Group Meeting_Fangjie (Last: 2019-03)

> Recent progress

> SELEX with different temperatures and solvents

> SELEX with a TF concentration gradient

> Analysis for the binding events around TSS

> SELEX with different temperatures and solvents

> Introduction

Trascription factor-DNA binding involves multiple types of interactions, such as electronic interaciton, H-bond, van der waals, and hydrophobic interactions. To understand the forces that underline a TF's binding specificity, SELEX has been performed with a series of different temperatures (0°C, 10°C, 20°C, 30°C, 40°C, 50°C) and different solvents (H2O, EG).

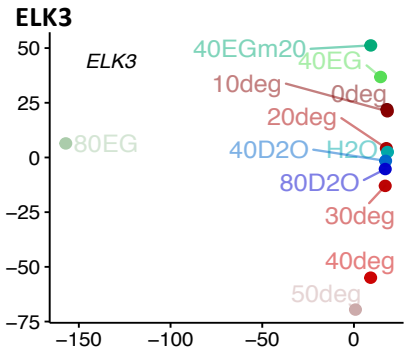
> Effect of temperature and solvent by top kmers

The specificity of a TF is described by the enriched kmers (9mers here). To explore the overall specificity change between different conditions, we used the top 100,000 9mers as the feature vector for each condition, and examined the specificity change with different dimensional reduction approaches. The rank of each 9 mer is used instead of frequency because the selection stringency can differ a lot between conditions.

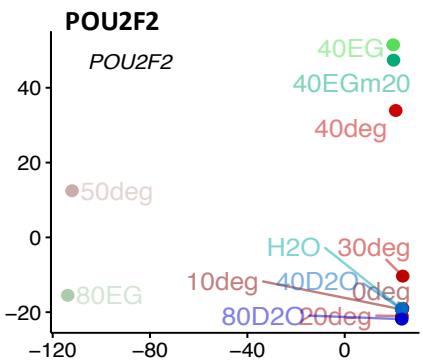
The PCA results show that
_ for most TFs (e.g. ELK3), increasing the temperature and changing the solvent from H2O to EG shift TF binding specificity towards opposite directions. The conclusion from PCA is supported by the raw xy-plots at the bottom.
_ for some TFs (e.g. POU2F2), increasing the temperature and changing solvent to EG shifts the TF binding specificity to the same direction.

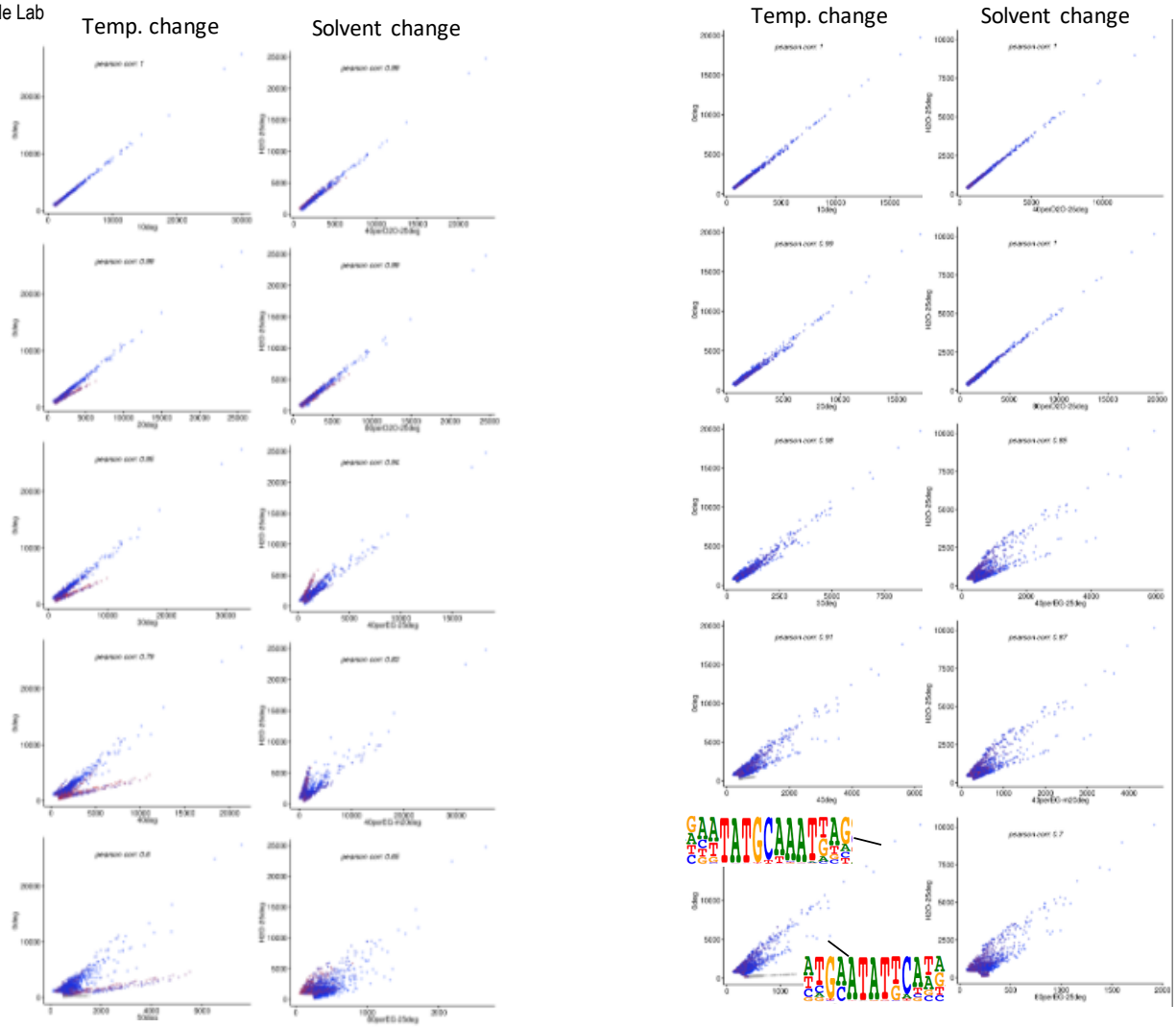
Increasing temperature and using EG have opposite effects on the electrostatic interactions and hydrophobic interactions of TF-DNA binding. The only common effect between them - if TF structure not significantly changed - is that a high temperature and using EG both destabilize the indirect H-bound. It is likely the two modes of POU2F2 differs in the amount of indirect H-bond.

