1 DISCUSSION

Since the publication of the Hi-C assay, there has been a flurry of follow-up papers largely focused on improving the technique and/or employing increasingly deep sequencing to achieve higher resolution analyses. A side-effect of this process is a growing bank of chromosome conformation datasets that, following proper reprocessing, can be compared and contrasted to gain biological insights into the importance of higher order chromatin organisation.

Despite the increasing availability of

Previous studies have compared regions of higher order structure between cell types[1] and species[2] but often anecdotally, on a selected region. Here we perform a multi-cell line comparison is a quantitative way, to

1.1 PREDICTIVE MODELS OF CHROMOSOME COMPARTMENTS

Integrative analyses of locus level chromatin data have allowed the prediction of functional chromatin states[3–6] but these states typically emcompass small regions such as the enhancers examined here. The prediction of higher order chromatin domains has received much less attention, and it was not clear until now that sufficient data existed to allow accurate predictions. Our data show that accurate predictions of Hi-C derived eigenvector values, and the nuclear compartment domains based upon them, are entirely feasible. Strong and significant correlations are seen between cell types for a variety of human higher order domains, deliniating variation in replication timing, lamin association and nuclear compartments derived from Hi-C eigenvectors[7]. The data presented here therefore suggest that a variety of such domains could be successfully modelled. Given the fact that the binding patterns of most human chromatin components have not yet been mapped the models presented here are remarkably successful, though will undoubtedly improve with further data and algorithm development. 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. It is not possible to distinguish cause and effect using the current data, but it seems likely that the alterations in domain organization occur prior to enhancer activity.

The recent abundance of epigenomic data in model cell types has enabled accurate modelling of the transcriptional output of human promoters, and a rigorously

quantitative assessment of the most influential chromatin features underlying gene expression[8]. 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. Our analysis exploits Hi-C datasets that have been re-analysed, from the intitial sequence read mapping onwards, identically for three different cell types. These data were collated with 35 locus level ENCODE chromatin datasets, also processed identically, and matched across the same cell types. In common with previous studies[2,7], we observed good concordance of higher order chromatin structure, reflected in Hi-C data, between different cell types. 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.

The current data suggest that the contributions of certain locus level chromatin features to higher order structures vary between cell types. Striking examples include the strong influence of H3K9me3 in K562 leukemia cells, and EGR1 binding in H1 hESC. EGR1 is a pivotal regulator of cell fate and mitogenesis with critical roles in development and cancer[9]. While the patterns of repressive H3K9me3 accumulation have been a focus in the cancer literature and have been proposed as a diagnostic marker in leukemia[10]. Similarly, the model for GM12878 (Epstein-Barr virus transformed lymphoblasoid) cells shows a disproportionate influence of ATF3 binding patterns, and ATF3 induction is a known consequence of virus transformed cells[11]. Thus, the most cell type specific features in these models may be important indicators of cell type specific functions. These cell type specific features present a paradox, in view of the strong correlations in organization genome wide across different cell types[2,7], and the demonstration that models trained in one cell type often perform well with data from other cell types. These contradictory observations are reconciled by the presence of inter-correlated clusters of features underlying A and B compartments. The shifting membership of these clusters evidently retains enough similarity between cell types to enable the cross application of models.

1.2 DOMAIN BOUNDARIES

Chromatin domains have been described at multiple scales, from 5 Mb chromosome compartments[1] down to 185 kb contact domains[12] 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,[13] 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[2]). Further it has been shown that the majority of enhancer–promoter contacts are tissue invariant,[14] hence these constitutive contacts could account for the high levels of domain conservation reported previously[1,2,7,12] and in this work (Chapter **??**).

In this work we find an array of chromatin features that, on average, are statistically associated or excluded from TAD or compartment boundaries. Among these are features with a long history of

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.

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. Peaks associated with active promoters were notable for both TAD and compartment boundaries in all cell types. Among the most influentual variables for the random forest models constructed for the two hematopoietic cell lines was the ubiquitous transcription factor YY1, which re-appeared in the analysis of chromatin boundary regions. Significant enrichments of YY1 were seen at TAD and nuclear compartment boundaries in all three cell types. Thus, the same protein was implicated at the level of broad genomic binding patterns (over 1 Mb intervals) and at the level of locally enriched peaks at boundary regions (spanning 100-500 Kb). This is intriguing as YY1 has recently been shown to co-localise with the architectural protein CTCF[15] and suggests that these proteins cooperate in the establishment of domain boundaries. The identification of such features, significantly enriched at boundary regions, provides potential targets for deletion in experimental studies further exploring the structure and function of domains (e.g.[16]). Both cell type specific and general constituents of boundaries may have utility in the biomedical interpretation of genomic variation in noncoding regions of the genome.

1.3 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.[17]

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.*[18] 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.[18] This is also in agreement with experimental results showing that TADs can remain intact with the depletion of CTCF over a timecourse.[13]

The predictive models generated in this thesis could also be applied to predicting the effects of experimental perturbations.

1.4 INSIGHTS INTO HIGHER ORDER CHROMATIN ORGANISATION

Our results agree with a functional model of genome architecture whereby a majority of the genome is arranged into large static compartments, be they Lamina associated, nucleolus associated or central and accessible chromatin. Indeed, it seems possible that such large, constitutive anchor points may be enough to generate a significant amount of concordance in nuclear architecture between cell types.[14] This broad overview is coupled with local changes in different tissues, allowing cell type specific regulation of gene environments through ”looping out”, detachment from the nuclear lamina and other conceivable mechanisms of variation. Whether these local changes are driven by DNA-binding proteins and chromatin remodellers or another biological mechanism remains unclear.

1.5 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.

1.6 FUTURE RESEARCH

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

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] Zwang Y, Oren M, Yarden Y (2012) Consistency test of the cell cycle: roles for p53 and EGR1. *Cancer research*, **72**(5): 1051–4.

[10] Müller-Tidow C, Klein HU, Hascher A, Isken F, Tickenbrock L, Thoennissen N, Agrawal-Singh S, Tschanter P, Disselhoff C, *et al.* (2010) Profiling of histone H3 lysine 9 trimethylation levels predicts transcription factor activity and survival in acute myeloid leukemia. *Blood*, **116**(18): 3564–71.

[11] Hagmeyer BM, Duyndam MC, Angel P, de Groot RP, Verlaan M, Elfferich P, van der Eb A, Zantema A (1996) Altered AP-1/ATF complexes in adenovirus-E1-transformed cells due to EIA-dependent induction of ATF3. *Oncogene*, **12**: 1025–1032.

[12] 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.

[13] 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.

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

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

[16] 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.

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

[18] 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.