Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

De Novo Spatial Reconstruction of Single Cells by Developmental Coalescent Embedding of Transcriptomic Networks

Yuxuan Zhao[1*], Shiqiang Zhang[2*], Carlo Vittorio Cannistraci[3,4#] and Jing-Dong J. Han[1,2#]

[1] Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Center for Quantitative Biology (CQB), Peking University, Beijing 100871, China

[2] CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences Center for Excellence in Molecular Cell Science, Collaborative Innovation Center for Genetics and Developmental Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai, 200031, China

[3] Center for Complex Network Intelligence (CCNI) at Tsinghua Laboratory of Brain and Intelligence (THBI), Department of Bioengineering, Tsinghua University, 60 Chengfu Road, Beijing, 100084, China

[4] Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Cluster of Excellence Physics of Live (PoL), Department of Physics, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany

[*]These authors contributed equally to this work
[#]Correspondence and requests for materials should be addressed to C.V.C (Email: kalokagathos.agon@gmail.com) and J.-D.J.H. (Email: jackie.han@pku.edu.cn)

Abstract

Single cell RNA-seq (scRNA-seq) profiles conceal temporal and spatial tissue developmental information. *De novo* reconstruction of single cell temporal trajectory has been fairly addressed, but reverse engineering single cell 3D spatial tissue localization is hitherto landmark based, and *de novo* spatial reconstruction is a compelling computational open problem. Here we show that a new algorithm - named D-CE - for coalescent embedding of single cell transcriptomic networks can address this open problem. We rely merely on the spatial information encoded in the expression patterns of developmental signal transcription factor (DST) genes, and we find that D-CE of cell-cell association DST-transcriptomic networks reliably reconstructs the Geo-seq or single cell samples' 3D spatial tissue distribution. Comparison to the novoSpaRC and CSOmap (recent and only available *de novo* 3D spatial reconstruction methods) on 16 datasets and 681 reconstructions, reveals a significantly distinctive superior performance of D-CE.

Main Text

Developmental events are orchestrated temporally and spatially by cell-cell interactions and molecular changes during pluripotency exit and cell fate determination lead to branching canalization of development lineages as depicted in Waddington's landscape[1]. Cell identity transition is precisely controlled and ordered, this implies that individual single cells are genetically fingerprinted and genomically programmed to evolve towards a 3D spatial tissue continuum. Single cell technologies - such as single cell RNA-seq (scRNA-seq) that simultaneously profile thousands and more single cells - have becoming powerful tools to capture such continuous

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

spatiotemporal changes during development[2]. Based on single cell profiles, the transition paths to the differentiated cells (or the developmental time trajectories) can be reconstructed by calculating transcriptomic similarities or dissimilarities between single cells. Various computational tools based on this assumption have been established to model the developmental time trajectories[3]. For instance, Monocle reduces the data dimensionality and uses the minimum spanning tree to model the developmental paths[4]. Diffusion pseudotime, which is based on diffusion-like random walk distances, is used to map developmental branching decisions[5]. As spatial information significantly contributes to the cellular developmental states, transcriptomic-based computational reverse engineering of 3D tissue distribution or 'pseudospace' could be potentially achieved[6]. A benchmark cell population timer/clock has been proposed to test the performance of pseudotime algorithms to approximate the real developmental time of each single cell[7]. Similarly a benchmark microdissection-based 3D transcriptome, termed Geo-seq, can be used to test the performance of algorithms to approximately map single cells onto *in vivo* positions[8]. Some landmark-based computational approaches have been proposed to reconstruct spatial distribution of single cell transcriptomes in zebra fish embryos and mouse liver based on preselected or verified spatially expressed landmark genes that are tied to specific lineage structures[9,10]. But, in truth, these approaches cannot be considered *de novo* reconstruction methods because the landmark genes are *ad hoc* expressed in certain specific 3D spatial positions of the considered *in vivo* tissue, or surrogate labels for the positions that are revealed and input to the algorithms together with the positions they mark. So far, effective, universally (tissue-wide) applicable, completely *de novo* approaches have yet to be developed, and only recently the first template structure constrained approach named novoSpaRC was proposed by Nitzan et al. [11] with encouraging yet improvable results. Here, to address the *de novo* reconstruction of single cell 3D spatial tissue ordering and localization, we design a novel algorithm according to Coalescent Embedding (CE), which is a model-free unsupervised machine intelligence methodology for network geometry embedding[12]. Mapping networks to their underlying geometric spaces helps to understand structure/function interplay in complex networked systems[12]. CE encloses under its name a class of machine intelligence algorithms for efficient embedding of large real networks to the latent geometric space, which have been proven to impact hyperbolic big-network-data analysis in biology, neuroscience and social science[12]. For instance, CE showed to boost the detection of community structural organization in social networks[12] and to reliably capture the original geometry of macroscale structural brain connectomes[13]. The name coalescent embedding derives from "angular coalescence", which is a term proposed to indicate that, as a result of this methodology of embedding, the individual network nodes geometrically aggregate together (from the Latin verb coalēscō: to join, merge, amalgamate single elements into a single mass or pattern) forming a pattern that is progressively ordered along the geometrical angular coordinates[12]. In CE algorithms, the node angular coordinates are ordered according to latent relations of topological homophily (similarity) between the network nodes[12], instead the node radial coordinates according to latent relations of topological hierarchy between the network nodes[12] (Fig. 1a). Our hypothesis is that, according to CE rationale, the embedding of a developmental network of transcriptomic topological similarity between cells (an association network derived from their gene expression) should produce an angular coalescent cell ordering that recapitulates the original single cell samples' 3D spatial tissue distribution.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

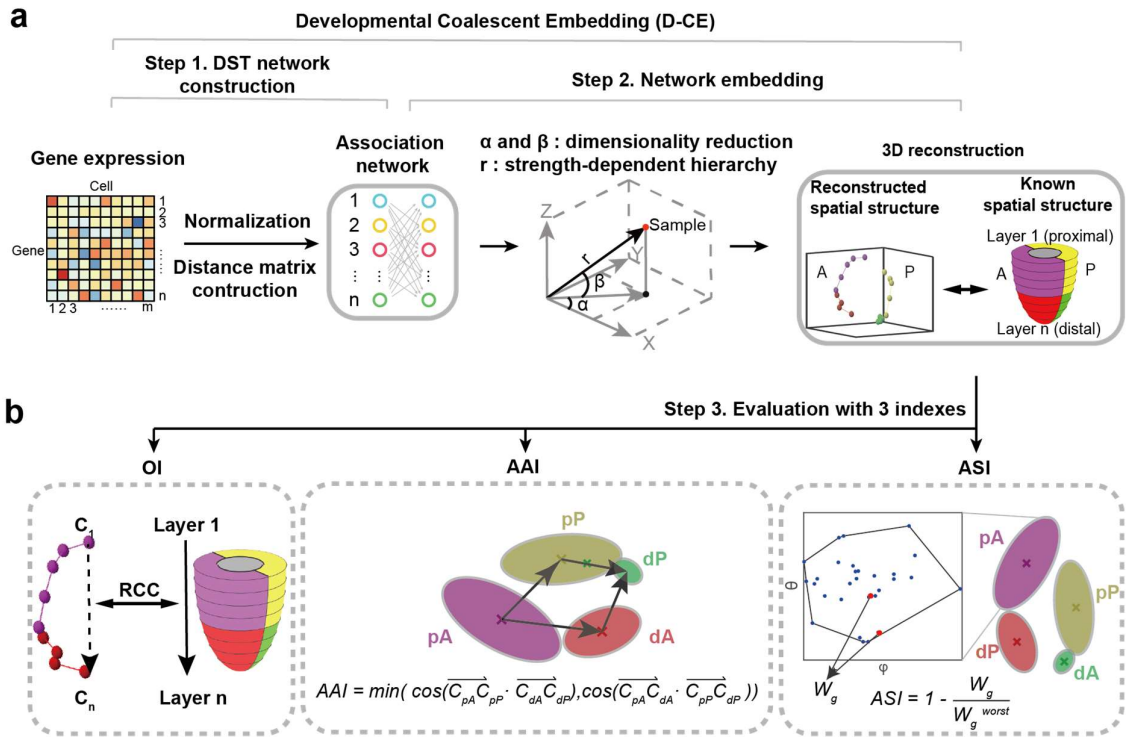

Figure 1. Method of spatial reconstruction of 3D localizations. **a,** Overview of spatial structure reconstruction by D-CE, which consists of angular and radial reconstruction. Angular reconstruction is based on dimensionality reduction by singular value decomposition (SVD), radial reconstruction is based on the strength-dependent hierarchy. Different normalization methods and network construction methods were applied to different gene set to construct an association network for D-CE. **b,** Then angular separation index (ASI), angular alignment index (AAI) and ordering index (OI) to the original sample layer order were used to assess the accuracy of spatial reconstruction of Geo-seq samples' positions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

