Functional transcription factor target discovery via compendia of binding and expression profiles

Chris Banks, Anagha Joshi, Tom Michoel

The Roslin Institute

9th October 2015

Introduction

This talk is on a method for predicting the functional targets of transcription related factors.

Introduction

This talk is on a method for predicting the functional targets of transcription related factors.

- What this is and why it's important?
  - Gene regulation—a computer scientist's view;
  - How to measure factor binding—ChIP-sequencing;
  - The challenge...

Introduction

This talk is on a method for predicting the functional targets of transcription related factors.

- What this is and why it's important?
  - Gene regulation—a computer scientist's view;
  - How to measure factor binding—ChIP-sequencing;
  - The challenge...
- Our method:
  - "Guilt by association" method;
  - Our data;
  - Power of different statistics and the "wisdom of crowds".

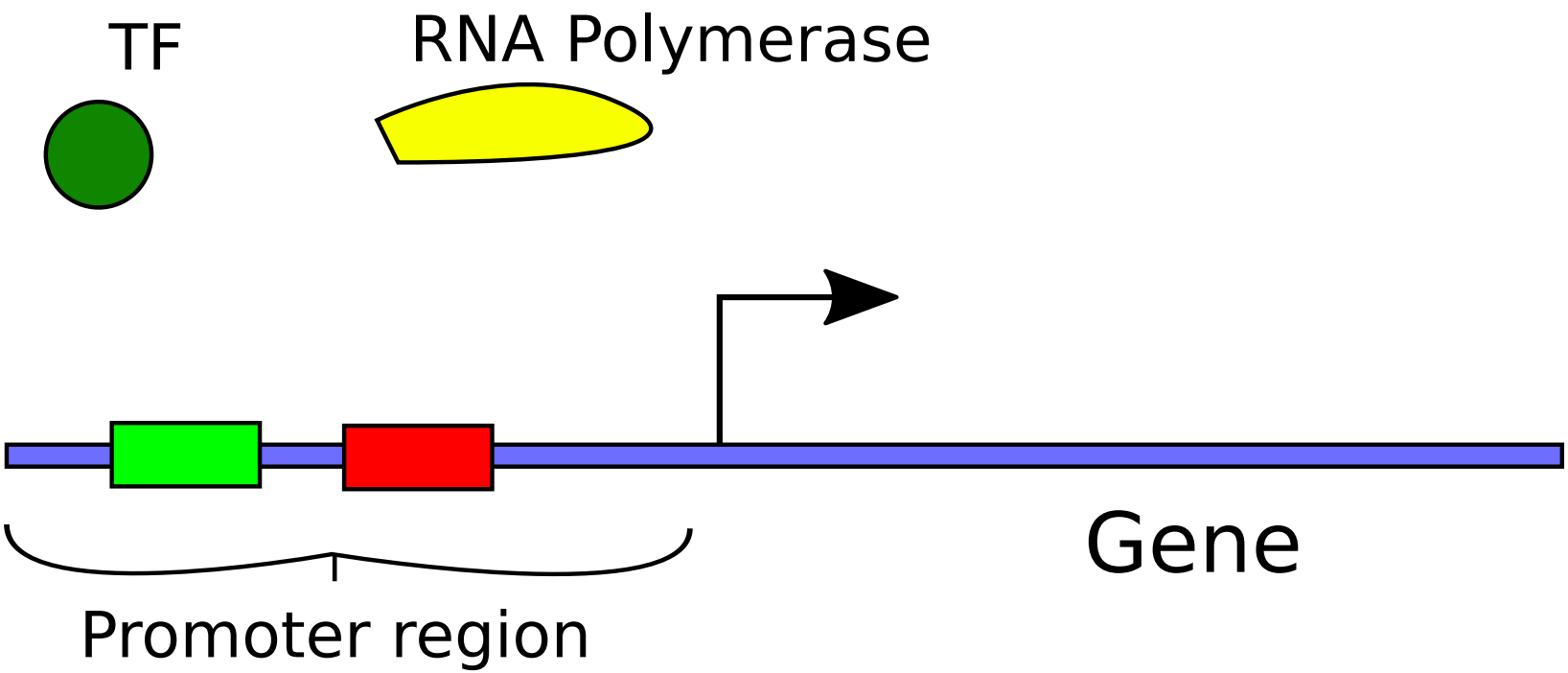
Introduction

This talk is on a method for predicting the functional targets of transcription related factors.

- What this is and why it's important?
  - Gene regulation—a computer scientist's view;
  - How to measure factor binding—ChIP-sequencing;
  - The challenge...
- Our method:
  - "Guilt by association" method;
  - Our data;
  - Power of different statistics and the "wisdom of crowds".
- Results.
- Future work.

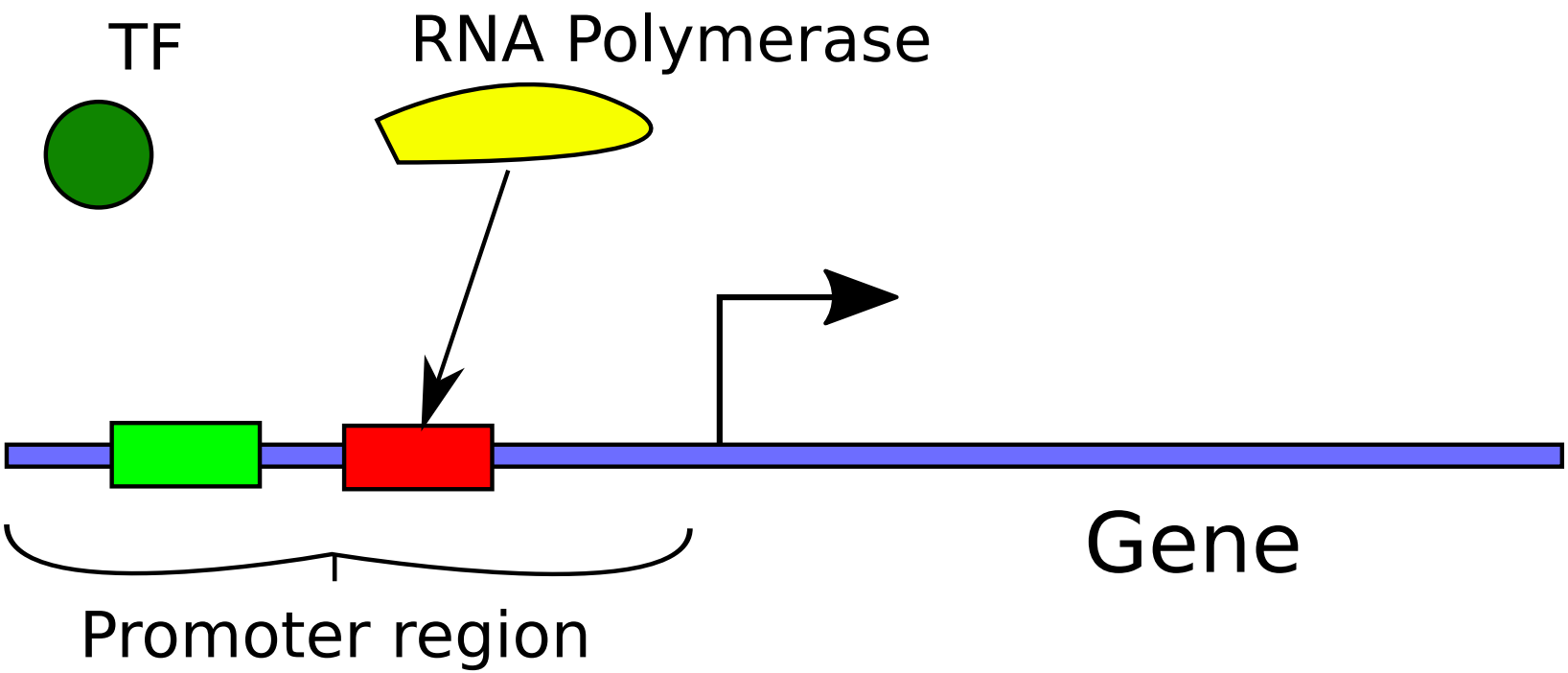
Gene regulation—a computer scientist's view

Transcriptional regulation by DNA-binding transcription factors (TFs) is a fundamental process governing all cell behaviour.



