Assessing Chromatin Hierarchy

Introduction

DNA is highly organised in the nucleus; if the nucleus of a cell was the size of a football then it would contain approximately 5 miles of DNA.

At the highest level of chromatin folding complexity, genomic DNA is organised into "A" and "B" compartments of euchromatin (typically open chromatin) and heterochromatin (typically closed, condensed chromatin), respectively. Within these large-scale compartments, megabase-scale topologically associated domains (TADs) are defined based on the high frequency of intra-domain interactions relative to inter-domain interactions.

At the primary-order of DNA folding complexity, DNA can be imagined as having a "beads on a string" structure, with the beads representing nucleosomes and the DNA as the string. Nucleosomes consist of 147bp of DNA wrapped 1.65 times around a complex of 8 histones. Also at a basic level, DNA forms loops via extrusion, whereby cohesion complexes bind to the DNA and the DNA is pulled through the complex in opposite directions to form a loop, until the DNA locates a CTCF motif that is pointing inward to the loop, where looping ceases (Sanborn et al., 2015). These loops facilitate physical contact between regions of the genome that are far in linear distance, for example bridging the gap between an enhancer and a promoter to regulate gene expression.

The interaction between chromatin regions and the accessibility of chromatin are mediated by protein complexes and epigenetic modifications which allows variability between tissues and cell-types and is highly non-static. For this reason, these structures should be explored in at least a tissue-specific manor, but ideally at biologically interesting times.

In this review, I summarise a selection the techniques commonly used to assess chromatin structure, focussing primarily on those which I am utilising in my current PhD work.

Evaluating DNA accessibility

DNA accessibility refers to whether the DNA is "open" and can therefore be bound by regulating factors or "closed" and likely inactive. There are four popular methods which have been developed for this purpose.

1. FAIRE-seq (2007)

Formaldehyde-Assisted Isolation of Regulatory Elements sequencing begins by crosslinking DNA to nucleosomes using formaldehyde and removing the regions bound to nucleosomes by phenol-chloroform extraction. The remaining DNA is sequenced to profile accessible regions. Although the experimental protocol is very simple, this method yields relatively low signal-to-noise ratios.

2. DNase-seq (2008)

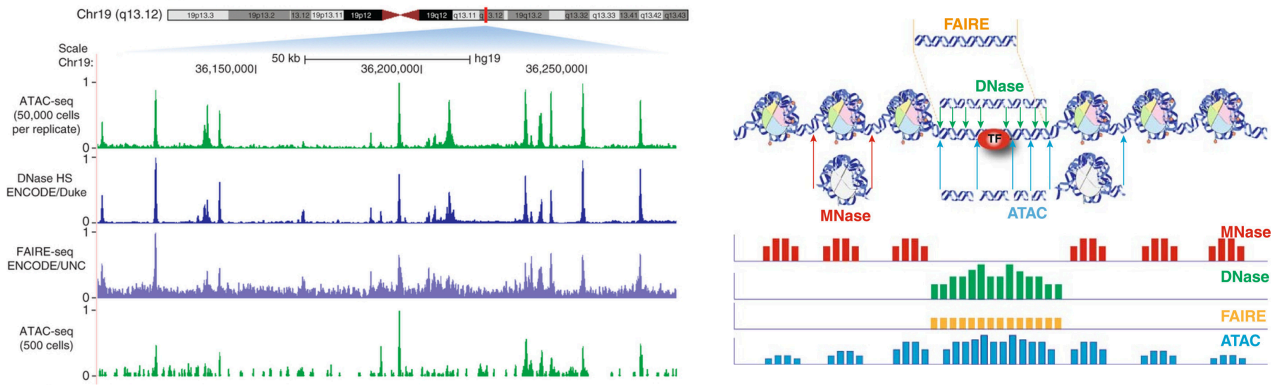
DNase I hypersensitive sites sequencing profiles accessible regions of the DNA utilising he enzyme DNase I which preferentially cleaves DNA at sites that are not wrapped around histones into ~ 150 bp fragments. This technique can also be used for transcription factor footprinting because the enzyme will not be able to cut the DNA where a protein is bound. DNase I requires lots of cells as input and has sequence-specific cutting biases, but these are largely known and so can be handled statistically.

