Stanford ChEM-H 2019 Seed Grant Competition: Postdocs at the Interface
Single-molecule methods for direct base-pair mapping of *in vitro* and *in vivo* protein-DNA interactions

GEORGI K. MARINOV[1,#,@], ZOHAR SHIPONY[1,#], AND ZHENG ZUO[1,2,*]

[*] *PI: Polly Fordyce*
[@] *PI: Anshul Kundaje*
[1] *PI: William J. Greenleaf*
[1] *Department of Genetics, Stanford University, Stanford, CA, 94305, USA*
[2] *Department of Bioengineering, Stanford University, Stanford, CA, 94305, USA*

Background and significance

Gene regulation is, in all organisms, primarily accomplished in one way or another through the interaction of proteins with nucleic acids. In most eukaryotes, the genome is packaged by nuleosomes (octamers composed of the four core nucleosomal histones), which have a refractory effect to transcription and to the binding of most regulatory factors to DNA. Active promoters and other regulatory elements are thus typically marked by decreased nucleosomal occupancy[1], by occupancy by various regulatory factors, and by specific chemical modifications of nearby histone molecules. The combined integrated action of these regulators determines the activity of the regulatory elements they associate with.

Profiling the occupancy of transcription factors (TFs) and histones on DNA has therefore always been of critical importance for the understanding of the mechanisms of gene regulation, and technological advancements enabling ever more detailed such measurements have accordingly played a major role for the development of the regulatory and chromatin biology fields. In particular, the ChIP (**Ch**romatin **I**mmuno**P**recipitation) technique has been indispensable for mapping protein-DNA interactions. It relies on the chemical crosslinking of proteins onto DNA, fragmentation of chromatin (usually using sonication), the specific pulldown of the proteins of interest using immune reagents, and the readout of the enriched DNA.

Since the invention of ChIP in the 1980s[2–4], advances in the field have been driven primarily by the evolution of methods for detection of the immunoprecipitated DNA. Coupling ChIP to qPCR[5] provided quantitative readout of occupancy but was limited to profiling only a few predetermined sites. Microarray technology allowed for probes to be designed covering large portions of the genome in the form of ChIP-Chip[6–9]. However, array-based methods suffered from low resolution and signal-to-noise ratio, poor reproducibility, and a number of other issues. The advent of high-throughput sequencing in the mid-2000s allowed for the unbiased sequencing-based readout of ChIP DNA and for relatively high-resolution, nearly truly genome-wide profiles, in the form of ChIP-seq[10–13].

Numerous further variations of the technique are now available, aimed at mapping the binding sites of RNA molecules (ChIRP-seq[14]), small molecules[15], the in vitro association of purified proteins with the genome (DAP-seq[16]), protein-mediated 3D DNA contacts (Hi-ChIP[17]), and others.

While highly powerful and informative, ChIP-seq is far from a perfect assay and efforts aimed at finding improvements and alternatives of it have continued apace. For example, artifactually enriched sites arising from sonication biases, copy number variation, indirect occupancy, and other sources[18,19] are a well known issue in ChIP datasets. Perhaps more importantly, the resolution of the assay is in many cases much coarser than desired. TFs typically occupy on the order of 10bp of DNA, however, the enriched regions derived from ChIP are on the order of 200bp. It is often difficult to identify precise occupancy sites, especially when multiple binding sites are closely clustered.

In order to address this issue, the ChIP-exo and ChIP-nexus assays were developed in recent years[20–22]. They are based on treating immunoprecipitated complexes with a processive exonuclease that is blocked when it encounters the cross-linked protein. The 5′ ends of sequencing reads correspond to these blockage points, providing a significantly higher-resolution of protein occupancy on DNA. However, these are still not direct base-pair maps of binding and are often difficult to interpret in cases of complex clustered binding to DNA.

Nuclease-based[23] alternatives to ChIP that do not use cross-linking have become popular recently, in particular in the form of CUT&RUN[24], which relies on MNAse recruitment to chromatin using a specific antibody and its subsequent activation. However, the resolution of these methods is not much higher than that of ChIP, and they suffer from significant off-target effects[24] as active MNAse often cleaves other regions that are close in 3D space, not just the actual target site[24].