As alternative visualization methods, I also tried to use the euclidean distance between kmer ranks of different samples and generated a 2-D projection plot, the results are very similar to the PCA plots (PC1 vs PC2). In addition, the output of the t-sne approach is more random, and gives distinctly different visuallizations. This is because whereas t-sne has higher power in separating less well-defined clusters, due to its non-linear feature and randomness of initialization, it is less preferred for the purpose to quantitatively explore the distance between conditions.



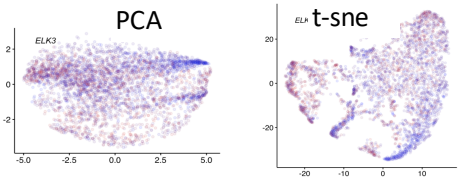
PCA



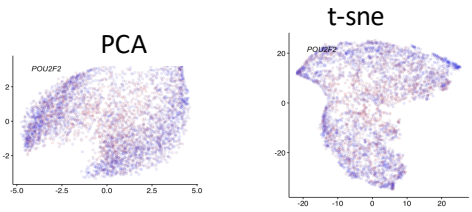


_ For each 9mer, I also used the counts across different conditions as features, and clustered all 9mers for each TF. There are clusters visible both with PCA and t-sne. With additional experiments planned on different conditions, this could be a very promising approach for estimating the number of binding modes for a TF (rather than picking lines from many xy-plots, direct counting of the cluster numbers here should do).

ELK3



POU2F2



> Future plan

_ Base the analysis on binding modes rather than kmers, then pick the extreme examples of modes that have prototypical behaviors across different conditions, and study the underlying mechanisms.

_ Study the hydration shell change of TFs upon binding

_ Salt concentration titration: decompose for electrostatic interaction

_ H2O-like molecules: decompose for H2O-mediated H-bound

_ Screen through the following small molecules

* solvent/solute hydrophilicity series

  Methanol, Ethanol, 1-propanol, n-butanol

  sodium formate/acetate/propionate/butyrate/sodium valproic acid

* additional solvents to test

  Acetonitrile, Dimethylformamide, Acetone, 2-Propanol, DMSO

* kosmo-chaotrope series

  - variation for anion

    Na3PO4, Na2SO4, NaF, NaCl, NaBr, NaI, NaClO4

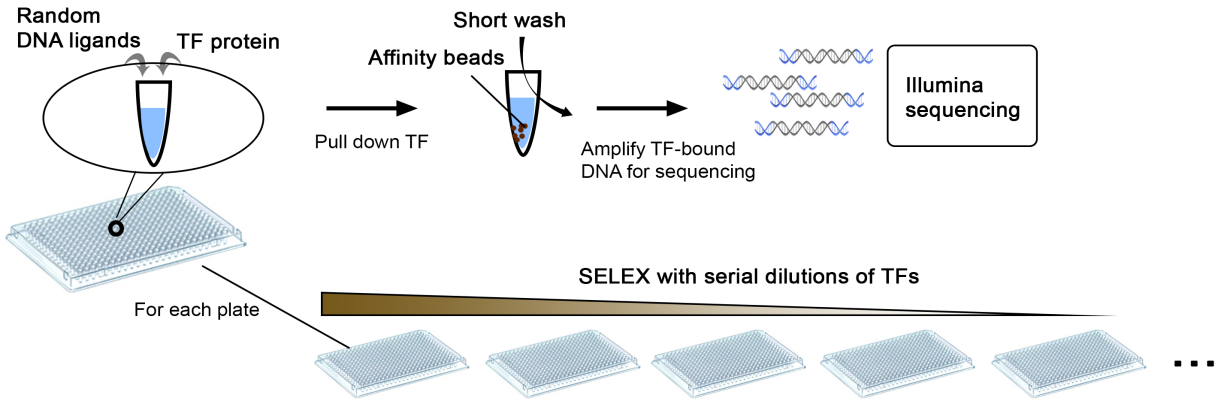
- variation for cation
  guanidinium SCN & Cl, tetramethylammonium Cl, CsCl, KCl, NaCl
- non-ionic
  trehalose, urea

