Dissecting cell state transitions by causal inference

Longchen Xu[1,2,#], TingTing Cong[3,#], Hengyi Xu[4,#], Naixue Yang[5], Chen Tian[6], Sijie Yang[1], Ming Zhu[6], Rahul Sinha[7], Ryo Yamamoto[8], Wei Zhang[1,2], Jianwei Wang[3,*] and Xun Lan[1,6,9,*]

[1]Tsinghua-Peking Joint Center for Life Sciences, Tsinghua University, Beijing 100084, China
[2]School of Life Sciences, Tsinghua University, Beijing 100084, China
[3]School of Pharmaceutical Sciences, Tsinghua University, Beijing 100084, China
[4]Eight-Year MD Program, Peking Union Medical College, Beijing 100084, China
[5]State Key Laboratory of Primate Biomedical Research, Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming 650500, China
[6]Department of Basic Medical Science, School of Medicine, Tsinghua University, Beijing 100084, China
[7]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA
[8]Institute for the Advanced Study of Human Biology (ASHBi), Kyoto University, Yoshida-Konoe-cho, Kyoto, Japan
[9]MOE Key Laboratory of Bioinformatics, Tsinghua University, Beijing 100084, China
[#]These authors contributed equally to this work
[*]Correspondence: jianweiwang@mail.tsinghua.edu.cn (J.W.), xlan@tsinghua.edu.cn (X.L.)

Abstract

Trajectory inference methods are frequently used for cell fate analysis, however, most of them are similarity-based and lack an understanding of the causality underlying differentiation processes. Here, we present CIBER, a Causal Inference–Based framework for the Evaluation of feature effects and the Reconstruction of cellular differentiation networks. CIBER provides a novel paradigm for dissecting cell state transitions other than trajectory inference and differential analysis. It is a versatile framework that can be applied to various types of data, including transcriptomic, epigenomic and microarray data. It can identify both known and potential cell-lineage structures with minimal prior knowledge. By integrating the CIBER-learned network with structural causal model and applying *in silico* perturbation as inventions, we generated an effect matrix that quantifies the impact of different features on each differentiation branch. Using this effect matrix, CIBER can identify crucial features involved in haematopoiesis, even if these features show no significant difference in expression between lineages. Moreover, CIBER can predict novel regulation associations and provide insight into the potential mechanism underlying the influence of transcription factors on biological processes. To validate CIBER's capabilities, we conducted *in vivo* experiments on *Bcl11b,* a non-differentially expressed transcription factor identified by CIBER. Our results demonstrate that

dysfunction of *Bcl11b* leads to a bias towards myeloid lineage differentiation at the expense of lymphoid lineage, which is consistent with our predictions.

Introduction

Cellular differentiation is a fundamental process in biology that underlies the development of complex organisms. Advances in single-cell omics data have provided unprecedented opportunities for elucidating developmental pathways and reconstructing cell lineages. Computational trajectory inference methods have been developed to study cellular differentiation. In most of these methods, cells are ordered along deterministic[1–5] or probabilistic[6,7] trajectories based on similarities in their expression patterns, and a numeric value known as the pseudotime is assigned to each cell to indicate how far it has progressed along a dynamic process of interest. However, these similarity-based methods may not capture the causal nature of cellular differentiation. Another types of trajectory inference methods[8–12] are RNA velocity-based, which model the differentiation process with the spliced-to-unspliced mRNA ratio in the form of differential equations, thereby restoring cell temporal information, and can then be utilized for causal gene regulatory network reconstruction[13]. However, RNA velocity need to represent cell dynamics in the transcriptome space[14] and is not suitable for other types of data such as epigenomic or microarray data alone. Recently, new methods[15,16] have been developed to infer how cell identity is governed by the complex regulation of gene expression through *in silico* transcription factor (TF) perturbation. However, they aim at inferencing the change in cellular differentiation process but not the process itself and need the pseudotime provided by other methods. More general approaches are required to identify causality between cellular states with minimal prior knowledge, not only through transcriptomic profiling but also through other types of data.

Granger causality[17–19] and Bayesian network[20–23] (BN) have been effectively applied for causal inference in studying biological processes such as gene regulatory networks and protein-protein interactions. However, these methods are rarely applied in scenarios such as the reconstruction of cellular differentiation networks and the identification of fate-determining features. BNs offer a computationally efficient approach for inferring causality among a set of variables and have been applied to infer cellular networks[20], protein-protein interactions[21], protein-signalling networks[22], gene regulatory networks[23,24], and long-range intrachromosomal interactions[25]. However, inferring feasible BNs from data is challenging because the structure learning of BNs is sensitive to parameters and requires extensive prior knowledge to obtain expected structures.

Here, we present CIBER, a novel causal inference–based framework, for reconstructing cellular differentiation networks and identifying fate-determining features. CIBER provides a hybrid BN structure learning algorithm to detect information flow among cell states to learn a directed acyclic graph (DAG) representing the differentiation network. In addition to single-cell omics data which is critical for existing trajectory inference methods, CIBER can be applied to microarray or bulk sequencing data in which the cell states are discrete. We adopt the structural causal model[26,27] to the learned DAG, utilize *in silico* perturbation as interventions for causal inference, and generate an effect matrix that quantifies the impacts of different features on each branch of the cellular differentiation network. Furthermore, we introduce differentiation driver feature (DDF) analysis to identify important features in the cellular differentiation process. DDF analysis can identify features crucial to the entire differentiation process or a specific lineage while

not differentially expressed among lineages and can complement the regular differential feature analysis to help us understand the biological processes comprehensively.

We applied CIBER framework to study haematopoiesis using various types of omics or microarray data and obtained expected cell differentiation networks. The application on a mouse microarray dataset showed that CIBER can identify both expected and potential branches that disagreed with the conventional hierarchy while reported in previous studies. In this study, we use regular differential feature (hereafter named RDF) analysis to denote the analysis that identifies differentially expressed features directly from the cell-by-feature matrices representing gene expression or chromatin accessibility. In contrast to RDF, DDF analysis is performed by identifying important features based on the edge-by-feature effect matrices inferred via CIBER from the cell-by-feature matrices. DDF analysis enriches the techniques for identifying important features related to cellular differentiation process and can complement RDF analysis to help us understand biological processes more comprehensively. Through DDF analysis, CIBER identifies fate-determining features that are not deemed statistically significant in RDF analysis, can predict novel regulation associations and provide insight into the potential mechanism underlying the influence of TFs on biological processes. We conducted experiments on *Bcl11b*, a non-differentially expressed TF identified by CIBER, and validated CIBER's prediction that it has a lymphoid-biased effect on the differentiation of haematopoietic stem cells.

Results

CIBER presents a novel framework for learning robust Bayesian network structures with minimal prior knowledge. Bayesian networks are directed acyclic graphical models that represent probabilistic dependencies between variables via edges and nodes[28]. The existing Bayesian network structure learning algorithms can be divided into three categories: constraint-based algorithms, score-based algorithms and hybrid algorithms[29,30] (see Methods for details). One major issue in Bayesian network structure learning is that the obtained graph is highly sensitive to parameters and the observed data, and small changes in input may result in considerably different outputs. To obtain a well-constructed network, extensive prior knowledge is usually required to adjust the learned structures.

To overcome these limitations, we developed CIBER framework to generate robust Bayesian network structures given minimal prior knowledge, such that the proposed method can be generalized to investigate understudied biological processes. CIBER presents a new hybrid algorithm that uses a mixed strategy of resampling over single cells and combining the Bayesian networks learned within each resampling (Fig. 1). Stratified resampling is used to select random subsets of cells to construct the network, resulting in a set of networks with various structures. Specifically, for each subset of cells, centroids of different cell types are utilized to learn a Bayesian network, with each centroid representing the mean expression level of its corresponding cell type. The centroids are connected to generate an undirected skeleton according to Spearman's correlation or the mutual information between them, with each edge representing the correlation between the two cell types it connects. Various algorithms for correlation estimation can be chosen, such as CMI2NI[31] and SPACE[32], under the support of GeNeCK[33], an R framework for constructing gene regulatory networks. Based on the undirected skeleton, we utilized the

hill climbing algorithm through the bnlearn[34] package to search for the directed acyclic graph (DAG) with the highest Bayesian information criterion (Methods).

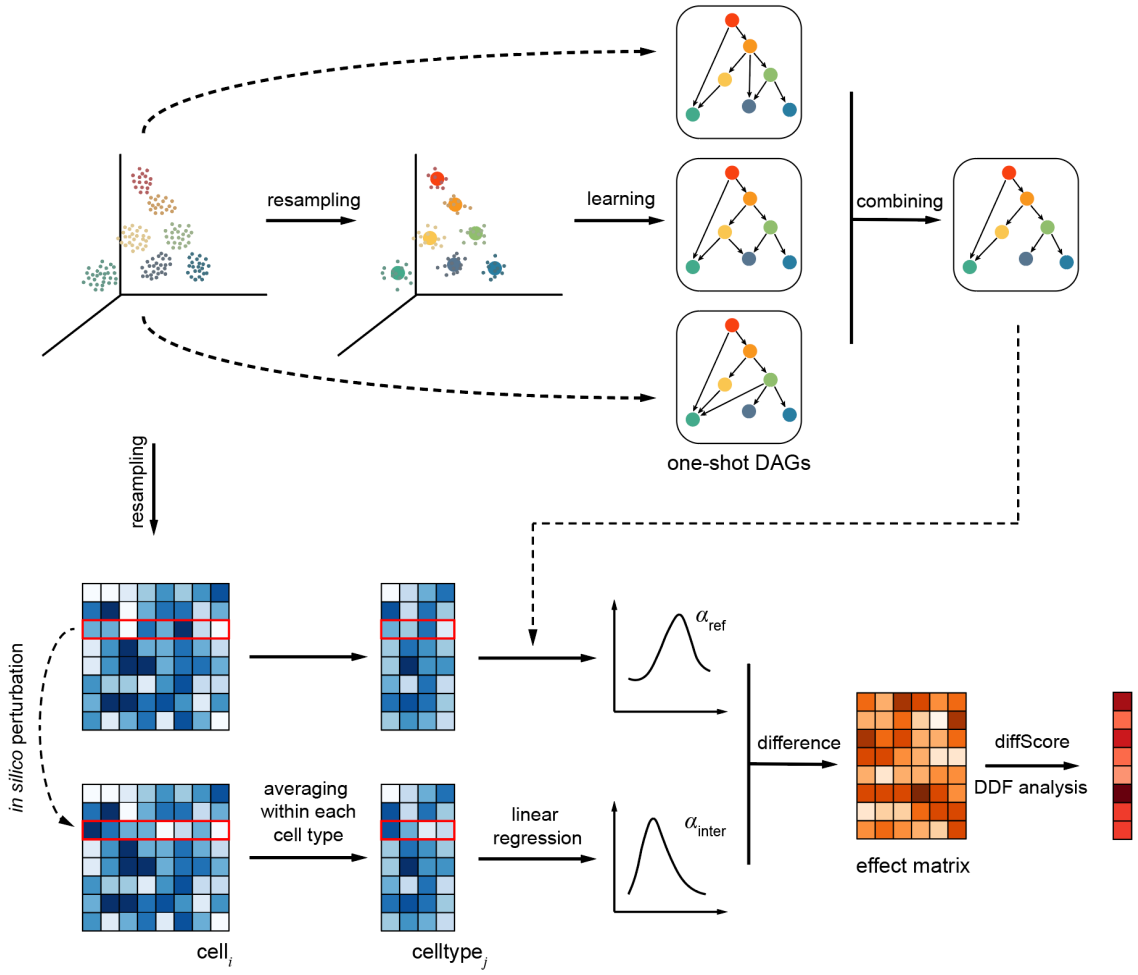


Fig. 1 | Simplified schematic of CIBER framework. The upper part shows the pipeline of the hybrid Bayesian network structure learning algorithm starting from cell resampling. Random subsets of cells are selected, and various DAGs are learned to represent the causation between the centroids of different cell types. Each DAG obtained within one resampling is referred to as a one-shot DAG for convenience in this study. Then, a combination strategy is applied over the one-shot DAGs to obtain the final Bayesian network. The bottom part shows the pipeline to generate effect matrix and to identify differentiation driving-genes. From the perspective of the structure causal model, a series of linear regressions are performed on the learned Bayesian network, with each child node taken as a dependent variable and its parent nodes taken as independent variables. A mixed strategy of resampling and interventions of *in silico* perturbation on a specific feature is utilized to obtain a pair of distributions for each edge, representing the reference and the intervened linear regression coefficients. The effect matrix is defined to quantify the difference between the two distributions and each element in the matrix reflects the effect of a feature on a differentiation branch. DDF analysis is performed on the effect matrix to identify fate-determining features.

