9244–9259 Nucleic Acids Research, 2017, Vol. 45, No. 16 Published online 17 July 2017
doi: 10.1093/nar/gkx618

Automatic identification of informative regions with epigenomic changes associated to hematopoiesis

Enrique Carrillo-de-Santa-Pau[1,†], David Juan[2,†], Vera Pancaldi[3,$], Felipe Were[1], Ignacio Martin-Subero[4], Daniel Rico[5,*], Alfonso Valencia[3,6,*] and on behalf of The BLUEPRINT Consortium[‡]

[1]Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, 28029, Spain, [2]Institut de Biologia Evolutiva, Consejo Superior de Investigaciones Científicas–Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, Barcelona, 08003, Spain, [3]Barcelona Supercomputing Centre (BSC), Barcelona, 08034, Spain, [4]Institut d'Investigacions Biomédiques August Pi i Sunyer (IDIBAPS), Department of Anatomic Pathology, Pharmacology and Microbiology, University of Barcelona, Barcelona, 08036, Spain, [5]Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, NE2 4HH, UK and [6]ICREA, Pg. Lluís Companys 23, Barcelona, 08010, Spain

Received March 16, 2017; Revised June 20, 2017; Editorial Decision June 27, 2017; Accepted July 06, 2017

ABSTRACT

Hematopoiesis is one of the best characterized biological systems but the connection between chromatin changes and lineage differentiation is not yet well understood. We have developed a bioinformatic workflow to generate a chromatin space that allows to classify 42 human healthy blood epigenomes from the BLUEPRINT, NIH ROADMAP and ENCODE consortia by their cell type. This approach let us to distinguish different cells types based on their epigenomic profiles, thus recapitulating important aspects of human hematopoiesis. The analysis of the orthogonal dimension of the chromatin space identify 32,662 chromatin determinant regions (CDRs), genomic regions with different epigenetic characteristics between the cell types. Functional analysis revealed that these regions are linked with cell identities. The inclusion of leukemia epigenomes in the healthy hematological chromatin sample space gives us insights on the healthy cell types that are more epigenetically similar to the disease samples. Further analysis of tumoral epigenetic alterations in hematopoietic CDRs points to sets of genes that are tightly regulated in leukemic transformations and commonly mutated in other tumors. Our method provides an analytical approach to study the relationship between epigenomic changes and cell lineage differentiation. Method availability: https://github.com/david-juan/ChromDet.

INTRODUCTION

Hematopoiesis is one of the most studied biological differentiation processes, in which different cell lineages arise from a common hematopoietic stem cell (HSC). This system can be seen as a hierarchical tree, where the more internal 'nodes' are the different lineage progenitors and the 'leaves' are the final mature cell types (1,2). This hierarchical tree with many 'nodes' and 'leaves' provides the best model to study chromatin remodeling during cell lineage differentiation (3–5).

Chromatin remodeling is a dynamic process that modulates the chromatin architecture and is vital to ensure proper functioning of the cell and maintenance of its identity (6). The de-regulation of chromatin remodeling factors often leads to diseases such as cancers (7) and neurodevelopmental disorders (8,9). A main role in this re-organization of chromatin is played by post-translational modifications of histone tails, which can affect many biological processes such as gene transcription, DNA repair, replication and recombination (10,11). Moreover, the cross-talk between different modifications affects the binding and function of other epigenetic elements, increasing the complexity of the chromatin remodeling process (12).

Despite great progress in our understanding of hematopoiesis during the last decades (13,14), we are still far from fully uncovering the details of the epigenetic

[*]To whom correspondence should be addressed. Tel: +44 1912087157; Fax: +44 1912085066; Email: daniel.rico@newcastle.ac.uk
Correspondence may also be addressed to Alfonso Valencia. Tel: +34 934134084; Fax: +34 934137721; Email: alfonso.valencia@bsc.es
[†]These authors contributed equally to the paper as first authors.
[‡]Member of International Human Epigenome Consortium (IHEC). For more information, see http://www.blueprint-epigenome.eu and http://ihec-epigenomes.org.
[$]Previously at Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, 28029, Spain.

mechanisms controlling this process. It is now widely accepted that the cell phenotype is directly related to its epigenetic makeup and that chromatin changes during differentiation contribute to the determination of cell fate. However, a major challenge in the field is to identify exactly where the epigenetic changes causing phenotypic changes occur. Similarly to the problem of distinguishing driver and passenger mutations in cancer, we can think of driver and passenger chromatin changes during cellular differentiation. Chromatin drivers of cellular differentiation would correspond to the subset of regions whose change is required to perform the different differentiation steps. As consequence, these regions must reflect one or more changes among cell types, while being fixed in any specific cell type. We therefore advocate the need to develop strategies identifying these key chromatin regions and their epigenetic changes that drive differentiation and determine cell fate. For this purpose, we take advantage of the large and comprehensive epigenomics datasets produced by the partners of the International Human Epigenome Consortium (IHEC; http://ihec-epigenomes.org/).

Here, we propose an approach to identify the key chromatin regions that undergo chromatin changes associated to cell differentiation during multiple differentiation steps in hematopoiesis (Figure 1). We define chromatin states based on the combinatorial patterns of 6 histone modifications in 42 human samples covering the myeloid and lymphoid differentiation lineages from HSCs. This framework establishes highly informative low-dimensional spaces based on a multiple correspondence analysis (MCA; (15)) of the profiles of histone modification combinations (chromatin states). Our integrative analysis of chromatin states in these samples recapitulates the human hematopoietic lineage differentiation tree from an epigenetic perspective. Moreover, our approach identifies 32,662 chromatin determinant regions (CDRs) in which chromatin changes are associated with the various differentiation steps the cells go through, possibly influencing their final lineage identities. The combination of chromatin states in these CDRs constitutes an epigenomic fingerprint that characterizes the different hematopoietic cell types. The method is available at https://github.com/david-juan/ChromDet.

MATERIALS AND METHODS

ChIP-Seq data processing

We retrieved data for 430 chromatin immunoprecipitation sequencing (ChIP-Seq) experiments from BLUEPRINT, ENCODE and NIH ROADMAP. We downloaded the hg19/GRCh37 alignments for 2 CD4+ and 1 CD8+ lymphocytes, 5 mature neutrophils, 3 CD14+ monocytes, 4 macrophages and 7 CD38- B cell samples from BLUEPRINT; 11 CD4+ and 3 CD8+ lymphocytes, 1 CD14+ monocyte and 3 CD34+ HSCs samples from NIH ROADMAP and 2 CD14+ monocytes samples from ENCODE described in Supplementary Table S1 and Figure 2A. In addition, the analysis including diseases was based on data from three acute myeloid leukemias (AML), six chronic lymphocytic leukemias (CLL) and three mantle cell lymphomas (MCL) from BLUEPRINT (see Supplementary Table S1). The BAM files were converted to BED format and duplicate reads were removed for all the experiments. We computed different quality control measures with phantompeakqualtools v1.10.1 (16) including total number of reads, normalized strand cross-correlation coefficient (NSC) and quality tag based on thresholded relative strand cross-correlation coefficient (RSC; see Supplementary Table S1). We flagged those histone experiments with less than 10^7 reads and no replicates; NSC < 1.05 and quality tag based on RSC < 0. Then, following a similar strategy used previously by the NIH Epigenomics Roadmap (17), we computed an overall quality rating per sample based on the six core histone modification quality experiments. We labeled samples as 'very high quality' if none or only one histone mark experiment failed in one out of the three quality criteria; 'high quality' if two or three histone experiments failed in one out of the three quality criteria; 'medium quality' if more than three histone marks failed in one out of the three criteria or up to two broad histone modifications (H3K36me3; H3K9me9; H3K27me3) failed in two out of the three quality criteria; 'low quality' if three or more histone experiments failed in two out of the three quality criteria or at least one histone experiment failed in the three quality criteria used. All the samples labeled as 'low quality' were discarded and not included in our study. The overall quality criteria for those histone experiments included in the analysis is shown in Supplementary Table S1.

Genome segmentation

The input information used to segment the genome into different chromatin states was derived from six histone modifications (H3K4me3; H3K4me1; H3K27me3; H3K9me3; H3K27ac and H3K36me3). We used the ChromHmm software (v1.10; 18) to define a 11 chromatin-states model (see Supplementary Figure S1) following the strategy proposed by the ChromHMM developers to set up the different parameters like number of states for training or posterior collapse (17,19,20). We evaluated the consistency and interpretability of chromatin states in models learnt with different numbers of chromatin states (5, 7, 9, 11, 13 and 20 states), quantified as the correlation of chromatin mark frequencies obtained for corresponding states across different models, as previously done by Ernst M. and Kellis M. (19). The results show that the 20 states model is recovered with correlations higher than 0.75 by the states trained in the model with 11 states, with little improvement in the 13 states model (Supplementary Tables S2 and S3). The 11 states model captures all the biological-interpretable states that were consistently found in larger models.

Importantly, a manual curation of the chromatin states based on available additional information (gene structures, CpG islands, Lamin B1, etc.) showed that the 11 states model retrieves all the main regulatory states (active promoter, bivalent promoter, enhancer, elongation, heterochromatin/low signal), without including any functionally unclassifiable chromatin state. Therefore, our approach of selecting 11 states to train the Hidden Markov Model (HMM) is aimed at striking an equilibrium between a low enough number of combinations and the biological interpretability of the states included in the analysis, based on the ChromHMM emission probabilities correla-