The biggest gap in our capabilities, however, derives from the fact that all immunoprecipitation-based methods for mapping protein occupancy require the targeted enrichment of proteins of interest using specific immune enrichment. But at any given time a mammalian cell may be expressing several hundred active TFs and it is currently practically impossible to directly evaluate what the activity of all of them is genome-wide, at all their binding sites, and to therefore derive a truly comprehensive picture of the regulatory landscape of the cell. General maps of chromatin accessibility can provide information about the enrichment of TF motifs in accessible chromatin, but they are too coarse-grained to be fully informative at the level of individual motif instances.

The long-term goals of the work described in this proposal are to, first, develop a novel high-resolution truly base-pair-level single-molecule improvement over the ChIP assay, second, to fill the gap described above by developing a method for mapping protein occupancy genome-wide at the single-molecule and base-pair level, and third, to develop single-molecule multiomics assays that will eventually map chromatin accessibility, protein-DNA contacts, and endogenous DNA methylation within the same chromatin fibers. The immediate short-term goals of the proposal cover the first and second of these aims.

To accomplish these objectives, we will take advantage of the ability of nanopore sequencing to directly read a wide variety of DNA modifications. Most ChIP assays are based on the chemical crosslinking of proteins to

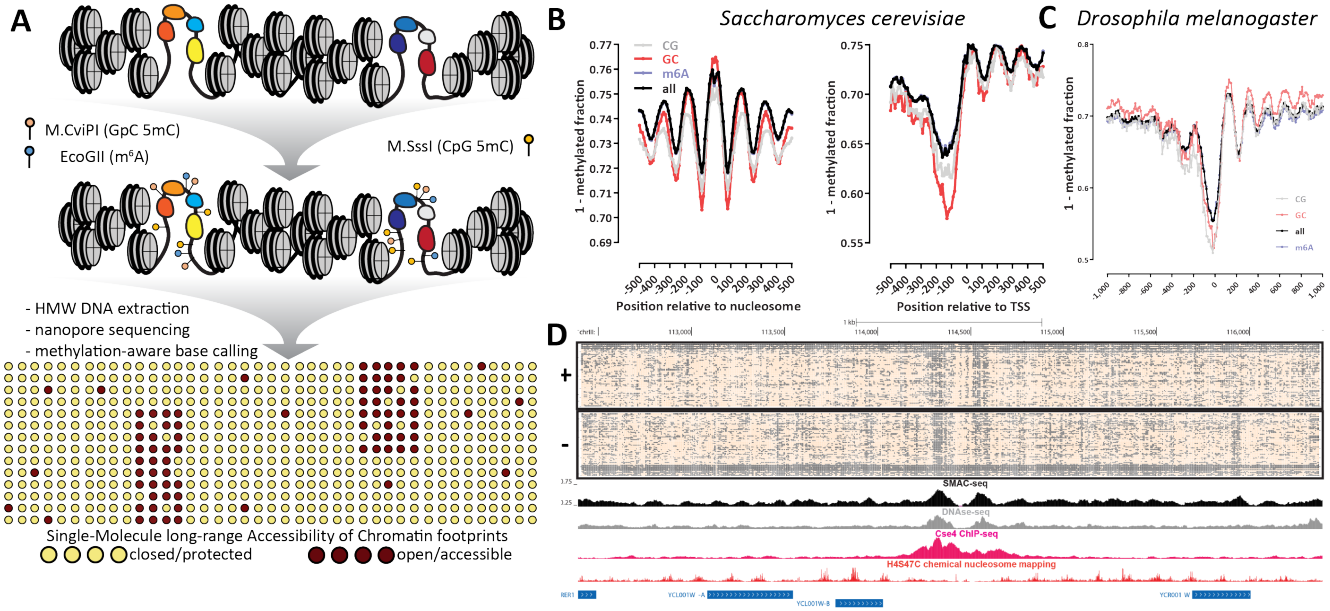

Figure 1: SMAC-seq maps chromatin accessibility within individual chromatin fibers using enzymatic methylation of exposed DNA and direct long-read single-molecule sequencing. The recently developed by us SMAC-seq assay serves as the starting point for our CLAPP-seq development efforts. (A) Overview of SMAC-seq; enzymatic DNA methylation is used to mark accessible DNA, which is then read out using nanopore sequencing; (B,C) SMAC-seq recovers known features of chromatin accessibility and nucleosome positioning in both unicellular eukaryotes such as yeast (B) and in metazoan cells (C); (D). SMAC-seq provides a single-molecule population-scale view of chromatin accessibility (shown is a region around one of the centromeres of *S. cerevisiae*).

