Systematic clustering algorithm for epigenetic data from high-throughput sequencing and its application to hematopoietic and leukemic cells

Azusa Tanaka1,2*, Yasuhiro Ishitsuka3,4*, Hiroki Ohta3,5*, Akihiro Fujimoto1, Jun-ichirou Yasunaga2,6 and Masao Matsuoka2,6

Abstract

The huge amount of data acquired by high-throughput sequencing requires data reduction for effective analysis. Here we develop a new data reduction method for genome-wide open chromatin data toward cell type classification. Regarding the genome as a string of 1s and 0s based on a systematically optimized set of peaks and calculating the Hamming distance enables us to quantitatively evaluate differences between samples of hematopoietic cells, classify cell types, and infer the origin of leukemic cells, potentially leading to a better understanding of leukemia pathogenesis.

Keywords: ATAC-seq; Hamming distance; Hematopoietic cell differentiation; Leukemia; Data reduction; Data normalization

Maintext

Cellular phenotypes are governed by epigenetic mechanisms. For example, to understand the differentiation and regulation of cells, information about how human DNA, which is 2 m long, is packed and chemically modified in the nucleus is desired [1, 2, 3, 4]. Methods such as chromatin immunoprecipitation sequencing (ChIP-seq) and assay for transposase accessible chromatin using sequencing (ATAC-seq) have proven to be useful for understanding the modification and detection of open chromatin on a genome-wide scale [5, 6, 7, 8, 9]. Those epigenetic data analysis techniques usually start with the detection of enrichment, also known as “peak calling”.

Compared to RNA-seq data analysis, whose target regions are mainly in certain loci or genes across samples, the target regions on epigenetic sequencing data are undetermined. To determine the target regions, peak calling with an appropriate tool is often performed first for the entire genome of every sample, and the target regions are defined as merged peaks among all samples. Then, the total number of reads or fragments present in each region is counted for each sample, leading to a matrix, \( X = (x_{i,j}) \), where \( x_{i,j} \) represents the number of reads/fragments from sample \( i \) in region \( j \). The matrix elements are normalized by quantile normalization to reduce the biases arising from variations in the data size over samples, followed by downstream processing [7, 8, 9].

However, this process raises two concerns. First, we do not fully understand the effect of merging all the peaks from different samples. For example, if two peaks from different samples slightly overlap, those two peaks are considered as one peak after the peak merging step. Therefore, the difference of the two peak positions, which may reflect cell identity, may be unintentionally ignored. The second concern is that we have no justification for applying quantile normalization over samples that are phenotypically different [10, 11].
Thus, the aim of the present study is to avoid these concerns by constructing an algorithm that systematically classifies epigenetic data obtained from high-throughput sequencing. In this analysis, we provide a systematic algorithm to determine the best parameters for peak calling and to select a set of peaks used for the downstream analysis, which greatly contributes to the data reduction. Further, to quantify the difference between samples, we introduce the Hamming distance, which often appears in information theory [12]. The Hamming distance is the number of mismatches between two strings of equal length. The string in our context is a binary sequence according to the peak presence or absence in an individual’s genome. This process contributes to data normalization.

In this report, we focused on 77 ATAC-seq datasets from 13 human primary blood cell types [7] as test data. The 13 cell types are comprised of hematopoietic stem cells (HSC), multipotent progenitor cells (MPP), lymphoid-primed multipotent progenitor cells (LMPP), common myeloid progenitor cells (CMP), megakaryocyte-erythroid progenitor cells (MEP), granulocyte-macrophage progenitor cells (GMP), common lymphoid progenitor cells (CLP), natural killer cells (NK), B cells, CD4\(^+\)T cells (CD4\(^+\)T), CD8\(^+\)T cells (CD8\(^+\)T), monocytes (Mono) and erythroids (Ery). For all 77 samples, we assigned ATAC-seq reads to reference genome hg19, and among them only those which had high mapping quality values (MQ \(\geq\) 30) were used for the peak calling by MACS2 (Additional file 1: Supplementary Note) [13]. The peak calling results consisted of the location with a width and the associated \(p\)-value. Note that we used MACS2 to call all ATAC-seq peaks with the following parameters (\(--\text{nomodel} \--\text{nolambda} \--\text{keep-dup all} -p p_G\)), where the number of peaks is affected by the peak calling parameter ‘\(-p\ p_G\)’. The parameter \(p_G\) is larger than any \(p\)-values of the peak results.

First, we ranked the peak results in the order of ascending \(p\)-values and then investigated the relationship between peak width and the corresponding ranking. We found that as the \(p\)-value increased, the width of the ATAC-seq peaks became shorter statistically, which suggested the feasibility of data reduction by selecting peaks with smaller \(p\)-values. Thus, we defined \(M_{\text{cut}}\) as the threshold such that only peaks with rankings not greater than \(M_{\text{cut}}\) were used for the analysis hereafter (Fig 1a). Then, for a given set of \((M_{\text{cut}}, p_G)\), we introduced \(B = \{h_x\}\), where \(h_x = 1\) when \(x\) is inside of a peak and otherwise 0. Finally, we evaluated the difference between two samples \((c, c')\) by using the Hamming distance \(H(B_c, B_{c'})\) between two binary sequences, \(B_c\) and \(B_{c'}\), which is the number of pairs of different values between \(B_c\) and \(B_{c'}\) at every position \(x\) (Fig 1b,c). Thus, we computed the distances between all pairs of 77 samples to build a distance matrix for various combinations of \((M_{\text{cut}}, p_G)\) and performed the unweighted pair group method with arithmetic mean (UPGMA) to draw dendrograms [14]. Next, to quantitatively evaluate the obtained dendrogram for each combination, we defined a type penalty, \(\lambda_\nu\), for a given cell type \(\nu\) for all 13 cell types above. Type penalty \(\lambda_\nu\) corresponds to the number of samples from different cell types in cluster \(\nu\) formed when all samples of cell type \(\nu\) meet together in the dendrogram (Fig 1d and Additional file 1: Supplementary Note). Additionally, we defined the penalty \(\lambda\) as \(\sum_{\nu \in T} \lambda_\nu\), where \(T\) is a set of the 13 cell types. We then sought the best parameter to optimize the dendrograms and
found that \((M_{\text{cut}}, p_G) = (12000, 10^{-2})\) gave the smallest penalty score (Fig 1e, f, g) in our searching resolution.

In order to understand why ATAC-seq data under the condition of \((M_{\text{cut}}, p_G) = (12000, 10^{-2})\) was well classified, we analyzed the properties of the top 50000 peaks. Specifically, in order to investigate the functional annotations of each ATAC-seq peak, we referred the ATAC-seq peaks in three cell types (HSC, B, and Mono) to the 15-state ChromHMM model data provided by the Roadmap Epigenomics Project [15, 16, 17]. ATAC-seq peaks were ranked according to \(p\)-values and divided into groups consisting of 1000 peaks. Then we calculated the average ratio of each group for each of the 15 states and also the standard deviation over all samples for each cell type. As shown in figure 1h and Additional file 1, though most of the peaks with higher rankings belonged to promoters, as the rank decreased, the ratio of the peaks from enhancer regions increased. The ratios of peaks from promoters and enhancers crossed at around peak rank 12000. This result suggests that the enhancer landscape plays an important role on cell identity, which is consistent with previous studies [7, 18, 19]. Moreover, when the peak number exceeded 20000, the standard deviation for enhancers greatly increased, demonstrating that the heterogeneity between samples became more dominant even in the same cell type, leading to a worse classification result (Fig 1h). This result indicates that the condition \(M_{\text{cut}} = 12000\) is adequate for achieving a balance between an enrichment of enhancer regions where cell characters are well reflected and a reduction of sample heterogeneity.

To evaluate the practicality of our algorithm on cancer research, we analyzed acute myelocytic leukemia (AML) by calculating the average distance over the same cell type between a leukemic cell sample and the above samples of the aforementioned 13 cell types. According to Ref. [7], these AML samples were divided into three stages. Preleukemic HSCs (pHSCs) were most closely related to HSCs and MPPs, leukemia stem cells (LSCs) were very similar to GMPs and LMPPs, and leukemic blasts showed a wider distribution. Some leukemic blasts were closest to GMP cells, while others clustered with monocytes. Table 1 summaries the clustering by providing the closest cell type for each AML sample (Additional file 1: Supplementary Note). We not only revealed the closest cell type, but also identified the second, third, and so on closest cell type, and quantified the difference between the characterization results of our algorithm and those from Ref. [7] as the “rank gap”. Overall, the results of our method had substantial overlaps with those of Ref. [7], validating our simple method.

In this paper, we have presented a novel algorithm to systematically classify cell types using the Hamming distance, which enabled us to find optimal parameters of the data reduction for the best cell classification. Further, this algorithm was found to quantitatively characterize the difference between hematopoietic cells and AML cells, which is broadly consistent with Ref. [7]. One of the remarkable points of our method is that we only used ATAC-seq data without gene expression data. Our analysis suggests that ATAC-seq data itself contains enough information to determine cell types even in the absence of regional annotation data such as promoters or enhancers in order to reveal elusive properties that significantly affect the phenotype of cell types. Another advantage of our method is that when adding new samples to the analysis, it is necessary to calculate the distances only between
newly appearing pairs of samples and not between preexisting samples. Thus, this systematic algorithm is flexible in terms of increasing sample numbers.

