Group Meeting_Fangjie_4 (last one in 2016-06)

> DNA bending related

> bending persistency analysis of nucleosome-favoring ligands

The purpose is to test whether nucleosome-favoring ligands are intrincically bent into a loop, that is, whether the bending direction of its constituent segments are consistent.

Using a 3rd-order markov model, the derived average sequence from the bound-library seemed to bend persistently (Fig1B, left), so we decided to test the bending of all individual sequences in SELEX libraries.

The python script from http://www.lfd.uci.edu/~gohlke/dnacurve/ predicts 3D-coordinates for all base-pairs within a single sequence of DNA. It is modified a bit to work with all ligands in SELEX libraries. Specifically, the MAXDIST (explained in Fig1A) for each ligand is calculated in the SELEX libraries. It makes sense as a qualitative estimate of the extent of persistent bending (Fig1B).

Fig1A

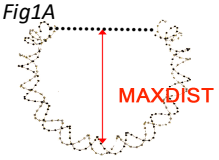
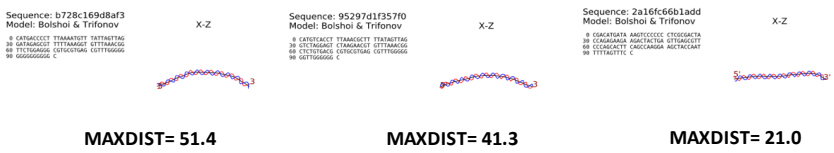


Fig1B



MAXDIST= 51.4 **MAXDIST= 41.3** **MAXDIST= 21.0**

The distribution of MAXDISTs across the cycles of the bound library is plotted in Fig2A. No significant change of MAXDIST is observed across cycles. The observed decrease could be due to the developed short-nucleotide bias. Comparing the MAXDISTS of the bound libraries to the unbound librarys (Fig2B) also showed little difference.

Fig2A

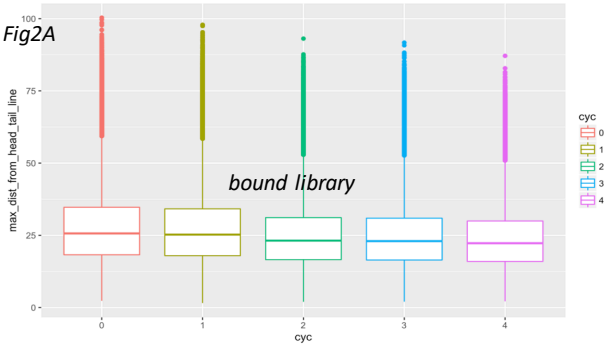
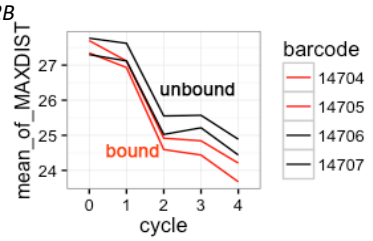
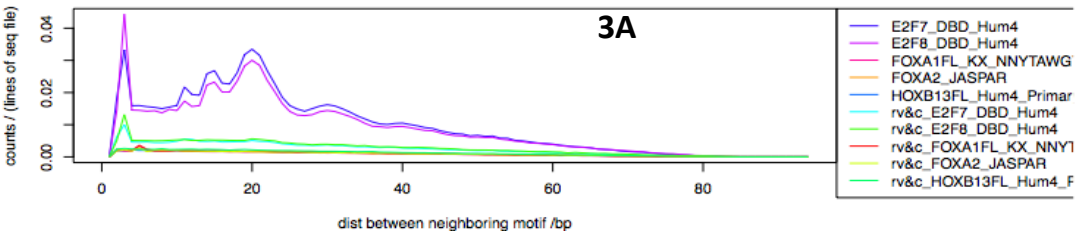


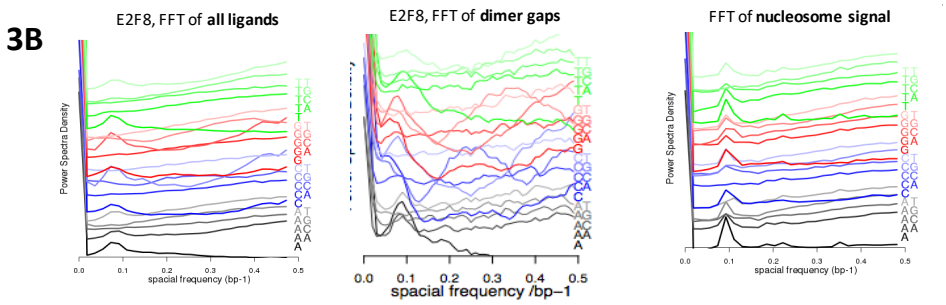
Fig2B



> evidence for DNA bending by TF dimer

TFs like E2F8, CDX, Neuro B/D dimer are know to form dimers which may bend DNA, E2F8 seems to be the most promising one as we have the structure and the spacing between 2motifs is showing peaks at long range (Fig3A). The FFT of the sequences from the gap region of E2F8 dimers are shown in FigB, gap is showing something but not really the same as the nucleosome one (FigB).





