Transcriptional decomposition reveals active chromatin architectures and cell specific regulatory interactions

Sarah Rennie[1], **Maria Dalby**[1], **Lucas van Duin**[1], **and Robin Andersson**[1,*]

[1]The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark
[*]Corresponding author, robin@binf.ku.dk

ABSTRACT

Gene transcription is influenced by favourable chromosome positioning and chromatin architectures bringing regulatory elements in close proximity. However, it is unclear to what extent transcription is attributable to topological organisation or to gene-specific regulatory programs. Here, we develop a strategy to transcriptionally decompose expression data into two main components reflecting the positional relationship of neighbouring transcriptional units and effects independent from their positioning. We demonstrate that the positionally dependent component is highly informative of topological domain activity and organisation, revealing boundaries and chromatin compartments. Furthermore, features derived from transcriptional components can accurately predict individual chromatin interactions. We systematically investigate regulatory interactions and observe different transcriptional attributes governing long- and short-range interactions. Finally, we assess differences in regulatory organisations across 76 human cell types. In all, we demonstrate a close relationship between transcription and topological chromatin architecture and provide an unprecedented resource for investigations of regulatory organisations across cell types.

The three dimensional organisation of a genome within a nucleus appears to be crucial for correct transcriptional activity[1]. On a global level, transcriptional activation or repression is often accompanied by nuclear relocation of chromatin, forming chromatin compartments[2] of coordinated gene transcription[3,4] or silencing[5]. Locally, chromatin forms sub-mega base pair domains of self-contained chromatin proximity, commonly referred to as topologically associating domains (TADs)[6]. TADs frequently encompass interactions between regulatory elements, such as between promoters and enhancers[7–10] as well as between co-regulated genes[11], which influence cell-type restricted transcriptional programs. In contrast, ubiquitously expressed promoters are enriched close to domain boundaries[6]. This organisation suggests that transcriptional activity depends on the highly non-random genomic positioning of regulatory elements and genes resulting in a compatible three-dimensional context for correct transcriptional regulation.

A dependency between transcriptional activity and chromosomal positioning is supported by positional clustering of co-expressed eukaryotic genes[12], a phenomenon that is preserved across taxa[13]. In addition, neighbouring gene expression correlation co-evolves and is particularly evident at distances below a mega base pair[14]. These observations are in line with coordinated gene expression within TADs[11,15]. The strong relationships between expression, chromosomal positioning and chromatin organisation have previously been exploited by us to predict enhancer-promoter (EP) interactions from expression data[16] and by others to infer compartments of transcriptional activity from genome-wide chromatin interaction data[2,8]. RNA expression was further among the top ranked features for predicting EP interactions in a recent machine learning approach[17]. However, it is unclear to what extent transcription is determined by chromosomal position and chromatin environment, and what fraction of transcription is due to independent regulatory programs not affected by the former. A way to systematically extract these components from expression data could lead to new insights into chromatin topology and in biological samples not susceptible to chromatin conformation capture techniques.

Here we investigate the coupling between transcriptional activity and chromosomal organisation by asking: what proportion of transcriptional output from a genomic region can be explained by its chromosomal position? And, what proportion can be explained as a result of independent effects not relating to its position? To this end, we hypothesised that a transcription unit (TU), i.e. of a gene or an enhancer[18], is likely to be more similar in terms of transcriptional output to its, linearly, proximal TUs than to distal loci, which are likely to be associated with different domains of chromatin interactions. By the modelling of this dependency, we show that transcription can be decomposed into a positionally dependent and an independent component, and that both effects are strongly represented, with contributions depending on cell type and location. We observe a tight coupling between transcriptional activity and chromatin organisation and suggest that much of the latter may be inferred by RNA expression data alone. We show that the positionally dependent component is highly reflective of chromatin organisation, revealing chromatin compartments and boundaries of transcriptionally active TADs. We further demonstrate how transcriptional components can be used to infer cell type-specific chromatin interactions. We demonstrate the accuracy of our approach in well-established cell lines and then decompose expression data from 76 human cell types in order to investigate their active chromatin architectures, providing an unprecedented resource of regulatory organisations across human cells.

Results

Decomposition of expression data reveals chromatin compartments and independent gene regulatory programs

We developed a transcriptional decomposition approach to separate the component of expression reflecting an underlying positional relationship between neighbouring genomic regions (positionally dependent (PD) component) from the expression dictated by TUs' individual regulatory programs (positionally independent (PI) component). Based on Bayesian modelling, our strategy decomposes normalised (zero-inflated negative binomial) aggregated RNA expression read counts in 10kb genomic bins into three components (Fig. 1A, Methods): a positional first order random walk component (PD), an independent and identically distributed component (PI) and a model intercept (α): $\nu_i = \alpha + \mathrm{PI}_i + \mathrm{PD}_i$, where ν_i is the library depth adjusted log mean expression in bin i.

We applied transcriptional decomposition to replicated Cap Analysis of Gene Expression (CAGE)[19] data[20], measuring transcription initiation sites and steady-state abundances of capped RNAs, from GM12878, HeLa-S3, and HepG2 cells. Upon close inspection (Fig. 1B), the PD component displayed considerably broader patterns than the PI component and appeared highly similar between cell types. Despite overall similarities, we identified large differences in individual loci in the PD component between cells, as exemplified by *KCNN3*, *NOS1AP*, and *PBX1* genes (Fig. 1C-E). For instance *NOS1AP* is surrounded by low PD signal and appears to reside in polycomb-repressed chromatin in GM12878 cells, as indicated by high levels of histone modification H3K27me3 and low levels of histone modification H3K36me3. HepG2 and HeLa-S3 cells displayed opposing signals for this locus, suggesting that the PD component contains information about chromatin compartments.

In order to understand the observed effects on a genome-wide scale, we compared the PD component with HiC-predicted chromatin compartments[8] in GM12878 cells. We observed that the PD signal in regions of active chromatin was higher than in those of facultative or constitutive chromatin, while constitutive chromatin states had the lowest signal (Fig. 2A). Overlaying the PD component on HiC compartment boundaries also showed clear shifts, many magnitudes stronger than what could be detected using the PI component (Fig. 2B and Supplementary Fig. S1A). In addition, we observed that the PD component clearly correlated within compartments more strongly than between compartments (Supplementary Fig. S1B-E). These results show that the states and boundaries of compartments are reflected by the PD signal and its relative change between consecutive bins.

To examine the localised patterns observed in the PI component (Fig. 1B), we trained a random forest model (Methods) on GM12878 transcriptional components to predict the presence or absence of DNase I hypersensitive sites (DHSs), histone variant H2A.Z and post-translational histone modifications H3K4me3 and H3K27ac (binarised DNase-seq and ChIP-seq data[21] in each bin), each associated with features of (transcriptionally) active promoters[22]. The resulting models allowed us to generate a probability distribution for each mark given each transcriptional component. For all tested marks we observed a clear bias with stronger predictive power from the PI component than the PD component (Fig. 2C). These results indicate that the PI component, in contrast to the PD component, carries information about promoter-localised and expression-level associated effects.

Overall, we found that both the PD and PI components explained considerable proportions of expression levels in GM12878 cells (Supplementary Fig. S2A). Each component explained on median roughly half of the expression levels in expressed bins, with contributions varying between loci (Supplementary Fig. S2B).