DNA, usually using formaldehyde, though a variety of other chemical crosslinking agents can also be used, as well as high-intensity UV lasers[25]. In order for DNA to be amplified, proteins are digested using proteinase treatment and crosslinks are simultaneously reversed through incubation at a high temperature. However, while proteinase enzymes cleave peptide bonds, the crosslink bonds are not peptide in nature; they are reversed by high temperature, while proteinase treatment alone leaves bulky adducts onto DNA[26]. These adducts can not only be directly read using nanopore sequencing but are expected to generate much stronger current shifts than DNA methylation marks due to their size and polarity (Figure 2E; while nanopore sequencing is very powerful for detecting DNA methylation even as it is, significant noise is still observed in base calls at the single-molecule level due to the small absolute differences between methylated and unmodified nucleotides). Base pair-level maps of direct protein-DNA contacts can be thus obtained (**C**ross-**L**inking **A**ssisted **P**rotein **P**ositioning sequencing, or CLAPP-seq), a property that, alone or in combination with other approaches, can be used to develop a new class of assays for mapping chromatin structure.

The work proposed here builds on previous efforts by us that resulted in the development of novel methods using single-molecule nanopore sequencing to map chromatin accessibility within individual chromatin fibers at a multikilobase scale[27] (Figure 1). Existing methods for profiling open chromatin genome-wide all rely on some combination of short-read sequencing and enzymatic cleavage, making it impossible to observe actual "chromatin haplotypes" and to evaluate the degree of co-accessibility between distal regulatory regions. We overcame this limitation by employing methyltransferases that preferentially modify accessible chromatin at a high density (in particular, the non-sequence-specific EcoGII enzyme that generates m^6A) and reading out methylation/open chromatin states using long-read nanopore sequencing (SMAC-seq, or **S**ingle-**M**olecule long-read **A**ccessible **C**hromatin mapping **seq**uencing assay). We have now successfully applied the method to evaluate long-range dependencies between regulatory elements in multiple eukaryote model systems (Figure 1). The research proposed here will build on the expertise and experience we have developed in the course of these studies.

Specific aims

Specific Aim 1: Development and optimization of an *in vitro* CLAPP-seq assay for mapping protein-DNA contacts. Our initial efforts will be focused on establishing proof-of-concept using *in vitro* experiments, on optimizing crosslinking conditions, and on developing analytical methods for mapping protein-DNA contacts in nanopore

