Group Meeting_Fangjie_3 (2015/11 to 2016/05)

>>>>>>>>> 1. Nucleosome SELEX >>>>>>>>>>>

1-1 EMSA-SELEX for nucleosome

> Infomation content for the enriched reads

The infomation content of kmers from different SELEX cycles were checked

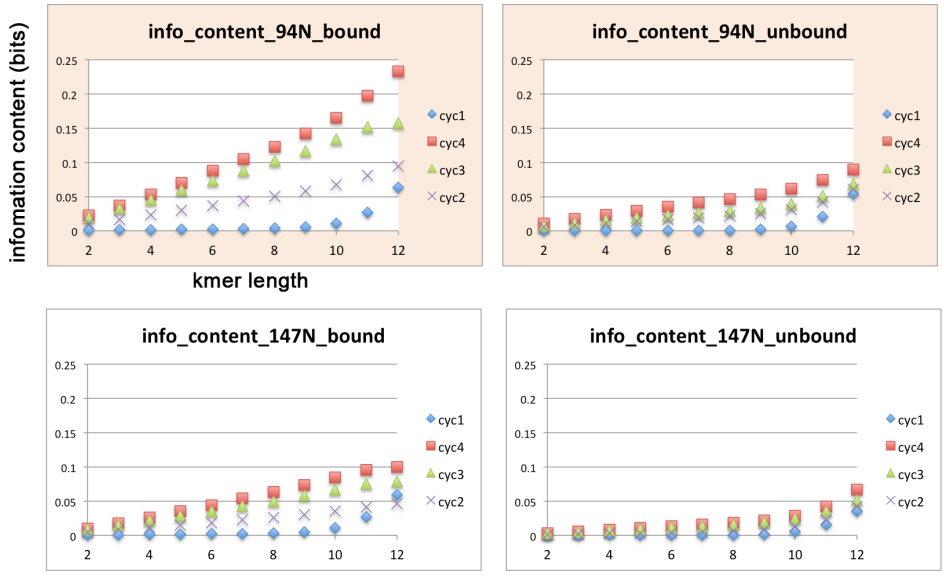


Fig1 info content of enriched libraries of nucleosome SELEX

_the short ligand (lig147, 94N, orange bk) is enriching more signals than the longer one (lig200, 147N, white bk);
_bound is more selective than unbound;
_cycle1 do not have much enrichment (Leena also mentioned cyc1 data worked worse than other cycs for her analysis), the info content rise when k>10 for cyc1 is more likely due to the noise when the count is not enough, the most informative kmer length seems to be around 10;
_maybe we can use ligands with even shorter random region to get higher selectivity.

I am a bit doubtful whether the bias of phusion can play an important role here. Because the infomation content of 147N_unbound should include the bias caused by phusion enzyme. However, it is still small compared to the enriched real signal of bound (94N_bound).

May be worth to try this analysis with gapped kmers to get some hint about the recognition mode of nucleosome.

> data analysis by Leena and Fan

Leena and Fan used the EMSA-SELEX results to map the genome and tried to figure out the binding possibilities for each genomic position. The SELEX data showed predictive power in their analysis. But some mysterious phenomena is yet to be understood:
_why introduing unbound data weakens the prediction power of the model;
_why cyc2-4 of the bound data show similar prediction power.

> scale down of the reconstitution reaction for highthroughput

The original reconstitution protocol for nucleosome uses a lot of materials (1µg DNA + 1µg Octamer). I failed a couple of times when I tried to use less octamers but similar amount of DNA. It turns out that the stoichiometry of DNA to octamer is important. I suppose too diluted histone components will be kinetically hindered to form correct octamer. But Lucas suggested the histone should already be in the octamer state even with 2M salt, so the reason is still unknown.

1-2 CAP-SELEX to study the interaction between TF and nucleosome

With Jianping's help, I set up the CAP SELEX protocol on biomek384. 384 TF DBDs were tested for both lig147 and lig200.

> workflow

_in order to get some signals as positive control first, this time I added TFs after the first purification of reconstituted nucleosome;
_a plate of TF monomer SELEX were also run with lig147;
_obtained reads were first filtered to remove primer dimer and wired concatanations.

DNA ligands + Histone octamer >>> nucleosome >>> purify for nucleosome (SBP-tag beads) >>> purify for TF (His-tag beads) >>> amplify enriched ligands for seq and next cycle

(add TFs and incubate)

> PHP based interface to view all related data of selex plate

As data from different analysis has been generated for each SELEX well, it becomes very inefficient when I am checking multiple types of data for one well to figure out what is happening or trying to find correlations between. I created a few PHP page which integrates infomation from our standard pipeline, and is also capable to display together for each well additional custom image features and text features, in a view of 384/96 plate .

> background of sticky nucleosome

Washing and binding for TF monomer ctrl and for the CAP-SELEX of lig200 were performed using 1x promega buffer (50mM NaCl, 1mM Mg) at the beginning. Then I noticed the high background of nucleosome in CAP-SELEX of lig200 (Fig2a, middle), so washing and binding for the CAP-SELEX of lig147 was performed using 1x low stringent buffer (150mM KCl, 2mM Mg) with 20mM imidazol. But it turns out to make the S/N ratio even worse (Fig2a, bottom), but if we think positively, this result can suggest that a higher nucleosome occupation is excluding the binding of TF. To verify I am planning to run CAP-SELEX of lig147 in the same buffer as lig200.

