V6 – Analyzing 3D chromatin conformation

Chromatin conformation has large implications on gene expression, but is usually ignored in expression analysis.

Program for today:

- 3D chromatin conformation
- Hi-C method
- Biases in Hi-C data analysis
- integrated analysis of multiple data sources

In lecture 6, we will discuss methods that characterize the three-dimensional conformation of chromatin in the cell nucleus.

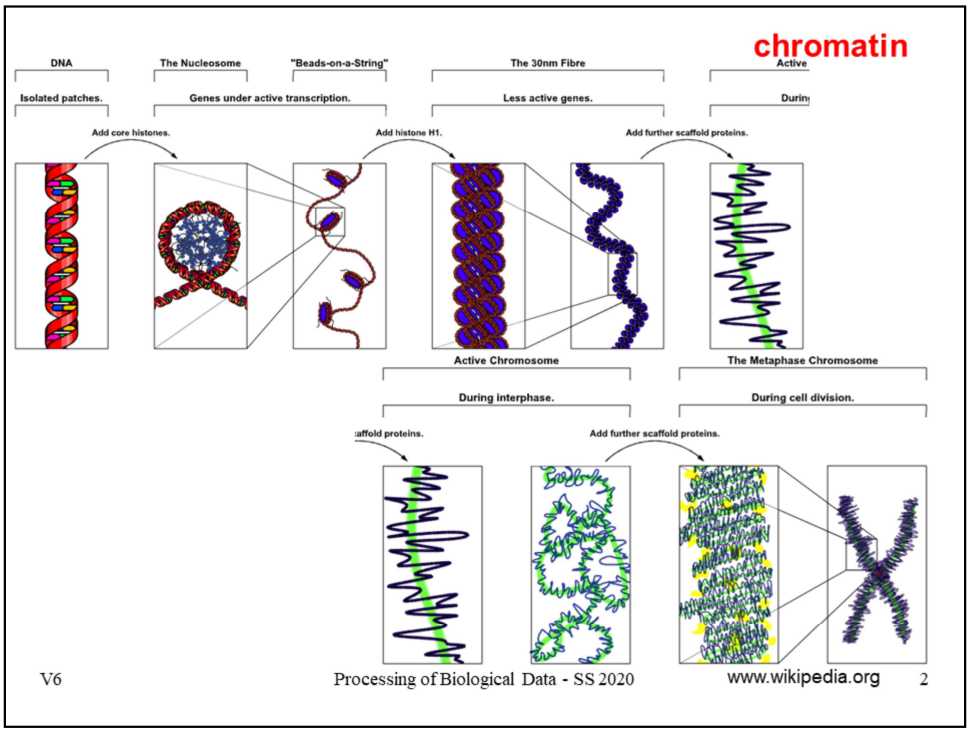
As you known, the 2 m long DNA needs to be drastically compacted in order to fit into a tiny nucleus of a eukaryotic cell (diameter ca. 6 micrometer in mammalian cells).

We will start with a short introduction of the three-dimensional conformation of chromatin.

Then, we will discuss the principles of the so-called Hi-C method that is able to provide information on the chromatin conformation.

Every experimental method may have biases. This is also the case for Hi-C. This means that bioinformaticians need to develop methods to correct for these biases.

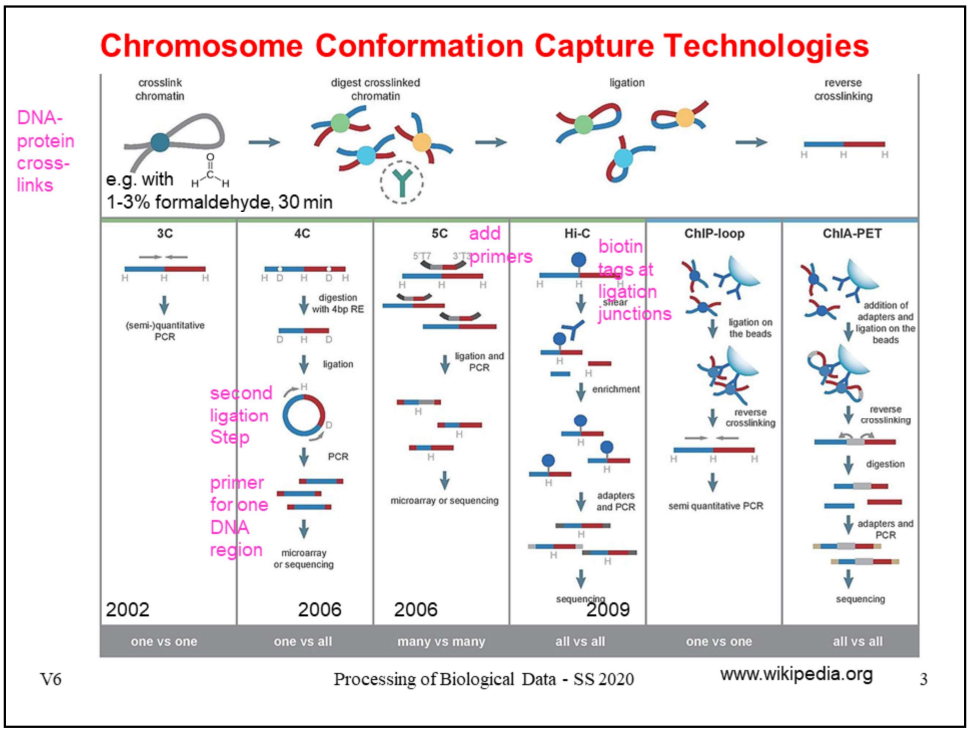
Finally, we will discuss a computational study that integrated evidence from multiple data sources to resolve details about the chromatin conformation.



This figure is taken from the „chromatin“ entry of Wikipedia.

When bioinformaticians speak about gene expression, they either think of the bare DNA strand (top left), or when DNA is wound around nucleosome particles consisting of histones (next figure to the right).

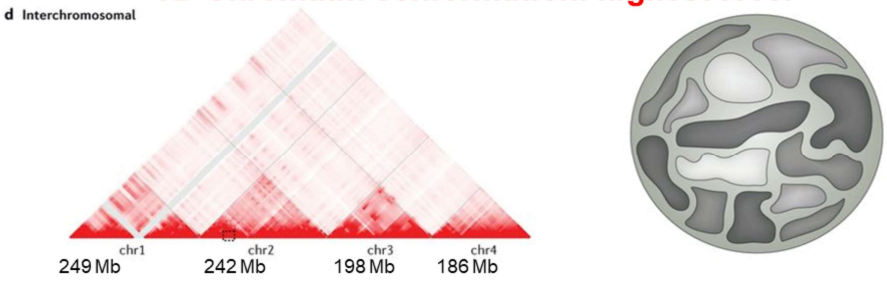
But DNA needs to be further compacted until the final structure of a chromosome pair (bottom right).



This figure is taken from the Wikipedia entry on „Chromosome_conformation_capture". As in the ChIP-seq method, the formation of DNA-protein crosslinks is induced by formaldehyde (see lecture 5, slide 4).

The genome is then cut (or: digested) into fragments with a restriction endonuclease enzyme. The size of restriction fragments determines the resolution of interaction mapping. Certain restriction enzymes (REs) such as EcoR1 or HindIII make cuts in 6bp recognition sequences. They cut the genome on average once every 4^6 = 2^12=4096 bp, giving ~ 1 million fragments in the human genome. (Hint: the recognition sequence of EcoR1 is G/AATTC. The cut is made after the initial Guanine base. Assuming a random sequence, where every nucleotide has frequency ¼, GAATTC sequences occur randomly every 4096 bps). For more precise interaction mapping, a 4bp recognizing RE may also be used, that will generate shorter fragments. In the next step, two ends are ligated by a DNA ligase enzyme. Cross-links are then reversed and the ligation mixture is purified. This is followed by quantitative detection of 3C or higherC ligation products, e.g. by PCR. There are many variants of the original 3C method. We will not discuss their differences here. In the Hi-C protocol, one uses high-throughput sequencing to determine the identity of the two ligated sequences.

