1 DISCUSSION

1.1 MODELLING HIGHER ORDER CHROMATIN ORGANISATION

Integrative analyses of locus level chromatin data have allowed the prediction of functional chromatin states[1–4] but these states typically emcompass small regions, on the order of hundreds of basepairs, rather than the larger-scale structures we are concerned with in this work.

Our data show that accurate predictions of Hi-C derived eigenvector values, and therefore chromosome compartments, are entirely achievable. Generalisation across cell types further suggests that chromosome compartments could be inferred for those cell types without any available Hi-C data but with available ChIP-seq for a handful of chromatin features. For example, the NIH Roadmap Epigenomics project has generated histone modification data in hundreds of cell lines, tissues and developmental stages.[5,6] If models in this work were adapted to use matched inputs, this would allow comprehensive comparisons of inferred chromosome compartments across a diverse range of conditions and cell types. In the same vein, chromosome compartments are known to be related to and recapitulate other aspects of higher order chromatin organisation, including replication timing domains, Lamina associated domains and nucleolus association domains. We therefore suggest a similar modelling approach could prove successful for each of these domains of interest. An exciting idea is that an integrative model capable of identifying these LADs and NADs could forward this information to a subsequent three-dimensional reconstruction algorithm, giving a truly *in situ* perspective onto nuclear architecture.

We had less success with the prediction of TAD boundaries (Section **??**). One reason for this is that the TAD calling algorithm used in this work[7] (Methods **??**), though a published and well-used method, produces observably-flawed domain calls in some contexts. In addition the sensitivity of this method is proportional to the sample sequencing depth, which varied across our three human Hi-C datasets. We resolved TAD domains to 40 kb bins, far below the resolution of individual CTCF motifs which can generate physical domains. Indeed, given the newly-available very highly-sequenced Hi-C datasets,[8] an improved method of predicting domains might start from individual ChIP-seq peaks and consider pairs of correctly-orientated CTCF motifs. In addition, any predictive model of such domains would do well to consider the hierarchical nature of chromatin organisation (exemplified by metaTADs, Section **??**) rather than seeking simple linear discretisation of chromatin fibre into adjacent domains. Finally, we note that an accurate predictive model of lower-levels of domain organisation, be they TADs or smaller physical domains, could likely recapitulate, on aggregate, broader domains such as compartments and metaTADs, culminating in a multi-scale model of nuclear architecture from the levels of kilobases up to entire chromosomes.

1.2 DOMAIN BOUNDARIES

Chromatin domains have been described at multiple scales, from 5 Mb chromosome compartments[9] down to 185 kb contact domains[8] in human cells. Across all domains, questions remain about how they are constructed and maintained. Two competing ideas are that boundary elements, akin to the classic chromatin insulators, block intra-domain contacts and the spread of heterochromatin and hence create chromatin domains; however, another suggestion is that boundary regions are rather less important and in fact the unavoidable consequence of adjacent self-interacting domains, perhaps instead held together through internal enhancer–promoter interactions, among contacts.

In favour of functional boundary elements, knockdown of CTCF has been shown to cause increased intraTAD contacts,[10] though the same study reported an orthogonal function for cohesin

The incidental boundary hypothesis is supported by data showing that deletion of specific boundary elements in insufficient to cause adjacent domains to merge,(ref XX) In addition, the majority of CTCF sites fall within TADs rather than at their boundaries (approximately 85% of human CTCF sites are non-boundary[7]). Further it has been shown that the majority of enhancer–promoter contacts are tissue invariant,[11] hence these constitutive contacts could account for the high levels of domain conservation reported previously[7–9,12] and in this work (Chapter **??**).

As with many biological phenomena the question of whether boundary regions or internal contacts are stabilising chromatin domains is a reductive false dichotomy, and it seems likely that both boundary insulation and interTAD contacts work together to maintain chromatin domains.