I checked the kmer counts of the sequences between two neighbouring motif matches (i.e. the dimer spacer). They are still enriched of TF motif-related sequences (even when using p=0.01 for MOODS), probably due to my use of all available previous motifs to pick dimer ligands rather than using only the monomer motif. Then there could be cases where a monomer motif is sitting between 2 dimer motifs, but still deemed as in the region of dimer spacer.

Further plans to address DNA bending

_test directly by X-ray scattering or maybe also light scattering?

_run PAGE for cyc4 SELEX library and search for TFs with anormalas shift

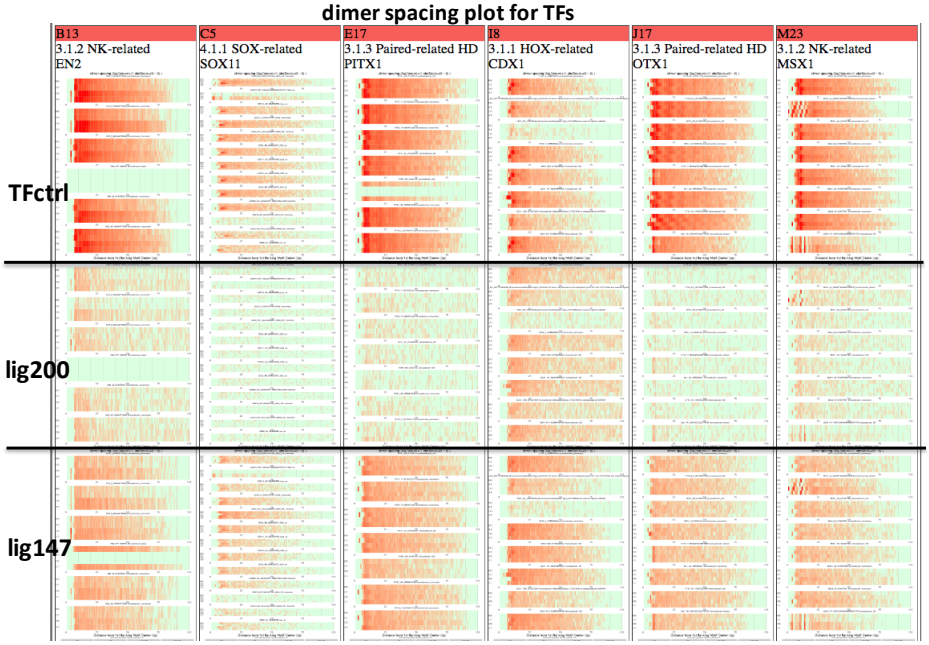
_for E2F8, if the bending exists it will be really sharp (the prefered spacing is 20bp from FigA). Such bending do not necessary give similar signal as the bending of nucleosome. We can try to analyse if any kmer is enriched in the dimer spacer, and also check the enrichment of physical features of kmers in the dimer spacer.

artificially place two motifs at the 2 ends of the long ligand, and see what enriches in the center

> nucleosome CAP-SELEX

> less allosteric effect in the presence of nucleosome

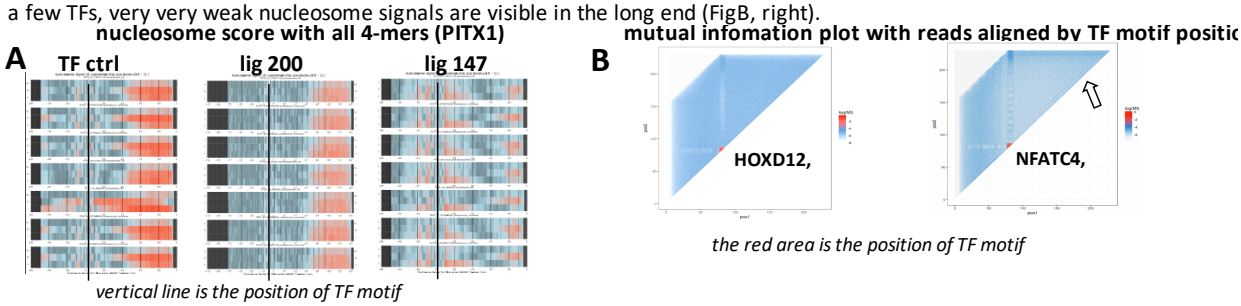
The following heatmap plots the number of TF dimer hits against the dimer spacer length. For the indicated TFs, less hits for ca. 10bp spaced dimer is observed when the TFs are assembled onto nucleosome. This could be because the mobility of DNA is already largely constrained by nucleosome, therefore no selection from the allosteric effect for the second binding event.



