bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Searching for the chromatin determinants of human hematopoiesis

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

[1]Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO)
[2]Institut d'Investigacions Biomédiques August Pi i Sunyer (IDIBAPS), Department of Anatomic Pathology, Pharmacology and Microbiology, University of Barcelona, Barcelona, Spain.
[3]Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom.

[4]Co-first authors
[5]Corresponding authors daniel.rico@newcastle.ac.uk, valencia@cnio.es
[6]Member of International Human Epigenome Consortium (IHEC). For more information, see http://www.blueprint-epigenome.eu and http://ihec-epigenomes.org

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 forty-two human healthy blood epigenomes from the BLUEPRINT, NIH ROADMAP and ENCODE consortia by their cell type. This approach recapitulates the human hematopoietic differentiation tree model from an epigenomic perspective. The analysis of the orthogonal dimension of the chromatin space allows us to 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 origin of these tumors. Our method provides an analytical approach to study the relationship between epigenomic changes and cell lineage differentiation.

Author summary

The changes in the chromatin configuration, or epigenome, have an essential role in development and differentiation of cells, but how these epigenomic changes are connected with the different cell phenotypes is not well understood. Hematopoiesis is a dynamic process in which hematopoietic stem cells give rise many different cell types, including lymphocytes, neutrophils, monocytes, platelets or erythrocytes. Here we propose a computational approach to derive an epigenomic space from epigenomes representing multiple different cell types and

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

describe its usefulness in the context of hematopoiesis. Concomitantly, our method recovers those genomic regions that are more informative about the segregation in different cell types. These regions are called Chromatin Determinant Regions (CDRs) as they present differential epigenetic states among different cell types. Further characterization of the detected hematopoietic CDRs shows that these genomic regions are associated to hematopoietic transcription factors and the expected functions that characterize each cell type. CDRs provide a valuable reference for the study of normal hematopoiesis as they allow to study epigenetic differences in the context of the whole system. Similarly, they help to explore the epigenomic differences of hematologic malignancies to the healthy cell types, providing additional clues about their cell of origin and about those epigenomic alterations interfering in normal hematopoiesis. Our approach will enable the integration of data from several resources for characterizing and comparing different differentiation systems when more data will be available.

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 (Rieger and Schroeder, 2012; Paul et al., 2015). This hierarchical tree with many ‘nodes’ and ‘leaves’ provides the best model to study chromatin remodeling during cell lineage differentiation (Ji et al., 2010; Bock et al., 2012; Lara-Astiaso et al., 2014).

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 (Ho and Crabtree, 2010). The de-regulation of chromatin remodeling factors often leads to diseases such as cancers (Kumar et al., 2016) and neurodevelopmental disorders (Ronan et al., 2013; Mirabella et al., 2016). 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 (Bannister and Kouzarides, 2011; Venkatesh and Workman, 2015). Moreover, the cross-talk between different modifications affects the binding and function of other epigenetic elements, increasing the complexity of the chromatin remodeling process (Juan et al., 2016).

Despite great progress in our understanding of hematopoiesis during the last decades (Doulatov et al., 2012; Rice et al., 2007), we are still far from fully uncovering the details of the epigenetic 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

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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; IHEC, 2016).

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. 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 hematopoietic stem cells (HSCs). This framework establishes highly informative low-dimensional spaces based on a Multiple Correspondence Analysis (MCA; Torres-Lacomba, 2006) 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.

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: haematological stem cells (HSC; n=3), neutrophils (n=5), monocytes (n=6), macrophages (n=4), naive and germinal center (GC) B-cells (n=4 and n=3) and CD4 and CD8 T-cells (n=13 and n=4), see Fig 2A and S1 Table 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, marking activation (Creyghton et al., 2010; Barski et al., 2007); H3K4me1, marking enhancers (Barski et al., 2007); H3K36me3, marking transcription (Barski et al., 2007); H3K27me3 and H3K9me3, marking polycomb and heterochromatin repression, respectively (Barski et al., 2007). 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).

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A multivariate Hidden Markov Model (HMM) was employed to learn combinatorial chromatin states based on the six histone marks using ChromHmm (Ernst and Kellis, 2012). Further, the genome of each sample was segmented using the 11 combinatorial chromatin states model generated. To facilitate biological interpretation, the 11 chromatin states were further collapsed into 5 functional chromatin states encompassing five main categories: transcription (T), heterochromatin (H), repressed promoter (R), enhancer (E) and active promoter (A; see Methods for details; S1 Fig). Thus, for each sample, we can create a vector representing the chromatin state of consecutive 200 bp windows along the whole genome, using this reduced 5-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 Methods for details).

In order to generate a high-dimensional chromatin space reflecting the major sources of epigenetic differences among hematopoietic samples (eg. changes in chromatin states), we applied a protocol based on Multiple Correspondence Analysis (MCA), which we have previously applied to protein sequence (Rausell et al., 2010) and gene expression (Martinez-Garcia et al., 2014) analysis. MCA is an analysis similar to Principal Component Analysis (PCA) but appropriate for categorical data. We created a 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 **Fig 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 (Fig 2B; see Methods 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 (3 different consortia and therefore different laboratories). The relative samples distribution and the clustering are robust, as shown by analysing each of the autosomal chromosomes independently (see **S2 Fig**).

As in PCA approaches, the interpretation of the two components selected by our approach 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 differentiation level of the cell types, as well as the sample environment, drawing 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* cultivated macrophages and germinal center B cells from tonsil (at the top of the plot), with less differentiated neutrophils, monocytes, T cells and naïve B cells from venous blood at the middle (see **Fig 2**). However, as previously observed based on single chromatin marks (Kundaje et al., 2015), 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

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

consistent with the known underlying biology of the system, showing that epigenomic states are an excellent source of information for discriminating these cell types.

