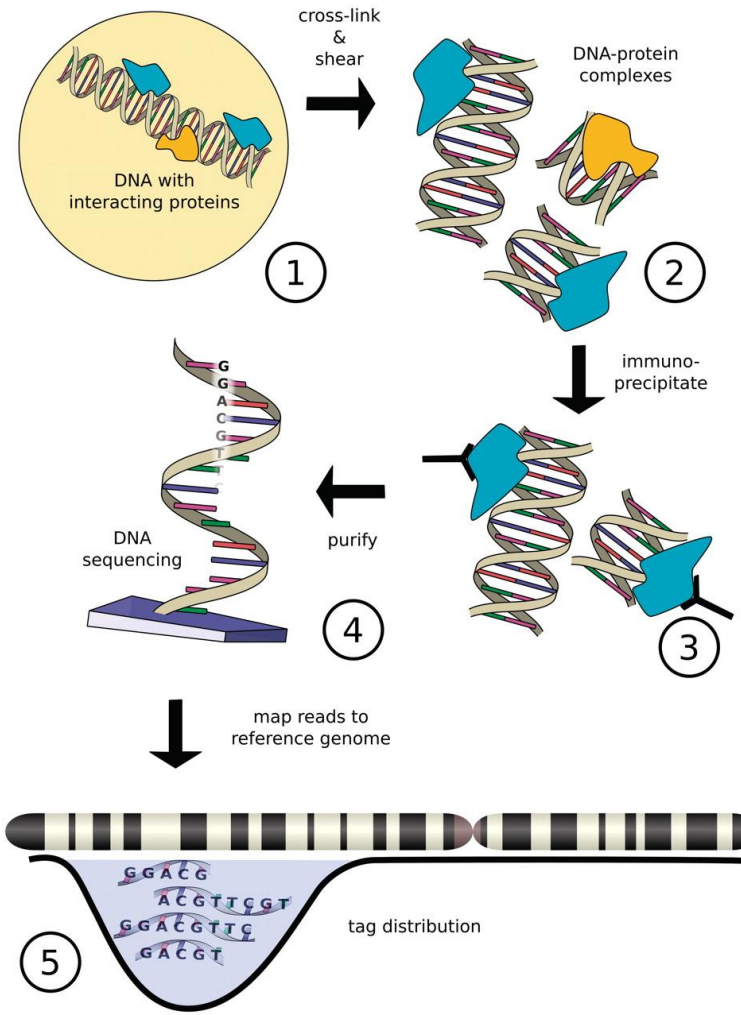
2. Combining transcriptional and post-transcriptional regulation to predict mutations altering the gene regulatory program in cancer cells

# Outline

1. Improving our capacity to predict transcription factor binding events

2. Combining transcriptional and post-transcriptional regulation to predict mutations altering the gene regulatory program in cancer cells

# Genome-scale data capturing TFBSs: ChIP-seq




```
>Seq 1
AGTTCAAAGTTCAAGTTCAAA
GTTCAAGTTCAAAGTTCAAGT
TCAAAGTTCA
>Seq 2
AGTCCAAAGTTCAAGTCCAAA
GTTCAAGTCCAAA
        ...
>Seq 33554
CTACCGGGGACCGGGGTGGA
ACCGGGG
>Seq 33555
ACCGGGGACCGGGGACCGGG
GACCGGGGGGCAAGGTTCATA
```

adapted from

A.M. Szalkowski and C.D. Schmid, 2010.

# You do not always ChIP what you expect

## Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins

Leonid Teytelman[a,b,1], Deborah M. Thurtle[c,1], Jasper Rine[c,2], and Alexander van Oudenaarden[a,b,d,2]

Nucleic Acids Research Advance Access published June 27, 2015

*Nucleic Acids Research, 2015* **1**
*doi: 10.1093/nar/gkv637*

## Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments

**Dhawal Jain, Sandro Baldi, Angelika Zabel, Tobias Straub and Peter B. Becker***

Worsley Hunt and Wasserman *Genome Biology* 2014, **15**:412
http://genomebiology.com/2014/15/7/412

Genome Biology

RESEARCH Open Access

## Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets

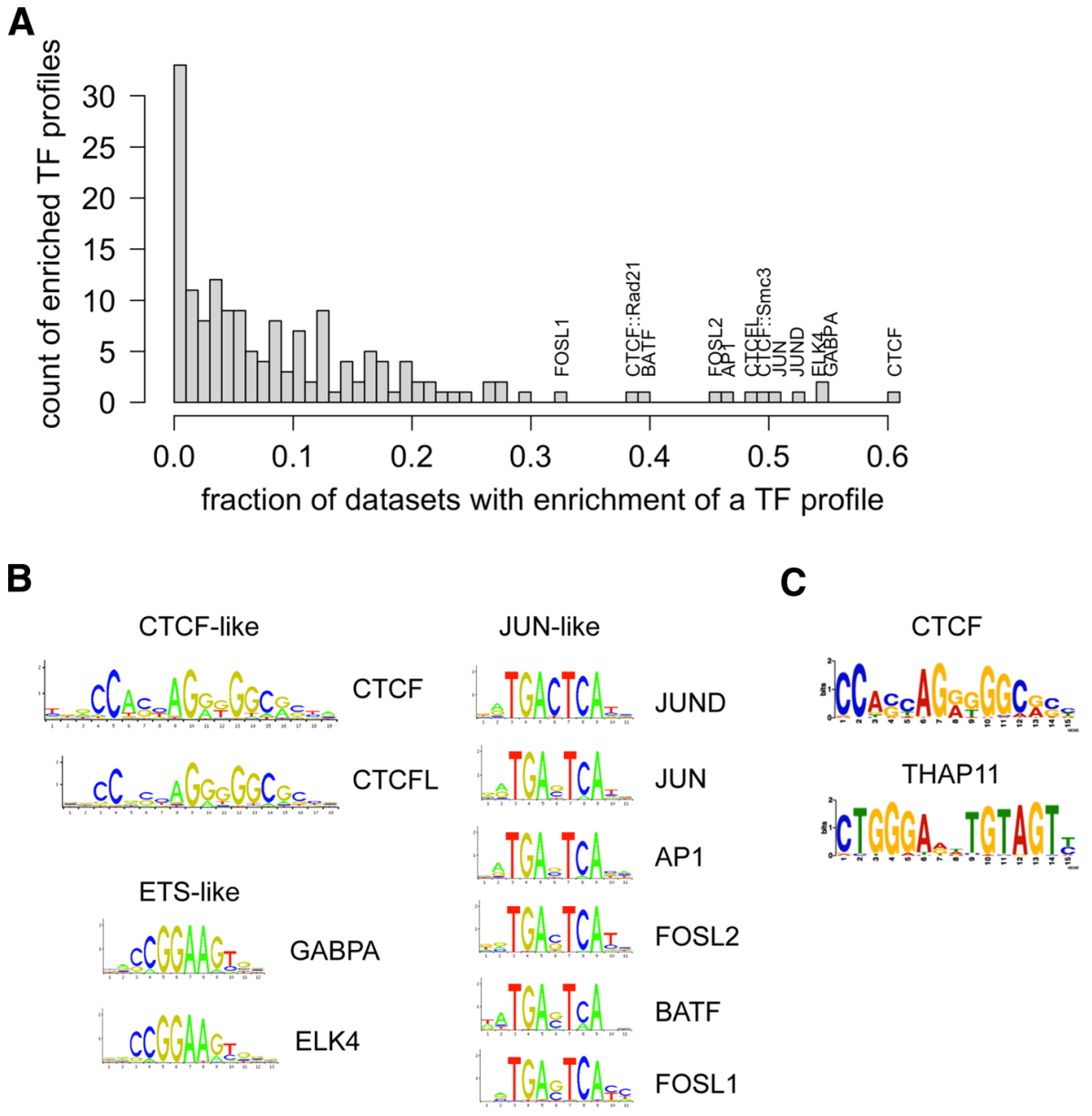
Rebecca Worsley Hunt[1,2] and Wyeth W Wasserman[1,3*]



## HOT or not: examining the basis of high-occupancy target regions

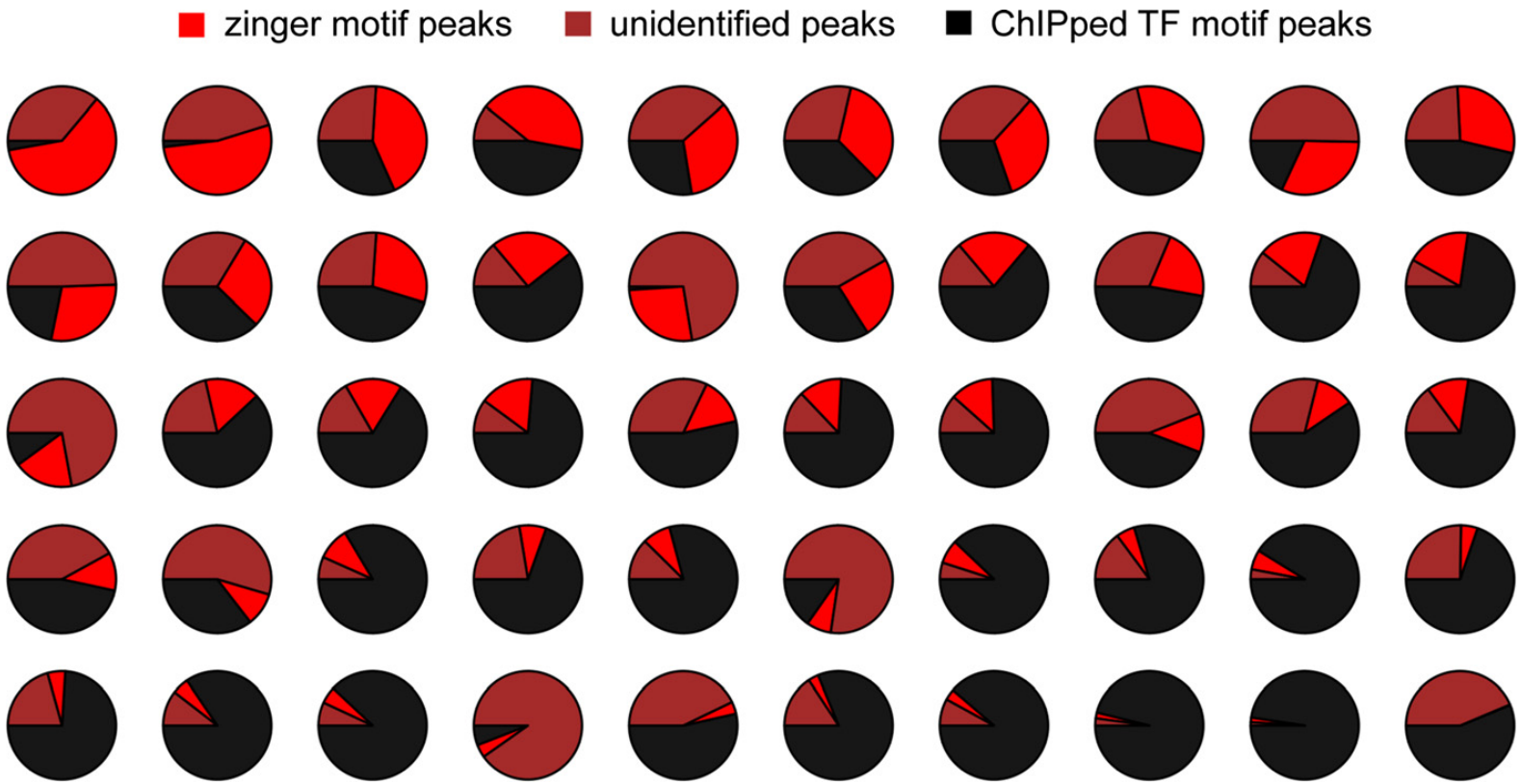
Katarzyna Wreczycka[1*], Vedran Franke[1*], Bora Uyar[1], Ricardo Wurmus[1], Altuna Akalin[1#]

# ChIP-seq peaks are enriched for zingers




Worsley-Hunt and Wasserman, 2014.

# ChIP-seq peaks are enriched for zingers




Worsley-Hunt and Wasserman, 2014.

# Modeling TFBSs

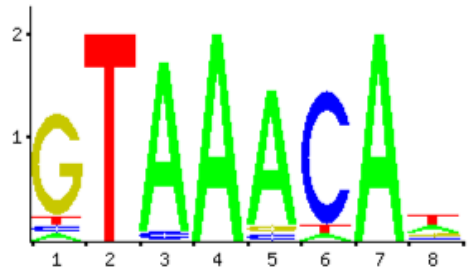
Known binding sites:

GTAACAAT
GTAAACAT
GTAAACAA
GTAAACAA
GTAAACAT
GTAAACAA
GTAAACAC
GTCAACAG
GTAAACAT
GTAAACAA
GTAAACAT
TTAAGTAA
ATAAACAA
CTAAACAG
GTAAACAT
GTAAACAA
GTAAACAT
GTAAACAC
GTAAACAT
GTAAACAG



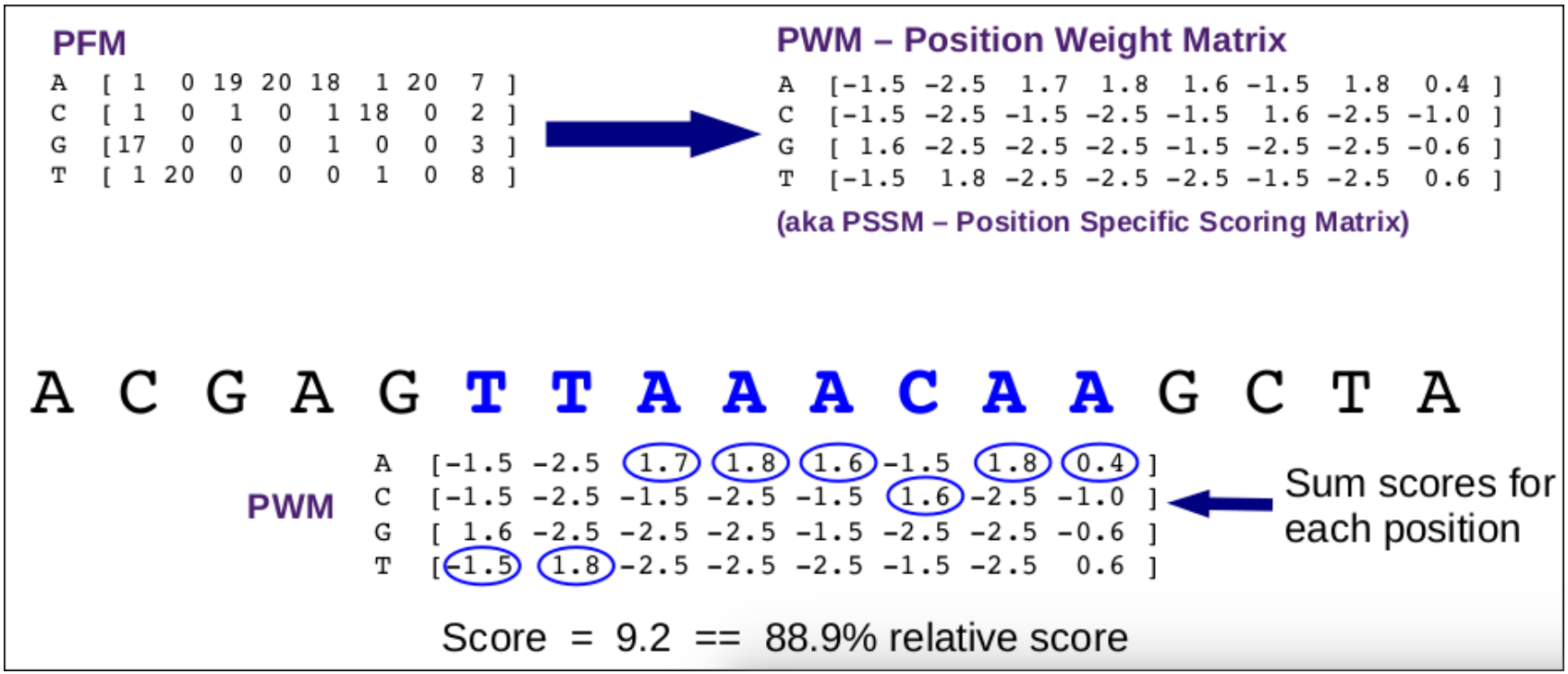
Position Frequency Matrix:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A | [ 1 | 0 | 19 | 20 | 18 | 1 | 20 | 7] |
| C | [ 1 | 0 | 1 | 0 | 1 | 18 | 0 | 2] |
| G | [17 | 0 | 0 | 0 | 1 | 0 | 0 | 3] |
| T | [ 1 | 20 | 0 | 0 | 0 | 1 | 0 | 8] |

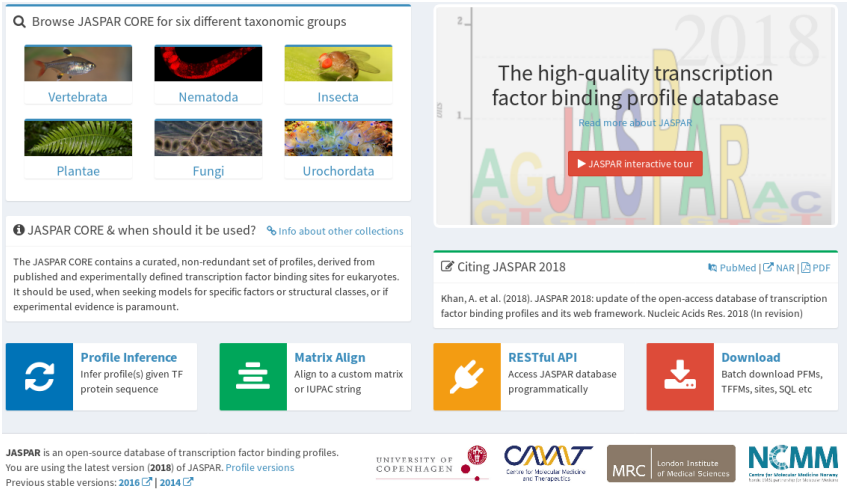


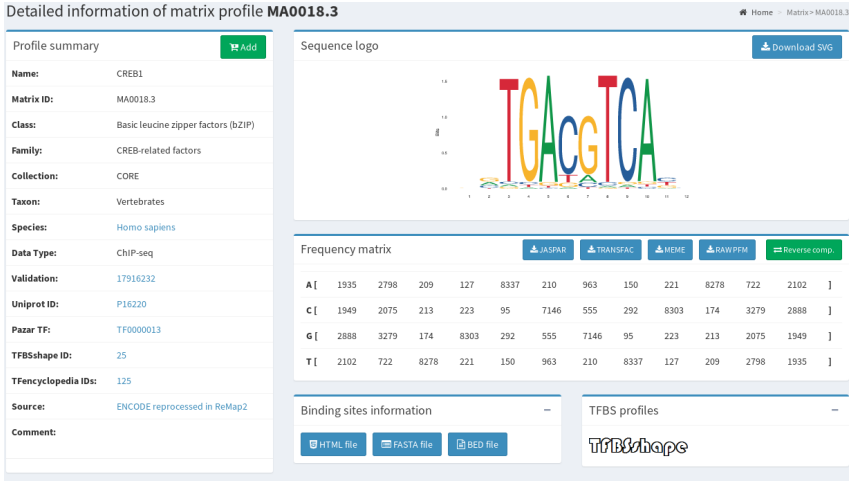

PFMs reflect the preferred binding motifs associated to TFs.