> the barrier effect by TF that positions nucleosome

The binding of nucleosomes to the TF-free region could be conceived more stable. Theoretically, if all the TF binding sites in nucleosome CAP-SELEX are aligned, a decrease of the nucleosome signal towards the TF binding site is expected.

I aligned the CAP-SELEX ligands according to the binding position of corresponding TF motifs, with the shorter end on the left side. However, the subsequent evaluation of the nucleosome signal was tricky. I tried to score with the spearman correlation between the rankings of all 4-mers. This method seems to be subject to the 1 or 2-nt bias that even the TF ctrl shows signal (FigA). Alternatively, I tried to score with the mutual infomation, as only nucleosome gives mutual infomation at long range. As indicated in FigB, left, the MI signal for most TFs is not dicernable for nucleosome. Probably due to too few counts of kmers. For a few TFs, very very weak nucleosome signals are visible in the long end (FigB, right).



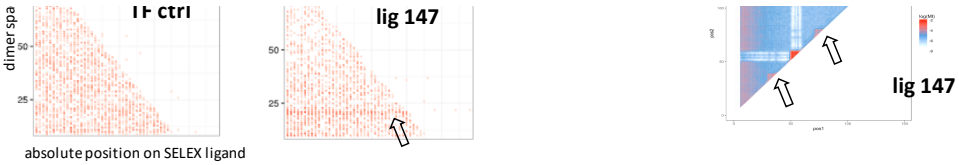
vertical line is the position of TF motif

the red area is the position of TF motif

To improve the detection of nucleosome signal
experimentally, maybe introducing in some footprinting tech, or adding a step of exo digestion will help;
it might also be worth to correct for 1nt bias to see if the spearman correlation score could be more reliable.

> new preferred spacing for CEBPB

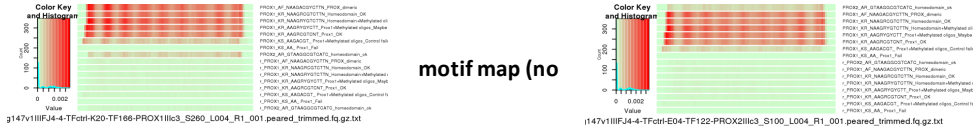
In addition to TFAP2C, in the presence of nucleosome, CEBPB also developed novel spacing rules with ca. 20 bp gap.



> periodic positioning without nucleosome

a few TFs are found to favor periodic positions on the SELEX ligand even without nucleosome. Probably there are 1 binding site on the adaptor, which controls the position of the 2nd binding event.

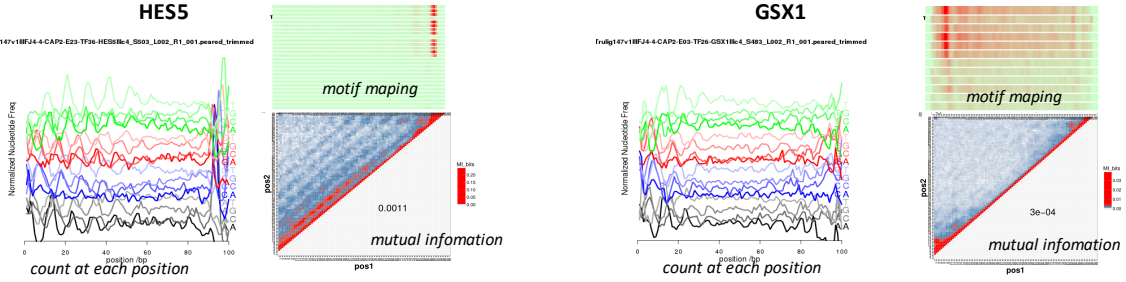
PROX1 **motif map (no** **PROX2**



> non homogeneously distribution of nucleosome signal on lig147

for many TFs, the signal of nucleosome is not homogeneously distributed even with the short ligand.
Some cases are easily understood as TF prefers the other end of the ligand, like HES5;
other cases, like GSX1, is more difficult to understand as the TF motif is enriched at the same side as nucleosome signal. In these cases, an overall decrease of nucleosome signal is also observed (MI plot of GSX1), probably the TFs can stabilize nucleosome when their motifs are positioned at the correct phase.

