Inferring cell trajectories of spatial transcriptomics via optimal transport analysis

Xunan Shen,[1,9] Ke Huang,[2,3,9] Lulu Zuo,[4,9] Zhongfei Ye,[1,9] Zeyu Li,[2,5] Qichao Yu,[2,5] Xuanxuan Zou,[2] Xiaoyu Wei,[2,6] Ping Xu,[2,7], Xin Jin,[2] Xun Xu,[2,8,]* Liang Wu,[2,3,]* Hongmei Zhu,[1,]* Pengfei Qin [2,3,10,]*

Affiliations:

[1]BGI Research, Beijing 102601, China

[2]BGI Research, Shenzhen 518083, China

[3]BGI Research, Chongqing 401329, China

[4]BGI, Tianjin 300308, China

[5]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China.

[6]BGI Research, Hangzhou 310030, China

[7]BGI College & Henan Institute of Medical and Pharmaceutical Sciences, Zhengzhou University, Zhengzhou 450000, China.

[8]Guangdong Provincial Key Laboratory of Genome Read and Write, Shenzhen, Guangdong, China.

[9]These authors contributed equally to this work

[10]Lead contact

*Correspondence: qinpengfei@genomics.cn (P.Q.), zhuhongmei@genomics.cn (H.Z.), wuliang@genomics.cn (L.W.), xuxun@genomics.cn (X.X)

Abstract

The integration of cell transcriptomics and spatial locations to accurately organize spatial trajectory is remaining challenges. Here we introduce spaTrack, a trajectory inference method incorporating both expression and distance cost of cell transition in spatial context. spaTrack is capable to construct fine local trajectories within single tissue section, as well as trace cell dynamics across multiple sections with temporal intervals. To capture potential dynamic drivers, spaTrack models the fate of a cell as a function of expression profile along temporal intervals driven by transcription factors. Applying spaTrack, we successfully disentangle spatiotemporal trajectories of axolotl telencephalon regeneration and mouse midbrain development. Furthermore, we uncover diverse malignant lineages expanding in a primary tumour. One of the lineages with upregulated extracellular matrix organization implants to the metastatic site and subsequently colonizes to a secondary tumour. Overall, spaTrack greatly facilitates trajectory inference from spatial transcriptomics, providing insights in cell differentiation of broad areas.

Introductions

Trajectory inference (TI) provides important insights in cell differentiation and biological process. Currently, there are numerous TI methods available, but most of them are designed for single cell RNA sequencing (SC) data and challenged by complex topologies. Frequently applied TI methods for SC data, e.g., Monocle2/3 [1,2], PAGA [3], Slingshot [4], stemID [5], Tscan [6], URD [7] et al. usually construct a skeleton frame of cell differentiation averaged or optimally extracted from a single cell embedding generated by dimension reduction (DR) techniques such as PCA, ICA, UMAP [8], Diffusion Map [9] and ForceAtlas2 [10]. Many of the existing approaches are limited to simple linear or branching topologies and overlook morphometric space. However, the dynamics of biological systems, such as embryogenesis or tumorigenesis, are often complex and spatially organized. Recent advances in spatial transcriptomics sequencing (ST) technologies provide an opportunity to simultaneously reveal both transcriptomic and spatial patterns of development, which the SC is unable to capture. Although TI methods for SC data can be directly applied to ST data, cell transition is spatially heterogenous due to their spatial location and surrounding cell environment, which conflicts with the continuous and averaged trajectory inferred from cell clusters using these methods.

To capture the single cell dynamics, RNA velocity has introduced alternative ways to study cellular differentiation of SC data [11-14]. It describes the rate of transcriptional dynamics for an individual gene at a given time point based on the ratio of its spliced and unspliced messenger RNA (mRNA). Great efforts have been made to develop various TI methods based on RNA velocity. However, estimation of RNA velocity is often found to be less robust for indicating cellular transitions due to several fundamental limitations. There are low contents of unspliced RNAs in SC data and intronic regions could not be fully captured; Conventional RNA-velocity methods assume constant transcription rates but the cell population/states are heterogeneous which usually leads to nonsensical backward trajectories. Metabolic-labeling data has been explored in velocity estimate, which measures the synthesis and degradation of labeled RNA within a known period of time directly and overcomes some drawbacks of conventional RNA-velocity methods [11]. However, the metabolic-labeling data is not always available in studies. In addition, most RNA-velocity methods do not reconcile the physical proximity of cells into trajectory inference. Besides, estimating RNA-splicing rate involves mapping BAM file which can be computationally time-consuming and require significant computational resources. Thus, there is a strong need for an alternative approach to efficiently generate single-cell spatial trajectories.

Optimal transport (OT) is a method to find least-cost schemes of coupling distributions and provide intuitively quantifications of their distance between multiple datasets or samples represented as distributions [15]. OT has recently been used for transcriptomic data analysis, including cell-cell communication inference [16,17], lineage study of hematogenesis [18] and annotation of ST data [17,19]. With the natural advantage of OT, it is capable to conveniently incorporate the profile of both gene expression and transition distance of cells into the cost matrix to solve the OT problem. A fully captured cell to cell transition matrix will facilitate the construction of trajectories with fine local details.

In this study, we introduced an efficient method, namely spaTrack, to construct cell trajectory at single cell resolution of spatial context, which utilizes OT frameworks and sensitively reconciles both gene expression and physical distance. When dealing with multiple samples from a time series, spaTrack can construct a dynamic map of cell migration and differentiation across all tissue sections, providing a comprehensive view of transition behavior over time. In our study, we have successfully applied spaTrack to reconstruct cell trajectories in spatial manners for a various of biological systems, including regeneration of injured axolotl telencephalon, development of dorsal midbrain of mouse embryo, and tumorigenesis and metastasis. Our approach has significantly facilitated the study of cell kinetics of ST data in various cases.

Results

Inferring cell trajectories from single ST data

spaTrack utilizes Optimal Transport (OT) as a foundation to infer the transition probability between cells of ST data in a single sample, by incorporating both gene expression profiles and cell location information. Cells that are distant from each other in expression level and physical space will have higher transfer costs, which indicates a lower transition probability or longer time interval in the biological process. The schematics of the algorithm and workflow of spaTrack in solving the TI problem in single ST data are demonstrated in Fig. 1a, which include 1) scaling both gene expression and physical distance into a cost matrix; 2) solving the optimal transport problem by incorporating an entropy term; 3) constructing the vector field of cell velocity from transferring probability; 4) organizing spatial trajectories from the vector field of cell velocity; 5) optimizing the path between a starting cell and corresponding ending cell using the least action path method; 6) identifying pseudotime-dependent genes using a generalized additive model (GAM) to fit the dynamics

of gene expression along a trajectory.

The cost matrix consists of two components: expression difference and physical distance, both of which employ Euclidean distance as the metric. The algorithm performs dimension reduction from original features and the Euclidean distance in the embedding space is used as the measurement of the difference of expression profile. By appropriately scaling these two distance measurements, we fine-tuned the relative importance of each one by adjusting the weight parameters of α_1 and α_2. The algorithm maximizes the distance of a cell to itself to prevent self-transitions. To solve the optimal transport problem using the cost matrix generated above, an entropic regularization term $\Omega(\gamma) =$ is introduced. This term helps to produce a smooth probability distribution [20]. The resulting OT probability matrix displays the likelihood of each cell being transferred to other cells with the minimum cost incorporating both expression profile and spatial coordinates (Fig. S1a and S1b). spaTrack makes the assumption that cells with higher transferring probabilities are more closely related during the development. Assigning the starting cells, the algorithm can reorder all other cells according to their transferring probabilities relative to the starting cells (see Methods). Subsequently, spaTrack creates a vector field of the cell velocity averaged from the transferring probabilities and directions of all cells in the neighbourhood. Streamlines of the vector field are finally organized and smoothed as the spatial trajectories. Furthermore, we adapted the least action path (LAP) algorithm[11] to construct the optimal path of differentiation between a starting cell and an ending cell. Specifically, the vector field of cell velocity of the transition probability in a spatial neighbourhood is computed to replace the RNA-splicing velocity. A set of neighbouring cells are mapped to the inferred optimal path to determine cell orders and pseudotimes along the differentiation. The arc length between the mapped anchor point and the starting cell is normalized as the pseudotime of a cell.

Sometimes, ST data from single sample will not capture all cell states of a complete biological process, especially when the process is time coordinated. However, integrating multiple samples will lose the spatial coordinates of each tissue section. spaTrack provides an integrating strategy to accommodate this situation. spaTrack separately computes the transition probability and cell velocity for each ST data; And then integrates the vector fields of all datasets to organize the overall trajectories in an UMAP embedding. This allows for a more accurate and complete representation of the cell trajectories without losing the spatial coordinates of each data.

Tracing cells across multiple ST data of a time series

Multiple ST data sampled from a time series will provide a spatiotemporal transcriptomic view of biological systems. In order to align cells and generate trajectories across samples of different time intervals, spaTrack adapts an unbalanced and structured OT algorithm considering the uneven expression mass and distributions of samples. ST data of different time intervals are treated as different distributions (Fig.1b). The optimization problem of this OT includes three terms: a measure of the expression profile differences between cells of the two samples; a measure of the spatial distance differences between paired-cells of the two samples; and a measure of the KL divergence (see Methods). spaTrack extends a distance consideration in the optimization problem: If a pair of cells (i, k) in time t_i is mapped to a pair of cells (j, l) in the next time t_{i+1} with high probability, the distance between cells i and k in the time t_i is close to the distance between cells j and l in the next time t_{i+1}. Therefore, three matrices are included in spaTrack to solve the optimization problem: a gene expression dissimilarity matrix, a spatial distance matrix for cells at time t_i , and a spatial distance matrix for cells at the next time t_{i+1}. The resulting transport plan depicts the transition probabilities of individual cells across ST samples.

