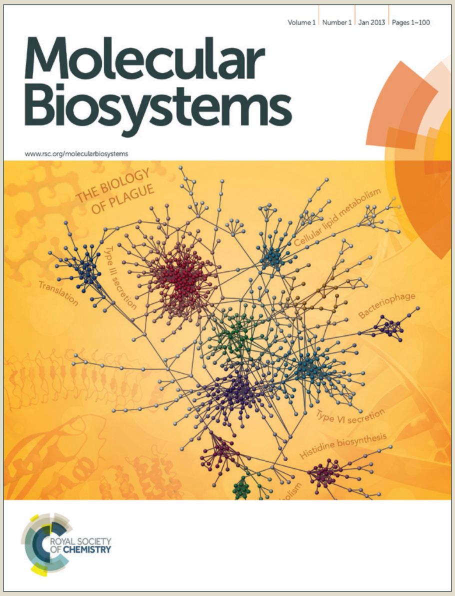
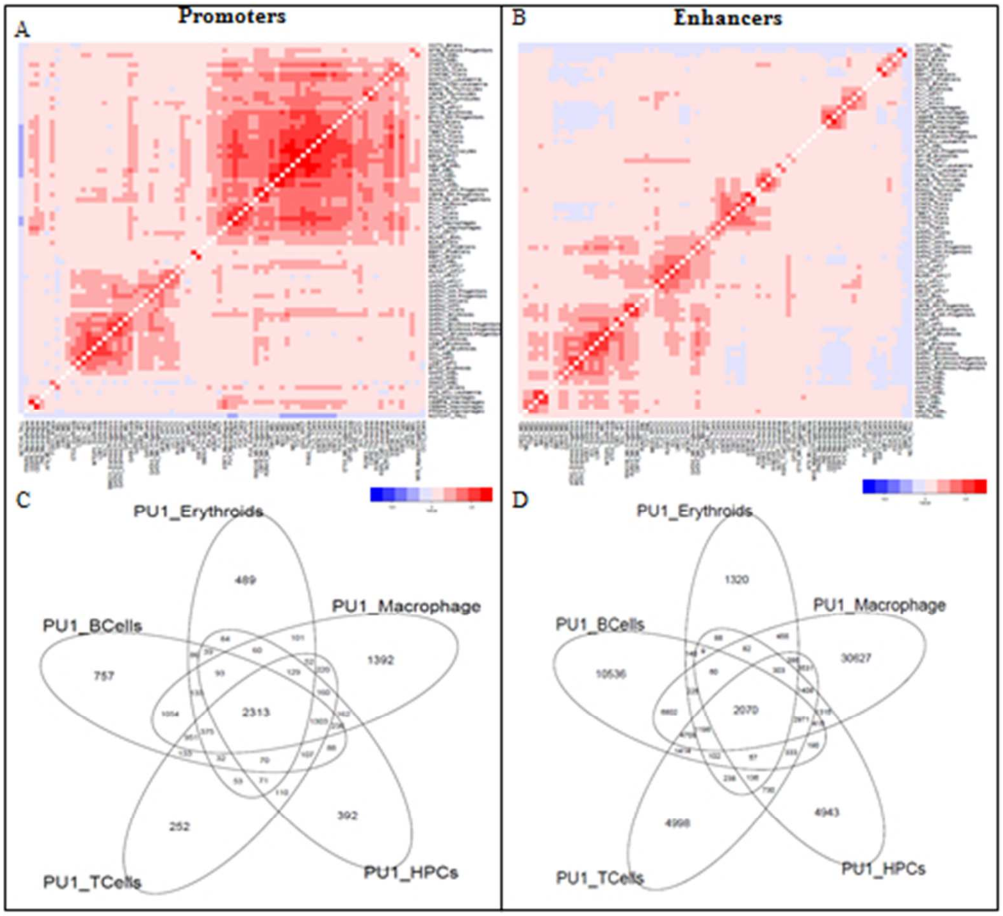
Molecular BioSystems

Accepted Manuscript


Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Molecular Biosystems
www.rsc.org/molecularbiosystems
THE BIOLOGY OF PLAGUE
Cellular lipid metabolism
Type III secretion
Translation
Bacteriophage
Type VI secretion
Histidine biosynthesis
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems



132x121mm (96 x 96 DPI)

Journal Name RSCPublishing

ARTICLE

Cite this: DOI: 10.1039/x0xx00000x

Received 00th January 2014,
Accepted 00th January 2014

DOI: 10.1039/x0xx00000x

www.rsc.org/

Concerted bioinformatic analysis of genome-scale blood transcription factor compendium reveals new control mechanisms

Anagha Joshi[a] and Berthold Gottgens[b]

Transcription factors play a key role in development and disease. ChIP-sequencing has become a preferred technique to investigate genome-wide binding patterns of transcription factors in vivo. Although this technology has led to many important discoveries, the rapidly increasing number of publicly available ChIP-sequencing data sets still remains a largely unexplored resource. Using a compendium of 144 publicly available murine ChIP-sequencing data sets in blood, we show that systematic bioinformatic analysis can unravel diverse aspects of transcription regulation; from genome-wide binding preferences, finding regulatory partners and assembling regulatory complexes, to identifying novel functions of transcription factors and investigating transcription dynamics during development.

