Supplementary Note 1 – Detailed notes on data analysis performed

Transcription factor-based analyses of ATAC-seq data:

To compute TF accessibility scores, across hematopoiesis we made pairwise cell type comparisons using normalized accessibility, described above, down the hematopoietic hierarchy, wherein we compared a given cell type to its progenitor cell type using all technical and biological replicates. We then used the TF deviations analysis code, as previously described[23], to determine the gain or loss of accessibility of a given TF motif. In brief, this analytical tool determines a bias-corrected "deviation" score, analogous to a z-score, of a given annotation by first summing normalized accessibility across all peaks sharing the given annotation, these scores are then normalized to the variance observed within a background signal representing the expected signal given GC sequence bias and peak intensity bias. This analytical framework is discussed in great detail in the supplementary methods of previously published work[23]. This measure of TF accessibility is highly robust to the number of sequenced reads, DNA sequence bias, and signal-to-noise bias. Here, values shown as "deviation" are the average across all replicates representing a given cell type. TF motifs used in this analysis are from the Jaspar database and were mapped using FIMO[23], a full description of the motif can be found on the Jaspar website using the Jaspar ID provided in Supplementary Table 3. To filter for hematopoiesis TFs, we filtered for TFs with a z-score >1.5 (max deviation for a given TF in each cell type, over all TFs in all cell types). High-variant TFs were then filtered to a unique hematopoietic TF set by hierarchical clustering of the motif locations in hematopoiesis peaks, TFs with the highest z-score across hematopoiesis were chosen to represent a unique TF motif set. To represent deviation scores across the hierarchy, we computed "relative deviation" scores, which represent scores relative to HSCs. HSCs are defined to have a deviation score of 0. Scores for terminally differentiated cell states are represented as the sum of "relative deviation" scores across all upstream progenitors.

GWAS association analysis:

To test for enrichment of GWAS variants in open chromatin and regulatory regions, we used all GWAS data sets in the Roadmap GWAS database (N=67)[18] and the GRASP database (N=178)[54]. We also included two larger GWAS studies for Type 1 diabetes[55] and Alzheimer's disease[56]. The GWAS SNPs were pruned to contain no variants in linkage disequilibrium by keeping the most significant p-value where there were multiple linked variants for the same trait. These were then expanded to all linked variants with European $R^2 \geq 0.8$ for all further analysis.

We performed a rank-based enrichment of GWAS variants in the distal elements of each cell type profiled in the Roadmap Epigenomics Project. We segmented each GWAS study into bins representing different tiers of significance. We set a minimum bin size of 50 and filled the first bin with the 50 most significantly associated variants for each study. We then filled the next bins with 2*50, 4*50 and 8*50 variants and then segmented the remaining variants into bins at the four quartiles of the remaining p-value distribution. We then computed the rank fold change enrichment of distal elements across the segmented GWAS[30]. For each bin we computed the fraction of GWAS variants less than or equal to the bin's p-value threshold that overlapped distal elements in each Roadmap cell type to determine a set of significant GWAS peak association, which varied by GWAS considered. We calculated the fold change enrichment by dividing this fraction by the fraction of all GWAS variants of any significance level overlapping distal elements. Using this approach, we generated a table of the mean fold change enrichment of the two most significant bins for each Roadmap cell type in each GWAS. To calculate blood-specific GWAS annotations, we performed hierarchical clustering of row-normalized z-scores and found two distinct clusters representing “Blood” and “Others”. Using this list of blood-enriched GWAS, we applied the “deviation” pipeline (as described in the previous section for TF motifs), using an identical approach wherein each GWAS disease is analogous to a TF motif and each GWAS peak association is analogous to an individual TF motif occurrence in a peak.

Single-cell ATAC-seq and enhancer cytometry analysis:

Preprocessing for single-cell ATAC-seq data was done as described in “ATAC-seq Data Analysis”. To compute “Myeloid”, “Erythroid” and “Lymphoid” differentiation scores, we first learned the PC’s from bulk samples across all normal cell-types producing 12 PCs using MATLAB (SVD PCA). To reduce the effect of technical biases in the PCs, we averaged over technical and biological replicates and filtered for distal elements, as described for CIBERSORT above. We then re-scored the bulk and single-cell ATAC-seq data by subtracting the mean followed by multiplying by the coefficients of the PCs learned using MATLAB’s PCA implementation. The centroids for each cell type from the rescored bulk samples were used for downstream processing. Samples were projected onto the hematopoiesis PCs with two methods. First, single cells were fit as the mixture of normal bulk cell types (linear least squares) using the PC scores of the corresponding developmental lineage (“Myeloid”, “Erythroid” or “Lymphoid”) and projected onto the bulk PCs (Fig. 6c,d, Supplementary Fig. 12b,d–e,g). Cells with a correlation coefficient (Pearson) of less than 0.9 compared to the least-squares mixture were excluded, notably these cells were enriched for cells with low read numbers and empty wells. Second, synthetic “Myeloid”, “Erythroid” or “Lymphoid” developmental trajectories were computed by fitting a line across the associated cell types of each PC score. Developmental scores were then assigned as the maximum similarity (Pearson) of single-cells to the developmental trajectory (Fig. 6e,f). Cells with a maximum correlation coefficient of <0.9 were excluded. This approach was tested by down-sampling bulk cells to 1,000 fragments (Supplementary Fig. 12b,c). This single-cell approach was also independently validated by their score similarity to CIBERSORT from bulk cell lines (Supplementary Fig. 12f–k).