# Scoring potential TFBSs



# JASPAR



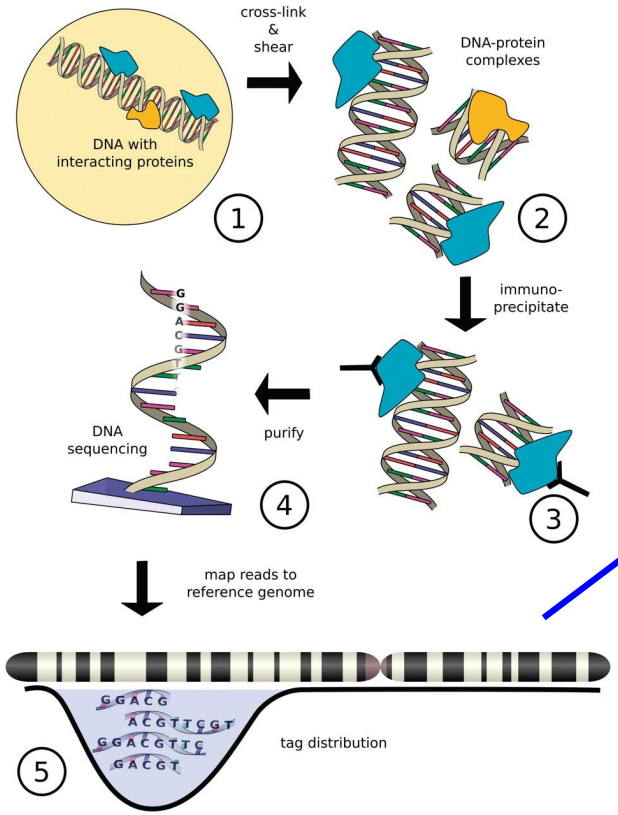




**Largest open-access database of manually curated TF binding profiles.**

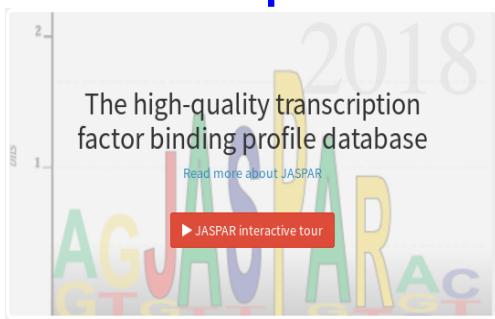
Khan, Fornes, *et al.*, 2018.

| Subset | # TF binding profiles |
|---|---|
| **Vertebrates** | **579** |
| Plants | 489 |
| Insects | 133 |
| Nematodes | 26 |
| Fungi | 176 |
| Urochordata | 1 |
| **Total** | **1404** |

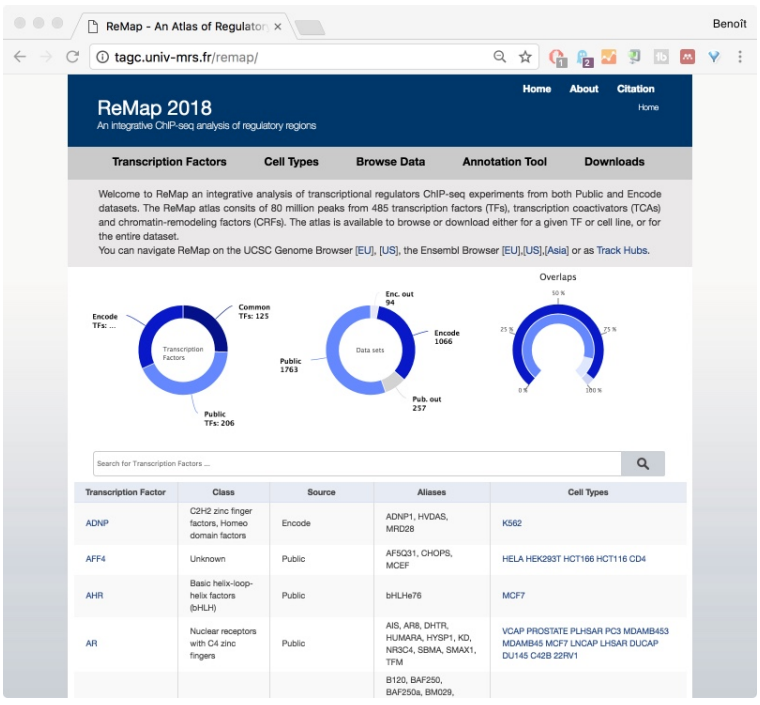
# Combining ChIP-seq peaks with JASPAR TF binding profiles



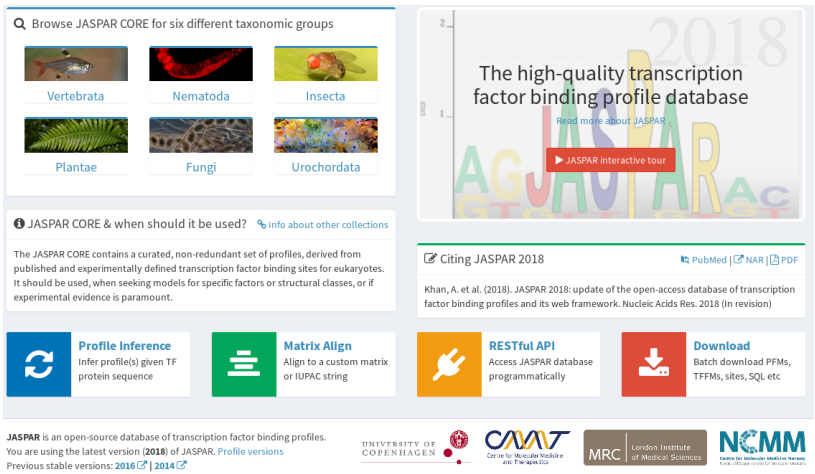

...gctaa**GTAACAAT**gagc...
...ctaaa**GTAAACAT**gcga...
...ccaat**GTAAACAA**acgg...



# Combining ChIP-seq peaks and TF-binding profiles

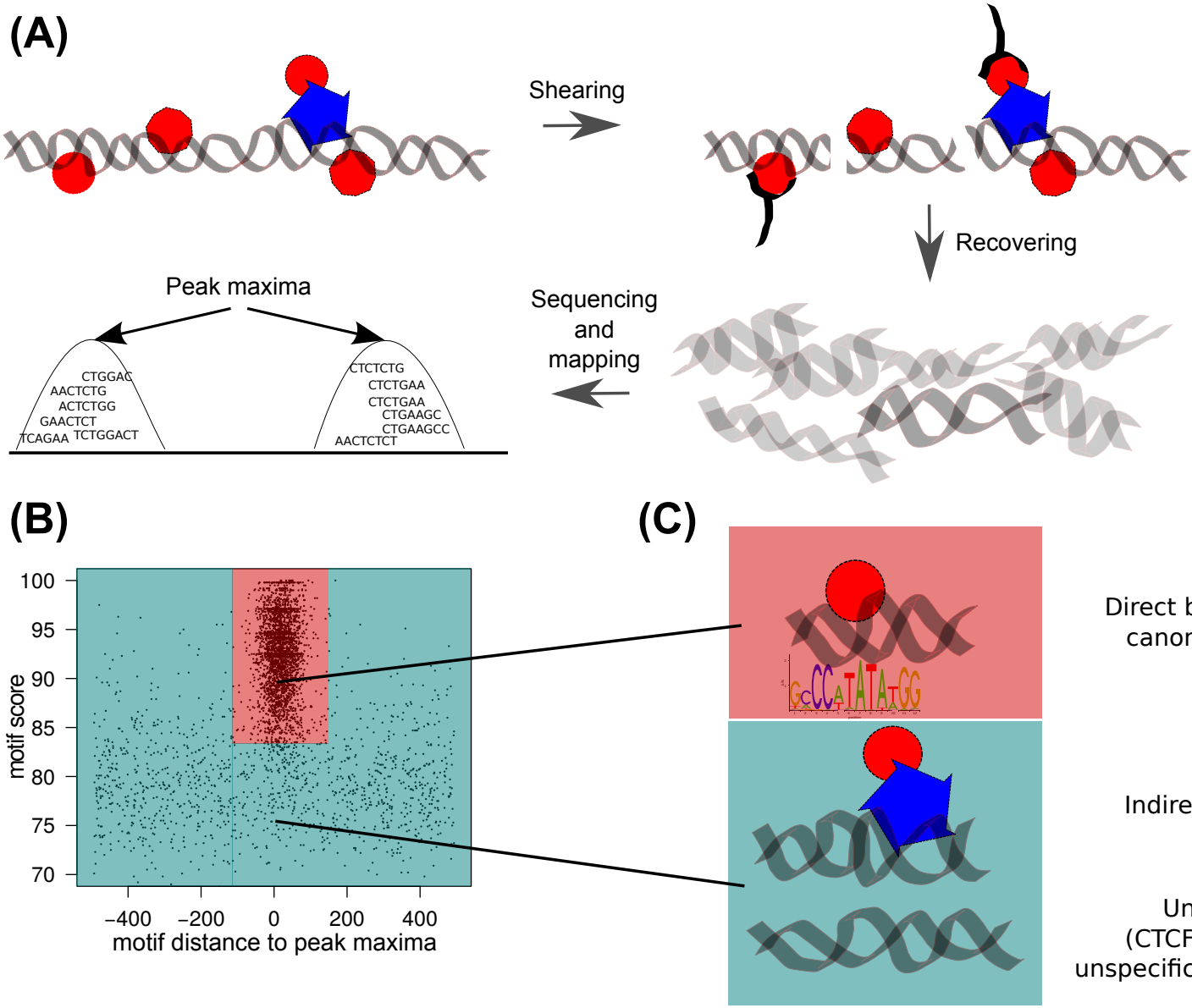



Chèneby *et al.*, 2018.



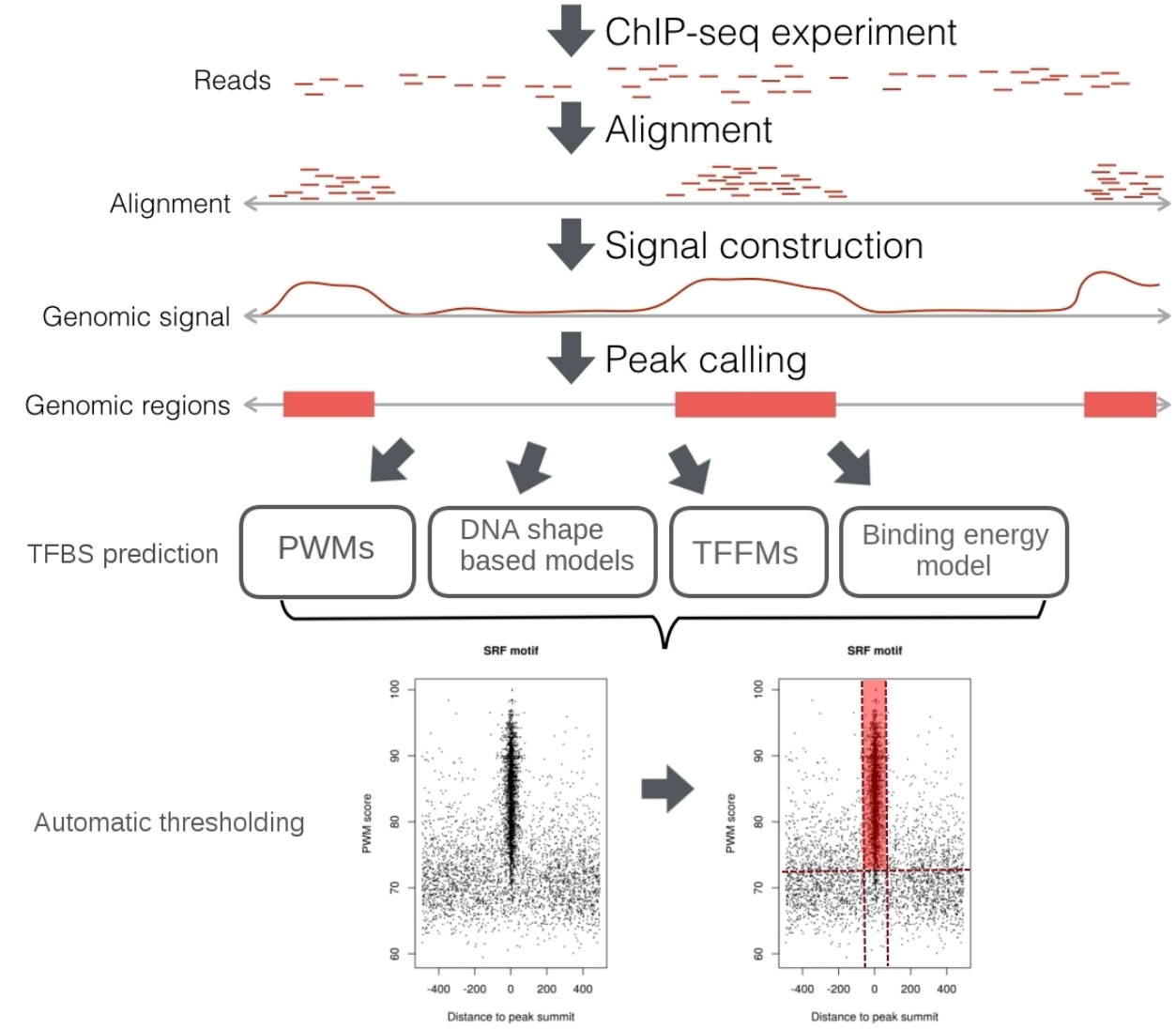

Khan, Fornes, *et al.*, 2018.

# What to expect when you are ChIP'ing



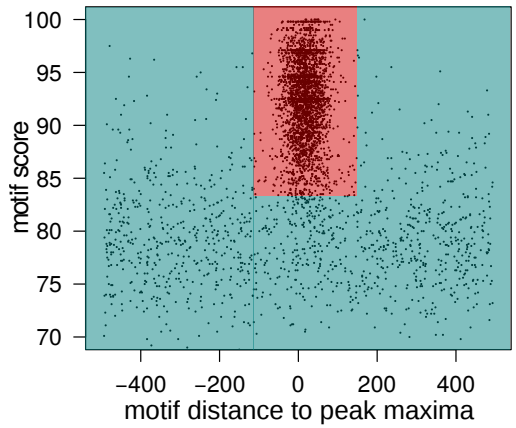

Mathelier, Shi, and Wasserman, 2015.

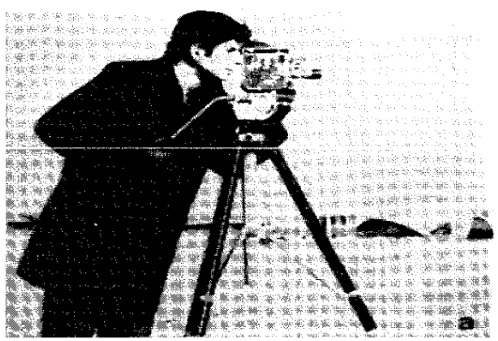
# ChIP-eat: from raw reads to high quality TFBSs

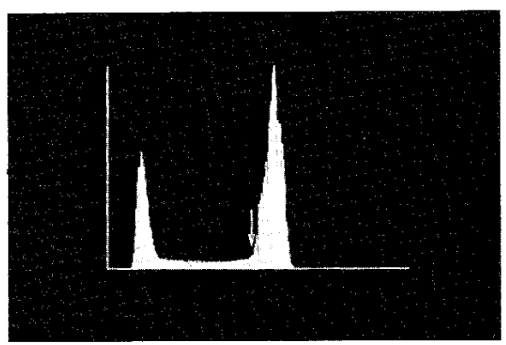



Gheorghe *et al.*, 2018.