1.3 DOMAIN EVOLUTION

In this work we find an array of chromatin features that, on average, are statistically associated or excluded from TAD or compartment boundaries (Section **??**). Among these are features with a long history of implications in chromatin organisation, including CTCF and cohesin subunit RAD21. We also report enrichments for Alu repeat elements (Section **??**) but no other repeat classes. Alu repeats and CTCF are linked by evidence that CTCF binding sites have in the past been dispersed through waves of retrosposon expansion.[13,14] Thus this suggests a model for the evolution of topological domains, whereby purifying selection removes those inserted CTCF sites which disrupt desirable regulatory environments, while those which bring-about efficient ”regulon” structures are favoured. Newly-released comparative Hi-C and CTCF datasets[15] provide an opportunity to investigate this proposed evolutionary model.

1.4 A NOTE ON CAUSALITY

Throughout this thesis we have probed correlative relationships: those between chromatin features and expression, or higher order chromatin structure, or domain boundaries. However even the most predictive correlations make no comment on the underlying chain of causality. Whether genome organisation is a cause of consequence of the functions of underlying genetic elements remains an open question.[16]

Two different approaches could be use to address the causality question. A standard rejoinder is to design wet-lab experiments, for example extending Hi-C studies to perturbation or differentiation time courses, such as that performed by collaborators in Chapter **??**. However, another approach is first develop theoretical models which, under simulation, recapitulate observed data, and then to use these models to generate falsifiable hypotheses about the effects of specific perturbations. This latter approach is exemplified in a study by Giorgetti *et al.*[17] where the authors applied physical polymer modelling to deconvolute population-level 5C data into single-cell conformations. The model suggests that population-level averages are explained by transient contacts in each cell, rather than persistent loops. Furthermore, these models were able to predict the effects of a genetic deletion of a CTCF site and found that contacts within a TAD contribute to maintenance of the domain, dispelling an insulation-only explanation.[17] This is also in agreement with experimental results showing that TADs can remain intact with the depletion of CTCF over a timecourse.[10]

The models built in this thesis could also be applied to predicting the effects of experimental perturbations. For example, an experiment decreasing the tri-methylation of H3K9, through down-regulation of SETDB1 or SV39H1 for example, might be expected to lead to heterochromatic regions to become more permissive or allow the transcription of marked tandem repeat sequences.[18] Our models further suggest the effect would be most pronounced in K562 cells (Section **??**). A previous experiment analysed the effects of losing H3K9me3 in SETDB1 knockout mice and found increased expression of a number of endogenous retroviruses,[19] but whether these expression changes were also coupled with alterations in chromosome compartment was not tested. Performing such an experiment over a number of timepoints could help to establish whether expression drives genomic regions to an active compartment or *vice versa*.

1.5 INSIGHTS INTO GENOME ORGANISATION

Overall our results agree with a functional model of genome architecture whereby a majority of the genome is arranged into large static compartments (Section **??**), be they Lamina associated, nucleolus associated or central and accessible chromatin. Indeed, it seems plausible that such large, constitutive anchor points may be enough to generate a significant amount of concordance in nuclear architecture between cell types.[11] These broad similarities are coupled with local structural changes in different cell lines (Section **??**), allowing cell type specific regulation of gene environments through ″looping out″, detachment from the nuclear lamina and other conceivable mechanisms of structural variation. Whether these local changes are driven by DNA-binding proteins and chromatin remodellers or functional contacts such as enhancer–promoter interactions remains unclear, though we report enrichments for both transcriptional activity and active enhancers in variable regions, but do not observe enrichments for CTCF elements, for example (Section **??**).

1.6 CONCLUSION

It has become commonplace to discuss the multi-layered, hierarchical organization of interphase chromosomes across strata ranging from nuclear compartments, down to the spectra of histone modifications and bound proteins at individual sub-genic regions. However we lack a detailed understanding of how these strata interact. We have shown that our perspectives of features occurring at different strata can be bridged by modelling approaches, and the models produced can used to explore the interrelationships between these different features quantitatively.

