Article

Multi-omics with dynamic network biomarker algorithm prefigures organ-specific metastasis of lung adenocarcinoma

Xiaoshen Zhang[1,2,3,6], Kai Xiao [4,6], Yaokai Wen[1,2,6], Fengying Wu[2], Guanghui Gao[2], Luonan Chen [4,5,7] ✉ & Caicun Zhou [2,7] ✉

Efficacious strategies for early detection of lung cancer metastasis are of significance for improving the survival of lung cancer patients. Here we show the marker genes and serum secretome foreshadowing the lung cancer site-specific metastasis through dynamic network biomarker (DNB) algorithm, utilizing two clinical cohorts of four major types of lung cancer distant metastases, with single-cell RNA sequencing (scRNA-seq) of primary lesions and liquid chromatography-mass spectrometry data of sera. Also, we locate the intermediate status of cancer cells, along with its gene signatures, in each metastatic state trajectory that cancer cells at this stage still have no specific organotropism. Furthermore, an integrated neural network model based on the filtered scRNA-seq data is successfully constructed and validated to predict the metastatic state trajectory of cancer cells. Overall, our study provides an insight to locate the pre-metastasis status of lung cancer and primarily examines its clinical application value, contributing to the early detection of lung cancer metastasis in a more feasible and efficacious way.

Lung cancer is the malignant tumor with the highest global mortality rate, with over 1.5 million deaths annually. More than 85% of the pathological types of lung cancer are non-small-cell lung cancer (NSCLC)[1], among which lung adenocarcinoma (LUAD) accounts for more than 40%[2]. Metastasis, the growth of cancer cells in organs distant from the one in which they originated, is the ultimate and most lethal manifestation of cancer, which is also the greatest contributor to death from cancer. For lung cancer, it is reported that the 5-year overall survival of patients combined with regional metastasis is less than around 20%, while fewer than 10% of patients with distant disease can survive for 5 years[3]. Consequently, effective strategies for early detection and

timely treatment of lung cancer metastasis are of significance for improving the survival of lung cancer patients.

Metastasis is an evolutionary process that involves multiple stages, such as the spread of cancer cells from the primary tumor and then intruding into the blood or lymphatic system, survival in the bloodstream and/or lymphatic system, embedding into a distant organ, survival in a new environment, and formation of a new metastatic tumor[4]. Clinically, the standard detection approaches for lung cancer metastasis are still imaging techniques. Nevertheless, lung cancer cells can metastasize to other sites even before the diagnosis of a primary tumor. Previous studies indicated that before the primary tumor diagnosis, the time taken for tumor cells to migrate to lymph

[1]School of Medicine, Tongji University, 200092 Shanghai, China. [2]Department of Medical Oncology, Shanghai Pulmonary Hospital, Tongji University School of Medicine, 200433 Shanghai, China. [3]Department of Respiratory Medicine, Shanghai Sixth People's hospital affiliated to Shanghai Jiao Tong University School of Medicine, 200233 Shanghai, China. [4]Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, 201100 Shanghai, China. [5]Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, 310024 Hangzhou, China. [6]These authors contributed equally: Xiaoshen Zhang, Kai Xiao, Yaokai Wen. [7]These authors jointly supervised this work: Luonan Chen, Caicun Zhou.
✉e-mail: lnchen@sibcb.ac.cn; caicunzhoudr@163.com

nodes was at an average of 4.26 ± 0.74 years[5]. For pleura metastases and distant metastases, the average dissemination time was -2.11 ± 0.33 years before the detection of the primary tumor[5]. These results were consistent with those of a previous pan-cancer study including breast, colorectal, and lung cancer, which indicated a seeding age of 3.6 years for overall lung cancer metastasis[6]. Therefore, it is important to identify lung cancer metastasis as early as possible, even when cancer cells are still in a pre-metastatic state without specific organotropism, which is difficult to achieve through conventional methods. Also, it is significant to capture early signals indicating cancer metastasis through a more clinically accessible approach due to the concealment of the disease.

Hematogenous routes account for a large proportion of bone and brain metastases of lung cancer[7,8]. Meanwhile, with the rapid development of liquid biopsy in recent years, various studies aimed to investigate potential biomarkers in peripheral blood for the prediction of lung cancer progression and metastasis. Previous studies have indicated that cancer cells would secrete some proteins to help create pre-metastatic niches in specific organs through peripheral blood, which could contribute to their subsequent metastasis[9]. For lung cancer, it has been reported that a significantly positive correlation was observed between serum myelin basic protein and lung cancer brain metastasis[10]. Also, increased serum levels of carcinoembryonic antigen (CEA) and osteopontin (OPN) were considered early warning signs of bone metastasis in lung cancer patients[11]. Another study enrolling 11 serum indicators were identified as independent predictive factors for NSCLC bone metastasis, and a random forest model based on them could predict the occurrence of bone metastasis -10.27 ± 3.58 months in advance in the prospective validation cohort[12]. Nonetheless, it was worth noting most studies were cross-sectional, with a small sample size, and thus, their predication efficacy should be further validated in prospective, large-scale cohort studies. In addition, most studies only evaluated certain types of serum markers based on current literature or examinations, while comprehensive profiling of the serum proteins was still absent, which limited our full understanding of the role of serum secretome promoting lung cancer organ-specific metastasis.

To figure out this question, we overlapped the data in two clinical cohorts, including four major types of lung cancer distant metastases, with single-cell RNA sequencing (scRNA-seq) of primary lesions and liquid chromatography–mass spectrometry (LC–MS) data of sera. Then, we introduced the dynamic network biomarker (DNB) method, which is an approach used for mathematically modeling gene expression networks on the basis of a temporally expressed sequence that can identify biomarkers for the early detection of the pre-metastatic state of lung cancer, on the basis of the theory that disease progresses through three states, namely, the normal state, the pre-disease state, and the disease state[13–15]. The pre-disease state is an unstable/critical state in which the normal state is changing into the disease state. At this time, gene expression levels and gene network structures have changed dramatically. DNB genes are at the core of these gene networks[16]. The DNB method has been used in studies of several fields, for instance, colorectal and liver cancer metastasis, the epithelial–mesenchymal transformation, lung adeno-squamous trans-differentiation, and breast cancer[17–21]. Compared with traditional molecular biomarkers that are used to detect disease states on the basis of their differential molecular levels measured at a single time point, the DNB method integrates temporal information, and this case is chosen for its superiority in identifying the critical time before the occurrence of the lung cancer metastasis[22]. The biomarkers inferred by the DNB system can foreshadow the initiation of distant metastasis, protruding the dramatic transition from the lung cancer in situ state to the organotropic metastasis state. Therefore, the DNB method can reveal the key genes with changed expression before lung cancer metastasis and biomarkers essential for early diagnosis of lung cancer in situ with potential to metastasize. By integrating the results of two

clinical cohorts, we can further obtain the serum key proteins indicating organ-specific metastasis, which can be easily detected through standard lab tests and translated into a clinically feasible prediction tool.

Most studies only focused on the disease state after cancer metastasis, while there was a lack of in-depth analysis of the pre-metastatic state of cancer. This is partly due to that the time evolution of many biological systems is not always smooth but occasionally abrupt[23]. Thus, the disease progression can be viewed as the evolution of a nonlinear dynamical system with a tipping point occurring after progressive changes in a certain organ or the whole organism. It is difficult to analyze through conventional biomarkers analysis that identifies the disease state by exploiting the information of the differential expressions of genes/proteins between the normal and disease states rather than signaling the pre-disease state since it is generally similar to the normal state in terms of phenotypes and gene expressions[23]. In contrast, the DNB approach has a solid background in nonlinear dynamical systems theory and can quantify the tipping point in an accurate manner due to their necessary conditions mathematically derived from the generic features of the tipping points corresponding to codimension-one local bifurcat ions[16]. Thus, the pre-metastatic state identified by the DNB method can update our understanding of cancer metastasis prediction, and it may help to diagnose and treat the disease in the ultra-early stage.

In this work, by observing the expression trend of key genes during the metastatic state trajectories, we determine the pre-metastatic state of each metastasis and depict the corresponding transcriptome features at the single-cell level. Furthermore, we successfully construct and validate an integrated single-cell classification system for the prediction of metastatic sites based on neural networks, and we further validate the results by integrated model classifier methods, which can more accurately distinguish the specific metastatic site of lung cancer.

Results

Detecting DNB protein modules of metastatic secretome prefiguring the organ-specific metastasis in LUAD through serum LC–MS

As mentioned, to identify characteristic metastatic secretome that could predict organ-specific metastasis, we performed LC–MS on sera from 117 LUAD patients, with or without metastases to four organs, including bone, lung (intrapulmonary metastasis), pleura, and brain (Fig. 1a). Those without metastasis were marked as "stage III" patients, and patients with brain or bone metastasis while also combined with concurrent metastasis were considered as "Brainplus" or "Boneplus" in our study. Overall, we detected 1492 proteins in sera from lung cancer patients, among the total 2125 secreted proteins (Fig. 1b), and the proteomic intensity across serum samples for patients with different metastases was illustrated, with no significant difference among the groups (Fig. 1c). With the DNB approach, several serum secreted proteins corresponding to each specific metastatic site were identified, with high DNB scores indicating the serum expression level of these protein modules fluctuated dramatically in patients with corresponding metastasis compared with those with other metastases or at stage III (left in Fig. 1d–h; Supplementary data 1). It has to be mentioned that lung cancer cells usually metastasize to the lung initially and then to other organs. Then, the KEGG annotation for these organ-specific DNB modules was also performed, with distinct signal pathways identified in different metastases (right in Fig. 1d–h). Furthermore, the protein set enrichment analysis indicated that these detected 1492 proteins could be clustered into eight blocks (C1–C8) (left in Fig. 2), with distinct involved biological processes and functions annotated by GO and KEGG analysis (right in Fig. 2). Interestingly, proteins in the DNB modules related to identical metastasis state were clustered into the same block, which were labeled alongside (left in Fig. 2).
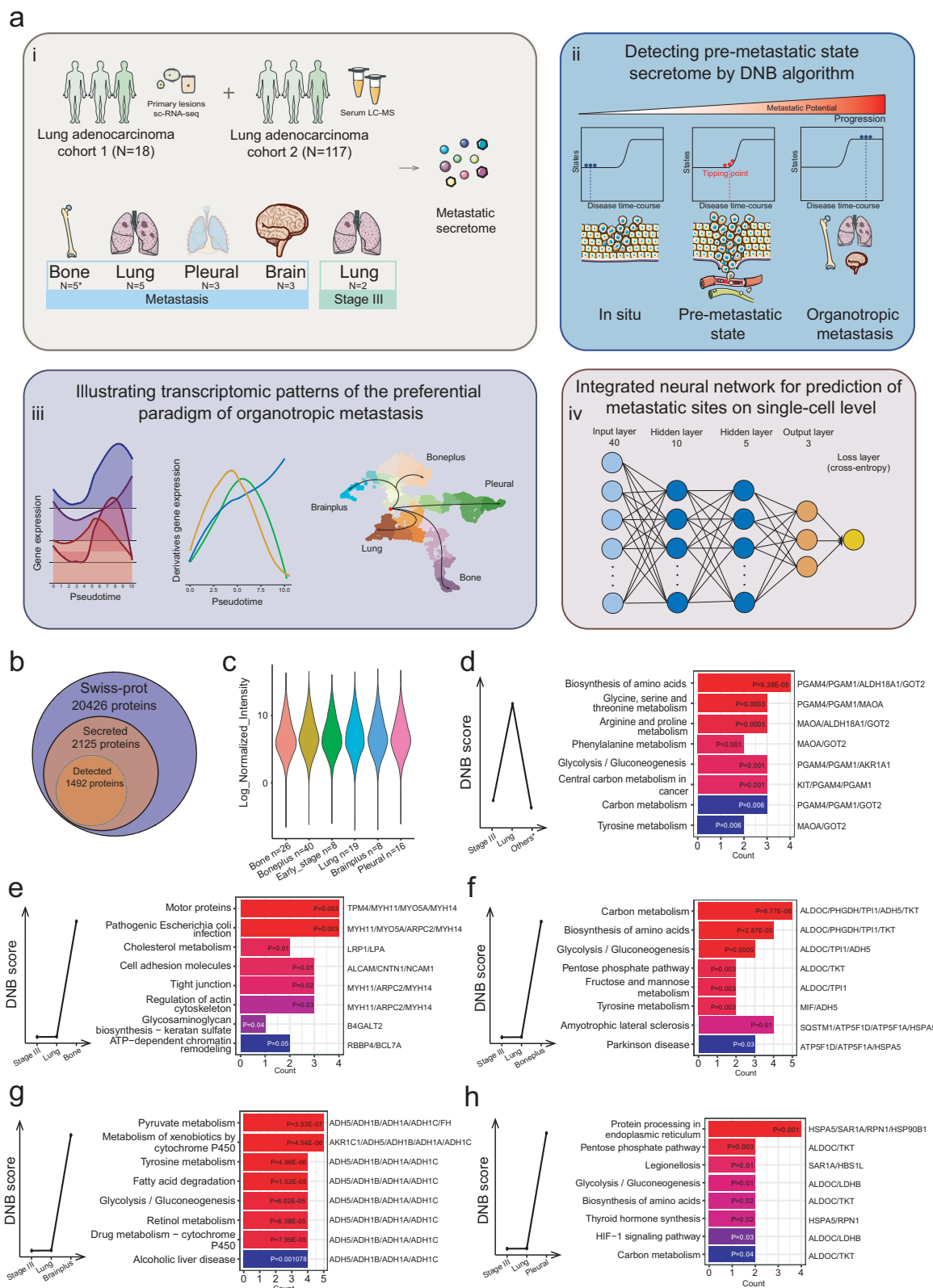
The discovery of DNB gene modules indicating the pre-metastasis states to different organs in LUAD utilizing scRNA-seq

