NOFIC-AICS Collaborative Project

Objective: Develop a general approach for studying the dynamic nuclear organization in mammalian cells. Ultimately, this will allow us to learn the general principles of dynamic nuclear organization of cells during the cell cycle, differentiation, and in response to extracellular signals.

Background: With the human genome sequence completely mapped, understanding the 3D genome organization in the nucleus has now become a new frontier in genomic research. Chromatin architecture can impact gene activities at multiple levels, and transcription processes in turn can also influence the local and global chromatin organization. Currently, the processes and principles governing dynamic chromatin organization are still poorly understood. The NIH 4D Nucleome (4DN) Consortium, launched in 2015, was intended to fill this knowledge gap. During the first three years of this Common Fund project, remarkable progress has been made in both the technologies and conceptual frameworks for 3D genome analysis. Yet major challenges remain, in particular in the study of the temporal dimension of 3D genome organization. This proposed collaborative project between the Nuclear Organization and Function Interdisciplinary Consortium (NOFIC) centers and the Allen Institute for Cell Sciences (AICS) is aimed to address this issue by capitalizing on the recent technological and conceptual advances made by both entities.

AICS is interested in understanding the principles by which cells reorganize in 3D space as they move from state to state in response to the cell cycle, differentiation, changes in environment, agonists/antagonists, etc. The initial focus of AICS is the 3D cellular organization. AICS has developed a panel of human pluripotent stem cell (hiPSC) lines expressing genetically engineered, EGFP-tagged proteins that can be used to map the dynamic localization of a particular organelle or structure using live cell imaging. They have further developed quantitative image-based assays and segmentation algorithms for quantitative analyses. AICS is now producing thousands of replicate images for each structure and making the images publicly available, along with corresponding analytical tools, reagents, and cell lines. Furthermore, scientists at AICS have used machine learning algorithms to model and integrate cells labeled with different EGFP markers. While their analysis is still in the early stage, they have already produced a relatively comprehensive picture of the changes in global cellular organization during mitosis (7 stages and 14 key intracellular structures). One of the most exciting new tools is the “label free imaging” of cellular structures. This method uses deep neural networks to learn to map from a brightfield image to an image with a particular structure fluorescently tagged. By applying multiple trained networks, each trained using a different tag, to the same cell the method enables the simultaneous observation of the nucleus, DNA, mitochondria, cell membrane, etc. from 3D brightfield images or movies. Additionally, the tool also predicts the organization of chromatin and how it changes during mitosis and has great potential for identifying key landmarks within the nucleus.

The NOFIC is a consortium of six research centers funded by the 4DN project to carry out integrative analysis of the chromatin organization in mammalian cells. Each center has

developed a unique set of experimental and computational approaches for analysis of the chromatin architecture and has been using them to study various cell types. The investigators within the NOFIC centers are leading scientists in the field of chromatin organization studies, and the tools they developed, such as single cell combinatorial indexing Hi-C, GAM, high throughput FISH, ChIA-PET, PLAC-seq, TSA-Seq, etc., represent state-of-the-art approaches to analysis of 3D genome organization. Deploying this diverse set of tools, in conjunction with the powerful methods developed by AICS, to a common cell system provides a unique opportunity to achieve the overall objective of the 4DN project.

Investigators participating in the collaboration:

- **NOFIC Centers:**
    - Shendure/Noble (UW)
    - Dekker (U Mass)
    - Ren/Murre/Pombo/Nicodemi (UCSD)
    - Alber/Chen (USC)
    - Belmont (UIUC)
    - Ruan (JAX)

- **AICS:**
    - **Susanne Rafelski**, PhD is Director of Assay Development at the AICS (http://alleninstitute.org/what-we-do/cell-science/).
    - **Ruwanthi Gunawardane**, PhD is Director of Stem Cell and Gene Editing at the AICS (same link)

Methods and approaches:

Aim 1. Imaging nuclear proteins in live cells during the cell cycle and stem cell differentiation (AICS)

