Systematic determination of transcription factor DNA-binding specificities in yeast

Lourdes Peña-Castillo, Gwenaël Badis

Systematic determination of transcription factor DNA-binding specificities in yeast.

Lourdes Peña-Castillo[1,2] and Gwenael Badis[3,4]

[1] Department of Biology, Memorial University of Newfoundland, St. John's, NL A1B 3X5, Canada

[2] Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada

[3] Institut Pasteur, Génétique des Interactions Macromoléculaires, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 3525, F-75724 Paris, France.

[4] corresponding author: gbreard@pasteur.fr

Summary

Understanding how genes are regulated, decoding their "regulome", is one of the main challenges of the post-genomic era. Here, we describe the *in vitro* method we used to associate cis-regulatory sites with cognate trans-regulators by characterizing the DNA-binding specificity of the vast majority of yeast transcription factors using Protein Binding Microarrays. This approach can be implemented to any given organism.

Key Words

Transcription regulation, transcription factors, DNA binding domain, Cis-regulatory element, enhancers, binding sites

1. Introduction

Decoding transcription factor (TF)-DNA interaction is one of the crucial steps to understand how genes are regulated. Most known transcription factor binding sites (TFBSs) are short (6-10 bases pairs) and degenerated. In addition, a particular TF may bind multiple binding sites with different affinity. Several factors such as combinatorial action of TFs and chromatin structure regulate gene expression but the first step to understand transcriptional regulation is to characterize individual binding sites.

To address this question, a variety of techniques have arisen in the last two decades, however, few of them are suitable for large-scale studies.

In vivo Chip-derived methods (Chip-Chip ***(1, 2)*** Chip-seq ***(3)***,Chip Pet ***(4)*** require immunoprecipitation of the TF of interest, and have all been used to characterize numerous TFBSs in several organisms. Drawbacks of these methods are the requirement of specific antibodies, the restriction to TFs expressed and active in experimental conditions, and the likely detection of indirect interactions, which can scramble the motif definition.

In vitro Selex ***(5)*** is the oldest low scale method that identifies a set of bound sequences from a random collection of sequences; however, this method is biased by multiple steps of PCR. More modern and powerful versions of Selex have recently been described ***(6, 7)***.

Universal Protein Binding Microarray (PBM ***(8)***) is an alternative *in vitro* method by which most yeast TFBSs have been characterized ***(9–11)***. PBMs have also been used to characterize TFBSs in other organisms ***(12–14)***.In standard PBM experiments, a GST-fused TF is allowed to bind a double stranded microarray containing a representation of all possible 10mer cut in 35mer pieces (see below and in ***(8)*** for details). A second step consisting of an antibody labelling highlights spots where the TF is bound. This technique requires no PCR

amplification and is highly sensitive and robust. This method is limited by the number of sequences that can be represented on a microarray, which determines the highest complexity of the motifs represented on the array. Consequently, TFs with long binding sites (>10 base pairs) may be difficult to characterize using this approach.

In this chapter, we provide details of the procedure we used to determine transcription factor DNA-binding specificities for numerous yeast TFs ***(9)*** using PBM experiments.
We explain how we rendered this large-scale study feasible, and describe how we computationally processed and analyzed the data.

2. Materials

2.1 Production and purification of GST-Tagged proteins

1. C41 DE3 cells
2. LB^{amp}: 10g/l bacto-tryptone, 5g/l bacto-yeast extract, 10g/l NaCl, pH7.0, $[\text{Ampicilline}]_{\text{final}} = 100\ \mu g/mL$
3. LB^{amp} + glucose : LB^{amp} +2g/l glucose
4. IPTG : stock solution at 100 mM
5. PBS pH 7.3: 137 mM NaCl, 2.7 mM KCl, 10 mM Na2HPO4, 2 mM KH2PO4
6. Lysozyme: stock solution at 80 mg/ml
7. Lysis Buffer: 50 mM Tris (pH 8.0), 150 mM NaCl, 2 mM DTT(add fresh).
8. Glutathione sepharose 4B (17-0756-01 Amersham)
9. Wash buffer 1X PBS + 2 mM DTT (add fresh)

10. Elution buffer: 50 mM Tris pH7.5, reduced glutathione 10 mM, Complete tablet (Roche), 2 mM DTT (add fresh)
11. Zinc Acetate 1 M
12. ActivePro Kit (Ambion)

2.2 Protein Binding Microarray

1. Stilt RC primer (see Table 1) HPLC-purified (Integrated DNA Technologies)
2. dNTP
3. Cy3 dUTP (GE Healthcare)
4. Thermo Sequenase™ DNA Polymerase (USB)
5. Microarray, stainless steel hybridization chamber (Agilent)
6. Four-chambers gasket cover slip (Agilent)
7. 10x sequenase reaction buffer (260 mM Tris-HCl, pH 9.5, 65 mM MgCl2) in a total volume of 900 µl.
8. PBS**:** (phosphate buffered saline) NaCl 137 mM, KCl 2.7 mM, Na_2HPO_4 10 mM, KH_2PO_4 1.8 mM, pH 7.4
9. Wash buffer A: PBS + 0.01% (vol/vol) Triton X-100
10. Wash buffer B: PBS + 0.1% (vol/vol) Tween-20
11. Wash buffer C: PBS + 0.5% (vol/vol) Tween-20
12. Wash buffer D: PBS + 0.05% (vol/vol) Tween-20
13. Blocking :2% (wt/vol) nonfat dried milk dissolved in PBS for 2 hours (or overnight) and filtered using a 0,45 µm filter.
14. Alexa488-conjugated rabbit polyclonal antibody to GST (Invitrogen)
15. salmon testes DNA
16. bovine serum albumin (New England Biolabs).

17. ZnAc 500X: 25 mM Zn Acetate, ZnAc 100X = 5 mM Zn Acetate

18. Stripping solution: 10 mM EDTA, 10% SDS, + 210 units Protease (Sigma CN P6911-1G, 5.8 units/mg) per 50 ml.

3. Methods

3.1. Experimental Design

The first step of a large-scale characterization of TF – DNA binding affinities is to determine the list of genes to assay and to generate a collection of GST-tagged TFs or DNA-binding domains (DBDs). In our study, we determined that a region containing the DBD plus 15 flanking amino acids (aa) is sufficient and appropriated for most TFs, as shorter domains are easier to clone and give proteins that are simpler to express and produce. We observe no difference between PBMs obtained from full length or truncated TFs when we compared both; however, the majority of our trials with full length TFs failed to give a sufficient yield to properly run a PBM experiment. Note that the dimerization domain has to be added in the design for TFs expected to dimerize (such as those containing a Helix-Loop-Helix domain). In order to define the domains to be tested, we selected a list of 36 distinct DBDs containing all the known examples of yeast specific DNA binding transcription factors ***(9)***. In order to catalog all possible yeast transcription factors, we employed the software HMMER (version 2.3.2, available at http://hmmer.janelia.org/) ***(15)*** to generate profile hidden Markov models for all DBDs and scanned the yeast genome to detect those DBDs. We also scanned the SMART Database (http://smart.embl-heidelberg.de/) ***(16)*** to extend the search and selected a total of 212 independent ORFs containing one or more of the 36 selected domains ***(9)***. Recent reviews ***(17, 18)*** estimates the number of known and putative yeast TF to 209.

For flexibility and cost, we created a Donor clone library compatible with the MAGIC system ***(19)*** using a ligation independent cloning strategy ***(20)***. Donor clones can be easily transferred by bacterial conjugation into a Gluthatione-S-transferase (GST) N-terminal tag Recipient vector such as pTH1137, a T7-GST-tagged variant of pML280 ***(19)***. Alternatively, a GATEWAY system ***(21)*** or any way to generate GST-fusion protein can be used.

3.2. Microarray design

Random universal PBM array is a 4X44K customed microarray (Agilent) containing all possible 10-mer within 35-nucleotide probes generated by a De Bruijn sequence of order 10.

