Published online 26 November 2014 *Nucleic Acids Research, 2014, Vol. 42, No. 22* ***13513–13524***
doi: 10.1093/nar/gku1254

Constrained transcription factor spacing is prevalent and important for transcriptional control of mouse blood cells

Felicia SL Ng, Judith Schütte, David Ruau, Evangelia Diamanti, Rebecca Hannah, Sarah J. Kinston and Berthold Göttgens[*]

Department of Haematology, Wellcome Trust and MRC Cambridge Stem Cell Institute & Cambridge Institute for Medical Research, Cambridge University, Cambridge CB2 0XY, UK

Received August 07, 2014; Revised November 11, 2014; Accepted November 12, 2014

ABSTRACT

Combinatorial transcription factor (TF) binding is essential for cell-type-specific gene regulation. However, much remains to be learned about the mechanisms of TF interactions, including to what extent constrained spacing and orientation of interacting TFs are critical for regulatory element activity. To examine the relative prevalence of the 'enhanceosome' versus the 'TF collective' model of combinatorial TF binding, a comprehensive analysis of TF binding site sequences in large scale datasets is necessary. We developed a motif-pair discovery pipeline to identify motif co-occurrences with preferential distance(s) between motifs in TF-bound regions. Utilizing a compendium of 289 mouse haematopoietic TF ChIP-seq datasets, we demonstrate that haematopoietic-related motif-pairs commonly occur with highly conserved constrained spacing and orientation between motifs. Furthermore, motif clustering revealed specific associations for both heterotypic and homotypic motif-pairs with particular haematopoietic cell types. We also showed that disrupting the spacing between motif-pairs significantly affects transcriptional activity in a well-known motif-pair—E-box and GATA, and in two previously unknown motif-pairs with constrained spacing—Ets and Homeobox as well as Ets and E-box. In this study, we provide evidence for widespread sequence-specific TF pair interaction with DNA that conforms to the 'enhanceosome' model, and furthermore identify associations between specific haematopoietic cell-types and motif-pairs.

INTRODUCTION

Transcription factors (TFs) are primary mediators of gene regulation, and they have long been known as essential regulators of cell fate decisions in the haematopoietic system. TF proteins form complexes, bind regulatory DNA sequences on enhancers and promoter regions and help to recruit the basic transcriptional machinery to control the expression of nearby genes. The interaction between two TFs and the DNA therefore represents the most basic component in understanding larger TF complex formations (1). However, the molecular mechanisms by which such complexes control gene expression are still largely unknown. One of the best-understood enhancers controls expression of the interferon-β gene, where it is now recognised that specificity in gene expression does not arise from the cumulative effect of individual TF binding events but from synergistic effects of multiple TFs mediating the assembly of a higher order enhanceosome complex (2). Precise combinations of TFs as well as the orientation and spacing between TFs are therefore requirements for assembly of a transcriptionally active enhanceosome in this particular instance. However, a recent study in Drosophila suggests that none of the above requirements are prevalent in the majority of functionally validated enhancers. Instead, tightly controlled gene expression is postulated to be achievable using flexible spacing between TFs and redundancy in TF interaction (3).

Recent advances and improved cost-effectiveness in next generation sequencing technology have greatly increased the number of publicly available genome-wide TF binding profiles generated by chromatin immunoprecipitation coupled with sequencing (ChIP-seq). To date, hundreds of datasets exist in the public domain that have been generated by the haematopoiesis research community. We have previously described the development of the HAEMCODE compendium and web interface (4), which provides access for the wider scientific community to several hundred, carefully curated ChIP-seq datasets in mouse blood cells, thus enabling comparative analysis of datasets generated in mul-

[*]To whom correspondence should be addressed. Tel: +44 1223 336 829; Fax: +44 1223 762670; Email: bg200@cam.ac.uk

tiple different laboratories. Similar resources have also been generated for embryonic stem cells (5,6).

In mouse haematopoiesis, specific spacing between two DNA binding motifs has previously been reported to be functionally important. Examples include (i) E-box and GATA motifs separated by 9bp and bound by TAL1 and GATA factors (7) that are important for the transcriptional activity of several erythroid gene regulatory elements and (ii) the Ets and IRF motif-pair separated by 2bp which occurs in gene regulatory sequences associated with genes important for lymphoid development (8). At genome-scale, similar analyses have been conducted on ENCODE datasets to show that binding of TF pairs can be spatially constrained (9,10). To assess the relative prevalence of spatially constrained binding versus the more relaxed model of the TF collective, genome-scale studies coupled with comprehensive statistical analysis and experimental validation will be required. Given the pivotal importance of combinatorial TF interactions in driving cell fate choices (11–13), research in this area not only has the potential to reveal new mechanistic aspects of TF function, but also inform our understanding of cell lineage specification during mammalian development.

Community efforts such as the development of HAEMCODE provide powerful new platforms for generating novel hypothesis that will lead to a better understanding of TF function. In this paper, we performed a comprehensive analysis of motif co-occurrence making use of all TF-datasets present in the HAEMCODE ChIP-seq compendium. By modelling TF binding to DNA with position weight matrices (PWMs), we were able to systematically predict binding sites and quantify the spacing between motifs. To infer TF pairs interacting with DNA, we chose an unbiased approach by considering not just sequence motifs that correspond to the TF precipitated in a particular ChIP-seq experiment, but also all other sequence motifs that were significantly enriched in the *de novo* motif discovery in a given sample. In this manner, we were able to comprehensively map instances of motif co-occurrence and quantify short range distances (±100 bp) between any pair of TFs across a large number of TFs and cell types. Statistical analysis indicated that TF partner choices are not random but are instead closely linked to cell-type-specific function. Moreover, experimental validation confirmed the functionality of two previously unknown motif-pairs and their spacing, involving the pairing of Ets with E-box and Ets with Homeobox TFs respectively.

MATERIALS AND METHODS

HAEMCODE ChIP-seq data processing

Public mouse ChIP-seq datasets from blood related cell types were obtained from the NCBI Gene Expression Omnibus and the EMBL-EBI European Nucleotide Archive (total: 289, Supplementary Table S1A). Raw reads were downloaded from the public repositories and then converted to *fastq* format and assessed for quality control using the FastQC software (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Adapter sequences were removed using the trimGalore software (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Reads from samples that pass the quality control were then aligned to the mm10 genome using Bowtie2 (14) and peak called using MACS2 (15) at different stringencies (*P*-value between 1e−4 and 1e−15). A suitable *P*-value was selected based on visual inspection of the ChIP-seq profiles in the UCSC genome browser. Supplementary Figure S1 shows the distribution of the number of peaks per sample for the 289 ChIP-seq samples.

PWM similarity and clustering

Analysis of large numbers of DNA-binding motifs is difficult for several reasons. First, comparisons across hundreds of motifs would be impossible if some form of summarization is not employed. Second, motif databases are often redundant and since different TFs may bind the same motifs, they are often stored as separate motifs. To deal with the large scale and redundancy of motifs, we propose a method to cluster a collection of motif PWMs by their similarity. The web-based tool, STAMP (16) provide a solution to compare a set of query motifs against a motif database to obtain top PWMs that is most similar to each query motif. In contrast, to cluster a collection of motifs, an 'all against all' comparison is necessary. We developed a pipeline for performing PWM similarity clustering in R (http://www.r-project.org/) as follows. A distance score matrix, $\boldsymbol{M}$, was calculated for 240 Jaspar (17) PWMs where $\boldsymbol{M}$ is a matrix of size 240 × 240. The value for position $(\boldsymbol{a},\boldsymbol{b})$ in the matrix is the mean Euclidean distance of the overlapping columns of a pair of PWMs, $\boldsymbol{A}$ and $\boldsymbol{B}$. See supplementary methods for details on computing the matrix, $\boldsymbol{M}$. Hierarchical clustering was applied to the score matrix and the resulting dendogram was written into a Newick tree file format. The Dynamic Hybrid algorithm in the *Dynamic Tree Cut* R package (18) was used to detect clusters in the hierarchical clustering results. Finally, the Newick tree file format and cluster memberships were uploaded to Evolview (http://www.evolgenius.info/evolview/) (19) to generate a circular cladogram. Source code for the PWM similarity clustering pipeline is available on bitbucket (https://bitbucket.org/feliciang/publication-motif-pair).