To visualize the dynamics of nuclear proteins in live cells, a number of human iPSC-lines have been developed in which nuclear proteins are tagged mono-allelically with fluorescent markers (eg. GFP). An initial set of nuclear proteins representing key nuclear landmarks were identified for tagging in collaboration with members of the 4DN Nuclear Organization and Functional Integration Centers (NOFIC) and are currently in progress at AICS (Table 1). These include lamin B1, a nuclear lamina protein; fibrillarin and nucleophosmin, nucleolar proteins localized to two separate subcompartments; H2B to label histones, NUP153, a nuclear pore protein; CTCF and SMC1A, proteins implicated in chromatin interactions and chromosome dynamics; HP1-beta (CBX1) representing heterochromatin. Additional proteins representing other key landmarks within the nucleus will be tagged. These may include EZH2, a silencing complex protein; CENP-A, a centromere protein; TRF2, a telomere protein as well as proteins representing nuclear speckles, enhancers, splicing, and epigenetic modifications. AICS has

already generated preliminary data for the first eight proteins, and some of this data is already publicly available for lamin B1 and fibrillarin (3D cell viewer at http://www.allencell.org).

Live-cell imaging will be used to follow changes in the location of the target nuclear proteins in two contexts, the cell cycle and differentiation, comparing changes between undifferentiated stem cells (WTC-11, a 4DN Tier 2 cell line) and cardiomyocytes derived from them. The AICS workflow for each protein imaged includes (1) creating the genome-edited lines, (2) developing a microscopy pipeline that generates high replicate data, and (3) analyzing, visualizing and modeling the data. CRISPR/Cas9 gene editing will produce clonal hiPS cells in which each protein is endogenously tagged with a fluorescent protein (mEGFP). Extensive quality control will include testing proper editing at the right locus, karyotype, morphology, pluripotency, protein localization, protein and RNA expression, and cell behavior.

Each cell line will be imaged using 3D live cell microscopy at high resolution and magnification (100x), and at high replicate numbers (300-1000 cells) for analysis of population variance. Microscope modalities include primarily spinning disk confocal microscopy as well as Zeiss AiryScan FAST super-resolution (2x increased resolution in x, y, and z) laser scanning confocal microscopy. Cells will be imaged as undifferentiated hiPSCs and as cardiomyocytes at day 22-29 of differentiation. Cells will first be imaged as individual 3D z-stacks to identify the location of each protein and how it changes as cells go through the cell cycle and differentiation. A DNA binding dye (NucBlue) will label the nucleus, and a cell membrane dye (CellMask Deep Red) will label the cell outline. Each cell image will be attributed to a stage of the cell cycle based on NucBlue staining.

The “label free imaging” tool will be used to train deep learning models [2] for each of the GFP-tagged proteins to develop an integrated view of key nuclear landmarks. This will greatly enhance the ability to identify nuclear state signatures based on multiple landmarks (instead of one landmark at a time) and may permit us to visualize these nuclear signatures based solely on transmitted light images, which would enable analysis of the dynamics of nuclear organization via time lapse imaging.

Table 1. A list of proteins representing key nuclear landmarks for fluorescent tagging and generation of clonal genome-edited human pluripotent stem cell lines at AICS.

	Structure	Gene/Protein	FP	Status	Data
1	Nuclear envelope	LaminB1	mEGFP	Available	Yes*
2	Nucleolus - DFC	Fibrillarin	mEGFP	Available	Yes*
3	Nucleolus - GC	NPM1	mEGFP	Available	By 12/19
4	Histones	H2B	mEGFP	Available	4/19

5	**Cohesin**	**SMC1A**	**mEGFP**	**Clonal QC (release expected in Dec 2018)**	
6	**Nuclear pores**	**NUP153**	**mEGFP**	**Available**	**4/19**
7	**Chromatin-binding**	**CTCF**	**mEGFP**	**Edited pool**	
8	**Heterochromatin**	**HP1-beta/CBX1**	**mEGFP**	**Edited pool**	
9	***Silencing complex***	***EZH2***	***mEGFP***	***future***	
10	***Centromere***	***CenpA***	***mEGFP***	***future***	
11	***Telomeres***	***TRF2***	***mEGFP***	***future***	
12	***Nuclear speckles***	***TBD***	***mEGFP***	***future***	
13	***Nuclear pores***	***NUP107***	***mEGFP***	***future***	
14	***Nuclear envelope***	***LaminA/C***	***mEGFP***	***future***	
15 +	***Nuclear speckles, enhancers, splicing, epigenetic modifications***		***mEGFP***	***future***	

*** Research imaging data (time lapse data, for example) will be made available to participating groups pending collaborative agreement. Processed data (segmentation, for example) will be made available and versioned.**