Introduction

The control of cell-type specific gene expression underlies development of all multicellular organisms, and is thought to be achieved through combinatorial interactions of transcription factors with gene regulatory sequences. Moreover, dysregulation of transcription has widely proven as a major contributor to human pathologies, with the recent development of small molecule drugs targeting protein interactions between transcriptional regulator generating much excitement [1, 2].

With the interaction between cis-regulatory DNA elements and trans-acting transcription factors (TFs) representing the fundamental basis of transcriptional control, the delineation of comprehensive collections of regulatory sequences together with knowledge of the TFs bound to them will be essential to gain global insights into transcriptional control mechanisms. Over the past 10 years, chromatin immunoprecipitation (ChIP) followed by microarray (ChIP-chip) or sequencing (ChIP-Seq) have become the most widely used approaches for genome wide identification and characterization of in vivo protein-DNA interactions. Due to the rapid drop in the cost of high throughput sequencing, ChIP sequencing has become the method of choice for the generation of high resolution maps of genome-wide protein-DNA interactions in mammalian systems [3].

To gain a holistic view of transcriptional control during development, it is essential to generate genome scale maps of key transcription factors across multiple cell types. However, generating such genome-scale maps in many different cell types remains a daunting task for individual research groups due to limiting human and financial resources. Moreover, each individual TF requires careful validation of antibody reagents, which limits the potential throughput of large-scale initiatives. Indeed, bespoke protocols are often developed by individual groups with specialist expertise, so that published ChIP-Seq studies commonly report binding maps for less than a handful of TFs [4–10] and only a few larger studies reporting 10 or more factors [11, 12] or a single factor across multiple cell types [13]. We have previously shown [14] that unlike gene expression data, ChIP-Seq datasets produced by different laboratories can be readily integrated. This analysis revealed that genome wide transcription factor binding profiles are largely governed by cellular context. We recently reported a TF ChIP-Seq compendium containing 144 publicly available studies pertaining to the mouse blood system [15]. Using this dataset, here we show how concerted bioinformatic analysis of such a high quality hand-curated compendium can reveal previously unknown aspects of transcriptional control. This includes identification of those TF-bound sites most likely to be functional, prediction of TF interactions and multicomponent complexes, specific functionality of individual TFs and the dynamics of transcriptional regulation during differentiation and development.

Results and discussion

Enhancers, unlike promoters cluster according to cell type

We collected genome-wide binding patterns (peaks) of 144 publicly available murine ChIP-sequencing data sets for 53 transcription factors in 15 major blood lineages and leukemia [15] to obtain 270,261 regulatory regions with at least one factor binding. We classified peaks in two groups: promoter and enhancer peaks by defining the peaks within 1kb of TSS as promoter peaks. 7.5% of the total peaks belonged to promoters and all non-promoter peaks were classified as putative enhancers. The hierarchical clustering of enhancers clustered them according to cell type (Figure1B, Supplementary figure 2) irrespective of the factor such as Fli1 in hematopoietic progenitor cells (HPC) clustered with other samples in HPCs and Fli1 in T cells clustered with T cell samples. There was an exception of one transcription factor, Pu.1. Pu.1 samples across multiple cell types clustered together [14].

The promoter regions did not show a strong cell type specific clustering but clustered into two major clusters (Figure 1A, Supplementary Figure 1). Cluster 1 consisted of Gata factors across multiple cell types with their known interacting partners such as Ldb1, Scl/Tal1 and Cluster 2 consists of a large agglomeration of over 35 samples of multiple factors in diverse cell types. More generally, the observation of lineage-specific pair-wise associations in distal but not promoter regions provides global confirmation for previous suggestions that tissue specific expression is largely mediated by distal elements (Heintzman et al., 2009).

