*New call info:
Phone:
+1-415-655-0002 US Toll
Access code: 800 040 314
URL:
https://4dnucleome.webex.com/join/4dnomics

At some point, the meeting ended. If that happens to you, please try to rejoin.

Goal: the purpose of this document is for the 4DN-omics standard group to come up with a set of recommendations for the Hi-C dataset to be generated and shared by the consortium.

Please provide feedbacks on the type information that should be considered in reporting the HiC data sets:

An annotated experimental protocol for Hi-C.

- An experimental protocol in use in Ren lab can be found here.
- Lin
- Job
- An editable version of the original *in situ* Hi-C protocol from the Aiden lab (Rao & Huntley et al., Cell, 2014) as well as a guide to QC standards can be found here. This document might serve as a good place to insert comments and alternative steps, since I would assume most extant in situ Hi-C protocols reflect modifications of this protocol.

  Note: Rao & Huntley et al., Cell, 2014 also reported Hi-C maps exploring ~100 different sets of experimental conditions; the conditions are listed here; the data can be found here and is extremely useful for making comparisons between protocol variants (ie tethered/dilution/insitu; crosslinking conditions; etc). In our experience, we have not found any conditions where nuclei can be isolated and where the results of *in situ* ligation are not as good or better. The pipeline that was used for this paper is available here: http://www.aidenlab.org/juicer/

Other protocols (such as TCC, ChIA-PET, 4C-seq, etc) will be discussed too. Please share these protocols here as well

- TCC
- DNase Hi-C
- ...

Data process procedures

- Ren lab
- Dekker lab
- Aiden Lab:

Juicer (Open Source Data Analysis Pipeline):
http://www.aidenlab.org/juicer/
https://github.com/theaidenlab/juicer

- Mirny lab: computational metrics for Hi-C data from Mirny Lab & DCIC can be found here

Job - the protocol optimal for one cell type might not be so for another; therefore variations are justified; flexibility should be permitted.

- Crosslinking procedures: length, concentration, temperature
- Ligation condition (volume)
- Inactivation conditions (SDS)

Job's input -

- cis - trans ratio - the range could be indicative of whether the experiments succeeded or not.
- PCR duplicates - 2%?
- Decay function - interphase or metaphase
- AB compartments
- Certain chromosomes interact with each other

Lin - needs **quantitative metrics to evaluate variations of different parameters**.

External benchmarks. - FISH. GAM (data yet not forthcoming).
List of known and universal chromatin interactions: some CTCF-CTCF interactions? Some of the strongest loops.

Metadata Standards

Information items	Should this be included? Why?	How to report the information?
Protocol information (e.g. in situ HiC, restriction enzyme used, crosslinking method, etc)		*Full reference protocol.*
Metadata (*who performed the*		

experiments, when, protocol specific information such as biotin labeling, amplification cycle numbers, etc)		
Cell sample information (*cell pellet characterization, tissue processing procedures, cell line passage numbers, etc*.)		
Sequencing information (instrument, sequencing depth and length, optic duplicates, etc)		
Sequencing QC information (read quality metric, etc) *[[[Sequenced Read Pairs, Normal Paired, Chimeric Paired, Chimeric Ambiguous, Unalignable, Ligation Motif Present, Alignable (Normal+Chimeric Paired), Unique Read Pairs, PCR Duplicates, Optical Duplicates, Library Complexity Estimate, Intra-fragment Read Pairs, Below MAPQ Threshold, Hi-C Contacts, Ligation Motif Present, 3' Bias (Long Range), Pair Type % (L-I-O-R), Inter-chromosomal, Intra-chromosomal, Short Range (<20Kb), Long Range (>20Kb)]]]*		
Initial data processing (alignment algorithms, *with parameters used and reference genome*)		
Sample description (Species, cell type, treatment condition, etc)		
Percentage of cis reads		
Reproducibility metric		

Resolution metric		
Raw chromatin contact matrix		
Normalized contact matrix		
Chromatin domain calls		
Chromatin loop calls		

Data Release

Geoff (Mirny Lab)
It would be useful to separate crucial upstream metadata items (protocol, sample, …) from intermediate dataset-quality items (cis reads, pcr duplicates, …) from downstream analysis/annotation items (domains, loops, …). Particularly as the latter are pipeline-dependent and may require some serious comparative analysis of pipelines (e.g. by AWG).

Nezar (Mirny Lab)
I suggest agreeing on a layout for ***data serialization***. Why? To minimize the assumptions any new codebase needs to make about the data + allow ease of ***interchange and interconversion*** into more application specific formats, pipelines, visualization tools, etc.
i.e. a “universal”, flat, common currency format (this is usually text)
1. decide on the minimal necessary fields to be provided
2. sort the data in “upper triangular” fashion (not absolutely necessary, but would be nice!)

