Reconstructing cellular differentiation networks and identifying cell fate–determining features with CIBER

Longchen Xu[1,2,#], TingTing Cong[3,#], Hengyi Xu[4,#], Sijie Yang[1], Rahul Sinha, Ryo Yamamoto, Wei Zhang[1,2], Jianwei Wang[3,*] and Xun Lan[1,5,6,*]

[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]Department of Basic Medical Science, School of Medicine, Tsinghua University, Beijing 100084, China

[6]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

Most existing trajectory inference methods are similarity-based and lack an understanding of the latent causality among 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 is suitable for various types of omics data, including profilings of gene expression and chromatin accessibility. We show that CIBER can capture known cell-lineage structures and identify potential novo differentiation branches. By combining the CIBER-learned network with the structural causal model and applying *in silico* perturbation, we constructed an effect-matrix that quantifies the impacts of different features on each branch. We subsequently identified features important to the network structure through differentiation driver feature (DDF) analysis. We demonstrated that DDF analysis can identify features crucial to haematopoiesis but show no significant

difference in expression between lineages. We further experimentally verified the effect of transcription factor *Bcl11b* on haematopoiesis, showing that it promotes lymphoid lineage differentiation while suppressing myeloid lineage differentiation, which is consistent with our predictions.

Introduction

Cellular differentiation is one of the most fundamental processes in biology. Single-cell omics data have provided unprecedented opportunities for elucidating developmental pathways and reconstructing cell lineages. Several computational trajectory inference methods have been developed for studying cellular developmental dynamics. 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. These similarity-based trajectory inference methods tend to identify correlations between different cell states; however, the process of cellular differentiation is driven by its causal nature and thus may not match these methods properly.

Recently, a series of approaches[8–11] for cell fate prediction based on RNA velocity[12] have emerged, which model the differentiation process with the spliced-to-unspliced mRNA ratio in the form of differential equations, thereby restoring considerable cell temporal information, and can then be utilized for causal gene regulatory network reconstruction[13]. However, RNA velocity is limited to representing cell dynamics in the transcriptome space[14] and is not suitable for other types of data. More general approaches are required to identify causality between cellular states, not only through transcriptomic profiling but also through different types of data, such as epigenomic profiling or microarray assays.

53 Methods including Granger causality[15–17], Boolean networks[18–20] and Bayesian
54 networks[21–24] have been effectively applied for causal inference in studying biological
55 processes such as gene regulatory networks and protein-protein interactions. However,
56 the application of these causal inference methods is rare in scenarios such as the
57 reconstruction of cellular differentiation networks and the identification of cell fate–
58 determining features. Bayesian networks (BNs) provide a computationally efficient
59 approach for inferring causality among a set of variables and have been applied to infer
60 cellular networks[21], protein-protein interactions[22], protein-signalling networks[23], gene
61 regulatory networks[24,25], and long-range intrachromosomal interactions[26]. However, data-
62 based inferences of feasible BNs are challenging because the structure learning of BNs is
63 sensitive to parameters and requires considerable prior knowledge to obtain the expected
64 structure.

65 Here, we present CIBER, a novo causal inference–based framework, to reconstruct the
66 cellular differentiation networks and to identify cell fate–determining features. CIBER
67 provides a hybrid structure learning algorithm for detecting information flow among
68 different cell types and is suitable for both gene expression and chromatin accessibility
69 data. In addition to single-cell transcriptomic and epigenome sequencing data, which
70 assume contiguous cell states and are critical for typical trajectory inference methods,
71 CIBER can use microarray or bulk sequencing data in which the states of various cell
72 types are discrete. We then adopt the framework of the structural causal model[27,28] to the
73 directed acyclic graph (DAG) structure of the learned cellular differentiation network,
74 utilize *in silico* perturbation as interventions for causal inference, and build an effect
75 matrix that quantifies the impacts of different features on each branch of the cellular
76 differentiation network. Furthermore, differentiation driver feature (DDF) analysis is
77 defined to identify important features in the cellular differentiation process. In contrast to
78 regular differential feature analysis, in which the identified features are markers of each
79 cell type, DDFs capture the features that significantly influence the entire differentiation

process or a specific lineage. We demonstrated that DDF analysis can identify differentiation-driving genes that are 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 trajectories. The application on a mouse microarray dataset with the most cell types showed that CIBER can identify 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 represent analysis identifying important features by directly comparing the values in cell × feature matrices, such as differentially expressed gene analysis with scRNA-seq datasets or differential deviation-score transcription factor (TF) analysis with scATAC-seq datasets. In contrast, DDF analysis is performed by identifying important driving features based on edge × feature matrices inferred by CIBER from the cell × 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 cell fate–determining features that are not deemed statistically significant in RDF analysis. We experimentally validated CIBER's prediction of the lymphoid-biased effect of TF B-cell lymphoma/leukaemia 11B (*Bcl11b*).

Results

CIBER offers a robust Bayesian network structure learning strategy. Bayesian networks are directed acyclic graphical models in which nodes represent variables and edges represent probabilistic dependencies between nodes[29]. One major issue in Bayesian network structure learning is that the obtained graph is extremely sensitive to parameters

105 and the observed data; thus, small changes in input may result in considerably different
106 outputs. To obtain a cellular differentiation network well-constructed for downstream
107 analysis, extensive expert knowledge is usually required as prior information to adjust the
108 learned structures. Here, we aim to develop a new framework that can generate robust
109 Bayesian network structures given minimal prior knowledge, such that the proposed
110 method can be generalized to investigate understudied biological processes.

111 The existing Bayesian network structure learning algorithms can be divided into three
112 categories: constraint-based algorithms, score-based algorithms and hybrid
113 algorithms[30,31]. Constraint-based algorithms construct a Bayesian network by first
114 generating an undirected skeleton via Markov blanket[32] or correlation and then orienting
115 the skeleton under acyclicity constraints to find the directed acyclic graph that best
116 satisfies the conditional independence test. The effectiveness and efficiency of constraint-
117 based algorithms are limited due to the lack of robustness and the high complexity of the
118 conditional independence test. Score-based algorithms evaluate how well a DAG fits the
119 observed data with a score such as the Akaike information criterion[33] or Bayesian
120 information criterion[34], with the goal of identifying the graph obtaining the highest score.
121 Major issues for score-based algorithms include the exponentially growing search space
122 as the number of variables increases and the search algorithm may lead to a local optimum.
123 One solution to the limitations of the abovementioned two types of algorithms is hybrid
124 algorithms, which first generate an undirected skeleton and then search for the DAG with
125 the highest score among the structures restricted to this skeleton. However, the result of
126 regular hybrid algorithms is sensitive to the parameters, and considerable prior knowledge
127 is required to obtain a proper structure.

128 To address the abovementioned issues and strengthen the robustness of Bayesian
129 network structure learning, we present in CIBER framework a new hybrid algorithm that
130 uses a mixed strategy of resampling over single cells and then combining the BNs learned

131 within each resampling (Fig. 1). The increasing size of single-cell omics data provides
132 hundreds or thousands of replicates of each cell type, allowing us to comprehensively
133 observe cell state transitions. CIBER takes an expression matrix as input and outputs a
134 feasible structure. A stratified resampling strategy is used to select random subsets of cells
135 to construct the network, resulting in a set of networks with various structures.
136 Specifically, for each subset of cells, we utilized the centroids of different cell types to
137 construct a Bayesian network, with each centroid representing the mean expression level
138 of its cell type. These centroids are connected to generate an undirected skeleton
139 according to Spearman's correlation or the mutual information between them, with each
140 edge representing the correlation between the two connected cell types. Several
141 algorithms for correlation estimation, such as CMI2NI[35] and SPACE[36], can be chosen
142 under the support of GeNeCK[37], an R framework for constructing gene regulatory
143 networks. Based on the undirected skeleton, we utilized the hill climbing algorithm
144 through the bnlearn[38] package to search for the DAG structure with the highest Bayesian
145 information criterion.