3. MNase-seq (2011)

Micrococcal Nuclease sequencing is conceptually the opposite of DNase-seq, in that it is used to find regions of the DNA that are bound to nucleosomes. To do this, it utilises MNase, an endo-exonuclease that cleaves DNA between nucleosomes and digests these regions of DNA that are unprotected by nucleosomes, thus only leaving those regions bound to nucleosomes. However, the length of DNA wrapped around histones is 147 bp but the size of DNA fragments after MNase digestion vary from 120-170 bp. MNase requires lots of cells as input and also has sequence-specific biases, for example it preferentially cuts A/Ts.

4. ATAC-seq (2013)

Assay for Transposase Accessible Chromatin with high throughput sequencing uses a hyperactive mutant transposase (Tn5) which cleaves DNA in open regions and also inserts adapter sequences (eliminating additional ligation steps). The DNA *between* these adapter sequences is then sequenced. ATAC-seq requires substantially less cells as input material than the other techniques, works well even with frozen samples and the whole protocol only takes approximately 3 hours. It is often the method of choice for profiling regions of open DNA.



(Left: Buenrostro et al, Nature Methods (2013). Right: Maria Tsompana and Michael J Buck, Epigenetics & Chromatin (2014))

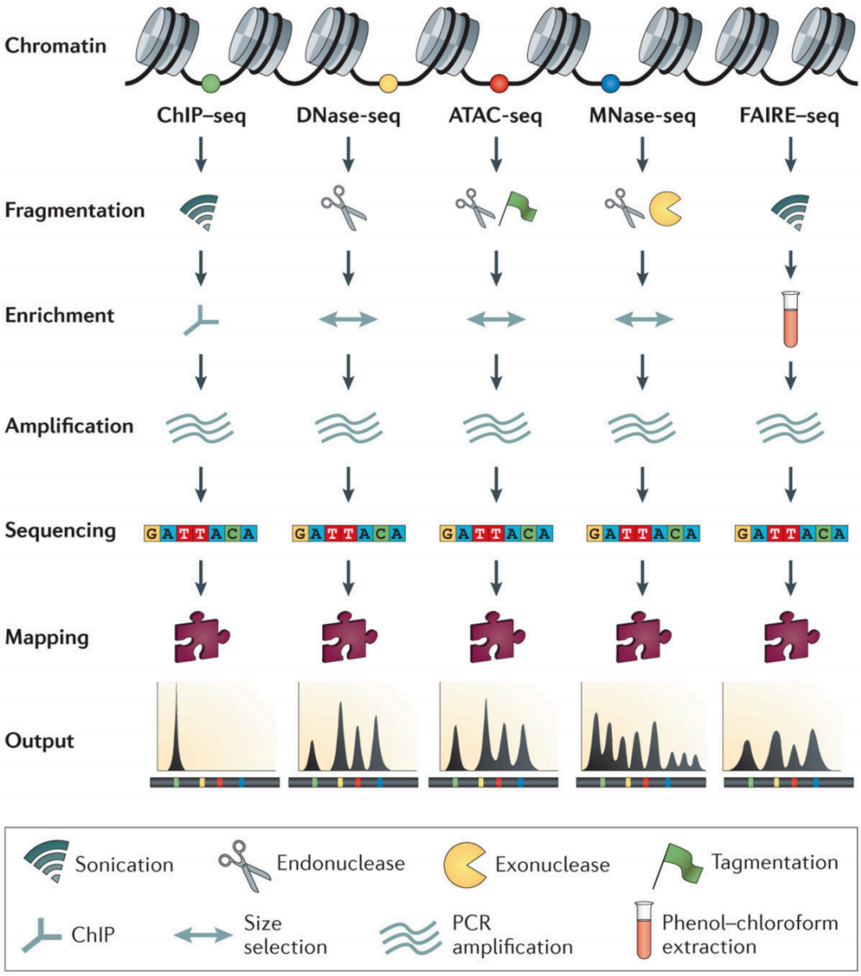
Evaluating chromatin state

The set of chromatin-associated proteins and epigenetic modifications at a given time in a genome region constitutes the chromatin state - for example the binding of a transcription factor or a histone modification (Chang et al, Computations and Structural Biotechnology Journal (2018)). We've seen above that some of the methods to evaluate chromatin accessibility can be used for transcription factor footprinting (namely DNase-seq and ATAC-seq). Here, I focus on chromatin immunoprecipitation sequencing (ChIP-seq, 2007) which aids the identification of regions in the genome that are bound by specific proteins.

Briefly, in the ChIP-seq experimental protocol, cells are first crosslinked using formaldehyde before being fragmented using restriction enzymes or sonication. DNA fragments that are bound to a protein of interest are pulled down using a specific antibody (immunoprecipitation) and these fragments are sequenced and aligned to a reference genome. Peak callers (e.g. MACS2) are used to derive peaks from the aligned reads, typically using input DNA from a control sample (e.g. "mock IP" – whereby DNA fragments undergo the immunoprecipitation step without the use of the antibody) to account for biases such as fragmentation bias in repetitive/open regions and uneven distribution of sequence tags across the genome. Advanced peak callers will take advantage of the directionality of the read and model the shift size between positive and negative strands; this is required because read count peaks do not line up with the exact position of the protein binding site because the DNA is protected from nuclease cleavage in these regions.