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 *Chromatin Determinant Regions* (CDRs; **Fig 1**).

In order to retrieve these CDRs we applied the second stage of our MCA-based protocol (Rausell et al., 2010). This involves building a *hematological chromatin regions space*, in which each genomic region can be located based on its patterns of chromatin states across cell types (see **Fig 1**). For this we projected the chromatin states of every region of the genome in 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 and thus interpret differences between the different cell types. For instance, a given region can show an enhancer state in lymphoid cell samples, and a heterochromatin 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 chromosomes size) (see S2 Table).

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 state in the others (see **S3 Fig** and **S3 Table**). 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 states in all other cell types (see **S3 Fig** and **S3 Table**). In general, CDRs related with Transcription states are larger than patterns with other states (see S4 Fig).

CDR chromatin state transitions across hematopoiesis

A more detailed analysis of the CDR transitions between cell types following the differentiation tree can provide insights about chromatin remodeling across lineages. From the first pluripotent stage (HSCs), 4 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). **Fig 2** shows transitions in CDR states across the various branches of the tree. We observe the transitions from the HSCs to the second stage to be characterized by a turning off

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of active and enhancer CDRs. In contrast, 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 the regulation the expression of nearby genes important for these cell type transitions. Thus, we associated each cell type specific CDR to its most proximal gene and carried out functional enrichment analysis for the CDRs defining each cell type by specifically active promoter, enhancer or transcription states (see **Fig 3B; S6-12 Figs; S4 Table**; see Methods for details). Indeed, genes proximal to the CDRs defining HSCs were mainly enriched in processes related with development and cell differentiation.

CDRs defining the myeloid lineage were close to genes related with tissue development, opsonization 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 (Luzina et al., 2012). CDRs defining the two different B cell types were associated to genes with functions in proliferation and differentiation.

We next asked whether CDRs involving cell-specific active promoter or enhancer states showed enrichment of transcription factor motifs (TFMs, see Methods 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 (**Fig 3C**).

We observed in HSCs specific motif enrichment of GATA factors, which have been related with regulation of the self-renewal of long-term hematopoietic stem cells and differentiation of bone marrow-derived mesenchymal stem cells (Kamata et al., 2014; Frelin et al., 2013; Ku et al., 2012; Bresnick et al., 2012). Enrichment in binding motives for factors like RUNX, implied in stem cell fate maintenance and normal function, was also observed in HSCs specific CDRs (Staber et al., 2014; Burns, 2005). Interestingly, motifs for the so far uncharacterized factor X gene family, known to regulate the major histocompatibility complex (MHC) class II (Reith and Mach, 2001), 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 (Bruhat et al., 2007; Gombart et al., 2007), which plays a key role during the differentiation of the monocyte lineage (Radomska et al., 1998; Yeamans et al., 2007). In contrast, EGR1 and EGR2 binding motifs, which are essential for macrophage but not for granulocyte differentiation (Laslo et al., 2006; Nguyen et al., 1993), are enriched in macrophages. Higher expression at RNA level is observed in macrophages compared with monocytes and mature neutrophils (see S13 Fig). In addition, enrichment for transcription factor binding sites related with macrophage differentiation like STAT3, JUND, MITF, NUR77 or ATF2 are observed in CDRs specific to macrophages

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Behmoaras et al., 2008; Coffer et al., 2000; Hanna et al., 2012; Hume, 2015; Miyata et al., 2013).

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 germinal centers affect their maintenance and B cell differentiation (Heise et al., 2014; Matsuyama et al., 1993). In addition, we observed enriched motifs for Early B-cell factor 1 (EBF1), a central transcription factor in B cells implicated in germinal center formation and class switch recombination (Gyory et al., 2012; Vilagos et al., 2012), Oct2 and Fli1, transcription factors expressed in B cells and related with normal B cells proliferation (Bradshaw et al., 2008; Sáez et al., 2002).

Finally, 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 (Scott, 1998).

Taken together, the functional 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 (Jones and Baylin, 2007) and epigenetic features such as DNA methylation or open chromatin have been shown to be useful to identify the cell of origin of tumours (George et al., 2016; Kulis et al., 2012). However, given the extensive genome-wide epigenetic alterations of tumour cells, matching tumoral cells with their healthy counterparts is a great challenge and an essential step to recapitulate 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 6 chronic lymphocytic leukemias (CLLs), 3 acute myeloid leukemias (AMLs) and 3 mantle cell lymphomas (MCLs) to explore the epigenetic similarity among healthy and cancerous samples.

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We projected the leukemic samples on the healthy hematopoietic chromatin space, based on their chromatin states at CDRs (see S14 Fig and Methods). Next, we used the distance of each leukemic sample to a reference healthy cell type (S14 Fig) 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 separate 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 S14 Fig). 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 Fig 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 tumours.

Each tumoral sample was projected into the healthy hematopoietic chromatin sample space using the CDRs whose chromatin states are represented in any of the healthy cell types (see Methods). 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 (Fig 4). AML samples appear to be epigenetically more divergent from the healthy states than those closer to the B cell derived cancer samples. Taken together, these results show the potential of our proposed CDR approach to characterise haematopoietic cell types in normal differentiation and disease .

Discussion

Chromatin remodeling is an essential process for determining the set of phenotypic conditions 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 (Juan et al., 2016). 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 (Ernst and Kellis, 2010; Kundaje et al., 2015), 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 (Rieger and Schroeder, 2012). 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 sources and

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

developed a new protocol based on a multivariate framework to define which cell types are epigenetically distinguishable and simultaneously detect the key genomic regions driving these differences.