De novo motif analysis and similarity to known motifs

The RepeatMasker program (http://www.repeatmasker.org) was used to assess the amount of repetitive elements in the 100bp sequences centred on the peak summit. Sequences containing >40% repeats were discarded and the remaining peak sequences were used for motif discovery. *De novo* motif analysis was carried out using the HOMER findMotifsGenome.pl program (20) and matches to known motifs were discovered using the TOMTOM software (21). Candidate *de novo* motifs were considered significant if they are present in at least 5% of the input sequences and have a p-value below the threshold of 1e−10. Matches to known motifs were discovered from the Jaspar v4 database (vertebrate CORE, FAM, PHYLOFACTS, POLII, CNE and SPLICE) (17) and only significant motifs (Q-value ≤ 0.05) were reported. In total, there were 1664 *de-novo* motifs and of these, 1152 had a significant match to Jaspar motifs.

Motif-pair discovery

Our approach to discovering motif-pairs and significant motif spacing can be divided into two main tasks as denoted by (I) and (II) below. The first part ('Scanning ChIP-seq peaks for known motifs') describes the procedure to identify high-confidence binding sites of not just the TF precipitated in each experiment but also binding sites of regulatory partners. The second part ('Finding motif-pairs and significant motif spacing') outlines the procedure to find pairs of binding sites predicted in (I) that are up to 100bp apart and to discover significant spacing between high-confidence binding sites. We developed a custom analysis pipeline to address the above points and applied it to ChIP-seq peak regions in the HAEMCODE datasets. This approach allowed us to interrogate motif pairs with and without significant spacing in regulatory regions of mouse blood cells.

(I) Scanning ChIP-seq peaks for known motifs. To obtain the genomic coordinates of true, high-confidence binding sites, the HOMER findMotifsGenome.pl program (20) was executed with the '–find' parameter on the 150bp region centred on the ChIP-seq peak summit. In contrast to the *de novo* motif analysis, we extended the search space by ±25 bp to take into account motifs that lie on the boundary of the 100 bp sequence. For each ChIP-seq sample, we scanned the peak sequences for Jaspar motifs identified as a significant match by TOMTOM in that sample so that only enriched motifs are taken into account. Collectively, all 289 samples encompass 354 enriched Jaspar motifs. We chose to use Jaspar motifs because PWMs of known motifs allowed comparisons to be made across datasets. The PWMs of *de novo* motifs, however, can appear in many 'flavours' (variations in PWM values for very similar sequence motifs) and presents a more difficult task for comparison.

When using the HOMER software to predict binding sites, two things needed to be considered for each PWM (i) 'core' motif and (ii) detection threshold. Firstly, the PWMs used to screen the input sequences are trimmed to obtain the 'core' motif. The 'core' motif is defined as the essential component of a motif that starts at the first position where information content (IC) $\geq$0.5 and ends at the last position where IC $\geq$0.5 (see Supplementary Methods for details on how to calculate IC). Therefore, PWMs are trimmed to obtain the motif 'core' by removing flanking regions with IC $<$0.5 and any motif 'core' $<$4 bp were discarded. The motif 'core' is then used for the screen rather than the full (original) motif to obtain a more accurate quantification of spacing between motifs. Without trimming, the flanking regions with low IC are treated as the full motif when screening and cannot be accounted for as true spacing (see Supplementary Figure S2). Second, a detection threshold value is calculated for each 'core' motif to predict high-confidence sites. See supplementary methods for details on calculating the detection threshold. For each ChIP-seq dataset, we scanned the 150bp peak sequences (centred on peak summit) using the 'core' motif and its detection threshold. Sequences that pass this threshold were considered high-confidence binding sites and used to find motif-pairs and motif spacing.

(II) Finding motif-pairs and significant motif spacing. First, the genomic locations for each Jaspar motif from the previous step ('Scanning ChIP-seq peaks for known motifs') with at least 50 occurrences in the genome were organized by Jaspar ID. By doing this, the motif regions are dissociated from the sample ID and unique coordinates were kept. By iterating over all pairwise combination of motifs, the offset, $\boldsymbol{s}$, between any two motifs was calculated using the BEDTools (22) function—windowBed—and then all motif-pair instances were grouped by its strand orientation (++, –, +–, –+) relative to the genome. For a pair of motifs—'query' and 'target', the offset value was obtained as follows. If the 'query' motif lies upstream of the 'target' motif, then the offset value is positive. On the other hand, if the 'query' motif lies downstream of the 'target' motif, then the offset value is negative. In both cases, offset is defined as the edge-to-edge distance between motifs and calculated starting from the first position flanking the 'query' motif to the first position (inclusive) of the 'target' motif (Supplementary Figure S2B). Without motif trimming, flanking regions may affect the offset value calculations.

The frequency, $\boldsymbol{f}$, of a particular motif-pair was then calculated for each offset value, $\boldsymbol{s} \in [-100, \ldots, -1, +1, \ldots, +100]$. Significance tests were performed separately for each motif-pair, relative orientation (++, –, +–, –+) and offset value. Then, the significance of the observed frequency, $\boldsymbol{f}$, at a particular offset, $\boldsymbol{s}$, was tested separately under the assumption that each offset value is independent and has no effect on neighbouring values. Let $\boldsymbol{N}$ be the total number of observed motif-pairs with a specific orientation. The significance of $\boldsymbol{f}$ was tested using the binomial distribution if $\boldsymbol{N} \leq 2000$ or the Poisson distribution if $\boldsymbol{N} > 2000$. Under the null hypothesis, a pair of motifs has no preferential spacing and, therefore, the expected probability of a motif-pair with a particular offset value (ranging from 1 to 100) in one of the four orientations is given by $\boldsymbol{q} = \frac{1}{4*100} = 0.0025$ and $\lambda = \frac{1}{100} \times N$, respectively. If there is no preference for motif spacing, the frequency of motif-pairs $\boldsymbol{f}$ with spacing $\boldsymbol{s}$ should follow a distribution

$$\boldsymbol{f} \sim \mathrm{Bin}(\boldsymbol{N},\ \boldsymbol{q}),\ \ \boldsymbol{N} \leq 2000$$

$$\boldsymbol{f} \sim \mathrm{Pois}(\lambda),\ \ \boldsymbol{N} > 2000$$

Significant results were filtered using a *Q*-value (23) (http://www.bioconductor.org/packages/release/bioc/html/qvalue.html) threshold of 1e−4. Logos of motif-pairs were generated using the WebLogo software (version 3.3) (24). 'Circos' plots (25) were used to display unique motif-pairs (multiple offset value not considered) by motif cluster (described in the 'PWM similarity and clustering' section).

Enrichment of motif-pair with significant spacing in ChIP-seq samples

To test the extent of overlap of motif-pair regions in HAEMCODE samples, the hypergeometric test was used to calculate the significance of the overlap. A ChIP-seq sample is significantly enriched for a motif-pair if the motif-pair genomic coordinates have a greater overlap with peak regions of a ChIP-seq sample than expected by chance. See supplementary methods for details on the hypergeometric test. Bonferroni correction (26) was applied to the p-value to correct for multiple motif-pairs and multiple ChIP-seq samples tested. For 7444 motif-pairs tested and 289 samples, we obtained a matrix of 7444 × 289 consisting of values 0 and 1: 1 denotes significance and 0 otherwise. A heatmap was generated in R using the *gplots* package (http://cran.r-project.org/package=gplots) by summarizing the matrix by cell type and motif-pair type. For each cell type category, we count the proportion of samples in a cell type that contain one or more motif-pairs in a specific motif-pair category (e.g. Ets + GATA) and coloured the heatmap cells based on this value. As a control cell type, we also performed the same enrichment analysis on 13 additional adipocyte ChIP-seq samples. Details of these 13 samples can be found in Supplementary Table S1B.

SNPs and Indels

Data in vcf format was downloaded from the Wellcome Trust Sanger Mouse Genome Project version 3 (27). We used vcftools (28) to extract SNP and Indel features that overlap TF bound regions (ChIP-seq peaks). Box plots were generated in R using the *ggplot2* package (29).

Candidate genes selection

For each candidate motif-pair, we searched for candidate genes for further testing as follows. Sequences of all motif-pair regions for a given candidate motif-pair were extracted in FASTA format and then aligned (without gaps) using the ClustalW2 software (30). New motif-pair PWMs were then calculated from the nucleotide frequencies of the aligned sequences. HOMER findMotifsGenome.pl was used to find the occurrences of motif-pairs based on the new motif-pair PWM. Finally, we mapped genomic regions of the motif-pairs to promoters of genes using annotation from MPromDb (31). For each candidate motif-pair, candidate genes were selected if they were expressed in any blood-related samples in the BioGPS dataset (32) and the conservation score (PhastCons or PhyloP) is within the top 10% amongst all mapped motif-pair occurrences. The BioGPS expression datasets was obtained from NCBI GEO (GSE10246). CEL files were processed by *gcrma* algorithm (33) in the gcrma R package. Conservation scores for 60-way Euarchontoglire multiple alignment were obtained from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/mm10/).