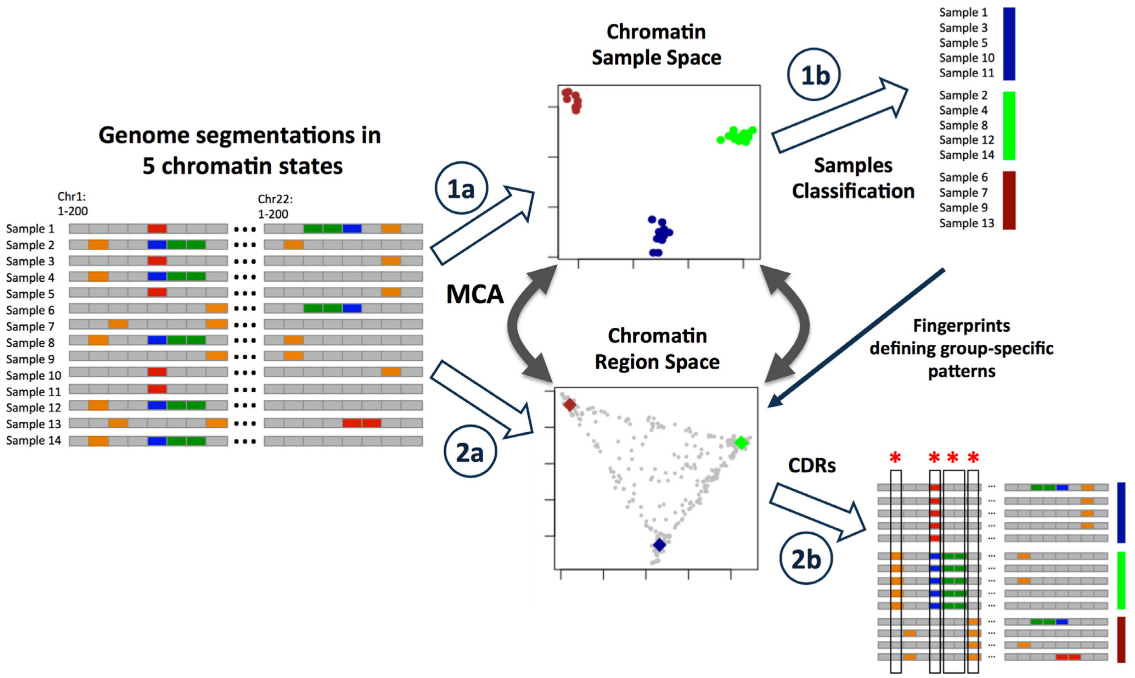

Figure 1. Framework to identify CDRs that determine cell or lineage identity based on chromatin state changes. (**1a**) A chromatin samples space is generated with MCA from the chromatin segmentations by each sample. (**1b**) Samples are classified depending on clusters derived from the MCA analysis. (**2a**) A second space is generated with MCA from the chromatin segmentations by each sample. (**2b**) The CDRs are obtained selecting those genomic regions that overlap with the cluster sample fingerprints, a reference sample representing each cell type cluster. These regions discriminate the different cell types classified in the samples space. (*Regions with chromatin changes among cell types → CDRs). See also Supplementary Figure S1 and Supplementary Data.

tion, prior knowledge regarding the function of these marks and our previous experience (12). In summary, the 11 states model selected captures the biologically interpretable states that were consistently found in larger models providing a suitable framework for our analysis.

We generated the model with the 'healthy' samples excluding B cells (see Supplementary Table S1 for details). The samples from B cells (naive and tonsil) and diseases (AML, CLL, MCL) were segmented with the model generated previously, as they were produced at the final stages of the BLUEPRINT project. Further, segmentations for each sample from the 11 states model were collapsed into 5 chromatin states summarizing similar states based on the emission probabilities, literature, biological knowledge and genomic feature enrichments: heterochromatin/low signal (H), enhancer (E), transcription (T), active promoter (A) and repressed promoter (R; Supplementary Figure S1A). Our a posteriori collapse into five chromatin states let us group dynamic states for a more robust representation of the epigenomic variability in cell types. In fact, differences in strength of enhancers, promoters or elongating regions can reflect more or less dynamic regions resulting in subtle differences between ChIP-seq experiments.

Therefore, for each sample, we have a vector of regions with their corresponding labels (chromatin states). In addition, we partitioned the genome into 200 bp, preserving the associated chromatin state labels in order to have the same number of regions in all samples and make them comparable. For further analysis, consecutive 200-bp intervals with the same labels pattern in all samples were merged, any change in one sample marking an interval transition.

Sample clustering in the chromatin sample space obtained by multiple correspondence analysis (MCA)

In this work we propose to use a methodological protocol based on MCA (15), previously applied to multiple sequence alignments of proteins, for the automatic extraction of relevant signatures (21) and to gene expression profiles for sample classification (22). MCA can be considered as an equivalent to principal component analysis (PCA) when working with qualitative data instead of continuous variables. MCA disentangles the sources of epigenomic variability among our samples into a set of principal components that form an orthogonal space which dimensions can be prioritized according to their corresponding eigenvalues. This MCA space can be reduced to a low dimensional one preserving most of the original information but filtering the main sources of noise. In brief, our protocol performs an MCA on a vectorial representation of multiple chromatin states sample vectors. It establishes the informative low dimensional space incorporating only those components with the highest eigenvalues, those explaining most of the total variance, where samples coordinates distribution is statistically different (P-value < 0.01, Wilcoxon test) between the tested component and the previously selected one, the one with the closer higher eigenvalue. In this work, we define the chromatin sample space as the space formed by this set of highly informative components coming from the MCA on the vectors of the chromatin states for the genomic regions analyzed samples. Robust unsupervised *k*-means clustering (23) is performed iteratively on this chromatin sample space for a range of pre-specified number of groups (from 2 to 50). Finally, optimal clustering solutions are detected as those maximizing the Calinsky's and Harabsz's (CH) index (24). In an analysis involving samples from different healthy cell types, as the one presented in this work, this protocol is intended to recover those cell types, or groups of cell types, whose epigenomic differences are able to discriminate them. These epigenetically robust groups of samples allow us to confidently address the detection of those regions that are important for establishing segregation of these samples.

A challenge of this approach was to deal with millions of regions within the same analysis. However, many of these regions will not be informative for discriminating the sample groups in our dataset. Highly variable regions and completely conserved regions are non-informative regions that increase the computational time cost, while sample-specific divergent regions can bias the results, being strongly influenced by the presence of sample outliers or sample-specific experimental noise. In order to reduce the influence of sample-specific patterns contributing to outlier effects, we focused on the set of regions presenting at least two different chromatin states in at least two samples each of them. Additionally, we filter out all the regions with change patterns (vectors of chromatin states for each genomic region across samples) that were poorly represented in our dataset. In particular, we filtered out those regions whose patterns were not shared by 10 regions (we obtain similar results for patterns shared by 5, 10 and 15 regions; data not shown). This step dramatically reduces the computational burden by removing regions with low influence in sample clustering. As a result of this filtering we run our MCA framework with 275, 825 regions from the 22 autosomal chromosomes of all the healthy samples.

Selection of chromatin determinant regions in the chromatin region space

Concomitantly to the detection of sample clusters, ideally equivalent to cell types, our framework allows the detection of the subset of regions better reflecting this inter-sample clustering. We called these regions CDRs and they are methodologically equivalent to the specificity determining positions detected (21) in protein families. First, we project the vectors reflecting every genomic region/state combination into the MCA space, generating the chromatin region space. Vectors representing chromatin patterns perfectly associated to every combination of sample clusters were used as fingerprints of the corresponding grouping. Every epigenomic region was associated to the closest fingerprint in the chromatin region space. Finally, CDRs were defined as those positions for which all their chromatin states were among the top 10 shortest distances to its fingerprints and the combination of these fingerprints form a perfect partitioning of the sample clusters (for a more detailed description see 21). In this situation, CDRs correspond to those regions with patterns of chromatin states along samples with very few intra-cluster epigenomic changes but with at least two clusters with different states. This definition of CDRs highlights the two key properties of these regions: the stability of their state is important for every single epigenomic cluster of samples and they define inter-cluster epigenomic changes. These properties point to the putative role of these regions in cell identity and cell fate respectively.

Chromatin determinant regions annotation, expression and enrichment analyses

Genomic annotation was carried out with HOMER software v4.7.2 (25). The tool annotatePeaks.pl was used with default parameters to annotate CDRs to genes with the following priority assigned: TSS (from −1 kb to +100 bp), transcription termination site (from −100 bp to +1 kb), protein coding exon, 5′-UTR exon, 3′-UTR exon, intron and intergenic. More detailed information is available in http://homer.salk.edu/homer/ngs/annotation.html. Gene ontology (GO) (Biological Process; 26) and Reactome (27) enrichment analysis were done adding the -go flag to the annotatePeaks.pl tool. Then, we calculated a P-value adjusted for multiple testing based on Benjamini–Hochberg correction using the p.adjust function in R (v3.2.2). All terms with an adjusted *P*-value < 0.05 were considered significant. We summarized the GO (Biological Process) significant terms with REVIGO (28).

The expression-associated analyses were carried out retrieving the RNAseq data for 60 483 protein-coding, ncRNA, pseudo, snoRNA and snRNA genes from The BLUEPRINT Data Analysis Portal (29). We took information from 12 macrophages, 8 monocytes, 6 neutrophils, 4 naive B cells, 3 germinal center (GC) B cells and 21 T cells, no data was found for HSCs from mature samples. We applied an ANOVA test to 7764 genes with CDRs associated and adjusted the *P*-value with Benjamini–Hochberg correction, adjusted *P*-value ≤ 0.05 was considered significant. Statistical analyses were carried out with aov and p.adjust functions from R.

The transcription factor motif (TFM) enrichments were performed with the findMotifsGenome.pl tool included in HOMER software (v4.7.2; 25). To determine the relative enrichment of known TFMs we excluded the CDRs referred to transcription, as they are related to polymerase elongation and not to transcription factors binding. The searches were done against a selected random background of windows adjusted to have equal GC content distribution in each of the input sequences. The region size was set up to 'given', other parameters were used by default. More detailed information is available in http://homer.salk.edu/homer/ngs/peakMotifs.html. The TFMs with a *q*-value $<$ 0.01 at least in one cell type were considered significant and selected to generate Figure 3C. We did not find enriched TFMs for T cells and neutrophils. The expression analyses for 28 of the transcription factor binding proteins of Figure 3C were performed with the same approach described above, the transcription factors in HSCs were not included in the expression analysis since BDAP does not provide data for HSCs.

Chromatin state transitions among cell types were represented with a Sankey diagram in Figure 3A using the 'makeRiver' and 'riverplot' functions included in the 'riverplot' R package (v0.5; https://cran.r-project.org/web/packages/riverplot/index.html).

Chromatin determinant regions in the context of disease

The 'healthy' hematopoietic chromatin sample space provides us a reference sample space, reflecting the informative epigenomic distances between normal hematopoietic cell types. As it is based on the major sources of information involved in hematopoiesis, it also serves us to study to what extend leukemic epigenomes retain features important to define the cell identity of the normal cell types.

In order to get a clearer view of these residual signals of 'normality', we focused on those CDRs for which the tumoral sample shows a chromatin state present in any cell type. For this, we projected the leukemic samples on the 'healthy' hematopoietic chromatin sample space, but considering only the influence of these CDRs. In practice, it means that every leukemic sample is projected based on a different number of regions and its position reflects the extent to which these regions correspond to patterns more related to one or other healthy cell type. This approach allows us to reduce the effect of tumor-related epigenetic changes and to weigh the contribution of patterns of chromatin states associated to more than one cell type according to their influence in the 'normal' chromatin sample space. We also projected the prototypic 'normal' cell types represented by the vectors presenting the chromatin states characteristic of the corresponding cell type for each CDR. Distances of leukemic samples to these prototypic 'normal' cell types reflect the similarity of the chromatin states in CDRs balancing the effect of chromatin states shared with other 'normal' cell types.