Lineage characterization has been the focus of many years of research (Rieger and Schroeder, 2012). The blood IHEC epigenomes (IHEC, 2016) provide a unique opportunity to investigate the epigenetic basis of lineage determination. Our results do not only establish an unsupervised epigenetic classification of the different cell types but additionally identify genomic regions and their associated to functional characteristics that might be on the basis of cell identity. These regions, named Chromatin Determinant Regions (CDRs), can be considered as the epigenetic signatures of human hematopoiesis, a set of reference regions that under epigenetic changes might be able to drive hematopoiesis.

The results are robust to the possible noise introduced by consortia specific protocols and the clusters were obtained without misclassified 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 H3K4m1 and H3K27me3 histone modifications independently was also unable to discriminate CD4+ from CD8+ T cell types (Kundaje et al., 2015), 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 to the HSC. Although both the classical and the more recent alternative hematopoiesis hierarchical differentiation models propose a similar differentiation distance for neutrophils and monocytes or T and B cells (Rieger and Schroeder, 2012; Paul et al., 2015), 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 for the B cells (Kawamoto et al., 2000).

The classical hematopoietic model establishes that the HSCs differentiates into the common myeloid progenitor (CMP) or the common lymphoid progenitor (CLP), divided in the myeloid and the lymphoid lineages (Rieger and Schroeder, 2012). However, this model is under discussion as it has been shown by Kawamoto et al. (Kawamoto et al., 2010) and other authors (Adolfsson et al., 2005; Igarashi et al., 2002; Lu et al., 2002; Wada et al., 2008) 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 (Kawamoto et al., 2010) 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. The future availability of complete epigenomes for more cell lineages (megakaryocytes, erythrocytes, NK cells, eosinophils, etc) and for the different intermediate progenitors will provide additional information to assess whether the myeloid-based differentiation model proposed by Kawamoto et al. (Kawamoto et al., 2010) is consistent with the chromatin landscape.

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Attema et al., 2007; Choukrallah et al., 2015; Maes et al., 2008; Orford et al., 2008).

A further characterization of these CDRs showed that they are enriched in DNA binding motifs associated to transcription factors with a key role in hematopoiesis. These results support the idea of CDRs as driver regions whose chromatin reconfiguration will be associated with 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 region.

As only a subset of blood cell types was used in this analysis, these CDRs have to be seen as the 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 (Barnkob et al., 2014) or mice models with DNA methylation (Bock et al., 2012) and transcriptomics (Paul et al., 2015). 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 (Rieger and Schroeder, 2012) or hematopoietic gene regulation programs (Parekh and Crooks, 2013) are not fully conserved between species.

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 (Vardiman et al., 2009; Papaemmanuil et al., 2016), which would explain why two of the AML samples cluster close to neutrophils, and the other one with monocytes. In summary, our proof of concept application of the epigenetic signature of hematopoiesis in the study of leukemia shows the power of the methodology that might turn out to be very useful when more leukemia and complete progenitor epigenomes will become available.

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 the human hematopoiesis, based on the CDRs that better describe the different cell types. This catalog, with further refinements by the inclusion of additional cell types and hematopoietic progenitors, might became the reference IHEC resource for human hematopoiesis studies.

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+ haematological stem cells samples from NIH ROADMAP and 2 CD14+ monocytes samples from ENCODE described in S1 Table and Fig 2A. In addition, the analysis including diseases was based on data from 3 Acute Myeloid Leukemias (AML), 6 Chronic Lymphocytic Leukemias (CLL) and 3 Mantle Cell Lymphomas (MCL) from BLUEPRINT (see S1 Table). The BAM files were converted to BED format and duplicate reads were removed for all the experiments.

Genome Segmentation

The input information used to segment the genome into different chromatin states was derived from 6 histone modifications (H3K4me3; H3K4me1; H3K27me3; H3K9me3; H3K27ac and H3K36me3). We used the ChromHmm software (v1.10; Ernst and Kellis, 2012) to define a 11 chromatin-states model consistent with prior knowledge regarding the function of these marks (see S1 Fig). We generated the model with the “healthy” samples excluding B cells (see S1 Table for details). The samples from B cells (naive and tonsil) and diseases (AML, CLL, MCL) were segmented with the model generated previously. Further, segmentations for each sample from the 11 states model were collapsed into 5 chromatin states: heterochromatin (H), enhancer (E), transcription (T), active promoter (A) and repressed promoter (R; S1 Fig). Therefore, for each sample, we have a vector of regions with their corresponding labels (chromatin states). In addition, we partitioned the genome into 200bp, 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 200bp 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 Multiple Correspondence Analysis (MCA; Torres-Lacomba, 2006), previously applied to multiple sequence alignments of proteins, for the automatic extraction of relevant signatures (Rausell et al., 2010) and to gene

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

expression profiles for sample classification (Martinez-Garcia et al., 2014). 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 a 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 analysed samples. Robust unsupervised k-means clustering (de Hoon et al., 2004) is performed iteratively on this Chromatin Sample Space for a range of pre-specified numbers of groups (from 2 to 50). Finally, optimal clustering solutions are detected as those maximizing the Calinsky's and Harabsz's (CH) index (Calinski and Harabasz, 1974). In an analysis involving samples from different healthy cell types, as the 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 important for establishing this samples segregation.

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. 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 (e.g. samples with the same or different states in the region) were poorly represented in our dataset. In particular, we filtered out those regions whose patterns were not shared by at least other 9 regions (we obtain similar results for patterns shared by 5, 10 and 15 regions; data not shown). The rationale of this step was to remove low informative regions that are both introducing noise and increasing the computational cost of the analysis. 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 Chromatin Determinant Regions and they are methodologically equivalent to the Specificity Determining Positions detected (SDPs; Rausell et al., 2010) in protein families.

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

First, we project the vectors reflecting every genomic regions/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 Rausell et al., 2010). 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 points to the putative role of these regions in cell identity and cell fate respectively.