ChIP-seq is commonly used to identify chromatin modifications which are indicative of various chromatin states. For example, the H3K4me3 modification is thought to take place in the adjacent nucleosome to the transcriptional start site, relaxing the structure and thereby marking promoter regions of the genome. The combination of histone modifications is also important in determining function, for example H3K4me1 alone marks primed enhancers whilst H3K4me1 co-occurring with H3K27ac marks active enhancers (Jiang and Morazavi, Briefings in Functional Genomics (2018)).

ChromHMM is a software program to capture combinations of multiple histone modifications, transitioning from chromatin marks (observables) to chromatin states (unobservables). The input to the algorithm is typically ChIP-seq data - specifically, the list of aligned reads for each chromatin mark which are automatically converted into presence or absence calls for each mark across the genome, based on a Poisson background distribution. A hidden Markov model (HMM) is then used to assign a chromatin ("emission") state to approximately 200 bp regions of the genome. It also outputs emission probabilities, which measure the probability of observing that histone mark given that you're in that state. ChromHMM can therefore be used to generate a single chromatin state track summary for various cell types using ChIP-seq data, but these tools only provide linear information on chromatin state.



(Meyer & Liu, Nature Reviews Genetics (2014))

Finding chromatin contacts

The invention of chromosome conformation technologies has allowed researchers to map regions of the genome which are physically contacting each other in 3D space. Briefly, DNA is crosslinked and digested with DNA restriction enzymes. The loose DNA fragment ends are then re-ligated to generate a hybrid DNA molecule formed of two fragments of DNA which may be very far apart in linear distance. Ligation junctions are detected, and the fragments are mapped back to a reference genome. In this review, I focus on the Hi-C approach and its variants. Table 1 is an extension of Table 1 in Ubelmesser and Papantonis (2019) and provides a summary of the recently developed techniques to analyse spatial genome organisation.

