Prac9: Background

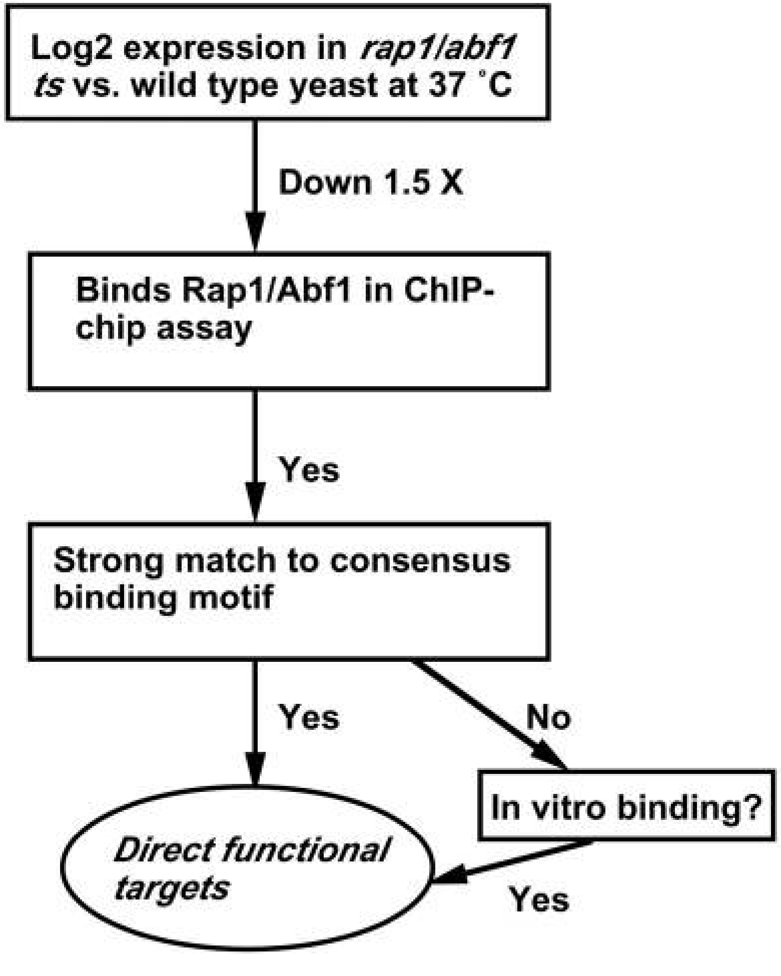
Abf1 is general regulatory factor (RFs) that contribute to transcriptional activation of a large number of genes, as well as to replication, silencing and telomere structure in yeast

In spite of their widespread roles in transcription, the scope of their functional targets genome-wide has not been previously determined

Yarragudi et al use microarrays to examine the contribution of these essential RFs to transcription genome-wide, by using mutants that dissociate from their binding sites at 37C

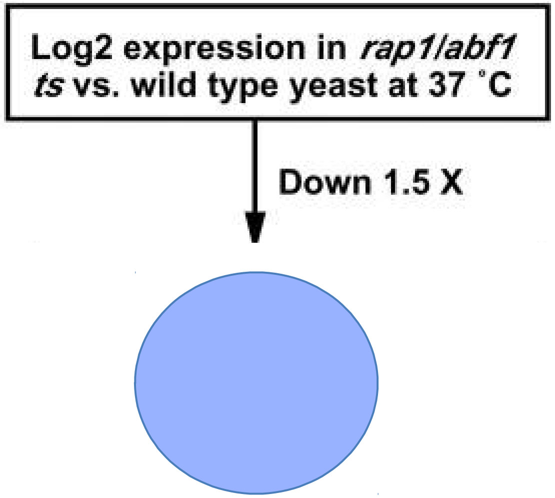
Yarragudi et al. Genome-wide analysis of transcriptional dependence and probable target sites for Abf1 and Rap1 in Saccharomyces cerevisiae. Nucleic Acids Res. 35(1) 2007.

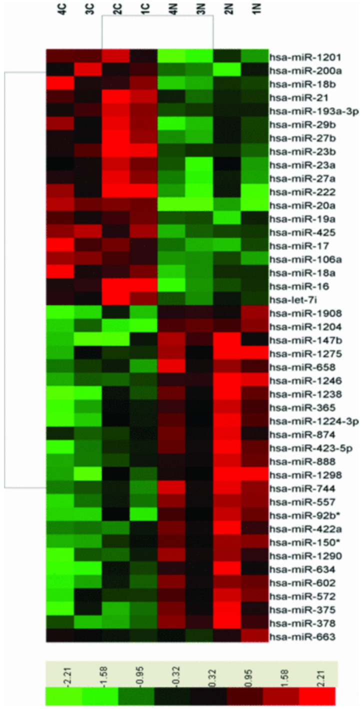
Bioinformatics: Identify targets


Log2 expression in *rap1/abf1 ts* vs. wild type yeast at 37 ˚C
Down 1.5 X
Binds Rap1/Abf1 in ChIP-chip assay
Yes
Strong match to consensus binding motif
Yes
No
In vitro binding?
Direct functional targets
Yes