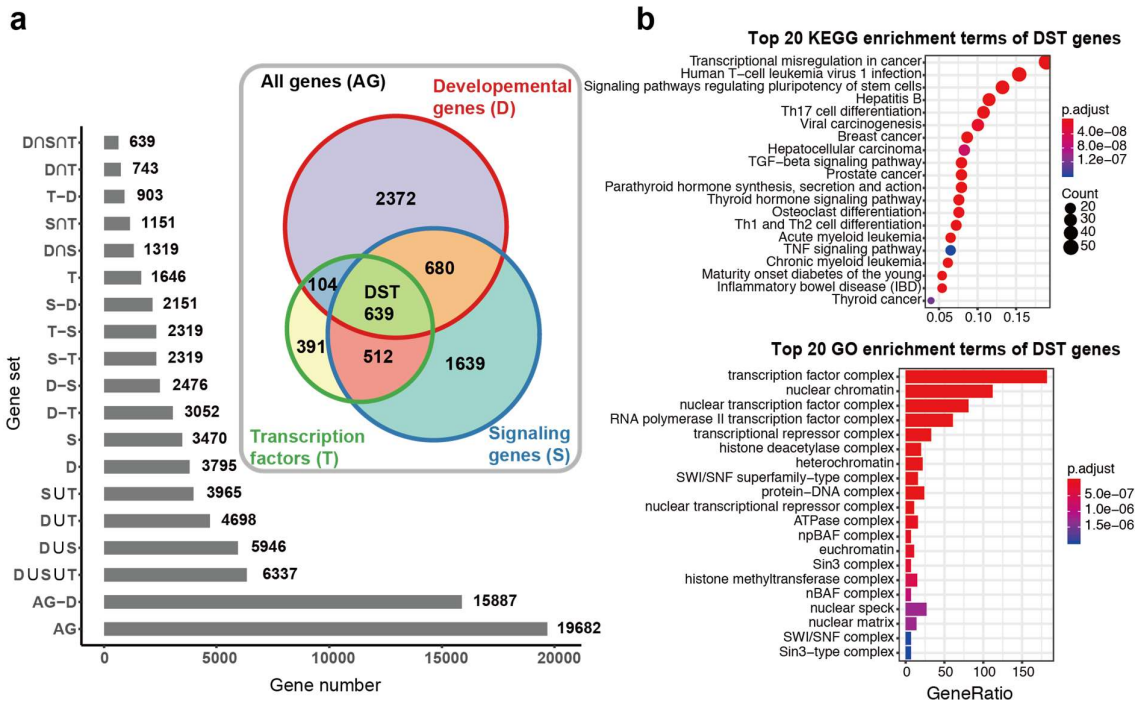

Figure 2. Knowledge based gene set selection. **a,** Various combinations of 3 different gene sets tested for Geo-seq data spatial reconstruction. These include 3795 developmental genes, 3470 signaling genes and 1646 transcription factors. Then the union, intersection and each unique part from these 3 gene sets were derived into a total of 19 gene meta-sets from the non-overlapping 7 sections in the Venn diagram. We tested all 19 gene sets for their performance in retrieving the spatial information of the samples. **b**, Overview of spatial structure reconstruction by D-CE. Different normalization methods and network construction methods were applied to each gene meta-set in panel (a) to construct an association network for D-CE. Then angular separation index (ASI), angular alignment index (AAI) and ordering index (OI) to the original sample layer order were used to assess the accuracy of spatial reconstruction of Geo-seq samples' positions. **c,** KEGG and GO terms enriched in the 639 DST genes over the union of input genes as shown in panel a.

Hence, in this study: i) we first develop a method to generate an association network of single cell samples starting only from their developmental signal transcription factor (DST) gene expressions (Fig. 1a and 2); ii) and then we embed this network in a 3D space. The union of these two steps represent a novel algorithm for *de novo* reconstruction of single cell 3D spatial tissue localization that we name "Developmental Coalescent Embedding" (D-CE, see methods for details, Fig. 1a).

To arrive at this final design of D-CE, we examined different possible strategies to implement D-CE. The critical point that required a thorough investigation was the appropriate way to build the association single cell network starting from their gene expressions. To facilitate our investigation, we reduced the gene expression list to only three sets: developmental genes, signaling genes and transcription factor genes; because previous studies have emphasized that those three gene sets often display spatial distribution patterns during early embryonic development[8,14]. However, we care to stress that this is fundamentally and conceptually different from the landmark genes used by previous methods. Landmark genes are preferentially expressed in a certain specific 3D spatial position of the original unfragmented tissue and are input together with their position to the algorithms as surrogate position labels. Whereas, those three gene classes – considered in our study - are associated to tissue development but are

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

not necessarily specifically expressed in any particular 3D spatial position of a certain type of tissue and most importantly the spatial positions they are associated with are not known to our algorithm. They cannot be adopted to offer orientation to the reconstruction, and therefore cannot be adopted for knowledge-driven reconstruction, which indeed is the typical strategy implemented by landmark methods.

In order to appraise the extent to which the three different gene sets would affect the performance of D-CE, we considered Geo-seq samples[14] whose 3D geometric coordinates are known and can be used as gold standards to evaluate the reliability of a *de novo* 3D reconstruction algorithm. Each Geo-seq sample is not a single cell but a portion of tissue constituted by a cohort of 10~20 single cells, however this is an ideal dataset to design and to test D-CE performance. In particular, we examined the impact of various combinations of 3 different gene sets, developmental (D), signaling (S) and transcription factor (TF, T) genes, to retrieve by D-CE the spatial information of the Geo-seq samples. We derived 19 gene meta-sets through union, intersection and each unique part between and among these 3 original gene sets (Fig. 2a, Methods). In addition, for each of the 19 gene meta-sets, we considered the sample-sample transcriptome distance matrix (weighted network) using each of 12 normalization methods (Supplementary Table 1) and each of 5 distance measures (Fig. 1a), resulting in 1140 (19x12x5) possible candidate association networks to test, and embedded each of them separately using the proposed D-CE algorithm (Fig. 1a, Methods). Specifically, we considered 5 distance measures, including: Spearman distance (SD) (1-Spearman rank coefficient (RCC)), Pearson distance (PD) (1-Pearson correlation coefficient (PCC)), Euclidean distance (ED), and PCC and RCC filtered by connectivity specificity index (CSI)[15], named PCC-CSI and RCC-CSI (Fig. 1a and Supplementary Fig. 1, Methods).

We tested each of these 1140 D-CE candidate strategies to spatially reconstruct the Geo-seq data of different germ layers in mouse early embryo development gastrulation stage (E6.5, E7.0 and E7.5). To evaluate the accuracy of the reconstruction, we grouped the Geo-seq samples into four groups: 1-2) proximal anterior and distal anterior (pA and dA); 3-4) proximal posterior and distal posterior (pP and dP). They can be further specified for each germ layer at each stage as: pA, dA, pP and aP of ectoderm (ect) and endoderm (en) at E6.5 and E7.0; pA, dA, pP and aP of ect, en and mesoderm (m) at E7.5 (Fig. 3a). Angular separation index[12] (ASI, ranging from 0 to 1 with 1 being perfect separation, see Methods for details) is used to test how well the four groups are separated according to angular coordinates in the embedding space. Angular alignment index (AAI) measures the parallelness between to vectors, ranging from -1 to 1, with 1 being perfect alignment and -1 being anti-parallel alignment (see Methods for details), and is used to test how well a relative spatial orientation is kept in among orthogonally distributed samples (e.g. proximal-distal orientation in anterior and posterior samples) in the embedding space. Ordering Index (OI, ranging from -1 to 1 with 1 being perfect agreement between the original order and the reconstructed order) to evaluate the accuracy of ordering the layers from distal to proximal (e.g. layer 1 to 11 in E7.0 embryo, Fig. 1b and Methods). Finally, to evaluate with one unique value for each of the 1140 embedding solutions, we consider the maximum rank of these three indices for each embedding (Methods), this means that a method that ranks 1 for each of these three indices will be the perfect candidate to be selected, because it has the lowest maximum ranking across the three possible indices. Indeed, the lowest maximum rank will indicate the best reconstruction of both

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

the anterior and posterior localization and the order of layers (Supplementary Fig. 1), and the best strategy to perform D-CE turns out to be based on the intersection of developmental genes, signaling genes and transcription factor genes (DST genes for short) for a total of 639 genes. This selected D-CE offers a *de novo* reconstruction of the spatial relationships of the Geo-seq and singe cell samples which is reported in Fig. 3, Supplementary Fig. 2 and 3. KEGG and GO enrichment analysis shows that DST gene set is highly enriched for pluripotency regulators and chromatin remodeling complexes which were known to play important roles in mouse embryo development (Fig. 2b, Supplementary Fig. 4). We found that D-CE based on DST gene set, with most normalization methods and PD, showed a very high correspondence to the original geometric locations of the samples in the mouse embryo from where the Geo-seq samples were derived (Fig. 3, Supplementary Fig. 2 and 3). The anterior and posterior samples of different germ layers in different stages are well separated into the opposing directions of the 3D space, similar to its original distribution in the developing mouse embryo, so are the proximal and distal samples, as evidenced by the significant ASI separating the pA, dA, pP and dP samples (ASI≥0.86 with ASI=1.00 for E6.5) with all germ layers combined together at stage E6.5, E7.0 and E7.5 (Fig. 3, Supplementary Fig. 2), or in each germ layer separately (ASI≥0.84 with ASI=1.00 for E6.5 ect, E7.0 en, E7.5 en and E7.5 m) (Supplementary Fig. 3) and the correct relative orientation of A to P and p to d (AAI are all above 0.50) (Fig. 3a, b and c, Supplementary Fig. 2 and 3). Remarkably, the distal to proximal sequence (1~11 horizontal layers) of the Geo-seq samples are also accurately captured by the embedded samples in the reconstructed structure (Fig. 3a, b and d, Supplementary Fig. 2 and 3, minimal OI for A, P and also E7.5 left or right (L/R) samples are all above 0.50). The E7.5 L/R samples are also correctly placed at the back of A and P samples (Fig. 3a, b and c, Supplementary Fig. 2 and 3). That the left and right samples of the same layers are aggregated together (Fig. 3, Supplementary Fig. 2 and 3) is expected because they are highly symmetric and have no expression differences as previously observed[2]. It should be noted that hierarchical clustering of the input PD matrix does not automatically separate the spatial domains (Supplementary Fig. 5). We also verified that by using: i) the first, second or third principle component correlated genes (PC1, 2 or 3) singularly or ii) the PC1, 2 and 3 correlated genes all together or iii) Scialdone *et al*'s pseudospace genes[6], satisfactory 3D reconstructions was not achieved (Supplementary Fig. 6). Their results were much inferior compared to the *de novo* reconstruction of D-CE based on DST genes (Supplementary Fig. 6). The successful 3D reconstruction of the Geo-seq samples confirm that the spatial information encoded in DST genes can be properly preserved by D-CE.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