* polyamines
  Ethylenediamine Cl, spermine Cl, spermidine Cl, Pentaethylenehexamine(PEHA)
* DNA intercalater
  SYBR_Gold, SYBR_Green, SYBR_Safe
* Amino acid binding competitor
  Arg, Asp, Lys, Asn

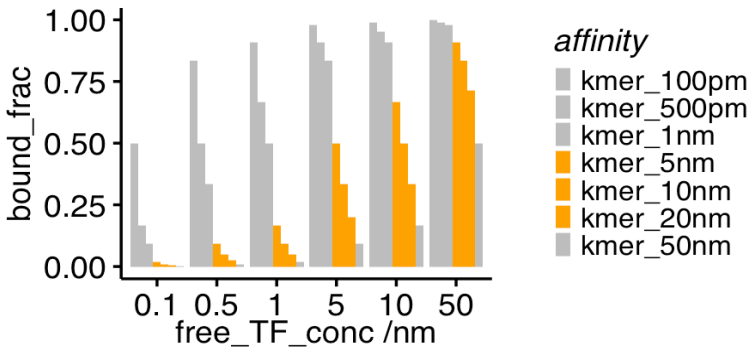
> SELEX with TF concentration gradient for TF-DNA affinities

> Introduction and experiment design

Sequencing provides accurate measurement of the ratio between different DNA sequences, however, it is not able to measure the absolute concentration of each sequence. Therefore, while it is straitforward to measure the relative affinity between different DNA sequences to a TF, to obtain the absolute affinity is difficult. Here, we try to use the ratios between different DNA sequences to figure out both the relative and absolute TF-DNA affinities. The experiment design is illustrated below:



During the titration of TF concentration, the ratios between kmers of different affinities will change, because binding of high-affinity kmers will saturate after all are bound. A simulated result is shown below, at 50nM of free_TF_conc, the bound_fraction of all kmers are much closer to each other than at 1nM of free_TF_conc. Such infomation can be used to fit for the absolute affinity of individual TFs.



To figure out the absolute affinities of TFs to DNA ligands using the saturation effect of the top kmers, a batch of SELEX is performed with the following precautions:

_ Ensure that DNA conc is lower than Kd of the consensus, so that the free TF conc can be approximated by the total TF conc. (we used 40 nM DNA in reaction, however, the DNA ligands are random, those with high-affinity motifs should be far lower than 10nM)

_ Involving titration points with higher TF conc than Kd of consensus, so that binding to the best sequences can be saturated

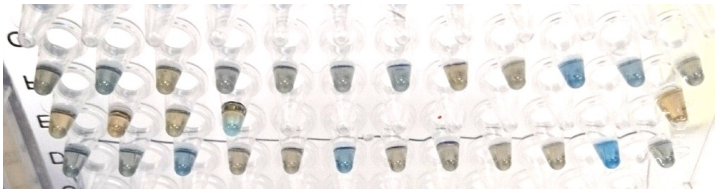
_ Wash is performed under low temperature and the presenece of EDTA, in order to minimize the dissociation and focusing on the binding equilibrium reaction

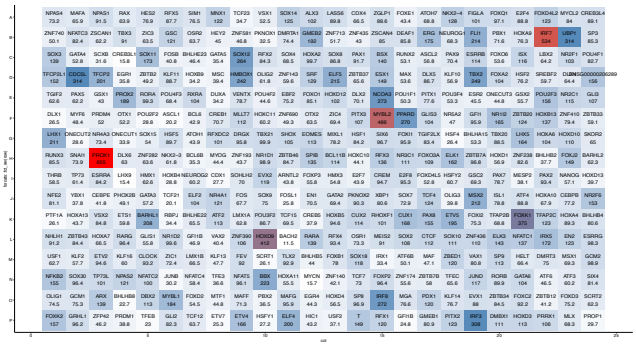
_ Use a short ligand, to reduce the carryover

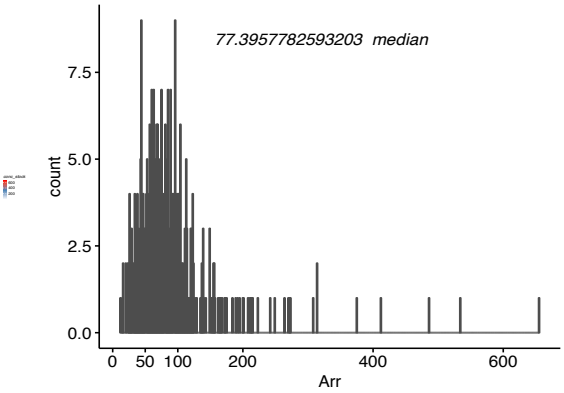
_ Ensure the lowest TF conc is far less than Kd of the consensus, so that no binding saturations occur, where the relative enrichment of different sequences accurately represent their relative affinity (after accouting for the background carryover)

> TF conc measurement

First I expressed 384 TF proteins and measured the concentrations by Bradford assay. The color difference is visible by eye and detected by a plate reader. To obtain more accurate The average stock concentration of TFs is 80ng/ul (histogram). That means in SELEX reaction the highest TF concentrations are on average 20ng/ul, this is approximately 550nM for an average recombinant TF, thus sufficient to saturate the binding of high-affinity sequences.