I tried to qualitatively evaluate the signal strength of nucleosome for each well as follows:

Fourier transformation >> baseline substraction >> calculate mean peak height corresponding to 10.2bp periodicity of 1-nt or 2-nt

I apply a threshold to call wells with considerable nucleosome signal and plotted in Fig2b. The nucleosome signal of lig147 is much stronger than lig200. Probably due to the use of more salty buffer. Some wells in TF monomer ctrl also showed clear periodicity (Fig2b bottom).

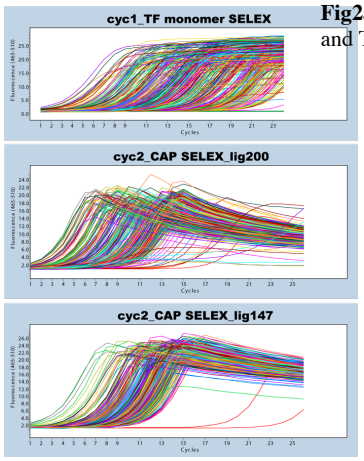


Fig2a qPCR data for CAP-SELEX and TF monomer ctrl

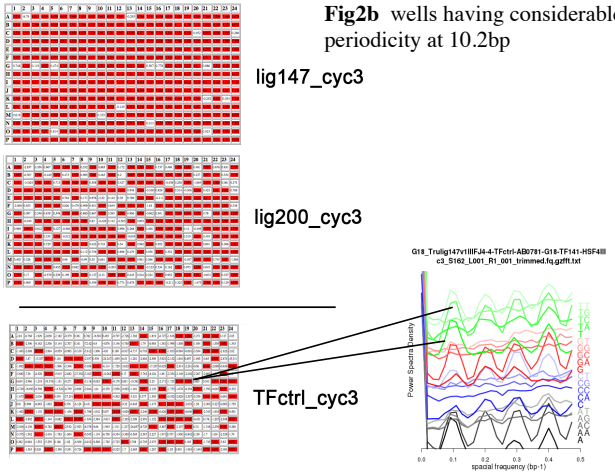


Fig2b wells having considerable periodicity at 10.2bp

The ratio of wells that enriched TF signals is plotted in Fig 3a

_less wells developed good TF signals for nucleosome CAP-SELEX than for TF monomer ctrl;

_more wells developed week signals for lig200 than for lig147, the narrower dynamic range in the qPCR of lig147 (Fig2a, bottom) may account for this.

To check whether the success of SELEX is related to higher background (of course we hope not so that we can correlate the success to the effect of nucleosome binding with less noise), the Cp values (mean of cyc1-4 qPCR data for each well) were plotted according to signal strength. Regretfully, as Fig3b indicates, it seems the background is affecting the success of SELEX.

_Wells with no TF signals have larger Cp values than wells with signals, suggesting the TFs pulling down less DNA more likely get their signals overwhelmed by the background;

_Wells with weak signals have smaller Cp values than wells with middle or strong signals, and have a larger standard deviation. This can be understood by thinking that the "weak" group actually consists of two populations. 1st population consists TFs having high specificity but low affinity (or simply with lower protein concentration), so their signal is partially buried by background; 2nd population consists TFs having low specificity but high affinity (or simply with higher protein concentration), which pull down lots of DNA without discrimination.

Fig3a overview of the successful rate

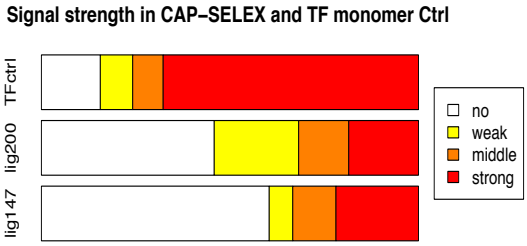


Fig3b correlation between qPCR data and TF signal strength

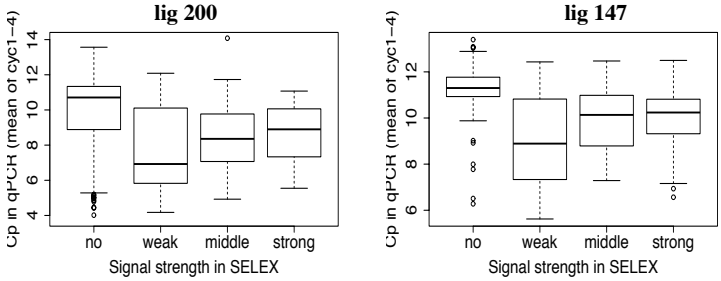
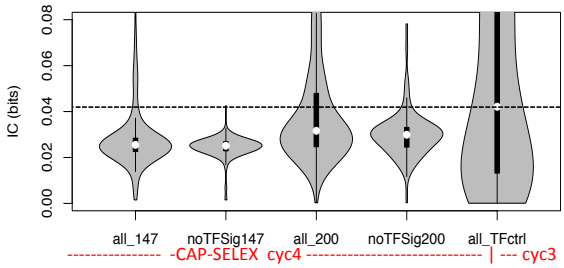


Fig4 presents the IC for the CAP-SELEX and the TF monomer ctrl:
_Selectivity in TF ctrl is much higher than in CAP-SELEX, IC of cyc3 of TF ctrl is already larger than IC of CAP-SELEX of cyc4.
_Contrary to the observasion in EMSA-SELEX (Fig1), IC for lig200 is slightly higher than IC of lig147 here, possibly due to the narrower dynamic range (thus less prominent TF signal) in the CAP-SELEX of lig147 (Fig2a). It can also be explained by the more complete occupation of nucleosome on lig147 which excludes TF binding.
_The IC developed in CAP-SELEX principally comes from the specificity of nucleosome. Because the overall IC is almost the same as the IC of the wells without TF signal. This can also be explained either by the exclusion of TF in the presence of nucleosome or just by the higher background.