HES5 **GSX1**



> ligand design for crystallization

ligands for structural strudy are selected according to the following steps:

1. extract only the unique ligands from the enriched library of nucleosome-TF-consecutive-purification-SELEX, perform motif mapping using MOODS with the corresponding motif (shown below), figure out the most frequent position of the motif
2. use all ligands of the enriched SELEX library, map them again with the corresponding motif (MOODS has a score for each motif hit), select only the ligands with a motif at the most frequent position
3. for selected ligands in step2, sort by duplication number of each ligand, take the top 50 most duplicated ligands
4. for selected ligands in step3, calculate nucleosome scores for each ligand, either with continuous 7mer or with (3mer + 7bp Gap + 3mer), take the average of the 2 scores as the nucleosome score of the ligand
5. calculate a finalScore for each ligand: finalScore= scaled(motif score by MOODS) + scaled(duplication number) + scaled(nucleosome score), sort the selected ligands in step3 by the finalScore

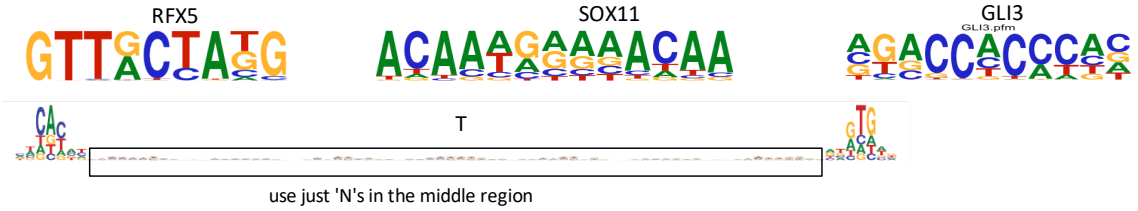
<u>now the top ligands should have preferable affinity to both nucleosome and TF, and with the TF motif located at the most preferred position relative to the nucleosome.</u>

6. also embed the motif hit sequence of each ligand into widom601 sequence, along with 4bp flankings (2bp for T)

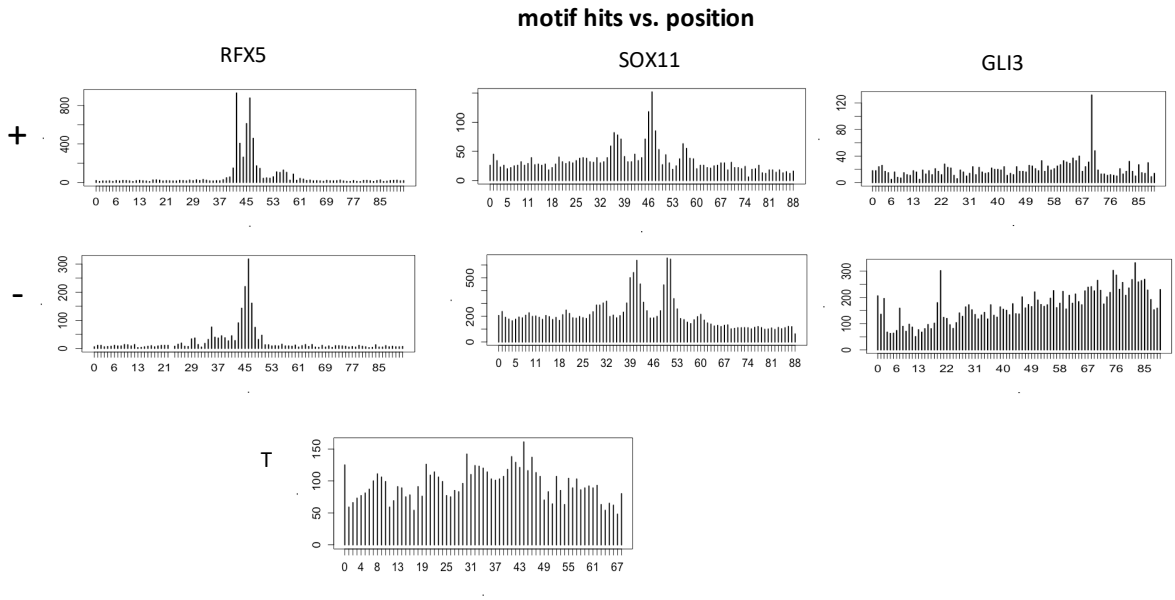
widom601:
CTGGAGAATCCCGGTCTGCAGGCCGCTCAATTGGTCGTAGACAGCTCTAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCGCCAAGGGGATTACTCCCTAGTCTCCAGGCA
CGTGTCAGATATATACATCCTGT

