

Identification of gene specific cis-regulatory elements during differentiation of mouse embryonic stem cells: An integrative approach using high-throughput datasets.

M S Vijayabaskar^{1§}, Debbie K Goode², Nadine Obier³, Monika Lichtinger^{3§}, Amber M.L Emmett¹, Fatin Z Abidin^{1‡}, Nisar Shar^{1#}, Rebecca Hannah², Salam A Assi³, Michael Lie-A-Ling⁴, Berthold Gottgens^{2*}, Georges Lacaud^{4*}, Valerie Kouskoff^{5*}, Constanze Bonifer^{3*} and David R Westhead^{1*}.

¹ School of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, LS2 9JT, United Kingdom

[§] Wellcome Sanger Institute, Wellcome Genome Campus, CB10 1SA, United Kingdom

[‡] Institute of Systems Biology (INBIOSIS), Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor D.E., Malaysia

[#] Department of Biomedical Engineering, NED University of Engineering & Technology, University Road, Karachi 75270, Pakistan.

² Wellcome Trust & MRC Cambridge Stem Cell Institute and Cambridge Institute for Medical Research, University of Cambridge, Hills Road, Cambridge CB2 0XY, United Kingdom

³ Institute for Cancer and Genomic Sciences, College of Medical and Dental Sciences

University of Birmingham. Birmingham B15 2TT, United Kingdom

⁴ CRUK Manchester Institute, University of Manchester, Manchester M20 4BX, UK

⁵ Division of Developmental Biology and Medicine, The University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT, United Kingdom

*To whom correspondence should be addressed. Tel: 00441133433116 Email: D.R.Westhead@leeds.ac.uk.

Abstract

Gene expression governs cell fate, and is regulated via a complex interplay of transcription factors and molecules that change chromatin structure. Advances in sequencing-based assays have enabled investigation of these processes genome-wide, leading to large datasets that combine information on the dynamics of gene expression, transcription factor binding and chromatin structure as cells differentiate. While numerous studies focus on the effects of these features on broader gene regulation, less work has been done on the mechanisms of gene-specific transcriptional control. In this study, we have focussed on the latter by integrating gene expression data for the *in vitro* differentiation of murine ES cells to macrophages and cardiomyocytes, with dynamic data on chromatin structure, epigenetics and transcription factor binding. Combining a novel strategy to identify communities of related control elements with a penalized regression approach, we developed individual models to identify the potential control elements predictive of the expression of each gene. Our models were compared to an existing method and evaluated using the existing literature and new experimental data using embryonic stem cell differentiation reporter assays. Our method is able to identify transcriptional control elements in a gene specific manner that reflect known regulatory relationships and to generate useful hypotheses for further testing.

Author summary

The inherited information in our DNA genomes is a code which defines both the functional units (proteins, nucleic acids etc.), and patterns of their usage, necessary to make life. The genome in mammals, such as man and mouse, has genes which code for about 20000 different proteins, but the usage of these proteins differs in each different type of cell within these complex multicellular organisms. How this differential usage is controlled is known as genetic regulation, and that is what we study here. We know that the details lie in how genes are turned on and off, but until the advent of high-throughput sequencing technology a genome-wide study was nearly impossible. Further complicating our efforts to understand genetic regulation is the involvement of parts of the genome that were previously deemed junk. In this work, we have focussed on how the genes are controlled at various developmental stages in mouse, by looking at the sequencing data from different regulatory mechanisms such as protein binding and local changes to DNA packaging etc. On a gene-by-gene basis, we have built statistical models that predict how genes are controlled when cells develop. These predictions provide a focus for future experimental studies of genetic regulation.

Introduction

The fate of a cell is determined by dynamics in the expression of genes, a process that is regulated at the highest level by the control of transcription [1, 2]. Cell differentiation at the genome level involves a complex interplay of processes [3], including DNA binding by transcription factors (TFs) [4] and changes in the structure of chromatin and DNA, for example, DNA methylation and epigenetic modifications of amino acids of the histones [5]. With the recent developments in high throughput sequencing (HTS) researchers have been able to study the genome-wide implications of these processes in various cell types and organisms [6]. From these studies, we have gained global insights into transcriptional regulation, such as the relationship between chromatin accessibility around the promoter region and gene expression [7, 8], the prevalence of histone modifications such as H3K27ac and H3K9ac near expressed genes [9], the presence of H3K27me3 modification near transcriptionally repressed genes [10] and the binding of master regulators to genes that are often associated with lineage differentiation [11, 12].

While patterns in regulatory mechanisms have been identified, much of the detail of the regulatory system still remains unknown, owing to the complex structure of the transcription machinery. Of note is the implication of non-coding regions of the genome of higher eukaryotes in transcriptional and post-transcriptional regulation of protein coding genes [4, 13]. Equally, it is becoming increasingly clear that much of the dynamics in gene expression is governed not by the regulatory input at promoters but to distal sites including enhancers [14, 15]. Enhancers have been a subject of interest in recent years because of their major role in transcriptional control of tissue specific gene expression programmes [14, 16, 17]. There are several proposed models for the mechanism of enhancer interaction with the RNA Polymerase II (RNAP) machinery over large distances to control gene expression [18], whereby the looping model is becoming widely accepted [19]. All the models propose binding of TFs to both enhancer and promoter regions and involve formation of multi-protein complexes. This aspect of combinatorial binding of TFs to control regions is well established [12, 20, 21].

Enhancer regions may be identified experimentally by reductionist approaches such as mutations in genomic regions associated with loss of expression of a nearby gene, and by using reporter assays to investigate transcriptional enhancement in cell lines or *in vivo*. There has also been significant interest in the identification of enhancers using theoretical methods. These can be based on characteristics of the DNA sequence, particularly conservation, and the large-scale availability of ChIP-seq data for well-known enhancer characteristics (the H3K27ac and H3K4me1 chromatin marks, binding of the EP300 coactivator protein etc.) has led to more sophisticated methods using machine learning to combine information from different sources. These efforts have been recently reviewed [2, 22-24], and it has been shown that binding by transcription factors is a highly specific indicator of enhancer activity.

Extending the problem of identification is the issue of mapping those enhancer elements to the genes they regulate. It is now widely accepted that physical enhancer-promoter (EP) interactions are required for transcriptional control [25, 26]. A range of related experimental techniques (3C, 4C, HiC, ChIA-PET, capture C) classed as chromosome conformational capture assays enable us to determine the three-dimensional conformations of specified regions or the whole of the genome [26] and EP interactions are likely to be a subset of the numerous chromosomal interactions identified [4, 27]. These techniques are being rapidly developed and the increasing availability of data sets, particularly in human cells, is leading to interesting insights [28, 29].

The enhancer mapping problem has also received theoretical attention, and extending earlier work [30-35] based on epigenetic activity correlations between enhancers and their putative target genes, more recent studies have taken advantage of the increasing availability of experimental data for three-dimensional genome interactions [36, 37]. For human cells, projects like ENCODE [38] have made

available genome scale data on chromatin structure, transcription factor binding and gene expression, enabling the development of methods to predict EP interactions from them, with training and validation employing appropriate 3D interaction data [30-32, 39]. For example, RIPPLE uses a Random Forest classifier with 5C data and a selected set of highly predictive features for identification, and characterisation of enhancers, and further uses an ensemble approach for cell lines with no 5C information. Recently, Cao et al. [40] developed JEME which considers the joint effects of multiple predicted enhancers on gene expression whereas earlier similar methods based predictions on different features [37, 41, 42]. Recently, Xi and Beer [43] have raised concerns about the validation and estimated accuracy of machine learning methods that rely on 3D interactions such as Hi-C to predict enhancers. They have focussed on Target Finder [42] which uses Hi-C data, and while encouraging results have been reported, they advocate proper cross-validation method since the coarser resolution of Hi-C data can lead to groups of potential enhancers all exhibiting the same pattern of promoter interactions. Alternative approaches have employed the correlation of enhancer activity and bidirectional enhancer transcription [44]. In other model organisms such as *Drosophila*, the emphasis has been on sequence features of the predicted enhancers linking to developmental gene expression patterns [45-47].

In this study, we address the related problem of identifying those gene specific cis control elements, many of which are expected to be enhancers, which are most relevant in controlling the expression of key genes involved in cellular lineage specification and progressive lineage restriction. Our focus is a data set from mouse, where three-dimensional interaction data are not available in most of the cell types. This method integrates chromatin immunoprecipitation, DNaseI-seq and RNA-seq data, and analyse two branching pathways of *in vitro* cellular differentiation, leading from embryonic stem (ES) cells to the myeloid blood lineage (macrophages) in one branch and to cardiac cells (cardiomyocytes) in the other [11, 25]. Our method is based on correlating a measure of the activity of a candidate cis-regulatory element (CRE), as indicated by transcription factor occupancy and chromatin structure/modification, with the pattern of gene expression. We introduce a new concept of cis-regulatory element communities (coCREs), which are genomic regions that show correlated patterns of activity and transcription factor binding and are considered together for robust model building. We compare the observations with known and predicted sets of enhancers and furthermore with the results from JEME trained on 3D interaction data from mouse embryonic stem cells.

Our study indicates a gene-specific coordinated binding of multiple master regulators during differentiation to control lineage specific expression of important genes, implying that the gene-centric approach may shed further light on the relationship between cell fate decisions and the underlying transcriptional landscape. We were able to successfully recapitulate regulation of cell type specific genes using previously known cis elements and also have been able to propose gene specific control mechanisms using novel regulatory regions identified through gene-specific expression modelling. Furthermore, we found that the genomic loci that are most predictive were characterised by high phylogenetic conservation, epigenomic activities and transcription factor binding. Finally, using reporter assays we were able to confirm two novel cis regulatory elements for genes *Nfe2* and *Sptbn1*. This statistical approach that works with fewer assays and does not require topological data, is simple to apply, adapt and understand than machine learning approaches. Furthermore, by employing *in vitro* cell types and focussing on mouse, an important model organism for which relatively less information on enhancers are available, the data provided here is a valuable resource to the research community.

Results and discussion

In order to study the mechanism of transcriptional control of genes involved in lineage specific cell fate decisions, we integrated gene expression and regulatory data from the two *in vitro* mouse differentiation pathways described above. Details of the associated cell types and data sets employed are given in Table 1. The results of integrating normalised expression data for protein coding transcripts are shown in

Figure S1(A-C): both clustering and principal components analysis indicate effective integration where related cell types from both series (ES cells, ESC and CESC; mesoderm cells, MES and CMES) show clear similarity, and terminally differentiated macrophages emerge as the most different cell type, nevertheless still showing clear lineage development from the intermediate haemopoietic cells. From this data, we identified 9854 differentially expressed genes (for details see Methods), and their patterns of expression were clustered using *k*-means to produce 391 clusters at the optimal BIC (Bayesian Information Criterion) score (Figure S1D). For the purpose of predictive modelling we considered 3 sets of genes: the 17 TFs for which we have ChIP-seq data (the ‘TF set’), the 437 genes that co-cluster with the TF set (‘TF cluster set’) and finally the remaining differentially expressed genes (‘DE gene set’). Details of the gene sets and clustering are provided in Table S1 and Figure S2.