# Entropy to automatically define TFBS enrichment zones

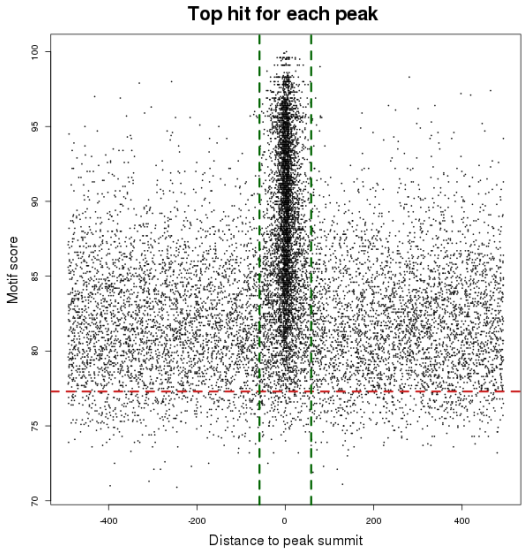


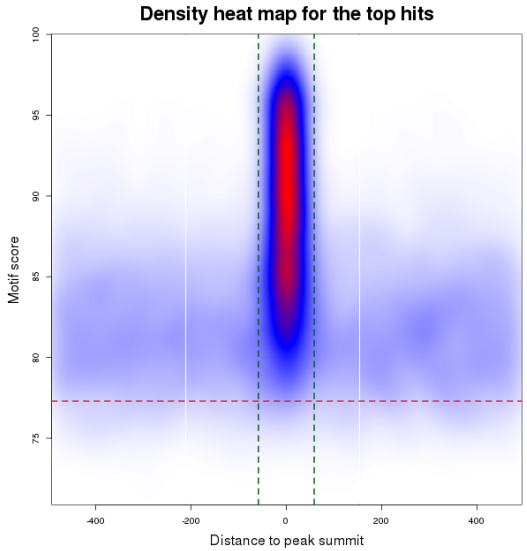



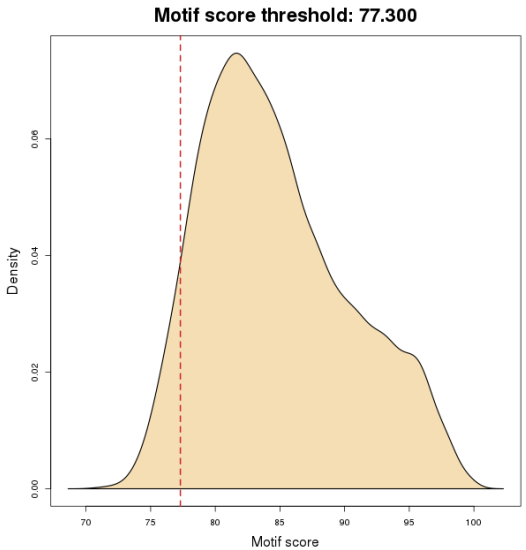


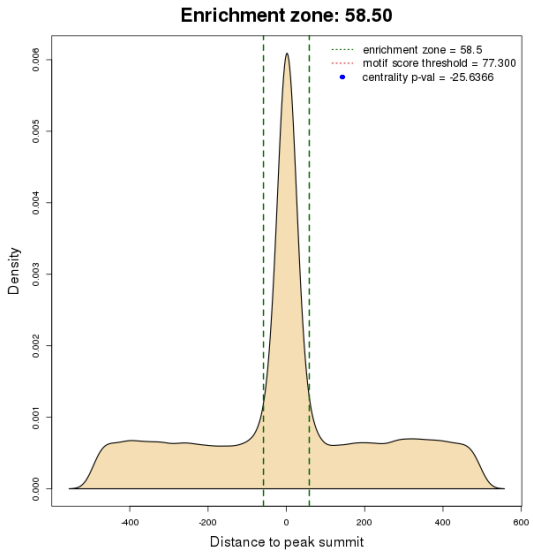
Kapur *et al.*, 1985.

# A map of TF-DNA interactions in the human genome



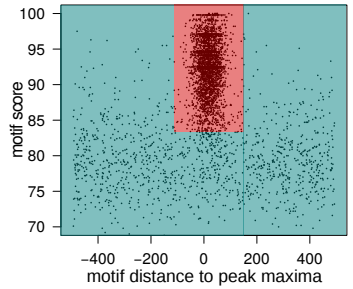



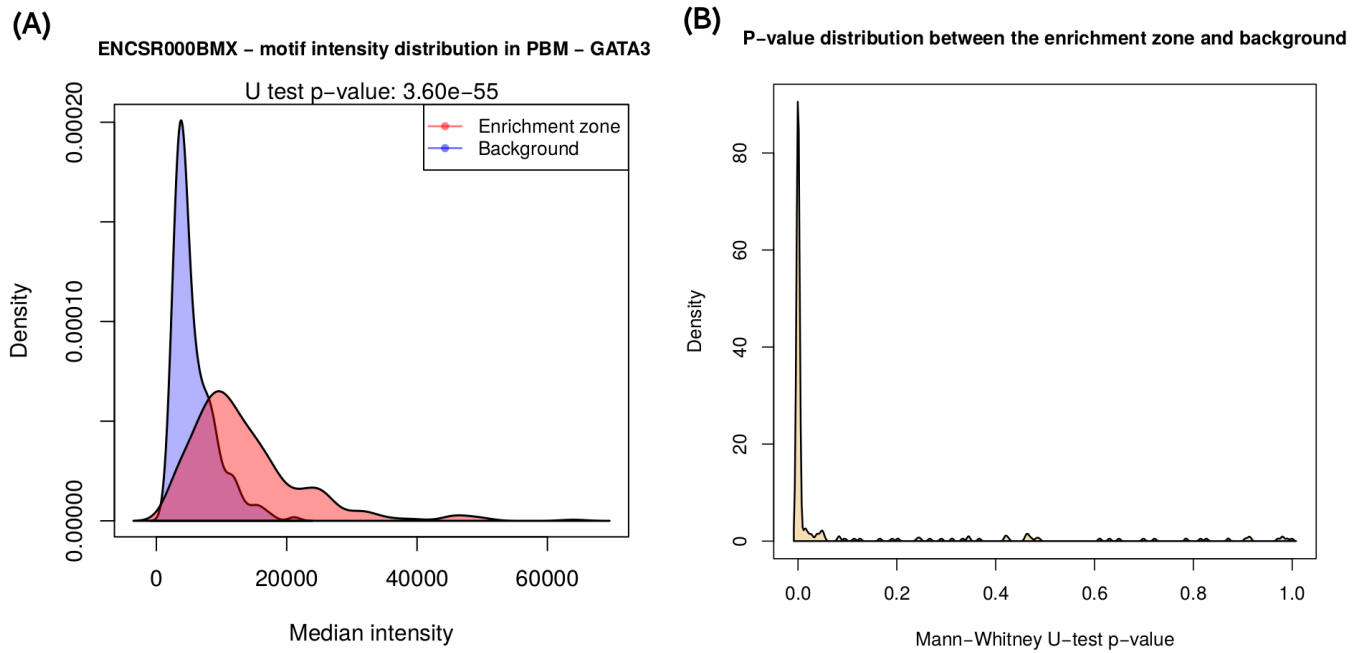




Gheorghe *et al.*, 2018.

We predict **direct TF-DNA interactions** covering $>$ **2%** of the human genome from 1,982 ChIP-seq data sets for 231 TFs.

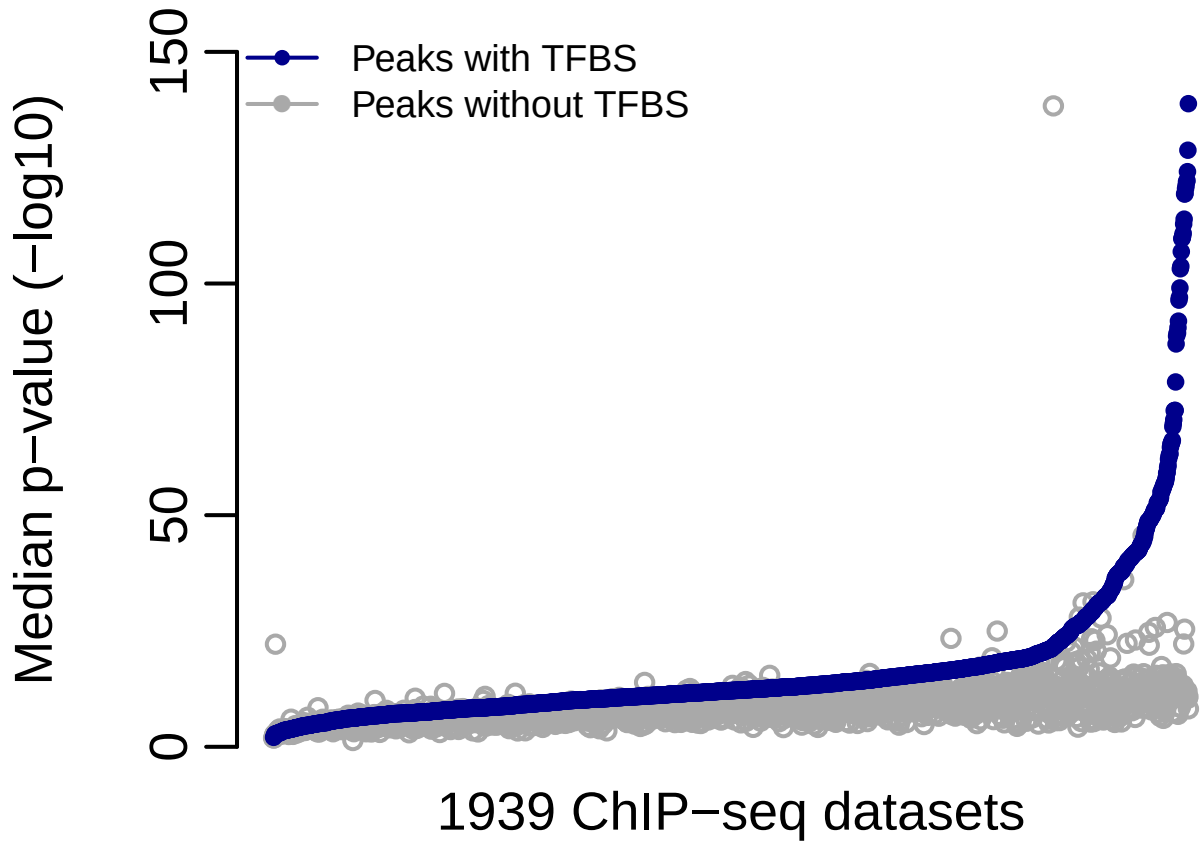
# The TFBS enrichment zone highlights higher binding affinity



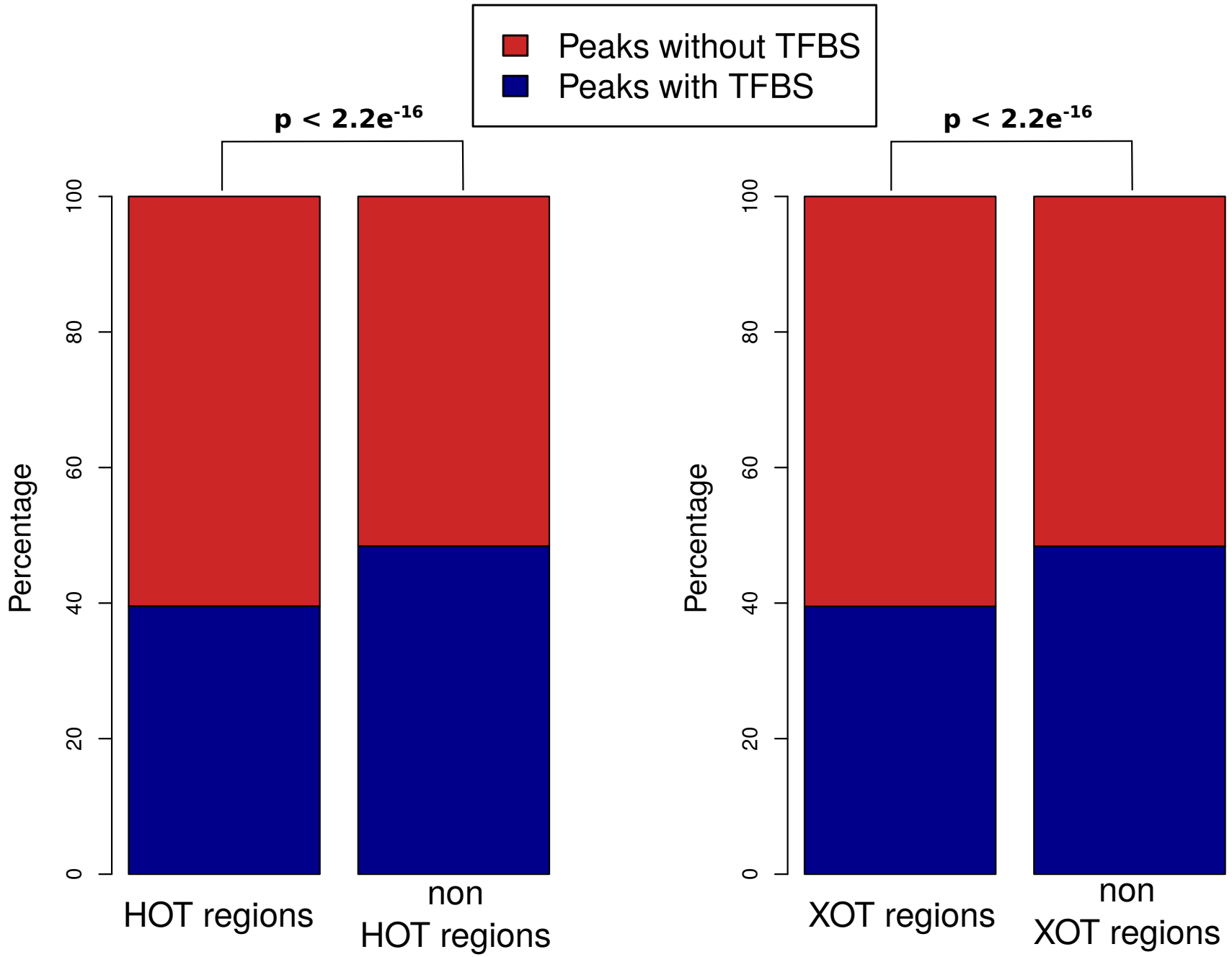




TFBSs in enrichment zones show higher PBM binding affinity than hits outside.

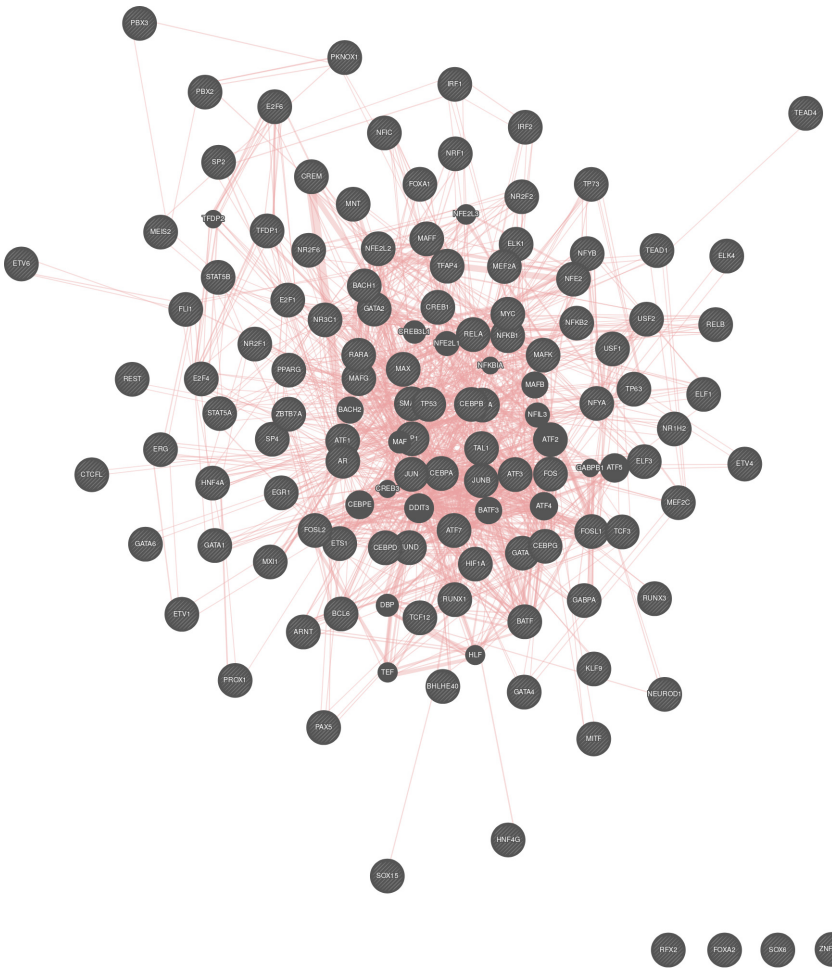
# Direct TF-DNA interactions are found in high confidence peaks



