

1 Inferring cell-cell interactions from pseudotime ordering of
2 scRNA-Seq data

3 Dongshunyi Li^{*1}, Jeremy J. Velazquez^{*2,3}, Jun Ding⁴, Joshua Hislop^{2,3,5}, Mo R.
4 Ebrahimkhani^{†2,3,5,6}, and Ziv Bar-Joseph^{†1,7}

5 ¹Computational Biology Department, School of Computer Science, Carnegie Mellon
6 University, Pittsburgh, PA 15213, USA

7 ²Department of Pathology, School of Medicine, University of Pittsburgh, Pittsburgh, PA
8 15213, USA

9 ³Pittsburgh Liver Research Center, University of Pittsburgh, Pittsburgh, PA 15261, USA

10 ⁴Meakins-Christie Laboratories, Department of Medicine, McGill University Health Centre,
11 Montreal, Quebec, H4A 3J1, Canada

12 ⁵Department of Bioengineering, Swanson School of Engineering, University of Pittsburgh,
13 Pittsburgh, PA 15261, USA

14 ⁶McGowan Institute for Regenerative Medicine, University of Pittsburgh, Pittsburgh, PA
15 15219, USA

16 ⁷Machine Learning Department, School of Computer Science, Carnegie Mellon University,
17 Pittsburgh, PA 15213, USA

18 June 30, 2021

19 **Abstract**

20 A major advantage of single cell RNA-Sequencing (scRNA-Seq) data is the ability to reconstruct
21 continuous ordering and trajectories for cells. To date, such ordering was mainly used to group cells and to infer interactions within cells. Here we present TraSig, a computational
22 method for improving the inference of cell-cell interactions in scRNA-Seq studies. Unlike prior
23 methods that only focus on the average expression levels of genes in clusters or cell types, TraSig
24 fully utilizes the dynamic information to identify significant ligand-receptor pairs with similar
25 trajectories, which in turn are used to score interacting cell clusters. We applied TraSig to several
26 scRNA-Seq datasets. As we show, using the ordering information allows TraSig to obtain unique
27 predictions that improve upon those identified by prior methods. Functional experiments validate
28 the ability of TraSig to identify novel signaling interactions that impact vascular development in
29 liver organoid.
30

31 Software: <https://github.com/doraadong/TraSig>

*These authors have contributed equally to this work.

†Correspondence: Mo R. Ebrahimkhani, mo.ebr@pitt.edu; Ziv Bar-Joseph, zivbj@cs.cmu.edu

32 **Keywords**

33 Cell-cell interactions, Development, Gene expression

34 Introduction

35 The ability to profile cells at the single cell level enabled the identification of new cell types and
36 additional markers for known cell types as well as the reconstruction of cell type specific regulatory
37 networks [1, 2]. Several methods have been developed to group or cluster cells in scRNA-Seq data
38 [3] and to reconstruct trajectories and pseudotime for time series scRNA-Seq data [4]. Such methods
39 have mainly focused on the expression similarity between cells in the same cluster or at consecutive
40 time points and on the differences in transcriptional regulation between cell types and over time [5].
41 More recently, a number of methods have been developed to infer another type of interaction from
42 scRNA-Seq data: signaling between cell clusters or cell types [6]. These methods attempt to identify
43 ligands in one of the clusters or cell types and corresponding receptors in another cluster and then
44 infer interactions based on the average expression of these ligand-receptor pairs. For example,
45 CellPhoneDB [7] scores ligand-receptor pairs using their mean expression values in two clusters and
46 assigns significance levels using permutations tests. SingleCellSingleR[8] designs a score based on
47 the product of ligand-receptors' mean expression values in two clusters and selects ligand-receptors
48 scoring above a predefined threshold.
49 While successful, most current methods for inferring cell-cell interactions from scRNA-Seq data only
50 use of the average expression levels of ligands and receptors in the two clusters or cell types they test
51 [6]. While this may be fine for steady state populations (for example, different cell types in adult
52 tissues), for studies that focus on development or response modeling, such averages do not take
53 full advantage of the available data in scRNA-Seq studies. Indeed, even cells on the same branch
54 are often ordered in such studies using various pseudotime ordering methods [9]. In such cases,
55 cells on the same branch (or cluster) cannot be assumed to be homogeneous with respect to the
56 expression of key genes. Using average analysis for such clusters may lead to inaccurate predictions
57 about the relationship between ligands and receptors in two different (though parallel in terms of
58 timing) branches. Specifically, Figure 1 presents four cases of pseudotime orderings for a ligand and
59 its corresponding receptor in two different branches. While the *average* expression of a ligand and
60 receptor in two different branches are the same, the first two cases are unlikely to strongly support an
61 interaction between these two cell types while the third and fourth, where both are either increasing
62 or decreasing in their respective ordering, are much more likely to hint at real interactions between

63 the groups. In other words, if two groups of cells are interacting, then we expect to see the genes
64 encoding signaling molecules in these groups co-express at a similar pace along the pseudotime.