Fig4 IC for CAP SELEX and TF monomer ctrl



If the relative affinity of TF classes is affected by the presence of nucleosome, and also assuming the TFs are not saturated with weak binders after wash, their Cp values in qPCR should change order when comparing cases in the absence and the presence of nucleosome.
I classified the TFs with their Cp values in the smallest 20% of each SELEX, to see if the classification changes between the absence and the presence of nucleosome. The result (Fig5) showed little effect. The HMG class (arrow) seemed to bind more with nucleosome, but still very unlikely since here is a multiple comparism.
I will fix the binding and washing conditions for both TF ctrl and CAP-SELEX in the next run, and if we pull down TF first in the CAP-SELEX and take an aliquot for qPCR to compare with the Cp of TF monomer ctrl, we should be able to figure out the absolute affinity change of each TF to DNA between the presence and the absence of nucleosome, which could be more interesting.

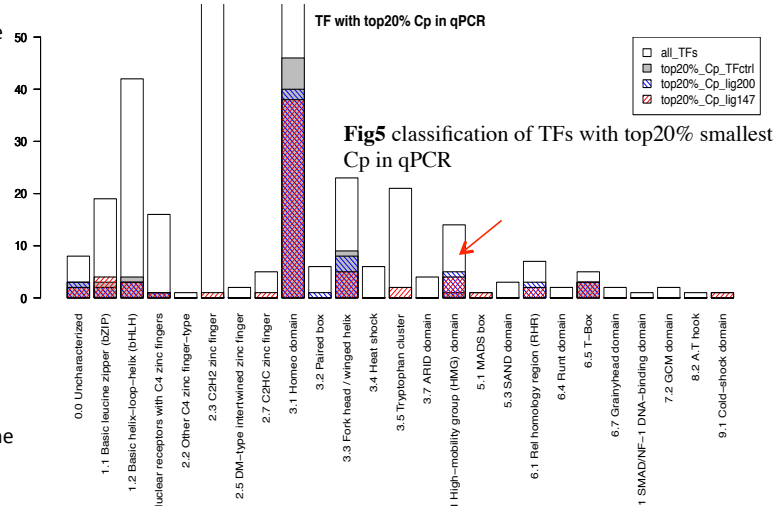


Fig5 classification of TFs with top20% smallest Cp in qPCR

> Signal changes to look at (suggestions are welcome)

>> First we can look at the localization bias of TF motifs on the whole ligand to check the preferential positions of TFs.
Workflow: For all TFs, first the corresponding PFM from Yimeng and Arttu's guide files were picked. Then the motifs were mapped against the uniq reads with MOODS (p=0.0001) to record all hit positions. Finally using the mapping result to plot. A typical plot for successful wells is shown in Fig6, with a failed one aside. It would be informative to further plot a track of TA (or other oligo nt) waves alongside to keep track of the phase of each position, and also to assign the hits with a position in its middle rather than at its head.

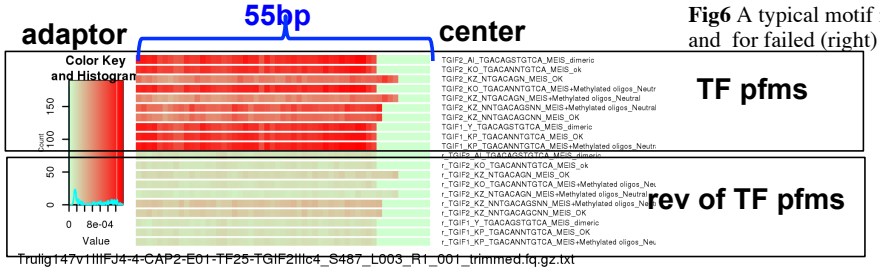
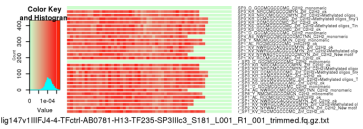


Fig6 A typical motif mapping plot for successful well (TGIF2, left), and for failed (right)



Because this batch is not yet sequenced for PE, only information of 55 bp is available on the map

A quite significant phenomenon to notice was that, for most successful TFs (especially the Homeodomain TFs), lig200 seems to enrich more motifs near the adaptor rather than near the center of the ligand (Fig7). Probably the higher relative occupancy of nucleosome at the center of lig200 may accounts for. Whereas the occupancy for lig147 could be more homogeneous.