# HOT regions are depleted of TFBSs in enrichment zones

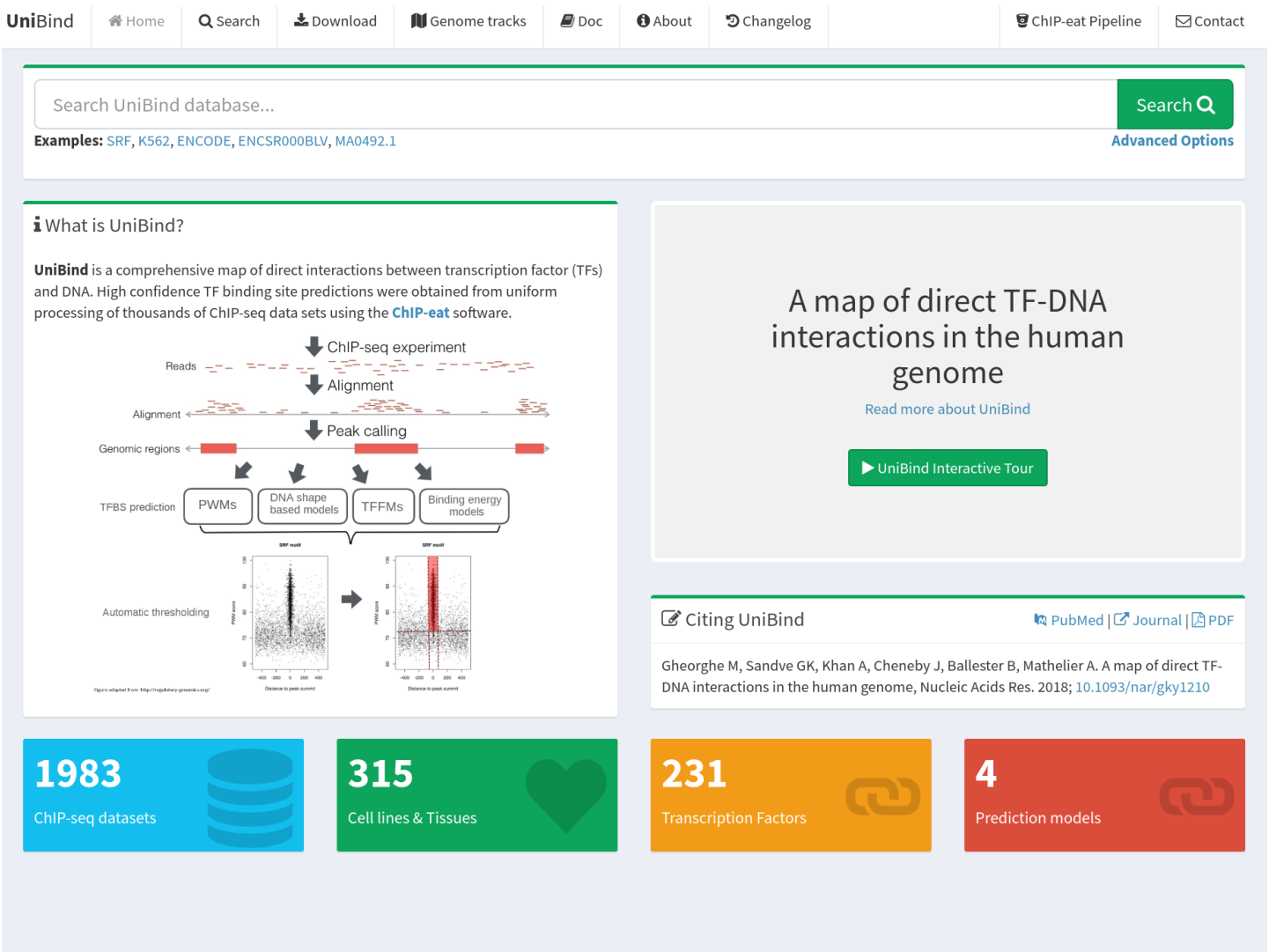


# Direct TF-DNA interactions reveal co-localizing TFs



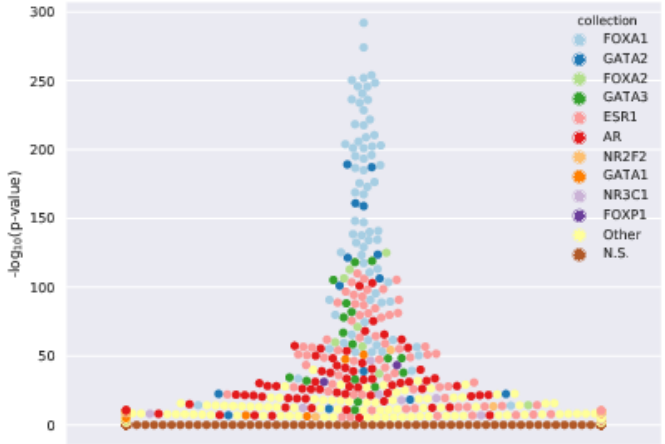

Out of 231 available TFs (26,796 pairs tested), 150 pairs of co-binding (112) TFs are predicted, 82% of which known in PPI database.

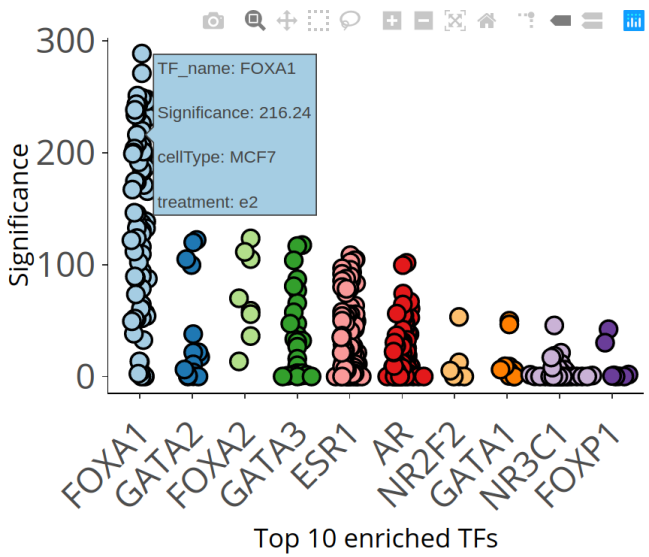
# UniBind

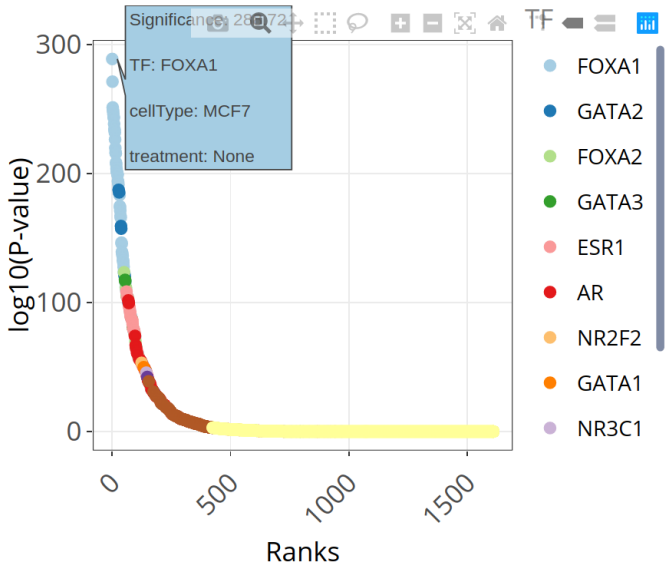



http://unibind.uio.no

# TFBS sets enrichment analyses




You can query UniBind to compute enrichment against your genomic regions.

# Summary

- We provide a genome-wide map of direct TF-DNA interactions by combining both experimental and computational evidences.
- TFBSs predicted in the enrichment zones cover $> 2\%$ of the human genome (1,983 ChIP-seq data sets - 231 TFs).
- TFBSs in enrichment zones show high PBM binding affinity and are found in high quality peaks.
- Direct TF-DNA interactions reveal co-binding TFs.
- *cis*-regulatory modules derived from TFBSs are enriched for disease- and trait-associated SNPs.

Gheorghe *et al.*. 2018.

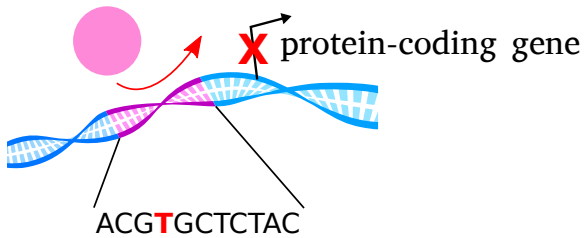
# Outline

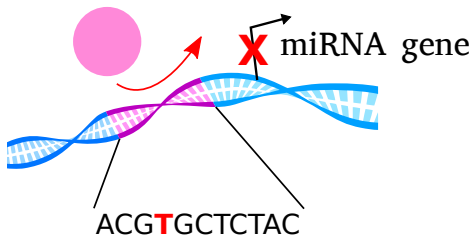
1. Improving our capacity to predict transcription factor binding events

2. Combining transcriptional and post-transcriptional regulation to predict mutations altering the gene regulatory program in cancer cells

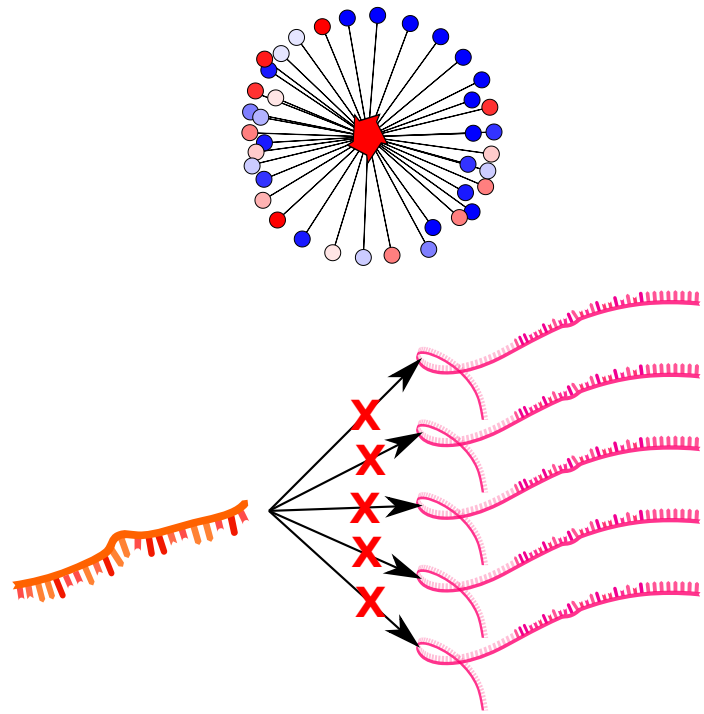
# Predicting somatic mutations altering the gene regulatory program in cancer cells

**Transcriptional dysregulation**



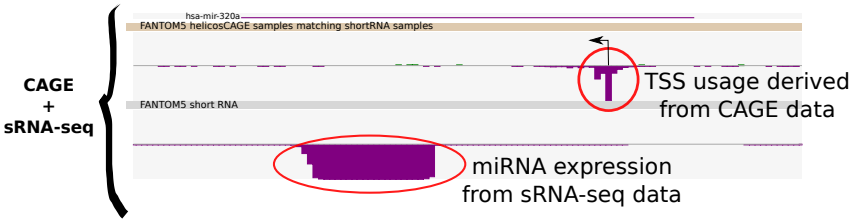




**Cascading effect**



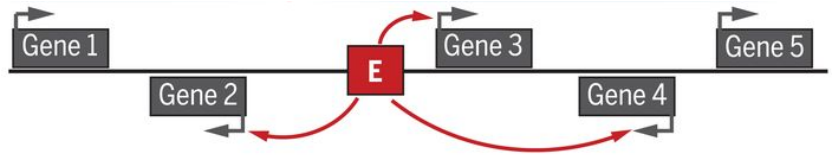
# Data to analyze transcriptional regulation of miRNAs and protein-coding genes

## miRNA TSSs




De Rie *et al.*, 2017.

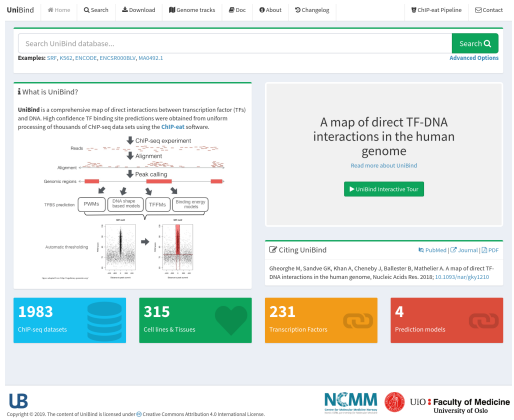
## Enhancer-TSS associations




Furlong *et al.*, 2018.

Fishilevich *et al.*, 2017.

## TFBSs from UniBind




Gheorghe *et al.*, 2018.

## miRNA - target networks


Agarwal *et al.*, 2015.

# Available cancer data cohorts

- Data requirements:
  - WGS (tumor and normal) to call SNVs and indels
  - RNA-seq
  - Copy number alterations

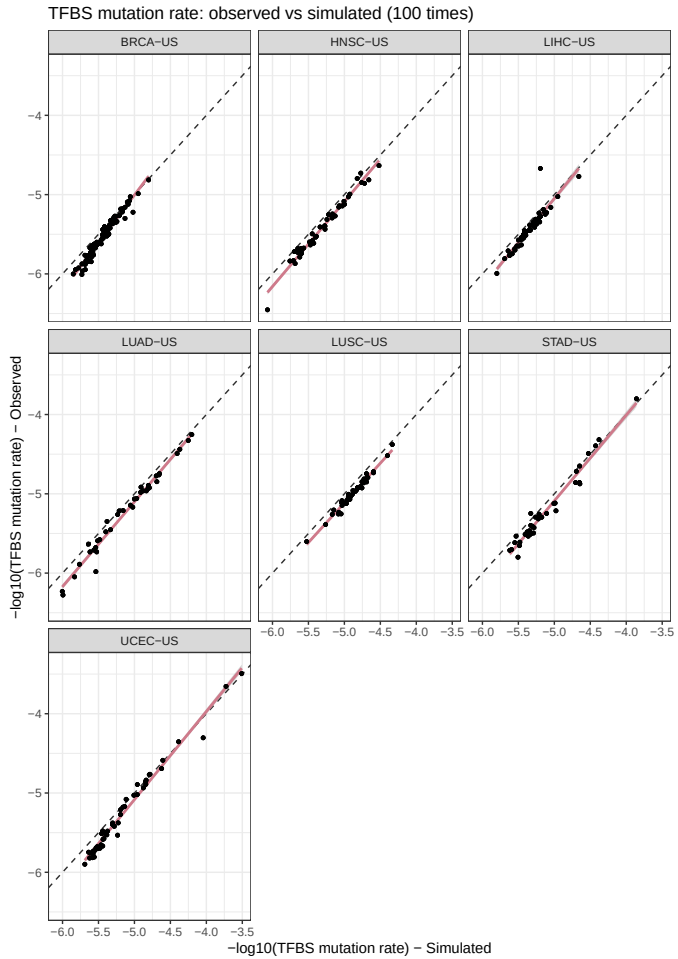
- Cohorts:
  - TCGA: 343 samples (7 cancer types)

| | | |
|---|---|---|
| BRCA | Breast cancer | 91 |
| LIHC | Liver Hepatocellular Carcinoma | 50 |
| UCUC | Uterine Corpus Endometrial Carcinoma | 48 |
| HNSC | Head and Neck Squamous Cell Carcinoma | 43 |
| LUSC | Lung Squamous Cell Carcinoma | 42 |
| LUAD | Lung Adenocarcinoma | 37 |
| STAD | Gastric Adenocarcinoma | 35 |

  - BASIS: 296 breast cancer samples

# TFBSs are less mutated than expected by chance

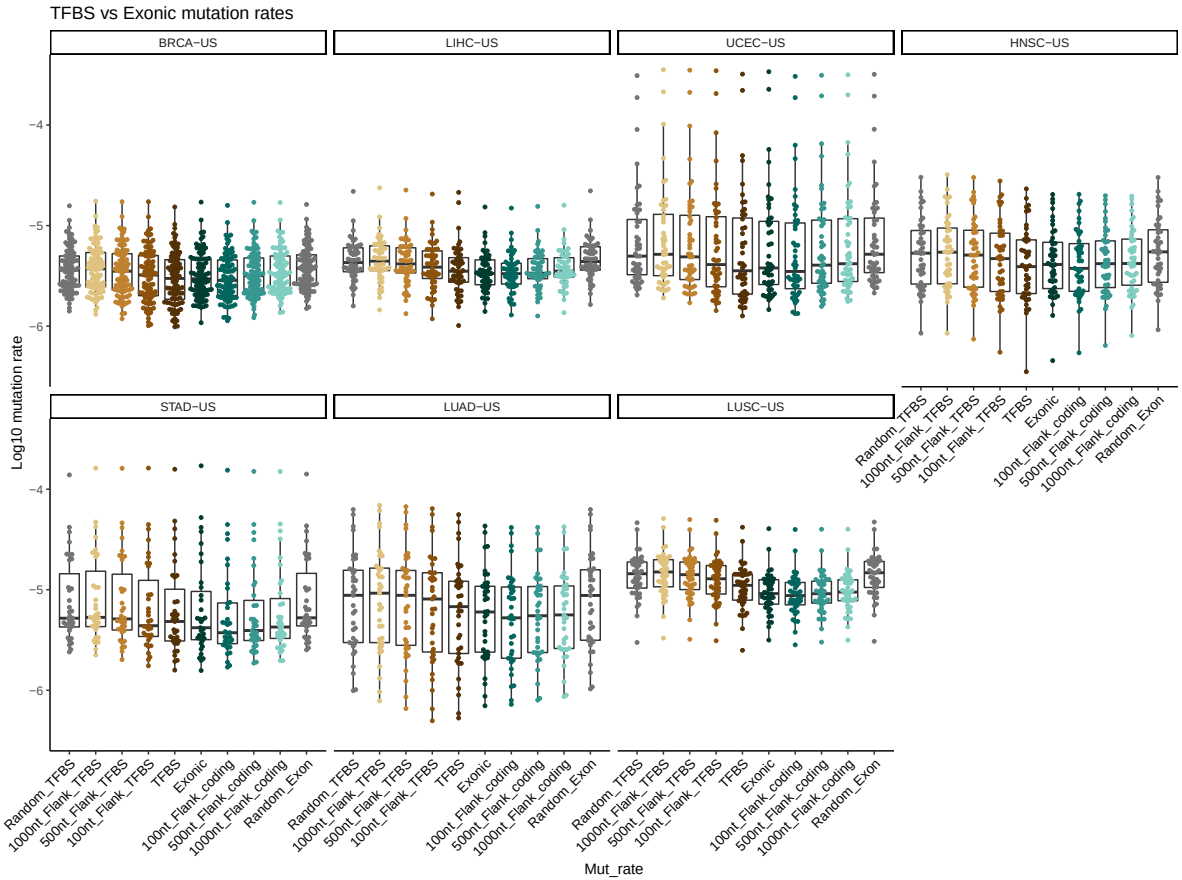
13,107,508 mutations in TCGA (from 211,421 to 2,141,178 per cohort).