After computing transport maps between two adjacent time points, the next step is to extend the transitions to the next time interval. To achieve this, we adopt the Markov assumption that the developmental process follows a memoryless property. Therefore, the long-range transitions could be inferred by composing transport maps using matrix multiplication (see Methods). At each time point, we start from the cells transported from the previous time point and infer their subsequent transitions. This progressive method has the advantage of avoiding direct inference of cell transition over long time intervals, and producing more coherent and credible results. Following this approach, spaTrack is able to establish a long-term and continuous transition map of cell trajectory.

Modeling dynamic driven factors

spaTrack explores the driven factors regulating cell trajectories, which helps to construct the regulatory network underlying cell differentiation. spaTrack establishes a global regulatory model to interpret the connection between the expression profiles of transcription factors (TFs) at current time and of targeted genes at later time. We propose to set up a regression model to learn the positive/negative regulation of TFs, which is implemented in a regression framework (Fig.1c). This regression model works for both data of discrete time points when the transfer

map has been constructed, and data of single sample assigned with continuous pseudotime. To validate the inferred regulation network of TFs and targeted genes, we extract TF-target pairs with top weights and examine their relationship of expression profiles across all cells. As expected, correlations are observed for those TF-target pairs with top weights (Fig.S1c).

According to the highlighted features of the algorithm (Fig. 1a-1c), spaTrack takes several advantages in TI. Firstly, spaTrack could accurately uncover spatial details of cell trajectory without the limitation of global topology. Secondly, a single sample may not capture all cell states involved in the complete developmental process. To overcome this limitation, spaTrack could fully generate and extend the complete trajectories by integrating the transition matrix of multiple samples, without losing their spatial information. Thirdly, direct cell mapping across multiple sections could vividly depicted cell trajectories along a time series. Fourthly, spaTrack captures potential driven factors and networks along the time intervals underlying cell differentiation. Furthermore, spaTrack exhibits low resource requirements in terms of both power consumption and computing memory while maintaining a high processing speed. Specifically, the generation of trajectories using 5k cells with 20,000 features can be accomplished within one minute, while utilizing a modest memory allocation of 6.9 GB (Fig.S1d).

Fine local trajectories of axolotl telencephalon regeneration

We applied spaTrack to reconstruct the spatially detailed trajectories of the regeneration of axolotl telencephalon after injury. Brain regeneration requires the coordination of complex responses in time and region-specific manners. Taking into consideration the spatial coordinates of cells, spaTrack can capture local details of cell trajectory that may be discontinuous in space. Axolotl is a model for studying brain regeneration as its ability to regenerate lost cortical cells after injury. We collected ST data of axolotl samples from a time series after injury[21], including samples of 5 days (D5), 10 days (D10), 15 days (D15), and 20 days (D20) (Fig.2a, Fig.S2a and S2b), to uncover the regeneration process in space.

D15 shows wound closure with enriched cell types of progenitors and immature neuron cells, including ependymoglial cells (EGC), ractive EGC (reaEGC), immature neuron (IMN), regeneration intermediate progenitor cell (rIPC) and *nptx*$^{+}$ lateral pallium excitatory neuron (nptxEX) (Fig.2b). reaEGC responses to injury and starts the tissue repair suggested by previous studies[21,22], presenting high proliferative activity (Fig.S2c and S2d). Adjacent layers

of intermediated cells were observed between reaEGC and nptxEX across the wound area, indicating their transitions during regeneration (Fig.2c-2e). spaTrack uncovered the local details of regeneration, generating the streamlines of cell differentiation as three spatial lineages in the tissue section of D15 (Fig. 2f). The first lineage ranged from wound center to the right-edge of telencephalon, which was reaEGC - rIPC - IMN - nptxEX. The second lineage was the regeneration of wound area on the dorsal region, which described the complex transitions of between reaEGC and IMN, rIPC, and dorsal palliumexcitatory neuron (dpEX). The third lineage presented the normal developmental trajectory from *Wnt*$^{+}$EGC (wntEGC) to medial pallium excitatory neurons (mpEX). These results depicted the differentiation from reaEGC to intermediate and mature neurons during regeneration after injury, which were consistent with previous reports[21,22]. Importantly, the trajectory of lineage 2 in the wound area was discontinuous with the normal developmental trajectory as shown in lineage 3, which were separate processes in time and space. Therefore, spaTrack reconstructed the spatial details of the regeneration trajectories, and explored different differentiations that were temporally and spatially discontinuous.

Single sample presents only a subset of cell types involved in the regeneration process, with sparse cell population and continuity (Fig.2g, Fig.S2e). However, integrating all ST samples of D5, D10, D15 and D20 will lose the spatial coordinates of each axolotl tissue section. To address this issue, spaTrack implements an integrating framework to separately calculate cell-transition probability of each axolotl section, and next integrate all transition matrix for the inference of complete trajectory of regeneration. Transition probability was calculated for each section by spaTrack and combined into a sparse probability matrix (Fig. 2h). Velocity vectors and streamlines were generated according to the integrated probability matrix and visualized on their UMAP embeddings (Fig.2i). The complete process of regeneration was presented by spaTrack starting from reaEGC to nptEX. Abundant intermediated cell states of rIPC and IMN were also captured showing better continuity than only one sample (Fig.2i, Fig.S2f).

Tracing neuron cells across mouse embryos of a time series

spaTrack provides a novel strategy to trace cells across multiple ST samples by direct mapping cells via an unbalanced OT strategy. Development of mouse embryos requires strict spatial-temporal organization. To probe the dynamics of early neurogenesis, we applied spaTrack on the ST data of developing dorsal midbrain of mouse embryos at day 12.5 (E12.5),

14.5 (E14.5), and 16.5 (E16.5) (Fig.3a, Fig.S3a and S3b). Radial glia cells (RGC) are reported as the progenitors of both neuroblasts (NeuB) and glioblasts (GlioB)[23], but their spatial-temporal transitions are not well characterized. RGC shows abundant and wide distribution in E12.5 but decreases in E14.5 and E16.5. NeuB and GlioB expand in E14.5 and E16.5, and are not evenly distributed along the spatial axis (Fig.3b). Using spaTrack, cells of E12.5 were optimally transported to E14.5, and the successfully transported cells of E14.5 were subsequently transported to E16.5 (Fig.3c), presenting their dynamic differentiation across time. At embryonic stage of E14.5, 81% of the successfully transported NeuB cells were found originating from RGC of E12.5, while the corresponding number for GlioB was 74%. Subsequently, at stage E16.5, 51% of the successfully transported NeuB and 42% of GlioB were transported from RGC of stage E14.5, indicating the RGC is the main source of NeuB and GlioB (Fig.3c). Visualizing these mappings in space, we directly observed the coordinately organization of differentiation in each time point (Fig.3d and 3e, Fig. S3c). Differentiations of RGC-NeuB and RGC-GlioB were restricted to different regions and embryonic stages. RGC in the rostral axis mainly differentiated into NeuB, and RGC in the dorsal and caudal regions differentiated into GlioB. RGC-NeuB differentiation occured from E12.5-E14.5, while RGC-GlioB differentiation arised between E14.5-E16.5. All of these results of spaTrack suggested neurogenesis and gliogenesis were asynchronous and spatially heterogeneous, consistent with previous findings [23].

Furthermore, driven factors of the neuron differentiation were investigated by a regression model in spaTrack. A regulatory network was built to present the connection between TFs and targets along the temporal intervals (Fig.3e). Several TFs were highlighted by our method. Ybx1 reported as a crucial factor for forebrain specification and restricting mid-hindbrain growth in mouse embryo, fine-tunes the spatiotemporal expression of neurodevelopmental genes [24]. Another TF Sox11 is required in embryonic neurogenesis and Sox11-depleted embryos develop small and disorganized brains, accompanied by transient deficits in neural progenitor cells [25]. Therefore, spaTrack could provide valuable reference and methodological support for the advancement of the neuroscience field.

Recovering the trajectories of tumourigenesis

Intratumoural heterogeneity manifests as spatial heterogeneity, which describes the uneven distribution of diverse malignant subclones within tumour, and as temporal heterogeneity, referring dynamic variations in tumour populations and molecules over time [26,27].

Intratumoural heterogeneity drives tumour progress and drug resistance, creating the need to quantitatively investigate tumour diversity and dynamics over space and time.

We collected ST data from a primary tumour section of intrahepatic cholangiocarcinoma (ICC) [28], covering the regions of both intratumour and tumour boundary (Fig.4a, Fig.S4a-S4c). Previous study[28] detected strong immunosuppression and metabolic reprogramming in the invasive zone of tumour boundary, suggesting the spatially diversity of tumour progress. Eight subclones were identified in the ST data of the primary ICC tumour (P0-P7) (Fig.4b and 4c), with P0 showed pronounced expression of stemness markers and cell cycling genes (Fig.4d and 4e, Fig.S4d). Applying spaTrack to reconstruct cell trajectory of malignant cells assigning P0 as starting cells (Fig.S4e), we identified three diverse lineages starting from P0 and spanning the tumour space in three directions (Fig. 4f and 4g). Lineage 1 (P0-P1-P2-P7) expanded to the border region between tumour and hepatic cells. Lineage 2 (P0-P3-P4) extended to the tumour bottom and lineage 3 (P0-P5-P6) elongated along the top area. To determine biological difference among the three lineages (Fig.4g), we identified trajectory-depended genes by fitting a generalized additive model between the pseudotime and gene expression along the optimal path, inferred using the LAP method (Fig.4h and 4i, Fig.S4f and S4g). In the associated genes of lineage 1, COL1A1 is a major component of the tumour extracellular matrix releated with tumour development and immune profile[29]. SAA1 and SAA2 lead to the recruitment and polarization of macrophages, promoting local immunosuppression [28]. Annotating all genes associated with lineage 1, we observed significant enrichment of ECM organization and regulation of platelet and neutrophils (Fig. 4j), which involve in tumour migration, metastasis, and immunosuppression[28,30]. The GSEA scores of ECM and EMT pathways further presented the metastatic potential of lineage 1 (Fig.4k). Additionally, spaTrack constructed the regulatory network underlying lineage 1 (Fig.4l), capturing the TFs and targets of tumour growth and metastasis, e.g. KLF7 [31] and ETS2 [32]. These characters were not observed in the other two lineages, indicating the spatial heterogeneity of tumour progress.