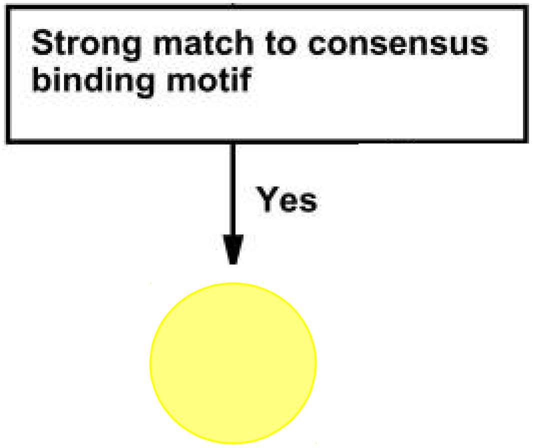
From: Yarragudi et al. Genome-wide analysis of transcriptional dependence and probable target sites for Abf1 in Saccharomyces cerevisiae. *Nucleic Acids Res*. 35(1) 2007.

Differential expression: Identify targets


Log2 expression in *rap1/abf1 ts* vs. wild type yeast at 37 ˚C
Down 1.5 X

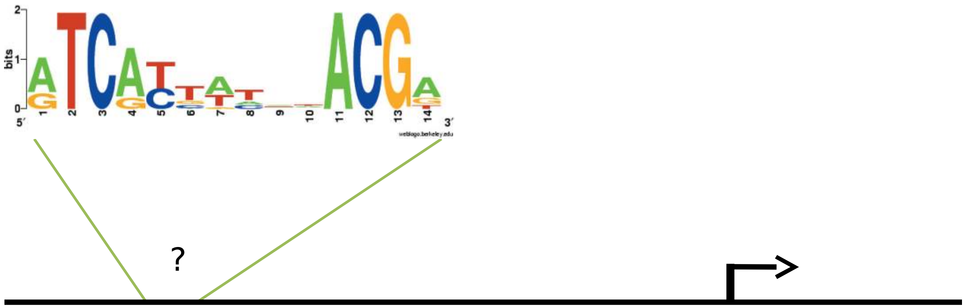


DNA binding: Identify targets



Strong match to consensus binding motif

Yes



```
>YAL067W
AGAGTACTGTTTTATGGCGCTTATGTGTATTCGTATGCGCAGA
ATGTGGG
>YPL242C
AAAACTTATTGCACCAGTTCAATTATATGTAACAAGGTGGTGC
AAAAACA
>YPR018W
TATGTTTTAGTGAACCTCAAGACAGAAGAGAATCGAAAGGA
AAAGGGAAA
>YAL065C
ATCCAACATGGAGGCCACAGACTACGAATGAAGAGTCTGTC
AGCTCTAAA
>YAL064W-B
TTGGATAGACCGTAACAACATCATTCACAGTAGCCGTGGCCG
```

Fishers Exact Test

- Quantify statistical significance of an association between two properties
- Used gene set enrichment

	Has Property	Does not have property	Row total
In gene set of interest	a	b	a+b
Not in gene set	c	d	c+d
Column total	a+c	b+d	a+b+c+d

Exercise 1

- α or significance level – a probability which is fixed in advance of making the hypothesis test.
- If the observed p-value is smaller than the significance level then the null hypothesis is rejected.
- Null hypothesis

- "Drug x is not indicative of chaperone regulator activity”

Exercise 1 Code

```
import stats
# 2 genes are annotated as negatives
c=2
#14 genes in our negative set
d= 14 -2
# Positive set of genes
Positives = set({"YPL106C", "YOL081W",
"YOR027W", "YOR299W", "YNL006W", "YNL007C",
"YLL039C", "YLR216C"})
# Genes annotated with GO Term
 has_property=set({"YER048C", "YIL016W",
"YLR090W", "YOR027W", "YMR161W", "YNL064C",
"YNL281W", "YDR214W", "YPL106C", "YNL007C",
"YNL227C"})
# We need to overlap Positives and has_property
a= # Fill me in here
```

```
#number of positives-a
b= Fill me in here
Print b
pval = stats.getFETpval(a,
b, c, d, left=False)

print pval
```

Exercise 1

- Provide the p-value and the significance level you are using.
- And a statement (reject or not reject null hypothesis)

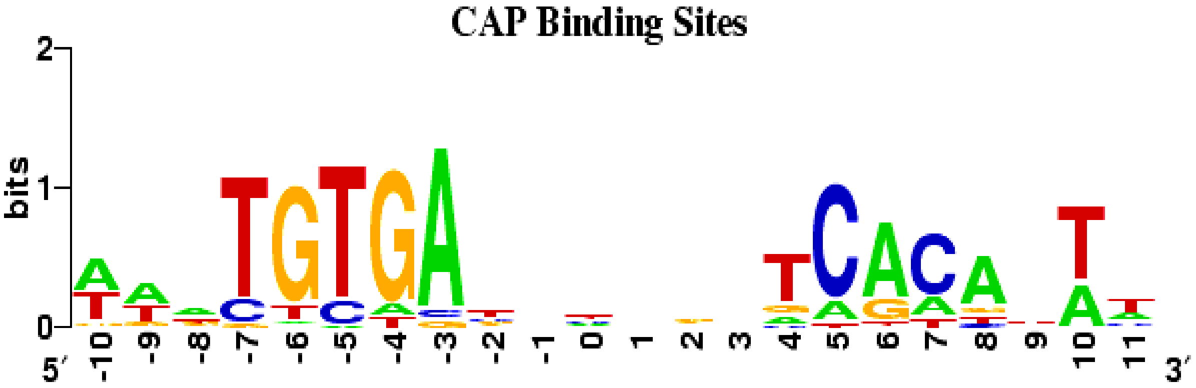
Exercise 2

- seqs=readFastaFile("yeast_promoters.fa”)
- print len(seqs)
- Hint: look at SCPD as a source
- 1-2 lines (How they are biologically sensible)