Table 1: Datasets and cell types used in this study

Pathway	Cell type	Description	ChIP-seq (H3K27Ac)	DNase1-seq	RNA-seq	ChIP-seq (TF)
mESC → Macrophages	ESC	ES cells	✓	✓	✓	NANOG, OCT4, SOX2, ESRRB
	MES	Mesoderm	✓	✓	✓	C/EBPB, ELK4, OCT4
	HB	Hemangioblast	✓	✓	✓	C/EBPB, GATA2, LMO2, TAL1
	HE	Hemogenic Endothelium	✓	✓	✓	C/EBPB, LMO2, TAL1, FLI1, MEIS1
	HP	Hematopoietic Progenitors	✓	✓	✓	C/EBPB, FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, PU.1, RUNX1, TAL1
	MAC	Macrophages	✓	✓	✓	C/EBPB, FLI1, LMO2, PU.1, RUNX1, TAL1
mESC → Cardio myocyt	CESC	ES cells	✓	✗	✓	✗
	CMES	Mesoderm	✓	✗	✓	✗

	CP	Cardiac Progenitors	✓	✗	✓	✗
	CM	Cardiomyocytes	✓	✗	✓	✗

The essence of the method presented here is to identify for each gene studied, a small set of CREs that best explain its pattern of expression (the ‘Gene Expression Profile’ (GEP)) from a relatively large set of candidate CREs. The methodology is illustrated in Figure 1A and full details can be found below and in Methods. In summary, for each gene, initial candidate CREs are genomic regions within 100 kBases (kB) of the transcription start site (TSS) with significant enrichment in H3K27ac ChIP-seq (a mark of active enhancers) data and/or a DNaseI-seq hypersensitive site (DHS). Candidate CREs are characterised by a chromatin activity profile (CAP, see Methods) across the cell types, derived from H3K27ac and DNaseI-seq data, and also a TF binding profile (TFBP). This information is integrated into a CRE network, from which community CREs (coCREs) are identified (see Methods). The CAP of a coCRE is the average of its constituent CREs. A gene-specific penalised regression is used to choose CREs that best predict the gene’s expression profile (GEP) from the set of singleton and community CREs.

Figure 1: Gene-specific predictive models. (A) Schematic representation of the methodology involved in developing gene specific predictive models. 1. Integration of DNaseI-seq and H3K27ac to quantify the chromatin activity profile (CAP) in candidate cis regulatory elements (CREs). TF ChIP-seq data is used to generate the transcription factor binding profile (TFBP) to quantify the community effect of candidate CREs mapped to a specific gene. 2. Gene wise expression values are obtained as RPKMs to form gene expression profiles (GEPs). 3. CAPs, TFBPs and GEPs are generated for all the regions and genes in the analysis. 4. CAP and TFBP are integrated in order to generate gene specific CRE networks. A greedy community detection is performed in order to identify the communities of CREs (coCREs) in the networks. A new set of CAPs involving aggregate CAPs of the coCREs along with the individual CAPs for singleton CREs are used to predict the GEP for a specific gene. (B) Histogram showing the distribution of candidate CREs per gene within 100kB of the transcription start site over all genes in the study. (C) The plot shows the change in cross-validated Mean Squared Error (MSE) as a function of increasing λ for a predictive model of *Runx1* gene expression. The two vertical dotted lines show the two cut offs λ_{min} and λ_{1se} . The total number of CREs with non-zero coefficients (β) at a given λ is shown above the plot.

Active chromatin regions as candidate CREs

The 262770 chromatin regions that are active (either H3K27ac enriched or a DNaseI Hypersensitive Site (DHS), see Methods sections), form our candidate CREs and are characterised in Figure S3. A ‘chromatin event’ in a CRE is a H3K27ac peak or DHS peak in a particular cell type, and Figure S3A is a frequency plot for chromatin events in all candidate CREs. Many CREs have only one chromatin event over all cell types: in order to use only high confidence CREs we retained only those with more than one chromatin event (118688 regions given in Table S2), and with enrichment level (either H3K27ac or DNaseI-seq) in at least one cell type higher than the lower quartile in the distribution of enrichment levels (Figure S3D and E, 82165 regions). Figure S3B shows that the sizes of the CREs used in this study are around 1-2 kB in width and Figure S3C shows the distribution of their conservation scores. It can be seen that majority of CREs have low conservation across the vertebrates, while a selective subset of regions is highly conserved. From this initial set of CREs, coCREs are identified as given below.

Community CREs (coCREs)

Figure 1B shows that genes may have many candidate CREs, ranging from zero to 100 with an average of around 18. This exemplifies the challenge of gene specific models that the number of potential predictors of gene expression exceeds the number of measurements, and calls for specific methods such as penalized regression (see below). However, preliminary data analysis revealed subsets of candidate CREs with highly correlated chromatin activity and TF binding profiles, and we developed a method to combine these first prior to model building. This served two purposes pertinent to model development; a technical complexity, where multiple correlated predictors can pose problems for penalized methods and a biological perspective, where correlated regions could represent collaborative regulation and could possibly be interacting in three-dimensional chromatin structure. Two candidate CREs were linked in a network graph (Figure 1A, step 4) if they had high correlations in both their chromatin activity and TF binding profiles, and from this graph we identified subsets of candidate CREs that form communities (dense sub-graphs), identifying these as coCREs (see Methods). The chromatin activity profile (CAP) of a coCRE is the mean of the CAPs of the community members. The coCREs along with the CREs near a gene were used to build a model predictive of the gene's GEP. The models based on coCREs performed better than models with all CREs used separately, in terms of drop-in-variances ($p \sim 10^{-16}$, see methods section) and model p-values ($p \sim 10^{-13}$) (Figure S3F), and this methodology was retained for the remainder of the study.

Gene specific predictive models

In order to generate gene-specific predictive models the candidate CREs within 100 kb of the TSS of a given gene were considered, and were reduced to a set of coCREs and singleton CREs (those not in any community). To affect gene expression, distal CREs such as enhancers need to interact with protein complexes nearer to the TSS, therefore singleton CREs were reduced to those proximal (within 20kb) to the TSS, limiting distal CREs to those that occur in communities (coCREs). The CAPs of this mixed set were considered as the initial predictors in the model building. Since the total number of candidate CREs mapped to genes are generally more than the number of cell types used (Figure 1B), we have used a penalized regression model (LASSO) [48, 49], which employs an additive penalty term with weight λ on the sum of the absolute sizes of the regression coefficients. Appropriate λ values were determined by cross-validation (Figure 1C) and two models were computed for each gene at $\lambda = \lambda_{min}$ (minimum cross-validated mean square error) and at the more conservative (fewer non-zero regression coefficients) value obtained by adding one standard error $\lambda = \lambda_{1se}$ [50]. The CREs or coCREs with non-zero regression coefficients β , were deemed to be the most predictive of the gene expression by the model (λ_{1se}) and were termed 'chosen CREs' in this study. Statistical significance in LASSO models is an open area of research [48-51], since inference must account for the fact that the method sequentially chooses the most predictive variables from a set of candidates. We adopted the covariance test [51] as a means to give a p value to each non-zero coefficient in the model.

An example of a predictive model for *Runx1* is given in Figure 2 (see Figure 1C for the cross-validation plot). *Runx1* is a gene encoding a transcription factor that is crucial for normal haematopoiesis [52] and is expressed during the later stages of the hematopoietic specification (Figure 2B). A set of 21 CREs and 3 coCREs were considered as the initial predictors, and two regions with non-zero β at λ_{1se} (Figure 2 A and D) were selected as chosen CREs by the model. One of these is a coCRE, a community of four CREs bound by a range of blood specific TFs in different haemopoietic cell types (Figure 2A) and the other a singleton. The expression that is predicted by the model correlates well with the predicted gene expression profile for *Runx1* as shown in Figure 2C. The covariance test gives p values of 0.00 and 0.42 (Table S3) for the chosen coCRE and singleton CRE above, showing strong evidence for a relationship between activity at the coCRE and gene expression.

Figure 2: Predictive model for an example gene, *Runx1*. (A) A network representation of the model, where the gene (here *Runx1*) for which the model is built (red octagon), chosen CREs (blue hexagons) and TFs bound to the chosen CREs (grey squares) are represented as nodes. Black arrows indicate the regulation of the gene by the CRE/coCRE and coloured arrows represent the binding of TFs to the CREs in different cell types. The colours corresponding to the cell types are given below the network. The TFBP of the CRE in a specific cell type is represented as a circular histogram and in the case of coCREs these represent the frequency of occurrence of a specific TF in the regions of that community (here the community comprises of 4 regions). The p -value of observing a combinatorial binding profile in that cell type is provided for each TFBP node and the methodology is given in Methods section. The abbreviations for the TFs in the circular histogram are: Esrrb (EB), Nanog (NG), Pou5f1 (O4), Sox2 (S2), Cebpb (CB), Elk4 (E4), Gata2 (G2), Lmo2 (L2), Tal1 (T1), Fli1 (F1), Tead4 (T4), Meis1 (M1), Gata1 (GA1), Gfi1 (G1), Gfi1b (GB), Runx1 (R1), Spi1 (P1). It should be noted that not all TFs in the circular histogram have supporting ChIP-seq data in all cell types (Table 1). In the absence of ChIP-seq data for a specific cell type, the bar for that TF in the histogram of that cell type is zero. (B) The gene expression profile (GEP) of *Runx1* with cell types along the horizontal axis and FPKM on the vertical axis. (C) The plot shows the best linear fit between the actual (X) and predicted (Y) GEP for *Runx1*. The spearman correlation coefficient is also provided. (D) The plot shows the tag density profile normalised as coverage per million aligned reads for the 10 cell types. *Runx1* gene structure is provided in blue below the coverage tracks. The predictor CREs that were used in the lasso model are given as grey boxes and the chosen CRE and the coCRE are given in red and yellow respectively. The super enhancers (SE) identified by Whyte et al.[53] are given as green bars and the enhancers given by SEA is in blue. The experimental enhancers identified by Schütte et al. and Dogan et al. are provided as well. In the case of *Runx1* there is no overlap with the Dogan et al. dataset, and hence the absence of any bars. It should be noted that the coCRE enhancer is represented as a composite of red boxes of member CREs.

Following the example of *Runx1*, models were generated for each member of the TF set. Full details can be found in Table 2 and Table S3, and details of the chosen (co)CREs for the TF set are shown in Figure S4. We chose to consider a model as potentially interesting if it had a least one non-zero β with adjusted Benjamini-Hochberg q -value based on the covariance test of < 0.05 . Focusing first on the TF set, models were successfully built with at least one predicted (co)CRE for 14 of the 17 TFs and 9 of these were considered interesting by this statistical criterion.

