Unravelling higher order genome organisation [working title]

Results 3: Domain boundaries

Benjamin L. Moore

July 17, 2015

igmm
INSTITUTE OF GENETICS
& MOLECULAR MEDICINE

MRC Medical Research Council

THE UNIVERSITY OF EDINBURGH

CANCER RESEARCH UK

1 CHROMATIN DOMAIN BOUNDARIES

1.1 INTRODUCTION

Multiple studies have defined chromatin domains of different types, for example: chromosome compartments;[?] topological associating domains (TADs);[?] contact and loop domains;[?] physical domains;[? ?] and others.[?] The existence of these domains necessitates ″boundary regions″ either between consecutive domains or bookending more sparsely-positioned domains, however the functional relevance of said boundary regions is still open to debate.

In their study of topological domains, Dixon *et al.* identified average enrichments over TAD boundary regions in both human and mouse for various features including CTCF and PolII.[?] Boundaries were also enriched for signs of active transcription, such as with the histone modification H3k36me3. These results, coupled with an observable enrichment for promoters at domain boundaries, have lead to the theory that boundaries may act as an additional layer of transcriptional control,[?] however an alternative theory could be that looping between enhancer elements and promoters results in an observable boundary through C-method experiments.[?] Another non-exclusive explanation is that if chromatin domains represent co-regulatory regions as is widely thought,[? ? ?] boundaries themselves could be mere side-effects and as such of limited biological interest.

An obvious experiment to resolve these opposing theories would be to delete a predicted boundary region and test for local changes in both contacts and expression. Such an experiment was performed on a region of the human X-chromosome containing the genes encoding the dosage-compensation long non-coding RNAs Xist and Tsix, which are separated by a TAD boundary.[?] This study found that while histone modifications within the body of a TAD could be removed without affecting the structure, deletion of a boundary did have an effect and lead to increased intradomain contacts.[?] Surpsingingly however, this effect was not total and some observable barrier remained, lending evidence that TADs may be centrally constrained, rather than by their borders.[?]

A second experiment used CRISPR genome editing to link TAD boundary changes with limb development disorders,[?] indicating that boundary changes could provide an underlying explanation for pathogenic non-coding structural variants.[?] Similarly, domain boundaries on X-chromosomes were found to be weakened following the disruption of condensation binding sites.[?] Together these studies suggest a complex scenario whereby TAD boundaries are an important structural feature, yet do not fully explain domain partitioning.

Computational analysis of boundaries has emerged during the time this work was completed. Border ″strength″, here defined by the ratio of total intra:inter-domain contacts, was found to correlate with increased occupancy of a combination of bound architectural proteins.[?]

Many questions remain about chromatin boundaries. For example, are the observed enrichments persistent across cell types and how do they compare across organisation strata, such as compartments and TADs? Through computational analysis of the set of boundaries re-called from published datasets, we can investigate these questions and probe boundary enrichments across a broad array of locus-level chromatin features.

1.2 TAD AND COMPARTMENT BOUNDARIES

The mammalian genome is organized into TADs, predominantly self-interacting chromatin domains, with boundary regions reportedly associated with pronounced peaks and troughs of particular features within 500 kb of the predicted boundary.[?] Exploration of this phenomenon using a set of 24 mouse ESC chromatin features (and a smaller number of human ESC features) reportedly revealed enrichment peaks of CTCF, H3K4me3 and H3K36me3, as well as a pronounced dip in H3K9me3, suggesting that high levels of transcription may contribute to boundary formation.[?] However, it was unclear whether other features show unusual patterns in TAD boundary regions, and whether the constellation of features involved changes between cell types. The features associated with boundaries separating A and B compartments calculated from Hi-C eigenvectors have not been studied to our knowledge. The datasets assembled here, consisting of 35 matched chromatin features across three cell types, allow us to conduct the first comparative study of the constituents of human TAD and compartment boundary regions.

We derived TAD boundaries according to established methods (see Methods XX) for all three cell types under study. We then sought evidence for significantly enriched or depleted features at TAD boundary regions using a conservative approach (a nonparametric statistical test and Bonferroni multiple testing correction, see Methods XX).

Our findings confirmed the previously reported peaks (CTCF and POL2) and dip (H3K9me3) in ESC data, but also revealed substantial heterogeneity between cell types. CTCF binding was found enriched at TAD boundaries across all cell types, but other features, including H3K36me3 and H3K4me3, show dramatic peaks of enrichment in H1 hESC cells that are not seen consistently in other cell types (Figure 6, Additional file 1: Figure S12). Although the dip in H3K9me3 at TAD boundaries is seen in all cell types, the extent of the depletion varies and is weakest in H1 hESC cells. Many other features show significant, though often modest, enrichments in a particular cell type. However, overall the complexity of TAD boundaries (measured as the number of strongly enriched features) is notably higher in H1 hESC than in the other two, more differentiated, cell types (Figure 6), involving large increases in the binding of sequence specific factors such as SP1 and JUND.

Across all three cell types several features demonstrate consistent and statistically significant patterns at TAD boundaries (Figure 6, Figure S12), including peaks associated with active transcription of genes (POL2, H3K9ac) and dips in H3K9me3, as previously reported.[?] However other novel feature peaks of interest emerge across cell types, such as peaks of H4K20me1, a modification previously implicated in chromatin compaction.[?] We also observe consistent increases in GC content at TAD boundaries, at a scale that is difficult to reconcile with the presence of smaller-scale features such as repeat elements or CpG islands (Additional file 1: Figure S12).

