1 CHROMATIN DOMAIN BOUNDARIES

1.1 INTRODUCTION

A succession of studies have defined chromatin domains of different types, for example: A and B chromosome compartments;[1] topologically associating domains (TADs);[2] contact and loop domains;[3] physical domains;[4,5] and others.[6] The existence of these domains necessitates ”boundary regions” either between consecutive domains or bookending more separated domains, however the functional relevance of said boundary regions is still open to debate.

In their study of topological domains, Dixon *et al.*[2] 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.[2] 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.[7] However an alternative theory is that if chromatin domains represent co-regulatory regions as is widely thought,[7–9] 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.[10] This study found that while histone modifications within the body of a TAD could be removed without affecting overall domain structure, deletion of a boundary did have an effect and led to increased intradomain contacts.[10] Surprisingly however, the two domains did not completely merge, lending credence to the alternative theory that TADs may be centrally constrained, rather than by their borders.[10]

A recent experiment used CRISPR genome editing to link TAD boundary changes with limb development disorders,[11] indicating that boundary changes could provide an underlying explanation for pathogenic non-coding structural variants.[12] Similarly, domain boundaries on *C. elegans* X-chromosomes were found to be weakened following the disruption of condensin binding sites.[13] Together these studies suggest a complex scenario whereby TAD boundaries are an important structural feature, yet do not fully explain domain partitioning.

Many questions remain about chromatin boundaries. For example, are the enrichments reported in Dixon *et al.*[2] 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 BOUNDARY ANALYSIS

The mammalian genome is organized into topologically associating domains (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.[2] Exploration of this phenomenon using a set of 24 mouse ESC chromatin features (and a smaller number of human ESC features) 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.[2] However, the peaks and dips of these features lacked any estimates of statistical significance. It was also unclear whether other features might show unusual patterns in TAD boundary regions, and how the constellation of features involved might vary between cell types. Moreover, 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.

1.2.1 TAD boundaries

We derived TAD boundaries from uniformly reprocessed Hi-C data (Chapter **??**) according to established methods (see Methods **??**) 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 **??**).

Our findings confirmed the previously reported peaks (CTCF and POL2) and dip (H3K9me3) in ESC data, but also revealed substantial heterogeneity between cell types and some novel boundary features. CTCF binding was found enriched at TAD boundaries across all cell types, but other features, including H3K27me3 and H3K4me3, show dramatic peaks of enrichment in H1 hESC cells that are not seen consistently in other cell types (Fig. 1). 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

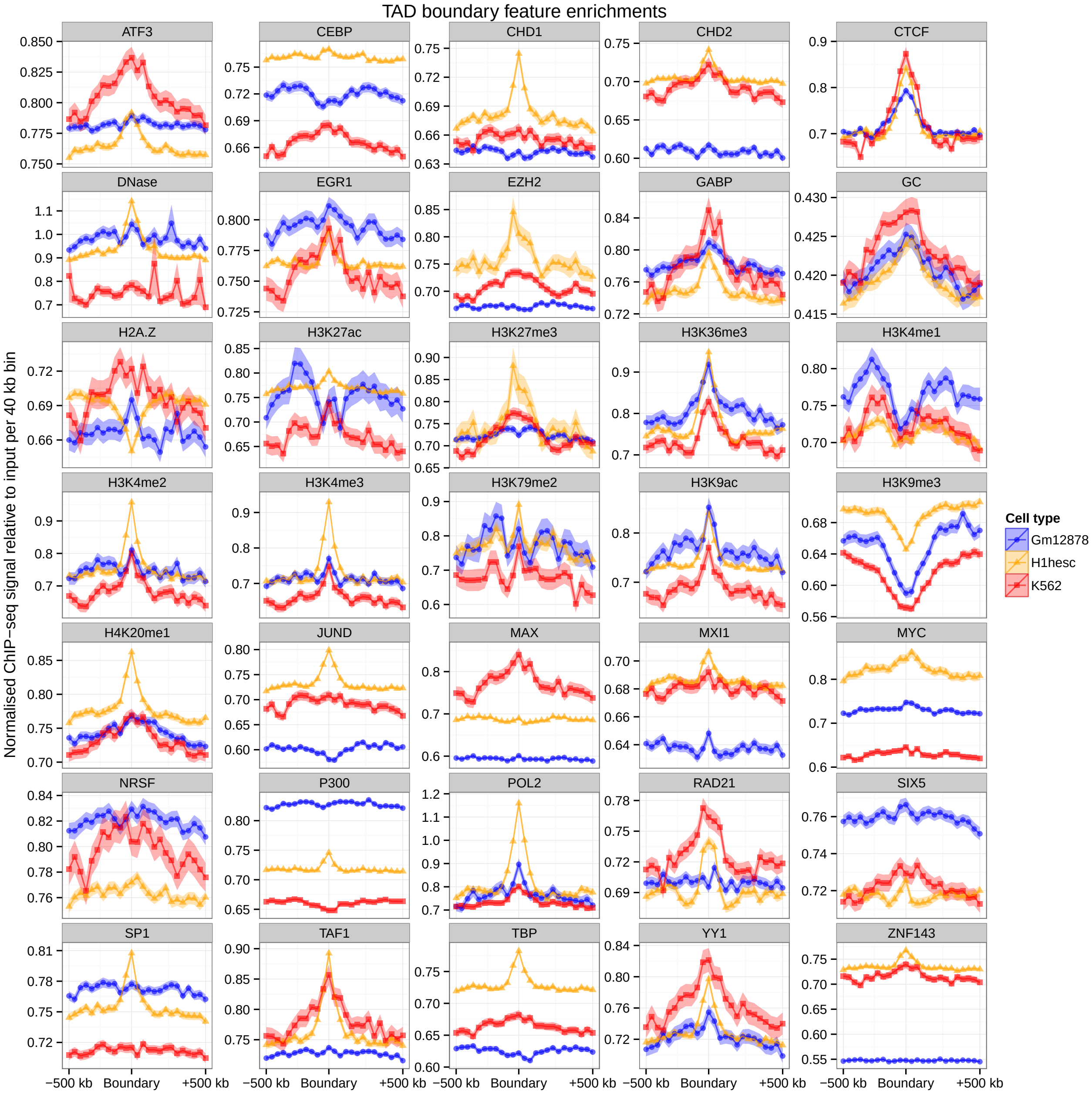

Figure 1: TAD boundary enrichments and depletions. 36 features were averaged over 1 Mb windows centred on TAD boundaries genome-wide (25 × 40 kb bins). Ribbons represent 95% confidence intervals of the mean at each position.

cells. We also note an apparent enrichment of H4K20me1 over TAD boundaries in H1 hESC cells, a modification previously implicated in chromatin compaction.[14] Finally we observe consistent increases in GC content at TAD boundaries, at a scale that is difficult to explain in terms of those much smaller GC-rich features such as binding motifs or CpG islands (Fig. 1). The statistical significance of these enrichments and depletions is considered in Section 1.2.3.

1.2.2 Compartment boundaries

Where neighbouring genomic regions occupy contrasting A and B nuclear compartments, the disparity implies the presence of a boundary region. We identified putative compartment boundaries using an HMM to infer the state sequence of A/B compartments across the genome based on observed principal component eigenvectors (Section **??**). Analogously to the TAD boundary analysis we then sought significant enrichments or depletions in our set of chromatin features over these compartment boundaries.

Compartment boundaries display similar spectra of enrichments to previously studied TAD boundaries[2] but at lower resolution, reflecting the different scales of these levels of organisation (Fig. 2). Peaks associated with active promoters (POL2, TAF1, H3K9ac) are again evident. Enrichments of CTCF and YY1 are again 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 a critical role in cellular reprogramming.[15] This histone modification has also recently been shown to mark the borders of small (hundreds of bp) regions of open chromatin,[16] hinting at similarities in chromatin boundaries at very different scales.

Certain features show intriguing contrasts between cell types. For example, the histone variant H2A.Z shows a clear enrichment over K562 compartment boundaries, but not the other two cell types (Fig. 2). 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.[17] Various other enrichments of modest effect size can also be seen (Fig. 2). In contrast with TAD boundaries, the composition of compartment boundaries appears least complex in H1 hESC, relative to the other two cell types (cf. Fig. 2).