Exercise 3

- Visualizing motifs using “logo”
- Shows sequence conservation
- Frequency of residue

Example



Exercise 3

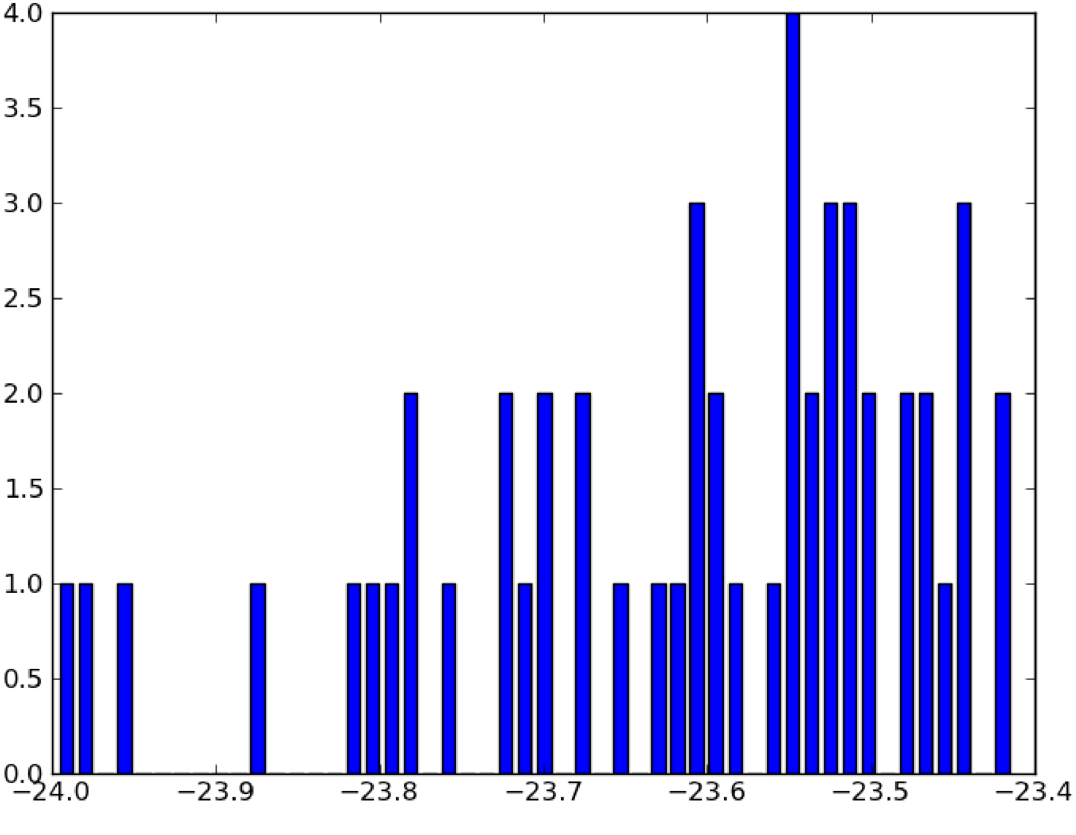
- Column 10
- Why do you only see G and A (few lines)
- Think about how PWMs are constructed
- Why total height is around 1bit (1bit is half of 2 bits)
- Think about what the height indicates.

Exercise 5 code

```
>>> bind_map = {}
>>> for s in yeast_prom: # yeast_prom is an array of sequences
    #Insert condition here !!!
        bind_map[s.name] = abf1_pwm.maxscore(s)[0] # save score
```