Despite the effect of focusing on 'conserved' states in CDRs, highly transformed leukemic samples could include a relevant number of changes to chromatin states characteristic of a different cell type. These effects will contribute to leukemic samples with less 'cell type-specific' CDRs. This situation can lead to less well-defined clusters of leukemic samples. Therefore, we decided to perform a hierarchical clustering (using Ward's method with euclidean distances as implemented in pheatmap v1.0.8 R package, http://CRAN.R-project.org/package=pheatmap) in this CDRs-based chromatin sample space, to illustrate the association of different leukemias to different cell types. As HSCs, macrophages and GC B cells 'prototypic' cell types were clearly very distant to the projections of all leukemic samples in this space, they were not considered in the hierarchical clustering, in order to improve the resolution of the relationship of tumoral and healthy samples.

We also define the ratio of CDRs with a chromatin state different to any healthy cell type as the tumoral epigenomic divergence. It represents how divergent a tumoral sample is from the space of healthy states calculated with the normal samples. Therefore, higher divergences imply higher probabilities that the cell type of origin of the tumoral sample is not represented in the healthy chromatin space or that the tumoral sample diverged so much than its projection on this space should be taken with care. The analyzed tumoral samples show epigenomic tumoral divergences ranging from 0.02 to 0.08 with higher values for AML samples, suggesting that they can be confidently analyzed in this space.

We defined CDRs altered in leukemias as those CDRs in which more than 50% of the tumoral samples show a chromatin state not observed in any normal sample. One of the advantages of this definition is that it is agnostic about the cell of origin of the tumor. Obviously, this definition, as any other, is limited to the cell types included in the study and some of these regions could be reclassified when more cell types (especially progenitor cell types) are available. In absence of more information, this criterium provides a simple definition of regions that are potentially important for tumoral progression.

Specifically altered regions in AML (or CLL or MCL) were defined as those CDRs with more than 50% of the AML (or CLL or MCL) samples presenting an unobserved state in normal cell types, but lower than 50% in the other two leukemia types. In both cases, CDRs altered in leukemia were analyzed using HOMER, as explained above. For exploring tumor-specific GO and Reactome enrichments, those terms enriched also in the whole set of CDRs were filtered out from altered CDR enrichments.

Resources

Method availability: https://github.com/david-juan/ChromDet

UCSC track hub to browse the CDRs and the chromatin states for all samples: http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&hubUrl=http://mcahematopoiesis.bioinfo.cnio.es/carrillo_et_al_NAR/hub.txt

Chromatin states model and segmentations: ftp://ftp.ebi.ac.uk/pub/databases/blueprint/paper_data_sets/chromatin_states_carrillo_build37

RESULTS

The chromatin space of human hematopoietic differentiation

We carried out a multi-group comparative analysis of chromatin states for representative cell types of the myeloid and lymphoid lineages to understand how epigenetic changes in chromatin are related to hematopoietic differentiation in humans. We focused our analysis on a set of 42 blood IHEC epigenomes from eight different cell types, with at least three independent biological replicates available: HSCs (HSC; $n = 3$), neutrophils ($n = 5$), monocytes ($n = 6$), macrophages ($n = 4$), naive and GC B cells ($n = 4$ and $n = 3$) and CD4 and CD8 T cells ($n = 13$ and $n = 4$), see Figure 2A and Supplementary Table S1 for details.

These epigenomes were assembled from ChIP-seq data generated by three IHEC consortia: BLUEPRINT ($n = 22$), NIH ROADMAP ($n = 18$) and ENCODE ($n = 2$). We integrated ChIP-Seq data experiments for the six core histone modification marks that are required to be included in IHEC epigenomes: H3K27ac marking active regulatory regions, H3K4me3 marking promoters (30,31); H3K4me1, related to enhancers (30); H3K36me3, marking transcription (30); H3K27me3 and H3K9me3, associated with polycomb and heterochromatin repression, respectively (30). Importantly, we only used histone mark sets where all six marks were profiled in the same individual (i.e. each epigenome corresponds to a unique individual).

A multivariate HMM was employed to learn combinatorial chromatin states based on the six histone marks using ChromHmm (18). However, others methods to segment the genome based on histone marks (or other features) could be used at this step, like Segway (32), EpicSeq (33), hiHMM (34), chromstaR (35), IDEAS (36) and others. In fact, the input for our method are the genome segmentations for the included samples. This means that users could use our method with the genome segmentations obtained by the software of his/her choice.

Further, the genome of each sample was segmented using the 11 combinatorial chromatin states model generated (see Supplementary Figure S1). To facilitate biological interpretation, the 11 chromatin states were further collapsed into 5 functional chromatin states encompassing five main categories: transcription (T), heterochromatin/low signal (H), repressed promoter (R), enhancer (E) and active promoter (A; see 'Materials and Methods' section for details; Supplementary Figure S1A). Thus, for each sample, we can create a vector representing the chromatin state of consecutive 200 bp windows along with the whole genome, using this reduced five-state alphabet. In order to reduce biases associated to the different size of each regulatory region, we collapse contiguous 200 bp windows having the same chromatin states pattern along all the samples (see 'Materials and Methods' section for details).

Our initial aim was to generate a low-dimensional chromatin space, a graphical representation of the structure and dimensionality of a complex and large data set, reflecting the major sources of epigenetic differences among hematopoietic samples (e.g. changes in chromatin states). To this end, we applied a protocol based on MCA, which we have previously applied to protein sequence (21) and gene expression (22) analysis. MCA is an analysis similar to PCA but appropriate for categorical data. We created an MCA-based multi-dimensional space in which the different samples are placed based on their vectors of chromatin states across the genome. The first stage of our protocol selects the minimal number of the most informative components that are relevant in this space, which already allows us to detect clusters of samples (see Figure 1).

Application of this approach to the matrix of collapsed chromatin states along the autosomal chromosomes in the 42 different samples results in a *hematological chromatin sample space* with the first two components as significantly informative according with a Wilcoxon test (Figure 2B; see 'Materials and Methods' section for details). Samples from the same cell type cluster together and the major blood cell types are clearly separated from each other, showing that the origin and technical biases of the samples are not affecting the results (three different consortia and therefore different laboratories). The relative samples distribution and the clustering are robust, as shown by analyzing each of the autosomal chromosomes independently (see Supplementary Figure S2).

As in PCA approaches, the interpretation of the two components selected by our method to separate the different cell types can lead to biological insight. Interestingly, the first component, represented on the horizontal axis, clearly separates myeloid (left side) from lymphoid cell types (right side) with HSCs situated in a central position. On the other hand, the second component on the vertical axis seems to reflect the lineage-independent epigenomic changes needed for the differentiation of the cell types from the HSCs, combined with the sample environment. We can draw a path from the pluripotent HSCs in bone marrow (at the bottom of the plot) all the way to the more mature cell types or subpopulations, such as *in vitro* cultured macrophages and GC B cells from tonsil (at the top of the plot). The central location of neutrophils, monocytes, T cells and naïve B cells from venous blood in this space suggests less epigenomic changes between these cell types and the HSCs (see Figure 2B). Interestingly, neutrophils and T cells are the cell types with least epigenomic changes from the HSCs. However, as in previous works based on single chromatin marks (37), we fail to discriminate CD4 and CD8 T cells, which form a tight cluster. In conclusion, our approach is able to capture the main biological differences between cell types, and is fully consistent with the known underlying biological process, showing that epigenomic states are an excellent source of information for discriminating these cell types.

Obviously, the value of the results obtained by our approach depends on the input information. Therefore, we strongly encourage introducing proper quality criteria to decide the inclusion of a sample in the analysis (see 'Materials and Methods' section). Interestingly, a detailed evaluation about the effect of using different input data from the segmentations (see Supplementary Data and Supplementary Figures S3–S5) supports the use of collapsed chromatin states to discriminate samples by cell type. These analyses identify elongation and enhancer states as the most informative sources of information, and illustrates the potential of our MCA-based approach for dealing with epigenomic data. Consequently, for studying CDRs associated to differentiation, we strongly recommend collapsing chromatin states into a small number of robustly defined states reflecting major functional shifts in transcription and enhancer activities, instead of more dynamic variations in the strength of the signal associated to these functions.

Chromatin determinant regions (CDRs)

So far we have shown how the MCA approach permits the generation of a space in which to robustly locate the different hematopoietic samples. Next, we aimed to identify the specific genomic regions that contribute most in defining specific cell types. We call these regions CDRs (Figure 1).

In order to retrieve these CDRs, we applied the second stage of our MCA-based protocol (21). This involves building a *hematological chromatin regions space*, in which each genomic region can be located based on its pattern of chromatin states across cell types (see Figure 1). For this we projected the chromatin states of every region of the genome on the same principal components of the hematological chromatin samples space. In this space we identify which regions have chromatin states that can discriminate the different cell types classified in the samples space (that is the different sample clusters). In practical terms, using this approach we find the CDRs that give rise to differences between cell types. For instance, a given region can show an enhancer state in lymphoid cell samples and a heterochromatin/low