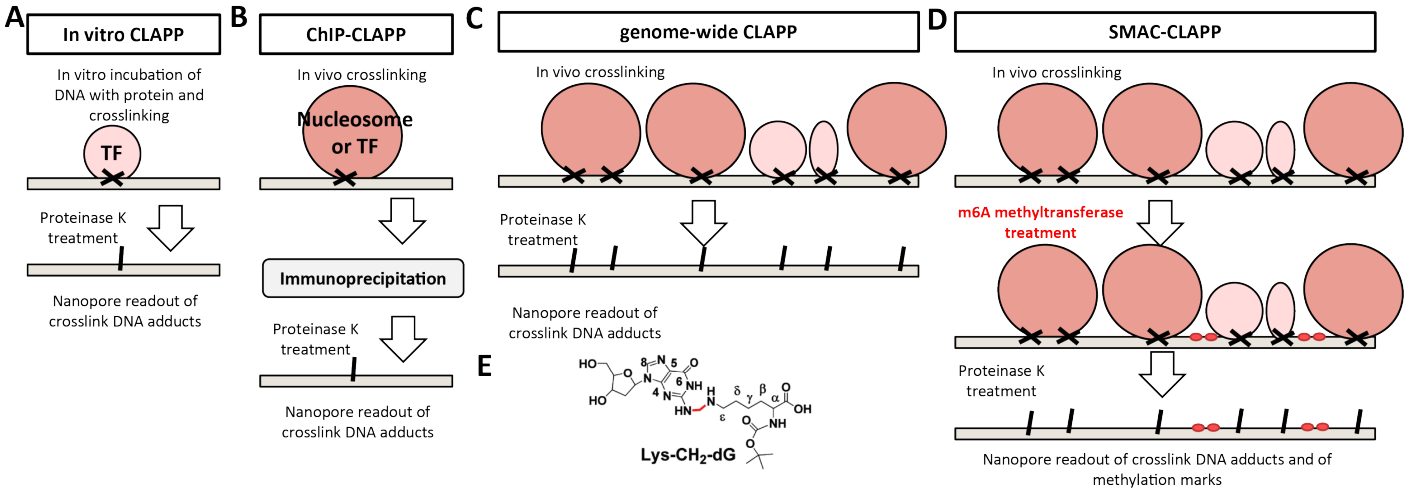

Figure 2: Single-molecule methods for mapping protein-DNA contacts to be developed as part of this proposal and as future work derived from it. (A) In vitro CLAPP (**C**ross-**L**inking **A**ssisted **P**rotein **P**ositioning); this is primarily a method development and validation part of the proposal though we foresee it as also a highly useful assay for studying *in vitro* nucleosome positioning and as a high-resolution replacement for methods such as DAP-seq[16]. Purified proteins are incubated with DNA, crosslinked and then digested with Proteinase K without reversing the crosslinks. The DNA adducts remaining on the DNA are then directly read out using nanopore sequencing; (B) In ChIP-CLAPP, crosslinking is carried out on live cells and the protein of interest is pulled down using immunoprecipitation as in standard ChIP protocols. (C) In genome-wide CLAPP, the immunoprecipitation step is omitted and all protein-DNA contacts in the genome are mapped in an unbiased fashion; (D) Longer-term, we aim to develop a combined chromatin accessibility and protein contact single-molecule multiomics assay by integrating CLAPP and SMAC (SMAC-CLAPP); cross-linked samples are treated with an m^6A methyltransferase and then digested with Proteinase K. Crosslink adducts and m^6A methylation are separately read out using nanopore sequencing. In principle, endogenous CpG-context cytosine methylation can also be simultaneously detected. A ChIP or another targeted enrichment step can also be added. (E) Structure of a DNA-amino acid crosslinking adduct[26] (in this case guanine-Lysin).

sequencing data. To this end we will employ purified TFs, which we will incubate with a panel of PCR-amplified genomic DNA segments carrying strong binding sites for these TFs, then crosslink, digest with Proteinase K, and subject to nanopore sequencing (Figure 2A). Such *in vitro* experiments will allow comparison against known ground truth observations (as DAP-seq measurements will also be performed side-by-side) and for optimization of reaction conditions. In particular, we will use the Sox2, Sox17 and Oct4 TFs, whose binding specificities are relatively well characterized as monomers or heterodimers), and for which we already have purified proteins. Optimization of reaction conditions will involve the identification of an optimal crosslinking agent and its concentration and duration (we will initially use formaldehyde, but it may be the case that more aggressive/longer-arm crosslinkers such as glutaraldehyde, chloroacetaldehyde, DSG, EGS, or some others are the ones that maximize crosslinking efficiency and/or detection power) and reaction quenching conditions[28] (the latter is important in order to make sure that no DNA adducts are formed in the absence of protein-DNA contacts). *In vitro* experiments will also enable training CLAPP-specific modified base calling algorithms.

We have already sequenced DNA modified with a variety of bulky adducts, and indeed observed much more robust DNA modification detection than what is obtained using plain DNA methylation marks (Figure 3).