Supplementary Note 2 - Using the accessibility profiles of hematopoietic subsets to chart the ontogeny of human diseases

In our work, we demonstrate the applicability of our data to understanding the cell types responsible for various human diseases. By measuring the activity of regulatory elements that overlap regions with predicted sites of functional variation from GWAS, it is now possible to more accurately predict the specific cell types impacted by genetic variants linked to diverse human diseases (Supplementary Fig. 10a–c; see methods)[28–30]. To do this we first filtered for GWAS that were significantly enriched in hematopoietic cells (Supplementary Fig. 10a,b), then calculated "deviation" scores for each GWAS across the hematopoietic hierarchy (see methods). We found that each of these associations can be traced through the hematopoietic lineage to predict the developmental point at which each variant may first exert its effects, thus enriching our understanding of the developmental origins of human disease (Fig. 4h-k and Supplementary Fig. 10c).

As an example, polymorphisms linked to mean corpuscular volume (MCV), a measure of the average volume of an erythrocyte cell, are most strongly enriched in erythroblasts (Fig. 4h). Intriguingly, many regions associated with MCV polymorphisms first become accessible at the CMP and MEP stages suggesting that these polymorphisms may exert their effects prior to full erythroid lineage commitment. As a second example, polymorphisms associated with rheumatoid arthritis (RA) show a strong enrichment in B cells (Fig. 4i), consistent with the known role of autoantibodies and pathogenic B cells in the pathogenesis of RA, as well as the documented success of B cell depletion therapy in the treatment of RA[31,32].

We find a more complex pattern in the disease alopecia areata, an autoimmune disease characterized by hair loss. The autoimmunity driving this disease has recently been associated with both innate and adaptive immune responses[33], a result consistent with the enrichment of polymorphisms for alopecia areata in both T cells and monocytes (Fig. 4j). B cells also harbor many active elements associated with alopecia areata but have not been studied in this disease, suggesting a new direction of investigation. Importantly, our results are not limited to diseases canonically associated with hematopoietic cells; polymorphisms linked to Alzheimer's disease show a strong enrichment in B cells and monocytes, two cell types that have predicted roles in the pathogenesis of the disease[28,34,35] (Fig. 4k).

Supplementary Note 3 – Extended discussion on the regulatory heterogeneity observed in AML

Our bulk cell measurements of AML cell types show stark patterns of regulatory heterogeneity, for example clonally-derived LSCs show contributions from multiple and distinct cell states, i.e. HSC, LMPP, CMP, and GMP. This ensemble measure suggests that at the single-cell level, either i) cells represent a clonal outgrowth of a rare cell type and/or intermediate differentiation state, ii) cells have coopted regulatory programs, and exist as stable intermediate cell states that are not normally observed in normal hematopoiesis, or iii) this ensemble measurement is actually a mixture cell types whereby samples represent a mixture of HSC, CMP, LMPP, GMP and monocyte-like states.

To definitively distinguish between options (i) and (ii), would likely require extensive single-cell ATAC-seq of tens-of-thousands of single-cells. Such an effort, would likely uncover a vast repertoire of rare regulatory heterogeneity, which may also include rare intermediate cell states encompassing our bulk measurements in AML. In our interpretations of these data, we believe that the regulatory heterogeneity observed in AML cells could arise from all three scenarios presented above, including cell states that are not normally stable in hematopoiesis. Importantly, each observed AML sample exhibits considerable and unique regulatory heterogeneity, which suggests that the regulatory diversity in AML cells are not a product of 1 or 2 rare and uncharacterized stable progenitor cell states. We find some cases (SU353 blast cells) where there appear to be two epigenetic clusters of cells, supporting the hypothesis that regulatory heterogeneity in AML can arise from inter-cellular epigenetic clonal heterogeneity (Fig. 6f). Additionally, we also find other cases (SU070 LSCs, SU070 blasts, and SU353 LSCs) where the AML cells appear to show intra- rather than inter-cellular heterogeneity with single cells harboring mixed regulatory contributions (Fig. 6c,d, and f).

Importantly, we have also provided single cell regulome data from LMPPs and monocytes which show that the regulatory heterogeneity seen in single cells from primary AML samples is not encompassed by the diversity of individual normal myeloid lineage cells (Fig. 6e and Supplementary Fig. 12b-e). Our results from primary patient AML are additionally supported by scATAC-seq analysis of the commonly used clonal HL60 cell line (Supplementary Fig. 12g-h). Therefore, our conclusion is that AML cells show regulatory heterogeneity arising from multiple sources including i) single cells harboring mixed regulatory contributions from multiple normal cells and ii) intercellular epigenetic heterogeneity.