The learned DAG is sensitive to the threshold used for evaluating correlations during the construction the undirected skeleton. In this study, we employed a range of sequential

correlation thresholds, such as from 0 to 1 in steps of 0.01, to obtain a set of DAGs varying from dense to sparse for each resampling of cells. Approximate upper and lower boundaries for the number of edges in the network can be assigned according to the expected complexity of the network. DAGs with the number of edges within the specified boundaries are selected and combined to generate a directed graph, where each edge is weighted according to its frequency of appearance. The weight of each edge reflects its robustness, which in turn signifies the confidence level of the causation between the nodes it connects. However, the graph may not necessarily be acyclic, necessitating trimming and binarization to produce an unweighted DAG that retains the most robust structure. We refer to this DAG as the "one-shot DAG" since it is obtained within a single resampling. This process is then repeated several times to get a set of one-shot DAGs, over which a second combination is applied to generate a final DAG that captures the most consistent causation between variables (see Methods for detailed description). Our approach is applicable not only to single-cell omics data but also to microarray and bulk sequencing data, omitting the resampling step due to limited replicates.

CIBER captures known differentiation structures in various datasets. We tested whether CIBER could obtain well-established cell lineage structures by applying it to study haematopoiesis using various types of data, including mouse microarray[35–37], mouse bone marrow scRNA-seq[38], human bone marrow scRNA-seq[39] and human bone marrow scATAC-seq datasets[40]. The mouse microarray dataset contains comprehensive blood cell types, from which we selected a subset of representative types of progenitors to reconstruct the early haematopoietic hierarchy. The subset includes 11 cell types — haematopoietic stem cell (HSC), multipotent progenitor subset A (MPPa), multipotent progenitor subset B (MPPb), common myeloid progenitor (CMP), common lymphoid progenitor (CLP), granulocyte-macrophage progenitor (GMP), megakaryocyte-erythroid progenitor (MEP), megakaryocyte progenitor (MkP), precolony-forming unit erythroid (pCFU-E), B-cell-biased lymphoid progenitor (BLP) and double negative 1 thymocytes (DN1) — each containing 3 or 4 biological replicates. In this microarray dataset, the expression levels of 21,678 genes are detected by 45,101 probes[35–37]. If a gene is simultaneously detected by several probes, we set its expression level to the median among all the probes. Because of the stable expression levels over the limited size of replicates, we omitted resampling and performed the analysis on a set of fixed points that are the centroids obtained by averaging the expression level over biological replicates for each cell type.

The results revealed that CIBER can extract robust lineage structures and provide information on potential differentiation branches (Fig. 2b). Compared to the results obtained with a typical hybrid algorithm consisting of mutual information and hill climbing (hereafter named MI-hc), the CIBER-learned network was more consistent to the conventional haematopoietic hierarchy and more accurately captured the bifurcation structure (Fig. 2b, c). In these figures, the edges contradicting prior knowledge are coloured in red. The CIBER-learned structure is robust without the need to manually set the exact correlation threshold, unlike MI-hc, which is sensitive to parameters, and small changes in the correlation threshold can result in considerably different structures (Extended Data Fig. 1). In addition to the well-known branches of the HSC→MPPa/MPPb→lymphoid/myeloid lineages and those inside the lymphoid or myeloid lineage, CIBER identified three unconventional interlineage edges that were reported to be possible in previous studies, CLP→GMP[41] and CMP→BLP/DN1[42,43].

We also applied CIBER to single-cell omics datasets, following the entire pipeline, including the resampling procedure. The structures learned by CIBER again identified larger numbers and higher percentages of conventional edges compared to the results of the typical hybrid algorithm MI-hc (Fig. 2d-f, Extended Data Fig. 2). Furthermore, the cell types in CIBER-learned structures were in relative positions consistent with the low-dimensional landscapes of diffusion maps[38], SPRING[39] or PCA[40] obtained in previous studies.

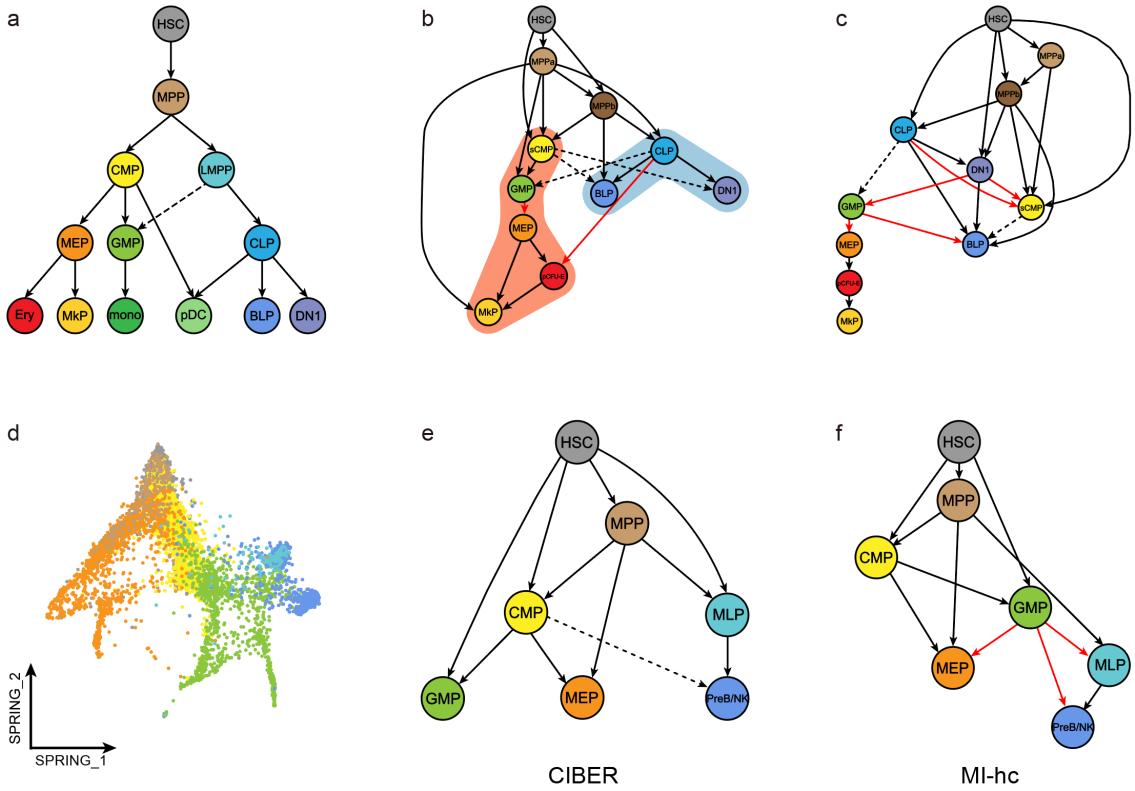


Fig. 2 | CIBER captures conventional differentiation structures in various datasets. a, Conventional haematopoietic hierarchy. **b**, CIBER result of the mouse microarray dataset, with the blue/red-shaded part representing the lymphoid/myeloid lineage. **c**, Best-tuned structure learned with MI-hc, a typical hybrid algorithm consisting of mutual information and hill climbing, using the same dataset as **b**; the structures learned under other correlation thresholds are shown in Extended Data Fig. 1. The dashed edges represent unconventional interlineage branches that were reported to be possible in previous studies, while the red edges represent less credible branches. **b** and **c** show that the CIBER result is consistent with conventional knowledge with a larger number and a higher percentage of edges and captures the bifurcation structure more accurately than MI-hc. **d-f**, Low-dimensional landscapes[39], the CIBER and MI-hc results of the human scRNA-seq dataset. Results for more datasets are shown in Extended Data Fig. 2.

CIBER quantifies the effects of features on the cellular differentiation hierarchy. CIBER enables quantification of the effects of features on the cellular differentiation hierarchy by utilizing the structural causal model (SCM), a conventional causal inference framework that combines probabilistic graphical models with the notion of interventions[26,27]. Similar to Bayesian networks, SCM considers a set of associated

observables in the form of a directed acyclic graph while viewing each node as a function of its parents in the graph. A natural restriction of SCM is the assumption of a Gaussian linear additive noise model, which limits the functions connecting each node to its parents to the form of linear regression (see Methods). In the SCM framework, interventions can be performed by changing the form of particular functions or the distribution of specific nodes to infer the effect of causality.

Here, we introduced the concept of SCM into the directed acyclic graph learned by CIBER and performed causal inference by applying *in silico* perturbation to the expression of specific features as interventions (Fig. 1). Specifically, we performed feature permutation on single-cell omics data and feature deletion on microarray or bulk sequencing data. We then recorded the coefficient associated with each edge in the Gaussian linear additive SCM, with and without interventions performed on specific features, and defined an effect matrix to measure the change in the coefficient due to the interventions. The effect matrix quantifies the impacts of the features on each edge in the cellular differentiation network, with positive values implying a possible promoting function on the corresponding differentiation branch and negative values implying a possible inhibiting function.

Differentiation driver feature (DDF) analysis identifies features determining cell fate. We present a novel method, called DDF analysis, which distinguishes itself from regular differential feature (RDF) analysis by identifying important features that are crucial to the cellular differentiation process via effect matrices obtained by *in silico* perturbation. The method quantifies the impact of individual features on the cellular differentiation network or specific lineages through diffScores (see Methods). The diffScore of a feature on the entire network is defined by summarizing the absolute values related to this feature in the effect matrix, while the diffScore on a specific lineage is defined as the sum of the absolute values related to the edges pointing to cell types in the lineage subtracting the original values of the edges pointing from the lineage to other lineages. Our results show that DDF analysis can identify crucial features which are not differentially expressed between lineages, and provides a paradigm to systematically investigate a feature's impact on the cellular differentiation process even in the absence of some important cell types. Additionally, DDF analysis can be combined with RDF analysis to provide a comprehensive understanding of biological processes due to the heterogeneous information captured by the two methods.

To evaluate the effectiveness of DDF analysis, we applied it to the mouse bone marrow scRNA-seq dataset and compared the results with that of the RDF analysis obtained by ranking genes according to the p value of Wilcoxon's signed-rank tests. To compare the results of the two analyses, we selected the same number of top genes on the two lists, and conducted a literature search on the top-ranked genes (Fig. 3a-c). Results showed that the RDF and DDF analyses both captured a part of the genes that have been proven to be important in haematopoiesis: a few important fate-determining genes identified by RDF analysis, such as *Gata2*[44], were not captured by DDF analysis; on the other hand, DDF analysis identified important genes that were not differentially expressed between the lymphoid and myeloid cell types, such as *Mpl*[45], *Ccl9*[46] and *Elane*[47]; nevertheless, some genes, such as *Flt3*[48], *Cd34*[49] and *Dntt*[50], were identified by both analyses.