Expansion of Table 1 in https://academic.oup.com/bfg/advance-article/doi/10.1093/bfgp/elz019/5586926

Method	Info	How?	Advantages	Disadvantages
In situ Hi-C	DNA-DNA proximity ligation is performed in intact nuclei, generating much denser Hi-C maps.	Crosslink cells, use a 4-cutter restriction enzyme to digest the DNA, fill the overhangs and incorporate biotinylated nucleotides, ligate the blunt-end fragmnets, shear the DNA, pulldown the biotinylated ligation junctions with streptavidin beads and sequence.	- High (kbp) resolutions due to 4-cutter - Variations allow application to samples with limited cell numbers (Low-C) - Reduces frequency of spurious contacts due to random ligation in dilute solution.	- Requires extreme sequencing depth - Mostly captures pairwise interactions.
Micro-C	Use micrococcal nuclease instead of restriction enzymes, which enables nucleosome resolution of chromatin contact maps.	Crosslink cells, digest using MNase, mononucleosomal end repair.	- High (nucleosome) resolution.	- Unreliable for capturing long-range interactions - Only captures pairwise interactions.
DNase Hi-C	Use DNase I instead of restriction enzymes, producing libraries of higher resolution.	Crosslink cells, lyse to liberate nuclei, treat with DNase I, end-repair chromatin ends and sequence.	- High (sub-kbp) resolution	- Bias towards DNase I hypersensitive sites - Only captures pairwise interactions.
ChIA-PET	Examine all loci bound by a specific protein. Combines ChIP with 3C.	Crosslink cells, sonicate DNA, use antibody of choice to enrich protein-of-interest-bound chromatin fragments. Continue as above.	- Good at detecting long-range interactions - Can focus on contacts from a factor of choice	- Requires large cell numbers - Generates short reads
CHi-C	Choose baits of interest for Hi-C.	As above, but includes a pull down step to only capture the interactions involving the bait of interest.	- Can be adapted for low cell counts - Can choose loci of interest (baits).	- Probes can be costly - Mostly captures pairwise interactions
HiChIP	Discover protein-centric *in situ* chromatin loops. Similar to ChIA-PET but with less input material	DNA contacts are established *in situ* in the nucleus before lysis, ChIP is performed on the contact library to directly capture long-range interactions associated with a protein of interest. Then use paired-end sequencing.	- Focused on contacts from a factor of choice - Works will low cell counts	- Dependent on antibody quality - Variable data complexity.
TSA-seq	Allows the distance of every gene from specific nuclear landmarks to be measured simultaneously.	The nuclear structure of interest is tagged with horseradish peroxidase which generates a highly reactive molecular called tyramide that labels the DNA. The closer the DNA is to the structure, the more it will be labelled.	- Provides information on physical distances	- Confined to particular nuclear features of interest.
Tri-C/ MC-4C	Tri-C characterises concurrent chromatin interactions at individual alleles.	Tri-C libraries are generated using an enzyme (NlaIII) selected to create relatively small DNA fragments (~200 bp) for the target fragment. The fragments are religated, MC-4C then uses Cas9 digestion of the sequence of interest.	- Long sequencing reads - Captures multi-way interactions	- Relatively low throughput - MC-4C hass a complicated Cas9 step.

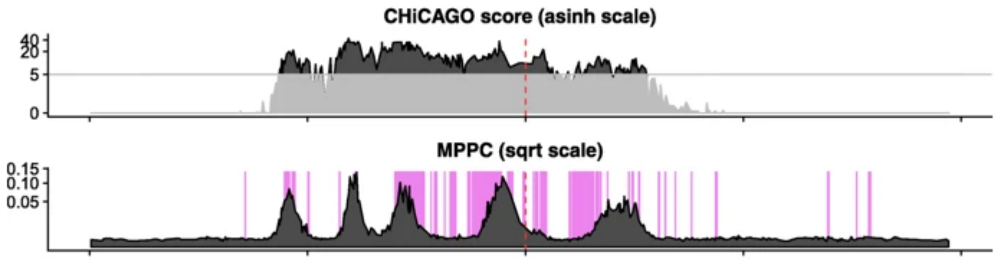
Hi-C is an all vs. all approach in that it generates contact maps among all parts of the genome. An enrichment step using streptavidin bead pull-down enriches fragments containing ligation junctions which are then sequenced. However, the resolution is relatively low (~ 1 Mb with 10 million paired-end reads) and a 10-fold increase in resolution requires a 100-fold increase in sequencing depth (E de Wit et al, Nature (2013)).

