Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays

Nature Genetics **36**, 1331 - 1339 (2004)

Sonali Mukherjee, Michael F Berger, Ghil Jona, Xun S Wang, Dale Muzzey, Michael Snyder, Richard A Young & Martha L Bulyk

Motivatsioon:
- Valk-DNA seondumised vajavad täpsemat modelleerimist kui allpool toodud joonisel

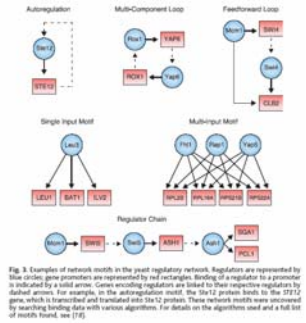


Fig. 3. Examples of network motifs in the yeast regulatory network. Regulators are represented by blue circles; gene promoters are represented by red rectangles. Binding of a regulator to a promoter is indicated by a solid arrow. Genes encoding regulators are linked to their respective regulators by dashed arrows. For example, in the autoregulation motif, the Ste12 protein binds to the *STE12* gene, which is transcribed and translated into Ste12 protein. These network motifs were uncovered by searching binding data with various algorithms. For details on the algorithms used and a full list of motifs found, see (7,8).

Katse skeem

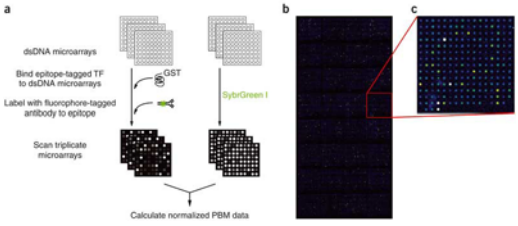
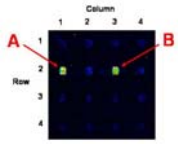


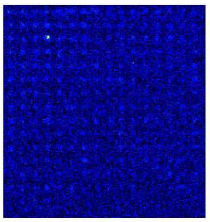
Figure 1. PBM schematic.
(**a**) Overview of PBM experiments.
(**b**) Whole-genome yeast intergenic microarray bound by Rap1. The fluorescence intensities of the spots are shown in false color, with white indicating saturated signal intensity, red indicating high signal intensity, green indicating moderate signal intensity and blue indicating low signal intensity.
(**c**) Magnification of a portion of the whole-genome yeast intergenic microarray bound by Rap1. ds, double-stranded; TF, transcription factor.

NB!
Kasutatakse kaheahelalist DNAd, mis on PCRiga võimendatud ja seejärel kiibile pandud.
Kaheahelalised DNAd vastavad ca 7000 pärmi geenidevahelisele piirkonnale (pikkus 100-2000 bp)
Kasutatakse puhastatud valku, millel küljes antikeha seostumiskoht

Positiivne ja negatiivne kontroll



a)



b)

a) Supplementary Figure 1
DNA microarray bound by CBP-FLAG-Rpn4, and labeled with Cy3-conjugated M2 anti(FLAG) antibody. Only the spots labeled A and B, which are the only spots in this portion of the microarray that contain good matches to the binding site motif for Rpn4, exhibit high signal intensity.

b) Supplementary Figure 2
PBM Negative Control Microarrays - scanned at very high laser power settings
rGST (Sigma)

PBM Data Quality Control Filters

Total # of spots on microarrays = 8192

(1) We considered only those spots with known sequence. We removed empty, blank, and control spots, and also spots for which the Research Genetics primers did not map well to the claimed intergenic region, thus possibly leading to poor PCR quality.
Total # of possible spots considered = 6814

(2) We employed a number of spot filters based on the SybrGreen I data. The following numbers of spots passed each of the following filters, applied in the following order:
1. Raw data spot quality (flags; intraspot SD/mean; 50% pixels > B+2SD): 6683
2. Variability (inter-array SD/mean <= 1): 6683
3. Length < 1500: 6654
4. Intensity (SybrGreenI intensity >= 10000): 6625
5. Density (SybrGreenI/length >= 61; this is 10% of the array average): 6494