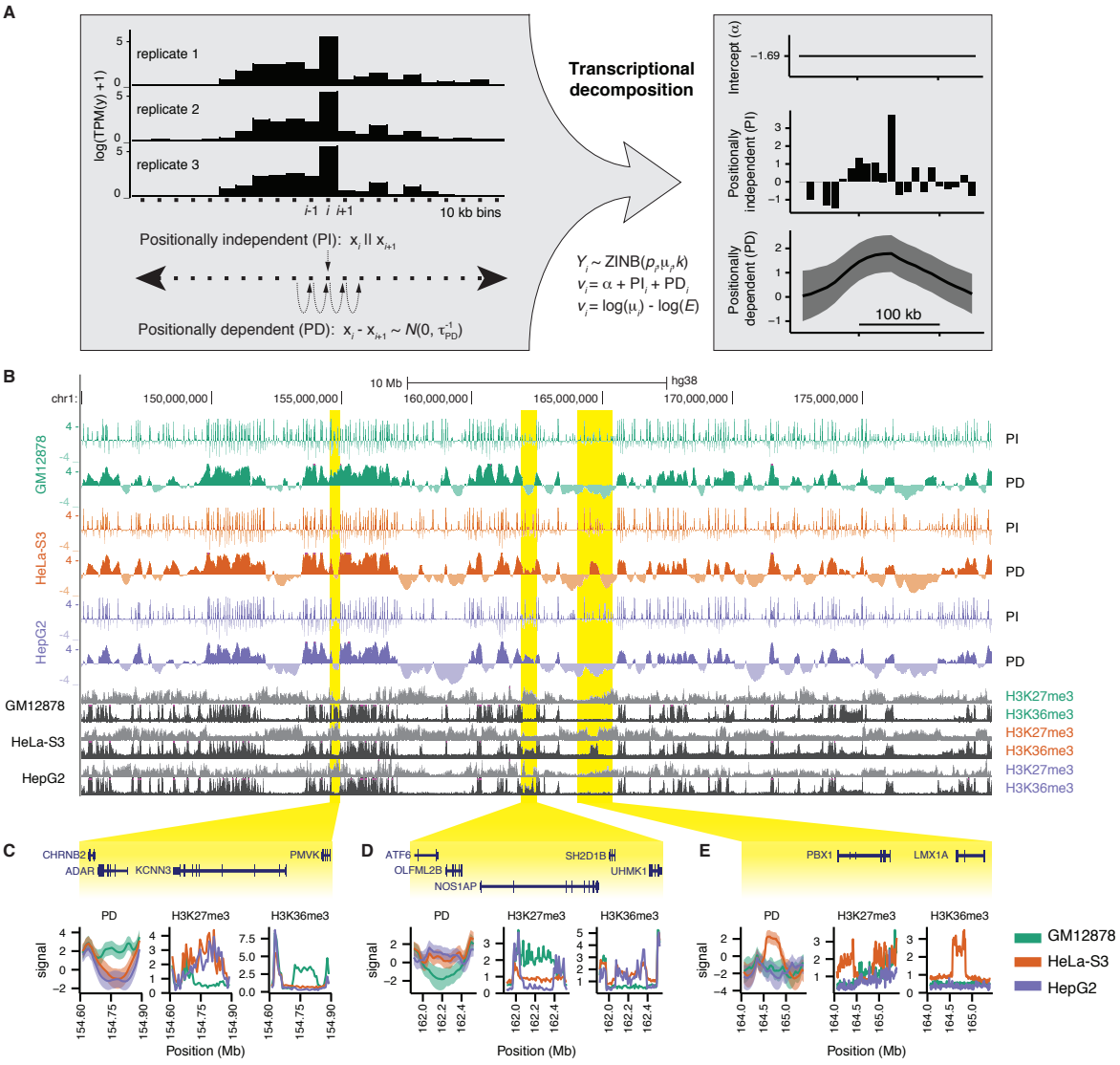

Figure 1. Transcriptional decomposition across chromosomes. A: Schematic illustrating how replicate samples are decomposed into transcriptional components. RNA expression count data, quantified in genomic bins (here 10kb), are normalised by a zero-inflated negative binomial. Via approximate Bayesian modelling, normalised expression data are decomposed into an intercept (α), a positonally independent (PI) component and a positionally dependent (PD) component. The PD compoenent is modelled as a first-order random walk, in which the difference between consecutive bins is assumed to be Normal and centred at 0. **B**: PI and PD componenets, as well as H3K27me3 and H3K36me3 ChIP-seq data for GM12878, HeLa-S3 and HepG2 cells at locus chr1:145,000,000-180,000,000. **C-E**: Loci (highlighted in B) around *KCNN3* (**C**), *NOS1AP* (**D**), and *PBX1* (**D**) genes showing cell-type specific PD signals. The PD signal and ChIP-seq data associated with repression (H3K27me3) and activation (H3K36me3) are shown.

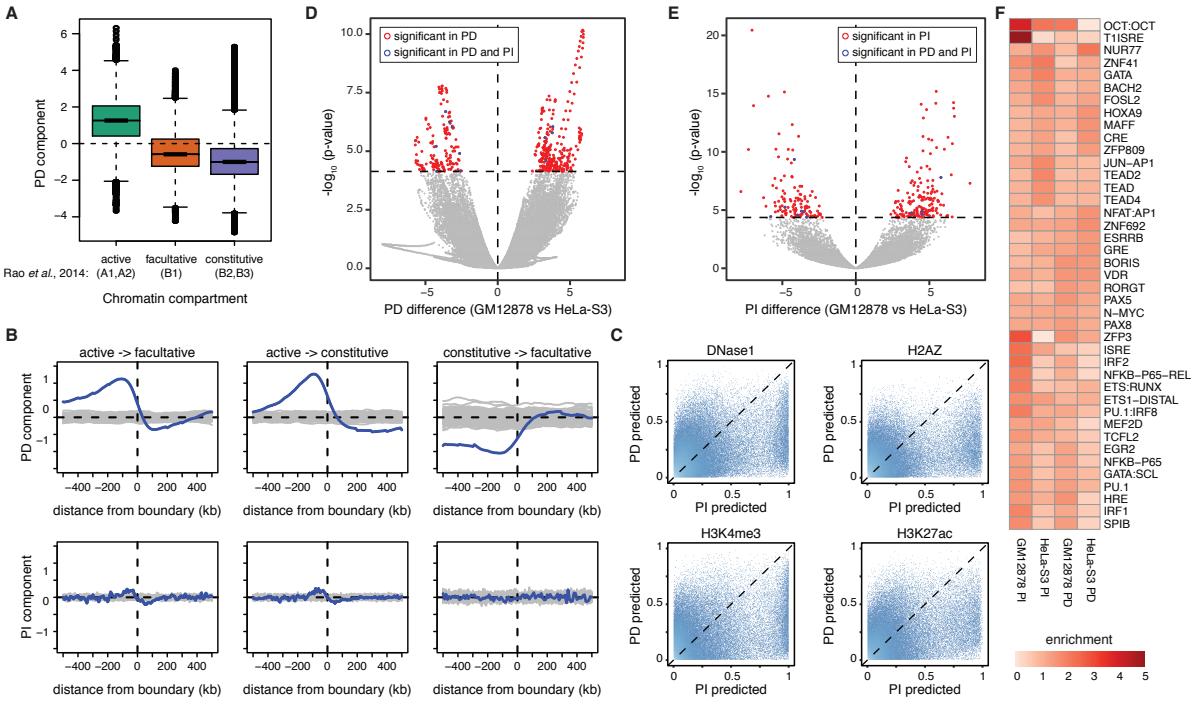

Figure 2. Transcriptional components reveal chromatin compartments and localised promoter-associated effects. **A**: Box-and-whisker plot of GM12878 PD signal grouped according to HiC-derived chromatin compartments[8]. **B**: Average PD signal (top row) and PI signal (bottom row) around boundaries of HiC-derived chromatin compartments. Horizontal dotted lines represent the transition between positive and negative and grey bands represent equivalent shifts across random compartment boundaries. **C**: Random forest class (presence/absence) probability of DNase1, H2AZ, H3K4me3, and H3K27ac as learned from the PI compoenent (x axes) and the PD component (y axes). **D**: PD difference (x axis) versus FDR-adjusted p-value (rescaled by -log10) for PD component differential expression between GM12878 and HeLa-S3. Red represents significant bins unique to the component, blue those common to both. **E**: As D but for PI component. **F**: TF motif enrichment (foreground vs background) around expressed CAGE-derived promoters associated with GM12878 or HeLa-S3 biased differentially expressed PD or PI bins. See also Supplementary Figure S2E.

When compared with HeLa-S3 cells, we observed clear differences in both the PD and PI components between cell types and that differences were orthogonal between components (Fig. 2D,E). Differentially expressed bins in the PD component (Supplementary Table S2) were, in line with its relationship with chromatin compartments, to a large degree associated with cell-type differences in chromatin states, changing from silent to H3K36me3-associated active chromatin in up-regulated bins (Supplementary Fig. S2C-D). The PI component, on the other hand, showed localised differential expression of bins (Supplementary Table S3), that were associated with highly cell-type specific enrichments of transcription factor motifs (Fig. 2F and Supplementary Fig. S2E), for instance NFKB and IRF in GM12878 cells.

Similar patterns of transcriptional components derived from GM12878 CAGE data were found when we applied transcriptional decomposition, guided by CAGE-estimated hyperparameters, to GM12878 RNA-seq data[22] (Supplementary Fig. S3, Supplementary Table S1, Methods), confirming the observed characteristics and differences between CAGE-derived components. Taken together, we conclude that expression data can be decomposed into a positionally dependent component, revealing chromatin compart-

ment activity, and a positionally independent component carrying information about localised, independent expression-associated events.

Expression associated domains show clear resemblance with active chromatin topology

Apart from displaying clear shifts at compartment boundaries, we noted that the PD component contained sub-patterns within broader consecutive regions of positive signals (Fig. 1B). We posited that such structures could represent finer, transcriptionally active chromatin organisation not necessarily reflecting chromatin compartments, but rather boundaries of active TADs. To test this hypothesis, we trained a generalised linear model (GLM) to predict HiC-derived TAD-boundaries[8] from features derived from transcriptional components (Supplementary Table S4). The GLM yielded an area under the ROC curve (AUC) of 0.73 (AUC 0.85 in regions of positive PD signal), indicating that there is information in the transcriptional components to infer TAD boundaries (Supplementary Fig. S4A,B). Among features considered, the model ranked the PD gradient (first order derivative), the PD inter-cell stability, and PD variance among the most important for predicting TAD boundaries (Supplementary Fig. S4C). Based on the GLM feature importance ranking, we devised a score to rank PD boundaries at significant gradients in the PD signal that also had low positional standard deviation (high stability) across cells (Fig. 3A, Methods). In GM12878 cells, we detected 5,109 boundary locations of PD sub-patterns. We refer to the regions demarcated by PD boundaries as expression associated domains (XADs, Fig. 3A).