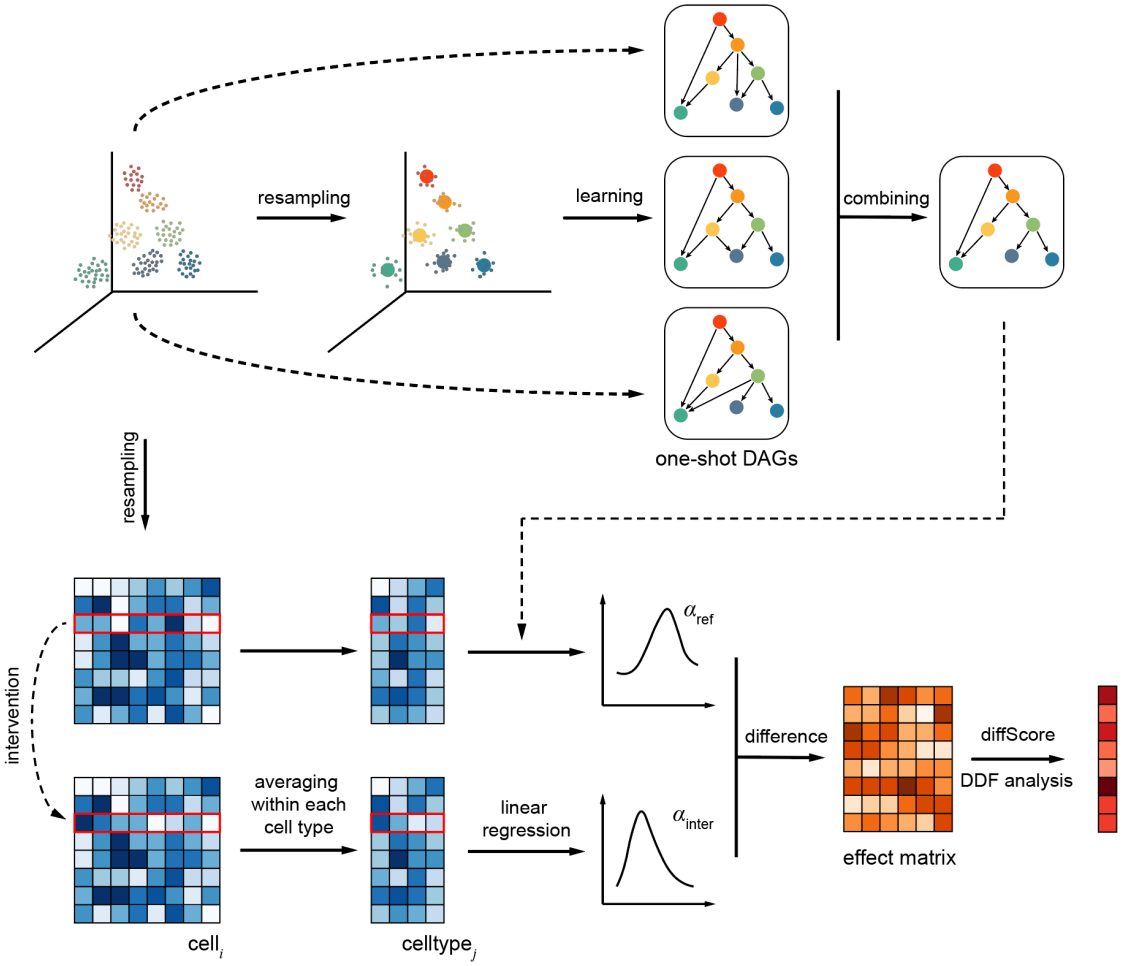


Fig. 1 | Simplified schematic of CIBER framework. The upper part shows the pipeline of the hybrid Bayesian network structure learning algorithm starting from cell resampling. Random subsets of cells are selected, and various DAGs are learned to represent the causation between the centroids of different 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 cell fate–determining features.

The learned DAG is affected by the threshold used to evaluate correlations while constructing the undirected skeleton. We used a set of sequential correlation thresholds, such as from 0 to 1 in steps of 0.01, to obtain a series of DAGs varying from dense to

166 sparse. Approximate upper and lower boundaries for the number of edges in the network
167 can be assigned according to the expected complexity of the network. DAGs with the
168 number of edges within the specified boundaries are selected and combined to generate a
169 directed graph in which each edge is weighted according to the frequency at which it
170 appears. Each weight reveals the robustness of an edge, thereby reflecting the confidence
171 of the causation between the nodes it connects. However, the graph is not necessarily
172 acyclic and should be trimmed and binarized to produce an unweighted DAG that
173 preserves the most robust structure within this batch of cells. This DAG is named one-
174 shot DAG hereafter because it is obtained within one resampling. This process is then
175 repeated for multiple times to generate a set of one-shot DAGs, over which a second
176 combination is applied to generate a final DAG that captures the most consistent causation
177 between variables (see Methods for detailed description). In addition to single-cell omics
178 data, CIBER is suitable for microarray and bulk sequencing data, but without resampling
179 due to limited replicates.

180 **CIBER captures known differentiation structures in various datasets.** To test
181 whether CIBER can obtain well-established cell lineage structures, we applied it to study
182 haematopoiesis with various datasets, including mouse microarray data[39–41], mouse bone
183 marrow scRNA-seq data[42], human bone marrow scRNA-seq data[43] and human bone
184 marrow scATAC-seq data[44]. The mouse microarray dataset contains comprehensive
185 blood cell types, from which we selected a subset of representative types of progenitors
186 to reconstruct the early haematopoietic hierarchy. The subset includes 11 cell types —
187 haematopoietic stem cell (HSC), multipotent progenitor subset A (MPPa), multipotent
188 progenitor subset B (MPPb), common myeloid progenitor (CMP), common lymphoid
189 progenitor (CLP), granulocyte-macrophage progenitor (GMP), megakaryocyte-erythroid
190 progenitor (MEP), megakaryocyte progenitor (MkP), precolony-forming unit erythroid
191 (pCFU-E), B-cell-biased lymphoid progenitor (BLP) and double negative 1 thymocytes
192 (DN1) — each containing 3 or 4 biological replicates. In this microarray dataset, the

193 expression levels of 21,678 genes are detected by 45,101 probes[39–41]. If a gene is
194 simultaneously detected by several probes, we set its expression level to the median
195 among all the probes. Because of the stable expression levels over the limited size of
196 replicates, we omitted resampling and performed the analysis on a set of fixed points that
197 are the centroids obtained by averaging the expression level over biological replicates for
198 each cell type.

199 The results revealed that CIBER can extract robust lineage structures and provide
200 information on potential differentiation branches (Fig. 2b). Compared to the results
201 obtained with a typical hybrid algorithm consisting of mutual information and hill
202 climbing (hereafter named MI-hc), the networks derived from CIBER conformed more
203 to the conventional haematopoietic hierarchy and captured the bifurcation structure more
204 accurately (Fig. 2b, c). In these figures, the edges contradicting prior knowledge are
205 coloured in red. MI-hc is sensitive to parameters, and small changes in the correlation
206 threshold can result in considerably different structures (Extended Data Fig. 1), while the
207 structure obtained by CIBER is robust without the need to manually set the exact
208 correlation threshold. In addition to the well-known branches of the HSC→MPPa/MPPb
209 →lymphoid/myeloid lineages and those inside the lymphoid or myeloid lineage, CIBER
210 identified three unconventional interlineage edges that were reported to be possible in
211 previous studies, CLP→GMP[45] and CMP→BLP/DN1[46,47].

212 The application of CIBER to the single-cell omics datasets followed the entire pipeline,
213 including the resampling procedure. The structures learned by CIBER again identified
214 larger numbers and higher percentages of conventional edges compared to the results of
215 the typical hybrid algorithm MI-hc (Fig. 2d-f, Extended Data Fig. 2). Furthermore, the
216 cell types in CIBER-learned structures were in relative positions consistent with the low-
217 dimensional landscapes of diffusion maps[42], SPRING[43] or PCA[44] obtained in previous
218 studies.

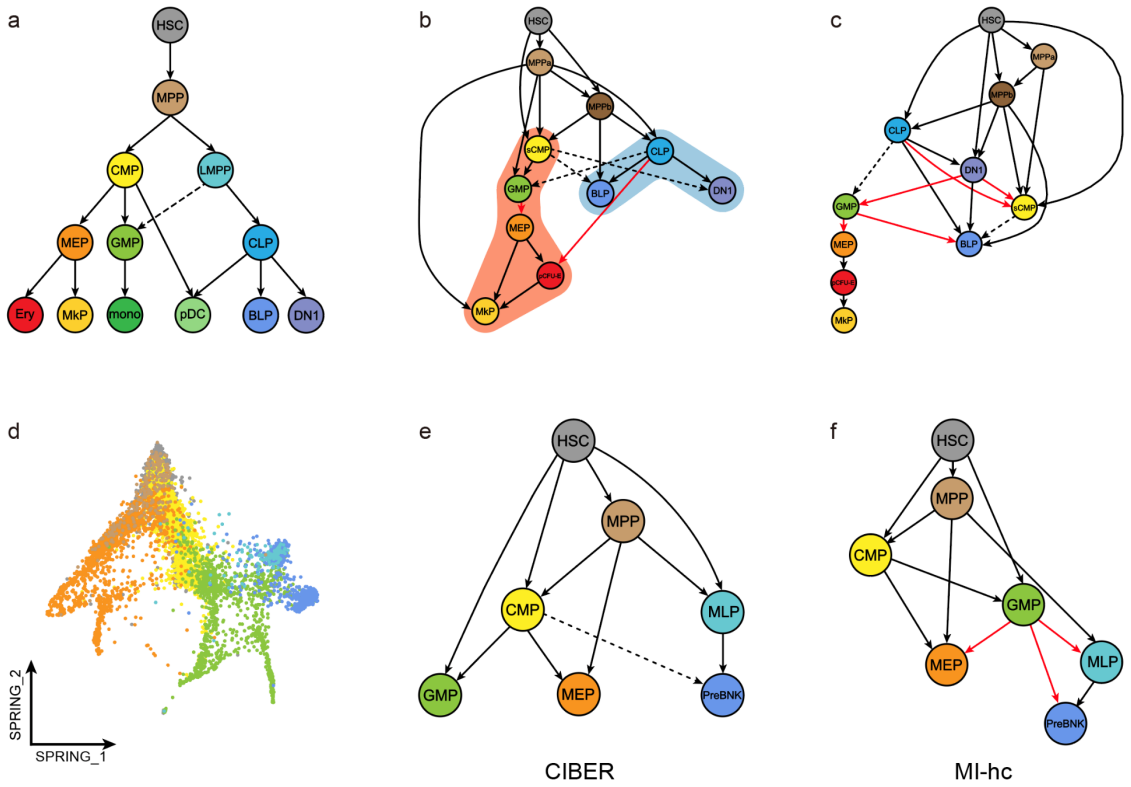


Fig. 2 | CIBER captures conventional differentiation structures in various datasets. a, Conventional haematopoietic hierarchy. **b**, CIBER result of the mouse microarray dataset, with the blue/red-shaded part representing the lymphoid/myeloid lineage. **c**, Best-tuned structure learned with MI-hc, 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[43], 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. The structural causal model (SCM) is a conventional causal inference framework that combines probabilistic graphical models and the notion of interventions[27,28]. 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 linear additive noise model, in which the functions connecting each node to its parents are limited to the form of linear

240 regression (Methods). In the SCM framework, interventions can be performed by
241 changing the form of particular functions or the distribution of specific nodes to infer the
242 effect of causality.