Transient luciferase assays

Wild-type and mutant DNA fragments for candidate regulatory regions (*Atf3*, *Cbfa2t3* and *Csf3* promoters) were obtained from GeneArt® by Life Technologies (see Supplementary Figure S3) and cloned into the pGL2 basic vector from Promega. K562 and 416b cell lines were transfected with the relevant vectors and a lacZ control vector by electroporation (220 V, 900 μF). Experiments were performed in triplicate and each experiment contained three technical replicates. The luciferase activity was analysed 24 h after transfection using the FLUOstar OPTIMA luminometer from BMG LABTECH. Significance was calculated by combining the *P*-values of each experiment (generated by using the *t*-test function in Excel) using the Fisher's method.

RESULTS

Public ChIP-seq datasets represent a rich resource for genome-wide motif-pair discovery

We previously reported the development of the HAEMCODE repository for curated public ChIP-seq datasets in mouse blood cells (4). This large collection of TF-binding maps is a rich resource for genome-wide analysis of gene regulation in haematopoiesis. Several hundred TF ChIP-seq samples covering all major blood cell types have been made publicly available on HAEMCODE for the scientific community. All datasets included in the HAEMCODE repository have been processed using a standardized ChIP-seq analysis pipeline (see 'Materials and Methods' section). At the start of this study, 289 ChIP-seq datasets (Supplementary Table S1A) covering 75 TFs across 14 blood cell types had been processed and so we focussed our analysis on this set of samples which in total corresponded to 528 545 genomic regions bound by at least one TF. For each ChIP-seq sample, we performed *de novo* motif analysis using HOMER (20) on the 100 bp sequences centred on the peak summit (Figure 1A). Enriched DNA motifs for each sample were then analysed using the TOMTOM program (21) to identify significant matches to known motifs in the Jaspar library (total of 913 PWMs) (17). Collectively, the enriched motifs in all 289 samples had significant matches to 354 Jaspar PWMs. These motif discovery results are the basis for the integrative analysis presented below.

The spacing between TF binding sites can be an important feature of TF co-occupancy. We reasoned that combining results from 289 ChIP-seq studies should allow us to perform a deep analysis of this phenomenon. Several studies in recent years have reported examples of preferred distance requirements between pairs of TFs binding on DNA using ENCODE datasets (9,10). Moreover, E-box and GATA binding sites separated by nine nucleotides have long been recognized as an important regulatory feature in erythroid and megakaryocytic cells (7). To examine TF co-occupancy and motif spacing preferences in the entire genome of haematopoietic cells, we developed a motif-pair discovery pipeline to identify motif co-occurrences and preferential distance(s) in all 528 545 TF-bound regions (Figure 1B). We used 'trimmed' Jaspar motifs (see methods) that are enriched in a particular sample to scan the TF-bound regions (150 bp centred on peak summits) of that sample and obtain genomic coordinates of motif occurrences. Of note, this procedure was performed independently for each sample using only those Jaspar motifs that were over-represented in the given sample in order to minimize the number of false positive results and to allow comparison across multiple datasets.

In total, 354 Jaspar motifs were used in the scan, so genomic coordinates from the scan were collated and organized by the 354 Jaspar IDs tested. Next, a comprehensive search was carried out for all combinations of motif-pairs and the relative orientations of the motif. For each motif-pair and orientation, the offset values between these two motifs (within ±100 bp) were calculated. Finally, all offset values from −100 to −1 and +1 to +100 were examined to identify distances that are over-represented and favoured by

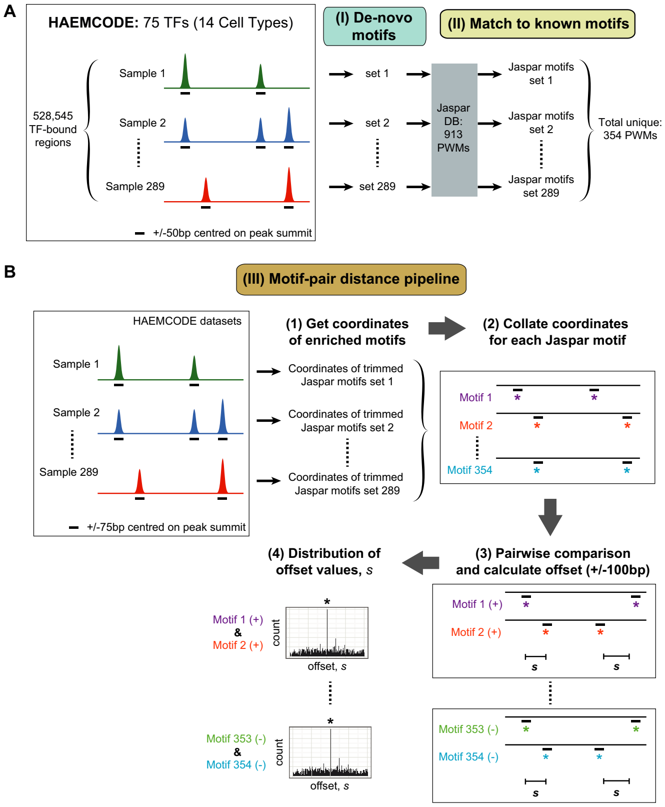
A
HAEMCODE: 75 TFs (14 Cell Types)
(I) De-novo motifs
(II) Match to known motifs
528,545 TF-bound regions
Sample 1
Sample 2
Sample 289
set 1
set 2
set 289
Jaspar DB: 913 PWMs
Jaspar motifs set 1
Jaspar motifs set 2
Jaspar motifs set 289
Total unique: 354 PWMs
+/-50bp centred on peak summit
B
(III) Motif-pair distance pipeline
HAEMCODE datasets
Sample 1
Sample 2
Sample 289
+/-75bp centred on peak summit
(1) Get coordinates of enriched motifs
Coordinates of trimmed Jaspar motifs set 1
Coordinates of trimmed Jaspar motifs set 2
Coordinates of trimmed Jaspar motifs set 289
(2) Collate coordinates for each Jaspar motif
Motif 1
Motif 2
Motif 354
(3) Pairwise comparison and calculate offset (+/-100bp)
Motif 1 (+)
Motif 2 (+)
s
Motif 353 (-)
Motif 354 (-)
(4) Distribution of offset values, s
Motif 1 (+) & Motif 2 (+)
Motif 353 (-) & Motif 354 (-)
count
offset, s

Figure 1. Workflow of this study. (**A**) Motif analysis procedure applied to 289 HAEMCODE datasets. *De novo* motif analysis was carried out separately for each sample in HAEMCODE. Each individual set of significant *de novo* motifs was then independently compared to the Jaspar collection of known motifs (total: 913 PWMs). Collectively, 354 unique PWMs were identified as significant matches within the 289 datasets. These unique PWMs were used in the motif-pair distance pipeline. (**B**) Procedure for identifying motif-pairs and over-represented distances between motifs. TF-bound regions (±150 bp centred on the peak summit) for each sample were individually scanned for the corresponding set of 'trimmed' Jaspar motifs so that only enriched motifs in that sample were considered (step 1). Then the coordinates of enriched motifs for all 289 samples were compiled and organized by motif ID (step 2). At this stage, the motif coordinates are dissociated from the sample ID. For all possible pairwise combinations of motifs and orientation, the offset values between motifs, ***s***, were calculated and significant offset values within ±100 bp were identified (steps 3 and 4).

the motif-pair (see 'Materials and Methods' section). Taking into account distinct motif-pairs, orientation and preferential offset values, 7444 significant results were obtained at a *Q*-value threshold of 1e−4 and these results (motif-pair logos and genomic coordinates) have been made available on this website—http://bioinformatics.stemcells.cam.ac.uk/publications/motifpair/motifpair.html. Processed ChIP-seq data, *de novo* motif and significant matches to Jaspar motifs for the 289 samples are also available and can be accessed from the HAEMCODE website—http://codex.stemcells.cam.ac.uk/.