As aforesaid, to fully understand the dynamic process from in situ cancer, to pre-metastatic state and then organotropic metastasis at the single-cell level, we performed scRNA-seq on primary lesions of 18 lung adenocarcinoma patients, diagnosed with stage III without metastasis or combined with metastasis to the brain, bone, pleura and lung (Fig. 1a), with a total of 25,421 cells obtained. The cell groups were classified into 22 individual clusters according to internal cellular characteristics inferred by the uniform manifold approximation and projection (UMAP) method, and a large proportion of cancer cells, including clusters C0, C2, C4, C5, C6, C7, C9, C12, C17, C18, C19 and C20 (namely, the Partition1), were identified and the analyses were mainly focused on them,

Fig. 1 | Discovery of metastatic secretome related to five different states based on the DNB algorithm. a Schematic illustration of study workflow. (i) Multi-omics screening of metastatic secretome based on two patient cohorts; (ii) Discovering the genomic trace of pre-metastatic state by DNB algorithm; (iii) Reconstructing the organotropic metastasis by implementing cell trajectory and pseudotime, in addition to elucidating the intermediate state of cancer cells with no specific organotropism; (iv) Distant metastasis prediction based on the neural network model, which was achieved on single-cell level; Please see Methods for detailed workflow; *the Bone group contains bone (*n* = 3) and "Boneplus" (*n* = 2). The figure was partly generated using Servier Medical Art, provided by Servier, licensed under

a Creative Commons Attribution 3.0 unported license. **b** Venn diagram illustrating the detected protein in the serum samples, and the relation of dependence with reviewed Swiss-prot proteins. **c** The proteomic intensity across serum samples with different metastatic states. The DNB modules discovered in the lung (intrapulmonary) metastasis (**d**), bone metastasis (**e**), bone metastasis with concurrent metastasis (Boneplus) (**f**), brain metastasis with concurrent metastasis (Brainplus) (**g**), and pleural metastasis (**h**), and their related KEGG annotation. Overrepresentation analysis was used for KEGG pathway enrichment and *P* value calculation (**d**–**h**); *Others indicates metastatic states except lung (intrapulmonary) metastasis. Source data are provided as a Source Data file.
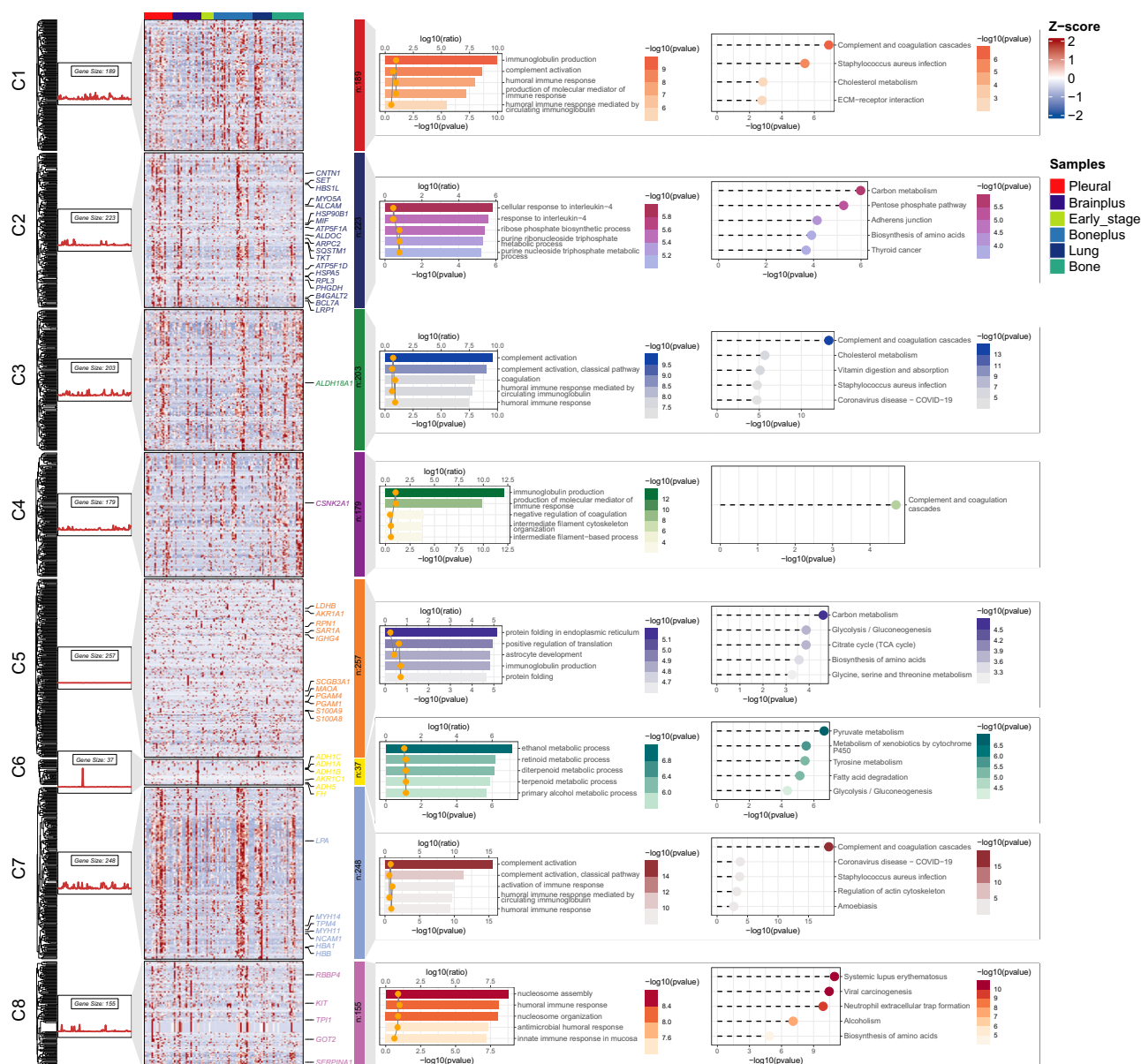

Fig. 2 | Proteomic signature of serum metastatic secretome. Proteins were clustered into eight blocks (C1–C8), followed by GO and KEGG annotation. Interestingly, proteins in the DNB modules related to identical metastasis states were clustered into the same block, which were labeled alongside. Over-representation

analysis was used for gene annotation (KEGG pathway enrichment, GO annotation), and Benjamini and Hochberg (BH) method was used for *P* value calculation. Source data are provided as a Source Data file.

with 11,881 cells totally (Fig. 3a). The pseudo temporal analysis first indicated the time sequence among these cancer cell trajectories (Fig. 3b), and then the origin (from which type of metastasis) of these cells was also shown (Fig. 3c), eventually leading to putative metastatic trajectories of cancer cell clusters with time and endpoints (metastatic organs) simultaneously (Fig. 3d). For subsequent DNB analysis, pseudotime were further divided into five parts as T0–T4 (Fig. 3d). Then, we utilized the DNB method in cancer cells according to different metastatic trajectories based on the pseudo temporal analysis, to locate the pre-metastasis states to different sites of LUAD. Successfully, for each metastatic site, we identified several gene modules according to the DNB score, which consisted of multiple genes and represented the
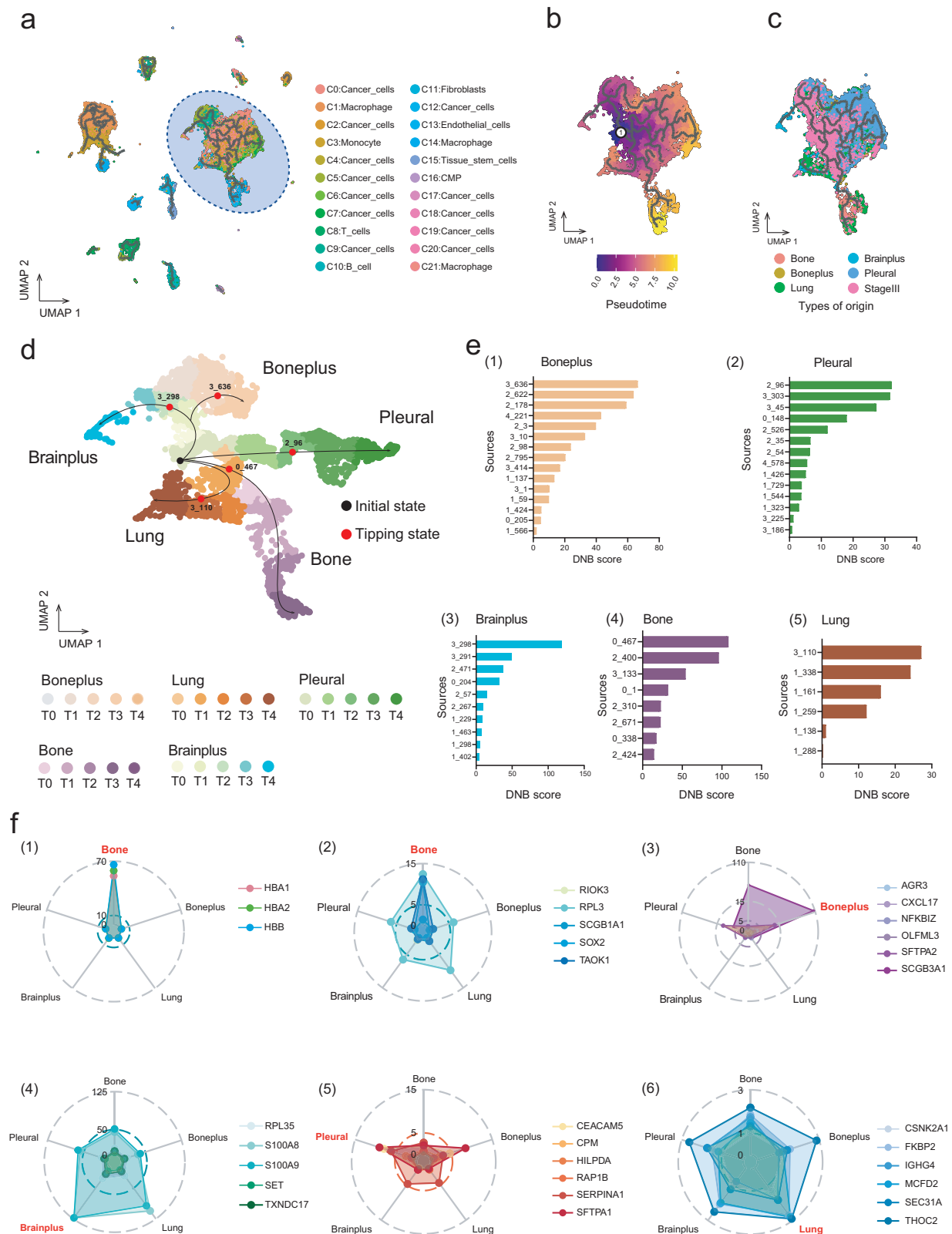
most dynamic variations in the critical time point just before metastasizing to each specific metastatic site (Fig. 3e(1)–(5), Supplementary data 2). The top genes in these gene modules for each metastatic site were then selected, respectively, and were used to define the pre-metastasis state of each organ inferred by the DNB algorithm (Fig. 3f(1)–(6)). Among them, *HBA1, HBA2, HBB* were the top three genes

for bone metastasis (Fig. 3f(1), (2)), *AGR3, CXCL17, NFKBIZ* were the top three genes for "Boneplus" metastasis (Fig. 3f(3)), *RPL35, S100A8, S100A9* were the top three genes for "Brainplus" metastasis (Fig. 3f(4)), *CECAM5, CPM, HILPDA* were the top three genes for pleural metastasis (Fig. 3f(5)), and *CSNK2A1, FKBP2, IGHG4* were the top three genes for lung metastasis (Fig. 3f(6)).

