Unravelling higher order genome organisation [working title]

Introduction

Benjamin L. Moore

July 20, 2015

igmm
INSTITUTE OF GENETICS
& MOLECULAR MEDICINE

1 REANALYSIS OF HI-C DATASETS

1.1 INTRODUCTION

Since the initial publication of the Hi-C technique in 2009,[1] there has been rapid advancement of both the technique itself and the resolution at which interaction frequencies have been analysed. From the proof-of-concept analysis at 1 megabase (Mb) and 100 killable (kb) resolution,[1] subsequent experiments achieved first 40 kb[2], then 10 kb[3] and most recently 1 kb[4], enabling bona fide fragment-level analysis for the first time.

Such rapid progression in the field has resulted in a wide variety of public Hi-C datasets being available, albeit with differing qualities. With proper correction and at a suitable resolution, these interaction frequencies can be compared and contrasted within and between species.

In this work I uniformly reprocessed publicly-available human Hi-C datasets, in order to address fundamental questions about the stability of higher order genome organisation within cell populations from the same species. Previously Hi-C studies have compared two samples per species, such as K562 against GM06990[1] or IMR90 against GM12878.[2] Here I make use of three Hi-C datasets corresponding to extensively-studied human cell lines: K562, GM12878 and H1 hESC. Together these make up the "Tier 1" cell lines studied by the ENCODE consortium,[5] hence have huge amounts of matched ChIP-seq and histone modification data available.

By combinatorial reanalysis of these cell-matched datasets, I can investigate the relationships between higher order chromatin structure and locus-level chromatin features.

1.2 HI-C REPROCESSING

Each Hi-C dataset used in this work was reprocessed using the same pipeline from raw sequencing reads. In each case, experiments used the same HindIII restriction enzyme.

1.3 COMPARTMENT PROFILES

After uniformly reprocessing each Hi-C dataset and calling compartment eigenvector profiles (see *Methods*), we can compare these between three human cell lines. Compartment profiles have a visibly high-correspondence (Fig. 1), despite the variable sources of both sample material and experimental data.

This close correspondence also validates our approach of combining these different datasets, and suggests our uniform pipeline is successfully accounting for differences in sequencing depth and other batch effects. The precise correlations of these independent measures are in the interval $R = [.75, .8]$ (Fig. **??**; Pearson correlation coefficients, PCC).