3D Chromatin conformation: highest level



Data from human GM12878 cells (lymphoblastoid cell line). **Nucleus**

At the highest-level of 3D organization *trans*-interactions are rare and individual chromosomes (chrs) occupy distinct **territories** (denoted by irregular shapes) within the nucleus (grey circle).

Gene-rich chromosomes are preferentially found inside the nuclear core and gene-poor chromosomes are localized close to the nuclear membrane.

Bonev & Cavalli, *Nature Rev Genet* **17**, 661–678 (2016) |

This is the link to the Bonev & Cavalli paper:
https://www.nature.com/articles/nrg.2016.112

We continue our review of the three-dimensional conformation of chromatin.

At this highest level of genomic contacts (left picture), one clearly sees that many contacts exist within individual chromosomes and few contacts exist between chromosomes.

The right picture symbolizes the nucleus. Distinct „territories“ are represented by darker or brigher colors.

Each chromosome is located in a particular territory.

Possibly, the nuclear core provides more conformational freedom to pack and unpack the chromatin. Here, one finds chromosomes containing many genes.

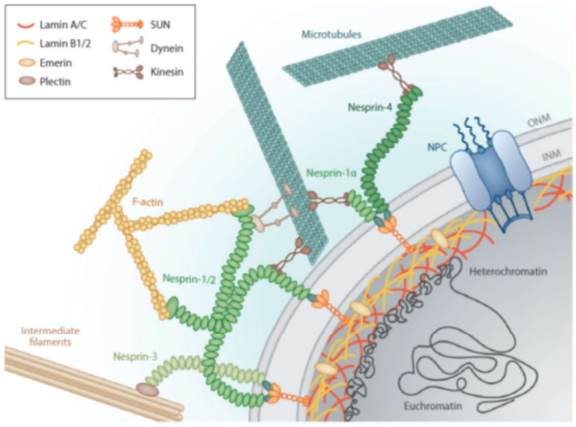
Gene-poor chromosomes tend to be at the periphery of the nucleus, close to the nuclear membrane.

Around the nuclear membrane

The nuclear envelope consists of outer and inner nuclear membranes (ONM and INM, respectively) separated by the 30–50-nm-wide perinuclear space (PNS).

Below the INM exists the 10–30-nm-thick, fibrous meshwork of the nuclear lamina.

The nuclear lamina is composed mostly of lamins, which are nuclear intermediate filament proteins.



Packed **heterochromatin** exists at the nuclear periphery.

Maurer & Lammerding, *Annu Rev Biomed Eng* **21**, 443 (2019) |

V6 Processing of Biological Data - SS 2020 5

This figure shows the double-layer composition of the nuclear membrane.

At the outside, microtubules (shown as sheets) and intermediate filaments connect to it.

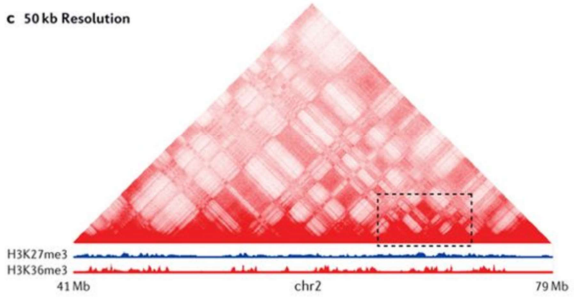
At the inside is a meshwork, the nuclear lamina containing lamin proteins.

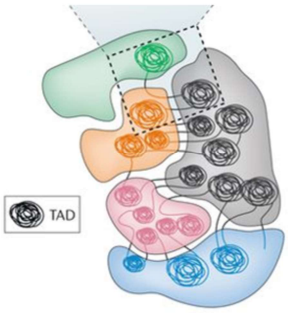
We will revisit these lamins at the end of this lecture.

This architecture suggests that the nuclear membrane will be quite stiff.

Any molecule that comes close to this stiff membrane will probably experience a reduced conformational flexibility.

3D Chromatin conformation: 50 kb resolution





Different topological domains with similar epigenetic signatures are characterized by stronger inter-domain interactions.
They are organized into **compartments**.

Here, blue and grey represent the active compartment, whereas interactions between green, orange and red **topologically associating domains** (TADs) form the inactive compartment.

Bonev & Cavalli, *Nature Rev Genet* **17**, 661–678 (2016) |

The left figure shows a 28 Mb region of the 242 Mb long chromosome 2.

Note the much higher resolution than on the previous slide 4.

On the next slide, we will zoom even further into the dashed area.

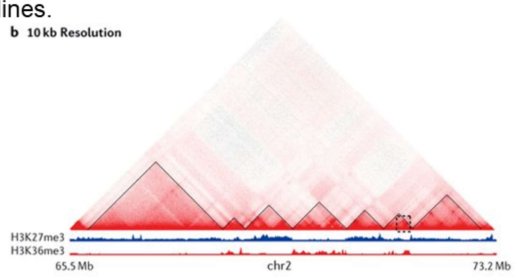
In the right figure, we see five differently colored so-called TAD domains.

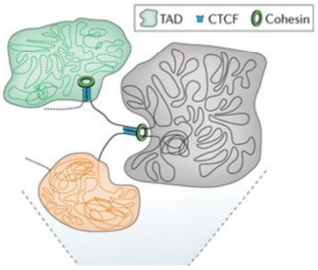
These are regions containing either actively expressed genes or inactive genes.

In the figure, this is represented by looser contacts between the balls in the blue and grey TAD domains.

3D Chromatin conformation: 10kb resolution

(left) ca. 8 Mb region containing several TADs that are manually annotated with solid lines.





(right) 3 different TADs, enriched for either active marks (H3K4me3 and H3K36me3; grey), Polycomb (H3K27me3; green) or heterochromatin (H3K9me3; orange) are schematically represented in the 3D space.
CTCF proteins are shown as blue rectangles and loop-extrusion complexes (potentially cohesin) are depicted as green circles.

Bonev & Cavalli, *Nature Rev Genet* **17**, 661–678 (2016) |

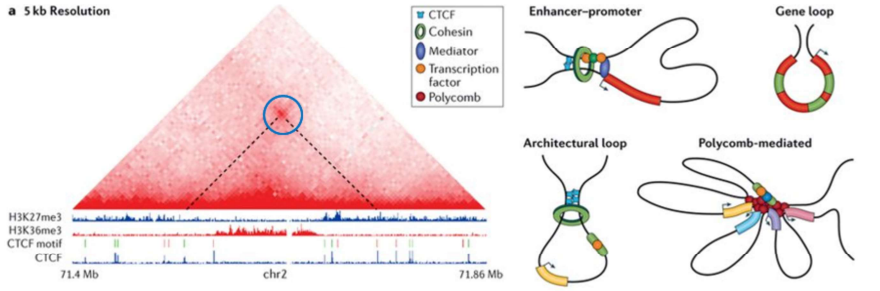
A very interesting recent discovery was that chromosomes are spatially segregated into sub-megabase scale domains, called topologically associating domains (TADs).

TADs typically manifest as triangles in Hi-C maps, in which regions within the same TAD interact with each other much more frequently than with regions located in adjacent domains.