> Ligand design

Because SELEX ligand with long randomized region has more carryover of the background kmers, in this study, a mixture of short ligands were used, including 8N, 14N, and 30N ligands. As the 30N ligand accounts for the major population, sequencing quality did not drop much after hitting the constant region of the shorter ligands.

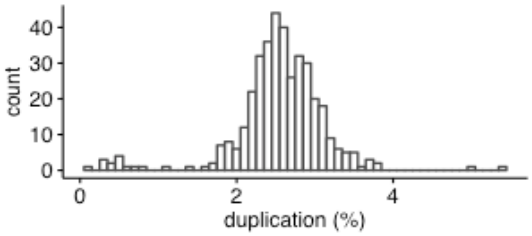
> The serial dilutions of TF and the temperatures

For the SELEX of each TF, 11 serial dilutions were used. We represent the conditions as 2_0 -- 2_10. The 2^0 condition represents the SELEX using original stock solution of the TF. For temperature variations, we tried both 25°C and 37°C.

Conditions:	2_0	2_1	2_2	2_3	2_4	2_5	2_6	2_7	2_8	2_9	2_10
Fold of Dilution:	2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7	2^8	2^9	2^10
Temperatures:	25°C, 37°C										

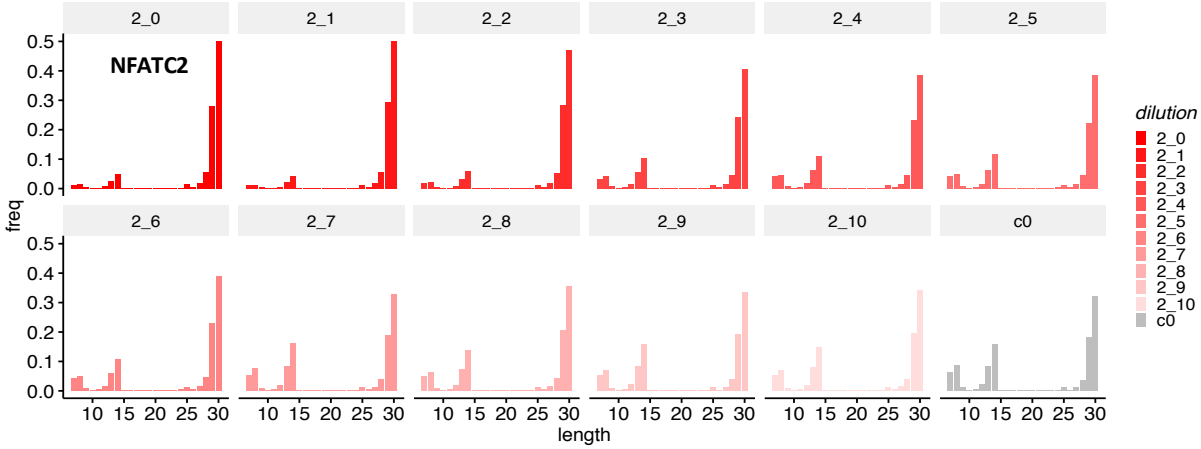
> Low duplication rate in the result library

To figure out whether the duplications need to be removed in the preprocessing, I calculated the duplication rates of the result library for all TFs, using the most diluted samples (2^10 dilution) under 37°C. The 2^10 libraries of all TFs only have ~3% duplication rate, therefore no duplication removal step is required for the preprocessing.