Exercise 5

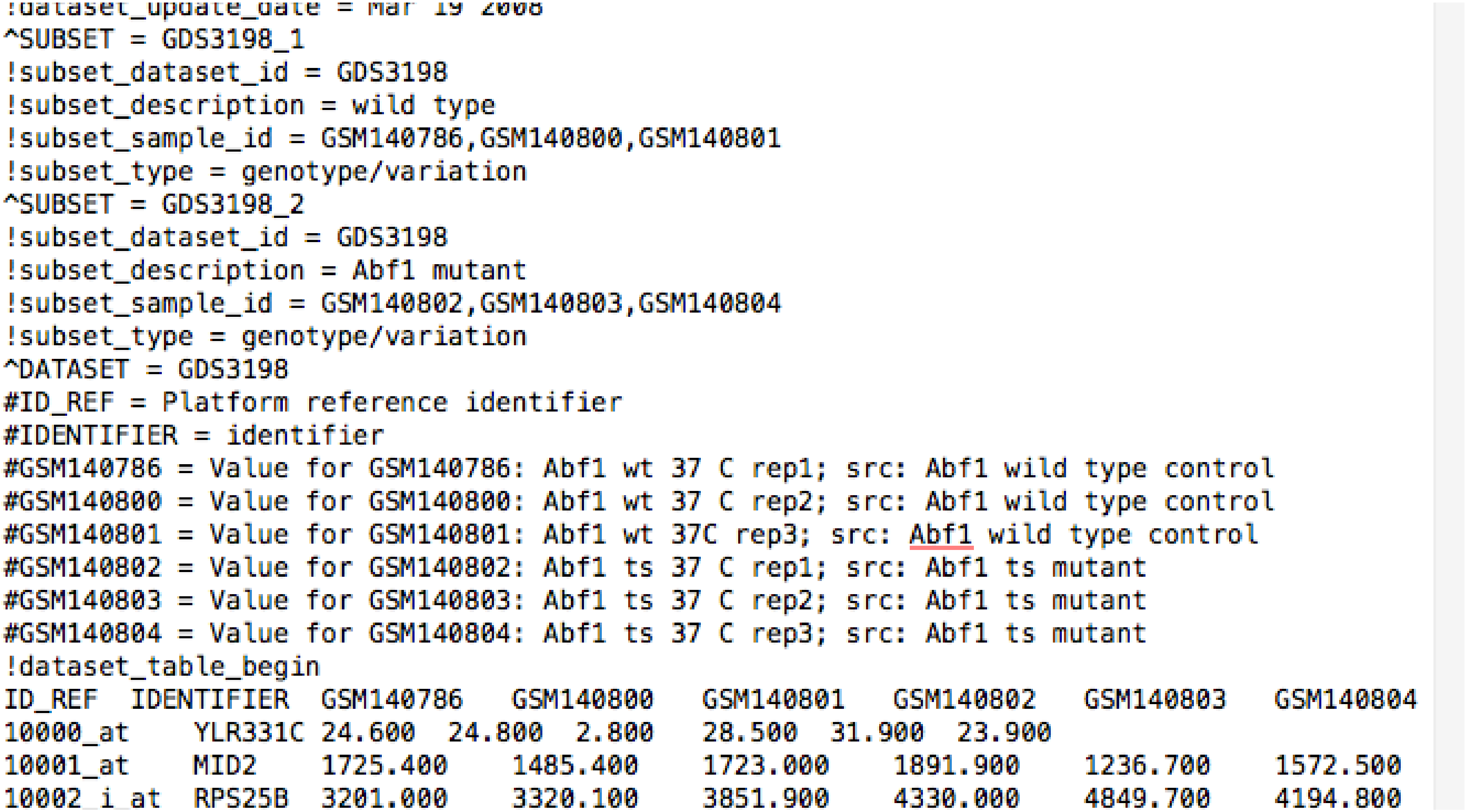
Try to aim for this graph



Exercise 5

- histogram
- Provide the list of 50 target genes
- Few lines (explaining your reason for the threshold)

Abf1 SOFT file



```
!dataset_update_date = Mar 19 2008
^SUBSET = GDS3198_1
!subset_dataset_id = GDS3198
!subset_description = wild type
!subset_sample_id = GSM140786,GSM140800,GSM140801
!subset_type = genotype/variation
^SUBSET = GDS3198_2
!subset_dataset_id = GDS3198
!subset_description = Abf1 mutant
!subset_sample_id = GSM140802,GSM140803,GSM140804
!subset_type = genotype/variation
^DATASET = GDS3198
#ID_REF = Platform reference identifier
#IDENTIFIER = identifier
#GSM140786 = Value for GSM140786: Abf1 wt 37 C rep1; src: Abf1 wild type control
#GSM140800 = Value for GSM140800: Abf1 wt 37 C rep2; src: Abf1 wild type control
#GSM140801 = Value for GSM140801: Abf1 wt 37C rep3; src: Abf1 wild type control
#GSM140802 = Value for GSM140802: Abf1 ts 37 C rep1; src: Abf1 ts mutant
#GSM140803 = Value for GSM140803: Abf1 ts 37 C rep2; src: Abf1 ts mutant
#GSM140804 = Value for GSM140804: Abf1 ts 37 C rep3; src: Abf1 ts mutant
!dataset_table_begin
ID_REF  IDENTIFIER  GSM140786   GSM140800   GSM140801   GSM140802   GSM140803   GSM140804
10000_at    YLR331C 24.600  24.800  2.800   28.500  31.900  23.900
10001_at    MID2    1725.400    1485.400    1723.000    1891.900    1236.700    1572.500
10002_i_at  RPS25B  3201.000    3320.100    3851.900    4330.000    4849.700    4194.800
```

Exercise 6

Provide probe and gene numbers...

```
g1 = ge.readGEOFile('GDS3198.soft',
id_column = 0)
```

```
g2 = ge.readGEOFile('GDS3198.soft',
id_column = 1)
```

Hint: getGenes() and len() may be useful.

Exercise 7

- Code
- Pairing (WT/mutants)
- Mention the transformations (ie. Log)
- How you filtered the top 100 and lowest 100
- Hint: indexing was useful.