243 Here, we introduced the concept of SCM into the directed acyclic graph learned by
244 CIBER and performed causal inference by applying *in silico* perturbation to the
245 expression of specific features as interventions (Fig. 1). Specifically, permutation of
246 features was performed on single-cell profiling data while deletion was performed on
247 microarray or bulk sequencing data. Then, we recorded the coefficient associated with
248 each edge in the linear additive SCM with/without interventions performed to a specific
249 feature and defined an effect matrix to measure the change in the coefficient due to the
250 interventions. Intuitively, the effect matrix quantifies the impacts of the features on each
251 edge in the cellular differentiation network, with a positive value implying a possible
252 promoting function on the corresponding differentiation branch and a negative value
253 implying a possible depressing function.

254 **Differentiation driver feature (DDF) analysis identifies features determining cell**
255 **fate.** Next, we developed DDF analysis, a novel method distinct from regular differential
256 feature (RDF) analysis, to identify important features crucial to the cellular differentiation
257 process via effect matrices obtained by *in silico* perturbation. Specifically, we defined
258 diffScores to assess the effect of individual features on the cellular differentiation network
259 or specific lineages. The diffScore of a gene on the entire network was defined by
260 summarizing the absolute values related to this gene in the effect matrix, while the
261 diffScore on a specific lineage was defined as the sum of the absolute values related to
262 the edges pointing to cell types in the lineage subtracting the original values of the edges
263 pointing from the lineage to other lineages (see Methods for detailed description).

264 DDF analysis provides a framework to systematically investigate a feature's impact on
265 the entire cellular differentiation hierarchy. Our study demonstrates that DDF analysis
266 can work effectively for scenarios in which cell types are limited. Additionally, DDF
267 analysis can be combined with RDF analysis to better identify important features due to
268 the independent information captured by the two methods.

269 To examine the effectiveness of DDF analysis, we first applied it to the mouse bone
270 marrow scRNA-seq dataset and compared the results with that of the RDF analysis
271 obtained by ranking genes according to the p value of Wilcoxon's signed-rank tests. To
272 compare the results of the two analyses, we selected the same number of top genes on the
273 two lists (Fig. 3a). The results showed that a few important cell fate–determining genes
274 identified by RDF analysis, such as *Gata2*[48], were not captured by DDF analysis; on the
275 other hand, DDF analysis identified important genes that were not differentially expressed
276 between the lymphoid and myeloid cell types, such as *Mpl*[49], *Ccl9*[50] and *Elane*[51];
277 nevertheless, some genes, such as *Flt3*[52], *Cd34*[53] and *Dntt*[54], were identified by both
278 analyses. We further conducted a literature search on the top-ranked genes and calculated
279 Spearman's correlation between the two gene lists (Fig. 3b-d). The results showed that
280 the RDF and DDF analyses both captured a part of the genes that have been proven to be
281 important in haematopoiesis.

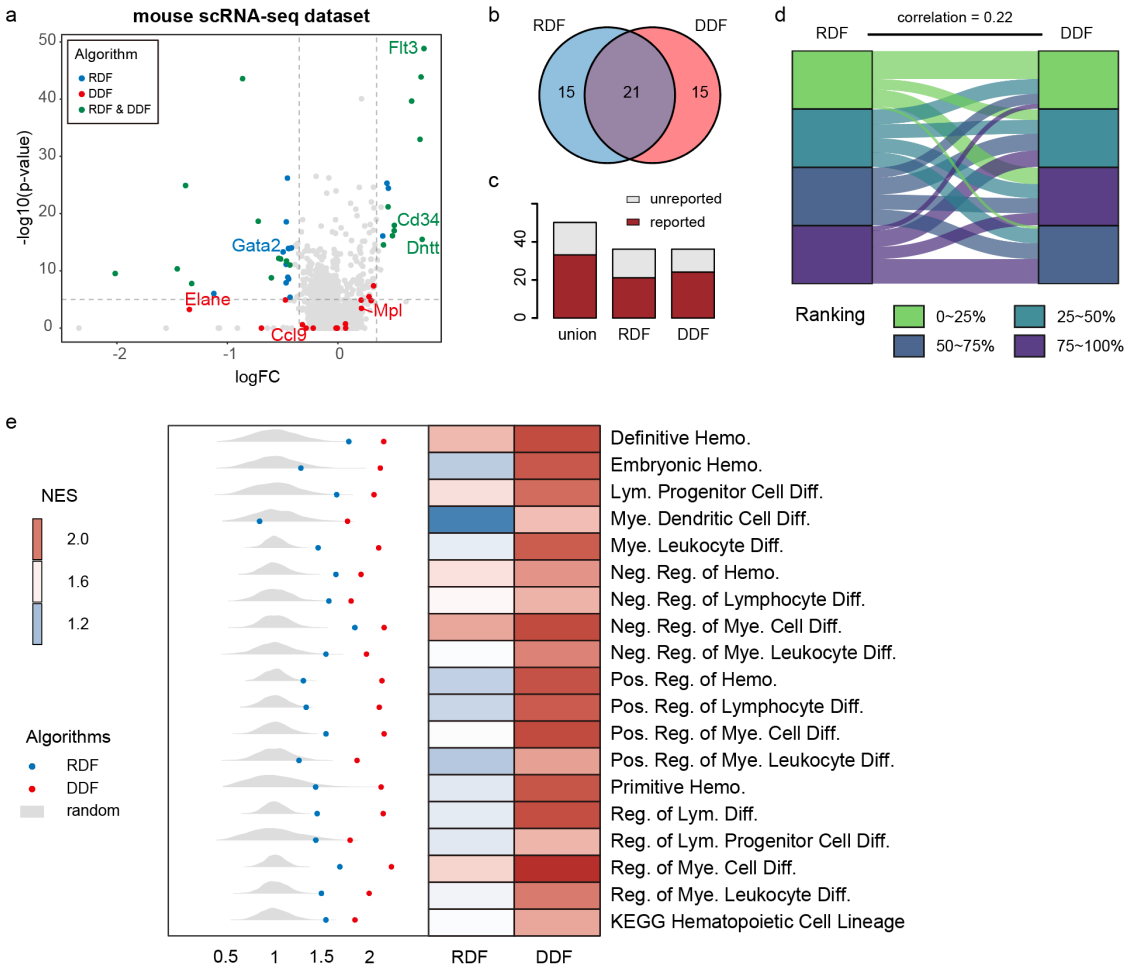


Fig. 3 | DDF identifies cell fate–determining features. The plots show the results on RDF and DDF gene lists obtained with the mouse scRNA-seq dataset. **a**, Features are allocated by log fold change and log p value obtained in the RDF analysis, from which the top differentially expressed genes are 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., Differentiation; Lym., Lymphoid; Mye., Myeloid; Reg., Regulation; Pos., Positive; Neg., Negative.)

Then, we evaluated the behaviours of RDF and DDF lists on enriching important genes related to haematopoiesis or haematopoietic differentiation. Specifically, we calculated the normalized enrichment scores (NES) through GSEA for gene lists with respect to all

302 the GO and KEGG terms related to haemopoiesis or haematopoietic differentiation (Fig.
303 3e). GSEA results showed that DDF surpassed RDF on almost all these pathways on NES.
304 Note that the dataset did not contain any lymphoid-lineage cell types other than
305 lymphomyeloid-primed progenitor (LMPP), an early lymphoid-biased progenitor. The
306 results demonstrated that DDF analysis performs well on datasets with insufficient cell
307 types, due potentially to the fact that it examines the entire cellular differentiation network
308 rather than being restricted to the difference between subgroups of cells, as RDF analysis
309 did.

310 Next, we applied the same analyses to the human scRNA-seq dataset and mouse
311 microarray dataset. Similar results were obtained for the human scRNA-seq dataset to the
312 results of mouse scRNA-seq dataset described above (Extended Data Fig. 3). The
313 microarray dataset offered us an opportunity to study the behaviour of DDF analysis in
314 scenarios where most cell types are available and are therefore friendly to RDF analysis.
315 The result showed the RDF list performed slightly better than the DDF list in 12 out of
316 22 terms (Extended Data Fig. 4e). However, the literature search still showed that DDF
317 analysis identified important genes which are not differentially expressed, demonstrating
318 that DDFs can complement RDFs to better identify genes crucial to cell differentiation
319 (Extended Data Fig. 4a-c).

320 **CIBER identifies driver TFs of biological processes**. We then applied CIBER and DDF
321 analysis to a human bone marrow scATAC-seq dataset[44], and identified TFs crucial to
322 haematopoiesis based on deviation score matrix calculated with chromVAR[55]. RDF
323 analysis was also performed to identify the TFs with significant difference in deviation
324 score between lymphoid and myeloid cell types. Top TFs identified by DDF only includes,
325 *IRF8*, *BCL11A*, *SPI1*, *SPIB* and *CEBPB*; TFs, such as *TCF3*, *LYL1* and *EBF1*, were found
326 by the RDF analysis alone, while both analyses uncovered TFs such as *GATA1*, *GATA2*,
327 *GATA3*, *ID3* and *LMO2* (Fig. 4c). With lineage-specific diffScores, DDF analysis can

328 further quantify the potential effects of each TF on promoting or suppressing specific
329 lineages (Fig. 4d).

