1 DISCUSSION

1.1 MODELLING HIGHER ORDER CHROMATIN ORGANISATION

Prior to the results presented in this thesis, much of the research into computational modelling of chromatin has been focused either on learning functional chromatin states from histone modifications and transcription factors (e.g. 1–10), spanning small regions on the order of hundreds of basepairs, or alternatively on the inference of the overall three-dimensional chromatin fibre trajectory based on conformation data (e.g. 11–19). In this work we attempt an intermediate approach, in which we use locus-level chromatin information to model higher order characteristics of nuclear architecture, such as chromosomal compartments and topological domains.

Our data show that accurate predictions of Hi-C derived chromosome compartment eigenvectors using locus-level chromatin features alone are entirely achievable (Section **??**). 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.[20,21] If the novel 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, nuclear lamina associated domains and nucleolus association domains.[22–24] 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, which could then use this information to generate a comprehensive, *in situ* perspective on 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[25] (Methods **??**), though a published and widely 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. Another consideration is that we resolved TAD domains to 40 kb bins, far removed from the approximately 15 basepair CTCF motifs which can generate physical domains. Indeed, given the recent release of some very-deeply-sequenced Hi-C datasets,[26] 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 non-overlapping 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: FUNCTIONAL OR INCIDENTAL?

Chromatin domains have been described at multiple scales, from 5 Mb chromosome compartments[22] down to 185 kb contact domains[26] in human cells. Across all domains, many 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 an unavoidable consequence of adjacent self-interacting domains, which are perhaps instead held together through internal enhancer–promoter interactions and other contacts. The importance of boundary elements has implications for the re-establishment process of domains during the cell cycle, for example, where it has been shown that domains are entirely absent during mitosis but then re-established in early G1 phase through an as-yet-unknown mechanism.[27,28] If boundary elements bring about domains, this may hint that key boundary-binding factors are retained through mitosis, else restored through sequence motifs. The alternative, rebuilding domains through internal contacts, would require a highly-reproducible and deterministic mechanism of reconnecting specific functional interactions in sequence.

In favour of functional boundary elements, both knockdown of CTCF[29] and deletion of a specific boundary element[30] have been shown to increase inter-TAD contacts, suggesting boundaries do indeed contribute to domain delineation. In this thesis we report an array of boundary enrichments and depletions (Section **??**), which at minimum suggests some directed biological process is in effect at boundaries. Nonetheless not all observed boundary enrichments and depletions are expected to have a detectable function; it has been shown for example that removal of the H3K27me3 mark had no effect on domain boundaries.[30] One potential functional consequence of boundaries could be that genes positioned adjacent to or over a domain boundary might be most amenable to dynamic regulation, for example by associating or disassociating from the nuclear lamina. Enrichments for gene promoters have been noted at domain boundaries in this work (e.g. Section **??**) and in previous studies.[25]

Alternatively, this boundary enrichment could be due to promoter–promoter looping inducing domain boundaries.[31–33]

The link between chromatin domain boundaries and transcription deserves additional consideration. Many of the boundary enrichments we report in this work are associated with transcriptional activity, such as POL2, H3K36me3 and H3K9ac (Section **??**), and as just discussed, boundaries are also enriched for gene promoters. Combined these results hint at a functional relationship between domain boundaries and transcriptional machinery, but it is not immediately clear why this should be the case. A biomechanical explanation could be that, in such cases where boundaries are formed from chromatin loops, a region of active transcription along with local histone acetylation would enable sufficient flexibility of the chromatin fibre to allow a loop hinge to form. However a study of chromatin domains over the *CFTR* locus found that TAD boundaries intersecting with promoters were present across cell types regardless of transcriptional state.[34] Another study reports that both transcribed and non-transcribed promoters are enriched at domain boundaries in *Drosophila*,[35] and so suggests gene density rather than transcriptional status could be a driver of domain boundary formation.

The incidental boundary hypothesis is supported by data showing that deletion of specific boundary elements, while increasing intra-TAD interactions, is insufficient to cause adjacent domains to completely merge,[30] suggesting the presence of other factors mediating domain stability. In addition, the majority of CTCF binding sites—currently thought to be the principal architects of domain boundaries—fall within TADs rather than at their boundaries (approximately 85% of human CTCF sites are non-boundary[25]). This strongly suggests CTCF binding alone is insufficient to bring about a domain boundary. Further it has been shown that the majority of enhancer–promoter contacts are tissue invariant,[28] hence if functioning as anchors of structural domains, these constitutive contacts could account for the high levels of domain conservation reported previously[22,25,26,36] and in this work (Chapter **??**).

As is the case with many biological phenomena, the question of whether boundary regions or internal contacts are responsible for chromatin domains is reductive, and it seems likely that both boundary insulation and intra-TAD 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 studies implicating them 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 retrotransposon expansion.[37,38] 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[39] offer an opportunity to investigate this proposed evolutionary model.

1.4 ON CAUSALITY

Throughout this thesis we have probed correlative relationships, including those between chromatin features and either expression (Section **??**), higher order chromatin structure (Section **??**), or domain boundaries (Section **??**). However even the most predictive correlations make no comment on the underlying chain of causality. Whether genome organisation is a cause or consequence of the functions of underlying genetic elements remains an open question.[33]

Two different approaches could be used 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 to first develop theoretical models which, under simulation, recapitulate observed data, and then to use these models to generate testable hypotheses about the effects of specific perturbations. This latter approach is exemplified in a study by Giorgetti *et al.*[19] 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. Subsequently these models were able to accurately predict the effects of a genetic deletion of a CTCF site separating the *Tsix* and *Xist* TADs.[19]

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, perhaps through down-regulation of SETDB1 or SV39H1, might be expected to lead to heterochromatic regions becoming more permissive and allow the transcription of marked tandem repeat sequences.[40] 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,[41] 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 transcriptional machinery 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.[28] These broad similarities are coupled with local structural changes in different cell lines (Section **??**, Chapter **??**), allowing cell type specific regulation of loci 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 by functional contacts such as enhancer–promoter interactions remains unclear.