* for T, as we only get signal from the 200bp ligand, which contains 154bp in random, the target 147bp ligand is produced by cutting out 4 bp from the left end and 3 bp from the right end of the 154bp. In this way the position of T motif seems to best agree with the phase preference of the double-

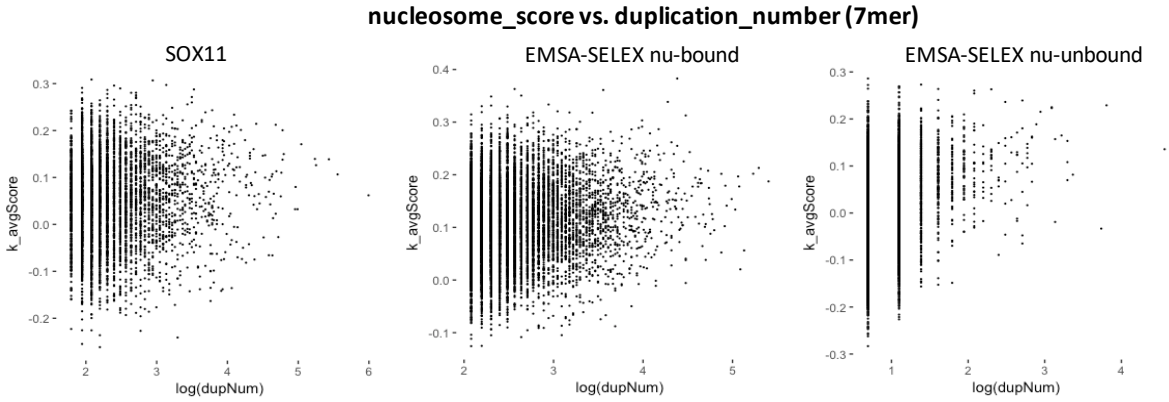
used motif



use just 'N's in the middle region



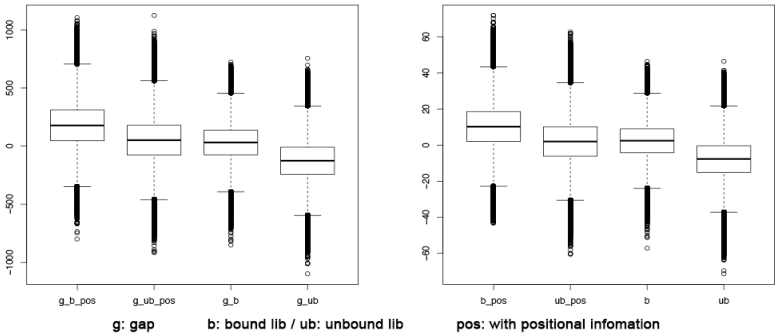
the scoring of nucleosome is not very well related with the duplication number of ligands in the library, either using continuous kmer or gapped kmer. This may suggest the duplication could be PCR bias, or the signal of nucleosome is weak, or the scoring strategy is not optimal.



I wondered if we can better discriminate the bound and unbound libraries of nucleosome, by including the fold-change infomation of more differently-gapped kmers, and by including positional infomations of each kmer. I compared the scoring results of only continuous 8mer and that of 8mers gapped 0-20bp in the center, both with and without positional infomation.

The result suggested that it does not make much sense to add gapped kmer and to consider the positional infomation. Probably most of the subtle deviations of kmer frequencies between bound and unbound libraries are just random rather than systematic. Maybe picking only the mostly biased kmers to score will suffer less from the background noise. Sequencing deeper may also work. (when not considering positional infomation, the overall score decreases is due to the same pseudo count normalized by different denominators, i.e. the sequence numbers of each library)

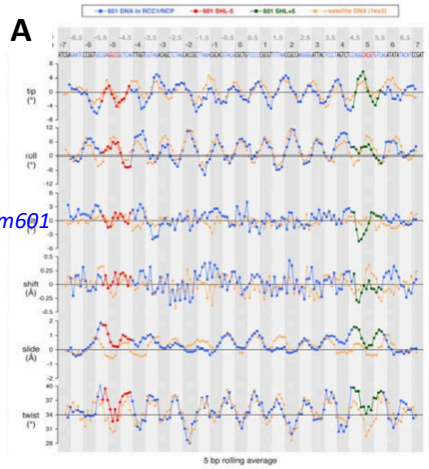
About the PCR bias, if the PCR polymerase isn't bias towards long kmer, can we just correct with 1nt or 2nt frequency to eliminate its effect?