Tracing tumour metastasis

spaTrack provides the ability to trace cells across tissues of different time/conditions and therefore could reconstruct the trajectory of tumour metastasis. Tumour metastasis refers to the process by which cancer cells detach from the primary tumour and spread through the bloodstream or lymphatic system to colonize distant organs [33]. Understanding the origins and

colonizing process of tumour metastasis, provides important insights in developing effective strategies to target metastatic relapse and improve patient outcomes [34].

We collected a metastatic tumour (Fig.5a, Fig.S5a and S5b) from the lymph node corresponding with the primary ICC tumour (Fig. 4a) in the same patient. Malignant cells at the metastatic site were categorized into four major clusters (M0-M3) (Fig. 5b and 5c), forming a layered structure. M0 located at the core site of the tumour; M1 lied in the middle layer; M2 and M3 covered the outer layer. We applied spaTrack to optimally transport the malignant cells from the primary tumour to the metastatic tumour. M3 showed numerous successful transports from P0/P1, which was significantly higher than any other pairs of clusters (Fig. 5d). By plotting the successful transports between the primary tumour and the metastatic tumour (Fig. 5e), we observed malignant cells of the primary tumour (mainly from P0-P1) implant to the bottom axis of the metastatic tumour, belonging to subclone M3, from where the metastatic cells putatively expanded to form a new tumour. To further investigate the origins and colonization of metastatic cells, we inferred and compared the genetic variants (SNP) of the malignant cells from both tumours. P0-P1 shares numerous variants with M3, more than any other pairs of subclones after adjusting the population size (Fig.5f), confirming the metastatic connection inferred by spaTrack. Furthermore, integrating the ST data of the primary tumour and the metastatic tumour together using Harmony in the SC manner, M3 approximated with P0 and P1 in the UMAP embedding space (Fig.S5c), which was consistent with the results of spaTrack transports. We additionally examined the stemness and cell cycling of metastatic cells, both of which indicated M3 presenting the activation in proliferation and expansion in the metastatic site (Fig. 5g and 5h, Fig.S5d and S5e). Constructing the regulatory network between the primary tumour and the metastatic tumour, we observed HMGA1, ID2, and CEBPG as the key factors driving the colonizing dynamics, all of which play important roles in tumour progression and metastasis [35-37] (Fig.S5f).

We subsequently applied spaTrack to generate trajectories of the metastatic cells, assigning the successfully transported cells of M3 as starting cells (Fig.S5g and S5h). It appeared that M3 initiated the colonization along the bottom axis and subsequently progressed to M2 where they formed the outer layer. Both M3 and M2 expanded towards the middle layer (M1) and formed the core site (Fig.5i). Overall, spaTrack vividly described the dynamic process of tumour metastasis including origination, colonization, and expansion (see model in Fig.5j). This comprehensive analysis would certainly provide us valuable implications in understanding tumour metastasis.

Discussion

Development of biological systems always requires strict spatial-temporal organization. Spatial coordinates and experimental time might be leveraged as important constraints supplementing to transcriptomic profiles in the TI work. spaTrack, presented as an innovative algorithm that uses the mathematical concept of OT, generates accurate and informative spatial trajectories by incorporating both gene expression profiles and spatial-temporal information from ST data. According to the highlighted features of the algorithm (Fig. 1a-1c), spaTrack could (1) reconstruct fine local trajectory from ST data; (2) integrate spatial transition matrix of multiple samples to generate complete trajectories; (3) trace cell trajectory across temporal samples via direct OT mappings; (4) capture developmental driven factors by modelling a function of predicting gene profile at later time by TF expression at current time.

spaTrack has been undergone extensively testing on both ST data and SC data, comparing with current widely used TI methods. OT framework has the natural advantage of incorporating spatial distance into the cost measure of cell transition and therefore will capture local details and generate spatial disconnected trajectories. Firstly, we applied spaTrack, scVelo, Dynamo, and monocle3, which could directly generate trajectories in spatial coordinates, on the ST data of axolotl telencephalon regeneration (Fig. S6a), which is a comprehensively studied model. As we described before, spaTrack uncovered three spatial trajectory lineages in the tissue section of D15 (Fig. 2f), consistent with previous studies [21,22]. Importantly, trajectories of lineage two are newly generated in the wound area and discontinuous with the trajectory in lineage three of normal tissue. scVelo presented confusing trajectories in the region of lineage 1 with multiple starting spots. Trajectories of scVelo could not be adjusted by simply reversing the velocity direction. Dynamo performed well in each lineage, but showed continuous trajectories between regions of lineage two and three, which is temporally disconnected. Monocle3 showed a skeleton along the tissue shape, without single cell trajectories and correct directions (Fig. S6b).

Secondly, we applied spaTrack on a SC data of primary human hematopoietic stem and progenitor cells (HSPCs) [38] (Fig. S6c), resulting more accurate trajectories than other methods. Human hematopoiesis is a continuously hierarchical process and is comprehensively investigated by previous studies [39,40]. The development of HSPCs follows a tree-like structure with HSC as the root. spaTrack successfully reconstructed the trajectories that closely recapitulate the established knowledge of hematopoiesis (Fig. S6d). In comparison, scVelo

generated several nonsensical reversing trajectories, starting from intermediated cell types of active cell-cycling states (e.g., erythrocyte progenitor (pro-ery), B cell progenitors (proB), and granulocyte and monocyte progenitor (GMP)). Without metabolic labels, Dynamo also showed reverse streamlines in erythrocyte and myeloid trajectories. Monocle3 generated a proper skeleton of hematopoiesis but missing the single cell details.

At last, spaTrack requires feasible computing power and memory, making it a fast and effective option for TI study of ST data. Under a standard CPU thread (Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz), spaTrack requires only minutes to finish the computation of 5k – 400k cells (with 20,000 features). The memory load depends seriously on the population size, which follows an exponential growth with 6.9 GB for 5k cells. (Fig.S1a).

ACKNOWLEDGMENTS

This study was supported by funding from the National Key Research and Development Program of China (No.2022YFA1103400 and 2021YFC2501900), and the Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515110832). This work was also supported by the China National GeneBank (CNGB).

AUTHOR CONTRIBUTIONS

P.F.Q., H.M.Z., L.W., and X.X. designed the study; X.N.S., L.L.Z., Z.F.Y., H.M.Z., and P.F.Q. developed the methodology; Z.F.Y., L.L.Z., X.N.S., and K.H. developed the software and online tutorials; K.H., L.L.Z., Z.F.Y., and P.F.Q. performed the data analysis; X.X.Z., Q.C.Y., Z.Y.L., L.W., and X.Y.W. assisted with data collection and analysis; P.F.Q., X.N.S., K.H., L.L.Z., Z.F.Y. wrote the manuscript; and P.F.Q., H.M.Z., L.W., X.X., and X.J. supervised the study.

DECLARATION OF INTERESTS

All financial interests are unrelated to this study. The authors declare no conflict of interests.

Reference

1 Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories.

14, 979-982 (2017).

2 Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. **566**, 496-502 (2019).

3 Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. **20**, 1-9 (2019).

4 Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. **19**, 1-16 (2018).

5 Grün, D. *et al.* De novo prediction of stem cell identity using single-cell transcriptome data. **19**, 266-277 (2016).

6 Ji, Z. & Ji, H. J. N. a. r. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. **44**, e117-e117 (2016).

7 Farrell, J. A. *et al.* Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. **360**, eaar3131 (2018).

8 McInnes, L., Healy, J. & Melville, J. J. a. p. a. Umap: Uniform manifold approximation and projection for dimension reduction. (2018).

9 Haghverdi, L., Buettner, F. & Theis, F. J. J. B. Diffusion maps for high-dimensional single-cell analysis of differentiation data. **31**, 2989-2998 (2015).

10 Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. J. P. o. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. **9**, e98679 (2014).

11 Qiu, X. *et al.* Mapping transcriptomic vector fields of single cells. **185**, 690-711. e645 (2022).

12 Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. J. N. b. Generalizing RNA velocity to transient cell states through dynamical modeling. **38**, 1408-1414 (2020).

13 Chen, Z., King, W. C., Hwang, A., Gerstein, M. & Zhang, J. J. S. A. DeepVelo: Single-cell transcriptomic deep velocity field learning with neural ordinary differential equations. **8**, eabq3745 (2022).

14 La Manno, G. *et al.* RNA velocity of single cells. **560**, 494-498 (2018).

15 Villani, C. *Optimal transport: old and new*. Vol. 338 (Springer, 2009).

16 Cang, Z. *et al.* Screening cell–cell communication in spatial transcriptomics via collective optimal transport. **20**, 218-228 (2023).

17 Cang, Z. & Nie, Q. J. N. c. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. **11**, 2084 (2020).

18 Schiebinger, G. *et al.* Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. **176**, 928-943. e922 (2019).

19 Nitzan, M., Karaiskos, N., Friedman, N. & Rajewsky, N. J. N. Gene expression

cartography. **576**, 132-137 (2019).

20 Cuturi, M. J. A. i. n. i. p. s. Sinkhorn distances: Lightspeed computation of optimal transport. **26** (2013).

21 Wei, X. *et al.* Single-cell Stereo-seq reveals induced progenitor cells involved in axolotl brain regeneration. **377**, eabp9444 (2022).

22 Lust, K. *et al.* Single-cell analyses of axolotl telencephalon organization, neurogenesis, and regeneration. **377**, eabp9262 (2022).

23 Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. **185**, 1777-1792. e1721 (2022).