1.6 SUMMARY

Work presented in this thesis began with the collection and uniform reprocessing of publicly-available genome-wide Hi-C datasets (Chapter **??**). While many studies present only their own novel data, we demonstrated the utility in making use of that which is already openly-available. We compared this chromosome conformation data across three human cell types of diverse origin (human embryonic stem cell H1 hESC, transformed lymphoblastoid cell line GM12878 and the chronic myelogenous leukemic line K562), and found strong conservation of higher order chromatin structure. Where we found regions of variable structure between cell types, these were enriched for cell type specific enhancer and transcriptional activity, and also showed dramatic changes in their long-range contact profiles. These results demonstrate the close relationship between genome structure and function across three human cell types.

In Chapter **??**, we reproduced and extended a predictive model of transcriptional output, before returning to our reprocessed Hi-C data to employ a similar machine learning and model dissection paradigm. Our models of compartment eigenvectors showed high predictive accuracy and in doing so learned general associative rules between locus-level chromatin features and chromosome compartments. Probing variable importance within these models revealed some differences consistent with the biology of the cell type in which a model was learned, whereas other dissimilarities appeared to be the result of collinear clusters within our feature space (Section **??**).

We also examine boundary composition across cell types and at varying levels of higher order chromatin structure, including TADs, chromosome compartments and those of a newly-proposed layer linking the two: metaTADs (Chapter **??**). Led by these observed enrichments and depletions, we report modest success with the prediction of TAD boundaries in the absence of Hi-C. Higher-resolution chromatin conformation capture data and improved domain calling algorithms will undoubtedly enable more powerful boundary-predictive models in the near future, which in turn could allow broad comparisons of inferred higher order chromatin structure without the application of costly and time-consuming genome-wide C-methods.

In summary, we show that integrative modelling of large chromatin dataset collections can generate useful insights into nuclear architecture and seed testable hypotheses for further study. As this thesis neared completion, another study was published on the prediction of chromosome compartments;[42] while just a month earlier, a separate publication reported a predictive model of TAD boundaries built from histone modifications.[43] These very recent studies, those presented throughout this thesis, and others no doubt soon to emerge, are proving machine learning and statistical analyses to be powerful and vital apparatus for advancing our understanding of higher order chromatin organisation.

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] Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, Ernst J, Kellis M, Ren B (2013) RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS computational biology*, **9**(3): e1002968.

[6] Song J, Chen KC (2015) Spectacle: fast chromatin state annotation using spectral learning. *Genome Biology*, **16**(1): 33.

[7] Arvey A, Agius P, Noble WS, Leslie C (2012) Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome research*, **22**(9): 1723–34.

[8] Luo C, Sidote DJ, Zhang Y, Kerstetter Ra, Michael TP, Lam E (2013) Integrative analysis of chromatin states in Arabidopsis identified potential regulatory mechanisms for natural antisense transcript production. *Plant Journal*, **73**(1): 77–90.

[9] Bednarz P, Wilczyski B (2014) Supervised learning method for predicting chromatin boundary associated insulator elements. *Journal of Bioinformatics and Computational Biology*, **12**(06): 1442006.

[10] Larson JL, Huttenhower C, Quackenbush J, Yuan GC (2013) A tiered hidden Markov model characterizes multi-scale chromatin states. *Genomics*, **102**(1): 1–7.

[11] 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.

[12] 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.

[13] 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.

[14] Lesne A, Riposo J, Roger P, Cournac A, Mozziconacci J (2014) 3D genome reconstruction from chromosomal contacts. *Nature methods*, (september).

[15] Trieu T, Cheng J (2014) Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic Acids Research*, **42**(7): 1–11.

[16] Peng C, Fu LY, Dong PF, Deng ZL, Li JX, Wang XT, Zhang HY (2013) The sequencing bias relaxed characteristics of Hi-C derived data and implications for chromatin 3D modeling. *Nucleic Acids Research*, **41**(19).

[17] 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*.

[18] Caudai C, Salerno E, Zoppè M, Tonazzini A (2015) Inferring 3D chromatin structure using a multiscale approach based on quaternions. *BMC Bioinformatics*, **16**(1): 234.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Pombo A, Dillon N (2015) Three-dimensional genome architecture: players and mechanisms. *Nature Reviews Molecular Cell Biology*, **16**(4): 245–257.

[24] Bickmore Wa (2013) The spatial organization of the human genome. *Annual review of genomics and human genetics*, **14**: 67–84.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**(7398): 381–5.

[31] Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**(1-2): 84–98.

[32] Sanyal A, Lajoie BR, Jain G, Dekker J (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**(7414): 109–13.

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

[34] Smith E, Lajoie B, Jain G, Dekker J (2016) Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus. *The American Journal of Human Genetics*, **98**(1): 185–201.

[35] Hou C, Li L, Qin ZS, Corces VG (2012) Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains. *Molecular Cell*, **48**(3): 471–484.

[36] 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.

[37] Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonalves Â, 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] Fortin JP, Hansen KD (2015) Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biology*, **16**(1): 180.

[43] Huang J, Marco E, Pinello L, Yuan GC (2015) Predicting chromatin organization using histone marks. *Genome Biology*, **16**(1): 162.