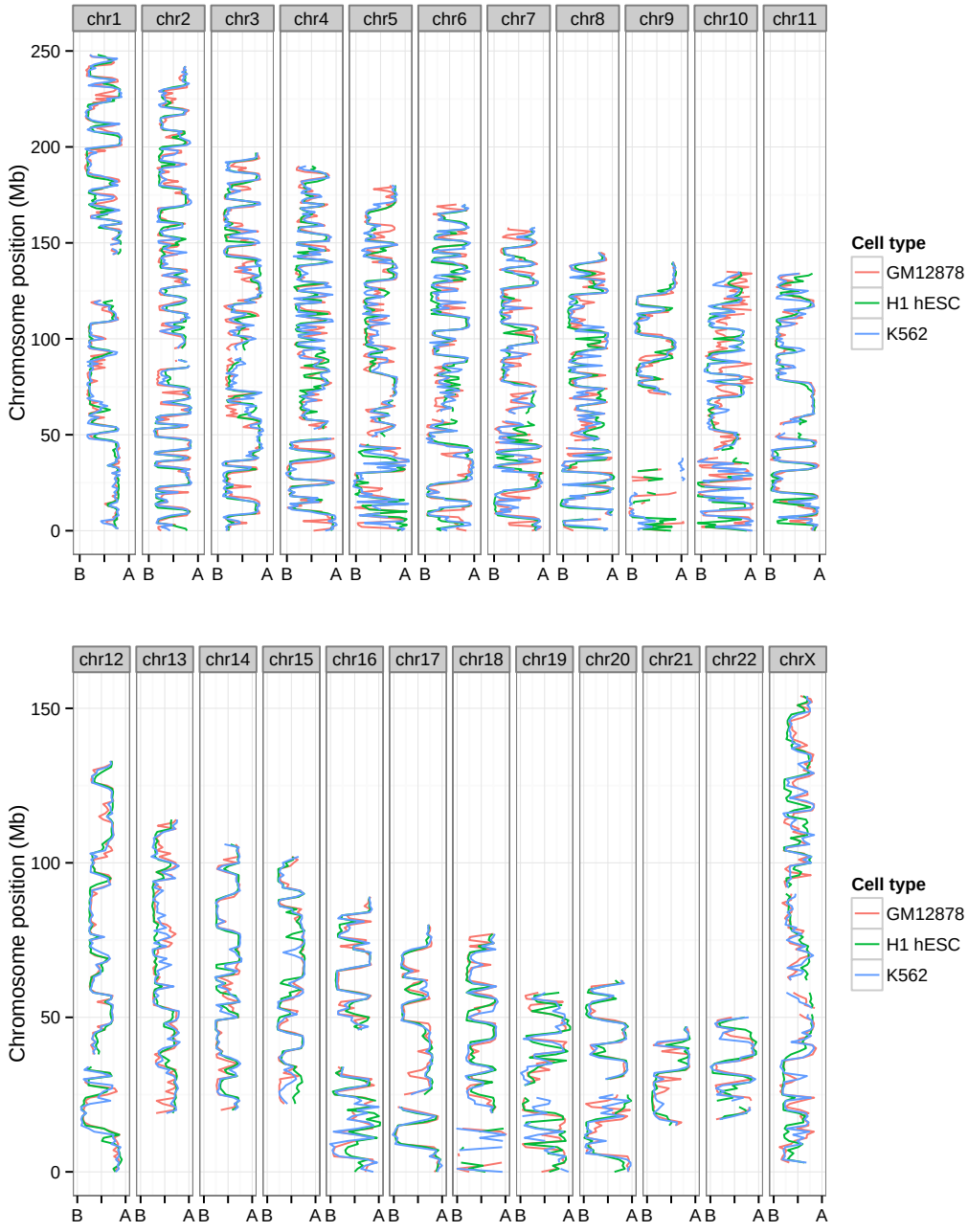

Figure 1: Compartment profiles are observably well-correlated between human cell types and across all chromosomes. Caption

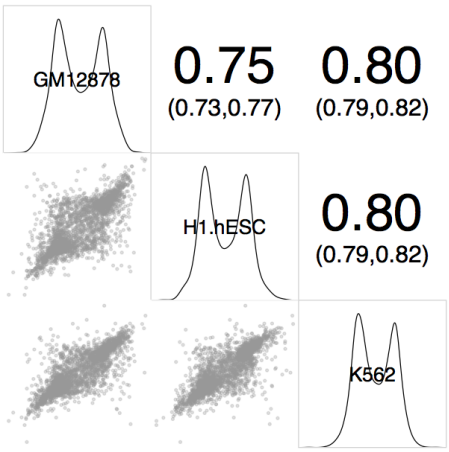

Figure 2: Compartment eigenvectors are well-correlated between human cell types Megabase resolution compartment eigenvector values are shown in a plot matrix. *Upper triangle*: Pearson correlation coefficients between pairs, with 95% confidence intervals; *diagonal* Kernel density estimates of eigenvector values per cell type; *lower triangle*: *x*-*y* scatterplot of values.

1.4 DOMAIN CALLS

1.4.1 Compartments

The continuous compartment eigenvector can be used as-is to classify A/B compartments, using positive and negative eigenvector values after first orientating the vector with respect to, for example, PolII Chip-seq data.[6] However, given the definition of compartments as generally broad and alternating domains along a chromosome, often matching other large domains of Lamin association, an improved classification method might penalise the calls of short compartment calls, which may be the result of noise.

For this reason, instead of using raw eigenvector values we consider observed values as emissions from unobserved underlying states. This can be modelled through a Hidden Markov Model (HMM), whereby we first parameterise models of state and their transitions, then infer the most likely state sequence to have emitted our observed data. This unobserved two-state sequence is then used for compartment calls (see Methods **??**).

In practice, this acts to de-noise our compartment calls. Where single sign-changes along the series would have resulted in a single-block compartment, these may now be modelled as noisy emissions from a single unobserved state. An examplar region is showing in Fig. 3. This shows an approximately 50 Mb region from chromosome 8 with eigenvector data from the H1 hESC cell line. A simple thresholding method in this region calls a total of 12 regions, whereas our HMM method finds only 6 larger regions in the same window.The disparity is caused by very short and single-bin compartments being disfavoured by the HMM-based method (Fig. 3).