These enrichments over compartment boundaries offer the opportunity to investigate the effects of the HMM boundary calling method used in this work with the previously-used simple thresholding approach.[1] To this end we compared that subset of boundaries retained through HMM calling (for details see Section **??**), with those that would have been called through simple thresholding when the compartment eigenvector changes sign. We find that boundary enrichments are largely reduced

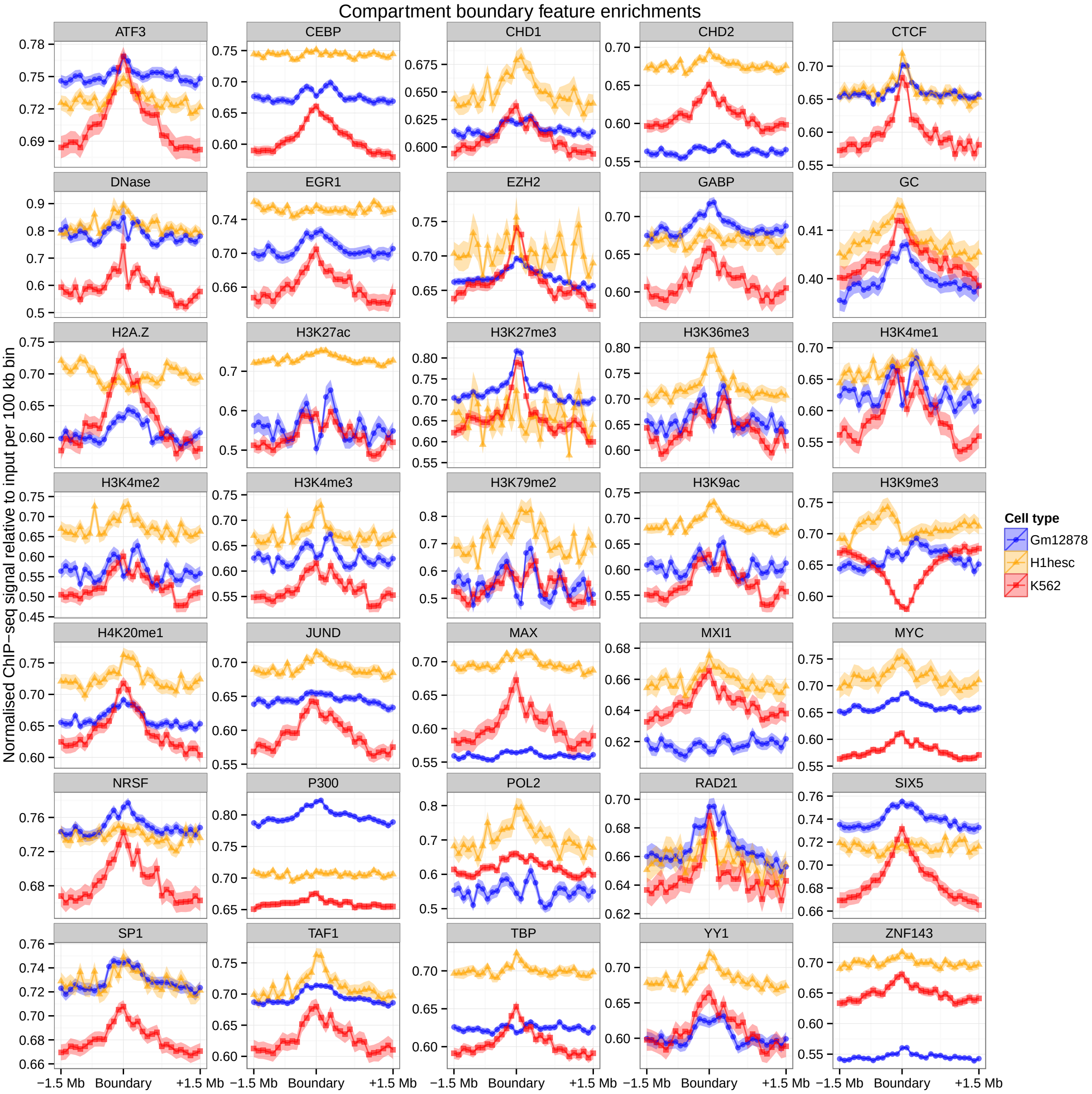

Figure 2: **Compartment boundary enrichments and depletions.** 36 features were averaged over 3 Mb windows centred on compartment boundaries genome-wide (30 × 100 kb bins). Ribbons represent 95% confidence intervals of the mean at each position.

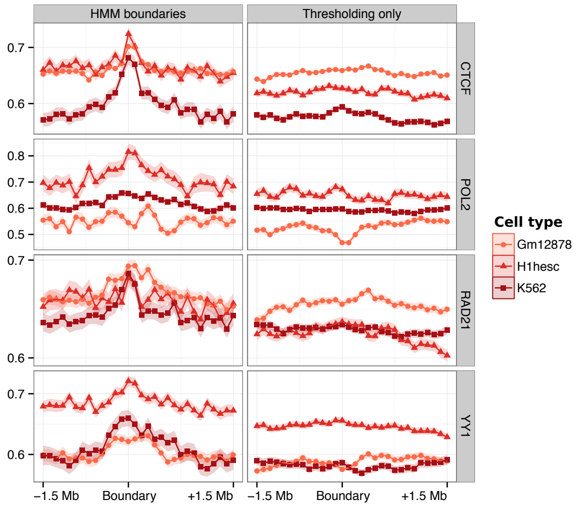

Figure 3: **Compartment boundary candidates filtered through HMM calling show reduced feature enrichments.** Boundary feature enrichments are compared for boundaries called via the HMM method described in this work (discussed in Section **??**), and boundaries omitted through the HMM method but present if a simple thresholding approach is used to delimit compartment domains. Boundary profiles shown as in Figure 2.

or absent in this latter set of boundaries (Fig. 3). This result suggests that the HMM boundary calling method is filtering noisy compartment boundary calls which would otherwise be present through a simple thresholding approach, however we note that the absence of boundary enrichment does not necessarily preclude the presence of a compartment boundary.

1.2.3 Significance testing of boundary associations

Domain boundary associations with other chromatin features are most frequently presented through the statistics-free ”average-o-grams” similar to those presented in this work (Figs. 1, 2). We then went on to perform a quantitative test of the associations between these features with compartment and TAD boundaries. In brief, we compare the values for each normalised signal directly over a boundary bin with those peripheral bins in each ≈ 1 Mb window (for TADs). We then perform a non-parametric, rank-based significance test to look for statistically significant enriched or depletions of each feature. Details of this procedure are given in Methods **??**.

Results of this significance testing, along with selected boundary profiles, are presented in Figure 4. Overall we find that compartment and TAD boundaries are associated with overlapping spectra of highly significant enrichments and depletions of chromatin features across cell types. Across all tests (Fig. 4), we frequently find boundary enrichments for DNA binding proteins implicated in chromosome architecture (e.g. CTCF, YY1, RAD21), but also note broad enrichments for classes of input features associated with active transcription (e.g. POL2, TBP, H3K9ac).

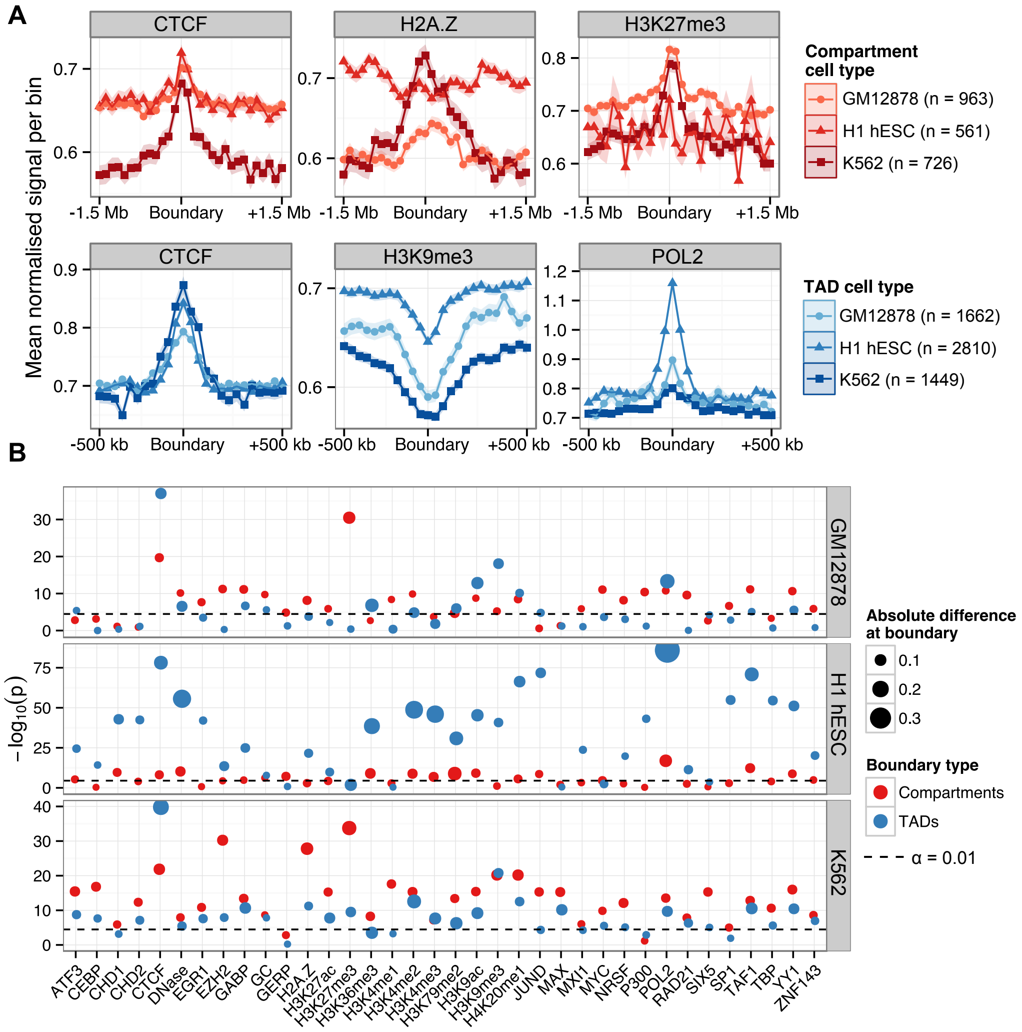

Figure 4: **Compartment and TAD boundary enrichment summary in three human cell types.** (A) Selected profiles for locus-level features are shown for compartment boundaries (CTCF, H2A.Z and H3K27me3) and TAD boundaries (CTCF, H3K9me3 and POL2), as mean normalized ChIP-seq signal relative to input chromatin per bin (±1 standard error). TAD boundaries were examined over 40 kb bins over the 1 Mb flanking each boundary; compartment boundaries were examined in 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.