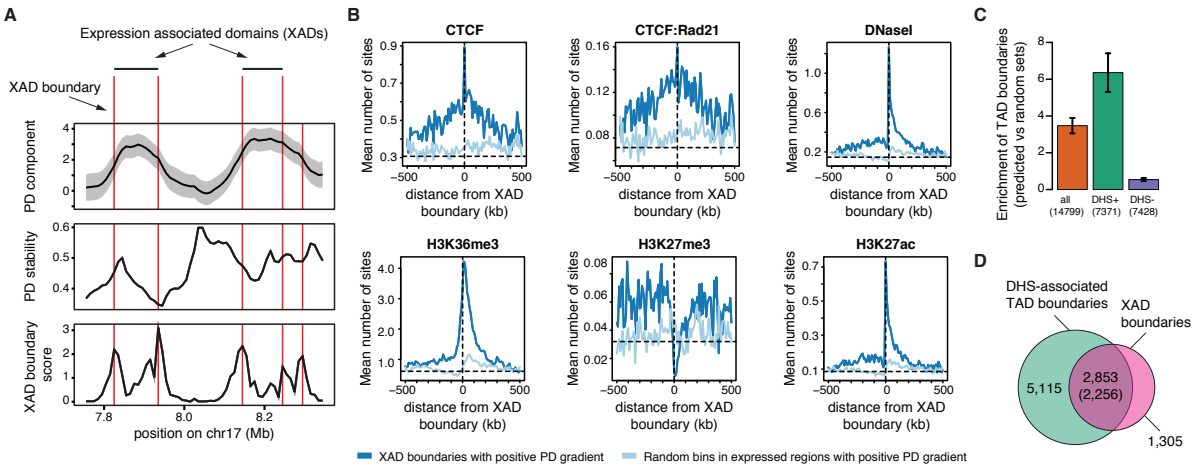

Figure 3. Expression associated domains mark regions of active topological domains **A**: Approach for identifying boundaries of expression associated domains (XADs) based on a PD boundary score. Shown are PD signal (mean +/- standard deviations), PD stability (across cell PD standard deviation) and the XAD boundary score. **B**: Average GM12878 profiles of binarised ChIP-seq data for CTCF, CTCF in combination with Rad21 (cohesin), DNaseI, H3K36me3 H3K27me3, and H3K27ac at XAD boundaries with positive PD gradient (dark blue) and at random expressed bins with postive PD gradient. Dotted vertical line represents boundary locations and horizontal dotted line represents background mean for given ChIP-seq mark. **C**: Enrichment of GM12878 TAD boundaries[8] among XAD boundaries compared to random bins proximal to expressed bins (DHS+ for DHS associated, DHS- for DHS non-associated). **D**: Venn diagram of association between GM12878 XAD boundaries and proximal (within 5 bins) DHS-assocated TAD boundaries.

We next assessed the occurrence of DHSs, ChIP-seq binding site peaks for architectural proteins CTCF and Rad21 (a subunit of cohesin) and histone modifications H3K36me3, H3K27me3, and H3K27ac

around GM12878 XAD boundaries (Fig. 3B). We observed an enrichment of DHSs, H3K27ac, and binding of CTCF alone and in combination with Rad21 around XAD boundaries. In addition, H3K36me3, H3K27ac, and DHSs were more enriched downstream than upstream of positive-gradient XAD boundaries. The opposite trend was observed for H3K27me3. The observed DHS and ChIP-seq patterns around XAD boundaries (Fig. 3B) resemble those around TAD boundaries in active compartments[23–26]. In line with these and the GLM results, we observed that GM12878 XAD boundaries were in general highly enriched in HiC-derived TAD boundaries[8] from the same cells (>3-fold enrichment over background, Fig. 3C). Specifically, XAD boundaries were highly enriched in DHS-associated TAD boundaries (>6-fold enrichment over background), but not in those distal to DHSs (Fig. 3C and Supplementary Fig. S5A,B). Out of 5,109 GM12878 XAD boundaries, 2,853 (56%) were proximal (within 5 bins) to DHS-positive TAD boundaries (Fig. 3D). Still, 69% (5,115 out of 7,371) of DHS-positive TAD boundaries were distal to XAD boundaries, suggesting that XADs represent a subset of DHS-positive TADs. Joint XAD and TAD boundaries were primarily found within active chromatin while XAD-unsupported DHS-positive TAD boundaries more frequently resided in facultative or constituitive heterochromatin (Supplementary Fig. S5C, p<1e-258, Chi-squared test). Furthermore, joint XAD and TAD boundaries displayed a higher HiC[8] chromatin interaction directionality[6] than DHS-positive TAD boundaries distal to XAD boundaries (Supplementary Fig. S5D,E), indicating that expression-associated chromatin is linked with stronger TAD boundaries (greater insulation). However, both sets were similarly associated with Rad21 (Supplementary Fig. S5F,G), whose co-binding with CTCF is believed to provide strong TAD boundaries[23–25]. Taken together, these results demonstrate that the PD component can be used to infer chromatin topology in active chromatin compartments.

Transcriptional components are informative of regulatory chromatin interactions

Since the PD component was strongly associated with structures of transcriptionally active TADs, we questioned the utility of transcriptional components in reflecting individual proximity based interactions. Namely, what does it mean to be proximal in the nucleus, from a transcriptional perspective?

To test the power of transcriptional components to classify chromatin interactions, we employed a random forest classification scheme with features derived from transcriptional components (Supplementary Table S5) as predictors on GM12878 promoter-capture HiC (CHiC) data[9] at 10kb resolution (Methods). A total of 1.8 million bin pairs were considered (bin-bin distance $>$ 50kb and $\leq$ 2Mb), for which each pair referred to a CHiC promoter bait and a potential target that overlapped a transcribed promoter[20] or transcribed enhancer[16] (>197,000 bin pairs with a potential enhancer target). Using a lenient interaction score threshold to define positives (CHiCAGO[27] score $\geq$ 3), we noted an overall high proportion of negatives (negative to positive (NP) ratio of 44.6:1, see Supplementary Table S6) and that the proportion of negatives increased over distance. In order to deal with such unbalanced data, we over-sampled[28] the positives and under-sampled the negatives in the training data to a fully balanced set across distances in a ten-fold cross-validation scheme. In each cross-validation round, we balanced the training data and predicted on held-out data at 20:1 or raw (unmodified) NP ratios.

Overall, we observed a good predictive performance (AUC: 0.74) on raw NP ratios for lenient interaction thresholds (Fig. 4A). Using a strict interaction threshold (CHiCAGO score $\geq$ 5) for evaluation increased the AUC (0.83) but reduced the recall (Fig. 4B). At a 20:1 NP test ratio, the overall predictive performances increased (AUC 0.82 and 0.88 for lenient and strict interaction thresholds, respectively). Interestingly, we observed a better precision and recall when evaluation of results was limited to enhancer targets (Supplementary Table S7, AUC of 0.85 and 0.89 for lenient and strict interaction thresholds at a 20:1 NP

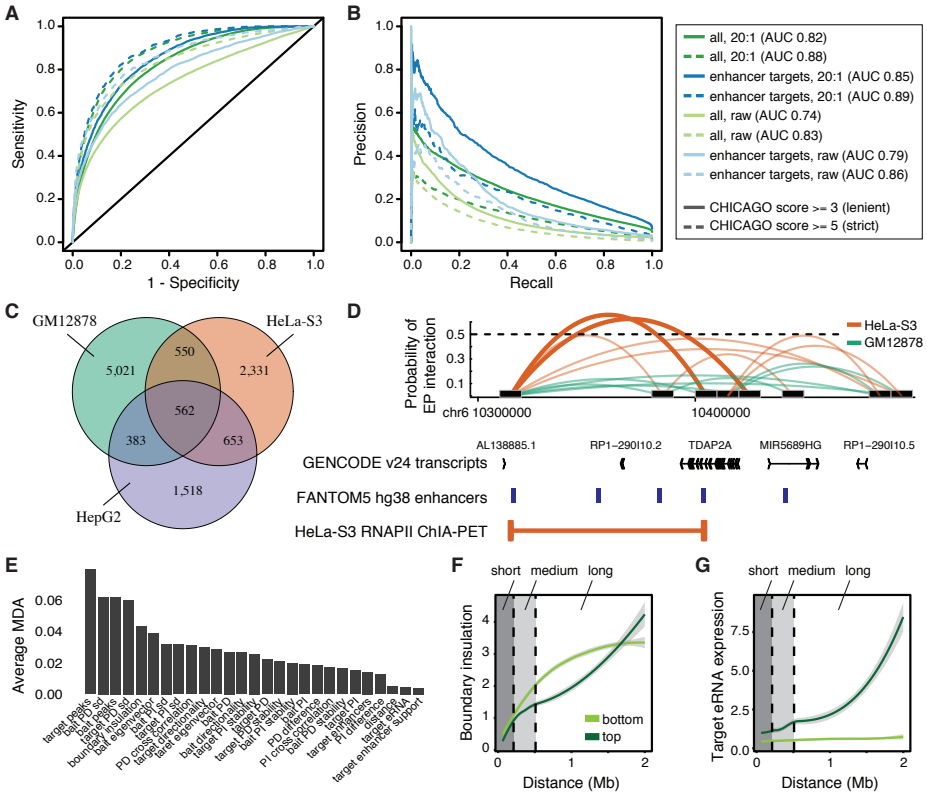

Figure 4. Transcription predicts cell type specific proximity interactions. **A-B**: Performance curves for predicting bait-target interactions from CAGE-universal features, split according to CHICAGO[27] score, testing negative to positive ratios and target feature type. **C**: Overlaps of predicted bait-enhancer interactions between GM12878, HeLa-S3, and HepG2 cells. **D**: An example of a loop predicted in HeLa-S3, but not in GM12878 cells, validated by HeLa-S3 RNAPII ChIA-PET interaction data. **E**: Features predictive of bait-target interactions, ordered by average mean decrease accuracy (MDA) across models from 10-fold cross validation. **F-G**: Loess curves representing feature separation over distance between high (top) and low (bottom) predicted probabilities, shown for features (**F**) *nbounds* (XAD boundary insulation) and (**G**) *eRNA_targ* (enhancer expression at target).

ratio, respectively; see Supplementary Fig. S6 for the effect of interaction thresholds on EP prediction performance). We observed varying predictive performances across chromosomes (Supplementary Fig. S7), likely reflecting differences in gene densities and transcriptional activities (data not shown). Despite the overall good EP predictive performance, both precision and recall decreased over increasing distances on held-out test data, in accordance with an increasing NP ratio (Supplementary Fig. S8). To circumvent the distance effect, we established a distance-dependent threshold in random forest voting, guided by the optimal F1 score, significantly improving the predictive performance over distance (Supplementary Fig. S9).