330 TF regulates cellular function by orchestrating the transcription of downstream target
331 genes. Next, we sought to test whether the target genes of the TFs identified in the
332 scATAC-seq dataset were also enriched in the DDF or RDF analysis of an independent
333 human bone marrow scRNA-seq dataset. we collected target genes for TFs uncovered by
334 both analysis from 5 databases (TRRUST[56], JASPAR[57], ChEA[58], Motifmap[59] and
335 ENCODE[60]) and ranked the target genes according to their significance in the DDF or
336 RDF analysis of the scRNA-seq dataset. Two types of metrics, normalized gene set
337 enrichment score (NES) and area under the receiver operating characteristic curve (AUC),
338 were calculated to measure the enrichment of targeted genes by each ranking algorithm
339 (Fig. 4e, Methods). When calculating AUC, the receiver operating characteristic (ROC)
340 curve was drawn by assigning TF-targeted genes as positive and all other genes as
341 negative samples. Background distributions of NES and AUC were obtained with 1,000
342 randomly shuffled ranking lists of all the genes. Results show that under both metrics,
343 DDF surpasses RDF on all gene sets targeted by different TFs (Fig. 4e, f, Extended Data
344 Fig. 5). For example, both DDF and RDF significantly surpass the random background,
345 and DDF scores higher for the *GATA1* target gene set; in the case of *BCL11B*, only the
346 DDF score is significantly higher than the random background (DDF NES p value=0.009,
347 RDF NES p value=0.103, DDF AUC p value<0.001, RDF p value=0.683). CIBER
348 quantifies the importance of TF target genes on cellular differentiation, thus provides
349 insight into the potential mechanism impacting the cellular differentiation process.

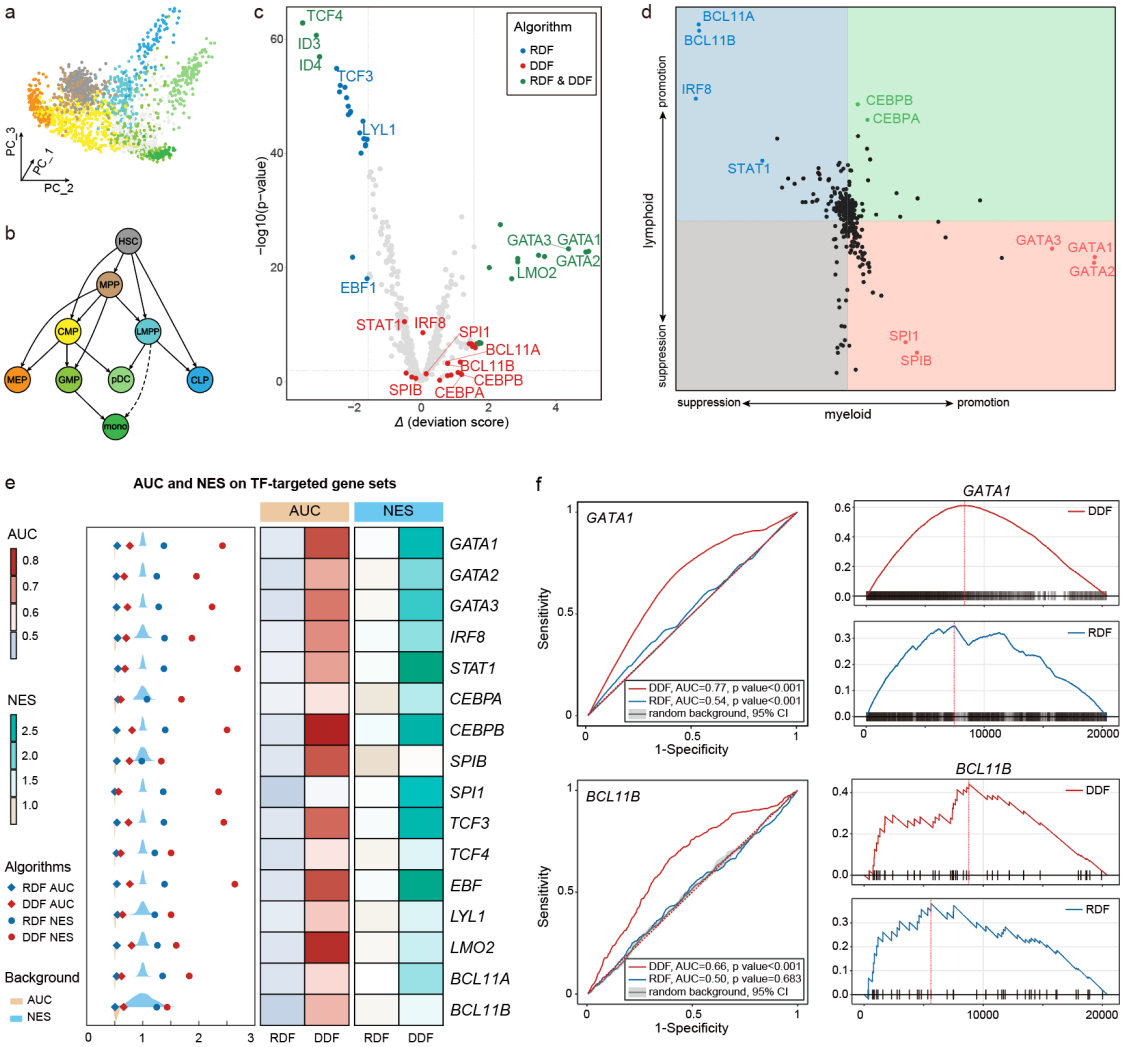


Fig. 4 | CIBER identifies driver TFs of biological processes. a, PCA reduction of a human bone marrow scATAC-seq dataset[44]. **b**, Differentiation hierarchy learned with CIBER. **c**, Top TFs identified by RDF and DDF analyses. Among them, several reported important TFs[44] are annotated. **d**, Lineage-specific 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<0.001, RDF p value=0.683). 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.

368 **In vivo experiments validate *BCL11B* as a crucial TF to fate determination in**
369 **haematopoiesis.** In addition to the previously reported TFs, we noted that DDF analysis
370 of bone marrow scATAC-seq data identified *BCL11B* which showed no significant
371 difference in deviation scores between lymphoid and myeloid cell types (Fig. 4c).
372 Lineage-specific diffScores indicate that *IRF8*, *BCL11A* and *BCL11B* promote the
373 differentiation process of the lymphoid lineage while suppress the myeloid lineage,
374 consistent with previous report of *IRF8*[61] and *BCL11A*[62,63] (Fig. 4d). GSEA and ROC
375 analyses of bone marrow scRNA-seq demonstrated that *BCL11B* target genes were
376 enriched in DDF analysis, providing further evidence that the TF may influence the
377 process of haematopoiesis (Fig. 4e, f).

378 To explore the effect of *Bcl11b* on the differentiation of haematopoietic stem cells, we
379 constructed *Bcl11b*-deficient mice (*Bcl11b*$^{flox/flox}$, Vav-iCre^{+}), hereafter named *Bcl11b*-
380 KO, to specifically knock out *Bcl11b* in the blood cells. Results showed that *Bcl11b*-KO
381 mice had a shorter survival time than WT mice in our cohort (Fig. 5b). Compared with
382 WT mice, the haematopoietic system of *Bcl11b*-KO mice after 5-fluorouracil (5-FU)
383 treatment had an increased proportion of reconstituted myeloid cells and a decreased
384 proportion of reconstituted lymphoid cells (Fig. 5d).