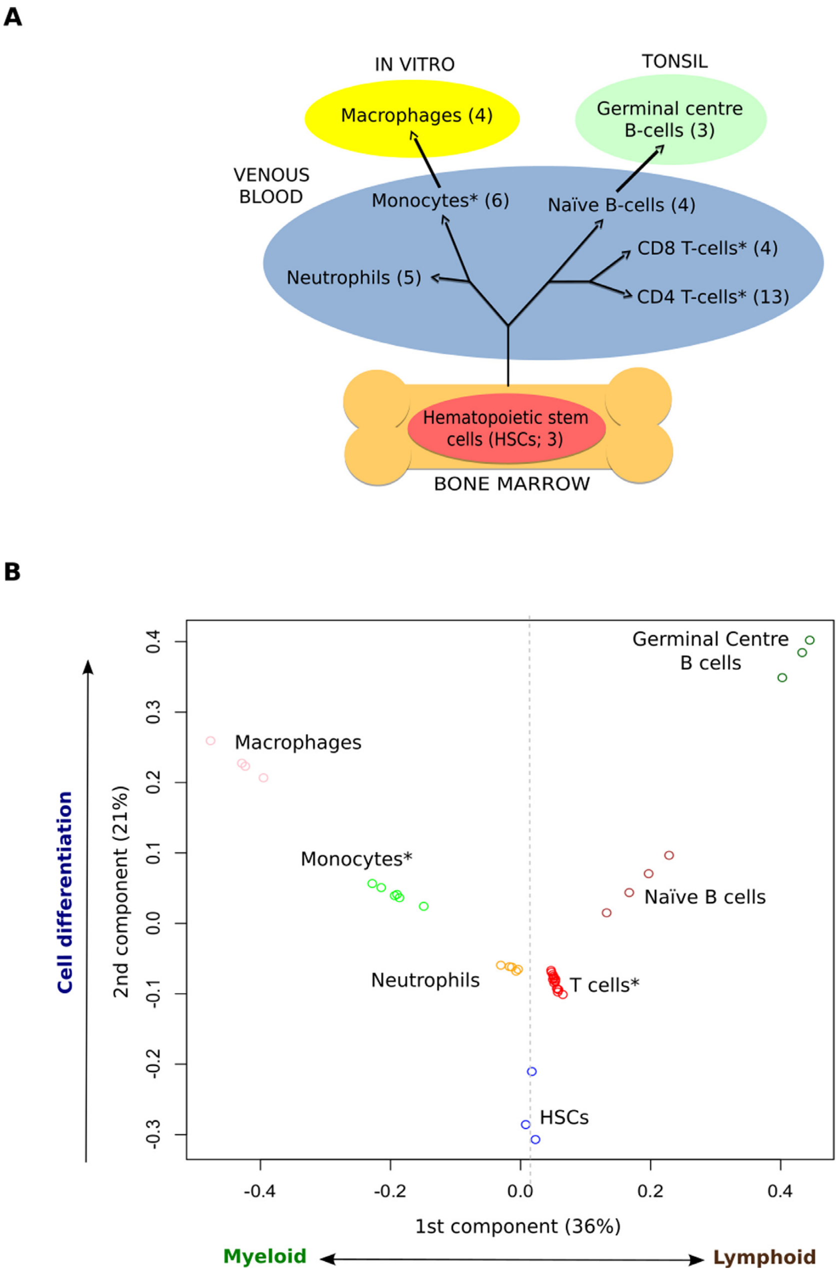

Figure 2. Hematopoietic cell types cluster based on chromatin states. (**A**) Schematic differentiation tree of the cell types considered, highlighting the tissue of origin and environment of each sample type. (**B**) Clustering of the samples in the MCA space recovers ontological relationships among cell types. (*Cell types with samples from different consortia) See also Supplementary Figure S2, Supplementary Table S1 and Supplementary Data.

signal state in the rest of the samples. In other cases, our protocol allows us to recover more complex patterns, such as those in regions able to discriminate more than two cell type groups. Starting from a total of 2,687,482 genomic regions for the 22 autosomal chromosomes included in the analysis, we recovered a total of 32,662 CDRs comprising 20,421,600 bp (a 0.71% of the canonical autosomal chromosome size) (see Supplementary Table S4).

As mentioned above, each CDR can be associated to a pattern of states across the different cell types, pointing to chromatin changes that might be drivers of cell differentiation. The most abundant CDR patterns we identified correspond to regions that have a transcription or enhancer state in one or two cell types, while having a heterochromatin/low signal state in the others (see Supplementary Figure S6 and Supplementary Table S5). The six most frequent patterns, that together comprise 61% of the CDRs, present transcription or enhancer states in GC B cells, HSC and macrophages, while having heterochromatin/low signal states in all other cell types (see Supplementary Figure S6 and Supplementary Table S5). In general, CDRs related to transcription states are larger than the ones showing patterns with other states (see Supplementary Figure S7). In addition, we can distinguish patterns that are cell type-specific (69, 3%), lineage-specific (16, 9%), which are shared by two or more close cell types, and others with more complex patterns between more distant cell types (13, 8%; see Supplementary Figure S8). We have included UCSC browser's screenshots of two interesting genes that show nearby CDRs, ABHD16B that shows transcription in the lymphoid lineage and LINC00494 active only in B cells (Supplementary Figure S9). As far as we know, these genes have not been previously related with hematopoiesis.

Recently, Corces *et al.* (38) generated ATAC-seq profiles to analyze the chromatin accessibility in a comprehensive collection of hematopoietic cell types, of which HSCs, B cells, T cells and monocytes are also included in our analysis. Around 10% of their defined set of 774 cell type-specific regions based on differential accessibility (38) overlapped with our defined CDRs. For these regions, we analyzed the ATAC-seq signal distributions for different CDR patterns (see Supplementary Figure S10). Importantly, we found that CDRs that show cell type-specific active state patterns in HSCs, B cells, T cells and monocytes respectively also show increased chromatin accessibility specifically for those cell types in the ATAC-seq data.

CDR chromatin state transitions across hematopoiesis

A more detailed analysis of the CDR transitions between cell types following the differentiation process can provide insights about chromatin remodeling across lineages. From the first pluripotent stage (HSCs), four possible second stages can be obtained (monocytes, neutrophils or naive B cells, T cells, according to the branch). After a further round of differentiation the third stage comprises macrophages (originating from monocytes) and GC B cells (originating from Naive B cells). Figure 3A shows transitions in CDR states across the various branches of the differentiation process. We observe the transitions from the HSCs to the second stage to be characterized by a turning off of active and enhancer CDRs. In contrast, in the second round of differentiation (from monocytes and naive B cells to macrophages and GC B cells, respectively) there is an increase in the activation of promoter and enhancer CDRs.

CDR association to genes and transcription factors binding sites

Chromatin state changes at CDRs might be pointing to drivers of cell differentiation and could be involved in regulating the expression of nearby genes that are important for these cell type transitions. We found most of the CDRs (94.5%) in intergenic and intronic regulatory regions, with an enrichment in the promoter and 5-UTR regions over the genomic background (see Supplementary Figure S11). A detailed annotation of each CDR is available in Supplementary Table S6. We associated each cell type-specific CDR to its most proximal gene and carried out a multi-group gene expression analysis of all the mature cell types, taking advantage of The BLUEPRINT Data Analysis Portal (BDAP; 29). The analysis was carried out on 7,764 genes with gene expression data available and associated CDRs, out of 60,483 included in BDAP, including protein-coding, ncRNA, pseudo, snoRNA and snRNA genes. The analysis showed that 81% had significant gene expression differences (adjusted P-value < 0.05) across the mature cell types (see Supplementary Table S7).

Further, functional-enrichment analyses were performed for the genes associated to each cell type-specific CDR having specifically active promoter, enhancer or transcription states (see Figure 3B; Supplementary Figures S12–S18; Supplementary Table S8; see 'Materials and Methods' section for details). As expected, genes proximal to the CDRs defining HSCs were mainly enriched in processes related to development and cell differentiation.

CDRs defining the myeloid lineage were close to genes related to tissue development and antimicrobial response among others. On the other hand, for CDRs defining the lymphoid lineage we found genes related to T-cell activation, cytokine production or response to interleukin-4, a cytokine produced by T cells involved in humoral and adaptive immunity (39). CDRs defining the two different B cell types were associated to genes with functions in proliferation and differentiation.

In addition, different neuron terms for differentiation and development were enriched for different cell types. These enrichments could be explained by the overlap in the molecular programs for hematopoiesis and neuropoiesis (40–42). The hematopoietic system is involved in many processes and genes related with neuronal development and function have been observed as expressed in different hematopoietic cell types (43). For example, we find a CDR overlapping with the gene encoding for Basp1 (Brain Abundant Membrane Attached Signal Protein) that belongs to many differentiation/morphogenesis-related GO terms, including 'central nervous system development'. This and other related neuronal genes were shown to be upregulated in GC B cells, where its pattern of gene expression is associated to the development of neurite-like projections of the membrane (44). Furthermore, interactions between the nervous and immune systems are required for organ function and homeostasis (45). A report has shown that primary CD34+ hHSCs express mRNA for a number of proteins that are used by neurons (among other cell types), including receptors for trophic factors and other mediators that are known to influence neuronal development (42). Finally, the similarity between these two differentiation programs could explain the fact that HSCs can differentiate to neural cells, albeit at relatively low efficiency (46–48).

We next asked whether CDRs involving cell type-specific active promoter or enhancer states were enriched in TFMs (see 'Materials and Methods' section for details). Hierarchical clustering based on the TFM enrichment patterns clearly separates the HSC TFMs profiles from those of the myeloid and the lymphoid cell types (Figure 3C). A detailed annotation for motifs in each CDR is available in Supplementary Table S6.

We observed in HSCs a specific motif enrichment for GATA factors, which have been related to regulation of the self-renewal of long-term HSCs and differentiation of bone marrow-derived mesenchymal stem cells (49–52). Enrichment in binding motifs for factors like RUNX, implied in stem cell fate maintenance and normal function, was also observed in HSCs-specific CDRs (53,54). GATA and RUNX factors were described by Corces *et al.* (38) as domi-