Chromatin Determinant Regions annotation and enrichment analyses

Genomic annotation was carried out with Hypergeometric Optimization of Motif EnRichment (HOMER software v4.7.2; Heinz et al., 2010). 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 (Biological Process; Ashburner et al., 2000) and Reactome (Fabregat et al., 2016) enrichment analysis were done adding the -go flag to the annotatePeaks.pl tool, all terms with a logP < -4.32 (p-value < 0.05) were considered significant (see S4 Table). We summarized the Gene Ontology (Biological Process) significant terms with REVIGO (Supek et al., 2011) to generate Figs S6-S12 and Fig 3B.

The transcription factor motif enrichments were performed with the findMotifsGenome.pl tool included in HOMER software (v4.7.2; Heinz et al., 2010). To determine the relative enrichment of known TFMs we excluded the CDRs referred to transcription, as they are related with 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 transcription factor motifs with a q-value<0.01 at least in one cell type were considered significant and selected to generate the Fig 3C.

Chromatin state transitions among cell types were represented with a Sankey diagram in Fig 3A using the “makeRiver” and “riverplot” functions included in the “riverplot” R package. (v0.5; RIVERPLOT, 2016)

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Chromatin Determinant Regions in the context of diseases

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 more clear 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 weight 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 (see S14 Fig).

Despite the effect of focusing on “conserved” states in CDRs, highly transformed leukemic samples could include a relevant number 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, PHEATMAP, 2016) 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 all leukemic samples, they were not considered in the hierarchical clustering.

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 much divergent is a tumoral sample 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. The analyzed tumoral samples show epigenomic tumoral divergences ranging from 0.02 to 0.08 with higher values for AML samples (see Fig 4).

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Funding

The research leading to these results was funded from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 282510 (BLUEPRINT). This work was carried out in the framework of the Platform for Biomolecular and Bioinformatics Resource PT 13/0001/0030 of the ISCIII and MINECO Project BFU2015-71241-R which are funded through the European Regional Development Fund (ERDF).

References

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. (2005b). Identification of Flt3+ Lympho-Myeloid Stem Cells Lacking Erythro-Megakaryocytic Potential. Cell *121*, 295–306.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics *25*, 25–29.

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. Proceedings of the National Academy of Sciences *104*, 12371–12376.

Bannister, A.J., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Research *21*, 381–395.

Barnkob, M.S., Simon, C., and Olsen, L.R. (2014). Characterizing the human hematopoietic CDome. Frontiers in Genetics *5*.

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.

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. Nature Genetics *40*, 553–559.

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. Molecular Cell *47*, 633–647.

Bradshaw, S., Zheng, W.J., Tsoi, L.C., Gilkeson, G., and Zhang, X.K. (2008). A role for Fli-1 in B cell proliferation: Implications for SLE pathogenesis. Clinical Immunology *129*, 19–30.

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 Research *40*, 5819–5831.

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bruhat, A., Cherasse, 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 Research *35*, 1312–1321.

Burns, C.E. (2005). Hematopoietic stem cell fate is established by the Notch-Runx pathway. Genes & Development *19*, 2331–2342.

Calinski, T., and Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics - Theory and Methods *3*, 1–27.

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. Nature Communications *6*, 8324.

Coffer, P.J., Koenderman, L., and de Groot, R.P. (2000). The role of STATs in myeloid differentiation and leukemia. Oncogene *19*, 2511–2522.

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. Proceedings of the National Academy of Sciences *107*, 21931–21936.

Doulatov, S., Notta, F., Laurenti, E., and Dick, J.E. (2012). Hematopoiesis: A Human Perspective. Cell Stem Cell *10*, 120–136.

Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature Biotechnology *28*, 817–825.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nature Methods *9*, 215–216.

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 Research *44*, D481–D487.

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. Nature Immunology *14*, 1037–1044.

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. Nature Communications *7*, 12166.

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. Journal of Leukocyte Biology *81*, 1535–1547.

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 & Development *26*, 668–682.

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. Circulation Research *110*, 416–427.

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. Molecular Cell *38*, 576–589.

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. The Journal of Experimental Medicine *211*, 2103–2118.

Ho, L., and Crabtree, G.R. (2010). Chromatin remodelling during development. Nature *463*, 474–484.

de Hoon, M.J.L., Imoto, S., Nolan, J., and Miyano, S. (2004a). Open source clustering software. Bioinformatics *20*, 1453–1454.

Hume, D.A. (2015). The Many Alternative Faces of Macrophage Activation. Frontiers in Immunology *6*.

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.

IHEC (2016). International Human Epigenome Consortium web page. http://ihec-epigenomes.org/

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.

Jones, P.A., and Baylin, S.B. (2007). The epigenomics of cancer. Cell *128*, 683–692.

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 Reports *14*, 1246–1257.

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.

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.

Kawamoto, H., Ikawa, T., Masuda, K., Wada, H., and Katsura, Y. (2010). A map for lineage restriction of progenitors during hematopoiesis: the essence of the myeloid-based model: Lineage-restriction pathways during hematopoiesis. Immunological Reviews *238*, 23–36.

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.

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. Nature Genetics *44*, 1236–1242.

Kumar, R., Li, D.-Q., Müller, S., and Knapp, S. (2016). Epigenomic regulation of oncogenesis by chromatin remodeling. Oncogene.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

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). Chromatin state dynamics during blood formation. Science *345*, 943–949.

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.

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. The Journal of Immunology *169*, 3519–3525.

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.” Journal of Leukocyte Biology *92*, 753–764.

Maes, J., Maleszewska, M., Guillemin, C., Pflumio, F., Six, E., Andre-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.

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 Medicine *6*, 27.