Where neighbouring genomic regions occupy contrasting A and B nuclear compartments, the disparity implies the presence of a boundary region. Putative compartment boundaries were identified by using an HMM to infer the state sequence of A/B compartments across the genome based on observed principal component eigenvectors. Analogously to the TAD boundary analysis we then sought significant enrichments or depletions in 36 chromatin features over these compartment boundaries (Figure 6, Figure S13). Compartment boundaries display similar spectra of enrichments to previously studied TAD boundaries[?] but at lower resolution, reflecting the different scales of these levels of organization (Figure 6B, Figure S13). Peaks associated with active promoters (POL2, TAF1, H3K9ac) are again evident. Parallel enrichments of CTCF, YY1 and H4K20me1 are also seen at compartment boundaries, as they were for TAD boundaries, in each cell type under study. In addition, compartment boundaries show enrichments of H3K79me2, which is known to play critical roles in cellular reprogram-

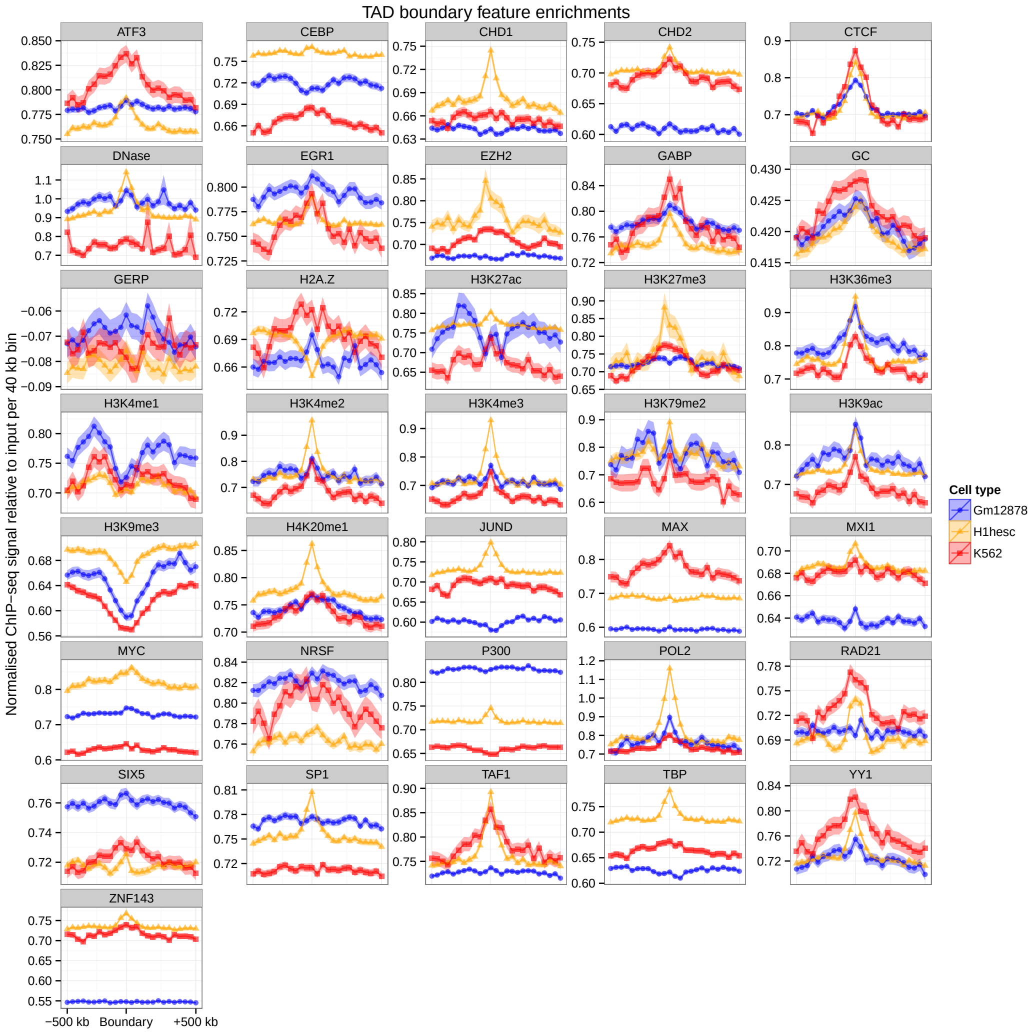

Figure 1: TAD boundary enrichments and depletions. 36 features were averaged over 1 Mb windows centred on TAD boundaries genome-wide (25 × 40 kb bins). Ribbons represent 95% confidence intervals of the mean at each position.