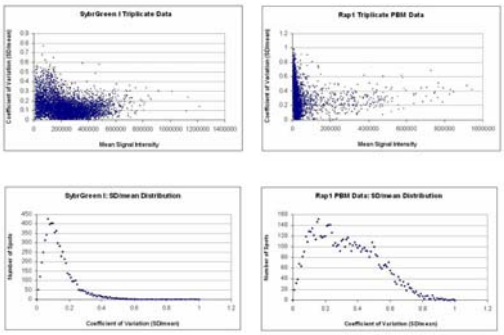
Total # of spots with acceptable SybrGreen I data = 6494

(3) We employed a number of spot filters based on the PBM data:
1. Raw data spot quality (flags; intra-spot SD/median > 2)
2. Variability (inter-array SD/mean > 1)
3. Must have SybrGreen data in order to calculate log ratio

(4) We removed duplicate spots. In cases where there were identical spots, we kept the one with PBM data and higher absolute SybrGreen I signal intensity.
Total # of possible non-duplicate spots = 6723

dataset	non-duplicate spots with SybrGreen I (and PBM) data	average SD/mean
SybrGreen I	6449 (95.9%)	0.130
Rap1	6431 (95.7%)	0.323
Abf1	6142 (91.4%)	0.282
Mig1	6442 (95.8%)	0.202

Punktide varieeruvus arvutatuna kolmest katsest



Tulemuste jaotusdiagrammil põhinevad saadud P-väärtused

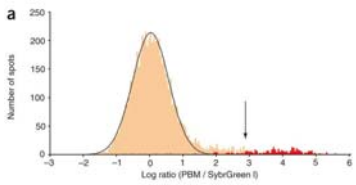
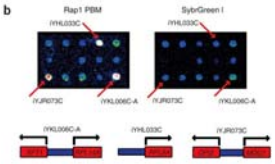


Figure 2. Identifying the specifically bound spots.
(**a**) Distribution of ratios of PDM data to SybrGreen I data for Rap1. The arrow indicates those spots passing a *P*-value threshold of 0.001 after correction for multiple hypothesis testing. Indicated in red are spots with an exact match to a sequence belonging to our discovered Rap1 binding-site motif.

NB! Enne mitmese testi korrektsiooni P-väärtus ca 10^{-7} (0.0000001)

Näited leitud valk-DNA seostumiste kohta

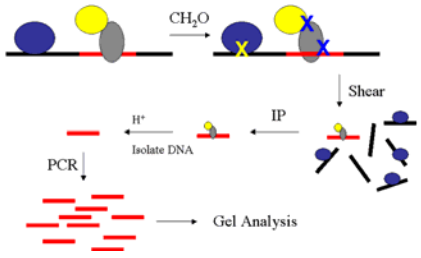


SGD ID	iYKL006C-A	iYHL033C	iYJR073C
PBM *P* value	4.6×10^{-12}	1.5×10^{-11}	1.2×10^{-9}
ChIP-chip*	Y	N	N
ChIP-chip#	Y	Y	N
Rap1 motif match	Y	Y	Y
Site conservation	2 s.d.	2 s.d.	Exact

Figure 2. Identifying the specifically bound spots.
(**b**) Magnification of intergenic regions, from both PBMs (left) and SybrGreen I−stained microarrays (right), upstream of *RPL14A*, *RPL8A* and *OPI3*, which are known to be direct targets of Rap1. The fluorescence intensities of the spots are shown in false color, color-coded as described for Figure 1. PBM *P* values are corrected for multiple hypotheses. Determination of binding in ChIP-chip experiments (Y, yes; N, no) is shown. All regions shown have an exact match to a sequence belonging to the discovered Rap1 motif. For each region, the binding site is conserved across five *sensu stricto* yeast strains, either to within two standard deviations or 100% identical at each position (Exact). The asterisk indicates Rap1 ChIP-chip data from Lee *et al.*[6]; the pound sign (#) indicates Rap1 ChIP-chip data from Lieb *et al.*[5]