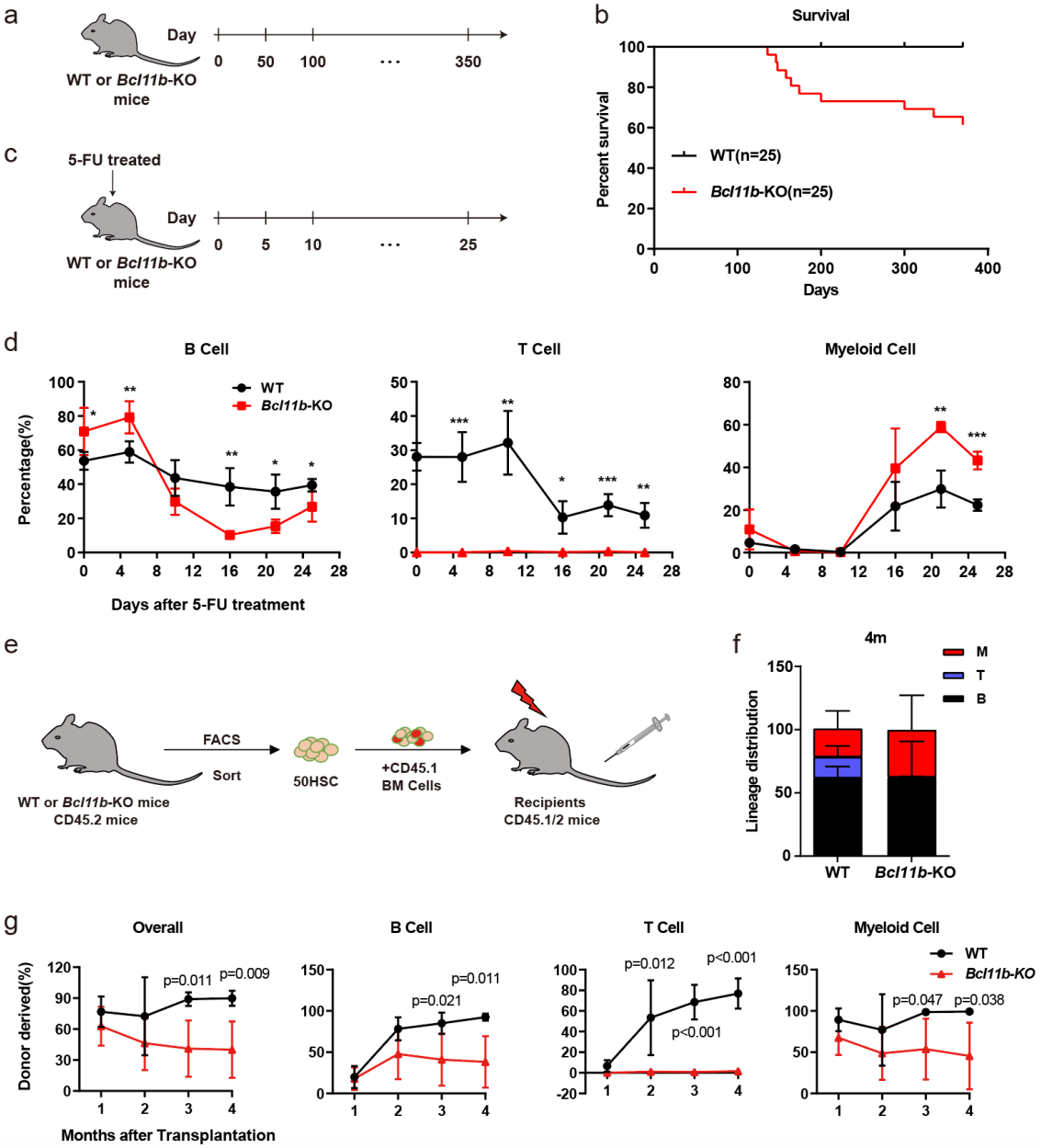


Fig. 5 | Experimental validation of CIBER predictions for *Bcl11b*. Survival experiments (**a**) and results (**b**) of WT and *Bcl11b*-KO mice in their native state show the survival capacity of *Bcl11b*-KO mice is hampered. Experiments (**c**) and percentages (**d**) of reconstituted B cells, T cells and myeloid cells in peripheral blood of WT and *Bcl11b*-KO mice after 5-FU treatment. **e**, Scheme of the transplantation strategy for *Bcl11b*-KO mice. Fifty HSCs freshly isolated from 2-month-old WT or *Bcl11b*-KO mice were transplanted into lethally irradiated recipients along with 3×10^5 competitor cells. Chimaera in peripheral blood was checked every month. **f**, Lineage distribution of B cells, T cells, and myeloid cells among donor-derived cells in peripheral blood four months after transplantation. (n=4-5 mice per group. Data are shown as the mean ± SD.). **g**, Line plots display the contribution of donor cells to peripheral blood overall (CD45.2$^+$), B cells (B220$^+$), T cells (CD3$^+$) and myeloid cells (Mac-1$^+$) every month after transplantation. (n=4-5 mice per group. Data are shown as the mean ± SD.)

398 To evaluate the impact of *Bcl11b*-deficiency on the long-term haematopoietic
399 reconstitution capacity and differentiation of HSCs, we further performed competitive
400 transplantation assays. Fifty freshly isolated HSCs from WT or *Bcl11b*-KO mice were
401 injected into lethally irradiated recipient mice together with 3×10^5 total bone marrow cells
402 as a competitor (Fig. 5e). Chimaera in peripheral blood was evaluated every month until
403 4 months. Peripheral blood analysis of recipient mice every month calculated the
404 percentage of cells reconstructed by donor-drives HSCs to all these in recipient peripheral
405 blood and showed that the reconstitution capacity of *Bcl11b*-KO HSCs was severely
406 impaired in various cells, including B, T and myeloid lineages, compared to WT
407 transplanting the same number of HSCs (Fig. 5g). At the same time, by analysing the
408 composition of donor-derived cells in recipient mice, we found that *Bcl11b*-deficient
409 HSCs enhanced myeloid differentiation and decreased T-cell differentiation (Fig. 5f). In
410 summary, the knockout of *Bcl11b* resulted in a reduced reconstitution capacity of HSCs,
411 however, in the offspring cells, the proportion of the myeloid lineage is increased
412 compared to that of the lymphoid lineage.

413 Discussion

414 Large-scale single-cell omics data have provided unprecedented opportunities to study
415 developmental pathways and have enabled the reconstruction of lineage relationships
416 between cells. Multiple computational trajectory inference methods have been developed
417 to investigate cellular developmental dynamics based on snapshot measurements. Most
418 trajectory inference methods are based on expression pattern similarities and provide cell
419 fate probabilities. However, cellular differentiation processes are complex causation-
420 driven phenomena that are naturally more suited for analysis with causal inference
421 methods. The existing approaches for inferring causality among cellular states are mainly
422 based on RNA velocity, while other methods, such as those based on Granger causality,

423 Boolean networks and Bayesian networks, have been effectively used for causality
424 inference in other biological processes, such as gene regulatory networks.

425 In this study, we present CIBER framework based on causal inference to reconstruct
426 cellular differentiation networks and identify cell fate–determining features. CIBER
427 improved the robustness of existing hybrid Bayesian network structure learning
428 algorithms to obtain feasible structures with minimal prior knowledge. The applications
429 of CIBER on several datasets of haematopoietic cells show that it is suitable to dataset
430 obtained through various techniques on different species and that it may help in
431 identifying potential differentiation branches. CIBER built an effect matrix based on the
432 concept of the structural causal model and interventions of *in silico* perturbation on the
433 expression level of specific features. The effect matrix quantifies the impacts of features
434 on differentiation branches and can be used to identify the cell fate–determining features
435 through DDF analyses developed in this study. Unlike regular differential feature (RDF)
436 analysis, which identifies markers of specific cell types, DDF analysis captures distinctive
437 information on differentiation lineages. The applications on different datasets revealed
438 that DDF analysis was effective in scenarios in which cell types were limited for RDF
439 analysis, and that DDF analysis could complement RDF analysis to identify important
440 features comprehensively.

441 CIBER differs from existing trajectory inference methods in several aspects in terms
442 of reconstructing the cellular differentiation process. Most existing trajectory inference
443 methods are suitable only for single-cell omics data on the assumption that the static
444 snapshot of the cellular states is continuously distributed, while CIBER is applicable
445 under more conditions, such as when the number of cells or cell types is limited, and is
446 even applicable to datasets obtained by microarray or bulk sequencing techniques. Except
447 for the approaches based on RNA velocity, most existing trajectory inference methods

are similarity-based, while CIBER is a causation-based framework that provides a new perspective on the cellular differentiation process.

Although CIBER can handle scenarios that existing trajectory inference methods cannot, it has several limitations. To better perform the analysis after building the effect matrix, at least two parents should exist for each child node in the obtained network. In addition, CIBER may not be suitable for cases in which the network is too simple, especially when the network is only a two-layer hierarchy, which prevents the construction of the effect matrix for the same reason as the previous problem and even impacts the robustness of BN structure learning. For extremely complex networks with too many nodes and edges, CIBER may be time consuming. Typically, CIBER is most effective for hierarchical structures with approximately 5 layers and less than 20 nodes.

Taken together, the CIBER framework provides a causal perspective for the reconstruction of cellular differentiation networks and the evaluation of how features influence differentiation branches, and DDF analysis enriches existing techniques for identifying important features and can be combined with RDF analysis to help us understand biological processes more comprehensively.

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. 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).
9. Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Nat Methods* **19**, 159–170 (2022).
10. Qiu, X. *et al.* Mapping transcriptomic vector fields of single cells. *Cell* **185**, 690-711.e45 (2022).
11. 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.
12. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
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. Deshpande, A., Chu, L.-F., Stewart, R. & Gitter, A. Network inference with Granger causality ensembles on single-cell transcriptomics. *Cell Reports* **38**, 110333 (2022).
16. 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).
17. 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).
18. Font-Clos, F., Zapperi, S. & La Porta, C. A. M. Classification of triple-negative breast cancers through a Boolean network model of the epithelial-mesenchymal transition. *Cell Systems* **12**, 457-462.e4 (2021).
19. Hamey, F. K. *et al.* Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. *Proc Natl Acad Sci USA* **114**, 5822–5829 (2017).
20. Weidner, F. M. *et al.* Capturing dynamic relevance in Boolean networks using graph theoretical measures. *Bioinformatics* **37**, 3530–3537 (2021).
21. Friedman, N. Inferring Cellular Networks Using Probabilistic Graphical Models. *Science* **303**, 799–805 (2004).
22. Jansen, R. A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science* **302**, 449–453 (2003).

23. 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).
24. 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).
25. Schütte, J. *et al.* An experimentally validated network of nine haematopoietic transcription factors reveals mechanisms of cell state stability. *eLife* **5**, e11469 (2016).
26. 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).
27. Pearl, J. Causal diagrams for empirical research. *Biometrika* **82**, 669–688 (1995).
28. Schölkopf, B. *et al.* Toward Causal Representation Learning. *Proceedings of the IEEE* **109**, 612–634 (2021).
29. Korb, K. & Nicholson, A. Bayesian Artificial Intelligence. in *Computer science and data analysis series* (2004). doi:10.5860/choice.41-5948.
30. Scutari, M. & Denis, J.-B. *Bayesian Networks: With Examples in R*. (Chapman and Hall/CRC, 2021). doi:10.1201/9780429347436.
31. Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. (MIT Press, 2009).
32. Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. (Morgan kaufmann, 1988).
33. Akaike, H. A new look at the statistical model identification. *IEEE transactions on automatic control* **19**, 716–723 (1974).
34. Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics* **6**, 461–464 (1978).
35. 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).
36. 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).
37. Zhang, M. *et al.* GeNeCK: a web server for gene network construction and visualization. *BMC Bioinformatics* **20**, 12 (2019).
38. Scutari, M. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software* **35**, 1–22 (2010).
39. Seita, J. *et al.* Gene Expression Commons: an open platform for absolute gene expression profiling. *PLoS One* **7**, e40321 (2012).
40. 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).
41. Ambrosi, T. H. *et al.* Aged skeletal stem cells generate an inflammatory degenerative niche. *Nature* **597**, 256–262 (2021).
42. Nestorowa, S. *et al.* A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**, e20–e31 (2016).