Unpublished

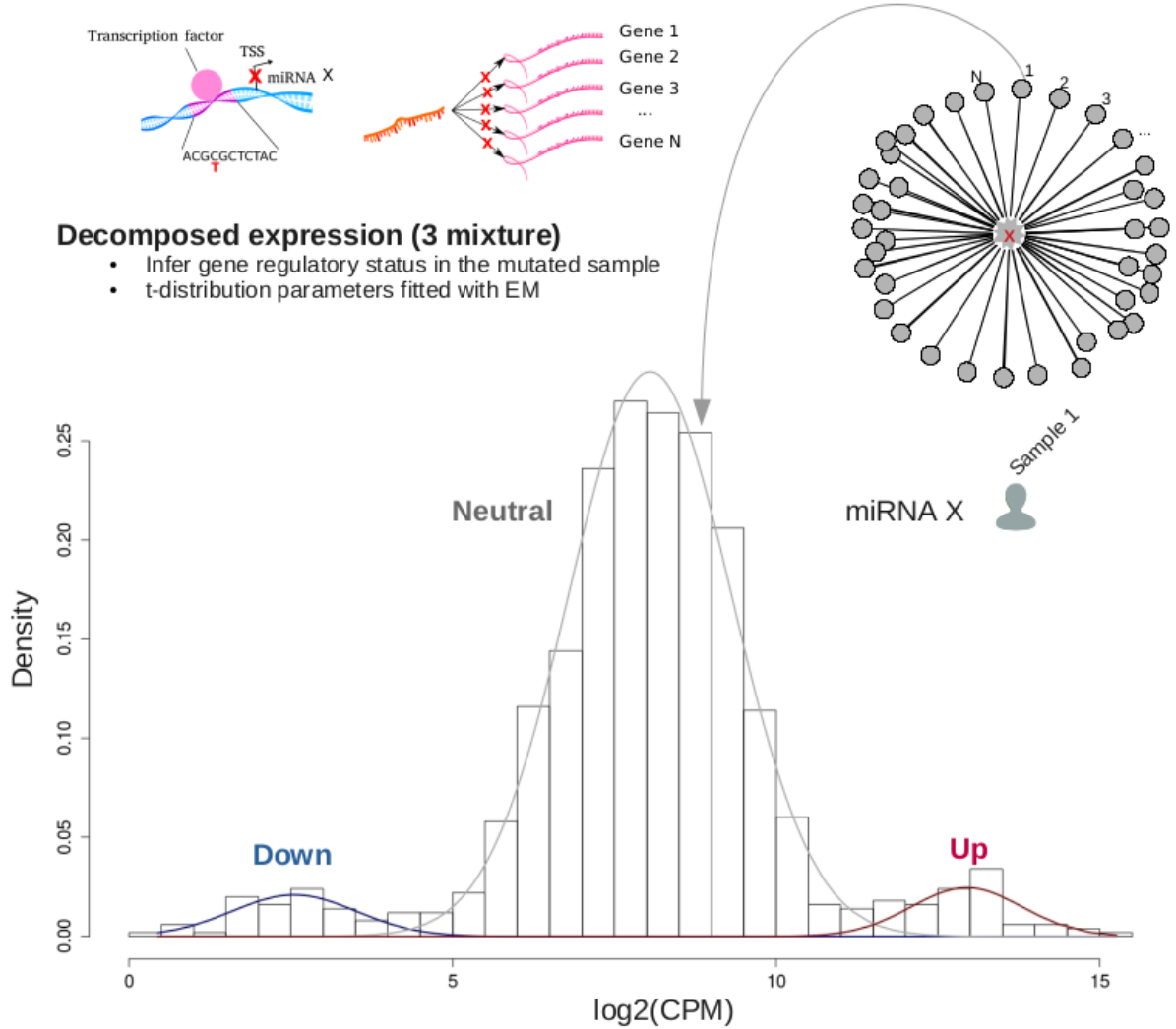
# TFBSs are less mutated than expected by chance

13,107,508 mutations in TCGA (from 211,421 to 2,141,178 per cohort).



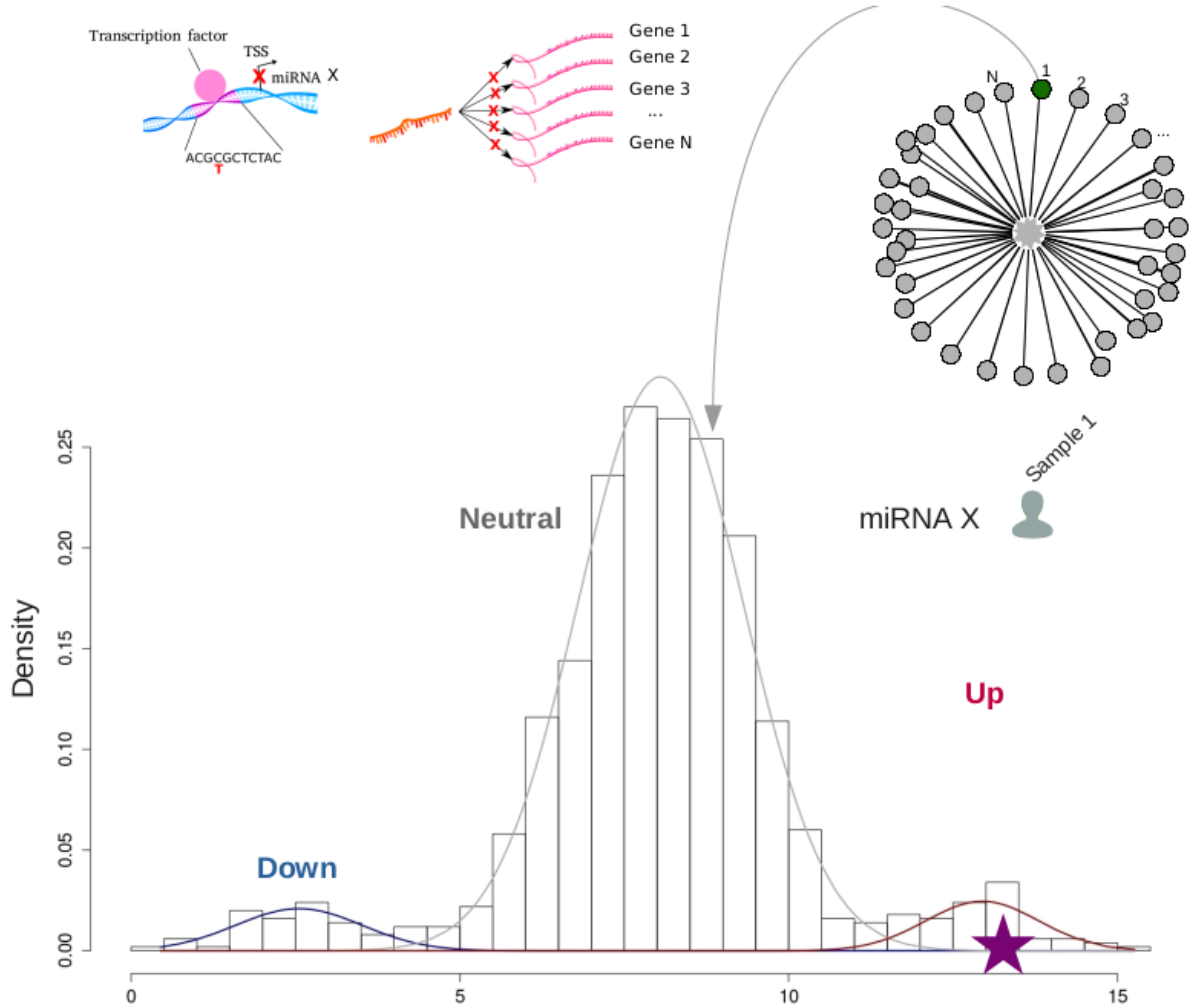

Unpublished

# Prediction of gene expression dysregulation



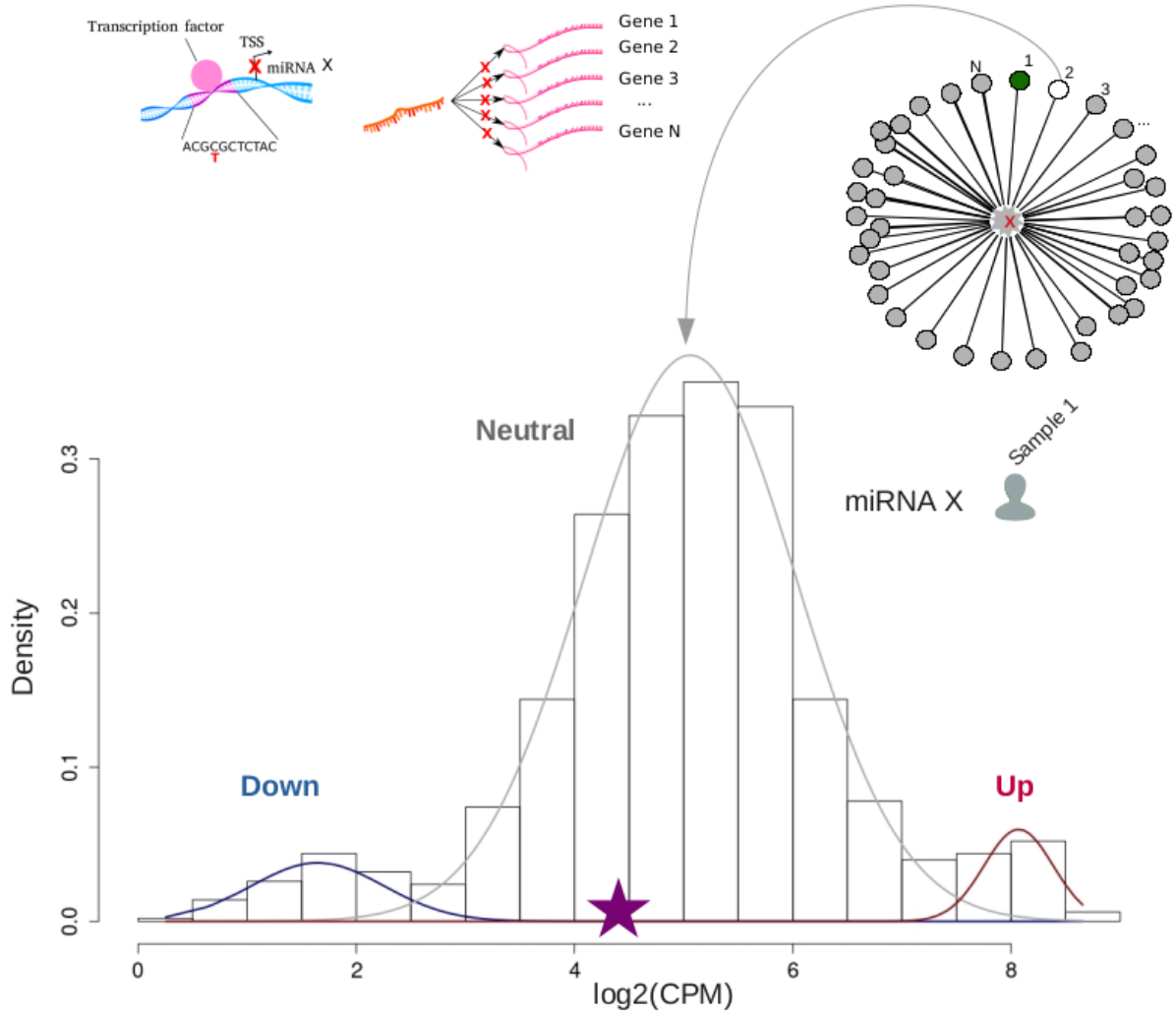

1. Ding et al. (2015) Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. Nat Comms

# Prediction of gene expression dysregulation



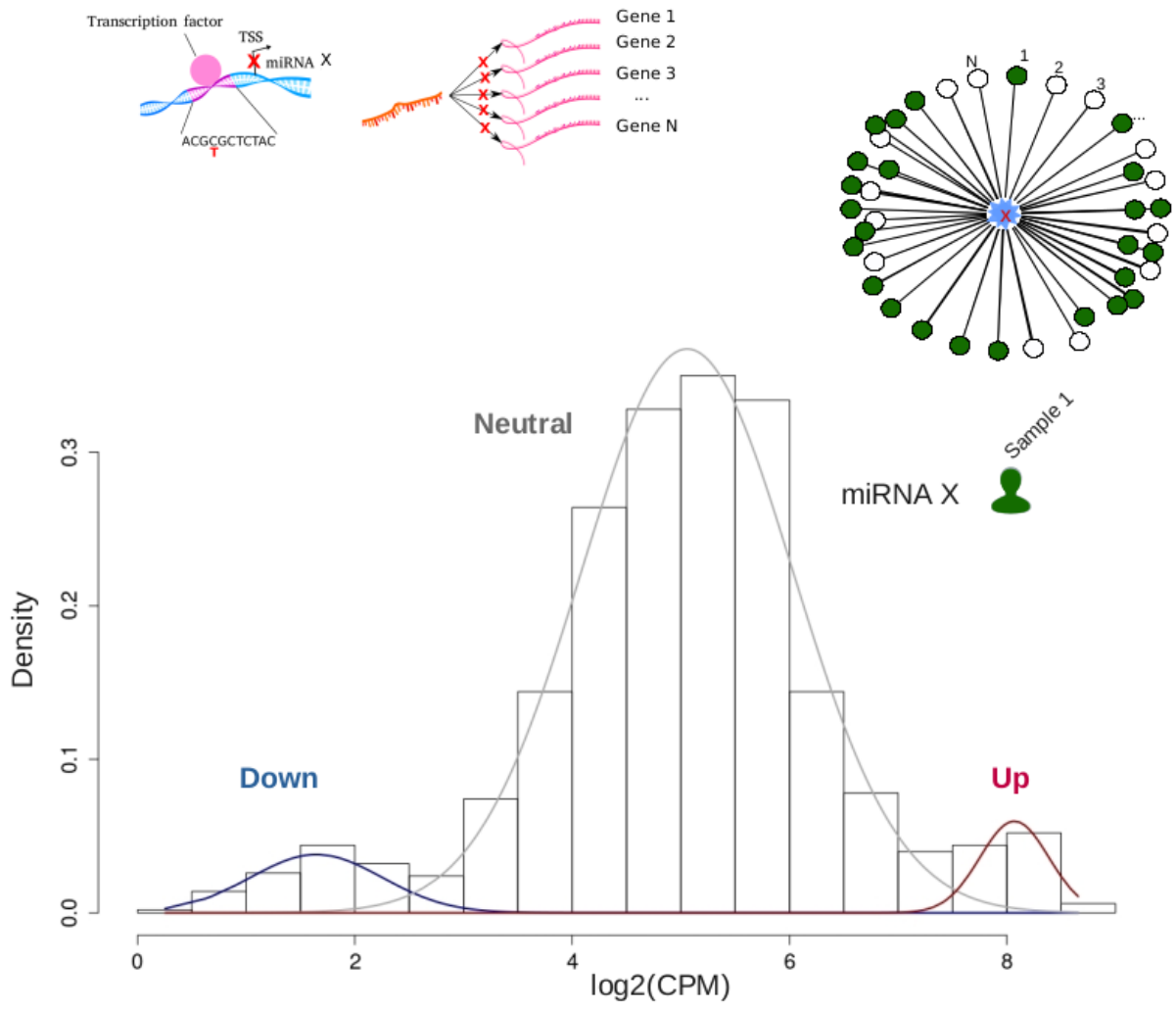

1. Ding et al. (2015) Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. Nat Comms

# Prediction of gene expression dysregulation



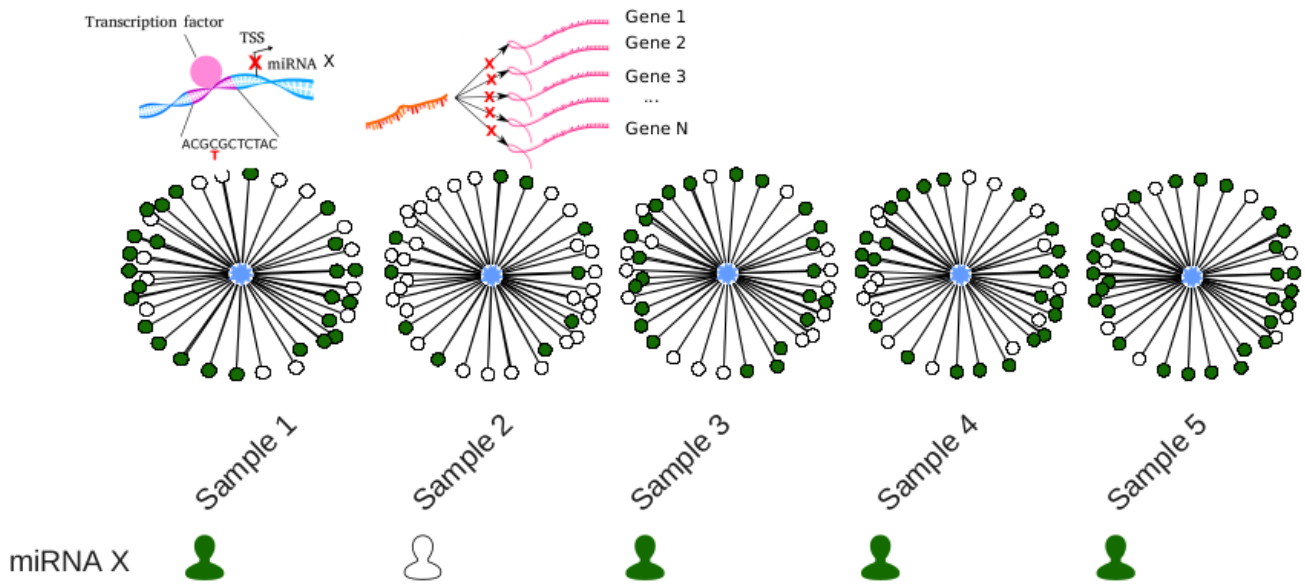

1. Ding et al. (2015) Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. Nat Comms

# Prediction of gene expression dysregulation



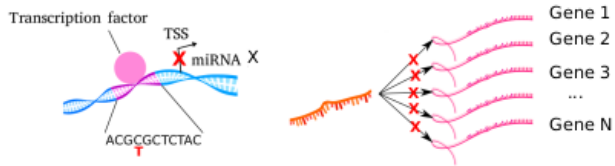

1. Ding et al. (2015) Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. Nat Comms

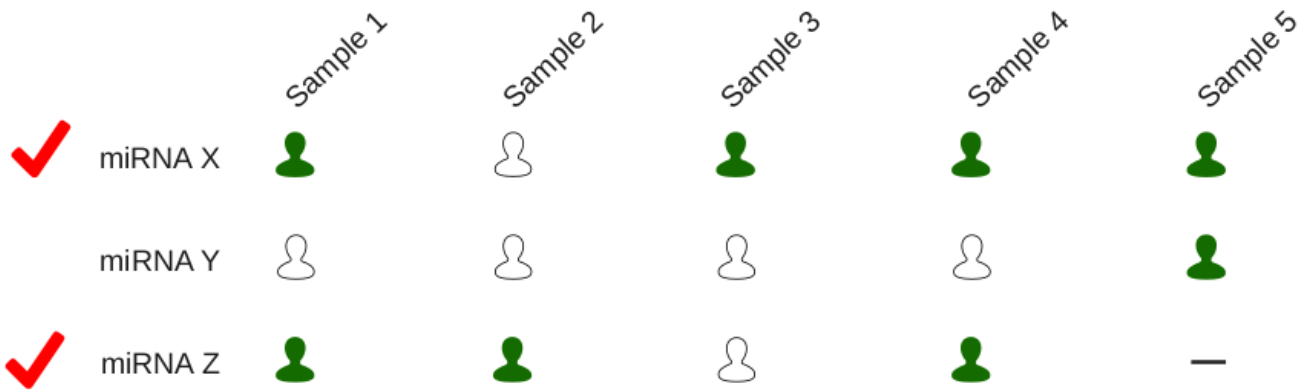
# Prediction of gene expression dysregulation