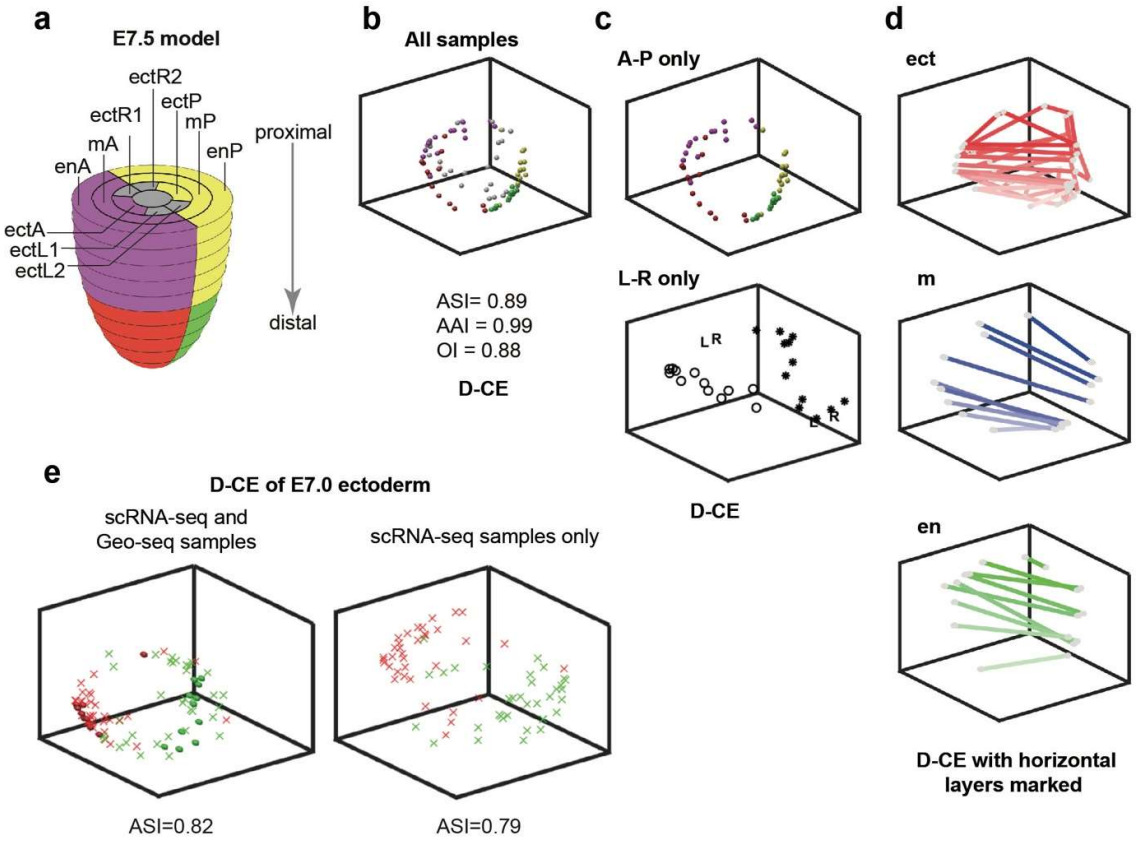

Figure 3. Spatial reconstruction of 3D localizations of Geo-seq samples from gastrulating mouse embryo. **a**, Illustration of the original locations of Geo-seq samples from the three germ layers in E7.5 mouse embryo. Sample positions are colored by spatial positions (purple, red, yellow and green for proximal anterior (pA), distal anterior (dA), proximal posterior (pP) and distal posterior (dP)). en, ect and m stand for endoderm, ectoderm and mesoderm. **b,** D-CE reconstructed 3D localizations of mouse Geo-seq samples from stage E7.5 all samples of all three germ layers. Samples' original positions are color coded as in the model in panel a. **c,** Upper panel, the same D-CE reconstruction of the E7.5 embryo samples as in panel b displaying only the A and P samples to visualize the clear distinction of proximal/distal anterior/posterior (pA, dA, pP, dP) samples. This is also reflected by the good separation and relative alignment of samples as measured by ASI and AAI. Lower panel, the same D-CE reconstruction of the E7.5 embryo samples as in panel b displaying only the L and R ectoderm samples (gray balls in panel b), which are highly symmetric and distributed to the back of the A and P samples. The L/R samples close to A and P (ectLa/ectRa and ectLp/ectRp) are labeled by open circles and stars, respectively, to visualize their closeness to anterior and posterior in the embedded structure. The most proximal and most distal L/R samples having no anterior or posterior bias are labeled as "L" or "R". **d,** The same D-CE reconstruction of the E7.5 embryo samples as in panel b displaying only ectoderm, mesoderm and endoderm samples in the upper, middle and lower panels, respectively. Samples in the same horizontally micro-dissected layer are connected by lines, which are color coded from dark to light in the proximal to distal order to visualize the recapitulation of the proximal to distal order in the samples. For each layer, samples are first linked into 3 units (A-P, ectLa-ectRa, ectLp-ectRp), then the pair of samples in different units with shortest distance are connected. This is also reflected by the high correlations to the original sample orders in the embryo as measured by OI. **e,** D-CE reconstruction of single cells from mouse embryo E7.0 ectoderm based on their scRNA-seq (marked by 'x') together with Geo-seq samples (balls) (left panel) or D-CE reconstruction using the scRNA-seq data alone (right panel). Samples' original positions are colored red for anterior and green for posterior. In the left panel, single cell samples derived from A and P regions are placed close to A and P Geo-seq samples, respectively. In the right panel, reconstruction with scRNA-seq alone maintained the similar spatial locations for these samples as in the left panel.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

Next we investigated whether D-CE based on DST genes is also applicable to single cell data. To investigate this problem, we embedded the network of Geo-seq and scRNA-seq samples from the same stage together based on the DST gene expression similarities among the samples. As expected, the posterior single cell samples and posterior Geo-seq samples are mapped close by on the sphere, and well separated from the anterior single cell and Geo-seq samples (ASI = 0.82 between anterior and posterior) (Fig. 3e left panel), suggesting the *de novo* reconstruction is effective at various resolution or granularity. Single cell embedding alone also well separated the known anterior and posterior single cells (ASI = 0.79) (Fig. 3e right panel). As most of the scRNA-seq data we obtained lack spatial labels or only have a label of A or P, we could only verify the correct reconstruction of all the A and P labels (Fig. 3e). Embedding of single cells alone largely retained the relative order (by shortest Euclidean distance) of the cells as in the single cell plus Geo-seq embedding (RCC=0.49), indicating that D-CE of single cells alone without any reference can *de novo* reconstruct the 3D position and spatial distribution of the single cells.

Using ASI, AAI and OI as performance measures, we compared D-CE versus 5 state-of-the art dimensionality reduction methods (PCA, tSNE, UMAP-corr, UMAP-cos, UMAP-euc) on the Geo-seq data. D-CE performs better than all these popular dimensionality reduction methods in 3D spatial reconstruction (Supplementary Fig. 7).

A single cell 3D position template-constrained spatial reconstruction method novoSpaRC[11] on the Drosophila embryo 3D gene expression dataset BDTNP[16] was recently developed. To compare our spatial reconstruction with novoSpaRC, we used the same 3D gene expression dataset BDTNP[16] to test spatial reconstruction accuracy. As only the expression data of 84 Drosophila embryo development regulating TFs (41 overlap with DST gene homologs in fly) are measured, D-CE of all these genes (and not exclusively DST genes as before) were used for network reconstruction with the normalization method optimized for the Geo-seq data. Here we found that, instead of PD (as before), its local-threshold-variation named PCC-CSI (which is a distance obtained by local threshold of Pearson correlation using CSI) performs the best for spatial reconstruction among all 5 distance options (Supplementary Fig. 8a, b and c). This difference in comparison to the Geo-seq data might be attributed to the large number of nodes in this network (3039 in BDTNP versus <100 in Geo-seq datasets). A gradient down-sampling of the BDTNP to data indeed shows that PD performs better than PCC-CSI when the sample number is <150, beyond that PCC-CSI performs better as shown by ASI and AAI (Supplementary Fig. 8d and e). OI is not evaluated because in the down-sampling, samples are randomly selected and hence rarely from a straight line along the x, y or z axis, and this disrupts the original distribution of the samples on 3 axes, making the OI inapplicable to the random sampled networks for evaluating the spatial reconstruction of down-sampled samples.

D-CE reconstructed spatial order is highly similar to the original 3D coordinates on the embryos structure (Fig. 4). For novoSpaRC, using the dot product of the optimal sample and location probabilistic coupling matrix $T_{m\times n}$ inferred by novoSpaRC and the original location $L_{n\times 3}$ (position template) as the reconstructed locations for each sample (Methods), we visualized its spatial reconstructions based on 0, 1 or 2 marker genes used. Marker genes are those having all expressions at all locations revealed to the algorithm, instead of *de novo* assigned or learned by the algorithm, so that other genes expression patterns can be compared to these genes for reference. In other words, with marker locations revealed, the novoSpaRC reconstruction is not *de novo*, but rather reference based, therefore it should offer the best performance possible for this algorithm. Judging from the ASI, AAI and OI, the *de novo* D-CE reconstruction

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

aligns better with the original spatial positions (without any predefined spatial marker genes) than using no marker, and even at least as good as the result of novoSpaRC using 1 marker gene (Fig. 4). These results are further confirmed by the expression patterns of 4 spatially distributed TF genes, *sna*, *Kr*, *eve* and *ken* (Supplementary Fig. 9). D-CE completely reconstructed the ventral expression pattern of the *sna* gene and the vertical bi-stripe pattern of *Kr*, recovered 6 out of 7 strips of *eve* and both stripes of *ken* with non-perfect placement. Whereas the *de novo* novoSpaRC without any marker gene only partially reconstructed the pattern of *ken*, but completely failed to recover the *sna* expression pattern, wrongly aggregated 7 stripes of *eve* into one broad stripe and recovered one of the two stripes of *Kr* (Supplementary Fig. 9). Notably, radius of the D-CE embedded structure which is designed to reconstruct the topological hierarchy among the single cells also reconstructed the ellipsoidal shape of the fly embryo together with larger curvature at the tips compared to the middle of the embryo (Fig. 4). A more recent de novo pseudospace reconstruction algorithm CSOmap[17] was developed based on ligand-receptor interaction pairs, which obviously is not applicable to the fly embryo developmental TF only dataset. Although the algorithm can be applied to the Geo-seq data, it apparently failed to reconstruct the spatial structure of the samples (Supplementary Fig. 10). It should be noted that novoSpaRC by default only provides 2D-grid template, thus only 2D reconstruction can be performed; CSOmap accepts only human or mouse data as input, hence it is not applicable to datasets of other species.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