571 43. Pellin, D. *et al.* A comprehensive single cell transcriptional landscape of
572 human hematopoietic progenitors. *Nature Communications* **10**, 2395
573 (2019).
574 44. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the
575 Continuous Regulatory Landscape of Human Hematopoietic Differentiation.
576 *Cell* **173**, 1535-1548.e16 (2018).
577 45. Graf, T. Differentiation plasticity of hematopoietic cells. *Blood* **99**, 3089–
578 3101 (2002).
579 46. Akashi, K., Traver, D., Miyamoto, T. & Weissman, I. L. A clonogenic
580 common myeloid progenitor that gives rise to all myeloid lineages. *Nature*
581 **404**, 193–197 (2000).
582 47. Prohaska, S. S., Scherer, D. C., Weissman, I. L. & Kondo, M.
583 Developmental plasticity of lymphoid progenitors. *Seminars in Immunology*
584 **14**, 377–384 (2002).
585 48. Single-cell ATAC-seq reveals GATA2-dependent priming defect in myeloid
586 and a maturation bottleneck in lymphoid lineages | Blood Advances |
587 American Society of Hematology.
588 https://ashpublications.org/bloodadvances/article/5/13/2673/475734/Single-
589 cell-ATAC-seq-reveals-GATA2-dependent.
590 49. The cytoplasmic domain of Mpl receptor transduces exclusive signals in
591 embryonic and fetal hematopoietic cells | Blood | American Society of
592 Hematology.
593 https://ashpublications.org/blood/article/100/6/2063/235068/The-
594 cytoplasmic-domain-of-Mpl-receptor-transduces.
595 50. ICSBP-mediated immune protection against BCR-ABL–induced leukemia
596 requires the CCL6 and CCL9 chemokines | Blood | American Society of
597 Hematology.
598 https://ashpublications.org/blood/article/113/16/3813/25078/ICSBP-
599 mediated-immune-protection-against-BCR-ABL.
600 51. CRISPR Mediated ELANE Single-Allele Knock-out Restores Proliferation
601 and Myeloid Differentiation of Neutropenia Patient Derived BM HSCs |
602 Blood | American Society of Hematology.
603 https://ashpublications.org/blood/article/136/Supplement%201/23/470437/C
604 RISPR-Mediated-ELANE-Single-Allele-Knock-out.
605 52. Sitnicka, E. *et al.* Key role of flt3 ligand in regulation of the common
606 lymphoid progenitor but not in maintenance of the hematopoietic stem cell
607 pool. *Immunity* **17**, 463–472 (2002).
608 53. Okuno, Y. *et al.* Differential regulation of the human and murine CD34
609 genes in hematopoietic stem cells. *Proceedings of the National Academy of*
610 *Sciences* **99**, 6246–6251 (2002).
611 54. Ng, S. Y.-M., Yoshida, T., Zhang, J. & Georgopoulos, K. Genome-wide
612 Lineage-Specific Transcriptional Networks Underscore Ikaros-Dependent
613 Lymphoid Priming in Hematopoietic Stem Cells. *Immunity* **30**, 493–507
614 (2009).
615 55. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR:
616 inferring transcription-factor-associated accessibility from single-cell
617 epigenomic data. *Nat Methods* **14**, 975–978 (2017).

56. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research* **46**, D380–D386 (2018).
57. 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).
58. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438–2444 (2010).
59. Xie, X., Rigor, P. & Baldi, P. MotifMap: a human genome-wide map of candidate regulatory motif sites. *Bioinformatics* **25**, 167–174 (2009).
60. Consortium, T. E. P. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLOS Biology* **9**, e1001046 (2011).
61. Differential IRF8 Transcription Factor Requirement Defines Two Pathways of Dendritic Cell Development in Humans. *Immunity* **53**, 353-370.e8 (2020).
62. Luc, S. *et al.* Bcl11a Deficiency Leads to Hematopoietic Stem Cell Defects with an Aging-like Phenotype. *Cell Reports* **16**, 3181–3194 (2016).
63. 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).

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 probes. For each gene detected by multiple probes, the median value was adopted as its expression level. The top variable features were first selected for Bayesian network structure learning; 3,000 features were selected in the mouse microarray dataset, 5,000 in the mouse bone marrow scRNA-seq dataset, 4,000 in the human bone marrow scRNA-seq dataset and 1,764 (all transcription factors) in the human bone marrow scATAC-seq dataset.

Construction of cellular differentiation networks. In general, Bayesian network structure learning is thought to be sensitive to both observations and parameters, and small changes in input can lead to considerably different outputs. Hence, expert

653 experience is needed as prior knowledge to obtain a structure applicable for further study.
654 In CIBER framework, we adopted several strategies to ensure a well-structured hierarchy
655 with minimal prior knowledge of assigning credible upper and lower boundaries for the
656 number of edges in the network and setting stem cells as the root.

657 CIBER utilizes a multistep strategy to obtain robust BN structures — resampling,
658 undirected skeleton learning, DAG learning, combining DAGs and trimming branches.
659 Particularly, for datasets with limited biological replicates, such as microarray datasets,
660 the resampling step is omitted. By combining the DAGs obtained with different
661 resampled subsets of single cells and various parameters, we can extract the most robust
662 associations between variables, which have a higher probability of being the true
663 causation. The details of the steps are described below.

664 **Resampling strategy.** The resampling strategy maximizes the efficacy of large single-
665 cell omics datasets. A stratified resampling strategy was performed for the single-cell
666 omics datasets to generate a subset of cells, while this strategy was omitted for the
667 microarray dataset. For the microarray dataset and each subset of single cells after
668 resampling, we calculated the mean expression level of each cell type to obtain a mean
669 expression matrix, with each column representing a cell type and each row representing
670 a specific feature. The resampling ratio can be set to approximately 0.3 to prevent high
671 randomness or homogeneity between the obtained subsets. Similarly, the cell number of
672 each cell type was suggested to be greater than 50 to prevent random effects.

673 **Undirected skeleton learning.** With the mean expression matrix as an input, an
674 undirected skeleton was built to describe the correlation between each pair of cell types,
675 with each node representing a cell type and each edge representing the correlation
676 between the connected cell types. Various metrics can be utilized to evaluate the
677 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[37]. The default algorithm used in CIBER is CMI2NI[35], 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[38] package to obtain the DAGs. The hill climbing algorithm was used to search for the DAG with the highest Bayesian information criterion (BIC)[34]. 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.

727 This framework considers a set of associated observables with the nodes of a directed
728 acyclic graph while viewing each node X_i as a function of its parents $\mathbf{PA}_i$ in the graph:

729 $$X_i = h_i(\mathbf{PA}_i, U_i),$$

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

732 $$X_i = \sum_k \beta_{ik}\mathbf{PA}_{ik} + U_i.$$

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

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

745 For the single-cell omics dataset, we used a resampling strategy similar to that
746 performed in reconstructing networks to record the distribution of the coefficient
747 associated with a particular edge e_0 obtained without and with interventions on a specific
748 feature f_0, which we termed α_{ref} and α_{inter}, respectively, where α represents the

749 abovementioned β_{ik}. For one resampling and intervention, the matrix $\mathrm{EffectMat}_{\mathrm{sampling}}$
750 is defined as

751 $$\mathrm{EffectMat}_{\mathrm{samping}}[f_0, e_0] = \mathrm{effect}(f_0, e_0) = \alpha_{\mathrm{ref}} - \alpha_{\mathrm{inter}},$$

752 where $\mathrm{EffectMat}_{\mathrm{samping}}[f_0, e_0]$ is the row f_0 and column e_0 element in the matrix, and
753 $\mathrm{effect}(f_0, e_0)$ is the effect of feature f_0 on edge e_0. By resampling and permuting several
754 times, we obtain a series of $\mathrm{EffectMat}_{\mathrm{samping}}$ matrices, which can be horizontally merged
755 to generate a final effect matrix in which the rows consist of all features and the columns
756 consist of all edges multiplied by the resampling and permutation times. For microarray
757 or bulk sequencing datasets, the effect matrix is constructed with the same procedures,
758 except that the resampling step is omitted and the intervention is performed with deletion
759 instead of permutation.

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

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

768 $$\mathrm{diffSum} = \sum_{e \in G} |\mathrm{effect}(f_0, e)|.$$

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

770 $$\mathrm{diffScore}_{\mathrm{lym}} = \sum_{e \in \{S \to L\} \cup \{L \to L\}} |\mathrm{effect}(f_0, e)| - \sum_{e \in \{L \to M\}} \mathrm{effect}(f_0, e)$$

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.

791 **Competitive HSC transplantation and analysis.** *Bcl11b*$^{flox/flox}$ mice were provided by
792 Dr. Yeguang Chen from Tsinghua University. In all experiments (unless otherwise
793 specified), mice were homozygous for the *Bcl11b* allele. *Bcl11b*-KO (C57BL/6, CD45.2)
794 mice, C57BL/6-SJL (CD45.1) (Stock No: 002014) and C57BL/6 WT (CD45.2) (Stock
795 No: 000664) mice were obtained from the Jackson Laboratory. CD45.1/2 mice were
796 heterozygotes from CD45.1 and CD45.2 mice. All mice were kept in specific-pathogen-
797 free (SPF), AAALAC-accredited animal care facilities at the Laboratory Animal
798 Research Center, Tsinghua University and all procedures were approved by the
799 Institutional Animal Care and Use Committee of Tsinghua University.

800 For the primary competitive HSC transplantation assay, 30 freshly isolated HSCs from
801 WT or *Bcl11b*-KO mice (CD45.2) were transplanted into lethally irradiated (10 Gy)
802 recipients (CD45.1/2) together with 3×10^5 competitor cells (CD45.1). Donor-derived
803 chimaerism (including B cells, T cells and myeloid cells) in the peripheral blood of
804 recipients was analysed at monthly intervals.

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

809 Data availability

810 The datasets used in this work were downloaded. Mouse microarray dataset[39–41] can be
811 found at GEO (GSE34723). Mouse bone marrow scRNA-seq data[42] can be found at GEO
812 (GSE81682). Human bone marrow scRNA-seq data[43] can be found at GEO (GSE117498).
813 Human bone marrow scATAC-seq data[44] can be found at GEO (GSE96772).

814 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 the High-Performance Computation platforms of Tsinghua University for providing computational resources.