Fig. 3 | The cell trajectory of five metastatic states demonstrated shared path and genomic disparities. a UMAP plot of 22 cell clusters with cell annotations, from merged samples with five different metastatic states. Blue circles indicated a large portion of cancer cells. **b** The pseudotime UMAP plot indicating the time sequence in the cell trajectory. The root cells were selected from the stage III samples. **c** The types of origin UMAP plot indicating the constituents of cell samples. **d** The inferred metastatic trajectory with time and endpoints (metastatic organs). For each metastatic trajectory, pseudotime was divided into five parts, T0–T4, for subsequent DNB analysis. Initial and tipping states were annotated on cell trajectories. **e** DNB modules of genes were calculated for five different metastatic states: bone metastasis with concurrent metastasis (Boneplus), pleura, brain metastasis with concurrent metastasis (Brainplus), bone, and lung. **f** Top genes determined the five different metastatic states inferred by the DNB algorithm. The perturbations of the selected genes indicated the intermediate stage of final metastases. The top eight genes for bone metastasis were illustrated in (1) and (2). The top three genes for bone metastasis were *HBA1*, *HBA2*, *HBB* (1), while the remaining five genes were separately listed in (2). The top three genes for bone metastasis with concurrent metastasis (Boneplus) were *AGR3*, *CXCL17*, *NFKBIZ* (3). The top three genes for brain metastasis with concurrent metastasis (Brainplus) were *RPL35*, *S100A8*, *S100A9* (4). The top three genes for pleural metastasis were *CECAM5*, *CPM*, *HILPDA* (5). The top three genes for lung metastasis were *CSNK2A1*, *FKBP2*, *IGHG4* (6). Source data are provided as a Source Data file.

Each metastatic state trajectory demonstrated distinct cellular components and transcriptomic characteristics

The 22 cell clusters from the unanimous cell trajectory were then annotated according to their origin of metastasis, inferred by UMAP method (Fig. 4a). Each metastatic state owned its distinct cellular content, with different proportion and number of 22 cell clusters (Fig. 4b, c), in which C12 and C4 were the main types for bone metastasis, C9 was the dominant type for both "Brainplus" and "Boneplus" metastasis and cluster C0 was the most enriched one in lung and pleural metastasis (Fig. 4d). The genomic signatures of each cluster were shown in heatmap (Fig. 4e), and then the dominant cell clusters in each metastatic state were chosen for differential expression analysis compared with their expression in other states (Fig. 4f(1)–(6)), as mentioned above. Interestingly, several DNB genes were also identified: *HBA1*, *HBA2*, *HBB*, and *FOM2* were significantly enriched in C12 cancer cells in bone metastasis, *VIM* was markedly enriched in C9 cancer cells in "Brainplus" metastasis, *SFTPA1* was highly expressed in "Boneplus" metastasis, *CEACAM5* was markedly enriched in C0 cancer cells while *CCDC80*, *CAV1*, *SERPINE1*, *HLA-DQA2*, *NNMT*, *COL1A2*, and *AKR1B* were down-regulated in C0 cancer cells in pleural metastasis, and *COL1A2* and *IGFBP7* were up-regulated while *SFTPA2* was down-regulated in C0 cancer cells in lung metastasis (Fig. 4f(1)–(6)). Meanwhile, most of the other DNB genes were still not found. Therefore, these results further supported that differentially expressed genes (DEGs) and DNB genes were two different concepts though intersections existed that DEGs were static results based on the comparison of two groups of gene expression data, which could only distinguish between two different states of disease while lack of ability to predict the "pre-disease" state during the process of disease.

## Overlapping the DNB modules derived from both sera and primary lesions and the identification of SAA1 as a putative biomarker of bone metastasis

To further validate the predictive role of DNB proteins detected from sera, defined as the metastatic secretome, they were compared and then overlapped with the DNB genes screened by scRNA-seq of primary lesions. Several co-existed genes/proteins among DNB modules from both sera and primary lesions were then identified in each metastasis state (Fig. 5a(1) and (2)). Among them, SAA1 was selected for a marker for pre-metastasis to the bone, PARP1 and NASP were selected for markers of "Boneplus" pre-metastasis, YWHAE, PSMB6, ACTR3 and AKR1C3 were markers for "Brainplus" pre-metastasis, and RPS3A, RACK1, EEF1B2, VDAC1 and EEF2 were markers for lung pre-metastasis, for their significantly higher expression compared with each other metastatic state. Then, we aimed to further validate the predictive role of these selected proteins, among which SAA1 was selected as the validation target since it was the only selected serum DNB protein for bone metastasis. First, we confirmed its high expression in cancer cells of bone metastasis trajectory in our scRNA-seq dataset of primary lesions (Fig. 5b). We also corroborated the clinical significance of *SAA1* in the Kaplan-Meier Plotter (KMplot) cohort[24] that lung cancer patients with high expression of *SAA1* had markedly worse overall survival (Fig. 5c(1)) and earlier first progression (Fig. 5c(2)) compared with those with low expression. Furthermore, we constructed a disseminated tumor cell (DTC) mice model by intracardiac injection of Lewis lung carcinoma (LLC) and $KRAS^{G12D}TP53^{-/-}$ (KP) lung carcinoma cell lines, and then compared the serum level of SAA1 of mice with or without bone metastasis (Fig. 5d). We found mice combined with bone metastasis exhibited significantly higher serum level of SAA1 compared with those without bone metastasis (Fig. 5e), further supporting the predictive role of SAA1 as well as the DNB modules selected by DNB algorithm.

Verification of the specific role of SAA1 in lung cancer bone metastasis supported by clinical cohorts

Furthermore, we aimed to confirm the clinical significance of SAA1 during bone metastasis in two human lung cancer cohorts. Initially, the scRNA-seq was performed in three lung cancer patients with bone metastasis, based on their samples of primary lung lesions and paired bone metastatic lesions (Fig. 6a). Then, we conducted the bulk RNA-seq ($n = 8$) and immunohistochemistry (IHC) staining ($n = 12$) on bone metastatic lesions from lung cancer patients ($n = 20$) (Fig. 6a). For the first cohort, we obtained a total of 12,521 cancer cells, which could be divided into 19 cell clusters (Fig. 6b), and we further distinguished the sources of these cells with primary lung or bone metastatic lesions (Fig. 6c). Then, we exhibited the proportion of each cell cluster in the primary lung or bone metastatic lesions, among which C1, C4 and C5 were the dominant cell types in primary lung lesions while C8 and C9 were the largest proportions in bone metastatic lesions (Fig. 6d). The expression level of *SAA1* in each cell cluster was also evaluated, and an overwhelming majority of cell clusters (18/19) expressed *SAA1*, with even more than 40% expression rate of *SAA1* observed in nine cell clusters (Fig. 6e). Furthermore, we explored the variation of the *SAA1* expression proportion in each cell cluster from primary lesions to bone metastatic lesions. The results indicated that 14/19 cell clusters showed an increased expression level of *SAA1* that its expression ratio reached nearly 100% among these clusters, and the overall expression of *SAA1* also increased from ~5 to ~15 among all cell clusters (Fig. 6f). These results further confirmed the sources of *SAA1* and strongly supported its critical role during the process of lung cancer bone metastasis. Also, to validate whether other DNB genes derived from primary lesions would also express in metastatic lesions, we investigated the expression level of the marker genes indicating bone pre-metastatic state in these bone metastatic lesions, and found that many of these genes were highly enriched in cell clusters, including *HBB*, *HBA2*, *HBA1*, *S100A8*, *SFTPA2*, *SCGB3A1*, and *SAA1* (Fig. 6g). Consequently, the results indicated the accuracy of DNB method and reliability of our identified marker genes based on primary lesions. Meanwhile, for the second cohort, through bulk RNA-seq, we compared the expression level of identified serum marker genes foreshadowing different metastatic sites, between bone metastatic lesions and normal bone tissues from healthy controls. The results indicated that only *SAA1* denoted a Log$_2$foldchange of ~10 times between bone metastatic lesions and normal bone tissues, while the Log$_2$foldchange of other serum proteins was all less than two, further supporting the feasibility of SAA1 as a pre-metastatic marker for bone metastasis as
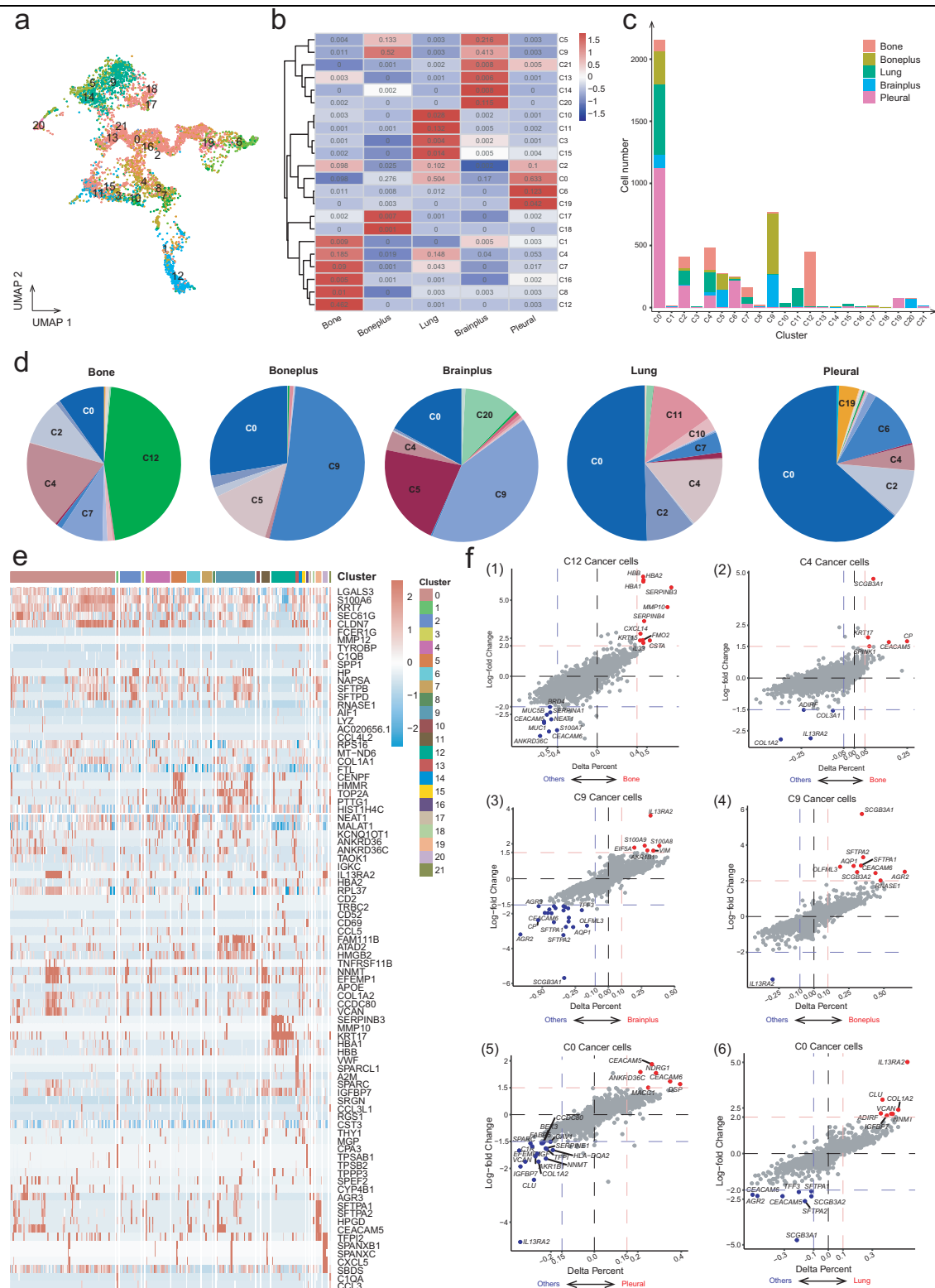
Fig. 4 | Genomic signatures in the subdivision of five metastatic state trajectories. a UMAP plot of five metastatic state trajectories, in addition, cells were re-clustered into 22 clusters. **b** Cellular proportions of five metastatic states in 22 cell clusters. **c** Cell numbers of five metastatic states in 22 cell clusters. **d** Pie charts illustrating the distributions of cell clusters in five different organ metastases. **e** Genomic signatures of top gene expression in 22 cell clusters. **f** The dominant cell clusters in each distant metastasis were chosen for differential expression analysis.