Matsuyama, T., Kimura, T., Kitagawa, M., Pfeffer, K., Kawakami, T., Watanabe, N., Kündig, T.M., Amakawa, R., Kishihara, K., Wakeham, A., et al. (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.

Mirabella, A.C., Foster, B.M., and Bartke, T. (2016). Chromatin deregulation in disease. Chromosoma *125*, 75–93.

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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. Developmental Cell *14*, 798–809.

Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V.I., Paschka, P., Roberts, N.D., Potter, N.E., Heuser, M., Thol, F., Bolli, N., et al. (2016). Genomic Classification and Prognosis in Acute Myeloid Leukemia. New England Journal of Medicine *374*, 2209–2221.

Parekh, C., and Crooks, G.M. (2013). Critical Differences in Hematopoiesis and Lymphoid Development between Humans and Mice. Journal of Clinical Immunology *33*, 711–715.

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.

PHEATMAP (2016). pheatmap: Pretty Heatmaps R-package. http://CRAN.R-project.org/package=pheatmap

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. Molecular and Cellular Biology *18*, 4301–4314.

Rausell, A., Juan, D., Pazos, F., and Valencia, A. (2010). Protein interactions and ligand binding: From protein subfamilies to functional specificity. Proceedings of the National Academy of Sciences *107*, 1995–2000.

Reith, W., and Mach, B. (2001). The bare lymphocyte syndrome and the regulation of MHC expression. Annu. Rev. Immunol. *19*, 331–373.

Rice, K.L., Hormaeche, I., and Licht, J.D. (2007). Epigenetic regulation of normal and malignant hematopoiesis. Oncogene *26*, 6697–6714.

Rieger, M.A., and Schroeder, T. (2012). Hematopoiesis. Cold Spring Harbor Perspectives in Biology *4*, a008250–a008250.

RIVERPLOT (2016). riverplot: Sankey or Ribbon Plots R-package. https://cran.r-project.org/web/packages/riverplot/index.html

Ronan, J.L., Wu, W., and Crabtree, G.R. (2013). From neural development to cognition: unexpected roles for chromatin. Nature Reviews Genetics *14*, 347–359.

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.

Scott, E.W. (1998). The Role of PU.1 in the Regulation of Lymphoid and Myeloid Hematopoietic Progenitors. In Molecular Biology of B-Cell and T-Cell Development, J.G. Monroe, and E.V. Rothenberg, eds. (Totowa, NJ: Humana Press), pp. 111–126.

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.

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLoS ONE *6*, e21800.

Torres-Lacomba, A. (2006). Correspondence Analysis and Categorical Conjoint Measurement. In Multiple Correspondence Analysis and Related Methods, M. Greenacre, and J. Blasius, eds. (Chapman and Hall/CRC), pp. 421–432.

Vardiman, J.W., Thiele, J., Arber, D.A., Brunning, R.D., Borowitz, M.J., Porwit, A., Harris, N.L., Le Beau, M.M., Hellstrom-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.

Venkatesh, S., and Workman, J.L. (2015). Histone exchange, chromatin structure and the regulation of transcription. Nature Reviews Molecular Cell Biology *16*, 178–189.

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. The Journal of Experimental Medicine *209*, 775–792.

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.

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.

Figure Legends

Fig 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.

See also S1 Fig

Fig 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 S2 Fig and S1 Table

bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig 3. Functional and Transcription Factor binding motifs characterization of Chromatin Determinant Regions

A) Sankey Plot representation of chromatin state transitions at CDRs during haematopoietic cell differentiation. Nodes for each cell type represent the five “collapsed” chromatin states (see Methods). For each pair of cell types in the haematopoietic 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 colour. The thickness of the lines is proportional to the number of regions that show a transition between a particular pair of states.

B) REVIGO summarized table with the 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 Methods).

See also S3-S13 Figs and S2-S4 Tables

Fig 4. Hierarchical Clustering of leukemias based on CDRs of healthy cell types suggests potential cell origin of tumors.

The healthy cell type clusters are summarized by each fingerprint, a reference sample representing each cell type cluster. The euclidian 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 S14 Fig

S1 Supporting Information

Supporting Figs

S1 Fig. Related to Fig 1. Chromatin States definitions, emission probabilities and genomic annotation enrichments.

S2 Fig. Related to Fig 2. MCA by chromosome

S3 Fig. Related to Fig 3. The most abundant Chromatin Determinant Region patterns

S4 Fig. Related to Fig 3. Boxplot with the sizes of the Chromatin Determinant Region patterns

S5 Fig. Related to Fig 3. Histogram with Chromatin Determinant Region distances to the Transcription Start Site

S6 Fig. Related to Fig 3. REVIGO treemap summarizing gene ontology biological process categories over-represented in active HSCs Chromatin Determinant Regions

S7 Fig. Related to Fig 3. REVIGO treemap summarizing gene ontology biological process categories over-represented in active T cell Chromatin Determinant Regions

S8 Fig. Related to Fig 3. REVIGO treemap summarizing gene ontology biological process categories over-represented in active Naïve B cells Chromatin Determinant Regions