Motivated by the good performance in predicting enhancer-promoter (EP) interactions, we next used the GM12878-trained random forest model to predict EP interactions also in HeLa-S3 and HepG2 cells (Supplementary Tables S8, S9). We noted that many predicted interactions were specific to each cell type (Fig. 4C, 48-78%), with only a small fraction (9-20%) of interactions shared between all three cell lines (see Supplementary Fig. S10 for results using a strict interaction threshold). For instance, the promoter of gene *TDAP2A* was predicted to interact with a 100kb downstream enhancer in HeLa-S3 cells, supported by HeLa-S3 RNAPII ChIA-PET interaction data[22] (Fig. 4D).

Transcriptional activity at the promoter target was ranked among the top features for predicting chromatin interactions (Fig. 4E). Separately training three random forest classifiers for short, medium, and long-range distances (covering bait-target distances within (50,200], (200,500], and (500,2000] kb, respectively) did not improve the predictive performance compared to the full model (data not shown), but allowed us to further investigate features driving long-range interactions (Supplementary Fig. S11A). As expected, boundary insulation (number of XAD boundaries between bait and target) had a higher influence on long-distance interactions than shorter ones (Fig. 4F). Interestingly, eRNA expression at the target enhancer clearly distinguished predicted positive from predicted negative EP interactions, and its importance increased over increasing EP distances (Fig. 4G).

In support of eRNA production at positive EP interactions, both enhancers and promoters in predicted cell-type specific interactions clearly showed an expression bias towards the cell type in which the interactions were identified, in contrast to shared EP interactions (Supplementary Fig. S11B,C). These results indicate that expression of regulatory elements is both reflecting their cell-type specific regulatory activity and their regulatory interactions. This is supported by observations that regulatory active enhancers that are interacting with promoters are more likely to be transcribed than non-interacting ones[29]. Taken together, we demonstrate that transcription is highly informative of regulatory interactions.

Transcriptional decomposition reveals regulatory differences between human cell types

We have above established that the PD component can be used to accurately infer boundaries of active TADs as well as differences in chromatin compartments and EP interactions between well-established and biologically distal cells lines. We continued with exploring what insights could be gained by transcriptional decomposition of primary cell CAGE data[20], for cell types for which chromatin topology is to a large degree unknown. We applied transcriptional decomposition to replicated primary cell CAGE data[20] expanding to a total of 76 human decomposed cell types (Supplementary Table S11, Methods). Using the resulting components, we extracted XAD boundaries and extended the previous set of pairs defined in the GM12878 training above to a common set of bin-pairs across all cell types (Supplementary Table S10), over which EP interactions were predicted.

Overall, the PD and PI components displayed opposing trends when compared across cell types. Many

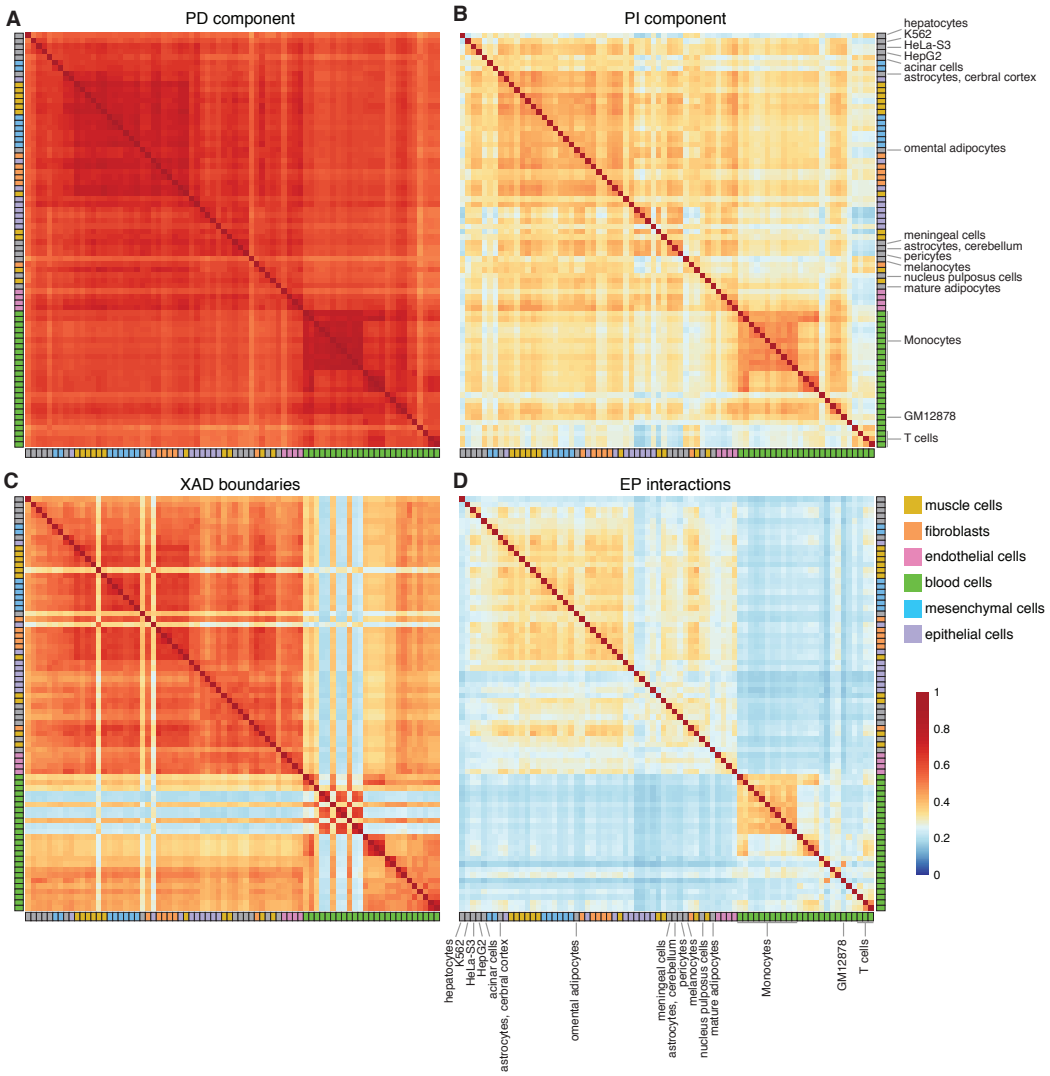

Figure 5. Transcriptional decomposition across 76 human cell types. **A-D**: Heat maps depicting pairwise similarities of the PD (**A**), PI (**B**) components, XAD boundary locations (**C**) and EP interactions (**D**) across cell types. All similarity scores calculated between cell types using $1 - L_1$-norm on binary data sets based on the sign of the PD, PI components or the presence/absence of XAD boundaries or EP interactions.

genomic bins showed highly similar PD signal across cells, while the PI component was rarely shared across more than a few cell types (Supplementary Fig. S12A,B). The PD component showed by far the strongest similarity with other cells (Fig. 5A), whereas the PI component grouped more closely according to cell type association and distinguished blood cells from mesenchymal, muscle, and epithelial cells (Fig. 5B).

Since many of the observed features around GM12878 XAD boundaries (Fig. 3B) tend to be cell-type specific (e.g. DHSs[30]), we wanted to further explore the cell-type specificity of XAD boundaries. We observed that XAD boundaries tended to be either highly cell-type specific or ubiquitous (Supplementary Fig. S12C), similar to that of TAD boundaries[31]. Samples associated with monocytes showed a particularly low similarity with other cell types (Fig. 5C).

In comparison, EP interactions exhibited very little agreement across cell types (Fig. 5D), with very few shared interactions across the full repertoire of samples (Supplementary Fig. S12D). Interestingly, cell type specific EP interactions were on average more distal than interactions shared across cells (Supplementary Fig. S12E). When examining EP interactions between groups of cells we observed clear differences at key identity genes. Leukocyte biased genes *ARHGAP25* (Fig. 6) and *CD48* (Supplementary Fig. S13) showed clear differences in predicted EP interactions, which were validated by GM12878 CHiC interaction data.

These results clearly demonstrates the value of our generated resource and how transcriptional decomposition can be used to gain new insights into the regulatory organisation of cells.

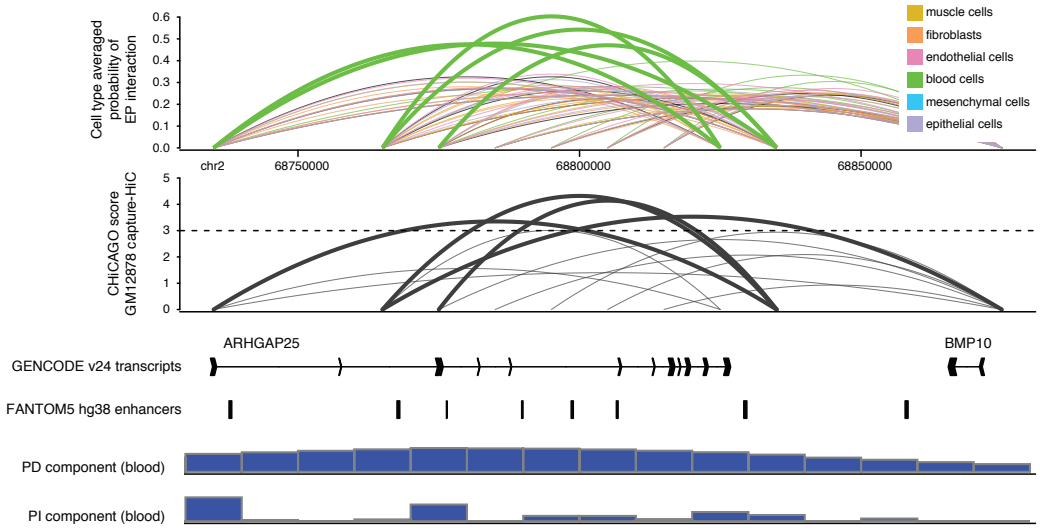

Figure 6. Transcriptional decomposition reveals cell type specific regulatory organisation around *ARHGAP25* gene. From top to bottom: predicted probability of EP interactions averaged across groups of cells (as in Figure 5), CHiCAGO score of interaction based on GM12878 capture HiC data. Below are displayed (in the following order) GENCODE v24 transcripts, FANTOM5 enhancers, the average PD and PI component across blood cells.