The design of this array is described in ***(8)***. The microarray designs we used in our study are variations of the original microarray. Details of the modifications can be found at http://hugheslab.ccbr.utoronto.ca/supplementary-data/yeastDBD/

For each TF, two versions of these arrays (A and B, corresponding to the same complexity) are used to perform replicate PBM experiments with two independently produced GST-tagged proteins. This allows testing the robustness of PBM reproducibility.

3.3. Expression and purification of GST tagged DBDs from *E. coli*

3.3.1. *E. coli* Cultures and induction

1- C41 DE3 cells are transformed with a plasmid expressing the GST-fusion gene of interest under the control of a PTAC promoter using standard procedure. 200 ml of LB ^{amp} + glucose (+ Zn acetate if necessary, see Note 1) are inoculated with 2 ml of an overnight LB ^{amp} grown preculture and grown at 25°C until OD600 is 0.5 to 0.8. 2ml of this “uninduced” culture is set-aside in a “negative control” tube.

2- IPTG is added to the main culture to a final concentration of 1 mM.

3- Both cultures are grown at 14°C overnight shaking. 2 ml of both cultures are saved for further control (see Note 2).

4- Cultures are centrifuged at 4°C 15 min at 3200 g. Pellet are resuspended in 30 ml ice-cold Wash buffer, transferred to a 50 ml Falcon tube and centrifuged at 4°C 15 min at 3200 g.

5- Pellets are decanted and flash frozen at -80°C if needed, or can be directly continued to the lysis step described below.

3.3.2. Lysis

1. Pellets are resuspended in 25 ml lysis buffer.
2. From a stock concentration at 80 mg/ml, 160 µl of lysozyme is added so that 12.8 mg of lysozyme is used for a pellet obtained from a 200 ml culture, and incubated in ice 20 min.
3. Cells are lysed by sonication (see Note 3). Lysates are centrifuged at 4°C 15 min at 3200 g. Cleared lysate are transferred to 50 ml Falcon tubes in ice and NaCl is added to obtain 250 mM final (see Note 4).

3.3.3 Purification

1. Two hundred microliters of gluthatione sepharose beads are equilibrated in 5 ml PBS, rotating at 4°C for 5 min and centrifuged at 4°C 5 min at 100 g. Supernatants are removed.
2. About 25 ml of lysate are incubated with equilibrated glutathione beads, one hour rotating at 4°C, centrifuged at 100 g and supernatants are carefully removed.
3. Beads are washed twice with 10 ml PBS wash buffer, 10 min on the rotating wheel at 4°C, spun down at 100 g and cleared from supernatant.

4. Beads are transferred into an Eppendorf tube, spun down at 100 g at 4°C and cleared from supernatant. GST-tagged proteins are eluted with 200 µl elution buffer, 30 min to 1 hour at 4°C rotating.
5. Eluates are collected in a new tube after centrifugation at 4°C 1 min at 100 g. Glycerol is added to each sample to 30% final. GST-proteins are stoked at a concentration of at least 500 nM when possible. An aliquot is saved for control (by western blot or SDS-PAGE) and samples are flash frozen at -80°C.

3.4. *In vitro* transcription/translation.

In vitro transcription/translation are performed or proteins unsuitable for *in vivo* purification (such as those forming aggregates). This approach is done using ActivePro Kit (Ambion) and following the Manufacturer's instructions. Glycerol is added to a final concentration of 30% to IVT samples prior to -80°C storage. Note that *in vitro* transcribed/translated proteins can be used non-purified (from the kit mixture) in the PBM hybridization.

Molar concentrations of all *in vitro* translated proteins are determined by Western blot utilizing a dilution series of recombinant GST (Sigma). Equal volumes of sample and known concentrations of GST are run on a standard Western blot procedure using anti-GST (G7781, Sigma dilution 1/5000) as a primary antibody, and anti-rabbit IgG-peroxydase (A0545, Sigma dilution 1/20 000). Concentrations are determined using Quantity One software version 4.5.0 (Bio-Rad) according to the GST standard curve.

3.5. PBM experiment

3.5.1. Making Agilent arrays double stranded.

1. Single-stranded oligonucleotide microarrays are double-stranded by primer extension using 1.17 µM RC stilt primer, 40 µM dATP, dCTP, dGTP, and dTTP, 1.6 µM Cy3 dUTP, 32 Units Thermo Sequenase™ DNA Polymerase, and 90 µl 10x reaction buffer. The common primer RC stilt may be labeled (Cy5) to check for uniformity of primer annealing.

2. The reaction mixture, microarrays, stainless steel hybridization chamber, and four-chambers gasket cover slip are pre-warmed to 85°C in a stationary hybridization oven and assembled according to the manufacturer's protocols.

3. After incubation at 85°C for 10 min, 75°C for 10 min, 65°C for 10 min, and 60°C for 90 min (see Note 5), the hybridization chamber is disassembled in 500 ml freshly made Wash buffer A at 37°C. Microarrays are transferred to a fresh dish, washed for 10 min in Wash buffer A at 37°C, washed once more for 3 min in PBS at 20°C, and spun dry by centrifugation at 40 g for 1 min (see Note 6).

4. Double stranded microarrays are scanned for Cy3 (using a resolution of at least 5 µm, excitation 542 nm, emission 570 nm), to check Cy3-dUTP incorporation homogeneity in the reverse strand. Double-stranded microarrays can be stored in dark and dry conditions for months before using for PBM experiments.

3.5.2. Protein Binding Microarray hybridization

1. Double-stranded microarrays are moistened in fresh Wash buffer A for 5 min. Microarrays are blocked with 150 µl Blocking solution under LifterSlip cover slips (Erie Scientific) for 1 h. During blocking, remove materials from freezer to thaw (zinc, BSA, DNA, protein, thaw on ice) and prepare the protein binding mixture.
2. The protein binding mixture is made of the purified TFs diluted to 100 nM (see Note 7) in a 175 µl final volume containing Blocking solution, 51.3 ng/µl salmon testes DNA and 0.2 µg/µl bovine serum albumin. Resulting mixtures are pre-incubated for 1 hour at room temperature
3. Blocking microarrays are washed once with Wash buffer B for 5 min and once with Wash buffer A for 2 min.
4. Pre-incubated protein binding mixtures are applied to individual chambers of a four-chamber gasket cover slip in a steel hybridization chamber (Agilent), and the assembled microarrays are incubated for 1 h at room temperature.
5. The hybridization chambers are individually disassembled in 500 ml freshly made Wash buffer A. Microarrays are washed again once with Wash buffer C for 5 min and once with Wash buffer A for 2 min.
6. Alexa488-conjugated rabbit polyclonal antibody to GST (Invitrogen) are diluted to 50 µg/ml in 1ml Blocking buffer and applied to a single-chamber gasket cover slip (Agilent).
7. The assembled microarrays are again incubated for 1h at room temperature, then individually disassembled in 500 ml freshly made Wash buffer D.
8. Microarrays are then washed twice with Wash buffer D for 3 min each, and once in PBS for 2 min. Slides are spun dry by centrifugation at 40 g for 5 min.

3.5.3. Microarray Stripping:

1. After scanning (described below), in order to re-use double stranded microarrays (see Note 8), bound proteins and antibodies are digested from double-stranded microarrays with 50 ml stripping solution, rotating overnight at 10 r.p.m. in a 50 ml Falcon tube at 37°C.
2. Microarrays are washed 3 times for 5 minutes each in Wash buffer C, once for 5 minutes in PBS, and rinsed in PBS in a 500 ml staining dish (slowly removed to ensure removal of detergent and uniform drying).
3. Before re-use, slides are scanned once at the highest laser power for Alexa488 (488 nm excitation (ex), 522 nm emission (em)) to confirm that no protein or antibody signal has remained.