Exercise 8

Code

```
result = sorted(meanfold.items(), key=lambda v: v[1])
print '========== Wildtype may down-regulate
=========='
for r in result[0:100]:
#Fill me in I am only one condition:
        print r[0]
print '========== Wildtype may up-regulate
=========='
for r in result[-1:-100:-1]:
    # fill me in I am only 1 condition
        print r[0]
```

Exercise 8 cont’d

Provide the gene list of 50 genes like so.

```
=========== Wildtype may down-regulate ===========
ATG29
YCLWOMEGA2
YLL067C
CDA1
YAL064W-B
YHR145C
YPR078C
RTG1
YOLCDELTA2
SPR3
YLR279W
…
```

Exercise 9

- Submit: A simple explanation (1-2 lines) why it is useful
- Hint: Consider multiple hypothesis testing
- (i.e. testing n terms)

Exercise 10 +11

- For Q10 Bind_map may be useful.
- For Q11 Store the gene symbols
- Provide
- Significant GO Terms

Exercise 12

- Helpful link:
- http://www.yeastgenome.org/cgi-bin/locus.fpl?locus=abf1
- Use get_GO_description

Prac9: Background

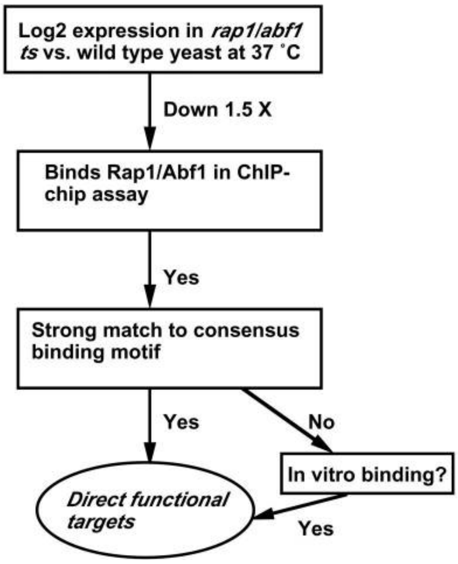
Abf1 is general regulatory factor (RFs) that contribute to transcriptional activation of a large number of genes, as well as to replication, silencing and telomere structure in yeast

In spite of their widespread roles in transcription, the scope of their functional targets genome-wide has not been previously determined

Yarragudi et al use microarrays to examine the contribution of these essential RFs to transcription genome-wide, by using mutants that dissociate from their binding sites at 37C

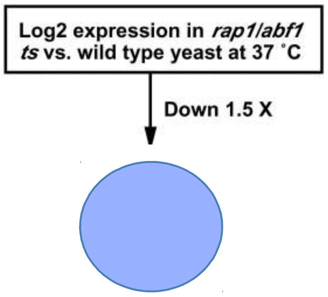
Yarragudi et al. Genome-wide analysis of transcriptional dependence and probable target sites for Abf1 and Rap1 in Saccharomyces cerevisiae. Nucleic Acids Res. 35(1) 2007.

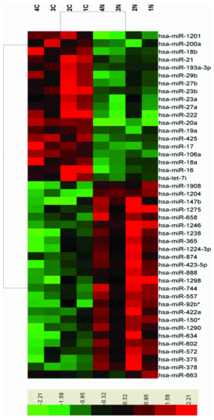
Bioinformatics: Identify targets


Log2 expression in *rap1/abf1 ts* vs. wild type yeast at 37 ˚C
Down 1.5 X
Binds Rap1/Abf1 in ChIP-chip assay
Yes
Strong match to consensus binding motif
Yes
No
In vitro binding?
Direct functional targets
Yes

From: Yarragudi et al. Genome-wide analysis of transcriptional dependence and probable target sites for Abf1 in Saccharomyces cerevisiae. *Nucleic Acids Res*. 35(1) 2007.

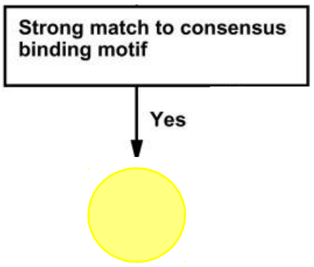
Differential expression: Identify targets