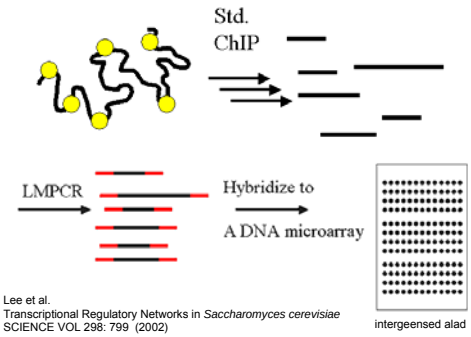
Chromatin IP (ChIP)

http://proteomics.swmed.edu/chiptochip.htm



NB!
ChIP puhul tuvastatakse valk-DNA seostumised *in vivo*

Chromatin IP on the chip (ChIP-chip)



Lee et al.
Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*
SCIENCE VOL 298: 799 (2002)

intergeensed alad

Millised seostumissaidid leiti?

Motiivide leidmisel kasutati programmi BioProspector

P-väärtus näitab sellise konserveerumisastmega motiivi leidmise tõenäosust samast arvust juhuslikult valitud intergeensetest aladest.

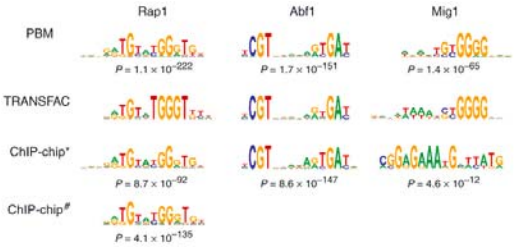
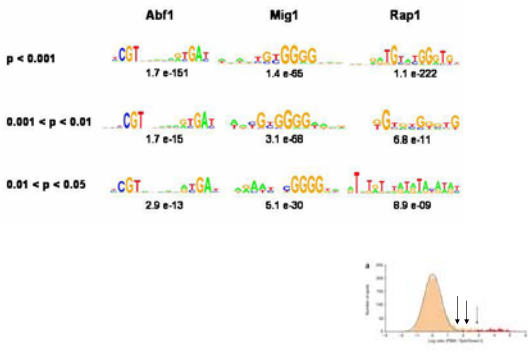


Figure 3. DNA binding site motifs as determined by PBMs compared with motifs derived from ChIP-chip data and from TRANSFAC.
Sequence logos were generated essentially as described previously[49]. Group specificity scores are shown. The asterisk indicates Rap1, Abf1 and Mig1 ChIP-chip data from Lee *et al.*[6]; the pound sign indicates Rap1 ChIP-chip data from Lieb *et al.*[5]. Although the Mig1 binding-site motif derived from the ChIP-chip data has a statistically significant group specificity score, it is not a match to either the TRANSFAC or the PBM Mig1 motif.

Millised seostumissaidid leiti erinevatel usaldusväärsuste vahemikel?



Kas teadaolevast konsensusjärjestusest erinevad DNA motiivid tõesti seovad valku ?

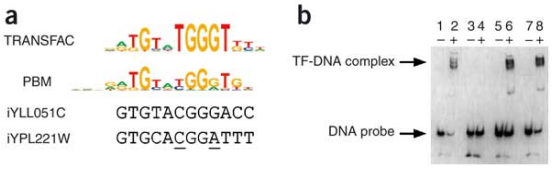


Figure 4. EMSAs of PBM-derived Rap1 binding-site sequences.
(**a**) Rap1 binding-site sequences present in the DNA probes corresponding to portions of the intergenic regions iYLL051C (P = 3.20*10^{-16}) and iYPL221W (P = 3.91*10^{-21}), aligned against the TRANSFAC and PBM-derived Rap1 binding-site sequence logos.

(**b**) Lanes 1 and 2, positive control DNA probe; lanes 3 and 4, negative control DNA probe; lanes 5 and 6, DNA probe corresponding to the best Rap1 binding-site sequence that could be identified in the iYLL051C intergenic region; lanes 7 and 8, DNA probe corresponding to the PBM-derived Rap1 binding-site sequence in the iYPL221W intergenic region. The presence (+) or absence (-) of Rap1 protein in the binding reaction is indicated. TF, transcription factor.