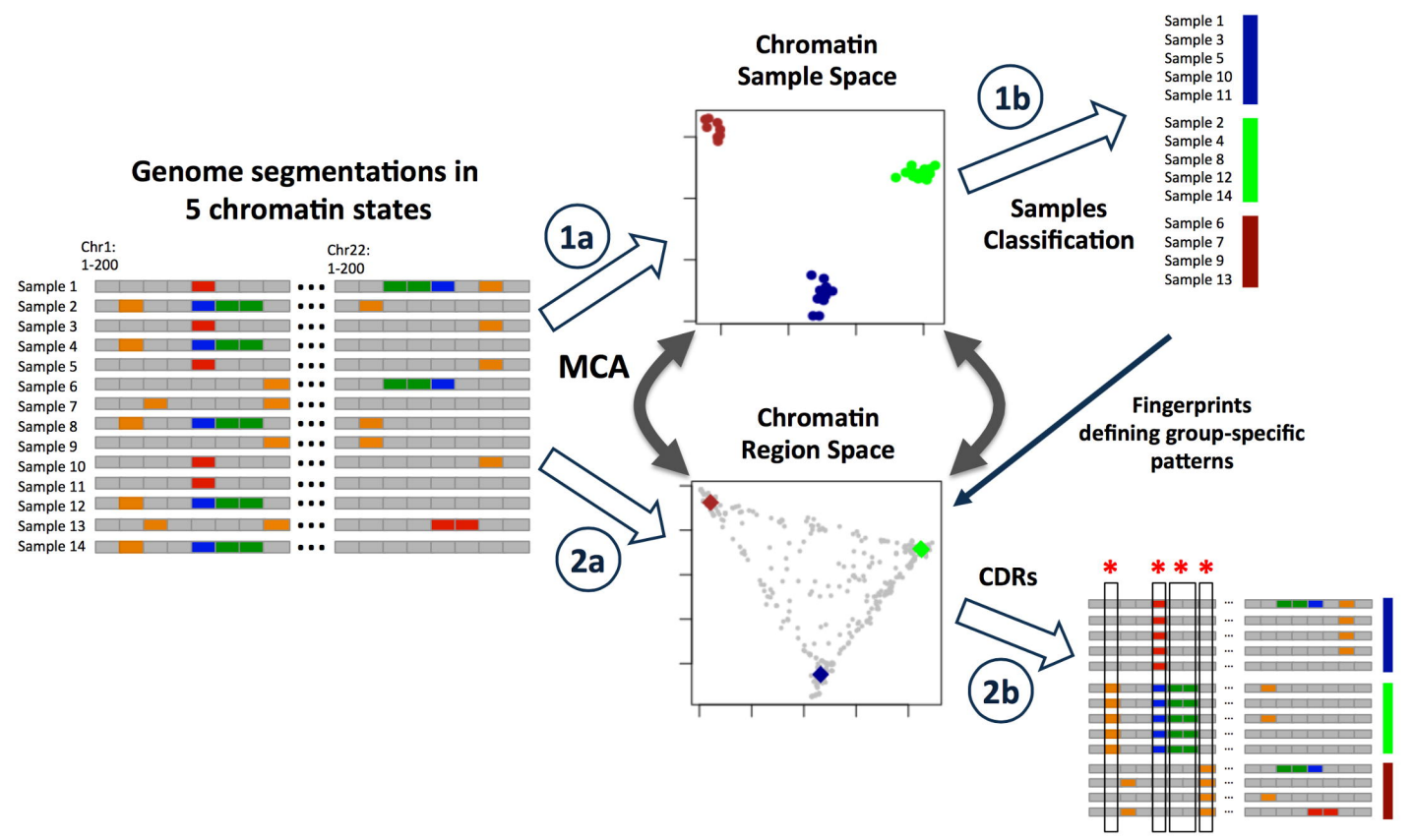
S9 Fig. Related to Fig 3. REVIGO treemap summarizing gene ontology biological process categories over-represented in active Germinal Centre B cells Chromatin Determinant Regions

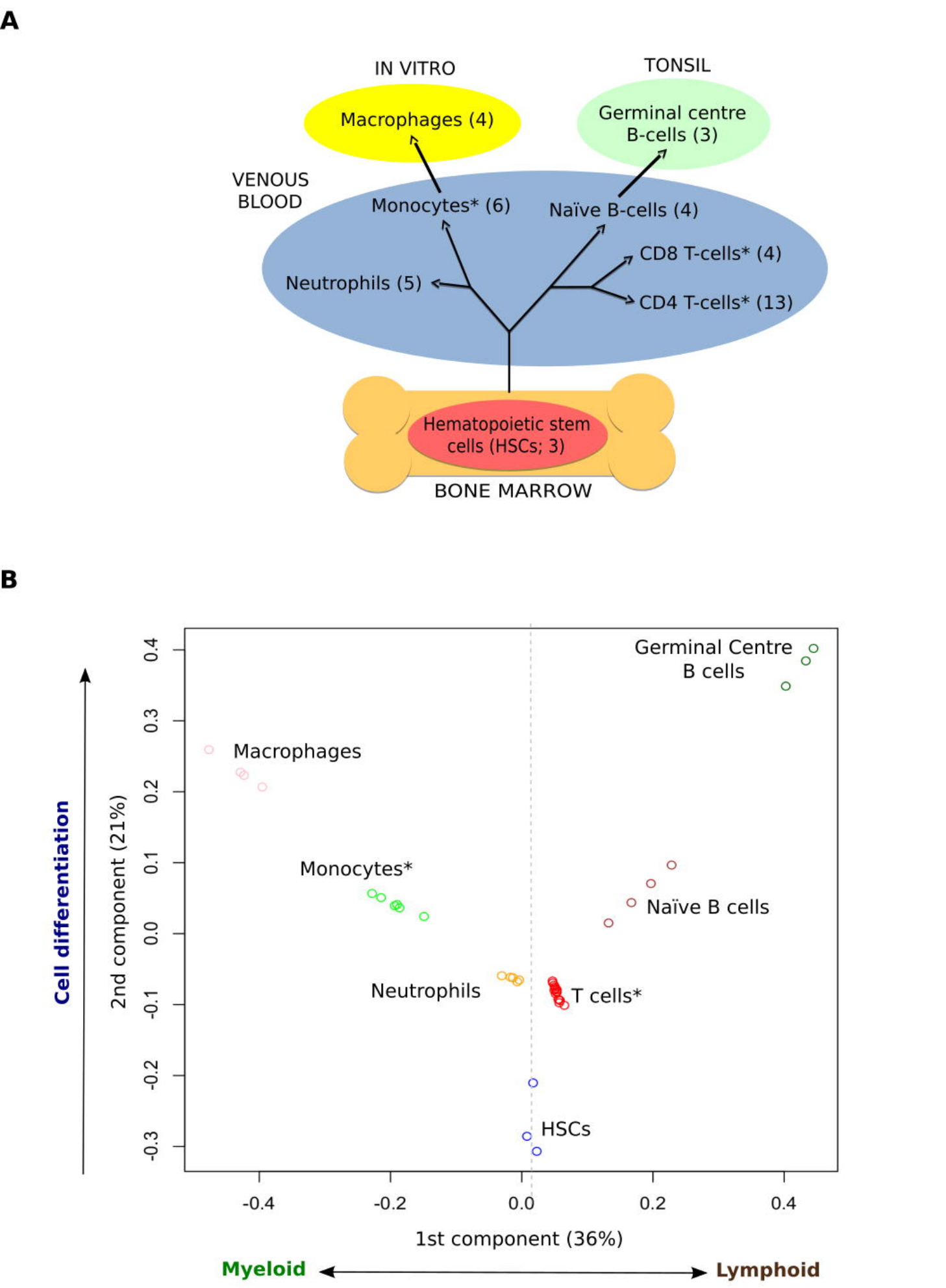
S10 Fig. Related to Fig 3. REVIGO treemap summarizing gene ontology biological process categories over-represented in active Neutrophil Chromatin Determinant Regions