Author Contributions

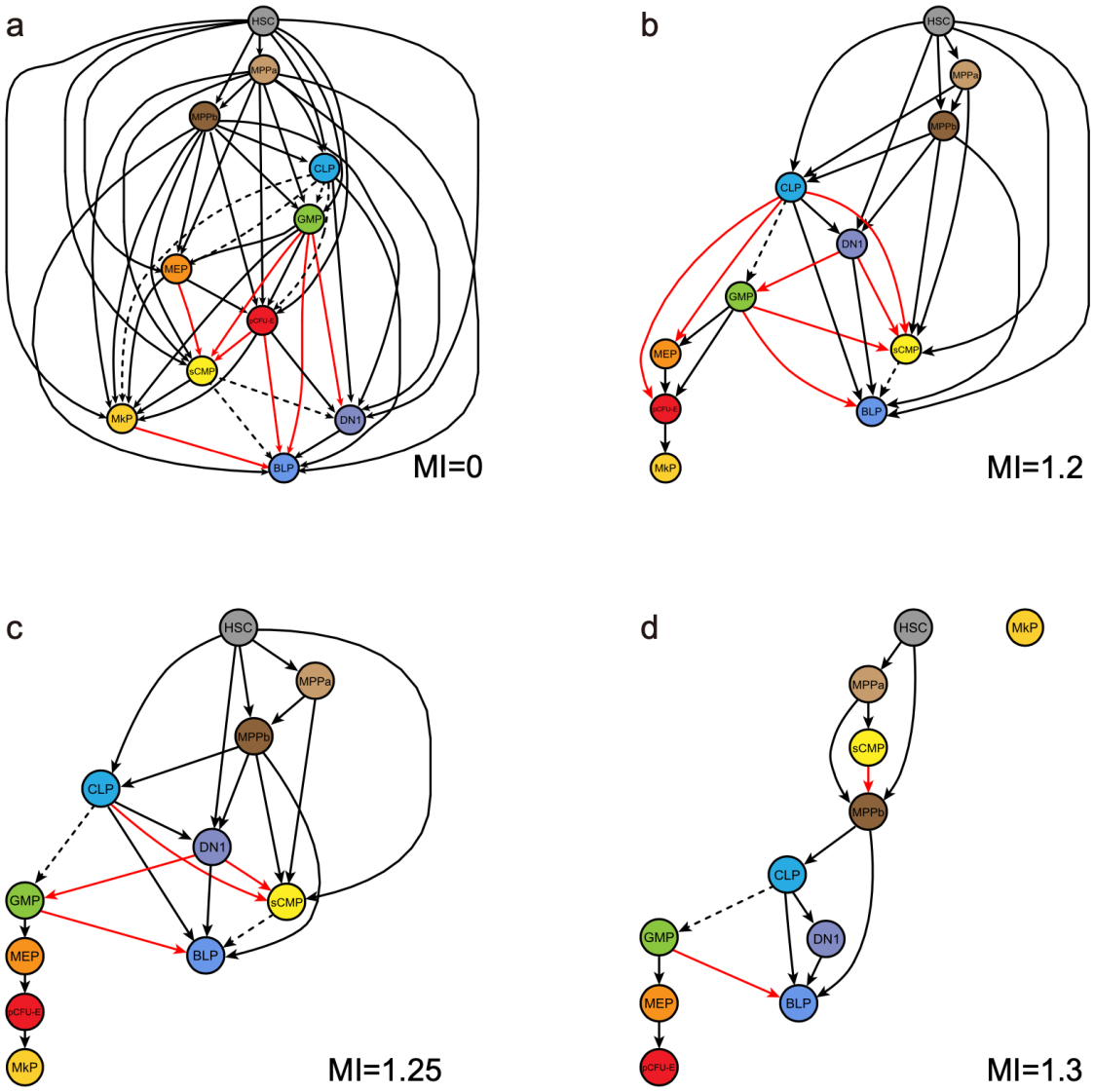
Funding

This work was supported by National Natural Science Foundation of China (Grant No. 81972680 to X.L.), Beijing Natural Science Foundation (Grant No. 20201100463 to X.L.), Tsinghua University-Peking University Jointed Center for Life Sciences (Grant No. 61020100119 to X.L.) and Alibaba innovation research programme.

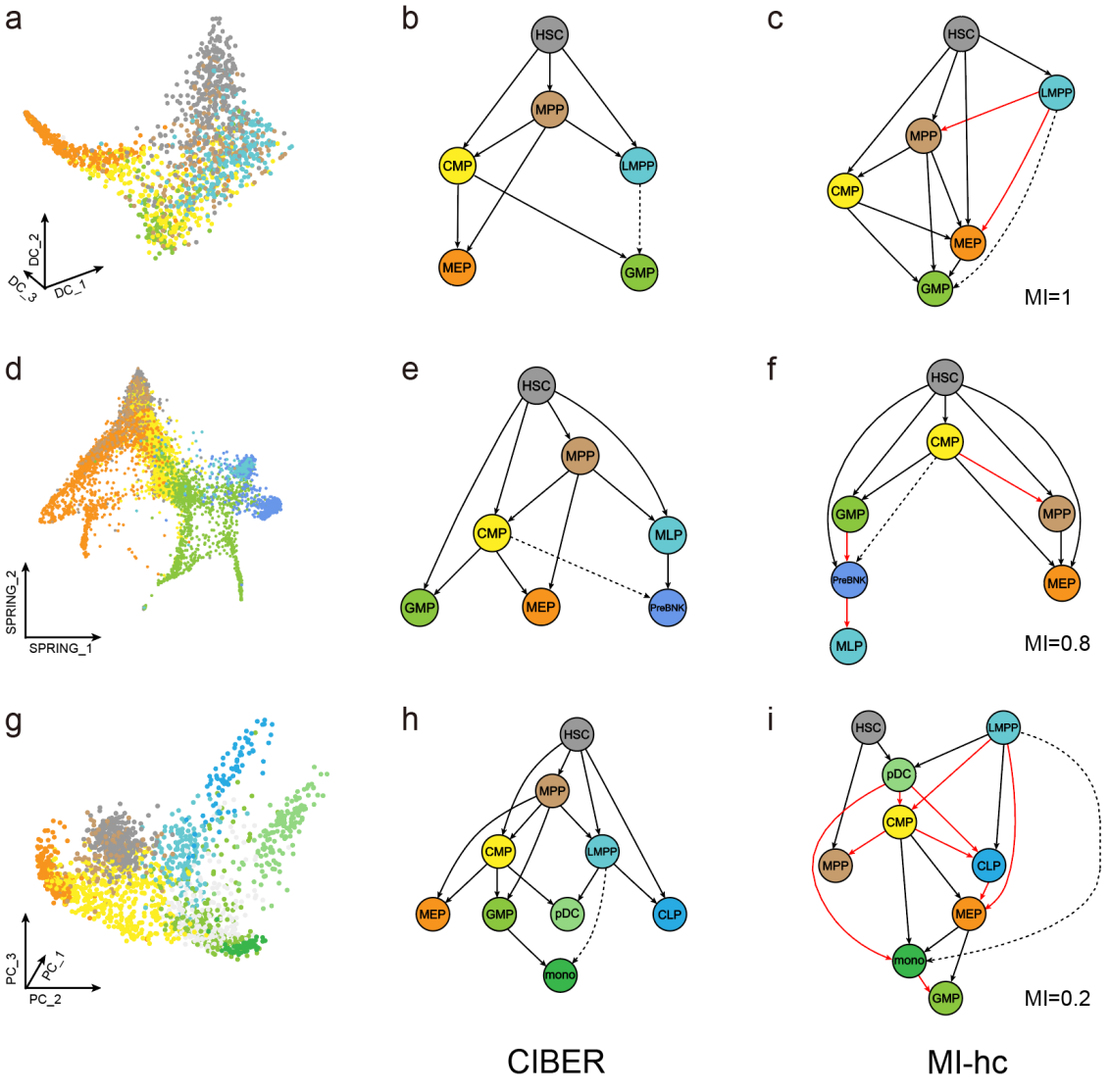
Competing interests

The authors declare no competing interests.