1. Ding et al. (2015) Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. Nat Comms

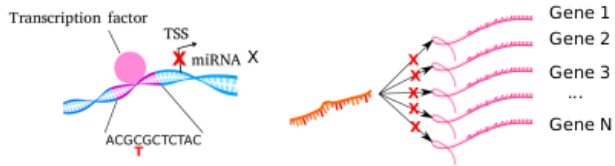
# Prediction of gene expression dysregulation



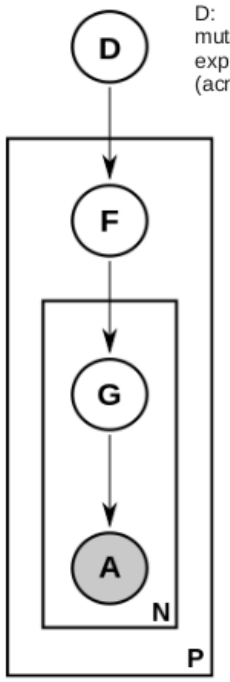




1. Ding et al. (2015) Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. Nat Comms

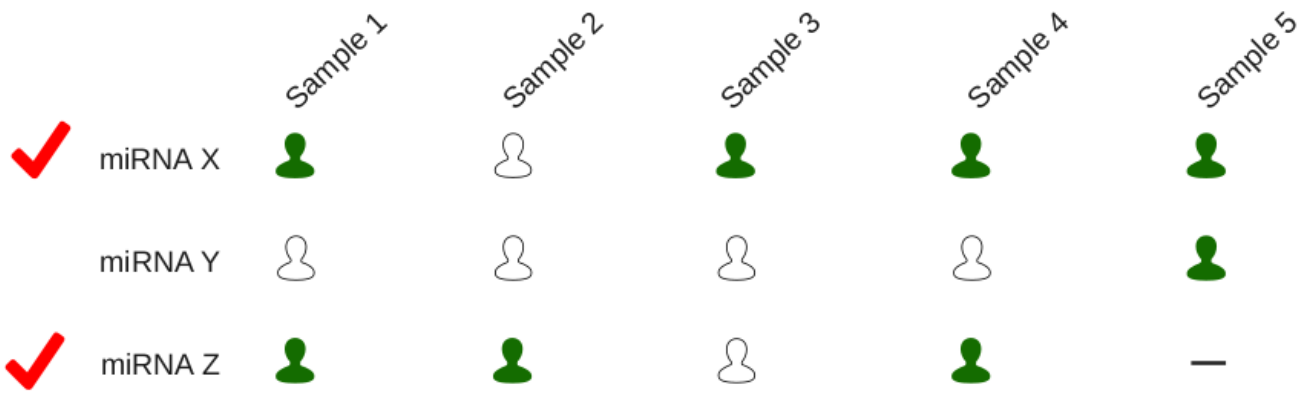
# Prediction of gene expression dysregulation




**Bayesian network**



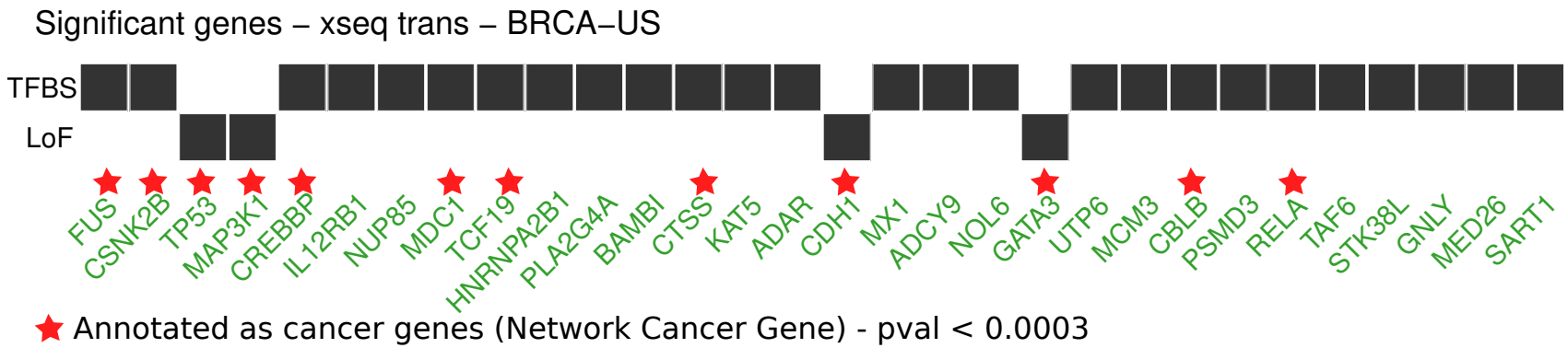

D: likely association between mutations and target gene expression dysregulation (across the cohort).




**N:** genes connected to gene A.
**P:** samples with mutations in gene A

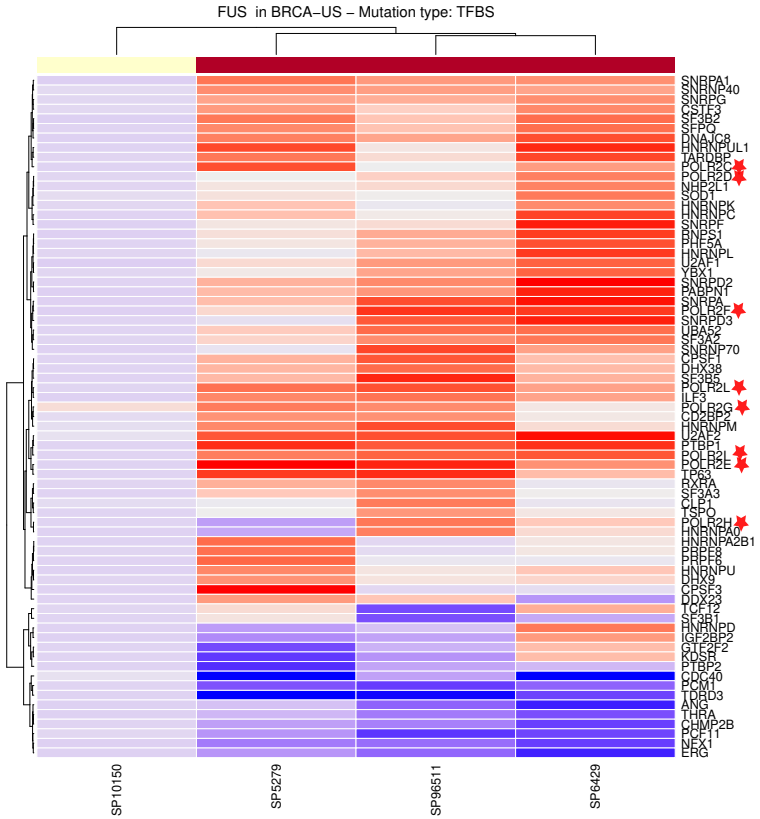
1. Ding et al. (2015) Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. Nat Comms

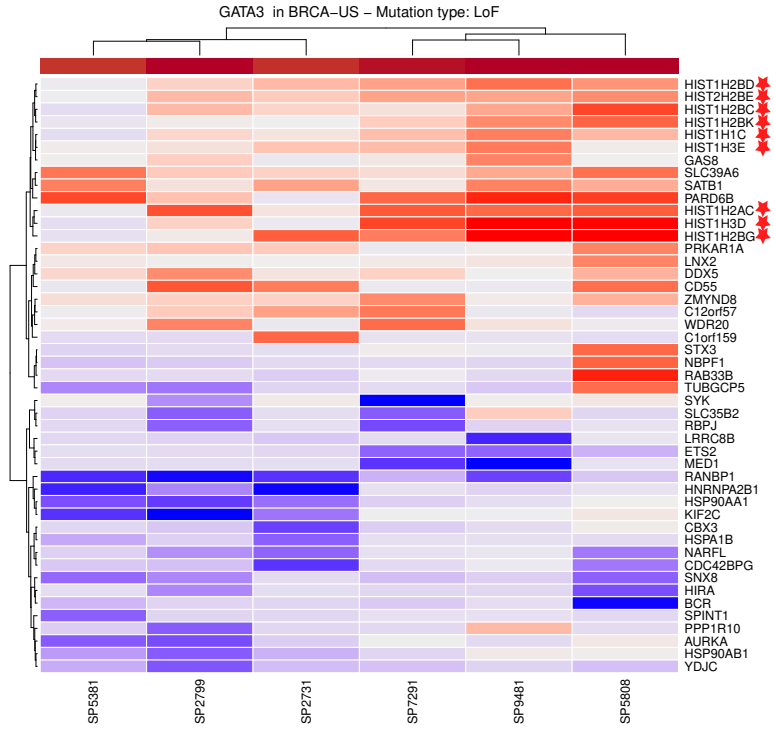
# Prediction results on protein-coding genes in breast cancer

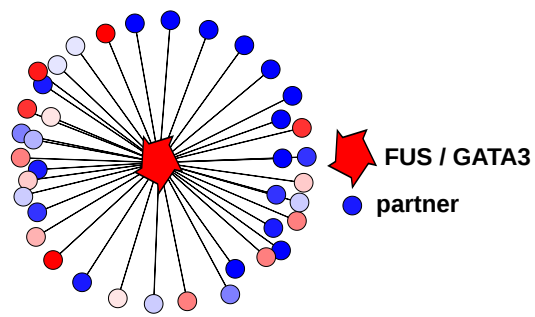



Unpublished

# Visualization of the regulatory network alteration

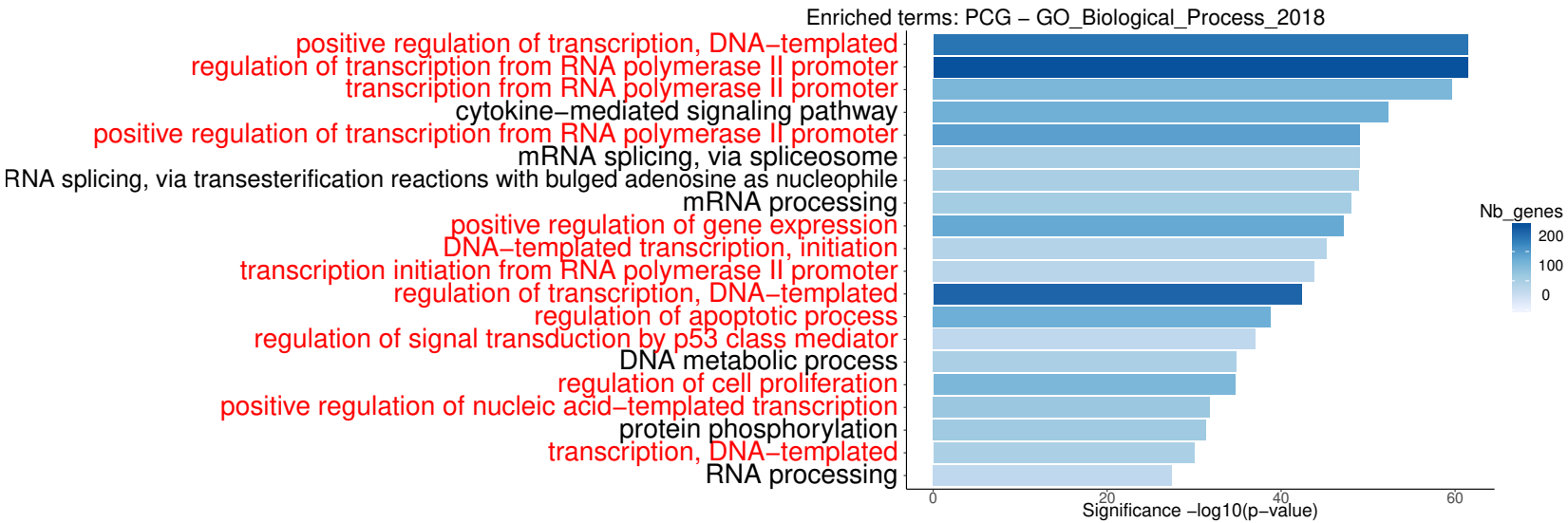


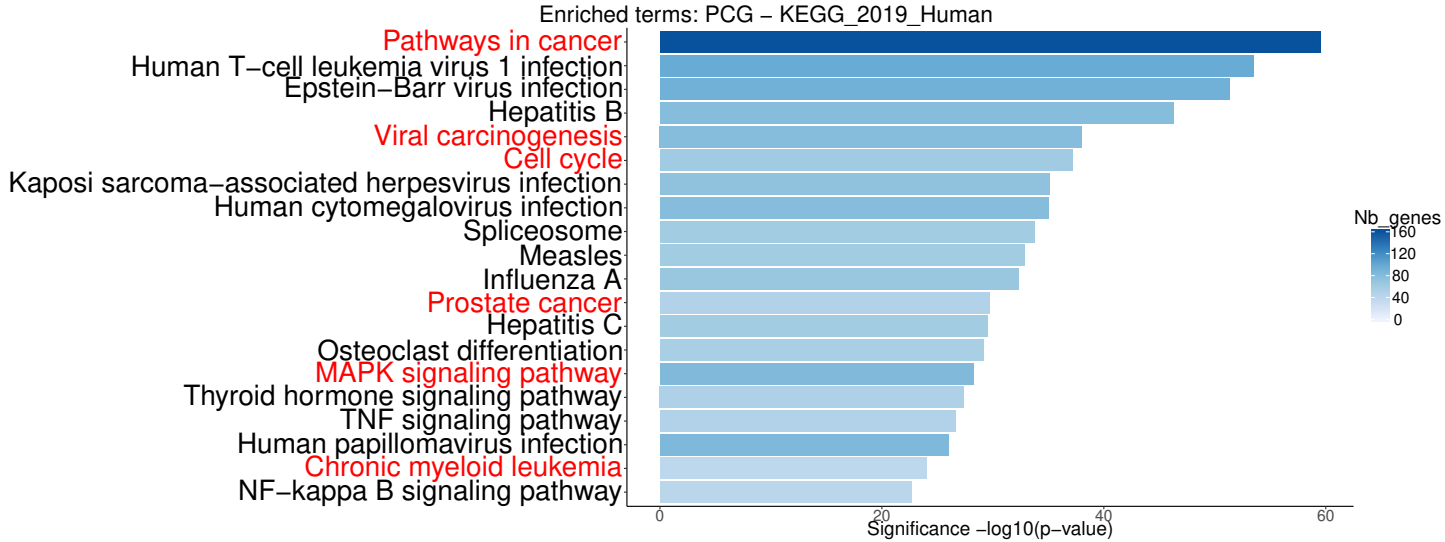

Unpublished

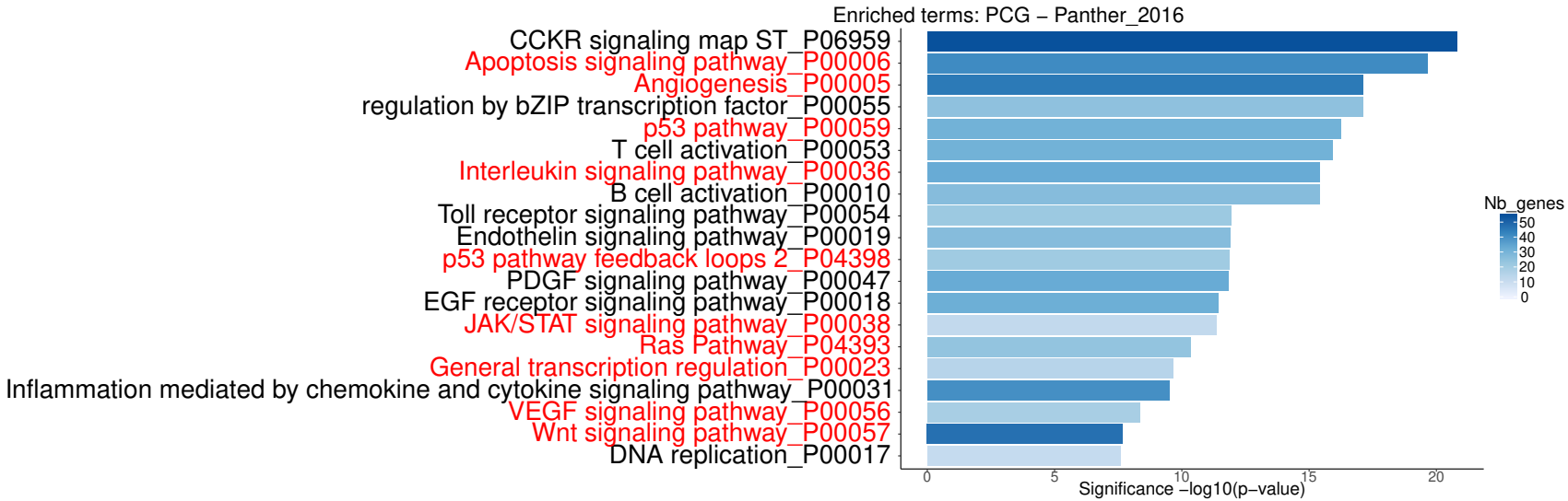
# Dysregulated networks for PCGs highlight key pathways