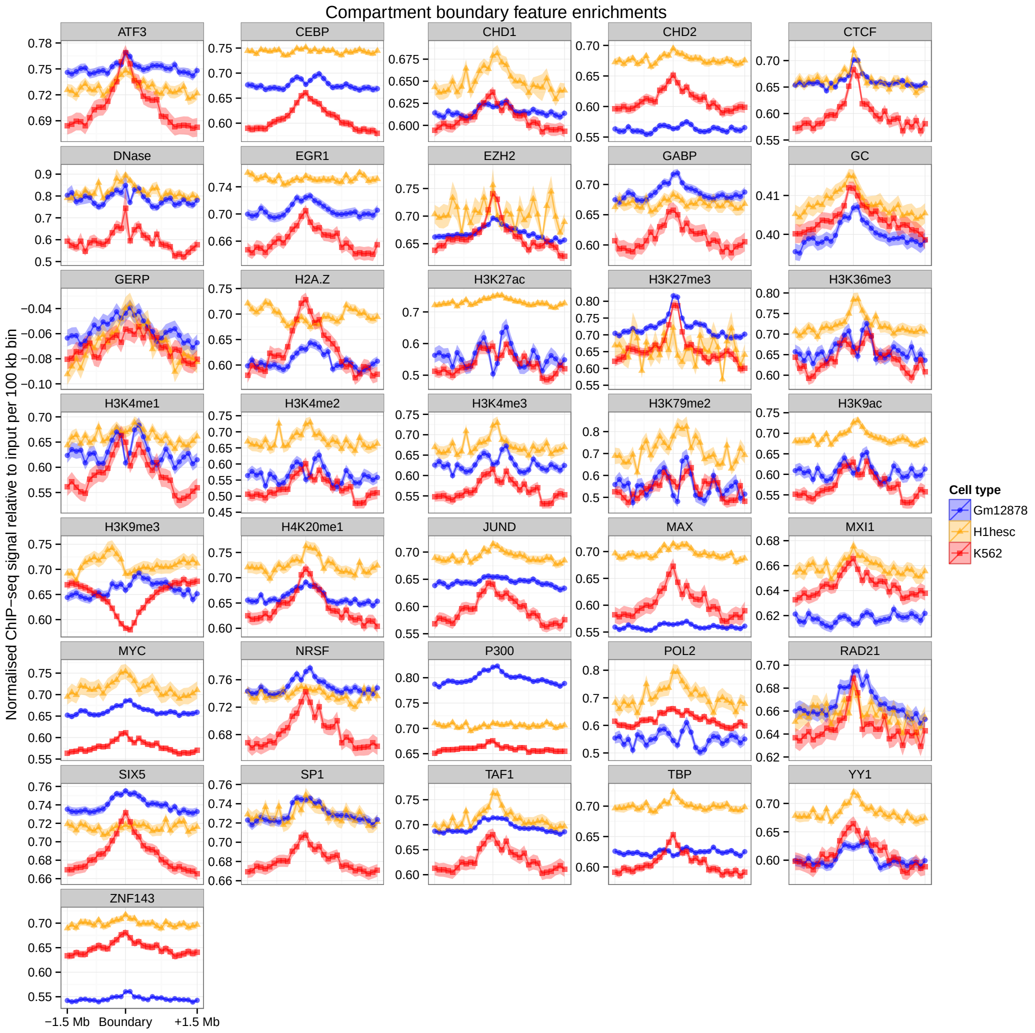

Figure 2: Compartment boundary enrichments and depletions. 36 features were averaged over 3 Mb windows centred on compartment boundaries genome-wide (30 × 100 kb bins). Ribbons represent 95% confidence intervals of the mean at each position.

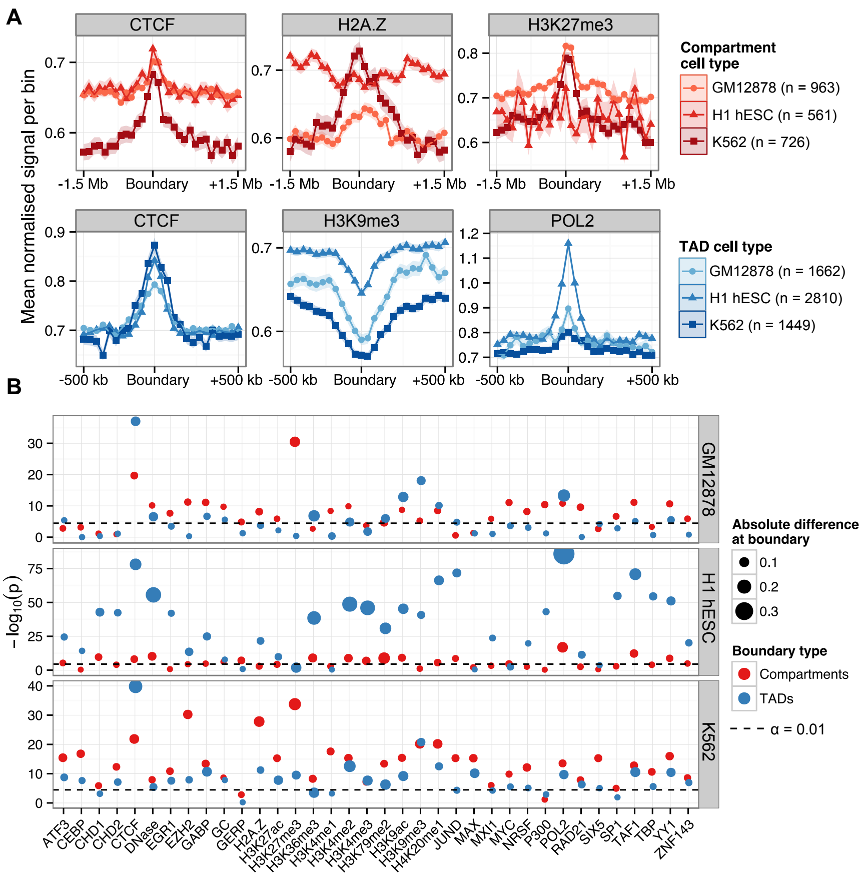

Figure 3: **Compartment and TAD boundary enrichment summary in three human cell types.** (A) Selected profiles for locus-level features are shown for TAD boundaries (CTCF, H3K9me3 and POL2) and compartment boundaries (H2A.Z, H3K4me2 and YY1), as a mean normalized ChIP-seq signal relative to input chromatin per bin (±1 standard error). TAD boundaries were examined over 40 kb bins over the 1 Mb flanking each boundary; compartment boundaries were examined over 100-kb bins over 3 Mb. (B) The significance of enrichment or depletion ($-\log_{10}(p)$ two-tailed Mann–Whitney test) of a feature was calculated as the boundary bin relative to the ten most peripheral bins (five either side). Points are scaled by the absolute mean difference in signal over the boundary relative to the mean of peripheral bins. ChIP-seq, chromatin immunoprecipitation sequencing; TAD, topological domain.

ming.[?] Remarkably, H3K79me2 has also recently been shown to mark the borders of small (hundreds of bp) regions of open chromatin.[?] Thus there may be similarities in chromatin compaction boundaries at very different scales.

Certain features show intriguing contrasts between cell types the histone variant H2A.Z lacks any trace of enrichment at H1 hESC compartment boundaries, but is significantly enriched in the other two cell types (Figure 6A), consistent with reports describing H2A.Z relocation during cellular differentiation.[?] Compartment boundaries also show enrichment for the cohesin complex subunit RAD21 in the two hematopoietic cell types , and cohesin is another factor implicated in modulating nuclear architecture in partnership with CTCF.[?] Various other enrichments with very modest effect sizes are also evident at compartment boundaries (Figure 6B, Figure S13). In contrast to TAD boundaries, the composition of compartment boundaries appears least complex in H1 hESC, relative to the other two cell types. Overall compartment and TAD boundaries are associated with overlapping spectra of chromatin features across cell types. These involve DNA binding proteins implicated in chromosome architecture (CTCF, YY1, RAD21), but also implicate the initiation and repression of transcription as critical to boundary formation. However these two boundary classes occur at different scales, with patterns of informative features typically spanning regions up to 500 Kb for TAD boundaries, and patterns associated with compartment boundaries often spanning more than 1 Mb.