Table 2: Generation of models

Gene set	Number of genes	Number of models with at least one $\beta \neq 0$	Number statistically significant ¹	Number of coCREs with $\beta \neq 0$	Number of singleton CREs with $\beta \neq 0$
TF set	17	14 (82%)	9 (53%)	34	22
TF cluster set	437	340 (78%)	226 (52%)	163	282
DE gene set	9854	6715 (68%)	3212 (30%)	2592	5567

¹A model is considered statistically significant if it has at least one regression coefficient β with Benjamini-Hochberg q value < 0.05 . The details of the model parameters and the CREs are given in Table S3.

Considering the TFs associated with pluripotency (see Figure S4), *Nanog* is mapped to four singleton CREs and two coCREs, but only the coCREs are statistically significant. The coCREs are located on either side of the *Nanog* promoter with one coCRE (a community of 4) overlapping with its TSS and

bound by all the four pluripotent TFs. *Pou5f1* that codes for OCT4 is mapped to one coCRE that includes the promoter region and a downstream region, and is bound to pluripotent TFs in ESC. Three CREs are mapped the *Sox2* gene, two of which are coCREs comprising 2 and 3 regions respectively. One of the regions (the community of 3 CREs), although in this case not statistically significant, is almost 75kB away from the promoter, bound by the OCT4/NANOG/ESRRB TFs on all the three sites and was identified as a super-enhancer by Whyte et al. [53].

Regarding TFs specific to the hematopoietic lineage, we commented on the chosen CREs for *Runx1* above (Figure 2A). The coCRE is related to the +23kb region that is experimentally verified [54]. It is bound mostly by LMO2 in HB, TAL1 in HE and the LMO2/TAL1/FLI1/GATA1/PU.1 complex in HP and the increasing recruitment of TFs and the increasing significances of the combinatorial binding closely follows its expression profile (Figure 2A). *Tal1* has a single coCRE encompassing the promoter region and two downstream regions, separated by around 30kB [15]. The coCRE is primarily bound by LMO2/TAL1/GATA2/TEAD4 in HB, LMO2/TAL1/FLI1 in HE, FLI1/PU.1 in HP and PU.1 in macrophages (Figure S4). *Gata2* has only one CRE mapped to it and it overlaps with its promoter region and is bound by LMO2/TAL1/FLI1 in HE and HP. It should be noted that although this locus fails to attain statistical significance (Figure S4 and Table S3) it is experimentally well-characterised [55]. *Lmo2* is mapped to a CRE lying on the proximal promoter [56] and the other coCRE bound predominantly by PU.1/CEBP β in macrophages lying on the distal promoter region. *Spi1* (PU.1) is a macrophage specific TF that is mapped to two coCREs upstream of its TSS, with one coCRE overlapping with its promoter and is bound by LMO2/FLI1/CEBP β in HP and with LMO2 possibly replaced by PU.1 in MAC. This coCRE comprises of the promoter and a -17kB upstream region, which has been shown to be involved in *Spi1* expression control [28]. *Cebpb* is mapped to multiple CREs with one CRE mapped to the promoter region and a coCRE that spans around 50kB and is bound by PU.1/CEBP β TFs. The coCRE is the most significant predictor and it overlaps with one of the super-enhancers upstream of *Cebpb* and with a region lying in an intron of the *Tmem189* gene downstream of *Cebpb* that shows high H3K27ac activity specifically in macrophages. This indicates the involvement of PU.1 or PU.1/CEBP β complex in upregulating *Cebpb* in macrophages.

These results show that the proposed method has sufficient statistical power to discover at least one significant predictor, which is a potential regulatory element, for more than half (9/17) of this small set of key transcription factor genes. It is noteworthy that in many cases this element is a coCRE, suggesting that these CRE communities are more likely than singleton elements to be predictive of gene expression.

Characteristics of the predicted cis-regulatory elements

The TF cluster and the DE gene sets represent much larger sets and the data in Table 2 show that the method still generates a useful proportion of statistically significant models, although this falls to 30% of genes in the largest set. Full details of the models are in Table S3. Figure 3 shows an analysis of the characteristics of the chosen CREs within statistically significant models. The phylogenetic conservation scores of the chosen CREs are significantly high compared to the score distribution of the all CREs (Figure 3A). Similarly, the occurrences of H3K27ac peaks and DHS across all cell types are higher for chosen CREs (Figure 3B) as are TF binding events (Figure 3C). These observations suggest that regions that are predictive of expression are hotspots for epigenomic activities and TF occupancies, confirming earlier observations [57].

Figure 3: Characteristics of chosen regions. Conservation, chromatin events and TF binding events in chosen CREs (red) compared to all candidate CREs (blue). Left is the TF gene set, middle the TF cluster set and right all differentially expressed genes (DE set). (A) The log of the phylogenetic conservation scores (see Methods), (B) the chromatin events (H3K27Ac peaks and DHS), and (C) TF binding events. The *p* values were obtained using student t-test for conservation and Kolmogorov-

Smirnov test for chromatin and binding events. All the p values were less than 0.01 except for conservation distribution in TF set ($p=0.07$).

Selected regulatory regions as enhancers

It is to be expected that a substantial number of the chosen (co)CREs are enhancers or include enhancers as sub-regions. In order to investigate this further, we considered existing datasets of enhancers from disparate sources including a dataset validated using *in-vivo* screening (VISTA) [58], two sets that were collated based on TF binding and further experimental validation (Schütte et al. and Dogan et al.) [22, 24], a set of super-enhancers identified using genome-wide binding profiles of TFs along with Mediator (Whyte et al) [53] and a dataset generated through integrative selection from various different NGS resources (SEA) [59]. It is to be noted that any comparison is likely to be affected considerably by the type of data used for integration in genome-wide enhancer datasets, the type of cells used in experimental validation of enhancer activity, and the experimental or computational methodology used by the authors in compiling the dataset. In this analysis, we have therefore attempted a conservative and unbiased approach of studying genomic overlaps between the chosen CREs/coCREs with each of the known enhancer datasets. The results are shown in Figure 4A and the total number of regions that overlap is provided in Table S4. As expected there is a significant overlap with genome-wide methods (Whyte et al. and SEA) that follow a similar approach. The overlap with Schütte et al. is high but the selected CREs show no significant overlap with datasets by Dogan et al. and VISTA. The likely reason for this discordance is the fact that Schütte et al. examined enhancers active in haemopoietic stem/progenitor cells which will be similar to the enhancers activated during haemopoietic specification (our study). Furthermore, with reductionist approaches the analysis is influenced by the genes studied in the experiments. For example, VISTA database has only 0 (0%), 6 (1%) and 75 (14.5%) of genes that overlap with our TF, TF Cluster and DE gene sets respectively. As a negative control, we chose to compare the overlaps of candidate CRE regions with similar H3K27ac levels that were not chosen as predictive by our method (Fig 4A, right hand panel). As expected, genome-wide data sets (SEA and Whyte) had significant overlaps with these negative control CREs. However, in the cell type specific data set of Schütte et al. we found that these negative control regions overlap less significantly with active elements and more significantly with inactive elements.

Figure 4: Validation of predicted enhancers. (A) Overlap of genomic regions between published sets of enhancers (vertical axis) and the CREs/coCREs chosen for genes in the three gene sets (horizontal axis). The dot plot indicates the significance ($-\log_{10}(p)$ with p adjusted for multiple testing) of the pairwise overlaps (red/large size = high significance, orange/small size = low significance). The absence of a dot signifies $p > 0.05$. The right panel (Enrich) shows a negative control of overlaps with candidate CREs that were not chosen as predictive by our method but with H3K27ac enrichment level similar to the chosen (co)CREs. (B) Expression of *Sptbn1* across stages of haematopoietic and cardiac differentiation shown as a blue line chart. The inset plot shows the chromatin activity profile (CAP) for a CRE predicted to be associated with this gene. (C) A screenshot of the predicted CRE within the *Sptbn1* gene body. The UCSC browser session shows this region shaded in blue, illustrating the dynamics of the active chromatin mark, H3K27ac (top), and chromatin accessibility (bottom) across the cell types. The coordinate considered for further validation is chr11:30166167-30166587 highlighted in transparent cyan box. (D) A 5-day time course of haematopoietic differentiation, tracking the expression of a YFP reporter gene driven by the predicted CRE. Expression peaks on day 5 (D5), which is equivalent to the haemogenic endothelium. The controls are the ESC line HM1 (black) and HM1 cells targeted with the reporter construct containing the minimal promoter (MP) only (grey).

We performed an in-depth gene-wise analysis of the overlap with Schütte et al. because of the significance of genomic overlap and also the similarity in the cell types mentioned provides us with an ideal, albeit small, dataset to check if the chosen CREs' roles are enhancers (details in Table S4). The

dataset overlaps with enhancers of 9 genes with 5 of the lasso models having a q value ≤ 0.05 . The coCRE predicted by our *Runx1* specific model overlaps with 3 active and 1 inactive enhancers by Schütte et al. (Figure 2D). Only one active enhancer (Runx1+204 [24]) was rejected by our model because of the 100kB distance cut-off (see Methods). In *Meis1* the singleton chosen CRE (Table S3) overlaps with the only active +48kB hematopoietic enhancer identified by Schütte et al. The coCRE of the *Spi1* (PU.1) overlaps with one enhancer and one inactive region, where the latter is a promoter region (Figure S5A). The coCRE of *Tal1* overlaps with all the six enhancers (3 active and 3 inactive) from Schütte et al. (Figure S5B). Interestingly one active enhancer was correctly mapped to the coCRE of *Tal1* although the CRE is in close proximity with the neighbouring gene, *Pdzk1ip1*. Although this enhancer is located far away from the other active enhancers, being part of a single coCRE may be an indication of interaction. In *Erg*, the model was able to map all the active enhancers to the single chosen coCRE (Figure S5C). In *Lyl1* the chosen CRE extends over a broad region comprising of one active and two inactive enhancers. However, in the case of *Fli1* the singleton CRE overlapped with two haemopoietically inactive elements.

While our method is intended to produce gene specific regulatory models, and does not aim directly to predict enhancer-promoter interactions, we considered that it would be useful to compare the information from our model with that from a contemporary method for the latter problem. We trained the JEME method [40] on 3D interaction data from ESC cells with the reduced feature set available in our data, and used this to predict interactions in other cell types (see Methods), focusing here on predictions in HP cells. An example of the *Runx1* gene is shown in Figure S5D and summary results over all nine Schutte et al. genes are in Table 3 and Figure S5E. In the case of *Runx1*, our retrained JEME predicts 10 potential enhancers within 100kB interacting with the promoter; these overlap all four identified as positive by Schutte et al., and only one of the four identified negatives. The remaining six JEME predictions represent untested regions, and may be false positives but equally could be regulatory. Our method on the other hand makes only five predictions, which overlap 3/4 positives and 2/4 negatives. Results for other genes are similar (Table 3 and full detail in Table S5). Overall the retrained JEME makes more predictions (Figure S5E), some of which may be false positives, and over the tested regions it is slightly more accurate than our method, although the difference is not statistically significant.