3.5.4. Image Quantification and Data Normalization:

1. Protein-bound microarrays are scanned on a ProScanArray HT Microarray Scanner (Perkin Elmer) to detect Alexa488-conjugated antibody (488 nm ex, 522 nm em) using three different laser power settings to best capture a broad range of signal intensities and ensure signal intensities below saturation for all spots.
2. Microarray TIFF images are analyzed using GenePix Pro version 6.0 software (Molecular Devices), bad spots are manually flagged and removed. The three Alexa488 scans obtained at different laser power settings are combined using masliner software ***(22)*** available at http://arep.med.harvard.edu/masliner/supplement.htm.

There are several approaches for normalizing microarray data. Different approaches may be appropriated to PBMs and yield comparable results. In ***(9)***, PBM data were normalized using the function justvsn() available in the Bioconductor package vsn ***(23)***. Another normalization procedure applied to PBM data is described in ***(24)***.

3.6. Obtaining probe sequences

To analyze PBM raw data, one needs to obtain the sequences corresponding to the probes on the microarray. The original universal 10-mer de Bruijn sequence microarrays described in ***(8)*** are available via a End-User License Agreement (EULA) at http://the_brain.bwh.harvard.edu/UPBMseqn/UPBMseqn_agreement.html. The microarray designs we used are variations of the original design. All steps henceforth refer to the modified microarray design used in our study ***(9)***.

1. Go to http://the_brain.bwh.harvard.edu/UPBMseqn/UPBMseqn_agreement.html and download the excel file if you agree with the EULA.
2. Save the probe identifiers and the probe sequences for array de Bruijn #1.
3. Go to http://hugheslab.ccbr.utoronto.ca/supplementary-data/yeastDBD/ and download the two files with the probe ID mapping.
4. Remove the 25 nucleotides at the end of each sequence (3' end) in de Bruijn #1 arrays corresponding to the common primer GTCTGTGTTCCGTTGTCCGTGCTGT.
5. Follow the instructions available at http://hugheslab.ccbr.utoronto.ca/supplementary-data/yeastDBD/README to obtain the probe sequences on the two arrays used in ***(9)***.
6. Extract the overlapping 8-mers represented on each probe sequence. Note that an 8-mer and its reverse complement are considered to represent the same feature. For example, probe sequences containing either “AAAAAACC” or “GGTTTTTT” are group together as containing the same 8-mer.
7. Write a tab-delimited text file containing the probe identifier in the first column and the 8-mers contained on each probe in the second column (one 8-mer per line). For example, the first six lines of such a file might look as follows:

ProbeID	Kmer
TRHyeSpot40330	AAAAAAAA
TRHyeSpot40330	AAAAAAAA
TRHyeSpot40330	AAAAAAAA
TRHyeSpot40330	CAAAAAAA
TRHyeSpot40330	TCAAAAAA
TRHyeSpot40330	TTCAAAAA

A Perl script to perform steps 4 to 7 is available in the supplementary material provided with this article.

3.7. Obtaining 8-mer affinity measurements

Preference of a transcription factor for each 8-mer is represented using three different values: median intensity, robust Z-score ***(25)***, and Enrichment-score (E-score, ***(8)***). To do all computational steps to obtain these 8-mer based values, we adopted R. Advantages of using R are an integrated interactive environment for analysis and visualization, and the availability of many functions and tools. Furthermore, R has often been adopted for bioinformatics protocols (e.g., ***(26)***). In what follows, all R commands and their output appear in Courier New font. Commands are preceded by a > sign. Note that in this protocol we use the <− notation for variable assignment in R. Computation time is based on a 2-core MacBook Air machine with 8 GB in RAM. If no time is given, the step takes less than 5 minutes to complete.

1. Read in the Probe to 8-mer mapping file such as the one produced in the previous section (here named `ArrayA_probesIDs_2_8mers.txt` and assumed to be in a directory called YeastData) by typing in the R console:

```
> probe_kmer_mapping <-
read.table("YeastData/ArrayA_probesIDs_2_8mers.txt", sep = "\t",
stringsAsFactors = FALSE, header = TRUE)
> head(probe_kmer_mapping)
```

```
R output:
```

```
         ProbeID     Kmer
1 TRHyeSpot40330 AAAAAAAA
2 TRHyeSpot40330 AAAAAAAA
3 TRHyeSpot40330 AAAAAAAA
4 TRHyeSpot40330 CAAAAAAA
5 TRHyeSpot40330 TCAAAAAA
6 TRHyeSpot40330 TTCAAAAA
```

2. Read in the probe intensities file. This file contains a table with the probe IDs as rows and the intensity measurements for each probe per microarray as columns. The file Array_A_35mer_raw_data.txt containing data for 118 arrays available at http://hugheslab.ccbr.utoronto.ca/supplementary-data/yeastDBD/ is used to demonstrate the following steps.

```
> Data <- read.table("YeastData/Array_A_35mer_raw_data.txt", sep =
"\t", stringsAsFactors = FALSE, header = TRUE, row.names = 1)
> dim(Data)
[1] 43803   118
> head(rawData[,1:2])
```

R output:

```
                      ABF1_4505.2_ArrayA ABF2_2116.1_ArrayA.1
TRHyeControl100_DT_100           1433.298             5184.860
TRHyeControl101_DT_101           2503.233             3372.940
TRHyeControl102_DT_102           2158.167             6091.378
TRHyeControl103_DT_103           1255.000             4197.835
TRHyeControl104_DT_104           1879.434             9360.506
TRHyeControl105_DT_105           1901.071             4519.405
```

3. Assemble a table with the probe IDs, corresponding 8-mers and probe intensities.

```
> fullTable <- merge(probe_kmer_mapping, Data, by.x = "ProbeID",
by.y = "row.names")
> head(fullTable[,1:3])
```

R output:

```
             ProbeID     Kmer ABF1_4505.2_ArrayA
1 TRHyeControl1_DT_1 CATCGACC           1843.926
2 TRHyeControl1_DT_1 CCATCGAC           1843.926
```

```
3 TRHyeControl1_DT_1 CCCATCGA           1843.926
4 TRHyeControl1_DT_1 CCCCATCG           1843.926
5 TRHyeControl1_DT_1 CCCCCATC           1843.926
6 TRHyeControl1_DT_1 ACCCCCAT           1843.926
```

4. Compute the median intensity and log median intensity for each 8-mer per experiment.

```
     > median_intensity <- sapply( 3:ncol(fullTable), function(i) {
          tapply(fullTable[,i], fullTable[,"Kmer"], median)
     })
> colnames(median_intensity) <-
colnames(fullTable)[3:ncol(fullTable)]
> dim(median_intensity)
```

R output:

```
[1] 32896   118
```

```
> head(median_intensity[,1:2])
```

R output:

```
         ABF1_4505.2_ArrayA ABF2_2116.1_ArrayA.1
AAAAAAAA           2044.320             7044.084
AAAAAAAC           1844.253             7246.297
AAAAAAAG           2107.263             6254.257
AAAAAAAT           1950.312             7073.151
AAAAAACA           1847.276             7350.742
```

```
AAAAAACC           1971.743             7378.000
```

```
> log_median_intensity <- log(median_intensity)
```

5. Calculate the robust Z-score per 8-mer per experiment. The robust Z-score is the number of median absolute deviations (MAD) away from the overall median intensity.

```
> getZscore <- function(mi){ (mi - median(mi)) / mad(mi)}
```

```
> zscore <- apply(log_median_intensity, 2, getZscore)
> dim(zscore)
```

```
R output:
[1] 32896   118
```

```
> head(zscore[,1:2])
```

```
R output:
```

```
         ABF1_4505.2_ArrayA ABF2_2116.1_ArrayA.1
AAAAAAAA           2.942674         -0.079967633
AAAAAAAC           0.468769          0.482131694
AAAAAAAG           3.671098         -2.441892575
AAAAAAAT           1.811884          0.001816074
AAAAAACA           0.508110          0.766348393
AAAAAACC           2.074392          0.839859899
```