C4 and C12 were analyzed due to their predominance in bone metastasis ((1) and (2)). C9 cluster was enriched both in "Brainplus" (brain metastasis with concurrent metastasis) and "Boneplus" (bone metastasis with concurrent metastasis); considering its multiple metastasis signatures, comparative analyses were conducted on both "Brainplus" and "Boneplus" ((3) and (4)). C0 cluster was enriched both in the lung and pleural metastases, and subsequent comparative analyses were conducted ((5) and (6)). Source data are provided as a Source Data file.
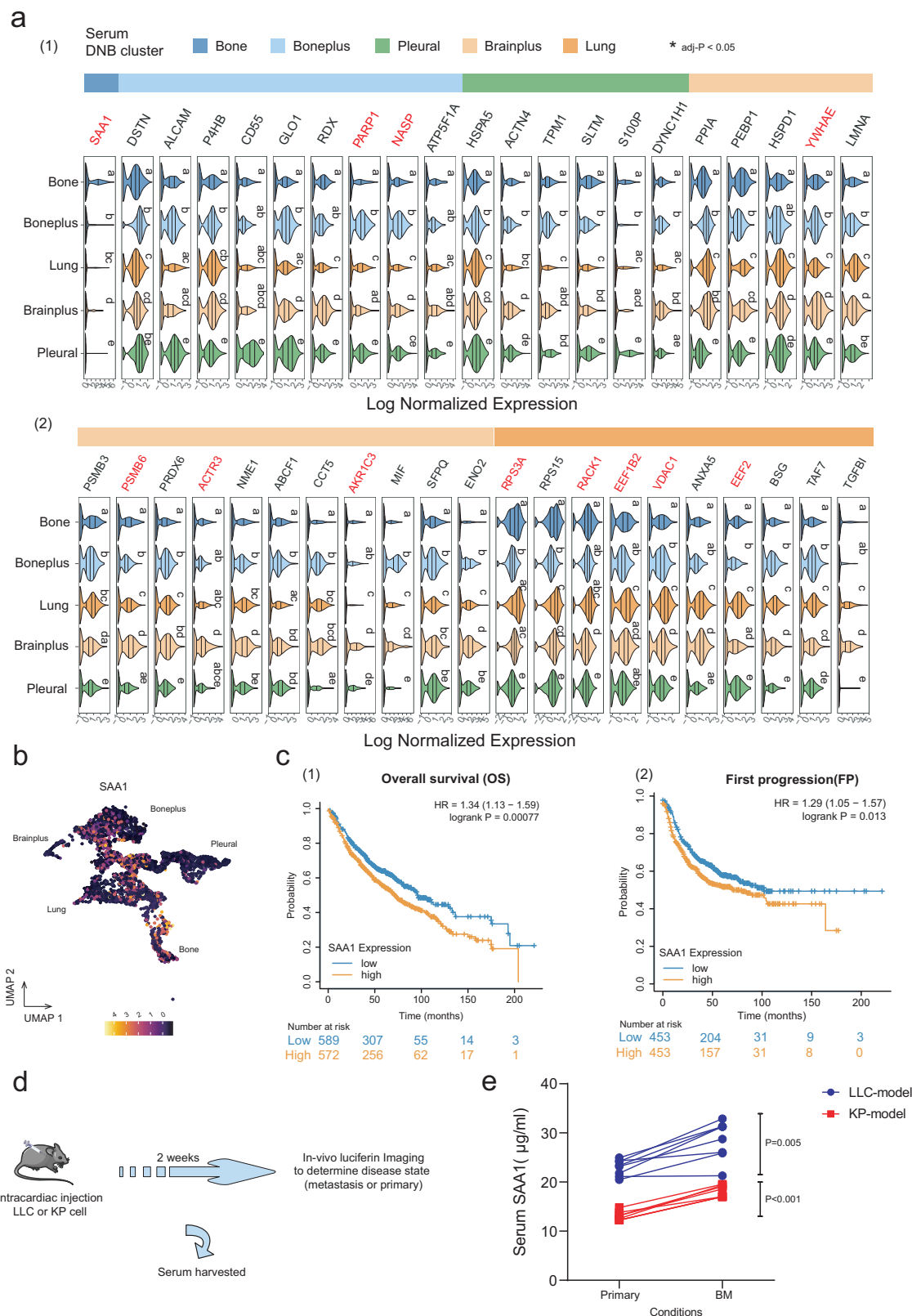
well as the organ-specificity of our identified DNB serum proteins (Fig. 6h). Then, we selected five representative DNB genes/proteins derived from Fig. 5a(1) and (2) that could be detected from both primary lesions and sera, including SAA1 (indicating bone metastasis), YWHAE (indicating "Brainplus" metastasis), EEF2 (indicating lung metastasis), PSMB6 (indicating "Brainplus" metastasis) and RPS3A (indicating lung metastasis), and aimed to validate their expression in

bone metastatic lesions in the second clinical cohorts through IHC staining. The results indicated that only SAA1 exhibited enhanced expression in bone metastatic lesions compared with other markers, with approximately 70% positive rate, while other markers were nearly not expressed in bone metastatic lesions (Fig. 6i), supporting SAA1 as an organ-specific predictive marker in lung cancer bone metastasis clinically. Eventually, we tested the serum level of SAA1 and YWHAE

Fig. 5 | Identification of SAA1 as a predictive marker for lung cancer bone metastasis inferred by the overlap results from sera and single-cell DNB algorithm. a Serum DNB proteins were overlapped with primary lesion DNB genes, in which comparative analyses were made to filter the most relevant organ-specific serum biomarkers ((1) and (2)). Different letter marks indicate significant differences ($P < 0.05$). In this case, proteins/genes marked in red indicate high specificity in their belonging groups (five different metastatic states). Kruskal–Wallis and Dunn's tests were performed to make inter-group comparisons. **b** *SAA1* expression distribution in the UMAP plot. **c** The clinical significance of *SAA1* as a biomarker in lung adenocarcinoma was illustrated, in which the overall survival and first progression results were significant. **d** Schematic figure indicating animal model validation of SAA1 as a serum biomarker for bone metastasis. The number of mice in each group was 16. The figure was partly generated using Servier Medical Art, provided by Servier, licensed under a Creative Commons Attribution 3.0 unported license. **e** The serum SAA1 was measured in two different cell lines-generated bone metastasis models. A total of 16 mice were included after model construction for each cell line (Lewis lung carcinoma (LLC)-model or *KRAS*^{G12D}*TP53*^{−/−} (KP)-model), and a comparison was made between primary ($n = 8$) and bone metastasis ($n = 8$) groups. The serum level of SAA1 was significantly higher in the samples of bone metastasis compared those without bone metastasis. One-way ANOVA was performed to make inter-group comparison. The *P* value for LLC-model comparison between primary and bone metastasis groups was 0.005, and the *P* value for KP-model comparison between primary and bone metastasis groups was <0.001. Source data are provided as a Source Data file.

among healthy controls and lung cancer patients with different metastases. The results indicated SAA1 showed markedly higher concentration in patients with bone metastasis, compared with healthy controls and those with "Brainplus", lung, and pleural metastases, further supporting the organ-specificity of the marker genes derived from the DNB approach (Fig. 6j(1)). Meanwhile, YWHAE also denoted a trend towards higher concentration in patients with "Brainplus" metastasis compared with patients with other metastases (Fig. 6j(2)).

The determination of intermediate status of cancer cells in five metastatic state trajectories

Cancer cells metastasize from primary to distant organs, involving a status of dramatic variation of cell organotropic metastatic potency, namely the intermediate status. To locate the intermediate status of cancer cells in each metastatic state trajectory, the top genes defining the pre-metastasis state of each organ were used as marker genes (Fig. 3f(1)–(6)). We first depicted the expression trends of the marker genes during the metastatic process, based on the time information derived from pseudo-temporal analysis (Fig. 7a, Supplementary Fig. 2). Previous studies have indicated the metastasis of cancer is preceded by a transitional phase characterized by drastic changes in gene expression through analyzing time-series transcriptomic data[17,20]. Thus, the intermediate status of the lung cancer cells was similarly decided by the points of maximum rate of change in expression of these marker genes (Fig. 7a, Supplementary Fig. 2), along with the determination of the 'before intermediate' and 'after intermediate' status, which were graphically shown together in UMAP plot (Fig. 7b). To further anatomize the characteristics of these three states in metastasis, we also compared the differentially expressed genes of three status in the heatmaps. The comparative results indicated each metastatic trajectory denoted distinct gene signatures in intermediate status. For instance, *S100A8*, *S100A9*, and *SFTPB* were the most highly expressed genes in bone metastasis, while *HLA-DQA2*, *PAEP*, and *IL-33* were the top gene signatures of lung metastasis. Likewise, the gene signatures of "before intermediate" and "after intermediate" also differed markedly (Fig. 7c, Supplementary Figs. 1, 3, 4).

Predicting the metastatic state in external validation cohorts based on the neural network model
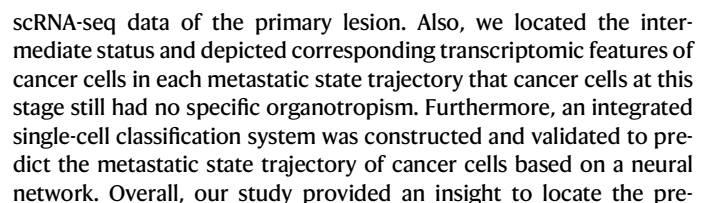
With the aim of constructing a feasible model to predict the cancer cell states of external input, an integrated neural network model was established (Fig. 8a). The training group data originated from the scRNA-seq data depicting five different metastatic trajectories, with only cancer cells under "after intermediate" status in each trajectory were included. Due to the rough classification of lung cancer metastasis by current literature, we adjusted the number of metastatic trajectories from five to three, combining the data of lung (intrapulmonary) metastasis and pleural metastasis into lung metastasis, as well as the bone metastasis and "Boneplus" metastasis into bone metastasis. The external data (GSE123902) was used as the validation set, in which we selected a LUAD patient with primary tumor (GSM3516665) and a LUAD patient with bone metastasis (GSM3516664) and identified cancer cells from them. These cancer cells were then input into the neural network for type prediction (Fig. 8a). Our results demonstrated that this neural network model exhibited superior performance in prediction efficacy for both primary tumor and bone metastasis, with area under curve (AUC) of 0.87 (precision of 80% and accuracy of 80%) for bone metastasis and AUC of 0.87 (precision of 79% and accuracy of 80%) for primary tumor (Fig. 8b). The neural network model distinguished these cells into three different metastatic states. Specifically, for cancer cells from bone metastasis, the overwhelming majority of them were labeled as bone metastasis status, while for cancer cells from primary tumor, nearly half of them were also matched to primary lesion status, indicating the predictive efficacy of this neural network model (Fig. 8c). We additionally enrolled a LUAD patient with brain metastasis from the external data (GSE123902) for further comparative analysis, and interestingly, the results for these three metastatic states indicated that *SAA1* was still significantly enriched in cancer cells labeled as bone metastasis compared with those labeled as primary tumor or brain metastasis (Fig. 8d), further supporting the predictive role of *SAA1* for bone metastasis in lung cancer patients.

Further, various machine learning models were used to construct ensemble classifiers, retaining only the previously analyzed significant genes in the original scRNA-seq data. The datasets were grouped into bone metastasis cells and primary lung cancer cells as inputs to the model, as mentioned in the "Methods" section. We first conducted cross-validation and found that the performance of the classifier was satisfactory. In 5 cross-validations, it showed good performance, which also indicated that the significant genes we analyzed were indeed meaningful between bone metastasis cells and primary lung cells (Supplementary Fig. 5a–c). Meanwhile, we also introduced the primary lung cancer patient dataset (GSM3516665) and lung cancer bone metastasis patient dataset (GSM3516664) from the external cohort (GSE123902) as the validation set. Initially, we removed batch effects between the external data and our data using the canonical correlation analysis (CCA) method. Then, we filtered out cells with significant differences between the external dataset and our data using the Mahalanobis distance. Then, we classified the two patient datasets separately using an integrated model classifier and visualized the accuracy of classification using the confusion matrix (Supplementary Fig. 5d, e). It could be seen that the integrated model also showed excellent performance on external datasets, verifying the reliability of the significant genes we analyzed. Eventually, we also utilized the receiver operating characteristic (ROC) curve to visualize the performance on the entire external dataset (Supplementary Fig. 5f), demonstrating that classifiers relying on the selected genes as feature inputs performed well, further re-validating our previous findings.

Discussion

One of the most important death causes of lung cancer is metastasis. Over the past two decades, although the treatment of NSCLC has been revolutionized by the use of tyrosine kinase inhibitors and immune treatments, metastatic disease remains largely incurable and presents

a, b, c, d, e, f, g, h, i, j figure panels

a 5-year survival rate of ~10%[25]. Therefore, early detection and timely treatment for lung cancer metastasis is critical. In this study, we successfully identified several serous metastatic secretomes as early-warning signals for five different metastatic states of lung cancer, whose expression patterns were also verified in primary lung lesions. Meanwhile, we demonstrated the marker genes defining the pre-metastatic state for each organ with the DNB algorithm utilizing the scRNA-seq data of the primary lesion. Also, we located the intermediate status and depicted corresponding transcriptomic features of cancer cells in each metastatic state trajectory that cancer cells at this stage still had no specific organotropism. Furthermore, an integrated single-cell classification system was constructed and validated to predict the metastatic state trajectory of cancer cells based on a neural network. Overall, our study provided an insight to locate the pre-

Fig. 6 | Verification of the specific role of SAA1 in lung cancer bone metastasis supported by the data of clinical cohorts. a Schematic figure validating the role in lung cancer bone metastasis clinically. For the first validation cohort, three paired lung adenocarcinoma samples of primary and bone metastatic lesions were collected and analyzed through scRNA-seq. For the second validation cohort, 20 lung adenocarcinoma bone metastatic lesions were collected, and eight of them were analyzed through bulk RNA-seq and compared with the normal bone tissues from healthy controls, while IHC staining was conducted on the remaining 12 samples to confirm the expression of SAA1, YWHAE, EEF2, PSMB6, and RPS3A. Created in BioRender. Wen, Y. (2024) BioRender.com/c79s395. **b** UMAP plot demonstrating 19 cancer cell clusters. **c** UMAP plot of cancer cells, grouped by cell origin. **d** The major cell cluster constituents in lung or bone metastatic samples. **e** The expression percentage of *SAA1* in each cancer cell cluster. **f** The variation of expression percentage of *SAA1* in each cancer cell cluster and total cancer cell clusters from primary to bone metastatic lesions. **g** The heatmaps illustrated that many of the DNB genes prefiguring bone metastasis, including *SAA1*, were highly enriched in cancer cell clusters of bone metastatic lesions. **h** For the second validation cohort, the expression Log_2foldchange of the DNB genes indicating different metastases was examined, and *SAA1* was the only gene denoting a Log_2foldchange of -10 times between bone metastatic lesions and normal bone tissues. **i** The IHC staining results of SAA1 (1), YWHAE (2) PSMB6 (3), RPS3A (4), and EEF2 (5) in bone metastatic lesions in the second validation cohort and their corresponding statistical analysis (6), with $n = 12$ in each group. The data were presented as mean ± SD; $P < 0.001$ for the comparison between SAA1 and others. One-way ANOVA was performed to make inter-group comparison. **j** The serum level of SAA1 in groups of healthy controls ($n = 30$), bone ($n = 10$), "Brainplus" ($n = 14$), lung ($n = 12$), pleura ($n = 11$); P value for inter-group comparison was listed as follows: bone-healthy volunteers: $P = 0.01$; bone-Brainplus: $P = 0.008$; bone-lung: $P = 0.005$; bone-pleura: $P = 0.005$ (1); and YWHAE in groups of bone ($n = 10$), brainplus ($n = 12$), lung ($n = 12$), pleura ($n = 11$) were comparatively analyzed, with P value for inter-group comparison was listed as follows: bone-Brainplus: $P = 0.305$; bone-lung: $P = 0.860$; bone-pleura: $P = 1.000$ (2). The data were presented as mean ± SD. One-way ANOVA and Dunnett's test corrections were performed to make inter-group comparison. $*P < 0.05$; $**P < 0.01$; $***P < 0.001$; ns: not significant. Bone mets.: Bone metastasis.

metastasis status of lung cancer and primarily examined its clinical application value, contributing to the early detection of lung cancer in a more clinically feasible and efficacious way.