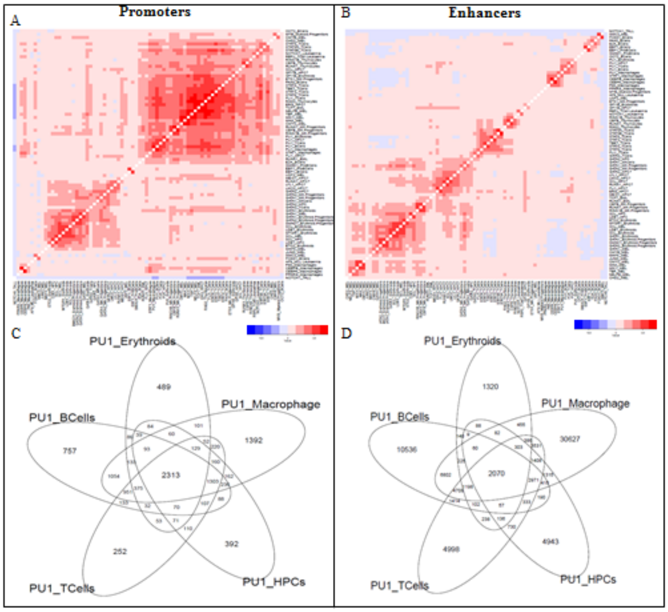


Figure 1: A,B Hierarchical clustering of pair-wise peak overlap of all in promoters and enhancers across all cell types red representing positive Pearson's correlation coefficient values and blue representing negative correlation coefficients C,D: 5-way Venn diagram of Pu.1 ChIP sequencing data from 5 cell types in promoters and enhancers representing higher overlap in promoters compared to enhancers.

As Pu.1 peaks in both promoters as well as enhancers cluster according to the factor rather than cell type, we characterized them in more detail. The 5-way Venn diagram of Pu.1 in promoter regions showed a high overlap of binding sites with about 50% of peaks present in all cell types (Figure 1C) whereas only about 10% of enhancer peaks were present in all cell types (Figure 1D). This shows that Pu.1 also agrees with the model where promoters mainly drive cell type invariant while enhancers drive tissue specific expression.

It is well established that transcription factors bind to different interacting partners in a cell type specific manner to drive gene expression [8]. To check if TFs have distinct interacting partners in promoter and enhancer regions, we calculated cis-regulatory motif enrichment in promoter and enhancer region separately for each factor using HOMER software. The sequence motif of the transcription factor chipped was enriched in both promoter and enhancers in most samples. Most samples also exhibited promoter-specific and enhancer-specific sequence motifs (Supplementary Figure 3). The GFY-STAF, NRF1 sequence motifs were enriched in promoters of most samples. Only a few motifs were sample specific in promoters such as Sp1 motif was enriched only in Scl/Tal1 promoter peaks. Sp1 is known to interact with Scl/Tal1 to drive expression of key gene loci such as Kit [16]. On the other hand, enhancers showed more sample specific motif enrichment. Ebf1 (early B cell factor) motif is enriched only for Pu.1 enhancers in B cells while MafA (macrophage activating factor) motif is only enriched in Pu.1 enhancers in macrophages.

Taken together, the data supports the suggestion that tissue-specificity is common feature of enhancers rather than promoters.

Transcription factor gene loci are enriched for peaks.

We mapped peaks across 15 blood lineages to their nearest genes resulting into an average of 13.5 peaks per gene. The 19869 unique gene loci were associated with peaks ranging from a single peak to over 200 peaks. The 726 genes with more than 50 peaks in their gene loci are enriched for functional categories 'transcription regulation' (p-value: 6.6E-18), 'hematopoiesis' (p-value: 1.9E-10) and 'blood vessel development' (p-value: 8.2E-8) demonstrating that hematopoietic regulatory genes have more binding sites in their gene loci. In an individual ChIP-sequencing experiment, most gene loci are associated with only one peak with an average of 1.8 peaks per gene. Genes with more than 5 peaks in their gene locus were enriched for hematopoietic functions. Transcription factor gene loci have an average of 2.5 peaks per gene, in agreement with previously reported suggestions that TF gene loci have a higher number of regulatory elements than average. This difference is statistically significant even after correcting for the gene length (p value: 2.2e-6).

It has been suggested that multiple peaks of a TF in a gene locus arise due to cross linking of multiple distant regulatory elements to the promoter, which might explain the lack of a consensus binding motif in many ChIP-seq peak regions [17]. We calculated the number of enhancer peaks for each factor with and without the presence of a peak at the promoter of a gene and did not observe any bias towards the presence of an enhancer peak with presence of a promoter peak.

Candidate regulatory regions bound by multiple factors might be functionally more relevant.