Unpublished

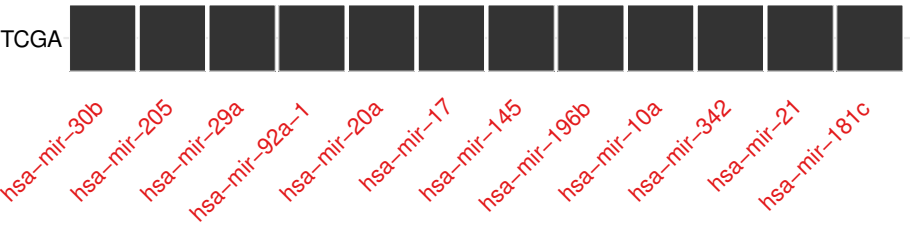
# Dysregulated networks for PCGs highlight key pathways

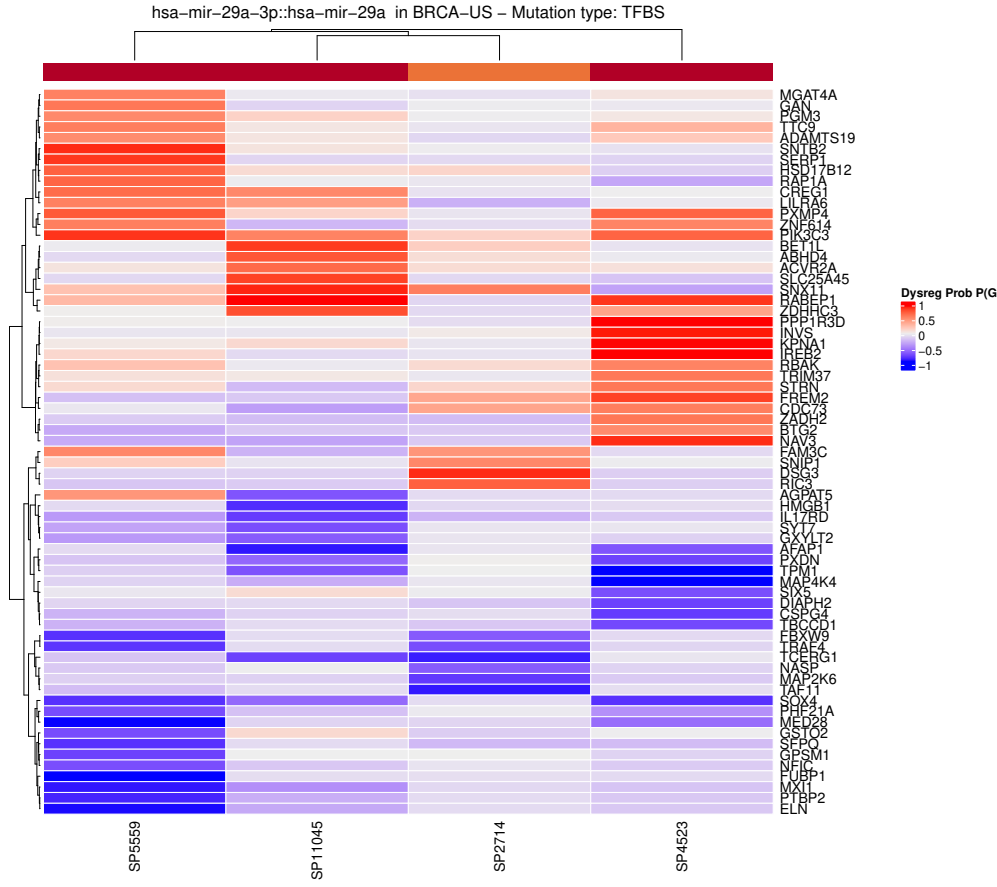



Unpublished

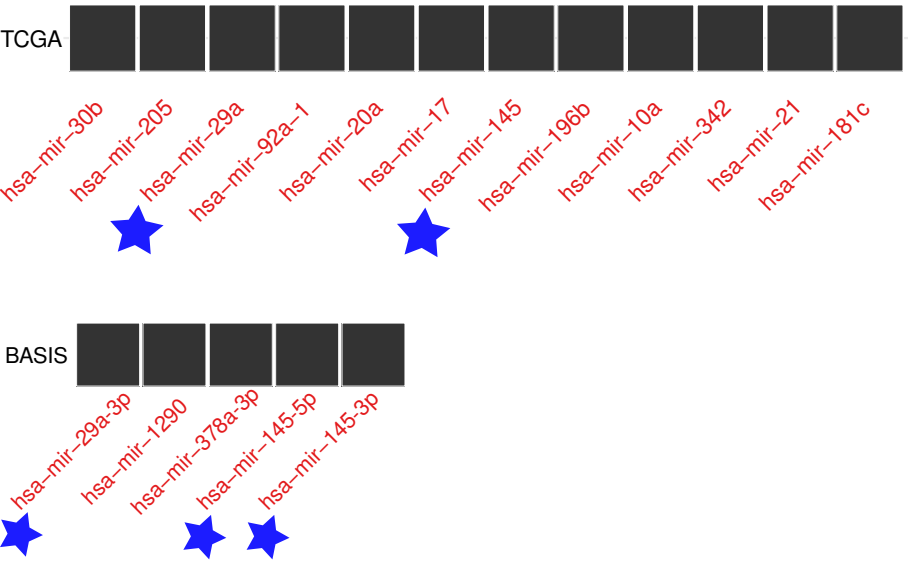
# Predictions on miRNA genes in breast cancer

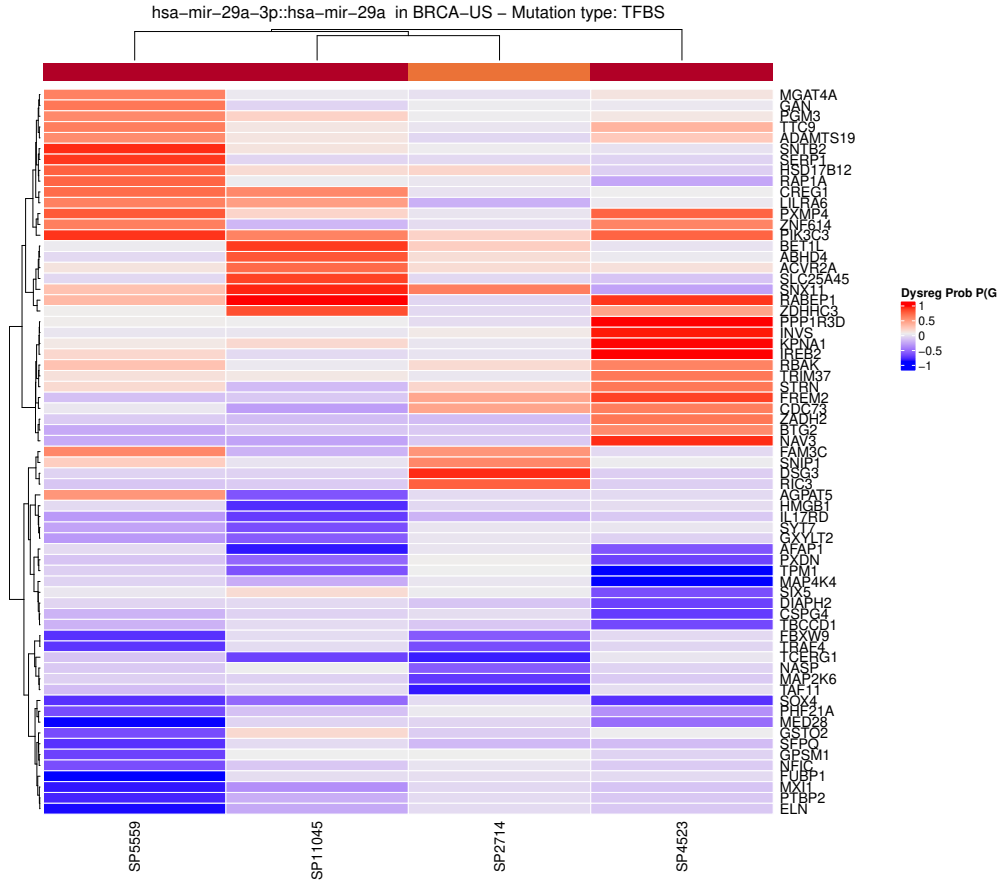



Unpublished

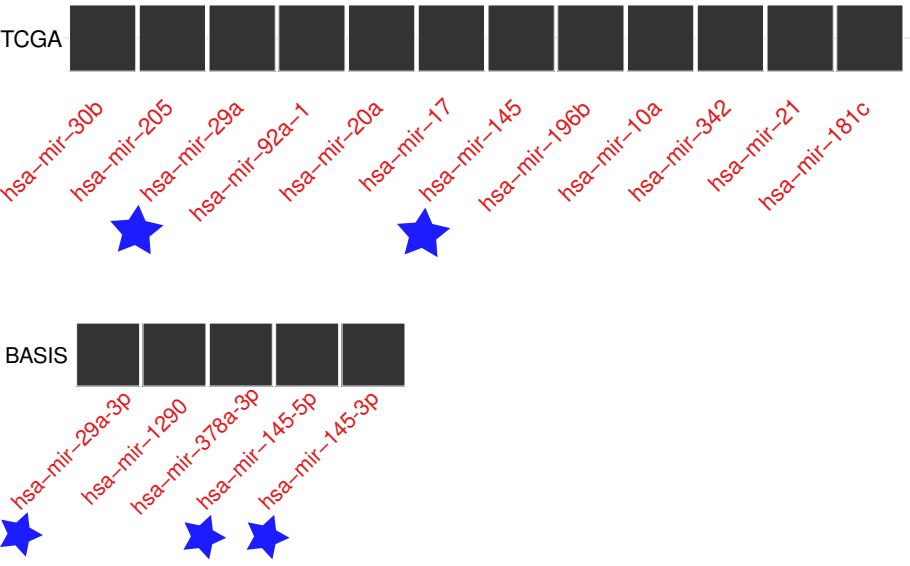
# Predictions on miRNA genes in breast cancer