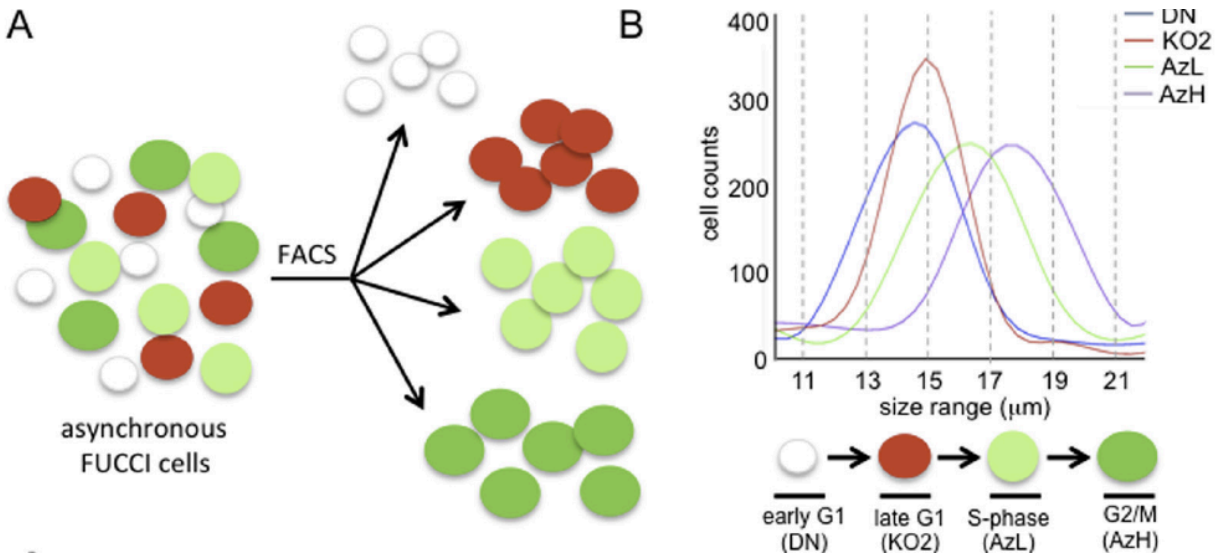
Aim 2. Analyze the chromatin structure, protein binding sites of the human iPSC lines during the cell cycle and stem cell differentiation (NOFIC).

Complementary sequence-based data will be produced to complement the live imaging analysis of state transitions and to enable in-depth understanding of dynamic chromatin organization during cell cycle and differentiation of the human iPSC system in Aim 1.

Common cell system: The human iPSC line WTC-11 will be used for this research aim. All genetically engineered cell lines at AICS are derived from this cell line. Two temporal processes will be studied:

- **Cell cycle**.
    - (For bulk assays - FUCCI system coupled with FACS; early G1, late G1, early S, late S +G2+M)
    - (For single cell based assays - either asynchronous cells or FACS sorted cells)

Comment (ASB): How will the synchronization be done? What type of synchrony will even be maintained for ensemble measurements after the initial synchronization and release? Is there a feasible protocol that will allow reasonable synchrony for ensemble measurements?



- **Differentiation**. A protocol for differentiation to cardiomyocytes has been established at AICS and would be used to study the dynamic chromatin re-organization in this process.

Comment (ASB): This is a very long differentiation protocol. How many time points are you considering?

Comment (ASB) - Could cell system be different? Serum stimulation, for example.

The following technologies will be employed on a common cell system:

- Hi-C (**Dekker**), micro-C possible
- Single cell Hi-C (**Shendure**)
- PLAC-seq (HiChIP) (**Ren**) (use GFP Ab?)
- GAM (**Pombo**): combined with GFP detection to take records of:
    - Option 1 (for GFP-fibrillarin/GFP-NPM1 cells) the DNA content (and contacts) of nuclear slices ± nucleoli, or
    - Option 2 (for GFP-lamin cells) the DNA content of nuclear slices that are apical or equatorial (based on whether the lamina pattern describes a spherical perimeter or a surface) to map lamina-associated DNA and the long-range contacts that occur in those nuclear positions.
- ChIA-PET (**Ruan**): Budget and priority issues. We may be able to contribute 2 datasets.
- TSA-seq (**Belmont**): Budget issues- right now we are on target for delivering TSA-Seq data from 4-5 cell lines for ~4 different compartments for the entire grant period. There are no funds for production- just technology development.
- I already see a problem: the proposed time points for just the cell cycle analysis (18) exceed by several fold the sum total of all data collection for TSA-Seq that we budgeted for during our entire 5-year grant period
- Repli-seq (**Gilbert**) - if AICS would track PCNA and place it into the context of its nuclear landmarks modeling scheme, then repli-seq would provide the DNA sequences that are replicating at the times and nuclear locations that are represented by the spatial patterns seen in real time by tracking PCNA - if other sub-nuclear bodies and landmarks are