The spatial partitioning of the genome into TADs correlates with many linear genomic features such as histone modifications and coordinated gene expression.

3D Chromatin conformation: 5kb resolution

(right) Examples of different types of chromatin loops that can potentially reside within a domain



(left) : example of an architectural loop (circled blue) as seen in high-resolution Hi-C data. Regions participating in loop formation are demarcated with dotted lines). Also shown are CCCTC-binding factor (CTCF)-binding profile and CTCF motif orientation.

Bonev & Cavalli, *Nature Rev Genet* **17**, 661–678 (2016) |

V6 Processing of Biological Data - SS 2020 8

In vertebrate genomes, *cis*-regulatory elements, such as enhancers, are separated from their target genes by relatively long distances along the linear genome.

In order to elicit its effect, an enhancer is brought into close spatial proximity with its target promoter through the formation of a 'chromatin loop‘.

The left figure shows dense contacts of neighboring regions along the x-axis and one peak (marked by a blue circle) between the two regions connected by dashed lines.

The right figure shows four examples how such loops can form. Long-range chromatin contacts can bring an enhancer region into close proximity of a promoter.

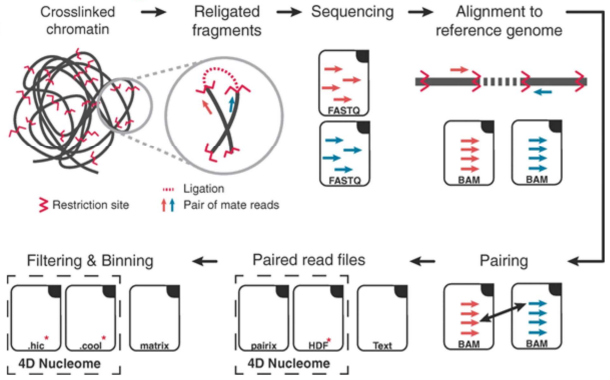
In a 'gene loop’ (primarily identified in yeast), the transcription termination site of a gene loops back to make contact with its own promoter. Gene loops have been suggested to reinforce the directionality of RNA synthesis from the promoter.

Anchors of cell-type-specific loops are often the promoters of differentially expressed genes and contain binding sites for the architectural protein CTCF.

Spatial associations between actively transcribed co-regulated genes in mice, between Polycomb-repressed genes in *Drosophila melanogaster* and more recently in mammalian cells have also been observed.

Processing data from HiC

Schematic representation of Hi-C data analysis, starting from a cartoon depicting cross-linked chromatin and a prototypic pair of mate reads positioned on the restriction fragments from which they originate.



Raw sequencing paired-end reads (in FASTQ files) are aligned to the reference genome considering the mate reads independently. Then, aligned reads (in BAM files) are assigned to their fragment of origin and paired. The paired reads are stored in a sorted file. Finally, after filtering and binning, the read counts are stored in contact matrix files.

Pal et al. Biophys Reviews 11, 67–78(2019)

Link to Pal et al: https://link.springer.com/article/10.1007/s12551-018-0489-1

The alignment of NGS reads to the genome is, in principle, a standard task. However, tor Hi-C reads, alignment may be challenging if the read spans the ligation junction.

Then, two portions of the read will match distinct genomic positions. These are also termed “chimeric reads”.

Aligned reads are then filtered to remove spurious signal due to experimental artifacts. Read filtering is particularly important for Hi-C data as multiple steps in the experimental protocol can generate biases in the sequencing results. Read level filters include the removal of reads with low alignment quality or PCR artifacts, i.e., multiple read pairs mapped in the same positions.

Then, read pairs filters are based on the distance of aligned reads to the downstream restriction site, which is used to estimate if the read pair is compatible with the expected size of sequenced fragment obtained from the ligation product (see slide 14).

Moreover, read pairs can be filtered if they are mapped on the same fragment, thus resulting from lack of ligation or self-ligation events, or if their orientation and distance in mapping positions is compatible with an undigested chromatin fragment.

Data from HiC

$n \times n$ contact matrix, where the genome is divided into n equally sized bins.

The value within each cell of the matrix indicates the **number of pair-ended reads** spanning between a pair of bins.

Depending on sequencing depths, the commonly used sizes of these bins can range from 1 kb to 1 Mb.

The bin size of Hi-C interaction matrix is also referred to as '**resolution**',

Owing to high sequencing cost, most available Hi-C datasets have relatively low resolution such as 25 or 40 kb, as the linear increase of resolution requires a quadratic increase in the total number of sequencing reads.

Zhang et al. *Nature Commun* **9**, 750 (2018)

Now we turn to the analysis of HiC-data. The data is typically represented as a contact matrix.

Although the reads are mapped and counted on individual restriction fragment ends, Hi-C data are usually not analyzed at single-fragment level. Instead, the read counts are generally summarized at the level of genomic bins, i.e., a continuous partitioning of the genome in intervals of fixed size ranging from 1 kb to 1 Mb. The rationale behind this approach is that genomic bins allow achieving a more robust and less noisy signal in the estimation of contact frequencies, at the expense of resolution.

Biases in computational analysis of Hi-C data

Procedures including crosslinking, chromatin fragmentation, biotin-labelling and re-ligation can all introduce **biases** that complicate the interpretation of observed contact frequencies.

Efficient and effective removal of multiple systematic biases is critical for the success of any subsequent analysis of C-data as well as for the proper interpretation of results.

Schmitt et al. Nature Rev Mol Cell Biol (2016) 17, 743

Link to this paper: https://www.nature.com/articles/nrm.2016.104

As mentioned, we need to remember that the Hi-C contact matrices are obtained by a complicated multi-step protocol.

All these steps can introduce biases that would lead to misleading interpretations if we do not correct for them.

Random collisions affect chromosome capture data

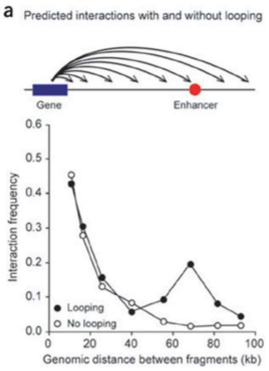
Detection of an interaction between two loci does not necessarily mean that they are engaged in a functional looping interaction.

-> loci along a chromatin fiber will also **randomly**, and quite frequently, **collide** as the result of the inherent flexibility of chromatin.

In general, the **frequency** of random collisions is inversely related to the **genomic distance** between loci (larger "search space" for larger radius).

Thus, relatively frequent but nonfunctional interactions are expected for loci separated by small distances.

For sites separated by larger genomic distances, this 'background' signal decreases rapidly, but remains detectable for sites separated by as much as 150 kb.



Job Dekker, *Nature Methods* **3**, 17–21 (2006)

https://bionumbers.hms.harvard.edu/bionumber.aspx?id=103112:

IJob Dekker is first author on a paper from 2002 (https://science.sciencemag.org/content/295/5558/1306) that presented the 3C method. This paper has been cited 3000 times.

Link to this Job Dekker paper: https://www.nature.com/articles/nmeth823

On this slide, we consider how the distance between two regions of the DNA affects the formation of contacts between them.

Job Dekker et al. reported (middle figure) that, on a length scale of many kb, the frequency decays with the inverse of the distance. For this, we consider DNA as a "cooked spaghetti".

But is this true?

Double-stranded DNA is a polymer. The stiffness of a polymer is typically characterized by its "persistence length" that defines the scale over which a polymer (such as DNA) remains roughly unbent in solution. For DNA, the persistence length has a value of ~50 nm (~150 bp). Thus, on length scales of kb, thermal fluctuations result in spontaneous bending of the DNA and the DNA can be considered as a cooked spaghetti.