S11 Fig. Related to Fig 3. REVIGO treemap summarizing gene ontology biological process categories over-represented in active Monocyte Chromatin Determinant Regions

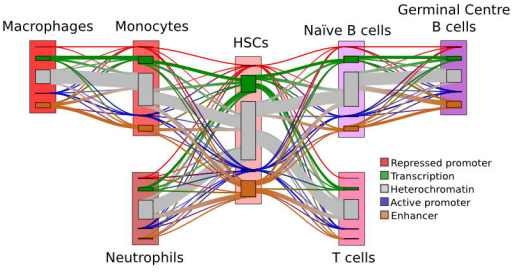
bioRxiv preprint doi: https://doi.org/10.1101/082917; this version posted October 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S12 Fig. Related to Fig 3. REVIGO treemap summarizing gene ontology biological process categories over-represented in active Macrophage Chromatin Determinant Regions
S13 Fig. Related to Fig 3. ERG2 expression in macrophages (red), monocytes (blue) and neutrophils (green) from BLUEPRINT data analysis portal (http://blueprint-data.bsc.es)
S14 Fig. Related to Fig 4. Epigenomes of leukemias projected into the “Healthy” blood cells chromatin space.
Supporting Tables
S1 Table. Related to Fig 2. Healthy and disease samples included in the analyses. (In S1_Table.xls)
S2 Table. Related to Fig 3. Chromatin Determinant Regions (In S2_Table.bed)
S3 Table. Related to Fig 3. Frequencies for Chromatin Determinant Region patterns (In S3_Table.xls)
S4 Table. Related to Fig 3. Gene Ontology (Biological Process) & REACTOME enrichment results from HOMER (In S4_Table.xls)


Genome segmentations in
5 chromatin states
Chr1:
1-200
Chr22:
1-200
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
Sample 9
Sample 10
Sample 11
Sample 12
Sample 13
Sample 14
1a
2a
MCA
Chromatin
Sample Space
Chromatin
Region Space
1b
Samples
Classification
Sample 1
Sample 3
Sample 5
Sample 10
Sample 11
Sample 2
Sample 4
Sample 8
Sample 12
Sample 14
Sample 6
Sample 7
Sample 9
Sample 13
Fingerprints
defining group-specific
patterns
CDRs
2b


A
IN VITRO
TONSIL
Macrophages (4)
Germinal centre
B-cells (3)
VENOUS
BLOOD
Monocytes* (6)
Naïve B-cells (4)
CD8 T-cells* (4)
Neutrophils (5)
CD4 T-cells* (13)
Hematopoietic stem
cells (HSCs; 3)
BONE MARROW
B
Germinal Centre
B cells
Macrophages
Monocytes*
Naïve B cells
Neutrophils
T cells*
HSCs
Cell differentiation
2nd component (21%)
1st component (36%)
Myeloid
Lymphoid
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.2
0.0
0.2
0.4