24 Evans, M. K. *et al.* Ybx1 fine-tunes PRC2 activities to control embryonic brain development. **11**, 4060 (2020).

25 Wang, Y., Lin, L., Lai, H., Parada, L. F. & Lei, L. J. D. D. Transcription factor Sox11 is essential for both embryonic and adult neurogenesis. **242**, 638-653 (2013).

26 Dagogo-Jack, I. & Shaw, A. T. J. N. r. C. o. Tumour heterogeneity and resistance to cancer therapies. **15**, 81-94 (2018).

27 Hausser, J. & Alon, U. J. N. R. C. Tumour heterogeneity and the evolutionary trade-offs of cancer. **20**, 247-257 (2020).

28 Wu, L. *et al.* An invasive zone in human liver cancer identified by Stereo-seq promotes hepatocyte–tumor cell crosstalk, local immunosuppression and tumor progression. 1-19 (2023).

29 Chen, Y. *et al.* Oncogenic collagen I homotrimers from cancer cells bind to α3β1 integrin and impact tumor microbiome and immunity to promote pancreatic cancer. **40**, 818-834. e819 (2022).

30 Winkler, J., Abisoye-Ogunniyan, A., Metcalf, K. J. & Werb, Z. J. N. c. Concepts of extracellular matrix remodelling in tumour progression and metastasis. **11**, 5120 (2020).

31 Gupta, R. *et al.* KLF7 promotes pancreatic cancer growth and metastasis by up-regulating ISG expression and maintaining Golgi complex integrity. **117**, 12341-12351 (2020).

32 Zhang, T., Liu, D., Wang, Y., Sun, M. & Xia, L. J. F. i. O. The E-Twenty-Six Family in Hepatocellular Carcinoma: Moving into the Spotlight. **10**, 620352 (2021).

33 de Visser, K. E. & Joyce, J. A. J. C. C. The evolving tumor microenvironment: From cancer initiation to metastatic outgrowth. **41**, 374-403 (2023).

34 Ganesh, K. & Massague, J. J. N. m. Targeting metastatic cancer. **27**, 34-44 (2021).

35 Sgubin, M. *et al.* HMGA1 positively regulates the microtubule-destabilizing protein stathmin promoting motility in TNBC cells and decreasing tumour sensitivity to

paclitaxel. **13**, 429 (2022).
36 Sikder, H. A., Devlin, M. K., Dunlap, S., Ryu, B. & Alani, R. M. J. C. c. Id proteins in cell growth and tumorigenesis. **3**, 525-530 (2003).
37 Huang, Y. *et al.* CEBPG promotes esophageal squamous cell carcinoma progression by enhancing PI3K-AKT signaling. **10**, 3328 (2020).
38 Qin, P. *et al.* Integrated decoding hematopoiesis and leukemogenesis using single-cell sequencing and its medical implication. **7**, 2 (2021).
39 Ranzoni, A. M. *et al.* Integrative single-cell RNA-seq and ATAC-seq analysis of human developmental hematopoiesis. **28**, 472-487. e477 (2021).
40 Buenrostro, J. D. *et al.* Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. **173**, 1535-1548. e1516 (2018).
41 Chizat, L., Peyré, G., Schmitzer, B. & Vialard, F.-X. J. J. o. F. A. Unbalanced optimal transport: Dynamic and Kantorovich formulations. **274**, 3090-3123 (2018).
42 Titouan, V., Courty, N., Tavenard, R. & Flamary, R. in *International Conference on Machine Learning.* 6275-6284 (PMLR).
43 Zappia, L., Phipson, B. & Oshlack, A. J. G. b. Splatter: simulation of single-cell RNA sequencing data. **18**, 174 (2017).
44 Li, H. *et al.* 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and samtools. **25**, 2078-2079 (2009).
45 Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. **28**, 1811-1817 (2012).

Data resources

The public ST data and SC data used in this study were collected as follows: ST data of Stereo-seq of axolotl telencephalon after injury was obtained from China National GeneBank DataBase (CNGBdb) with accession number CNP0002068. We selected four samples of 5, 10, 15, and 20 days after injury. ST data (Stereo-seq) of mouse midbrain development were collected from CNGBdb with accession number CNP0001543. Three ST samples of mouse embryo sections at 12.5 day, 14.5 day, and 16.5 day were downloaded and applied. ST data (Stereo-seq) of tumours were collected from CNGB with accession number CNP0002199, including one primary tumour of intrahepatic cholangiocarcinoma (ICC), and one corresponding metastatic tumour. One SC sample (10x genomics) of human hematopoietic stem and progenitor cells (HSPCs) were downloaded from the Genome Sequence Archive of

CNCB-NGDC (National Genomics Data Center of China National Center for Bioinformation), with accession number HRA000084.

Code availability

The open-source software is available at https://github.com/yzf072/spaTrack. The tutorial is deposited at https://spatrack-tutorials.readthedocs.io/en/latest/index.html.

Methods

Inferring cell trajectories from single ST data

Construction of cost matrix

To reduce computational burden, we perform Principal Component Analysis (PCA) to reduce the dimensionality of the data. Subsequently, we select the top 10 PCA components (defaulting to 30) for downstream analysis. To construct the cost matrix, we incorporate both gene expression profiles and physical distances. Differences in gene expression profile and cell coordinates are quantified using Euclidean distance. For cell i and cell j, the Euclidean distances of gene expression (g_ij) and physical distance (d_ij) are calculated as follows:

$$\begin{aligned} g_{ij} &= \sqrt{\left(x_{1j} - x_{1i}\right)^2 + \left(x_{2j} - x_{2i}\right)^2 + \cdots + \left(x_{nj} - x_{ni}\right)^2}, \\ d_{ij} &= \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \end{aligned} \quad (1)$$

Where n represents the number of principal components selected in the previous step of PCA.

To balance the contributions of the two distance measurements, we first normalize the distances. We then integrate the normalized gene expression distance g_{ij} and normalize spatial distance d_{ij} by scaling factors α_1 and α_2 to compute the cost matrix C_{ij} of cell transition. These factors control the relative importance of each distance measurement, with suggesting values for α_1 and α_2 are between 0 and 1. To prevent self-transitions, the cost matrix is re-defined. When $i = j$, we set the cost to the maximum of C_{ij} times 10□, so that the cost of a self-transition is maximized:

$$C_{ij} = \begin{cases} \alpha_1 \cdot \frac{g_{ij}}{\sum g_{ij}} + \alpha_2 \cdot \frac{d_{ij}}{\sum d_{ij}} & \text{for } i \neq j \\ \mathrm{Max}\left(C_{ij}\right) \times 10^7 & \text{for } i = j \end{cases} \quad (2)$$

Transition probability between cells

Adapting the concept of optimal transport (OT), we calculate the transition matrix by solving the following optimization problem:

$$\gamma = \arg\min_{\gamma} \langle \gamma, \mathbf{M} \rangle_F + \mathrm{reg} \cdot \Omega(\gamma)$$
$$\text{s.t. } \gamma \mathbf{1} = \mathbf{a} \quad (3)$$
$$\gamma^T \mathbf{1} = \mathbf{b}$$
$$\gamma \geq 0$$

Where **M** is the cost matrix calculated above, **Ω** is the entropic regularization term $\Omega(\gamma) = \sum i, j \quad \gamma_{i,j} \log(\gamma_{i,j})$, **a** and **b** are source and target weights (both sum to 1).

Cell order assignment

To utilize spaTrack for cell trajectory analysis, we specify the starting cells as ancestral cells. This can be achieved through various means including importing cell coordinates, cell type, or by manual selection using the interactive user interface to create lasso spots. Once the starting cells are defined, spaTrack will assign cell orders or directions relative to the starting cells for other cells in the dataset. This can be achieved by calculating the transferring probability of each cell to the starting cells. The probability can be determined by summing the transition probabilities of the cell to each of the starting cells.

Let $cell_1$, $cell_2$, … $cell_s$ as starting spots, for $cell_i$, the sum of transition probabilities is calculated as:

$$P_i = p_{1,i} + p_{2,i} + \cdots + p_{s,i} \quad (4)$$

Where the P_i is the probability of starting spots transferred to cell i. We assign a cell order to each cell by assuming that cells with a higher probability of transferring to their ancestors are closer to the ancestors in the trajectory. The probabilities of starting cells transferring to each cell were ranked ascending as:

$$r_i = R_{(P_i)} \quad (5)$$

Assuming the same interval, cell orders are normalized using the following formula:

$$R_{\mathrm{i}} = \frac{r_i - 1}{N} \quad (6)$$

Where N is the total cell number.

Cell velocity and organizing trajectory

Cell velocity is defined as the overall transition probability and direction from a cell to its neighbours. Before calculating cell velocity, the neighbouring cells are determined using their spatial coordinates and their cell PCA embedding matrix. Users are allowed to choose the number of neighbouring cells to consider. The spatial neighbours are identified using the K nearest neighbours (KNN) algorithm, while the PCA matrix neighbours are determined using Euclidean distance between cells. The final set of neighbouring cells is obtained by taking the intersection of the two sets of neighbours.

For each cell i, the n neighbours are selected. The velocity between cell i and cell j ($j \leq n$) is defined as following:

$$v_{ij} = \begin{cases} p_{i,j} & \text{for } R_i > R_j \\ -p_{i,j} & \text{for } R_i < R_j \end{cases} \quad (7)$$

Then the final velocity of cell i is calculated by averaging the velocities of cell i in its neighbourhood:

$$V_i = \frac{\sum v_{ij}}{n} \quad (8)$$

The trajectory is organized from the vector field of cell velocities, which is adapted from sctour [19]. Briefly, the optimal transition probability matrix is used as weights to calculate the unitary displacement vector for each cell. Only n KNNs of each cell are considered (n = [total spot number/50]):

$$\Delta u = \sum_{j \neq i} \left(p_{ij} - \frac{1}{n}\right) \frac{u_j - u_i}{\|u_j - u_i\|} \quad (9)$$

Where u_i and u_j were the coordinates of cell i and j.