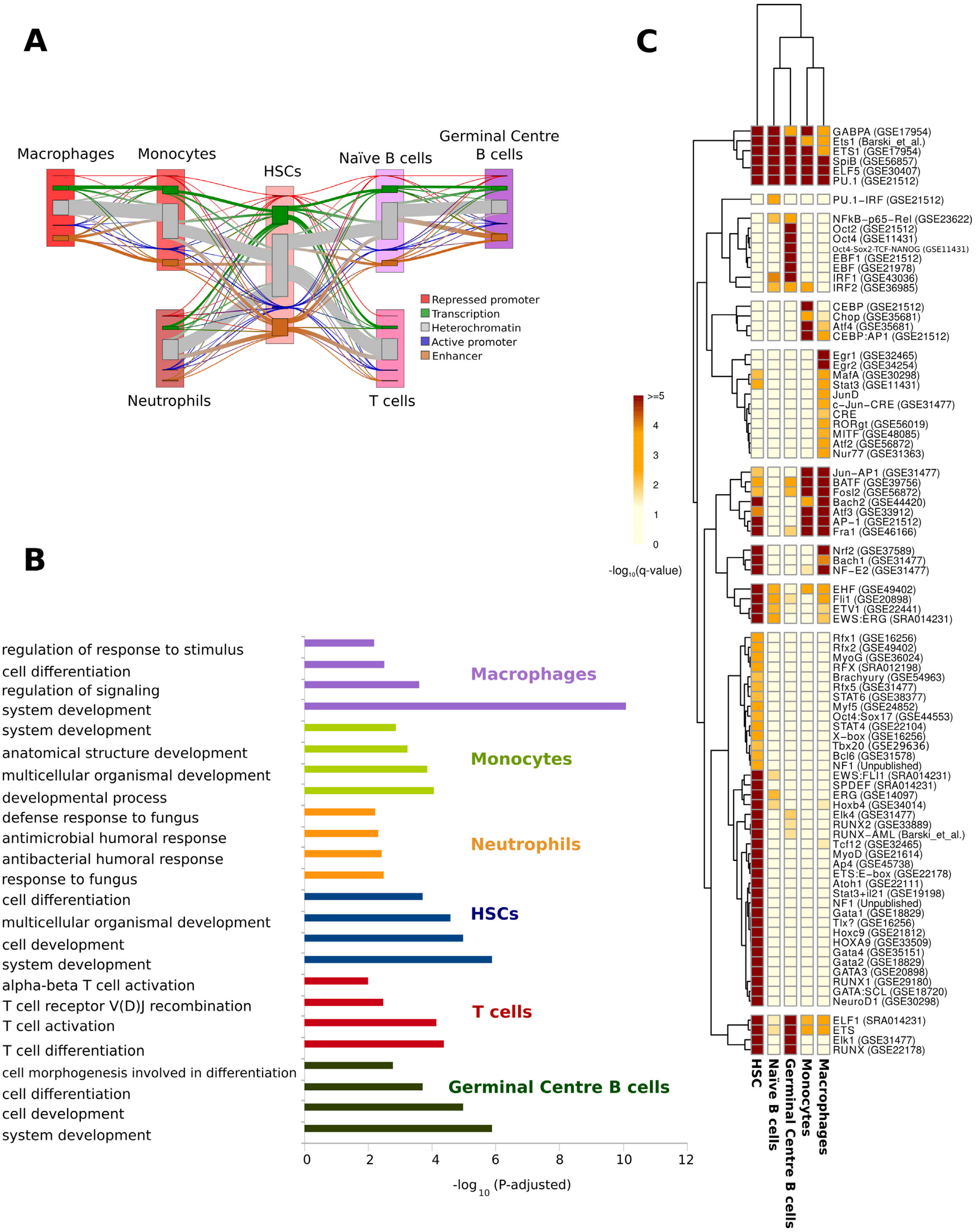

Figure 3. Functional and transcription factor binding motifs characterization of chromatin determinant regions (CDRs). (**A**) Sankey Plot representation of chromatin state transitions at CDRs during hematopoietic cell differentiation. Nodes for each cell type represent the five 'collapsed' chromatin states (see 'Materials and Methods' section). For each pair of cell types in the hematopoietic differentiation pathway, flows, represented by line thickness, are proportional to the number of regions that show a transition between a particular pair of states. Changes in chromatin states between two stages of differentiation are shown with lines that change color. The thickness of the lines is proportional to the number of regions that show a transition between a particular pair of states. (**B**) Enriched ontology terms from the genes related to the CDRs that characterize each cell type. (**C**) Heatmap and hierarchical clustering based on transcription binding proteins enriched in the CDRs that characterize each cell type (see 'Materials and Methods' section). See also Supplementary Figures S3–15, Tables S2–6 and File S1.

nant regulators of chromatin accessibility in hematopoiesis. Interestingly, motifs for the so far uncharacterized factor X gene family, known to regulate the major histocompatibility complex class II (55), were also exclusively enriched in CDRs specific for HSCs.

In myeloid cell types, CDRs specific to monocytes are enriched in binding motifs for the C/EBP homologous protein (CEBP/CHOP) and its interactor ATF4 (56,57), which plays a key role during the differentiation of the monocyte lineage (58,59). In contrast, EGR1 and EGR2 binding motifs, which are essential for macrophage but not for granulocyte differentiation (60,61), are enriched in macrophages. Higher expression at RNA level is observed in macrophages compared with monocytes and mature neutrophils (see Supplementary Figure S19). In addition, enrichment for transcription factor binding sites related to macrophage differentiation like STAT3, JUND, MITF, NUR77 or ATF2 (62–66) is observed in CDRs specific to macrophages (Figure 3C).

Binding motifs for members of the NF–KB complex (NF–KB, RELA, IRF2), implicated in stimulus response, were enriched in CDRs characterizing GC B cells. It is known that defects of this complex in GCs affect their maintenance and B-cell differentiation (67,68). In addition, we observed enriched motifs for Early B-cell factor 1 (EBF1), a central transcription factor in B cells implicated in GC formation and class switch recombination (69,70), Oct2 and Fli1, transcription factors expressed in B cells and related to normal B-cells proliferation (71,72).

TFMs from the ETS transcription factor family genes (GABPA, ETS1, SpiB, PU.1 and ELF5) were enriched in all cell type-specific CDRs. These gene families are ubiquitously expressed in the different blood cell types, although they are known to play specific roles in different cell types. For example, in monocytes, PU.1 regulates the transcription of a large proportion of myeloid-specific genes, while in B cells it is involved in regulating the transcription of the heavy and light immunoglobulin chain genes (73).

Finally, we took advantage of BDAP expression data for 28 transcription factors whose DNA binding motifs were enriched in CDRs (Figure 3C) and for which expression data were available. We excluded transcription factors whose binding motifs were specifically enriched in HSCs as this immature cell type is not included in BDAP. A subset of 96.5% (27/28) of them showed differential expression between cell types (see Supplementary Table S7). In addition, we observed that 60% (16/27) of the transcription factors with changes in expression also have a CDR associated to them by proximity, suggesting a central role for chromatin regulatory regions in the hematopoietic regulatory network.

Taken together, the gene expression, GO and TFM-enrichment analyses suggest that the identified CDRs are indeed important functional regions, where chromatin remodeling is linked to cell fate. Overall, we have shown that our approach is useful to identify key and potentially driver local changes in the epigenomes of healthy cells across different hematopoietic lineages.

Clustering of healthy and leukemic samples based on CDRs

The framework explained above allowed us to identify specific genomic regions that are under epigenetic control and might contribute to define blood cell types. This framework can be further exploited to analyze the relationships between leukemia and healthy cell types.

Extensive epigenetic changes are common in most leukemias and solid tumors (74) and epigenetic features such as DNA methylation or open chromatin have been shown to be useful to identify the cell of origin of tumors (75,76). However, given the extensive genome-wide epigenetic alterations of tumor cells, matching tumoral cells with their healthy counterparts is a great challenge and an essential step to identify the chromatin changes leading to malignancy.

The CDRs constitute an epigenetic signature of hematopoiesis. Therefore, we reasoned that they should be useful to classify blood cancer samples according to their similarity to normal cell types. We used the data generated by The BLUEPRINT consortium for three hematopoietic neoplasms, including six CLLs, three AMLs and three MCLs to explore the epigenetic similarity among healthy and cancerous samples.

We projected the leukemic samples on the healthy hematopoietic chromatin space, based on their chromatin states at CDRs (see Supplementary Figure S20 and 'Materials and Methods' section). Next, we used the distance of each leukemic sample to a reference healthy cell type (Supplementary Figure S20) to quantify the similarities and differences observed at the CDRs level between healthy and disease epigenomes.

The distribution of the leukemia samples in the CDRs healthy hematopoietic chromatin sample space separates them into two main groups. The AML samples localized into the myeloid region of the space, while the CLL and MCL samples were in the lymphoid region (see Supplementary Figure S20). A hierarchical clustering based on the distances of each leukemia sample to each reference healthy cell type shows that CLL and MCL samples both cluster with the reference Naïve B cell (see Figure 4; cluster I). In contrast, AML samples are distributed in more than one cluster, with two samples clustering within the reference neutrophil cluster IV and the other one within the reference monocyte cluster II, suggesting a different origin for these tumors.

Each tumoral sample was projected onto the healthy hematopoietic chromatin sample space using the CDRs whose chromatin states are represented in any of the healthy cell types (see 'Materials and Methods' section). However, there is a variable number of CDRs per tumoral sample whose chromatin state is not represented in the normal cell types. We can view these chromatin states either as features related to maturation stages of cells not included in our analyses, or as changes that have occurred specifically in the malignant transformation. Interestingly, we can observe characteristic divergence patterns for the different neoplasms (Figure 4). AML samples appear to be epigenetically more divergent from the healthy states than those closer to the B-cell derived cancer samples.

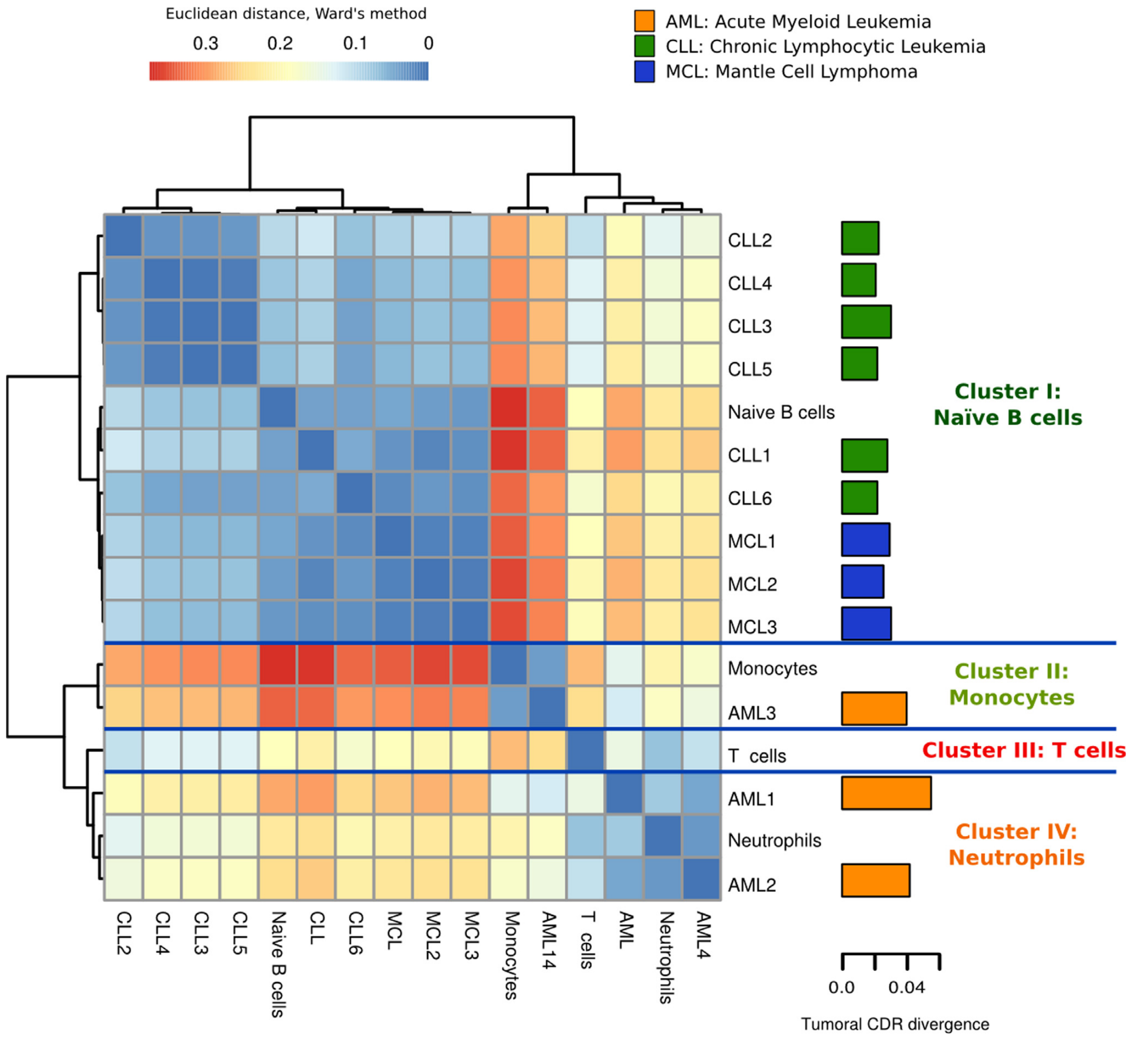

Figure 4. Hierarchical clustering of leukemias based on CDRs of healthy cell types suggests potential lineage origin of tumors. The healthy cell type clusters are summarized by each fingerprint, a reference sample representing each cell type cluster. The euclidean distances between samples and fingerprints are calculated with the Ward's method. The barplot in the right shows the epigenomic divergence (ratio of chromatin changes in CDRs) of each cancer sample to the healthy states. See also Supplementary Figures S16–20 and Tables S7–9.