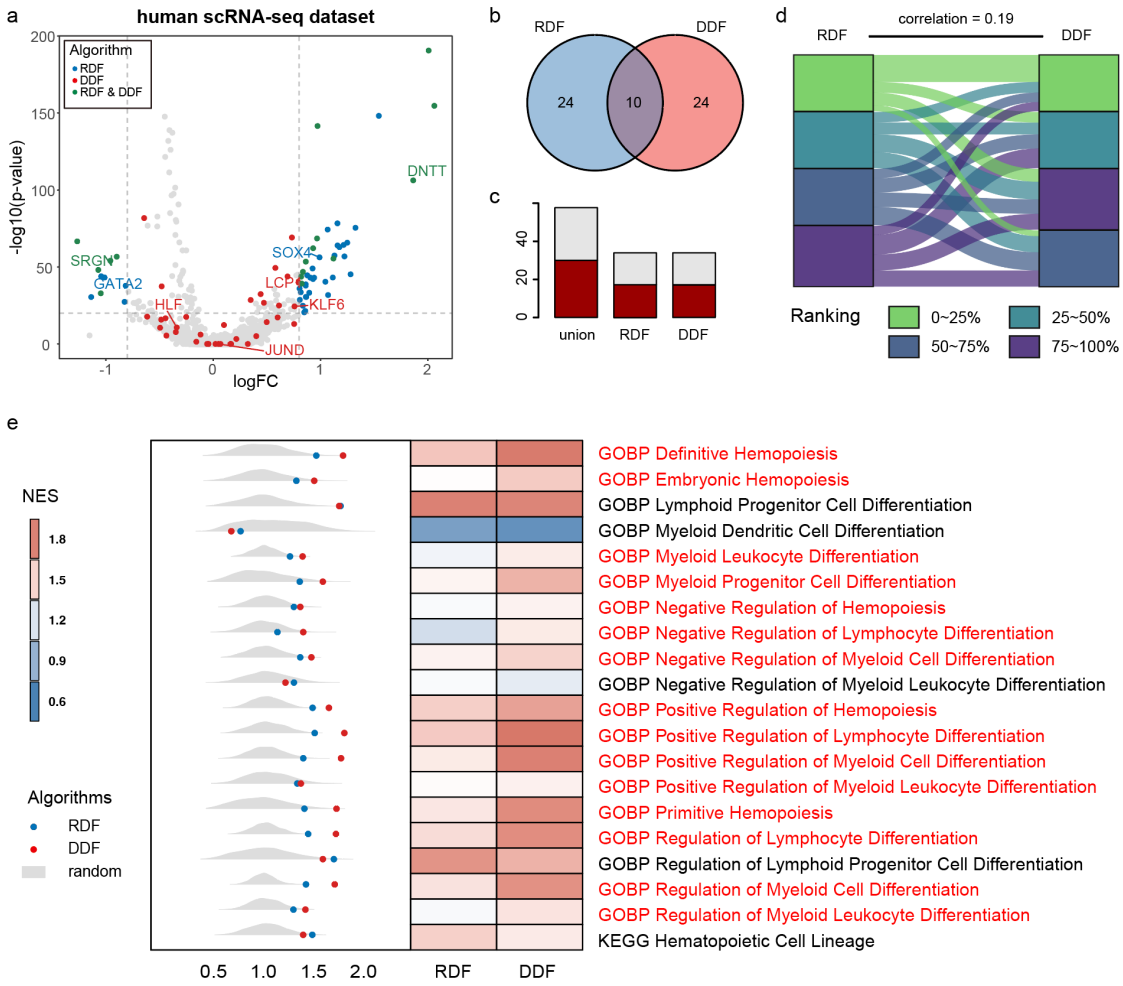
Extended data



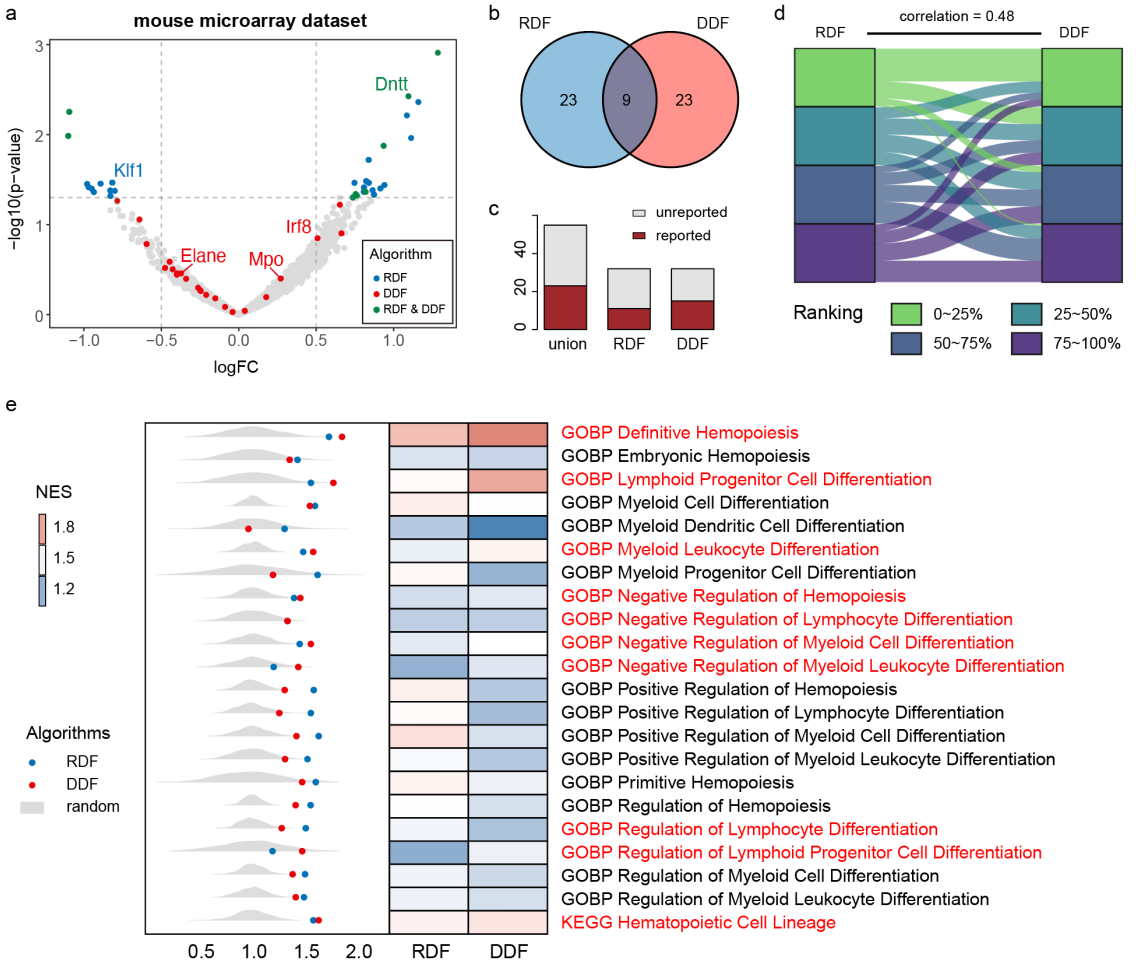
Extended Data Fig. 1 | Typical Bayesian network structure learning algorithm is sensitive to parameters. a-d, Networks learned by the typical Bayesian network structure learning strategy MI-hc given different correlation thresholds with the same microarray dataset. The dashed edges represent possible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



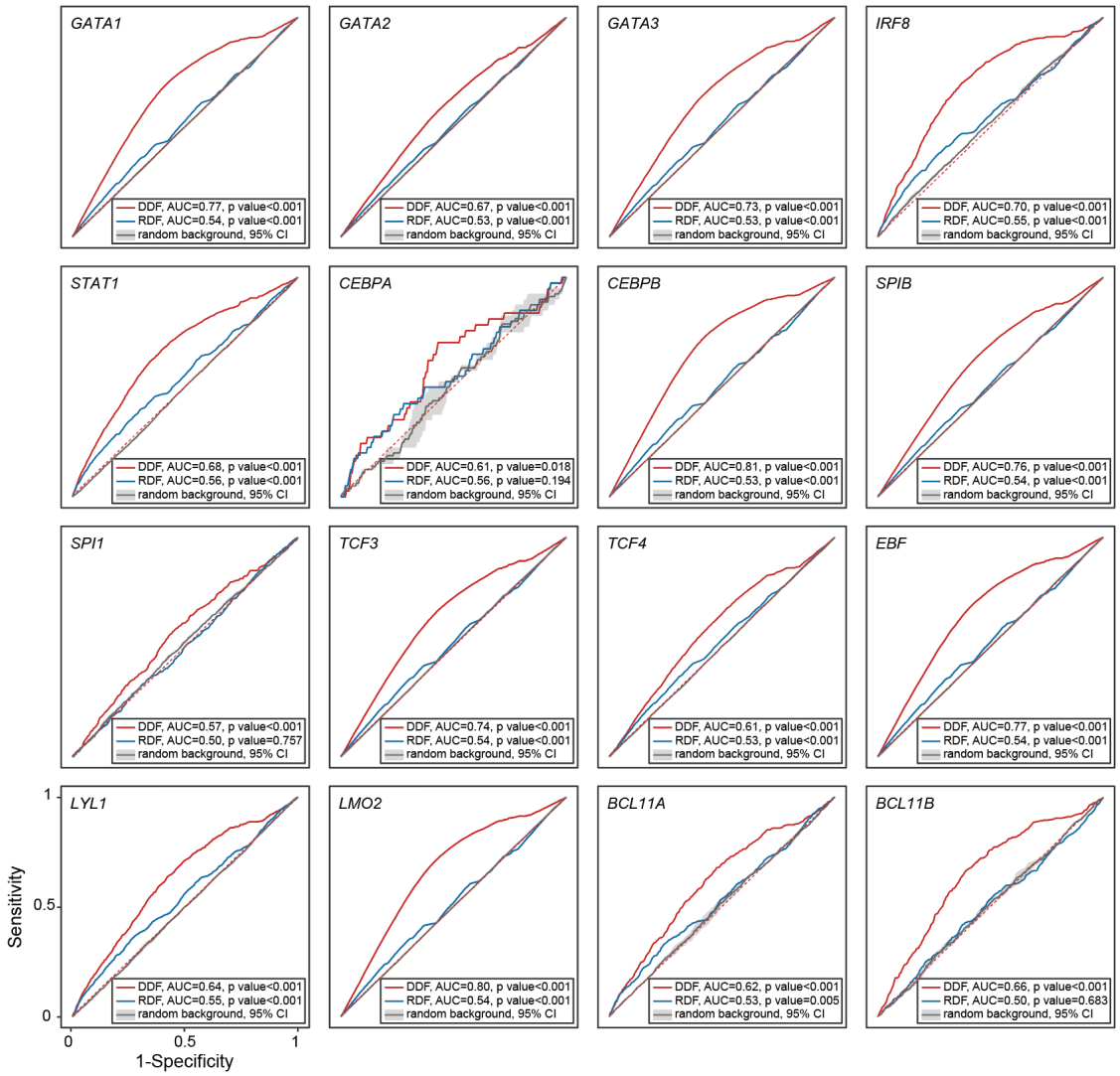
Extended Data Fig. 2 | CIBER-learned structures conform to prior knowledge and reduction landscapes better than the MI-hc algorithm. Plots show low-dimensional landscapes[42–44], CIBER and MI-hc results of (**a-c**) mouse scRNA-seq[42], (**d-f**) human scRNA-seq[43] and (**g-i**) human scATAC-seq[44] 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 DDF ranking list obtain the higher NES are coloured in red. 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 DDF list obtain the higher NES are coloured in red. 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 of each TF-targeted gene set.