Our study utilized DNB algorithm based on muti-omics clinical data with both serum secretome and single-cell transcriptome of primary lesion to predict lung cancer metastasis. Compared with conventional approaches based on the differential expression of molecular biomarkers in a static manner used to detect the metastatic state of cancers, DNB-based methods can be used to identify the pre-metastatic state during the whole metastatic process with the support of dynamics-based data science[16]. Previous studies with the DNB method successfully revealed *CALML3* as a predictive biomarker and an early warning indicator of the initiation of hepatocellular carcinoma lung metastasis, and *DHX9* in mature B cells was identified as a predictable biomarker before lymph node metastasis in colorectal cancer[17,20]. These results have demonstrated the feasibility and efficacy of the DNB method for predicting tumor progression and metastasis. Nevertheless, there is still a lack of reports investigating the pre-metastatic state of lung cancer using the DNB method up till now. Given the high incidence rate of metastasis in lung cancer, we suggested that the DNB method had great clinical application value in the early detection of lung cancer metastasis and could eventually lower the death rate of lung cancer. In addition, previous DNB studies were mainly focused on the genomic or transcriptome profiling of primary lesions, while the great potential of serum secretome combined with the DNB method has not been revealed. Through overlapping the DNB modules of both serum secretome and transcriptome of the primary lesion, several co-existed marker genes/proteins were filtered as serum indicators for five metastatic states, among which the critical role of SAA1 in bone metastasis was successfully validated in both clinical data and in vivo mice experiment. As is known, it is currently unrealistic to perform scRNA-seq for cancer patients to guide their therapy options due to the cost-effectiveness. Thus, our study focused on the genes/proteins based on the intersection of scRNA-seq of primary lesions and serum LC–MS in order to identify those secretory proteins in sera (secreted by lung cancer lesions and transported into the blood) that could be easily detected by standard lab tests, such as radio-immunoassay and enzyme-linked immunosorbent assay (ELISA) methods. These secretory proteins not only indicate the pre-metastatic state with organotropism but also can be detected in a clinically feasible and affordable way. We further validated the reliability of the DNB approach using *SAA1* as an example that we confirmed its high expression in bone metastatic sites versus normal bone tissues, as well as its expression in both primary tumors and bone metastatic lesions at the RNA level based on clinical cohorts. Meanwhile, the IHC results supported its high expression in bone metastatic lesions at the protein level, and its serum level would specifically increase among

patients with bone metastasis compared with healthy controls or those with other metastases detected by ELISA. Thus, we believed our findings were of translational value. Nonetheless, a prospective validation in a large patient cohort of these serum secretory proteins was absent in our study. Thus, the clinical application of these serum proteins should be further examined in a multi-center, prospective cohort in the future. Overall, the results further supported the reliability of the DNB approach in predicting cancer metastasis as previous studies, and more importantly, provided an insight into predicting lung cancer metastasis in a more clinically accessible way utilizing peripheral blood.

To further construct a viable model to predict the metastatic states of lung cancer in clinics, we built an integrated neural network model originating from the scRNA-seq data after selection and further validated its performance utilizing external lung cancer metastatic data. Eventually, we successfully built a prediction model for primary tumor and bone metastasis, with exceptional performance on both precision and accuracy, supporting its feasible clinical application in external validation. We also validated these biological phenotypes through an ensemble learning model. Meanwhile, our study revealed several significant regulators underlying each metastatic state trajectory of lung adenocarcinoma as we further located the intermediate status of each trajectory, shadowing an insight into understanding the initiation of distant metastasis. As demonstrated in previous studies, a phase transition precedes the metastasis of cancer characterized by dramatic variation in gene expression[17,20]. Compared with their time-series transcriptomic data derived from different sampling times or TNM stages, we quantified this process utilizing the pseudotime information based on single-cell transcriptome data, with more accuracy. Upon this status, no specific organotropic metastasis occurs while its future trajectory has been established. Nonetheless, conventional DEGs in previous studies were static results based on the comparison of two groups of gene expression data, which could only distinguish between in situ cancer and organotropic metastasis while lacking the ability to describe the "pre-metastasis" state during the metastatic process. Consequently, several critical genes that initiate the metastasis could be easily neglected with the DEG analysis. For example, *SAA1*, also known as serum amyloid A1, was initially detected by the DNB algorithm rather than the DEG analysis in our study, and our further clinical data and in vivo mice experiment also suggested its crucial regulatory role in lung cancer bone metastasis. *SAA1* is a member of the serum amyloid A family of apolipoproteins and also a sensitive acute phase high-density lipoprotein, playing a critical role in cholesterol homeostasis and high-density lipoprotein metabolism, whose expression is up-regulated when the body is stressed by inflammation and tissue damage[26–28]. Recently, the role of *SAA1* in the occurrence and development of tumors has aroused increasing
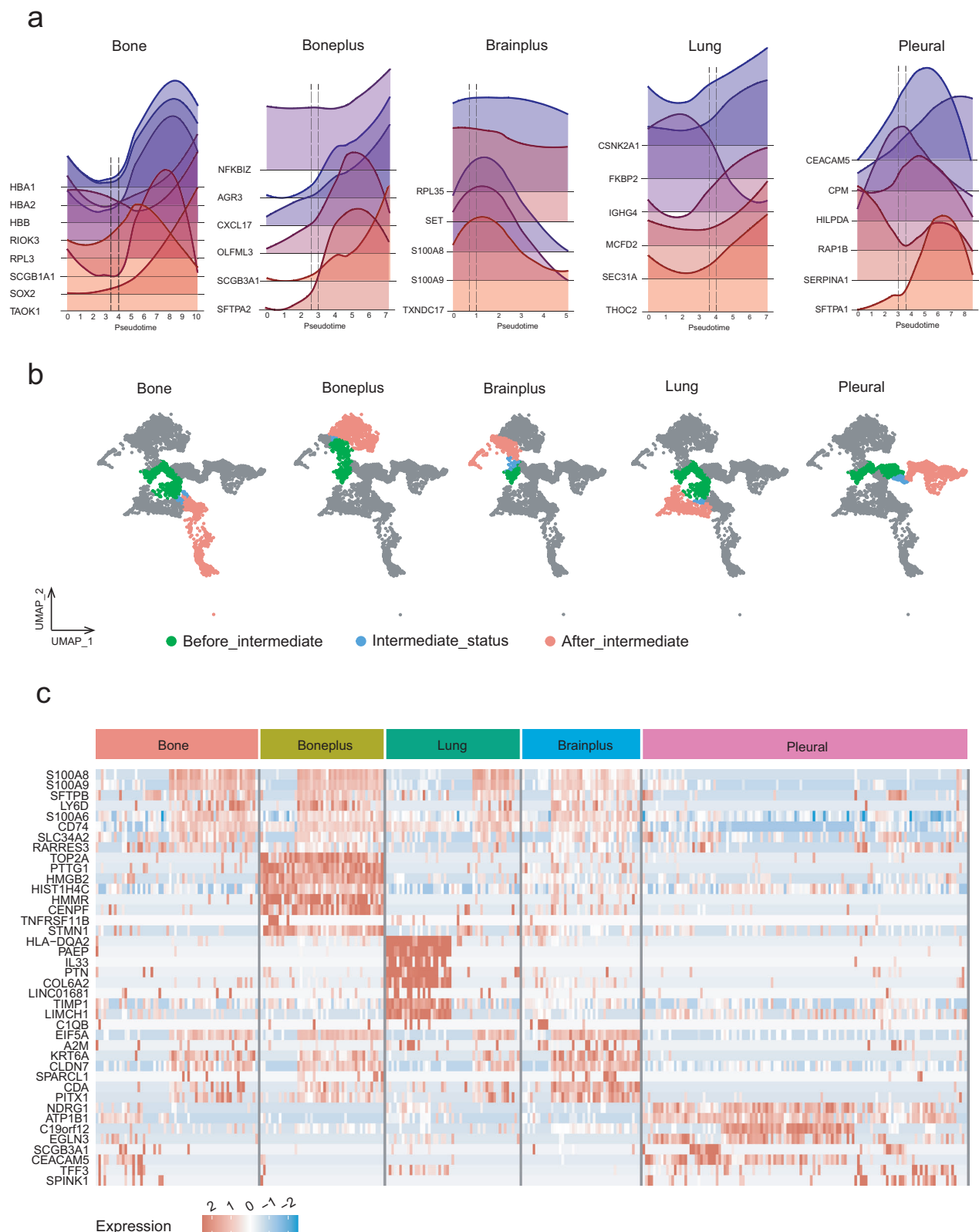
Fig. 7 | Determination of pre-metastatic states by defining the intermediate-status cells in five metastatic state trajectories. a The expression trend of marker genes in five metastatic state trajectories. **b** The graphical mapping of the intermediate-status cell of metastatic cell trajectory in UMAP plot. **c** Top gene signatures of intermediate-status cells of five metastatic states. Source data are provided as a Source Data file.