Specific contacts affect neighboring loci

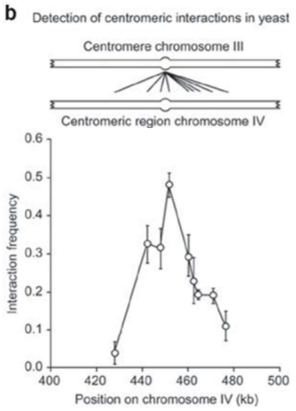
A **specific contact** between two elements located on two different chromosomes—in this example between centromeres—will also **bring neighboring fragments** into **closer proximity**.

Then, they may nonspecifically interact.

Failure to determine a local peak in interaction frequencies may result in incorrectly concluding that two elements specifically interact,

whereas in reality it is their neighbors that are engaged in a specific interaction.

In this example, only the interaction between the two centromeres may be specific (-> highest peak) , whereas interactions with neighboring loci are likely the result of random collisions.



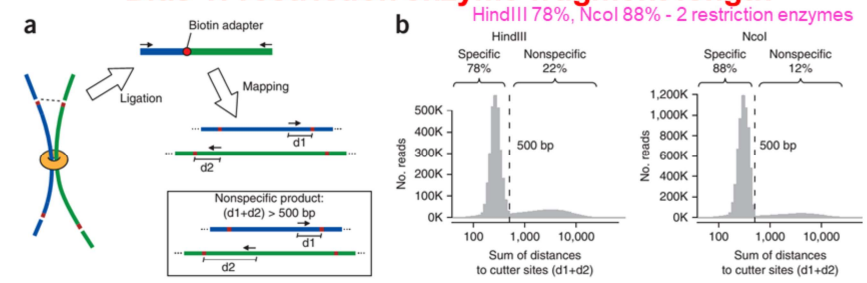
Job Dekker, *Nature Methods* **3**, 17–21 (2006)

If a specific contact is formed in one location, neighboring regions are also close to the „opposite“ DNA regions.

This may lead to the formation of non-specific contacts between adjacent regions which would not form if the specific contact had not formed.

Dekker suggests that only the highest peak should be considered in the bottom figure and the other peaks should be omitted from the analysis.

Bias 1: restriction enzyme fragment length



Hi-C ligation products (shown schematically in **a**) are expected to map near restriction sites because of size selection.

(**b**) For each Hi-C paired read, the sum of distances is computed from mapped Hi-C sequences to the nearest restriction sites. Shown is the distribution of distances.

Two distinct populations of reads are observed, one distributed as expected for normally ligated and size-selected products and one including reads mapped farther away from restriction sites.

Solution: discard reads with distance > 500 bp

Yaffe, Tanay Nature Genet (2011) 43, 1059

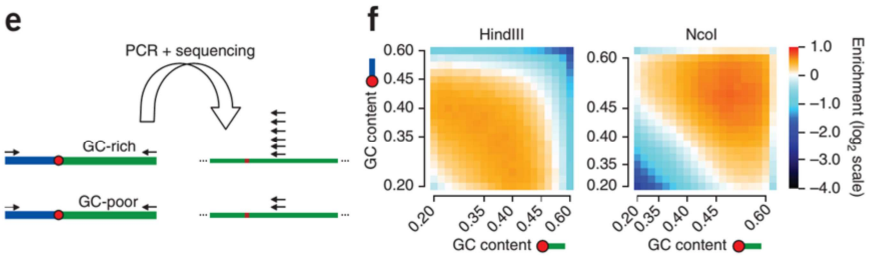
Yaffe & Tanay paper: https://www.nature.com/articles/ng.947

Some Hi-C sequence pairs likely represent **ligation products between nonspecific cleavage sites rather than restriction fragment ends**. This means that the DNA ligase did not merge the blue and green fragments shown in (a) that are connected by a crosslink. Rather, the ligase merged two arbitrary fragments. Such cases are not useful for the analysis of chromatin contacts.

As shown in (b), 22% of the *trans* read-pairs in the HindIII experiment and 12% in the NcoI experiment were mapped with a generally uniform distribution over the restriction fragments, in contrast to the majority of reads that mapped with the expected distribution within 500 bp (the size selection parameter) of the nearest restriction site.

The cleavage and ligation events that generated these reads are unlikely to have occurred on cutter sites. Yaffe and Tanay therefore suggest to discard them from downstream analysis.

Bias 2 : GC content



(**e**) Ligation product processing and sequencing may be biased by GC content. In this schematic example, the GC-rich region produces many more reads.

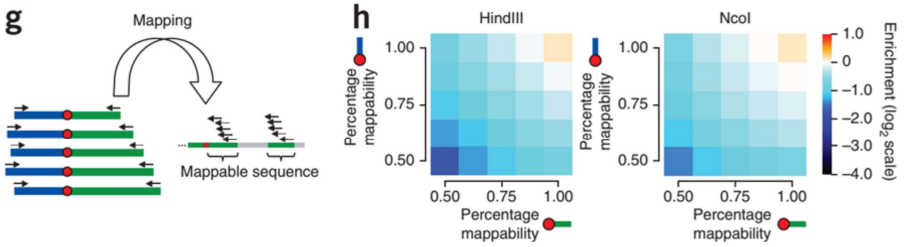
(**f**) Plotting the GC content of the 200 bp near the restriction fragment ends for *trans*-contacts shows intense and contrasting GC biases for the HindIII and NcoI experiments:

NcoI "prefers" GC-rich sequences, HindIII disfavors them.

Another known major source of bias in sequencing experiments is the nucleotide composition of the DNA under study.

Also in Hi-C, some key steps are likely to be affected by the GC content near the ligated fragment ends (e). Analysis of the correlation between the GC content of the 200 bp next to the restriction site and the probability of *trans* contact (f) shows that GC content is a source of incompatibility between the replicates. The GC-content bias maps for the HindIII and NcoI data sets were inversely correlated (element-wise $\rho = -0.14$), providing a partial explanation for a global low correlation between the derived *trans*-contact maps.

Bias 3 : sequence mappability



(**g**) Effect of sequence uniqueness. Different fractions of uniquely mappable short tags are observed next to restriction sites.

As shown in **h**, this has a direct empirical linear effect on Hi-C coverage.

Mappability is predicted and confirmed (**h**) to have a linear effect on the estimated *trans*-contact probabilities.

Yaffe & Tanay correct for biases 2 & 3

by a maximum likelihood approach.

Yaffe, Tanay Nature Genet (2011) 43, 1059

Another genomic variable affecting *trans*-contact probabilities in a purely technical fashion is the mappability (or genomic uniqueness) of the fragment ends (g).

To compute the fragment end mappability score, the whole-genome sequence was split into artificial reads (50-bp reads, starting every 10 bp) and then mapped back to the genome using MAQ. For each fragment end the mappability score was then defined to be the portion of artificial reads mapped uniquely to the genome (MAQ quality > 30) within a 500-bp window starting at the fragment end toward the fragment.

Poisson regression

Poisson regression is a generalized linear model form of regression analysis used to model count data and contingency tables.

Poisson regression assumes that the response variable Y has a Poisson distribution,

$$f(n, \lambda) = (\lambda^n e^{-\lambda})/n!$$

and assumes that the logarithm of its expected value can be modeled by a linear combination of unknown parameters.