Finally, we expect that our approach with slight modifications will be applicable to other epigenetic sequencing data such as ChIP-seq and Bisulfite sequencing, whose target regions for analysis are not uniform between samples. Also, our algorithm could extend its application to leukemic samples whose cell of origin is uncertain. In the future, we will apply this method to adult T cell leukemia (ATL), a T cell leukemia in which the T cells in some cases morphologically resemble monocytes, to unveil the mechanism of this mysterious phenomenon [20].

| sample          | “closest normal cell” calculated in Fig.6i in Ref. [7] | “closest cell type” calculated by our algorithm | rank gap |
|-----------------|--------------------------------------------------------|--------------------------------------------------|----------|
| SU654-pHSC      | MPP                                                    | HSC                                              | 1        |
| SU353-pHSC      | MPP                                                    | MPP                                              | 0        |
| SU351-pHSC      | MPP                                                    | CMP                                              | 3        |
| SU209-pHSC1     | GMP                                                    | Mono                                             | 2        |
| SU209-pHSC2     | GMP                                                    | Mono                                             | 1        |
| SU209-pHSC3     | GMP                                                    | Mono                                             | 1        |
| SU070-pHSC1     | HSC                                                    | HSC                                              | 0        |
| SU070-pHSC2     | HSC                                                    | HSC                                              | 0        |
| SU048-pHSC      | MPP                                                    | HSC                                              | 1        |
| SU583-pHSC1     | GMP                                                    | GMP                                              | 0        |
| SU583-pHSC2     | GMP                                                    | Mono                                             | 2        |
| SU577-pHSC      | MPP                                                    | HSC                                              | 1        |
| SU501-pHSC      | MPP                                                    | HSC                                              | 1        |
| SU496-pHSC      | MPP                                                    | CMP                                              | 3        |
| SU484-pHSC      | MPP                                                    | HSC                                              | 1        |
| SU444-pHSC      | MPP                                                    | HSC                                              | 1        |
| SU654-LSC       | LMPP                                                   | LMPP                                             | 0        |
| SU583-LSC       | GMP                                                    | GMP                                              | 0        |
| SU577-LSC       | GMP                                                    | GMP                                              | 0        |
| SU496-LSC       | GMP                                                    | GMP                                              | 0        |
| SU444-LSC       | GMP                                                    | GMP                                              | 0        |
| SU353-LSC       | GMP                                                    | GMP                                              | 0        |
| SU209-LSC       | GMP                                                    | GMP                                              | 0        |
| SU070-LSC       | GMP                                                    | GMP                                              | 0        |
| SU654-Blast     | GMP                                                    | GMP                                              | 0        |
| SU444-Blast     | Mono                                                   | Mono                                             | 0        |
| SU353-Blast     | GMP                                                    | GMP                                              | 0        |
| SU351-Blast     | Mono                                                   | Mono                                             | 0        |
| SU209-Blast     | GMP                                                    | GMP                                              | 0        |
| SU070-Blast1    | Mono                                                   | Mono                                             | 0        |
| SU070-Blast2    | Mono                                                   | Mono                                             | 0        |
| SU048-Blast1    | GMP                                                    | Mono                                             | 0        |
| SU048-Blast2    | GMP                                                    | Mono                                             | 2        |
| SU048-Blast3    | GMP                                                    | Mono                                             | 1        |
| SU048-Blast4    | GMP                                                    | Mono                                             | 2        |
| SU048-Blast5    | GMP                                                    | Mono                                             | 1        |
| SU048-Blast6    | GMP                                                    | Mono                                             | 2        |
| SU583-Blast     | GMP                                                    | Mono                                             | 1        |
| SU577-Blast     | GMP                                                    | Mono                                             | 0        |
| SU501-Blast     | Mono                                                   | Mono                                             | 0        |
| SU496-Blast     | GMP                                                    | GMP                                              | 0        |
| SU484-Blast     | Mono                                                   | Mono                                             | 0        |

Table 1: Comparison between the “closest normal cell” of Fig.6i in Ref. [7] and “closest cell type” computed by our algorithm. The second and the third “closest type” were also identified by our algorithm. The “rank gap” represents the difference of the result between the two analytical methods. For example, the “closest normal cell” of sample SU-654-pHSC is MPP in Ref. [7], but is HSC by our algorithm. “MPP” was the second “closest cell type”. Thus, the rank gap was calculated as 2-1 (=1). If the results from the two analytical methods coincide with each other, the rank gap is 0.

Funding
This study was supported by JSPS KAKENHI Grant Numbers JP19K16740 (AT), JP15J04853 (AT) and a grant from the Naito Foundation (AT).
Author's contributions
AT, YI and HO performed the analysis and wrote the manuscript. All authors discussed the results and approved the final manuscript.

Acknowledgements
The authors thank P. Karagiannis for valuable comments and proofreading of this manuscript.

Author details
1 Department of Human Genetics, Graduate School of Medicine, The University of Tokyo. 2 Laboratory of Virus Control, Institute for Frontier Life and Medical Sciences, Kyoto University. 3 Center for Science Adventure and Collaborative Research Advancement, Graduate School of Science, Kyoto University. 4 Department of Mathematics, Graduate School of Science, Kyoto University. 5 Department of Physics, Graduate School of Science, Kyoto University. 6 Department of Hematology, Rheumatology and Infectious Disease, Faculty of Life Sciences, Kumamoto University.

References
1. Klemm, S.L., Shipony, Z., Greenleaf, W.J.: Chromatin accessibility and the regulatory epigenome. Nat Rev Genet 20, 207–220 (2019)
2. Gaspar-Maia, A., Alajem, A., Mesher, E., Ramalho-Santos, M.: Open chromatin in pluripotency and reprogramming. Nat Rev Mol Cell Biol 12, 36–47 (2011)
3. John, S., Sabo, P.J., Thurman, R.E., Sung, M.H., Biddle, S.C., Johnson, T.A., Hager, G.L., Stamatoyannopoulos, J.A.: Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genetics 43, 264–268 (2011)
4. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stengachis, A.B., Wang, H., Vrnout, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T.K., Diegel, M., Dunn, D., Ebersol, A.K., Frum, T., Giste, E., Johnson, A.K., Johnson, E.M., Kutyavin, T., Lajoie, B., Lee, B.K., Lee, K., London, D., Lotakis, D., Nephi, S., Neri, F., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L, Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Stamatoyannopoulos, G., Iyer, V.R., Lieb, J.D., Sungaev, S.R., Akey, J.M., Sabo, P.J., Kaul, R., Furey, T.S., Dekker, J., Crawford, G.E., Stamatoyannopoulos, J.A.: The accessible chromatin landscape of the human genome. Nature 489, 75–82 (2012)
5. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., Greenleaf, W.J.: Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10, 1213–1218 (2013)
6. Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzalez, M.L., Snyder, M.P., Chang, H.Y., Greenleaf, W.J.: Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523, 486–490 (2015)
7. 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., Majeti, R., Chang, H.Y.: Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genetics 48, 1193–1203 (2016)
8. Qu, K., Zaba, L.C., Satpathy, A.T., Giresi, P.G., Li, R., Jin, Y., Armstrong, R., Jin, C., Schmitt, N., Rahbar, Z., Ueno, H., Greenleaf, W.J., Kim, Y.H., Chang, H.Y.: Chromatin Accessibility Landscape of Cutaneous T Cell Lymphoma and Dynamic Response to HDAC Inhibitors. Cancer Cell 32, 37–414 (2017)
9. Rendeiro, A.F., Schmidt, C., Strefford, J.C., Walewska, R., Davis, Z., Farlik, M., Oscier, D., Bock, C.: Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenetic signatures and transcription regulatory networks. Nat Commun 7, 11938 (2016)
10. Meyer, S.U., Pfaff, M.W., Ulbrich, S.E.: Normalization strategies for microRNA profiling experiments: A 'normal' way to a hidden layer of complexity? Biotechnol Lett 32, 1777–1788 (2010)
11. Hicks, S.C., Irizarry, R.A.: QUANTO: A data-driven approach to guide the choice of an appropriate normalization method. Genome Biol 16(1), 1–8 (2015)
12. MacKay, D.J.C.: Information Theory, Inference and Learning Algorithms, p. 628. Cambridge University Press, Cambridge (2011)
13. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., Liu, X.S.: Model-based Analysis of ChIP-Seq (MACS). Genome Biol 9, 137 (2008)
14. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis, 5th edn. Wiley Series in Probability and Statistics, p. 330. John Wiley & Sons, Ltd., Chichester (2011)
15. Ernst, J., Kellis, M.: ChromHmm: automating chromatin-state discovery and characterization. Nat Methods 9, 215–216 (2012)
16. Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J.A., Birney, E., Hardison, R.C., Dunham, I., Kellis, M., Noble, W.S.: Integrative annotation of chromatin elements from encode data. Nucleic Acids Res 41, 827–841 (2013)
17. Li, D., Hsu, S., Purushotham, D., Sears, R.L., Wang, T.: WashU Epigenome Browser update 2019. Nucleic Acids Res 47, 158–165 (2019)
18. Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D.A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., Friedman, N., Amit, I.: Immunogenetics. Chromatin state dynamics during blood formation. Science 345, 943–949 (2014)
19. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., Lee, W., Lilje, B., Drabløs, F., Lennartsson, A., Rönnbлад, M., Hrydziuszko, O., Vitezic, M., Freeman, T.C., Alhendi, A.M.N., Arner, P., Axton, R., Baillie, J.K., Beckhouse, A., Bodega, B., Briggs, J., Brombacher, F., Davis, M., Detmar, M., Ehrlund, A., Endoh, M., Eslami, A., Fagioliini, M., Fairbairn, L., Faulkner, G.J.: Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science 347, 1010–1015 (2015)
20. Matsuoka, M., Jeang, K.T.: Human T-cell leukaemia virus type 1 (HTLV-1) infectivity and cellular transformation. Nat Rev Cancer 7, 270–280 (2007)