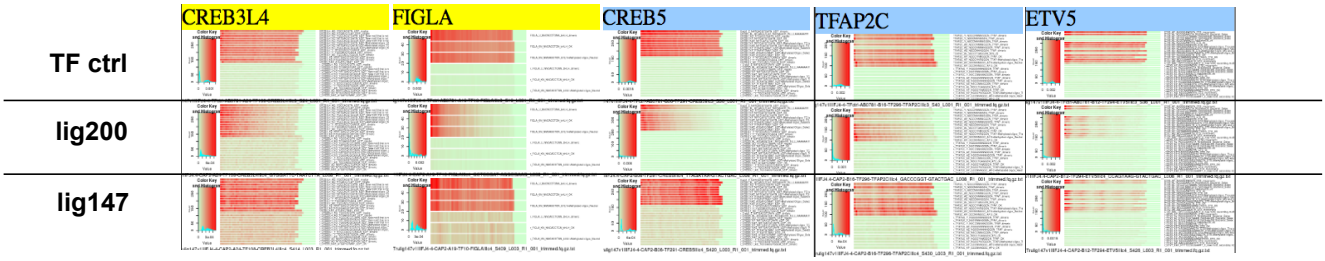


Fig7 more motif hits near the adaptor of lig200

But a few closely related POU factors, and bHLH-ZIP factors seem to result in more bias toward the adaptor for lig147 (Fig7).

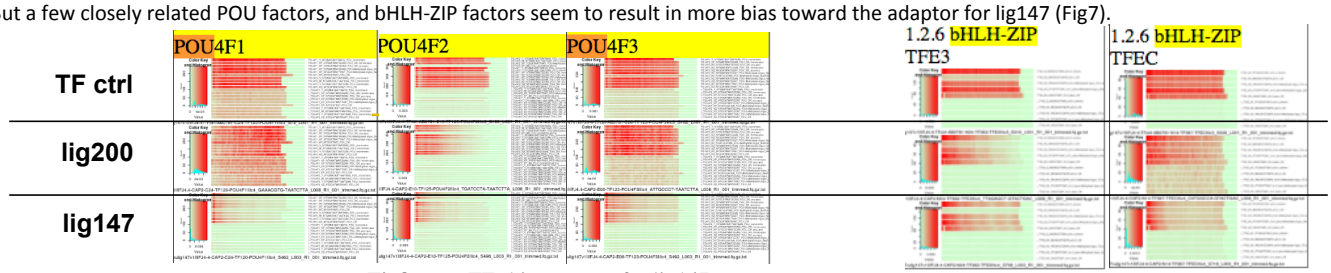


Fig8 some TFs biases more for lig147

We have been expecting to see TFs prefering specific phases of nucleosome DNA. This indeed happens (Fig9). Such preference may be partially explained by the nucleosome selection that enriches periodic ligands (thus with biased subsequences at different phases). As shown in Fig9b, the top panel, when mapping the TF motifs against nucleosome-favored ligands, we already can see some contrast of a similar pattern. However, the contrast of the periodic pattern became much more distinct for the ligands favored by CAP-SELEX (Fig9b, bottom panel). Therefore, part of the contrast should be contributed by other factors, conceivable factors are something like the periodic steric hiderence or the periodic bending direction change of DNA.,

Such preferences of phase is seen for most of the Homeodomain factors and some Forkhead factors (FigS2b,c). Also, some TF familys like Ets and HD-PROS already show a periodic preference pattern in the absence of nucleosome (FigS2e,f). The adaptor of our SELEX ligand should somehow participated into such patterning, probabaly by dictating the position of one binding event, then subsequent binding events can be biased by the allosteric effect. However, nucleosome still seem to strengthen the contrast of the patterning for Ets factors (FigS2e).

Fig9a

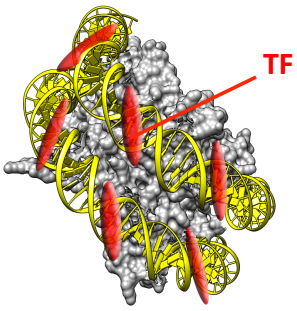


Fig9b

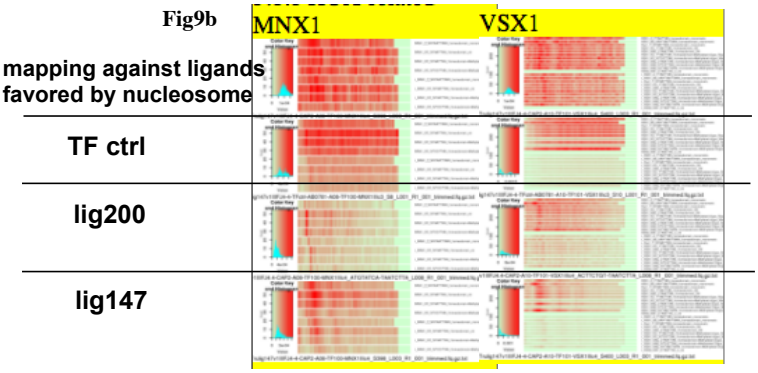


Fig9 TF binding with preferred phase

We also expect some regions of the nucleosome DNA to be specific (e.g. near the dyad axis), either due to the surface of nucleosome can provide additional interactions (Fig10a), or due to the inhomogenous looseness of DNA wrapping, or to that the nucleosome DNA can be more easily single-stranded at specific points (Anderson et al., PLosOne, 2010). In this batch we observed RFX5 and SOX11 (Fig10b), and probably also SOX12 and TBX4 (FigS3), to have such preferred regions of binding. The binding motif of RFX5 is also changed (Fig10c). The signal development of them are shown in FigS3. Hopefully we can see more region preferences with full length TFs, especially for those TFs which have low affinity to nucleosome by their DBD. If their affinity would raise in full-length, then it probably would suggest they managed to find an anchor point somewhere on histones. TFs with chromatin-modifier-like domains should be paid particular attention.