> Distribution of ligand length after selection

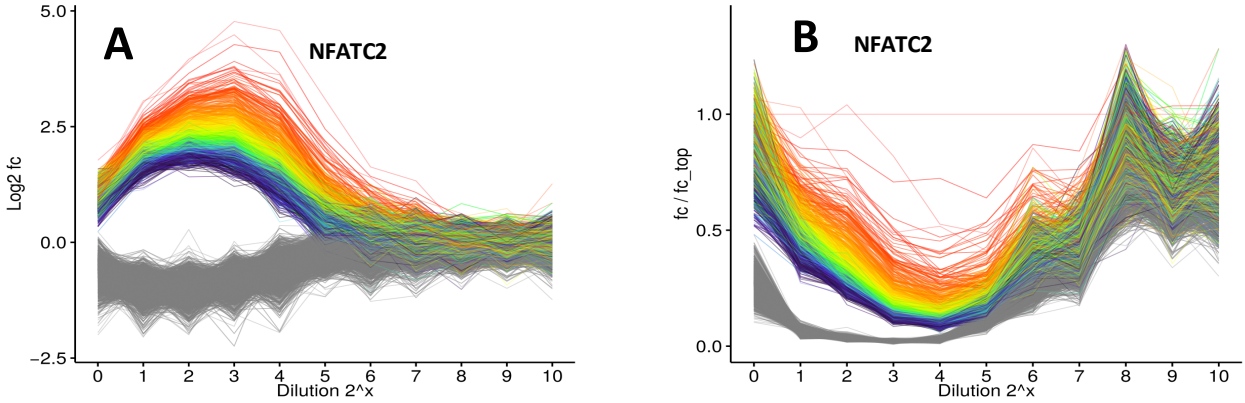
8N, 14N, and 30N ligands are mixed at 1:1:2 (vol) in the input c0 (the grey plot). After SELEX selection, the fraction of 30N ligands increases with the concentration of the TF. This suggest that TF binding is still considerably selective at very high concentrations. In turn, because 30N ligands has much higher probability to contain a binding site than the shorter ones, TF binding will bias and enrich the 30N ligands. For highly diluted samples (2^7-2^10 dilution), the ligand length distribution is similar to c0, indicating that most ligands there are from non-specific carryover.



> Saturated binding of high-affinity kmers at high TF concentrations

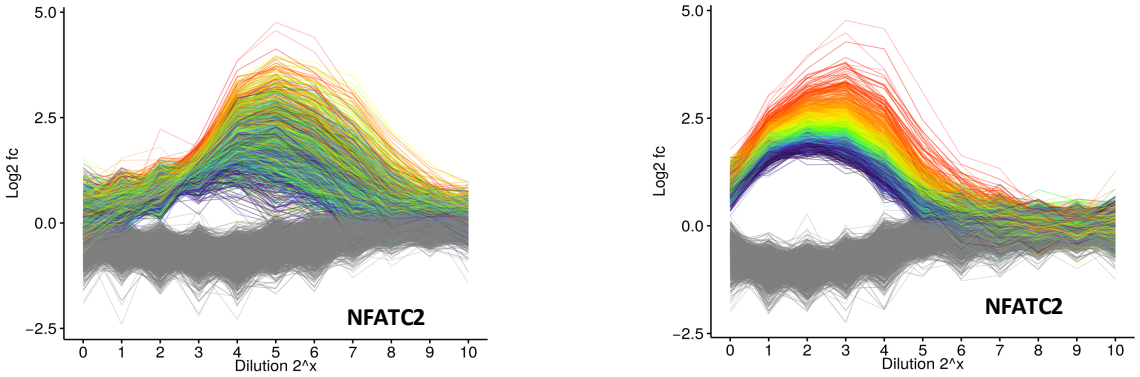
We next examined whether the binding saturation of high affinity kmers is observed in the TF concentration titration. The following figures show an example of NFATC2. In **Fig. A** the log2 enrichment of each 8 mer (compared to c0 input) is plotted for all serial dilutions. The top 3000 most enriched 8mers are assigned with rainbow colors according to their enrichment rank. The tail 3000 8mers are colored grey. In this plot we observed the signal enrichment of NFATC2 for dilutions of 2^0-2^6, and not for further dilutions because all 8mers there have similar extent of enrichment. The binding saturation for high-affinity sequences is observed for 2^3-2^0.

In **Fig. A**, too high or too low TF concentrations both decrease the selection stringency, respectively due to saturated binding and background carryover. As the ratios between 8mers and the consensus will be used to fit for the absolute affinity of a TF, in **Fig. B**, we also plotted similarily for the ratios between 8mers and the consensus. The plot likely suggests the saturation starts from 2^3 dilution, and that 2^4 dilution is the best cycle to estimate the relative affinity between all 8mers, if without background substraction.



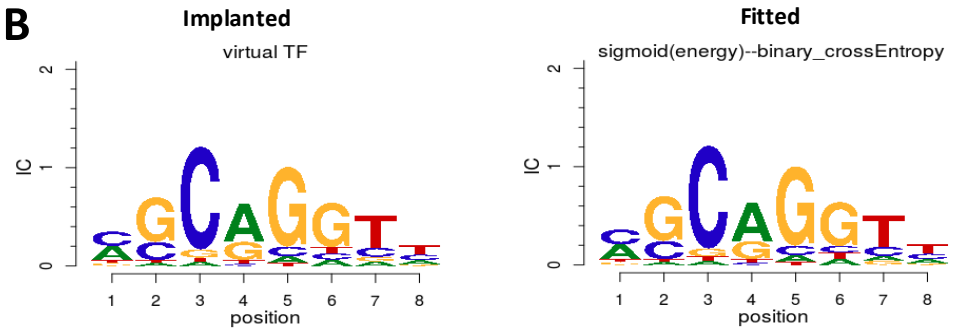
> Saturation starts at higher TF conc at 37℃ compared to 25℃

As expected, most TFs have weaker affinities to DNA at higher temperatures.