We analyzed these divergent CDRs as a potential source of information about epigenomic alterations that might be important for tumoral transformation. To this end, we focused on those 297 CDRs where most of the leukemic samples have a potentially unhealthy chromatin state (a chromatin state that is never observed in the healthy samples, see Materials and Methods' section and Supplementary Table S9). We proceeded by associating these CDRs to genes by proximity and investigated whether these genes were commonly regulated or mutated in tumors.

Functional enrichment analysis of the 177 genes associated to these CDRs by proximity showed that they are mutated in a large number of tumors, including AML and CLL (COSMIC tumoral signatures (77), P-value < 0.05, see Table 1 and Supplementary Table S10). An analysis of those CDRs specifically altered in each of the leukemias (282 CDRs in AML, 591 in CLL and 727 in MCL) shows similar results. However, we observed that mutational signatures associated specifically to AML, CLL and MCL were enriched only in tumors different from the leukemia where we detected the epigenomic alteration (see Table 1 and Supplementary Table S10).

On the contrary, when comparing with gene signatures regulated in tumors, we found that genes associated to altered CDRs in AML and CLL respectively are enriched in the expression signatures of the corresponding leukemias (MSigDB gene expression signatures (78), P-value < 0.05, see Table 1 and Supplementary Table S10). This result supports that these alterations in CDRs are linked to the detected gene expression changes in the associated genes (see Table 1 and Supplementary Table S10).

We also found that the three sets of genes specifically altered in the different leukemias are all enriched in the same general processes: differentiation and development, cell–cell adhesion, endocytosis and phagocytosis or metabolic processes (GO biological process, P-value < 0.05, see Supplementary Table S11 and Figures S21–S24). Although these sets of genes are related to similar processes, they con-

Table 1. COSMIC and MSigDB enrichments for leukemia-altered CDRs

Terms COSMIC	**GLOBAL_0.5**	**AML_0.5**	**CLL_0.5**	**MCL_0.5**
diffuse_large_B_cell_lymphoma	0.00	0.03	0.00	0.00
acute_myeloid_leukaemia_therapy_related	0.01	1.00	0.00	0.00
haematopoietic_and_lymphoid_tissue-haematopoietic_neoplasm-acute_myeloid_leukaemia_therapy_related	0.01	1.00	0.00	0.00
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-diffuse_large_B_cell_lymphoma	0.01	0.03	0.00	0.00
acute_myeloid_leukaemia	0.01	1.00	0.00	0.02
haematopoietic_and_lymphoid_tissue-haematopoietic_neoplasm-acute_myeloid_leukaemia	0.01	1.00	0.00	0.02
haematopoietic_and_lymphoid_tissue-haematopoietic_neoplasm	0.02	1.00	0.00	0.00
haematopoietic_neoplasm	0.06	1.00	0.00	0.00
haematopoietic_and_lymphoid_tissue-lymph_node-lymphoid_neoplasm-acute_lymphoblastic_leukaemia	0.12	1.00	0.00	0.00
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm	0.19	0.00	0.01	0.00
Terms MsigDB	**GLOBAL_0.5**	**AML_0.5**	**CLL_0.5**	**MCL_0.5**
GUTIERREZ_CHRONIC_LYMPHOCYTIC_LEUKEMIA_DN	0.02	1.00	0.00	1.00
HUTTMANN_B_CLL_POOR_SURVIVAL_DN	1.00	1.00	0.00	1.00
VALK_AML_WITH_CEBPA	1.00	1.00	1.00	0.02

The values in orange background are statistically significant. See also Supplementary Table S10.

tain different genes (only three genes in common among the three leukemias) and they are related to different detailed functions. In fact, genes associated to CDRs altered in most AML samples are mainly enriched in membrane transporters and metabolic pathways, those altered in CLL are enriched in many signal transduction pathways (VEGF, WNT, FGFR, ERBB or MAPK signaling) and those in MCL in morphogenetic and developmental processes (*P*-value < 0.05, see Supplementary Table S11). These observations draw a scenario where leukemic mutations and epigenomic alterations point to the same processes that are key for tumor progression, but involve different genes in a leukemia-specific way. Taken together, these results show the potential of our proposed CDR approach to characterize hematopoietic cell types in normal differentiation and disease.

DISCUSSION

Chromatin remodeling is an essential process for determining the set of phenotypes deployed by eukaryotic cells. Chromatin regulation is based on combinatorial associations among proteins and complex communication networks, which define the functional states of the different genomic/chromatin regions (12). These functional states play a determinant role to define cell identity during the differentiation process. Despite the great efforts made in the last few years to generate functional chromatin maps for many cell types (19,37), we are still far from identifying the genomic regions where driver chromatin changes occur, their association with functional changes that give the cell its identity during development or their implications in disease.

Hematopoiesis is possibly the best characterized differentiation process, usually represented by a hierarchical tree based on morphological criteria and refined with surface markers (1). Hematopoiesis provides a well-defined model to study cell differentiation from an epigenetic perspective. We face the challenge of studying this process by integrating epigenomic information from multiple human blood cell types and different data sources. The blood IHEC epigenomes provide a unique opportunity to investigate the epigenetic basis of lineage determination.

We have developed a new protocol, based on a useful and powerful multivariate framework based on a rigorous statistical approach, to define in an unsupervised manner which cell types are epigenetically distinguishable. Importantly, we simultaneously identify the key genomic regions driving these differences. These regions, named CDRs, can be considered as the epigenetic signatures of human hematopoiesis, a set of reference regions that through their epigenetic changes might be able to drive hematopoiesis.

The results are robust to the possible noise introduced by consortia-specific protocols and the clusters obtained provide perfect classification of samples in the different cell types. We observed clear clusters for seven cell types plus an additional cluster for CD4+ and CD8+ T cells. Interestingly, a recent work using H3K4me1 and H3K27me3 histone modifications independently was also unable to discriminate CD4+ from CD8+ T cell types (37), supporting the hypothesis that the epigenomes of these cell types are very similar.

The sample space, in addition to clearly separating the myeloid from the lymphoid lineages, reflects the epigenetic distance of each cluster from the HSC. Although both the classical and the more recent alternative hematopoietic hierarchical differentiation models propose a similar differentiation distance for neutrophils and monocytes or T and B cells (1,2), our space shows clearly very different epigenetic differentiation distances for neutrophils and monocytes, as well as for T and B cells. These differences suggest that cell types with shorter epigenetic distances from HSCs may reach the mature state earlier. In the case of murine fetal liver T and B cells, it is known that the T-cell progenitors appear earlier than the B cells ones (79).

The classical hematopoietic model establishes that the HSCs differentiate into the common myeloid progenitor (CMP) or the common lymphoid progenitor (CLP), divided in the myeloid and the lymphoid lineages (1). However, this model is under discussion, as it has been shown by Kawamoto *et al*. (79) and other authors (80–83) that the T- and B-cell progenitors retain the potential to differentiate into myeloid cells. These results have led to the proposal of an alternative 'myeloid-based' model for hematopoiesis (79), which would suggest that the two main branches are not as well separated as initially thought. Interestingly, we found that the epigenetic distance between neutrophils and T cells is very short in our model, both cell types being very close to the HSC group.

Unfortunately, although our CDRs refer to chromatin changes during all the lineage differentiation steps, data for progenitors (GMP, CMP, CLP, MPP, ...) do not meet the IHEC standards and could not be used in our analysis. Therefore, we can not assign each CDR to the precise intermediate cell type in which it was originated. The future availability of complete epigenomes for more cell lineages, including intermediate progenitors, will provide additional information to assess whether the myeloid-based differentiation model proposed by Kawamoto *et al*. (79) is consistent with the chromatin landscape.

The strength of our protocol, beyond providing a classification of cell types, is to identify the CDRs that drive human hematopoiesis. We detected 32,662 CDRs that represent the epigenetic signature of hematopoiesis for the cell types included in the analysis. Interestingly, we observed that all the transitions starting from HSCs to other cell types were enriched in epigenetic inactivation, while the Monocytes-to-macrophages and naive-to-GC B-cells transitions are enriched in epigenetic activation. These results suggest that the differentiation process involves a first phase characterized by loss of stemness through epigenetic repression of the HSC processes, followed by activation of more specific regulatory programs that define specific differentiated cell types (84–87).

A further characterization of these CDRs showed that they are enriched in DNA binding motifs of transcription factors with a key role in hematopoiesis. These results support the idea of CDRs as driver regions whose chromatin reconfiguration is associated to cell type-specific regulatory programs. Moreover, we also observed that these regions are proximal to genes with functions in cell differentiation and cell type- or lineage-specific processes, coherent with the transitions reflected by the epigenetic pattern of the regions.

As only a subset of blood cell types was used in this analysis, these CDRs have to be seen as only a first approximation to understand human hematopoiesis from an epigenetic perspective. It is important to note that other previous models were proposed based on surface markers (88) or mice models with DNA methylation (4) and transcriptomics (2). Although the human hematopoietic differentiation model closely resembles the murine one, accumulated evidence has shown that they differ in important aspects. For example, the HSC immunophenotypes (1) or hematopoietic gene regulation programs are not fully conserved between species (89).

In addition to providing a useful epigenetic signature of hematopoiesis, we have also shown that the CDRs could provide useful information about disease-related epigenetic features. We applied our method to study the epigenetic similarities between leukemias and healthy cell types by projecting the leukemia samples in the space generated with the CDRs. We hypothesized that leukemia derived from certain healthy cell types would maintain the epigenetic CDR signature of its cell of origin. Indeed, our approach recovers a coherent distribution of hematological cancers, with B-cell neoplasms clustering close to B naïve cells, and a more heterogeneous classification of the AML samples. AML is known to be a very heterogeneous disease with many different subtypes and a difficult clinical classification (90,91), which would explain why two of the AML samples cluster close to neutrophils, and the other one with monocytes. In addition, we also performed a functional analysis of the CDRs more recurrently epigenetically changed in different leukemias, showing that they tend to target general processes (such as differentiation and development, cell–cell adhesion, endocytosis and phagocytosis). Interestingly, different genes within these pathways are either epigenetically altered or mutated in the specific leukemias, suggesting mutual exclusivity of the two types of alterations in the same genes. In summary, our proof of concept application of the epigenetic signature of hematopoiesis in the study of leukemia shows the power of our methodology. Only when more leukemia and complete progenitor epigenomes will become available, we will be able to exploit the full potential of this approach.