Also, we have to develop an approach to discover the position-specific binding for TFs with changed specificity.

Fig10a

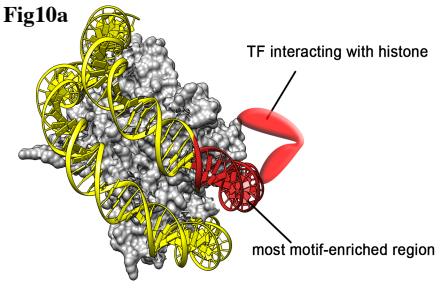


Fig10b TF binding with preferred region

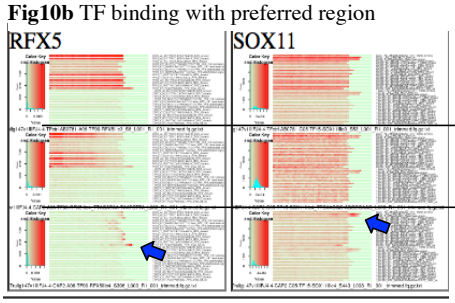
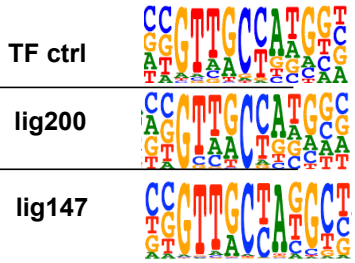


Fig10c motif change for RFX5



Mapping with previous motifs only addresses the reported binding modes. In order to capture all potential bindings, I checked the positional IC (of 4mers) accross the ligand. As the binding by nucleosome only gives a flat background (Fig11a), any region with IC peaks is more likely to have contacted with TF. If we find regions with IC peaks but not revealed in the TF motif map, then the TF should have invoked other specificities for such regions.
For lig147, many TFs have a flat IC profile similar to Fig11a. Some developed peaks at the flanking and agreed with the motif map (Fig11b). Homeodomain, Forkhead, E2A and a few other TFs developed a periodic pattern (Fig11c), where the IC peaks correspond to the AA/TA/TT di-nt frequency peaks (major groove facing outward). This IC pattern accords with the motif pattern for some of the TFs (like VSX1) but not always (TLX2, TCF12), so should contain additional specificity infomation. I checked a few of them, and found that the peaks are biasing GT_nA_n kmers (Fig11d), which is in accord with the original preference of nucleosome but just the bias is much stronger than without TF (Fig11a), thereby raises peaks. The reason why TF binding invokes such IC peaks is yet to be understood.

Fig11a IC map for nucleosome signal has a flat bk

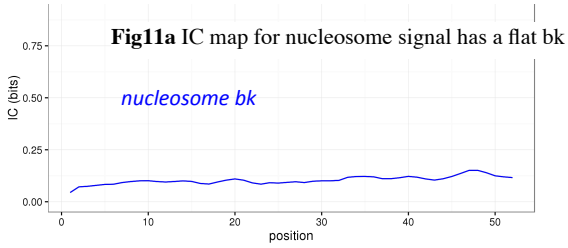


Fig11b IC map & motif map biasing flank

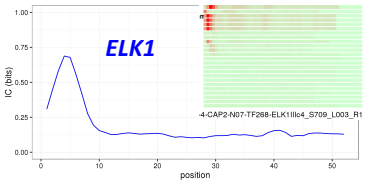


Fig11c periodic IC maps (E.g. Homeodomain, Forkhead, E2A)

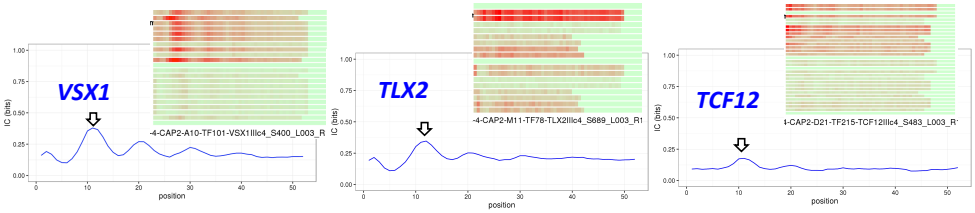


Fig11d mostly biased kmers at IC peaks (arrow)

GTTTTT
GTTTTA
TTTTAA
TGTTTT
TTTTTA
GTTTAA
TTTAAA
CGTTTT
...

GTnAn

The motif mapping approach is sensitive but do not consider novel binding specificities. IC map captures all specificities but seems not sensitive enough (many TFs with clear signal pattern in motif mapping still give flat IC profile). Probably mapping with the actually enriched kmers would do better. Fan helped me pick out the "real" enriched kmers from each well, free from the frame-shift background of the most enriched kmers. We can also try to map with these kmers.