mapped relative to PCNA by microscopy, that would then put the repli-seq into the context of the 3D organization at unprecedented resolution, and thus, indirectly all the things that replication timing seems to proxy. In other words, if we knew where in the nucleus relative to all the measurable landmarks DNA synthesis was taking place at various times (via PCNA), then we could overlay repli-seq to infer what sequences are at those locations throughout interphase, knowing that interphase chromatin does not re-position (<0.5uM motion), and we could do it for all sequences genome-wide since replication is a whole genome process.

- High throughput FISH (**Dekker/Misteli**)

Aim 3. Integrating nuclear protein-based landmarks with chromatin sites and gene expression (NOFIC and AICS).

Given the diverse collection of data to be collected in Aim 1, we will develop, validate and apply multiple analytic methods for data integration, all of which aim toward the ultimate goal of producing dynamic 3D models of individual cells, in which genomic coordinates are registered relative to other nuclear landmarks in Euclidean space. Six 4DN labs plan to contribute analyses toward this end.

Alber lab (University of California Los Angeles)

We will integrate volumetric data about (i) nuclear volumes, (ii) the locations and shapes of nucleoli, (iii) as well as speckles into our genome structure modeling protocol. In other words, we will include single cell volumetric grids from about 1000-10,000 cells as spatial constraints with data from Hi-C, lamina DamID and potentially TSA-seq data to generate a population of genome structures consistent with all data. Imaging data will improve genome structure modeling and will allow prediction of chromosome and gene locations in imaged cells. For instance, we would generate several hundred or more genome structures per individual nuclear image to generate probability density distributions of genes for each imaged cells.

Ma lab (Carnegie Mellon University)

We will primarily explore three directions based on our ongoing work.

- We will further develop and apply a new neural network architecture to better capture biological imaging data that we published recent (Chidester et al. ISMB 2019). The effectiveness of neural networks (e.g., CNN) can be further enhanced by encoding invariance to uniformative augmentations of the data. In many bioimaging data types, especially microscopy data, a key invariance is rotation. However, the encoding of rotation equivariance and invariance into CNNs to learn meaningful features for cell phenotyping was not explicitly explored before. We plan to consider the integration of rotation equivariance and invariance to analyze the localization of markers in context of nuclear organization from the microscopy images. We will first apply our efficient rotation-equivariant convolutional scheme, called conic convolution as an effective

alternative to group convolution. The information captured by the neural network can be used for various other tasks, including the integration with genomic data.

- We will further develop and apply our recent work in graph representation learning to analyze single cell Hi-C (scHi-C) data by modeling the cell-to-cell variation of chromatin interaction as “hyperedges”. Specifically, we have developed a self-attention-based graph neural network called Hyper-SAGNN for the representation learning and hyperedge prediction of hypergraphs (Zhang et al. arXiv:1911.02613 2019). By applying our algorithm to several scHi-C datasets, we found that our method achieves great performance in embedding scHi-C datasets with continuous or discrete states. It also has the ability to incorporate other related single-cell datasets to study 3D genome variability at single-cell resolution.
- We will apply the integrative probabilistic framework we developed in our NOFIC project to reveal genome-wide compartmentalization by combining TSA-seq, DamID, and Hi-C. These spatial localization states and their associations with various types of functional genomic properties (e.g., DNA replication timing) can be further compared with structural modeling results from the Alber lab and the imaging data from the AICS.

Libbrecht Lab (Simon Fraser University)

We are interested in developing machine learning and statistical methods that draw connections between omics, imaging and genomic sequence. Some potential lines of research include:

1. Produce annotations of domain state that integrate Hi-C, histone modifications and localization (such as DamID and TSA-seq), using a joint statistical model.

2. Reduce a 2D Hi-C matrix to a 1D reduced representation such that the full 2D matrix can be inferred from the 1D representation. Such a reduced representation by construction captures all the features of a given genomic position that contributes to 3D conformation, and therefore identifies the structure-defining elements in the genome.

3. Use a pairwise prior to integrate imaging data into a model of omics data. An imaging data set may tell us that two structures are nearby without indicating which specific genomic positions take part in that structure. Such information can be incorporated into a statistical model using a pairwise prior, which states that two variables are expected to be equal. This technique is applicable to several types of models, including domain annotation models.