6. Obtain a table with the ranks of the probes in descending order by their intensity (i.e., the rank of probe with the highest intensity is 1) per experiment.

```
> assignRanks <- function(intensities){
      length(intensities) - rank(intensities, ties.method = "first")
+ 1
 }
> ranksTable <- apply(Data, 2, assignRanks)
> dim(ranksTable)
```

R output:

```
[1] 43803   118
```

```
> head(ranksTable[,1:2])
```

R output:

	ABF1_4505.2_ArrayA	ABF2_2116.1_ArrayA.1
TRHyeControl100_DT_100	41017	39360
TRHyeControl101_DT_101	3093	43704
TRHyeControl102_DT_102	8135	32012
TRHyeControl103_DT_103	43547	43029
TRHyeControl104_DT_104	18268	5116
TRHyeControl105_DT_105	17199	42280

7. Assemble a table with the probe IDs, corresponding 8-mers and probe ranks.

```
> ranksTableFull <- merge(probe_kmer_mapping, ranksTable, by.x =
"ProbeID", by.y = "row.names")
> dim(ranksTableFull)
```

R output:

```
[1] 1226484     120
```

```
> head(ranksTableFull[,1:3])
```

R output:

```
             ProbeID     Kmer ABF1_4505.2_ArrayA
1 TRHyeControl1_DT_1 CATCGACC              19982
2 TRHyeControl1_DT_1 CCATCGAC              19982
3 TRHyeControl1_DT_1 CCCATCGA              19982
4 TRHyeControl1_DT_1 CCCCATCG              19982
5 TRHyeControl1_DT_1 CCCCCATC              19982
6 TRHyeControl1_DT_1 ACCCCCAT              19982
```

8. Calculate the E-score per 8-mer per experiment. The E-score of an 8-mer is the subtraction of the average rank of the top half of the probes in which the 8-mer is absent minus the average rank of the top half of the probes in which the 8-mer occurs divided by the total number of probes in both top halves ***(8)***. For example, suppose we have 200 probes from which 10 contain a given 8-mer. The E-score of this 8-mer is obtained by

subtracting the average rank of the 95 brightest probes in which the 8-mer is absent minus the average rank of the 5 brightest probes in which the 8-mer occurs, and dividing the result of this subtraction by 100 (see Note 9).

```
#Exact calculation - slow
> get_Escores_exact <- function(ranks, numProbes){
      keepFraction <- 0.5

      sortRanks <- sort(ranks)
      #ranks of background probes; i.e., those without the 8mer
      ranksb <- setdiff(1:numProbes, sortRanks)

      n <- trunc(length(sortRanks) * keepFraction)
      m <- trunc((numProbes - length(sortRanks)) * keepFraction)

      pf <- sortRanks[1:n]
      pb <- ranksb[1:m]

      (sum(pb) / m - sum(pf)/n) / (m+n)

}
```

9. Alternatively the E-score can be approximated by the area under the receiver operating characteristic curve (AUC) minus 0.5 as it is done in the "seed_and_wobble.pl" program accompanying ***(24)***. The following R code is based on the E-score

approximation done in the “seed_and_wobble.pl” program. This function is much faster than the exact calculation done in the previous step. (Timing ~ 15 min)

```
> get_Escores_approx <- function(ranks, numProbes){
      keepFraction <- 0.5
      sortRanks <- sort(ranks)
      n <- trunc(length(sortRanks) * keepFraction)
      m <- trunc((numProbes - length(sortRanks)) * keepFraction)
      ranksum <-   sum( sapply(1:n, function(i) {
            if (sortRanks[i] - i > m) {
                m+i-1
            } else {
                sortRanks[i]
            }
      }))

      ((n^2+n)/2 + n*m/2 - ranksum) / (n*m)
}

> E_score <- sapply( 3:ncol(ranksTableFull), function(i) {
      tapply(ranksTableFull[,i], ranksTableFull[,"Kmer"],
get_Escores_approx, nrow(ranksTable)  )
})
> colnames(E_score) <-
colnames(ranksTableFull)[3:ncol(ranksTableFull)]
> dim(E_score)
```

```
R output:
[1] 32896   118

> head(E_score[,1:2])

R output:

         ABF1_4505.2_ArrayA ABF2_2116.1_ArrayA.1
AAAAAAAA          0.2518629          -0.06358507
AAAAAAAC          0.2134302          -0.02259507
AAAAAAAG          0.3254102          -0.14840729
AAAAAAAT          0.2198637           0.09712511
AAAAAACA          0.0913399           0.04167966
AAAAAACC          0.2695739          -0.08644905
```

10. As a sanity check, check the E-score and Z-score distribution. A PBM experiment is considered successful if it has at least one 8-mer with an E-score above 0.45 and the Z-score distribution shows a long right tail (Figure 1A). Additionally, Z-scores of independent PBM experiments, done with the same TF, exhibit positive correlation (Figure 1B).

All 8-mer based values for TFs studied in ***(9)*** are available in NCBI Gene Expression Omnibus (GEO) under accession GSE12349.

3.8. Comparing 8-mer profiles between TFs of the same family

Using the 8-mer profiles of various TFs, we can compare DNA binding specificities of TFs of the same family. For example, Figure 2 shows a comparison of the 8-mer E-scores for two yeast TFs of the GATA family with distinct motifs, GAT3 and GZF3; while Figure 3 shows a comparison of the 8-mer E-scores for two yeast TFs of the same GATA family that share the same primary motif, GLN3 and GZF3.

There is a relation between 8-mer profiles and sequence similarity for TFs of the same family. Figure 4 shows this relation for TFS of the yeast zinc finger GATA family. Observation of this fact and the availability of 8-mer profiles produced by PBMs allows to apply machine learning techniques that infer binding preference of a TFs using the k-mer affinity information available for other family members (e.g., ***(27, 28)***). R offers several packages to apply techniques such as random forests (RFs), k-nearest neighbour (KNN) and support vector machines.

3.9. Obtaining DNA sequences motifs from top-scoring 8-mers.

There are several models to represent the DNA sequence specificity of a TF and several methods to obtain such a model from a set of sequences. Position weight matrices (PWMs) are the predominant paradigm to represent DNA motifs bound by a TF. A PWM models the DNA sequence preference of a TF as matrix with a row for each symbol in the alphabet (i.e. A, C, G and T) and a column for each position of the TFBS (i.e., number of columns is equal to the length of the TFBS). Each column provides a score per nucleotide representing the relative preference for the given base at that position in the binding site. State of the art algorithms and paradigms to represent TFBS have recently been evaluated ***(29)***. Based on this evaluation, the best performing PWM-based method is

BEEML-PBM ***(30)*** and the best 8-mer based method is FeatureREDUCE (http://bussemakerlab.org/people/ToddRiley/featurereduce.html). BEEML-PBM is available at http://stormo.wustl.edu/beeml/. This method is written in R and requires as input a two-column table with the normalized intensities and probe sequences, and a PWM as a seed. This seed PWM can be either one obtained by another method, one available in the literature or one from a TF of the same family.