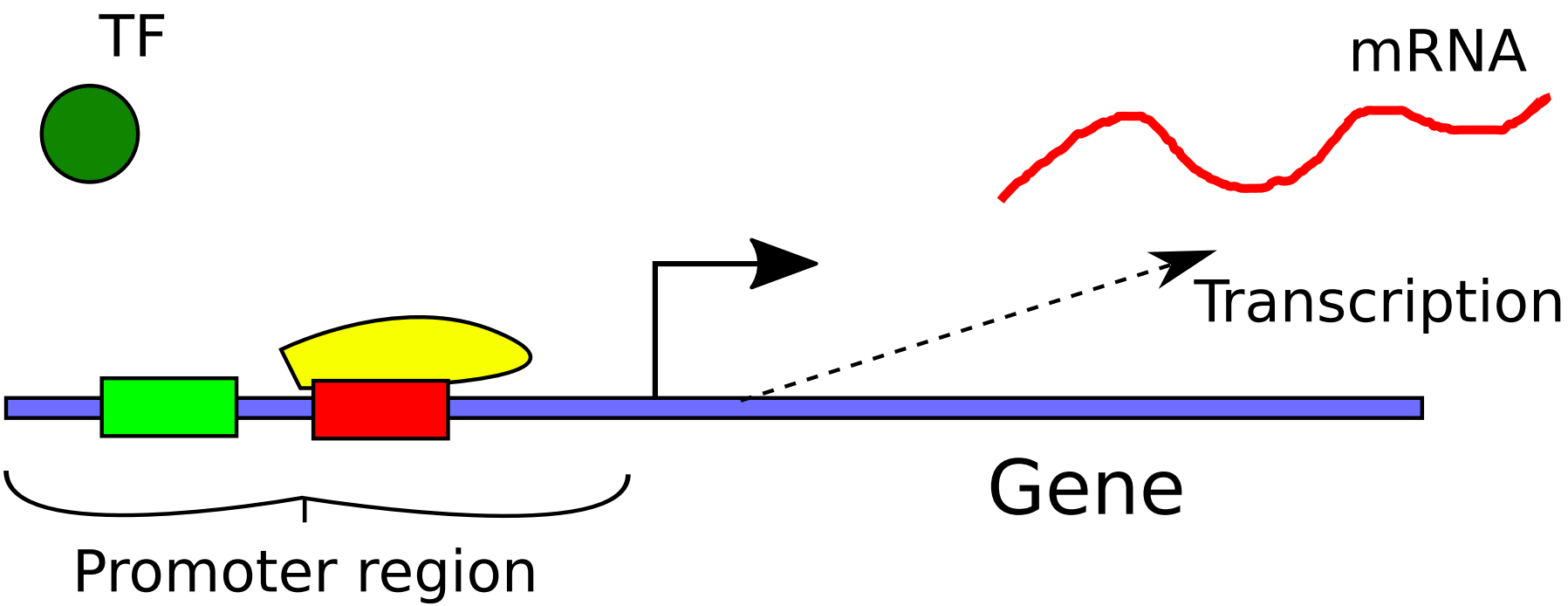
Gene regulation—a computer scientist's view

Transcriptional regulation by DNA-binding transcription factors (TFs) is a fundamental process governing all cell behaviour.



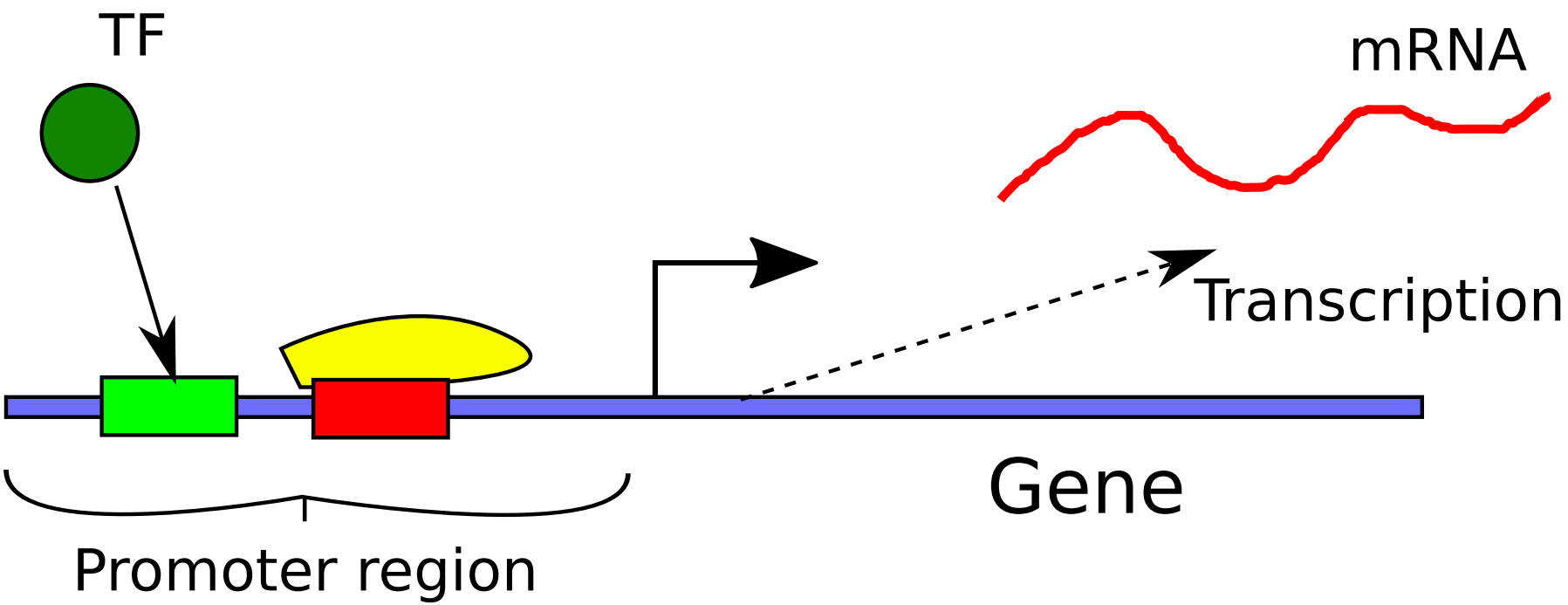
Gene regulation—a computer scientist's view

Transcriptional regulation by DNA-binding transcription factors (TFs) is a fundamental process governing all cell behaviour.