1.4.2 TADs

1.5 DOMAIN EPIGENETICS

The use of well-studied human cell types allows intersection with publicly-available epigenomics datasets, such as those produce by the ENCODE consortium.[5] In total, 35 cell-matched ChIP-seq datasets were available for all three of the tier 1 ENCODE cell lines: GM12878, H1 hESC and K562 (see Methods XX). Domain calls from our reprocessed

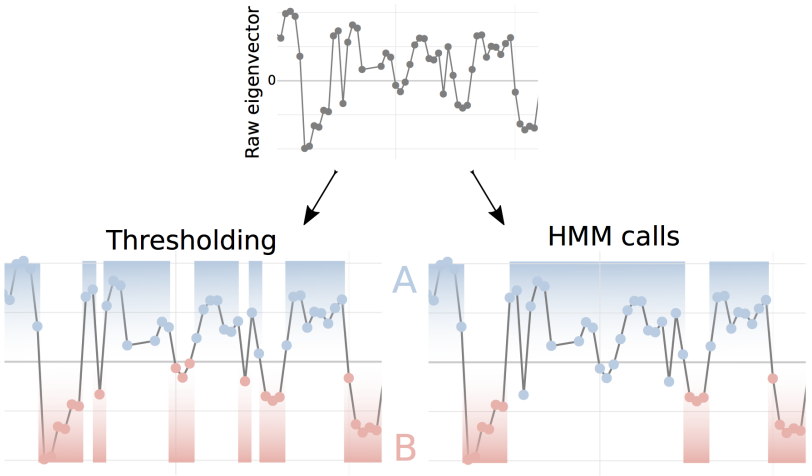

Figure 3: Compartment calls by simple thresholding method or context-aware HMMs. Chromosome compartments have previously been called through simple thresholding at 0,[1] in this work we also use an HMM-based method to call unobserved states that have emitted our noisy observed values (*right*).

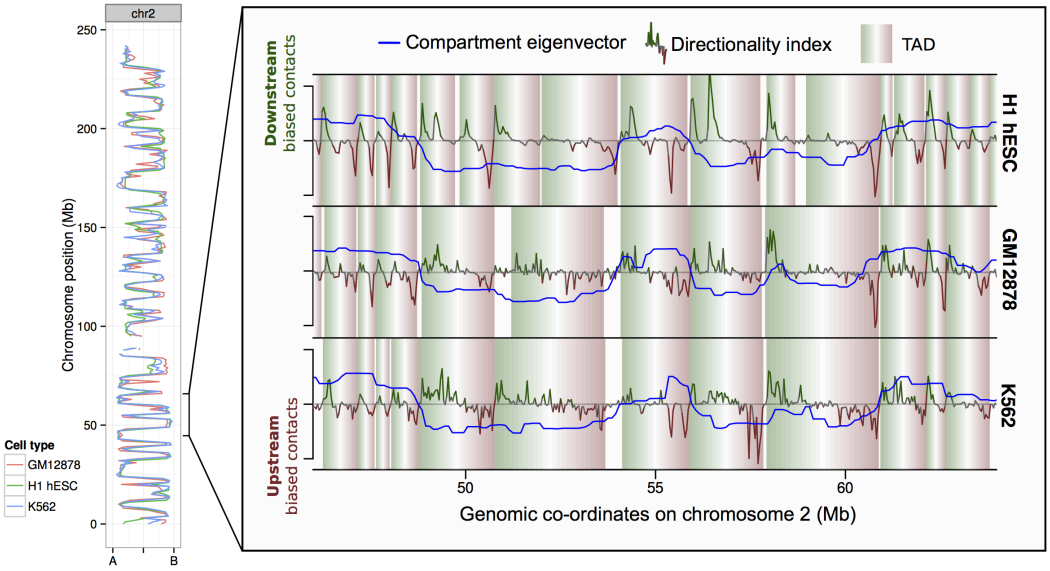

Figure 4: Regions of variable structure alter their long-range contacts. The eigenvector compartment profile is shown for chromosome 2 for three human cell types (*left*). At higher resolution, the zoomed region illustrates conservation of topological domains (TADs) over 20 Mb of the same chromosome.