1.2.1 CTCF and YY1

Significant peaks in YY1 are evident in all cell types, which is intriguing given the evidence that YY1 and CTCF cooperate to affect long distance interactions.[?] Co-binding of CTCF with YY1 has also been shown to identify a subset of highly conserved CTCF sites.[?] Co-binding of CTCF and YY1 may also therefore be a contributing factor in the establishment of TAD boundaries, which appear to be broadly conserved across mammals.[?] To test this, we split our sets of TAD boundaries into those possessing ChiP-seq peaks (region peaks called by ENCODE[?]) for CTCF, YY1, both CTCF and YY1 (overlapping peaks) and neither. We then tested each boundary subset for genome-wide enrichments of the other features in our dataset (Figure S14). Unexpectedly, we found that boundaries marked by YY1 (without overlapping CTCF peaks) were generally most strongly-enriched for other features in our dataset. We also found that boundaries lacking both CTCF and YY1 peaks showed instead the strongest enrichments for RAD21 in each cell type (Figure S14), reinforcing previous findings that describe the distinct influences of CTCF and cohesin in organizing chromatin structure.[? ? ?]

1.2.2 Repeats

Dixon *et al.*'s study of TAD boundaries identified short interspersed element (SINE) repeats as being enriched over domain boundaries and suggested roles for these repeats in altering genome organisation, in line with prior evidence.[? ?] Interestingly, SINE elements are thought to be responsible for spreading CTCF binding sites through mammalian genomes[?] (thought not in primates[?]). Analysis of recent high-resolution Hi-C data again reported a SINE B2 link with CTCF loops in mice.[?] Together these results suggest repeats could be a key component in the makeup of domain boundaries.

To investigate this, we used the `RepeatMasker`[?] software package to call repeat classes and families in the `hg19` and `mm10` genome assemblies. Counts for each annotated feature were then average over boundaries as described previously (Methods XX).

At the level of repeat class, we corroborate the findings of **?**] that the majority of repeat classes show no enrichment or depletion at TAD boundaries,

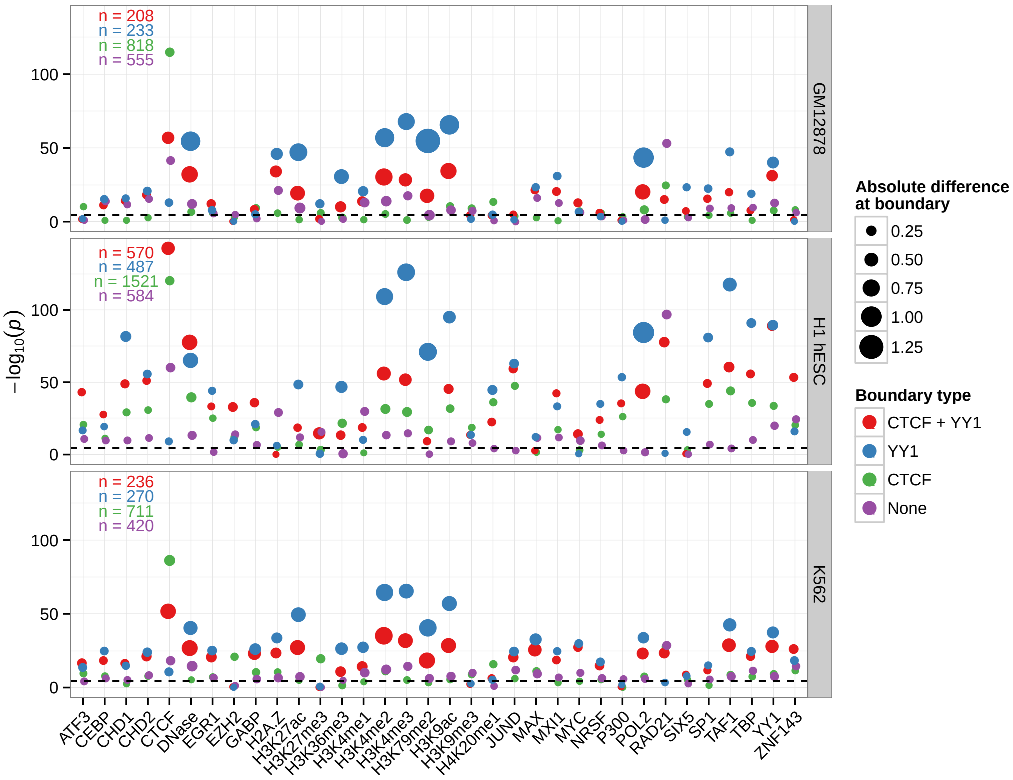

Figure 4: Distinct enrichments of CTCF and YY1 boundaries. TAD boundary feature enrichments are shown (as in Fig. 3) for boundaries split into classes based on specific enrichments: CTCF and YY1 groups are those boundaries with at least one ENCODE region peak[?] for their respective features, while CTCF + YY1 is the group of boundaries which had one or more overlapping peaks for these two factors. Boundaries in the none group has neither a CTCF or YY1 region peak called (but can still be enriched for their respective features in terms of raw signal).

