Unravelling higher order genome organisation [working title]

Thesis plan

Benjamin L. Moore

October 28, 2014

igmm
INSTITUTE OF GENETICS
& MOLECULAR MEDICINE

1 INTRODUCTION

Need to introduce the seminal studies of genome organisation, particularly those using C methods (Lieberman-Aiden *et al.*[1], Dixon *et al.*[2], older 3C stuff[3] etc.). Detail the fractal globule[1] model of genome organisation, as well as counter theories like strings, binders and switches (SBS)[4]. There's also growing computational literature in calling domains and various sub-domains (e.g. Bing Ren's D.I. + HMM[2], but also alternative domains[5], dynamic programming solutions[6] etc.).

Also mention the criticisms of C methods[7,8] and discuss the intricacies of processing the data (biases,[9] normalisation[10–12] etc.).

Machine learning has been successful at calling chromatin states (`ChromHMM`, `SegWay`, others) and generally building models of complex biological phenomena. Explain how this type of computational approach has led to biological insights, with particular deference to ENCODE studies such as Dong *et al.*[13], as well as the most recent ENCODE papers.

Potentially too far off topic but there's also a lot of literature on the inference of 3D conformation[14–18] — that is, actual three-dimensional trajectories of the DNA polymer — using data generated by C based methods as input. Current results could be extended by looking at 3D spatial data.

Finally discuss the most recent and overlapping work published, such as that by Comoglio and Paro[19] in which replication timing was predicted by combinatorial modelling.

2 METHODS

I already have some text for this from first year report and the paper methods section. Things to cover include:

- Processing raw reads, mapping
- ENCODE data processing, MACSv2 (do I mention this?)
- ICE, Hi-C normalisation
- Calling boundaries, HMMs
- Modelling, random forests, variable importance
- GLASSO, regularisation
- Citations for R packages used
- Package code for entire thesis?

3 | RESULTS

Here each section would be a separate chapter.

3.1 **EARLY STUFF:** MODELLING TRANSCRIPTION AND CHROMATIN

I replicated the work of Dong *et al.*[13] in modelling of transcriptional output based on a large set of ENCODE features. I extended their work by adding new features, and dissected the "best bin" approach to discover where (relative to a gene) influential variables correlated best with expression.

We then applied the same approach to modelling a different set of data: the A / B compartment profiles reported in Lieberman-Aiden *et al.*[1]. Noting that Hi-C datasets were available for the three tier 1 ENCODE cell lines, I applied this modelling approach to each in turn, with their own corpus of ENCODE features. Models of this facet of higher order structure proved roughly as accurate as those of Dong *et al.*[13] were for transcriptional output.

3.2 **MODEL DISSECTION:** REGULARISED MODELS, INFLUENTIAL VARIABLES, CROSS-APPLICATION

Having reasonably accurate models of chromatin organisation, it's then of interest to understand why they are successful and if improvements can be made. Firstly, rankings of variable importance were looked at per cell type model. Models were also cross-applied from one cell type to another.

We were interested in building minimal viable models, or those suitably regularised such that accuracy was maintained while the dimensionality of input features was minimised. To this end, we employed the Graphical LASSO algorithm, a tuneable L1 regulariser, to reduce each model of 36 variables down to approximately 5 with little loss in predictive power. However, this "wrapper" method of regularisation was independent from the learning algorithm, hence may not represent a truly optimal subset of features.

In order to resolve this, we employed a regularised Random Forest algorithm[20,21], as well as a brute-force process of constructing all possible subset models with varying numbers of features. For example, all combinations of five variables from the original 36 were passed to the learner and the accuracy was compared. From this we discovered that while model performance was affected by the number of input features, due to the pervasive multi-collinearity any subset model of five variables would perform almost equally well. This signalled that generated minimal viable models may provide little additional understanding of the relationships between higher order chromatin organisation and our locus level features.

3.3 **ODDS 'N' ENDS:** TADS, BOUNDARIES, SUPER BOUNDS, G-BANDS

TADs are a well-described facet of higher order chromatin organisation at a scale below that of nuclear compartments. We recalled these domains in each cell type under study and compared the results. Unsurprisingly we found

TAD boundaries to be well-matched between these cell types, confirming them as a relatively invariant level of organisation.