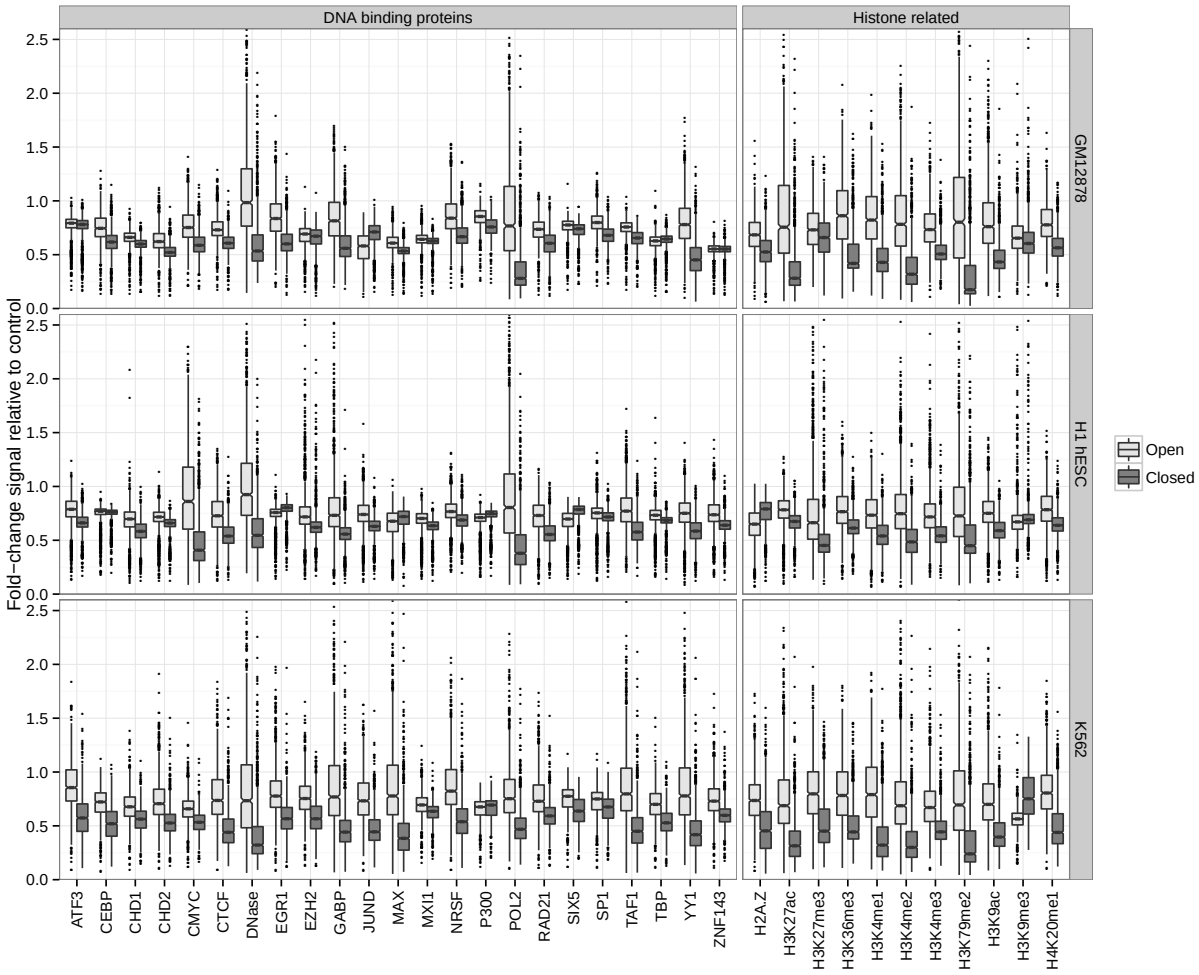

Figure 5: The chromatin signatures of A/B compartments. Notched boxplots summarise the distribution of each feature over 1 Mb bins in open (A) and closed (B) compartments genome-wide.