Additional File
Additional file 1:
Supplementary Note, Tables S1, S2, S3 and Figures. S1-S13.
Figure 1 The systematic clustering algorithm and its application to cell classification. 

a) Distribution of peaks depending on $p$-values. 

b) Schema of the Hamming distance calculation from the peak locations. Each locus is converted to 1 or 0 based on the peak overlapping status. 

c) Matrix of Hamming distances $d_{ij}$ between samples $i$ and $j$. This matrix is used for the downstream analysis. 

d) Schema of calculating the type penalty $\lambda_\nu$ (for $\nu=\text{HSC}$). First, we find the lowest node under which we can find all samples belonging to cell type $\nu$ (for HSC, this is $n_{10}$). The type penalty for HSC is the number of samples under node $n_{10}$ that are not HSC (in the example shown, $\lambda_\text{HSC} = 1$). e) The penalty score $\lambda$ depends on $M_\text{cut}$ and $p_G$. $\lambda$ represents the sum of the type penalty scores for all cell types. 

f) UPGMA clustering of cell samples from all replicates of 13 normal hematopoietic cell types. Values shown are Hamming distances. 

g) Unrooted UPGMA dendrogram of cell samples showing correlations between cell types. 

h) Functional annotation results of ATAC-seq peaks from Monocyte (6 samples). Horizontal axis represents the peak rank, and longitudinal axis represents the percentage of each of the 15 functional states for every group of 1000 peaks (only percentages $\geq 12\%$ at the maximum are shown).
This supplementary note explains in detail the construction of the systematic clustering algorithm for epigenetic data in terms of cell types and also discusses the details of the outputs from its application to samples of hematopoietic cells and leukemic cells.

Contents

I. Background and preprocessing of epigenome data  
   A. Cell type classification by cell surface markers  
   B. ATAC-seq: Analysis for open chromatin regions based on Tn5-transposase  
   C. Reads  
   D. Alignment of reads onto the reference genome  
   E. Pilings of reads

II. Algorithm: Systematic data reduction and hierarchical clustering optimization  
   A. How to systematically “normalize” epigenome data of different sizes  
      1. Peak-calling with ranking  
      2. Parameterized binarization by cutting off low-ranked peaks  
   B. Quantifying differences between two binary sequences  
      1. Hamming distance  
   C. Review: Hierarchical clustering  
      1. UPGMA algorithm  
      2. How to draw dendrograms  
   D. Optimization of hierarchical clustering for cell-type classification  
      1. Global penalty as a cost function  
   E. Computational cost of the algorithm

III. Application to hematopoietic cells  
   A. Determination of the best parameters for the best cell-type classification  
   B. How to relate the best parameters to genomic context  
      1. Irrelevance to genomic regions corresponding to a cell surface marker  
      2. Relevance to enhancer regions  
   C. Robustness of the clustering

IV. Application to Leukemic cells

References

A. Information of samples
I. BACKGROUND AND PREPROCESSING OF EPIGENOME DATA

The main objective of this supplementary note is to explicitly construct a systematic algorithm that classifies given data of ATAC-seq (assay for transposase accessible chromatin using sequencing) in terms of cell types as determined by cell surface markers. In this paper, we mainly target ATAC-seq data for 13 types of hematopoietic cells [1].

First, we start with a basic background about hematopoietic cells and ATAC-seq in Sections I A and I B. Then, we explain a rather conventional way to preprocess ATAC-seq in Sections I C to I E.

A. Cell type classification by cell surface markers

Hematopoietic stem cells (HSC) are multipotent progenitors and the origin of all functional blood cells. Though extensive studies over decades have sought to determine a unified picture of the differentiation pathways, there remains no consensus [2]. In this paper, we consider only 13 cell types of hematopoietic cells by setting

\[ T := \{B, CD4^+T, CD8^+T, CLP, CMP, Ery, GMP, HSC, LMPP, MEP, Mono, MPP, NK\} \]

These cell types are experimentally categorized by the cell surface markers as shown in TABLE S1 [1].

| Cell type | Cell surface marker |
|-----------|---------------------|
| HSC       | Lin-, CD34+, CD38-, CD10-, CD90+ |
| MPP       | Lin-, CD34+, CD38-, CD10-, CD90- |
| LMPP      | Lin-, CD34+, CD38-, CD10-, CD45RA+ |
| CMP       | Lin-, CD34+, CD38+, CD10-, CD45RA-, CD123+ |
| MEP       | Lin-, CD34+, CD38+, CD10-, CD45RA-, CD123- |
| GMP       | Lin-, CD34+, CD38+, CD10-, CD45RA+, CD123+ |
| CLP       | Lin-, CD34+, CD38+, CD10+, CD45RA+ |
| NK        | CD56+ |
| B         | CD19+, CD20+ |
| CD4^+T    | CD3+, CD4+ |
| CD8^+T    | CD3+, CD8+ |
| Mono      | CD14+ |
| Ery       | CD71+, GPA+, CD45-low |

TABLE S1: Types of hematopoietic cells and their corresponding cell surface markers in Ref. [1]. For example, CD34+ and CD38- for cell type \( \nu \) means that a cell of type \( \nu \) expresses CD34 but not CD38 at its surface. In this paper, we study hematopoietic stem cells (HSC), multipotent progenitor cells (MPP), lymphoid-primed multipotent progenitor cells (LMPP), common myeloid progenitor cells (CMP), megakaryocyte-erythroid progenitor cells (MEP), granulocyte-macrophage progenitor cells (GMP), common lymphoid progenitor cells (CLP), natural killer cells (NK), B cells (B), CD4^+T cells (CD4^+T), CD8^+T cells (CD8^+T), monocytes (Mono), and erythrocytes (Ery).

B. ATAC-seq: Analysis for open chromatin regions based on Tn5-transposase

Throughout this supplementary note, we use hg19 as the human reference sequence [3]. It consists of 24 groups of symbol sequences, which corresponds to chromosomes labeled as \( \mathbb{X} = \{1, 2, \ldots, 22, X, Y\} \). The underlying structure of a chromosome is a long chain of DNA and the DNA is represented as a sequence of elements in set

\[ \mathbb{D} := \{A, T, G, C\}, \]

where each symbol corresponds to the nucleotides adenine (A), thymine (T), guanine (G), and cytosine (C).

For the \( \gamma \)-th chromosome \( (\gamma \in \mathbb{X}) \), the length of the corresponding DNA sequence is written as \( L_\gamma \), where \( 5.0 \times 10^7 \leq L_\gamma \leq 2.5 \times 10^8 \) and the total length is \( L = \sum_{\gamma=1}^{22} L_\gamma + L_X + L_Y \sim 3.1 \times 10^9 \). To position the \( x \)-th base pair in the \( \gamma \)-th chromosome, we set

\[ b^\gamma_x \in \mathbb{D}, \]

with \( 1 \leq x \leq L_\gamma \). In this paper, for set \( \mathbb{S} \), we write the number of elements in \( \mathbb{S} \) as \( |\mathbb{S}| \). For example, we have \( |\mathbb{D}| = 4 \) and \( |\mathbb{X}| = 24 \).
Chromatin is a complex of DNA and associated proteins such as histones. A chromatin has “open” regions, around which the density of the DNA and the associated proteins are rather low and also “closed” regions, around which the opposite situation happens. Gene expressions are largely regulated by the interactions between DNA and transcription factors depending on the open and closed regions. The analysis of open/closed chromatin regions is necessary for the understanding of cell differentiation and phenotype [4, 5]. ATAC-seq was developed for the genome-wide detection of open chromatin regions. One of the features of ATAC-seq is that it uses Tn5 transposase. At a certain proper condition, Tn5-transposase mainly cut DNA in open chromatin regions and the sequences of those DNA fragments of open chromatin regions. One of the features of ATAC-seq is that it uses Tn5 transposase. At a certain proper condition, Tn5-transposase mainly cut DNA in open chromatin regions and the sequences of those DNA fragments are obtained by sequencers [6]. ATAC-seq has several advantages compared to the other epigenomic sequencing methods [7]. For example, to analyze open chromatin regions, DNase-seq needs about $10^2$-$10^3$ cells and takes 4-5 days to obtain the data. ATAC-seq, on the other hand, requires only about $10^3$-$10^4$ cells and takes half a day.