A typical ChIP-sequencing experiment generates millions of reads and hundreds to thousands of peaks. It is widely assumed that not all binding events are of equal functional significance. However, dissecting out functionally important binding events from potentially opportunistic binding events still remains an unsolved problem. Approximately 60% of the 270 thousand peaks of TFs across multiple cell types in blood are bound by more than one factor. We investigated whether the binding of multiple TFs provides any clues towards the functional implications of a binding event. As sequence conservation of a DNA fragment across species is predictive of functionality, we calculated human-mouse sequence conservation scores for all peaks. The sequences underlying peaks bound by multiple factors were more conserved across mammals than those bound by a single factor (Figure 2A). Moreover, peaks bound by multiple factors were enriched in the VISTA enhancer database (Figure 2B), a collection of over 700 enhancer regions functionally validated in transgenic mouse assays [18]. Taken together, these observations suggest that peaks bound by multiple factors might be more likely to be functional. Studies in mammalian cell types indeed have shown that the densely occupied regions tend to lie in the vicinity of genes characteristic of that particular cell type [11, 19]. In addition to functionality of peaks bound by multiple TFs, it has also been shown that gene loci with multiple binding events are more likely to be functionally significant targets [20]. Genes bound at multiple locations in most samples are over-represented for developmental processes including 'muscle tissue development' and 'cell fate commitment', as well as for 'transcription factor activity'.

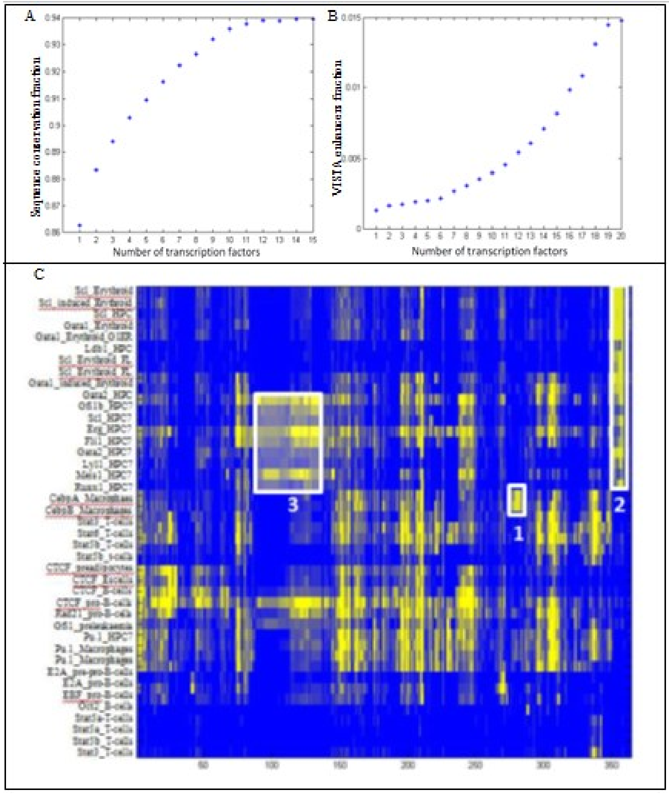


Figure 2: A,B. Fraction of conserved peaks across human and mouse and fraction of in vivo validated peaks (Visel et al., 2007) respectively classified according to the number of transcription factors bound. C. Heatmap of all ChIP-seq samples against over-represented (yellow) JASPAR motifs showing sequence motifs over-represented in at least one of the samples. Box 1 represents variants of Cebp motif, box 2 represents variants of GATA motif, while box 3 represents variants of homeo-domain motif.

Prediction of new candidate regulatory partners using enriched cis-regulatory motifs

Combinatorial transcriptional control is a key aspect of eukaryotic transcription as it provides cell type specificity as well as an ability to integrate multiple signals at a transcriptional level. In order to find over-represented cis-regulatory sequence motifs in each ChIP sequencing sample, we used a list of approximately 1300 sequence motifs with known or unknown associated TFs from the JASPAR data-base [26]. Figure 2C shows all significantly enriched motifs (x axis) for all samples (y axis) highlighted in yellow. The enriched motifs are useful in three ways. Firstly, they validate the chipped TF e.g. the Cepb motif is enriched in the two samples CebpA and CebpB (Figure 2C (1)). Secondly, they indicate important binding motifs for a particular cell type, such as enrichment of the GATA motif in HPC7 and erythroid cells (Figure 2C (2)). Important regulators such as Runx1 and Tal1 are thought to be recruited indirectly to many regulatory regions with the help of GATA factors [11]. Thirdly and most importantly, new candidate regulatory partners can be predicted, for example a homeodomain box motif is overrepresented only in the binding sites of all factors chipped in hematopoietic progenitor cells (Figure 2C (3)). Hox proteins, known to play key roles in governing proliferation and differentiation of haematopoietic progenitor cells, can therefore be nominated as new candidate interacting partners with the other blood stem cell factors present in the compendium.