Preferred spacing of TF binding site pairs is prevalent in haematopoietic TF-bound regions

By integrating Jaspar motif analysis results from the 289 HAEMCODE TF-binding maps, we were able to study the TF co-occupancy patterns and preferential spacing that are relevant to haematopoiesis. In total, we analysed 528 545 regions that are bound by at least one TF in any of the 289 datasets (overlapping regions were merged). Out of all these regions, 435 558 (82.4%) regions contain an over-represented motif and 364 911 (69%) regions contain motif-pairs (Figure 2A). This suggests that a majority of experi-

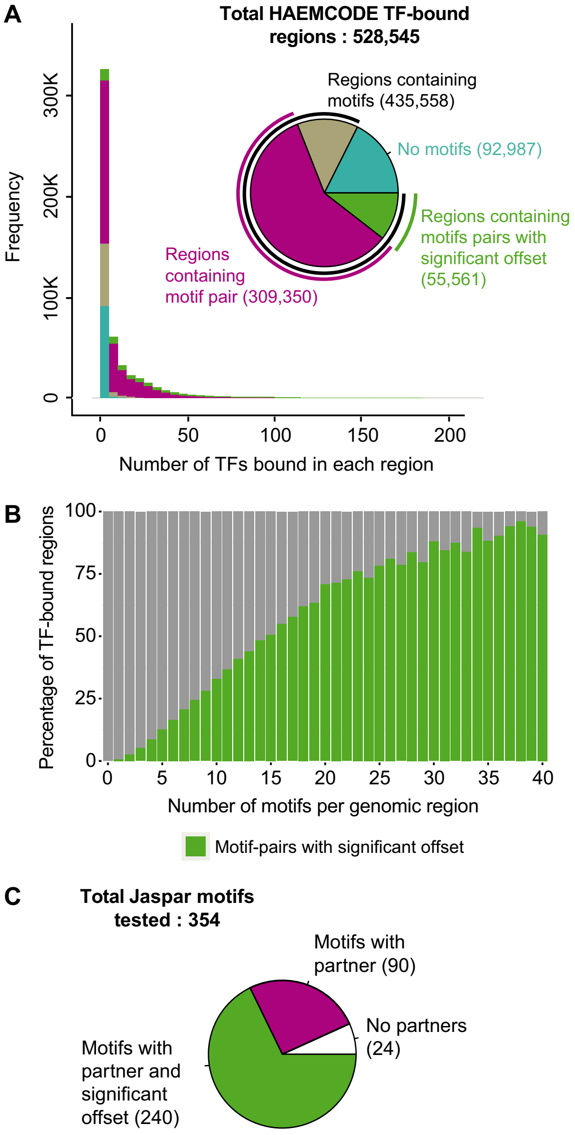

Figure 2. Motif-pairs with preferential spacing are prevalent in haematopoietic TF-bound regions. (**A**) The pie chart shows all genomic regions in this study bound by at least one TF and the proportion of regions containing motifs or motif-pairs (with and without significant spacing). The histogram shows the distribution of the number of TFs (binwidth = 5) in each genomic region and each bar is coloured to indicate the presence or absence of motifs and motif-pairs (with and without significant spacing). (**B**) Barchart indicates the percentage of TF-bound regions (*y*-axis) that contain motif-pairs with significant offset. Percentages are shown for TF-bound regions containing zero up to 40 motifs (*x*-axis). (**C**) All unique Jaspar motifs identified as significant matches to *de novo* motifs in Figure 1A (step II). These motifs were used to scan TF-bound regions for binding sites and the pie chart highlights the proportion that participates in motif-pairs (with and without significant spacing).

mentally verified TF binding regions are most likely also bound by another TF. Of the 364 911 regions containing motif-pairs, 55 561 regions (15.2%) contain motif-pairs that have a recurring, significantly over-represented preferential spacing between the two motifs. Although a large proportion of the TF-bound regions were only bound by one TF (218 332, 41.3%), motif-pairs with significant spacing were also discovered in these regions (4170, 1.9%). This result indicates that experimental data for certain TFs are still missing and, therefore, not yet included in the compendium. Interestingly, we also found that 17.6% (92 987 regions) of all TF-bound regions contain no match to any of the 354 motifs analysed. The vast majority (92.5%, 86 056) of these 'motif-less' TF-bound regions was bound by just a single TF.

TF-bound regions with more motifs are more likely to contain motif-pairs with significant offset (Figure 2B) suggesting that preferential spacing is frequently found in regions where many TFs co-operate. A large proportion of regions that contain motif-pairs with preferential spacing are located within intragenic regions (24 751, 38.1%), but a large proportion is also found in intergenic regions (23 578, 36.3%). Motif-pairs are also common in promoter regions (10 755, 16.6%), especially considering that only 5.4% of all TF peak regions overlap with promoters (Supplementary Figure S4A and B). A small proportion of motif-pair regions are found within UTRs and exons (Supplementary Figure S4B). When compared to data from the Vista Enhancer Browser (34), 13.6% (172/1261) of known enhancers in this database overlapped motif-pair regions with constrained spacing from our dataset. Presumably the overlap is not larger because the Vista database contains a relatively small set of validated enhancers mainly related to brain, muscle and limb development. In total, 354 Jaspar motifs were enriched in at least one sample and of these, 330 motifs (93.2%) were found to participate in a pair (Figure 2C). Of these 330 motifs, 240 motifs (72.7%) contributed to motif-pairs with significant preferential spacing and this subset covers all TF classes in the Jaspar library (Supplementary Figure S4C). Overall, the distribution of all significant offset values shows a symmetric distribution with preferential offset values of ∼±1 and ∼±11 bp (Figure 3A); with the former consistent with directly adjacent binding of two TFs, whereas the latter corresponds to one helical turn around the DNA. Furthermore, motif-pairs appear to require a strict spacing with the majority of pairs having just one preferred offset value (3216, 69.1%) (Figure 3B).

To facilitate interpretation of this large number of results, we next clustered the Jaspar PWMs into groups of similar motifs. Jaspar's collection of motifs is highly redundant and often more than one PWM exists for the same type of binding site. The 240 motifs that participate in a motif-pair with preferential spacing were processed through a newly developed clustering procedure (see methods for algorithm and link to R source code) to obtain 19 clusters of similar PWMs (Supplementary Figure S5). Thirty four PWMs did not fall into any of the 19 clusters of major motif classes and were removed from further analysis, but are available in Supplementary Table S3. By grouping all significant motif-pairs by the cluster they belong to, we were able to display all the interactions between motif clusters in a 'circos' plot (Figure 3C). The ribbon thickness in the plot is proportional to the number of unique motif-pairs (multiple offset value not counted) found between the two clusters of motifs and some of the most abundant types of motif-pairs are shown by thick ribbons. This also reflects the abundance of certain types of motif-pairs in haematopoietic transcription. Examples of heterotypic motif-pairs (two different motifs) include 'Ets + Homeobox' (182 unique pairs), 'Cebp + Homeobox' (71 unique pairs), 'CTCF + Homeobox' (67 unique pairs) and 'Ets + zinc-coordinating' (86 unique pairs). Partnering between motifs from the same cluster (e.g. homotypic motif-pairs), is also possible and this is shown by semi-circles within the same cluster of motif. For example, the 'Ets + Ets' type of motif-pairs constitute ∼15.1% of all the motif-

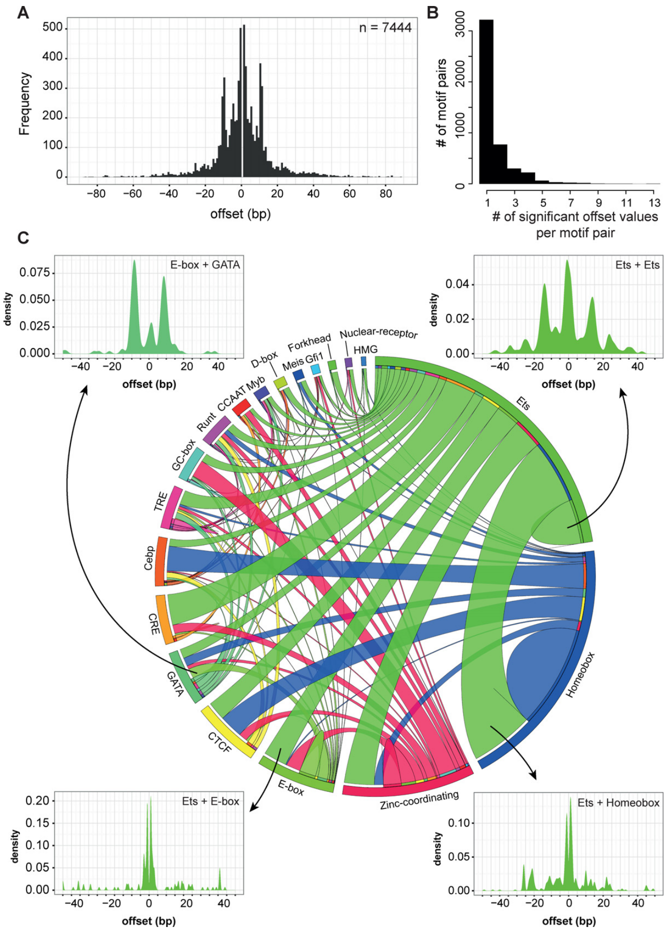

Figure 3. Motif-pairs with significant spacing. (**A**) Distribution of all significant offset values for all motif-pairs and motif orientations. (**B**) Number of significant offset values per motif-pair. The number of offset value refers to the total unique offset values for a specific motif-pair of a specific orientation. (**C**) 'Circos' plot showing all unique motif-pairs and orientations with significant offset value grouped by motif cluster. Clusters were defined by similarity and clustering of Jaspar PWMs (see Supplementary Figure S5) and arranged on the plot based on the number of pairs per cluster. Ribbon thickness is proportional to the number of unique motif-pairs (multiple offset values not counted) belonging to the clusters. Data used to generate 'circos' plot can be found in Supplementary Table S2. Example offset density plots in the range of ±50 bp are shown for four types of motif-pairs to illustrate distinct preferences for spacing between motifs: 'E-box + GATA' (28 unique pairs), 'Ets + E-box' (67 unique pairs), 'Ets + Ets' (126 unique pairs), and 'Ets + Homebox' (182 unique pair).

pairs involving an Ets motif, while ‘Homeobox + Homeobox’ make up 42% of all Homeobox motif-pairs.