Discussion

To what extent transcriptional activity is attributable to chromosomal organisation, and vice versa, is debatable. The decomposition of expression data into positionally dependent and independent components along chromosomes, as introduced in this study, suggests that positionally attributable expression levels closely reflect chromatin compartments and domain architecture, and account for a sizeable overall fraction of TU expression levels. We observe that the level of positional dependency of expression data varies between loci and cell types, suggesting locus- and cell-dependent effects of topological organisation and varying levels of individual gene regulatory programs. Our modelling also has merit in the prediction of fine-scale chromatin interactions in a manner broadly scalable across large numbers of samples, potentially paving the way for large-scale cost- and time-effective computational analyses across atlases of high quality expression data sets, such as FANTOM5 or GTex[20,32].

The usage of expression data to infer topological chromatin organisation is limited by the inability to inform on transcriptionally silenced, closed or poised states. However, by focussing only on expressed TUs, such as those observed with CAGE, and their relative relationships, we can attempt to understand patterns that are highly relevant to various cell types of interest. Interestingly, our predictions follow general properties commonly observed across chromatin conformation datasets. XAD boundary locations were seen to be informed by the presence of stable PD signal across cell types, reflected in their high degree of sharing, which closely corresponds to the nature of TADs, whose locations appear largely cell type invariant[6,31]. On the other hand, predicted EP-interactions show strong cell type specificity across our repertoire of analysed cell types and are biased towards the cell types in which they are actively transcribed.

Our results concur with, and significantly extend, the previous finding that CAGE is an important predictor of chromatin interactions[17]. Notably, enhancer RNA expression level at predicted targets rapidly increased with the distance between the bait and the target, suggesting further work could be merited in teasing out the relationship between distance and enhancer selection. Our predictions show strong performance in distinguishing interacting from non-interacting examples, however a more robust prediction could be gained by incorporating other data types as they become available for cell types of interest, such as histone modifications and transcription factor binding, which have previously been seen to be predictive[17].

In summary, we suggest that the relative positioning, activities and stability of active TUs form a vital part of their correct and organised functioning within their target cell type. In addition, we hope that the generation of predicted components, XAD boundaries and EP-interactions across a total of 76 cell types from the FANTOM5 consortium[20] allows gaining a deeper understanding of the dynamic regulation at key identity genes in a wide diversity of cellular states not yet subjected to methods informing on their higher order structures.

...

References

1. Gorkin, D. U., Leung, D. & Ren, B. The 3D Genome in Transcriptional Regulation and Pluripotency. Cell stem cell **14**, 762–775 (2014).

2. Lieberman-Aiden, E. et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science (New York, N.Y.) **326**, 289–293 (2009).

3. Osborne, C. S. et al. Active genes dynamically colocalize to shared sites of ongoing transcription. Nature genetics **36**, 1065–1071 (2004).

4. Schoenfelder, S. et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nature genetics **42**, 53–61 (2010).

5. Guelen, L. et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature **453**, 948–951 (2008).

6. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature **485**, 376–380 (2012).

7. Noordermeer, D. et al. Temporal dynamics and developmental memory of 3D chromatin architecture at Hox gene loci. eLife **3**, e02557 (2014).

8. Rao, S. S. P. et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell **159**, 1665–1680 (2014).

9. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nature genetics **47**, 598–606 (2015).

10. Heidari, N. et al. Genome-wide map of regulatory interactions in the human genome. Genome research **24**, 1905–1917 (2014).

11. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature **485**, 381–385 (2012).

12. Cohen, B. A., Mitra, R. D., Hughes, J. D. & Church, G. M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nature genetics **26**, 183–186 (2000).

13. Fukuoka, Y., Inaoka, H. & Kohane, I. S. Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. BMC genomics **5**, 4 (2004).

14. Ghanbarian, A. T. & Hurst, L. D. Neighboring Genes Show Correlated Evolution in Gene Expression. Molecular biology and evolution **32**, 1748–1766 (2015).

15. Le Dily, F. et al. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes & development **28**, 2151–2162 (2014).

16. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. Nature **507**, 455–461 (2014).

17. Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. Nature genetics **48**, 488–496 (2016).

18. Andersson, R., Sandelin, A. & Danko, C. G. A unified architecture of transcriptional regulatory elements. Trends in Genetics **31**, 426–433 (2015).

19. Kanamori-Katayama, M. et al. Unamplified cap analysis of gene expression on a single-molecule sequencer. Genome research **21**, 1150–1159 (2011).

20. FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. A promoter-level mammalian expression atlas. Nature **507**, 462–470 (2014).

21. Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The Ensembl Regulatory Build. Genome Biology **16**, 56 (2015).

22. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature **489**, 57–74 (2012).

23. Li, Y. et al. Characterization of constitutive CTCF/cohesin loci: a possible role in establishing topological domains in mammalian genomes. BMC genomics **14**, 553 (2013).

24. Tang, Z. et al. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell **163**, 1611–1627 (2015).

25. de Wit, E. et al. CTCF Binding Polarity Determines Chromatin Looping. Molecular cell **60**, 676–684 (2015).

26. Pope, B. D. et al. Topologically associating domains are stable units of replication-timing regulation. Nature **515**, 402–405 (2014).

27. Cairns, J. et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biology **17**, 1 (2016).

28. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002).

29. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. Nature cell biology **489**, 109–113 (2012).

30. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. Nature cell biology **489**, 75–82 (2012).

31. Schmitt, A. D. et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell reports **17**, 2042–2059 (2016).

32. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nature genetics **45**, 580–585 (2013).

Acknowledgements

This project has received funding from the Danish Council for Independent Research and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 638173). The authors would like to thank Hideya Kawaji for remapping and DPI tag

clustering of FANTOM CAGE data and members of the Robin Andersson lab and Albin Sandelin lab for rewarding comments and discussions.

Author contributions statement

S.R and R.A. conceived the project; S.R conducted most analyses, with support from L.D, M.D and R.A; S.R and R.A wrote the paper; all authors reviewed the final manuscript.

Additional information

Data generated in this project have been made available at https://zenodo.org/record/556727 (DOI:10.5281/zenodo.556727) and https://zenodo.org/record/556775 (DOI:10.5281/zenodo.556775). The authors declare no competing financial interests.

Methods

Most analyses were carried out in *R*[33]. All liftovers from hg19 to hg38 were carried out using the chain `hg19ToHg38.over.chain` downloaded from UCSC[34] using the `R` package `rtracklayer`[35], unless otherwise stated. All overlaps between genomic regions were carried out using the `R` package `GenomicRanges`[36], unless otherwise stated.

Processing of CAGE data sets

Source of CAGE data

Data was produced by the FANTOM5 project[37]. The data was mapped to hg38 and CTSSs (CAGE tag starting site) were clustered into tag clusters (TCs) according to decomposition peak identification (dpi) generation as per FANTOM5[38]. Only samples with more than 0.5 million tags mapping within the TCs were included in the analysis.

Enhancer calling

Enhancers were called based on bidirectional balanced RNA signatures as per the FANTOM5 consortium[39]. Enhancers were only identified distal to known exons (+/-100bp region from boundaries) and transcription start sites +/-300bp, defined by GENCODE v24 annotation. In total, 63,285 enhancers were identified across 1,829 libraries.

Due to varying noise levels across FANTOM libraries and the intrinsic low expression levels of transcribed enhancers, library-specific noise levels were estimated to define of robust set of enhancers in each sample. For each library, expression was quantified in randomly sampled genomic regions distal to assembly gaps, DNase hypersensitive sites (ENCODE), known exons and gene TSSs (GENCODE) to create a genomic background expression distribution. For each library, we called an enhancer active (used) if its expression was above the 99.9th quantile of the library's genomic background expression distribution. The robust set of enhancers consist of 60,215 over 1,829 libraries, being significantly expressed in at least one library. The expression was quantified and TPM (tags per million) normalised according to the total number of mapped reads within the full set of TCs.

Data binning into 10kb regions

CTSS files containing positions of raw CAGE counts from libraries for the ENCODE tier 1 cell lines (GM12878, K562, HeLa-S3 and HepG2) were intersected with non-overlapping 10kb regions across the genome. For each chromosome (chr1-chr22 and chrX), regions were defined from coordinate 1 (1-based) and in consecutive complete 10kb blocks, up to two blocks after the last bin containing a single CAGE tag across the set of libraries. Region sizes of 40kb and 100kb were also considered, but 10kb was exclusively used in the analyses for comparability to high resolution chromatin capture data sets.

Data distribution

Due to the sparsity of transcription in bins associated with non-structural regions and regions poorly mapped, including the case of libraries with poor sequencing depth, a zero-inflated negative binomial distribution was assumed to hold for the counts across the 10kb bins on each chromosome, given by, for bin i, $Y_i \sim \text{ZINB}(p_i, \mu_i, k)$, where Y_i is bin count, p_i represents the zero probability parameter, μ_i represents the mean and k the size, or dispersion, parameter. The mean is given by $\mu_i = k\frac{1-p_i}{p_i}$, with

variance $\sigma_i^2 = \mu_i(1+\frac{\mu}{k})$. The model assumes a zero-truncated negative binomial distribution on the non-zero counts ($Y_i \sim NB(\mu_i, k)$ for $y_i = 1, 2, \ldots$) with probability equal to $1 - p_i$.