### C. Reads

As briefly reviewed above, one Tn5-transposase cuts and splits DNA into two parts or fragments. If there are two Tn5-transposases, two locations of DNA are cut to make three fragments.

Thus, we can view fragment $f$ as a subsequence of a DNA sequence consisting of successive symbols in $\mathbb{D}$. Since we refer to the same DNA sequence of the human genome in this study, fragment $f$ can be also represented by three coordinates: the chromosome number $\gamma \in \mathbb{X}$, the start position $s = s(f)$, and the end position $e = e(f)$, where $1 \leq s \leq e \leq L_{\gamma}$. In other words, $f$ is a sequence $(b_{\gamma}^{s}, b_{\gamma}^{e+1}, \ldots, b_{\gamma}^{e})$, that can be expressed as $f = (\gamma, s, e)$. A sample is, in our settings, a product generated by a certain experimental procedure through ATAC-seq library preparation from a set of cells [6].

The input of a sequencer is the set of the obtained fragments $\{f_i\}_{i=1}^{N_{i}}$, where a fragment $f_i$ is $(\gamma_i, s_i, e_i)$, its length $L(f_i)$ is equal to $e_i - s_i + 1$, and the number of fragments is denoted as $N_{i}$. A sequencer with “paired-end sequencing” outputs a DNA sequence of the two edges of a fragment as two reads $(R_i^e, R_i^s)$ where

$$R_i^e := (R_j)_{j=1}^{e_i}, \quad R_i^s := (R_j')_{j=1}^{s_i} \in \mathbb{D},$$

meaning that each length of the two reads $(R_i^e, R_i^s)$ is $\ell_i$.

We consider read as a sequence of four symbols in $\mathbb{D}$ of length less than or equal to $\ell_0$, where $\ell_0$ can be changed as a parameter controlled by the sequencer. Note that for the case of “single-end sequencing”, where one gets only a read from one edge, we obtain read $R_i = \{R_j\}_{j=1}^{e_i}$. In the end, we obtain the data of reads $R := \{R_i\}_{i=1}^{N_{i}}$ where the number of reads is denoted as $N_{i}$. Note that in the case of “paired-end sequencing”, one may regard both $R_i^e$ and $R_i^s$ as $R_i$. This is the starting point of our analysis because sequencers do not directly give the actual values of $f_i$.

Summarizing the relationship between fragments and reads, let us assume that all reads are obtained from “paired-end sequencing” and that the sample preparation and the sequencer output are “ideal” as follows. If we denote fragment $f_i$ as sequence $(b_{\gamma_i}^{s_i}, b_{\gamma_i}^{s_i+1}, \ldots, b_{\gamma_i}^{e_i})$, then the beginning read $R_i^e$ and the terminal read $R_i^s$ corresponding to $f_i$ are

$$R_i^e = \begin{cases} (b_{\gamma_i}^{s_i}, b_{\gamma_i}^{s_i+1}, \ldots, b_{\gamma_i}^{s_i+\ell_0-1}) & \text{for } \ell_0 \leq L(f_i), \\ (b_{\gamma_i}^{s_i}, b_{\gamma_i}^{s_i+1}, \ldots, b_{\gamma_i}^{e_i}) & \text{for } \ell_0 > L(f_i), \end{cases}$$

$$R_i^s = \begin{cases} (b_{\gamma_i}^{e_i-\ell_0+1}, b_{\gamma_i}^{e_i-\ell_0+2}, \ldots, b_{\gamma_i}^{e_i}) & \text{for } \ell_0 \leq L(f_i), \\ (b_{\gamma_i}^{s_i}, b_{\gamma_i}^{s_i+1}, \ldots, b_{\gamma_i}^{e_i}) & \text{for } \ell_0 > L(f_i). \end{cases}$$

In other words, if the length $L(f_i)$ of fragment $f_i$ is greater than or equal to $\ell_0$, the beginning read $R_i^e$ is the direct inference of the first $\ell_0$ symbols of the fragment $f_i$. The condition for terminal read $R_i^s$ is similar. If length $L(f_i)$ is less than $\ell_0$, we directly infer $R_i^e = R_i^s = f_i$ as a sequence of four symbols, where we see that the two reads have the same length. However, the situation above is “ideal” and there are unexpected errors that stochastically flip symbols in the ideal situation. Thus, we need to infer the information of fragments in a statistical manner. Note that this inference can be straightforwardly applied to the case of “single-end sequencing”.

### D. Alignment of reads onto the reference genome

Hereafter, for simplicity, we consider single-ended reads $R = \{R_i\}_{i=1}^{N_{i}}$ because similar processes can be done for paired-end reads. We perform mapping of the reads data $R$ from a sequencer onto the DNA sequence. We use the
BWA-MEM algorithm of the software *BWA* (v0.7.16a) with no options. This algorithm aligns each read onto the reference sequence \((b^*_\gamma)_{\gamma \in \mathbb{X}, 1 \leq x \leq L_\gamma}[3]\) and gives an estimate of the quality of the alignment (for details, see [8] and references therein). Then we obtain the following data:

- Chromosome number \(?\gamma(R_i) \in \mathbb{X} \cup \{U\}\) with the start position \(?\hat{o}(R_i)\) of read \(R_i\) mapped onto the DNA sequence, where \(1 \leq \hat{o}(R_i) \leq L_\gamma(R_i)\). Note that \(U\) is a set of unplaced sequences in any elements in \(\mathbb{X}\). Hereafter, \(\mathbb{X}\) includes \(U\) with \(L_U \approx 3.7 \times 10^6\).
- The mapping quality score \(MQ(R_i) \geq 0\) of read \(R_i\) calculated by using the Phred quality score.

Therefore, \((\hat{\gamma}(R_i), \hat{o}(R_i))\) infers the coordinates \((\gamma_i, s_i, s_i + \ell_i) = (\hat{\gamma}(R_i), \hat{o}(R_i), \hat{o}(R_i) + \ell_i)\) of read \(R_i\) onto the DNA sequence. For \(R_i\), we define \(\hat{T}(R_i)\) as

\[
\hat{T}(R_i) := (\hat{\gamma}(R_i), \hat{o}(R_i), MQ(R_i)) .
\]

To select reliable data with \(\hat{T}(R_i)\), we preprocess the outputs obtained above as follows:

(i) In order to reduce duplicated reads, which could be produced artificially in the sequence sample preparation, we apply the command *MarkDuplicates* in PICARD software (v1.119) [9] with the *REMOVE_DUPLICATE* option.

(ii) Then we cut off reads with a mapping quality score \(MQ(R_i)\) less than 30. We used *samtools* for this purpose [10].

After processes (i) and (ii), we obtain

\[
\hat{P}(R') := \{(\hat{\gamma}(R'_i), \hat{o}(R'_i), \ell'_i)\}_{i=1}^{N_r},
\]

where \(\ell'_i\) is the length of \(R'_i\) and \(N_r\) denotes the number of reads after preprocessing. \(\{R'_i\}_{i=1}^{N_r}\) can be straightforwardly determined by \(\{R_i\}_{i=1}^{N_r}\). This is part of the information obtained by the preprocessing. Note that \(MQ(R'_i) \geq 30\) holds for any \(i\) with \(1 \leq i \leq N_r\) and there are no duplicated pairs in \(\hat{P}\). For simplicity, hereafter, we sometimes express \(\hat{P}(R)\) as \(P\). We use similar abbreviations for other symbols.

### E. Pilings of reads

From the data \(\hat{P}\), we can calculate how many reads are on position \((\gamma, x)\) in the DNA sequence. We consider the set of reads located on position \((\gamma, x)\) symbolically by defining

\[
\forall_{\gamma,x}(\hat{P}(R)) := \{1 \leq i \leq N_r \mid \hat{\gamma}(R_i) = \gamma \text{ and } (\hat{o}(R_i) \leq x \leq \hat{o}(R_i) + \ell_i)\} .
\]
FIG. S2: Histogram for the width of peaks ($\beta_k - \alpha_k$) obtained from $g_k$ vs the ranking $k$ determined by the associated $p$-value $p_k$ in ascending order with $p_G = 10^{-2}$ for a CD4$^+$ T cell. The bin size is 400 × 400.

FIG. S3: (a) The number $Y_\gamma = Y_{\gamma,x}$ with $\gamma = 3$ of reads as a function of position $x$ in the DNA sequence and the peak region ($\alpha_k, \beta_k$) determined by the MACS2 algorithm. Both values are calculated by ATAC-seq. The peak regions and the associated $p$-values ($\alpha_k, \beta_k, p_k$) in the left and right peaks with $p_G = 10^{-2}$ are ((188271079, 188271985), $10^{-422.5872}$) and ((188286401, 188287077), $10^{-329.52139}$), respectively. The width of the peaks ($\beta_k - \alpha_k$) in the left- and right-hand sides are 906 and 676, respectively. (b) Binary sequence ($h_x$) determined by the peak regions seen in (a).