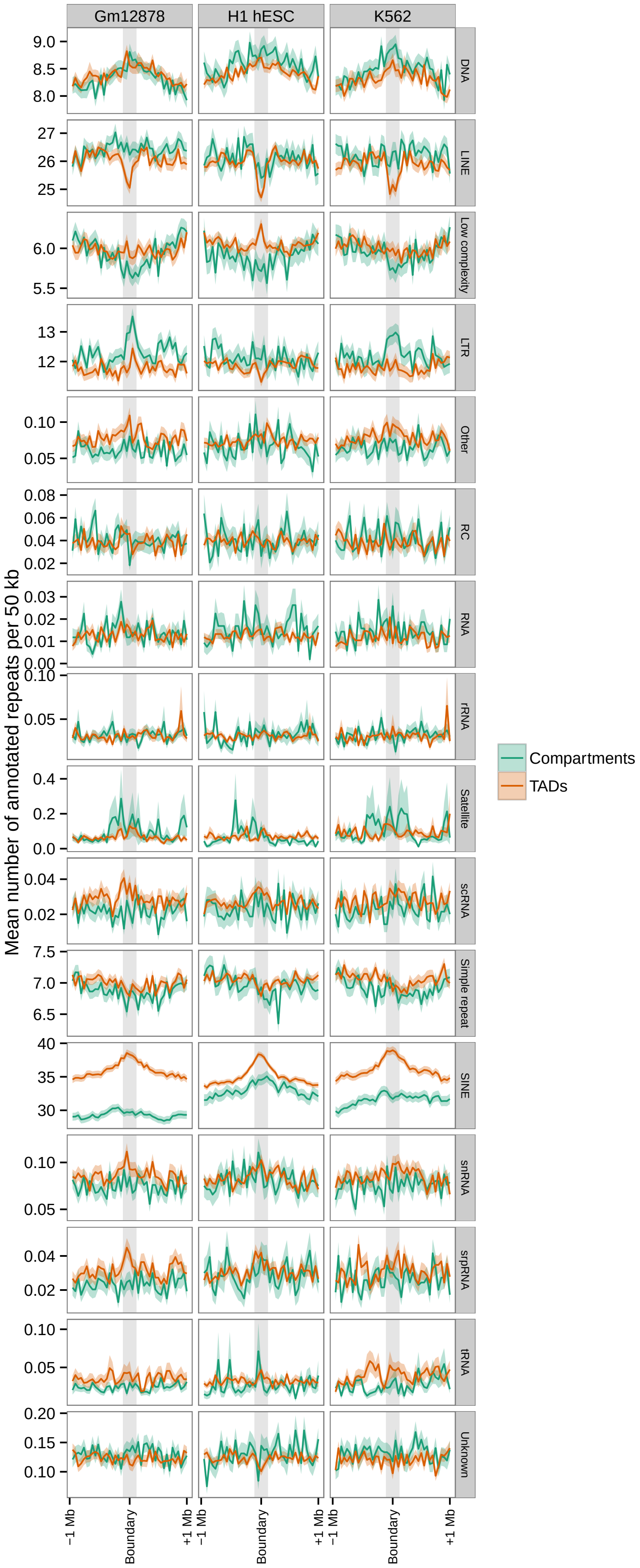

Figure 5: Repeat class average-o-grams over all TAD and compartment boundaries. RepeatMasker repeat annotations are counted per 50 kb for 1 Mb either side of each TAD and compartment boundaries. The mean count genome-wide is plotted with ±95% confidence intervals.

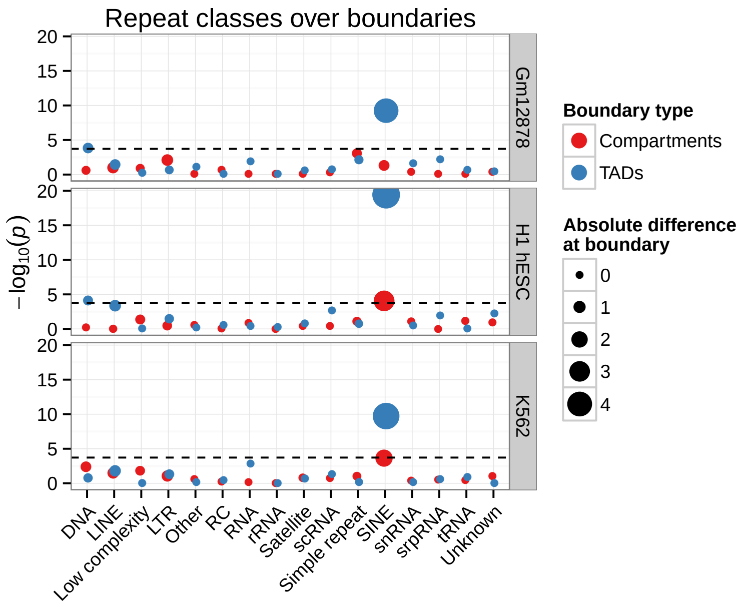

Figure 6: **Significance and effect sizes of repeat class enrichments/depletions over boundaries.** Boundary profiles (Fig. 5) were tested for enrichment or depletion of each factor at the boundary bin relative to peripheral non-boundary bins (see Methods XX). Bubble area is proportional to the raw effect size of an enrichment or depletion. The Bonferroni-corrected significance threshold is highlighted with a dashed line.

and we find that this also holds for compartment boundaries (Fig. 5). A notable exception is the short interspersed element (SINE) repeat class which appears to be enriched at TAD boundaries in each cell type. Testing the significance of this observed peak confirms this to be the case, with SINEs significantly enriched at TAD boundaries in each cell type, and borderline significant enrichments can also be observed at compartment boundaries (Fig. 6).

Repeat class profiles also suggest LINEs may be depleted over TAD boundaries and DNA repeats may be enriched at both boundary types (Fig. 5), however statistically these observations do not surpass our pre-defined significance threshold ($\alpha = 0.05$) after multiple testing correction (Fig. 6).

Repeat classes can be broken into smaller repeat families. **?**] reported that the Alu (or B1 in mouse) repeat families are enriched over TAD boundaries. Again we can reproduce this finding and extend it to compartment boundaries (Fig. 7).

1.3 DE NOVO BOUNDARY PREDICTION

We have shown TAD and compartment domain boundaries to be well-marked by a variety of features. Compartment boundaries are successfully predicted as a side-effect of modelling the continuous compartment profile eigenvector (Section XX) however a related measure of activity and repression does not exist for TADs.

We attempted to model TAD boundaries in a variety of ways: firstly a using a class-balanced classification framework and secondly through indirect models of directionality index and the downstream domain-caller HMM state.[?]