Noble lab (University of Washington)

We will begin by building coarse-scale, consensus models that assume that the relative locations of key nuclear landmarks are invariant across cells. For example, such a model might only attempt to resolve DNA at the level of chromosome territories, using an optimization procedure that takes into account the sequence-based assays -- Hi-C, sciHi-C, PLAC-seq and GAM -- while also incorporating restraints from AICS imaging data. Note that data from one or more experimental assays may optionally be excluded from the optimization, for the purposes of

validation of the resulting model. Furthermore, one benefit of this coarse consensus approach is the feasibility of gathering additional, independent measurements via FISH to confirm aspects of the inferred model.

One challenge associated with the ensemble modeling approach is that the number of parameters in the inferred ensemble potentially exceeds the number of observations in the data. In such cases, unless aggressive regularization is employed, the resulting optimization may be under-determined. This project will directly address this challenge by bringing to bear a much larger collection of data than has been employed previously, thereby further restraining the models.

We will also explore an alternative way to reduce the number of parameters in the model by creating a representative set of consensus models. The key idea here is to identify, from the live cell imaging or sciHi-C data, a representative set of cells whose statistical properties mimic, as closely as possible, the statistical properties of the entire collection of cells. This type of selection can be carried out using submodular optimization techniques that have been used successfully in a variety of other fields. We will then build one consensus 3D model for each of the cells in the representative set, taking into account data derived from that cell as well as nearby cells in the population. This aggregation will be carried out in a weighted fashion, so that for each model the data from neighboring cells is given more importance than data from distal cells.

A key challenge in the representative set consensus modeling approach is finding a mapping between single cells in the sciHi-C, GAM and live cell imaging data sets. This is a particularly hard problem because the data to be integrated do not share any dimension; i.e., the cells being assayed are different, and the measurements produced by each assay are also different. This challenge will be addressed by optimization methods being developed in the Noble lab that jointly learn multiple embeddings, one for each data type, into a shared latent space. The key idea is that the distributions of cellular locations in the embedded space should be similar across multiple data types. This optimization can be solved by employing a differentiable function that measures the distance between distributions, such as the Wasserstein or earth mover's distance. Given these learned embeddings, we will select representative cells from, e.g., the live cell imaging data and then build a consensus model for each one, while also taking into account the sciHi-C and GAM data from cells that are nearby in the latent space. We will also incorporate restraints that capture properties derived from the population assays (Hi-C and PLAC-seq).

Nicodemi lab (University of Naples "Federico II")

The Nicodemi lab has developed models of polymer physics that can describe chromatin conformations at the single-molecule level and resolutions down to 10nm. They consider the scenario where contacts between distal DNA sites are mediated by molecular binders such as TFs. They have been shown to explain Hi-C/GAM maps genome-wide with good accuracy (correlation 0.95). By Machine Learning we infer the location of the minimal set of putative DNA

binding sites required to explain folding, based only on contact matrices (e.g., Hi-C, GAM, etc.), without previous knowledge of TFs or epigenetic tracks. Next, we can integrate information on our predicted binding sites and binding molecules with microscopy, combined with epigenetic and genomic data. In particular, in an approach similar to those described for consensus polymer models, we can integrate the microscopy derived spatial positioning of nuclear proteins and factors with the ensemble of 3D structures computationally predicted. Additionally, the above results will be combined with Bayesian methods developed to infer the network of the traversed single-cell states during cell transformation processes. The aim is to produce a comprehensive description of the molecular mechanisms underlying chromatin regulation at the single-molecule level, and its dynamics during the cell cycle and stem cell differentiation. A strong synergy with the other involved labs would maximize results and impact of the different approaches.

Neretti lab (Brown University)

The Neretti lab, in collaboration with the Srivastava lab at FSU, has developed a methodology to infer the 3D structure of chromosomes from single cell Hi-C data by leveraging information contained in bulk Hi-C data. This approach utilizes a Bayesian framework, with the bulk contact matrix as a prior to overcome the sparsity of single cell Hi-C contact matrices. We will apply this methodology (SIMBA3D) to single cell Hi-C data of WTC-11 cells to infer the 3D structure of all individual chromosomes. We will generate an ensemble of solutions for each cell and will characterize cell-to-cell variability.