Kuidas kattuvad erinevate katsete tulemused ?

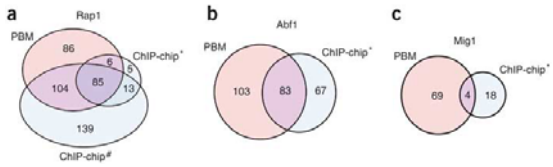


Figure 5. Comparison of bound intergenic regions derived from PBM data as compared with those derived from ChIP-chip[5, 6].
Venn diagrams depicting the results of the comparison for Rap1 (**a**), Abf1 (**b**) and Mig1 (**c**). The Venn diagrams depict data for only those intergenic regions for which data were available for both ChIP-chip and PBMs. The asterisk indicates Rap1, Abf1 and Mig1 ChIP-chip data from Lee *et al.*[6]; the pound sign (#) indicates Rap1 ChIP-chip data from Lieb *et al.*[5]

Kas leitud motiivid on konserveerunud erinevates pärmiliikides ? (Kas in vitro PBM meetod on parem kui ChIP)

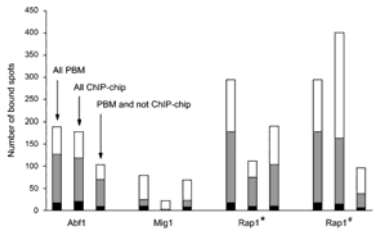


Figure 6. Cross-species sequence conservation of binding sites identified from PBM data as compared with those identified from ChIP-chip data.
From left to right for a single transcription factor, bars represent all spots bound in PBMs, all spots bound in ChIP-chip and spots bound in PBMs and not ChIP-chip. The subset of bound spots with *S. cerevisiae* binding sites conserved to within two standard deviations of the motif average across all five *sensu stricto* species is shown in dark gray. The subset of *S. cerevisiae* bound spots with conserved sites 100% identical across all five species is shown in black. The remaining bound spots are shown in light gray. The asterisk indicates Rap1, Abf1 and Mig1 ChIP-chip data from Lee *et al.*[6]; the pound sign (#) indicates Rap1 ChIP-chip data from Lieb *et al.*[5]

Probleemid:

Palju infot ühe katsega, katseid on lihtne korrata

Iga uuritava valgu peab enne puhastama (ja leidma sobivad antikehad)

Ei detekteerita kõiki seondumiskohti vaid ainult neid, mis on kiibil.
Kiibi disain kallis ja seda on raske laiendada kogu genoomile

Katse toimub in vitro, valk ei pruugi olla samasugune kui rakus, DNA ei pruugi olla samasugune kui rakus

Kuigi info on semi-kvantitatiivne, ei võimalda ta hinnata seondumise tugevust. Valk-DNA seondumisi ei saa kirjeldada SEONDUB / EI SEONDU

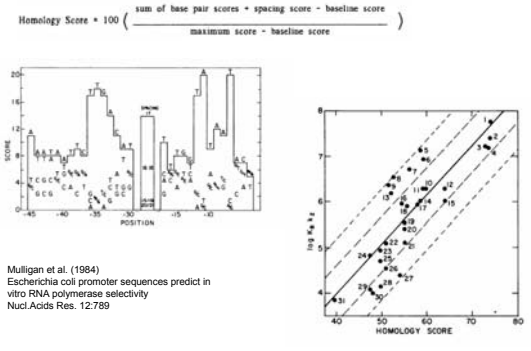
Kuidas mõõta seondumissaitide tasakaalukonstante K_{eq}?

Kas seda saaks ennustada seostumissaidi motiivi kirjeldava maatriksi abil?

Kuidas oleks sel juhul õige koostada esialgne maatriks?