The motif-mapping and IC map patterns seem to be pretty conservative for each TF family. Of note is the SOX-related TFs (FigS2d), they enrich their motifs homogeneously in TF monomer ctrl, but developed very specific pattern on lig200 (not seen for other TFs), and the pattern disappeared again on lig147 (or failed). Could it be that the SOX proteins are favoring some specific positions next to but not occupied by nucleosome?

>> Secondly, we can check the affinity change of TFs to DNA upon the binding of nucleosome. As discussed above, we can use qPCR to address this issue by using the same amount of starting materials (nucleosome-bound DNA and free DNA) in a paralleled run.
In addition, the successful rate of each TF family in CAP-SELEX may also offer some hints to their affinity change. E.g., Tal/Fos/Jun-related bHLH/bZIP factors (FigS4, cf FigS2a,d,g for mapping results) that are successful in TF ctrl generally also succeeded in CAP-SELEX of lig200 (FigS4a), but almost all fails in CAP-SELEX of lig147 (FigS4b). Whereas other families have higher success rate.

>> changes on binding specificity
A few wells do developed obviously different specificity, whereas most of the minor change remains tricky to be called a significant one. I am thinking to check the kmer-kmer plot for wells of strong signals and see if anything pops up. We can also perform motif discovery respectively for the flanking and for the center part of CAP-SELEX libraries, and see if the outcome differs (because many TFs severely biases the positions near adaptor, but probably their weak enrichment near the center actually contains new motifs).
It has been reported that pioneer TFs may target nucleosome DNA on partial / degenerated motifs. Like SOX2 will have its "G" position became degenerated thereby imposes less distortion on DNA when binding to nucleosome. We have Sox2 DBD in this batch but the dimer motif with a distinctive "G" is still of a high count.

>> cross-strand binding
Grooves of the 2 DNA strands wrapped on nucleosome are aligned. This may facilitate a cross-stand binding for potential TFs (Fig12a). An ideal TF with long and rigid DNA recognition motif should strongly favor this cross-strand mode, but probably does not exist. Maybe it is more reasonable to expect cross-strand bindings by TF dimers. Katja has suggested me with some multidomain TFs and TFs forming dimers independent of DNA (e.g. E2F7&8, FOXP, TEAD, ZNF, OneCut2&3, PAX).
About the data analysis, I am thinking to look at the correlation of kmers at one site and after ca.70bp on the same ligand.

>> inactivation of binding modes requiring long available major groove
As represented in Fig12b, due to the steric hinderance when DNA major groove faces histone, the available major groove is not consecutive. If a TF (or a rigid TF dimer at protein level) requires long, consecutive major groove to bind, it will be more likely for such TF to change binding affinity or enrich different binding modes in the presence of nucleosome. I spotted a few TFs to enrich less dimer in the presence of nucleosome (Fig12b). The failure of Tal/Fos/Jun-related TFs (FigS4) for lig147 seems to also support such hypothesis as they are obligate dimers. But not explained is that bHLH-ZIP still has a high successful rate.

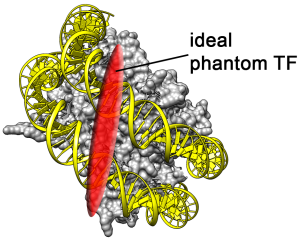


Fig12a cross-strand binding

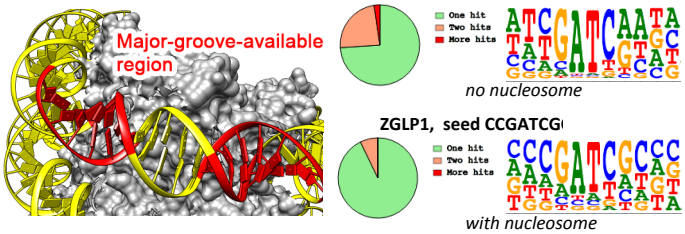


Fig12b non-consecutive available major groove

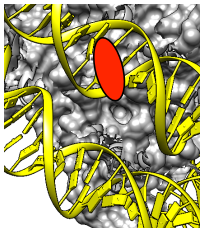


Fig12c break of symmetry for the 2-fold axes of DNA

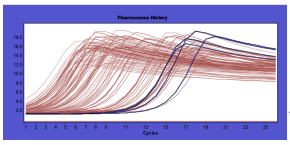
>> strand-selective binding
Fig12c indicates the original 2-fold rotation axis of DNA. We can see that in the presence of nucleosome, it is no longer symmetric for the regions with 2 parallel ds-DNA helices. Bulky TFs or TFs that can make additional contact with the other DNA helix are likely to enrich more motifs on 1 strand than the other, for each half of the lig147, and average out when summing the 2 halves together.

>> different binding impairment for TFs with A/T-rich motifs and G/C-rich motifs
When wrapped to nucleosome, high-affinity sequences are having their major grooves facing outward for AT-rich segments, and their minor grooves facing outward for GC-rich segment. Thus TFs having AT-rich motifs may have more advantage to bind to nucleosome. It seems to me that TFs having GC-rich motifs are showing weaker signals or having signals more biased to the very flanking regions of lig147, but further systematic examination is required.

> Future plan
_To address the above-mentioned potential changes of TF binding in presence of nucleosome.
Of note is that all these signals could be buried into TF monomer signals if the nucleosome is not fully (or not tightly) wrapped. It might be better to first run 4 cycs or more of nucleosome monomer SELEX with lots of starting materials to ensure the library complexity, then based on this library run CAP-SELEX with TFs for additional cycles.