We calculated Spearman's correlation between the two gene lists and found that they were weak correlated (Fig. 3d), and further evaluated their behaviours on enriching important genes related to haematopoiesis or haematopoietic differentiation. Specifically, we calculated GSEA normalized enrichment scores (NES) for gene lists with respect to

all the GO and KEGG terms related to haemopoiesis or haematopoietic differentiation (Fig. 3e). GSEA NES results showed that DDF surpassed RDF on almost all these pathways. Note that the dataset did not contain any lymphoid-lineage cell types other than lymphomyeloid-primed progenitor (LMPP), an early lymphoid-biased progenitor. This illustrates that DDF analysis performs well even in the absence of important cell types, due potentially to the fact that it examines the entire cellular differentiation network rather than being restricted to the difference between subgroups of cells, as RDF analysis did.

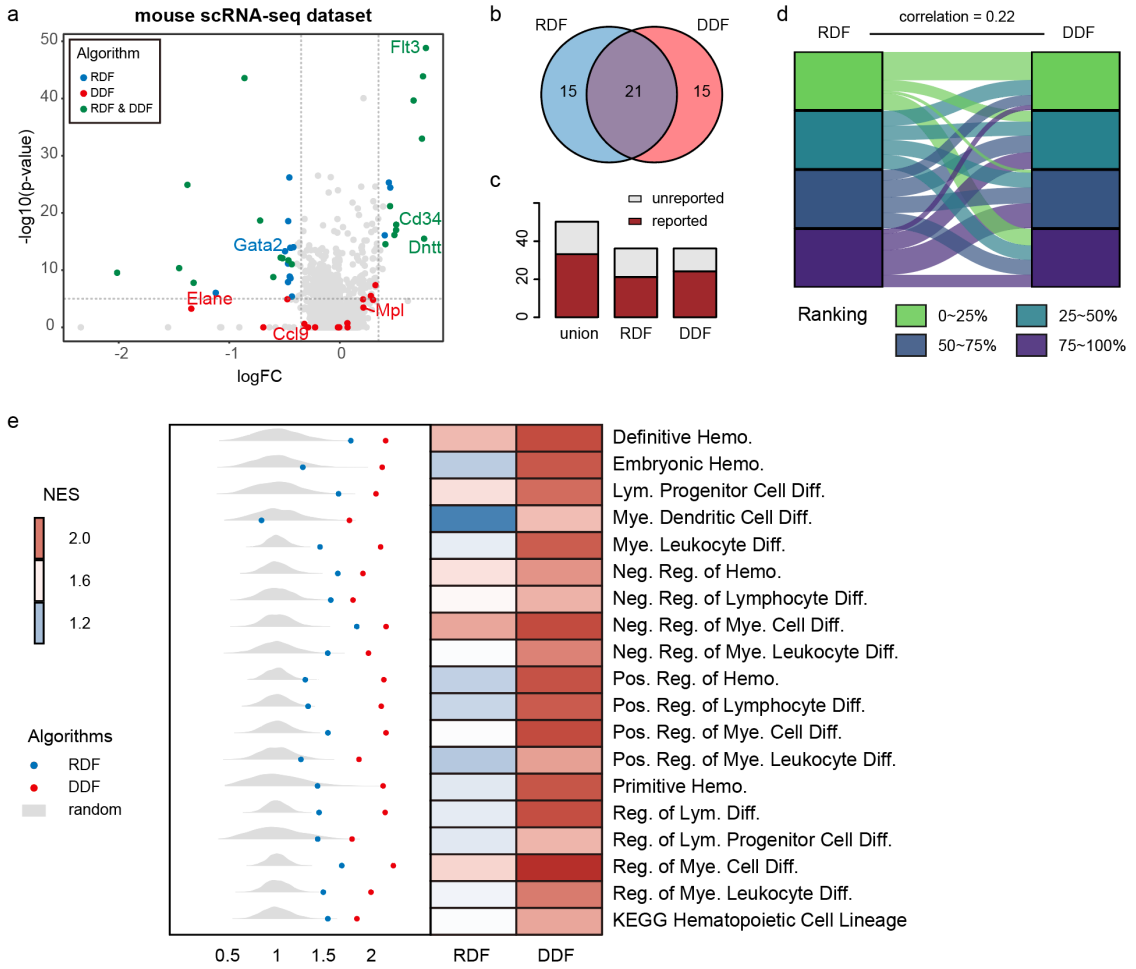


Fig. 3 | DDF identifies fate-determining features. The plots show the results on RDF and DDF gene lists obtained with the mouse scRNA-seq dataset. **a**, Features are allocated by log fold change and log p value obtained in the RDF analysis, from which the top differentially expressed genes are coloured, as well as the same number of DDF-selected features important to the entire network. **b**, The number of shared top features between the two feature sets. **c**, Literature search result shows that the DDF and RDF analyses identify different important features, in which the red parts represent the reported features. **d**, Correlation between the ranked lists shows that the DDF analysis captures different information from the RDF analysis. The tiles represent the top 25%, 25~50%, 50~75% and bottom 25% features in RDF or DDF ranking lists. **e**, GSEA normalized enrichment scores (NES) with terms related to haematopoiesis or haematopoietic differentiation in the GO and KEGG databases. DDF list obtains higher NES than RDF on all the terms. Dots in the left part represent the NES of RDF and DDF ranking lists, while the distributions show the background NES given randomly ranked gene lists. (Hemo., Hemopoiesis; Diff.,

298 Differentiation; Lym., Lymphoid; Mye., Myeloid; Reg., Regulation; Pos., Positive; Neg.,
299 Negative.)

300 Next, we applied the same analyses to the human scRNA-seq dataset and mouse
301 microarray dataset. Similar results were obtained for the human scRNA-seq dataset to the
302 results of mouse scRNA-seq dataset described above (Extended Data Fig. 3). The
303 microarray dataset offered us an opportunity to study the behaviour of DDF analysis in
304 scenarios where most cell types are available and are therefore friendly to RDF analysis.
305 The result showed the RDF list performed slightly better than the DDF list in 12 out of
306 22 terms. However, the literature search still showed that DDF analysis identified
307 important genes that are not differentially expressed, demonstrating that DDF analysis
308 can complement RDF analysis to better identify genes crucial to cell differentiation
309 (Extended Data Fig. 4).
310 **CIBER identifies crucial TFs regulating biological processes**. We then applied CIBER
311 and DDF analysis to a human bone marrow scATAC-seq dataset[40], and identified TFs
312 crucial to haematopoiesis based on deviation score matrix calculated with chromVAR[51].
313 RDF analysis was also performed to identify the TFs exhibiting significant differences in
314 deviation score between lymphoid and myeloid cell types. The top TFs identified by DDF
315 alone include *SPI1*, *SPIB* and *CEBPB*; whereas TFs like *TCF3*, *LYL1* and *EBF1* were
316 exclusively identified by RDF analysis. Both analyses uncovered TFs such as *GATA1*,
317 *GATA2*, *GATA3*, *ID3* and *LMO2* (Fig. 4c).
318 With lineage-specific diffScores, DDF analysis can further predict the potential effects
319 of each TF on promoting or inhibiting specific lineages (Fig. 4d). In addition to the
320 previously reported TFs, we noted that DDF analysis exclusively identified *BCL11B*, and
321 predicted that it may have similar impacts on haemopoiesis of promoting the lymphoid
322 lineage while inhibiting the myeloid lineage differentiation with previous reported *IRF8*[52]
323 and *BCL11A*[53,54].
324 We further examined if the target genes of the TFs identified in the scATAC-seq
325 dataset were also enriched in the DDF or RDF analysis of an independent human bone
326 marrow scRNA-seq dataset since TFs regulate cellular function by orchestrating the
327 transcription of downstream target genes. To achieve this, we collected target genes for
328 15 reported TFs[40] and *BCL11B* from 5 databases (TRRUST[55], JASPAR[56], ChEA[57],
329 Motifmap[58] and ENCODE[59]) and ranked the target genes via DDF or RDF analysis
330 performed on the scRNA-seq dataset. To measure the enrichment of TF-targeted genes,
331 we calculated two types of metrics, the normalized gene set enrichment score (NES) and
332 the area under the receiver operating characteristic curve (AUC) (Fig. 4e, Methods).
333 When calculating AUC, the receiver operating characteristic (ROC) curve was drawn by
334 assigning TF-targeted genes as positive and all other genes as negative samples.
335 Background distributions of NES and AUC were obtained with 1,000 randomly shuffled
336 ranking lists of all the genes and were used for calculating p value. Results show that
337 under both metrics, DDF list surpasses RDF list on all gene sets targeted by different TFs
338 (Fig. 4e, Extended Data Fig. 5). Therefore, CIBER can predict novel regulation
339 associations and provide insight into the potential mechanism underlying the influence of
340 TFs on biological processes. For example, for the *GATA1*-targeted gene set, DDF list
341 scores higher than RDF list, and both lists significantly surpass the background; in the
342 case of *BCL11B*, only the DDF list significantly surpasses the background (DDF NES p
343 value=0.009, RDF NES p value=0.103, DDF AUC p value=0.018, RDF p value=0.194)
344 (Fig. 4f).