Reflecting their different scales, we find enrichment and depletion profiles typically spanning regions of up to 500 kb for TAD boundaries but those over compartment boundaries often span more than a megabase (Fig. 4). We expect part of the reason for this discrepancy is the resolutions at which domains were called: TADS are resolved to 40 kb bins while compartment boundaries fall between megabase-sized bins. A further consideration is that larger numbers of TAD boundaries were called in the H1 hESC cell line, due its more deeply-sequenced Hi-C library (Section **??**), giving greater statistical power to detect enrichments thus resulting in smaller *p*-values (Fig. 4). The lower resolution megabase bins used in compartment calling do not suffer from this issue.

1.2.4 CTCF and YY1

Significant boundary enrichments for both CTCF and YY1 are evident in all cell types at both compartment and TAD boundaries (Fig. 4), which is intriguing given the evidence that YY1 and CTCF cooperate to affect long distance interactions.[19] Co-binding of CTCF with YY1 has also been shown to identify a subset of highly conserved CTCF sites.[20] This colocation may also therefore be a contributing factor in the establishment of TAD boundaries, which appear to be broadly conserved across mammals.[2]

To test this, we split our sets of TAD boundaries into those possessing ChiP-seq peaks (region peaks as called by the ENCODE data processing pipeline[18]) 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 (Fig. 5). Unexpectedly, we found that those boundaries marked by YY1 but without overlapping CTCF peaks were generally most strongly-enriched for other features in our dataset. This result potentially highlights YY1 as an under-appreciated contributor to boundary demarcation, particularly relative to the well-studied CTCF. We also found that boundaries lacking both CTCF and YY1 peaks showed instead the strongest enrichments for RAD21 in each cell type (Fig. 5), reinforcing previous findings that describe the distinct influences of CTCF and cohesin in organising chromatin structure.[17,21,22]

1.2.5 Repeats

Dixon *et al.*[2] identified short interspersed element (SINE) repeats as being enriched over TAD boundaries and suggested roles for these repeats in altering genome organisation, in line with prior evidence.[2,23] For example, SINE elements are thought to be responsible for spreading CTCF binding sites through mammalian genomes during evolution.[24] Analysis of recent high-resolution Hi-C data again reported a SINE B2

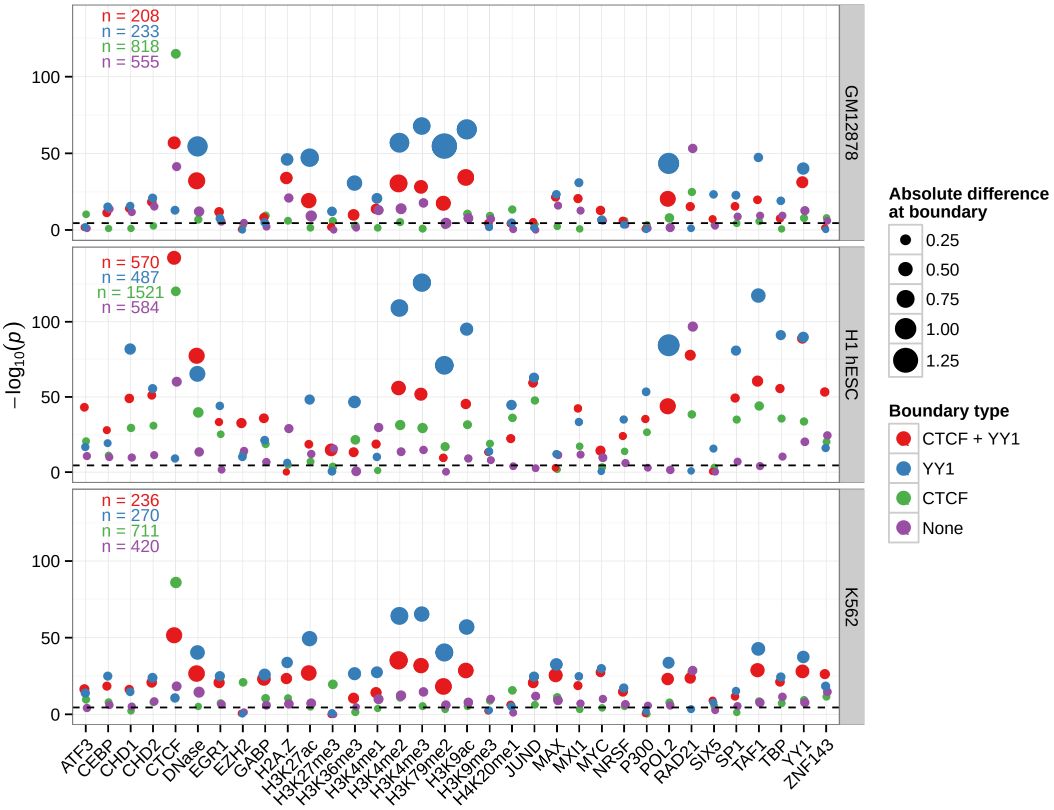

Figure 5: Distinct enrichments of CTCF and YY1 boundaries. The significance of TAD boundary enrichments and depletions are shown (as in Fig. 4) for boundaries split into classes based on the presence or absence of ChIP-seq peaks within boundary bins. CTCF and YY1 groups are those boundaries with at least one ENCODE region peak[18] 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 had neither a CTCF or YY1 region peak called (but can still be enriched for their respective features in terms of raw signal).

link with CTCF loops in mice,[3] and examples have been reported of human genes whose expression has been altered by CTCF sites inserted through Alu repeats.[25] Together these results suggest repeats could be a key component in the makeup of domain boundaries.

To investigate this, we used the `RepeatMasker`[26] software package to call repeat classes and families in the `hg19` and `mm10` genome assemblies. Counts for each annotated feature were then averaged over boundaries as described previously (Methods **??**).

At the level of repeat class, we corroborate the findings of Dixon *et al.*[2] that the majority of repeat classes show no enrichment or depletion at TAD boundaries, and we find that this also holds for compartment boundaries (Fig. 6). 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. 7).

We also find long interspersed elements (LINEs) are significantly depleted over TAD boundaries in two cell types, and borderline significant in the third, though with modest effect sizes in each case (Figs. 6, 7). DNA repeats appear to be enriched at both boundary types (Fig. 6), however these observations do not surpass our significance threshold ($\alpha = 0.05$) after multiple testing correction (Fig. 7).

Repeat classes can be broken into smaller repeat families. Dixon *et al.*[2] reported that the Alu repeat family of the SINE repeat class (or SINE B2 in mouse) is enriched over TAD boundaries. Again we can reproduce this finding and extend the analysis to compartment boundaries, where we do not see a significant enrichment for Alu repeat elements (Fig. 8). Surprisingly almost all other repeat families show no significant departure from their expected levels over TAD or compartment boundaries.

1.3 TAD BOUNDARY PREDICTION

We have shown TAD and compartment domain boundaries to be reliably and significantly marked by a variety of features. Compartment boundaries are successfully predicted as a side-effect of modelling the continuous compartment profile eigenvector (Section **??**), reflecting global patterns of transcriptional activation and repression, however a related measure of activity and repression, or other comparable profile we might want to predict, does not exist for TADs. Instead then, in this section we attempt to predict TAD boundaries within a class-balanced classification framework.