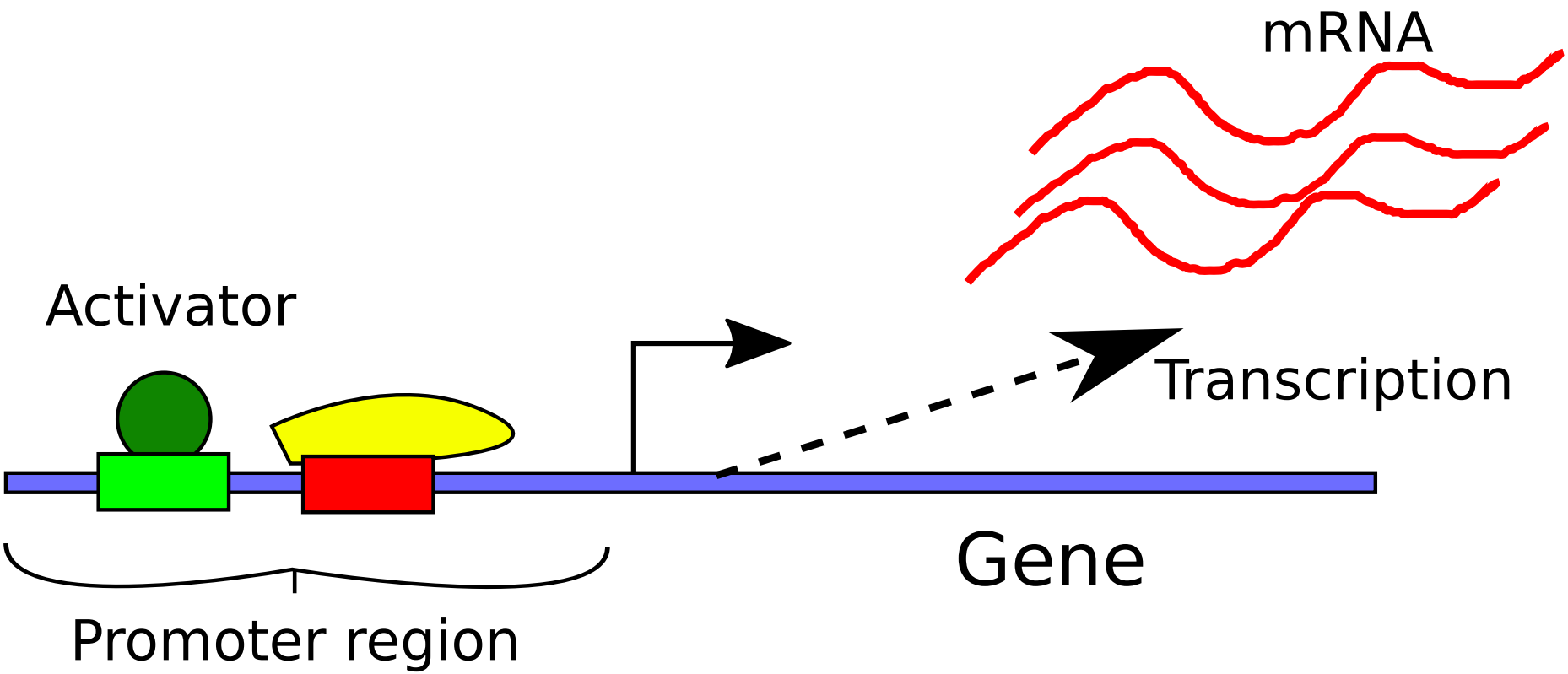
Gene regulation—a computer scientist's view

Transcriptional regulation by DNA-binding transcription factors (TFs) is a fundamental process governing all cell behaviour.



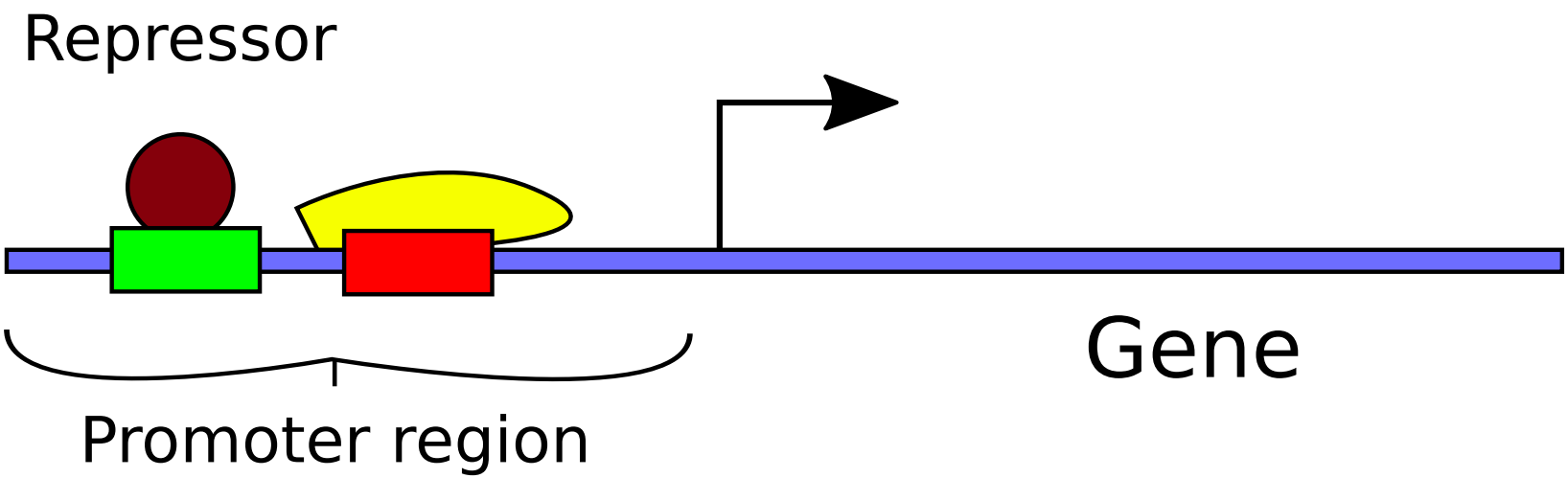
Gene regulation—a computer scientist's view

Transcriptional regulation by DNA-binding transcription factors (TFs) is a fundamental process governing all cell behaviour.