While our main goal for the these experiments is to develop and validate the CLAPP assay, we expect in the long term *in vitro* CLAPP to also be highly informative when applied to biological questions regarding protein-DNA interactions that can be studied in vivo. For example, Sox17 is known to exist in mouse ES cells and share the same binding specificity with Sox2, but can promote the differentiation to endoderm cell lineage. If nanopore sequencing can identify the unique crosslinking signatures for Sox2 and Sox17 respectively, we can use it to address some important questions related to stem cell maintenance and differentiation. Also we can generate base-pair-resolved maps of genome-wide TF *in vitro* occupancy (unlike the coarser-grained resolution of DAP-seq), and most intriguingly, studying the behavior of in vitro reconstituted nucleosomes subjected to posttranscriptional modifications, *in vitro* reconstituted transcription systems, and others.

Specific Aim 2: Development and optimization of a ChIP-CLAPP assay for targeted mapping of protein-DNA contacts. Our second goal is to adapt the CLAPP method to *in vivo* conditions by coupling it to ChIP (Figure 2B). This will provide the desired high-resolution truly base-pair alternative of the ChIP assay and will also allow us to work in a more localized, less complex context in terms of development and optimization of data analysis

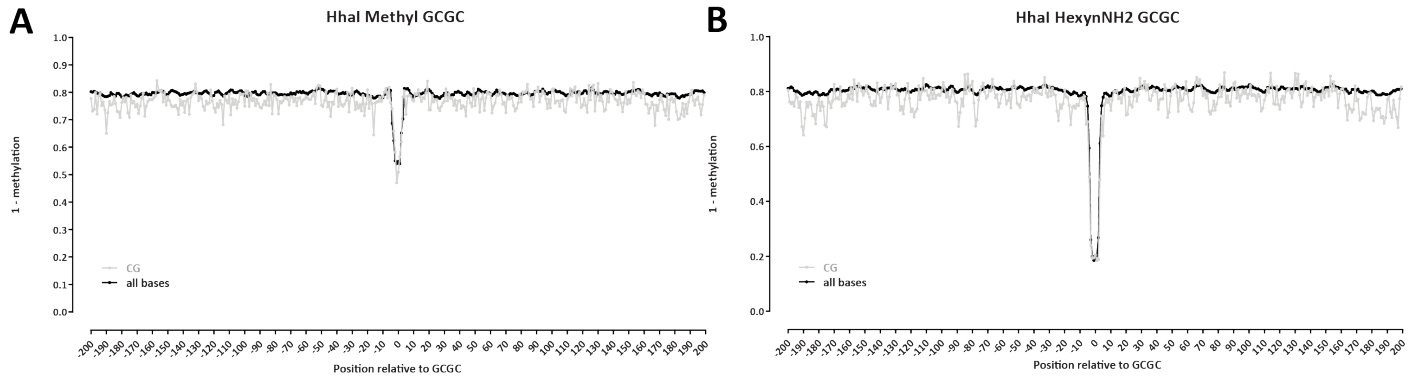

Figure 3: Bulky DNA modifications are robustly detectable using nanopore sequencing. Lambda DNA was methylated using the HhaI methyltransferase (which methylates C nucleotides in GCGC sequence conexts) and either the SAM cofactor (through which a simple methyl group is deposited) or modified SAM carrying a bulkier adduct (such as Hexyn-NH_2). DNA was then sequenced using the Oxford Nanopore MinION platform. Though the absolute methylation levels are similar as measured by protection against restriction digestion (not shown), detection of DNA modification levels is much more robust for the Hexyn-NH_2 adduct (B) than it is for plain methylation (B).

methodology. One limitation of this approach is that at present nanopore sequencing requires more than 100 ng of DNA as input. These amounts significantly exceed what is obtained from the typical TF ChIP reaction but are easily reached when histone marks are ChIP-ed, thus our development experiments will target histone modifications. Another challenge is that nanopore-based ChIP-seq has not been reported before, and that nanopore sequencing does not read sequences shorter than $\sim$200 bp, which is longer than the length of many sonicated fragments. We will apply restriction digestion instead (using a 5-cutter enzyme leaving overhangs) together with a ligation step that generates longer fragments in order to circumvent this limitation. These experiments will be piloted in fruit fly S2 cells as the *Drosophila* genome is relatively compact and high depth coverage can be achieved without sequencing on multiple nanopore flowcells.