Table 3. Overlaps of predicted regulatory elements with experimentally tested regions

Gene	Predictions ¹		Overlap +ve ²		Overlap -ve ³		No overlap ⁴	
	JEME	CRE	JEME	CRE	JEME	CRE	JEME	CRE
<i>Erg</i>	12	4	1/3	3/3	1/2	0/2	10	2
<i>Fli</i>	15	1	2/2	0/2	0/3	2/3	13	0
<i>Gata2</i>	7	1	2/4	2/4	0/2	0/2	6	0
<i>Gfi1b</i>	15	4	3/3	0/3	1/1	0/1	11	4
<i>Lyl1</i>	13	4	1/1	1/1	1/2	2/2	12	3
<i>Meis1</i>	10	1	1/1	1/1	2/3	0/3	6	0
<i>Spi1</i>	14	4	0/1	1/1	0/2	1/2	14	2
<i>Runx1</i>	10	5	4/4	3/4	1/4	2/4	6	1

<i>Tall</i>	10	3	2/3	3/3	0/3	3/3	8	0
Totals	106	27	16/22	14/22	6/22	10/22	86	12

1. The number of predicted regulatory regions within 100kB of the TSS for each method
2. ‘True positives’ – the number of regulatory regions experimentally verified as positive and overlapped by a predicted region. Numbers in the denominator indicate the number of experimentally verified positive regions for this gene.
3. ‘False positives’ – the number of regions experimentally verified as negative and overlapped by a predicted region. Numbers in the denominator indicate the number of experimentally verified negative regions for this gene.
4. The number of predicted regions that do not overlap experimentally tested regions. Either false positives or novel discoveries.

Novel enhancers for *Sptbn1* and *Nfe2* genes verified experimentally

We next sought to assess the utility of our gene specific chosen CREs by experimental testing of previously unknown potential enhancer elements contained within them. We selected two of our predicted CREs that exhibit discrete CAPs and that are associated with genes that are important for haematopoiesis, namely *Nfe2* and *Sptbn1*. *Nfe2* encodes for NF-E2 transcription factor expressed in HSCs, erythroid, myeloid and megakaryocytic lineages, and is known to be involved in epigenetic modification thereby regulating blood cell maturation programmes. Abnormal expression of this gene is linked to pathogenesis of myeloproliferative neoplasms (MPNs) [60]. *Sptbn1* encodes for β spectrin, a cytoskeletal protein and has been implicated in the determination of cell shape, organelle organisation and cellular traffic [61]. A fusion gene SPTBN1-FLT3 has been observed in a small population of BCR-ABL-negative CML [62].

For both of these genes the CREs were obtained from their respective gene-specific models using λ_{1se} . We segmented the predicted CRE region into 500 bp windows and identified the window with the highest number of TF binding sites within it. This TF hotspot within the predicted CRE was then cloned upstream of a reporter gene and used for single site targeted integration into a mouse ESC line as described previously by Wilkinson et al. [63]. This enabled us to follow the dynamics of reporter gene expression during blood specification.

Out of 11 possible predictors for CREs associated with *Sptbn1*, one is chosen within the gene body and it is evident that its H3K27ac and DHS profile across the cell types coincides with the expression profile of the gene (Figure 4 B and C). The TF hotspot for this CRE shows binding of TEAD4, TAL1, LMO2 in HB, FLI1, TAL1, LMO2 in HE and RUNX1, GFI1, GATA2, TAL1, LMO2 in HP to this 500bp region (highlighted by cyan box in Figure 4C). Consistent with our CRE prediction, we see enhancer activity with the reporter gene assay (Figure 4D) and this shows an expression profile that reflects profile in Figure 4B with high expression in HE and HP cells. In the case of *Nfe2* (see Figure S6) we find two predicted CREs and this presents an interesting case because the q value of the model unlike *Sptbn1* is high (Table S3). CRE1 shows a promoter like profile, having a high H3K27ac peak showing a characteristic bimodal distribution that dips where there is enrichment for hypersensitive sites (Figure S6). CRE2 presents an interesting case because it does not have a strong H3K27ac signal, but has a very specific DHS at the HB stage that progressively opens up until HP. Also, the CAP correlates well with the GEP of *Nfe2*. Furthermore, its TF hot spot is bound by the LMO2, TAL1 complex from HB to MAC and is joined by the binding of GATA1, GATA2, GFI1 and GFI1B in HP where the chromatin activity is the highest. The reporter gene assay shows a very similar expression profile to the CAP, with expression as early as HB, peaking on day 5 when HPs begin to emerge. Therefore, both enhancer studies demonstrate that the dynamics of expression regulated by the predicted CREs follows the CAP used in the model building.

Combinatorial binding and transcriptional regulatory network

The complex structure of transcriptional regulatory networks in higher eukaryotes reflects the combinatorial control of genes by multiple transcription factors. It is poorly understood owing to the multiplicity of potential regulatory elements and the associated difficulty in assigning distal elements to the genes they control. The methods introduced here allow the construction of networks where transcription factors link only to genes for which they bind to chosen CREs and with which they are co-expressed. These networks contain putative causal regulatory relationships and here are called cis regulatory networks (CRNs); they are approximate and likely to reflect only the strongest aspects of control for each gene, but nevertheless they represent a substantial simplification compared to alternative approaches. Figure S7 illustrates this in comparison to simple co-expression networks (CENs, where TFs are connected to all genes whose expression is correlated). This shows that the CRN is simpler in terms of vertex degree and connectivity. It is also clear that those TFs with high betweenness centrality, a measure of the importance of the node to the overall network structure, stand out much more clearly in the CRN, and are enriched for known critical haemopoietic transcription factors, TAL1, RUNX1 and FLI1. Figure 5 illustrates CRNs for the TF cluster set (Figure 5A) and the TF set (Figure 5B). It is evident that the regulation involved in maintaining pluripotency is disconnected with other systems and that there are at least four different sub-graphs of regulation: (i) the pluripotency network controlled by Nanog, Oct4, Sox2 and Esrrb, (ii) the HE dominant network controlled by Fli1, (iii) the HP network comprising of the Runx1, Scl/Tal1-Lmo2 TFs and (iv) the MAC specific network with Pu.1 (Spi1) forming the nexus between the HE/HP networks and the Cebp network.

Figure 5. Cis regulatory networks (CRNs) and joint clustering of expression and regulation. (A) CRN for the TF cluster set. The genes are represented as nodes and directed edges show the genes that are co-expressed and also bound by the TF of one gene (source) to a predicted CRE of the other gene (target). The cell type at which the expression of the gene is highest is shown as colours on the node. (B) The CRN for the TF set (colours as A). (C) Joint clustering of genes in the TF cluster set. Genes cluster together according to the relatedness of both gene expression patterns (red-blue heatmap) and the binary pattern of TF binding (green-white heatmap) at their main predicted (co)CRE. Each cluster is distinguished by a colour coded bar above the GEP heatmap (highlighted as “Joint cluster ID”). For each cluster, the average of the TF binding profile is shown as the TF binding propensity, where 0 represents absence of TF binding and 1 represents binding of that specific TF in all the regions belonging to that cluster (green=TF binding; white= no TF binding).

Network complexity in higher eukaryotes has so far made it difficult to answer the fundamental question of how many different genes are regulated in the same way, or if a similar pattern of gene expression can be produced by different regulatory processes. Figure 5C shows a preliminary analysis of this question where genes are assigned to clusters where they share common patterns of both gene expression (GEP) and transcription factor binding (TFBP) to their single most significant/predictive (co)CRE, using a recently introduced joint clustering method [64] (see Methods section). This analysis uses the “*TF cluster set*” with genes having chosen (co)CREs bound to only one TF binding event removed, reducing the gene set from 437 to 222 genes, in order to focus on genes with clear combinatorial TF binding.

Clusters obtained from joint clustering are rather different to those from “expression only” clustering shown in Figure S2B. Clusters based on joint information (TF binding and gene expression) are approximately two times larger than those based on gene expression only, and represent a broader range of expression patterns: in the case of using expression information only, most of the TFs lie in separate clusters (with the exception of Tal1 and Gfi1b), while including TF binding information draws together TFs into shared clusters. This suggests that shared TF regulation leads to a broad range of similar gene

expression profiles, and that other regulatory processes, beyond the information in this data set, could define the detailed differences shown between the smaller “expression only” gene clusters.

Among the interesting aspects of biology suggested by joint clustering is that the pluripotency factors Esrrb and Sox2 form a separate cluster to Nanog and Oct4, which show an expression pattern that persists longer and is associated to different TF binding profiles. In relation to haematopoiesis, genes in the cluster containing Fli1, Lmo2 and Elk4 are bound predominantly by FLI1 at their CREs in HE and HP suggesting a tightly controlled regulatory mechanism by Fli1 underlining its importance in haemopoiesis and vasculogenesis [32]. Genes in the cluster with Runx1 express in the later stages of haemopoiesis (HE, HP, and MAC) and exhibit a more complex binding pattern of factors including FLI1, SCL/TAL1, RUNX1 and GATA2 in HB, HE and HP. Thus, the hematopoietic differentiation from HB to MAC seems to be tightly controlled by distinct regulatory mechanisms at various stages of the pathway that start with the FLI1 controlled genes co-expressing with Fli1 and Lmo2 in HB resulting in the initial activation of hematopoietic genes, followed by the tightly controlled cluster of Runx1, Tal1, Gata2 and Meis1 involved in the formation of HE and later the transition from HE to HP dictated by CEBP/ β bound genes and finally PU.1 specifying the macrophage transcriptional network.

Conclusion

Transcriptional regulation can be investigated at the level of a single gene, where studies lead to detailed understanding of all or most relevant cis-control elements, or at genome-scale where high-throughput studies can reveal many general aspects of regulation. Here we have attempted to create methodology that can bridge the gap between these contrasting types of study. One of the challenges of genome-scale study is that the data sets involved are indeed very large, but the information per gene is relatively small. Coupled with the plethora of potential regulatory elements for any gene in higher eukaryotes, it is a significant challenge to identify those elements most relevant to its regulation. We have shown that careful integration of high-throughput data of a range of types and sources, coupled with the building of gene specific predictive models, can identify a few (typically 1-4) statistically significant potential regulatory elements for a large proportion of genes. We view the method as a means of hypothesis generation that will feed into further detailed study of individual genes and regulatory elements, and here we have illustrated this with experimental confirmation for two example genes.

The CREs identified by our method, show many of the expected characteristics, including higher than average conservation, binding of transcription factors and activity in multiple cell types. Counter-point to this is that many of the CREs are not conserved, perhaps reflecting the fact that regulation of genes may diverge more quickly than the genes themselves in closely related species, showing that conservation is not an effective way of reducing the number of candidate CREs in model building. Introducing the concept of a community of CREs, related by similar profiles of transcription factor binding and activity measures across the cell types involved, has aided this work technically by reducing the number of potentially predictive variables and reducing multi-co-linearity. Equally importantly, it may have deeper biological significance, for instance in three-dimensional genome structure, and this will be a subject of future investigations. This feature of using network-based construction of collaborating CREs prior to model building, to the best of our knowledge, is presented for the first time in this work. From our analysis of haemopoietic master regulators, we found that many of the chosen coCREs contain elements close to promoters, but also link in distal CREs (≥ 20 Kb) with similar profiles of transcription factor binding and activity.