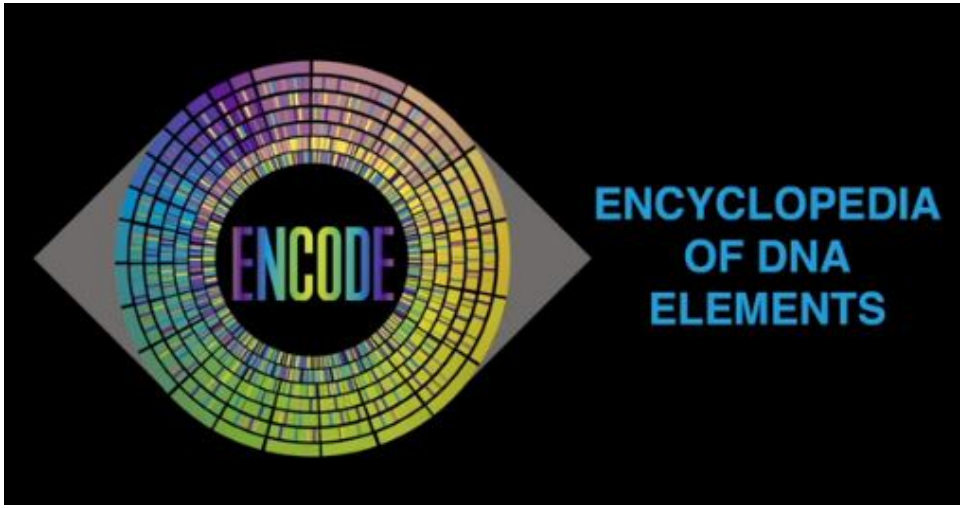


Gene regulation—a computer scientist's view

Transcriptional regulation by DNA-binding transcription factors (TFs) is a fundamental process governing all cell behaviour.

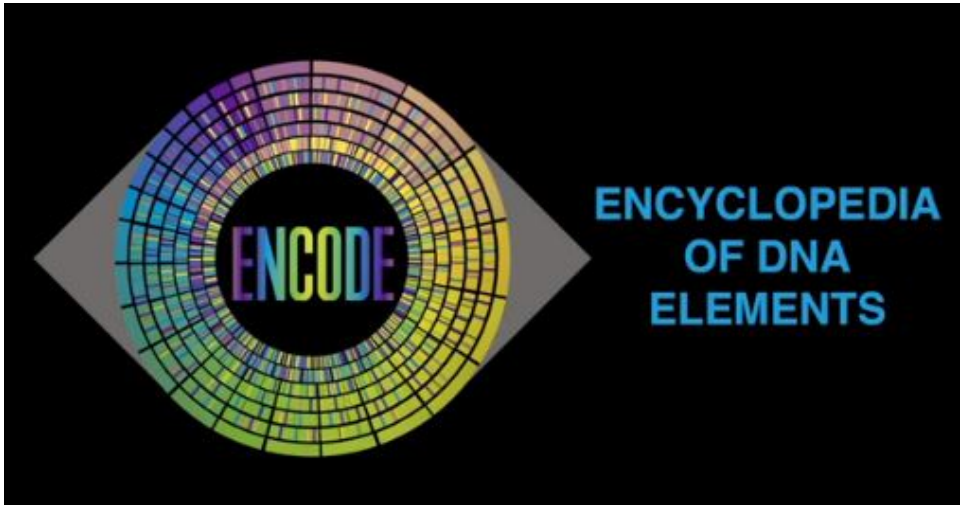


ChIP-sequencing



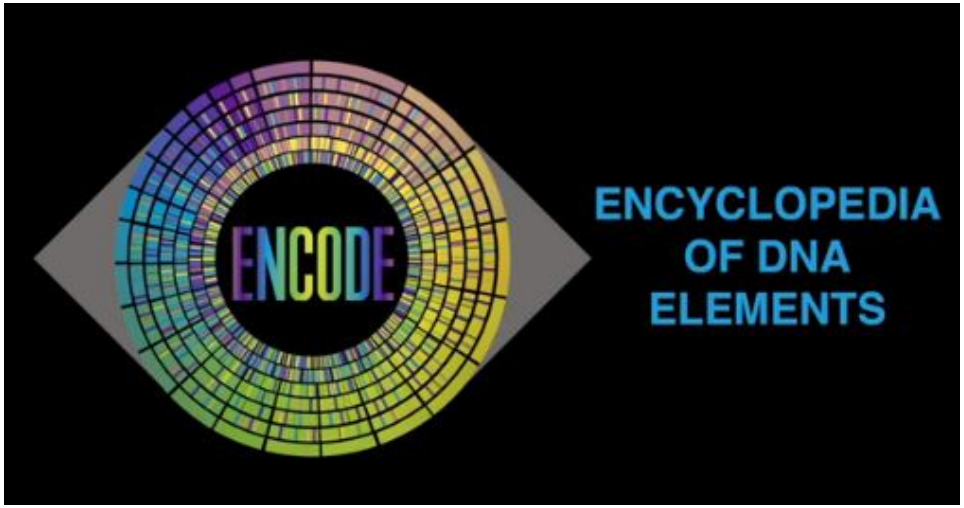
- The ENCODE project has extensively mapped the genomic locations where these TFs bind—

ChIP-sequencing



- The ENCODE project has extensively mapped the genomic locations where these TFs bind—
  - using ChIP-sequencing technology.

ChIP-sequencing



- The ENCODE project has extensively mapped the genomic locations where these TFs bind—
  - using ChIP-sequencing technology.

- **However:** it has been found that **most** TF binding within the promoter region of a gene **does not cause differential expression** of that gene. (*Cusanovich et al., 2014*)

The challenge

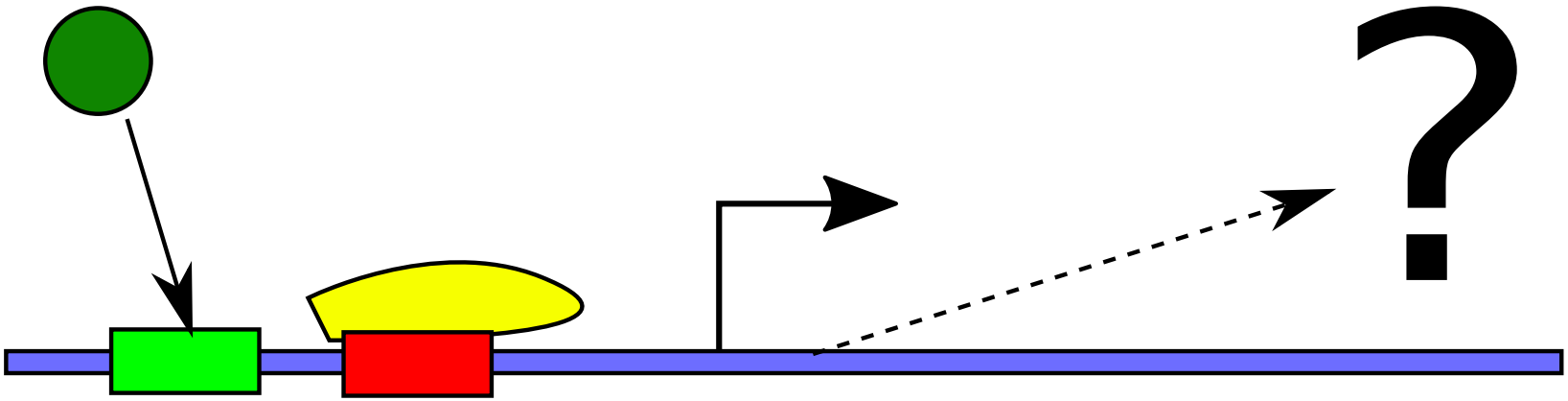
- So, the **big challenge** is:

The challenge

- So, the **big challenge** is:
- Given the emerging view is that binding sites are **redundant** (*Spivakov, 2014*) and **cumulative binding** is the key to **robustness**.
  - I.e. if some site is lost by mutation then function is not lost.