3.10. Seeking transcription-factor binding sites (TFBS) onto promoter region

In addition to determine the sequence specificities of a TF and represent this specificities as a PWM, one usually wants to identify genes being regulated by this TF. Putative targets of a TF can be determined by finding genes whose promoter region contains the motif bound by that TF. It is possible to do all computational steps to identify TFBSs within R. In the following steps, we continue using the same notation as in section 3.4. These steps were adapted from the Bioconductor ***(31)*** workflow available at http://www.bioconductor.org/help/workflows/generegulation/

1. Read into R the PWMs of the TFs of interest. An excel file with PWMs for the yeast TFBS determined in [9] is available at http://hugheslab.ccbr.utoronto.ca/supplementary-data/yeastDBD/. Assume we have extracted PWMs from this excel file into tab-delimited text files ending with "_PWM.txt" in the directory YeastData. We can then read all these files and converted the PWMs into count matrices by typing into the R console:

```
> files <- list.files("YeastData", pattern = "*_PWM.txt",
include.dirs = TRUE, recursive = TRUE, full.names = TRUE)
```

```
> PWMs <- sapply(files, read.table, sep = "\t", stringsAsFactors =
FALSE, header = TRUE, row.names = 1)
> names(PWMs) <- gsub(".*/", "", gsub("_PWM.txt", "", files), perl =
TRUE)
> PCMs <- lapply(PWMs, function(pwm) {round(100 * pwm) })
> names(PCMs)
[1] "GAT3" "GLN3" "GZF3"
> PCMs[["GAT3"]]
```

R output:

```
  X1 X2 X3 X4 X5 X6 X7 X8 X9
A 65  8 92 12  3  6 51 28 32
C  9  2  0  0 94 27 19 33 25
G 20 88  0  1  1  4 18 21 21
T  6  1  8 87  2 63 12 17 21
```

2. Obtain the promoter sequence of the genes in whose promoter region one wants to look for a TFBS. Note that the genes must be listed using their systematic name. In this example, we are using 15 genes listed as targets of GZF3 in Saccharomyces Genome Database ***(32)***.

```
> ORFs <- read.table("YeastData/GZF3_ORF_targets.txt", header =
FALSE, stringsAsFactors = FALSE)
> ORFs<- ORFs[,1,drop = TRUE]
> ORFs[1:5]
```

R output:

```
 [1] "YCL025C"   "YPR171W"   "YBL042C"   "YKR039W"   "YFL021W"
```

```
> library(GenomicFeatures)
> library(BSgenome.Scerevisiae.UCSC.sacCer3)
> library(TxDb.Scerevisiae.UCSC.sacCer3.sgdGene)
> transcripts_coordinates <-
transcriptsBy(TxDb.Scerevisiae.UCSC.sacCer3.sgdGene, by =
"gene")[ORFs]
> promoter.seqs <- getPromoterSeq(transcripts_coordinates,
Scerevisiae, upstream = 1000, downstream = 0)
> head(promoter.seqs, n=3)
```

R output:

```
DNAStringSetList of length 6
[["YCL025C"]]
CTGAAAGAGCGCCTTTACCTCAACCTACCATGGCAAACATAACAGAAAACATAAAAAAATTATCCTAG
AGCCCAATGTTCCATGAAAAGAGCTGTGGCAAGGACAGAAACAAAAAAAAAATCAAGAACTCAACATT
A...
[["YPR171W"]]
CTGATGTTCAGTAAAGCCGCCTAGCTTTACGTGCCGAAATATTGATAATATGTCTCAGCCACTTCCTG
GCTTAACTATTTAAATGATATTTCTGCATCCATCGGTATGGCGCACAATAAACGGTATCTGAGAATAT
C...
[["YBL042C"]]
GCAATAGTGGCCATATTTTGTTTAACTTTATAGTTCAATAGTCTTGGCTACTCTCTTTCCAACTCAGT
```

```
TCACCTTGTATTATACCGCTTGTTTTTGCCACCCTTTGAGTTTCCTCGATCCTTTAAGTTGGAAAAGA
T...

> promoter.seqs <- unlist(promoter.seqs)
> head(promoter.seqs, n = 3)

R output:

   A DNAStringSet instance of length 3
    width seq
names
[1]  1000
CTGAAAGAGCGCCTTTACCTCAACCTACCATGGCAAACATAACAGAAAACATAAAAAAAT...GTTTA
TTATGTAATCTTTATAGAAGAAGCACGCTAATATAGACAAAGATAGCTTCGCACA YCL025C
[2]  1000
CTGATGTTCAGTAAAGCCGCCTAGCTTTACGTGCCGAAATATTGATAATATGTCTCAGCC...ATTCT
AATCAATAAAAGTCACAGTAACCAGCTTTTCCTAGCTTTTCGAAGTTTCGGAAGT YPR171W
[3]  1000
GCAATAGTGGCCATATTTTGTTTAACTTTATAGTTCAATAGTCTTGGCTACTCTCTTTCC...CATTG
CGGAAATAAAAGGCGGTAACTAGTCCTCTCATTCATTAATTCTATATAAGAGAAA YBL042C
```

3. Find matches of the motifs in the promoter sequences obtained in the previous step. After executing the first two commands, pwm.hits contains a list per TF containing the locations of putative TFBSs per gene.

```
> library(Biostrings)
> pwm.hits <- lapply(PCMs, function(pwm) {
      sapply(promoter.seqs, function(pseq, pwm) {matchPWM(pwm,
```

```
    pseq, min.score = "90%")}, as.matrix(pwm))
     })
```

```
> names(pwm.hits)
```

R output:

```
[1] "GAT3" "GLN3" "GZF3"
```

```
> head(pwm.hits[["GAT3"]], n = 2)
```

R output:

```
$YCL025C
  Views on a 1000-letter DNAString subject
subject:
CTGAAAGAGCGCCTTTACCTCAACCTACCATGGCAAACATAACAGAAAACATAAAAAAATTATCCTAG
AGC...ATGTAGAACAAGTTTATTATGTAATCTTTATAGAAGAAGCACGCTAATATAGACAAAGATAG
CTTCGCACA
views: NONE

$YPR171W
  Views on a 1000-letter DNAString subject
subject:
CTGATGTTCAGTAAAGCCGCCTAGCTTTACGTGCCGAAATATTGATAATATGTCTCAGCCACTTCCTG
GCT...TTTATATATGAATTCTAATCAATAAAAGTCACAGTAACCAGCTTTTCCTAGCTTTTCGAAGT
TTCGGAAGT
views: NONE
```

```
> head(pwm.hits[["GZF3"]], n = 2)

R output:

$YCL025C
  Views on a 1000-letter DNAString subject
subject:
CTGAAAGAGCGCCTTTACCTCAACCTACCATGGCAAACATAACAGAAAACATAAAAAAATTATCCTAG
AGC...ATGTAGAACAAGTTTATTATGTAATCTTTATAGAAGAAGCACGCTAATATAGACAAAGATAG
CTTCGCACA
views:
    start end width
[1]   571 578     8 [AGATAAGC]
[2]   747 754     8 [TGATAAGA]

$YPR171W
  Views on a 1000-letter DNAString subject
subject:
CTGATGTTCAGTAAAGCCGCCTAGCTTTACGTGCCGAAATATTGATAATATGTCTCAGCCACTTCCTG
GCT...TTTATATATGAATTCTAATCAATAAAAGTCACAGTAACCAGCTTTTCCTAGCTTTTCGAAGT
TTCGGAAGT
views:
    start end width
[1]    43  50     8 [TGATAATA]

> sessionInfo()
```

R output:

```
R version 3.0.2 (2013-09-25)
Platform: x86_64-apple-darwin10.8.0 (64-bit)

locale:
[1] en_CA.UTF-8/en_CA.UTF-8/en_CA.UTF-8/C/en_CA.UTF-8/en_CA.UTF-8

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets
methods   base

other attached packages:
 [1] BSgenome.Scerevisiae.UCSC.sacCer3_1.3.19    BSgenome_1.28.0
Biostrings_2.30.1
 [4] TxDb.Scerevisiae.UCSC.sacCer3.sgdGene_2.9.0
GenomicFeatures_1.12.4                      AnnotationDbi_1.24.0
 [7] Biobase_2.22.0
GenomicRanges_1.14.3                        XVector_0.2.0
[10] IRanges_1.20.6                              BiocGenerics_0.8.0

loaded via a namespace (and not attached):
 [1] biomaRt_2.18.0     bitops_1.0-6       DBI_0.2-7
RCurl_1.95-4.1     Rsamtools_1.12.4   RSQLite_0.11.4
rtracklayer_1.20.4
 [8] stats4_3.0.2       tools_3.0.2        XML_3.95-0.2
zlibbioc_1.6.0
```

PWM is a classical model to represent TFBS. It allows summarizing sequence binding information of a TF obtained by various methods into a single motif (see JASPAR database as an example ***(33)***, and it is easily represented as a sequence logo to visualize the motif with the highest affinity.