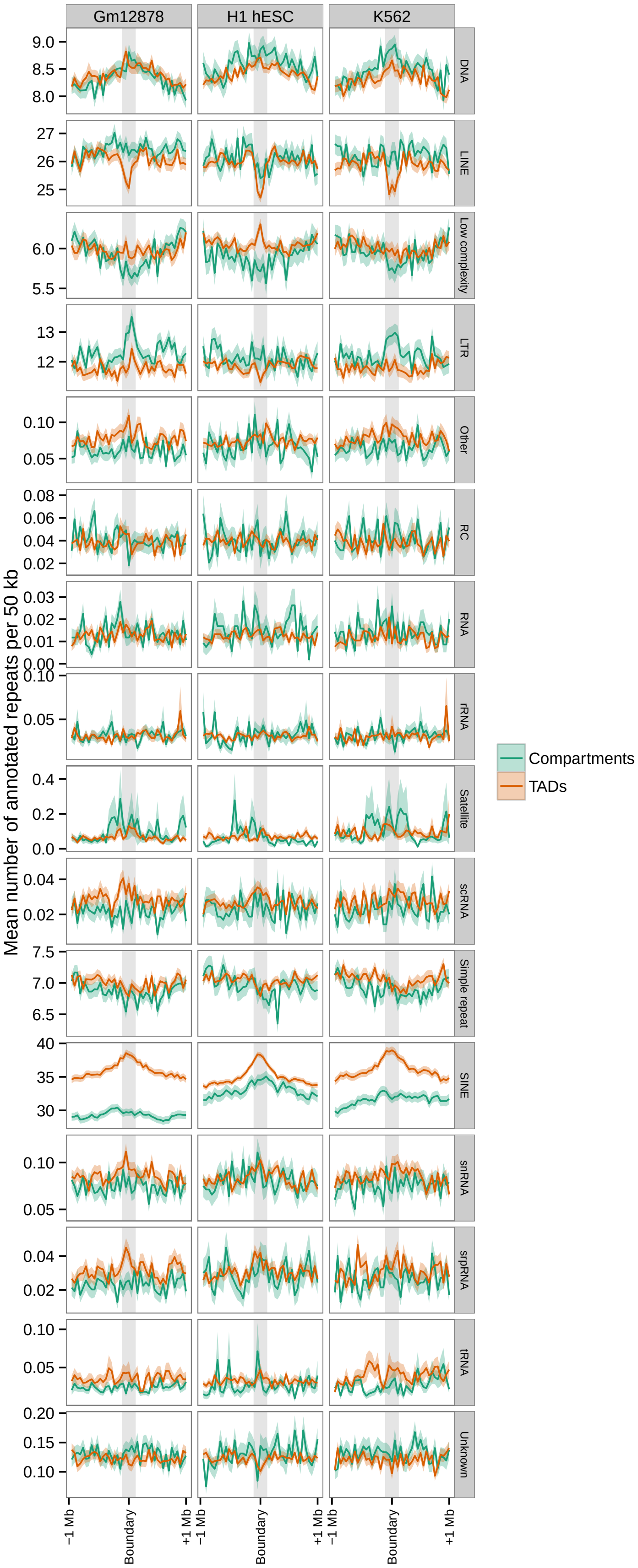

Figure 6: Repeat class average-o-grams over all TAD and compartment boundaries. RepeatMasker repeat annotations are counted per 50 kb for 1 Mb either side of all TAD and compartment boundaries. The mean counts per bin genome-wide are plotted with ±95% confidence intervals.

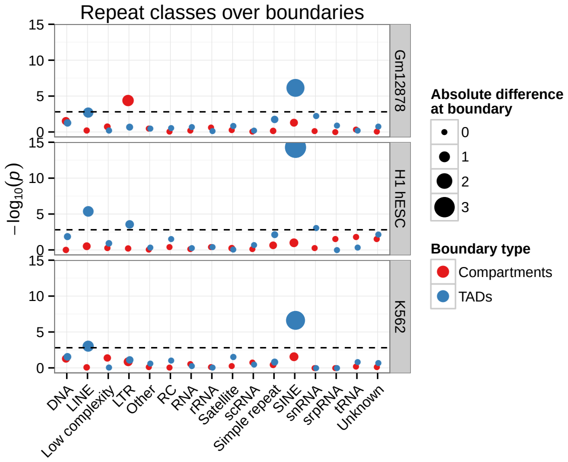

Figure 7: Significance and effect sizes of repeat class enrichments and depletions over boundaries. Boundary profiles (as shown in Figure 6) were each tested for enrichment or depletion of each repeat class at the boundary bin relative to peripheral non-boundary bins (see Methods **??**). Bubble area is proportional to the raw effect size of an enrichment or depletion. The Bonferroni-corrected significance threshold is highlighted with a dashed line.

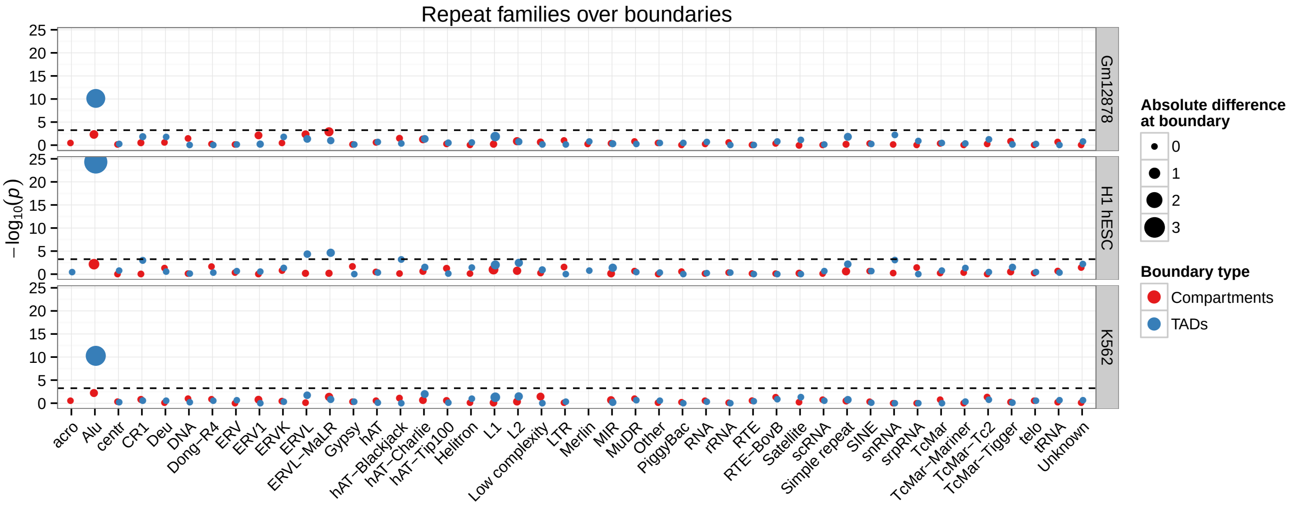

Figure 8: Significance and effect sizes of repeat family enrichments and depletions over boundaries. As per Figure 7 but for a more specific repeat classification.

1.3.1 Learning boundary classification

Results presented in this thesis describe the spectra of chromatin marks over TAD boundaries (Figs. 1, 4), thus it is of interested to test if we can build a predictive model (in the manner of Chapter **??**) that can call boundary regions from these marks alone. Such a model, if successful, could have broad utility in domain prediction in metazoan organisms where Hi-C data is not available.

A straightforward approach to this modelling task is to build a supervised classifier that learns the associations between two classes of genomic region: those labelled TAD boundaries and those which are not. To this end, we again turn to a Random Forest (RF) model, due to its many attractive properties discussed previously (Methods **??**). Our input feature set is made up of the same 35 matched features used in models of compartment eigenvector (Section **??**), with the addition of Alu repeat element counts (Section 1.2.5). Domain calls are those produced by the Dixon *et al.*[2] TAD calling algorithm (Methods **??**), therefore TAD boundaries were resolved to 40 kb bins. We class TAD boundary bins as boundary true positives (TP), and select matched bins 500 kb upstream as boundary true negatives (TN) for our training set.

To build parsimonious and accurate models (as discussed in Section **??**), we used the `AUC-RF` algorithm.[27] This is a form of stepwise model selection which optimises feature subset selection relative to the area under the receiver operating characteristic (AUROC), a metric which captures both the specificity and sensitivity of a classifier. The AUC-RF algorithm was applied to a training set of 80% of boundaries per cell type, with predictions assessed on out-of-bag (OOB) data as each forest was constructed. Selected models were then applied to the remaining held-out test set of 20% of TAD boundaries, with their matched non-boundary bins (full details are given in Methods **??**).

Predictive performance of these models is shown as ROC plots (Fig. 10) and in each case an AUROC of around 0.67–0.71 was achieved. In practice, this means that each classifier has around a 70% probability of ranking a random boundary region more highly than a random non-boundary region.[28] According to a commonly-used AUROC rule of thumb, this performance falls between the ranges of ”poor” to ”moderate” classification accuracy.[29]

Despite this sub-optimal classification accuracy, it is still of interest to analyse the variable importances in each cell type model. Strikingly we find that CTCF stands out as the most informative variable in each classifier by some margin (Fig. 11). This is in agreement with our results that CTCF showed among the highest and most consistent enrichments at TAD boundaries (Fig. 4). RAD21 is also highly ranked in each case and this ties in with our previous results suggesting orthogonal boundary enrichments for either CTCF or RAD21 (Section 1.2.4). Surprisingly YY1 does not feature as highly ranked in any model despite our observed consistent enrichments at TAD boundaries (Fig. 4). In fact YY1 was pruned from the optimal H1 hESC and GM12878 models by

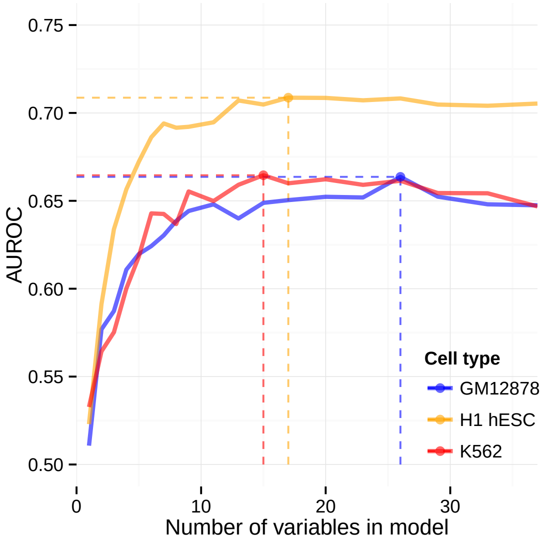

Figure 9: The AUC-RF stepwise algorithm finds that variable subset which maximises the AUROC on out-of-bag data. The AUROC is calculated at each stage of a stepwise method for selecting the best subset model from a full featureset. Here AUC-RF selected the following combinations: GM12878: 26 vars, 0.66 AUC; H1 hESC: 17 vars, 0.71 AUC; K562: 15 vars, 0.66 AUC. The AUC-RF algorithm is described in Methods **??**.