> Fitting PWM for relative affinity between kmers

If each base contribute independently to binding, the PWM model represents the relative affinity for all sequences to a TF. Harmen Bussemaker group developed a method to fit physical PWM for the 1st cycle SELEX data (PNAS April 17, 2018 115 (16) E3692-E3701), where the binding probability of a ligand is described by the sum of the binding probabilities across the whole ligand. Similarily to that, we used a neural network model to fit the binding PWM from the 1st cycle SELEX data (for NFATC2 in **Fig. A**). The current neural network model is not yet completely physical due to the inability to tune the loss function. However, the result is quite close, as shown in **Fig. B**, the fitted motif is highly similar to the reads generated according to the physical model using the implanted motif. For the downstream analysis, we assume the fitted motif for NFATC2 here represents the physical affinity.



> Absolute TF affinity by fitting saturated binding

According to the equilibrium equation, the bound fraction of an arbitrary kmer is described by the following **formula A**. Where the Kd of the current kmer is represented by its relative affinity, times the Kd of the consensus kmer as a reference.

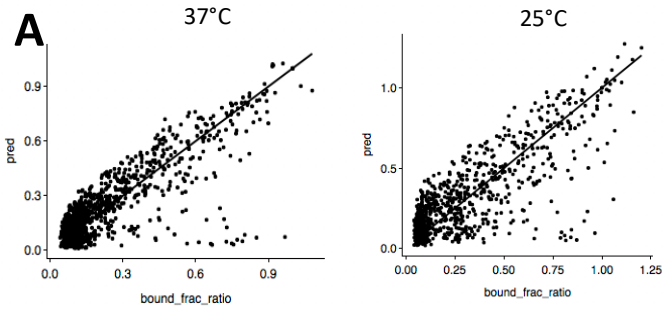
$$Bound\_Fraction = \frac{[TF]}{[TF] + rel\_affinity \times K_d^{ref}} = 1 - \frac{1}{1 + [TF]/(rel\_affinity \times K_d^{ref})} \quad \textbf{A}$$

Then with **formula B**, the enrichment ratio between any kmer and the reference kmer after the binding assay can be related to Kd(ref)

$$\frac{kmer / kmer_{c0}}{kmer_{ref} / kmer_{ref,c0}} = \frac{1 - \frac{1}{1 + [TF]/(rel\_affinity \times K_d^{ref})}}{1 - \frac{1}{1 + [TF]/K_d^{ref}}} \quad \textbf{B}$$

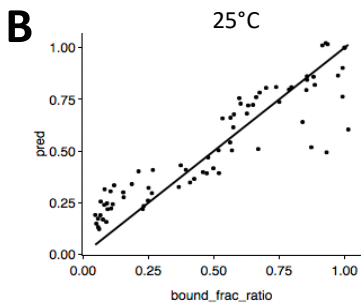
The left side of the function, the enrichment ratios are from direct counts of the sequencing libraries. [TF] is from Bradford measurement. The relative affinities of kmers are calculated from the PWM fitted above. Currently the only unknown variable is the Kd(ref). I used the libraries of 2^0-2^4 dilution to fit for the Kd of the consensus of NFATC2 ("TGGAAAAA"). The fitting is also performed with neural network. Currently it seems more like an overkill to fit such simple model with neural network, however, in the future, theoretically it is possible to fit PWM together with the absolute Kd, for which purpose NN can be of much advantage.

Using the enrichment ratio of all huddinge distance 2 kmers, we fitted the binding affinity of "TGGAAAAA" to be 89nM at 37°C, and 24nM at 25°C. The prediction for low-affinity kmers are less accurate due to the non-specific carry over at the flankings. They have higher abundance than expected. Also, the frame-shift kmers (dots along the x-axis) have large deviations from expected.