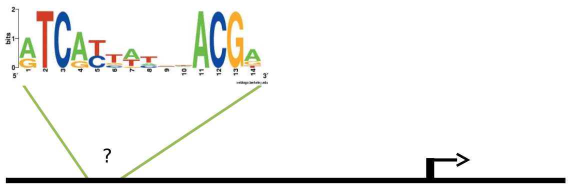

Log2 expression in *rap1/abf1 ts* vs. wild type yeast at 37 ˚C
Down 1.5 X



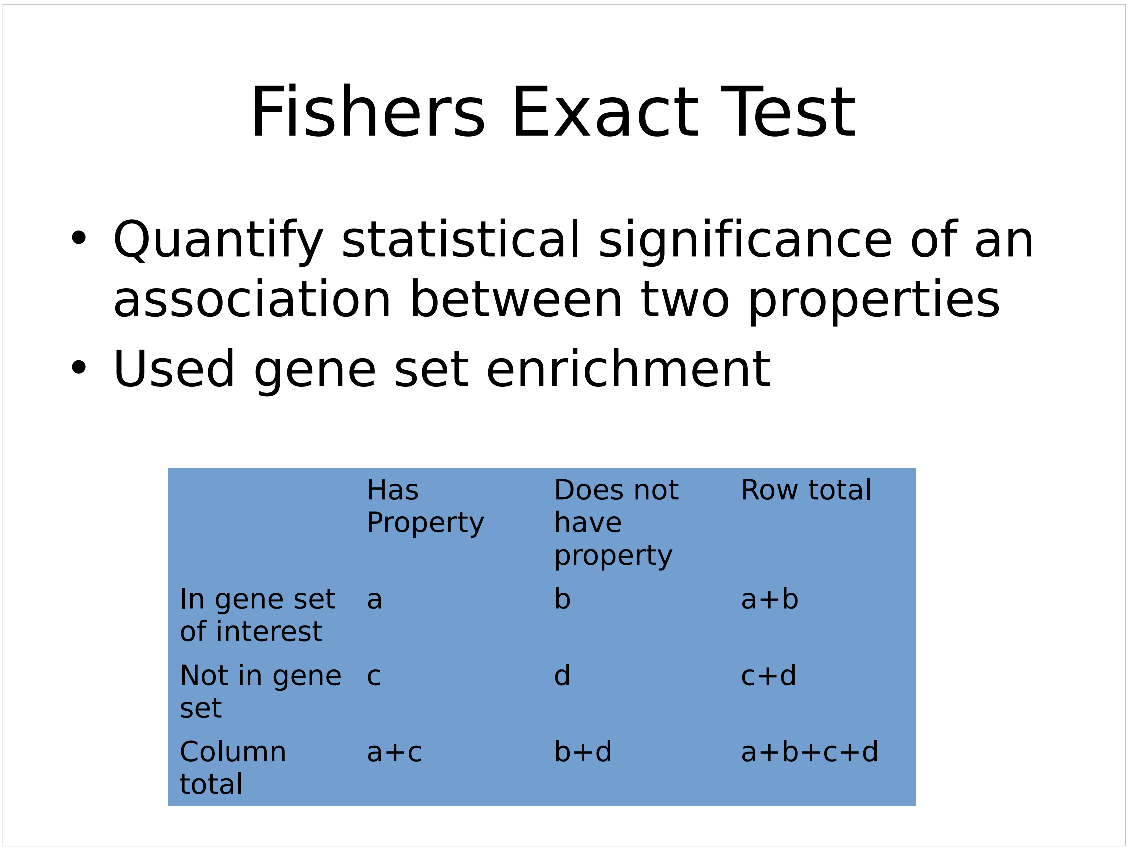
[illegible]

DNA binding: Identify targets





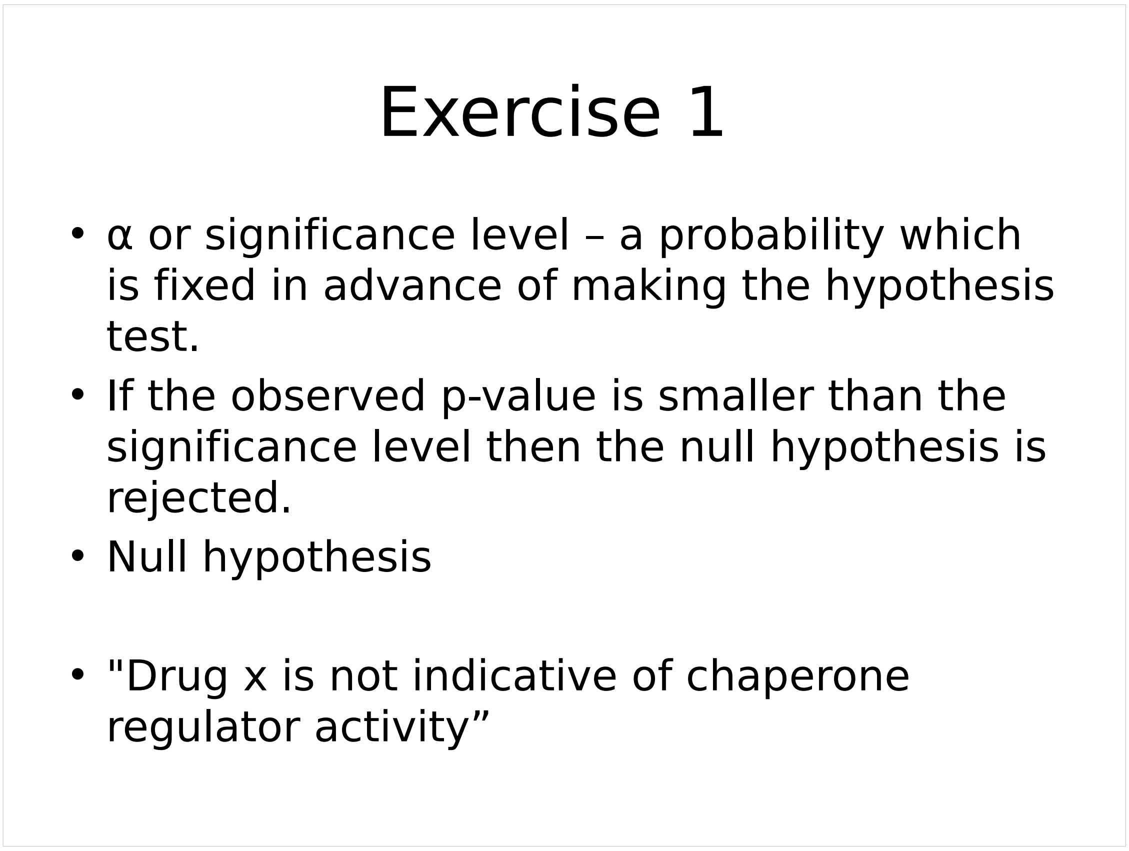
```
> YAL067W
AGAGTACTGTTTTATGGCGCTTATGTGTATTCGTATGCGCAGA
ATGTGGG
> YPL242C
AAAACTTATTGCACCAGTTCAATTATATGTAACAAGGTGGTGC
AAAAACA
> YPR018W
TATGTTTTAGTGAACCTCAAGACAGAAGAGAATCGAAAGGA
AAAGGGAAA
> YAL065C
ATCCAACATGGAGGCCACAGACTACGAATGAAGAGTCTGTC
AGCTCTAAA
> YAL064W -B
TTGGATAGACCGTAACAACATCATTCACAGTAGCCGTGGCCG
```