Specific Aim 3: Development of a genome-wide CLAPP assay for mapping protein-DNA contacts. Our final goal for this proposal is to generate pilot genome-wide CLAPP data, without an enrichment step (Figure 2C). These experiments will be carried out in the yeast *Saccharomyces cerevisiae*, as it has a very small for a eukaryote genome, it has no endogenous methylation, and a wealth of functional genomic information, such as high-resolution nucleosome positioning maps and comprehensive large-scale TF occupancy mapping datasets, is available for it[27]. We also plan to explore a prokaryote genome that lacks nucleosomes and is expected to be less tightly associated with proteins, using *E. coli* as a model system. *S. cerevisiae* experiments will allow us to study nucleosome positioning at the level of protein-DNA contacts, and to evaluate the relationship between base pair-level contacts and sequence recognition motifs for most yeast TFs; this will prepare us for transitioning to studying more complex metazoan genomes.

Long-term goals and future directions

The successful development of the approaches outlined here will also open the door for the development of a variety of novel methods for studying chromatin structure in previously unavailable detail. We are particularly excited about the prospect of obtaining single-molecule multiomic measurements of protein-DNA contacts, chromatin accessibility and endogenous methylation (in systems where it exists). This will be possible through the integration of the SMAC and CLAPP methods (SMAC-CLAPP; Figure 2D), although it will likely require careful development of much more sophisticated basecalling algorithms and models than the ones that exist at present.

Finally, nanopore sequencing can directly read out not just DNA molecules but also RNA, and the CLAPP approach will in principle be also applicable at the RNA level. This is an even longer-term prospect, as at present the accuracy of nanopore base calling at the RNA level leaves a lot to be desired, but is nevertheless a highly intriguing one. RNAs spend much of their life in the cell associated with a variety of proteins that regulate their stability, translation, non-coding activities, and many other aspects of their function. Comprehensive full-mRNA-length long-read base-pair mapping of RNA-protein interactions would provide invaluable information about these processes, that is at present impossible to obtain.

References

1. Klemm SL, Shipony Z, Greenleaf WJ. 2019. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20**(4):207–220.
2. Gilmour DS, Lis JT. 1984. Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proc Natl Acad Sci U S A* **81**(14):4275–4279.
3. Gilmour DS, Lis JT. 1985. In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster. *Mol Cell Biol* **5**(8):2009–2018.
4. Solomon MJ, Larsen PL, Varshavsky A. 1988. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* **53**(6):937–947.
5. Hecht A, Strahl-Bolsinger S, Grunstein M. 1996. Spreading of transcriptional repressor SIR3 from telomeric heterochromatin. *Nature* **383**(6595):92–96.
6. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290**(5500):2306–2309.
7. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**(6819):533–538.
8. Lieb JD, Liu X, Botstein D, Brown PO. 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet* **28**:327–334.
9. Weinmann AS, Yan PS, Oberley MJ, Huang TH, Farnham PJ. 2002. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev* **16**(2):235–244.
10. Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**(5830):1497–1502.
11. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**(4):823–837.
12. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**(7153):553–560.
13. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**(8):651–657.
14. Chu C, Qu K, Zhong FL, Artandi SE, Chang HY. 2011. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol Cell* **44**(4):667–678.
15. Anders L, Guenther MG, Qi J, Fan ZP, Marineau JJ, Rahl PB, Lovn J, Sigova AA, Smith WB, Lee TI, Bradner JE, Young RA. 2014. Genome-wide localization of small molecules. *Nat Biotechnol* **32**(1):92–96.
16. Bartlett A, O'Malley RC, Huang SC, Galli M, Nery JR, Gallavotti A, Ecker JR. 2017. Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat Protoc* **12**(8):1659–1672.
17. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, Chang HY. 2016. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13**(11):919–922.
18. Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, Lefrançois P, Struhl K, Gerstein M, Snyder M. 2009. Mapping accessible chromatin regions using Sono-Seq. *Proc Natl Acad Sci U S A* **106**(35):14926–14931.
19. Teytelman L, Ozaydin B, Zill O, Lefrançois P, Snyder M, Rine J, Eisen MB. 2009. Impact of chromatin structures on DNA processing for genomic analyses. *PLoS One* **4**(8):e6700.
20. He Q, Johnston J, Zeitlinger J. 2015. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat Biotechnol* **33**(4):395–401.
21. Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**(6):1408–1419.
22. Rossi MJ, Lai WKM, Pugh BF. 2018. Simplified ChIP-exo assays. *Nat Commun* **9**(1):2842.
23. Schmid M, Durussel T, Laemmli UK. 2004. ChIC and ChEC; genomic mapping of chromatin proteins. *Mol Cell* **16**:147–157.
24. Skene PJ, Henikoff S. 2017. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**. pii: e21856.
25. Steube A, Schenk T, Tretyakov A, Saluz HP. 2017. High-intensity UV laser ChIP-seq for the study of protein-DNA interactions in living cells. *Nat Commun* **8**(1):1303.
26. Lu K, Ye W, Zhou L, Collins LB, Chen X, Gold A, Ball LM, Swenberg JA. 2010. Structural characterization of formaldehyde-induced cross-links between amino acids and deoxynucleosides and their oligomers. *J Am Chem Soc* **132**(10):3388–3399.
27. Shipony Z, Marinov GK, Swaffer MP, Sinott–Armstrong NA, Skotheim JM, Kundaje A, Greenleaf WJ. 2018. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *bioRxiv* 504662.
28. Wu CH, Chen S, Shortreed MR, Kreitinger GM, Yuan Y, Frey BL, Zhang Y, Mirza S, Cirillo LA, Olivier M, Smith LM. 2011. Sequence-specific capture of protein-DNA complexes for mass spectrometric protein identification. *PLoS One* **6**(10):e26217.