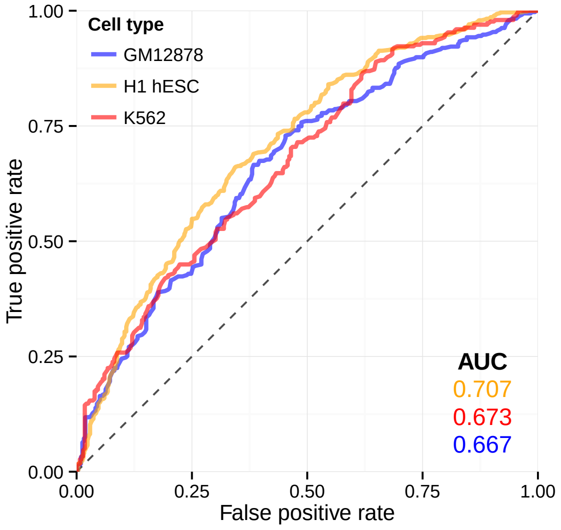

Figure 10: Receiver operating characteristics for TAD boundary classifications. ROCs are shown for three cell type specific classifiers of TAD boundary bins. The area under each cure is also shown (*inset*).

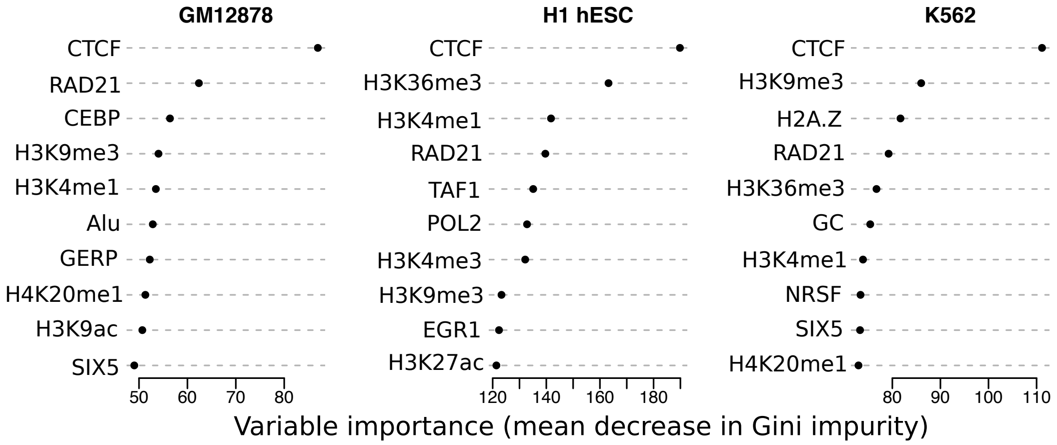

Figure 11: Variable importance for boundary classification Random Forest models. The ten most important variables in each boundary classification model are shown and ranked by their variable importance as defined by mean decrease in Gini impurity (Methods **??**).

the AUC-RF algorithm due to its low variable importance (*data not shown*). This could be due to a redundancy between information provided by YY1 and the CTCF variable as, for example, co-binding of YY1 and CTCF is thought to occur at sites of long-range chromatin interactions.[19]

1.4 METATAD BOUNDARIES

Our collaborators in the Pombo lab (Max Delbrück Center, Berlin) proposed the concept of ”metaTADs”: sequential aggregations of adjacent and strongly-interacting TADs that form a hierarchy of domain organisation covering each chromosome.

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

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 testable hypothesis, for example, is that boundaries of larger metaTAD structures could display greater enrichments for boundary-defining features.

1.4.1 MetaTAD boundary comparison

Due to the manner in which metaTADs were constructed by our collaborators, by sequential aggregation of TADs (Methods **??**), boundary comparisons between TADs and metaTADs are not completely straightforward. Every metaTAD boundary is, at a

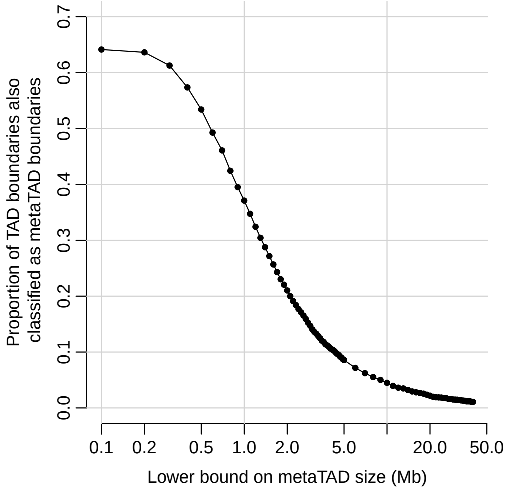

Figure 12: Calibration of metaTAD size selection bounds. As the lower bound on metaTAD size is increased, the proportion of all TAD boundaries which are also metaTAD boundaries decreases.

lower level, also a TAD boundary hence for a meaningful comparison we compared only a subset of metaTAD boundaries. Selection of this threshold involved a trade-off between maintaining a sufficient sample size of metaTADs and minimising the overlap between metaTAD and TAD boundaries to increase the discriminative power of our comparison (Fig. 12).

From the calibration plot (Fig. 12), a lower bound metaTAD size cut-off of 10 Mb was selected for comparison with TAD boundaries. This left a reasonable sample size of 263 metaTAD boundaries, while reducing the overlap with the set of all TAD boundaries to approximately 5% (Fig. 12). We also used an upper bound for size selection, based on observations by our collaborators that interactions between metaTADs larger than around 40 Mb were no higher than expected background signal (*data not shown*). In practice, almost all boundaries making up metaTADs larger than 10 Mb are also present in those larger than 40 Mb, but as hierarchical clustering continued up to the whole chromosome level, this upper bound may exclude a small number of edge-case peripheral TADs which aggregated into chromosome-wide metaTADs without evidence of heightened intra-TAD interactions.

Next a comparison between metaTAD boundaries for metaTAD size (s) in the range 10 Mb $< s <$ 40 Mb was performed. Our collaborators generated several ChIP-seq datasets, including for CTCF and three PolII variants, as well as expression data in the form of CAGE (Methods **??**). We calculated the average profiles of each of these features over regions surrounding the set of all metaTAD and TAD boundaries (Fig. 13). These average profiles show heightened enrichment for PolII variants, CTCF and DNase, with non-overlapping 95% confidence intervals of the mean over the boundary bin. Profiles also suggest increased enrichments of gene density and the

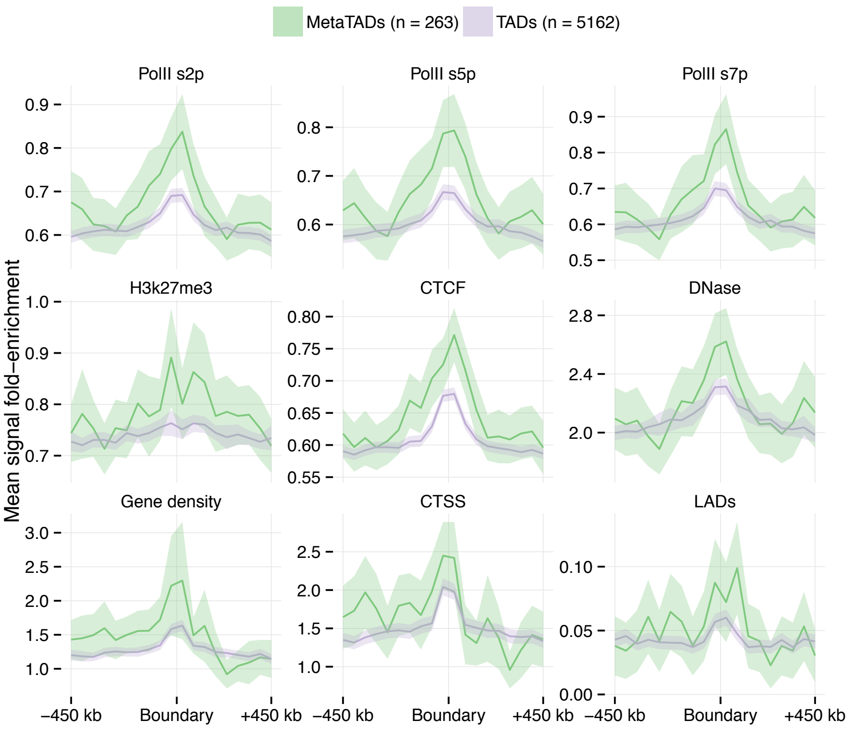

Figure 13: **Large metaTADs show greater enrichments than TADs for an array of boundary features.** Genome-wide profiles of epigenomic features and gene densities averaged over all TAD and metaTAD (10 – 40 Mb) boundaries (ribbons show 95% confidence intervals of the mean).

histone modification H3K27me3 at metaTAD boundaries relative to TAD boundaries (Fig. 13). The co-incidence of metaTAD boundaries and lamina associated domains (LADs) is explored further in Section 1.4.2.

Increased enrichment at metaTAD boundaries relative to TAD boundaries lends evidence to the functional importance of metaTADs, and suggests boundaries become increasingly well-demarcated at higher levels of organisation. However, if this is a genuine biological phenomenon, we may expect the trend not just to be observable in a comparison between two selected sets, but to increase monotonically as we ascended the metaTAD hierarchy from TADs to chromosomes.