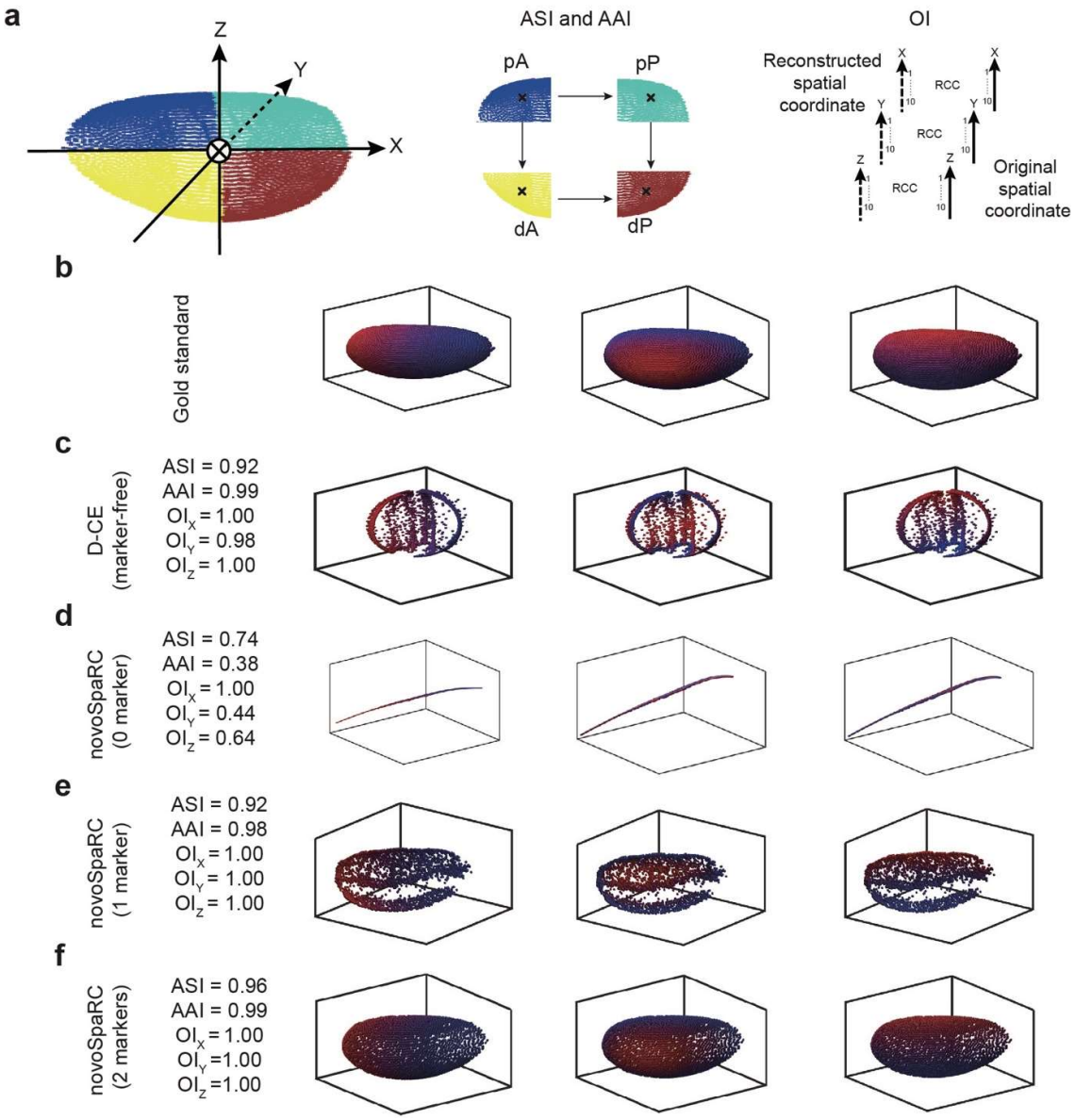

Figure 4. Comparison of spatial reconstruction of BDTNP dataset using D-CE and novoSpaRC. **a,** Illustration of Drosophila embryo segmentation for ASI, AAI and OI calculation. The embryo was divided into 4 groups along the x and z coordinates (a, middle) by spatial locations for ASI and AAI calculation. Each coordinate was sorted and divided into 10 groups (a, right) and OI was calculated based on the gold standard original spatial coordinate. **b,** Original spatial positions in Drosophila embryo examined by the BDTNP dataset, which is colored by spatial coordinates on x (b, left), y (b, middle) and z-axis (b, right), respectively. Reference coordinates of x, y and z axis are labeled in ascending order with a color gradient from blue to red, which is also used to paint the samples in the reconstructed structures to visualize their 3D orders in the following panels. **c,** D-CE spatial reconstruction of BDTNP dataset visualized by sample color code designated by the gold standards (panel b) for x, y and z axis, respectively. **d, e and f,** Same as panel b, except that the reconstruction was done by novoSpaRC with 0 marker (d), 1 marker (e) and 2 markers (f). Indexes of spatial reconstruction evaluation are shown next to each reconstructed embryo.

To further test whether D-CE can reconstruct spatial gene expression patterns of different cell types directly using scRNA-seq data, we applied it to a *Drosophila* embryo scRNA-seq dataset[11] (Supplementary Fig.11) and a zebrafish embryo blastoderm cap scRNA-seq dataset[9] (Supplementary Fig.12). In the reconstructed

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

Drosophila embryo, the expression pattern of dorsal/ventral specific gene (such as *ush*, *twi* and *sna*) are highly correlated with the FISH images downloaded from BDGP[18] dataset, with OI > 0.5. For anterior/posterior specific genes (such as *ImpE2* and *Adgf-1*), the pattern is not as good as dorsal/ventral pattern. For zebrafish embryo blastoderm cap, the gene expression pattern of 9 spatial specific gene shows high correlation (OI > 0.5) with the FISH images (downloaded from the ZFIN database[19] Compare to novoSpaRC, D-CE reconstructed gene expression patterns have significantly higher OI to the FISH images (Wilcoxon signed rank test p = 0.03 and 0.04 for the 6 and 9 spatially expressed Drosophila and zebrafish genes previously tested[11]).

For a more comprehensive comparison of D-CE with novoSpaRC and CSOmap, we applied them to 6 additional transcriptome datasets with annotated spatial ordering of samples or cell types and 8 additional transcriptome datasets[16,20-26] with annotated spatial coordinates for spatial reconstruction using the provided coordinates as evaluation gold standards. These included the human thymus[27] scRNA-seq (Fig.5a to d), zebrafish tail Geo-seq[28] (Fig. 5e,f and g), the head and neck cancer (HNC) scRNA-seq[29] (Supplementary Fig.13), human embryonic cerebral cortex scRNA-seq[30] (Supplementary Fig.14), mouse embryonic brain digitized ISH image derived gene expression data[31] (Supplementary Fig.15) and mouse neocortical layers RNA-seq[32] (Supplementary Fig.16), the mouse olfactory bulb and human breast cancer[24], the cancerous prostate[21], the melanoma lymph node[22] (Supplementary Fig. 17) and the postmortem lumbar and cervical spinal cord tissue barcoded microarray-based spatially resolved transcriptome datasets[23] (Supplementary Fig. 18), the mouse hippocampus[20] and the mouse brain[25] seqFISH datasets, a mouse medial ganglionic eminence LCM-seq dataset[26] and the mouse hypothalamic preoptic region MERFISH dataset[16] (Supplementary Fig. 19). Together with the datasets tested above (Fig. 1-4), in total 16 datasets (Table 1), and 681 reconstructions (one dataset my contain multiple experiments, e.g., barcoded-microarray based spatial transcriptomic dataset contains more than 400 arrays (Table S1)) were tested using D-CE, novoSpaRC and when applicable CSOmap. The OI, ASI and AAI indexes of the D-CE reconstructions were compared with those by novoSpaRC and CSOmap using log2 fold change and Wilcoxon signed rank test. For all 3 methods, we tested using all genes, DST or LR genes for reconstruction. Across all of the reconstructed structure, D-CE is significantly superior to the other methods in ASI and AAI (Supplementary Fig. 20a), and when focusing only on the well reconstructed structure by at least 1 method (OI or ASI p value < 0.05, Fig. 5h and Supplementary Fig. 20b). As mouse spinal cord dataset has 407 arrays, more than half of the total number of all reconstructions, we also made the comparisons without this dataset (Supplementary Fig. 20c), and found a similar superior performance of D-CE. Across all the reconstructions, all expressed genes perform better than LR and DST, suggesting that the latter are not sufficient to capture all spatial information for all types of tissues.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

Table 1. Spatially labeled transcriptome datasets tested for spatial reconstruction and number of reconstructions per dataset

Dataset	Number of reconstructions
Mouse embryo Geo-seq[14]	10
BDTNP[16]	1
Zebrafish tail Geo-seq[28]	1
Human thymus scRNA-seq[27]	1
Mouse cerebral cortex scRNA-seq	2
Mouse embryo brain ISH[31]	1
Mouse neocortical layer microsurgical RNA-seq[32]	1
Mouse olfactory bulb and human breast[24]	16 (arrays)
Head and neck cancer scRNA-seq[30]	1
Cancerous prostate[21]	12 (arrays)
Melanoma lymph node[22]	8 (arrays)
Postmortem lumbar and cervical spinal cord tissue[23]	407 (arrays)
Mouse hippocampus[20]	21 (arrays)
Mouse brain[25]	14 (arrays)
Mouse medial ganglionic eminence[26]	4 (arrays)
MERFISH[16]	181 (arrays)
Total	681

In conclusion, we developed D-CE which is an effective landmark free and model free *de novo* 3D reconstruction method for single cell analysis. We demonstrated - through comprehensive analysis of currently available spatial transcriptomes - the superior performance of D-CE over the existing reconstruction methods on nearly 700 reconstructions. We also found developmental signaling transcription factor genes (not necessarily only the current DST gene set, as not all developmental genes have been uncovered) can often serve as a spatial signature for network embedding (in particular for embryo structures). However, when not all developmental genes are known for the tissue or process, using all genes can more reliably reconstruct all tissues and developmental structures. On one side, this enables designing an effective strategy to implement *de novo* 3D reconstruction by D-CE. We find that filtering PCC using CSI performs better than directly using PD network for large datasets (more than 150 samples) as indicated by gradient down-sampling of the BDTNP dataset (Supplementary Fig. 8), which indicates that when using D-CE *de novo* spatial reconstruction method, both PD and PCC-CSI need to be tested and, in particular, for a dataset with more than 150 samples, CSI filtering should be implemented.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