Fishers Exact Test

- Quantify statistical significance of an association between two properties
- Used gene set enrichment

	Has Property	Does not have property	Row total
In gene set of interest	a	b	a+b
Not in gene set	c	d	c+d
Column total	a+c	b+d	a+b+c+d



Exercise 1

- α or significance level – a probability which is fixed in advance of making the hypothesis test.
- If the observed p-value is smaller than the significance level then the null hypothesis is rejected.
- Null hypothesis

- "Drug x is not indicative of chaperone regulator activity”

Exercise 1 Code

```
import stats
# 2 genes are annotated as negatives
c=2
#14 genes in our negative set
d= 14 -2
# Positive set of genes
Positives = set({"YPL106C", "YOL081W",
"YOR027W", "YOR299W", "YNL006W", "YNL007C",
"YLL039C", "YLR216C"})
# Genes annotated with GO Term
 has_property=set({"YER048C", "YIL016W",
"YLR090W", "YOR027W", "YMR161W", "YNL064C",
"YNL281W", "YDR214W", "YPL106C", "YNL007C",
"YNL227C"})
# We need to overlap Positives and has_property
a= # Fill me in here
```

```
#number of positives-a
b= Fill me in here
Print b
pval = stats.getFETpval(a,
b, c, d, left=False)

print pval
```

Exercise 1

- Provide the p-value and the significance level you are using.
- And a statement (reject or not reject null hypothesis)

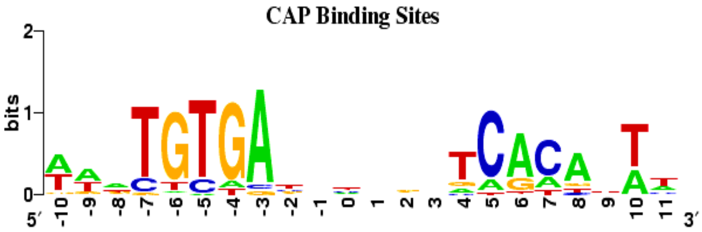
Exercise 2

- seqs=readFastaFile("yeast_promoters.fa”)
- print len(seqs)
- Hint: look at SCPD as a source
- 1-2 lines (How they are biologically sensible)

Exercise 3

- Visualizing motifs using “logo”
- Shows sequence conservation
- Frequency of residue

Example



Exercise 3

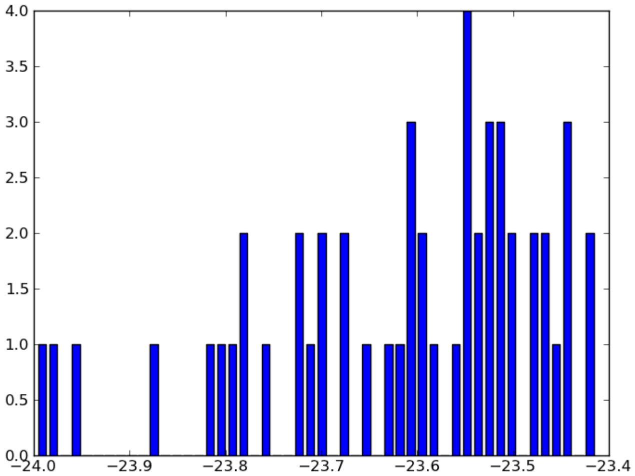
- Column 10
- Why do you only see G and A (few lines)
- Think about how PWMs are constructed
- Why total height is around 1bit (1bit is half of 2 bits)
- Think about what the height indicates.