In conclusion, our results have shown the value of our multivariate framework in investigating the differentiation processes. We propose a catalog of epigenetic signatures of human hematopoiesis, based on the CDRs that best describe the different cell types. This catalog, with further refinements by the inclusion of additional cell types and hematopoietic progenitors, could become the reference IHEC resource for human hematopoiesis studies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the joint BSC-CRG-IRB Research Program in Computational Biology, José María Fernández and Jon Sánchez (CNIO and BSC) and the BLUEPRINT Data Coordination Centre team for technical support. Vera Pancaldi acknowledges a CNIO-Friends fellowship.

FUNDING

European Union's Seventh Framework Programme [FP7/2007–2013, 282510 (BLUEPRINT)]; Spanish Ministry of Economy, Industry and Competitiveness and European Regional Development Fund [Project Retos BFU2015–71241-R]. Funding for open access charge: Project Retos BFU2015–71241-R (to A.V.).

Conflict of interest statement. None declared.

REFERENCES

1. Rieger,M.A. and Schroeder,T. (2012) Hematopoiesis. *Cold Spring Harb. Perspect. Biol.*, **4**, a008250.
2. Paul,F., Arkin,Y., Giladi,A., Jaitin,D.A., Kenigsberg,E., Keren-Shaul,H., Winter,D., Lara-Astiaso,D., Gury,M., Weiner,A. *et al.* (2015) Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, **163**, 1663–1677.
3. Ji,H., Ehrlich,L.I.R., Seita,J., Murakami,P., Doi,A., Lindau,P., Lee,H., Aryee,M.J., Irizarry,R.A., Kim,K. *et al.* (2010) Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature*, **467**, 338–342.
4. Bock,C., Beerman,I., Lien,W.-H., Smith,Z.D., Gu,H., Boyle,P., Gnirke,A., Fuchs,E., Rossi,D.J. and Meissner,A. (2012) DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. *Mol. Cell*, **47**, 633–647.
5. Lara-Astiaso,D., Weiner,A., Lorenzo-Vivas,E., Zaretsky,I., Jaitin,D.A., David,E., Keren-Shaul,H., Mildner,A., Winter,D., Jung,S. *et al.* (2014) Immunogenetics. Chromatin state dynamics during blood formation. *Science*, **345**, 943–949.
6. Ho,L. and Crabtree,G.R. (2010) Chromatin remodelling during development. *Nature*, **463**, 474–484.
7. Kumar,R., Li,D.-Q., Müller,S. and Knapp,S. (2016) Epigenomic regulation of oncogenesis by chromatin remodeling. *Oncogene*, **35**, 4423–4436.
8. Ronan,J.L., Wu,W. and Crabtree,G.R. (2013) From neural development to cognition: unexpected roles for chromatin. *Nat. Rev. Genet.*, **14**, 347–359.
9. Mirabella,A.C., Foster,B.M. and Bartke,T. (2016) Chromatin deregulation in disease. *Chromosoma*, **125**, 75–93.
10. Bannister,A.J. and Kouzarides,T. (2011) Regulation of chromatin by histone modifications. *Cell Res.*, **21**, 381–395.
11. Venkatesh,S. and Workman,J.L. (2015) Histone exchange, chromatin structure and the regulation of transcription. *Nat. Rev. Mol. Cell Biol.*, **16**, 178–189.
12. Juan,D., Perner,J., Carrillo de Santa Pau,E., Marsili,S., Ochoa,D., Chung,H.-R., Vingron,M., Rico,D. and Valencia,A. (2016) Epigenomic co-localization and co-evolution reveal a key role for 5hmC as a communication hub in the chromatin network of ESCs. *Cell Rep.*, **14**, 1246–1257.
13. Doulatov,S., Notta,F., Laurenti,E. and Dick,J.E. (2012) Hematopoiesis: a human perspective. *Cell Stem Cell*, **10**, 120–136.
14. Rice,K.L., Hormaeche,I. and Licht,J.D. (2007) Epigenetic regulation of normal and malignant hematopoiesis. *Oncogene*, **26**, 6697–6714.
15. Torres-Lacomba,A. (2006) Correspondence analysis and categorical conjoint measurement. In: Greenacre,M and Blasius,J (eds). *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall/CRC Press, London, pp. 421–432.
16. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
17. Roadmap Epigenomics Consortium, Kundaje,A, Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

18. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
19. Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.
20. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
21. Rausell,A., Juan,D., Pazos,F. and Valencia,A. (2010) Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 1995–2000.
22. Martinez-Garcia,R., Juan,D., Rausell,A., Muñoz,M., Baños,N., Menéndez,C., Lopez-Casas,P.P., Rico,D., Valencia,A. and Hidalgo,M. (2014) Transcriptional dissection of pancreatic tumors engrafted in mice. *Genome Med.*, **6**, 27.
23. de Hoon,M.J.L., Imoto,S., Nolan,J. and Miyano,S. (2004) Open source clustering software. *Bioinformatics*, **20**, 1453–1454.
24. Calinski,T. and Harabasz,J. (1974) A dendrite method for cluster analysis. *Commun. Stat.*, **3**, 1–27.
25. 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.
26. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Michael Cherry,J., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
27. Fabregat,A., Sidiropoulos,K., Garapati,P., Gillespie,M., Hausmann,K., Haw,R., Jassal,B., Jupe,S., Korninger,F., McKay,S. *et al.* (2016) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **44**, D481–D487.
28. Supek,F., Bošnjak,M., Škunca,N. and Šmuc,T. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, **6**, e21800.
29. Fernández,J.M., de la Torre,V., Richardson,D., Royo,R., Puiggròs,M., Moncunill,V., Fragkogianni,S., Clarke,L. and BLUEPRINT ConsortiumBLUEPRINT Consortium and Flicek,P. *et al.* (2016) The BLUEPRINT data analysis portal. *Cell Syst.*, **3**, 491–495.
30. Barski,A., Cuddapah,S., Cui,K., Roh,T.-Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
31. Creyghton,M.P., Cheng,A.W., Welstead,G.G., Kooistra,T., Carey,B.W., Steine,E.J., Hanna,J., Lodato,M.A., Frampton,G.M., Sharp,P.A. *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 21931–21936.
32. Hoffman,M.M., Buske,O.J., Wang,J., Weng,Z., Bilmes,J.A. and Noble,W.S. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.
33. Mammana,A. and Chung,H.-R. (2015) Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biol.*, **16**, 151.
34. Sohn,K.-A., Ho,J.W.K., Djordjevic,D., Jeong,H.-H., Park,P.J. and Kim,J.H. (2015) hiHMM: Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics*, **31**, 2066–2074.
35. Taudt,A., Nguyen,M.A., Heinig,M., Johannes,F. and Colome-Tatche,M. (2016) chromstaR: tracking combinatorial chromatin state dynamics in space and time. *bioRxiv*, doi:10.1101/038612.
36. Zhang,Y., An,L., Yue,F. and Hardison,R.C. (2016) Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res.*, **44**, 6721–6731.
37. Roadmap Epigenomics Consortium, Kundaje,A, Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
38. Corces,M.R., Buenrostro,J.D., Wu,B., Greenside,P.G., Chan,S.M., Koenig,J.L., Snyder,M.P., Pritchard,J.K., Kundaje,A., Greenleaf,W.J. *et al.* (2016) Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.*, **48**, 1193–1203.
39. Luzina,I.G., Keegan,A.D., Heller,N.M., Rook,G.A.W., Shea-Donohue,T. and Atamas,S.P. (2012) Regulation of inflammation by interleukin-4: a review of ‘alternatives’. *J. Leukoc. Biol.*, **92**, 753–764.
40. Terskikh,A.V., Easterday,M.C., Li,L., Hood,L., Kornblum,H.I., Geschwind,D.H. and Weissman,I.L. (2001) From hematopoiesis to neuropoiesis: evidence of overlapping genetic programs. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 7934–7939.
41. Goolsby,J., Marty,M.C., Heletz,D., Chiappelli,J., Tashko,G., Yarnell,D., Fishman,P.S., Dhib-Jalbut,S., Bever,C.T. Jr, Pessac,B. *et al.* (2003) Hematopoietic progenitors express neural genes. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 14926–14931.
42. Steidl,U., Bork,S., Schaub,S., Selbach,O., Seres,J., Aivado,M., Schroeder,T., Rohr,U.-P., Fenk,R., Kliszewski,S. *et al.* (2004) Primary human CD34+ hematopoietic stem and progenitor cells express functionally active receptors of neuromediators. *Blood*, **104**, 81–88.
43. Niemi,J.P., DeFrancesco-Lisowitz,A., Roldán-Hernández,L., Lindborg,J.A., Mandell,D. and Zigmond,R.E. (2013) A critical role for macrophages near axotomized neuronal cell bodies in stimulating nerve regeneration. *J. Neurosci.*, **33**, 16236–16248.
44. Yu,D., Cook,M.C., Shin,D.-M., Silva,D.G., Marshall,J., Toellner,K.-M., Havran,W.L., Caroni,P., Cooke,M.P., Morse,H.C. *et al.* (2008) Axon growth and guidance genes identify T-dependent germinal centre B cells. *Immunol. Cell Biol.*, **86**, 3–14.
45. Veiga-Fernandes,H. and Pachnis,V. (2017) Neuroimmune regulation during intestinal development and homeostasis. *Nat. Immunol.*, **18**, 116–122.
46. Mezey,E., Key,S., Vogelsang,G., Szalayova,I., Lange,G.D. and Crain,B. (2003) Transplanted bone marrow generates new neurons in human brains. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 1364–1369.
47. Wagers,A.J., Sherwood,R.I., Christensen,J.L. and Weissman,I.L. (2002) Little evidence for developmental plasticity of adult hematopoietic stem cells. *Science*, **297**, 2256–2259.
48. Cogle,C.R., Yachnis,A.T., Laywell,E.D., Zander,D.S., Wingard,J.R., Steindler,D.A. and Scott,E.W. (2004) Bone marrow transdifferentiation in brain after transplantation: a retrospective study. *Lancet*, **363**, 1432–1437.
49. Kamata,M., Okitsu,Y., Fujiwara,T., Kanehira,M., Nakajima,S., Takahashi,T., Inoue,A., Fukuhara,N., Onishi,Y., Ishizawa,K. *et al.* (2014) GATA2 regulates differentiation of bone marrow-derived mesenchymal stem cells. *Haematologica*, **99**, 1686–1696.
50. Frelin,C., Herrington,R., Janmohamed,S., Barbara,M., Tran,G., Paige,C.J., Benveniste,P., Zuñiga-Pflücker,J.-C., Souabni,A., Busslinger,M. *et al.* (2013) GATA-3 regulates the self-renewal of long-term hematopoietic stem cells. *Nat. Immunol.*, **14**, 1037–1044.
51. Ku,C.-J., Hosoya,T., Maillard,I. and Engel,J.D. (2012) GATA-3 regulates hematopoietic stem cell maintenance and cell-cycle entry. *Blood*, **119**, 2242–2251.
52. Bresnick,E.H., Katsumura,K.R., Lee,H.-Y., Johnson,K.D. and Perkins,A.S. (2012) Master regulatory GATA transcription factors: mechanistic principles and emerging links to hematologic malignancies. *Nucleic Acids Res.*, **40**, 5819–5831.
53. Staber,P.B., Zhang,P., Ye,M., Welner,R.S., Levantini,E., Di Ruscio,A., Ebralidze,A.K., Bach,C., Zhang,H., Zhang,J. *et al.* (2014) The Runx-PU.1 pathway preserves normal and AML/ETO9a leukemic stem cells. *Blood*, **124**, 2391–2399.
54. Burns,C.E. (2005) Hematopoietic stem cell fate is established by the Notch-Runx pathway. *Genes Dev.*, **19**, 2331–2342.
55. Reith,W. and Mach,B. (2001) The bare lymphocyte syndrome and the regulation of MHC expression. *Annu. Rev. Immunol.*, **19**, 331–373.
56. Bruhat,A., Chérasse,Y., Maurin,A.-C., Breitwieser,W., Parry,L., Deval,C., Jones,N., Jousse,C. and Fafournoux,P. (2007) ATF2 is required for amino acid-regulated transcription by orchestrating specific histone acetylation. *Nucleic Acids Res.*, **35**, 1312–1321.
57. Gombart,A.F., Grewal,J. and Koeffler,H.P. (2007) ATF4 differentially regulates transcriptional activation of myeloid-specific genes by C/EBP and C/EBP. *J. Leukoc. Biol.*, **81**, 1535–1547.
58. Radomska,H.S., Huettner,C.S., Zhang,P., Cheng,T., Scadden,D.T. and Tenen,D.G. (1998) CCAAT/Enhancer Binding Protein α Is a Regulatory Switch Sufficient for Induction of Granulocytic Development from Bipotential Myeloid Progenitors. *Mol. Cell. Biol.*, **18**, 4301–4314.