We constructed cell type specific models of nuclear organization, as reflected in Hi-C derived eigenvector profiles, to discover the most influential features underlying higher order structures. We found open and closed compartments to be well-correlated with combinatorial patterns of histone modifications and DNA binding proteins, enabling accurate predictive models. These models could be cross-applied successfully between cell types highlighting constellations of common structural features associated with different nuclear compartments as expected. Dissection of the most influential variables also revealed important differences between models, consistent with the known biological contrasts among these cell types, such as the prominence of EGR1 in embryonic stem cells and H3K9me3 in the leukaemia cell line. Investigation of regions showing variable nuclear organization across the three cell types under study, revealed enrichments for cell type specific enhancer activity, often nucleated at genes with known roles in cell type specific functions. Finally we used model predictions to examine boundary composition between higher order domains across cell types. Among enrichments of a large number of factors observed at different boundaries in different cell types, CTCF and YY1 were found consistently and may cooperate to establish domain boundaries. In summary, we show that integrative modelling of large chromatin dataset collections using random forests can generate useful insights into chromosome structure and seed testable hypotheses for further experimental studies.

REFERENCES

[1] Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, **9**(3): 215–6.

[2] Ram O, Goren A, Amit I, Shoresh N, Yosef N, Ernst J, Kellis M, Gymrek M, Issner R, *et al.* (2011) Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell*, **147**(7): 1628–39.

[3] ENCODE (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414): 57–74.

[4] Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes Ja, *et al.* (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research*, **41**(2): 827–41.

[5] Bernstein BE, Stamatoyannopoulos Ja, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra Ma, Beaudet AL, *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*, **28**(10): 1045–8.

[6] Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**(7539): 317–330.

[7] 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.

[8] Rao S, Huntley M, Durand N, Stamenova E, Bochkov I, Robinson J, Sanborn A, Machol I, Omer A, *et al.* (2014) A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, **159**(7): 1665–1680.

[9] 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.

[10] Zuin J, Dixon JR, van der Reijden MIJa, Ye Z, Kolovos P, Brouwer RWW, van de Corput MPC, van de Werken HJG, Knoch Ta, *et al.* (2013) Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, pp. 1–6.

[11] Bouwman BA, de Laat W (2015) Getting the genome in shape: the formation of loops, domains and compartments. *Genome Biology*, **16**(1): 154.

[12] Chambers EV, Bickmore Wa, Semple CA (2013) Divergence of mammalian higher order chromatin structure is associated with developmental loci. *PLoS computational biology*, **9**(4): e1003017.

[13] Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, **148**(1-2): 335–348.

[14] Nikolaev LG, Akopov SB, Didych Da, Sverdlov ED (2009) Vertebrate Protein CTCF and its Multiple Roles in a Large-Scale Regulation of Genome Activity. *Current genomics*, **10**(5): 294–302.

[15] Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S (2015) Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Reports*, **10**(8): 1297–1309.

[16] Sexton T, Cavalli G (2015) Review The Role of Chromosome Domains in Shaping the Functional Genome. *Cell*, **160**(6): 1049–1059.

[17] Giorgetti L, Galupa R, Nora EP, Piolot T, Lam F, Dekker J, Tiana G, Heard E (2014) Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, **157**(4): 950–963.

[18] Kim J, Kim H (2012) Recruitment and biological consequences of histone modification of H3K27me3 and H3K9me3. *ILAR journal / National Research Council, Institute of Laboratory Animal Resources*, **53**(3-4): 232–9.

[19] Karimi MM, Goyal P, Maksakova Ia, Bilenky M, Leung D, Tang JX, Shinkai Y, Mager DL, Jones S, *et al.* (2011) DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mescs. *Cell Stem Cell*, **8**(6): 676–687.