> learning nucleosome affinity model by considering the physical parameters of DNA base pairs

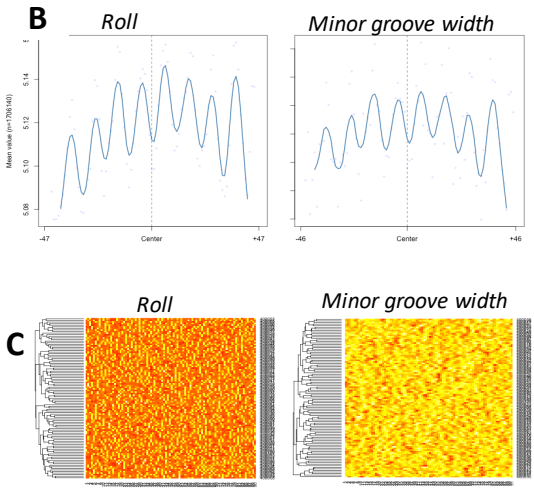
The text-based approach of sequence analysis suffers from the fluctuation caused by sparse data, especially when we subdivide the whole data set. E.g. In the boxplot above, the IQR becomes larger when we take into consideration the positional information. Maybe we can try to derive physical parameters from base-pairs or short oligo-nucleotides, and use them to augment the classification model for bound and unbound like Remo's lab has been doing.

periodicity of these physical features are already obvious for individual sequences (FigA), it is promising to use such features to discriminate bound and unbound, and probably also locate the center of nucleosome.
I used Remo's DNAshapeR package to check the nucleosome bound EMSA-library. As an ensemble, the signal is pretty strong (FigB), however the pattern is not obvious for individual sequences (FigC), possibly due to much less cycles of selection compared to widom sequences



blue: widom601

Nature. 2010 Sep 30;467(7315):562-6



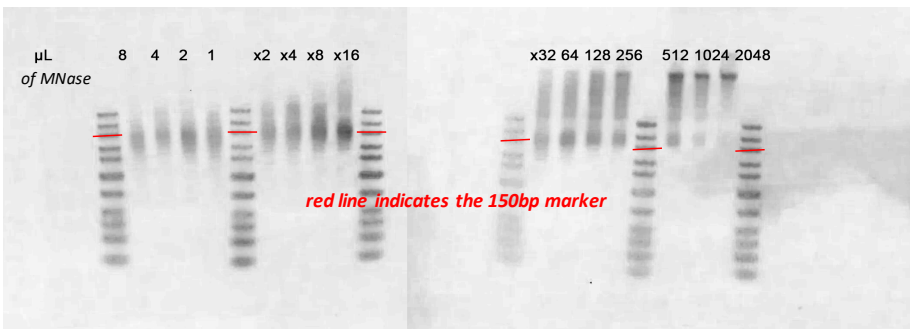
> MNase mapping for nucleosome positions

To study the motif and preferred positions of TFs on nucleosomal DNA in vivo, the plan is to combine MNase and ChIP data
Two batches of MNase mapping are done for GP5d and LoVo with the following design

condition		
batch2	GP5d	
batch2	GP5d_fixed	
batch2	LoVo	a conc titration (12 serial dilutions) of MNase is done for each condition
batch2	LoVo_fixed	
batch1	LoVo	