attention. Previous studies demonstrated that *SAA1* could contribute to cancer development and accelerate tumor progression and distant metastasis[29–33]. Nonetheless, the relationship between *SAA1* and lung cancer metastasis remains largely uncertain, and our study figured out this question with the DNB method. Likewise, other regulators in the intermediate status detected by the DNB algorithm should be focused, and they might also provide novel insights into the mechanism research of lung cancer metastasis. For instance, the marker identified
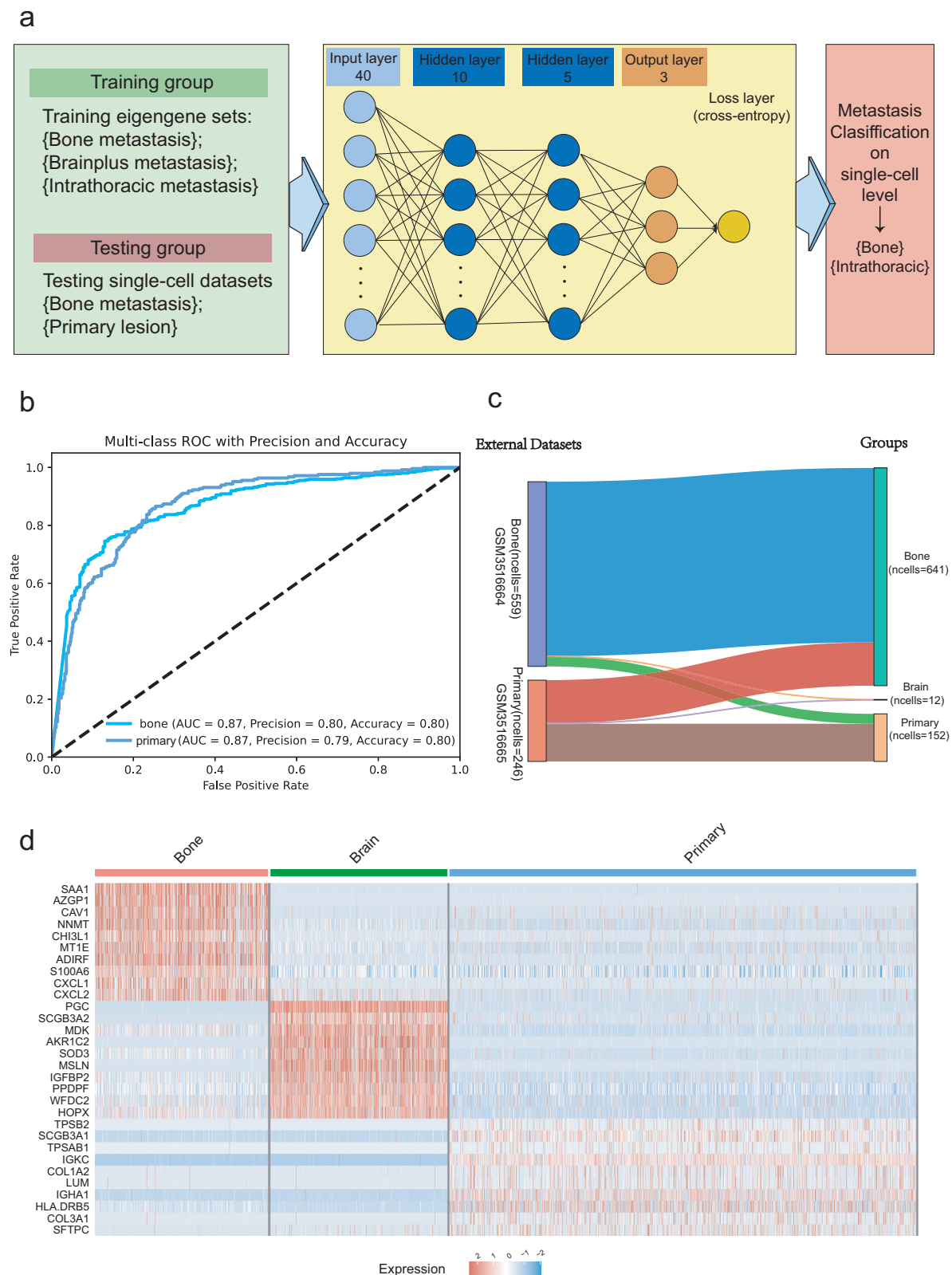
for predicting "Brainplus" metastasis, *PSMB6*, was considered a novel target for Bortezomib resistance in multiple myeloma[34]. *PSMB6* was also essential for multiple myeloma cell survival, and the upregulation or activating mutation of *PSMB6* conferred proteasome inhibitor resistance[35]. For lung adenocarcinoma, a six-gene-based risk prediction score, including *PSMB6*, could predict the overall survival. The authors further indicated that the knockdown of *PSMB6* could

significantly downregulate the proliferation of lung cancer cells[36]. Second, for *YWHAE* also identified for "Brainplus" metastasis, previous studies demonstrated that *YWHAE* expression was associated with tumor size, lymph node metastasis, and poor patient survival in patients with breast cancer, supported by the cell model results indicating that its overexpression significantly increased the proliferation, migration, and invasion abilities of breast cancer cells[37]. It was also

Fig. 8 | Metastatic state prediction of external single-cell datasets by neural network model, demonstrating preferable classification. a Schematic figure of neural network establishment and data analysis workflow. External labeled validation single-cell datasets (primary tumor and bone metastasis) were used for validation of genomic trace of metastatic signature. For the training set, bone and "Boneplus" were combined as a bone metastasis group. Lung (intra-pulmonary) and pleural metastasis were combined as a lung metastasis group. The reason for the signature combination was due to limited consensus on the discrete metastatic classification of public datasets. The figure was partly generated using Servier Medical Art, provided by Servier, licensed under a Creative Commons Attribution 3.0 unported license. **b** The receiver operating characteristic (ROC) curve indicating the effectiveness of the trained neural network. **c** The chart illustrating connecting information of verified group (left) and labeled metastatic status (right). **d** Transcriptomic signatures of primary tumor ($n = 1$), brain metastasis ($n = 1$) and bone metastasis ($n = 1$) from the external dataset (GSE123902), in which *SAA1* was in the top gene signature of bone metastasis. Source data are provided as a Source Data file.

considered a reliable predictive biomarker for gastric cancer peritoneal metastasis[38]. For lung cancer, its overexpression was validated to promote the malignant behaviors of NSCLC cells and boost tumor growth[39]. Third, for *RPS3A* identified from lung pre-metastatic state, previous studies indicated its biological role including favoring apoptosis and enhancing the malignant phenotype. Also, they considered *RPS3A* as a player that might regulate the responses of leukemia cells to chemotherapy[40]. In addition, high *RPS3A* expression in hepatocellular carcinoma was indicated to promote the biological processes related to tumorigenesis, metastasis, and immunosuppression. And *RPS3A*-based nomograms showed better accuracy for hepatocellular carcinoma prognosis when compared with traditional prognosis-prediction staging systems. Therefore, *RPS3A* might serve as a therapeutic target in and predict the efficacy of immunotherapy for hepatocellular carcinoma[41]. For lung cancer, previous studies claimed *RPS3A* as a highly informative marker for human squamous cell lung cancer, since its expression in squamous cell lung cancer increased by 70% compared with normal tissues[42]. Eventually, as for *EEF2* identified from lung pre-metastatic state, previous studies illustrated that LUAD patients with high *EEF2* expression had a significantly higher incidence of early tumor recurrence and a significantly worse prognosis. In vitro study further demonstrated that silencing of *EEF2* expression increased mitochondrial elongation, cellular autophagy, and cisplatin sensitivity. Moreover, *EEF2* was sumoylated in lung cancer cells, and *EEF2* sumoylation correlated with drug resistance, suggesting that *EEF2* was an anti-apoptotic marker in LUAD[43]. Previous studies also indicated that *PRMT7* contributed to the metastasis phenotype in human NSCLC cells, possibly through the interaction with *HSPA5* and *EEF2*[44]. Most of the previous studies postulated the DNB genes related to drug resistance and cellular stress, with the potential for migration and metastasis. Overall, though the correlation between these marker genes and specific metastatic types might be identified for the first time, their cancer-related behaviors might have been reported before, indicating such a DNB-detected correlation was biologically relevant.

Several limitations existed in our study. Initially, since the scRNA-seq was based on the cancer cells from primary lung lesions rather than metastatic sites, the high expression of background genes of lung tissue could not be neglected, which might lead to bias in the classification system we built. Partially due to this reason, the prediction model indicated that a relatively high proportion of primary cancer cells (50.1%) of the verified group would metastasize to bone, and a certain proportion of bone metastatic cancer cells were classified as primary cancer cells. Thus, a future adjusted model based on metastatic site was required. Also, the biological behaviors between the primary and metastatic sites were not always consistent. It was possible that the expression of genes increased in both primary lesions and sera, while their expression was not elevated in metastatic sites. To validate our findings, we further collected and analyzed bone metastasis samples from LUAD patients through bulk RNA-seq. The results also indicated that *SAA1* was significantly up-regulated in bone metastatic samples compared with normal bone tissues, and it also showed strong expression in bone metastatic lesions directly detected by IHC staining, supporting the accuracy of our findings based on the status of primary lung lesion when distant metastasis occurred. The scRNA-seq

data based on paired samples of primary and bone metastatic lesions also indicated the cell percentage of *SAA1* expression would also increase from 5 in primary lesions to 15 in bone metastatic lesions in total. It was worth mentioning that there were usually no indications for surgery in treating other distant metastases of lung cancer, which might even violate ethics sometimes, leading to the difficulty in obtaining the samples. Nonetheless, though metastatic samples were difficult to obtain, future studies were still warranted to confirm our findings in other metastatic samples if possible. Second, our study only enrolled stage III or IV patients, while lung cancer metastasis could occur at an earlier stage, especially under the condition of surgery. Since the biological behaviors between early and advanced lung cancer could be totally different, the conclusions of the study should be interpreted with caution when considering early-stage lung cancer. Another DNB study focusing on early-stage lung cancer should be conducted to validate our findings in the future. Specifically, approximately 30% of patients with early-stage lung cancer receiving surgery will still suffer from the cancer recurrence or metastasis[45]. Since there were usually no metastatic sites in patients when they underwent surgery, a follow-up to confirm metastasis after surgery in a prospective cohort study was essential to investigate the potential metastatic capability of early-stage lung cancer, which would be another research emphasis in our serial studies focusing on lung cancer metastasis. Then, the classification system could be further validated and enhanced based on the DNB genes derived from early-stage lung cancer samples. Third, since our work was not mainly focusing on specific markers on each metastasis, further investigation on the function of marker genes in lung cancer metastasis was absent. In our study, the serum biomarker SAA1 discovered by DNB presaging lung cancer bone metastasis has been fully validated by clinical data and bone metastasis mice models, proving the accuracy of the DNB approach and the feasibility of using secretory proteins as biomarkers to predict metastatic sites. Nonetheless, it remains unknown how it functions in detail during this process, and future researchers should conduct more mechanism studies surrounding these marker genes on the process of organ-specific metastasis, and thereby the relevant drugs can be developed to avoid metastasis at the ultra-early stage. Meanwhile, it would be more ideal if the predictive value of SAA1 could be validated in pre-metastasis mice models or prospective cohorts. However, several technical standards have not been established in mice models with pre-metastatic bone, such as the construction approach and definition of the pre-metastatic period of bone, and there was still a lack of prospective clinical cohorts to validate SAA1 as a pre-metastatic bone signal among lung cancer patients. Thus, future studies should focus more on the constructions of pre-metastasis models and prospective cohorts. Furthermore, the neural network model we utilized to build the classification system might lead to overfitting. Thus, it was essential to validate its accuracy in a larger external cohort to avoid potential bias. Nevertheless, due to the availability of public datasets, it was difficult to attain such ideal data with enough amount. Consequently, we also used the ensemble learning model to further confirm the findings derived from the neural network model, and the results supported the reliability of this classification system. Still, future larger external cohorts were urgently required to further verify these findings. Also, other

proteomic tools, like SignalP, might also be used in the future, which would be beneficial to further validation of secreted proteins of our LC-MS data[46–51].

Above all, our study used the DNB approach to identify the marker genes initiating lung cancer organ-specific metastasis based on multi-omics clinical data. Moreover, an integrated single-cell classification system for prediction of metastatic sites based on the neural network model was successfully built and validated externally, which could be utilized clinically to distinguish not only the metastatic potential but also the metastatic site of lung cancer.

Methods

Ethical statements
Ethical approvals for this study were obtained. This work was carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki). This study was approved by the ethical committee of Shanghai Sixth People's Hospital of Shanghai Jiao Tong University School of Medicine (2024-0941, 2018-074) and the ethical committee of Shanghai Pulmonary Hospital (No. L20-351-2). In both committees, the maximal tumor size for the mice model was 1500 mm³, and we confirmed our experiments did not exceed it. No sex or gender-related analysis was performed. Written consents were obtained from all enrolled participants.

Human serum sample preparation and LC−MS analysis
Human vein blood samples were harvested and stored in the collecting tube under 4 °C, centrifuged at $1900 \times g$ for 8 min to separate sera. Peptides were extracted from the serum samples as follows[52]. Briefly, 1 μL of each serum sample was diluted to 50 μL with 50 mM ammonium bicarbonate (ABC). Proteins were reduced with 10 mM dithiothreitol (DTT) for 30 min at 56 °C, then alkylated with 20 mM iodoacetamide (IAA) for 30 min in the dark at 37 °C. Next, proteins were diluted with 70 μL 50 mM ABC at the estimated end concentration of 0.5 μg/μL. For digestion, the proteins were digested by trypsin at a mass ratio of 1:50 (enzyme: protein) for 12 h at 37 °C on the shaker at $140 \times g$. Six microliters of 10% formic acid (FA) were added to quench the reaction. Digested peptides were cleaned-up with Sep PAKTM C18 column (Waters, USA), then centrifuged at $14,000 \times g$ for 30 min and the supernatants were collected for future use. Peptide quantification was performed using Pierce Quantitative Peptide Assays & Standards (Thermo Scientific, No. 23275). After protein digestion, samples were desalted using Pierce C18 Spin Columns (Thermo Scientific, Cat No. 87777) and then lyophilized in a vacuum drier. Peptide separation was carried out using a nanoElute liquid chromatography system (Bruker Daltonics, Billerica, USA). A 200 ng peptide fragment was used for LC−MS analysis. All samples were analyzed using a hybrid trapped ion mobility spectrometry (TIMS) quadrupole time-of-flight mass spectrometer (MS) (TIMS-TOF Pro, Bruker Daltonics, Billerica, USA) equipped with a nanoelectrospray ion source. The intense threshold was set to 5000, and the ion mobility was scanned from 0.7 to 1.3 V s/cm². The spray voltage was 1500 V, and the ion source temperature was 180 °C. The auxiliary gas flow rate was 3 L/min. The accumulation and ramp time were set to 100 ms each, and mass spectra were recorded in the range of $100–1700 m/z$ in positive electrospray mode. A single cycle collection time was 1.16 s, which included one full TIMS-MS scan and 10 parallel accumulation-serial fragmentation (PASEF) secondary scans. For scheduled data-independent acquisition PASEF (DIA-PASEF) analysis, the files were analyzed using DIA-NN (1.8)[53] against a plasma spectral library containing 5102 peptides and 819 unique proteins from the Swiss-Prot database of Homo sapiens. In the DIA-NN settings, the software automatically sets the retention time extraction window. Protein and peptide false discovery rates were set not to exceed 1%. Protein inference was set to the protein names (from the FASTA file), and the cross-run normalization was set as "RT-dependent".