For two samples in reads data $R$ obtained from SRA (SRR2920495.sra and SRR2920466.sra) used in Ref. [13], we show $Y_{\gamma,x} := |Y_{\gamma,x}|$, which is the number of reads on each position ($\gamma, x$) in the DNA sequence in Fig. S1. Hereafter, we will use reads data $R$ from Ref. [13] as the initial input of the analysis.

II. ALGORITHM: SYSTEMATIC DATA REDUCTION AND HIERARCHICAL CLUSTERING OPTIMIZATION

A. How to systematically “normalize” epigenome data of different sizes

When we analyze the preprocessed ATAC-seq data with $P(R)$, we have to care for biases caused by the fact that the amount of reads $N_r$ depends on the setting for the samples preparation and on the sequencers used. Normalization is done to remove such biases.

A conventional way to perform normalization is to use quantile normalization, where the distribution of the reads number on certain regions in the DNA is assumed to be the same for all samples [11, 12]. However, there is no strong reason to support this assumption, particularly for sample sets of different cell types. Furthermore, under this
assumption, there is a risk that we overlook important differences between different cell types. Therefore, in this paper, we do not assume this property.

An alternative way to perform normalization is to reduce the data into a simple binary value \( h_{\gamma,x} \in \{0, 1\} \) on each genomic position \((\gamma, x)\), where \(h_{\gamma,x}\) depends on the data size \(N_x\) as little as possible. For example, one could determine the state of \(h_{\gamma,x} = 1\) and \(h_{\gamma,x} = 0\) as an “open” and “closed” chromatin status, respectively, on genomic position \((\gamma, x)\).

In this direction, our ultimate purpose is to look for the “best” principle that determines two states for \(h_{\gamma,x}\), by which a set of samples including different cell types are completely classified into groups of the same cell type. We use no information about cell types when determining the value of \(h_{\gamma,x}\), because we would like to have an algorithm that can be applied without knowing the cell types.

1. Peak-calling with ranking

Currently we do not have the “best” solution to properly determine two effective states for \(h_{\gamma,x}\). As a candidate to approach the “best” solution, we use the MACS2 algorithm, which was originally invented to analyze ChIP-seq data [14, 15] but is now widely used to estimate the location of open chromatin regions from ATAC-seq data [16, 17].

We would like to find the set of position \((\gamma, x)\), where the number of the associated reads, \(Y_{\gamma,x}(P(R))\), is relatively high in the neighborhood \((\gamma, x)\). The MACS2 algorithm is likely to detect those positions from the data of the reads described by \(P(R)\) (for details of this algorithm, see [14, 15]). In our calculation, we use the MACS2 (v2.1.2) callpeak command with option “--nomodel --nolambda --keep-dup all -p ph"", where we need to set parameter \(pG\) as a parameter of peak inference (for details, see [14, 15]). We will later come back to how to determine a better value for parameter \(pG\).

By applying MACS2 to the input ATAC-seq data, we obtain the following output data structure:

- The label \(\gamma_k \in X\) of chromosome where the \(k\)-th peak belongs, the start position \(1 \leq \alpha_k \leq L_\gamma\) of \(k\)-th peak and the end position \(1 \leq \beta_k \leq L_\gamma\) for \(1 \leq k \leq M\) (here \(M\) is the number of peaks). We call \(g_k = (\gamma_k, \alpha_k, \beta_k)\) the \(k\)-th peak region.

- For each \(g_k\), \(p\)-value \(p_k\) with \(p_k \leq pG\) associated to \(k\)-th peak. Note that MACS2 outputs \(\log_{10}(1/p_k) = -\log_{10} p_k\) instead of \(p_k\).

We define \(A\) as

\[
A := (g_k, p_k)_{k=1}^M,
\]

\[
g_k := (\gamma_k, \alpha_k, \beta_k).
\]

By reordering the terms of \(k\), we can set \(p_k \leq p_{k'}\) for any \(k < k'\) without loss of information.

In Fig. S2, we show the distribution of peak width \(|\beta_k - \alpha_k|\) versus ranking \(k\). Note that \(g_k\) with high \(p_k\) could be affected significantly by the conditions of the experiments including sequencing, because the data above the rank value 40000 unnaturally touches the value of the lower limit of width 200, which is predetermined by the MACS2 algorithm. Thus, there is a possibility that peaks with higher \(p\)-values could strongly depend on both the inference algorithm and the number of reads \(N_x\), and thus would presumably not contribute to the detection of cell phenotypes. This observation suggests we should remove peaks with higher \(p\)-values.

2. Parameterized binarization by cutting off low-ranked peaks

Next, let us reconsider how to alleviate biases in the data by introducing threshold number \(M_{cut}\), such that

\[
\overline{A}(M_{cut}) := \{g_k\}_{k=1}^{M_{cut}},
\]

which leads to the removal of \(\{g_k\}_{k=M_{cut}+1}^M\) as a candidate for the normalization of the ATAC-seq data. Note that \(\overline{A}(M_{cut} = \infty) = \{g_k \mid (g_k, p_k) \in A\}\). Then, by using \(\overline{A}\), we may introduce a binary sequence

\[
B := \{h_{\gamma,x}\}_{\gamma \in \mathbb{X}, \alpha \leq x \leq \beta钾},
\]

such that \(h_{\gamma,x} = 1\) if there is \(k\) satisfying \(\alpha_k \leq x \leq \beta_k\) with \((\alpha_k, \beta_k) \in \overline{A}\) and otherwise \(h_{\gamma,x} = 0\) as shown in Fig. S3.

Note that \(pG\) and \(M_{cut}\) can be regarded as parameters for determining the value of \(h_{\gamma,x}\) within the MACS2 algorithm and what part of the data is taken into account, respectively. Thus, our task under the principle above turns out to be how to determine a proper set of \((M_{cut}, pG)\) for better cell classification. This issue is studied later.
B. Quantifying differences between two binary sequences

Let us move onto the situation when one considers a set of samples to evaluate the difference between two binary sequences $B$. Here, our strategy is to find the proper distance that can be measured from the normalized ATAC-seq data of two samples. Using that distance, we try to obtain hierarchical clustering of a set of hematopoietic cell samples to quantitatively characterize the relationship among those samples. Note that $p_G$ and $M_{cut}$ do not depend on the sample.

Preliminarily, we introduce conventions. Let $N_s$ be the number of samples. We then write the set of samples as $\mathbb{S} := \{1, 2, \ldots, N_s\}$, where $N_s = 77$ in this study [1]. For sample $c \in \mathbb{S}$, we add index $c$ to related objects as a superscript. For example, we use the symbol $N^c_s$ as the number of reads of sample $c$. Similarly, we write the truncated data of peaks associated to sample $c$ as $B^c := \{h^c_{\gamma,x}\}_{\gamma \in \mathbb{X}, 1 \leq x \leq L_\gamma}$.

There are many methods to evaluate the difference between a binary sequence $B^c$ from sample $c \in \mathbb{S}$ and $B^{c'}$ from sample $c' \in \mathbb{S}$. In this paper, we take the Hamming distance for all sample pairs. We use the distance as an initial condition for the hierarchical clustering and then use the Unweighted Pair Group Method with Arithmetic mean (UPGMA) to complete the hierarchical clustering [18]. Examples of hierarchical clustering with $(M_{cut}, p_G) = (2000, 10^{-2})$ and $(50000, 10^{-2})$ are shown in Fig. S4. We will later come back to the dependence of the results on different hierarchical clustering methods from UPGMA in Section III C.

1. Hamming distance

The Hamming distance is often used to compare two binary sequences in information theory (it is also called $p$-distance in cladistics) and is equal to the number of positions on which two symbols have different values.

The Hamming distance between two binary sequences $B^{c_1}$ and $B^{c_2}$ with $c_1, c_2 \in \mathbb{S}$ is defined as

$$H(B^{c_1}, B^{c_2}) := \sum_{\gamma \in \mathbb{X}, 1 \leq x \leq L_\gamma} \delta(h^{c_1}_{\gamma,x}, h^{c_2}_{\gamma,x}),$$

where we define

$$\delta(h^{c_1}_{\gamma,x}, h^{c_2}_{\gamma,x}) = \begin{cases} 1 & (h^{c_1}_{\gamma,x} \neq h^{c_2}_{\gamma,x}) \\ 0 & (h^{c_1}_{\gamma,x} = h^{c_2}_{\gamma,x}). \end{cases}$$

C. Review: Hierarchical clustering

In this subsection, we give a detailed explanation of the algorithms used to perform UPGMA (Group Averaging) and draw dendrograms. For more details about the algorithms and the implementation, see [18].