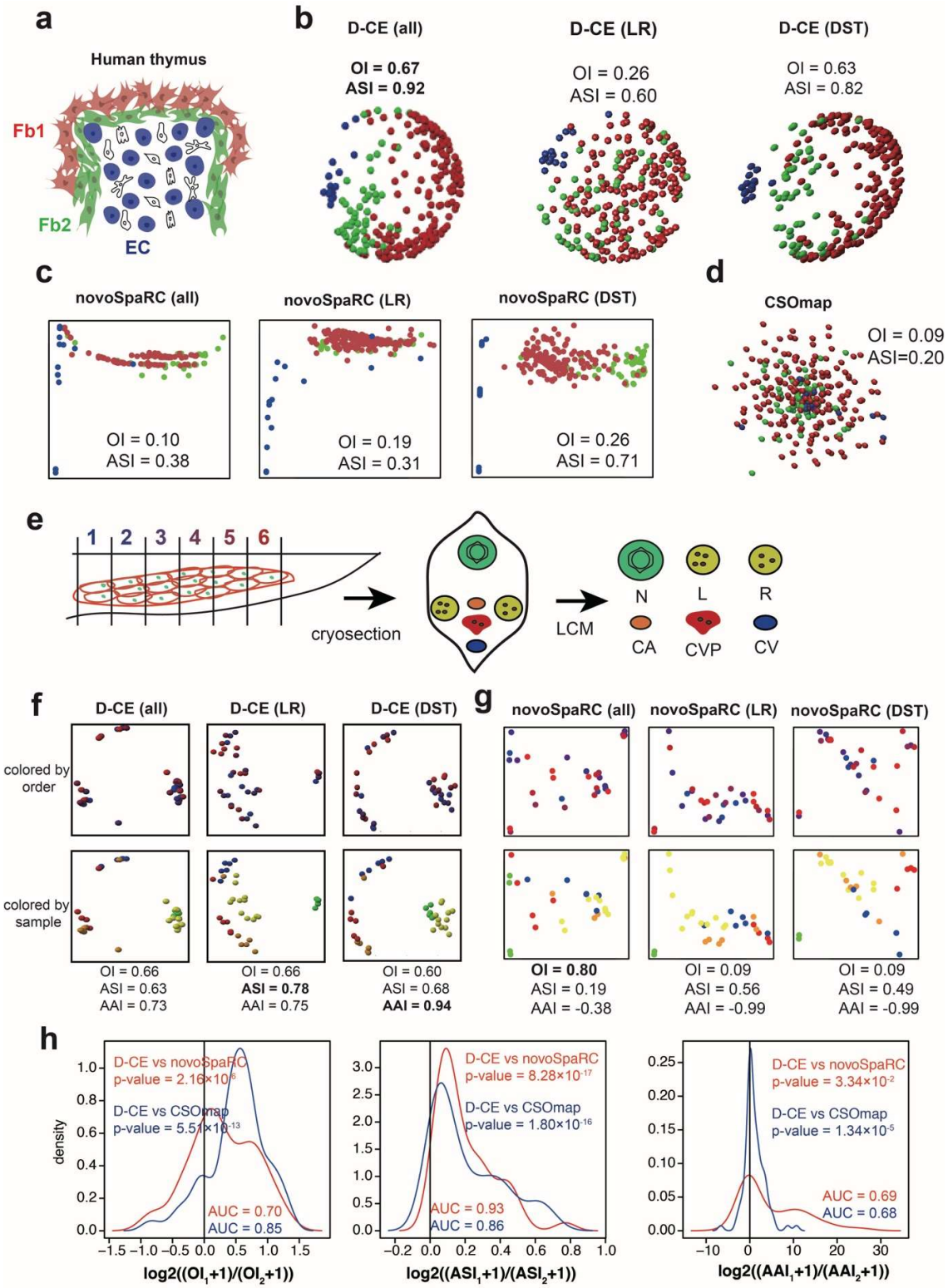

Figure 5. Spatial reconstruction of human thymus and zebrafish caudal tissue samples. a, Illustration of spatial structure of Fb1 (fibroblasts cell type1), Fb2 (fibroblasts cell type2) and EC (endothelial cells) in human thymus. **b,** D-CE reconstructed structure using all genes (left), LR (middle) and DST genes (right). For each index, the best performing values among all three methods and three gene lists are highlighted by bold font. **c**, novoSpaRC reconstructed structure using the same gene list as c, d and e. **d** CSOmap reconstructed structure of thymus. **e**, Positions of Geo-seq of laser capture

microdissection (LCM) zebrafish caudal hematopoietic tissue (CHT) samples. CHT region at 55 hpf was embedded for cryo-section, and subsequently six regions, including neuro (N), left muscle (L), right muscle (R), caudal artery (CA), caudal vein (CV), and caudal vein plexus (CVP). **f**, D-CE reconstructed structure using all genes (left), LR (middle) and DST genes (right). The top rows were colored by Geo-seq layer orders from 1 (blue) to 6 (red), the second row were colored by cell types. **g**, The same reconstructions as in f but by NovaSpaRC. **h**, Pairwise comparisons of 681 reconstructions from 16 datasets. We filtered out the un-reconstructable datasets with no method's reconstruction getting a OI p-value < 0.05 by the "cor.test" function in R. The left panel is the distribution of $\mathbf{log_2}\left(\frac{OI_1+1}{OI_2+1}\right)$. OI_1 means OI_{D-CE}. OI_2 means $OI_{novoSpaRC}$ (red line) or OI_{CSOmap}(blue line). The AUC represents the area under curve of density plot with $\mathbf{log_2}\left(\frac{OI_1+1}{OI_2+1}\right)$ > 0, that is when D-CE performs better than the other method. For each comparison, pairwise Wilcoxon signed rank test p-values are labeled on the plot. The distributions for ASI and AAI are similarly shown in the middle and right panels. CSOmap is only applicable to 14 datasets and 679 reconstructions due to its limit to human and mouse data and LR genes only.

Methods

Datasets for gene set selection

186 Geo-seq samples with known positions in mouse embryo E6.5, E7.0 and E7.5 and 69 scRNA-seq datasets (GSE120963) in E7.0 mouse embryo with A and P spatial labels were used to develop the 3D reconstruction method. For better comparison with novoSpaRC, the same expression matrix as novoSpaRC is used, which is downloaded from https://github.com/rajewsky-lab/novoSpaRC[11].

Genes for cell-cell network construction were selected based on 3 gene lists: 4512 developmental genes based on GO database[33,34], which are genes with GO terms containing keywords of 'differentiation', 'development' and 'morphogenesis', 2302 transcription factors obtained from the AnimalTFs[35] and RIKEN databases[36], and 4895 signaling genes obtained from our previous curation[37]. All samples are first subjected to batch effect correction by ComBat. Then the batch effect corrected RPKMs (Reads Per Kilobase of exon model per Million mapped reads) are used for further analysis. All expressed genes are defined as RPKM>1 in at least 2 samples. Among them, there are 3795 developmental genes, 1646 transcription factors and 3470 signaling genes for downstream analysis. The union, intersection and difference of each pair of datasets or among the 3 datasets, all expressed genes and all expressed genes minus the developmental genes were used to generate a total of 19 gene lists for spatial reconstruction, gene expression levels are transformed by $\log_{10}(FPKM+1)$. For each dataset, 12 different normalization methods were applied to each gene set (Supplementary Table 1), which give rise to a final of 228 datasets for network construction using D-CE. In order to embed the scRNA-seq and Geo-seq data together, ComBat was first used to eliminate batch effects.

The top PC loading genes are selected using function 'dimdesc' R package 'FactoMineR'[38] , the genes with p value $<10^{-10}$ are selected as top PC loading genes which result in 5731, 898 and 585 genes for PC1, 2 and 3, respectively. These 3 gene sets, individually and combined, were compared to DST genes on the performance of D-CE. DST genes were also compared to Scialdone *et al*'s pseudospace genes[6], which is a set of genes displaying a gradient along pseudospace axis, 334 assigned to anterior and 87 to posterior, the union of these genes are used for spatial reconstruction.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

Developmental Coalescence Embedding (D-CE)

We propose a new algorithm that we name D-CE and is designed under the framework of CE methodology[12], according to which the network nodes in the embedded space are ordered preserving hidden relations of: i) homophily (similarity) on the angular coordinates and ii) hierarchy on the radial coordinates[12]. In this study, our main hypothesis is that, according to CE rationale, the embedding of a developmental network of transcriptomic topological similarity between cells (an association network derived from their gene expression) should produce an angular coalescent cell ordering that recapitulates the original single cell samples' 3D spatial tissue distribution. While the cell hierarchy on the radial coordinates is obtained via a measure of node centrality in the network topology. The details of how to implement angular and radial inference is in the network embedding sub-section below. Indeed, any CE algorithm such as D-CE consists of two steps (see Fig. 1b): 1) network construction; 2) network embedding. In the next two sub-sections we will describe respectively the specific design of each of these two steps for our proposed D-CE. The Matlab code of D-CE for *de novo* 3D reconstruction is an open access tool downloadable at https://github.com/JackieHanLab/D-CE.

Step 1: network construction

In this section we describe how to build a weighted association network between Geo-seq or cell samples in order to perform the first step of D-CE. The final weighted association network is represented as a distance adjacency matrix that is obtained from the conversion of node similarities in node distances. This association network is used in step 2 in order to perform the network embedding which provides the angular coordinates that allow the 3D spatial reconstruction. We propose two different strategies that can be used to build the distance matrix.

The first strategy is the following. For each normalized gene set (normalization is first done within the gene set) the pairwise distance matrix between samples is generated by using Spearman distance (SD):

$$SD=1-RCC,$$

where RCC is the Spearman correlation coefficient, or Pearson distance (PD):

$$PD=1-PCC,$$

where PCC is the Pearson correlation coefficient, or directly the Euclidean distance (ED) between each pair of samples.

In addition, we also considered a second strategy that is designed to apply a soft-threshold that penalizes non-topological-specific low correlations and rewards local connectivity similarities that are associated to high correlations. This second strategy can be applied only to adjust correlation networks, hence we will apply it only to RCC and PCC. The first step is based on computing the Connectivity Specificity Index (CSI)[15]. For instance, in the case of the Pearson correlation PCC, CSI sparsifies (removing negligible links that are put to zero) the PCC similarity network according to this formula:

$$PCC\_CSI_{i,j} = \frac{number\ of\ nodes\ connected\ to\ i\ and\ j\ with\ PCC < PCC_{i,j} - 0.05}{number\ of\ nodes\ in\ the\ network}$$

where i and j are two samples (nodes in the correlation network). The same formula can be used to compute $RCC\_CSI_{i,j}$.

The result of this first step is a similarity matrix where a zero element indicates that the similarity between two samples is negligible according to CSI. Then, the nonzero

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

elements (x_+) of this similarity matrix are 'reversed' to obtain a distance matrix according to this reverse function:

$$f(x_+) = abs(x_+ - min(x_+) - max(x_+))$$

where abs is the absolute value and min and max are respectively the minimum and maximum. Finally, after this distance matrix is created, in order to assign a distance also to nonadjacent node pairs (which are the zero elements), the shortest path between each pair of nonadjacent nodes is computed and its value is stored as their distance. This generate a PCC-CSI distance matrix. Applying the same strategy, we can generate also the RCC-CSI by substituting PCC with RCC in the procedure above.

In summary, we propose 5 different distance matrices which represent 5 different network construction options for D-CE: Pearson distance (PD), Spearman distance (SD), Euclidean distance (ED), PCC-CSI distance and RCC-CSI distance.

Step 2: network embedding

After getting the sample-sample distance matrix (network) according to one of the 5 different options described above, the network embedding step of the D-CE algorithm consists of two routines.