If $\mathbf{x} \in \Re^n$ is a vector of independent variables, then we formulate

$$\text{Log } (E(Y|\mathbf{x})) = a + b\,\mathbf{x} = \boldsymbol{\theta}\,\mathbf{x}$$

with coefficients *a* and *b* which can be summarized into $\boldsymbol{\theta}$.

The predicted mean of the associated Poisson distribution is then

$$\lambda = E(Y|\mathbf{x}) = e^{\theta \mathbf{x}}$$

www.wikipedia.org

On the next slide, we will introduce the HiCnorm tool for bias correction.
HiCnorm utilites a mathematical technique termed Poisson regression.
On this slide, we provide some brief background on this method.

HiCnorm tool

HiCnorm corrects for the 3 biases (effective length feature, the GC content feature and the mappability feature) using Poisson regression.

Let $\boldsymbol{U^i} = \{u^i_{jk}\}_{1 \le j,k \le n_i}$ represent the $n_i \times n_i$ Hi-C *cis* contact map for chromosome i, where n_i is the number of consecutive, disjoint 1 MB bins in chromosome i.

u^i_{jk} : number of detected paired-end reads spanning two bins L^i_j and L^i_k ("**raw data**")
x^i_j and x^i_k : effective length feature at loci j and k for chromosome i,
y^i_j and y^i_k : GC content feature at loci j and k for chromosome i,
z^i_j and z^i_k : mappability feature at loci j and k for chromosome i.

Hu et al. Bioinformatics 28, 3131-3133 (2012)
www.wikipedia.org

Link to HiCnorm paper:
https://academic.oup.com/bioinformatics/article/28/23/3131/192582

HiCnorm is an explicit bias correction method.

Here, we will look at the basic steps how biases are estimated and removed.

HiCnorm attempts to correct 3 types of biases. Each one of them is modeled by an independent variable x, y and z.

HiCnorm tool

We assume that u_{jk}^i follows a Poisson distribution with rate θ_{jk}^i:

$$\log(\theta_{jk}^i) = \beta_0^i + \beta_{len}^i \log(x_j^i x_k^i) + \beta_{gcc}^i \log(y_j^i y_k^i) + \log(z_j^i z_k^i).$$

Here β_0^i is the intercept term.
β_{len}^i and β_{gcc}^i represent the effective length bias and the GC content bias, respectively. $\log(z_j^i z_k^i)$ is the Poisson offset term of the mappability bias.

We fit this Poisson regression model, and let $\hat{\beta}_0^i$, $\hat{\beta}_{len}^i$ and $\hat{\beta}_{gcc}^i$ represent the corresponding parameter estimates.
We further define the estimated Poisson rate $\hat{\theta}_{jk}^i$ as following:

$$\hat{\theta}_{jk}^i = exp\{\hat{\beta}_0^i + \hat{\beta}_{len}^i \log(x_j^i x_k^i) + \hat{\beta}_{gcc}^i \log(y_j^i y_k^i) + \log(z_j^i z_k^i)\}.$$

The residual $e_{jk}^i = u_{jk}^i / \hat{\theta}_{jk}^i$ is the **normalized interaction** between two bins L_j^i and L_k^i. This is done separately for *cis* and *trans* interactions.

Hu et al. Bioinformatics 28, 3131-3133 (2012)
www.wikipedia.org

Link to HiCnorm paper:
https://academic.oup.com/bioinformatics/article/28/23/3131/192582

Shown at the bottom is the normalization of the raw data by the estimated Poisson rate of loci j and k.

Cis interactions take place on the same chromosome.

Trans interactions are contacts between DNA regions that are located on different chromosomes.

Biases in computational analysis of Hi-C data

In general, there exist two types of approaches to account for biases in C-data.

(1) account for biases in an **explicit fashion** — by assuming that all sources of systematic biases are known based on biases determined empirically from the observed data.

(2) account for biases in an **implicit way** — by assuming no known source (or sources) of bias, and assuming that the cumulative effect of the bias is captured in the sequencing coverage of each locus (or 'bin').

As Hi-C is a genome-wide assay, the implicit models assume that each locus should receive **equal sequence coverage** after biases are removed.

Implicit models all rely on some implementation of **matrix-balancing algorithms**.

Schmitt et al. Nature Rev Mol Cell Biol (2016) 17, 743

Link to this paper: https://www.nature.com/articles/nrm.2016.104

Schmitt et al. recommend that researchers should analyse their data using both the explicit and implicit approaches to ensure the biological relevance of their findings.

Matrix balancing

A matrix is **unbalanced** if the L2 norm of some rows and their corresponding columns are different by orders of magnitude.

$$\|\boldsymbol{x}\|_2 := \sqrt{x_1^2 + \cdots + x_n^2}.$$

Some computations such as the computation of eigenvalues are numerically unstable if the matrix is unbalanced.

Generally, given an unbalanced matrix A, the goal of **matrix balancing** is to find an invertible diagonal matrix D such that DAD^{-1} is balanced or approximately balanced in the sense that every row and its corresponding column have the same norm.

Here, we describe what characterizes unbalanced and balanced matrices.

Matrix balancing approaches

Implicit matrix-balancing approaches are widely used to account for biases in Hi-C data. They rely on two different assumptions.

(1) the combinatorial-bias effect between two interacting loci can be simplified as the product of the two locus-specific bias effects.

(2) if there is no bias effect (that is, when all bias has been accounted for), the total genome-wide contact summation for each locus will be a constant, implying that each locus has 'equal visibility' to the Hi-C assay.

Schmitt et al. Nature Rev Mol Cell Biol (2016) 17, 743

No comments.

Matrix balancing approaches

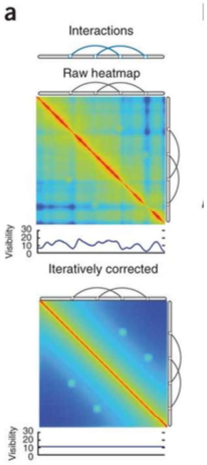
Two matrix-balancing algorithms used together with HiC-data are:

Vanilla coverage: To account for bias, the observed contact frequency between locus A and locus B is divided by the product of the total genome-wide contact frequency at locus A and the total genome-wide contact frequency at locus B. This ratio is used as the normalized contact frequency.

Iterative correction and eigenvector decomposition (ICE): this process iterates through the vanilla coverage procedure (using updated total genome-wide contact frequencies!) until there is convergence of the normalized contact frequency.

+ reduced coverage variability from locus to locus

- greatly increased computational cost.



Schmitt et al. Nature Rev Mol Cell Biol (2016) 17, 743
Imakaev et al. Nature Methods 9, 999–1003 (2012)

The idea of the first method („Vanilla coverage“) is that two DNA loci having each a high contact frequency in principle also have a relatively high chance of making contacts to eachother.

Thus, one normalizes the contact frequency A-B by the product of the individual contact frequencies.

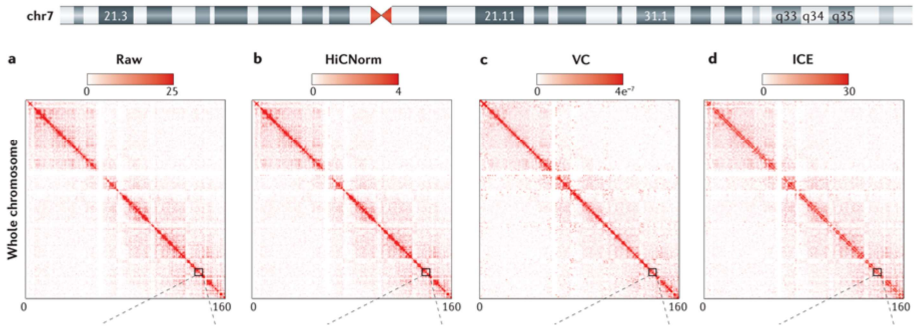
The second method builds upon the first method but adds further iterations.