Fitted Kd: 89 nM

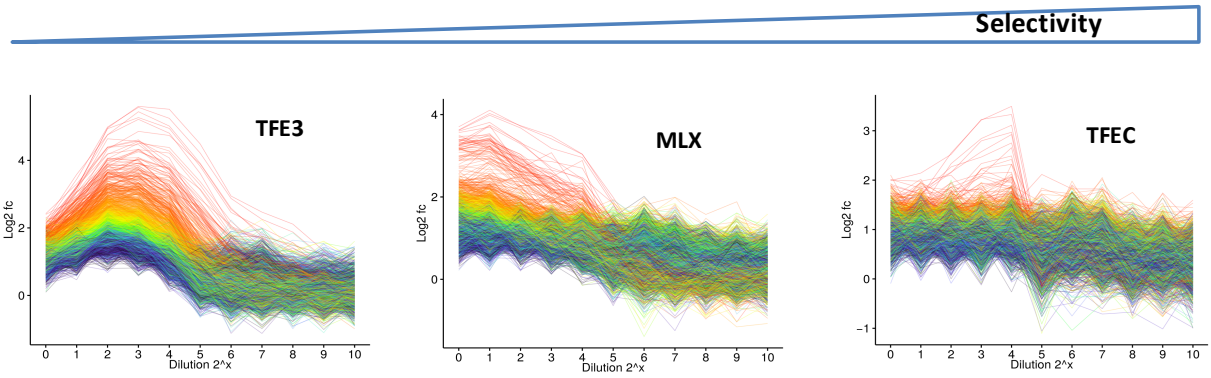
Fitted Kd: 24 nM



Fitted Kd: 23 nM

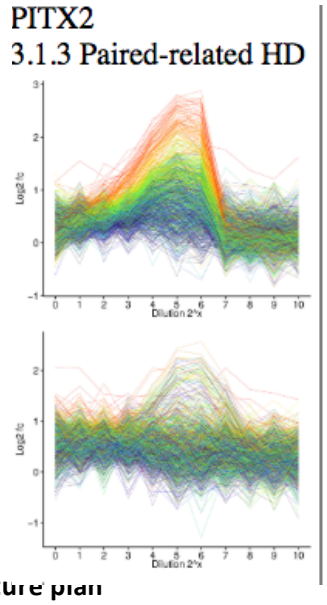
> TFs have different selectivity

The binding assay also allows comparing the selectivity between different TFs, as only the equilibrium binding is allowed and little dissociation is expected in the presence of EDTA. In addition, the single round of selection results in no exponetial bias, the difference between TFs are more reliable.



> Different selectivity at different temperatures

PITX has different selectivity at different temperatures. Its specificity is higher at 37°C and lower at 25°C. More 8mers have higher level of binding than background at 37°C. With the conc titration we know that this is not from the saturated binding of the high-affinity 8mers. Likely a TF conformational change has accompanied the temperature change.



37°C

25°C

> Future plan

> apply the calculation of absolute Kd to all TFs

// not HT because we need to manually pick the cycles used for fitting for different TFs

> figure out a bk model to substract in the fitting

> mechanisms of selectivity differece and its change with temperature

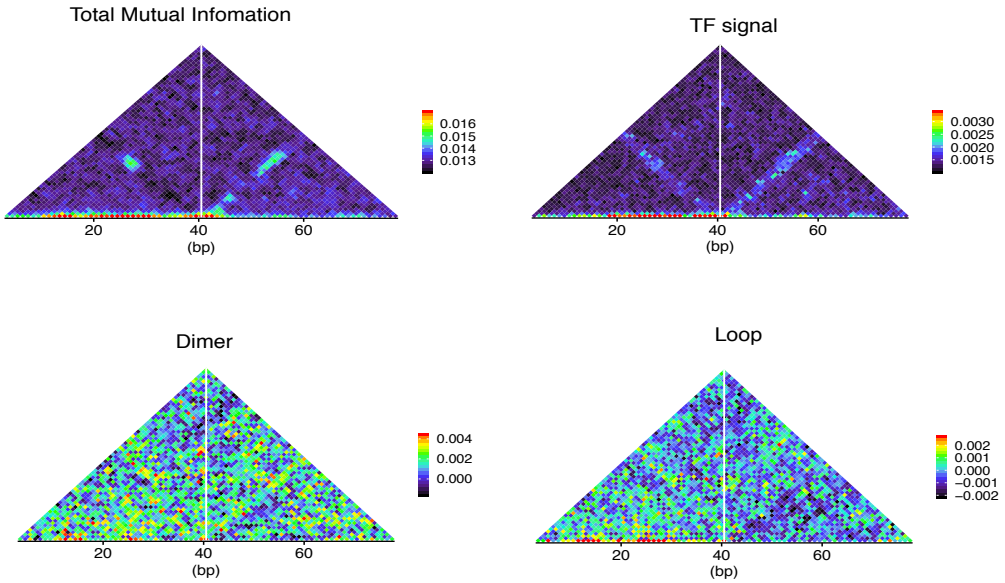
> fit PWM and background (lamda) together with the Kd, they are separable variables

> if the unbound can be collected, math and experiment can be much easier

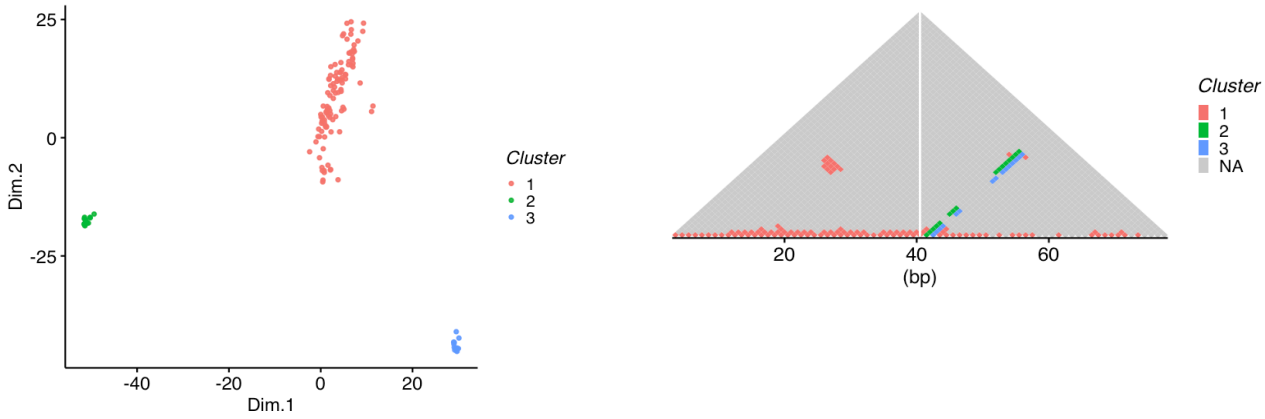
> Analysis for the binding events around TSS

The Starr-seq data produced by Sahu for the first time provided a clean data set to explore the binding events nearby the transcription starting sites (TSS). Teemu arranged and shared the preprocessed file containing the enriched 40+40bp sequeces centered at TSS. Analyses were performed on this data.