LUAD primary and bone metastatic samples and scRNA-seq
Lung primary lesions were harvested by fine needle biopsy or bronchoscopy, and 18 scRNA-seq datasets were utilized for subsequent analysis. Single-cell preparation, gene expression matrix generation, and data QC can be found in our previous article[54]. Meanwhile, for the first clinical validation cohort, we collected and analyzed three paired samples of primary lung lesions and bone metastatic lesions from LUAD patients with bone metastasis and the experimental procedures were the same as above.

LUAD bone metastatic lesions and bulk RNA-seq
For the second clinical validation cohort, the bulk RNA-seq was performed on bone metastatic lesions from LUAD patients. Specifically, LUAD samples were collected and stored in RNA later at −80 °C. Total RNA was extracted using the RNeasy mini kit (Qiagen, 74104, Germany). Strand-specific libraries were prepared using the VAHTS Universal V6 RNA-seq Library Prep Kit for Illumina® (Vazyme, NR604-02, China) according to the manufacturer's instructions. The purified libraries were quantified using a Qubit 2.0 Fluorometer (Life Technologies, Q32866, USA) and validated using the Agilent 2100 bioanalyzer (Agilent Technologies, 2100 bioanalyzer, USA) to confirm the insert size and calculate the mole concentration. Clusters were generated by cBot using libraries diluted to 10 pM and sequenced on the NovaSeq 6000 sequencing system (Illumina, NovaSeq 6000, USA). For data analysis, FastQC (Version 0.11) was used for raw data quality control. Cutadapt (Version 3.6) was used for adapter sequence removal. HISAT2 (Version 2.2.0) was used for sequence alignment with the human genome hg38, to acquire the sequencing alignment BAM file. HTseq-count (Version 0.11.1) was used for the generation of the gene counts file. Deseq2 (Version 1.38.3) was used for differential transcriptome analysis.

IHC staining
For the second clinical validation cohort, the IHC staining was also performed in bone metastatic lesions in LUAD patients. The expression of SAA1, YWHAE, PSMB6, RPS3A, and EEF2 was examined. Serial paraffin sections were subjected to H&E staining and IHC. Briefly, sections were de-paraffin with xylene, rehydrated, and then blocked with 1% BSA at room temperature for 30 min. After the incubation with primary antibodies at 4 °C overnight, the sections were incubated with HRP-conjugated secondary antibodies at 37 °C for 30 min. Positive reactions were visualized with DAB (Dako, USA), followed by hematoxylin counterstaining. The information of the antibodies we used was listed as follows: (i) EEF2 Polyclonal antibody (Proteintech, 20107-1-AP, dilution 1:1000); (ii) RPS3A Polyclonal antibody (Proteintech, 14123-1-AP, dilution 1:1000); (iii) SAA1 Polyclonal antibody (Proteintech, 16721-1-AP, dilution 1:1000); (iv) 14-3-3E Monoclonal antibody (Proteintech, 66946-1-Ig, dilution 1:3000); (v) PSMB6 Polyclonal antibody (Proteintech, 11684-2-AP, dilution 1:1500).

ELISA
For mice models, we detected the serum level of SAA1 by ELISA kit, provided by EIAab (EIAab, E0885m), and performed our experiments strictly according to the manufacturer's protocol (https://www.eiaab.com.cn/product-cn-elisa_kit/saa1_mouse/). Meanwhile, for human serum samples, we also conducted ELISA for the comparison of serum levels of SAA1 and YWHAE by the corresponding ELISA kit, also provided by EIAab (EIAab, E0885h; E14290h). We strictly followed the manufacturer's protocol as well (https://www.eiaab.com.cn/product-elisa_kit/saa1_human/) (https://www.eiaab.com.cn/product-elisa_kit/saa1_human/; https://www.eiaab.com.cn/product-cn-elisa_kit/1433e_human/).

Cancer cell subtype identification
To isolate cancer cells within our samples, we initiated a multi-step approach leveraging the monocle3 package. Initially, batch-corrected

clustering was executed to unravel inherent cellular heterogeneity. Subsequently, we amalgamated proximate clusters into partitions, enhancing the precision of our analyses. Using the singleR package, we performed cell type annotation, a critical step in pinpointing the cell subtypes of interest. Notably, our scrutiny revealed that cells within Partition 1 encompassed the specific cancer cell subtype we aimed to investigate. This systematic procedure, combining batch-corrected clustering, partition integration, and singleR-based annotation, proved instrumental in accurately identifying and isolating the cancer cell subtypes within the complex cellular landscape of our samples.

Constructing trajectories and pseudo-temporal information with Monocle3

To input temporal information into our single-cell data for subsequent analyses, we employed the monocle3 package to infer pseudo-temporal trajectories. Initially, the learn_graph function was applied to establish a differentiation trajectory. Subsequently, we identified the earliest time point interval, corresponding to the position with the highest cell density in stage III, as the root node. Using the order_cells function, we derived pseudo-temporal information for the dataset, assigning each cell a position along the inferred trajectory. To facilitate subsequent Differential Network Biology analysis, the temporal information needed to be partitioned into distinct stages. This was achieved through the choose_graph_segments function in the monocle3 package. An interactive interface was utilized to segment each trajectory into five distinct temporal subsets, crucial for subsequent Differential Network Biology analysis. This monocle-based approach allowed us to not only construct trajectories but also provided a pseudo-temporal framework for understanding the developmental progression of cells in our single-cell dataset.

DNB selection

To identify DNB, we employed the DNBr package from the laboratory of Professor Luo Nan Chen[15]. The computation of DNB scores was performed using the DNBcompute function, taking the five distinct temporal subsets as inputs. This function calculated DNB scores for various modules, each comprising several genes, across different time stages. Following the computation, the DNB filter function was applied to filter and select the top five modules with the highest DNB scores for each time stage. These modules, characterized by sets of genes ranging from a few to several dozen, represented the most dynamically changing components within the network. Finally, the resultAllExtract function was utilized to export the results, focusing on modules with the most significant changes in DNB scores. These selected modules and their constituent genes represent the dynamic components of the network with the most pronounced variations.

This approach, utilizing the DNBr package, facilitated the systematic identification and extraction of DNB, providing valuable insights into the temporal evolution of network dynamics in the biological system under investigation.

Bone metastatic murine model establishment

LLC and KP lung carcinoma cells were acquired from our institution and authenticated through STR profiling. 4–5 weeks aged male C57BL/6 mice were purchased from Shanghai Slac Laboratory Animal Company, China. They were kept in specific pathogen-free laboratory animal facilities under standard conditions with temperatures of 21–23 °C, 40–60% humidity, and 12 h light/dark cycles, with rodent chow and water ad libitum, and randomized to form the metastasis model. Both cell lines were used for model construction, and totally two animal models were established ($n = 16$ for each group). Briefly, cells were counted as $2 \times 10^5/0.1$ ml for intracardiac injection, and followed by in-vivo luciferin imaging after two weeks (Perkin Elmer). Fluorescence on the limbs indicating bone metastasis was dissected for visual verification.

Murine serum SAA1 detection in bone metastasis models

Initially, we anesthetized mice and used a 1 ml syringe to align with the midpoint line of the mouse's upper limbs for cardiac puncture and blood collection. Then, we collected blood into a 1.5 ml EP tube and let it stand at room temperature for 2 h until blood clots. The blood samples were then centrifuged at 4 °C at 1400×g for 10 min to harvest the serum samples. Subsequently, we detected serum SAA1 by ELISA, as mentioned above.

Intermediate-state metastatic cell selection

To identify intermediate status, we initially exported the expression data of genes of interest along with pseudo-temporal information as a data frame. Subsequently, we employed the ggplot package to visualize the gene expression over time using the Locally Weighted Scatterplot Smoothing (LOESS) method for data fitting. Afterward, we computed the derivatives of the gene expression curves and applied the LOESS method once again to fit and visualize the obtained derivatives. This process allowed us to capture the changes in gene expression trends and identify points of maximum rate of change. Among the multitude of genes considered, we observed numerous points of extremum in the derivatives along the pseudo-temporal axis. The intersection of these extremum points was selected as the set of intermediates. This method, utilizing LOESS for data fitting and derivative analysis, enabled the systematic identification of intermediate status in the gene expression profiles over pseudo-temporal progression, providing key insights into critical transition points within the biological system under investigation.

Integrated neural network for single-cell metastatic status prediction

The raw data of the scRNA-seq, originally comprising tens of thousands of features, was dimensionally reduced to a few dozen features identified along different trajectories we previously discovered. Given the prior segmentation of pseudo-temporal intermediate statuses, the dataset was filtered to retain only cancer cells under post-intermediate status. This filtered scRNA-seq dataset was then employed as the training set for a multi-class neural network. The model's initial layer, namely the input layer, contained nodes equal to the number of features and accepted input data. This would be followed by two hidden layers; the first comprised ten nodes, and the second contained five nodes, both utilizing the ReLU activation function to learn and extract features from the input data. The final layer was the output layer, where, informed by an extensive literature review, we noted that researchers generally did not further distinguish between bone metastasis and "Boneplus" metastasis, as well as lung metastasis (intra-pulmonary metastasis) and pleural metastasis. Therefore, we merged the five trajectories into three categories: bone metastasis (including both bone and 'Boneplus' metastasis), brain metastasis, and lung metastasis (including both intra-pulmonary/lung and pleural metastasis). The number of nodes in this layer equaled the number of categories, using the softmax activation function to output a probability for each category based on the features learned in the hidden layers. The model utilized categorical cross-entropy as the loss function, a standard choice for multi-class classification problems, and employed the Adam optimizer, a widely used algorithm capable of automatically adjusting the learning rate. After training, the external data (GSE123902) was used as the validation set. We selected a LUAD patient with primary tumors (GSM3516665) and a LUAD patient with bone metastasis (GSM3516664) from this set, identified cancer cells through singleR, and then input these into the neural network for type prediction. The results were visualized using Sankey diagrams and pie charts, and the performance of the neural network, trained with our data, was evaluated using ROC curves.

Construction of an integrated model classifier

We utilized various machine learning models to construct classifiers for ensemble models, using the significant genes we selected as features to classify scRNA-seq data and verify whether the significant genes we previously analyzed could serve as markers for classifying different developmental trajectories. The classifier of the integrated model could improve the performance and robustness of the overall model by combining the prediction results of multiple single models (base classifiers). The integrated model could effectively improve the prediction accuracy. Different models might perform better on different sub-data sets. By integrating these models, we could comprehensively take advantage of each model to improve the overall performance. Single models (especially complex models) might perform well on training data but perform poorly on test data, resulting in overfitting. By combining the prediction results of multiple models, the integration method could reduce the risk of overfitting and improve the generalization ability of models on new data. The integrated models were more robust because they did not depend on the performance of a single model. If a base classifier performed poorly, the prediction results of other classifiers could compensate for its shortcomings, making the prediction of the ensemble model more stable and reliable. We created Random Forest (RandomForestClassifier)[55], Gradient Boosting Tree (GradientBoostingClassifier)[56], XGBoost (XGBClassifier)[57], and LightGBM (LGBMClassifier)[58] models. The parameters used were all n_estimators=100, random_state=42, and then the soft voting (VotingClassifier) was used[59] to integrate multiple models, which was an ensemble learning technique that made final classification decisions by averaging the prediction probabilities of multiple base classifiers. It not only considered the prediction results of the base classifiers but also the prediction probabilities of each classifier. This enabled the model to evaluate the possibility of each category more carefully, so as to make more accurate decisions. Because the soft voting method integrated the probability prediction results of multiple models, it could provide higher prediction accuracy than the single model or hard voting method in most cases. This was because it better captured the prediction information of each model. By integrating the prediction probability of multiple models, the soft voting method could effectively reduce the uncertainty and volatility in the prediction of a single model. This helped to improve the stability and robustness of the model on different data sets. By integrating the prediction of multiple models, the soft voting method could effectively reduce the risk of a single model overfitting training data, and improve the performance of the model on test data.

Then, we used cross-validation (cross_val_score) to evaluate the accuracy of the integrated model, specifically using *K*-fold cross-validation, and outputting the results of cross-validation and average accuracy. Cross-validation was a method of evaluating the performance of machine learning models, which involved repeatedly training and testing a model by dividing the dataset into multiple subsets. Its advantage lies in the ability to reduce the randomness and volatility of model evaluation results through multiple partitions and repeated training and testing. Cross-validation provided more stable and reliable performance estimates, and could fully utilize available data, especially in situations with limited data volume. Each data point would be used for testing in a certain partition, making the evaluation more comprehensive. At the same time, cross-validation helped detect and reduce model overfitting by training and testing on different subsets of data. The model needed to perform well on multiple partitioned test sets in order to be considered reliable. Through multiple verifications, cross-validation could provide the expected performance of the model on unseen data, thereby better estimating the model's generalization ability. We used a confusion matrix and ROC to visualize the results and average accuracy of cross-validation.