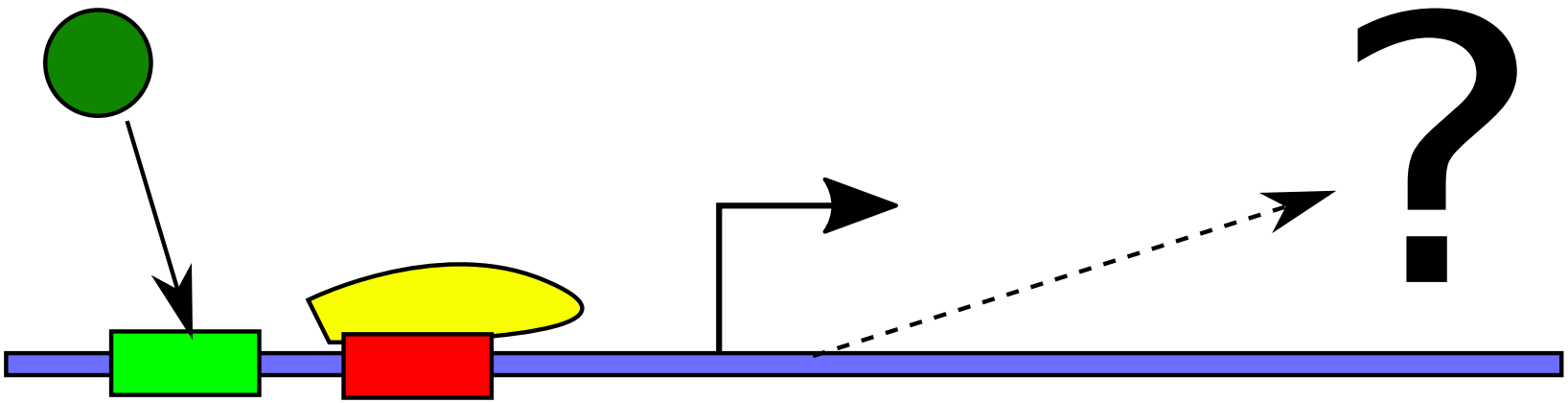
The challenge

- So, the **big challenge** is:
- Given the emerging view is that binding sites are **redundant** (*Spivakov, 2014*) and **cumulative binding** is the key to **robustness**.
  - I.e. if some site is lost by mutation then function is not lost.



The challenge

- So, the **big challenge** is:
- Given the emerging view is that binding sites are **redundant** (*Spivakov, 2014*) and **cumulative binding** is the key to **robustness**.
  - I.e. if some site is lost by mutation then function is not lost.



- **Predicting** when a **TF binding** will cause a **biologically significant response** in the **neighbouring gene**.

"Guilt-by-association"

- In other areas of genomics **function** is often **predicted** using the **"guilt-by-association" principle**.

“Guilt-by-association”

- In other areas of genomics **function** is often **predicted** using the **“guilt-by-association” principle**.

- E.g. genes with **similar expression profiles**, or interaction profiles, **often share the same function**.

"Guilt-by-association"

- In other areas of genomics **function** is often **predicted** using the **"guilt-by-association" principle**.
- E.g. genes with **similar expression profiles**, or interaction profiles, **often share the same function**.
- **Can we apply the same principle to our ChIP-sequencing data?**

"Guilt-by-association"

- In other areas of genomics **function** is often **predicted** using the **"guilt-by-association" principle**.
- E.g. genes with **similar expression profiles**, or interaction profiles, **often share the same function**.
- **Can we apply the same principle to our ChIP-sequencing data?**
- Can the relationship between binding and expression predict function?

The data

- Eight factors (not all strictly TFs, but all involved in regulation):
  - CEBPB, EP300, EZH2, MYC, RAD21, REST, TAF1, YY1.

The data

- Eight factors (not all strictly TFs, but all involved in regulation):
  - CEBPB, EP300, EZH2, MYC, RAD21, REST, TAF1, YY1.
- ENCODE data for 10 or more cell-lines:

The data

- Eight factors (not all strictly TFs, but all involved in regulation):
  - CEBPB, EP300, EZH2, MYC, RAD21, REST, TAF1, YY1.
- ENCODE data for 10 or more cell-lines:
  - ChIP-sequencing data for several peak-to-gene models:
    - 1/5/10/50kb around TSS and 1/5kb around TSS and in gene body.
    - Our binding score is the sum of peaks in this region.

The data

- Eight factors (not all strictly TFs, but all involved in regulation):
  - CEBPB, EP300, EZH2, MYC, RAD21, REST, TAF1, YY1.
- ENCODE data for 10 or more cell-lines:
  - ChIP-sequencing data for several peak-to-gene models:
    - 1/5/10/50kb around TSS and 1/5kb around TSS and in gene body.
    - Our binding score is the sum of peaks in this region.
  - RNA-sequencing data—normalised to compare expression levels for each cell-line.

The data

- Eight factors (not all strictly TFs, but all involved in regulation):
  - CEBPB, EP300, EZH2, MYC, RAD21, REST, TAF1, YY1.
- ENCODE data for 10 or more cell-lines:
  - ChIP-sequencing data for several peak-to-gene models:
    - 1/5/10/50kb around TSS and 1/5kb around TSS and in gene body.
    - Our binding score is the sum of peaks in this region.
  - RNA-sequencing data—normalised to compare expression levels for each cell-line.
- Differential expression in a knock-out cell-type for five factors—
  - proxy gold-standard for functional DNA-binding,
  - for one of the ENCODE cell-lines.

Data matrices

For each factor and each peak-to-gene model:

	Cell-line 1 $\cdots$ Cell-line n
Gene 1 $\vdots$ Gene m	ChIP Peak counts

	Cell-line 1 $\cdots$ Cell-line n
Gene 1 $\vdots$ Gene m	Expression levels

	Knock-out
Gene 1 $\vdots$ Gene m	Diff. Expr.

Data matrices

For each factor and each peak-to-gene model:

	Cell-line 1 $\cdots$ Cell-line n
Gene 1 $\vdots$ Gene m	ChIP Peak counts

	Knock-out
Gene 1 $\vdots$ Gene m	Diff. Expr.

	Cell-line 1 $\cdots$ Cell-line n
Gene 1 $\vdots$ Gene m	Expression levels

- Does correlation between ChIP and RNA predict KO expression?

The short answer

- Yes.

The short answer

- Yes.
- Weakly overall, but an improvement on the previous approaches.

The short answer

- Yes.
- Weakly overall, but an improvement on the previous approaches.
- High confidence for the top few hundred of targets.

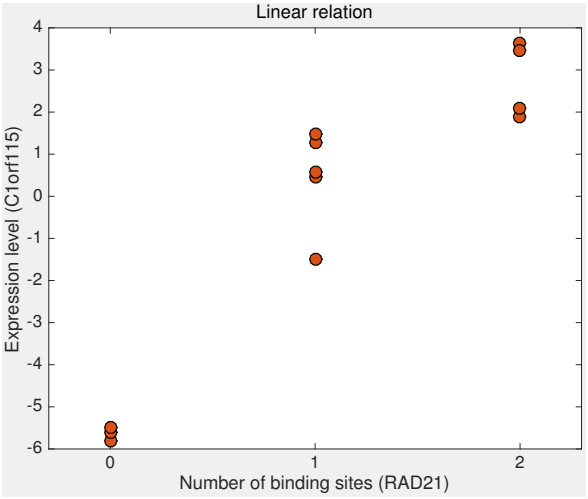
The short answer

- Yes.
- Weakly overall, but an improvement on the previous approaches.
- High confidence for the top few hundred of targets.
- And we did find a way to improve this further...