- (chrom1, start1) <= (chrom2, start2)
- lexsort by (chrom1, start1, chrom2, start2)

contact records (read pairs): e.g. BEDPE is an informal standard
binned contact frequencies: similar tsv file for one resolution
normalization weights: e.g. fits nicely in a BED file of normalization weights for genomic bins

Basically, tools should be expected to be able to consume and produce these serializations, even if they use other optimized formats or databases. These need to be as simple as possible and not require specialized tools to parse -- bioinformaticians love flat text files for good reason.

The counterparts for other omics file formats is tab-delimited files like BED (which can be compressed and indexed with tools like bgzip+tabix). Jim Kent’s big binary files (bigWig, bigBed) are associated directly with a flat text counterpart. The same applies to SAM/BAM, VCF/BCF.

Bryan (Dekker Lab):
I agree 100% with Nezar above - this is how we _*already*_ handle data internally in the Dekker Lab (**BEDPE - upper triangle, tsv format**). Normally we compress via gzip.

We keep:

itx_classification, readID, chr_1, pos_1, strand_1, chr_2, pos_2, strand_2

Deciding on file format standards seems to be our first priority. TEXT format with data serialization is a good option. Standard input/output file formats would make comparing different "modules" of each pipeline much easier.

One further idea would be to only distribute RAW contact matrices along with a vector (BED) file containing the bias factors for each bin. Then balanced matrices (data) can be extracted/computed on the fly from the combination of the bias factors + raw contact matrix. This could also facilitate different balancing methods from multiple groups.

Aiden Lab

I'm including an overview of the primary file formats we use, as well as a the codebases that we use to generate, manipulate, and visualize *hic* data. All of these formats and codebases are available in an open-source fashion.

Note on contact matrices. My lab does have a file format for individual contact matrices. Our infrastructure is built around being able to zoom in and out across resolutions, for which you need to have multiple resolutions available in a single file.

Juicer (Data Analysis):

(*fastq* format-->*hic* format)
Multi-resolution contact matrix ensembles stored in *hic* format
http://www.aidenlab.org/juicer/
https://github.com/theaidenlab/juicer

(hic-->various annotation formats)
Includes *arrowhead* (domain caller); *hiccups* (loop caller)

Juicebox (Data Visualization):

Reads *hic* format
Google Earth style display:
https://github.com/theaidenlab/juicebox
http://www.aidenlab.org/juicebox/
VR display:
https://play.google.com/store/apps/details?id=com.visor.JuiceboxVR&hl=en

Job Dekker

I think we need to set up files such that the datafiles are separate from (but compatible with) analysis and visualization formats. The first files that need to be shared are just the data itself (and not some analyzed/transformed form of the data), in a format that can be then used for any later analysis or visualization. Analysis and visualization is no doubt dramatically going to change in the years to come. Hence I am in strong support of Nezar's proposal.

Metadata suggestions (Mirny Lab)

We can learn from the experience of Structural Biology that uses PDB database to store coordinates of protein/DNA/RNA structures. Here is a description of their file standards with the metadata
http://www.wwpdb.org/documentation/file-format-content/format33/sect2.html

Working Group Scope for Omics Data Standards (ODS) and Data Analysis (DA)

ODS will cover all issues pertaining to defining vocabulary in light of what is already known in the field. This includes specifying standards for protocols (What can be called a "Hi-C experiment"? When can a particular experiment be called "successful" as opposed to “failed”? What forms of bias is each experiment susceptible to?) and for the data that they produce (such as metadata reporting requirements and standardized data formats). It also includes creating actionable definitions of features including loops and domains, thus defining "what" these features are.

DA will cover all issues pertaining to entirely new classes of features as well as issues pertaining to how to best use 4DN data to make new discoveries. DA will also be responsible for identifying the best methods for identifying features as defined by ODS. Therefore DA will be responsible for the question of "How" we can best identify features given how the feature has been defined by ODS. (Thus the ODS discussion of a loop would look like: "What is a loop?" "A loop is a pair of loci that satisfy requirements X, Y, Z"; the subsequent discussion in DA would be: "How do we identify pairs of loci that satisfy requirements X, Y, Z accurately and efficiently at consortium scales?" "Algorithm A and implementation B are the best.")

In addition to the currently scheduled meetings for ODS and DA, we will add an additional, monthly meeting, which will be joint between ODS and DA (ODS+DA). Aside from providing an opportunity for shared interaction, this meeting will have its own scope, and will address two specific issues that require the closest coordination among the WGs. These two issues are: (1) creating gold standard examples of feature annotations (such as a genome-wide loop map that would be used as the basis for subsequent algorithm development); or at the very least creating an approach and sets of criteria that can be used to create such standards; and (2) Computational assessment of emerging protocols in terms of how well they perform with reference to these gold standards.