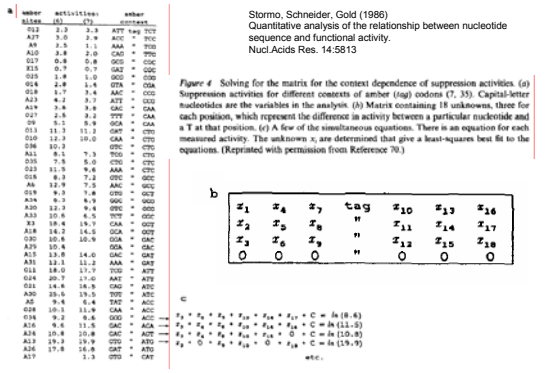
Kas siinkirjeldatud tehnoloogia võimaldaks koostada sellist maatriksit?

Geeni tegelik aktiivsus on heas korrelatsioonis promootori järjestuse motiiviga



Mulligan et al. (1984)
Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity
Nucl.Acids Res. 12:789

Geeni tegeliku aktiivsuse andmete abil saab leida aktiivsust kirjeldava maatriksi regressiooni abil.



Stormo, Schneider, Gold (1986)
Quantitative analysis of the relationship between nucleotide sequence and functional activity.
Nucl.Acids Res. 14:5813

Figure 4 Solving for the matrix for the context dependence of suppression activities. (*a*) Suppression activities for different contexts of amber (*tag*) codons (7, 35). Capital-letter nucleotides are the variables in the analysis. (*b*) Matrix containing 18 unknowns, three for each position, which represent the difference in activity between a particular nucleotide and a T at that position. (*c*) A few of the simultaneous equations. There is an equation for each measured activity. The unknown x_i are determined that give a least-squares best fit to the equations. (Reprinted with permission from Reference 70.)

b

x_1	x_4	x_7	tag	x_{10}	x_{13}	x_{16}
x_2	x_5	x_8	"	x_{11}	x_{14}	x_{17}
x_3	x_6	x_9	"	x_{12}	x_{15}	x_{18}
0	0	0	"	0	0	0

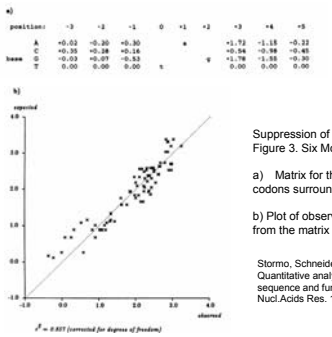
c

$x_2 + x_5 + x_8 + x_{10} + x_{14} + x_{17} + C = \ln(8.6)$
$x_2 + x_5 + x_8 + x_{10} + x_{14} + x_{16} + C = \ln(11.5)$
$x_2 + x_5 + x_8 + x_{11} + 0 + C = \ln(10.8)$
$x_2 + 0 + x_8 + x_{10} + 0 + x_{18} + C = \ln(19.9)$
etc.

Geeni tegeliku aktiivsuse andmete abil saab leida aktiivsust kirjeldava maatriksi regressiooni abil.

observed = ln(tegelik aktiivsus)

expected = regressioon-maatriksi abil ennustatud



Suppression of nonsense codons by a suppressor tRNA
Figure 3. Six Mono-Nucleotides vs. Suppression Activity.

a) Matrix for the ln(suppression activity) using the two codons surrounding the amber codon.

b) Plot of observed ln(activities) versus those predicted from the matrix of part a.

Stormo, Schneider, Gold (1986)
Quantitative analysis of the relationship between nucleotide sequence and functional activity.
Nucl.Acids Res. 14:5813

Kuidas oleks parem ennustada seostumise tugevust? Infosisaldusega maatriksid.

Fields et al. (1997) JMB 271: 178-194.
Quantitative Specificity of the Mnt Repressor

	11	12	13	14	15	16	17	18	19
	–	–	–	–	–	–	–	–	–
A	4	1	0	0	7	93	3	2	17
C	58	0	0	0	0	19	117	113	54
G	58	123	0	124	117	3	3	2	3
T	4	0	124	0	0	9	1	7	50

$$\vec{\mathbf{W}}(b,j) = \log_2 \frac{f(b,j)}{p(b)}$$

b 4 erinevat nukleotiidi
f(b,j) nukleotiidi b sagedus positsioonis j
p(b) nukleotiidi b sagedus genoomis