(2.1) The first routine is associated with inferring the 3D angular coordinates of the samples and consists of two subroutines. The first subroutine is the n-by-n distance matrix doubly-centering operation given by the formula:

$$\bar{X} = X - \frac{1}{n} \cdot O \cdot X - \frac{1}{n} \cdot X \cdot O - \frac{1}{n^2} \cdot O \cdot X \cdot O$$

where O is an n-by-n matrix of all 1's.

The second subroutine is the spectral decomposition of the doubly-centered distance matrix by means of the singular value decomposition (SVD):

$$\bar{X} = U \cdot S \cdot V'$$

$$D_{n,3} = \left(sqrt(S_{3,3}) \cdot (V_{n,3})'\right)'$$

where S is an n-by-n diagonal matrix with singular values of $\bar{X}$ on its diagonal and $sqrt$ is the square root operation. U and V are two unitary matrices, the columns of which are singular vectors of $\bar{X}$. V' is the Hermitian transpose (the complex conjugate of the transpose) of V. $D_{n,3}$ is the score matrix where each row is a node of the network and each column is a different dimension of embedding that can be used to assign to each network node respectively the x, y, z coordinate of the 3D embedding. Then, the coordinates are transformed into polar coordinates and the angular coordinates are kept.

(2.2) The second routine is associated with inferring the radial coordinates as a function of the node strength, which is the sum of the edge similarities incident on a node. A node with high strength is very similar to many other nodes in the network[12], therefore it is high in the topological hierarchy[12]. Indeed, being similar to many nodes means that many nodes consider you at the center of the connectivity structure. Hence, according to the CE methodology[12], nodes with higher strength should be located towards the center of the embedding and therefore have lower radial coordinates; whereas nodes with lower strength should be located towards the periphery of the embedding and therefore have higher radial coordinates. On the basis of this rationale we design a procedure to infer the radial coordinates that is described step by step below. For a certain node i, its similarity is defined as:

$$S_i = \sum_{j=1}^{N} s_{i,j}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

where $s_{i,j}$ is the similarity metric between node i and j, and N denotes all nodes connected to i. For Spearman, Pearson and Euclidean distance network, $s_{i,j} = 1 - d_{i,j} / \max_{i,j} d_{i,j}$, where $d_{i,j}$ is the distance between node i and j, and for CSI network, $s_{i,j} = CSI_{i,j}$.

Nodes are first sorted in descending order by strength with nodes with highest strength ranked first. Then, the radial coordinate of the i-th node is determined by the following formula that we term *heterogeneity-adaptive radius (HAr)* and is specifically designed for D-CE embedding in order to capture the node hierarchy and according to the rationale that we clarify below:

$$HAr_i = 1 - \frac{\beta}{\ln(o_i) + 1}$$

o_i is the ranking value of i-th node, and the logarithm adjustment $\ln(o_i)$ of the ranking value is introduced to mitigate the growth of the denominator when the strength of a node is low and, possibly, many nodes with similarly low strength are arbitrarily ordered in the high value zone of the ranking. The adjustment coefficient $\beta = \frac{RSD}{1+RSD}$, where $RSD = \frac{std(S)}{mean(S)}$ is the relative standard deviation of the strength among all nodes, is a measure of heterogeneity of the node strength distribution.

HAr_i is confined to the interval]0,1[. The reversed square brackets indicate that the value 0 and 1 are respectively the inferior and superior limits of the interval, but in practice they cannot be reached. For a networked system with high hierarchical organization, the node strength will have large heterogeneity because the distribution of the node strength will have high RSD and, as a consequence, $\beta \to 1$. Hence, $\beta \to 1$ means that the network has very high hierarchical organization and, to reflect this feature in the embedding visualization, the node with highest strength (for which $o_i = 1$) takes $r_i \to 0$ and is located more towards the center of the embedding, then all the other nodes following it will assume larger radial values in the range]0,1[. In contrast, if the network hierarchical organization is lower than the previous case, for example, we assume that $\beta = 1/2$, then the node with the highest strength (for which $o_i = 1$) takes $r_i = 1/2$, and the radial coordinates space available for the representation will be squeezed to the radial interval [0.5,1[. In conclusion, the proposed formula to infer radial node coordinates in D-CE will visually represent networked systems with very high hierarchical organization as a 3D distribution of points occupying all the radial space from the periphery to the center of the radial coordinates. Instead, in case of networked systems with very low hierarchical organization, all the nodes will be compressed and equally distributed towards the periphery of the polar coordinate representation, and occupy only a reduced and peripheral portion of the radial space.

Note that some networked systems might have a hierarchical organization that changes across the angular coordinates with a certain pattern. In this case, the 3D embedding might result in not spherical and assume other shapes, for instance an ellipsoidal shape.

Optimizing gene list, normalization and edge weight for reconstruction

The 3D coordinates of each sample from each of the 1140 possible strategies of D-CE (we considered each possible combination of 19 gene meta-sets, 12 normalization methods and 5 edge weighting distances) were obtained to evaluate the performance of each of them in spatial reconstruction as follows.

For spatial reconstruction, the anterior and posterior samples are further divided into 4 groups according the spatial locations: dA, dP, pA and pP, which means

the distal/proximal part of anterior/posterior, the four groups are colored with red, green purple and yellow, respectively, in Fig. 3, Supplementary Fig. 2 and 3. For each germ layer in each stage, only the angular coordinates of the samples (with the radius all set to 1) are used to calculate the following three indexes. 1) The angular separability index (ASI) of each part from the rest of the samples is defined and implemented by Muscoloni et al.[12]; 2) The angular alignment index (AAI) is the minimum of two cosine values between two pairs of vectors $\mathrm{AAI} = \min\left(cos\left(\overrightarrow{C_{dA}C_{dP}} \cdot \overrightarrow{C_{pA}C_{pP}}\right), cos\left(\overrightarrow{C_{pA}C_{dA}} \cdot \overrightarrow{C_{pP}C_{dP}}\right)\right)$ to measure the parallelness of A to P orientations in p and d samples and p to d orientation in A and P samples, the AAI ranges from -1 to 1, with 1 indicating the two vectors perfectly parallel, and -1 perfectly anti-parallel. C_{pA}, C_{pP}, C_{dA} and C_{dP} denote the geometric center of pA, pP, dA and dP samples. For each group of samples, the coordinates of the geometric center of the group is defined as $\left(\frac{\sum_{i=1}^{N} x_i}{N}, \frac{\sum_{i=1}^{N} y_i}{N}, \frac{\sum_{i=1}^{N} z_i}{N}\right)$, where N is the total number of samples in this group and (x_i, y_i, z_i) is the 3D coordinate of the i-th sample in the group. With the coordinates of the 4 centers $C_{gi} = \left(x_{gi}, y_{gi}, z_{gi}\right)\ i = 1, 2, 3\ and\ 4$ determined, the cosine value between $\overrightarrow{C_{g1}C_{g2}}$ and $\cdot \overrightarrow{C_{g3}C_{g4}}$ is calculated as

$$cos\left(\overrightarrow{C_{g1}C_{g2}} \cdot \overrightarrow{C_{g3}C_{g4}}\right) = \frac{(x_{g2}-x_{g1})\cdot(x_{g4}-x_{g3})+(y_{g2}-y_{g1})\cdot(y_{g4}-y_{g3})+(z_{g2}-z_{g1})\cdot(z_{g4}-z_{g3})}{\sqrt{(x_{g2}-x_{g1})^2+(y_{g2}-y_{g1})^2+(z_{g2}-z_{g1})^2}\cdot\sqrt{(x_{g4}-x_{g3})^2+(y_{g4}-y_{g3})^2+(z_{g4}-z_{g3})^2}}$$

3) The ordering index (OI) is used to compare the order of samples to the known order of samples. The OI of A, P, L or R samples was calculated in each reconstruction between their original proximal to distal orders and their orders in reconstructed coordinates. The original order is determined by the labels of samples, and the rank/order of samples in the reconstructed spatial structures are determined by:

a) We first calculate OI for the Geo-seq samples in each spatial domain, A, P, L or R, separately. For each domain, the reconstructed spatial positions of samples from each layer were obtained, if there are more than one sample in the same layer, the geometric center of all the samples in this layer is used as the spatial position;

b) The center of the first layer samples is designated as the reconstructed position of the first layer, then rank the sequential layers by the shortest spatial distance to the last layer.

c) The RCC of the original and reconstructed ranks are calculated as OI.

d) The minimal OI in A, P, L and R is taken as the overall OI of the reconstruction.

Giving the 1140 possible embedding strategies, we use the 3 indexes to evaluate each of the 1140 spatial reconstruction of samples in different developmental stages and germ layers to select the best embedding as follows:

a) For each dataset, the strategies are ranked in the descending order of each of the 3 indexes to put the first to be the best performance, and the index are substituted with the ranking value, in which 1 is the highest and 1140 the lowest.

b) For each strategy, the maximum rank of each index of reconstruction of all the different stages and germ layers are calculated, and the 3 maximum ranks form an array with 3 values. For instance, Strategy12 = [45,4,115] means it perform top in AAI, well on ASI, and so-so on OI, in this way, the best method would have a good performance for each stages and germ layers.

c) A max of each vector is taken for each strategy as a measure of robustness of spatial reconstruction. For instance, in the case above, Strategy 201 = 115, in this way the best method should be robust and have a good performance for all indexes.

d) The minimum of those 1140 values (each value is the maximum rank of each strategy) is considered as the best strategy for 3D spatial reconstruction, because it ranks towards the first positions according to the three indexes.

For the BDTNP dataset, the embryo is cut into 4 groups according to the x and z coordinates in original locations, then the angular coordinates are used to calculate ASI and AAI as described above. Same as the mouse embryo, the fly embryo is bilaterally symmetric along y axis, therefore only half of the embryo is used as novoSpaRC did. As the order of cells on y axis was not considered by novoSpaRC[11], we also evaluate the reconstructed spatial order in x and z axes using the ASI and AAI of the 4 groups according to x and z axes. For OI, the x, y or z coordinate is sorted from high to low and divided into 10 groups, the original order of other groups is determined by the ascending order of coordinate, the geometric center of each groups in the reconstructed structure is calculated as the spatial location of this group. The group with lowest mean coordinate is designated as the first group, then the rest are calculated as described above for Geo-seq data.

Comparing dimensionality reduction methods in spatial reconstruction

Besides D-CE, 5 other dimensionality reduction methods, PCA, t-SNE and UMAP-corr, UMAP-cos, UMAP-euc are also tested for spatial reconstruction in Cartesian coordinates, and then they are compared with D-CE using ASI, AAI and OI. The 3 indexes of all the results using D-CE, PCA, t-SNE and UMAP are ranked in descending order, and the maximum ranks are calculated as described above.