1.5.1 A/B compartments

The overwhelming majority of intersected chromatin features are significantly enriched in active A compartments relative to B compartments (Fig. 5). This is expected, A compartments represent the actively-transcribed and accessible portions of the genome, and have previously been shown to positively correlate with many of the features shown.[1,7]

Exceptions to this rule are few. However the repressive histone modification H3k9me3 is found more often in B compartments in two cell types, as is the P300 transcription factor (Fig. 5). Also of note is the histone variant H2A.Z which is significantly enriched in A compartments in GM12878 and K562, but this relationship is reversed in the embryonic stem cell line (Fig. 5). Recent evidence suggests specialised roles for H2A.Z in regulating both repression and activation during embryonic stem cell differentiation, acting as a "general facilitator".[8] Additionally H2A.Z has been reported to mark histone octamers for depletion, thereby permitting gene activation during differentiation.[9] Potentially, then, the H2A.Z enrichment in B compartments could be driven by regions soon to be de-repressed as the stem cell differentiates.

1.5.2 TAD classes

Unlike compartments, initially TAD were not observably correlated with blocks of chromatin features (e.g. 2). Later studies have linked TADs with such enrichments, first in *Drosophila*[10] and later in human cells, where it was argued TADs are merely a low-resolution window to smaller "sub-compartments", bearing similar active and inactive marks to their much-larger namesakes.[4]

Here we look for evidence of the Sexton *et al.*[10] characterisation of TADs called in our human cell types.

1.6 VARIABLE REGIONS

Despite the vast majority of the genome being in matched chromatin compartments, there are also regions of disagreement. Reasons for observable differences include technical errors and bias, but also more interesting functional explanations, where cell-type specific activation or repression is reflected in changes in higher order structure.

To conservatively call regions of variable structure (RVS), we used HMM-called compartment states and selected those which were either: i) open in one cell type and closed in both others or ii) closed in one cell type and open in both others. This left sets of RVS which could be considered as "flipped open" or "flipped closed" in a given cell type.

1.6.1 Chromatin state enrichment

1.6.2 Contact changes

A defining characteristic of active A compartment regions is a preferential bias in contacting other A compartment regions.[1] However, it is not clear whether cell-type-specific transitions in higher-order structure are solely compartment-level phenomena, or involve other structural strata. We therefore examined the genome-wide contact profiles of each region of variable cell-type-specific chromatin structure in detail. If these cell-type-specific structures are mediated by finer-scale structural levels (such as TADs) we might expect to see predominantly short-range contacts in their underlying contact profile. Instead, we found that variable regions preferentially interact with other A compartment regions in the cell types in which they are active (**??**), but not in the other cell types in which they are inactive. This supports the idea that these cell-type-specific regions are undergoing compartment-level transitions, disproportionately mediated by the formation of long-range contacts, while also not precluding additional changes at lower levels such as TADs.

1.7 NUCLEAR POSITIONING

Chromosome positioning within the nucleus is known to reflect gene density, with the most gene-dense chromosomes occupying the centre of the nucleus in human cells.[11] Kalhor *et al.*[6] used a Hi-C variant to recreate probability density maps of chromosome positions which again reflected this feature of higher order chromatin organisation, and also reported active regions were more diffuse than inactive. A testable expectation with the eigenvector data used in this work is that active A compartments are enriched in the central nucleus of our human cell types, and B compartments are preferentially located in the nuclear periphery.

To test this, published data on chromosome positioning preference within the nucleus was used to label chromosomes as "central" or or "edge".[12]

