Unravelling higher order genome organisation [working title]

Discussion

Benjamin L. Moore

June 23, 2015

igmm
INSTITUTE OF GENETICS
& MOLECULAR MEDICINE

MRC Medical Research Council

THE UNIVERSITY OF EDINBURGH

CANCER RESEARCH UK

1 DISCUSSION

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[?]. 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[? ?], we observed good concordance of higher order chromatin structure, relected 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.

Integrative analyses of locus level chromatin data have allowed the prediction of functional chromatin states[? ? ? ?] 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[?]. 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 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[?]. 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[?]. 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[?]. 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[? ?], 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.

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[?] 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.[?]). 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.

1.1 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.