In comparison of regulatory regions chosen by the models with experimentally and computationally known enhancers we find that coCREs overlap more significantly than CREs, further emphasising that collaborations between distal regulatory elements could be a prevalent mode of transcriptional regulation. In addition, we were able to exploit the loci identified by the models to transform

incomprehensible co-expression networks (CENs) to tangible cis-regulatory networks (CRNs). The derived CRNs more clearly define the networks and subnetworks of TFs and master regulators responsible for different stages of differentiation.

The comparison of our method with JEME provides an interesting perspective on two completely different computational approaches to predicting how genes are regulated, one (ours) designed to be independent of 3D chromosome interaction data and to identify a small number of the most predictive elements for the expression of a gene, and the other trained on, and designed to predict 3D interactions. Comparison of methods in this area for mouse is difficult owing to a lack of gold standard data for real regulatory relationships. Three-dimensional interaction is often used in this way, but not all interactions are regulatory [65]. Comparison on the Schutte et al. data set, albeit relatively small, showed the methods to have similar performance. The JEME related method produced more predicted elements for each gene and may have a higher false positive rate for this reason. On the other hand, for known elements, performance of our method was marginally worse, but this difference was not statistically significant. Overall these results show that our rather simple method, based on the limited data sets available for mouse cells in this case, is useful and able to perform comparably to a method developed for the larger data sets typically available for human cell lines.

Materials and Methods

Dataset

For this study, high-throughput data from two *in vitro* differentiation pathways of mouse embryonic stem cells (ESCs) were considered: differentiation to macrophages [15] and to cardiomyocytes [66]. The fastq files for H3K27ac, DNaseI, RNA-seq and Transcription Factor (TFs) experiments were downloaded from Gene Expression Omnibus (GEO identifiers: GSE69101 and GSE47950 respectively) [36]. Gene and transcript definitions from RefSeq [37] were obtained using the UCSC Table Browser utility [67, 68]. The conservation scores are the phastCons evolutionary conservation scores for 60 vertebrate species obtained from the UCSC genome browser using its Table Browser. Enhancers defined by Whyte et al. [53] [40], Schütte et al. [24] and Dogan et al. [22] were obtained from their respective publications. The enhancers from SEA [59] and VISTA [58] were downloaded from their respective websites.

Data processing

The sequences were processed uniformly by first trimming reads for sequence quality of 20 using Cutadapt [69]. The trimmed reads were aligned to mm10 (UCSC genome browser) using Bowtie2 [70] for letter-space reads and SHRiMP [71] for colour-space reads. Only reads aligned to unique chromosomal positions and with a mapping quality of at least 20 were retained for further calculations. Total reads that overlap the exons of the genes were calculated using HTSeq-count [72]. Gene level expression values were computed as reads per kilobase of exons per million mapped reads (RPKM) [73] (equation 1), and a standardised expression value for a gene was computed as $(x-\mu)/\sigma$ where x is the expression value, μ the mean and σ the standard deviation over cell types. For the initial data analysis, the expression data for protein coding genes in each cell type were analysed by principal component analysis (PCA) and hierarchical clustering to confirm the expected biological relationships between the cell types (see Figure S1).

$$RPKM = \frac{n \times 10^9}{s \times N} \quad (1)$$

where n = total number of reads aligned to exons of size s bp and N is the total reads aligned to the genome.

DNaseI Hypersensitive Sites (DHSs) were called using DFilter with default parameters (-bs=100 -ks=50 -refine and $FDR \leq 0.05$) while the TF peaks were called using MACS (default parameters) since the method of generating shifting tags is well suited for DNA binding proteins. For TF peaks manual assessed p -values were used and are described in Goode et al. [15]. H3K27ac enriched regions in hematopoietic lineage were the “gapped peaks” called by MACS2 [74] using q -value cutoffs of 0.01 for narrow peaks and 0.1 for broad peaks [39] and the H3K27ac enriched regions in the cardiac lineage were obtained from Wamstad et al. [66]. DHSs and H3K27ac peaks that had an overlap of at least 125 bps were merged and the read abundances in these peak regions were computed as RPKMs (equation 1, where s is the size of the region in bp) using BedTools [75]. Conservation scores for the regions were obtained as the 3rd quartile value (top 75%) using bwtool [76].

Differentially expressed genes were identified by all against all pairwise cell type comparisons using DESeq [72]. A gene was defined as differentially expressed (DE) if its expression has changed at least four-fold with an adjusted p -value of less than 0.05 in at least one comparison. We identified 9854 genes that were differentially expressed and with expression value greater than 1 in at least one cell type. The standardised expression values of these DE genes were then clustered using k -means and the Bayesian Information Criterion (BIC) scores were obtained as a function of k using the adegenet R package [77, 78]. The optimal number of clusters ($K = 391$) was thus determined as the k with the lowest BIC score (Figure S1D). For a given DE gene its Gene expression profile (GEP) is an m -dimensional vector where m is this the total number of cell types analysed (here $m=10$) standardized gene expression values over all cell types, GEPs) were generated for all the differentially expressed genes considered in this study.

Identification of candidate *cis* regulatory elements (CREs)

The DHS and H3K27ac peaks that had an overlap of at least 125 bps were merged to obtain the initial set of candidate *cis*-regulatory elements (candidate CREs). For each candidate CRE, a “chromatin activity profile” (CAP) is a m -dimensional vector representing the chromatin activity of the CRE in the m cell types (here $m=10$). For a given CRE and a given cell line the chromatin activity is the $\log_2(\text{RPKM})$ of H3K27Ac computed as given in equation 1 where s here is the width of the H3K27ac peak, and if DNaseI-seq data was available for the cell type, the average of the $\log_2\text{RPKM}$ of the H3K27ac and the DNaseI-seq data (Table 1), since these two measures have been shown to be highly correlated in cell type specific DHSs [79]. For each gene the predictors are the set of candidate CREs within 100kb of a gene’s transcription start site (TSS) identified using ChIPpeakAnno [80].

Identification of community CREs

The CREs were overlapped with the 33 TF ChIP-seq peak sets (Table 1) and for each region the “TF binding profile” (TFBP) was generated as a binary vector indicating the overlap (1) or non-overlap (0) of the 33 peak sets (Figure 1A, Step: 2). For a given gene, we identified the N CREs with 100 kB of the gene and in order to generate community CREs (coCREs) two matrices of $N \times N$ dimensions were generated: (i) the activity correlation matrix (C) that represents the pairwise similarity between the CAPs of the N CREs, and (ii) the TFBP correlation matrix (B) that represents the pairwise similarity of the TFBPs of the N CREs. Each element of matrix C (C_{ij}) is the Spearman correlation coefficient of the CAPs of the two CREs (i and j) and similarly B_{ij} is the Jaccard index of the TF binding profiles of i and j . From these two matrices, an adjacency matrix (A) of $N \times N$ dimensions were computed as given in equation 2. A network then is a representation of A where CREs within 100kb of the gene represent the nodes and an edge between two nodes (if $A_{ij} = 1$) represent highly correlated activity of both their CAPs and TFBPs.

$$A_{ij} = \begin{cases} 1 & \text{if } B_{ij} \cdot C_{ij} > 0.5, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

For such a network, communities are the dense sub-networks of sets of nodes with relatively large numbers of edges among them and were identified using a greedy optimization of modularity score [81]. The coCREs are defined as communities of more than one region with at least two TFs binding to them. The CAP of a coCRE is the average of the activity profiles of the constituent CREs. To differentiate them from coCREs, candidate CREs not assigned to communities are referred to as ‘singleton CREs’.

Penalized linear models

All coCREs within 100kB of a transcript’s start site (TSS), and singleton CREs within 20kb, were considered as potential predictor variables in building the linear model. The CAPs of the CREs were considered as a column in the independent term matrix (X). The response vector Y , consisting of expression values corresponding to the cell type in X was modelled using linear regression as

$$y = \beta_0 + \sum_{i=1}^n \beta_i x_i, \quad (3)$$

where β_i are the regression coefficients. For a penalized linear model, the aim is to minimize the penalized likelihood (assuming a normal distribution), $\min_{\beta} (\|y - \hat{y}\|_2^2 + \lambda \sum_{i=1}^n |\beta_i|)$, where the penalty is on the summed absolute values of the regression coefficients and λ is the regularisation parameter. The l_1 norm form of the penalty leads to zero values for a subset of the β coefficients, depending on the size of this parameter. A leave-one-out cross validation was performed for a sequence of λ values to obtain the optimal value of λ where mean square error is minimized λ_{min} and also $\lambda_{1se} = \lambda_{min} + 1 \times stdev$, which is a more conservative estimate for λ (Figure 1C). The CREs/coCREs with $\beta \neq 0$ at the optimal λ are the most predictive for that gene, and are termed ‘chosen CREs’. If at the λ cutoff the total number of predictors with non-zero regression coefficients is zero, then the highest λ value at which at least one predictor has a non-zero coefficient was considered. The penalized regressions and cross-validations were performed using glmnet package in R [82]. In order to generate p values for the predictors with $\beta \neq 0$ we have used the covariance testing as implemented by the covTest package in R [51] and adjusted for multiple testing using Benjamini-Hochberg method (q values). The covariance test tests the significance of a predictor as it enters the active set at a given step in the lasso path [51]. The p -value for a predictor takes into account the adaptive nature of lasso and also the shrinkage effects due to the ℓ_1 -norm penalty. The q -value for each gene model is the q -value of the most significant predictor at λ_{1se} . The models generated, the λ values, the chosen predictors along with the q -values are provided in Table S3. It can be seen from Table S3 and Figure S8 that for the selected models there are non-zero beta with $p > 0.05$. They should be viewed with caution in comparison to better supported interactions, because on one hand the cross validation indicates they have a significant contribution to prediction accuracy but on the other hand the covariance test, which is more conservative, considers them insignificant.

Apart from using covariance tests for assessing a model’s significance one can obtain empirical p -values by randomising either the CAP or GEP. Here for a given gene, we retained the CAP of the predictors and randomised the GEP $N=10000$ times and generating N models. The spearman’s correlation (C_r) between the randomised GEP and the GEP predicted by the model is calculated. If the correlation derived for the actual model is C_r , and x the total number of iterations where $C_r \geq C_r$, then an empirical

p -value can be generated as x/N . The comparison between the two tests for model significance is given in Figure S7E. Although there is general agreement with the exception of Runx1 and Spi1, the randomisation approach the p -value is time consuming (limited by the number of iterations), and we preferred the covariance test method which is explicitly designed for LASSO models and therefore likely to be more powerful.