The reasons is that normalization of all matrix entries of e.g. locus A (one row or one column) will affect its total contact frequency.

Then, the normalization factor in the next iteration will be somehow different.

This element is similar to the SVDimpute method (lecture 3, slide 20).

Application of 4 bias removal methods: full chromosome



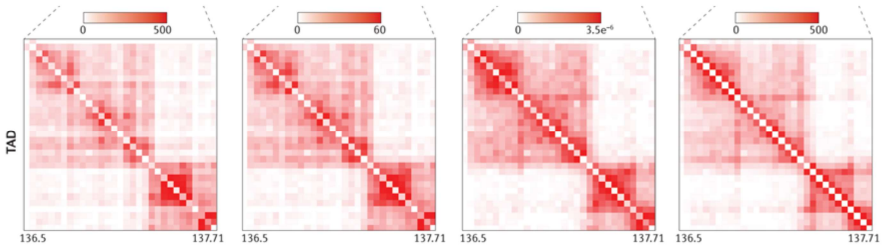
High-resolution Hi-C data from IMR90 cells were processed uniformly until the bias-removal step. Then either raw contact matrices were generated (a) or normalization was conducted with one of three methods (b) to (d).

Shown is data for whole human chromosome 7 for a raw Hi-C contact matrix (part **a**), after normalization with HiCNorm (**b**), or with vanilla coverage (VC) (**c**) and iterative correction and eigenvector decomposition (ICE) (part **d**).

Schmitt et al. Nature Rev Mol Cell Biol (2016) 17, 743

So far, no extensive comparisons of the different methods have been reported.

Application of 4 bias removal methods: TAD domains



Pairwise interactions observed at higher frequency are depicted as a darker red colour along the colour gradient, whereas light red coloration represents very few observed interactions in the Hi-C data.

Different normalization methods yield slightly differences but very different numbers.

It is currently unclear which method works best.

Schmitt et al. Nature Rev Mol Cell Biol (2016) 17, 743

Another bias that is not explicity considered by HiCnorm is that restriction enzymes used in library preparation are biased towards cutting at open chromatin regions.

Schmitt et al. further recommend „It is also good practice to conduct Hi-C data analyses using both types of bias-removal approaches, as this eliminates the possibility of making a discovery that is dependent on the type of bias-removal method."

Integration of multiple data sets

The group of Frank Alber/USC has originally constructed a 3D model of the nuclear pore complex via data integration.

They now work on 3D models of chromatin.

lamina-DamID experiments identify specific chromatin domains with a high propensity to be located at the nuclear envelope (NE).

Chromosome conformation capture experiments (Hi-C and variants) detect chromatin interactions at a genome-wide scale.

Li et al. Genome Biology (2017) 18:145

Now we will turn to a very different approach.

In 2007, Frank Alber was leading author or a pioneering study that determined the molecular structure of the nuclear pore complex (https://www.nature.com/articles/nature06405). The team integrated diverse experimental observables and then used molecular simulations to generate molecular conformations that are compatible with the observables. His own group at the University of Southern California (http://web.cmb.usc.edu/people/alber/Group.html) now utilizes similar approaches to study the three-dimensional conformation of the genome. For this, they utilize here two sorts of experimental information: lamina-DamID and Hi-C.

Link for the Li et al. paper:
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1264-5

lamina-DamID experiments

Schematic illustration of DNA adenine methyltransferase identification (DamID) experimental pipeline.

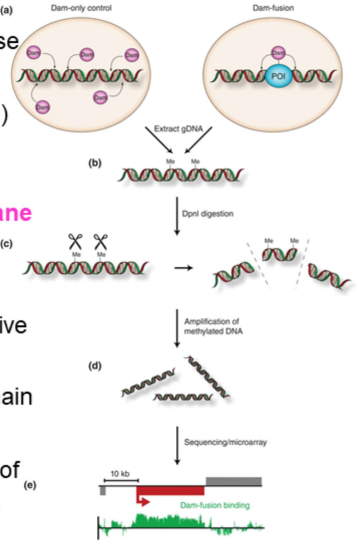
(a) Dam only or Dam fused to a protein of interest (POI) (blue) is expressed in a suitable cell type or transgenic organism. **Here: POI is lamin B1 that is part of the nuclear lamina → DAM localizes to nuclear membrane**

(b) Genomic DNA is extracted. DNA obtained includes N6-adenine methylation sites (Me) catalyzed by Dam.

(c) Genomic DNA is digested by the methylation sensitive restriction enzyme, DpnI.

(d) Digested fragments are amplified by polymerase chain reaction (PCR).

(e) Representative output indicating chromatin binding of a protein of interest at an individual locus. Vertical bars indicate the $\log_2$ ratio of Dam-fusion/Dam only.



WIREs Dev Biol (2016) 5:25 – 37.

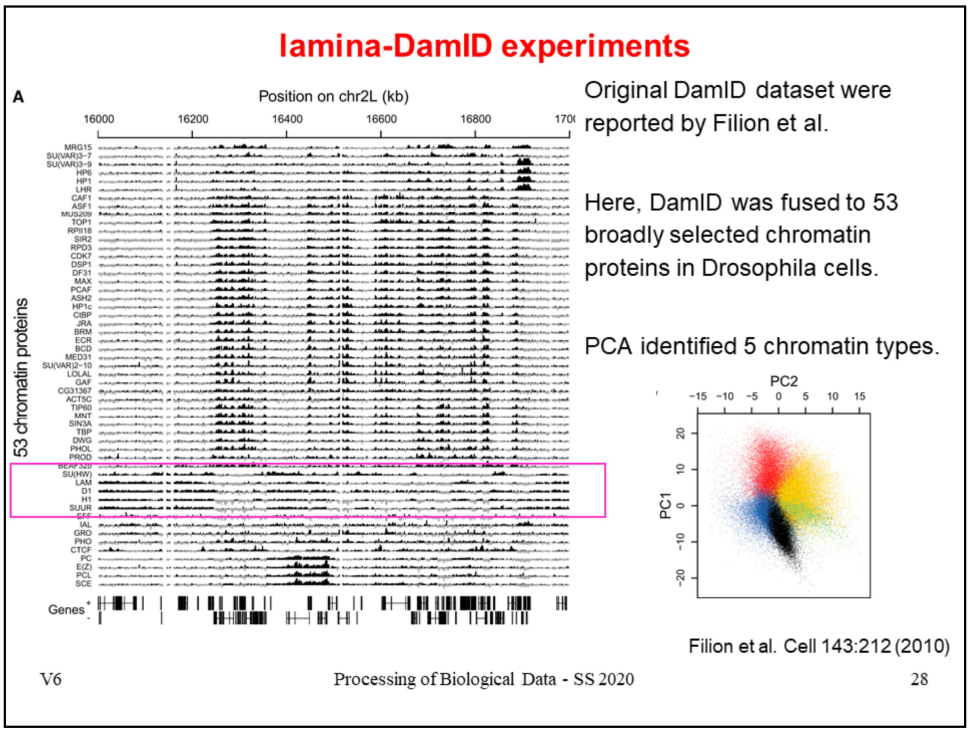
This slide illustrates the principles of the lamina-DamID experiments. „Dam“ is an abbreviation of the enzyme DNA adenine methyltransferase that methylates adenine bases at the N6 position.

The idea behind this is that Dam will methylate adenine bases in the genome that it can access. By sequencing the DNA one can then find out which regions these are.