It is also worth noting that four motif clusters—Ets, Homeobox, Zinc-coordinating and E-box, have high numbers of interacting partners and make up >50% of all motif-pairs. For example, Ets motifs can form pairs with all the other clusters of motifs and with itself while, in contrast, CRE motifs form pairs almost exclusively with Ets motifs. When examining specific motif-pairs, we found distinct preferences in the significant offset values. As shown in the four example density plots (Figure 3C), the ‘E-box + GATA’ pair are separated by ∼9 bp, ‘Ets + E-box’ pairs are separated by ∼1 bp, ‘Ets + Ets’ are separated by ∼1 or ∼15 bp, and ‘Ets + Homeobox’ are separated by 1–3 bp. Importantly, this analysis also demonstrated the consistency of our results despite utilizing different ‘flavours’ of the same motifs in the Jaspar library because the same offset values are frequently significant across independent motifs within each cluster pair. Since we analysed the motifs in relation to genome orientation, our results are symmetric in that offset values that were significant in the ‘plus’ strand were also significant in the ‘minus’ strand. The full results for all heterotypic and homotypic motif-pairs and their preferential spacing can be found in Supplementary Figure S6.

TF-bound regions containing motif-pairs are functionally important

Having shown that the spacing between motif-pairs is commonly constrained, we next made use of the organization of results by motif clusters to determine patterns that are consistent and therefore most likely to be biologically relevant. To investigate further the functional role of motif-pairs and their preferred spacing, we examined single nucleotide polymorphism (SNP) and insertion/deletion (Indel) data from the genomes of 18 key mouse strains (27). We reasoned that if motif-pairs with preferential spacing are functionally important, these regions are under stronger selective pressure compared to regions without motif-pairs. Indeed, we found that regions containing one or more overlapping motif-pairs have fewer SNPs and Indels than regions without motif-pairs (Figure 4A (i) and (ii)). Moreover, there are fewer SNPs/Indels found in regions in which more motif-pairs are present. The conservation score on these regions also shows that the number of motif-pairs is directly proportional to higher conservation score (Figure 4A (iii)). Of note, this trend was not observed when we examined TF-bound regions containing zero up to ten motifs regardless of the number of motif-pairs with constrained spacing (Supplementary Figure S8A). To test the significance of the observed trend, we used the Kolmogorov–Smirnov test to compare the distribution of values (SNP, Indel, PhyloP) in TF-bound regions containing 0 motif pairs against TF-bound regions containing 1 or more motif-pairs. Results showed that the two distributions are indeed very different (Supplementary Figure S8B) and, therefore, represent two different types of regulatory regions where regions with more motif-pairs are more conserved. These results suggest that regulatory regions that contain spatially constrained motif-pairs are less likely to be mutated, and therefore are likely to be particularly important for haematopoietic gene regulation.

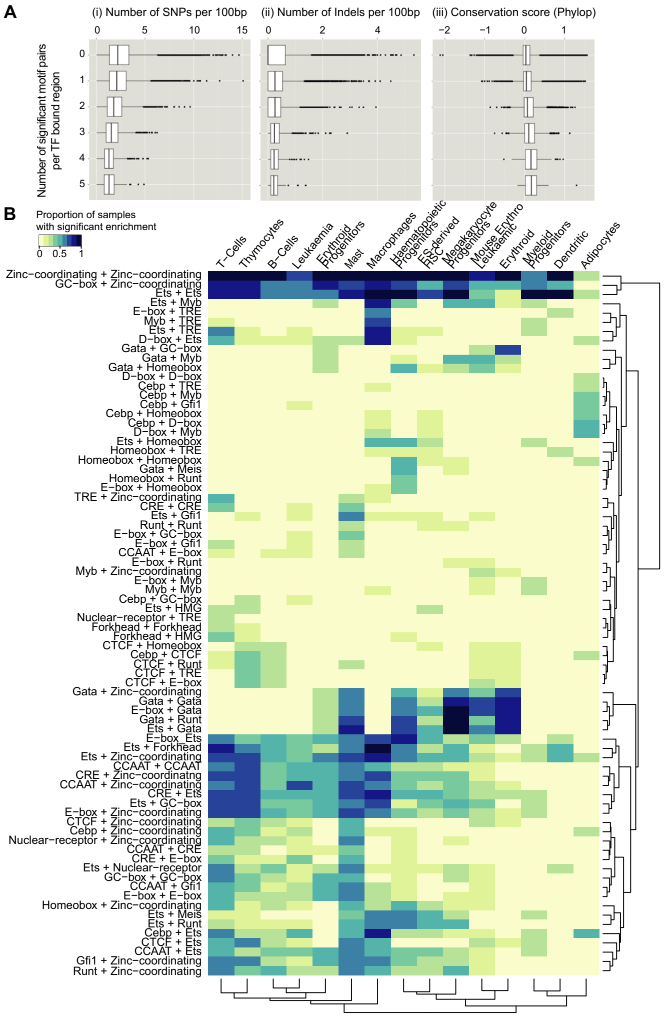

Figure 4. Properties of motif-pairs and orientation with significant spacing. (**A**) (i) Box plot showing the distribution of the number of SNPs overlapping TF-bound regions (ChIP-seq peaks) containing zero to five motif-pairs. (ii) Same as in (i) but for Indels. (iii) Same as in (i) but for PhyloP conservation scores. Positive PhyloP scores indicate conservation in 60way Euarchontoglire multiple alignment. (**B**) Heatmap of significantly enriched motif-pairs by cell type. Pairs of motifs were grouped by motif clusters they belong to (see Supplementary Figure S5). For a particular pair of motif cluster, row elements are coloured based on the proportion of samples in each cell type that are enriched for at least one motif-pair in that category. Only motif-pair categories that are significantly enriched in ≥20% samples in at least one cell type are shown. Dark blue cells denote higher proportion; light yellow cells denote lower proportion.

We also investigated the role of motif-pairs with preferential spacing in relation to cell-type-specific functions in blood development by analysing the enrichment of motif-pairs in peak regions of all 289 samples. For each motif-pair with significant offset value, we tested each of the 289 samples independently for over-representation of the motif-pair in the sample and summarized the results in a heatmap (Figure 4B). The 289 samples were categorized by cell types (heatmap columns) and motif-pairs were categorized by motif clusters (heatmap rows) to illustrate the proportion of samples in a cell-type enriched with a motif-pair. The heatmap revealed several interesting patterns about the relationships between motif-pairs and stages in blood development. Three motif-pairs—'zinc coordinating + zinc coordinating', 'GC-box + zinc coordinating', and 'Ets + Ets' (heatmap rows 1–3)—are significantly enriched across all haematopoietic cell types. In contrast, five pairs involving GATA motifs—'GATA + zinc coordinating', 'GATA + GATA', 'E-box + GATA', 'GATA + Runt' and 'Ets + GATA' (heatmap rows 46–50)—are more cell-type-specific, as shown by their enrichment in haematopoietic progenitors as well as cells of the erythroid and myeloid lineage. Moreover, when we performed the same analysis on adipocyte ChIP-seq samples, blood-related motif-pairs have very low enrichment in this cell type. To further examine the cell type-specific properties of motif-pairs, we explored an alternative visualization of the motif-pairs enriched in different cell types by using 'circos' plots to display the abundance of motif-pairs in each cell type (Supplementary Figure S7). We find that 'Homeobox + Homeobox' motif-pairs are predominantly found in multipotent haematopoietic progenitors and ES-derived HSPCs. In all cell types, CTCF is an important TF binding partner as shown by the ribbons extending from CTCF to several other motifs. We also observed that motif-pairs involving 'GATA' are mostly found in haematopoietic progenitors and cells of the erythroid and myeloid lineage.