It should be noted that if there are more than one highly correlated CREs for a given gene that are predictive of the expression, then LASSO is designed to select one of the CREs. In order to ascertain the ability of the model to accurately choose the best correlated regions, we removed the chosen regions from the list of predictors for a given gene and generated a new model. We performed this analysis for all the genes in the three gene sets and found only 15 new models with at least one chosen region ($\beta \neq 0$) at λ_{1se} and a p -value of less than 0.05 (Table S3). While this may remain an issue with LASSO regression, it indeed shows that in this study such effects are minimal possibly due to the construction of coCRE prior to model building.

For each gene, two models (λ_{1se}) were generated, one with regions before the construction (pre-coCRE) and the other with regions after the construction of community CREs (post-coCRE), as predictors. The difference in drop-in-variances ($\Delta div = div(\text{post-coCRE}) - div(\text{pre-coCRE})$) and in p -values ($\Delta pv = pv(\text{post-coCRE}) - pv(\text{pre-coCRE})$) for the best predictor (highest div), were computed for any given gene if $p < 0.05$ in at least one of the two models. In order to ascertain the technical impact of coCREs, we used paired Student t-tests for computing the significances for higher drop in variances ($+\Delta div$, Figure S3F, red), and decrease in p -values ($-\Delta pv$, Figure S3F, red) in post-coCRE models compared to pre-coCRE models. Figure S3F shows that in the *TF Cluster* set, both the div , and p -values for post-coCRE models are significantly better than pre-coCRE models. This trend remains true when applied to all the genes considered in this study.

Overlaps with data sets of experimentally determined enhancers

We considered the overlap of the set of (co)CREs chosen in our models with several external data set of enhancers described above (see Figure 4A), using the hypergeometric test. To form a negative control set for these studies, for each gene, a set of CREs/coCREs were selected from the available set of rejected predictors (P) that match the chromatin activity level of the predictors chosen by the model (C). If CAP_r is the chromatin activity profile vector of a CRE/coCRE (r), then $m_r = \max(CAP_r)$. The set of rejected regions with matching chromatin activity for that gene is $\{r \in P \mid m_r > m_c - 0.5\}$ where m_c is the $\max(CAP)$ of any of the chosen predictors. Such sets were generated for the three gene sets (Table 2) and their overlaps with known enhancers are given in Figure 4A.

Application of JEME to the myeloid lineage data set

Cao *et al.*'s Joint Effects of Multiple Enhancers (JEME) method employs a two-step machine learning framework to predict cell line specific enhancer-TSS interactions [40]. Firstly, LASSO linear regression models calculate the ability of DNase, H3K27ac, H3K27me3 and H3K4me1 signals to predict the expression of a TSS within 1Mb of an enhancer. Secondly, LASSO error terms are input, along with DNase and histone enhancer, promoter and window features, into a Random Forest classifier trained on chromosome conformation data.

JEME was applied to the myeloid lineage (ESC to MAC) dataset, for which DNase-seq, CHIP-seq and RNA-seq data were available. As we did not have data for H3K4me1, this mark was substituted for H3K4me3. Input files were processed as described in Cao *et al.* 2015 [40]. TSS co-ordinates were obtained from the mm10 RefSeq Curated annotation [37] using the UCSC Table Browser [67, 68]. Cell specific active enhancers were defined by the 4-state ChromHMM predictions from Goode *et al.* 2016 [15]. JEME code for Cao *et al.*'s 'Roadmap + ENCODE' dataset, was downloaded from

<https://github.com/yiplabcuhk/JEME> and adapted for our dataset. JEME's Random Forest classifier was retrained in WEKA [83] on a set of 18,353 positive and 122,353 background pairs, assembled from the ESC HiC data (GEO identifier: GSM2026260) [84] using the 'random targets' method [40]. Predictions were made for enhancer-TSS interactions in MES, HB, HE, HP and MAC cells, using the default threshold of 0.35. JEME predicted 64.2% of selected CRE-gene pairs, accounting for 7.9% of all JEME predictions within 100kb of a TSS considered by our method (p value for overlap $\ll 10^{-65}$). The results shown in Table 3 do not change significantly by varying the threshold.

Joint clustering of TFBPs and GEPs

In order to jointly cluster gene expression and regulatory input we employed a method we recently developed [64]. This adopts a probabilistic mixture model-based clustering with integration of binary (TF binding) and continuous (expression) data using Bernoulli and Gaussian distributions respectively. The mixture components (clusters) represent sets of genes with related TF binding and expression patterns. In this context, a TF binding pattern is a binary string reflecting binding (1) or not (0) for a set of TFs to the chosen (co)CRE in question in each cell type. A data matrix was constructed with the GEPs of the genes in the TF cluster set (Table 2) and the TF binding pattern (TFBP) of the statistically best predictive (co)CRE for each gene based on their p -values (Table S3). In case of ties between two or more predictors, the (co)CRE with the most regulatory input in terms of total number of bound TFs was considered. Gene-(co)CRE pairs with only one TF binding event were further removed. The method determines a clustering solution that minimizes an information criterion related to the standard Bayesian Information Criterion ($BIC = k\ln(n) - 2\ln(L)$), where k is the number of parameters in the model, n the number of data points and L the maximized likelihood) or Akaike criterion ($AIC = 2k - 2\ln(L)$). In this case we used a criterion related to the latter, $AIC_{2.5} = 2.5k - 2\ln(L)$, which was shown to be optimal in similar problems [64].

Statistical significance of combinatorial binding of TFs

We calculated the statistical significance of multiple TF occupancy of M potential CREs as follows. Let the i^{th} TF (TF _{i}) have N_i binding peaks overlapping the M CREs, then the probability that any CRE is occupied by at least one TF _{i} peak is $1 - \left(\frac{M-1}{M}\right)^{N_i}$. The probability a CRE is occupied by at least one peak of each of N independently binding TFs is $\prod_{i=1}^N \left(1 - \left(\frac{M-1}{M}\right)^{N_i}\right)$. The expected number of CREs occupied by at least one peak of each of the N TFs is $M \times \prod_{i=1}^N \left(1 - \left(\frac{M-1}{M}\right)^{N_i}\right)$ and this was used with the Poisson distribution to obtain p values for the occurrence of CREs occupied by a combination of N TFs.

Core regulatory network

TF gene set and TF cluster gene sets were used for generating two different types of network, namely the co-expression network (CEN) and cis-regulatory network (CRN). For the CEN, a pairwise Spearman correlation matrix (M) was generated for the gene expression values (GEP) in the set and an adjacency matrix (A) was derived from M where $A_{ij} = 1$ for the top 25% of the node pairs with highest correlation coefficient. The resultant matrix was represented by an undirected graph where an edge was drawn between two genes i and j if $A_{ij} = 1$. CRNs were generated with a directed edge from gene i to gene j if and only if they were connected in the CEN and gene i is a TF that binds to a chosen (co)CRE in the expression model for gene j . For generation of CRNs only selected CREs/coCREs with $q \leq 0.05$ were used.

For computing the network parameters both CRNs and CENs are considered as undirected. A degree of a node is the total number of edges incident on it. For a node "x" in a network of N nodes, if "o" is

the total number of all possible shortest paths, and if “ m ” is the total number of shortest paths from all $N \setminus x$ (all nodes except x) to $N \setminus x$ that traverse through x , then betweenness centrality is $\frac{m}{o}$. A node is termed as a neighbour of another node if there exists an edge between them and connectivity of a node is the total number of neighbours of the node. Neighbourhood connectivity of a node “ x ” is the sum of the connectivities of its neighbours [85, 86].

Availability of data and materials

All the datasets used in this study are publicly available from Gene Expression Omnibus (GEO) with GSE69101 and GSE47950 accession numbers. The code is available in github as an R package (<https://github.com/vjbaskar/lenhancer>)

Supplementary Data

Supplementary Data are available online. Table S1-S5 are given as separate Microsoft Excel sheets and Figure S1-S8 are provided as a single Microsoft Powerpoint file.

Acknowledgements

MSV and DRW thank Walter Gilks and Christopher Yau for discussions on statistical significance in LASSO models. MSV thanks Dr. Matthew Care for discussions on gene expression analyses.

References

1. Van Nostrand EL, Kim SK. Seeing elegance in gene regulatory networks of the worm. *Current opinion in genetics & development*. 2011;21(6):776-86. doi: 10.1016/j.gde.2011.08.007. PubMed PMID: 21963133; PubMed Central PMCID: PMC3224770.
2. Ouyang Z, Zhou Q, Wong WH. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*. 2009;106(51):21521-6. doi: 10.1073/pnas.0904863106. PubMed PMID: 19995984; PubMed Central PMCID: PMC2789751.
3. Walter J. An epigenetic Tet a Tet with pluripotency. *Cell stem cell*. 2011;8(2):121-2. doi: 10.1016/j.stem.2011.01.009. PubMed PMID: 21295264.
4. Bulger M, Groudine M. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Developmental biology*. 2010;339(2):250-7. doi: 10.1016/j.ydbio.2009.11.035. PubMed PMID: 20025863; PubMed Central PMCID: PMC3060611.
5. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*. 2003;33 Suppl:245-54. doi: 10.1038/ng1089. PubMed PMID: 12610534.
6. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. *Trends in genetics : TIG*. 2014;30(9):418-26. doi: 10.1016/j.tig.2014.07.001. PubMed PMID: 25108476.
7. Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, et al. Modeling gene expression using chromatin features in various cellular contexts. *Genome biology*. 2012;13(9):R53. doi: 10.1186/gb-2012-13-9-r53. PubMed PMID: 22950368; PubMed Central PMCID: PMC3491397.
8. Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America*. 2010;107(7):2926-31. doi: 10.1073/pnas.0909344107. PubMed PMID: 20133639; PubMed Central PMCID: PMC2814872.
9. Kouzarides T. Chromatin modifications and their function. *Cell*. 2007;128(4):693-705. doi: 10.1016/j.cell.2007.02.005. PubMed PMID: 17320507.
10. Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, Lee TI, et al. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*. 2006;441(7091):349-53. doi: 10.1038/nature04733. PubMed PMID: 16625203.
11. Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EEM. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*. 2009;462(7269):65-U72. doi: 10.1038/nature08531. PubMed PMID: ISI:000271419200032.
12. Wilson NK, Foster SD, Wang X, Knezevic K, Schutte J, Kaimakis P, et al. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell stem cell*. 2010;7(4):532-44. doi: 10.1016/j.stem.2010.07.016. PubMed PMID: 20887958.
13. Burgess DJ. Gene regulation: Finding genetic target sites. *Nature reviews Genetics*. 2016;17(6):314-5. doi: 10.1038/nrg.2016.53. PubMed PMID: 27087501.
14. Creighton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*. 2010;107(50):21931-6. PubMed PMID: Medline:21106759.
15. Goode DK, Obier N, Vijayabaskar MS, Lie ALM, Lilly AJ, Hannah R, et al. Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. *Developmental cell*. 2016;36(5):572-87. doi: 10.1016/j.devcel.2016.01.024. PubMed PMID: 26923725; PubMed Central PMCID: PMC4780867.

16. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*. 2011;470(7333):279-83. doi: 10.1038/nature09692. PubMed PMID: 21160473; PubMed Central PMCID: PMC4445674.
17. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74. doi: 10.1038/nature11247. PubMed PMID: 22955616; PubMed Central PMCID: PMC3439153.
18. Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. *Cell*. 2011;144(3):327-39. doi: 10.1016/j.cell.2011.01.024. PubMed PMID: 21295696; PubMed Central PMCID: PMC3742076.
19. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*. 2012;13(9):613-26. doi: 10.1038/nrg3207. PubMed PMID: WOS:000308064000009.
20. Gering M, Yamada Y, Rabbitts TH, Patient RK. Lmo2 and Scf/Tal1 convert non-axial mesoderm into haemangioblasts which differentiate into endothelial cells in the absence of Gata1. *Development*. 2003;130(25):6187-99. doi: 10.1242/dev.00875. PubMed PMID: 14602685.
21. Davis RL, Weintraub H, Lassar AB. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell*. 1987;51(6):987-1000. PubMed PMID: 3690668.
22. Dogan N, Wu W, Morrissey CS, Chen KB, Stonestrom A, Long M, et al. Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics & chromatin*. 2015;8:16. doi: 10.1186/s13072-015-0009-5. PubMed PMID: 25984238; PubMed Central PMCID: PMC4432502.
23. Hardison RC, Taylor J. Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews Genetics*. 2012;13(7):469-83. doi: 10.1038/nrg3242. PubMed PMID: WOS:000305464800009.
24. Schutte J, Wang HG, Antoniou S, Jarratt A, Wilson NK, Riepsaame J, et al. An experimentally validated network of nine haematopoietic transcription factors reveals mechanisms of cell state stability. *Elife*. 2016;5. doi: ARTN e11469. PubMed PMID: WOS:000371884500001. 10.7554/eLife.11469.
25. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. *Nature*. 2009;461(7261):199-205. doi: 10.1038/nature08451. PubMed PMID: 19741700; PubMed Central PMCID: PMC2923221.
26. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews Genetics*. 2013;14(6):390-403. doi: 10.1038/nrg3454. PubMed PMID: 23657480; PubMed Central PMCID: PMC3874835.
27. Woodcock CL, Ghosh RP. Chromatin higher-order structure and dynamics. *Cold Spring Harbor perspectives in biology*. 2010;2(5):a000596. doi: 10.1101/cshperspect.a000596. PubMed PMID: 20452954; PubMed Central PMCID: PMC2857170.
28. Freire-Pritchett P, Schoenfelder S, Varnai C, Wingett SW, Cairns J, Collier AJ, et al. Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *eLife*. 2017;6. doi: 10.7554/eLife.21926. PubMed PMID: 28332981; PubMed Central PMCID: PMC5407860.
29. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*. 2016;167(5):1369-84 e19. doi: 10.1016/j.cell.2016.09.037. PubMed PMID: 27863249; PubMed Central PMCID: PMC5123897.
30. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. 2011;473(7345):43-9. doi: 10.1038/nature09906. PubMed PMID: 21441907; PubMed Central PMCID: PMC3088773.

31. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. *Nature*. 2012;489(7414):75-82. doi: 10.1038/nature11232. PubMed PMID: 22955617; PubMed Central PMCID: PMC3721348.
32. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317-30. doi: 10.1038/nature14248. PubMed PMID: 25693563; PubMed Central PMCID: PMC4530010.
33. Ziller MJ, Edri R, Yaffe Y, Donaghey J, Pop R, Mallard W, et al. Dissecting neural differentiation regulatory networks through epigenetic footprinting. *Nature*. 2015;518(7539):355-9. doi: 10.1038/nature13990. PubMed PMID: 25533951; PubMed Central PMCID: PMC4336237.
34. Tsankov AM, Gu H, Akopian V, Ziller MJ, Donaghey J, Amit I, et al. Transcription factor binding dynamics during human ES cell differentiation. *Nature*. 2015;518(7539):344-9. doi: 10.1038/nature14233. PubMed PMID: 25693565; PubMed Central PMCID: PMC4499331.
35. Gjoneska E, Pfenning AR, Mathys H, Quon G, Kundaje A, Tsai LH, et al. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature*. 2015;518(7539):365-9. doi: 10.1038/nature14252. PubMed PMID: 25693568; PubMed Central PMCID: PMC4530583.
36. Ron G, Globerson Y, Moran D, Kaplan T. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. *Nature communications*. 2017;8(1):2237. doi: 10.1038/s41467-017-02386-3. PubMed PMID: 29269730; PubMed Central PMCID: PMC5740158.
37. Roy S, Siahpirani AF, Chasman D, Knaack S, Ay F, Stewart R, et al. A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic acids research*. 2015;43(18):8694-712. doi: 10.1093/nar/gkv865. PubMed PMID: 26338778; PubMed Central PMCID: PMC4605315.
38. Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*. 2004;306(5696):636-40. doi: 10.1126/science.1105136. PubMed PMID: 15499007.
39. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*. 2014;111(17):6131-8. Epub 2014/04/23. doi: 10.1073/pnas.1318948111. PubMed PMID: 24753594; PubMed Central PMCID: PMC4035993.
40. Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, et al. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nature genetics*. 2017;49(10):1428-36. doi: 10.1038/ng.3950. PubMed PMID: 28869592.
41. He B, Chen C, Teng L, Tan K. Global view of enhancer-promoter interactome in human cells. *Proceedings of the National Academy of Sciences of the United States of America*. 2014;111(21):E2191-9. doi: 10.1073/pnas.1320308111. PubMed PMID: 24821768; PubMed Central PMCID: PMC4040567.
42. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature genetics*. 2016;48(5):488-96. doi: 10.1038/ng.3539. PubMed PMID: 27064255; PubMed Central PMCID: PMC4910881.
43. Xi W, Beer MA. Local epigenomic state cannot discriminate interacting and non-interacting enhancer-promoter pairs with high accuracy. *PLoS computational biology*. 2018;14(12):e1006625. Epub 2018/12/19. doi: 10.1371/journal.pcbi.1006625. PubMed PMID: 30562350; PubMed Central PMCID: PMC6298642.
44. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. *Nature*. 2014;507(7493):455-61. doi: 10.1038/nature12787. PubMed PMID: 24670763; PubMed Central PMCID: PMC5215096.
45. Hafez D, Karabacak A, Krueger S, Hwang YC, Wang LS, Zinzen RP, et al. McEnhancer: predicting gene expression via semi-supervised assignment of enhancers to target genes. *Genome*

biology. 2017;18(1):199. doi: 10.1186/s13059-017-1316-x. PubMed PMID: 29070071; PubMed Central PMCID: PMC5657048.

46. Yang W, Sinha S. A novel method for predicting activity of cis-regulatory modules, based on a diverse training set. *Bioinformatics*. 2017;33(1):1-7. doi: 10.1093/bioinformatics/btw552. PubMed PMID: 27609510; PubMed Central PMCID: PMC6075022.

47. Wilczynski B, Liu YH, Yeo ZX, Furlong EE. Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state. *PLoS computational biology*. 2012;8(12):e1002798. doi: 10.1371/journal.pcbi.1002798. PubMed PMID: 23236268; PubMed Central PMCID: PMC3516547.

48. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. *J R Stat Soc B*. 2011;73:273-82. PubMed PMID: WOS:000290575300001.

49. Tibshirani RJ, Taylor J, Lockhart R, Tibshirani R. Exact Post-Selection Inference for Sequential Regression Procedures. *J Am Stat Assoc*. 2016;111(514):600-14. PubMed PMID: WOS:000381326700012.

50. Lee JD, Sun DL, Sun YK, Taylor JE. Exact Post-Selection Inference, with Application to the Lasso. *Annals of statistics*. 2016;44(3):907-27. PubMed PMID: WOS:000375175200001.

51. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A Significance Test for the Lasso. *Annals of statistics*. 2014;42(2):413-68. doi: 10.1214/13-AOS1175. PubMed PMID: 25574062; PubMed Central PMCID: PMC4285373.

52. Lichtinger M, Ingram R, Hannah R, Muller D, Clarke D, Assi SA, et al. RUNX1 reshapes the epigenetic landscape at the onset of haematopoiesis. *Embo J*. 2012;31(22):4318-33. doi: 10.1038/emboj.2012.275. PubMed PMID: ISI:000311157500008.

53. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*. 2013;153(2):307-19. doi: 10.1016/j.cell.2013.03.035. PubMed PMID: 23582322; PubMed Central PMCID: PMC3653129.

54. Bee T, Ashley EL, Bickley SR, Jarratt A, Li PS, Sloane-Stanley J, et al. The mouse Runx1 +23 hematopoietic stem cell enhancer confers hematopoietic specificity to both Runx1 promoters. *Blood*. 2009;113(21):5121-4. doi: 10.1182/blood-2008-12-193003. PubMed PMID: 19321859.

55. Gao X, Johnson KD, Chang YI, Boyer ME, Dewey CN, Zhang J, et al. Gata2 cis-element is required for hematopoietic stem cell generation in the mammalian embryo. *The Journal of experimental medicine*. 2013;210(13):2833-42. doi: 10.1084/jem.20130733. PubMed PMID: 24297994; PubMed Central PMCID: PMC3865483.

56. Landry JR, Bonadies N, Kinston S, Knezevic K, Wilson NK, Oram SH, et al. Expression of the leukemia oncogene Lmo2 is controlled by an array of tissue-specific elements dispersed over 100 kb and bound by Tal1/Lmo2, Ets, and Gata factors. *Blood*. 2009;113(23):5783-92. doi: 10.1182/blood-2008-11-187757. PubMed PMID: 19171877.

57. Kvon EZ, Stampfel G, Yanez-Cuna JO, Dickson BJ, Stark A. HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. *Genes & development*. 2012;26(9):908-13. doi: 10.1101/gad.188052.112. PubMed PMID: 22499593; PubMed Central PMCID: PMC3347788.

58. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser - a database of tissue-specific human enhancers. *Nucleic Acids Research*. 2007;35:D88-D92. doi: 10.1093/nar/gkl822. PubMed PMID: WOS:000243494600019.

59. Wei YJ, Zhang SM, Shang SP, Zhang B, Li S, Wang XY, et al. SEA: a super-enhancer archive. *Nucleic Acids Research*. 2016;44(D1):D172-D9. doi: 10.1093/nar/gkv1243. PubMed PMID: WOS:000371261700022.

60. Kaufmann KB, Grunder A, Hadlich T, Wehrle J, Gothwal M, Bogeska R, et al. A novel murine model of myeloproliferative disorders generated by overexpression of the transcription factor NF-E2.