Transcription factors show preference to a particular genomic location

In order to investigate whether TFs have a preference for specific genomic contexts, we used HOMER [8] to calculate enrichment with respect to 9 categories defining gene structure such as 3' UTR, 5' UTR, Exon, Intron, Inter-genic, and Promoter regions as well as repeat elements such as LINE, SINE and LTR. All transcription factors were enriched for promoter binding as expected. The components of Ldb1 complex in Erythroid cells were specifically enriched for intronic regions while Chd2 and Smc3 in MEL and Notch1 in T-ALL samples were enriched for 3’ UTR regions (Supplementary Figure 4). All Pu.1 samples were enriched for LTR repeat elements whereas CebpA and CebpB in macrophages were enriched for SINE repeat elements (Supplementary Figure 5). Bourque et. al. [21] showed that binding sites of five transcription factors ESR1, TP53, POU5F1, SOX2, and CTCF are embedded in distinctive families of transposable elements which facilitate dynamics in transcriptional network during evolution such as new locations of CTCF binding generated by SINE repeat element expansion in mammals [22]. The repeat region enrichment analysis thus provides clues towards how these transcription factors might have gained new regulatory sites during evolution.

Another genomic feature thought to be important for transcription control are CpG islands which facilitate promoter function by destabilising nucleosomes and attracting proteins that create a chromatin state suitable for transcription [23]. Rozenberg et. al [24] observed that the frequency of six TFBS (ETS, NRF1, BoxA, SP1, CRE and E-box) can accurately predict presence of CpG islands in promoters suggesting that they are structural elements critical for CpG island function. In line with this, transcription factors such as the three ETS factors Erg, Fli1 and Pu.1were enriched for CpG rich regions. Interestingly, peaks of components of the Ldb1 complex (Gata1, Gata2, Ldb1, Mtgr1 and Scl) occurred significantly less often than expected by chance in CpG rich regions.

Taken together, we found binding biases of transcription factors with respect to genomic locations, repeats and CpG islands. The functional relevance of these observations remains to be investigated.

Table 1: Top 5 over-represented and 5 under-represented ChIP-seq samples with peaks in CpG rich regions along with corresponding P value.

#	Sample	Prefer/ Avoid	P value
1	Erg_HPC7	Prefer	<1e-256
2	Fli1_T-cells	Prefer	<1e-256
3	Gfi1b_HPC7	Prefer	<1e-256
4	Pu.1_B-cells	Prefer	<1e-256
5	Rag2_thymocytes	Prefer	<1e-256
5	Ldb1_Erythroid	Avoid	3.4e-4

4	Gata1_Erythroid_progenitors	Avoid	8.9e-8
3	Lmo2_HPC7	Avoid	9.6e-5
2	Lyl1_HPC7	Avoid	5.2e-8
1	Smad1_Erythroid_progenitors	Avoid	<1e-256

TF complexes can be predicted using ChIP sequencing datasets

Physical interaction of TFs is an important aspect in determining tissue specific gene expression, and cooperative binding to DNA may be subject to spatial constraints. For each TF pair, we mapped the sequence motifs to peaks bound by both TFs and calculated the distance between two motifs. We selected motif pairs displaying a specific distance preference in at least two independent ChIP-seq experiments. Importantly, this analysis recovered previously known spacing of 8-10 bps between GATA and E-box motifs involved in binding of Gata1/Scl/E2A/Lmo2 multiprotein complexes [25]. Of interest, additional preferred pair-wise spacing could be recovered such as 20bp spacing between the motifs for CTCF and Pu.1. The functional significance of this remains to be explored. The co-ordinate binding between a major fate determining factor such as Pu.1 with a more architectural transcription factor such as CTCF, does however provide tantalizing clues as to how interactions between such factors may potentially be involved in stabilizing cell type specific transcription programs. We also find an overlapping joint motif - CANNTGGAW between Scl and ETS factors (Pu.1 and Fli1).

To investigate any new motifs showing distance specificity with respect to TF binding sites from our compendium, we calculated distances between each sample and all possible 3mers (43=64 patterns). We found 3 binding distance preferences; the first pattern, GATA and GAT, had a 3/4bp gap consistent with Gata factors binding as homo-dimers validated by crystal structure (Bates et al., 2008). The second pattern, GATA and CTG or GTC, had a 9bp gap mapping to GATA and a half Ebox binding as a part of the Ldb1 complex. The final pattern, Gfi1b and (A/T)GC, had a 2bp gap.

Lineage priming in progenitor cells

TFs are major determinants of cell fate and lineage choice. However, most lineage determining TFs are expressed across multiple lineages, suggesting that combinatorial interactions are critical in determining cell type specificity. By merging data sets from different studies, the TF ChIP-seq compendium serves as an excellent resource in the study of genome wide binding patterns of the same TF in multiple cell types. Grouping the genome wide binding patterns of Pu.1 in haematopoietic progenitor cells (HPCs) along with two mature cell types (macrophages and B cells) highlights that cell type specific, as well as ubiquitous binding events are present in both promoters and enhancers with ubiquitous binding events being more common in promoters. T and B cells specific functional categories such as 'lymphocyte activation (p value: 1.9e-6)' , 'immune system development (p-value: 5.1e-4)', 'B cell receptor signalling pathway (p-value: 1.2e-2)' are enriched in genes near Pu.1 peaks in HPC7 and B cells and not in macrophages while macrophage specific functional categories such as 'endocytosis (p-value: 2.0e-5)', 'inflammatory response (p-value: 6.2e-3)' are over-represented in genes near Pu.1 peaks in HPC7 and macrophages and not in B cells. This is a strong indicator of lineage priming in the progenitor cells and therefore provides global confirmation for one of the most hotly debated topics in stem cell biology.