Comparison to the existing *de novo* spatial reconstruction method

For CSOmap, the TPM of all LR genes were used for spatial reconstruction. Only human and mouse LR interaction were provided.

As novoSpaRC method is developed specifically for Berkeley Drosophila Transcription Network Project (BDTNP) dataset, which only contains the expression level of 84 TFs. For comparison to novoSpaRC, the D-CE of the BDTNP data used all 84 TFs' profiles with the same normalization method and distance metric as for to the 3D reconstruction of the Geo-seq data.

To determine the coordinates of samples reconstructed by novoSpaRC, the probabilistic coupling matrix $T_+^{m\times n}$ between m samples and n locations is calculated using novoSpaRC, then the dot product $T_+^{m\times n}\cdot L^{n\times 3}$, where $L^{n\times 3}$ is the original 3D coordinates of the locations, is used to determine the reconstructed sample locations. Specifically, $T_{i,j}$ is the probability of sample i mapping to location j, and as the distribution of $\sum_j T_{i,j}$ follow a uniform distribution, the weighted sum coordinates of all the locations for sample $i\,(\sum_{j=1}^{n} T_{i,j}\cdot x_j\,,\ \sum_{j=1}^{n} T_{i,j}\cdot y_j\,,\ \sum_{j=1}^{n} T_{i,j}\cdot z_j)$ is is used to as novoSpaRC reconstructed location (with x, y and z coordinates) of sample i, where(x_j, y_j, z_j) is the coordinate of the location j.

Down-sampling of the BDTNP dataset and comparison of PCC versus PCC-CSI network

The samples in the BDTNP dataset are randomly selected to 1/n, with n=2 to 100, of the total number of samples. The sampling is repeated 20 times at each sampling rate. For each down-sampled BDTNP sample set, the PD network and PCC-CSI networks are used for D-CE. ASI and AAI are calculated and the mean ASI and AAI of the 20 repeats at each sampling rate are plotted to compare the performance PD network and PCC-CSI network, on samples of different sizes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

Reconstruction of gene expression pattern using scRNA-seq data

The *Drosophila* and zebrafish embryo scRNA-seq data were downloaded from GEO (GSE95025, GSE66688). DST genes were converted to the homologenes of *Drosophila* and zebrafish with DAVID and R package 'Homologene'. The union of these 2 gene sets is used for downstream 3D reconstruction.

To evaluate the reconstructed structure, the matching FISH image is first converted to gray scale, and then cut into 10 layers along either anterior-posterior or dorsal-ventral axis. The gray-scale density of each layer was used as the gold standard expression level in each layer. The direction of the reconstructed structure is corrected by rotating around x and y axis with different angles (with π/30 as step size, sampled from 0 to 2π) and cut into 10 layers according to the x and z axis. Then the RCC between FISH order and each rotated reconstructed structure layer order were calculated. The orientation with max RCC is defined as the optimal orientation, whose OI is used as the final OI.

Reconstruction of transcriptome data with spatial coordinates

For 8 transcriptome data with spatial coordinates, such as barcoded microarray-based spatially resolved transcriptome, LCM-seq, seqFISH and MERFISH data, we use the spatial order or coordinates as gold standard to evaluate the reconstructed structure of each method. These 8 datasets are the mouse olfactory bulb and human breast cancer dataset[24], the cancerous prostate dataset[21] and the melanoma lymph node dataset[22] downloaded from www.spatialtranscriptomicsresearch.com; the postmortem lumbar and cervical spinal cord tissue dataset[23] downloaded from https://als-st.nygenome.org; the mouse hippocampus dataset[20] downloaded from the supplementary information of their article, the mouse brain dataset[25] downloaded from GEO with accession number GSE98674, mouse medial ganglionic eminence dataset[26] downloaded from GEO with accession number GSE60402 and the MERFISH dataset[16] downloaded from https://science.sciencemag.org/content/suppl/2018/10/31/science.aau5324.DC1. The log2 transformed gene expression values were used for reconstruction.

For each experiment, samples are labeled by their x and y coordinates and used as gold standards to compute OI, ASI and AAI after the embedding of the samples.

Reconstruction of transcriptome data with known spatial orders

We also use 6 transcriptome data with known spatial order to test different method. The human thymus atlas dataset[27] was downloaded from Zenodo repository (DOI: 10.5281/zenodo.3572422) and only the structural cells (excluding immune cells) in 7 weeks thymus tissue were used for reconstruction. The mouse embryonic brain RNA-seq dataset[31] was downloaded from the supplementary information of the paper. The zebrafish tail Geo-seq dataset[28] was downloaded from GEO with accession numbers GSE120581. The HNC scRNA-seq dataset was downloaded with accession numbers GSE103322. The mouse neocortical layer RNA-seq dataset[32] was downloaded with the accession number of GSE27243. The human embryonic cerebral cortex scRNA-seq dataset[30] was downloaded from GEO with the accession numbers GSE103723. 3D reconstruction and OI and ASI calculation were as same as the HNC dataset.

Log2 transformed FPKM (Fragments Per Kilobase of exon model per Million mapped fragments) was used for D-CE and novoSpaRC. TPM (Transcripts Per Kilobase of exon model per Million mapped reads) was used for CSOmap

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

reconstruction as required. The spatial order of the samples is reconstructed using D-CE, novoSpaRC and when applicable CSOmap (only applicable to human and mouse LR-including gene sets) and the OI of the reconstructed structure is defined as the RCC between the original order of the samples and the reconstructed order of the samples. Since AAI needs 4 samples to calculate, it is not applicable to the HNC dataset, which contains only 3 cell types.

Suppl. Algorithmic procedure for spatial reconstruction, and AAI and OI calculations

INPUT: $x_{n,g}$ (expression matrix of n samples and g genes)
OUTPUT: $D_{n,xyz}$(the cartesian coordinate of each sample)

(1) Network construction
A normalized expression matrix $xn_{n,g}$ in which each element is the square root of each element in the $x_{n,g}$ is first calculated, and a PCC network $P_{n,n}$ is built as follow:
For i=1...n
  For j=1...n
    $P_{i,j}$ is the Pearson correlation coefficient of row i and j
    $P_{i,j}=\frac{E(xn_{i,}\cdot xn_{j,})-E(xn_{i,})E(xn_{j,})}{\sqrt{E(xn_{i,}^2)-E(xn_{i,})^2}\sqrt{E(xn_{j,}^2)-E(xn_{j,})^2}}$
Then, the distance adjacency network $dist_{n,n}$ is calculated:
if CSI matrix is used
  CSI matrix is calculated as:
  for i=1..n
    for j=1...n
      $CSI_{i,j}=\frac{sum(P_{i,}<(P_{i,j}-0.05)\&P_{,j}<(P_{i,j}-0.05)}{number\ of\ nodes\ in\ the\ network}$
  The nonzero elements (CSI_+) of this similarity matrix are 'reversed' to obtain a distance matrix:
  for i=1...n
    for j=1...n
      if $CSI_{i,j}$=0
        $dist_{i,j}$=0
      else
        $dist_{i,j}$=$|CSI_{i,j}-\max(CSI_+)-\min(CSI_+)|$
  The distance between nonadjacent nodes is set as the shortest path:
  for i=1..n
    for j=1...n
      If $CSI_{i,j}$==0
        $dist_{i,j}$=shortest path between node i and j
else
  Just use Pearson distance as the weight directly:
  $dist_{n,n}=1-P_{n,n}$

(2) Network embedding
(2.1) Then, centered distance matrix $\overline{dist_{n,n}}$ is built:
$\overline{dist_{n,n}}=dist_{n,n}-\frac{1}{n}\cdot O\cdot dist_{n,n}-\frac{1}{n}\cdot dist_{n,n}\cdot O-\frac{1}{n^2}\cdot O\cdot dist_{n,n}\cdot O$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0196.v1

where O is an n-by-n matrix of all 1's;

%Then, apply SVD on the centered distance matrix and get the 3D coordinate:

$\bar{X} = U \cdot S \cdot V'$

$D_{n,xyz} = \left(sqrt(S_{3,3}) \cdot \ (V_{n,3})'\right)'$

(2.2) Radial coordinate adjustment:

if CSI matrix is used

for i = 1...n

$S_i = \sum_{j=1}^{N} CSI_{i,j}$

else

for i = 1...n

$S_i = \sum_{j=1}^{N}(1 - dist_{i,j} / \max_{i,j} dist_{n,n})$

$RSD = \frac{std(S_n)}{mean(S_n)}$

$\beta = \frac{RSD}{1 + RSD}$

The final Cartesian coordinate $D_{n,xyz}$ is calculated:

Sort the nodes according to S_n in descending order to ranks $r_{1\ldots n}$;

for i = 1...n

for j in x, y, z

Compute the original radius $R_i^{svd} = \sqrt{(D_{i,x})^2 + (D_{i,y})^2 + (D_{i,z})^2}$

Compute the final Cartesian coordinates:

$$\widetilde{D}_{i,j} = \frac{D_{i,j}}{R_i^{svd}} * (1 - \frac{\beta}{\ln(r_i)+1})$$

INPUT: $D_{n,xyz}$, L_n(the Cartesian coordinate and the spatial location (dA, dP, pA, pP) of each sample)

OUTPUT: AAI (angular alignment index)

First, the Cartesian coordinates of the geometric center of each spatial part is calculated:

for i in dA, dP, pA, pP

$\mathrm{D1}_{ni,xyz} = D_{L_n==i,}$

$$C_i = (\frac{\sum_{j=1}^{ni} D1_{j,x}}{ni}, \frac{\sum_{j=1}^{ni} D1_{j,y}}{ni}, \frac{\sum_{j=1}^{ni} D1_{j,z}}{ni})$$

The cosine value below is calculated to measure the parallelness of A to P orientations in p and d samples:

$$cos(\overrightarrow{C_{dA}C_{dP}} \cdot \overrightarrow{C_{pA}C_{pP}}) = \frac{(C_{dP_x} - C_{dA_x}) \cdot (C_{pP_x} - C_{pA_x}) + (C_{dP_y} - C_{dA_y}) \cdot (C_{pP_y} - C_{pA_y}) + (C_{dP_z} - C_{dA_z}) \cdot (C_{pP_z} - C_{pA_z})}{\sqrt{(C_{dP_x} - C_{dA_x})^2 + (C_{dP_y} - C_{dA_y})^2 + (C_{dP_z} - C_{dA_z})^2} \cdot \sqrt{(C_{pP_x} - C_{pA_x})^2 + (C_{pP_y} - C_{pA_y})^2 + (C_{pP_z} - C_{pA_z})^2}}$$