The Journal of experimental medicine. 2012;209(1):35-50. doi: 10.1084/jem.20110540. PubMed PMID: 22231305; PubMed Central PMCID: PMC3260873.

61. Hayes NV, Scott C, Heerkens E, Ohanian V, Maggs AM, Pinder JC, et al. Identification of a novel C-terminal variant of beta II spectrin: two isoforms of beta II spectrin have distinct intracellular locations and activities. *Journal of cell science*. 2000;113 (Pt 11):2023-34. PubMed PMID: 10806113.
62. Grand FH, Iqbal S, Zhang L, Russell NH, Chase A, Cross NC. A constitutively active SPTBN1-FLT3 fusion in atypical chronic myeloid leukemia is sensitive to tyrosine kinase inhibitors and immunotherapy. *Experimental hematology*. 2007;35(11):1723-7. doi: 10.1016/j.exphem.2007.07.002. PubMed PMID: 17764812.
63. Wilkinson AC, Goode DK, Cheng YH, Dickel DE, Foster S, Sendall T, et al. Single site-specific integration targeting coupled with embryonic stem cell differentiation provides a high-throughput alternative to in vivo enhancer analyses. *Biol Open*. 2013;2(11):1229-38. doi: 10.1242/bio.20136296. PubMed PMID: WOS:000209206800017.
64. Zainul Abidin FN, Westhead DR. Flexible model-based clustering of mixed binary and continuous data: application to genetic regulation and cancer. *Nucleic Acids Res*. 2016. doi: 10.1093/nar/gkw1270. PubMed PMID: 27994031.
65. Jager R, Migliorini G, Henrion M, Kandaswamy R, Speedy HE, Heindl A, et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature communications*. 2015;6:6178. Epub 2015/02/20. doi: 10.1038/ncomms7178. PubMed PMID: 25695508; PubMed Central PMCID: PMC4346635.
66. Wamstad JA, Alexander JM, Truty RM, Shrikumar A, Li F, Eilertson KE, et al. Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell*. 2012;151(1):206-20. doi: 10.1016/j.cell.2012.07.035. PubMed PMID: 22981692; PubMed Central PMCID: PMC3462286.
67. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. *Nucleic acids research*. 2004;32(Database issue):D493-6. doi: 10.1093/nar/gkh103. PubMed PMID: 14681465; PubMed Central PMCID: PMC308837.
68. Karolchik D, Hinrichs AS, Kent WJ. The UCSC Genome Browser. *Current protocols in bioinformatics / editorial board, Andreas D Baxevanis [et al]*. 2009;Chapter 1:Unit 4. Epub 2009/12/04. doi: 10.1002/0471250953.bi0104s28. PubMed PMID: 19957273; PubMed Central PMCID: PMC2834533.
69. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011. 2011;17(1). doi: 10.14806/ej.17.1.200
pp. 10-12.
70. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*. 2009;10(3):R25. doi: 10.1186/gb-2009-10-3-r25. PubMed PMID: 19261174; PubMed Central PMCID: PMC2690996.
71. Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. SHRiMP: accurate mapping of short color-space reads. *PLoS computational biology*. 2009;5(5):e1000386. doi: 10.1371/journal.pcbi.1000386. PubMed PMID: 19461883; PubMed Central PMCID: PMC2678294.
72. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome biology*. 2010;11(10):R106. doi: 10.1186/gb-2010-11-10-r106. PubMed PMID: 20979621; PubMed Central PMCID: PMC3218662.
73. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*. 2008;5(7):621-8. doi: 10.1038/nmeth.1226. PubMed PMID: 18516045.
74. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). *Genome biology*. 2008;9(9):R137. doi: 10.1186/gb-2008-9-9-r137. PubMed PMID: 18798982; PubMed Central PMCID: PMC2592715.

75. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-2. doi: 10.1093/bioinformatics/btq033. PubMed PMID: 20110278; PubMed Central PMCID: PMC2832824.
76. Pohl A, Beato M. bwtool: a tool for bigWig files. *Bioinformatics*. 2014;30(11):1618-9. doi: 10.1093/bioinformatics/btu056. PubMed PMID: 24489365; PubMed Central PMCID: PMC4029031.
77. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*. 2008;24(11):1403-5. doi: 10.1093/bioinformatics/btn129. PubMed PMID: 18397895.
78. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*. 2011;27(21):3070-1. doi: 10.1093/bioinformatics/btr521. PubMed PMID: 21926124; PubMed Central PMCID: PMC3198581.
79. Shu W, Chen H, Bo X, Wang S. Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains. *Nucleic acids research*. 2011;39(17):7428-43. doi: 10.1093/nar/gkr443. PubMed PMID: 21685456; PubMed Central PMCID: PMC3177195.
80. Zhu LJ, Gazin C, Lawson ND, Pages H, Lin SM, Lapointe DS, et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC bioinformatics*. 2010;11:237. doi: 10.1186/1471-2105-11-237. PubMed PMID: 20459804; PubMed Central PMCID: PMC3098059.
81. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. *Phys Rev E*. 2004;70(6). PubMed PMID: WOS:000226299200018.
82. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010;33(1):1-22. PubMed PMID: WOS:000275203200001.
83. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*. 2009;11(1).
84. Krijger PH, Di Stefano B, de Wit E, Limone F, van Oevelen C, de Laat W, et al. Cell-of-Origin-Specific 3D Genome Structure Acquired during Somatic Cell Reprogramming. *Cell stem cell*. 2016;18(5):597-610. doi: 10.1016/j.stem.2016.01.007. PubMed PMID: 26971819; PubMed Central PMCID: PMC4858530.
85. Maslov S, Sneppen K. Specificity and stability in topology of protein networks. *Science*. 2002;296(5569):910-3. doi: 10.1126/science.1065103. PubMed PMID: 11988575.
86. Assenov Y, Ramirez F, Schelhorn SE, Lengauer T, Albrecht M. Computing topological parameters of biological networks. *Bioinformatics*. 2008;24(2):282-4. doi: 10.1093/bioinformatics/btm554. PubMed PMID: 18006545.

Supplementary Information

S1 Table: The three gene sets for which predictive models were generated and the gene expression data.

S2 Table: Table containing CREs used in this study.

S3 Table: Gene-wise predictive models for the genes and the chosen CREs/coCREs.

S4 Table: Overlap between chosen CREs/coCREs and known enhancers.

S5 Table: (A) Positive enhancer elements for the genes in Schutte et al. data and their overlap with predictions from our (co)CRE method and retrained JEME method. (B) Negative enhancer elements for the genes in Schutte et al. data and their overlap with predictions from our (co)CRE method and retrained JEME method.

S1 Figure: **Gene expression analysis.** (A) Principal components analysis of gene expression data where the cell types are projected on the first two principal components (PCs). (B) The cumulative contribution of the PCs to the variance observed. (C) Heatmap showing the hierarchically clustered cell types based on the correlation (Pearson) of their gene expression profiles. (D) BIC scores as a function of number of clusters (K) when clustering gene expression profiles for differentially expressed genes. The vertical line corresponds to the K with the lowest BIC score.

S2 Figure: **Gene sets used in this study.** (A) The normalised expression values of the genes in the “*TF set*”. (B) shows the hierarchically clustered heatmap of the genes in the “*TF cluster set*”. (C) is a table containing the total number of genes that are differentially up (red) and down (blue) when comparing two cell types (rows to columns). See Table S1 for information on the genes used in this analysis.

S3 Figure: **Statistical overview of the cis-regulatory elements (CREs).** Frequency of the CREs with a given number of chromatin events (A), size of CREs in (\log_{10} of bp) (B), conservation scores (C), H3K27ac enrichment (D) and DNaseI-seq enrichment (E). (F) Gene specific models are built for the *TF Cluster* set with coCRE (post-coCRE) and without construction of coCREs (pre-coCRE, see Methods). A gene is considered if $p < 0.05$ in either of the two models. For a given gene the predictor with best drop in variance (*div*, from covariance test [49,50]) and p values (pv) for each of two models are considered and compared. The differences in the *div* and p values, Δdiv and Δpv respectively, were computed for each gene and the frequencies are plotted as bar charts (lower panel). The logged *div* and pv for both models for *TF Cluster* genes are plotted with lines coloured as given in the legend (upper panel). A $+\Delta div$ or a $-\Delta pv$ (red) indicates that the post-coCRE model is better than that of pre-coCRE and a $-\Delta div$ or a $+\Delta pv$ (blue) indicates vice versa. A paired t-test shows that post-coCRE models are significantly better than pre-coCRE models (div : $1.8e-06$ and pv : $3.5e-06$).

S4 Figure: **Gene expression models for the genes in the TF set.** The gene-specific models showing the chosen coCREs and singleton CREs that are most predictive of the transcription factor’s gene expression profile. The method of generating the network and annotating the nodes and the edges is same as Figure 2A. Table S3 contains all the relevant details of the models and the CREs shown here. Please note that not all the TFs have supporting ChIP-seq data (Table 1). If a TF does not have ChIP-seq data in a given cell type, the value in that histogram is zero (see also Figure 2A)

S5 Figure: **Comparison of CREs/coCREs of key genes with known enhancer datasets.** The genome browser snapshot for the tag density tracks and the datasets of known enhancers are generated similar to Figure 1D. The overlap between chosen (co)CREs (red bar below the density tracks) for Spi1 (A), Scl/Tal1 (B) and Erg(C) with Schütte et al. (blue and purple bars) are shown. (D) Browser snapshot showing enhancers predicted by JEME in the Hematopoietic Progenitors (HP) in blue, alongside the

tested regions investigated by Schütte et al. in green and dark red, and CREs/coCREs selected by our method in red. (E) Overlap of predicted CREs of our method, and JEME, with Schütte et al (active).

S6 Figure: **Experimental testing of predicted enhancer for *Nfe2l3*.** (A) Gene expression profile of *Nfe2l3* and chromatin accessibility profile of the predicted enhancer (inset). (B) Chosen CRE containing the predicted enhancer is highlighted as transparent cyan box. (C) Reporter gene investigation of the enhancer activity.

S7 Figure: **Network parameters for the GRNs.** Network parameters such as the degree (A), Betweenness centrality (B) and Neighbourhood connectivity (C) for the key genes (*TF set*) (left) and all the genes (right) in the *TF Cluster set*. CRNs (yellow) and CENs (black) are generated as described in Methods section. (D) The co-expression networks (CEN) of genes in the *TF set*. The edges are drawn if the correlation coefficients (spearman) of GEP between two genes were in the top 25%. (E) Comparison of p values generated by the covariance test (covTest) and those from GEP randomisation for predictive models of the indicated TF genes.

S8 Figure: **Covariance tests' significance values of CREs and coCREs in gene-wise models.** The $-\log_2P$ (adjusted) cutoffs on the x-axis and the total number of CREs or coCREs with the $-\log_2P$ (adjusted) better than a given cut off for all the gene models with at least one $\beta \neq 0$ in blue and only for the significant models in red i.e. with $q \leq 0.05$ (Table 2), on the y-axis.