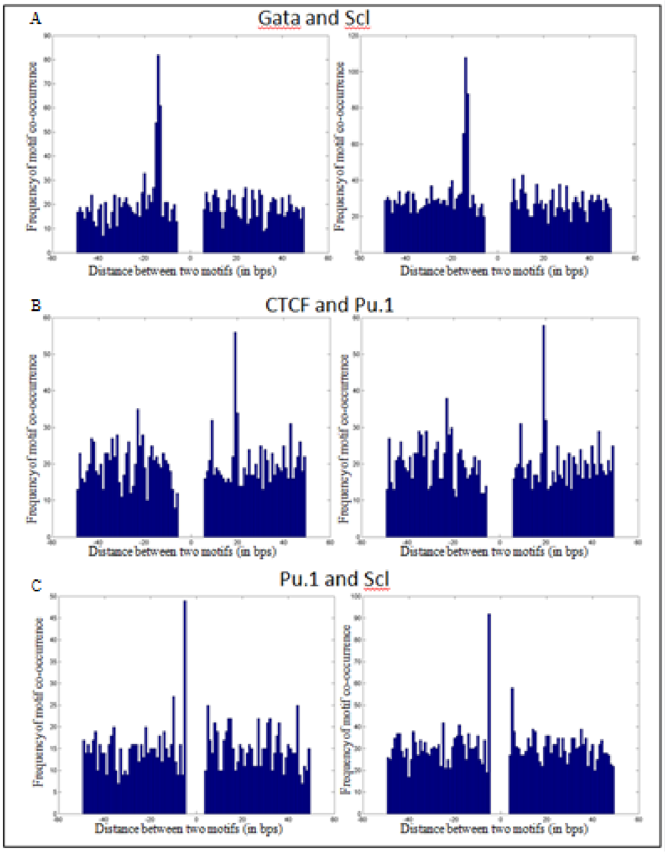


Figure 3: A. Frequency of distance between the Scl motif and GATA motif in peaks occupied by both Gata1/Gata2 and Scl, plotted such that GATA motif is at position zero. A peak with 8-10 bps gap between the two sequence motifs is over-represented. B. Similarly there is a preferred gap of 20 bps between the CTCF and Pu.1 motifs C. A gap of -1 bp between the Pu.1 and Scl motifs is significantly enriched. Each motif pair was validated by at least two independent ChIP-seq experiments.

Methods

The Genome-wide binding patterns of 53 transcription factors in 15 major blood lineages and leukaemia were obtained from [15]. Peaks within a 1kb region from a gene TSS, based on RefSeq gene annotation, were classed as promoter peaks. For each transcription factor pair, the significance of peak overlap was calculated using 1000 randomisations. Human-mouse orthologous regions were downloaded from the MGI database. The overlaps between peaks and human-mouse orthologous regions, as well as experimentally validated enhancers in mouse [27] were calculated using BEDtools [28]. For the two groups, we calculated whether the pair-wise overlap of promoter and non-promoter peaks was significantly over-represented (red) or under-represented (blue) compared to 100 randomizations. Using HOMER [8] and based on gene context or repeat elements, peaks were sorted into 9 categories: 3' UTR, 5' UTR, exon, intron, intergenic, promoter, LINE, SINE and LTR. CpG islands were downloaded from UCSC. A list of transcription factors in mouse was downloaded from RIKEN [29]. To find distance preferences between pairs of TFs, the sequences for peaks bound by both transcription factors were obtained using UCSC

Galaxy and the binding locations of each sequence motif was determined using TFSBsearch [30]. Cis-regulatory sequence motifs were downloaded from the JASPAR library [26] and the motifs were searched in peaks using TFSBsearch [30]; over-representation was calculated with respect to 100 random sequence sets of the same number and lengths of real peak sequences. Functional enrichment was calculated using DAVID [31]. Most analysis was done using Perl, MATLAB and R scripts.

Conclusions

The advent of next generation sequencing technologies has led to a dramatic shift in modern biological research, where bioinformatic processing and interpretation of large-scale data sets is rapidly replacing data generation as the major bottleneck. Moreover, bioinformatic analysis of genome-scale data sets is often restricted to the particular context of the paper that first reported them, even though the raw data are made publicly available in online repositories. Consequently, a whole potential treasure trove of biological insights remains essentially unexplored.