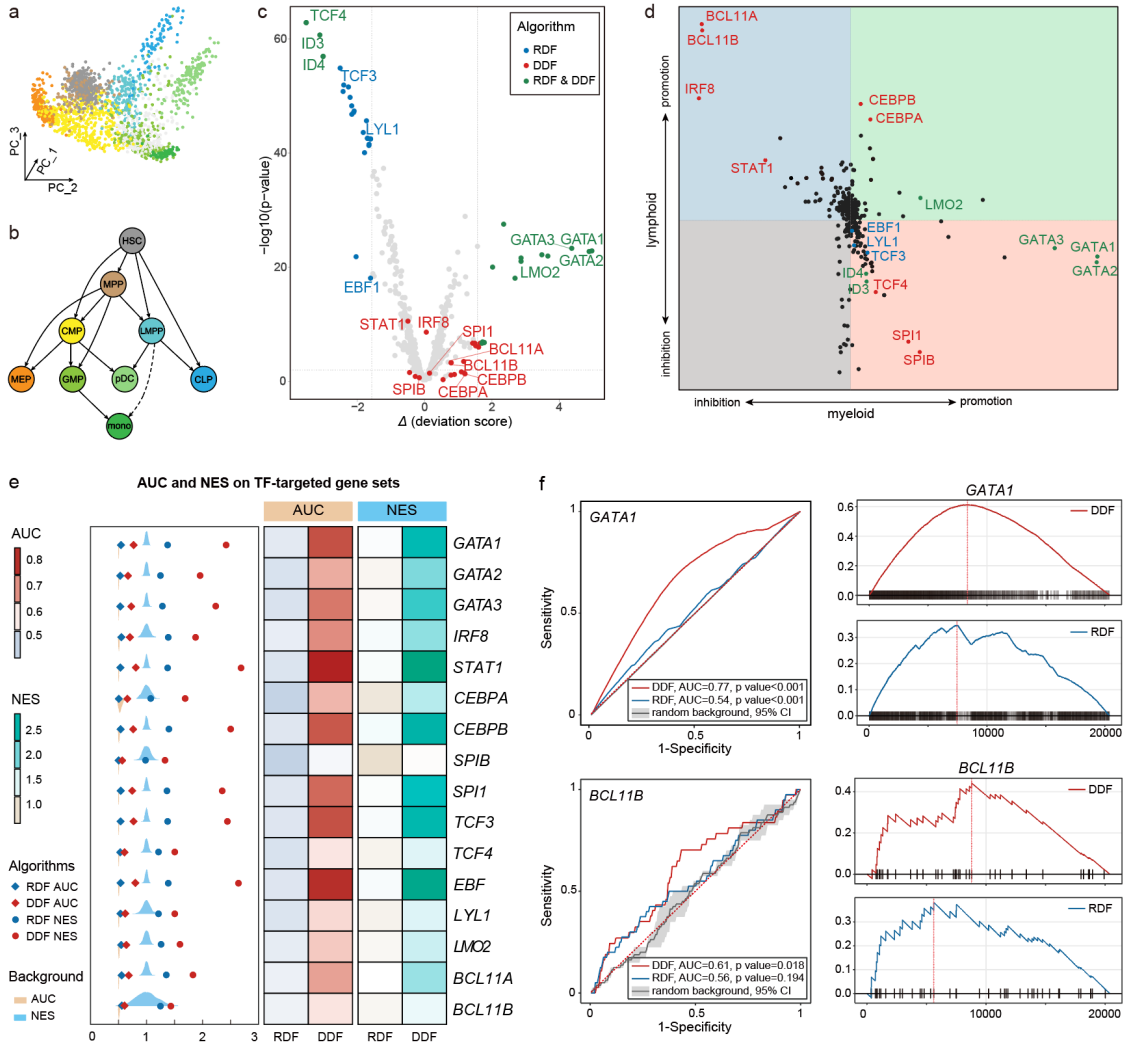


Fig. 4 | CIBER identifies driver TFs of biological processes. a, PCA reduction of a human bone marrow scATAC-seq dataset[40]. **b**, Differentiation hierarchy learned with CIBER. **c**, Top TFs identified by RDF and DDF analyses. Among them, several reported important TFs[40] are annotated. **d**, Lineage-specific diffScores quantify the effects of each TF on haematopoiesis under a higher resolution of promoting or suppressing specific lineages. **e**, Comparison of rankings obtained with DDF and RDF analyses using ROC or GSEA analyses applied with the set of genes regulated by each TF. AUC and GSEA NES results show DDF surpasses RDF on each gene set and significantly surpasses random background; background is set as the distribution of 1,000 random gene rankings. **f**, AUC and NES Results on *GATA1*-targeted and *BCL11B*-targeted gene sets respectively. For *GATA1*-targeted gene set, both DDF and RDF significantly surpass random background, and DDF scores higher; for *BCL11B*-targeted gene set, DDF significantly surpasses random background, while no statistical difference exists between RDF and the background (DDF NES p value=0.009, RDF NES p value=0.103, DDF AUC p value=0.018, RDF p value=0.194). Thus, compared with regular differential analysis, CIBER can better identify important TFs and enrich reported regulated genes, providing insight into potential regulated genes through which a TF influences the process of haematopoiesis.

364 **In vivo experiments validate *BCL11B* as a crucial TF to fate determination in**
365 **haematopoiesis.** The above mentioned results indicate that the TF *Bcl11b* may influence
366 the differentiation of haematopoietic stem cells. To validate this, we constructed *Bcl11b*-
367 deficient mice (*Bcl11b*$^{flox/flox}$, Vav-iCre^{+}), hereafter named *Bcl11b*$^{-/-}$ mice, by specifically
368 knocking out *Bcl11b* in the blood cells. Results showed that the lifespan of *Bcl11b*$^{-/-}$ mice
369 was significantly shorter than that of wild-type (WT) mice in the same cohort, indicating
370 that *Bcl11b* is essential for the maintenance of hematopoietic system (Fig. 5a). To
371 examine the haematopoietic dynamicity of *Bcl11b*$^{-/-}$ mice, we treated *Bcl11b*$^{-/-}$ and WT
372 mice with 5-fluorouracil (5-FU) and analyzed the lineage composition in peripheral blood
373 (Fig. 5b). The result showed that myeloid cells were significantly reduced in both groups
374 five days after 5-FU injection, indicating that they are sensitive to DNA damage challenge
375 (Fig. 5c). Notably, T cells were almost undetectable in *Bcl11b*$^{-/-}$ mice, while remaining
376 static in the WT group five days after 5-FU injection compared to non-treated mice,
377 suggesting that the *Bcl11b* plays an essential role in maintaining T cell in response to
378 DNA damage. Between days 10 to 20, we observed a significant decrease in the
379 percentage of both T and B lineages compared to WT mice, indicating that *Bcl11b* is
380 crucial for lymphoid differentiation, consistent with our prediction (Fig. 4b).
381 To assess the impact of *Bcl11b*-deficiency on the long-term haematopoietic
382 reconstitution capacity and differentiation of HSCs, we conducted competitive
383 transplantation assays by injecting 50 freshly isolated HSCs either from *Bcl11b*$^{-/-}$ or WT
384 mice into lethally irradiated recipients together with 3×10^{5} total bone marrow cells as
385 competitor (Fig. 5d). Peripheral blood chimerism was monitored monthly until the 4th
386 month. Our results demonstrate that the targeted dysfunction of *Bcl11b* severely impairs
387 the reconstitution capacity of HSCs (Fig. 5e). Notably, the differentiation potential of
388 *Bcl11b*$^{-/-}$ HSCs to T lineage is severely compromised from the 2nd month after
389 transplantation. By the end of the 4th month, *Bcl11b*$^{-/-}$ HSCs exhibit differentiation bias
390 towards the myeloid lineage at the expense of T lineage (Fig. 5f), which is consistent with
391 our prediction (Fig. 4b).

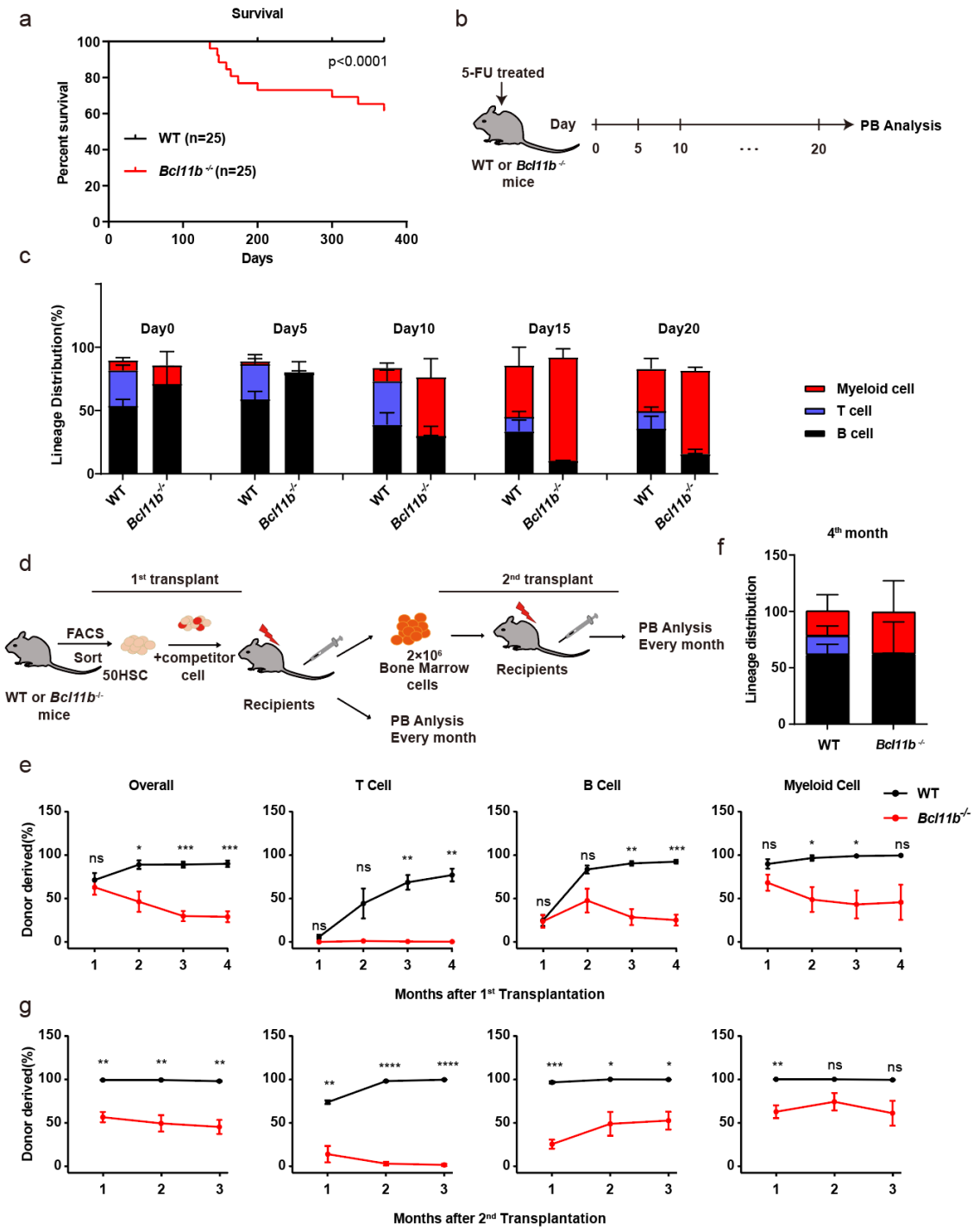


Fig. 5 | Experimental validation of CIBER predictions for *Bcl11b*. a, Results of survival experiments of WT and *Bcl11b*$^{-/-}$ mice in their native state show the survival capacity of *Bcl11b*$^{-/-}$ mice is hampered. Experiments (**b**) and Lineage distribution (**c**) of reconstituted B cells, T cells and myeloid cells in peripheral blood of WT and *Bcl11b*$^{-/-}$ mice after 5-FU treatment. **d**, Scheme of the transplantation strategy for *Bcl11b*$^{-/-}$ mice. Fifty HSCs freshly isolated from 2-month-old WT or *Bcl11b*$^{-/-}$ mice were transplanted into lethally irradiated recipients along with 3×10^5 competitor cells. For the 2nd transplants, 2×10^6 whole bone marrow cells were transferred from primary recipient mice. Chimera in peripheral blood was examined every month for three months. **e,** Contributions of donor cells to peripheral blood

402 overall (CD45.2$^+$), B cells (B220$^+$), T cells (CD3$^+$) and myeloid cells (Mac-1$^+$) every month
403 after 1st transplantation. (n=4-5 mice per group. Data are shown as the mean ± SD.) **f**,
404 Lineage distribution of B cells, T cells, and myeloid cells among donor-derived cells in
405 peripheral blood four months after transplantation. (n=4-5 mice per group. Data are shown
406 as the mean ± SD.). **g**, Contributions of donor cells to peripheral blood overall (CD45.2$^+$), B
407 cells (B220$^+$), T cells (CD3$^+$) and myeloid cells (Mac-1$^+$) every month after 2nd transplantation.
408 (n=4-5 mice per group. Data are shown as the mean ± SD.)

409 To further investigate the influence of *Bcl11b* dysfunction on the self-renewal capacity
410 of HSCs, we performed secondary transplantation assays (Fig. 5d). Our findings
411 demonstrate that the chimera of *Bcl11b*$^{-/-}$-derived cells in the peripheral blood of
412 recipients is significantly lower than that of the control group, including B, T and myeloid
413 lineages (Fig. 5g), suggesting that the self-renewal capacity of *Bcl11b*$^{-/-}$ HSCs is impaired.
414 In summary, our results indicate that *Bcl11b* plays a critical role in promoting T
415 lymphocyte differentiation and that its dysfunction impairs the reconstitution and self-
416 renewal capacity of HSCs, resulting in a differentiation bias towards the myeloid lineage.