Decomposition using random effects model

The decomposition models the mean log count as a combination of an intercept and two random effects, PD and PI, set up as $v_i = \alpha + \mathrm{PD}_i + \mathrm{PI}_i$ where α is the intercept, PD_i and PI_i are random effects for the PD and PI components respectively and v_i is the linear predictor given by $v_i = \log(\mu_i) - \log(E)$ for library depth E and $\log(E)$ is the offset term. The PD component is modelled as a first order random walk $\mathrm{PD}_i - \mathrm{PD}_{i+1} \sim N(0, \tau_{\mathrm{rw}}^{-1})$ where $\mathrm{PD}_i - \mathrm{PD}_{i+1}$ represents the component difference between successive bins and τ_{rw} is the precision of the normally distributed differences, with mean 0. The PI component assumes that bins are represented by a vector of independent and Gaussian distributed random variables with precision τ_{iid}.

Model fitting

The model was fit in the form of a Bayesian mixed model using the software $R-INLA$[40] which implicitly assumes a Gaussian field on the parameter space and uses a Laplacian approximation to allow for fast and deterministic convergence of parameters. Hyperparameters are defined for the size and zero probability parameters (gamma distributed and gaussian dstributed prior distributions respectively) of the negative binomial distribution, and precision parameters τ_{rw} and τ_{iid} (log-gamma distributed priors) for the PD and PI components respectively. Three replicates for each of the ENCODE cell lines were included, assuming library depths equal to the sum of the tags in their respective CTSS files. ENCODE cell lines were modelled con-currently for each chromosome, assuming a common distribution for the hyperparameters.

Model scaling and convergence

In order to allow for efficient prior estimations we scaled the random walk components such that the average variance (measured by the diagonal of the generalized inverse) is equal to 1. In order to achieve efficient convergence and avoid precisely defining priors beforehand, the option `diagonal=1` was set within the INLA call (to avoid falling into sparse errors). The posteriors based on the converged model were then fed into a second model specifying `diagonal=0` in order to achieve more accurate estimates.

Applying the model to RNA-seq data

To see if chromosomal transcriptional decomposition is broadly applicable to RNA datasets, we applied the same modelling procedure to deeply sequenced GM12878 RNA-seq samples (https://www.encodeproject.org/experiments/ENCSR843RJV/)[41]. The libraries were mapped to hg38 using hisat[42], using default parameters, multimappers were removed using samtools[43] and reads were binned at 10kb resolution (based on bins defined previously for CAGE) using bamCoverage[44]. Transcriptional decomposition was applied to the resulting counts, assuming library depth to be the total of the genome-wide bin counts. A range of hyper-parameters for the PD and PI components was tested (corresponding to the CAGE defined values of the parameters, and -3 to +3 relative) and the combination selected, per chromosome, whereby the PD component correlated the most strongly with the PD component in GM12878 CAGE (ignoring regions not within 25 bins of a transcription unit in RNA-seq).

Analysis of PD and PI datasets

Proportions of transcription units represented by PD and PI components

To address the proportions of total mRNA levels allocated to each of the PD and PI components, the raw CAGE counts in GM12878 were used to identify 10kb bin regions with a mean count greater than 50 tags across the three replicates, thus ensuring most selected bins were positive in both components (but removing those which were not). For each of these bins the fraction $\frac{PD}{PD+PI}$ was calculated, where *PD* and *PI* are the respective estimates of the PD and PI components in the bin. These fractions were plotted as a histogram and the median value was identified as an average ballpark for relative allocation to PD component vs the PI component.

ChIP-seq data for ENCODE cell lines

ChIP-seq GRCh38 bigwig and peak data were downloaded from Ensembl FTP and Ensembl biomart based on Ensembl regulatory evidence v 84[45].

Compartment data

Compartment coordinates for GM12878[46] were lifted over from hg19 to hg38 using liftOver tool with default settings. For simplicity of interpretation, the five compartments types from (Rao et al, 2014)[46] were associated with Active (A1,A2), Facultative (B1) and Constitutive (B2,B3) chromatin environments.

Association between PD and chromatin environment

Overlaps between the 10kb bin regions and compartment regions were used to assign bins to chromatin environments, removing bins not having a corresponding overlap. Bins representing shifts between two different chromatin environments were deemed boundary bins and for sets corresponding to each possible shift (e.g. active to constitutive), the mean of the PD component signal was calculated for bins at an increasing distance either side of the boundary bins, in steps of 10kb up to and including 500KB. Cases at a given distance whereby another boundary bin was encountered were removed. For each set, a random sampling was used to generate a list of random boundary bins of length equal to the set size. Background sets were each generated 100 times to form distributions.

Correlations and differences within and between compartments

Intra-chromosomal bin pairs which overlapped with an annotated compartment were assigned an integer according to how many boundaries compartment boundary bins (see above) were between them (boundary insulation). For all bin pairs, both the absolute first order difference in the PD signal and the correlation between the four ENCODE cell lines was calculated. The differences and the correlations were averaged using the median either at each distance from 10kb apart to 2MB apart or across all distances, separately for each possible boundary insulation.

ChIP-seq biases in PD vs PI components

For each ChIP-seq binding mark (DNase1, H2AZ, H3K4me3, H3K27ac), 10kb bin regions were overlapped with binding locations to give the presence or absence of the mark in each bin. For each of the PI and PD components, the bin estimate, first order difference between bin estimates, standard deviation of the bin estimate, stability of the bin estimate (standard deviation across cell type PD component standard deviations) were calculated. For the two components separately, a random forest model was trained with the ChIP-seq binding presence or absence as the response. Out of the bag probabilities from the models based on the PD data and the PI data were compared directly between the two components.

Differentially expressed (DE) bins between cell types in PD and PI components

Posterior estimates and standard deviations of the linear combinations representing the difference in the PD or PI component between GM12878 and HeLa-S3 equivalent bins were generated from the transcriptional decomposition models (described above). Since posterior distributions of the estimates are approximately Gaussian, approximate p-values from z-scores were generated in order to produce standardised scores for the differences. A Benjamini-Hochberg correction was applied according to the number of bins containing an active transcription unit (s.t. there were $\geq$10 tags across the three replicates in at least one of HeLa-S3 or GM12878), and using an FDR < 0.01 to generate a list of significant DE bins in each component.

Bias towards ChIP-seq binding in PD vs PI differentially expressed bins

H3K27me3 and H3K36me3 ENCODE Broad Institute bigwig data (from Ensembl Regulatory Build v 84) were quantified in 10kb genomic bins. The aggregated signal values in each bin were TPM normalised (accroding to all genomic bins). The TPM values for H3K27me3 and H3K36me3 were then inspected at DE bins between GM12878 and HeLa-S3 cells (see above).

Transcription factor targets in PD vs PI differentially expressed bins

Active regions within the DE PD and PI bins were tested for transcription factor binding enrichment using Homer[47]. Regions were defined as actively transcribed FANTOM5 DPI TCs[38], requiring > 1 tpm in at least two of the six libraries (three GM12878 and three HeLa-S3). DPI TCs were extended to regions of -500/+100 around TSSs. DU regions of the PI components with up-regulation in Hela-S3 or in GM12878 (HeLa-S3 PI, GM12878 PI), and DU regions of the PI components with up-regulation in Hela-S3 or in GM12878 (HeLa-S3 PD, GM12878 PD), were tested individually against the universe of actively transcribed TC regions. The tests were performed using a stranded search with GC bias correction. Significantly enriched known motifs (FDR < 0.01) were selected in each of the four tests and plotted together in a heatmap using the *R* package `pheatmap`[48] with Euclidean clustering of transcription factors (columns).

XAD boundaries predictions in ENCODE cell lines

List of TAD boundaries

TAD boundaries regions based on 1kb resolution HIC data in GM12878 were downloaded[46] and lifted over to hg38, requiring a 1:1 correspondence between regions defined in each of the two builds. Boundaries were assigned to 10kb bins based on overlaps. Adjacent bins with boundaries were dealt with by assigning the boundary to the bin with the largest overlap, so that no two adjacent bins contained boundaries, resulting in a total of 14799 distinct bin-sized regions containing TAD boundaries.

GLMnet model for finding features associated with TAD boundaries

Features generated per bin according to those listed in Supplementary Table S4. The set of 10kb bins containing HIC boundaries (see above) was extended either side by 1 bin to supply boundary regions to predict on. Due to lack of CAGE information in non-structural regions, the set of bins was reduced to regions with a potential for a boundary to be predicted by using only the set within 250KB of a bin containing 5 tags or more in GM12878 (replicate sum). The response (presence or absence of a TAD boundary with a bin) was generated in two ways, first for all bins within this set, and secondly under the requirement that the boundary had to sit in a positive random walk region.

In order to assess features which might distinguish bins containing or not containing boundaries, we fit a logistic regression model using the `glmnet`[49] package in *R*. The predictors were scaled before modelling

for generating scores of relative importance, using `caret`[50] package in *R*. To test the performance of the model at predicting boundary regions, a 2-fold cross validation was applied, where the data was randomly split into two equal sized parts and the total performance was assessed according to the combined predictions on the halves of the data which were held out of the modelling after the corresponding half had been trained on. ROC and precision-recall statistics were generated on a per-chromosomal basis, and for all of the chromosomes together, using the `ROCR` package[51] in *R* and plotted using custom functions.