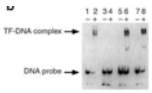
	11	12	13	14	15	16	17	18	19
	–	–	–	–	–	–	–	–	–
A	−2.64	−3.97	−4.97	−4.97	−1.97	1.59	−2.97	−3.38	−.80
C	.92	−4.97	−4.97	−4.97	−4.97	−.64	1.92	1.87	.82
G	.92	1.99	−4.97	2.00	1.92	−2.97	−2.97	−3.38	−2.97
T	−2.64	−4.97	2.00	−4.97	−4.97	−1.64	−3.97	−1.97	.71

Infosisaldusega maatriks sobib hästi seostumistugevuse hindamiseks

Stormo and Fields (1998) TIBS 23: 109-113.
Specificity, free energy and information content in protein–DNA interactions.

$$K_{eq} = \frac{[\mathbf{T} \bullet \mathrm{X}_i]}{[\mathbf{T}][X_i]}$$

$$\frac{K(X_1)}{K(X_2)} = \frac{[X_1 \bullet \mathbf{T}]}{[X_1][\mathbf{T}]} \frac{[\mathbf{T}][X_2]}{[X_2 \bullet \mathbf{T}]} = \frac{[X_1 \bullet \mathbf{T}]}{[X_2 \bullet \mathbf{T}]} \frac{[X_2]}{[X_1]}$$



$$\Delta G_s(b) = -RT \ln K_s(b).$$

K_s on saadud K_{eq} väärtuste normaliseerimisel nii et K_s=1 vastab mittespetsiifilisele seondumisele antud genoomis

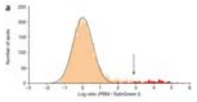
(a)	11	12	13	14	15	16	17	18	19
	–	–	–	–	–	–	–	–	–
A	0.26	0.37	0.91	0.31	0.058	**−0.29**	0.55	0.14	0.39
C	−0.16	0.87	0.58	1.90	0.74	−0.11	**−0.72**	**−0.33**	−0.33
G	**−0.16**	**−0.68**	0.36	**−0.68**	**−0.56**	0.45	1.30	0.30	0.45
T	0.26	1.20	**−0.63**	0.68	0.85	0.13	0.74	0.11	**−0.16**

deltaG negatiivne – tugev seondumine
deltaG 0 – juhuslik seondumine
deltaG positiivne – energeetiliselt ebasoodne, välditud seondumine

Kas iga maatriks sobib seostumistugevuse hindamiseks? Millised järjestuste alusel peaks selle maatriksi koostama?

Tüüpiliselt valitakse maatriksi koostamiseks "teadaolevalt seostuvad järjestuse motiivid". Seostumistugevuse erinevus võib aga olla kuni 10 000 kordsete erinevustega.

Kas on võimalik koostada sagedusmaatriks ChIP katse tulemustest nii, et tegelik seondumistugevus oleks arvesse võetud?



BioProspector
Liu X, Brutlag DL, Liu JS. (2001)
BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.
Pac Symp Biocomput. 2001:127-138

MDScan
Liu et al. (2002)
An algorithm for finding protein-DNA binding sites with applications to chromatin immunoprecipitation microarray experiments.
Nature Biotechnology 20: 835-839

Motif Regressor
Conlon et al. (2003)
Integrating regulatory motif discovery and genome-wide expression analysis.
PNAS 100: 3339-3344

(a)

	11	12	13	14	15	16	17	18	19
	–	–	–	–	–	–	–	–	–
A	0.26	0.37	0.91	0.31	0.058	**−0.29**	0.55	0.14	0.39
C	−0.16	0.87	0.58	1.90	0.74	−0.11	**−0.72**	**−0.33**	−0.33
G	**−0.16**	**−0.68**	0.36	**−0.68**	**−0.56**	0.45	1.30	0.30	0.45
T	0.26	1.20	**−0.63**	0.68	0.85	0.13	0.74	0.11	**−0.16**

acacttgtatgtagcgcatctttctttacggtcaatcagcatggtgttaaattg