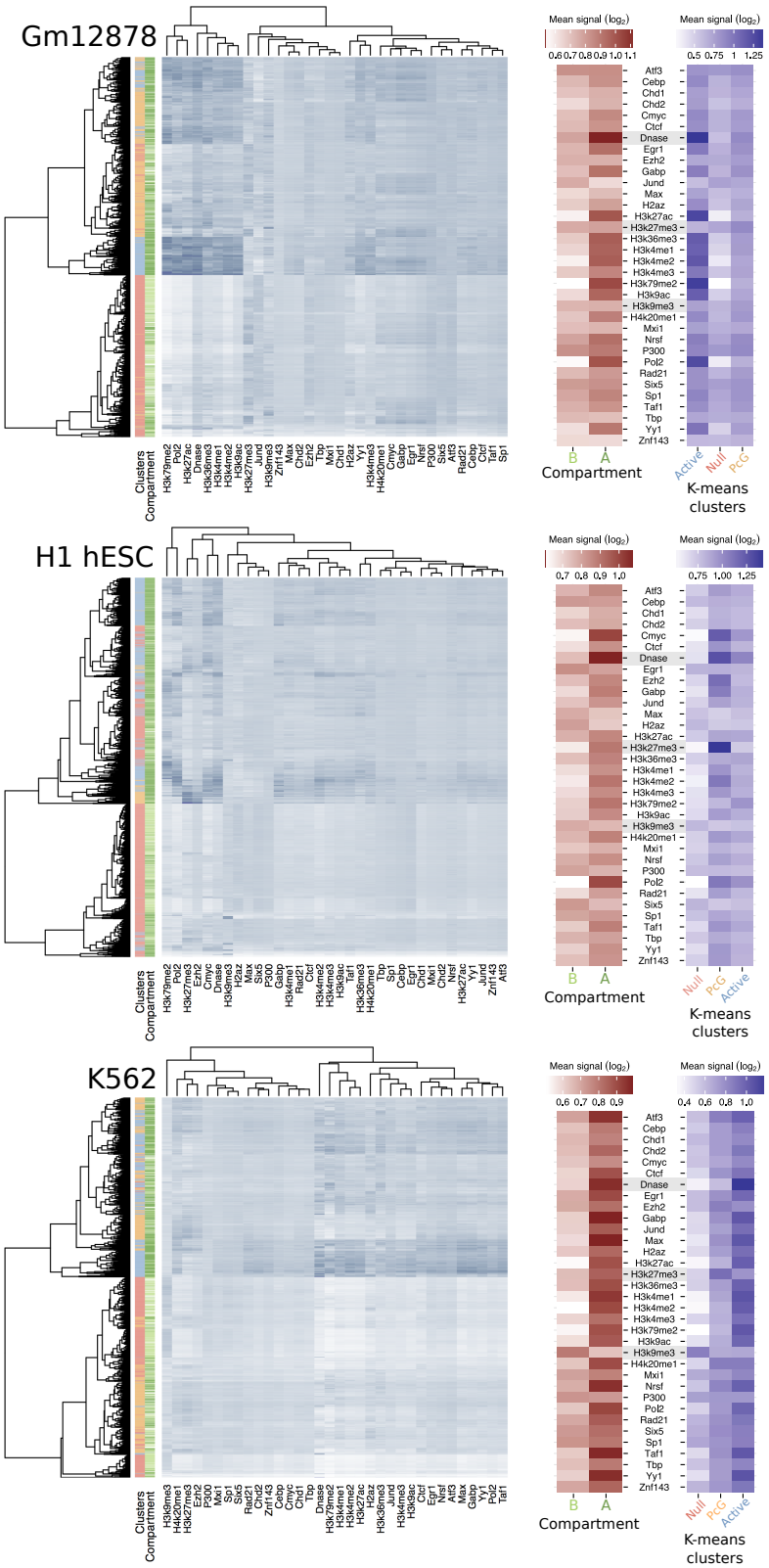

Figure 6: TADs reflect epigenetic domains. Following the Drosophila results of Sexton *et al.*[10], clustering of TAD domains by mean log2 signal of 34 ENCODE features distinguishes null, active and polycomb-associated (PcG) domains, as well as reflecting the encompassing A/B compartments.