417 Discussion

418 Recent advancements in single-cell omics data have allowed researchers to study
419 developmental pathways and lineage relationships with unprecedented resolution.
420 However, most trajectory inference methods are based on expression pattern similarities
421 and provide cell fate probabilities, which may not accurately reflect the complex
422 causation-driven phenomena of cellular differentiation. In this study, we present CIBER
423 framework based on causal inference to reconstruct cellular differentiation networks and
424 identify fate-determining features, which provides a novel paradigm for dissecting cell
425 state transitions other than trajectory inference and differential analysis.
426 Compared to trajectory inference methods, CIBER has several advantages. It is
427 applicable under a broader range of conditions, including datasets with limited cells or
428 cell types, and even those obtained by microarray or bulk sequencing techniques.
429 Furthermore, unlike most similarity-based methods, CIBER provides a causation-based
430 perspective on the cellular differentiation process.
431 CIBER providess a robust Bayesian network structure learning strategy to obtain
432 feasible structures with minimal prior knowledge. The applications of CIBER show that
433 it is suitable to different types of datasets and that it may help in identifying potential
434 differentiation branches. Based on the concept of the structural causal model and
435 interventions of *in silico* feature perturbation, CIBER builds an effect matrix quantifying
436 the impacts of features on differentiation branches and can be used for identifying fate-
437 determining features through DDF analysis. DDF analysis can identify features crucial to
438 the entire differentiation process or a specific lineage while not differentially expressed
439 among lineages and can complement the regular differential feature analysis to help us
440 understand the biological processes comprehensively. Besides, DDF analysis is effective
441 in scenarios even in the absence of important cell types.
442 Despite its advantages, CIBER has some limitations. CIBER may be time-consuming,
443 and should take the centroid of cell types or clusters as nodes rather than cells, and it may
444 not be suitable for extremely complex networks with too many nodes and edges. We
445 found that CIBER was most effective for hierarchical structures with approximately 5
446 layers and less than 20 nodes.

In conclusion, the CIBER framework provides a causal perspective for the reconstruction of cellular differentiation networks and the evaluation of how features influence differentiation branches. DDF analysis enriches existing techniques for identifying important features and can be combined with differential analysis to help us understand biological processes more comprehensively. Overall, CIBER has the potential to revolutionize the study of cellular differentiation and provides insights into the underlying mechanisms of developmental pathways.

References

1. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* **13**, 845–848 (2016).
2. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods* **14**, 979–982 (2017).
3. Setty, M. *et al.* Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol* **34**, 637–645 (2016).
4. Welch, J. D., Hartemink, A. J. & Prins, J. F. SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biology* **17**, 106 (2016).
5. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* **32**, 381–386 (2014).
6. Schiebinger, G. *et al.* Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* **176**, 928-943.e22 (2019).
7. Setty, M. *et al.* Characterization of cell fate probabilities in single-cell data with Palantir. *Nature Biotechnology* **37**, 451–460 (2019).
8. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
9. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* **38**, 1408–1414 (2020).
10. Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Nat Methods* **19**, 159–170 (2022).
11. Qiu, X. *et al.* Mapping transcriptomic vector fields of single cells. *Cell* **185**, 690-711.e45 (2022).
12. Li, C., Virgilio, M. C., Collins, K. L. & Welch, J. D. Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction. *Nat Biotechnol* 1–12 (2022) doi:10.1038/s41587-022-01476-y.
13. Qiu, X. *et al.* Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe. *Cell Systems* **10**, 265-274.e11 (2020).
14. Bergen, V., A. Soldatov, R., V. Kharchenko, P. & J. Theis, F. RNA velocity—current challenges and future perspectives. *Molecular Systems Biology* **17**, e10282 (2021).

15. Kamimoto, K. *et al.* Dissecting cell identity via network inference and in silico gene perturbation. *Nature* 1–10 (2023) doi:10.1038/s41586-022-05688-9.
16. Wang, L. *et al.* Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multi-omics. 2022.09.14.508036 Preprint at https://doi.org/10.1101/2022.09.14.508036 (2022).
17. Deshpande, A., Chu, L.-F., Stewart, R. & Gitter, A. Network inference with Granger causality ensembles on single-cell transcriptomics. *Cell Reports* **38**, 110333 (2022).
18. Finkle, J. D., Wu, J. J. & Bagheri, N. Windowed Granger causal inference strategy improves discovery of gene regulatory networks. *Proceedings of the National Academy of Sciences* **115**, 2252–2257 (2018).
19. Papili Gao, N., Ud-Dean, S. M. M., Gandrillon, O. & Gunawan, R. SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* **34**, 258–266 (2018).
20. Friedman, N. Inferring Cellular Networks Using Probabilistic Graphical Models. *Science* **303**, 799–805 (2004).
21. Jansen, R. A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science* **302**, 449–453 (2003).
22. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A. & Nolan, G. P. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science* **308**, 523–529 (2005).
23. Liu, F., Zhang, S.-W., Guo, W.-F., Wei, Z.-G. & Chen, L. Inference of Gene Regulatory Network Based on Local Bayesian Networks. *PLOS Computational Biology* **12**, e1005024 (2016).
24. Schütte, J. *et al.* An experimentally validated network of nine haematopoietic transcription factors reveals mechanisms of cell state stability. *eLife* **5**, e11469 (2016).
25. Zhang, X., Branciamore, S., Gogoshin, G., Rodin, A. S. & Riggs, A. D. Analysis of high-resolution 3D intrachromosomal interactions aided by Bayesian network modeling. *PNAS* **114**, E10359–E10368 (2017).
26. Pearl, J. Causal diagrams for empirical research. *Biometrika* **82**, 669–688 (1995).
27. Schölkopf, B. *et al.* Toward Causal Representation Learning. *Proceedings of the IEEE* **109**, 612–634 (2021).
28. Korb, K. & Nicholson, A. Bayesian Artificial Intelligence. in *Computer science and data analysis series* (2004). doi:10.5860/choice.41-5948.
29. Scutari, M. & Denis, J.-B. *Bayesian Networks: With Examples in R*. (Chapman and Hall/CRC, 2021). doi:10.1201/9780429347436.
30. Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. (MIT Press, 2009).
31. Zhang, X., Zhao, J., Hao, J.-K., Zhao, X.-M. & Chen, L. Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. *Nucleic Acids Research* **43**, e31 (2015).
32. Peng, J., Wang, P., Zhou, N. & Zhu, J. Partial Correlation Estimation by Joint Sparse Regression Models. *Journal of the American Statistical Association* **104**, 735–746 (2009).

33. Zhang, M. *et al.* GeNeCK: a web server for gene network construction and visualization. *BMC Bioinformatics* **20**, 12 (2019).
34. Scutari, M. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software* **35**, 1–22 (2010).
35. Seita, J. *et al.* Gene Expression Commons: an open platform for absolute gene expression profiling. *PLoS One* **7**, e40321 (2012).
36. Fathman, J. W. *et al.* Upregulation of CD11A on hematopoietic stem cells denotes the loss of long-term reconstitution potential. *Stem Cell Reports* **3**, 707–715 (2014).
37. Ambrosi, T. H. *et al.* Aged skeletal stem cells generate an inflammatory degenerative niche. *Nature* **597**, 256–262 (2021).
38. Nestorowa, S. *et al.* A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**, e20–e31 (2016).
39. Pellin, D. *et al.* A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nature Communications* **10**, 2395 (2019).
40. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).
41. Graf, T. Differentiation plasticity of hematopoietic cells. *Blood* **99**, 3089–3101 (2002).
42. Akashi, K., Traver, D., Miyamoto, T. & Weissman, I. L. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* **404**, 193–197 (2000).
43. Prohaska, S. S., Scherer, D. C., Weissman, I. L. & Kondo, M. Developmental plasticity of lymphoid progenitors. *Seminars in Immunology* **14**, 377–384 (2002).
44. Single-cell ATAC-seq reveals GATA2-dependent priming defect in myeloid and a maturation bottleneck in lymphoid lineages | Blood Advances | American Society of Hematology. https://ashpublications.org/bloodadvances/article/5/13/2673/475734/Single-cell-ATAC-seq-reveals-GATA2-dependent.
45. The cytoplasmic domain of Mpl receptor transduces exclusive signals in embryonic and fetal hematopoietic cells | Blood | American Society of Hematology. https://ashpublications.org/blood/article/100/6/2063/235068/The-cytoplasmic-domain-of-Mpl-receptor-transduces.
46. ICSBP-mediated immune protection against BCR-ABL–induced leukemia requires the CCL6 and CCL9 chemokines | Blood | American Society of Hematology. https://ashpublications.org/blood/article/113/16/3813/25078/ICSBP-mediated-immune-protection-against-BCR-ABL.
47. CRISPR Mediated ELANE Single-Allele Knock-out Restores Proliferation and Myeloid Differentiation of Neutropenia Patient Derived BM HSCs | Blood | American Society of Hematology. https://ashpublications.org/blood/article/136/Supplement%201/23/470437/CRISPR-Mediated-ELANE-Single-Allele-Knock-out.

48. Sitnicka, E. *et al.* Key role of flt3 ligand in regulation of the common lymphoid progenitor but not in maintenance of the hematopoietic stem cell pool. *Immunity* **17**, 463–472 (2002).
49. Okuno, Y. *et al.* Differential regulation of the human and murine CD34 genes in hematopoietic stem cells. *Proceedings of the National Academy of Sciences* **99**, 6246–6251 (2002).
50. Ng, S. Y.-M., Yoshida, T., Zhang, J. & Georgopoulos, K. Genome-wide Lineage-Specific Transcriptional Networks Underscore Ikaros-Dependent Lymphoid Priming in Hematopoietic Stem Cells. *Immunity* **30**, 493–507 (2009).
51. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* **14**, 975–978 (2017).
52. Cytlak, U. *et al.* Differential IRF8 Transcription Factor Requirement Defines Two Pathways of Dendritic Cell Development in Humans. *Immunity* **53**, 353-370.e8 (2020).
53. Luc, S. *et al.* Bcl11a Deficiency Leads to Hematopoietic Stem Cell Defects with an Aging-like Phenotype. *Cell Reports* **16**, 3181–3194 (2016).
54. Tsang, J. C. H. *et al.* Single-cell transcriptomic reconstruction reveals cell cycle and multi-lineage differentiation defects in Bcl11a-deficient hematopoietic stem cells. *Genome Biol* **16**, 178 (2015).
55. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research* **46**, D380–D386 (2018).
56. Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **50**, D165–D173 (2022).
57. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438–2444 (2010).
58. Xie, X., Rigor, P. & Baldi, P. MotifMap: a human genome-wide map of candidate regulatory motif sites. *Bioinformatics* **25**, 167–174 (2009).
59. Consortium, T. E. P. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLOS Biology* **9**, e1001046 (2011).
60. Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. (Morgan kaufmann, 1988).
61. Akaike, H. A new look at the statistical model identification. *IEEE transactions on automatic control* **19**, 716–723 (1974).
62. Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics* **6**, 461–464 (1978).