1. UPGMA algorithm

To describe the algorithm, we define the distance between two subsets $C_1, C_2 \subset \mathbb{S}$ as the average of all distances between samples in $C_1$ and samples in $C_2$. Equivalently, we define

$$\mathcal{H}(C_1, C_2) := \frac{1}{|C_1||C_2|} \sum_{c_1 \in C_1} \sum_{c_2 \in C_2} H(B^{c_1}, B^{c_2}).$$

If $C_1$ or $C_2$ is empty, we set $\mathcal{H}(C_1, C_2) = 0$. We sometimes identify sample $c \in \mathbb{S}$ and subset $\{c\}$ of single element $c$. For example, we write $\mathcal{H}(C_1, \{c\})$ for $\mathcal{H}(C_1, \{c\})$. Note that $\mathcal{H}(\{c_1\}, \{c_2\}) = \mathcal{H}(c_1, c_2) = H(B^{c_1}, B^{c_2})$ by definition.

We define a cluster as subset $C$ of $\mathbb{S}$ with a specified order of elements. UPGMA is an algorithm that can construct set $M_N$ of clusters and order the elements in $\mathbb{S}$ to draw dendrograms.

1. We set $C_\tau := \{\tau\}$ for $1 \leq \tau \leq N_s$. We do not consider the order of the elements in $C_\tau$ since they are sets of a single element.
(ii) We define the list of uncombined clusters as $L_1 := \{C_1, C_2, \ldots, C_{N_s}\}$, and set the historical list of clusters as $M_1 = L_1$.

(iii) At the $t$-th step ($1 \leq t \leq N_s - 1$), we define $C_{t+N_s}, L_{t+1}$ and $M_{t+1}$ inductively.

(a) We look up the pair $C_{\tau'}, C_{\tau''}$ with $\tau' < \tau''$ in $L_t$ such that their distance is a minimum; that is,

$$\mathcal{H}(C_{\tau'}, C_{\tau''}) = \min_{C', C'' \in L_t} \mathcal{H}(C', C'').$$

Note that $1 \leq \tau' < \tau'' < t + N_s$ by construction. We consider only the case when the pair is uniquely determined.
(b) We define a new cluster \( C_t + N_s = C_{\tau'} \cup C_{\tau''} \). If the elements of \( C_{\tau'} \) are ordered as \( c_1, c_2, \ldots, c_z \) and the elements of \( C_{\tau''} \) are \( c'_1, c'_2, \ldots, c'_z \), the elements of \( C_t + N_s \) are ordered as

\[
e_1, e_2, \ldots, e_z, c'_1, c'_2, \ldots, c'_z.
\]

(c) We define

\[
\mathcal{L}_{t+1} := (\mathcal{L}_t \setminus \{C_{\tau'}, C_{\tau''}\}) \cup \{C_t + N_s\},
\]

\[
\mathcal{M}_{t+1} := \mathcal{M}_t \cup \{C_t + N_s\}.
\]

If \( t < N_s - 1 \), go to the \((t+1)\)-th step.

We can easily see that if we do not consider the ordering, then we have \( C_{2N_s - 1} = \mathcal{S} \) as a set. Thus we finally obtain a list of \( 2N_s - 1 \) clusters \( \mathcal{M}_{N_s} = \{C_1, C_2, \ldots, C_{2N_s - 1}\} \), and an ordering of all elements of \( \mathcal{S} \) from \( C_{2N_s - 1} \).

2. How to draw dendrograms

The (rooted) dendrogram displays how our clustering combines pairs of clusters and the distance of the pairs. In the following, we explain an algorithm that introduces new symbols.

(i) If sample \( \tau \in \mathcal{S} \) appears in the ordering of \( C_{N_s - 1} \) as the \( a_{\tau} \)-th element, then we associate point \( n_\tau = (a_\tau, 0) \) in two-dimensional coordinate space to cluster \( C_{\tau} \). We call point \( n_\tau \) the leaf, which corresponds to \( C_{\tau} \).

(ii) For \( 1 \leq t < N_s - 1 \), we inductively associate point \( n_{t + N_s} \) to cluster \( C_{t + N_s} \). If \( C_{t + N_s} \) is constructed as the union of \( C_{\tau'} \) and \( C_{\tau''} \), with \( 1 \leq \tau' < \tau'' < t + N_s \), we associate to \( C_{t + N_s} \) the node

\[
n_{t + N_s} = \left( a_{t + N_s} = \frac{a_{\tau'} + a_{\tau''}}{2}, \mathcal{H}(C_{\tau'}, C_{\tau''}) \right).
\]

Note that \( C_{\tau'} \) and \( C_{\tau''} \) are uniquely determined. We call \( n_{t + N_s} \) the node associated to the \((t + N_s)\)-th cluster \( C_{t + N_s} \).

(iii) We connect \( n_{t + N_s} \) with \( n_{\tau'} \) and \( n_{\tau''} \).

Since each node or leaf \( n \) corresponds to cluster \( C \), we can define the offspring set \( \mathcal{B}_n \) of \( n \) as set \( C \) without ordering. Graphically, the offspring set of node \( n \) is the set of samples corresponding to leaves branched from node \( n \), as displayed in Fig. S5. This intuitive explanation is justified, since the \( y \)-coordinate of the “mother node” \( n_{t + N_s} \) is larger than or equal to those of the “child nodes” \( n_{\tau'}, n_{\tau''} \) if we use UPGMA. Note that there are many choices to draw dendrograms; for example, at any branching node, we can exchange two branches without any essential change in the data structure.

D. Optimization of hierarchical clustering for cell-type classification

By using the methods explained above, we can obtain a clustering diagram that depends on \((M_{\text{cut}}, p_G)\). Thus, our next task is to quantitatively characterize better clustering.

Recall that each cell \( c \in \mathcal{S} \) has type \( \nu \in \mathcal{T} \). We denote by \( \mathcal{S}_\nu \subset \mathcal{S} \) the set of all samples of the same cell type \( \nu \). Note that \( \mathcal{S} = \cup \mathcal{S}_\nu \) holds by definition. We then need to systematically determine the best clustering that satisfies all samples \( c \in \mathcal{S}_\nu \) being in the same type \( \nu \) within the clustering. The condition can be restated as an optimization problem by introducing a cost function “penalty” for the performance of clustering in this subsection.

Briefly speaking, for a given tree, we define \( \mathcal{O}_\nu \) as the offspring of set \( \mathcal{B}_n \), which includes \( \mathcal{S}_\nu \) and gives the minimum number of elements. Thus, we define the type penalty as the number of samples of types different from type \( \nu \) in \( \mathcal{O}_\nu \); in other words, \( \lambda_\nu(M_{\text{cut}}, p_G) := |\mathcal{O}_\nu| - |\mathcal{S}_\nu| \), where \( \mathcal{O}_\nu \) is determined by only a set of Hamming distances \( \{H(B^c, B^e)\}_{c, e \in \mathcal{S}} \). We define the global penalty \( \lambda(M_{\text{cut}}, p_G) := \sum_{\nu \in \mathcal{T}} \lambda_\nu(M_{\text{cut}}, p_G) \) as a “cost function” of optimization. See the next subsection for another expression of the global penalty and its properties. Thus, the optimization problem that we have to solve is find \((M^*, p^*_G)\) that minimizes the cost function \( \lambda(M_{\text{cut}}, p_G) \).
1. Global penalty as a cost function

In this section, we discuss the global penalty, a quantity that measures how the obtained hierarchical clustering differs from our knowledge of cell type classifications. We also give examples displaying the computation of the penalties and extreme situations that represent the theoretical bounds of the penalties. Note that these examples are just for explanation and not obtained from actual data.

In our settings, each sample is previously classified by types. Explicitly, set \( T \) consists of thirteen types:
\[
T = \{ B, \text{CD4}^+T, \text{CD8}^+T, \text{CLP}, \text{CMP}, \text{Ery}, \text{GMP}, \text{HSC}, \text{LMPP}, \text{MEP}, \text{Mono}, \text{MPP}, \text{NK} \}.
\]

For each type \( \nu \in T \), we denote the set of samples classified to type \( \nu \) as \( S_\nu \). This set could be empty, though it is not our case. For every pair \( \nu, \nu' \) of distinct types, there are no common elements in \( S_\nu \) and \( S_\nu' \), and the union of \( S_\nu \) among all types \( \nu \in T \) coincides with \( S \). Equivalently,
\[
S = \bigcup_{\nu \in T} S_\nu.
\]

For a given hierarchical clustering constructed in the manner of the previous section, the type penalty for type \( \nu \) is the quantity \( \lambda_\nu \) defined as follows. If \( S_\nu \) is empty, we set \( \lambda_\nu = 0 \). Otherwise, since the cluster grows step by step, there is the minimum \( \tau \) for \( 1 \leq \tau \leq 2N_s - 1 \) such that \( S_\nu \subset C_{\tau} \). We denote the minimal \( \tau \) by \( \tau(\nu) \). Thus, \( O_\nu \), which was defined above, is equal to \( C_{\tau(\nu)} \). Then we define \( \lambda_\nu \) as the number of elements in \( C_{\tau(\nu)} \) that are not of type \( \nu \). In other words, we set
\[
\lambda_\nu := |C_{\tau(\nu)}| - |S_\nu|.
\]

Since \( C_\tau \) includes all elements of type \( \nu \), we find \( \lambda_\nu \geq 0 \). Also since \( C_\tau \) is a subset of \( S \), we find \( \lambda_\nu \leq |S| - |S_\nu| \). Thus we have
\[
0 \leq \lambda_\nu \leq |S| - |S_\nu|.
\]

See Fig. S5 for an explanation of type penalties.