To test this, we reran the metaTAD boundary analysis (Fig. 13) but at a range of metaTAD size cut-offs. Results of this analysis are shown in Figure 14. Generally we find increasing enrichments in metaTAD boundaries relative to TAD bounds through the range of lower bound cutoffs from 0 to 20 Mb, possibly with a slight decreased effect size at the highest cutoff of 30 Mb, where the sample size of boundaries decreases to just 62 (Fig. 14). This analysis strengthens the evidence for heightened functional enrichment of metaTAD boundaries (Figs. 13, 14) and suggests the metaTAD aggregation procedure is capturing boundaries of increasing strength, in terms of enrichment

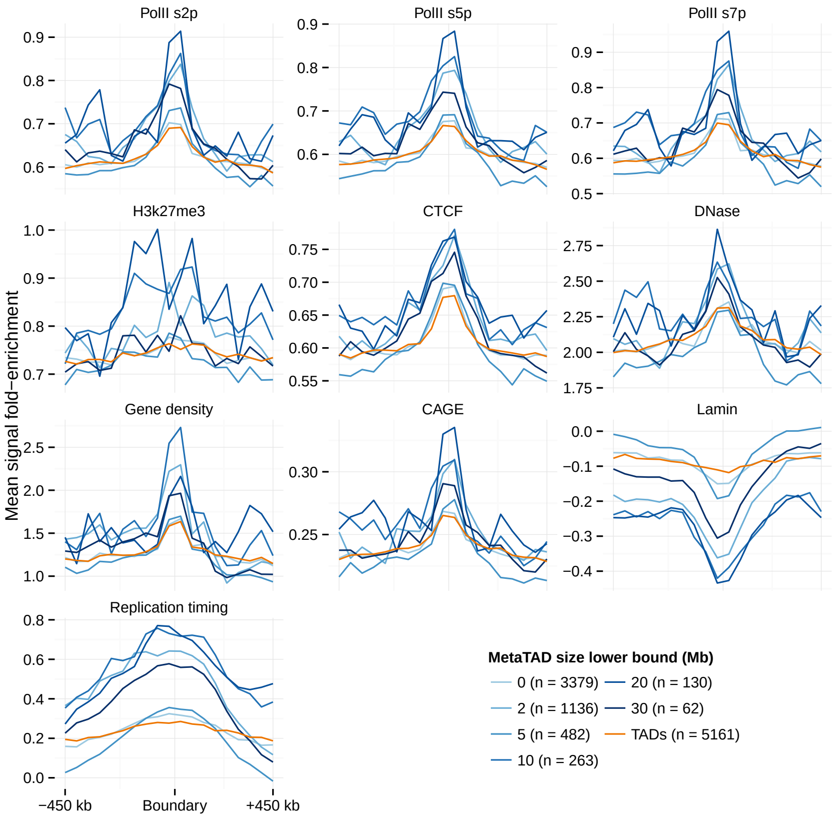

Figure 14: MetaTAD boundary enrichments and depletions with varying size selection. Boundary average-o-grams are shown for TAD and metaTAD boundaries with variable lower bound cut-offs. Sample sizes for each threshold are shown in the legend (*inset*).

of boundary associated features, through the metaTAD hierarchy. Though the nature and significance of these boundary enrichments are currently open to debate, there exist precedents where greater enrichments over boundaries have been invoked as evidence that a novel TAD calling algorithm outperforms previous efforts.[6,30]

1.4.2 Lamina associated domains

In the previous section we report a colocation of metaTAD boundaries and lamina associated domains (LADs), and at a greater level than that observed with smaller TADs (Fig. 13). This hints at an association between metaTADs and LADs which merits further investigation.

High resolution LAD data in mouse embryonic stem cells were retrieved from Peric-Hupkes *et al.*[31] in the form of continuous measures of lamin-B1 association produced by the DamID technique, known to reflect proximity to the nuclear lamin.[32] This measure of lamina association was then processed in windows around each metaTAD and TAD boundary, and profiles were combined to form a heatmap (Fig. 15).

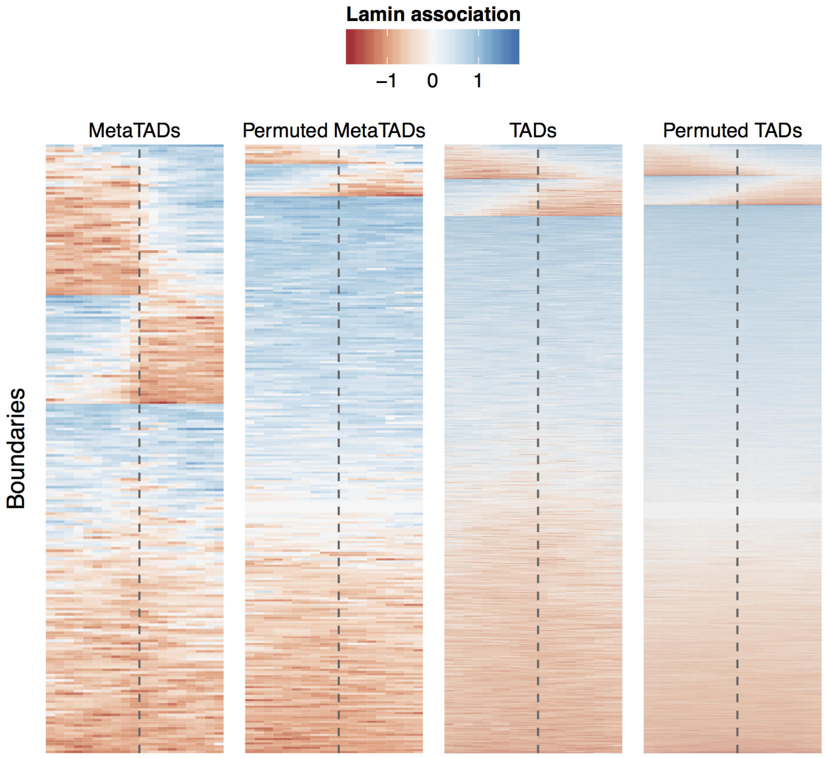

Figure 15: MetaTAD boundaries align with those of lamina associated domains. Heatmaps of LaminB1 association microarray probe intensity values over MetaTAD boundaries (from domains of size 10 – 40 Mb) and TAD boundaries, are displayed beside examples of circularly-permuted boundaries (Methods **??**). Profiles are shown ±450 kb from each boundary.

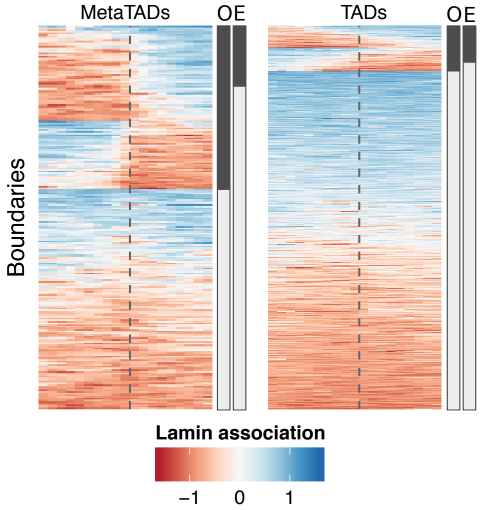

Figure 16: **MetaTAD boundaries co-occur with LAD boundaries more often than is expected by chance.** Heatmaps of lamina association over MetaTAD boundaries are shown as in Figure 15. Sidebars labelled O and E reflect observed and expected proportions of metaTAD / LAD boundary overlaps (Methods **??**).

Boundaries were seriated in order to separate out those that coincide with a LAD boundary, indicated by a transition in lamina association values. This was achieved by fitting a linear regression model across the lamina association profile of each boundary, then using a coefficient cutoff heuristic to select those that represent a boundary transition (Methods **??**). Using this approach, we found a markedly large proportion of metaTAD boundaries (43%) co-occur at a LAD boundary (Fig. 15). The same comparison using TAD boundaries found a coincidence of just 12%. However LADs are large domains and there were over 5,000 TADs called using these Hi-C data by our collaborators, thus in absolute numbers of boundaries this is still represents a large overlap.

To test the significance of these observations, we apply a permutation-based statistical test where our observed coincidences are compared with those produced by 1,000 circular (per-chromosome) permutations (Methods **??**). We found that the metaTAD and LAD boundary coincidence is around a 2.7-fold increase above null expectation (observed: 42.6%; expected: 15.8%; empirical *p*-value: $p < 1 \times 10^{-4}$; Fig. 16). Meanwhile TAD boundaries were found to have a smaller, yet still significant, 1.2-fold increase in coincidence with LAD boundaries relative to a null model (observed: 11.8%; expected: 9.5%; empirical *p*-value: $p < 1 \times 10^{-4}$; Fig. 16).