The boundaries of TADs have previously been reported as bound by numerous factors, some of which (e.g. CTCF[22]) have previously implicated roles in organising genome conformation. With a larger set of ChIP-seq datasets available, we quantitatively tested for enrichment or depletion of 36 DNA binding proteins and histone modifications. This enabled us to compare enrichments across cell types and identify those that were consistently marking these boundaries. Further, we applied the same methodology to boundaries of compartments and discovered similar spectra of enrichments and depletions, but at a lower resolution — in agreement with a "fractal globule" view of genome organisation.

We investigated the idea of "*super boundaries*" which were both TAD and compartment boundaries. It emerged that these boundaries, though present, did not display stark differences from non-overlapping TAD or compartment boundaries.

We also found an agreement between A and B compartments with the long-known Giemsa stain bands. Both were previously known to correlated with patterns of high and low GC content ("isochores") and more recent work made a similar observation but did not quantitatively test the association ([cite Casey's student's thesis]). This also hints at a link between interphase genome organisation (i.e. that measured on average by Hi-C) and metaphase organisation, as assayed by Giemsa stain. This potentially contrasts with prominent research which found total structural rearrangement to a homogenous state during mitosis[23].

3.4 ADDITIONAL CHAPTERS FROM MY NEXT PROJECT

Now that a paper from my initial project is submitted, I'll be starting a new project which should in theory fill ~2 (?) chapters. This project will continue with the genome organisation theme and likely continue to make use of some of the reprocessed datasets I have generated.

The initial exploratory steps of my next project will look at expression in the context of my existing higher order structure datasets. Cap analysis of gene expression (CAGE) data is available for each cell line used so far—this data combines quantitative expression output with precise positional information so offers the opportunity to broadly investigate expression and genome organisation. TADs have been hypothesised to function as "regulons"[24], co-ordinating expression response amongst spatially-collocated genomic regions. Can we find evidence for this proposed function between our three human cell lines? Similarly, how is domain conservation reflected in transcription: does e.g. expression breadth correlate with domain conservation?

Other questions to investigate include:

- Do predicted co-locating domains share some aspect of expression? After normalising for genomic distance are they significantly co-regulated or functionally-enriched?
- Probe the evidence for "transcription factories", particularly with newer, high-resolution Hi-C datasets. Can we reliably find these structures? Are they all intra-domain? What types and proportion of all genes are involved? How specific are they to a given cell type?
- There's also the opportunity to analyse mouse Hi-C data for comparative work, such as relating synteny and higher order genome organisation (specifically with respect to domain structures).
- Finally there's some single-cell Hi-C data available which may be of interest but is probably too sparse for any of the above purposes.

3.5 COLLABORATIONS AND SIDE PROJECTS

I've also analysed related C-methods data produced by researchers in wet lab groups:

- Adam Douglas (**Hill group**) 4C, Capture-C, FAIRE-seq: Analysis of 4C contacts between the ZRS enhancer and the SHH gene in mouse developing limb bud cells with and without trichostatin A (TSA) treatment. Preliminary results appear to suggest that with an inactive SHH gene, ZRS makes multiple non-specific contacts but after treatment forms a specific and significant contact with SHH, located approximately 1 Mb away. A newer C-method, Capture-C, will also be used across this region to backup findings from the 4C and FISH experiments.
- Iain Williamson (**Bickmore group**) 5C: Comparing contacts for anterior and posterior developing limb over the HoxD locus. Also comparing with existing mouse Hi-C data to visualise a potentially changing TAD structure. (Possibly won't include this minor analysis.)

4 DISCUSSION

Summarise key results and place into broader (particularly biological) context.

Some points from the submitted manuscript:

- We have shown that it is possible to construct comparable models describing the features underlying higher order chromatin structure, and that their predictive accuracy can be high.
- Random Forest models summarised the important relationships among these many variables, providing insights into the quantitative contributions of locus level chromatin features to higher order structures. Although certain features were notably more influential in a particular cell type, the models shared overlapping constellations of informative features, allowing the cross application of models between cell types.
- These models also allowed us to probe the features underlying regions with variable higher order structure between cell types, revealing enrichments of cell type specific enhancer activity, and suggesting links between functional chromatin states and higher order domain dynamics.
- Chromatin boundaries, separating TADs and nuclear compartments at different scales, also showed cell type specific enrichments of various locus level chromatin features. Across cell types, the complexity of boundary composition varies considerably so that only a few features were seen consistently enriched or depleted at boundaries.