Methods

Data preprocessing. Normalized mouse microarray, mouse bone marrow scRNA-seq, human bone marrow scRNA-seq and human bone marrow scATAC-seq datasets were obtained from the data sources. In the microarray dataset, each cell type had 3 or 4 biological replicates, and the expression levels of 21,678 genes were detected by 45,101

633 probes. For each gene detected by multiple probes, the median value was adopted as its
634 expression level. The top variable features were first selected for Bayesian network
635 structure learning; 3,000 features were selected in the mouse microarray dataset, 5,000 in
636 the mouse bone marrow scRNA-seq dataset, 4,000 in the human bone marrow scRNA-
637 seq dataset and 1,764 (all TFs) in the human bone marrow scATAC-seq dataset.
638 **Construction of cellular differentiation networks.** The existing Bayesian network
639 structure learning algorithms can be divided into three categories: constraint-based
640 algorithms, score-based algorithms and hybrid algorithms[29,30]. Constraint-based
641 algorithms construct a Bayesian network by first generating an undirected skeleton via
642 Markov blanket[60] or correlation and then orienting the skeleton under acyclicity
643 constraints to obtain the directed acyclic graph (DAG) that best satisfies the conditional
644 independence test. The effectiveness and efficiency of constraint-based algorithms are
645 limited due to the lack of robustness and the high complexity of the conditional
646 independence test. Score-based algorithms evaluate how well a DAG fits the observed
647 data with a score such as the Akaike information criterion[61] or Bayesian information
648 criterion[62], with the goal of identifying the graph obtaining the highest score. Major issues
649 for score-based algorithms include the exponentially growing search space as the number
650 of variables increases and the search algorithm may lead to a local optimum. One solution
651 to the limitations of the abovementioned two types of algorithms is hybrid algorithms,
652 which first generate an undirected skeleton and then search for the DAG with the highest
653 score among the structures restricted to this skeleton. However, the three types of BN
654 structure learning face a common problem: they are sensitive to observations and
655 parameters, and small changes in input can lead to considerably different outputs. Hence,
656 considerable prior knowledge is needed to tune and obtain expected structures.
657 To overcome these limitations, CIBER employs several strategies to ensure a well-
658 structured hierarchy with minimal prior knowledge of assigning credible upper and lower
659 boundaries for the number of edges in the network and setting stem cells as the root.
660 CIBER utilizes a multistep strategy to obtain robust BN structures — resampling,
661 undirected skeleton learning, DAG learning, combining DAGs and trimming branches.
662 Particularly, for datasets with limited biological replicates, such as microarray datasets,
663 the resampling step is omitted. By combining the DAGs obtained with different
664 resampled subsets of single cells and various parameters, we can extract the most robust
665 associations between variables, which have a higher probability of being the true
666 causation. The details of the steps are described below.
667 **Resampling strategy.** The resampling strategy maximizes the efficacy of large single-
668 cell omics datasets. A stratified resampling strategy was performed for the single-cell
669 omics datasets to generate a subset of cells, while this strategy was omitted for the
670 microarray dataset. For the microarray dataset and each subset of single cells after
671 resampling, we calculated the mean expression level of each cell type to obtain a mean
672 expression matrix, with each column representing a cell type and each row representing
673 a specific feature. The resampling ratio can be set to approximately 0.3 to prevent high
674 randomness or homogeneity between the obtained subsets. Similarly, the cell number of
675 each cell type was suggested to be greater than 50 to prevent random effects.
676 **Undirected skeleton learning.** With the mean expression matrix as an input, an
677 undirected skeleton was built to describe the correlation between each pair of cell types,
678 with each node representing a cell type and each edge representing the correlation
679 between the connected cell types. Various metrics can be utilized to evaluate the
680 correlation, such as Spearman's correlation or the mutual information. In addition to the

correlation, several algorithms developed for undirected skeleton construction can be conveniently chosen under the support of GeNeCK[33]. The default algorithm used in CIBER is CMI2NI[31], an approach that was developed for reverse-engineering gene regulatory networks by quantifying the conditional mutual inclusive information between two genes given a third gene by calculating the Kullback–Leibler divergence. The undirected skeleton is determined by the threshold used for judging the correlation, and a sparser skeleton will be obtained given a higher correlation threshold. A set of sequential correlation thresholds was used to obtain a series of undirected skeletons for further DAG learning.

DAG learning. Typically, two causally associated variables should also be correlated. The undirected skeleton provides us with a reference that can be used to determine the causal DAG. We excluded absent edges in the undirected skeleton in the alternative DAGs and utilized the existing score-based structure learning algorithm offered in the bnlearn[34] package to obtain the DAGs. The hill climbing algorithm was used to search for the DAG with the highest Bayesian information criterion (BIC)[62]. BIC is commonly used in Bayesian network structure learning to evaluate how well a graph G fits the observed data D and can be formulated as follows:

$$\mathrm{BIC}(G{:}D) = LL(G{:}D) - \|G\| \cdot \log|\mathrm{D}|/2,$$

where $LL(G{:}D)$ describes the likelihood that G fits D, and $\|G\|$ and $\log|\mathrm{D}|$ represent the complexity of the graph structure and dataset, respectively. Thus, hill climbing seeks to identify the least complex DAG that best explains the observed data.

Combination of DAGs and trimming branches. A DAG can be generated with each correlation threshold under a resampled subset of cells. By changing the correlation threshold, we can obtain a series of different DAGs. The DAGs with the number of edges located between the prior-set upper and lower boundaries are selected and combined to obtain a directed graph, with each edge weighted according to its frequency of occurrence. According to the assumption that the true causation should be robust to different parameters, a combined weighted DAG can be generated by extracting the highest weighted edges and removing the lowest weighted edges in cycles.

Then, this DAG is binarized to obtain an unweighted DAG named one-shot DAG in this study, which indicates the robust causation within this batch of cells. For microarray or bulk sequencing datasets, the one-shot DAG is the final learned Bayesian network structure. However, for single-cell omics datasets, the one-shot DAG reflects only the possible causation within a specific subset of the whole dataset. Therefore, a similar combination is applied over different one-shot DAGs to generate a quasi-final DAG. Under some special conditions, when generating the quasi-final DAG, additive prior knowledge may be needed to determine which edge should be severed if two edges with the same weights occur in one cycle. The quasi-final DAG may contain some edges that are not robust while being kept for the reason that their opposite edges occur even less.

An optional procedure can be performed on the quasi-final DAG to limit the maximum number of parents for each node and to remove redundant low-weighted edges. Specifically, under the structural causal model framework, we fit the quasi-final DAG to the mean expression matrix of all the cells in the original dataset twice, once with the top 25% variable features and once with the top 25~50% variable features, and maintain the edges that remain stable while removing edges with low-weighted coefficients.

Construction of the effect matrix. Based on the learned Bayesian network, the SCM framework can be utilized to build an effect matrix that quantifies the effects of different features on various differentiation branches. SCM is a conventional causal inference

framework that combines probabilistic graphical models and the notion of interventions. This framework considers a set of associated observables with the nodes of a directed acyclic graph while viewing each node X_i as a function of its parents $\mathbf{PA}_i$ in the graph:

$$X_i = h_i(\mathbf{PA}_i, U_i),$$

where U_i represents noise or exogenous variables. Under a linear additive noise assumption, X_i can be viewed partially as a linear regression of $\mathbf{PA}_i$:

$$X_i = \sum_k \beta_{ik}\mathbf{PA}_{ik} + U_i.$$

Interventions are performed to infer causality based on the SCM equations. Typically, by intervening in the form of a function or distribution of $\mathbf{PA}_i$, a latent expression of X_i can be obtained. The difference between the observed and latent expressions of X_i is measured to estimate the effect of causality.

In the construction of the effect matrix, we perform interventions by *in silico* perturbation of permuting (or deleting) the expression of a specific feature in all cell types and estimate the effect of causality by concentrating on the change in β_{ik} instead of the change in X_i. The coefficient β_{ik} represents the association between two connected cell types and reflects the information flowing from $\mathbf{PA}_{ik}$ to X_i and the information of the corresponding differentiation branch. Thus, the change in the coefficients caused by intervening in a specific feature quantifies the effect of that feature on the cellular differentiation process.

For the single-cell omics dataset, we used a resampling strategy similar to that performed in reconstructing networks to record the distribution of the coefficient associated with a particular edge e_0 obtained without and with interventions on a specific feature f_0, which we termed α_{ref} and α_{inter}, respectively, where α represents the abovementioned β_{ik}. For one resampling and intervention, the matrix $\text{EffectMat}_{\text{sampling}}$ is defined as

$$\text{EffectMat}_{\text{samping}}[f_0, e_0] = \text{effect}(f_0, e_0) = \alpha_{\text{ref}} - \alpha_{\text{inter}},$$

where $\text{EffectMat}_{\text{samping}}[f_0, e_0]$ is the row f_0 and column e_0 element in the matrix, and $\text{effect}(f_0, e_0)$ is the effect of feature f_0 on edge e_0. By resampling and permuting several times, we obtain a series of $\text{EffectMat}_{\text{samping}}$ matrices, which can be horizontally merged to generate a final effect matrix in which the rows consist of all features and the columns consist of all edges multiplied by the resampling and permutation times. For microarray or bulk sequencing datasets, the effect matrix is constructed with the same procedures, except that the resampling step is omitted and the intervention is performed with deletion instead of permutation.

DDF analysis. We developed DDF analysis to identify features important to specific lineages or the entire cellular differentiation network according to the effect matrix. When applying DDF analysis, diffScores are defined to evaluate the effects of different features on the differentiation process or a specific lineage.

For the haematopoiesis network, we defined diffSum, $\text{diffScore}_{\text{lym}}$ and $\text{diffScore}_{\text{myl}}$ to quantify the effects of the features on the entire haematopoietic hierarchy, the lymphoid-lineage and the myeloid-lineage, respectively. Let e represent an edge; the effect of feature f_0 on the entire haematopoietic cellular differentiation network G is defined as

$$\text{diffSum} = \sum_{e \in G} |\text{effect}(f_0, e)|.$$

The lymphoid-lineage and myeloid-lineage effects are defined as

$$\text{diffScore}_{\text{lym}} = \sum_{e \in \{S \to L\} \cup \{L \to L\}} |\text{effect}(f_0, e)| - \sum_{e \in \{L \to M\}} \text{effect}(f_0, e)$$

and

$$\text{diffScore}_{\text{myl}} = \sum_{e \in \{S \to M\} \cup \{M \to M\}} |\text{effect}(f_0, e)| - \sum_{e \in \{M \to L\}} \text{effect}(f_0, e),$$

where S represents early cell types, including stem cells and multipotent cells, L represents lymphoid-lineage cell types, M represents myeloid-lineage cell types, and $\{S \to L\}$ represents the set of edges pointing from cell types in set S to cell types in set L.

Transcription factor–targeted gene set GSEA and ROC analyses. For each TF, we collect the regulated genes from 5 databases (TRRUST, JASPAR, ChEA, Motifmap and ENCODE) and apply GSEA and receiver operating characteristic (ROC) analyses on this gene set. GSEA and ROC analyses are performed with the ranked gene lists obtained by DDF and RDF analyses.

Specifically, normalized enrichment score (NES) is calculated in GSEA analysis. When applying ROC analysis, we plot ROC curve by assigning the TF-targeted genes as positive and the others as negative samples. In ROC curve, the true positive rate is referred to as sensitivity and the true negative rate is referred to as specificity,

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

and

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}.$$

AUC score is calculated as the area under ROC curve.

Background distributions are obtained by calculating NES and AUC of 1,000 randomly ordered lists of all the genes.

Competitive HSC transplantation and analysis. In all experiments (unless otherwise specified), mice were homozygous for the *Bcl11b* allele. *Bcl11b*$^{-/-}$ (C57BL/6, CD45.2) mice, C57BL/6-SJL (CD45.1) (Stock No: 002014) and C57BL/6 WT (CD45.2) (Stock No: 000664) mice were obtained from the Jackson Laboratory. CD45.1/2 mice were heterozygotes from CD45.1 and CD45.2 mice. All mice were kept in specific-pathogen-free (SPF), AAALAC-accredited animal care facilities at the Laboratory Animal Research Center, Tsinghua University and all procedures were approved by the Institutional Animal Care and Use Committee of Tsinghua University. Bone marrow cells were harvested by crushing the femur and tibias with pestle and mortal in Hanks balanced salt solution (HBSS) with 2% fetal bovine serum and 1% N-2-hydroxyethylpiperazine-N'-2-ethanesulfonic acid buffer (HBSS+). Cells were stained with antibodies labelled with fluorochromes. For hematopoietic stem and progenitor cells enrichment, bone marrow cells were stained with c-Kit-APC. Hematopoietic populations were identified by flow cytometry using BD LSRFortessa (BD Biosciences). Data were analysed using FlowJo software.

For the primary competitive HSC transplantation assay, 50 freshly isolated HSCs from WT or *Bcl11b*$^{-/-}$ mice (CD45.2) were transplanted into lethally irradiated (10 Gy) recipients (CD45.1/2) together with 3×10^5 competitor cells (CD45.1). Donor-derived chimerism (including B cells, T cells and myeloid cells) in the peripheral blood of recipients was analysed at monthly intervals.