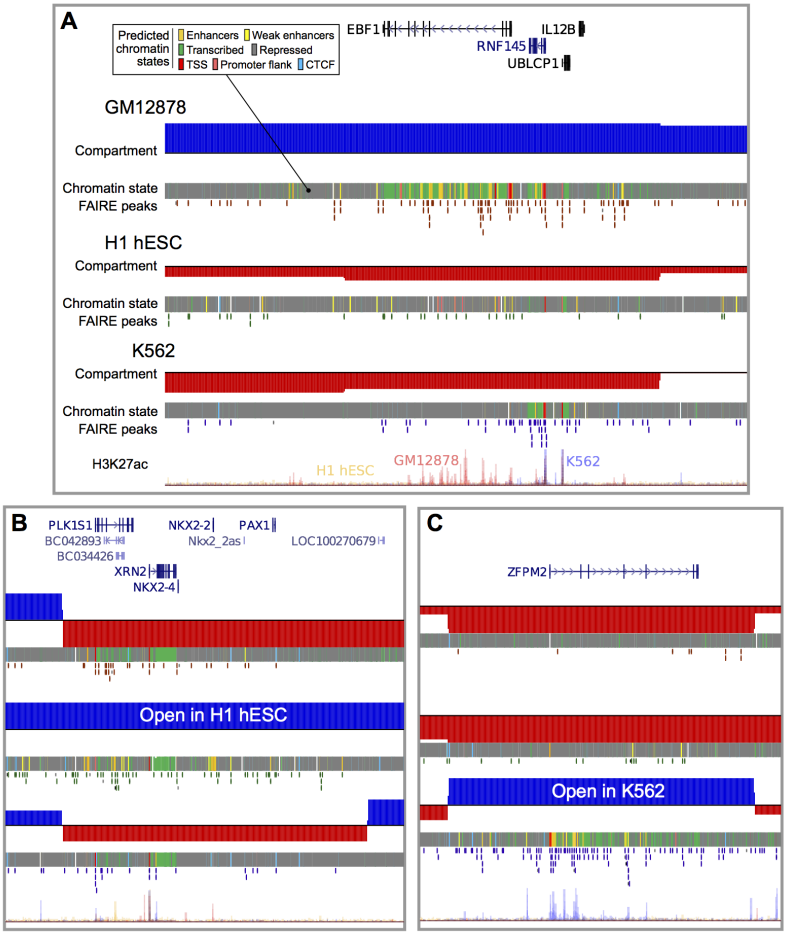

Figure 7: Structurally variable regions indicate cell type specific biology. Regions occupying the active A nuclear compartment in one cell type, but the repressive B compartment in the other two, were selected and ranked by the number of predicted active enhancers. Examples of particular interest from the top 5 regions per cell type are shown: (A) the chr5:158-159 Mb region which occupies the open A compartment in GM12878 cells, (B) the chr20:21-22 Mb region which is open in H1 hESC, (C) the chr8: 106-107 Mb region which is open in K562. Displayed tracks are: known (UCSC) genes, compartment eigenvectors, chromHMM/Segway combined chromatin state predictions, open chromatin FAIRE peaks, and H3K27ac signal.