Budget

Equipment and supplies

Based on detailed projections of expenditures, we request funds totaling $50,000 for the following supplies:

- $2,500 for expressing and preparing purified transcription factors
- $7,500 for purchasing four ONT Flongle starter packs (consisting of 1× Flongle Adapter and 12× Flongle Flow Cells each). Having these at our disposal will allows us to quickly and in parallel test and optimize a wide number of crosslinking conditions for our in vitro CLAPP experiments.
- $25,000 for purchasing ONT MinION and PromethION flowcells for sequencing in vivo CLAPP and ChIP-CLAPP samples at a sufficient depth
- $5,000 for purchasing Illumina NextSeq flowcells (for parallel sequencing of control DAP-seq and ChIP-seq experiments)
- $5,000 for general lab supplies
- $5,000 for the purchase of additional data storage capacity for raw and processed nanopore datasets

Participants roles and expertise

Georgi K. Marinov

G.K.M.'s PhD work included, as one of its main areas of focus, the development of best practices and protocols for carrying out and analyzing ChIP-seq experiments, in particular as part of the ENCODE Consortium Project. His most recent research has concentrated on the development of single-molecule long-read methods for profiling chromatin accessibility and other aspects of chromatin structure. He will be contributing to the project his expertise in designing and optimizing ChIP experiments, and in generating and analyzing nanopore sequencing datasets.

Zohar Shipony

Z.S PhD work studied the behavior of epigenetic memory, with a focus on DNA methylation, in different cell type, including cancer cells, normal cells and emryonic stem cells. These efforts led to the discovery that while somatic and cancer cells maintain their epigenetic memory between cell divisions, with a high rate of epimutation calculated as 1/500 bases per cell division, embryonic stem cells maintain a dynamic epigenetic landscape that can be rewritten between cell cycles. He will be contributing to the project his expertise in working with modified DNA and his knowledge of Nanopore sequencing.

Zheng Zuo

Z.Z. did his PhD work to characterize many aspects of protein-DNA interactions, including specificity, cooperativity, and methylation sensitivity. Currently he is combining sequencing, microfulidics, and chemical biology approaches to study the post-translational modifications(PTMs) effect on protein-DNA and protein-protein interactions. In this project, he will be contributing to the design of DNA templates and constructs for in vitro test, expression of various transcription factors, including Sox2, Oct4, CTCF etc, perform crosslinking, proteinase digestion, and help analyze the specificity of those studied TFs.