5 END MATERIAL

- Appendix
- References

REFERENCES

[1] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**(5950): 289–93.

[2] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**(7398): 376–80.

[3] van Steensel B, Dekker J (2010) Genomics tools for unraveling chromosome architecture. *Nature biotechnology*, **28**(10): 1089–1095.

[4] Barbieri M, Chotalia M, Fraser J, Lavitas LM, Dostie J, Pombo A, Nicodemi M (2012) Complexity of chromatin folding is captured by the strings and binders switch model. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(40): 16173–8.

[5] Filippova D, Patro R, Duggal G, Kingsford C (2014) Identification of alternative topological domains in chromatin. *Algorithms for molecular biology : AMB*, **9**: 14.

[6] Levy-Leduc C, Delattre M, Mary-Huard T, Robin S (2014) Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics*, **30**(17): i386–i392.

[7] Gavrilov Aa, Golov AK, Razin SV (2013) Actual ligation frequencies in the chromosome conformation capture procedure. *PloS one*, **8**(3): e60403.

[8] Gavrilov Aa, Gushchanskaya ES, Strelkova O, Zhironkina O, Kireev II, Iarovaia OV, Razin SV (2013) Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. *Nucleic acids research*, **41**(6): 3563–75.

[9] Yaffe E, Tanay A (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, **43**(11): 1059–65.

[10] Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods*, **9**(10): 999–1003.

[11] Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS (2012) HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics (Oxford, England)*, **28**(23): 3131–3.

[12] Hu M, Deng K, Qin Z, Liu JS (2013) Understanding spatial organizations of chromosomes via statistical analysis of Hi-C data. *Quantitative Biology*, **1**(2): 156–174.

[13] Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M, Guigó R, *et al.* (2012) Modeling gene expression using chromatin features in various cellular contexts. *Genome biology*, **13**(9): R53.

[14] Varoquaux N, Ay F, Noble WS, Vert Jp (2014) A statistical approach for inferring the 3D structure of the genome. *Bioinformatics (Oxford, England)*, **30**(12): i26–i33.

[15] Ay F, Bunnik EM, Varoquaux N, Bol SM, Prudhomme J, Vert JP, Noble WS, Le Roch KG (2014) Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome research*.

[16] Dekker J, Marti-Renom Ma, Mirny La (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews. Genetics*, **14**(6): 390–403.

[17] Baù D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom Ma (2011) The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nature structural & molecular biology*, **18**(1): 107–14.

[18] Hu M, Deng K, Qin Z, Dixon J, Selvaraj S, Fang J, Ren B, Liu JS (2013) Bayesian inference of spatial organizations of chromosomes. *PLoS computational biology*, **9**(1): e1002893.

[19] Comoglio F, Paro R (2014) Combinatorial modeling of chromatin features quantitatively predicts DNA replication timing in Drosophila. *PLoS computational biology*, **10**(1): e1003419.

[20] Deng H, Runger G (2012) Feature selection via regularized trees. In: *Neural Networks (IJCNN), The 2012 International Joint Conference on*. IEEE, pp. 1–8.

[21] Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SaFT (2013) Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Briefings in bioinformatics*, **14**(3): 315–26.

[22] Ong CT, Corces VG (2014) CTCF: an architectural protein bridging genome topology and function. *Nature reviews. Genetics*, **15**(4): 234–46.

[23] Naumova N, Imakaev M, Fudenberg G, Zhan Y, Lajoie BR, Mirny La, Dekker J (2013) Organization of the mitotic chromosome. *Science (New York, N.Y.)*, **342**(6161): 948–53.

[24] Le Dily F, Bau D, Pohl a, Vicent GP, Serra F, Soronellas D, Castellano G, Wright RHG, Ballare C, *et al.* (2014) Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes & Development*, **28**(19): 2151–2162.