Two-tailed unpaired Student's *t* test was applied for analysing experiment data after testing for normal distribution. All histograms were plotted by GraphPad Prism 8 software, and p value <0.05 was regarded as significant for all tests. All data are depicted as the mean ± SD.

Data availability

The datasets used in this work were downloaded. Mouse microarray dataset[35–37] can be found at GEO (GSE34723). Mouse bone marrow scRNA-seq data[38] can be found at GEO (GSE81682). Human bone marrow scRNA-seq data[39] can be found at GEO (GSE117498). Human bone marrow scATAC-seq data[40] can be found at GEO (GSE96772).

Code availability

CIBER toolkit is implemented in R and is available on Bioconductor and GitHub (https://github.com/Lan-lab/CIBER). Tutorials of CIBER usage cases are also available through GitHub (https://github.com/Lan-lab/CIBER/tutorials).

Acknowledgements

The authors are grateful to Dr. Yeguang Chen from Tsinghua University for providing *Bcl11b*$^{flox/flox}$ mice. We thank the High-Performance Computation platforms of Tsinghua University for providing computational resources.

Author Contributions

L. X., J. W. and X. L. conceived the project. L. X. designed the models and theories. L. X. and H. X. implemented the code. L. X., H. X., N. Y., C. T. and S. Y. performed the analyses. J. W. designed the experiments and T. C. conducted the experiments. L. X. and T. C. generated the figures. M. Z., R. S., R. Y., W. Z. helped in designing the analyses and explaining the results. X. L. supervised the project. All authors contributed to the writing of the manuscript.

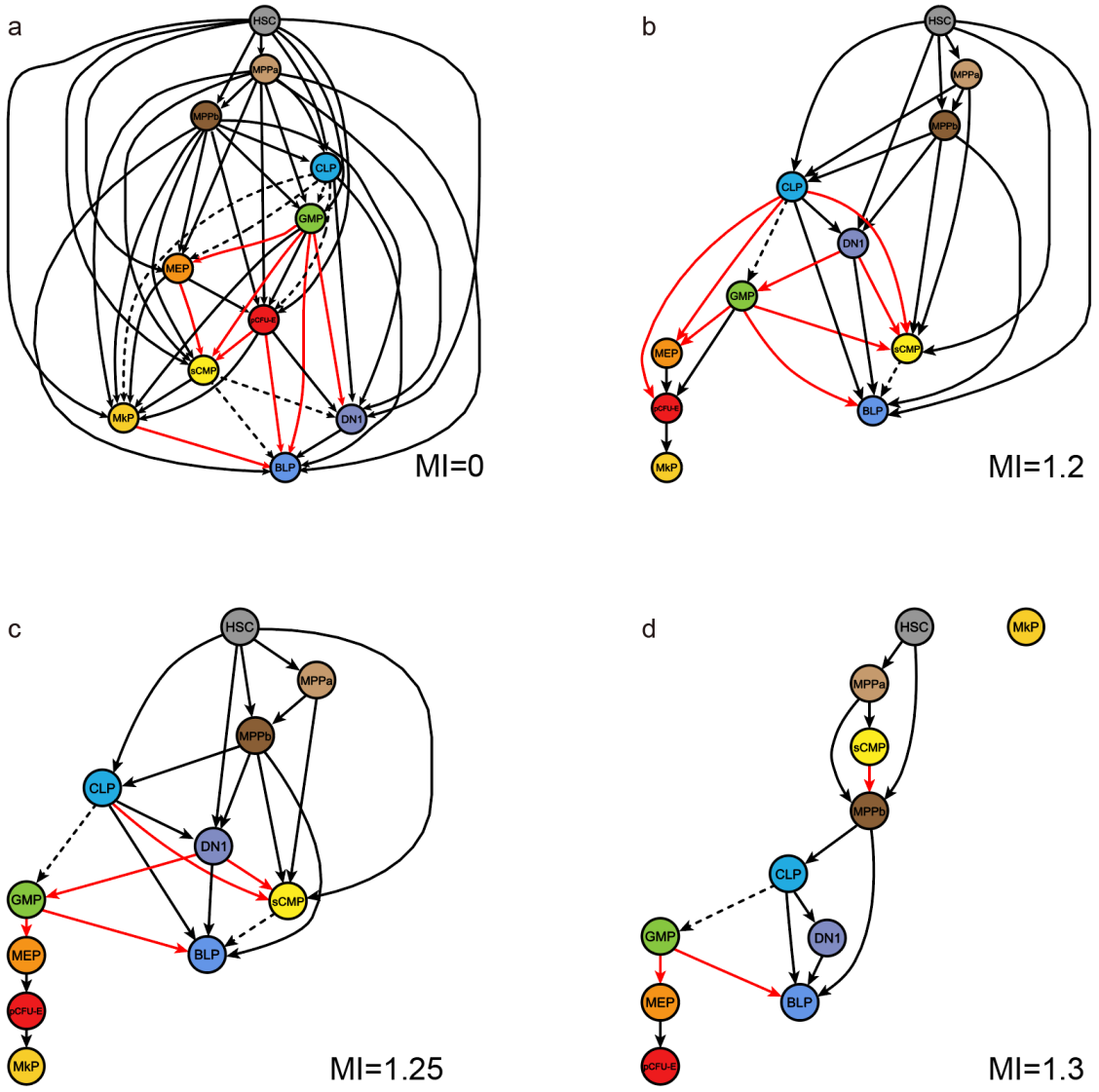
Funding

This work was partially supported by the grants (Grant No. 81972680 to X.L.) from National Natural Science Foundation of China, the grants (Grant No. 61020100119 to X.L.) from Tsinghua University-Peking University Jointed Center for Life Science, the grants (Grant No. 20201100463 to X.L.) from Beijing Natural Science Foundation, a start-up fund from Tsinghua University-Peking University Joined Center for Life Science and Alibaba innovation research programme.

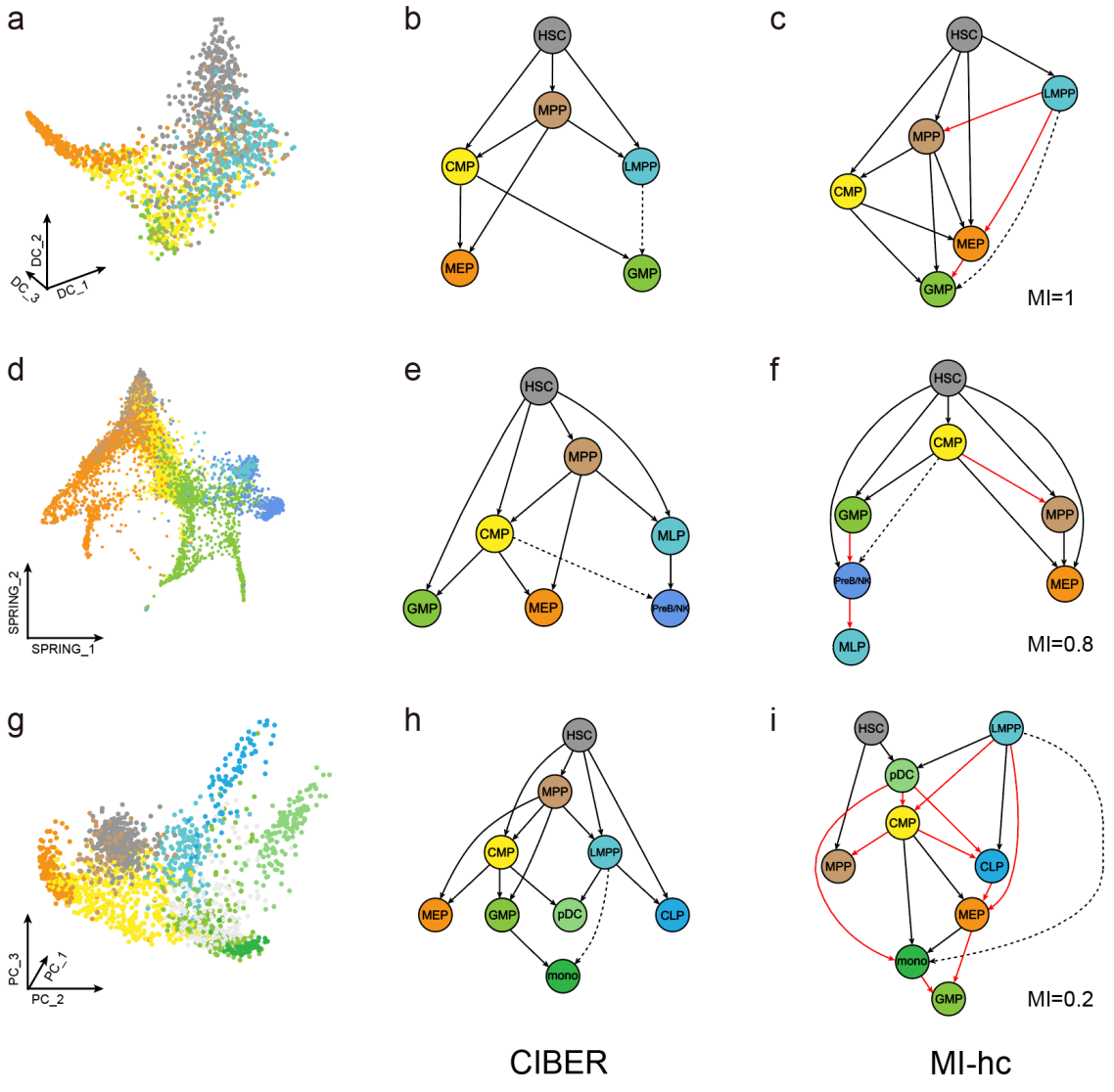
Competing interests

The authors declare no competing interests.