As with other enrichments (e.g. Fig. 14), we went on to verify that this result was not specific to the choice of boundary size cutoff. Recall that as metaTADs are derived from TADs, selecting metaTADs within a size range is a trade-off between minimising overlap between boundaries assigned to TADs and metaTADs to enable more powerful comparison, while retaining a sufficiently large sample size of metaTAD boundaries

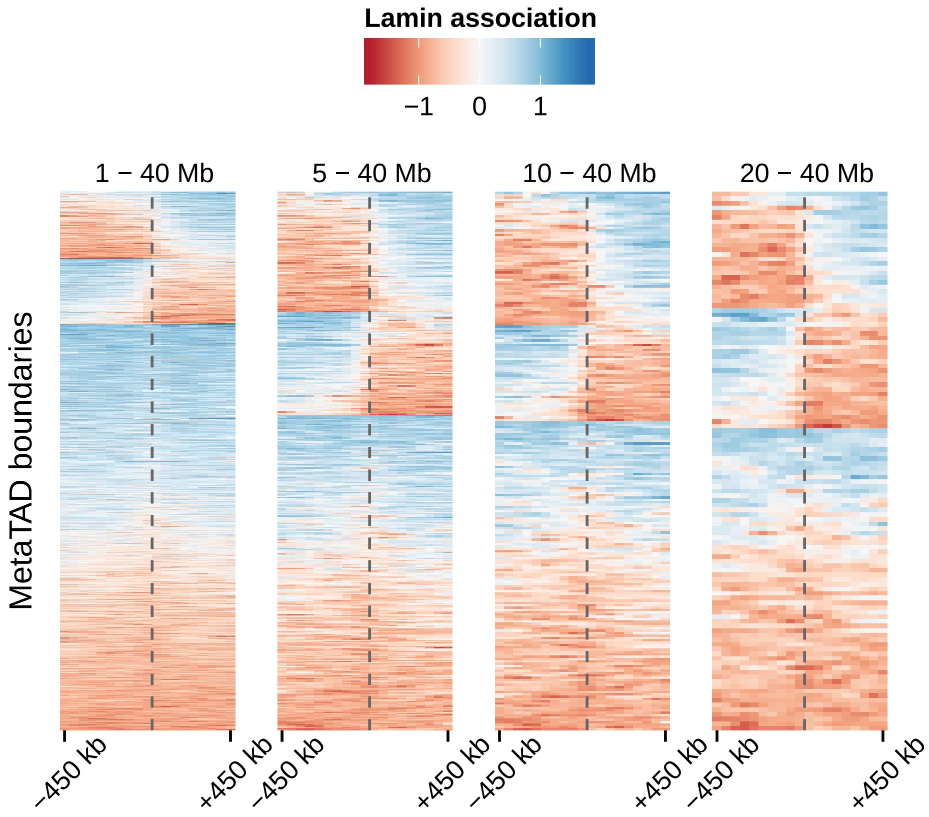

Figure 17: **MetaTAD–LAD coincidence increases at higher levels of the metaTAD hierarchy.** Heatmaps of lamina association over MetaTAD boundaries are shown as in Figure 15. Panels show the size selection cutoffs for metaTAD domains considered in each case.

(Section 1.4.1). In the case of LAD–metaTAD coincidence, we again find this result is insensitive to the selection of domain size cutoff, and indeed that there is evidence for an increasing proportion of co-occurrence as we ascend the metaTAD hierarchy (Fig. 17).

This result suggests again that metaTADs seems to offer useful perspectives onto higher order genome organisation. In this case, it appears TADs will often neatly aggregate within LADs and together these constitute what we observe as a metaTAD.

1.4.3 Boundaries over a time series

For the first time, our collaborator's applied the Hi-C technique over a differentiation time course from mouse embryonic stem cells, to neural progenitors and finally fully-differentiated neuron cells. Successive expression measures were also taken alongside this Hi-C data in the form of CAGE data, produced by the FANTOM5 consortium.[33]

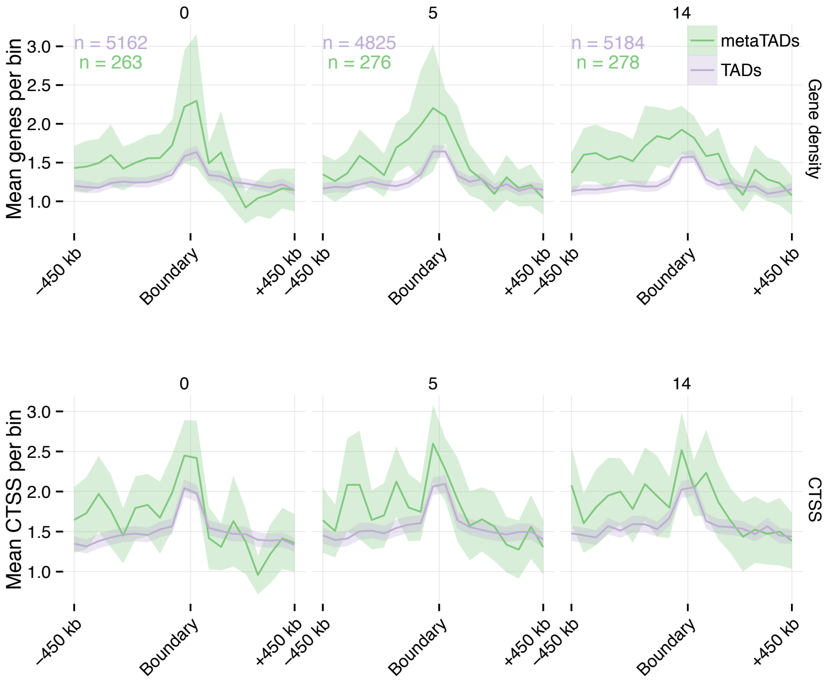

Figure 18: Observed enrichments persist over a time series. CAGE-defined active TSS (CTSS) were counted per 50 kb bin across each TAD and MetaTAD (10 – 40 Mb) boundary and averaged (ribbons show 95% confidence intervals of the mean). Gene densities refer to mean counts of annotated genes per bin, with an overlap of at least 250 bp (Methods **??**).

Together these datasets offer a unique perspective onto how higher order genome organisation varies with expression during differentiation.

Collaborators explored changes in both the overall tree structure between timepoints, and aggregate expression changes between TADs and metaTADs at successive timepoints. They identified rewiring events in the metaTAD tree over the timecourse, and found corresponding changes in expression (*data not shown*). Given metaTAD structures appear to differ and matched CAGE data exists for each timepoint, it was of interest to test how observed boundary enrichments for gene expression (Figs. 13, 14) might vary over this timecourse.

We find actively-transcribed CAGE-defined TSS (CTSS) to be consistently enriched over the shifting boundaries through the differentiation timecourse (Fig. 18). We coupled this with a static measure of gene density in order to distinguish expression changes from genic overlap, however both series show similar patterns so it does not seem that boundary expression varies at a global scale over this timecourse. At each timepoint, peak heights over boundary bins suggest modestly stronger enrichments at metaTAD boundaries relative to TAD boundaries, as seen with other features (Fig. 13).

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 (e.g. Fig. **??**) are effectively abolished in metaphase.[34] This raises the question of how the structural organisation 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.[35] These G-band patterns are constant over human cell types at metaphase, but all traces of interphase higher order structure were reported to be absent at metaphase.[34] We would therefore not necessarily expect an agreement between G-bands, labelled in metaphase cells, and nuclear compartments, called from cell populations which were mostly in interphase.

We examined the genome wide concordance of interphase compartment boundaries with metaphase G-band boundaries, relative to an expected distribution derived by permutation (Methods **??**). We found a significant, though modest, excess of compartment boundaries within close proximity of G-band boundaries, such that 13.9% of compartment boundaries are within 500 kb of a G-band boundary (expectation = 10.5%, K-S test: $D = 0.076$, $p < 3 \times 10^{-12}$). This is seen for compartment boundaries calculated for all three cell types independently (a full comparison for GM12878 is shown in Figure 19).

The genome wide overlap of compartment A and B regions with particular G-band classes is nonrandom, and suggests a greater correspondence than that of simple boundary distances. 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 (Fig. 20). Approximately 40% of the genome jointly occupies interphase compartment A as well as the lightly-stained `gneg` or `gpos25` metaphase G-bands, or occupies the interphase B compartment in addition to the well-stained `gpos75` or `gpos100` bands at metaphase (Fig. 20). Similar trends are seen across all three cell types as expected given the high correlations seen in A/B compartment profiles between cell types (*data not shown*).

This agreement is not wholly unexpected given the known characteristics of G-negative and G-positive bands, with contrasting gene density, GC content and replication timing[35] that is strongly reminiscent of the contrasts between interphase A and B compartments.[1] Despite showing an association, these data agree with experimental evidence that many domain boundaries are not well preserved between interphase and metaphase. However there is evidence for relationships between