The main advantage of PBM experiments is the possibility to generate comprehensive k-mer profiles, which provide more detailed and extensive information on binding affinities. Secondary motifs may be thus revealed by the k-mer profile, exhibiting different sequences and affinities than the main motif (see Figure 3 for an example). These secondary motifs might be excluded from the PWM representation. Such secondary motifs, possibly enhanced by cofactors under physiological conditions, might be relevant *in vivo*.

4. Notes

1. For Zinc Finger proteins only, add Zinc Acetate to all buffers (including LB media, PBS, Wash buffer, *etc*) to a final concentration of 50 μM.
2. It is important to check on a SDS-PAGE gel the the-GST fusion protein inductions in the IPGT induced and uninduced sample running the crude extract obtain from 2 ml of both cultures on a SDS-PAGE gel.
3. Sonication settings:

Automatic setting: Pulse 1sec; Rest 3sec. Total pulse time: 2min, Amplitude 60min.

Note that sonication settings depend on the model of sonicator being used. The probe size is usually ½ or ¾ inches. Sonication process should be modified for the type of probe, cell, etc.

4. The NaCl is used to decrease unspecific ionic binding of proteins to the GSTbeads.

5. Temperature is gradually decreased to ensure proper annealing of the RC stilt primer to template DNA probed on the array. Hybridization can be performed on a Tecan Hybridization station is available, in this case, all the buffers must be filter-sterilized.

6. All washes are performed in a 50 ml Falcon tube at room temperature on a wheel rotating at 10 r.p.m.

7. The optimal molarity depends on the Kd of each protein. 100nM is an optimized concentration that we determined experimentally and apply to all our TFs but a range from 5 to 200nM (empirically determined) is possible, depending proteins.

8. Microarray can be re-used two –without any loss- to up to four times to keep a good quality of signal.

9. 9. The function in step 3.4.8 does the exact E-score calculation; note however that this exact calculation is quite slow (timing ~8 min per TF). We recommend to use instead the function defined in step 3.4.9.

5. References

1. B. Ren, F. Robert, J.J. Wyrick, et al. (2000) Genome-wide location and function of DNA binding proteins, Science (New York, N.Y.). 290, 2306–2309.

2. V.R. Iyer, C.E. Horak, C.S. Scafe, et al. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF, Nature. 409, 533–538.

3. D.S. Johnson, A. Mortazavi, R.M. Myers, et al. (2007) Genome-wide mapping of in vivo protein-DNA interactions, Science (New York, N.Y.). 316, 1497–1502.

4. C.-L. Wei, Q. Wu, V.B. Vega, et al. (2006) A global map of p53 transcription-factor binding sites in the human genome, Cell. 124, 207–219.

5. A.R. Oliphant, C.J. Brandl, and K. Struhl (1989) Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein, Molecular and cellular biology. 9, 2944–2949.

6. A. Zykovich, I. Korf, and D.J. Segal (2009) Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing, Nucleic acids research. 37, e151.

7. A. Jolma, T. Kivioja, J. Toivonen, et al. (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities, Genome research. 20, 861–873.

8. M.F. Berger, A.A. Philippakis, A.M. Qureshi, et al. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities, Nat Biotech. 24, 1429–1435.

9. G. Badis, E.T. Chan, H. van Bakel, et al. (2008) A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters, Molecular cell. 32, 878–887.

10. R. Gordân, K.F. Murphy, R.P. McCord, et al. (2011) Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights, Genome biology. 12, R125.

11. C. Zhu, K.J.R.P. Byers, R.P. McCord, et al. (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors, Genome Research. 19, 556–566.

12. M.F. Berger, G. Badis, A.R. Gehrke, et al. (2008) Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences, Cell. 133, 1266–1276.

13. G. Badis, M.F. Berger, A.A. Philippakis, et al. (2009) Diversity and complexity in DNA recognition by transcription factors, Science (New York, N.Y.). 324, 1720–1723.

14. B.W. Busser, D. Huang, K.R. Rogacki, et al. (2012) Integrative analysis of the zinc finger transcription factor Lame duck in the Drosophila myogenic gene regulatory network, Proceedings of the National Academy of Sciences of the United States of America. 109, 20768–20773.

15. R.D. Finn, J. Clements, and S.R. Eddy (2011) HMMER web server: interactive sequence similarity searching, Nucleic acids research. 39, W29–37.

16. J. Schultz, R.R. Copley, T. Doerks, et al. (2000) SMART: a web-based tool for the study of genetically mobile domains, Nucleic acids research. 28, 231–234.

17. T.R. Hughes and C.G. de Boer (2013) Mapping yeast transcriptional networks, Genetics. 195, 9–36.

18. C.G. de Boer and T.R. Hughes (2011) YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities, Nucleic Acids Research. gkr993.

19. M.Z. Li and S.J. Elledge (2005) MAGIC, an in vivo genetic method for the rapid construction of recombinant DNA molecules, Nature genetics. 37, 311–319.

20. C. Aslanidis and P.J. de Jong (1990) Ligation-independent cloning of PCR products (LIC-PCR), Nucleic acids research. 18, 6069–6074.

21. A.J. Walhout, G.F. Temple, M.A. Brasch, et al. (2000) GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes, Methods in enzymology. 328, 575–592.

22. A.M. Dudley, J. Aach, M.A. Steffen, et al. (2002) Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range, Proceedings of the National Academy of Sciences. 99, 7554–7559.

23. W. Huber, A. von Heydebreck, H. Sueltmann, et al. (2002) Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression, Bioinformatics. 18 Suppl. 1, S96–S104.

24. M.F. Berger and M.L. Bulyk (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors, Nature protocols. 4, 393–411.

25. A. Birmingham, L.M. Selfors, T. Forster, et al. (2009) Statistical methods for analysis of high-throughput RNA interference screens, Nature methods. 6, 569–575.

26. S. Anders, D.J. McCarthy, Y. Chen, et al. (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor, Nature protocols. 8, 1765–1786.

27. T.M. Alleyne, L. Pena-Castillo, G. Badis, et al. (2009) Predicting the binding preference of transcription factors to individual DNA k-mers, Bioinformatics (Oxford, England). 25, 1012–1018.

28. R.G. Christensen, M.S. Enuameh, M.B. Noyes, et al. (2012) Recognition models to predict DNA-binding specificities of homeodomain proteins, Bioinformatics (Oxford, England). 28, i84–9.

29. M.T. Weirauch, A. Cote, R. Norel, et al. (2013) Evaluation of methods for modeling transcription factor sequence specificity, Nature biotechnology. 31, 126–134.

30. Y. Zhao and G.D. Stormo (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity, Nature biotechnology. 29, 480–483.

31. R.C. Gentleman, V.J. Carey, D.M. Bates, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics, Genome biology. 5, R80.

32. J.M. Cherry, E.L. Hong, C. Amundsen, et al. (2012) Saccharomyces Genome Database: the genomics resource of budding yeast, Nucleic acids research. 40, D700–5.

33. A. Sandelin, W. Alkema, P. Engström, et al. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles, Nucleic Acids Research. 32, D91–D94.

34. C.T. Workman, Y. Yin, D.L. Corcoran, et al. (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos, Nucleic acids research. 33, W389–392.

6. Table and figure legends

Figure 1.

A. Distribution of Z-scores of a successful array for the TF GLN3. Note the long right tail of the distribution. **B.** Correlation of 8-mer E-scores for the TF GLN3 obtained from two PBM experiments performed on microarrays of different designs. The red line is the loess-smoothed line. The vertical and horizontal gray lines indicate the 0.45 E-score.

Figure 2.