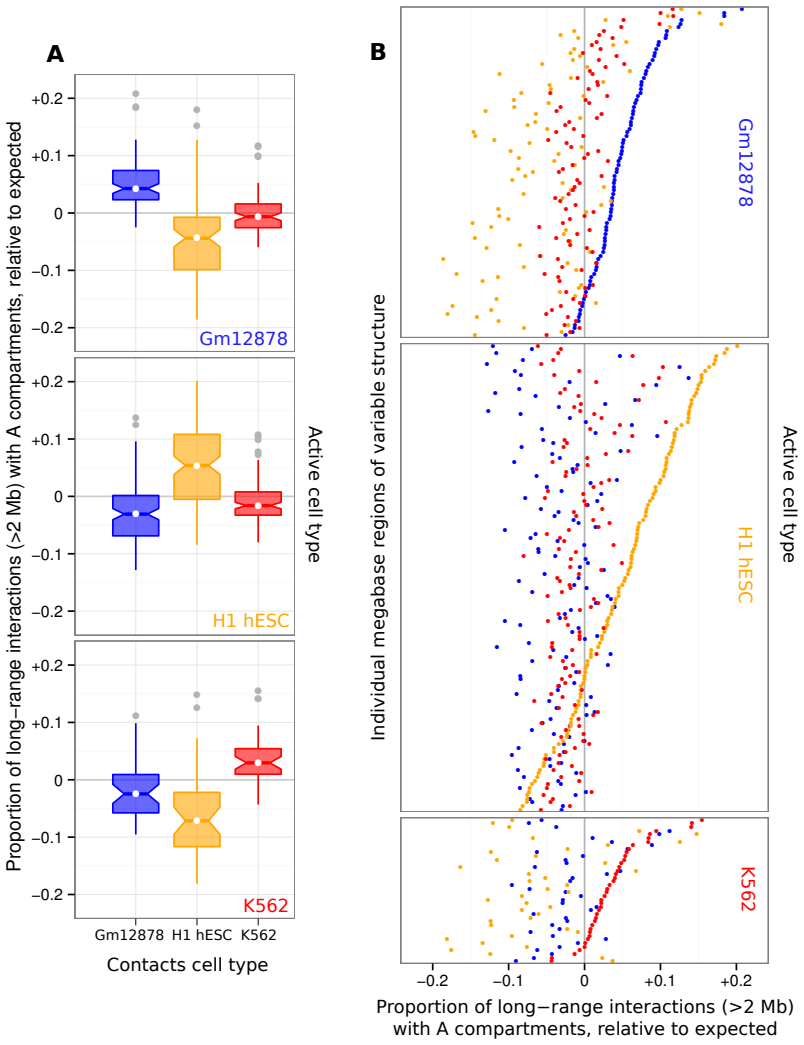

Figure 8: Regions of variable structure alter their long-range contacts. Placeholder

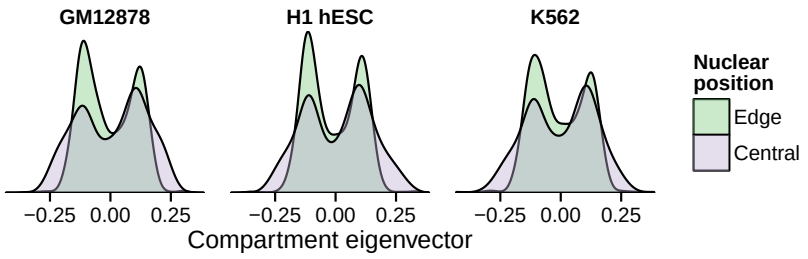

Figure 9: Nuclear positioning of chromosomes relative to compartment eigenvectors. Kernel density estimates showing peripheral chromosomes have a greater proportion of B compartments (negative eigenvectors) relative to centrally-position chromosomes. Positioning data from Boyle *et al.*[12] (Methods XX).

Chromosomes whose DAPI hybridisation signals were significantly enriched ($p \leq 2 \times 10^{-2}$) in the inner nuclear shell, as defined by Boyle *et al.*[12], made up the "central" group and included chromosomes 1, 16, 17, 19 and 22. Similarly the "edge" group had enriched signals ($p \leq 5 \times 10^{-3}$) in the outer shell relative to the inner nuclear shell and included chromosomes 2, 3, 4, 7, 8, 11, 13 and 18. The remaining chromosomes showed no significant preference to either inner or outer nuclear shells at $\alpha = 0.05$.[12]

We found that positive eigenvectors (reflecting A compartments) did show a modest relative enrichment in centrally-positioned chromosomes relative to those located at the nuclear periphery (Fig. 9). The significance of the difference in distribution of eigenvectors in the central and edge of the nucleus was determined by a two-sided Kolmogorov-Smirnov (K-S) test, with the null hypothesis that there is no difference between the empirical cumulative density functions of the central chromosome eigenvectors ($F_{central}$) and peripheral (F_{edge}). The difference was found to be statistically significant in each cell type ($H_0 : F_{edge} = F_{central}$; GM12878: $D = 0.11$, $p < 6 \times 10^{-4}$; H1 hESC: $D = 0.12$, $p < 8 \times 10^{-8}$; K562: $D = 0.10$, $p < 5 \times 10^{-3}$)

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] Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen Ca, Schmitt AD, Espinoza CA, Ren B (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**(7475): 290–4.

[4] 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.

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

[6] Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology*, **30**(1): 90–8.

[7] Dekker J, Marti-Renom Ma, Mirny La (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews. Genetics*, **14**(6): 390–403.

[8] Hu G, Cui K, Northrup D, Liu C, Wang C, Tang Q, Ge K, Levens D, Crane-Robinson C, Zhao K (2013) H2A.Z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation. *Cell Stem Cell*, **12**(2): 180–192.

[9] Li Z, Gadue P, Chen K, Jiao Y, Tuteja G, Schug J, Li W, Kaestner KH (2012) Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. *Cell*, **151**(7): 1608–1616.

[10] 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.

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

[12] Boyle S, Gilchrist S, Bridger JM, Mahy NL, Ellis Ja, Bickmore Wa (2001) The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human molecular genetics*, **10**(3): 211–9.