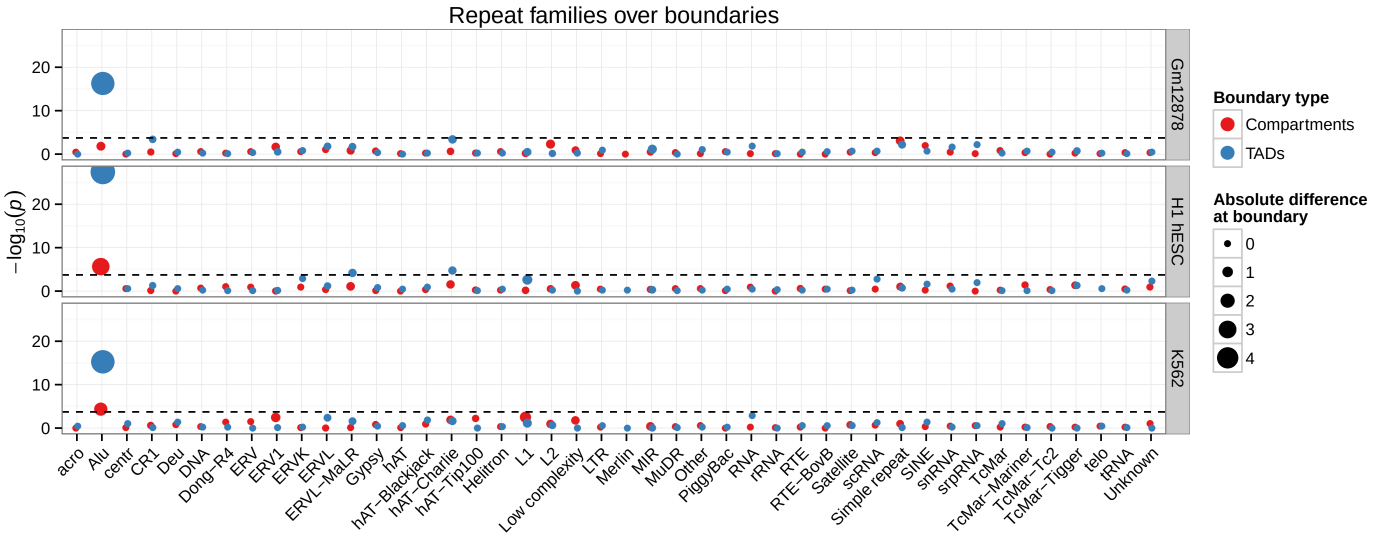

Figure 7: Significance and effect sizes of repeat family enrichments/depletions over boundaries. As per Fig. 6 but for a higher resolution repeat classification.

1.4 METATAD BOUNDARIES

Our collaborators uncovered the concept of ”metaTADs”: sequential aggregations of adjacent and strongly-interacting TADs to form a hierarchy of domain organisation covering each chromosome.

MetaTADs are constructed simply by performing constrained heretical clustering based on inter-TAD contacts. That is, those two neighbour TADs that have the largest number of interTAD contacts are linked to form a metaTAD and this process is recursed until all TADs on a chromosome are joined into a tree-like network which fully describes the hierarchical nature of domain organisation.

My contribution to this work was to explore these newly-described metaTAD structures and perform boundary analysis as was done with TADs and compartments (Section 1.2). A hypothesis to test could be that boundaries of larger metaTAD structures could display greater enrichments for boundary-defining features.

1.4.1 Lamin associated domains

1.4.2 Boundaries over a time series

1.5 OTHER BOUNDARIES

1.5.1 Giemsa bands

A recent analysis of Hi-C datasets examined the hierarchy of nuclear compartment and TAD organisation in human HeLa cells across the cell cycle. They found that interphase and metaphase chromatin structure are highly distinct, such that the TADs and compartments observed here are effectively abolished in metaphase.[?] This raises the question of how the structural organization seen in (and often shared between) interphase cells is inherited through the cell cycle.

Human Giemsa metaphase banding (G-band) pattern data have been integrated with the human genome assembly, and although such data are widely used, they are also necessarily of low resolution.[?] These G-band patterns are constant over human cell types at metaphase, but all traces of

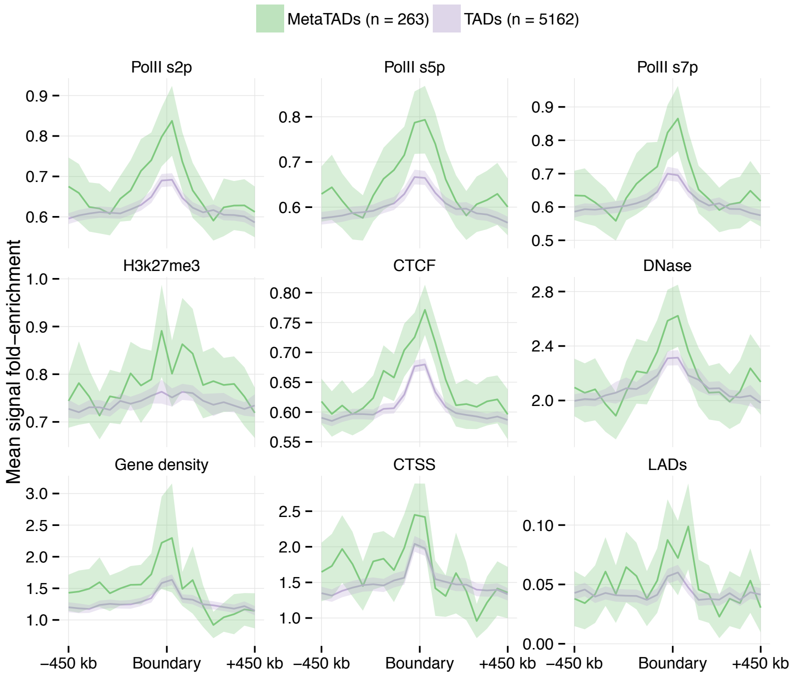

Figure 8: Large metaTADs show greater enrichments for an array of boundary features. Genome-wide profiles of epigenomic features and gene densities averaged over all TAD and metaTAD (10 – 40 Mb) boundaries (ribbons show 95% confidence intervals of the mean).

interphase higher order structure were reported to be absent at metaphase.[?] We would therefore expect no agreement between metaphase G-bands and the patterns of interphase TADs and A/B nuclear compartments defined here, over all three cell types.