59. Yeamans,C., Wang,D., Paz-Priel,I., Torbett,B.E., Tenen,D.G. and Friedman,A.D. (2007) C/EBP binds and activates the PU.1 distal enhancer to induce monocyte lineage commitment. *Blood*, **110**, 3136–3142.
60. Laslo,P., Spooner,C.J., Warmflash,A., Lancki,D.W., Lee,H.-J., Sciammas,R., Gantner,B.N., Dinner,A.R. and Singh,H. (2006) Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell*, **126**, 755–766.
61. Nguyen,H.Q., Hoffman-Liebermann,B. and Liebermann,D.A. (1993) The zinc finger transcription factor Egr-1 is essential for and restricts differentiation along the macrophage lineage. *Cell*, **72**, 197–209.
62. Behmoaras,J., Bhangal,G., Smith,J., McDonald,K., Mutch,B., Lai,P.C., Domin,J., Game,L., Salama,A., Foxwell,B.M. *et al.* (2008) Jund is a determinant of macrophage activation and is associated with glomerulonephritis susceptibility. *Nat. Genet.*, **40**, 553–559.
63. Coffer,P.J., Koenderman,L. and de Groot,R.P. (2000) The role of STATs in myeloid differentiation and leukemia. *Oncogene*, **19**, 2511–2522.
64. Hanna,R.N., Shaked,I., Hubbeling,H.G., Punt,J.A., Wu,R., Herrley,E., Zaugg,C., Pei,H., Geissmann,F., Ley,K. *et al.* (2012) NR4A1 (Nur77) deletion polarizes macrophages toward an inflammatory phenotype and increases atherosclerosis. *Circ. Res.*, **110**, 416–427.
65. Hume,D.A. (2015) The many alternative faces of macrophage activation. *Front. Immunol.*, **6**, 370.
66. Miyata,Y., Fukuhara,A., Otsuki,M. and Shimomura,I. (2013) Expression of activating transcription factor 2 in inflammatory macrophages in obese adipose tissue. *Obesity*, **21**, 731–736.
67. Heise,N., De Silva,N.S., Silva,K., Carette,A., Simonetti,G., Pasparakis,M. and Klein,U. (2014) Germinal center B cell maintenance and differentiation are controlled by distinct NF-κB transcription factor subunits. *J. Exp. Med.*, **211**, 2103–2118.
68. Matsuyama,T., Kimura,T., Kitagawa,M., Pfeffer,K., Kawakami,T., Watanabe,N., Kündig,T.M., Amakawa,R., Kishihara,K. and Wakeham,A. (1993) Targeted disruption of IRF-1 or IRF-2 results in abnormal type I IFN gene induction and aberrant lymphocyte development. *Cell*, **75**, 83–97.
69. Gyory,I., Boller,S., Nechanitzky,R., Mandel,E., Pott,S., Liu,E. and Grosschedl,R. (2012) Transcription factor Ebf1 regulates differentiation stage-specific signaling, proliferation, and survival of B cells. *Genes Dev.*, **26**, 668–682.
70. Vilagos,B., Hoffmann,M., Souabni,A., Sun,Q., Werner,B., Medvedovic,J., Bilic,I., Minnich,M., Axelsson,E., Jaritz,M. *et al.* (2012) Essential role of EBF1 in the generation and function of distinct mature B cell types. *J. Exp. Med.*, **209**, 775–792.
71. Bradshaw,S., Jim Zheng,W., Tsoi,L.C., Gilkeson,G. and Zhang,X.K. (2008) A role for Fli-1 in B cell proliferation: Implications for SLE pathogenesis. *Clin. Immunol.*, **129**, 19–30.
72. Sáez,A.-I., Artiga,M.-J., Sánchez-Beato,M., Sánchez-Verde,L., García,J.-F., Camacho,F.-I., Franco,R. and Piris,M.A. (2002) Analysis of octamer-binding transcription factors Oct2 and Oct1 and their coactivator BOB.1/OBF.1 in lymphomas. *Mod. Pathol.*, **15**, 211–220.
73. Scott,E.W. (1998) The role of PU.1 in the regulation of Lymphoid and Myeloid Hematopoietic Progenitors. In: Monroe,JG and Rothenberg,EV (eds). *Molecular Biology of B-Cell and T-Cell Development*. Humana Press, NJ, pp. 111–126.
74. Jones,P.A. and Baylin,S.B. (2007) The epigenomics of cancer. *Cell*, **128**, 683–692.
75. George,J., Uyar,A., Young,K., Kuffler,L., Waldron-Francis,K., Marquez,E., Ucar,D. and Trowbridge,J.J. (2016) Leukaemia cell of origin identified by chromatin landscape of bulk tumour cells. *Nat. Commun.*, **7**, 12166.
76. Kulis,M., Heath,S., Bibikova,M., Queirós,A.C., Navarro,A., Clot,G., Martínez-Trillos,A., Castellano,G., Brun-Heath,I., Pinyol,M. *et al.* (2012) Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet.*, **44**, 1236–1242.
77. Forbes,S.A., Beare,D., Gunasekaran,P., Leung,K., Bindal,N., Boutselakis,H., Ding,M., Bamford,S., Cole,C., Ward,S. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.
78. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
79. Kawamoto,H., Ikawa,T., Ohmura,K., Fujimoto,S. and Katsura,Y. (2000) T cell progenitors emerge earlier than B cell progenitors in the murine fetal liver. *Immunity*, **12**, 441–450.
80. Adolfsson,J., Månsson,R., Buza-Vidas,N., Hultquist,A., Liuba,K., Jensen,C.T., Bryder,D., Yang,L., Borge,O.-J., Thoren,L.A.M. *et al.* (2005) Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. *Cell*, **121**, 295–306.
81. Igarashi,H., Gregory,S.C., Yokota,T., Sakaguchi,N. and Kincade,P.W. (2002) Transcription from the RAG1 locus marks the earliest lymphocyte progenitors in bone marrow. *Immunity*, **17**, 117–130.
82. Lu,M., Kawamoto,H., Katsube,Y., Ikawa,T. and Katsura,Y. (2002) The common myelolymphoid progenitor: a key intermediate stage in hemopoiesis generating T and B cells. *J. Immunol.*, **169**, 3519–3525.
83. Wada,H., Masuda,K., Satoh,R., Kakugawa,K., Ikawa,T., Katsura,Y. and Kawamoto,H. (2008) Adult T-cell progenitors retain myeloid potential. *Nature*, **452**, 768–772.
84. Attema,J.L., Papathanasiou,P., Forsberg,E.C., Xu,J., Smale,S.T. and Weissman,I.L. (2007) Epigenetic characterization of hematopoietic stem cell differentiation using miniChIP and bisulfite sequencing analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 12371–12376.
85. Choukrallah,M.-A., Song,S., Rolink,A.G., Burger,L. and Matthias,P. (2015) Enhancer repertoires are reshaped independently of early priming and heterochromatin dynamics during B cell differentiation. *Nat. Commun.*, **6**, 8324.
86. Maës,J., Maleszewska,M., Guillemin,C., Pflumio,F., Six,E., André-Schmutz,I., Cavazzana-Calvo,M., Charron,D., Francastel,C. and Goodhardt,M. (2008) Lymphoid-affiliated genes are associated with active histone modifications in human hematopoietic stem cells. *Blood*, **112**, 2722–2729.
87. Orford,K., Kharchenko,P., Lai,W., Dao,M.C., Worhunsky,D.J., Ferro,A., Janzen,V., Park,P.J. and Scadden,D.T. (2008) Differential H3K4 methylation identifies developmentally poised hematopoietic genes. *Dev. Cell*, **14**, 798–809.
88. Barnkob,M.S., Simon,C. and Olsen,L.R. (2014) Characterizing the human hematopoietic CDome. *Front. Genet.*, **5**, 331.
89. Parekh,C. and Crooks,G.M. (2013) Critical differences in hematopoiesis and lymphoid development between humans and mice. *J. Clin. Immunol.*, **33**, 711–715.
90. Vardiman,J.W., Thiele,J., Arber,D.A., Brunning,R.D., Borowitz,M.J., Porwit,A., Harris,N.L., Le Beau,M.M., Hellström-Lindberg,E., Tefferi,A. *et al.* (2009) The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood*, **114**, 937–951.
91. Papaemmanuil,E., Döhner,H. and Campbell,P.J. (2016) Genomic classification in acute myeloid leukemia. *N. Engl. J. Med.*, **375**, 900–901.