If Dam could distribute freely in the nucleus, one would probably not learn much from this experiment beside the general accessibility of open/chosed chromatin that can also be studied by DNase experiments.

However, one can try to localize Dam to the nuclear membrane. Then it would only be able to methylate DNA fragments that are in contact with the nuclear membrane. This is exactly what is done here.

Dam is fused to the protein lamin B1 that is part of the nuclear lamina. For comparison, one also runs a control experiment (top left figure) where Dam is expressed alone.



Traditionally, chromatin was divided into heterochromatin and euchromatin. Filion et al. wanted to study how many states a finer classification needs to contain.

They determined high-resolution binding profiles of 53 chromatin proteins in the embryonic *Drosophila melanogaster* cell line Kc167.

These include proteins from most known chromatin protein complexes (e.g. histone-modifying enzymes), proteins that bind specific histone modifications, general transcription machinery components, nucleosome remodelers, insulator proteins, heterochromatin proteins, structural components of chromatin, and several DNA-binding factors.

They found that the majority of silent genes in the genome are located in BLACK chromatin.

BLACK chromatin is almost universally marked by four of the 53 mapped proteins: histone H1, D1, IAL, and SUUR, whereas SU(HW), LAM, and EFF are also frequently present

Integration of multiple data sets

So far, most population convolution models of genome structures have typically relied on just **one data type**, such as Hi-C, even though a single experimental method cannot capture all aspects of the spatial genome organization.

However, data are available from several technologies with complementary strengths and limitations.

Integrating all these different data types should increase the accuracy and coverage of genome structure models.

Moreover, such models would offer a way to cross-validate the consistency of data obtained from complementary technologies.

Li et al. Genome Biology (2017) 18:145

No comments.

Integration of multiple data sets

For example, lamina-DamID experiments show a chromatin region's probability to be close to the lamina at the nuclear envelope,

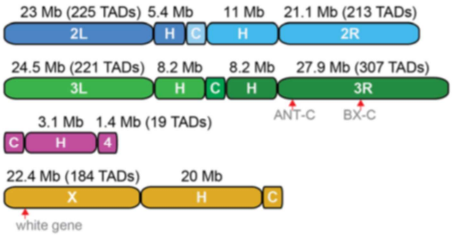
whereas Hi-C experiments reveal the probability that two chromatin regions are in spatial proximity.

3D fluorescence in situ hybridization (FISH) experiments show the distance between loci directly, and can be used to measure the distribution of distances across a population of cells.

Li et al. Genome Biology (2017) 18:145

Li et al. also performed independent FISH experiments to test the predictions from the data integration approach.

Drosophila melanogaster



The genome of *D. melanogaster* (sequenced in 2000, and curated at the FlyBase database) contains 139.5 million base pairs on four pairs of chromosomes:
an X/Y pair, and three autosomes labeled 2, 3, and 4.

It contains around 15,682 genes.
The euchromatin genome was divided into **1169 physical domains**
based on Hi-C interaction profiles.

www.wikipedia.org

Frank Alber and co-workers wanted to characterize the three-dimensional structure from *Drosophila melanogaster*, the fruit fly, because both data sets (Hi-C and lamina-DamID) were available.

Drosophila is an extremely well-known model organism for studying animal development.

Around 1980, Eric Wieschaus and Christiane Nüsslein-Volhard succeeded in identifying and classifying the 15 genes that direct the cells to form a new fruit fly. For this discovery, they receive the Nobel Prize in Physiology or Medicine in 1995.

Integration of multiple data sets

Suppose **A** is a probability matrix derived from Hi-C data.
Its elements describe how frequently a given pair of TADs
are in contact with each other in an ensemble of cells.

E is a probability vector derived from lamina-DamID data.
Its entries describe how frequently a given TAD is in contact
with the nuclear envelope (NE).

The goal is to generate a population of genome structures **X**, whose TAD–TAD and TAD–NE contact frequencies are statistically consistent with both **A** and E.

We formulate the genome structure modeling problem
as a maximization of the likelihood P(**A**, E|**X**).

Li et al. Genome Biology (2017) 18:145

Two independent experiments (Hi-C and lamina-DamID) generated two sets of observations, **A** and E.

A is a matrix describing contacts between pairs of DNA regions.

E is a vector with entries for each DNA region.

The task is now to generate chromatin 3D conformations that are compatible with A and E.

Consider population of chromatin conformations

The **structure population** is defined as a set of M diploid genome structures $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M\}$, where the *m*-th structure $\mathbf{X}_m$ is a set of 3D vectors representing the center coordinates of 2 *N* domain spheres.

The contact probability matrix $\mathbf{A} = (a_{IJ})_{N \times N}$ for N TAD domains is derived from the Hi-C data. Each element a_{IJ} is the probability that a direct contact between domains I and J exists in a structure of the population.

The contact probability vector $E = \{e_I | I = 1, 2, \ldots, N\}$ is derived from the lamina-DamID data and defines the probability for each TAD to be localized at the NE.

Li et al. Genome Biology (2017) 18:145

Chromatin is modelled as a linear sequence of N spheres representing N domains.

A diploid genome consists of 2 sets of chromosomes. Hence, each chromatin conformation has 2N spheres.

Likely, there does NOT exist a single chromatin conformation where every genomic region only occupies a single, fixed spot.

Instead, we can imagine that the DNA shows dynamic flexibility so that we should rather speak of an ensemble of conformations that can interconvert and will be visited over time.

Li et al. model this ensemble by a population of M genome structures.

Not every single structure needs to be compatible with the observed data A and E, but rather the full population of structures needs to be compatible.

Integration of multiple data sets

Thus, the optimization problem is expressed as:

$$\hat{\mathbf{X}} = \arg \max_{\mathbf{X},\mathbf{W},\mathbf{V}} \log P(\mathbf{A}, E, \mathbf{W}, \mathbf{V}|\mathbf{X})$$

subject to
- spatial constraint I : nuclear volume constraints
- spatial constraint II : excluded volume constraints
- spatial constraint III : chromosome pairing upper bound
- spatial constraint IV : consecutive domain constraint