Liu lab (University of Michigan)Jointly embedding multiple single-cell omics measurements:

The Liu Lab has developed a *dynamic graph embedding approach* to characterize the dynamics of chromatin organization over a sequence of chromatin contact maps, such as those from differentiation. This approach uses dynamic graph wavelet transformation (DGWT) to map the sequence of graphs into a Fourier domain where we can evaluate the connectivity of the vertices (a.k.a., genomic loci) at different scales efficiently and effectively. Denote the embedding of vertex v_i at time t by $z_{i,t}$, which is a vector of length k, corresponding to k scales used to evaluate the structural role of node v_i. Therefore, the trajectory of vertex v_i at these T time points in this k-dimensional Euclidean space characterizes the dynamic of chromatin organization at the genomic locus corresponding to vertex v_i. The Liu Lab is interested in applying this approach on the sequence of chromatin contact maps and identifying the changing structures within the process.

Another potential contribution from the Liu Lab is *an improved MMD-MA* algorithm, which relaxes previous assumption used in the MMD-MA algorithm that the cells from different views are drawn from the same population, and these cells form similar manifold structures in their own views. This assumption may not hold in the situations that (1) the single cell sequencing platforms have different selection bias towards different cell types which will result the cell populations from different views have different compositions, and (2) manifold structures from

different views are quite different (e.g. two cell types have moderate transcriptional difference but their chromatin accessibility profiles are quite similar). Therefore, we plan to apply the improved MMD-MA algorithm on AICS-NOFIC single cell datasets, and evaluate whether it can provide a better alignment than previous MMD-MA algorithm.

Additional unstructured text

What are the scientific questions that could be addressed by integrative analysis?

- Could we use replication foci to anchor all the rest of the data (omics and imaging) to obtain view of genome reorganization?
- What is the shape of the compartments, domains, loops in live cells?
- Is there a disease angle? For example, cardiomyopathy?

General comments for discussion (ASB):

1. How does iPS differentiation into cardiomyocytes conceptually differ from differentiation of H1 cells into definitive endoderm? Which should be a priority if a group can only choose one to investigate in depth?
2. It seems we are back to a discovery mode to collect all data possible for cell cycle and one differentiation model. But this is sort of looking to me like just a different flavor of the JAWG project- especially the differentiation system.. A few subprojects will be qualitatively different and benefit greatly from the Allen Institute live cell data- for instance the Frank Alber ensemble modeling in which realistic parameters can be used for modeling. But much of the proposed work seems like it will be just another version of the JAWG project, with just a different differentiation model. Do the different groups have adequate budgets and manpower, or will this dilute efforts into two projects, both underfunded in support and people. The cell cycle project could be interesting but it raises budget issues again.
3. Both the beginning and end states have problems of significant cell and nuclear shape changes that will complicate live-cell imaging. (I assume the cardiomyocytes will start beating at some point near the end state of differentiation.) Is there a way to pick specific scientific questions we want to answer and choose a subset of data points to solve these questions?
4. What is the ultimate motivation for this NOFIC Center collaboration? I thought part of it was to act as a bridge for another 5 years of funding.
5. Are there any alternatives that still exploit the special resources of the Allen Institute that might fit better into our time line and budgets?
    a. For example: Could we focus on two genotypes and two states?
        i. Wild type versus mutant iPS cells? Here the idea would be as proof of principle to pick a mutation that is known to perturb nuclear organization- for example a lamin A mutation associated with a cardiomyopathy manifesting itself during aging.
        ii. Undifferentiated versus differentiated? Example- use the cardiomyocyte differentiation system with the lamin A cardiomyopathy mutation.

Susan - not sure how well F-net works with nuclear structure;

- Possible solutions - FISH (super-res)?
- Machine learning may pick up information that we do not see by eye, but existing
- Machine learning may also pick up indirect connections between entities.
- Landmarks such as ends of chromosomes or repetitive sequences?
- How to tag a loci or multiple loci? - dCAS9, TALE-proteins,
- Cell state - iPSC, cardiomyocytes, mitosis,

To do - draft a plan for this,

Lower hanging fruits - David G (PCNA); Jay/Bill, stereotypical localization (in population); currently, nucleolus could be imaged very nicely.

Question - how to use deep learning to incorporate the stereotypical localization data from population? Could this be a mini-project? Can we combine CTCF/Cohesin localization data with live cell fluorescence data?

Additional information - some genes are on or off, how that relates to other things? Live imaging followed by FISH on the same cells. This is happening now at AICS.