1-3 HT nucleosome monomer SELEX (with methylation)

Only the lig147 (94N) was used here because the signal is stronger (as suggested by Fig1).
Making advantage of the nonspecific binding of nucleosome to Ni beads, SELEX with original octamer can also be run in high-throughput. As shown in the qPCR result (fig to the right), the S/N ratio and the dynamic range looked good.



> well design
5 cycles of SELEX were run for 96 wells of nucleosome monomer SELEX, with a methylated copy and an unmethylated copy, designed as follows:
_**basic design** (nucleosome + DNA at different stoichiometry, with or without 2mM Mg2+);
_collecting the unbound and the aggregated precipitation by EMSA for basic designs;
_with competitor ligands of different concentrations and of different strength;
relative strength of the competitors: widom_601 >CAG> random
_methyl+ and methyl- TFs (ATF4 and LHX9) as control;
_background controls (no protein)

Methyl SELEX was run according to Yimeng's protocol but without the final cycle that compares the motif strengths. Because I did not have online barcodes with my Truseq type ligand, which has been paralleled with illumina index. But now I already ordered the online-barcoded ones and should be able to run it the next time.
Another problem for the methylation SELEX is that the ligand compatible with bisulfite conversion requires custom primers to sequence. Adding read1 custom primers is hard for Hiseq 4000, and adding the read2 custom primer is hard for both Hiseq 2000 and 4000 (requires to sequence a whole flow cell at once). I hope to order 2 PCR primers for standard Truseq library, but with all Cs methylated (cost a bit more than 10k SEK) , then we do not have to use ligands that requires custom primer, and do not have to sacrifice the infomation from 1 of the 2 strands.

> positive control

Both LHX9 and ATF4 showed methylated motif but not very strong (the following figure), ATF4 failed for the normal ligand.

> nucleosome signal

The signal and complexity for most wells look pretty good (Fig13). This time only SR 55bp is sequenced, may be worth to pick a few wells to sequence deeper for PE. But the unbound data have weaker signals than the previous EMSA batch (I haven't carefully titrated histone/DNA ratio for unboud this time).

	6mer IC	peakHeight of FFT
this batch	0.043	22
EMSA_library	0.067*	14
SBP nucleosome in CAP-SELEX	0.027	17.9

* higher mono-nt bias due to DNA purification every round?

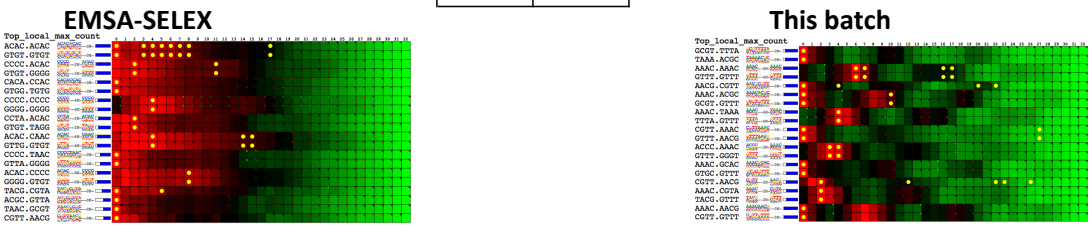


Fig13 comparison of the batches of nucleosome SELEX

Adding strong competitor (widom601) actually resulted in worse signals (Fig14 left). Probably becuase of the incurred complexity bottleneck in early cycles. However, adding medium (CAG sequence) or weak competitor (random) did strengthen the signal. Interestingly, methylation decreased the specificity of nucleosome (Fig14 right).

Fig14 competitor and methylation effects on IC

> slightly methyl-minus nucleosome

We can see from the k-k plot (Fig15) that kmers enriched with normal and methyl ligands is kind of correlated. The plot also suggests nucleosome is slightly methyl-minus, which corresponds to the observation that nucleosome-associated DNA displays less methylation. I found 2 CG-containing motif (Fig15 b,c) enriched only for normal ligands, they are weak but robust (identifiable difference in most of the 32 pair of methyl-nonmethyl conditions). There are also CG-containing motifs (Fig15 d) that enriched in both normal and methyl ligands.

Fig15a. k-k plot

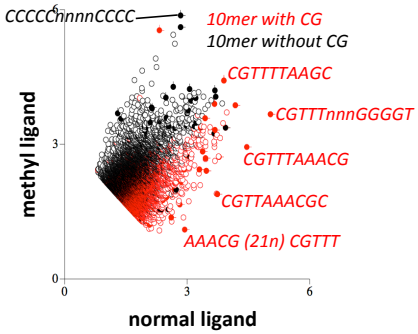
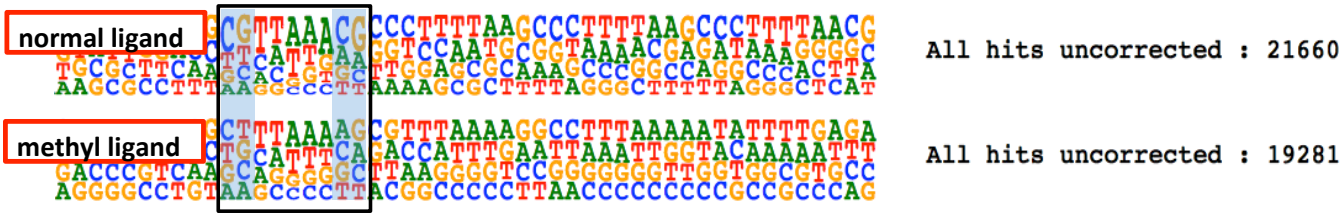