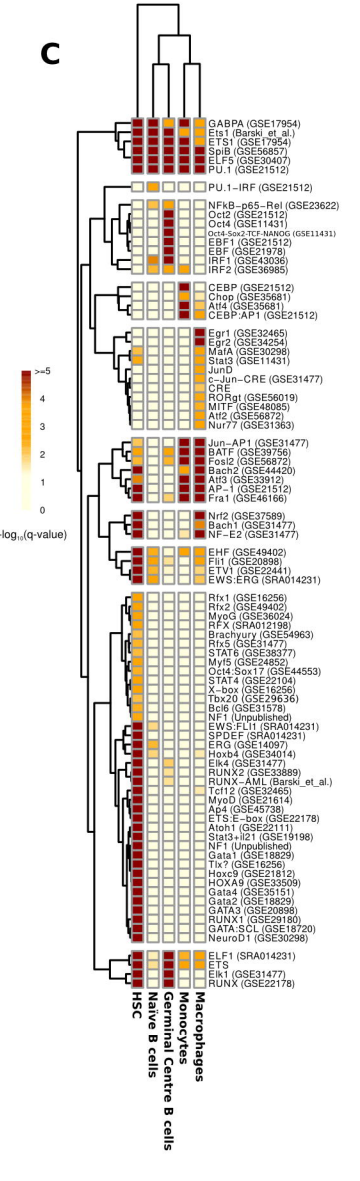
A

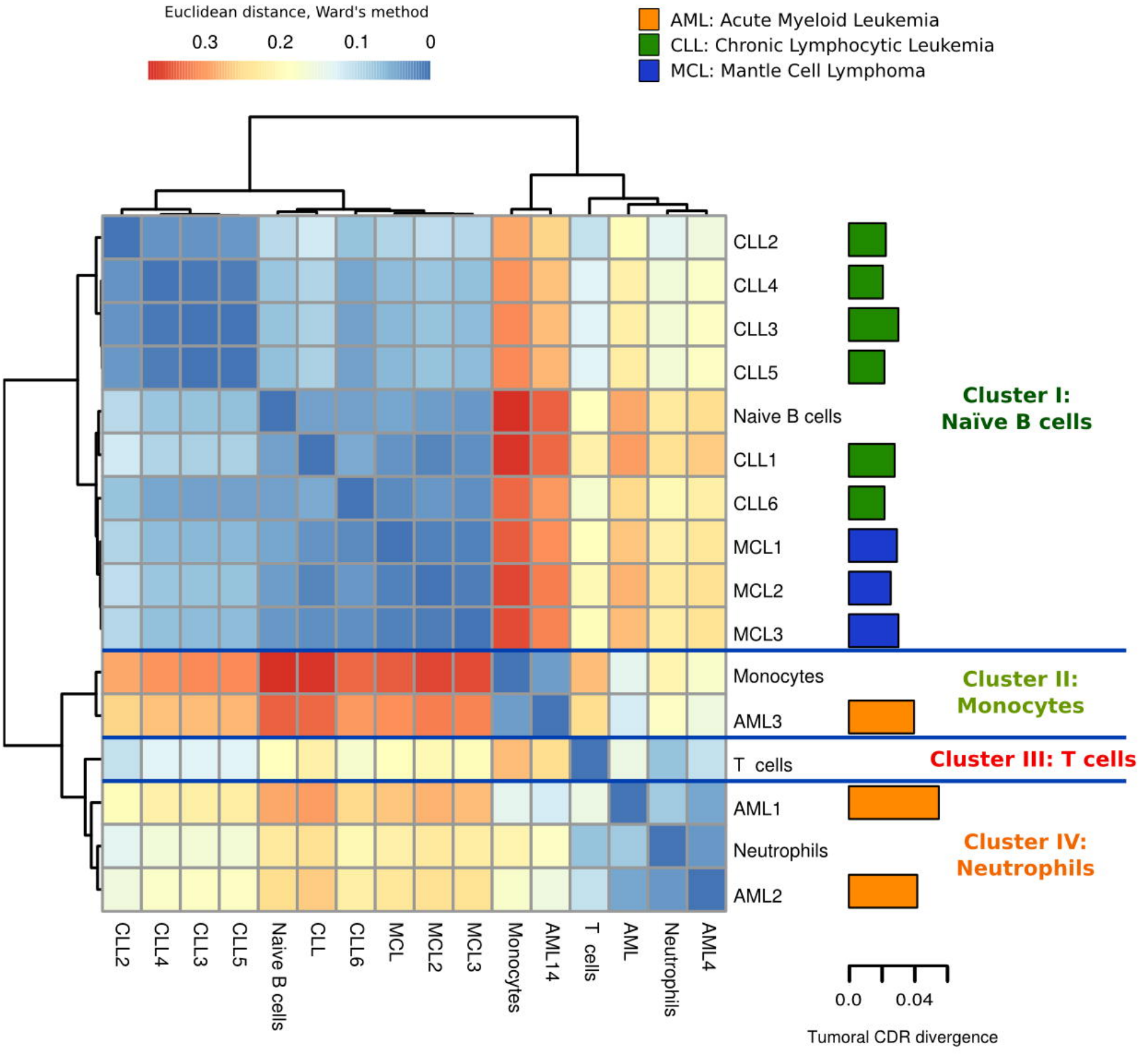


B

Cell Type	Summarized GO terms	N° GO terms	N° genes
Germinal centre B cells	neuron differentiation	93	935
	regulation of phosphate metabolism	25	1203
	positive regulation of sodium ion transport	9	260
Naïve B cells	regulation of lymphocyte proliferation	21	94
	columnar/cuboidal epithelial cell differentiation	16	87
	cellular modified amino acid metabolism	13	27
T-cells	sprouting angiogenesis	28	66
	T cell activation	24	59
	cellular response to interleukin-4	23	56
HSCs	nervous system development	56	569
	cell-cell signaling	18	617
	regulation of localization	13	636
Neutrophils	opsonization	22	76
	positive regulation of hepatic stellate cell activation	19	35
	response to fungus	17	33
Monocytes	activation of transmembrane receptor protein tyrosine kinase activity	56	213
	regulation of cardiac muscle contraction	33	167
	regulation of metal ion transport	8	55
Macrophages	regulation of localization	54	786
	nervous system development	53	440
	cell-cell signaling	19	506

C




Euclidean distance, Ward's method
0.3
0.2
0.1
0
AML: Acute Myeloid Leukemia
CLL: Chronic Lymphocytic Leukemia
MCL: Mantle Cell Lymphoma
CLL2
CLL4
CLL3
CLL5
Naive B cells
CLL1
CLL6
MCL1
MCL2
MCL3
Monocytes
AML3
T cells
AML1
Neutrophils
AML2
Cluster I:
Naïve B cells
Cluster II:
Monocytes
Cluster III: T cells
Cluster IV:
Neutrophils
CLL2
CLL4
CLL3
CLL5
Naive B cells
CLL
CLL6
MCL
MCL2
MCL3
Monocytes
AML14
T cells
AML
Neutrophils
AML4
0.0
0.04
Tumoral CDR divergence