For a given hierarchical clustering, the global penalty \( \lambda \) is defined to be the total sum of type penalties,
\[
\lambda := \sum_{\nu \in T} \lambda_\nu.
\]
FIG. S6: Dendrograms with extreme penalties. Both of these dendrograms have six leaves (|S| = 6), that are classified into three types (in these examples, |T| = 3). (a) This example gives the lowest global penalty 0. (b) In this example, we have \( \tau(CD4^+T) = \tau(CD8^+T) = \tau(NK) = 11 \). Since the corresponding cluster \( C_{11} \) is the whole set \( S \), the local penalty is \( 6 - 2 = 4 \) for each type, and the global penalty is \( 4 \times 3 = 12 \). This result gives the upper bound \( (|T| - 1) \cdot |S| = (3 - 1) \cdot 6 = 12 \) in equation (1).

\[ 0 \leq \lambda \leq \sum_{\nu \in T} (|S| - |S_\nu|) = (|T| - 1) \cdot |S|. \]  

(1)

In our case, since \(|T| = 13\) and \(|S| = 77\), we have \( 0 \leq \lambda \leq (13 - 1) \cdot 77 = 924 \). Note that for a certain class of trees, these upper and lower bounds are not achieved. Fig. S6 displays examples of the upper and lower bounds. Further, we write \( \lambda(M_{cut},p_G) \) as \( \lambda \) to point out that \( \lambda \) depends on \( (M_{cut},p_G) \).

E. Computational cost of the algorithm

After obtaining data of the reads positions described by \( \hat{P}(R) \), we can produce a hierarchical clustering diagram. Here we consider the computation cost, except for the cost done by the MACS2 algorithm, based on the order of the sample number \( N_s \) and the DNA length \( L \).

Indeed, the dominant part of the computational cost is on the order of \( K N_s^2 L_0^2 \), which mainly comes from the calculation of the Hamming distance with cost \( O(N_s^2 L_0^2) \). Note that constant \( K \) depends on neither \( N_s \) nor \( L \). \( L_0 \) is \( O(M_{cut}) \ll L \).

Moreover, when we wish to add one sample to given samples for which we have already calculated the distance matrix, then the calculation cost is \( K'N_s L_0 \), where another constant \( K' \) also depends on neither \( N_s \) nor \( L \). Thus, the systematic algorithm in this paper is feasible in terms of increasing sample number.

Note that in the context of estimating the best optimization parameter \((M_{cut}^*,p_G^*)\), by using \( M_m \) different values for \( M_{cut} \) and \( M_p \) different values for \( p_G \), the computational cost becomes on the order of \( M_m M_p K N_s^2 L_0^2 \). However, these two numbers \((M_m,M_p)\) do not depend on \( N_s \) or \( L \) and can be adjusted according to the searching resolution of the optimization. Typically, we set \((M_m,M_p) \simeq (30, 10)\) in our optimization.

III. APPLICATION TO HEMATOPOIETIC CELLS

A. Determination of the best parameters for the best cell-type classification

First, we take into account all the peaks and check how the diagram and \( \lambda(M_{cut} = \infty, p_G) \) depend on \( p_G \), as shown in Fig. S7. Considering the accuracy of the parameter searching, we conclude that \( 1.5 \leq -\log_{10} p_G^* \leq 4 \). We also performed the above optimization for \( 1.5 \leq -\log_{10} p_G \leq 4 \), as shown in Fig. S8. We found that \((M_{cut}^*,p_G^*)\) is close to \((12000, 10^{-2})\) and obtained the best clustering, as shown in Fig. S9, for which the penalty \( \lambda \) is smallest in our
FIG. S7: Global penalty $\lambda$ dependent on $p_G$ ($M_{\text{cut}} = \infty$).

FIG. S8: (left) Global penalty $\lambda$ and (right) type penalty $\lambda_\nu$ for each type $\nu$ with $p_G = 10^{-2}$.

searching resolution. Hereafter, to investigate the property of the best clustering, we set $(M_{\text{cut}}^*, p_G^*)$ as $(12000, 10^{-2})$. Our searching resolution in terms of increasing $M_{\text{cut}}$ was typically 2000.

At the best parameters, $(M_{\text{cut}}^*, p_G^*) = (12000, 10^{-2})$, we computed distance matrix $H(S_\nu, S_\nu')$ for types and drew non-rooted hierarchical clustering with UPGMA and Neighbor joining (NJ) [19], as shown in Fig. S10. Though the details of the two clustering diagrams are different, both diagrams seem to be consistent to hematopoietic differentiation diagram in Ref. [1].

B. How to relate the best parameters to genomic context

In the following, we study what this best clustering tells us about the regulation of cell types at the chromatin level. The result of Section III A suggests that peaks of $\{g_k\}_{k=1}^{M_{\text{cut}}^*}$ with $M_{\text{cut}}^* = 12000$ include key regions for characterizing cell types. Therefore, we investigate whether these top 12000 peaks are related to the expression of a cell surface marker gene or functional genomic regions such as promoters and enhancers.

1. Irrelevance to genomic regions corresponding to a cell surface marker

When we evaluated the quality of the clustering results, we used the information of cell types for each sample. Since these cell samples were collected by flow cytometry using antibodies for the cell surface markers, we investigated
whether the genomic regions of a given cell surface marker overlap with the regions of the top 12000 peaks.

Let us introduce a set of genes $D := \{[\sigma_z, \epsilon_z]\}_{z=1}^{N_g}$, where $N_g$ is the number of genes in the human genome, $\sigma_z$ is 2kb upstream of gene $z$, and $\epsilon_z$ is the gene’s end position. Note that we obtained data $D$ for the location of genes on the DNA sequence through [20]. Then we define

$$M_{\leq z} := \{1 \leq k \leq M_{\text{cut}} \mid ||(\alpha_k, \beta_k) \cap [\sigma_z, \epsilon_z]|| > 0\},$$
$$M_{> z} := \{k > M_{\text{cut}} \mid ||(\alpha_k, \beta_k) \cap [\sigma_z, \epsilon_z]|| > 0\},$$

where $g_k = (\alpha_k, \beta_k)$ is the peak position and gene $z$ is a cell surface marker; in this case, $z = CD38$.

Finally, we computed the ratio $|M_{\leq z}|/|M_{> z}|$ as shown in Fig. S11. When we compare CD38-expressing cells with non-expressing cells, there is no clear difference.
There are no raw texts in the image.
FIG. S12: Percentage \((100 \times |E^z_\text{peak}|/1000)\) of individual biologically functional annotations \(z\) in every 1000 peaks along the peak ranking for Bcells (top), monocytes (middle), and HSCs (bottom). Only the functional annotations with percentages \(\geq 12\%\), \(z \in \{\text{FlankingActiveTSS}, \text{ActiveTSS}, \text{Enhancers}, \text{Quiescent Low}\}\), are shown.
C. Robustness of the clustering

In general, when one performs data clustering, the effect of variations of the clustering algorithms and the loss of data on the clustering output should be considered.

First, we consider the dependence of the clustering results on the variations of the clustering algorithms. Besides UPGMA which we used here, there are several hierarchical clustering methods including Ward method, WPGMA (Weighted Pair Group Method with Arithmetic Mean), UPGMC (Centroid Clustering or Unweighted Pair Group Method with Centroid Averaging), and WPGMC (Median Clustering or Weighted Pair Group Method with Centroid Averaging). Indeed, at least for \((M^{*}_{\text{cut}}, p^*_G) = (12000, 10^{-2})\), we have found that UPGMA and Ward method give the same global penalty score of 36. The penalty in the same condition was 338 for WPGMA, 303 for UPGMC, and 150 for WPGMC. These results suggest UPGMA was a better choice for our purpose.

Regarding the loss of data, let us consider to make new reads data \(\hat{R}\) from \(R\). Specifically, we set \(r\) with \(0 \leq r \leq 1\) as the probability of randomly removing \(\lceil rN_r \rceil\) reads from \(R\) with the uniform distribution, where \(\lceil \chi \rceil\) means the minimum integer larger than or equal to \(\chi\). Then we obtain \(\hat{R} = \{R'_i\}_{i=1}^{N_r\lfloor rN_r \rfloor}\), where \(R'_i\) is one of reads in \(R\). Then we obtained \(P(\hat{R})\) and computed \(\lambda\) for \((M^{*}_{\text{cut}}, p^*_G) = (12000, 10^{-2})\). As shown in Fig. S13, we found that when ratio \(r\) is increased from 0 to 0.5, \(\lambda\) is distributed between 36 and 127 and the average of \(\lambda\) is relatively constant. Note that \(r = 0\) gives \(\lambda = 36\) and the highest possible value of \(\lambda\) for 77 samples is 924. On the other hand, in the region \(r \geq 0.7\), \(\lambda\) increases obviously. Thus, we conclude that, for small \(r\), the average penalty tends to be close to that of \(r = 0\) although there are some variations depending on the samples.