To ameliorate this situation, progress on two fronts will be vital. Firstly, significant efforts need to be invested into the generation of data integration platforms that facilitate cross-referencing between the multiple independent studies. Secondly, bioinformatic analysis strategies need to be developed to facilitate extraction of novel biological hypotheses from integrated genome-scale resources.

In this paper, we have addressed this latter issue and provide seven examples of bioinformatic analysis, that together have allowed us to develop a number of new hypotheses on transcriptional control mechanisms with the potential to transform our understanding of blood cell development. Importantly, both the procedures outlined as well as the take-home messages learned should be readily transferable to the exploitation of ChIP-Seq datasets in other cellular systems, and thus have the potential to significantly advance our understanding of a wide range of both normal and pathological cellular processes.

Acknowledgements: AJ is a Chancellor's fellow at the Roslin institute, University of Edinburgh. The Gottgens' laboratory is funded by grants from Leukaemia and Lymphoma Research, the Leukaemia and Lymphoma Society, Medical Research Council, BBSRC, CRUK and the National Institute for Health Research Cambridge Biomedical Research Centre. The author(s) declare that they have no competing interests.

Notes and references

[a] The Roslin Institute, University of Edinburgh, Easter Bush Campus, Midlothian, EH25 9RG, UK

[b] Department of Haematology, Cambridge Institute for Medical Research, Cambridge University, Hills Road, Cambridge, CB2 0XY, UK

Electronic Supplementary Information (ESI) available: See DOI: 10.1039/b000000x/

REFERENCES

1. Delmore JE, Issa GC, Lemieux ME, Rahl PB, Shi J, Jacobs HM, Kastritis E, Gilpatrick T, Paranal RM, Qi J, Chesi M, Schinzel AC, McKeown MR, Heffernan TP, Vakoc CR, Bergsagel PL, Ghobrial IM, Richardson PG, Young RA, Hahn WC, Anderson KC, Kung AL, Bradner JE, Mitsiades CS: **BET bromodomain inhibition as a therapeutic strategy to target c-Myc**. *Cell* 2011, **146**:904–917.
2. Dawson MA, Prinjha RK, Dittmann A, Giotopoulos G, Bantscheff M, Chan W-I, Robson SC, Chung C, Hopf C, Savitski MM, Huthmacher C, Gudgin E, Lugo D, Beinke S, Chapman TD, Roberts EJ, Soden PE, Auger KR, Mirguet O, Doehner K, Delwel R, Burnett AK, Jeffrey P, Drewes G, Lee K, Huntly BJP, Kouzarides T: **Inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia**. *Nature* 2011, **478**:529–533.
3. Ho JWK, Bishop E, Karchenko PV, Nègre N, White KP, Park PJ: **ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis**. *Bmc Genomics* 2011, **12**:134.
4. Yu M, Mazor T, Huang H, Huang H-T, Kathrein KL, Woo AJ, Chouinard CR, Labadorf A, Akie TE, Moran TB, Xie H, Zacharek S, Taniuchi I, Roeder RG, Kim CF, Zon LI, Fraenkel E, Cantor AB: **Direct recruitment of polycomb repressive complex 1 to chromatin by core binding transcription factors**. *Mol Cell* 2012, **45**:330–343.
5. Trowbridge JJ, Sinha AU, Zhu N, Li M, Armstrong SA, Orkin SH: **Haploinsufficiency of Dnmt1 impairs leukemia stem cell function through derepression of bivalent chromatin domains**. *Genes Dev* 2012, **26**:344–349.
6. Nakayamada S, Kanno Y, Takahashi H, Jankovic D, Lu KT, Johnson TA, Sun H, Vahedi G, Hakim O, Handon R, Schwartzberg PL, Hager GL, O'Shea JJ: **Early Th1 cell differentiation is marked by a Tfh cell-like transition**. *Immunity* 2011, **35**:919–931.
7. Ng S-L, Friedman BA, Schmid S, Gertz J, Myers RM, Tenoever BR, Maniatis T: **IκB kinase epsilon (IKK(epsilon)) regulates the balance between type I and type II interferon responses**. *Proc Natl Acad Sci U S A* 2011, **108**:21170–21175.
8. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities**. *Mol Cell* 2010, **38**:576–589.
9. Wontakal SN, Guo X, Smith C, MacCarthy T, Bresnick EH, Bergman A, Snyder MP, Weissman SM, Zheng D, Skoultchi AI: **A core erythroid transcriptional network is repressed by a master regulator of myelo-lymphoid differentiation**. *Proc Natl Acad Sci U S A* 2012, **109**:3832–3837.
10. Wilson NK, Miranda-Saavedra D, Kinston S, Bonadies N, Foster SD, Calero-Nieto F, Dawson MA, Donaldson IJ, Dumon S, Frampton J, Janky R, Sun X-H, Teichmann SA, Bannister AJ, Göttgens B: **The transcriptional program controlled by the stem cell leukemia gene Scl/Tal1 during early embryonic hematopoietic development**. *Blood* 2009, **113**:5456–5465.
11. Wilson NK, Foster SD, Wang X, Knezevic K, Schütte J, Kaimakis P, Chilarska PM, Kinston S, Ouwehand WH, Dzierzak E, Pimanda JE, de Bruijn MFTR, Göttgens B: **Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators**. *Cell Stem Cell* 2010, **7**:532–544.
12. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh Y-H, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung W-K, Clarke ND, Wei C-L, Ng H-H: **Integration of external signaling pathways with the core transcriptional network in embryonic stem cells**. *Cell* 2008, **133**:1106–1117.
13. Wei G, Abraham BJ, Yagi R, Jothi R, Cui K, Sharma S, Narlikar L, Northrup DL, Tang Q, Paul WE, Zhu J, Zhao K: **Genome-wide analyses of transcription factor GATA3-mediated gene regulation in distinct T cell types**. *Immunity* 2011, **35**:299–311.
14. Hannah R, Joshi A, Wilson NK, Kinston S, Göttgens B: **A compendium of genome-wide hematopoietic transcription factor maps supports the identification of gene regulatory control mechanisms**. *Exp Hematol* 2011, **39**:531–541.
15. Joshi A, Hannah R, Diamanti E, Göttgens B: **Gene set control analysis predicts hematopoietic control mechanisms from genome-wide transcription factor binding data**. *Exp Hematol* 2013, **41**:354–366.e14.
16. Lécuyer E, Herblot S, Saint-Denis M, Martin R, Begley CG, Porcher C, Orkin SH, Hoang T: **The SCL complex regulates c-kit**