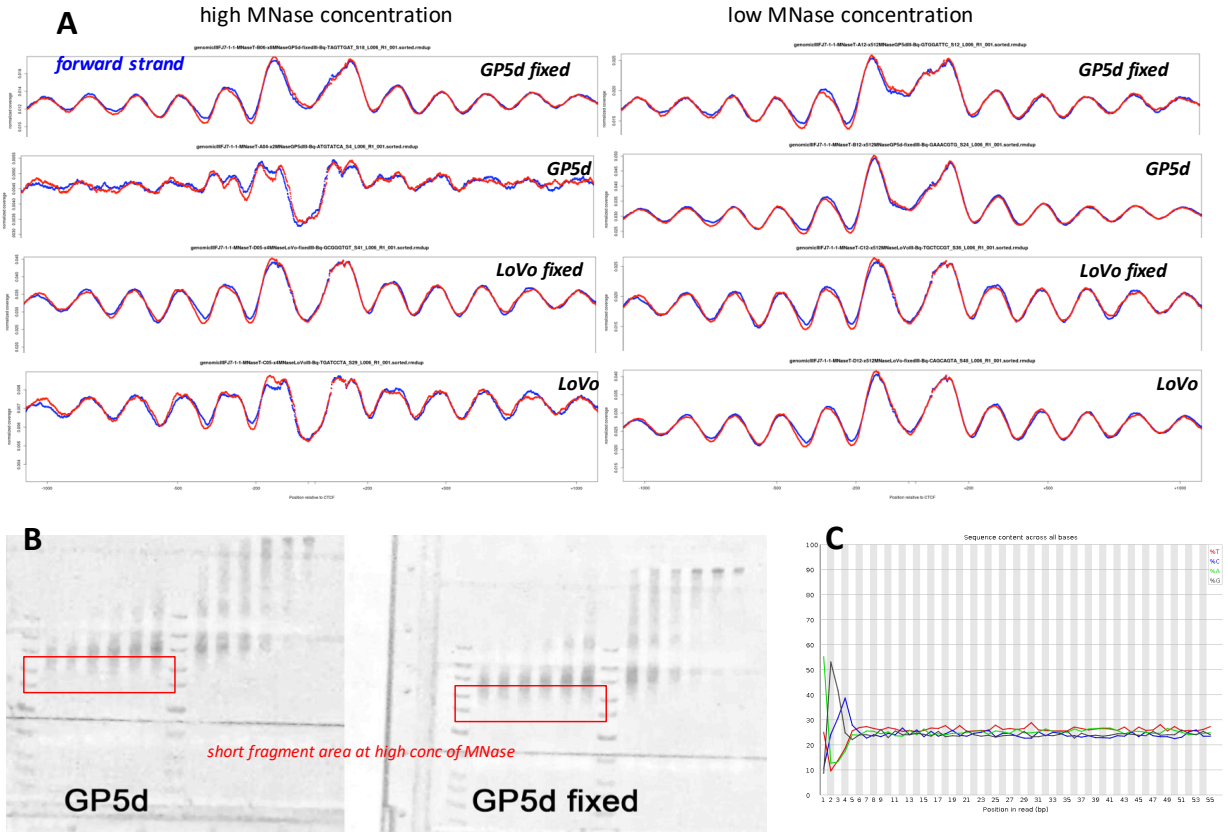
after MNase digestion, the gel image of the purified DNA is as follows

An example of MNase digested DNA with concentration titration of Mnase (batch 1 LoVo)



Levi performed the data-processing of MNase map. The occupancy of nucleosome around ChIP-seq CTCF sites are plotted. As shown in FigA, the non-fixed samples show a split in the middle of each nucleosome peak, at high MNase concentrations. This could be due to the additional cutting of MNase around the dyad axis of nucleosome at high concentration. But as shown in FigB, the red area, when comparing between the DNA gel images between the fixed and non-fixed cells, no significant increase of shorter DNA fragment is observed for non-fixed cells even at the high concentrations of MNase.

also, MNase showed sequence biases at the cutting site (FigC).
The gel images of MNase digestion do not have very sharp bands probably due to the lack of sonication (not moved to Solna yet



Introducing in UMIs to the ligation adaptor can help remove the effect from PCR duplicateds. With the suggestions from Kashyap, it seems that using illumina barcodes as UMI is most convenient as it requires no additional effort on processing. I can use i7 barcode as UMI and i5 barcode to demultiplex.
However, the MNase library is expected to have pretty high complexity, I wonder if we really need UMIs if not sequencing extremly deeply.

>> others

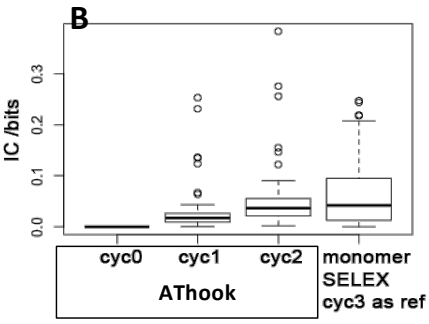
> infomation content analysis of AT-hook factors

Yimeng started to address the binding of AT-hook factors. A few preliminary, infomation-content based methods show that the binding of AT-hook factors are having comparable selectivity to normal TFs. FigB represents the development of IC for AT-hooks, if the stringency between batches are comparable, the IC of cyc3 of AT-hooks is comparable to previous monomer SELEX of normal TF. FigC is the mutual infomation plot for all gapped 3+3mers, AT-hook also seem to induce MI bias for shortly-gapped kmers (arrow). It could be interesting to further look into the binding model of AT-hook factors. Probably first check the bias of all kmers and then see if we can generalize the rule to make a simplified model.