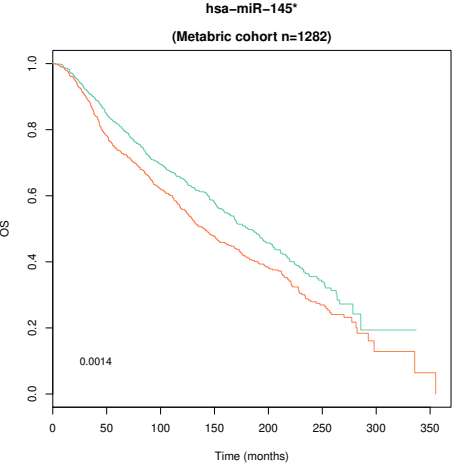


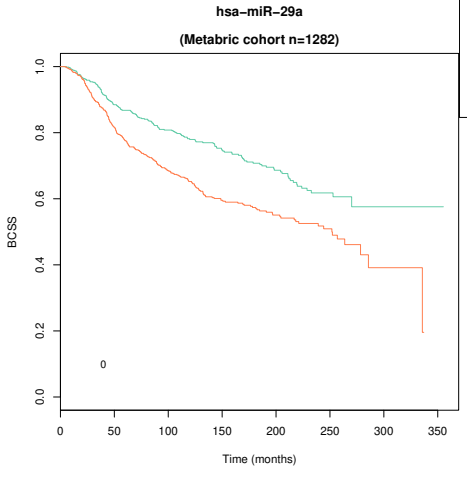



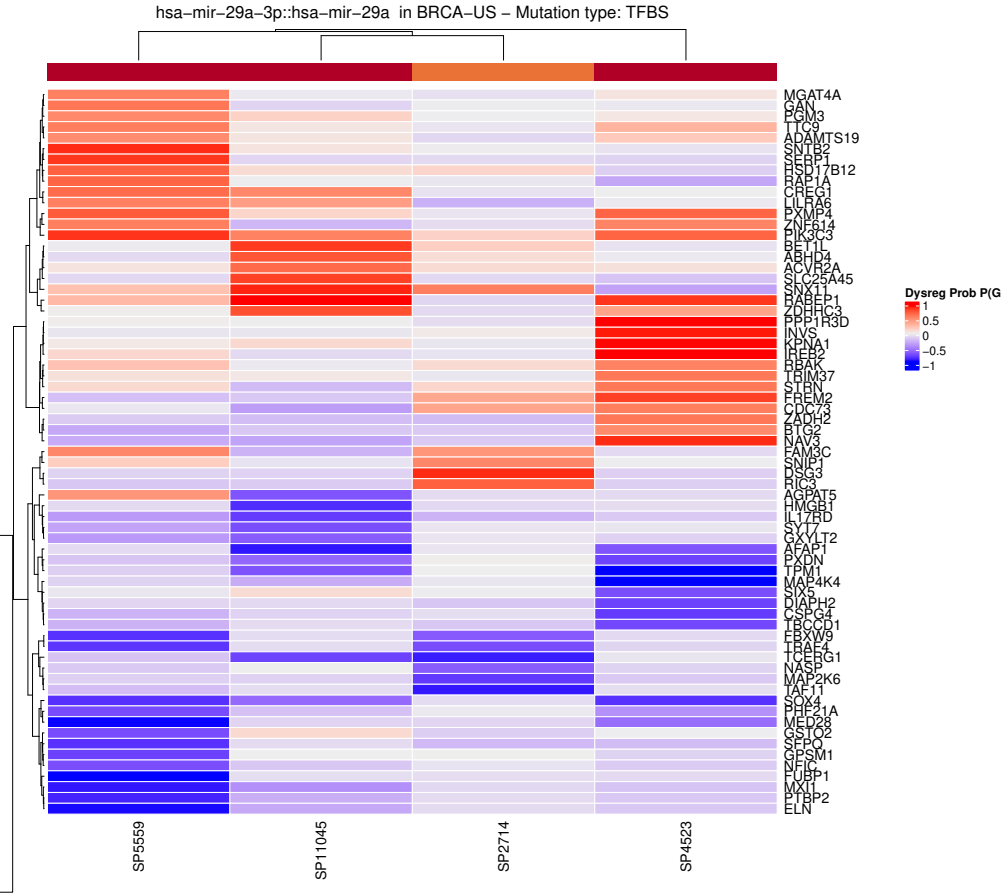

Unpublished

# Predictions on miRNA genes in breast cancer



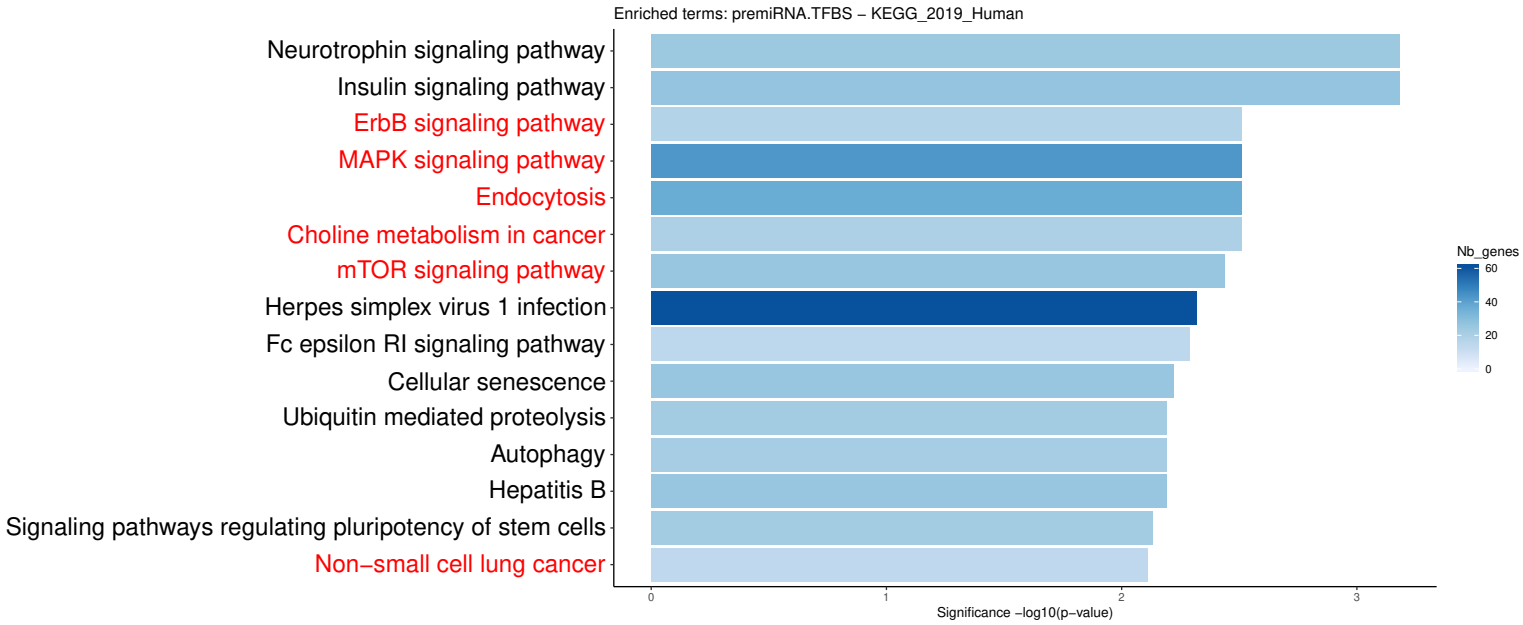



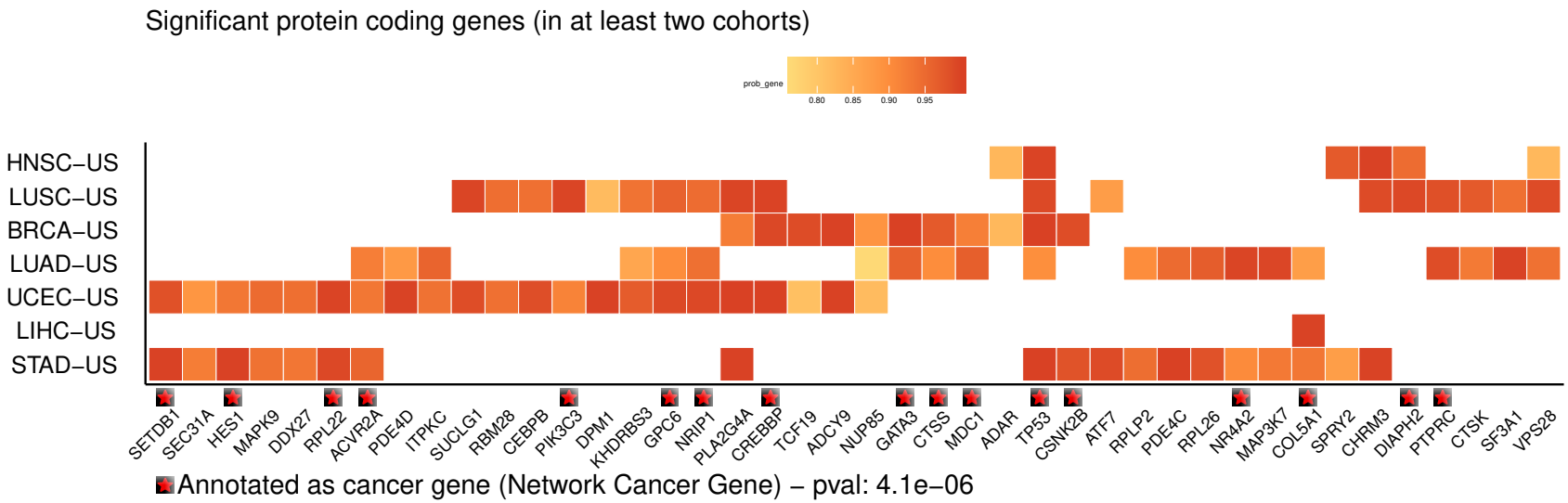

Unpublished

# Dysregulated networks for miRNAs highlight key pathways



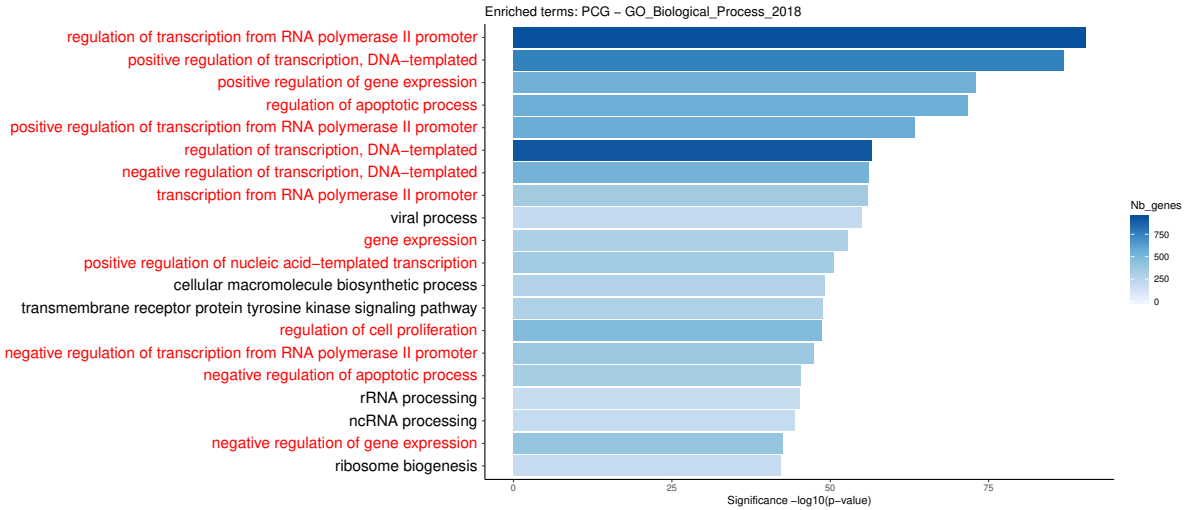

Unpublished

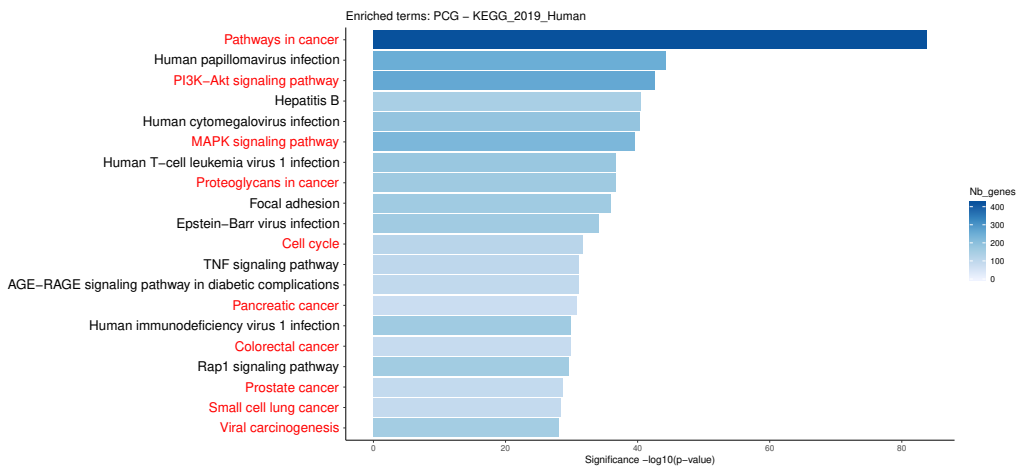
# Pan-cancer analysis of PCG networks dysregulation

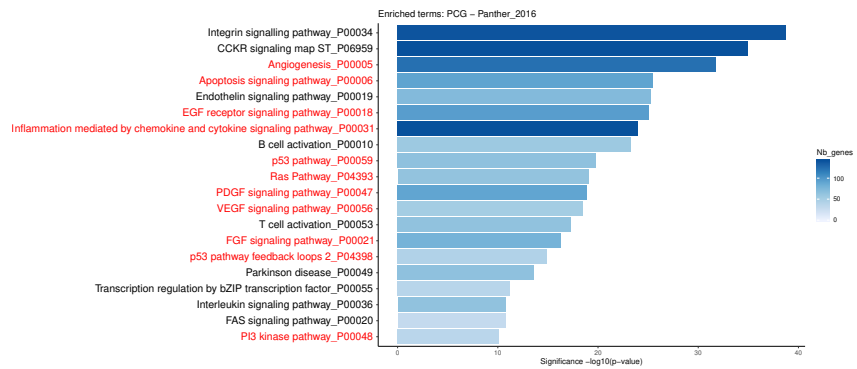



Unpublished

# Pan-cancer analysis of PCG networks dysregulation

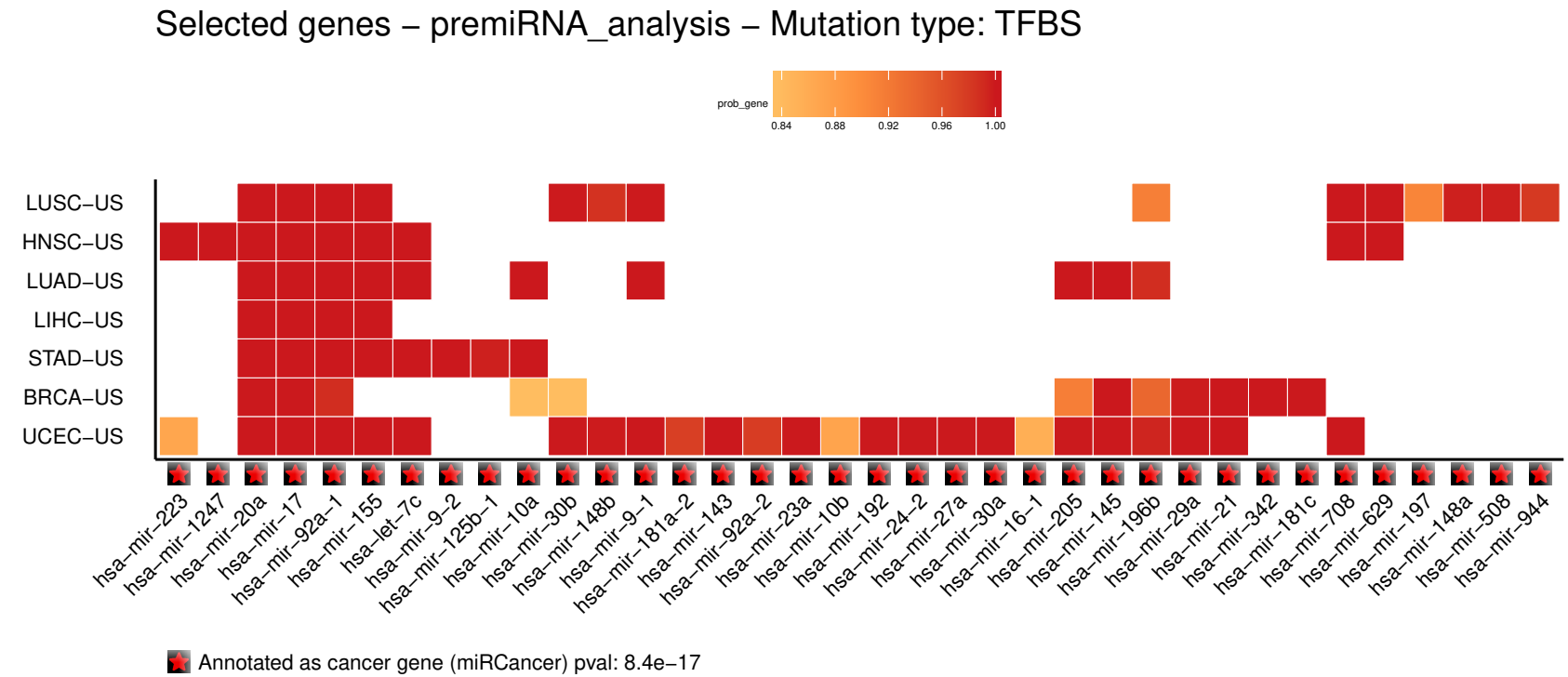


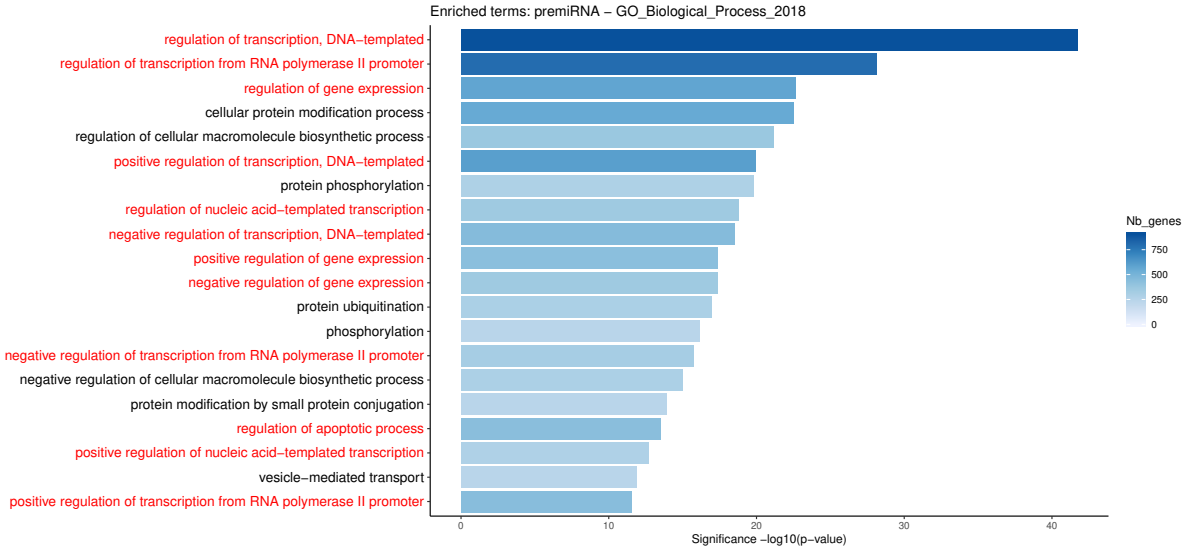

Unpublished

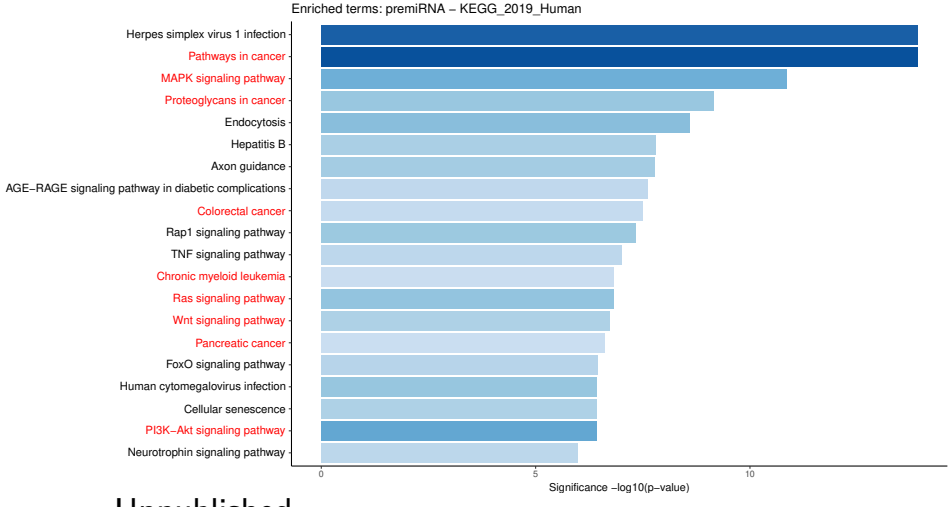
# Pan-cancer analysis of miRNA networks dysregulation

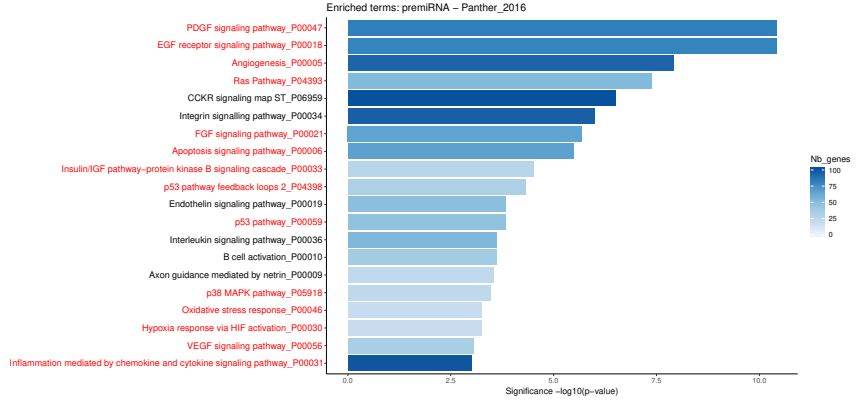



Unpublished

# Pan-cancer analysis of miRNA networks dysregulation




Unpublished

# Summary

- Combining LoF mutations with mutations in TFBSs highlights candidate driver PCGs pan-cancer

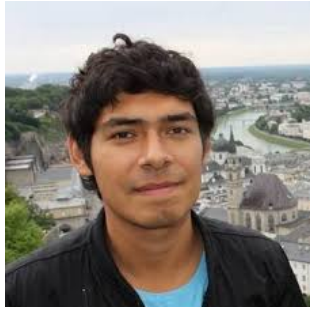
- dysmiR predicts candidate driver miRNAs that are dysregulated through cis-regulatory mutations with cascading effects on the gene expression regulation program.

- We highlight candidate regulatory-disrupting variations dysregulating the gene expression regulatory program in cancer pathways

Unpublished

# Acknowledgements

**Marius Gheorghe**
ChIP-eat - UniBind
JASPAR - ReMap



**Jaime Castro-Mondragon**
Cancer deregulation
JASPAR

**Miriam Ragle-Aure**
Cancer deregulation

**Aziz Khan**
JASPAR
UniBind

**JASPAR**:

- Oriol Fornes
- Arnaud Stigliani
- Robin van der Lee
- Adrien Bessy
- Jeanne Cheneby
- Shubhada Kulkarni
- Ge Tan
- Damir Baranasic
- David Arenillas
- Albin Sandeling
- Klaas Vandepoele
- Boris Lenhard
- Benoit Ballester
- Wyeth Wasserman
- Francois Parcy

**ChIP-eat - UniBind - ReMap**:

- Geir Kjetil Sandve
- Jeanne Cheneby
- Marie Artufel
- Benoit Ballester

**Cancer deregulation**:

- Vessela Kristensen
- Anne-Lise Borresen-Dale
- Anita Langerod
- BASIS consortium

NCMM

UNIVERSITAS OSLOENSIS MDCCCXI

The Research Council of Norway

HELSE SØR-ØST