Algorithm for detecting XAD boundaries from CAGE data

The top three features from the generalised linear model outlined in the previous section were selected, namely the PD component (PD), the difference in the PD component (PD_{diff}) and the PD component stability (PD_{stab}). Based on on these three features, the following algorithm was implemented for the detection of XAD boundaries:

1. Calculate $X = \frac{PD|PD_{\text{diff}}|}{PD_{\text{stab}}}$.
2. Calculate local maxima of X and rank in order of largest to smallest values of X.
3. For each chromosome, calculate the proportion of bins of positive PD, p_k, and take the top $\lceil N_k \rceil$ of the ranked maxima such that $N_k = (p_k / \sum_{\text{chr}} p_k)N$, where N is the total target number of XAD boundaries.
4. Split the boundaries into "up jumps" and "down jumps" according to positive and negative values of PD_{diff}, respectively.
5. Shift the "up jumps" right by one bin (to account for the discrepancy in which of the bins on either end of the PD_{diff} should be called the bound. Leave the "down jumps" as is.
6. Return the vector sort(down jumps, up jumps).
7. Repeat above for the ENCODE cell lines.
8. For the final set of bounds in GM12878, choose the set such that the bound is in GM12878 + at least one other ENCODE cell line (and similarly for the set in the other cell types).

The above algorithm was applied to the ENCODE random walks, using the `EMD`[52] package in *R* to generate the local maxima and specifying a target of 5500 XAD. This resulted in a final list of 5109 XAD boundaries after the final filtering step.

Comparing XAD boundaries to ChIP-seq data

XAD boundaries were calculated at bins where the PD component has either a positive or negative change (first order difference in PD component). To generate enrichment of ChIP-seq marks around XAD boundaries, only those with a positive change are considered, in order to avoid biases from signal averaging. The results for the negative gradient boundaries are similar or opposing to that of the positive gradient boundaries.

For each of CTCF, DNaseI, H3K36me4, H3K27me3 and H3K27ac, the list of 10kb bins was overlapped with significantly bound sites to determine how many sites appeared in each bin. The mean number of sites overlapping the bins containing the predicted (positive change) XAD boundaries was calculated, then for bins one away from the boundary, and so forth in steps of 10kb up to a distance of 500KB. Cases at a given distance whereby another boundary bin was encountered were removed to avoid contaminated signal. For the background set, a set of random XAD boundaries of equal cardinality to the real boundaries

were generated, under the conditions of positive gradient and within the set of bins plausible for being predicted as boundaries as previously defined. The same analysis was calculated for CTCF:Rad21, which was based on the number of CTCF sites multiplied by the number of Rad21 sites which overlapped a given bin.

Comparing boundary predictions to TAD boundaries

To calculate the overlap between the XAD set and the TAD set, regions around the XAD set were extended by 5 bins (50KB) on either side and then overlapped with the set of 10kb regions deemed to contain TAD boundaries. The number of TAD boundaries associated with a DHS site which fall within these regions were then counted, together with the number of predicted boundaries which fell in the vicinity of a DHS associated TAD boundary bin (defined as at least one DNase1 site from ChIP-seq overlapping the TAD boundary bin or one or more of the two adjacent bins).

To calculate the enrichment of XAD boundaries at the locations of boundary bins overlapping TAD boundaries, we calculated the proportion bins with predicted boundaries which overlapped the TAD boundary bin set and divided by the same value generated from a random set of boundaries (of the same length as the XAD set, falling within the set of plausible bins). We repeated the randomisation step 100 times and calculated the mean enrichment and standard deviation. The same analysis was repeated according to where the TAD boundary was DHS associated and where the TAD boundary was not associated with a DHS site (defined in the same way as with the overlaps above).

HiC directionality around XAD boundary predictions

Processed intrachromosomal HIC data for GM12878 at 10kb resolution[46] was downloaded and normalised, according to supplied recommendations, using the KR method. The directionality score[53] was implemented and applied to locations of TAD boundaries whose corresponding liftovers in hg38 were either supported or not supported by XAD boundaries (within +/- 5 bins of the boundary). The Directionality score was calculated for +/-25 bins around the location of the boundary bins and for a jump size of 50 bins to calculate the direction bias of interactions. Means were plotted to obtain patterns of global directionality bias.

Modelling of interactions in GM12878

Reprocessing of CaptureHIC data

CaptureHIC data for GM12878 was downloaded and fastq files extracted using the SRA toolkit[54]. The data was mapped and corrected using HICUPS[55], specifying bowtie2[56] for the mapping and GRCh38 for the genome (downloaded from `ftp://ftp.ensembl.org/pub/release-85/`). CHiCAGO[57] was then applied to the remapped BAM files to call significant interactions against each bait, using default settings except for specifying 10kb binning and with the baits defined as the 10kb bin containing the defined sequences from the CaptureHIC protocol[58] lifted over from hg19 to hg38. Score cut-offs of $\geq$3 or $\geq$ 5 were applied to the final output in order to determine which bait-target pairs were considered as interacting, with the rest assigned as non-interacting.

Bait target pairs generation

Binned 10kb regions were overlapped with CaptureHIC protocol defined baits[58], lifted over to hg38. For each such 'bait bin', potential targets within 2MB (200 bins) were assigned and their numbers reduced according to the presence of CAGE tags in more than one replicate in at least one of the ENCODE cell

lines. For analyses based on enhancer associated targets, targets were considered based on overlap with at least one CAGE defined enhancer which is active (more than one tag in more than one replicate) in at least one of the ENCODE cell lines.

Distance cut off

Only bait-target pairs which fall into the distance range of between 6 and 200 bins were considered (corresponding to greater than 50KB and up to 2MB).

Feature generation

The full list of features are listed in Table S5. Features from the PI and PD profiles were added for each bait-target pair separately for the bait and the target (see Table S4 for their descriptions). Enhancer information was added for the bait and targets based on the FANTOM5 enhancer set for hg38 (see above) - including the total eRNA output produced at the enhancer (replicate sum), the number of enhancers deemed active within the bin in the given cell line (at least one tag in at least one replicate) and the number of cell lines supporting the target enhancer (number of ENCODE cell lines with at least one enhancer active within the bin). Bin directionality was calculated based on pooled replicates, using a directionality score[39] and assigning a value of 0 where no tags were present. The boundary insulation between bait and target was calculated according to the number of XAD boundaries observed in the intervening bins. Boundary insulation was defined up to 3; all values greater than 3 were given a score of 3. The peaks detected were the number of CAGE peaks overlapping the bait or target bin, according the hg38 DPI set from FANTOM5, which were transcribed in that cell type (at least one tag in at least one replicate). The cross correlation for the PI and PD components was calculated per chromosome by correlating the respective PI and PD profiles over four ENCODE cell lines between the bait and target bins, specifying Kendall for the correlation method. The first eigenvector was calculated from the cross correlation matrices for each chromosome and its value supplied for the bait and target bins. Distance was defined as the number of 10kb bins separating the bait and the target.

Dataset generation

Two datasets from GM12878 were analysed. For assessing the applicability of the model trained in GM12878 for predicting interactions in other cell types, the dataset as generated was kept in its current format (termed 'raw' format). In addition to this, since the number of positive interactions decays sharply with larger distances, a second dataset was generated where the ratio of negative to positive cases over distance was balanced by randomly sampling 20 negatives to each positive in the dataset, with replacement to account for cases where the negative rows were not at least 20 fold in number to the positive rows.

Data balancing for training sets

To generated a balanced dataset for training, SMOTE, as implemented in the `unbalanced`[59] package in *R*, was applied to the data, specifying parameters `percOver=200` and `percUnder=150` to genere new positives together with under sampling the negatives to achieve a balance of 1:1 in the dataset. In order to balance the data set most fairly over distances, SMOTE was applied separately across each possible bait-target separation.

10 fold cross validation procedure

The model was implemented in R using the `randomForest`[60] package supplemented with `foreach`[61] and `doParallel`[62] to run on multiple cores. We used 10 fold cross validation, whereby we split the dataset randomly into 10 equal sized pieces and held out a single piece as a testing set on each of the

10 runs of the model which was trained on the remaining 90%. All training was carried out based on a ChiCAGO score cut-off of 3. All performance statistics and probability estimates in GM12878 are based on predictions made across the held out runs over the full data set.

Splitting into short, medium and long models for feature importance

To assess whether there are different features which are important for specific distances without the bias of most examples being weighted towards short distances, the above analysis outline was repeated for the same data set but restricted to three possible distances: (50KB,250KB], (250KB,500KB] and (500KB,2MB]. For each set of distances, the MDA was averaged across the 10 runs in order to obtain a final feature performance.

Most efficient cut off for assessing predictions

To find the optimal probability cut-off for calling a predicted interaction, the value for which the F1 statistic was maximised was calculated using `optim` in *R*, according to the desired score cut-off. Since the most efficient cut-off is not fixed according to distance, the F1 statistic was maximised separately for five sets of bait-target distances: (50kb,100kb], (100kb,250kb], (250kb,500kb], (500kb,1MB], (1MB,2MB], and performance analysed for predictions generated using the resulting cut-offs. To calculate the effect of the CHiCAGO score cut-off on precision and recall, we optimised the F1 statistic separately for the five sets based on a range of score cut-offs (0.1, 0.25, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6).

Assessing model performance

We used the `pROC`[63] package in *R* to generate AUC statistics and the `caret`[50] package in *R* to generate the precision, recall and F1 statistics. Plots were generated using custom functions based on statistics generated from the `ROCR`[51] package.