The cosine value below is calculated to measure the parallelness of p to d orientations in A and P samples:

$$cos(\overrightarrow{C_{pA}C_{dA}} \cdot \overrightarrow{C_{pP}C_{dP}}) = \frac{(C_{dA_x}-C_{pA_x})\cdot(C_{dP_x}-C_{pP_x})+(C_{dA_y}-C_{pA_y})\cdot(C_{dP_y}-C_{pP_y})+(C_{dA_z}-C_{pA_z})\cdot(C_{dP_z}-C_{pP_z})}{\sqrt{(C_{dA_x}-C_{pA_x})^2+(C_{dA_y}-C_{pA_y})^2+(C_{dA_z}-C_{pA_z})^2}\cdot\sqrt{(C_{dP_x}-C_{pP_x})^2+(C_{dP_y}-C_{pP_y})^2+(C_{dP_z}-C_{pP_z})^2}}$$

Then the final AAI is calculated as the minimum of the 2 cosine values:

$$\mathrm{AAI} = \min\left(cos(\overrightarrow{C_{dA}C_{dP}} \cdot \overrightarrow{C_{pA}C_{pP}}), cos(\overrightarrow{C_{pA}C_{dA}} \cdot \overrightarrow{C_{pP}C_{dP}})\right)$$

INPUT: $C_{n,xyz}$, D_n, L_n (the cartesian coordinate, the spatial domain each sample belonged to and the known spatial layer each sample belonged to)
OUTPUT: OI (order index)

for each domain d in D_n

$\mathrm{C1}_{nd,xyz} = C_{D_n==d,}$

$\mathrm{L1}_{nd} = L_{D_n==d}$

for i = 1... $(\max_{nd} L1 - \min_{\mathrm{nd}} L1 + 1)$

$C2_{n^i,3} = C1_{\mathrm{L1}==(\min_{nd} \mathrm{L1}-1+\ ),}$

$Cen_i = (\frac{\sum_{j=1}^{n^i} C2_{j,x}}{n^i}, \frac{\sum_{j=1}^{n^i} C2_{j,y}}{n^i}, \frac{\sum_{j=1}^{n^i} C2_{j,z}}{n^i})$

The rank of reconstructed spatial order Rr is calculated as:

for i = $1 \ldots (\max_{n^d} L1 - \min_{n^d} \mathrm{L1} + 1)$

$Rr_i = 0$

$Rr_1 = 1$

While min(r^R)=0

$\mathrm{Cmax} = Cen_{Rr==\max(Rr)}$

for each i that $Rr_i = 0$

$$\mathrm{d}_i = \sqrt{(Cen_{i_x} - \mathrm{Cmax}_x)^2 + (Cen_{i_y} - \mathrm{Cmax}_y)^2 + (Cen_{i_z} - \mathrm{Cmax}_z)^2}$$

$\mathrm{Rr}_{\mathrm{d}_i==\min_{Rr_i=0} \mathrm{d}_i} = \max(Rr) + 1$

The rank of original spatial order $Ro = 1 \ldots (\max_{nd} L1 - \min_{\mathrm{nd}} L1 + 1)$

The OI is defined as the rank correlation of Rr and Ro

$$\mathrm{OI}_a = \frac{E(Rr\cdot Ro)-E(Rr)E(Ro)}{\sqrt{E(Rr^2)-E(Rr)^2}\sqrt{E(Ro^2)-E(Ro)^2}}$$

Finally, the overall OI is defined as the minimal OI in all of the domains

$$OI_{final} = \min_d OI_d$$

Author Contributions

J.D.J.H. conceived the project. J.D.J.H. designed, with C.V.C.'s help, the project and analyses. J.D.J.H. and C.V.C invented the network construction part of D-CE. C.V.C. invented the network embedding part of D-CE and designed the algorithm. Y.Z. and S.Z. wrote, verified and tested the codes with occasional help from D.L and C.V.C. J.D.J.H., C.V.C. and Y.Z. wrote the paper with help from S.Z.

Acknowledgement

This work was supported by grants from National Natural Science Foundation of China (91749205) and China Ministry of Science and Technology (2016YFE0108700) to J.D.J.H. We thank Denghui Liu and Shengbao Suo for technical assistance.

References

1 Ferrell, J. E., Jr. Bistability, bifurcations, and Waddington's epigenetic landscape. *Curr Biol* **22**, R458-466, doi:10.1016/j.cub.2012.03.045 (2012).

2 Peng, G. *et al.* Spatial transcriptome for the molecular annotation of lineage fates and cell identity in mid-gastrula mouse embryo. *Developmental cell* **36**, 681-697 (2016).

3 Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *bioRxiv*, doi:10.1101/276907 (2018).

4 Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386, doi:10.1038/nbt.2859 (2014).

5 Haghverdi, L., Buttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* **13**, 845-848, doi:10.1038/nmeth.3971 (2016).

6 Scialdone, A. *et al.* Resolving early mesoderm diversification through single-cell expression profiling. *Nature* **535**, 289-293, doi:10.1038/nature18633 (2016).

7 Sun, N. *et al.* Inference of differentiation time for single cell transcriptomes using cell population reference data. *Nat Commun* **8**, 1856, doi:10.1038/s41467-017-01860-2 (2017).

8 Peng, G. *et al.* Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo. *Dev Cell* **36**, 681-697, doi:10.1016/j.devcel.2016.02.020 (2016).

9 Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502, doi:10.1038/nbt.3192 (2015).

10 Halpern, K. B. *et al.* Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352-356, doi:10.1038/nature21065 (2017).

11 Nitzan, M., Karaiskos, N., Friedman, N. & Rajewsky, N. Gene expression cartography. *Nature* **576**, 132-137, doi:10.1038/s41586-019-1773-3 (2019).

12 Muscoloni, A., Thomas, J. M., Ciucci, S., Bianconi, G. & Cannistraci, C. V. Machine learning meets complex networks via coalescent embedding in the hyperbolic space. *Nat Commun* **8**, 1615, doi:10.1038/s41467-017-01825-5 (2017).

13 Cacciola, A. *et al.* Coalescent embedding in the hyperbolic space unsupervisedly discloses the hidden geometry of the brain. *arXiv preprint arXiv:1705.04192* (2017).

14 Peng, G. *et al.* Molecular architecture of lineage allocation and tissue organization in early mouse embryo. *Nature* **572**, 528-532, doi:10.1038/s41586-019-1469-8 (2019).

15 Fuxman Bass, J. I. *et al.* Using networks to measure similarity between genes: association index selection. *Nat Methods* **10**, 1169-1176, doi:10.1038/nmeth.2728 (2013).

16 Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, doi:10.1126/science.aau5324 (2018).

17 Ren, X. *et al.* Reconstruction of cell spatial organization from single-cell RNA sequencing data based on ligand-receptor mediated self-assembly. *Cell Res*, doi:10.1038/s41422-020-0353-2 (2020).
18 Tomancak, P. *et al.* Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biol* **8**, R145, doi:10.1186/gb-2007-8-7-r145 (2007).
19 Ruzicka, L. *et al.* The Zebrafish Information Network: new support for non-coding genes, richer Gene Ontology annotations and the Alliance of Genome Resources. *Nucleic Acids Res* **47**, D867-D873, doi:10.1093/nar/gky1090 (2019).
20 Shah, S., Lubeck, E., Zhou, W. & Cai, L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron* **92**, 342-357, doi:10.1016/j.neuron.2016.10.001 (2016).
21 Berglund, E. *et al.* Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat Commun* **9**, 2419, doi:10.1038/s41467-018-04724-5 (2018).
22 Thrane, K., Eriksson, H., Maaskola, J., Hansson, J. & Lundeberg, J. Spatially Resolved Transcriptomics Enables Dissection of Genetic Heterogeneity in Stage III Cutaneous Malignant Melanoma. *Cancer Res* **78**, 5970-5979, doi:10.1158/0008-5472.CAN-18-0747 (2018).
23 Maniatis, S. *et al.* Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science* **364**, 89-93, doi:10.1126/science.aav9776 (2019).
24 Stahl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78-82, doi:10.1126/science.aaf2403 (2016).
25 Eng, C. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235-239, doi:10.1038/s41586-019-1049-y (2019).
26 Zechel, S., Zajac, P., Lonnerberg, P., Ibanez, C. F. & Linnarsson, S. Topographical transcriptome mapping of the mouse medial ganglionic eminence by spatially resolved RNA-seq. *Genome Biol* **15**, 486, doi:10.1186/s13059-014-0486-z (2014).
27 Park, J. E. *et al.* A cell atlas of human thymic development defines T cell repertoire formation. *Science* **367**, doi:10.1126/science.aay3224 (2020).
28 Xue, Y. *et al.* A 3D Atlas of Hematopoietic Stem and Progenitor Cell Expansion by Multi-dimensional RNA-Seq Analysis. *Cell Rep* **27**, 1567-1578 e1565, doi:10.1016/j.celrep.2019.04.030 (2019).
29 Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e1624, doi:10.1016/j.cell.2017.10.044 (2017).
30 Fan, X. *et al.* Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-seq analysis. *Cell Research* **28**, 730-745, doi:10.1038/s41422-018-0053-3 (2018).
31 Huang, Y. *et al.* Single-cell-level spatial gene expression in the embryonic neural differentiation niche. *Genome Res* **25**, 570-581, doi:10.1101/gr.181966.114 (2015).
32 Belgard, T. G. *et al.* A transcriptomic atlas of mouse neocortical layers. *Neuron* **71**, 605-616, doi:10.1016/j.neuron.2011.06.039 (2011).
33 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).

34 The Gene Ontology, C. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330-D338, doi:10.1093/nar/gky1055 (2019).
35 Zhang, H. M. *et al.* AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res* **40**, D144-149, doi:10.1093/nar/gkr965 (2012).
36 Kanamori, M. *et al.* A genome-wide and nonredundant mouse transcription factor database. *Biochem Biophys Res Commun* **322**, 787-793, doi:10.1016/j.bbrc.2004.07.179 (2004).
37 Xu, C. *et al.* Accurate Drug Repositioning through Non-tissue-Specific Core Signatures from Cancer Transcriptomes. *Cell Rep* **25**, 523-535 e525, doi:10.1016/j.celrep.2018.09.031 (2018).
38 Husson, F. o., Lê, S. b. & Pagès, J. r. m. *Exploratory multivariate analysis by example using R*. (CRC Press, 2011).