Optimal path between two cells and pseudotime calculation.

To study the differentiation trajectory between two cells over space, we adapt the least action path (LAP) algorithm [11] to construct the optimal path between a starting spot and ending spot. Firstly, we construct a vector field of cell velocity from transition probability instead of RNA velocity, as described in *Formula 7* and *8*, which enables the estimation of cell velocity at any coordinate point. Secondly, given a starting cell and an ending cell, the initial path will be a line connecting the two points. The path is adjusted according to the cell velocity following

the LAP algorithm. We will get an optimal path that best fits the transition between the two cells. Afterwards, we need to map all the cells around the optimal path to assign cell orders and pseudotimes along the differentiation. We use the k-Nearest Neighbor (KNN) method to search cells spatially around the path. The neighbouring cells are vertically mapped to the optimal path, and the order of the cells is determined according to the mapped anchor point relative to the starting cell. Pseudotimes are defined as arc length between the mapped anchor point and the starting cell. Pseudotimes are normalized to a 0-1 range by dividing the total length of the path.

Tracing cells across multiple ST data with time-intervals

Unbalanced transport across multiple ST data

To compute the transport map between cells at time t_i and t_{i+1}, assuming that there are m cells at time t_i and n cells at time t_{i+1}, we solve the following optimization problem:

$$\begin{aligned}&\underset{\boldsymbol{\pi}\in\mathbb{R}_+^{n\times m}}{\operatorname{argmin}}[(1-\alpha)\langle\boldsymbol{\pi},\mathbf{M}\rangle_{\mathrm{F}}\\&+\alpha\textstyle\sum_{i,j,k,l}\boldsymbol{D}\big(\mathbf{d}_1(i,k),\mathbf{d}_2(j,l)\big)\boldsymbol{\pi}_{i,k}\boldsymbol{\pi}_{j,l}\\&+\lambda_1\mathrm{KL}(\boldsymbol{\pi}\mathbf{1}^m\mid p_1)+\lambda_2\mathrm{KL}(\boldsymbol{\pi}^T\mathbf{1}^n\mid p_2)]\end{aligned}\tag{10}$$

Where $\mathbf{M}\in\mathbb{R}^{n\times m}$ measures the gene expression dissimilarity between cells of two samples, and $<A,B>_{\mathrm{F}}$ denotes Frobenius inner product of matrices A and B, and $\mathbf{d}_1(i,k),\mathbf{d}_2(j,l)$ are the spatial distances between cells i, k and their corresponding cells j, l at different times respectively, and $\boldsymbol{D}$ measures the difference between scaled distances (Euclidean norm $\|\cdot\|^2$). In addition, λ_1 and λ_2 are regularization parameters and p_1 and p_2 are weight vectors of each cell. By default, $p_1=\frac{1}{m}\cdot\mathbf{1}^m, p_2=\frac{1}{n}\cdot\mathbf{1}^n$,where$\mathbf{1}^m,\mathbf{1}^n$denotes a column vector of length m,n containing all ones. The transport problem is solved with following considerations:

1. If cell i in time t_i is mapped to cell j in the next time t_{i+1} with a high weight $\boldsymbol{\pi}_{\mathrm{ij}}$, then the expression profile x_i of cell i is similar to the expression profile x_j of cell j.
2. If a pair of cells (i,k) in time t_i is mapped to a pair of cells (j,l) in the next time t_{i+1} with high weights $\boldsymbol{\pi}_{\mathrm{ij}}$ and $\boldsymbol{\pi}_{\mathrm{kl}}$, then the distance $\boldsymbol{d_1}(i,k)$ between cells i and k in the first time t_i is close to the distance $\boldsymbol{d_2}(j,l)$ between cells j and l in the next time t_{i+1}.

3. Unbalanced optimal transport [41], is with a more realistic approach to solving practical problems, for instance, it is suitable for scenarios where batch effects are present at different time points or when investigating the impact of the varying numbers of cells with value-added differentiation.

The sum of the first two terms in *Formula 10* represents a classic Fused Gromov-Wasserstein algorithm[42]. By introducing the last term, we extend the structured transport to handle unbalanced transport problems, where the equality constraints are relaxed to impose bounds on the marginals of the transport plan using of KL-divergence measure.

Computing trajectories of interest cells

At a given time point, a collection of starting cells can represent a specific cell type or any region of interest in space. Then the distribution of descendant cells at the next time point t_{i+1} can be calculated based on the transition matrix,

$$p_{t_i}(x) = \begin{cases} \frac{1}{|S|} & x \in S \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

In which S is the set of starting cells. The descendant distribution can be calculated as following

$$p_{t_{i+1}}^T = p_{t_i}^T \boldsymbol{\pi}_{t_i, t_{i+1}} \quad (12)$$

where $\boldsymbol{\pi}_{t_i, t_{i+1}}$ is the optimal transport map between t_i and t_{i+1} calculating from (10).

Learning gene regulatory models

OT has the capability to capture potential driven dynamics. We interpret the vector field as a model of gene regulation, which establishes functional relationships between the expression of transcription factors (TFs) at current time point and the expression of genes in a period of time. We propose to set up a regression model to learn the positive/negative regulation of genes by TFs. For ST data of two time points, we sample pairs of cells with expression (X_i, X_{i+1}) from the transport map and calculate gene changes Δ:

$$\Delta_i = X_{i+1} - X_i \quad (13)$$

Then, we extract TF expression from time $i+1$ and construct the following regression:

$$f: R^m \rightarrow R^t,$$

$$f(\Delta_i) = Y_{i+1},$$

$$\min_{W \in R^{t \times m}} \frac{1}{n} \sum_{i=1}^{n} \| Y_{i+1} - \Delta_i W^T \|^2 \quad (14)$$

where $Y_{i+1} = X_{i+1}T$, $X_i \in R^{1 \times m}$, $X_{i+1} \in R^{1 \times m}$ denotes the gene expression of pairwise mapped cells at two time points with m genes, f is the learned linear continuous function, representing the relationship between genes and TFs, $W \in R^{t \times m}$ denotes weight matrix with TFs and genes, $T \in m \times t$ stands for one-hot encoding matrix of TFs and genes. Here, we perform min-max normalization for gene changes and TF expression respectively and put data into the regression model to get weights.

For ST data of a single tissue section, we format the data to adapt the regression model. Cells are sorted according to their inferred pseudotime and are averagely grouped according to the setting bins. Then, cells of each pair of adjacent bins could be inputted to the model, which are processed in the same way.

To solve the problems of sparse data and reduce slow convergence, we used a meta-analysis method. We repeatedly select random cells and calculate the mean expression as new data, to improve data efficiency and increase sample size. To avoid the instability caused by random initialization of the model, we take the average of ten random training as the final result. Finally, top positive/negative correlation weight pairs are sorted from high to low and stored in data frame format. The regulatory network of TFs and genes can be displayed visually.

Simulations of SC and ST data.

We applied Splatter [43] to simulate single cells with temporal orders and spatial coordinates. Parameters to restrain the transcriptome of single cells were assigned as nGenes=3000, batchCells=1000, mean.shape=0.6, mean.rate=0.3, bcv.common=0.2, dropout.mid=0, dropout.shape=-1, out.prob=0.05, de.prob=1. We simulated 100 batches of SC data to demonstrate the connection between transition probability and temporal orders of cell trajectory, with additional parameters as method = "paths", path.from=c(0), path.length=c(100), path.skew=c(0.5), path.nonlinearProb=0.1.

Another 100 batches of ST data were simulated to show the connection between transition probability and spatial distance, with additional parameter group.prob = c(1). We assigned 1000 cells of each simulation to a 5000 μm × 5000 μm square, assuming each cell

taking up a 50 μm × 50 μm spot. For cell i and cell j, the Euclidean distances of spatial distance (d_{ij}) were calculated as:

$$d_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \quad (15)$$

Processing of raw data of ICC and its metastatic tumour

We downloaded the Stereo-seq GEM file of the primary tumour of ICC and its metastatic tumour from a previous study[22]. The GEM file includes the DNB coordinates and gene UMI counts in each DNB (220 nm). It was difficult to segment cells of tumour tissue and assemble the reads of single cells. We therefore merged 100 × 100 DNBs into a single informative 'bin' as a pseudo-cell (50 μm × 50 μm in square). To remove low-quality data, cells with expressed genes number < 500, expressed genes UMI <500 and a proportion of mitochondrial UMI > 20% were removed from downstream analysis. Finally, we obtained a total of 19908 and 28609 cells for the primary tumour and metastatic tumour respectively. The quality details of the data showed by violin and heatmap plots were presented in Fig. S4a, S4b, Fig.S5a, and S5b.

Cell type deconvolution of ST data and identification of malignant cells

ST data were deconvoluted using the seeded NMF method implemented in SPOTlight v0.17 [27]. SC data were used as references to infer the composition of each ST bin [22]. A default threshold of 0.08 was applied to filter the composition of cell type. The distribution of malignant cells was further examined by marker genes. We removed bins with high expression of marker genes relating with T cell, B cell, macrophage, fibroblast and endothelia cells. Finally, we obtained a total of 6,470 and 7,927 malignant cells of the primary tumour and metastatic tumour respectively. BayesSpace[28] was performed to cluster cells with spatial coordinate.

Identifying the genetic variants in ST data

In order to reliably detect single-cell expressed variants, we pooled all reads of tumour cells together to call variants (SNP). Tumour cells were determined by annotation of SPOTlight. Both Samtools [44](v1.16) and Strelka [45] are applied to do the variants calling. Successful callings from both methods were used for downstream analysis. Due to the sparsity of the ST data, we used following criteria of filtration to reduce the artifacts and false positives: variants

covered by at least 70 reads; reads with alternative variant take up >5% of all reads; variants observed in at least 3 tumour cells.