Capture Hi-C (CHi-C) is a variation of Hi-C that had an additional pull-down step so that interacting regions with a fragment of interest ("bait") are enriched. For example, promoter capture Hi-C (PCHi-C) can be used to generate libraries enriched with promoter-containing fragments. CHi-C methods have unique statistical properties that follow-on statistical pipelines need to consider, including:

- Asymmetry of CHi-C interaction matrices whereby there are fewer baits but many possible "other-end" preys.
- Uneven capture efficiency in that the bait fragments are selected *a priori*.
- Multiple testing problems where many more tests are conducted at large distances, where fewer true interactions may lie.

CHiCAGO is an algorithm which attempts to take into account these unique statistical properties to provide a "score" for each interaction. This method uses a background correction procedure that models both the technical (using a Poisson random variable) and biological noise (using a negative binomial random variable) and implements a weighted false discovery control procedure. Generally, a CHiCAGO score > 5 is deemed biologically interesting. However, the resolution from a CHiCAGO analysis is generally low, in that it identifies many adjacent prey fragments that all seem to be interacting with the bait of interest.

Peaky uses a Bayesian sparse variable selection approach to fine-map chromatin contacts by the joint modelling of read counts at neighbouring prey fragments. The statistic of interest used in the Peaky pipeline is the marginal posterior probability of a contact (MPPC), which is the proportion of sampled models in the RJMCMC algorithm where the estimated contact strength is strictly greater than 0.



Aggregation

In this review I have commented on methods to profile open regions of the DNA, to characterise chromatin state across the genome and to identify physically contacting regions. Often researchers will use one specific tool to answer their very specific research question, however there is merit to be gained by integrating these chromatin assessment methods.

Incorporating information on chromatin accessibility is commonly used to confirm experimental results. For example, in the figure above, the pink regions show active chromatin derived by aggregating states from ChromHMMM analysis of ChIP-seq data. These have been used to show that the contacts identified using Peaky align to open regions of the genome, increasing the reliability of their results.

Aggregating information on chromatin interactions (e.g. Hi-C) with information on chromatin state (e.g. ChIP-seq) may be also useful in uncovering the regulation patterns in the genome. ChIA-PET combines ChIP-seq with 3C technologies to find regions of the genome that are bound to the same factor (Fullwood et al, 2009). However, the interpretation from the analysis is limited because it cannot easily be used to detect whether the regions are interacting *because* of the protein of interest (cannot rerun the analysis with the protein knocked out as this is required to “pull-down” the regions of interest).

Incorporating chromatin accessibility information (e.g. DNase-seq) with ChIP-seq data has been used in the ENCODE project to define chromatin states using HMMs. Histone modification ChIP-seq data is also often mapped to open chromatin peaks to confirm the chromatin state of regulatory elements. ATAC-seq-like experiments are cheaper and easier to run than ChIP-seq and so could potentially be used to see when/where a region of DNA bound to a TF becomes open/closed.

Uncovering the hierarchical structure of chromatin in specific cell types at biologically interesting time points is useful in its own right but it also has the potential to improve inferences from GWAS. GWAS coupled with fine-mapping finds genetic variants that are likely causal for specific diseases. By mapping these variants to regions of the genome and assessing the chromatin structure at various time points and cell types, researchers can build up a better idea of what specific biological functions this variant effects. The software programme “CHEERS” aims to identify SNP enrichment across cell states by accounting for subtle changes in the chromatin landscape (i.e. how many reads form the peak) (Soskic et al, Nature Genetics (2019)), and it is used to find immune disease SNP enrichments across a variety of activation states of T cells and macrophages.

In summary, integrating chromatin interactions and accessibility profiles with genetic and epigenetic features is needed to untangle disease mechanisms. Rather than the generation of endless amounts of data, robust statistical methods need to be developed which can reap the benefits from the aggregation of these data sources.