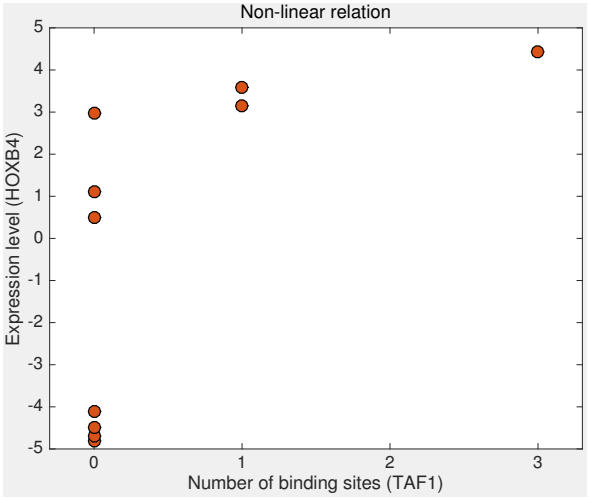
Correlation methods

- We found a number of different patterns for correlation:

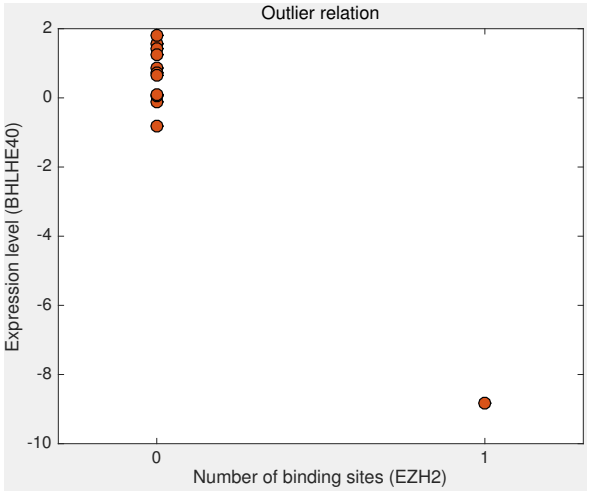
a



b



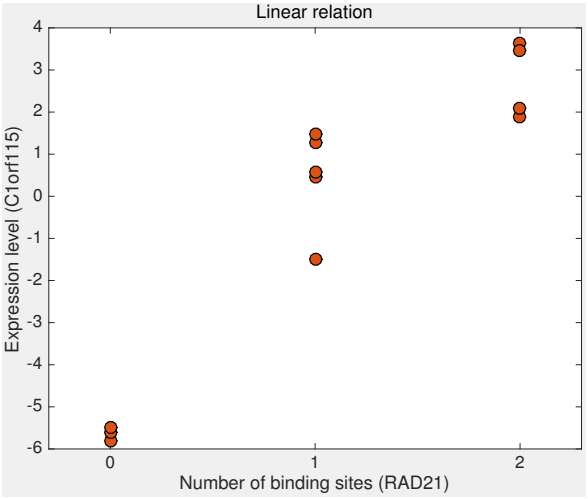
c



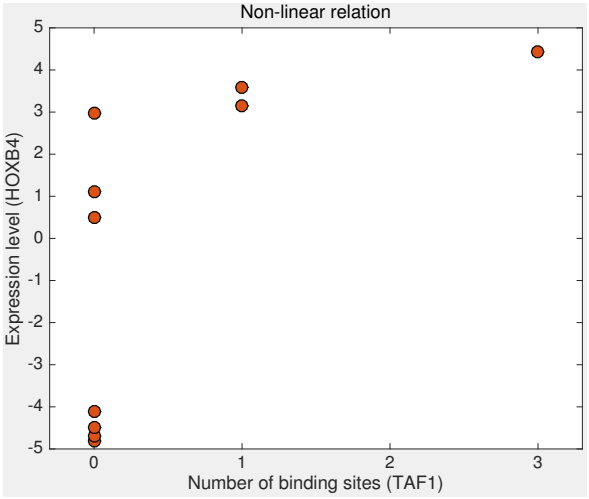
Correlation methods

- We found a number of different patterns for correlation:

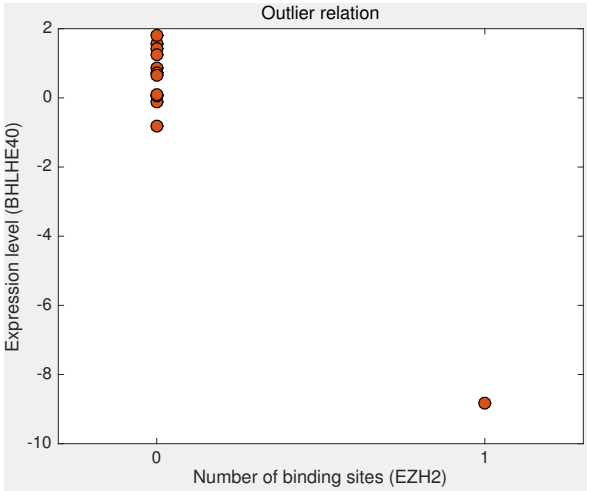
a



b



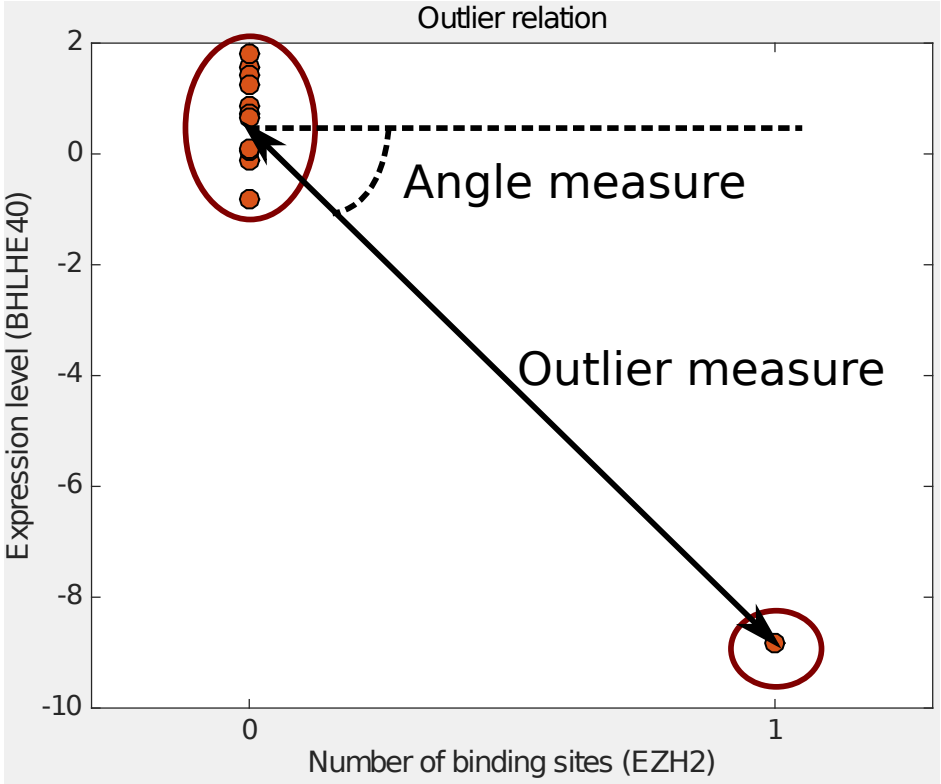
c



- Best accounted for by:
  - (a) Pearson,
  - (b) Spearman,
  - (c) Combined Angle Ratio Statistic.
    - Based on ARS (*Marstrand and Storey, 2014*), but with negative associations.