Further, we investigated \(\lambda\) for higher values of \(M_{\text{cut}}\) to see robustness against random selections. Specifically, we investigated the behavior of \(\lambda\) by varying \(r\) for \(M_{\text{cut}} = 20000\) and 30000 with \(p^*_G = 10^{-2}\). The minimum value of \(\lambda\) is 39 for \(M_{\text{cut}} = 20000\) at \(r = 0.005\) and 0.3. For \(M_{\text{cut}} = 30000\), the minimum is 31 at \(r = 0.7\). These quick observations imply that the clustering method could be improved by searching a larger parameter space \((M_{\text{cut}}, p_G, r)\). Though this effort would require intensive computer power, it would be worth exploring for better clustering in the future.
FIG. S13: Global penalty $\lambda$ against ratio $r$ of reads removed randomly from the uniform distribution. $M_{\text{cut}} = 12000$ (top), $M_{\text{cut}} = 20000$ (middle), and $M_{\text{cut}} = 30000$ (bottom).
IV. APPLICATION TO LEUKEMIC CELLS

In this section, we apply the constructed algorithm with parameters \((M_\text{cut}^*, p_G^*) = (12000, 10^{-2})\) to analyze ATAC-seq data of acute myelogenous leukemia (AML) cells [1]. These AML samples are divided into three stages: Preleukemic HSCs (pHSCs), leukemia stem cells (LSCs) and leukemic blasts (Blast). We added AML ATAC-seq data to healthy hematopoietic ATAC-seq data and calculated the Hamming distance to see which cell type is closest to the added leukemia samples.

Specifically, we compute \(H(S_\nu, \{\ell\})\) as the average distance between cell type \(\nu \in \mathbb{T}\) and AML sample \(\ell\); in this case, sample \(\ell\) was extract from one patient (see TABLE. S2). We define the qth closest cell type of sample \(\ell\) as type \(\nu_q(\ell)\) to provide the qth minimum of \(H(S_\nu, \{\ell\})\) in terms of \(\nu\). Using this quantity, we define the rank gap between type \(T_0 \in \mathbb{T}\) and sample \(\ell\) as

\[G_{T_0, \ell} = q - 1,\]

such that \(T_0 = \nu_{q(\ell)}\). In particular, we call \(\nu_{q(\ell)}\) the closest type of sample \(\ell\). Note that rank gap \(G_{T_0, \ell} = 0\) holds when \(T_0 = \nu_{q(\ell)}\).

| Type of sample | Marker expression |
|----------------|-------------------|
| AML pHSC       | Lin-, CD34+, CD38-, TIM3-, CD99- [1] |
| AML LSC        | Lin-, CD34+, CD38-, TIM3+, CD99+ [1] |
| AML Blast      | Non-LSC; CD45-Intermediate, SSC-High [1] |

TABLE S2: How to extract leukemic cell samples

We computed \(\nu_{q(\ell)}\) as shown in TABLE 1 in the main text to compare with the “closest normal cell” \(T_0\) of the target leukemia sample, which was determined by a previous study [1]. We found rank gap \(G_{T_0, \ell} = 0\) for LSC, but \(G_{T_0, \ell}\) can be non-zero for pHSC and Blast depending on the samples. For all cases with non-zero \(G_{T_0, \ell}\), in the differentiation process considered in Ref. [1], the difference between two differenlly characterized types is only “one differentiation step”; that is, one characterization is one step either before or after another characterization in the differentiation process.

[1] Corces MR, Buenrostro JD, Greenside PG, Chan SM, Koenig JL, Snyder MP, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet. 2016 Oct;48(10):1193–1203. doi:10.1038/NG.3646.
[2] Laurenti E and Götgens B. From haematopoietic stem cells to complex differentiation landscapes. Nature. 2018;7689(533):418–426. doi:10.1038/nature25022.
[3] National Human Genome Research Institute: The human genome project. https://www.genome.gov/human-genome-project. Accessed 13 Dec 2019.
[4] Bradbury J. Human Epigenome ProjectUp and Running. PLoS Biol. 2003 Dec;1(3):e82. doi:10.1371/journal.pbio.0000082.
[5] Gasper-Maia A, Alajem A, Meshorer E, and Ramalho-Santos M. Open chromatin in pluripotency and reprogramming. Nat Rev Mol Cell Biol. 2011 Jan;12(1):36-47. doi:10.1038/nrm3036.
[6] Buenrostro JD, Giresi PG, Zaba LC, Chang HY, and Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013 Dec;10(12):1213–1218. doi:10.1038/nmeth.2688.
[7] Meyer CA and Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat Rev Genet. 2014 Nov;15(11):709–721. doi:10.1038/nrg3788.
[8] Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint, arXiv:1303.3997 [g-bio.GN].
[9] Broad Institute: The webpage of Picard Toolkit. http://broadinstitute.github.io/picard/. Accessed 13 Dec 2019.
[10] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug;25(16):2078–2079. doi:10.1093/bioinformatics/btp352.
[11] Meyer S U, Pfaffl M W, and Ulbrich S E. Normalization strategies for microRNA profiling experiments: A ‘normal’ way to a hidden layer of complexity? Biotechnol Lett. 2010;32:1777–1788. doi:10.1007/s10529-010-0380-z.
[12] Hicks S C and Irizarry R A. quantro: A data-driven approach to guide the choice of an appropriate normalization method. Genome Biol. 2015;16(16):1–8. doi:10.1186/s13059-015-0679-0.
[13] The ATAC-seq data in Ref. [1] is available through https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74912. Accessed 13 Dec 2019.
[14] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137. doi:10.1186/gb-2008-9-9-r137.
[15] Liu T: The webpage of MACS in Ref. [14]. https://github.com/taoliu/MACS. Accessed 13 Dec 2019.
[16] Denny S K, Yang D, Chuang C H, Brady J J, Lim J S S, Grüner B M, et al. Nifib Promotes Metastasis through a Widespread Increase in Chromatin Accessibility. Cell. 2016;2(166):328–342. doi:10.1016/j.cell.2016.05.052.
[17] Corces M R, Granja J M, Shams S, Louie B H, Seoane J A, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science. 2018;6413(362):1–58. doi:10.1126/science.aav1898.
[18] Müllner D. Modern hierarchical, agglomerative clustering algorithms. Preprint, arXiv:1109.2378 [stat:ML].
[19] Saitou N and Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987 Jul;4(4):406–425. doi:10.1093/oxfordjournals.molbev.a040454.
[20] University of California: The webpage for locations of genes in DNA sequence, http://genome.ucsc.edu/cgi-bin/hgTables Accessed 13 Dec 2019.
[21] Ernst J and Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012 Feb;9(3):215-216. doi:10.1038/nmeth.1906.
[22] The webpage for biological functions in DNA sequence in Ref. [21]. https://egg2.wustl.edu/roadmap/web_portal/, where one can obtain the corresponding data through “Core 15 state model of Chromatin state learning”. 
Appendix A: Information of samples

In TABLE S3, we show the information of samples listed in TABLE 1 of the main text.

| Sample       | SRR number          |
|--------------|---------------------|
| SU654-pHSC   | SRR2920595          |
| SU353-pHSC   | SRR2920571          |
| SU351-pHSC   | SRR2920568          |
| SU209-pHSC1  | SRR2920564          |
| SU209-pHSC2  | SRR2920562          |
| SU209-pHSC3  | SRR2920561          |
| SU070-pHSC1  | SRR2920557          |
| SU070-pHSC2  | SRR2920556          |
| SU048-pHSC   | SRR2920552          |
| SU583-pHSC1  | SRR2920588          |
| SU583-pHSC2  | SRR2920587          |
| SU575-pHSC   | SRR2920584          |
| SU501-pHSC   | SRR2920581          |
| SU496-pHSC   | SRR2920579          |
| SU484-pHSC   | SRR2920576          |
| SU444-pHSC   | SRR2920574          |
| SU654-LSC    | SRR2920594          |
| SU583-LSC    | SRR2920586          |
| SU575-LSC    | SRR2920583          |
| SU496-LSC    | SRR2920578          |
| SU444-LSC    | SRR2920573          |
| SU353-LSC    | SRR2920570          |
| SU209-LSC    | SRR2920559          |
| SU070-LSC    | SRR2920555          |
| SU654-Blast  | SRR2920593          |
| SU444-Blast  | SRR2920572          |
| SU353-Blast  | SRR29205690         |
| SU351-Blast  | SRR2920567          |
| SU209-Blast  | SRR2920558          |
| SU070-Blast1 | SRR2920554          |
| SU070-Blast2 | SRR2920553          |
| SU048-Blast1 | SRR2920551          |
| SU048-Blast2 | SRR2920550          |
| SU048-Blast3 | SRR2920549          |
| SU048-Blast4 | SRR2920548          |
| SU048-Blast5 | SRR2920547          |
| SU048-Blast6 | SRR2920546          |
| SU583-Blast  | SRR2920585          |
| SU575-Blast  | SRR2920582          |
| SU501-Blast  | SRR2920580          |
| SU496-Blast  | SRR2920577          |
| SU484-Blast  | SRR2920575          |

TABLE S3: Sample name and its corresponding SRR number [13].