Exact spacing between motifs is important for transcriptional activity

To investigate the role of strict spacing between TFs, we selected three candidate motif-pairs: 'E-box + GATA', 'Ets + E-box', and 'Ets + Homeobox' (Figure 5A). We selected these motif-pairs because 'E-box + GATA' is a well-known motif-pair in haematopoiesis, while the latter two are previously unknown motif-pair spacings. Moreover, these three candidates are among the most frequently occurring results. As shown by the density plots in Figure 3C, these motif-pairs favour ~9, ~1 and ~3bp offset values, respectively. GREAT analysis on these motif-pair regions showed enrichment of blood-related processes, phenotypes and diseases (see Supplementary Table S4).

We chose TF-bound regions for functional validation by performing a screen on all HAEMCODE regulatory regions for occurrences of the candidate motif-pairs using the newly generated PWMs for each candidate motif-pair (Figure 5B). We calculated each motif-pair PWM from the aligned sequences of motif-pair regions and the new probability matrices can be found in Supplementary Table S5. Candidate genomic regions were prioritised based on two criteria: (i) high evolutionary sequence conservation and (ii) expression of the adjacent gene in one or more blood-related cell types. Following the screen, we chose the promoter regions of the following three genes for further analysis: *Cbfa2t3*, *Atf3* and *Csf3* (Figure 6A). The *Cbfa2t3* transcriptional co-repressor interacts with TAL1 to regulate erythrocyte differentiation (35,36) while *Atf3* is a key transcriptional repressor of target genes in lipopolysaccharide-stimulated macrophages (37). Macrophage progenitor function has also been shown to be dependent on *Csf3* because

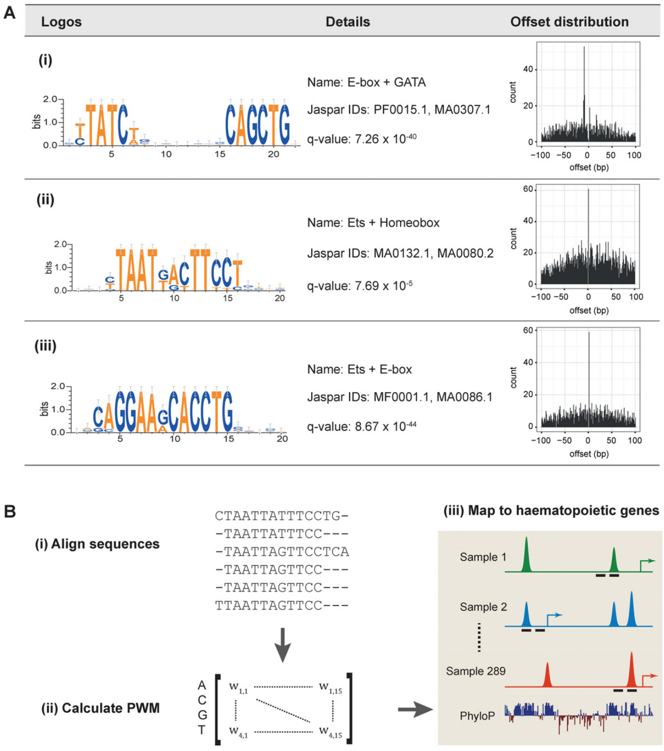

Figure 5. Candidate motif-pairs. (**A**) Logos, details and offset distribution of three candidate motifs. Each candidate was named based on the motif clusters it belongs to. (**B**) Procedure for identifying candidate genes for functional analysis. For each candidate motif-pair, a PWM was calculated based on the alignment of all the corresponding motif-pair regions. The new motif-pair PWMs were then used to map to conserved haematopoietic genes (see 'Materials and Methods' section).

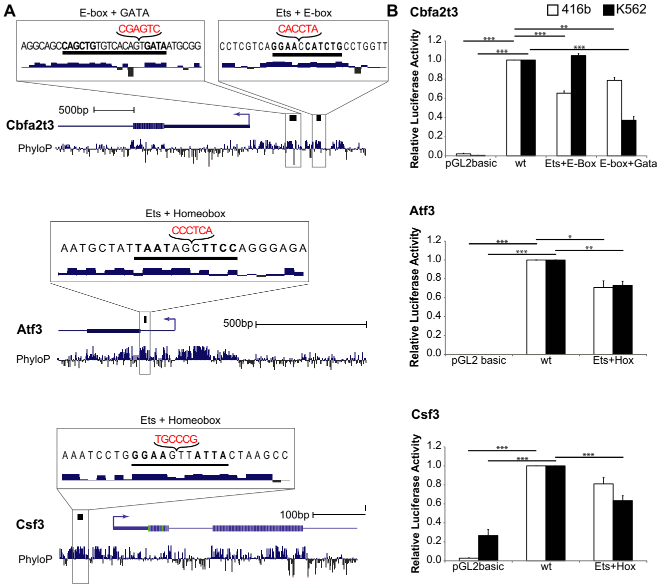

Figure 6. Functional validation of candidate genes. (**A**) UCSC genome browser tracks showing promoter regions of Cbfa2t3, Atf3 and Csf3 genes. The candidate motif-pairs for each gene are highlighted on top of each gene and enlarged in the genome browser track inset. The nucleotides that were introduced in order to disrupt the spacing between the two motifs are shown in red. (**B**) Luciferase assays in transiently transfected 416b or K562 cells. Results indicate the average and standard error of the mean of three independent repeats and are shown relative to the luciferase activity of the wild-type promoter (*P*-values: * ≤0.05, ** ≤0.01, *** ≤0.001).

mice with *Csf3* deficiency produced lower levels of granulocyte and macrophages compared to mice without the deficiency (38). To investigate the importance of motif spacing, we generated luciferase reporter constructs containing wild-type and motif-spacing mutant versions of these three promoters. Motif-pair spacing was disrupted by introducing a 6bp random sequence (see 'Materials and Methods' section) and the random sequences were designed without generating additional motifs while maintaining the flanking nucleotides, and in all cases maintained at least one nucleotide flanking the motifs. Retaining one nucleotide may not be sufficient in situations where the offset is more than one base pair, suggesting that the design of the random sequence may be improved in future analysis by retaining two or more nucleotides. Luciferase reporter gene assays demonstrated that the 9 bp spacing between the GATA and E-box motifs within the *Cbfa2t3* promoter region is important for transcriptional activation in human (K562) as well as murine (416 bp) cell lines because disruption of the 9 bp spacing caused a significant reduction in the luciferase activity in both cell lines (Figure 6B). These results established the *Cbfa2t3* promoter region as a new example of haematopoietic regulatory elements that depend on the 9bp-spaced 'E-box + GATA' motif-pair. Importantly, similar analysis of the newly identified motif-pairs between Ets and Homeobox as well as Ets and E-box motifs revealed important functions for gene activation of *Atf3*, *Csf3* and *Cbfa2t3* in mouse and/or human cells. Indeed, luciferase activity of motif-spacing mutants was significantly reduced for all tested promoter regions in at least one of the two cell lines (Figure 6B). In the *Atf3* promoter region, luciferase activity for the Ets and E-box pair did not change in K562 cells although a significant reduction was observed in 416 bp cells, therefore suggesting that Ets and E-box TFs may not form a complex on the *Atf3* promoter in K562 cells. Luciferase assays therefore provided functional validation of two previously unrecognized constrained motif-pairs, namely the 'Ets + Homeobox' and 'Ets + E-box' motif-pairs. Going back to the HAEMCODE compendium, we also found evidence for binding of E-box, GATA and Ets TFs on the motif-pair regions on the *Cbfa2t3* gene locus (Supplementary Figure S9).

DISCUSSION

The widespread uptake of ChIP-seq technology has undoubtedly improved our understanding of multiple aspects of TF function in mouse haematopoiesis, such as combinatorial aspects of transcriptional control in HSCs (39), function of key regulators such as Gata3 across the T-lymphoid lineage (40), early acquisition of blood progenitor transcriptional programs (41,42) and the identification of different TF binding modes for a given factor such as TAL1 (43). Focussed studies by individual laboratories are complemented by large consortia efforts such as ENCODE (44), which also generate significant numbers of haematopoietic TF ChIP-Seq datasets. Collectively, these datasets provide a powerful resource for performing complex integrated analysis and thereby uncovering new insights into the role of TFs in cellular development.

For this paper, we made use of the HAEMCODE ChIP-seq compendium and its large collection of uniformly processed haematopoietic TF ChIP-seq datasets. In contrast to individual projects which provide snapshots of specific developmental time-points, the HAEMCODE compendium integrates large numbers of these snapshots to provide coverage across a wide range of blood cell development. Following identification of pairs of co-occurring motifs, we were therefore able to interrogate co-operative binding events across different cell types. We therefore not only report lists of preferential motif-pairs, but also identify specific motif-pair combinations that are associated with single or subsets of blood cell types. Moreover, we demonstrate that spatially-constrained haematopoietic motif-pairs display elevated levels of sequence conservation, and that two previously unrecognized motif-pairs are critical for full promoter activity when tested by reporter gene assays.