Exercise 5 code

```
>>> bind_map = {}
>>> for s in yeast_prom: # yeast_prom is an array of
sequences
    #Insert condition here !!!
        bind_map[s.name] = abf1_pwm.maxscore(s)[0]
# save score
```

Exercise 5

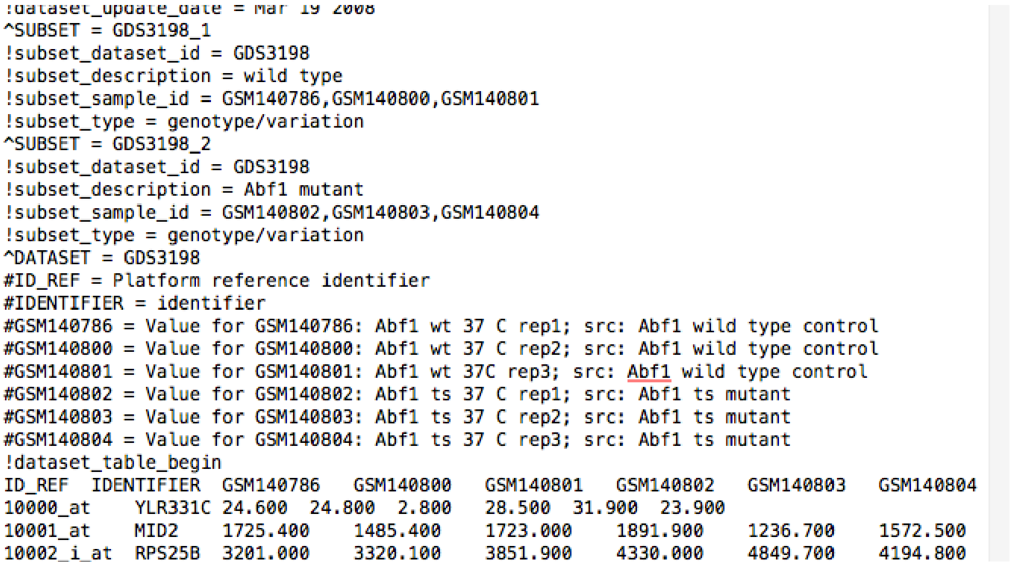
Try to aim for this graph



Exercise 5

- histogram
- Provide the list of 50 target genes
- Few lines (explaining your reason for the threshold)

Abf1 SOFT file



```
!dataset_update_date = Mar 19 2008
^SUBSET = GDS3198_1
!subset_dataset_id = GDS3198
!subset_description = wild type
!subset_sample_id = GSM140786,GSM140800,GSM140801
!subset_type = genotype/variation
^SUBSET = GDS3198_2
!subset_dataset_id = GDS3198
!subset_description = Abf1 mutant
!subset_sample_id = GSM140802,GSM140803,GSM140804
!subset_type = genotype/variation
^DATASET = GDS3198
#ID_REF = Platform reference identifier
#IDENTIFIER = identifier
#GSM140786 = Value for GSM140786: Abf1 wt 37 C rep1; src: Abf1 wild type control
#GSM140800 = Value for GSM140800: Abf1 wt 37 C rep2; src: Abf1 wild type control
#GSM140801 = Value for GSM140801: Abf1 wt 37C rep3; src: Abf1 wild type control
#GSM140802 = Value for GSM140802: Abf1 ts 37 C rep1; src: Abf1 ts mutant
#GSM140803 = Value for GSM140803: Abf1 ts 37 C rep2; src: Abf1 ts mutant
#GSM140804 = Value for GSM140804: Abf1 ts 37 C rep3; src: Abf1 ts mutant
!dataset_table_begin
ID_REF  IDENTIFIER  GSM140786   GSM140800   GSM140801   GSM140802   GSM140803   GSM140804
10000_at    YLR331C 24.600  24.800  2.800   28.500  31.900  23.900
10001_at    MID2    1725.400    1485.400    1723.000    1891.900    1236.700    1572.500
10002_i_at  RPS25B  3201.000    3320.100    3851.900    4330.000    4849.700    4194.800
```

Exercise 6

Provide probe and gene numbers...

```
g1 = ge.readGEOFile('GDS3198.soft',
id_column = 0)
```

```
g2 = ge.readGEOFile('GDS3198.soft',
id_column = 1)
```

Hint: getGenes() and len() may be useful.

Exercise 7

- Code
- Pairing (WT/mutants)
- Mention the transformations (ie. Log)
- How you filtered the top 100 and lowest 100
- Hint: indexing was useful.

Exercise 8

Code

```
result = sorted(meanfold.items(), key=lambda v: v[1])
print '========== Wildtype may down-regulate
=========='
for r in result[0:100]:
#Fill me in I am only one condition:
        print r[0]
print '========== Wildtype may up-regulate
=========='
for r in result[-1:-100:-1]:
    # fill me in I am only 1 condition
        print r[0]
```

Exercise 8 cont’d

Provide the gene list of 50 genes like so.

```
========== Wildtype may down-regulate ==========
ATG29
YCLWOMEGA2
YLL067C
CDA1
YAL064W-B
YHR145C
YPR078C
RTG1
YOLCDELTA2
SPR3
YLR279W
…
```

Exercise 9

- Submit: A simple explanation (1-2 lines) why it is useful
- Hint: Consider multiple hypothesis testing
- (i.e. testing n terms)

Exercise 10 +11

- For Q10 Bind_map may be useful.
- For Q11 Store the gene symbols
- Provide
- Significant GO Terms

Exercise 12

- Helpful link:
- http://www.yeastgenome.org/cgi-bin/locus.fpl?locus=abf1
- Use get_GO_description