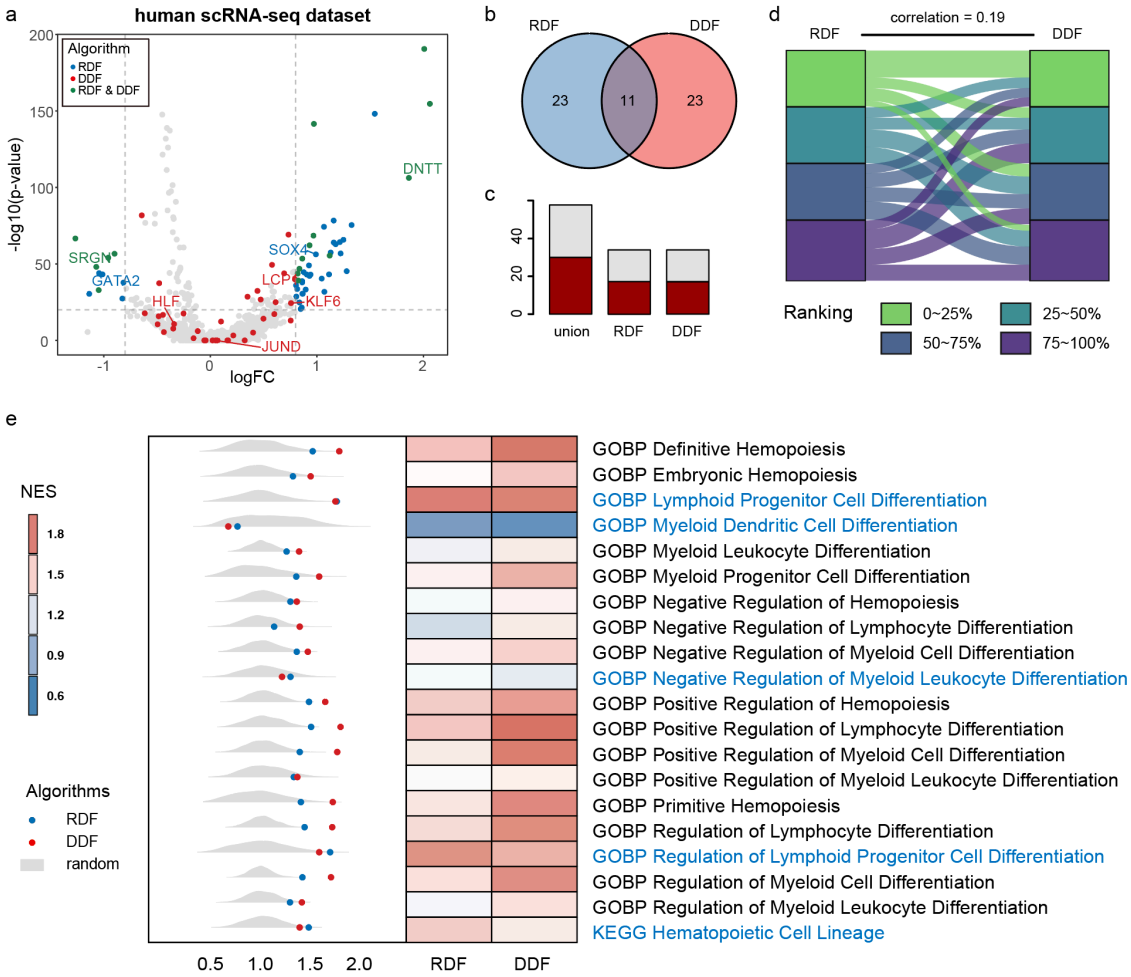
Extended data



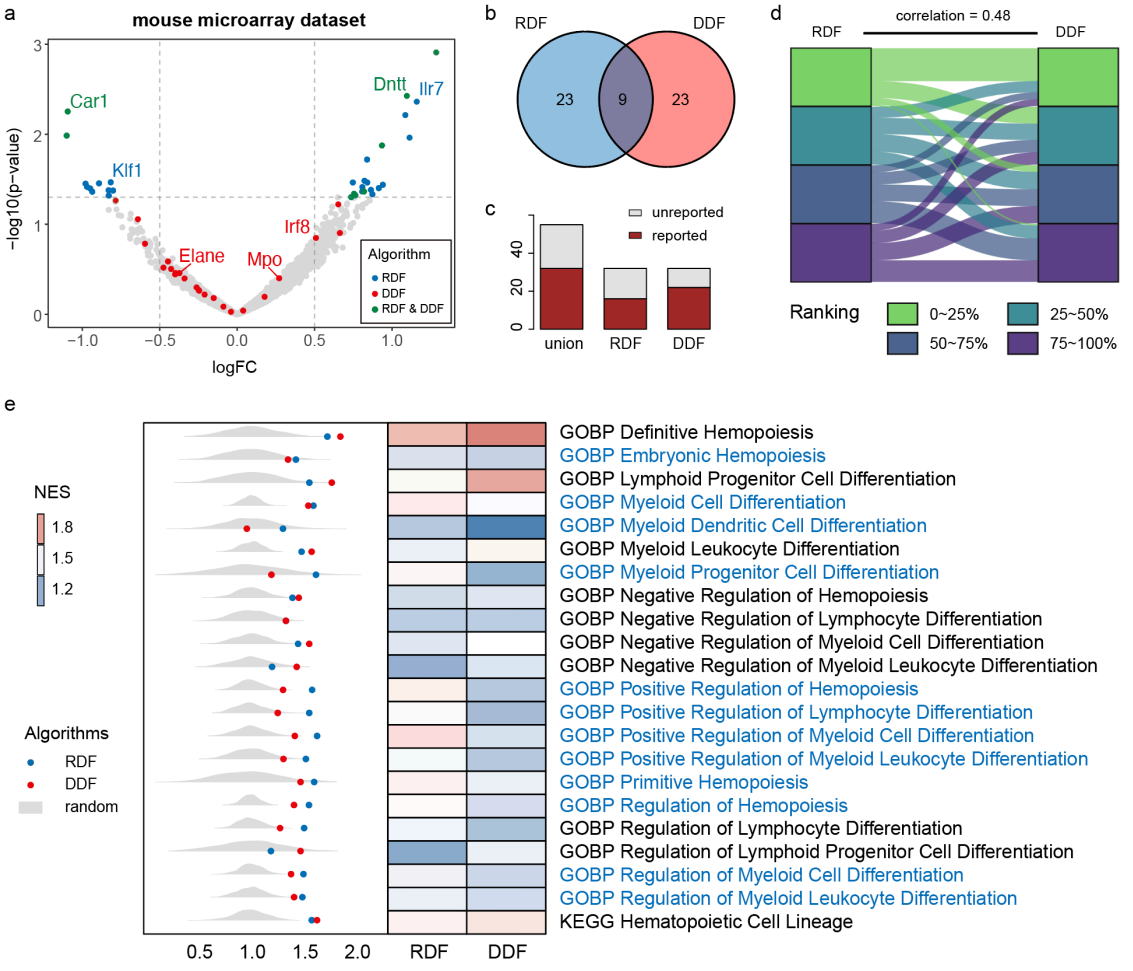
Extended Data Fig. 1 | Typical Bayesian network structure learning algorithm is sensitive to parameters. a-d, Networks learned by the typical Bayesian network structure learning strategy MI-hc given different correlation thresholds with the same microarray dataset. The dashed edges represent possible unconventional interlineage branches, while the red edges represent less credible branches.



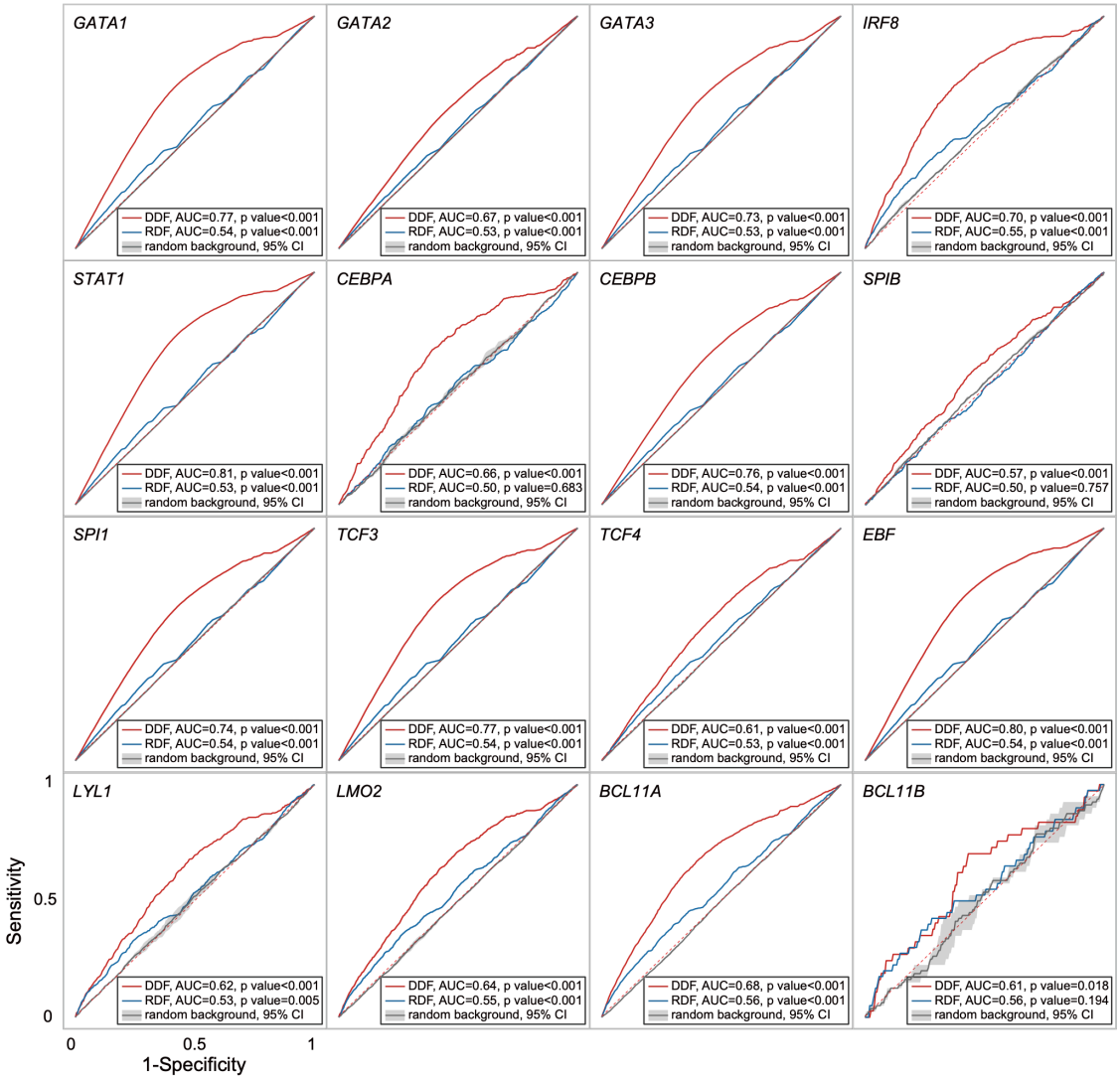
Extended Data Fig. 2 | CIBER-learned structures conform to prior knowledge and reduction landscapes better than the MI-hc algorithm. Plots show low-dimensional landscapes[38–40], CIBER and MI-hc results of (**a-c**) mouse scRNA-seq[38], (**d-f**) human scRNA-seq[39] and (**g-i**) human scATAC-seq[40] datasets respectively. The dashed edges represent unconventional interlineage edges that were reported to be possible in previous studies, while the red edges represent less credible edges. The results show that compared to the typical Bayesian network structure learning algorithm MI-hc, the CIBER-learned results were consistent with conventional knowledge with a larger number and a higher percentage of edges and captured the bifurcation structure more clearly.



Extended Data Fig. 3 | DDF identifies lineage driving features. The plots show the results on RDF and DDF gene lists obtained with the human scRNA-seq dataset. **a**, Features are allocated by log fold change and log p value obtained through RDF analysis; top RDF and DDF features are coloured. **b**, The number of shared features between the two feature sets. **c**, Literature search result shows that the DDF and RDF analyses identify different important features, in which the red parts represent the reported features. **d**, Correlation between the ranked lists. **e**, GSEA NES with terms related to haematopoiesis or haematopoietic differentiation in the GO and KEGG databases. The terms for which RDF list obtains the higher NES are coloured in blue. Dots in the left part represent the NES of RDF and DDF ranking lists, while the distributions show the background NES given randomly ranked gene lists.



Extended Data Fig. 4 | DDF results on the mouse microarray dataset. a, Features are allocated by log fold change and log p value obtained through RDF analysis; top RDF and DDF features are coloured. **b**, The number of shared top features between the two feature sets. **c**, Literature search result shows that the DDF and RDF analyses identify different important features, in which the red parts represent the reported features. **d**, Correlation between the ranked lists. **e**, GSEA NES with terms related to haematopoiesis or haematopoietic differentiation in the GO and KEGG databases. The terms for which RDF list obtains the higher NES are coloured in blue. Dots in the left part represent the NES of RDF and DDF ranking lists, while the distributions show the background NES given randomly ranked gene lists. The RDF list performed slightly better than the DDF list in 12 out of 22 terms. However, the literature search still showed that DDF analysis identified some important genes which are not significantly expressed, demonstrating that DDFs can complement RDFs to better identify genes crucial to cell differentiation.



Extended Data Fig. 5 | ROC curves on each gene set targeted by different TFs. Results show that DDF analysis ranks the genes targeted by the reported important TFs at higher positions compared to regular differential analysis, indicating that DDF analysis may provide insight into the potential mechanism of how important TFs impact the cellular differentiation process.