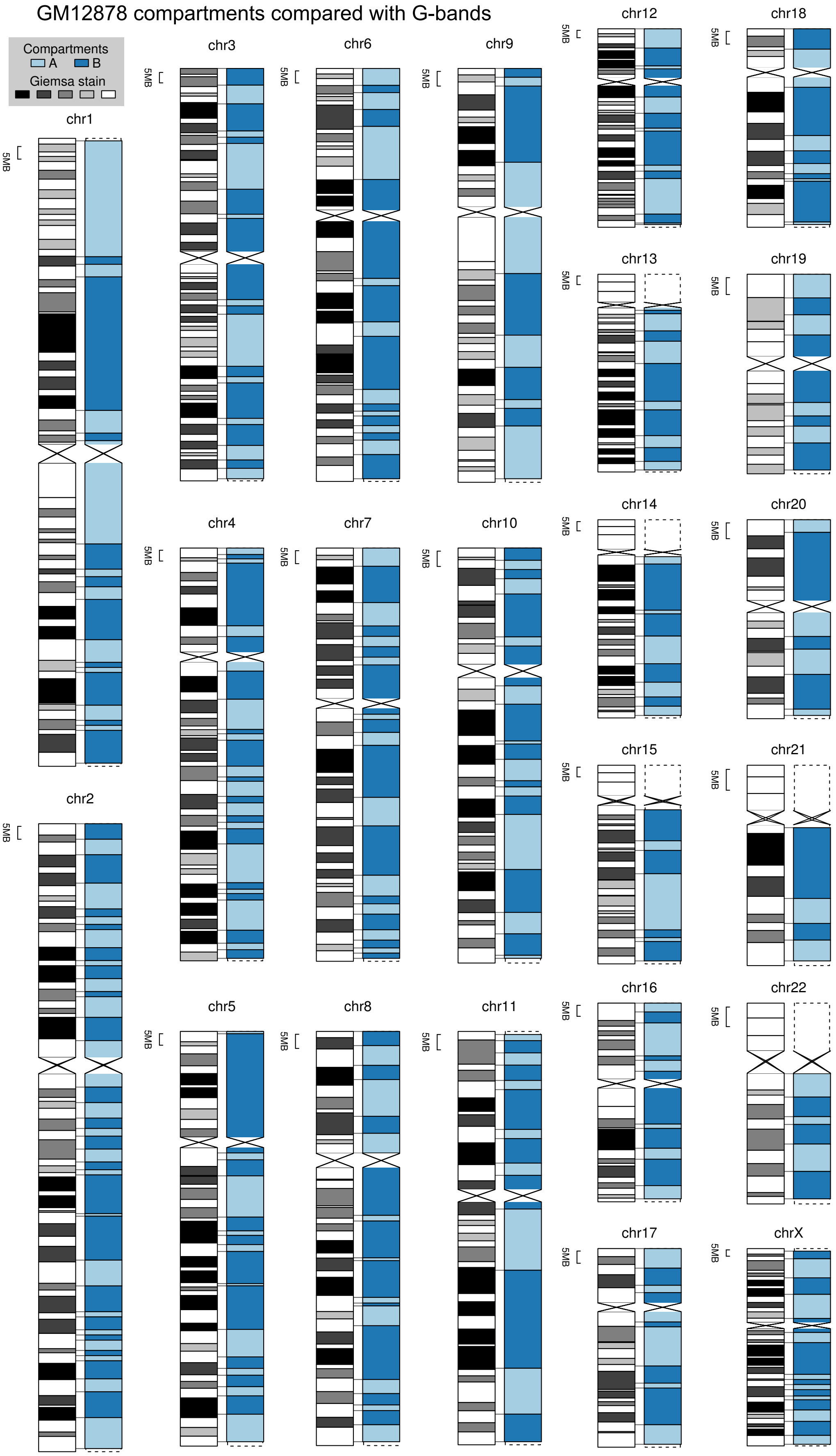

Figure 19: **Genome-wide agreement between Giemsa bands and A/B compartments.** G-bands, assayed at metaphase, often correspond with interphase A/B compartments across all chromosomes. Data for cell type GM12878 is shown.

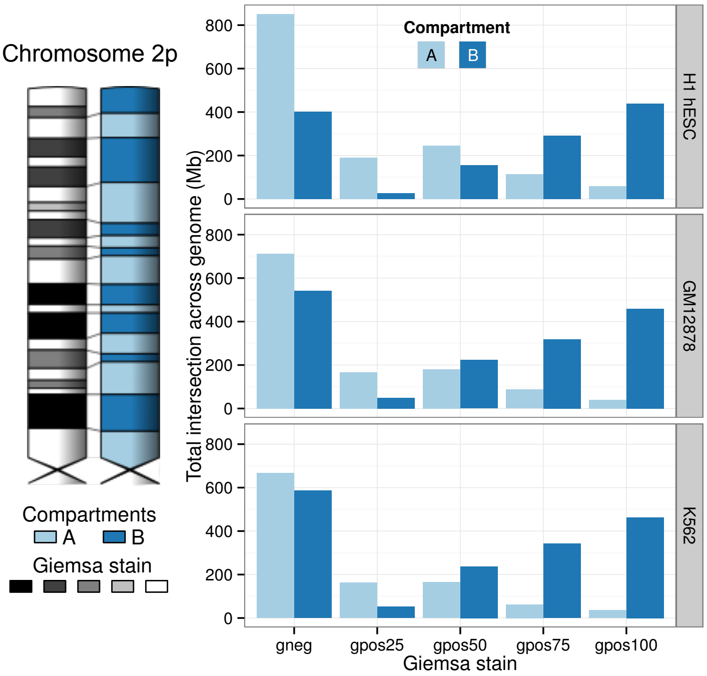

Figure 20: Giemsa–stain bands correspond to A/B compartments. The correspondence between G-bands and A/B compartments is broken down into the five levels of Giemsa stain. A compartments largely match *gneg* staining, while *gpos75* and *gpos100* are enriched in B compartments.

the broad structural categories of compartments and G-bands which may reflect similarities in the degree of compaction throughout the cell cycle.

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] 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.

[4] Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, **148**(3): 458–72.

[5] 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.

[6] Filippova D, Patro R, Duggal G, Kingsford C (2014) Identification of alternative topological domains in chromatin. *Algorithms for molecular biology : AMB*, **9**: 14.

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

[8] Le Dily F, Bau D, Pohl a, Vicent GP, Serra F, Soronellas D, Castellano G, Wright RHG, Ballare C, *et al.* (2014) Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes & Development*, **28**(19): 2151–2162.

[9] Nora EP, Dekker J, Heard E (2013) Segmental folding of chromosomes: A basis for structural and regulatory chromosomal neighborhoods? *BioEssays*, **35**(9): 818–828.

[10] 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.

[11] Lupiáñez D, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz J, *et al.* (2015) Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, pp. 1012–1025.

[12] Ren B, Dixon J (2015) A CRISPR Connection between Chromatin Topology and Genetic Disorders. *Cell*, **161**(5): 955–957.

[13] Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ (2015) Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*.

[14] Evertts AG, Manning AL, Wang X, Dyson NJ, Garcia Ba, Coller Ha (2013) H4K20 methylation regulates quiescence and chromatin compaction. *Molecular biology of the cell*, **24**(19): 3025–37.

[15] Onder TT, Kara N, Cherry A, Sinha AU, Zhu N, Bernt KM, Cahan P, Marcarci BO, Unternaehrer J, *et al.* (2012) Chromatin-modifying enzymes as modulators of reprogramming. *Nature*, **483**(7391): 598–602.

[16] Chai X, Nagarajan S, Kim K, Lee K, Choi JK (2013) Regulation of the boundaries of accessible chromatin. *PLoS genetics*, **9**(9): e1003778.

[17] 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.

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

[19] Atchison ML (2014) Function of YY1 in Long-Distance DNA Interactions. *Frontiers in immunology*, **5**(February): 45.

[20] Schwalie PC, Ward MC, Cain CE, Faure AJ, Gilad Y, Odom DT, Flicek P (2013) Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. *Genome biology*, **14**(12): R148.

[21] Seitan VC, Faure AJ, Zhan Y, McCord RP, Lajoie BR, Ing-Simmons E, Lenhard B, Giorgetti L, Heard E, *et al.* (2013) Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome research*, **23**(12): 2066–77.

[22] Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, Ong CT, Hookway Ta, Guo C, *et al.* (2013) Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. *Cell*, **153**(6): 1281–95.

[23] Lunyak VV, Prefontaine GG, Núñez E, Cramer T, Ju BG, Ohgi KA, Hutt K, Roy R, García-Díaz A, *et al.* (2007) Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science (New York, N.Y.)*, **317**(5835): 248–251.

[24] 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.

[25] Ruiz-Narváez Ea, Campos H (2008) Evolutionary rate heterogeneity of Alu repeats upstream of the APOA5 gene: Do they regulate APOA5 expression? *Journal of Human Genetics*, **53**(3): 247–253.

[26] Tarailo-Graovac M, Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, (SUPPL. 25).

[27] Calle ML, Urrea V, Boulesteix AL, Malats N (2011) AUC-RF: A new strategy for genomic profiling with random forest. *Human Heredity*, **72**(2): 121–132.

[28] Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognition Letters*, **27**(8): 861–874.

[29] Streiner DL, Cairney J (2007) What's under the ROC? An introduction to receiver operating characteristics curves. *Canadian Journal of Psychiatry*, **52**(2): 121–128.

[30] Weinreb C, Raphael BJ (2015) Identification of hierarchical chromatin domains. *Bioinformatics (Oxford, England)*, pp. 1–8.

[31] Peric-Hupkes D, Meuleman W, Pagie L, Bruggeman SWM, Solovei I, Brugman W, Gräf S, Flicek P, Kerkhoven RM, *et al.* (2010) Molecular Maps of the Reorganization of Genome-Nuclear Lamina Interactions during Differentiation. *Molecular Cell*, **38**(4): 603–613.

[32] Pickersgill H, Kalverda B, de Wit E, Talhout W, Fornerod M, van Steensel B (2006) Characterization of the Drosophila melanogaster genome at the nuclear lamina. *Nature genetics*, **38**(9): 1005–1014.

[33] Consortium TF, Pmi R, Dgt C (2014) A promoter-level mammalian expression atlas. *Nature*, **507**(7493): 462–70.

[34] 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.

[35] Furey TS (2003) Integration of the cytogenetic map with the draft human genome sequence. *Human Molecular Genetics*, **12**(9): 1037–1044.