We examined the genome wide similarity of all interphase domain structure boundaries to metaphase G-band boundaries, relative to an expected distribution derived by permutation (see Methods) (Figure S9). There is a significant, though extremely modest, excess of compartment boundaries within close proximity of G-band boundaries, such that 13.90% of compartment boundaries are within 500 kb of a G-band boundary (expectation = 10.50%, K-S test: $D = 0.076$, $p < 3 \times 10^{-12}$). This is seen for compartment boundaries calculated for all three cell types independently. The genome wide overlap of compartment A and B regions with particular G-band classes is nonrandom, and suggests much greater correspondence. Regions assigned to compartment A are significantly over-represented within lighter staining (especially G-negative) bands, while compartment B regions are over-represented in the most darkly staining (G-positive) bands. Approximately 40% of the genome jointly occupies interphase compartment A as well as gneg/gpos25 metaphase G-bands, or occupies the interphase B compartment as well as gpos75/gpos100 at metaphase. Again, the same trends are seen significantly across all three cell types. This agreement is not unexpected given the broad differences in G-negative and G-positive bands, with contrasting gene density, GC content and replication timing[?] that is strongly reminiscent of the contrasts between interphase A and B compartments,[?] but to our knowledge has not been directly studied before. These data suggest that across the genome most fine structure, reflected in domain boundaries, is not well preserved between interphase and metaphase. However there is evidence for conservation of broader structural categories across a substantial fraction of the genome, which may reflect broad similarities in the degree of compaction seen at many regions across the cell cycle.

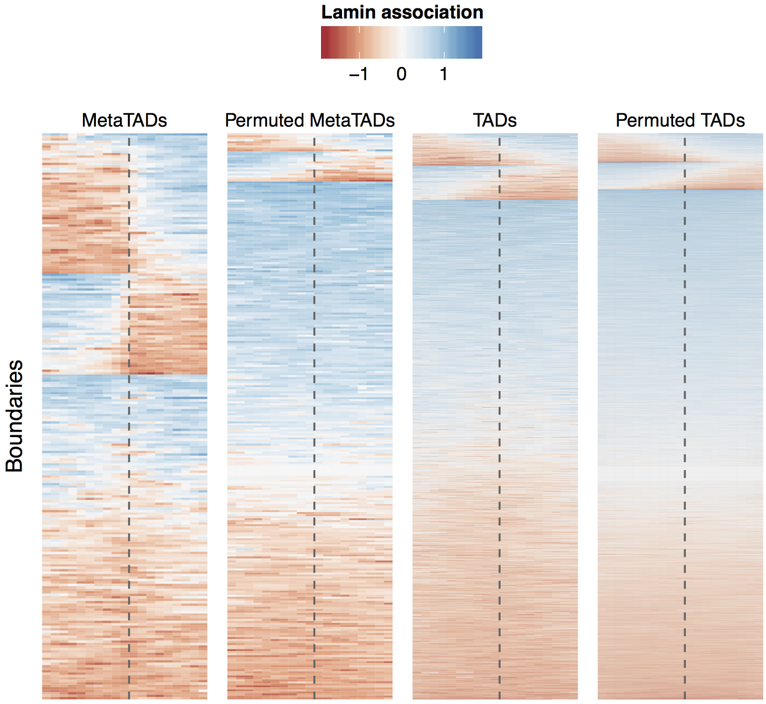

Figure 9: MetaTADs align with lain associated domains. Heatmaps of LaminB1 association microarray probe intensity values over MetaTAD boundaries (from domains of size 10 – 40 Mb) and TAD boundaries, are displayed alongside examples of circularly-permuted boundaries. 42.6% of MetaTAD boundaries (10 – 40 Mb) had an absolute linear regression coefficient $> .05$ of lamin association intensities, indicating a boundary transition (versus 15.8% expectation from 1000 circular permutations, $p < 1 \times 10^{-4}$). TAD boundaries were also significantly more associated with lamin association transitions (Observed: 11.8%, Expected: 9.5%; empirical p-value: $p < 1 \times 10^{-4}$). Profiles are shown ±450 kb from each boundary.

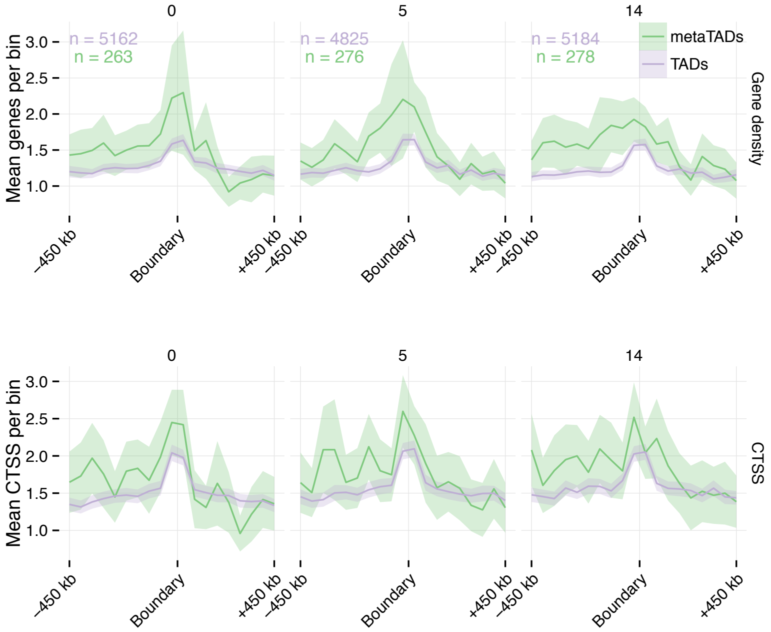

Figure 10: Observed enrichments persist over a time series. CAGE-defined active TSS (CTSS) were counted per 50 kb bin across each TAD and MetaTAD (10 – 40 Mb) boundary and averaged (ribbons show 95% confidence intervals of the mean). Gene densities refer to mean counts of annotated genes per bin, with an overlap of at least 250 bp. Peak heights suggest modestly stronger enrichments at MetaTAD boundaries relative to TAD boundaries.