Shared SNPs between clusters of the primary tumour and the metastatic tumour were calculated. To yield more dependable comparisons, we performed 30 repetitions of counting of shared SNPs by random cell sampling from clusters. For each counting and comparison, cells of clusters were sampled to equal population size.

Identifying pseudotime-dependent genes

spaTrack applies generalized additive model to fit the dynamics of gene expression along a trajectory. For each gene, spaTrack fits the expression changes and the corresponding pseudotime value of cells using the generalized additive model in pyGAM package. The formula of the model is as:

$$g \sim s(t, k) \quad (16)$$

Where g represents the gene expression in cells; t denotes pseudotime value of all cells along a trajectory; The function k is a spline function used as a piecewise polynomial to fit smooth curves. P-values are adjusted for multiple testing using the BH method.

To determine whether the dynamics of gene expression across trajectory is decreasing or increasing, spaTrack calculates JS score between actual expression and standard downward/upward trends using following formula:

$$M = \frac{1}{2}(G + S) \quad (17)$$

$$JS(P||S) = \frac{1}{2}KL(G||M) + \frac{1}{2}KL(S||M) \quad (18)$$

Where, G represents predicted gene expression from the model, S represents a set of standard downward- or upward-trend values. KL is calculated by python SciPy packages.

Figure legends

Fig. 1 Frameworks of spaTrack.

a. Construct cell trajectories from single ST data. In brief, we scale the gene expression difference g_{ij} and spatial distance d_{ij} to construct the cost matrix of cell transition C_{ij}. The transition probabilities are estimated by solving the optimal transport problem; Cells are ranked according to their transition probabilities relative to the starting cells; A vector field of cell velocity is built to organize the optimal trajectories; The optimal path between a starting cell and an ending cell is constructed using the least action path method; To identifying pseudotime-dependent genes, we use a generalized additive model (GAM) to fit the dynamics of gene expression along a trajectory.

b. Trace cells across multiple samples of a time series. To compute the transport map between cells at time t_0 and t_1, we solve the unbalanced optimal transport (uOT) problem by adapting the Fused Gromov-Wasserstein (FGW) algorithm. $d_1(i,k)$, $d_2(j,l)$ are the spatial distances between any pair of cells i, k at t_0, and their corresponding cells j, l at t_1; D measures the scaled difference of d_1 and d_2.

c. Capture dynamic driven factors. A neural network framework is implanted in the algorithm, with expression profile of genes at t_1 time as input layer, prediction of TF expression of t_0 time as output layer. TF-gene pairs with high weights are screened to build regulatory network.

Fig.2 Fine local trajectories of axolotl telencephalon regeneration

a. Collections of the ST data of regenerative stages after injury of axolotl telencephalon at 5 days (D5), 10 days (D10), 15 days (D15), and 20 days (D20).

b. Spatial distribution of cell types around the regenerating site on the section of D15.

c. Heatmap of transition probability relative to starting cells.

d. Vector field of cell velocity.

e. Regenerative trajectories of D15 inferred by spaTrack.

f. Three major lineages of regenerative trajectories, and the optimal path of each lineage.

g. Spatial distribution of regeneration-related cell types in the ST data of D5, D10, D15, and D20.

h. Integration of transition matrixes from multiple ST samples.

i. The complete regenerative trajectories integrating from multiple samples.

Fig.3 Tracing neuron cells across mouse embryos of a time series

a. Collections of mouse embryo ST data and their dorsal midbrain regions, at day 12.5 (E12.5), 14.5 (E14.5), and 16.5 (E16.5)
b. Spatial distribution of radial glia cells (RGC), neuroblasts (NeuB), and glioblasts (GlioB) of dorsal midbrain.
c. Sankey diagram of the cell tracing of temporal sections. Blue segment represents RGC, brown for NeuB, red for GlioB.
d. Plots of the transition probability of RGC of E12.5 (left) to cells of E14.5 (middle) and E16.5 (right). Probabilities of successful OT cells are plotted.
e. The differentiation trajectories of RGC across samples of different time.
f. Regulatory network underlying the RGC differentiation over time.

Fig.4 Recovering the diverse trajectories of tumourigenesis

a. A ST sample of primary intrahepatic cholangiocarcinoma (ICC). H&E staining (middle), and distribution of malignant cells (right).
b. Spatial distribution of tumour subclones.
c. UMAP plot of tumour subclones.
d. Spatial expression of cancer stem cell markers. Expression of gene CD44, ID1, CDH1, and FOSL1 are summed up.
e. The cell cycling score of tumour subclones.
f. Trajectories of tumour expansion. The transition probability relative to starting cells (left top) and vector field of cell velocity (left bottom) are visualized.
g. Three lineages of tumour trajectories in space.
h. The optimal path of the lineage 1 (P0-P1-P2-P7).
i. Pseudotime-dependent genes of lineage 1, screened by fitting a generalized additive model.
j. Functional annotation of pseudotime-dependent genes of the three lineages.
k. Gene scores of ECM and EMT pathways. Gene score is calculated as the averaged expression of the genes in a pathway.
l. Regulatory network underlying tumour expansion of lineage 1.

Fig.5 Tracing tumour metastasis

a. A ST sample of metastatic tumour in lymph node, corresponding with the primary ICC in Fig. 4a. H&E staining (middle), and distribution of malignant cells (right)
b. Spatial distribution of tumour subclones in metastatic site.

c. UMAP plot of tumour subclones.

d. Counts of optimally transported cells from primary tumour to metastatic tumour. The successful transports are determined by their transition probability.

e. Plots of the successful transports between primary tumour and metastatic tumour.

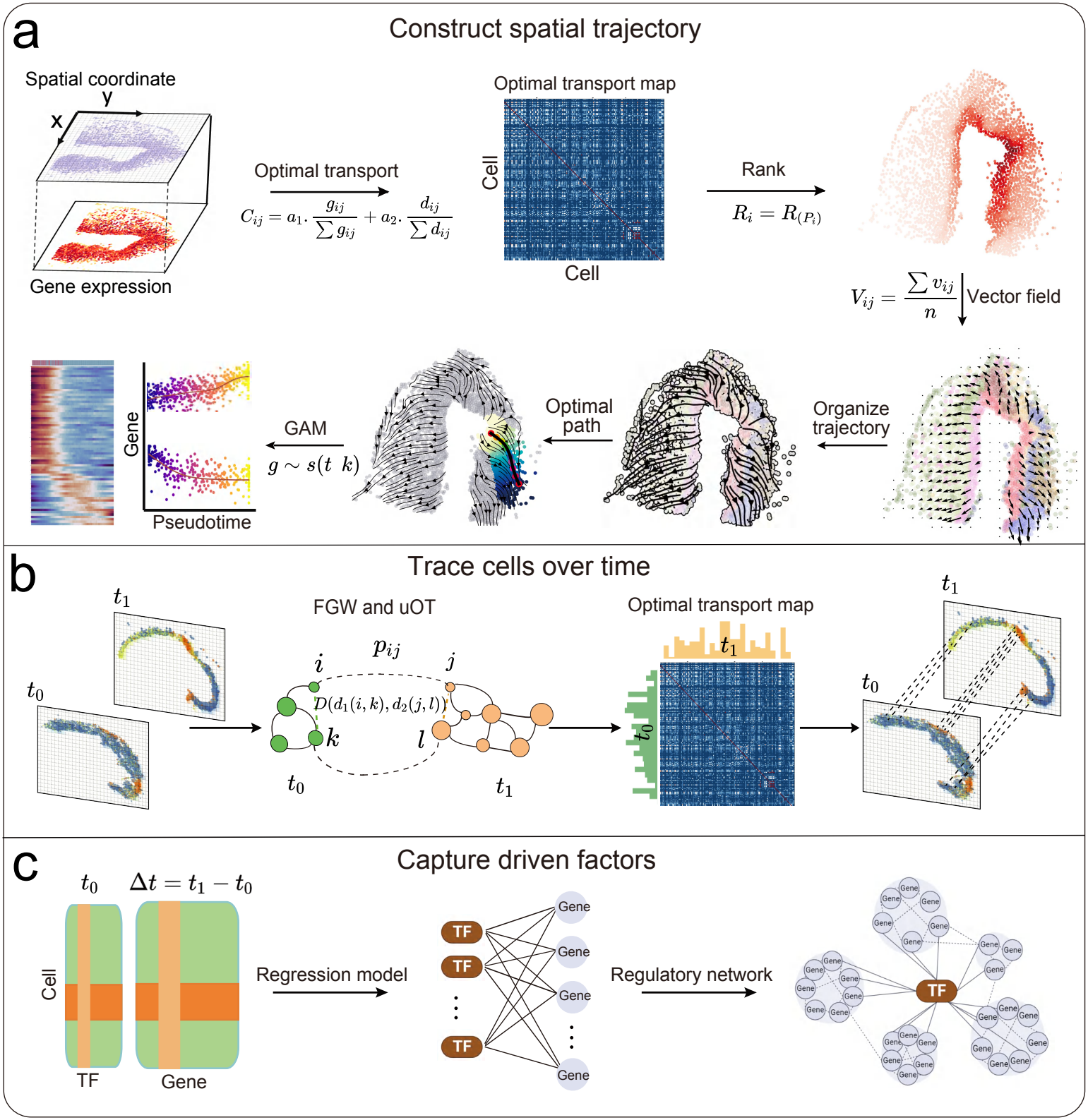
f. Expression variants (SNP) shared between primary tumour and metastatic tumour. The number of shared variants is adjusted by population size of clusters. The sharing numbers are significantly higher between P0-P1 and M3 than any other pair of clusters ($p<0.05$ in one sample t-test).