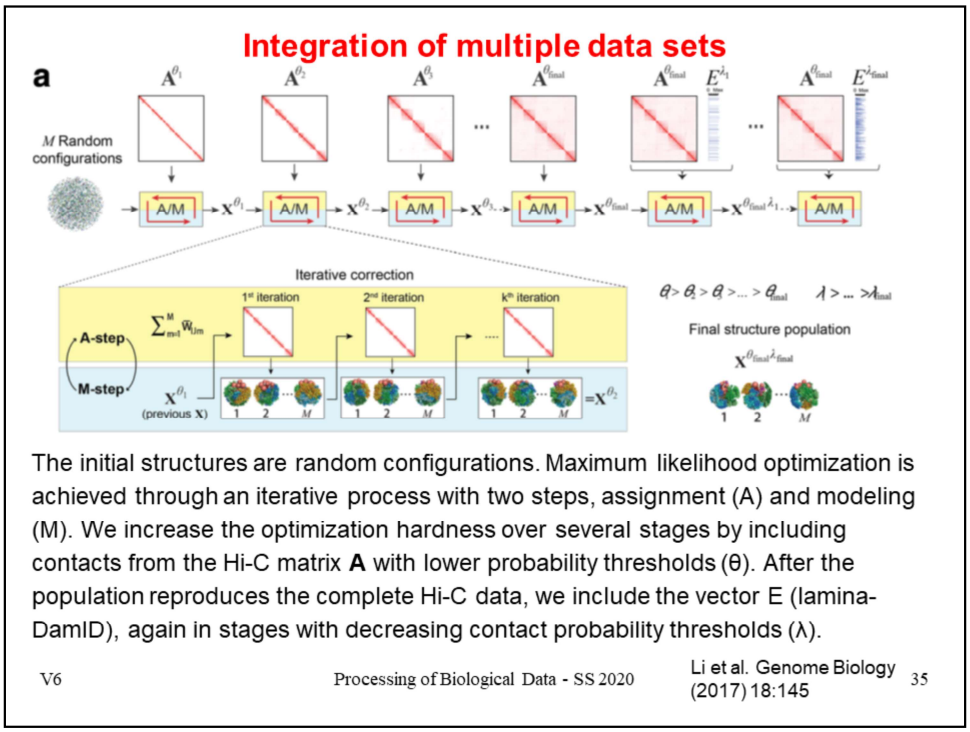
The log likelihood can be expanded as

$$\begin{aligned} \log P(\mathbf{A}, E, \mathbf{W}, \mathbf{V}|\mathbf{X}) &= \log P(\mathbf{A}, E|\mathbf{W}, \mathbf{V}) P(\mathbf{W}, \mathbf{V}|\mathbf{X}) \\ &= \log P(\mathbf{A}|\mathbf{W}) P(E|\mathbf{V}) P(\mathbf{W}, \mathbf{V}|\mathbf{X}) \end{aligned}$$

The "contact indicator tensor" **W** = $(w_{ijm})_{2N \times 2N \times M}$ is a binary, third-order tensor. It contains the information missing from the Hi-C data **A**, namely which domain contacts belong to each of the M structures in the model population and also which homologous chromosome copies are involved.

V = $(v_{im})_{2N \times M}$ specifies which domain is located near the NE in each structure of the population and also distinguishes between the two homologous TAD copies

One interesting problem is to assign which of the M structures belongs to which chromatin contacts.



The initial structures are random configurations. Maximum likelihood optimization is achieved through an iterative process with two steps, assignment (A) and modeling (M). We increase the optimization hardness over several stages by including contacts from the Hi-C matrix **A** with lower probability thresholds (θ). After the population reproduces the complete Hi-C data, we include the vector E (lamina-DamID), again in stages with decreasing contact probability thresholds (λ).

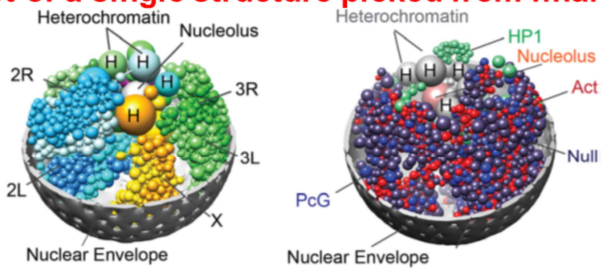
The approach taken here is similar to the approach used previously when Frank Alber modeled the structure of the nuclear pore complex.

Li et al. argue that it is practically impossible to generate genome structures „ab initio“ (without prior knowledge) that simultaneously fulfil all experimental constraints.

Instead, they introduce contact distance restraints A piecewise (upper row, from left to right) followed by adding the membrane distance restraints E.

The colored spaghetti balls in the bottom row illustrate the populations of M genome structures.

Snapshot of a single structure picked from final population



(left) The full diploid chromosomes are shown in colors: blue, chr2; green, chr3; magenta, chr4; orange, chrX.

The two homologs of the same chromosome are distinguished by the color tone, with one homolog copy with lighter and one with darker color. The heterochromatin spheres are larger than the euchromatin domains. The nucleolus is colored in silver.

(right) euchromatin domains are colored to reflect their epigenetic class:

red, active;

blue, PcG;

green, HP1;

dark purple, null.

Heterochromatin spheres are shown in grey and the nucleolus in pink

V6 Processing of Biological Data - SS 2020 Li et al. Genome Biology (2017) 18:145 36

In these figures, physical domains (which would be referred to as TADs in mammalian cells) are represented as spheres.

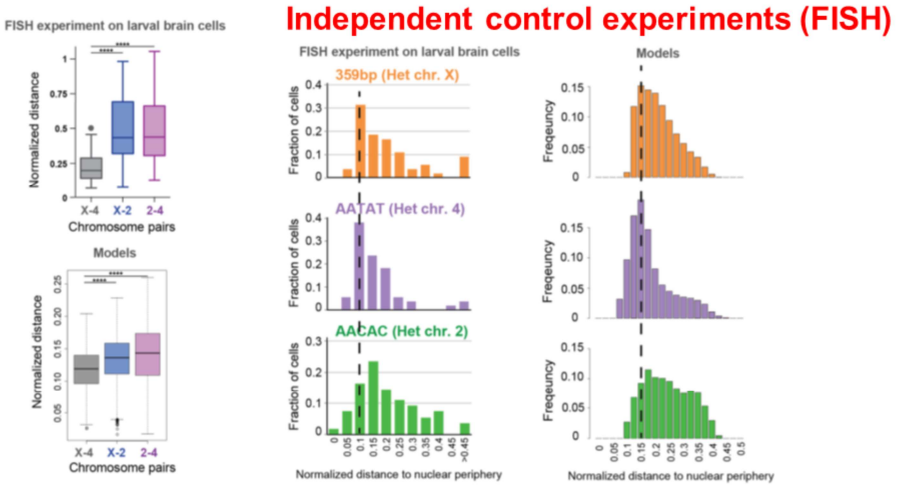
In the left figure, each chromosome is colored differently.

In the right figure, the domain spheres are colored differently.

It is unclear whether this structure represents the same conformation as in the left figure.

Coloring represents the functional classes of the physical domains. Four functional classes based on their epigenetic signatures are assigned: null, active, Polycomb-group (PcG), and HP1/centromere.

Note that this figure only represents a single structure snapshot of the conformational population.



The model predicts certain location preferences for pericentromeric heterochromatin of individual chromosomes. We confirmed these predictions using FISH staining of heterochromatic repeated sequences (satellites) in *Drosophila* cells of larval brains.

(Left panel) FISH experiments showed that the satellite repeats of chromosomes X and 4 (grey) are more often closer to each other than those of chromosomes X and 2 (blue), or 2 and 4 (magenta((top), in agreement with the computational models (bottom).

(middle panel) The satellite repeats of chromosomes X (top) and 4 (middle) are more often closer to the nuclear periphery than those of chromosome 2 (bottom).

This matches the conformations of the model population (right panel).

Summary

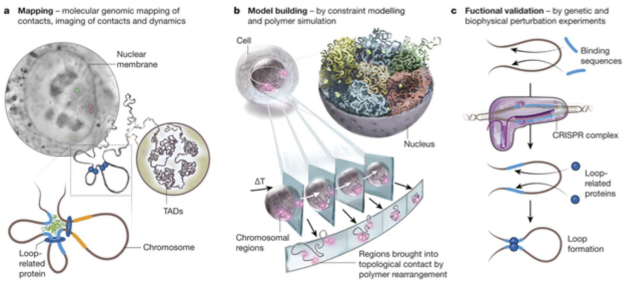
Chromosome capture techniques enable to obtain information on **contacts** along one chromosome and between chromosomes.

Experimental design introduces various **biases** that must be corrected before analysis.

Data integration has great potential.

Considering **populations** of different structures helps to resolve conflicts between data.

An important activity in this area is the **4D Nucleome** project.



https://www.4dnucleome.org/index.html

Paper on 4D Nucleome project: https://www.nature.com/articles/nature23884
https://www.4dnucleome.org/index.html