DNA sequence motifs can be represented using a variety of approaches ranging from simple IUPAC strings to position weight matrices, and also more complicated statistical models of TF binding affinities (e.g. dinucleotide weight matrices, K-mers). PWMs are intuitive representations of nucleotide preferences for each base position, involve few parameters and in most cases outperform these alternative models (45). Compared with simple IUPAC strings however, working with PWMs can be complicated by the fact that multiple very similar versions may be present in motif collections such as JASPAR, or be discovered in *de novo* motif discovery applications. Further analysis such as the identification of specific motif-pairs reported here therefore faces the significant problem of many redundant results, where result files are dominated by motif-pairs that consist of virtually identical component motifs. To address this issue, we implemented a new pipeline for PWM similarity and clustering analysis, which enabled us to discover new biologically meaningful and consistent constrained motif-pairs. The use of Jaspar motifs for TF binding site discovery may have excluded pairs involving currently uncharacterised *de novo* motifs or specific preferences in the motif flanking positions, suggesting that future analysis may include those *de novo* motifs without a match to Jaspar motifs. Of note, our current analysis might have recovered some instances of preferential flanking positions not present in Jaspar as new constrained motif-pairs, in line with the notion that an extended motif consensus sequence may indicate the binding of multiprotein complexes.

Despite the rapid increase in the number of published ChIP-seq datasets, the task of correctly identifying functionally significant binding events and predicting transcriptional regulatory mechanisms has proved difficult. Our analysis of all HAEMCODE TF-bound regions revealed that a significant fraction of peaks contained no match to any of the 354 motifs obtained by *de novo* motif discovery of all the individual ChIP-Seq peak lists in HAEMCODE. Furthermore, these 'motif-less' peaks predominantly corresponded to regions with only one TF bound. Although it has been reported previously that TF binding can occur in regions without any enriched sequence motifs, we would argue that at least some of these regions may represent false positives identified by the peak calling algorithm. Systematic bias in the form of low complexity regions, repetitive sequences, mis-assembled reference genome and unannotated regions can contribute to false positive ChIP-seq peaks (46,47). Another plausible explanation for some of the 'motif-less' peaks is DNA looping, where regions without motifs were in close vicinity to regions with motifs, and therefore have become cross-linked in the ChIP-Seq protocol. Nevertheless, the observation that 'motif-less' peaks are predominantly found in regions only bound by a single TF in the compendium would argue that many of them represent false positives, especially as most TF ChIP-Seq samples so far do not include biological replicates. We would argue therefore that (i) future ChIP-Seq studies should include biological replicates to permit identification of high-confidence binding peaks for each individual TF, and (ii) analysis of existing datasets will be more robust when focussing on peak regions bound by at least two TFs. Nonetheless, our current understanding of 'functionality' in relation to TF binding to DNA remains limited (48), and a substantial fraction of reproducible binding events may represent 'opportunistic' binding to DNA sequences that happen to be readily accessible.

It is widely accepted that key regulatory TFs function as components of co-operative multimeric complexes sometimes referred to as an enhanceosome. Assembly of multiple TFs and accessory proteins on a given DNA sequence may require precise spacing between the binding sites to facilitate TF interactions on contiguous segments of the DNA. Here, we examined the promoter of the *Cbfa2t3* gene, which encodes a major regulator of blood cell development (35,36). The Cbfa2t3 promoter contains two motif-pairs with preferential spacing—'E-box + GATA' (offset 9bp) and 'Ets + E-box' (offset 1 bp). Both motif-pairs are present within a 79 bp contiguous stretch, and disrupting the spacing of each motif-pair significantly reduces transcriptional activity. Incorporation of six random nucleotides increases spacing by one half turn of the double helix so that each TF now binds opposing faces of the double helix and thereby allowing us to test the synergistic effect of co-operating TFs. Functional analysis of constrained motif-pairs in two additional haematopoietic promoters not only supported the importance of TF interaction in mediating transcriptional activation, but also validated two previously unknown haematopoietic motif-pairs with constrained spacing, namely the 'Ets + E-box' (offset 1bp) and 'Ets + Homeobox' (offset 3 bp) motif-pairs. When designing the 6bp insertion, we took into account the binding energies of the bases immediately adjacent to the predicted binding sites by maintaining at least one nucleotide flanking the motifs in the 6bp insertion. In the case of 'Ets + Homeobox' and 'E-box + Gata' motif-pairs, the spacing is >1 bp and, therefore, an improved design of the inserted random sequence may be considered for future analysis.

Interestingly, our analysis of HAEMCODE datasets revealed that only a small subset of all possible pair-wise combinations of motifs showed evidence for constrained spacing. We found 240 motifs that participated in pairs with preferential spacing, which theoretically could give rise to 28 920 possible combinations of motif-pairs (240 homotypic + 28 680 heterotypic pairs). However, only 2303 motif-pairs were actually found to display preferences for a strict spacing. These observations suggest that a subset of TF combinations is particularly relevant for transcriptional control during haematopoietic development. Preferences for motif-pair combinations also extended to cell type-specificity within the haematopoietic hierarchy in that certain combinations are more prevalent than others in distinct cell types. The seemingly sparse 'circos' plot for the dendritic cell type is likely the consequence of small sample numbers for this cell type, which will be improved by increasing the coverage of different TFs in dendritic cell samples. Of note, the HAEMCODE compendium is an ongoing effort to add newly published datasets and to increase the coverage across different blood cell types (49). Features of motif-pairs in our dataset also revealed four clusters (Ets, Homeobox, zinc-coordinating, and E-box) that were highly promiscuous and formed many pairs with most (if not all) of the other motif clusters while, in comparison, the remaining 15 motif clusters had smaller numbers of interacting partners. In some ways, this observation is analogous to 'hubs' in regulatory networks, and suggests that a subset of 'hub TFs' may be able to recruit a wide range of other TFs during the assembly of multimeric TF complexes on DNA.

Binding by multiple TFs to a given region has previously been proposed as a criterion for selecting the most biologically relevant ChIP-seq peaks (50,51). Here, we provide evidence for widespread sequence-specific DNA binding of TF complexes with constrained spacing on haematopoietic gene regulatory elements. Our findings contrast with the recently proposed 'TF collective' model (3), where co-occupancy of Drosophila cardiac enhancers was reported not to require specific spacing or orientation of binding sites. Instead, we would argue that both variable and constrained spacing are prevalent in mouse haematopoietic gene regulatory sequences. Moreover, we demonstrate that TF-bound regions containing spatially constrained motif-pairs display elevated sequence conservation both within and across species. This observation is consistent with important transcriptional functions in haematopoietic gene regulation, which we validated using three different promoters and their corresponding motif-spacer mutants in two different cell line models.

Taken together, we highlight specificity of interacting partners, potentially constrained spacing, and cell type-selectivity as important properties of combinatorial transcriptional control processes, with likely roles in the establishment of distinct cell type identities within the haematopoietic system and beyond.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Leukaemia and Lymphoma Research [12029]; the Medical Research Council [g0900951]; Cancer Research UK [c1163/A12765]; National Institute for Health Research Cambridge Biomedical Research Centre [RP-PG-0310-1002]; Biotechnology and Biological Sciences Research Council [BB/I00050X/1]; Wellcome Trust to the Cambridge Institute for Medical Research [100140/Z/12/Z] and Wellcome Trust & MRC Cambridge Stem Cell Institute