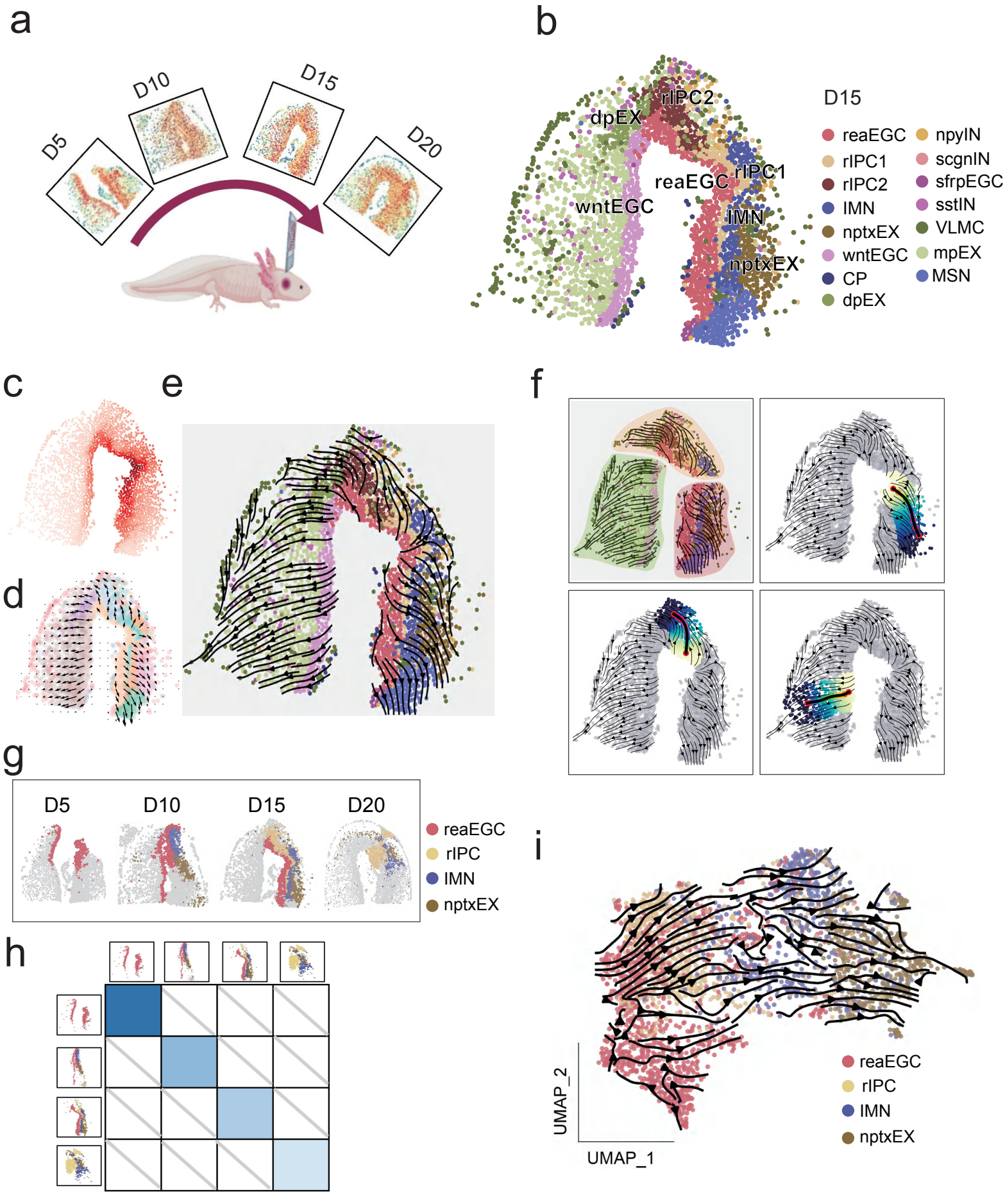
g. The expression of stemness markers in subclones of metastatic tumour.

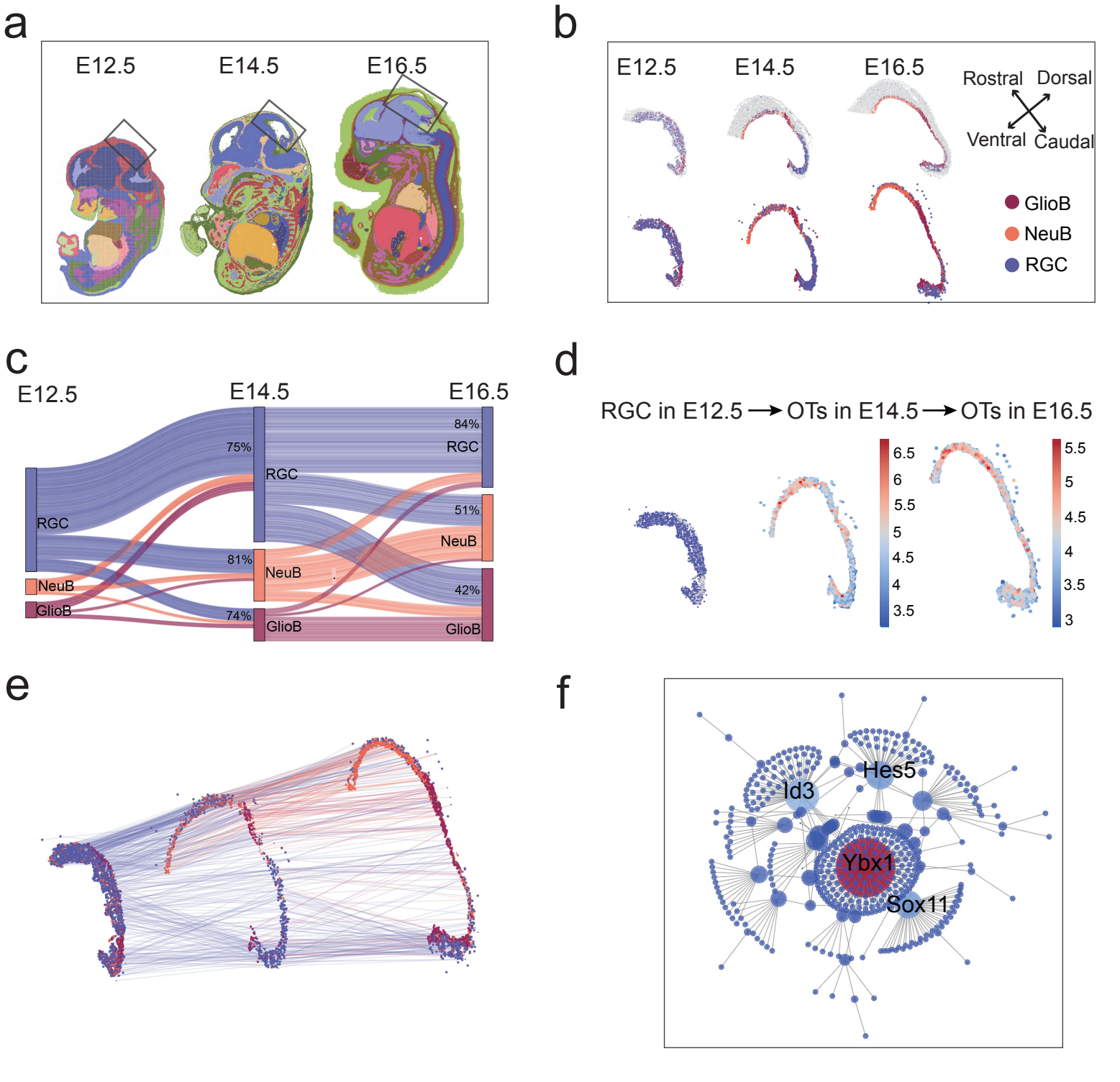
h. The G2M scores in subclones of metastatic tumour.

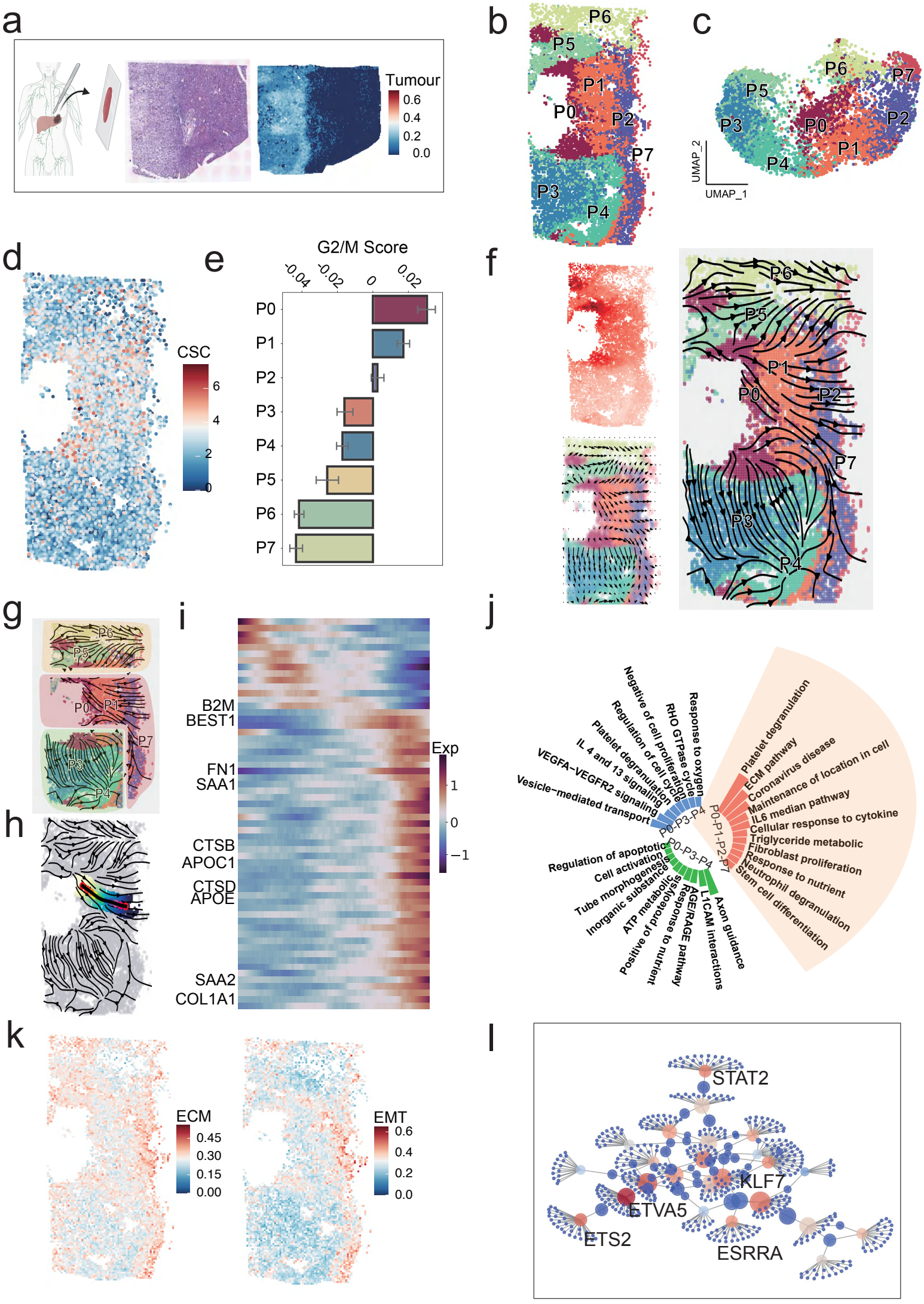
i. Trajectories of tumour colonization in the metastatic site. The transition probability relative to start cells (left top) and vector field of cell velocity (left bottom) are visualized.