expression in hematopoietic cells through functional interaction with Sp1. *Blood* 2002, **100**:2430–2440.
17. Farnham PJ: **Insights from genomic profiling of transcription factors**. *Nat Rev Genet* 2009, **10**:605–616.
18. Visel A, Minovitsky S, Dubchak I, Pennacchio LA: **VISTA Enhancer Browser--a database of tissue-specific human enhancers**. *Nucleic Acids Res* 2007, **35**(Database issue):D88–92.
19. Tijssen MR, Cvejic A, Joshi A, Hannah RL, Ferreira R, Forrai A, Bellissimo DC, Oram SH, Smethurst PA, Wilson NK, Wang X, Ottersbach K, Stemple DL, Green AR, Ouwehand WH, Göttgens B: **Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators**. *Dev Cell* 2011, **20**:597–609.
20. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G: **GREAT improves functional interpretation of cis-regulatory regions**. *Nat Biotechnol* 2010, **28**:495–501.
21. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng HH, Liu ET: **Evolution of the mammalian transcription factor binding repertoire via transposable elements**. *Genome Res* 2008, **18**:1752–1762.
22. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT: **Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages**. *Cell* 2012, **148**:335–348.
23. Deaton AM, Bird A: **CpG islands and the regulation of transcription**. *Genes Dev* 2011, **25**:1010–1022.
24. Rozenberg JM, Shlyakhtenko A, Glass K, Rishi V, Myakishev MV, FitzGerald PC, Vinson C: **All and only CpG containing sequences are enriched in promoters abundantly bound by RNA polymerase II in multiple tissues**. *Bmc Genomics* 2008, **9**:67.
25. Wadman IA, Osada H, Grütz GG, Agulnick AD, Westphal H, Forster A, Rabbitts TH: **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* 1997, **16**:3145–3157.
26. Bryne JC, Valen E, Tang M-HE, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A: **JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update**. *Nucleic Acids Res* 2008, **36**(Database issue):D102–106.
27. Visel A, Minovitsky S, Dubchak I, Pennacchio LA: **VISTA Enhancer Browser--a database of tissue-specific human enhancers**. *Nucleic Acids Res* 2007, **35**(Database):D88–D92.
28. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinforma Oxf Engl* 2010, **26**:841–842.
29. Kanamori M, Konno H, Osato N, Kawai J, Hayashizaki Y, Suzuki H: **A genome-wide and nonredundant mouse transcription factor database**. *Biochem Biophys Res Commun* 2004, **322**:787–793.
30. Chapman MA, Donaldson IJ, Gilbert J, Grafham D, Rogers J, Green AR, Göttgens B: **Analysis of multiple genomic sequence alignments: a web resource, online tools, and lessons learned from analysis of mammalian SCL loci**. *Genome Res* 2004, **14**:313–318.
31. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* 2009, **4**:44–57.Citations here in the format A. Name, B. Name and C. Name, *Journal Title*, 2000, **35**, 3523; A. Name, B. Name and C. Name, *Journal Title*, 2000, **35**, 3523.