Finally, we used an external dataset as the validation set to validate the classifier of the ensemble model, demonstrating the classifier's ability to classify external data. We utilized ROC to visualize the results of cross-validation and external validation.

Statistics and reproducibility

For the integrated single-cell data, the standard Seurat quality control workflow was applied during preprocessing. When creating the Seurat object, the parameters were set to min.cells = 3 and min.features = 200 to filter out low-count cells. The percentage of mitochondrial gene counts for each cell was calculated using Seurat's PercentageFeatureSet function. Further filtering of low-quality cells was performed by selecting cells with nFeature_RNA > 200 & nFeature_RNA < 2500 & percent.mt <20 using the subset function. After clustering the entire dataset with Monocle3, Partition 1, which corresponds to the cells required for the study, was subsetted for further analysis. For analyzing differential changes along the cell trajectories and conducting DNB analysis, only the cells along the trajectory were selected, and the choose_graph_segment() function in Monocle3 was used for subsetting.

R 4.3.1 was used for statistical analysis and bioinformatics analysis. $P < 0.05$ was considered as statistically significant. Unless specifically described, the Wilcoxon test was used for comparative analysis.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The mass spectrometry proteomics data generated in this study have been deposited in the ProteomeXchange Consortium through the iProx partner repository database under accession code IPX0009723000. The bone metastasis bulk RNA-seq data generated in this study have been deposited in the NCBI GEO database under accession code GSE225208. The bone metastasis single-cell RNA-seq data generated in this study have been deposited in the NCBI SRA database under accession code SRR29634661,SRR29634663, SRR29634662, SRR29634658, SRR29634659, SRR29634660, under the BioProject AccessionPRJNA1129208 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1129208]. The lung cancer Single-cell RNA-seq data used in this study are available in the NCBI GEO database under accession code GSE148071. Raw single-cell RNA sequencing data and raw protein data can be found on Zenodo. The processed single-cell RNA data and metadata can be downloaded as a Seurat RDS file from Zenodo at https://zenodo.org/records/13764759. Source data are provided with this paper.

Code availability

Codes related to this article can be accessed via the link https://github.com/zxsanalytics/MODNB.

References

1. Sung, H. et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
2. Leiter, A., Veluswamy, R. R. & Wisnivesky, J. P. The global burden of lung cancer: current status and future trends. *Nat. Rev. Clin. Oncol.* **20**, 624–639 (2023).
3. Esposito, M., Ganesan, S. & Kang, Y. Emerging strategies for treating metastasis. *Nat. Cancer* **2**, 258–270 (2021).
4. Gerstberger, S., Jiang, Q. & Ganesh, K. Metastasis. *Cell* **186**, 1564–1579 (2023).
5. Tang, W.-F. et al. Timing and origins of local and distant metastases in lung cancer. *J. Thorac. Oncol.* **16**, 1136–1148 (2021).
6. Hu, Z., Li, Z., Ma, Z. & Curtis, C. Multi-cancer analysis of clonality and the timing of systemic spread in paired primary tumors and metastases. *Nat. Genet.* **52**, 701–708 (2020).

7. Popper, H. H. Progression and metastasis of lung cancer. *Cancer Metastasis Rev.* **35**, 75–91 (2016).

8. Milovanovic, I. S., Stjepanovic, M. & Mitrovic, D. Distribution patterns of the metastases of the lung carcinoma in relation to histological type of the primary tumor: an autopsy study. *Ann. Thorac. Med.* **12**, 191–198 (2017).

9. Wu, K. et al. Exosomal miR-19a and IBSP cooperate to induce osteolytic bone metastasis of estrogen receptor-positive breast cancer. *Nat. Commun.* **12**, 5196 (2021).

10. Liu, W., Zhao, J. & Wei, Y. Association between brain metastasis from lung cancer and the serum level of myelin basic protein. *Exp. Ther. Med.* **9**, 1048–1050 (2015).

11. Ayan, A. K. et al. Is there any correlation between levels of serum osteopontin, CEA, and FDG uptake in lung cancer patients with bone metastasis? *Rev. Esp. Med Nucl. Imagen Mol.* **35**, 102–106 (2016).

12. Teng, X. et al. Development and validation of an early diagnosis model for bone metastasis in non-small cell lung cancer based on serological characteristics of the bone metastasis mechanism. *EClinicalMedicine* **72**, 102617 (2024).

13. Liu, X. et al. Detection for disease tipping points by landscape dynamic network biomarkers. *Natl Sci. Rev.* **6**, 775–785 (2019).

14. Wu, X., Chen, L. & Wang, X. Network biomarkers, interaction networks and dynamical network biomarkers in respiratory diseases. *Clin. Transl. Med.* **3**, 16 (2014).

15. Chen, L., Liu, R., Liu, Z.-P., Li, M. & Aihara, K. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci. Rep.* **2**, 342 (2012).

16. Aihara, K., Liu, R., Koizumi, K., Liu, X. & Chen, L. Dynamical network biomarkers: Theory and applications. *Gene* **808**, 145997 (2022).

17. Liu, H. et al. Single-cell transcriptomics reveal DHX9 in mature B cell as a dynamic network biomarker before lymph node metastasis in CRC. *Mol. Ther. Oncolytics* **22**, 495–506 (2021).

18. Jiang, Z. et al. SMAD7 and SERPINE1 as novel dynamic network biomarkers detect and regulate the tipping point of TGF-beta induced EMT. *Sci. Bull. (Beijing)* **65**, 842–853 (2020).

19. Chen, P., Liu, R., Chen, L. & Aihara, K. Identifying critical differentiation state of MCF-7 cells for breast cancer by dynamical network biomarkers. *Front. Genet.* **6**, 252 (2015).

20. Yang, B. et al. Dynamic network biomarker indicates pulmonary metastasis at the tipping point of hepatocellular carcinoma. *Nat. Commun.* **9**, 678 (2018).

21. Fang, Z. et al. Oxidative stress-triggered Wnt signaling perturbation characterizes the tipping point of lung adeno-to-squamous transdifferentiation. *Signal Transduct. Target Ther.* **8**, 16 (2023).

22. Liu, R., Wang, X., Aihara, K. & Chen, L. Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. *Med. Res. Rev.* **34**, 455–478 (2014).

23. Liu, R., Chen, P. & Chen, L. Single-sample landscape entropy reveals the imminent phase transition during disease progression. *Bioinformatics* **36**, 1522–1532 (2020).

24. Győrffy, B. Discovery and ranking of the most robust prognostic biomarkers in serous ovarian cancer. *Geroscience* **45**, 1889–1898 (2023).

25. Goldstraw, P. et al. The IASLC Lung Cancer Staging Project: proposals for revision of the TNM stage groupings in the forthcoming (Eighth) edition of the TNM Classification for Lung Cancer. *J. Thorac. Oncol.* **11**, 39–51 (2016).

26. Deetman, P. E., Bakker, S. J. L. & Dullaart, R. P. F. High sensitive C-reactive protein and serum amyloid A are inversely related to serum bilirubin: effect-modification by metabolic syndrome. *Cardiovasc. Diabetol.* **12**, 166 (2013).

27. Prüfer, N., Kleuser, B. & van der Giet, M. The role of serum amyloid A and sphingosine-1-phosphate on high-density lipoprotein functionality. *Biol. Chem.* **396**, 573–583 (2015).

28. Gabay, C. & Kushner, I. Acute-phase proteins and other systemic responses to inflammation. *N. Engl. J. Med.* **340**, 448–454 (1999).

29. Cho, W. C. S., Yip, T. T., Cheng, W. W. & Au, J. S. K. Serum amyloid A is elevated in the serum of lung cancer patients with poor prognosis. *Br. J. Cancer* **102**, 1731–1735 (2010).

30. Milan, E. et al. SAA1 is over-expressed in plasma of non small cell lung cancer patients with poor outcome after treatment with epidermal growth factor receptor tyrosine-kinase inhibitors. *J. Proteom.* **76**, 91–101 (2012).

31. Wang, J.-Y. et al. Elevated levels of serum amyloid A indicate poor prognosis in patients with esophageal squamous cell carcinoma. *BMC Cancer* **12**, 365 (2012).

32. Findeisen, P. et al. Serum amyloid A as a prognostic marker in melanoma identified by proteomic profiling. *J. Clin. Oncol.* **27**, 2199–2208 (2009).

33. Kosari, F. et al. Clear cell renal cell carcinoma: gene expression analyses identify a potential signature for tumor aggressiveness. *Clin. Cancer Res.* **11**, 5128–5139 (2005).

34. Yuan, C. et al. The STAT3 inhibitor stattic overcome bortezomib-resistance in multiple myeloma via decreasing PSMB6. *Exp. Cell Res.* **429**, 113634 (2023).

35. Shi, C.-X. et al. Proteasome subunits differentially control myeloma cell viability and proteasome inhibitor sensitivity. *Mol. Cancer Res.* **18**, 1453–1464 (2020).

36. Bian, Y. et al. Identification and validation of a proliferation-associated score model predicting survival in lung adenocarcinomas. *Dis. Markers* **2021**, 3219594 (2021).

37. Yang, Y.-F. et al. YWHAE promotes proliferation, metastasis, and chemoresistance in breast cancer cells. *Kaohsiung J. Med. Sci.* **35**, 408–416 (2019).

38. Jiang, Y. et al. RNA-Binding Protein COL14A1, TNS1, NUSAP1 and YWHAE are valid biomarkers to predict peritoneal metastasis in gastric cancer. *Front. Oncol.* **12**, 830688 (2022).

39. Li, Z. et al. The novel miR-873-5p-YWHAE-PI3K/AKT axis is involved in non-small cell lung cancer progression and chemoresistance by mediating autophagy. *Funct. Integr. Genom.* **24**, 33 (2024).

40. Hu, Z. B., Minden, M. D., McCulloch, E. A. & Stahl, J. Regulation of drug sensitivity by ribosomal protein S3a. *Blood* **95**, 1047–1055 (2000).

41. Zhou, C. et al. High RPS3A expression correlates with low tumor immune cell infiltration and unfavorable prognosis in hepatocellular carcinoma patients. *Am. J. Cancer Res.* **10**, 2768–2784 (2020).

42. Slizhikova, D. K., Vinogradova, T. V. & Sverdlov, E. D. [The NOLA2 and RPS3A genes as highly informative markers for human squamous cell lung cancer]. *Bioorg. Khim.* **31**, 195–199 (2005).

43. Chen, C.-Y. et al. Sumoylation of eukaryotic elongation factor 2 is vital for protein stability and anti-apoptotic activity in lung adenocarcinoma cells. *Cancer Sci.* **102**, 1582–1589 (2011).

44. Cheng, D. et al. PRMT7 contributes to the metastasis phenotype in human non-small-cell lung cancer cells possibly through the interaction with HSPA5 and EEF2. *Onco Targets Ther.* **11**, 4869–4876 (2018).

45. Uramoto, H. & Tanaka, F. Recurrence after surgery in patients with NSCLC. *Transl. Lung Cancer Res.* **3**, 242–249 (2014).

46. Nielsen, H., Teufel, F., Brunak, S. & von Heijne, G. SignalP: the evolution of a web server. *Methods Mol. Biol.* **2836**, 331–367 (2024).

47. Teufel, F. et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–1025 (2022).

48. Shrivastava, K. et al. Identification and In silico analysis of proline-glutamate/proline-proline-glutamate proteins of *Mycobacterium tuberculosis* complex: a comparison of computational web-based tools. *Int. J. Mycobacteriol.* **12**, 248–253 (2023).

49. Singhal, N. et al. Efficacy of signal peptide predictors in identifying signal peptides in the experimental secretome of *Picrophilous*

torridus, a thermoacidophilic archaeon. *PLoS ONE* **16**, e0255826 (2021).

50. Almagro Armenteros, J. J. et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).

51. Vakili, O. et al. Finding appropriate signal peptides for secretory production of recombinant glucarpidase: an in silico method. *Recent Pat. Biotechnol.* **15**, 302–315 (2021).

52. Hu, A. et al. Cancer serum atlas-supported precise pan-targeted proteomics enable multicancer detection. *Anal. Chem.* **95**, 862–871 (2023).

53. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 (2020).

54. Wu, F. et al. Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nat. Commun.* **12**, 2540 (2021).

55. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

56. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).

57. Chen, T. & Guestrin, C. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794 (Association for Computing Machinery, San Francisco, California, USA, 2016).

58. Ke, G. et al. in Proceedings of the 31st International Conference on Neural Information Processing Systems 3149–3157 (Curran Associates Inc., Long Beach, California, USA, 2017).

59. Kuncheva, L. I. Combining Pattern Classifiers: Methods and Algorithms. (Wiley-Interscience, 2004).

Acknowledgements

Author contributions

Conceptualization: X.Z., K.X., L.C., C.Z.; Investigation: X.Z., K.X., Y.W., G.G., F.W.; Funding acquisition: C.Z.; Supervision: L.C., C.Z.; Writing—original draft: X.Z., Y.W.; Writing—review & editing: C.Z., L.C.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-53849-3.

Correspondence and requests for materials should be addressed to Luonan Chen or Caicun Zhou.

Peer review information *Nature Communications* thanks Nguyen Quoc Khanh Le and the other, anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.