Top: Scatter plot comparing 8-mer E-scores for two yeast TFs of the GATA zinc finger family, GAT3 and GZF3. The highlighted dots representing 8-mers containing the 6-mers indicated on the top left corner of the plot show a clear difference in the sequence preference of these TFs. Bottom: Sequence logos of the TFBS of both TFs. Sequence logos were created using enoLOGOS [34].

Figure 3.

Top: Scatter plot comparing 8-mer E-scores for two yeast TFs of the GATA zinc finger family, GLN3 and GZF3. Blue dots represent 8-mers containing either “AGATAA”, “AGATAG”, “CGATAA”, “CGATAG”, “TGATAA”, or “TGATAG” and with an E-score > 0.45

for either of the two TFs. These TFs show identical preferences for the same highest-scoring 8-mers. Green and yellow dots represent 8-mers containing respectively "AATCT" and "ATATC", with an E-score > 0.3 for either of the two TFs. The distribution of these dots in the scatter plot indicates a difference in lower affinity sequence preferences between GLN3 and GZF3. Bottom: Sequence logos of the TFBS of both TFs. Sequence logos were created using enoLOGOS ***(34)***.

Figure 4.

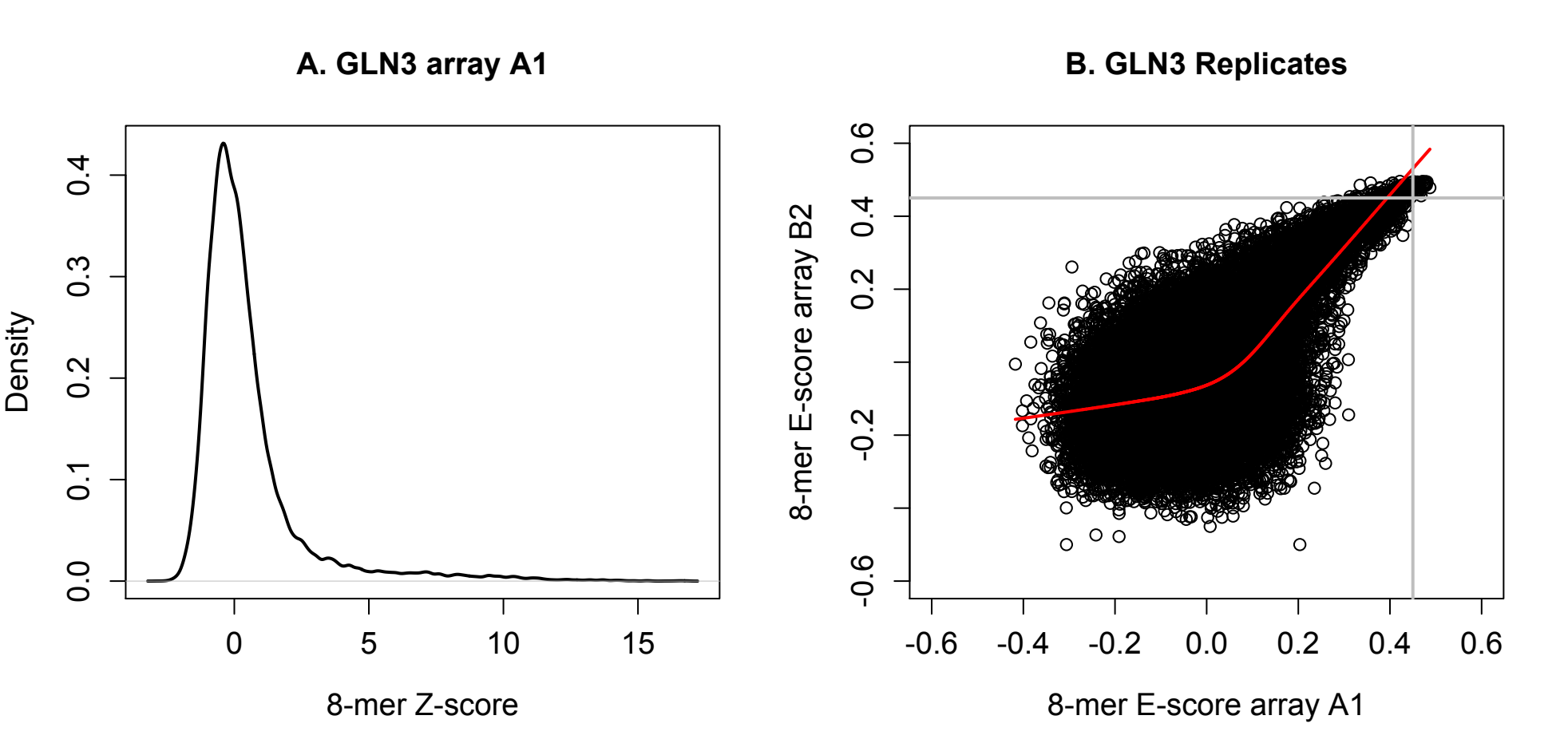
Similarity between 8-mer profiles across TFs of the GATA zinc finger family as a function of the percentage of sequence identity across the DNA-binding domains of these TFs. The more similar the sequences of the DBDs are, the more similar the 8-mer profiles. The red line is the loess-smoothed line.

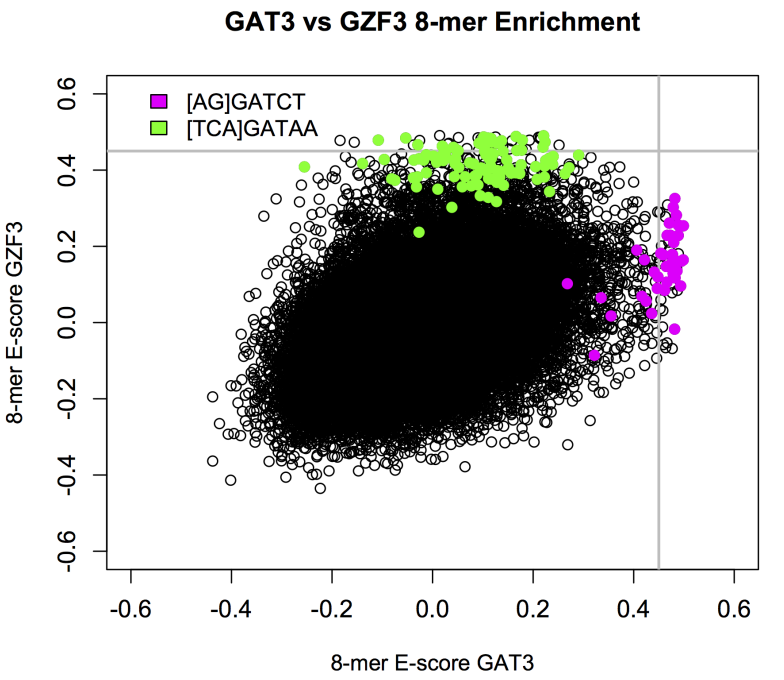
Table 1: Oligonucleotide sequences

Name	Sequence
Stilt sequence	5'- CTCACAATCTTGACGGCAGGCATGT-3'
RC Stilt	5'- ACATGCCTGCCGTCAAGATTG-3'

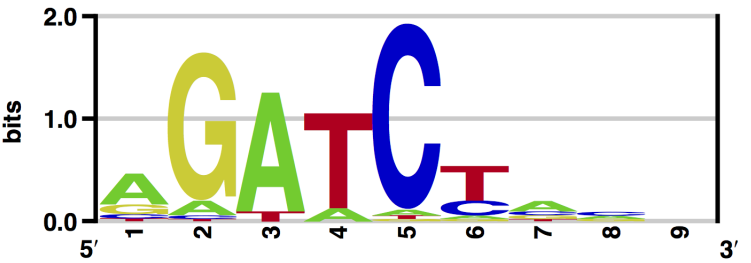
Aknowledgement

We thank Shaheynoor Talukder for standard operating procedure and Timothy R. Hughes for data availability. We also thank Esther T. Chan for useful comments. GB work was supported by the CIHR, the Institut Pasteur and the Centre National pour la Recherche Scientifique. LPC's work was supported by a NSERC Discovery Grant and Memorial University of Newfoundland.