CARS

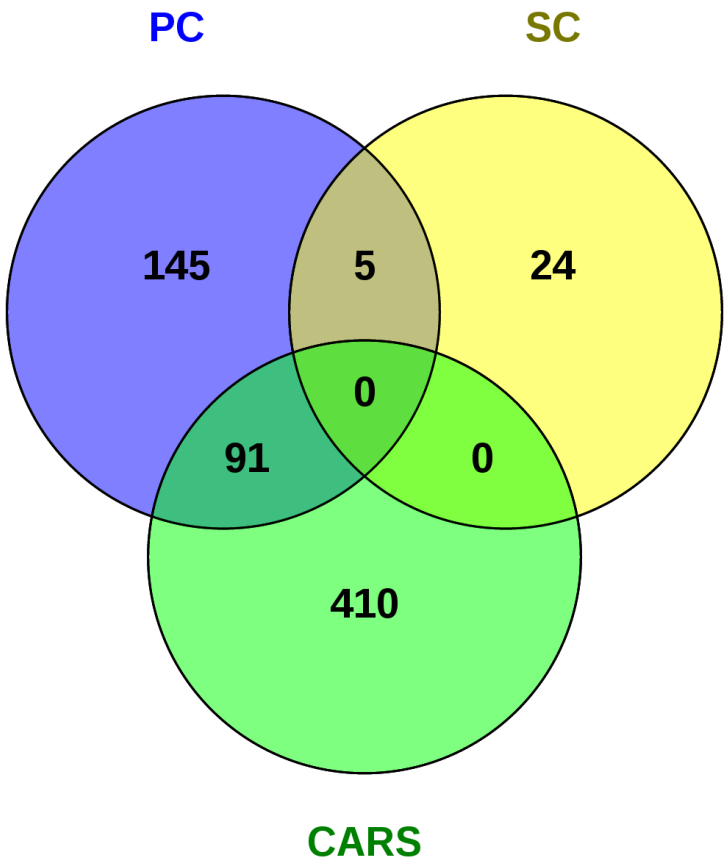
- Combined Angle Ratio Statistic (CARS)



- For each point:
- $\text{Angle}_i \times \text{Outlier}_i$
- CARS = max score for any point.

Wisdom of crowds

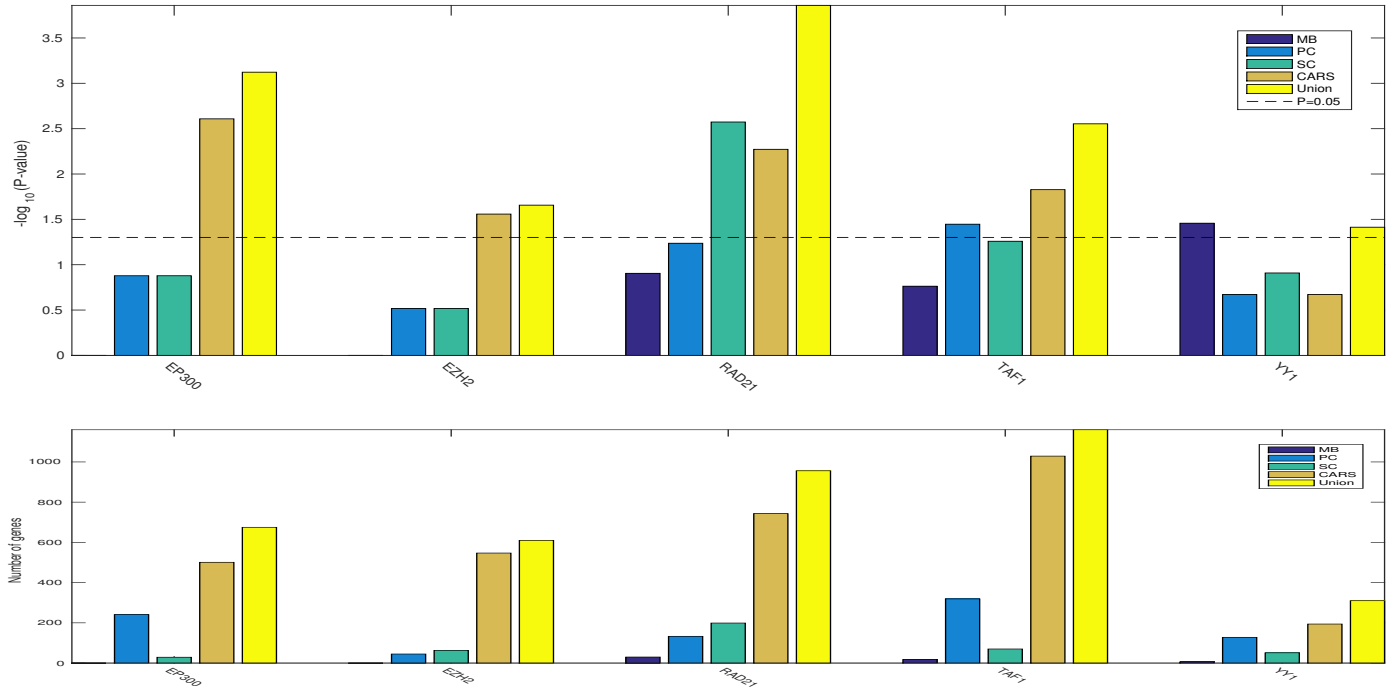
EP300:



- Different correlation measures find different relationships, so why not combine their strengths?

Wisdom of crowds

- Different correlation measures find different relationships, so why not combine their strengths?
- Taking the union of the top results from each measure:



- Significance is the hypergeometric overlap with the gold standard.

Other findings

- Correlation across time points of a defined biological process is also predictive of functional effects.