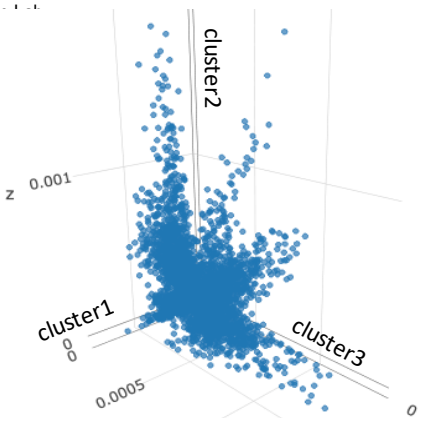
The mutualInfo plot (following figure) of the Starr-seq reads (40+40bp surrounding TSS) shows clear signals at the diagonal, together with additional long-distance cooperations. However, the underlying binding specificities that give rise to the signal is unclear. I tried to add constraints to the kmer-pairs that gives the mutual infomation, but non of the preformed hypothesis explains the MI signal. The MI signal is not from TF binding, dimer binding, or stem-loop sturctures.



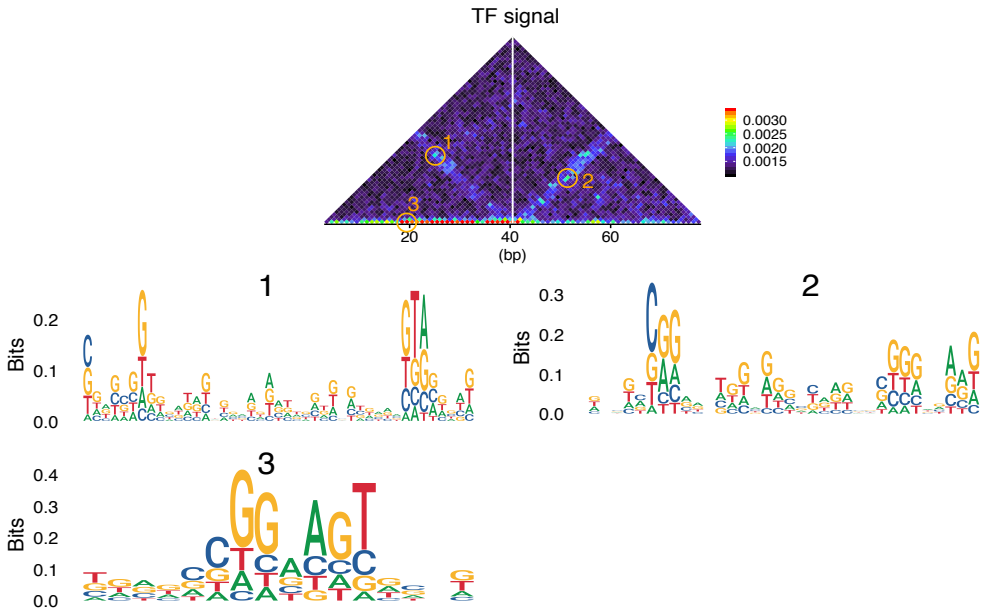
To address this, kmer were counted for all position-pairs with considerable MI signal, the kmer ranks were then used for clustering. The PCA result (left fig) shows that there are 3 distict clusters. When mapped back (right fig), we can see that both cluster 2 and 3 are from the long-distance MI signals downstream of TSS. As cluster 2 and 3 are located next to each other, they likely represent the same well-positioned binding event.



To get more insight, the 6mer counts between the 3 clusters are plotted. The results suggest that cluster1 consists of a mixture of weak specificities, while cluster 2 and 3 consist of stronger specificities. An examination on the preferred kmers in cluster 2 and 3 also suggest that cluster 2 and 3 indicate the same binding event.



However, planting the most enriched kmers as seed for the long-range correlation signals does not converge into PFM, the MI signal would represent binding of other less-specific transcription machineries. In control, the binding events close to the diagonal (position 3 in the figure) gives a stronger PWM suggesting binding of ETS factors.



> Next

> Progress further for the projects above

> NCAP-SELEX with chromatin architecture proteins and TF dimers

_ Dimers of TFs representing typical nucleosome binding modes (including SOX2 and Oct4 for Patrick)

_ ChAHP and CHD4 only as a control

_ HMGB1

_ H1.4

_ HP1

_ ZNF