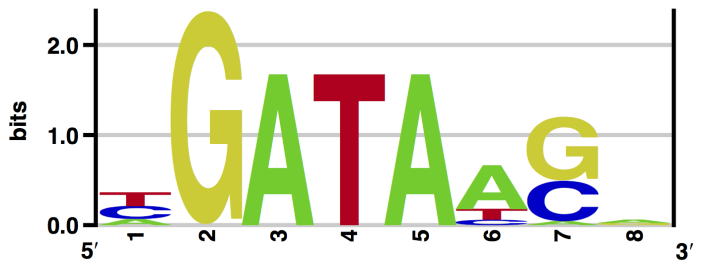

A. GLN3 array A1
Density
0.0
0.1
0.2
0.3
0.4
0
5
10
15
8-mer Z-score
B. GLN3 Replicates
8-mer E-score array B2
-0.6
-0.2
0.2
0.4
0.6
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
8-mer E-score array A1

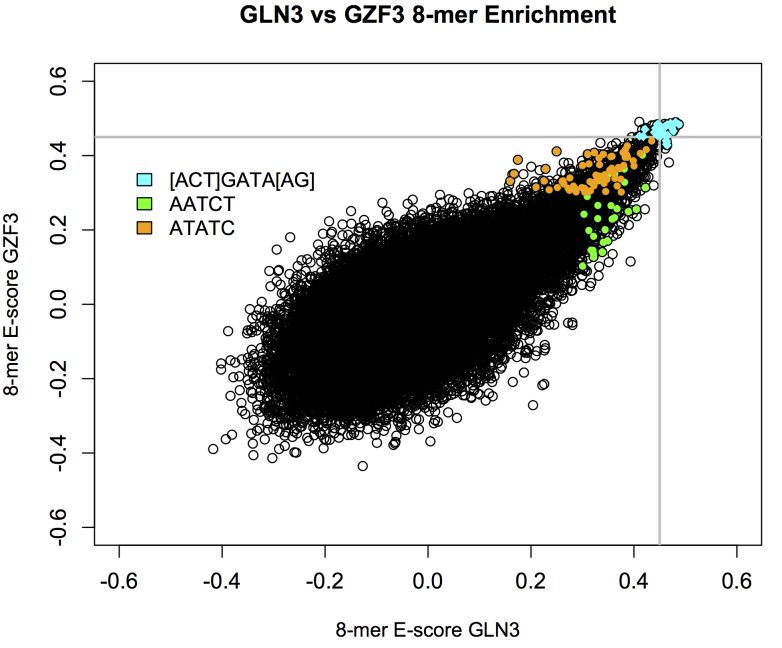

GAT3 vs GZF3 8-mer Enrichment
[AG]GATCT
[TCA]GATAA
8-mer E-score GZF3
8-mer E-score GAT3
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6

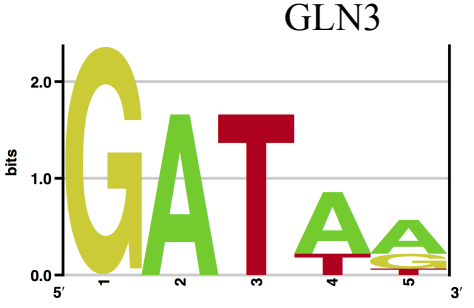
GAT3

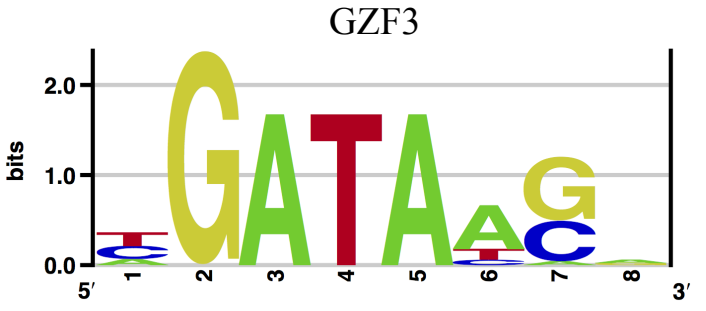

bits
2.0
1.0
0.0
5′
3′
1
2
3
4
5
6
7
8
9

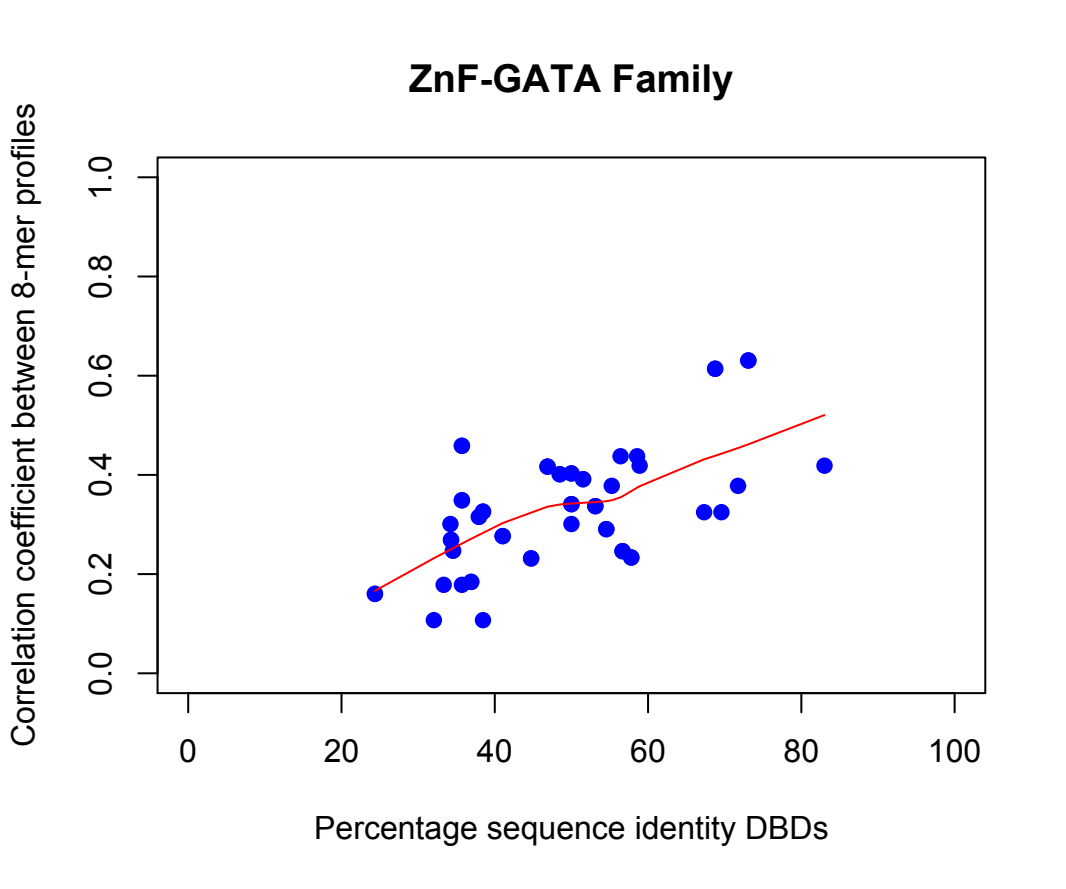
GZF3


bits
2.0
1.0
0.0
5′
3′
1
2
3
4
5
6
7
8


GLN3 vs GZF3 8-mer Enrichment
[ACT]GATA[AG]
AATCT
ATATC
8-mer E-score GZF3
8-mer E-score GLN3
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6


GLN3
bits
2.0
1.0
0.0
5′
3′


GZF3
bits
2.0
1.0
0.0
5′
3′


ZnF-GATA Family
Correlation coefficient between 8-mer profiles
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Percentage sequence identity DBDs