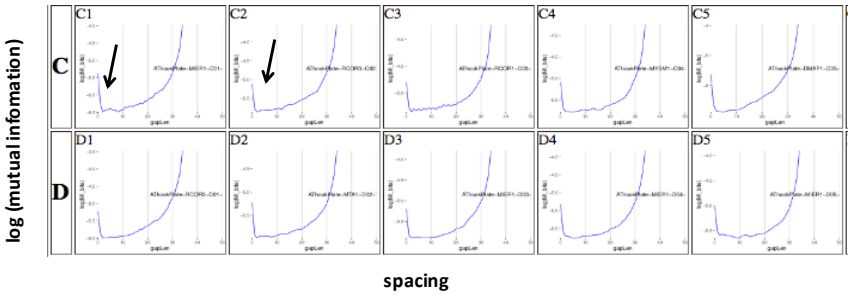
A IC of AT-hook factors after 3 cyc of SELEX

RCOR3	cyc3	0.38343	RERE	cyc3	0.04253	MYST1	cyc3	0.02205
ARID3C	cyc3	0.27593	RCOR3	cyc3	0.04128	PHF20	cyc3	0.0217
NFKB2	cyc3	0.25616	CXXC1	cyc3	0.0404	IRX1	cyc3	0.02099
HMGA1b	cyc3	0.15532	MIER1	cyc3	0.03863	MTA2	cyc3	0.0191
FOXB2	cyc3	0.14675	KMT2B_2	cyc3	0.038	MIER1	cyc3	0.01639
HMGA2	cyc3	0.12217	MTA1	cyc3	0.03778	MIER3	cyc3	0.01296
HMGA1a	cyc3	0.09093	MYSM1	cyc3	0.0372	MTA1	cyc3	0.01294
PATZ1	cyc3	0.07593	DNTTIP1	cyc3	0.03578	MIER1	cyc3	0.01048
RCOR1	cyc3	0.06455	MIER1	cyc3	0.03407	MIER2	cyc3	0.00918
AKNA_5	cyc3	0.06337	MTA1	cyc3	0.03266	MIER3	cyc3	0.00908
CBX2_1	cyc3	0.06147	NCOR1	cyc3	0.03177	MIER1	cyc3	0.00734
ZBTB24	cyc3	0.05733	AKNA_4	cyc3	0.03093	ARID3A	cyc3	0.00606
MIER2	cyc3	0.05293	DMAP1	cyc3	0.02848	PHF21A	cyc3	0.00525
CEBPB	cyc3	0.05153	GLYR1	cyc3	0.02712	ZBTB24	cyc3	0.00197
RFX5	cyc3	0.04668	MTA2	cyc3	0.02586			
SCML4	cyc3	0.04486	HMGA1c	cyc3	0.02531			
RCOR2	cyc3	0.04308	RCOR3	cyc3	0.0238			

B



C mutual infomation plot for 3mers vs. gap length



PLAN

> nucleosome CAP-SELEX project

_ repeat with DBD (ongoing)

_ run CAP-SELEX with Yimeng's FL proteins and AT hook

_check with Eevi and Sahu to see if we can verify any of the new TF binding modes in nucleosome CAP-SELEX by combining ChIP-nexus and Mnase-map data. Also, it might be worth to try an association study between the TF binding events and the nucleosome slidability nearby, to see if the binding of any TF locks nucleosomes there.

_check if most TF motifs are scored better with nucleosome-unbound library than bound, to see if the evolution of TF is favoring nucleosome depleted region

_check if the high concentration of nucleosome in early embryo contributes to restrict the binding thus allow only pioneer TFs

_SELEX to address the binding to dinucleosome or nucleosome linker

_ **need a more sensitive method to detect nucleosome signal**: the lack of a sensitive way to call nucelosome signal (from less reads and shorter oligonucleotides) is impeding many analysis, e.g., to check the presence/position/phasing of nucleosomes relative to TF motifs.

> quantitative binding models for TFs

> the binding model of AT-hooks with Yimeng