Fig15c. seed CGTTTnnnGGGGT

Fig15d. seed GCGTTTTAAG

Fig15b. seed CGTTWAACG



It seems nucleosome favors CG-containing subsequences when they are not methylated (Fig16a left). But such selectivity disappears when the CGs are methylated (Fig16a right, the poping out CG-containing branch is still robust, however). These k-k plots seem to suggest the efficiency of methylation is not bad because the bias toward CG-containing susequences is quite completely supressed for the methylated ligand (Fig16a, right). Comparison of unbound or precipitation between normal and methyl ligands (Fig16b) revealed little difference. Although difference is expected between the unbounds, probably the low S/N ratio for unbound selection rendered it invisible.

Fig16 a normal vs. unbound normal & methyl vs. unbound methyl

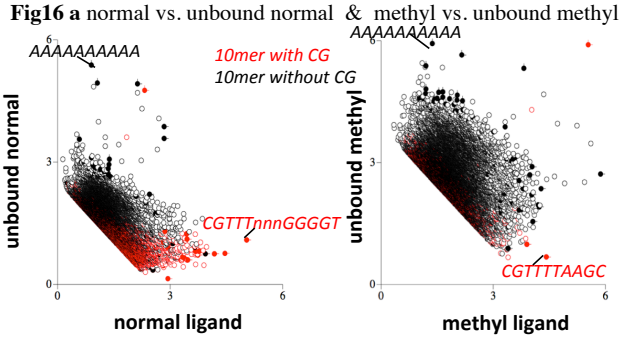
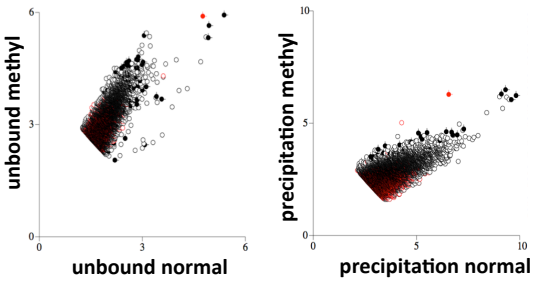


Fig16 b normal vs. methyl ligands (for unbound and precipitation)



The phase-amplitude plot of all Fourier-transformed dinucleotide signals (Fig S5) suggested an overall phase-shift caused by methylated adaptor. We observed a decrease of CG's amplitude in the presence of methylation. As the amplitude has been normalized against the counts, it should be suggesting the decrease of CG's periodicity rather than the decrease of its absolute count (but still need verification). There are also some minor but robust effects by methylation like the shift of the relative phase of CC, and decrease of the relative amplitude of TT.

In the absence of Mg2+, the signal of nucleosome is stronger, and the DNA pulled down is much less. However when I checked the fraction of DNA reconstituted into nucleosome, it was similar to that in the presence of Mg2+. Probably in the absence of Mg2+, only well-packed nucleosome particles can bind non-specifically to Ni beads.

>>>>>>>>> 2. HT measurement of TF binding kinetics and thermodynamics >>>>>>>>>>

Recently I have been concentrating on the nucleosome SELEX and haven't had much time to look at this project. But I found a recent work (Riley and Bussemaker, 2015, eLife), which fits an energy-based model to PBM results, can also be adapted to quantitatively fit our SELEX data.

affinity of a subsequence

$$y(S) = \beta_0 + \beta_1 \sum_v e^{-\Delta\Delta G(S_v)/RT} \qquad \Delta\Delta G(S) = \sum_{\varphi \in \Phi(S)} \Delta\Delta G_\varphi$$

Intensity — bk — const — sum over all subsequences — sum over all 1nt and 2nt features

They also corrected for the localization of the motifs on PBM ligand; probably not necessary for SELEX ligands. However, we have to add correction for >=2 occurence of motif on a single SELEX ligand. Because multiple binding in PBM only causes linear increase in signal, but in SELEX it will cause much more significant enrichment for such ligands.
I should be able to design a model for our HT quantitative measurement, but still it would be nice if I may get help or advices on the fitting process of the model.

> primary analysis of HT-dissociation data

With the data from qPCR and sequencing, I tried to plot the dissociation curve for a few kmers of different abundance in cyc3 SELEX library. The plot gives a quick overview and helps to get idea of the appropriate background model. Ideally we expect kmers with low abundance to have a faster dissociation. As the dissociation of a ligand is dictated by the strongest kmer on it, counting all kmers on each ligand for the dissociation plot (Fig17 left) will bring in much background, rendering the dissociation of all kmers similar. I tried to count only the strongest kmer for each ligand, then the plot became discriminative for kmers of different abundance (Fig17 right). Fan has been developing a program to better address this issue, and looking at dissociation of ligands with multiple binding sites. I will also explore the background model experimentally, by mixing ligands from different cycles and test their dissociation.

Fig17 dissociation curve with different kmer-counting strategy