[097922/Z/11/Z]; Yousef Jameel scholarship [to F.S.L.N.]. Funding for open access charge: Wellcome Trust & MRC Cambridge Stem Cell Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Klemm,J.D., Schreiber,S.L. and Crabtree,G.R. (1998) Dimerization as a regulatory mechanism in signal transduction. *Annu.Rev. Immunol.*, **16**, 569–592.
2. Maniatis,T., Falvo,J.V., Kim,T.H., Kim,T.K., Lin,C.H., Parekh,B.S. and Wathelet,M.G. (1998) Structure and function of the interferon-beta enhanceosome. *Cold Spring Harbor Symp. Quant. Biol.*, **63**, 609–620.
3. Junion,G., Spivakov,M., Girardot,C., Braun,M., Gustafson,E.H., Birney,E. and Furlong,E.E. (2012) A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell*, **148**, 473–486.
4. Ruau,D., Ng,F.S., Wilson,N.K., Hannah,R., Diamanti,E., Lombard,P., Woodhouse,S. and Gottgens,B. (2013) Building an ENCODE-style data compendium on a shoestring. *Nat. Methods*, **10**, 926.
5. Martello,G., Sugimoto,T., Diamanti,E., Joshi,A., Hannah,R., Ohtsuka,S., Gottgens,B., Niwa,H. and Smith,A. (2012) Esrrb is a pivotal target of the Gsk3/Tcf3 axis regulating embryonic stem cell self-renewal. *Cell Stem Cell*, **11**, 491–504.
6. Halbritter,F., Vaidya,H.J. and Tomlinson,S.R. (2012) GeneProf: analysis of high-throughput sequencing experiments. *Nat. Methods*, **9**, 7–8.
7. Wadman,I.A., Osada,H., Grutz,G.G., Agulnick,A.D., Westphal,H., Forster,A. and Rabbitts,T.H. (1997) The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *EMBO J.*, **16**, 3145–3157.
8. Li,P., Spolski,R., Liao,W., Wang,L., Murphy,T.L., Murphy,K.M. and Leonard,W.J. (2012) BATF-JUN is critical for IRF4-mediated transcription in T cells. *Nature*, **490**, 543–546.
9. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y. *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.
10. Jankowski,A., Szczurek,E., Jauch,R., Tiuryn,J. and Prabhakar,S. (2013) Comprehensive prediction in 78 human cell lines reveals rigidity and compactness of transcription factor dimers. *Genome Res.*, **23**, 1307–1318.
11. Pimanda,J.E., Chan,W.Y., Wilson,N.K., Smith,A.M., Kinston,S., Knezevic,K., Janes,M.E., Landry,J.R., Kolb-Kokocinski,A., Frampton,J. *et al.* (2008) Endoglin expression in blood and endothelium is differentially regulated by modular assembly of the Ets/Gata hemangioblast code. *Blood*, **112**, 4512–4522.
12. Wilson,N.K., Timms,R.T., Kinston,S.J., Cheng,Y.H., Oram,S.H., Landry,J.R., Mullender,J., Ottersbach,K. and Gottgens,B. (2010) Gfi1 expression is controlled by five distinct regulatory regions spread over 100 kilobases, with Scl/Tal1, Gata2, PU.1, Erg, Meis1, and Runx1 acting as upstream regulators in early hematopoietic cells. *Mol. Cell. Biol.*, **30**, 3853–3863.
13. Elcheva,I., Brok-Volchanskaya,V., Kumar,A., Liu,P., Lee,J.H., Tong,L., Vodyanik,M., Swanson,S., Stewart,R., Kyba,M. *et al.* (2014) Direct induction of haematoendothelial programs in human pluripotent stem cells by transcriptional regulators. *Nat. Commun.*, **5**, 4372.
14. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
15. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
16. Mahony,S., Auron,P.E. and Benos,P.V. (2007) DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput. Biol.*, **3**, e61.
17. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C.Y., Chou,A., Ienasescu,H. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–147.
18. Langfelder,P., Zhang,B. and Horvath,S. (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, **24**, 719–720.
19. Zhang,H., Gao,S., Lercher,M.J., Hu,S. and Chen,W.H. (2012) EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. *Nucleic Acids Res.*, **40**, W569–572.
20. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
21. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.
22. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
23. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 9440–9445.
24. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
25. Krzywinski,M., Schein,J., Birol,I., Connors,J., Gascoyne,R., Horsman,D., Jones,S.J. and Marra,M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.
26. Bonferroni,C.E. (1936) Teoria statistica delle classi e calcolo delle probabilità. *Pub. R Ist. Super. Sci. Econ. Commerc. Firenze*, **8**, 3–62.
27. Keane,T.M., Goodstadt,L., Danecek,P., White,M.A., Wong,K., Yalcin,B., Heger,A., Agam,A., Slater,G., Goodson,M. *et al.* (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, **477**, 289–294.
28. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
29. Wickham,H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.
30. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
31. Gupta,R., Bhattacharyya,A., Agosto-Perez,F.J., Wickramasinghe,P. and Davuluri,R.V. (2011) MPromDb update 2010: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-seq experimental data. *Nucleic Acids Res.*, **39**, D92–97.
32. Wu,C., Orozco,C., Boyer,J., Leglise,M., Goodale,J., Batalov,S., Hodge,C.L., Haase,J., Janes,J., Huss,J.W. 3rd *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**, R130.
33. Wu,Z., Irizarry,R.A., Gentleman,R., Martinez-Murillo,F. and Spencer,F. (2004) A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.*, **99**, 909–917.
34. Visel,A., Minovitsky,S., Dubchak,I. and Pennacchio,L.A. (2007) VISTA enhancer browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, D88–D92.
35. Goardon,N., Lambert,J.A., Rodriguez,P., Nissaire,P., Herblot,S., Thibault,P., Dumenil,D., Strouboulis,J., Romeo,P.H. and Hoang,T. (2006) ETO2 coordinates cellular proliferation and differentiation during erythropoiesis. *EMBO J.*, **25**, 357–366.
36. Schuh,A.H., Tipping,A.J., Clark,A.J., Hamlett,I., Guyot,B., Iborra,F.J., Rodriguez,P., Strouboulis,J., Enver,T., Vyas,P. *et al.* (2005) ETO-2 associates with SCL in erythroid cells and megakaryocytes and provides repressor functions in erythropoiesis. *Mol. Cell. Biol.*, **25**, 10235–10250.
37. Gilchrist,M., Thorsson,V., Li,B., Rust,A.G., Korb,M., Roach,J.C., Kennedy,K., Hai,T., Bolouri,H. and Aderem,A. (2006) Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. *Nature*, **441**, 173–178.
38. Lieschke,G.J., Grail,D., Hodgson,G., Metcalf,D., Stanley,E., Cheers,C., Fowler,K.J., Basu,S., Zhan,Y.F. and Dunn,A.R. (1994) Mice lacking granulocyte colony-stimulating factor have chronic neutropenia, granulocyte and macrophage progenitor cell deficiency, and impaired neutrophil mobilization. *Blood*, **84**, 1737–1746.

39. Wilson,N.K., Foster,S.D., Wang,X., Knezevic,K., Schutte,J., Kaimakis,P., Chilarska,P.M., Kinston,S., Ouwehand,W.H., Dzierzak,E. *et al.* (2010) Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell*, **7**, 532–544.
40. Wei,G., Abraham,B.J., Yagi,R., Jothi,R., Cui,K., Sharma,S., Narlikar,L., Northrup,D.L., Tang,Q., Paul,W.E. *et al.* (2011) Genome-wide analyses of transcription factor GATA3-mediated gene regulation in distinct T cell types. *Immunity*, **35**, 299–311.
41. Lichtinger,M., Ingram,R., Hannah,R., Muller,D., Clarke,D., Assi,S.A., Lie,A.L.M., Noailles,L., Vijayabaskar,M.S., Wu,M. *et al.* (2012) RUNX1 reshapes the epigenetic landscape at the onset of haematopoiesis. *EMBO J.*, **31**, 4318–4333.
42. Tanaka,Y., Joshi,A., Wilson,N.K., Kinston,S., Nishikawa,S. and Gottgens,B. (2012) The transcriptional programme controlled by Runx1 during early embryonic blood development. *Dev. Biol.*, **366**, 404–419.
43. Kassouf,M.T., Hughes,J.R., Taylor,S., McGowan,S.J., Soneji,S., Green,A.L., Vyas,P. and Porcher,C. (2010) Genome-wide identification of TAL1′s functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res*, **20**, 1064–1083.
44. Encode Project Consortium, Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and Snyder,M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
45. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.
46. Pickrell,J.K., Gaffney,D.J., Gilad,Y. and Pritchard,J.K. (2011) False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics*, **27**, 2144–2146.
47. Treangen,T.J. and Salzberg,S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, **13**, 36–46.
48. DeVilbiss,A.W., Sanalkumar,R., Johnson,K.D., Keles,S. and Bresnick,E.H. (2014) Hematopoietic transcriptional mechanisms: from locus-specific to genome-wide vantage points. *Exp. Hematol.*, **42**, 618–629.
49. Sanchez-Castillo,M., Ruau,D., Wilkinson,A.C., Ng,F.S., Hannah,R., Diamanti,E., Lombard,P., Wilson,N.K. and Gottgens,B. (2014) CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res.*, 10.1093/nar/gku1895.
50. Goke,J., Jung,M., Behrens,S., Chavez,L., O'Keeffe,S., Timmermann,B., Lehrach,H., Adjaye,J. and Vingron,M. (2011) Combinatorial binding in human and mouse embryonic stem cells identifies conserved enhancers active in early embryonic development. *PLoS Comput. Biol.*, **7**, e1002304.
51. Handstad,T., Rye,M., Mocnik,R., Drablos,F. and Saetrom,P. (2012) Cell-type specificity of ChIP-predicted transcription factor binding sites. *BMC Genomics*, **13**, 372.