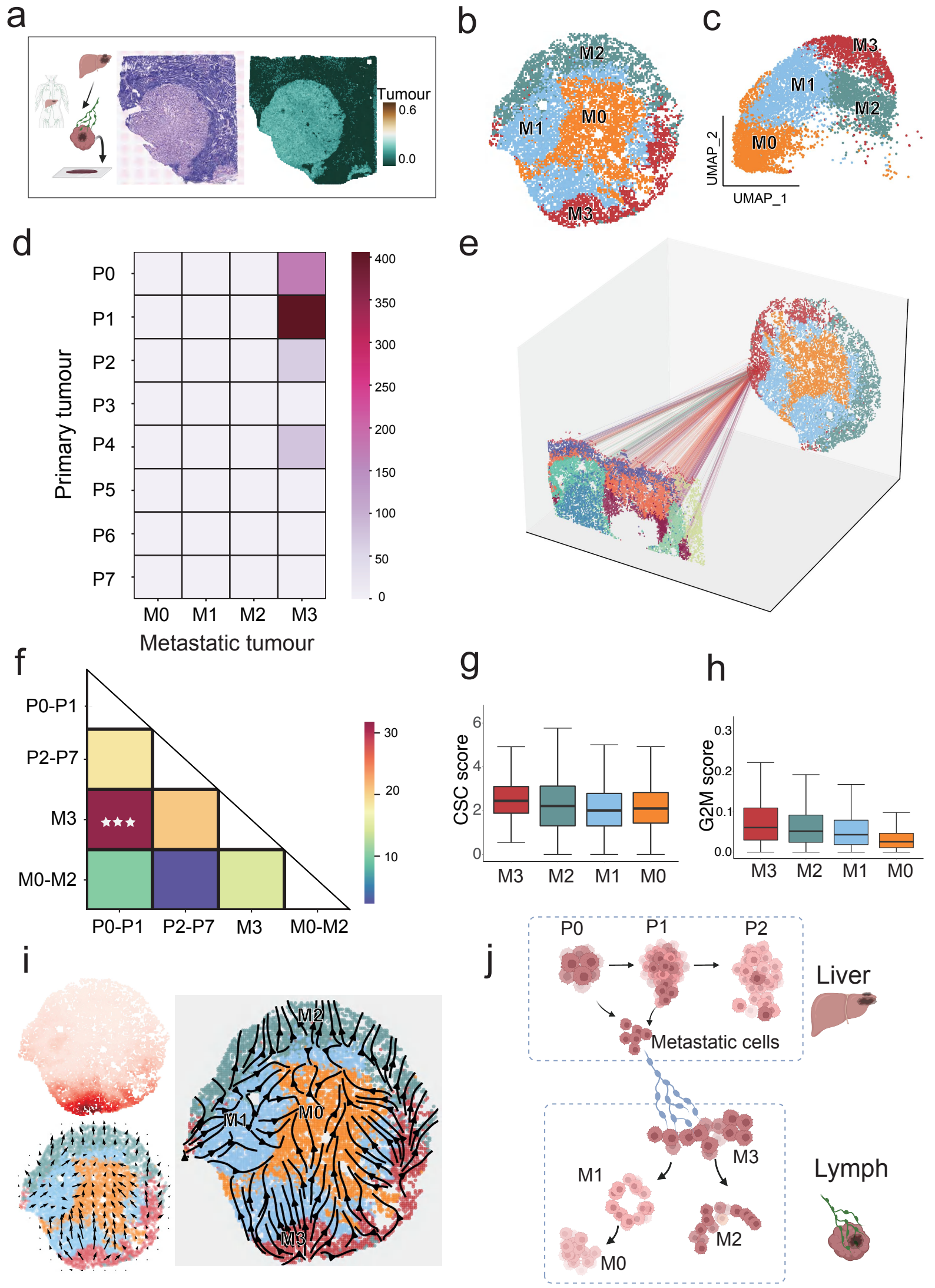
j. A model of the tumour metastasis revealed by spaTrack.


a
Construct spatial trajectory
Spatial coordinate
x
y
Gene expression
Optimal transport
$C_{ij} = a_1 . \frac{g_{ij}}{\sum g_{ij}} + a_2 . \frac{d_{ij}}{\sum d_{ij}}$
Optimal transport map
Cell
Cell
Rank
$R_i = R_{(P_i)}$
$V_{ij} = \frac{\sum v_{ij}}{n}$
Vector field
Organize trajectory
Optimal path
GAM
$g \sim s(t\ \ k)$
Gene
Pseudotime
b
Trace cells over time
t_1
t_0
FGW and uOT
p_{ij}
i
j
$D(d_1(i,k), d_2(j,l))$
k
l
t_0
t_1
Optimal transport map
t_1
t_0
t_1
t_0
c
Capture driven factors
t_0
$\Delta t = t_1 - t_0$
Cell
TF
Gene
Regression model
TF
TF
TF
Gene
Gene
Gene
Gene
Regulatory network
TF
Gene


a
D5
D10
D15
D20
b
D15
rIPC2
dpEX
reaEGC
rIPC1
wntEGC
IMN
nptxEX
reaEGC
rIPC1
rIPC2
IMN
nptxEX
wntEGC
CP
dpEX
npyIN
scgnIN
sfrpEGC
sstIN
VLMC
mpEX
MSN
c
d
e
f
g
D5
D10
D15
D20
reaEGC
rIPC
IMN
nptxEX
h
i
UMAP_2
UMAP_1
reaEGC
rIPC
IMN
nptxEX


a
E12.5
E14.5
E16.5
b
E12.5
E14.5
E16.5
Rostral
Dorsal
Ventral
Caudal
GlioB
NeuB
RGC
c
E12.5
E14.5
E16.5
RGC
NeuB
GlioB
75%
81%
74%
84%
51%
42%
d
RGC in E12.5 → OTs in E14.5 → OTs in E16.5
6.5
6
5.5
5
4.5
4
3.5
5.5
5
4.5
4
3.5
3
e
f
Hes5
Id3
Ybx1
Sox11


a
Tumour
0.6
0.4
0.2
0.0
b
P6
P5
P1
P0
P2
P7
P3
P4
c
P6
P7
P5
P3
P0
P2
P1
P4
UMAP_2
UMAP_1
d
CSC
6
4
2
0
e
G2/M Score
-0.04
-0.02
0
0.02
P0
P1
P2
P3
P4
P5
P6
P7
f
P6
P5
P1
P0
P2
P7
P3
P4
g
P6
P5
P0
P1
P7
P3
P4
h
i
B2M
BEST1
FN1
SAA1
CTSB
APOC1
CTSD
APOE
SAA2
COL1A1
Exp
1
0
−1
j
Platelet degranulation
ECM pathway
Coronavirus disease
Maintenance of location in cell
IL6 median pathway
Cellular response to cytokine
Triglyceride metabolic
Fibroblast proliferation
Response to nutrient
Neutrophil degranulation
Stem cell differentiation
P0-P1-P2-P7
Axon guidance
L1CAM interactions
RAGE/RAGE pathway
Response to nutrient
Positive of proteolysis
ATP metabolic
Inorganic substance
Tube morphogenesis
Cell activation
Regulation of apoptotic
P0-P3-P4
P0-P3-P4
Vesicle-mediated transport
VEGFA-VEGFR2 signaling
IL 4 and 13 signaling
Platelet degranulation
Regulation of cell cycle
Regulation of cell proliferation
RHO GTPase cycle
Response to oxygen
k
ECM
0.45
0.30
0.15
0.00
EMT
0.6
0.4
0.2
0.0
l
STAT2
KLF7
ETVA5
ETS2
ESRRA


a
Tumour
0.6
0.0
b
M2
M0
M1
M3
c
M3
M1
M2
M0
UMAP_2
UMAP_1
d
P0
P1
P2
P3
P4
P5
P6
P7
Primary tumour
M0
M1
M2
M3
Metastatic tumour
400
350
300
250
200
150
100
50
0
e
f
P0-P1
P2-P7
M3
M0-M2
***
30
20
10
P0-P1
P2-P7
M3
M0-M2
g
CSC score
6
4
2
0
M3
M2
M1
M0
h
G2M score
0.3
0.2
0.1
0.0
M3
M2
M1
M0
i
M2
M1
M0
M3
j
P0
P1
P2
Liver
Metastatic cells
M3
M1
Lymph
M2
M0