Generating predictions for HeLa-S3 and HepG2

The dataset for GM12878 was trained as described above against a score of ≥ 3 and used to predict on the equivalent set of bin pairs in HeLa-S3 and HepG2, which were subsequently reduced to those with enhancer targets only (using the same criteria as described above). Final probabilities were calculated based on the mean of probabilities over the 10 runs. The distance based F1-maximising cut-offs were applied to obtain a final list of interactions. Enhancer-promoter interaction sharing between GM12878, HeLa-S3 and HepG2 was calculated based on whether the interactions were present in 1, 2 or 3 of the cell types and venn diagrams were generated using the *R* package `VennDiagram`[64].

Example domains predicted in HeLa-S3 but not GM12878

Processed PolII ChIA-PET interactions data for HeLa-S3 was downloaded from ENCODE[41] and start and end regions were lifted over to hg38, removing those without a 1:1 correspondence between the two builds. Bait-target pairs for predicted enhancer-promoter interactions in HeLa-S3 were selected according to a probability of at least 0.6 in HeLa-S3 and a probability of less than 0.4 in GM12878. Both ends were intersected with the lifted over start and end regions in the ChIA-PET data in order to generate a list of candidate examples. We selected an example on chromosome 6 due that chromosomes high performance statistics from the modelling described above. The *R* package `Sushi`[65] was used to generate plots of the resulting loops and lines together with annotations and enhancer locations for the data sets in the example region.

Analysis and predictions across 76 cell types from FANTOM5

Library selection and tanscriptional component generation

A total of 249 CAGE libraries FANTOM5 were selected according to the availability of sample replicates. Transcriptional decomposition was applied to generate PD and PI components for a total of 76 cell types (including 4 ENCODE cell lines and 72 primary cells - see Table S10 for list of library identifiers and names). For the purposes of consistency between the ENCODE generated data sets and the primary cell type generated data sets, the hyperparameters were fixed for the random walk and independent components to the same values which were generated from the models in ENCODE (whilst allowing the hyperparameters for the zero inflated negative binomial distribution to vary).

Hierarchical clustering of raw data samples

Raw binned data at 10kb resolution was normalised into tags per million (tpm) and the mean was taken across replicates to obtain a matrix of 76 columns against the total genomic bin count (303,065). Only regions potentially transcribed in the given set of CAGE assays were considered by asking for bins which had more than one cell type containing tags. The matrix was transformed into log_{10} values (adding a pseudo-count of 1) and hierarchical clustering using `hclust` was applied in *R*. The ordering from the clustering was used to guide the row and column ordering in the heatmaps for Figure 5. The function `cutree` was used to find 10 groups from the clustering, which were annotated and merged manually to generate the most biologically relevant cell type groupings from the data.

Heatmaps for comparison PD and PI components

A common set of bins (34,953) was derived for comparison between cell types of the PD and the PI components by choosing the set of bins where the sign of the PI component was positive in more than one cell type. To generate similarity matrices, the PD and PI signals were converted into binary according to whether the signal was positive (1) or negative (0) (PI - 0.1 was used for the independent component to avoid non-expressed bins with very small positive estimates). The resultant matrix of 76 columns and rows according to the common bin set was used to calculate 1 - L_1-norm between each pair of cell types to calculate a similarity matrix. The same metric to cell type boundary sharing and cell type enhancer promoter sharing, with methods described in the sections below. The *R* package `pheatmap`[48] was used to generate cell type group annotated heatmaps of the similarity matrices.

Extending boundary predictions for 76 cell types from FANTOM5

For all cell types, XAD boundaries were calculated from the algorithm described above for ENCODE datasets, supplying the stability scores across the full set of cell types. To calculate boundary sharing across all cell types, non-overlapping bins of size of 100kb were generated and calculated for each expanded bin the number of cell types within which a boundary was found. In order to cover all possible windows the expanded bins were shifted by a 10kb bin 10 times and the final number of shared boundaries was calculated according to an average of the 10.

Predicting interactions for 76 cell types from FANTOM5

Datasets with features were generated as described for the ENCODE cell lines for all cell types, with the bin selection also extended to the full set, thus creating a common bin set which is larger than that for the ENCODE cell lines alone. Features non-specific to a cell type were also calculated more broadly with consideration to the 76 cell types. The model was trained as above using the GM12878 dataset using the broader bin set. Similar model performance was noted or this model when testing on the raw dataset

using 10-fold cross validation. The held out set predicted probabilities were used for GM12878 and the predicted probabilities for the other 75 cell types were generated by averaging over 10 trained models across the whole dataset (to robustly account for random differences in the data balancing for the training).

Analysis of interaction data across samples

The datasets for the 76 cell types were reduced according to whether at least one dataset had an active CAGE enhancer annotated to it, in order to obtain a list of potential enhancer promoter interactions. To generate lists of predicted interactions, we generated distance based probability cut-offs in GM12878, using the same method for the ENCODE data sets above, using score cut-offs ≥ 3 and ≥ 5.

To calculate enhancer-promoter interaction sharing statistics, the ≥ 3 score cut off was used and for each possible number of cell types (from 1 to 76) the number of significant interactions which were present in exactly that number of cell types was calculated. To generate the enhancer-promoter sharing heatmap, all interactions which were present in more than one cell type were considered and pairwise cell type similarity calculated from (1 - L_1-norm) between columns of the binary (1 predicted / 0 not predicted) matrix with cell types as columns and interaction set as rows. This generated a similarity matrix which we plotted using `pheatmap`[48] based on the cell type ordering from the hierarchical clustering in the raw data.

Blood specific EP interactions

To isolate examples of interactions specific to blood, interactions meeting the criteria of present (≥ 3 score distance based efficient cut offs) in at least half of the cell types labelled as 'blood cells' (see Supplementary Table S11) and in fewer than a total of up to a maximum 3 more than the number of blood cells within the 76 cell types were classified as 'blood specific'. For further trimming of examples and plotting of loops, probabilities assigned to all cell types assigned as blood cells were averaged into one using the mean. Merged probability vectors were also generated for endothelial cells, epithelial cells, muscle cells, mesenchymal and fibroblasts. Examples of blood specific interactions were selected based on the loop being still significant according to the averaged probabilities. The *R* package `Sushi`[65] was used to generate plots of selected examples, including tracks for annotations and the averaged PD, PI components for blood cells.

Resource generation of 76 cell types from FANTOM5

Values for the PD, PI components, together with locations of XAD boundaries and predicted enhancer-promoter interactions were saved out as BED files, using UCSC zero-based coordinates, and supplied as supplementary data files.

References

33. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2016). URL `https://www.R-project.org/`.

34. Kent, W. J. et al. The human genome browser at UCSC. Genome research **12**, 996–1006 (2002).

35. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an r package for interfacing with genome browsers. Bioinformatics **25**, 1841–1842 (2009).

36. Lawrence, M. et al. Software for computing and annotating genomic ranges. PLoS Computational Biology **9** (2013).

37. FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. A promoter-level mammalian expression atlas. Nature **507**, 462–470 (2014).

38. Kawaji, H. Transcription initiation peaks based on fantom5 cage data on hg38 and mm10 [data set]. Zenodo (2017).

39. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. Nature **507**, 455–461 (2014).

40. Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the royal statistical ... (2009).

41. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature **489**, 57–74 (2012).

42. Kim, D., Langmead, B. & Salzberg, S. L. Hisat: a fast spliced aligner with low memory requirements. Nature methods **12**, 357–360 (2015).

43. Li, H. et al. The sequence alignment/map format and samtools. Bioinformatics **25**, 2078–2079 (2009).

44. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deeptools: a flexible platform for exploring deep-sequencing data. Nucleic acids research **42**, W187–W191 (2014).

45. Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The Ensembl Regulatory Build. Genome Biology **16**, 56 (2015).

46. Rao, S. S. P. et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell **159**, 1665–1680 (2014).

47. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. Molecular cell **38**, 576–589 (2010).

48. Kolde, R. pheatmap: Pretty Heatmaps (2015). URL https://CRAN.R-project.org/package=pheatmap. R package version 1.0.8.

49. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software **33**, 1 (2010).

50. from Jed Wing, M. K. C. et al. caret: Classification and Regression Training (2016). URL https://CRAN.R-project.org/package=caret. R package version 6.0-73.

51. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. Rocr: visualizing classifier performance in r. Bioinformatics **21**, 7881 (2005). URL http://rocr.bioinf.mpi-sb.mpg.de.

52. Kim, D. & Oh, H.-S. Emd: a package for empirical mode decomposition and hilbert spectrum. The R Journal **1**, 40–46 (2009).

53. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature **485**, 376–380 (2012).

54. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. Nucleic acids research gkq1019 (2010).

55. Wingett, S. et al. Hicup: pipeline for mapping and processing hi-c data. F1000Research **4** (2015).

56. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with bowtie 2. Nature methods **9**, 357–359 (2012).

57. Cairns, J. et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biology **17**, 1 (2016).

58. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nature genetics **47**, 598–606 (2015).

59. Pozzolo, A. D., Caelen, O. & Bontempi, G. unbalanced: Racing for Unbalanced Methods Selection (2015). R package version 2.0.

60. Liaw, A. & Wiener, M. Classification and regression by randomforest. R News **2**, 18–22 (2002).

61. Analytics, R. & Weston, S. foreach: Provides Foreach Looping Construct for R (2015). R package version 1.4.3.

62. Analytics, R. & Weston, S. doParallel: Foreach Parallel Adaptor for the 'parallel' Package (2015). R package version 1.0.10.

63. Robin, X. et al. proc: an open-source package for r and s+ to analyze and compare roc curves. BMC Bioinformatics **12**, 77 (2011).

64. Chen, H. VennDiagram: Generate High-Resolution Venn and Euler Plots (2016). URL https://CRAN.R-project.org/package=VennDiagram. R package version 1.6.17.

65. Phanstiel, D. H. Sushi: Tools for visualizing genomics data (2015). R package version 1.10.0.