Other findings

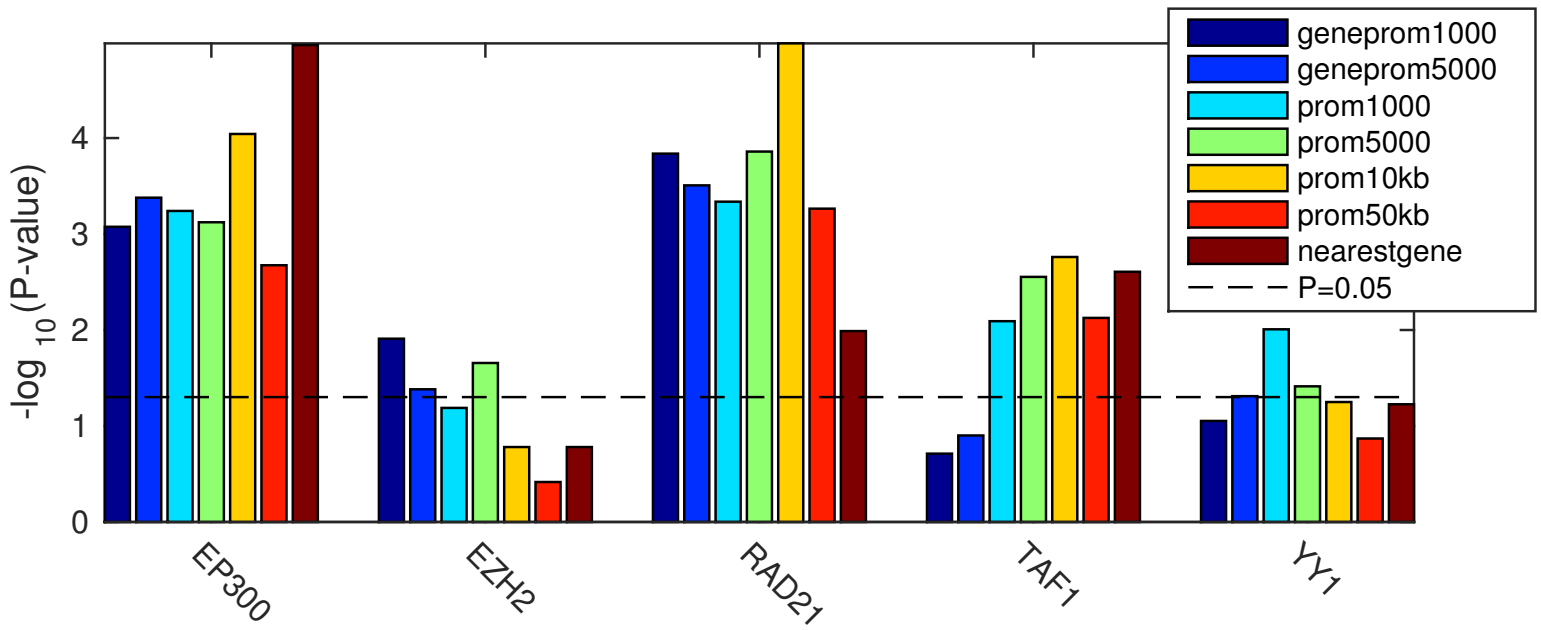
- Correlation across time points of a defined biological process is also predictive of functional effects.
    - Significant results using some mouse liver time-series data for circadian clock related factors.

Other findings

- Correlation across time points of a defined biological process is also predictive of functional effects.
    - Significant results using some mouse liver time-series data for circadian clock related factors.
- Different binding target models work better for different factors:

Other findings

- Correlation across time points of a defined biological process is also predictive of functional effects.
  - Significant results using some mouse liver time-series data for circadian clock related factors.
- Different binding target models work better for different factors:



- EZH2 binds closer to the TSS; EP300 binds further away.

Summary

- Correlation of binding and expression data for multiple cell-lines improves prediction of functional TF binding.

Summary

- Correlation of binding and expression data for multiple cell-lines improves prediction of functional TF binding.
- Plenty of data is available from the ENCODE resource—
  - though not many cell-lines for a lot factors.

Summary

- Correlation of binding and expression data for multiple cell-lines improves prediction of functional TF binding.
- Plenty of data is available from the ENCODE resource—
  - though not many cell-lines for a lot factors.
- A wisdom-of-crowds approach to correlation measures gives better precision.

Summary

- Correlation of binding and expression data for multiple cell-lines improves prediction of functional TF binding.
- Plenty of data is available from the ENCODE resource—
  - though not many cell-lines for a lot factors.
- A wisdom-of-crowds approach to correlation measures gives better precision.
- The method works with time-series data, as well as multiple cell-lines.

Future work

- The on-off model (CARS) seems to be the best predictor:

Future work

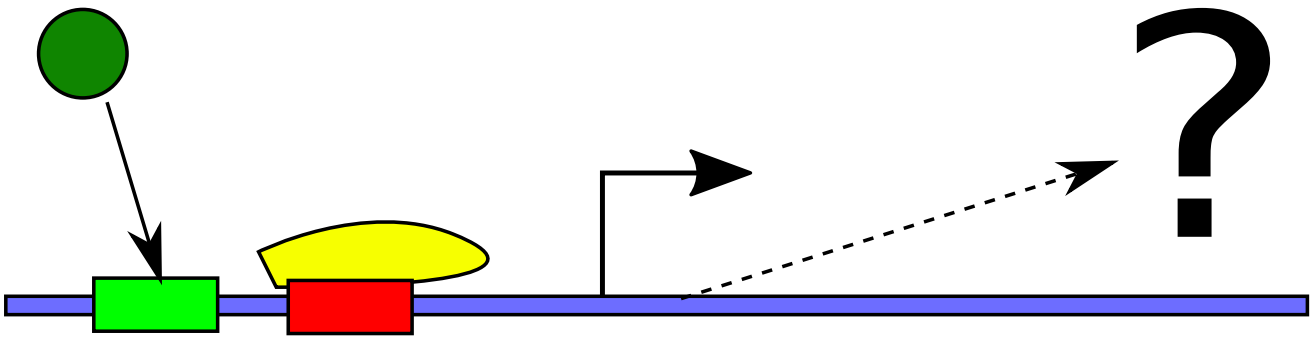
- The on-off model (CARS) seems to be the best predictor:
  - Does this mean we can use fewer cell-lines?

Future work

- The on-off model (CARS) seems to be the best predictor:
  - Does this mean we can use fewer cell-lines?
  - If so, there is data available for many more factors.

Future work

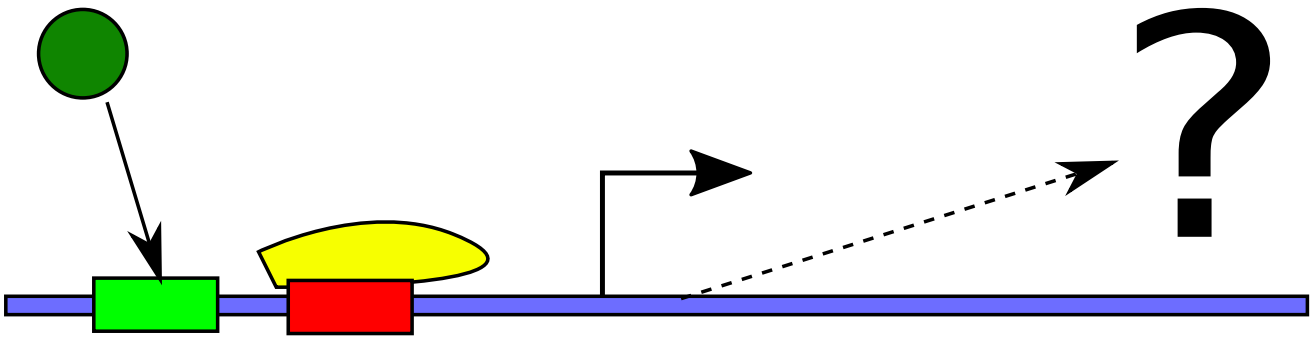
- The on-off model (CARS) seems to be the best predictor:
  - Does this mean we can use fewer cell-lines?
  - If so, there is data available for many more factors.



- What makes our predictions functional?

Future work

- The on-off model (CARS) seems to be the best predictor:
  - Does this mean we can use fewer cell-lines?
  - If so, there is data available for many more factors.



- What makes our predictions functional?
  - More investigation required on the idea of cumulative/combinatorial binding.