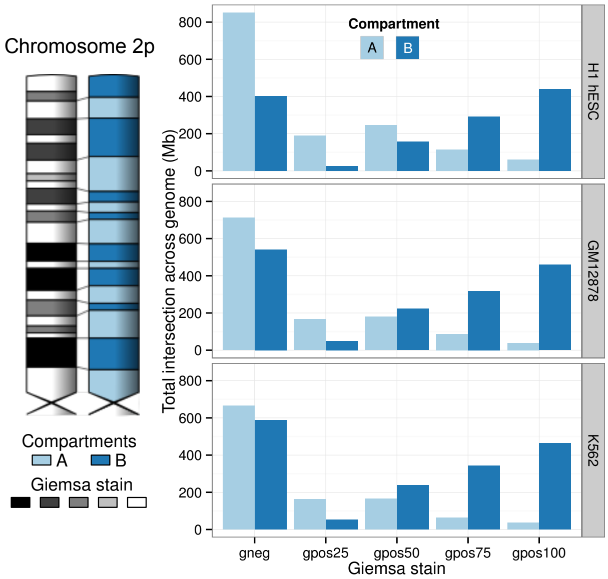

Figure 11: Giemsa =-stain bands correspond to A/B compartments. Placeholder

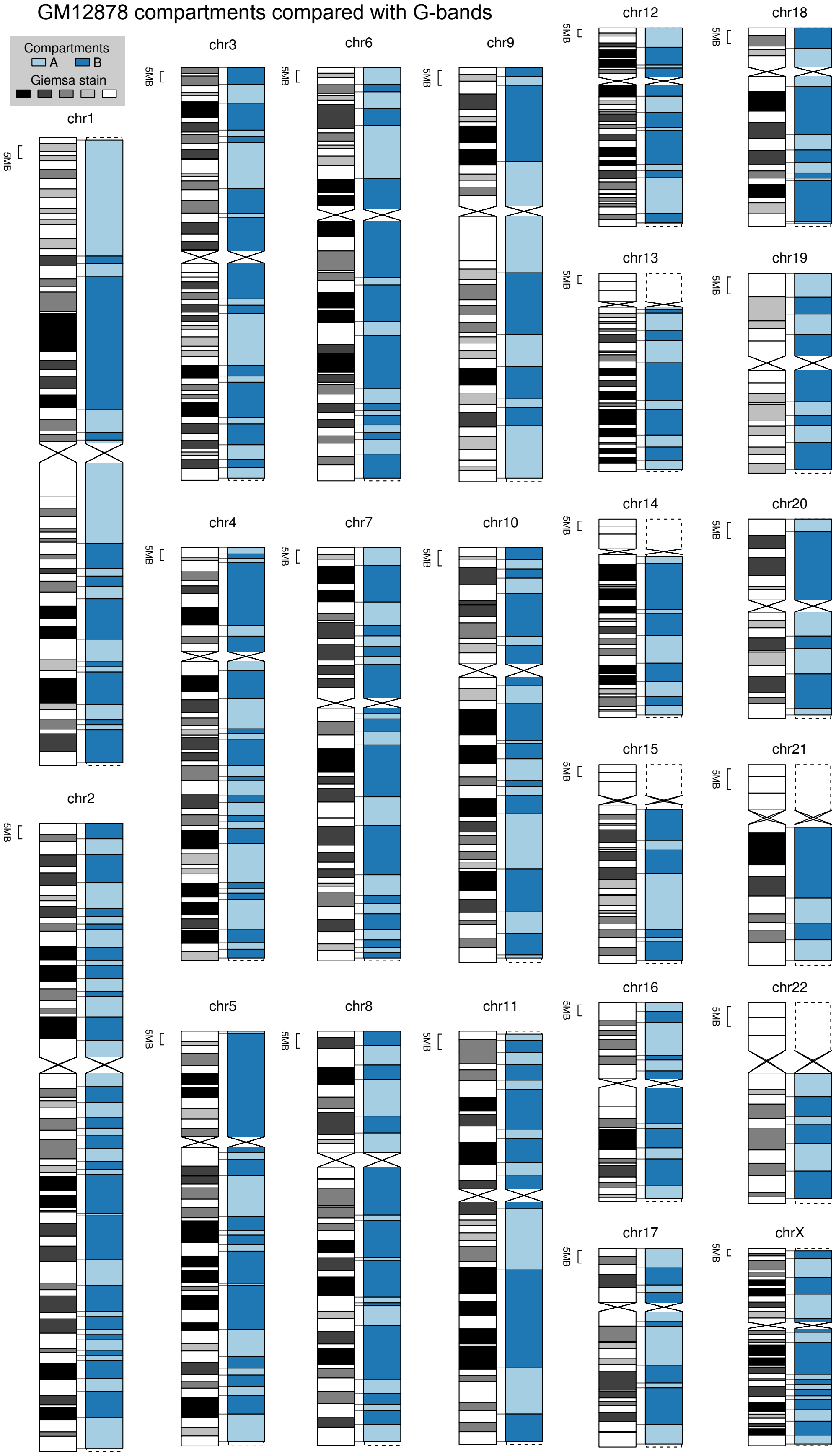

Figure 12: **Genome-wide agreement between Giemsa bands and A/B compartments in the Gm12878 cell type** Placeholder

1.5.2 Superboundaries

Thus far compartment and TAD boundaries have been considered separately, however it is of interest to consider how these boundary regions interact across scales. Open questions remain about the co-occurence of these two boundary regions, and whether