65 To enable the use of pseudotime ordering for predicting cell type interactions, we developed TraSig.
66 TraSig can use several of the most popular pseudotime ordering and trajectory inference methods
67 to extract expression patterns for ligands and receptors in different edges of the trajectory using a
68 sliding window approach. It then uses these profiles to score temporal interactions between ligand
69 and their known receptors in different edges corresponding to the same time. Permutation testing is
70 used to assign significance levels to specific pairwise interactions and scores are combined to identify
71 significant cluster-cluster interactions.

72 We applied TraSig to a number of scRNA-Seq datasets and compared its performance to a number
73 of popular methods for inferring signaling interactions from scRNA-Seq data. As we show, the
74 ability to utilize the temporal information in the analysis improves the accuracy of predicted relevant
75 pairs and leads to distinct predictions that are not identified by other methods that rely on average
76 expression. We experimentally validated a number of interaction predictions from TraSig for liver
77 organoid differentiation data.

78 Results

79 We developed a computational method, TraSig for inferring cell-cell interactions from pseudotime
80 ordered data. Figure 2 presents an overview of the method. We start by using a trajectory inference
81 method to obtain grouping and pseudotime ordering for cells in the dataset. Here we use continuous
82 state Hidden Markov model (CSHMM) [10] for this, though as discussed below, TraSig can be
83 applied to results from other pseudotime ordering methods. We then reconstruct expression profiles
84 for genes along each of the edges using sliding windows summaries. Next we compute a dot product
85 score for pairs of genes in edges (clusters) sampled at the same time or those representing the
86 same pseudotime. Finally, we use permutation analysis to assign significance levels to the scores we
87 computed. See Methods for details on each of the steps of TraSig.

88 Reconstructing dynamic liver development model using CSHMM

89 We first applied TraSig to a liver organoid differentiation scRNA-seq dataset composed of 11,083 cells
90 sampled at two time points: day 11 and day 17 [11]. The data was preprocessed using a standard
91 Seurat V3 [12] pipeline and cell types were assigned as previously discussed [11]. These were used to
92 initialize trajectory inference using CSHMM [10]. Following filtering to remove genes not expressed
93 in any of the cells, 26,955 genes were used to learn the CSHMM model. Figure 3a presents the
94 resulting model learned for this data. As can be seen, the method identifies 12 clusters (edges) for
95 these data. These agree very well with the clustering assignments from the Seurat single cell analysis.
96 Specifically, CSHMM assigns separate edges for hepatocyte- (edge 3, 5, 9 and 10), endothelial- (edges
97 7 and 11), stellate- (edges 2 and 8), and ductal/cholangiocyte-like (edges 4 and 6) cells. In addition,
98 the model also presents informative pseudotime ordering of cells as we discuss below based on the
99 reconstructed expression profiles for key marker genes.

100 Inferring cell type interactions for liver development

101 We next applied TraSig to the model reconstructed by CSHMM in order to gain insight into
102 developmental signaling of co-differentiating liver cells from multiple germ layers. Such data is
103 severely lacking for humans and so the use of the trajectory learned for liver organoid differentiation
104 can provide valuable information on interactions regulating liver development. We thus tested all
105 pairs of edges for which the assigned cells were from the same time point (Supplementary Notes).
106 Figure 6a presents the results for scoring interactions between edges representing the same time
107 (Methods). For the day 11 clusters (edge 1, 2, 3, 4, 5, 7), we find strong interactions between
108 stellate-like 1 cells (edge 2) and endothelial-like cells (edge 7) and between ductal/cholangiocyte-like
109 cells (edge 4) and endothelial-like cells (edge 7). For the day 17 clusters (edge 6, 8, 9, 10, 11), we
110 find that the strongest interactions are between the ductal/cholangiocyte-like cells (edge 6) and
111 stellate-like cells (edge 8). We also find high scoring interactions between stellate-like cells (edge
112 8) and endothelial-like cells (edge 11) and between ductal/cholangiocyte-like cells (edge 6) and
113 endothelial-like cells (edge 11) for the day 17 clusters. The detection of significant interactions
114 between the endothelial, stellate, and cholangiocyte cell types is further supported by their proximity
115 in the liver. The stellate cells wrap around the endothelial cells and are bordered by the cholangiocyte

116 comprised bile ducts [13].

117 **TraSig identifies ligand-receptor interactions important to vascular development**

118 We evaluated the significant ligand-receptor pairs that were ranked highly by TraSig for the high
119 scoring cluster pairs. We found that many agree with known functions and signaling pathways
120 activated during liver development. Figure 3 presents a few examples of identified ligand-receptor
121 pairs. We next studied the top scoring edges predicted to interact with endothelial-like cells.
122 Endothelial cells play a major role in vascular development in liver [14]. To study the interactions of
123 such cells, we looked for cluster pairs for which the receiver (receptor) cluster is the day 17 endothelial-
124 like cell cluster (edge 11). GO term analysis of the identified ligands and receptors for these cluster
125 pairs identifies several relevant functional terms related to vascular development including “blood
126 vessel development” (minimum p-value among cluster pairs $3.91939e - 65$), “regulation of endothelial
127 cell proliferation” (p-value $3.76500e - 27$) and “vascular process in circulatory system” (p-value
128 $7.27963e - 12$).

129 Many of the ligand-receptor pairs identified for interactions involving the endothelial-like cells
130 are known to play a role in endothelial cell specification, migration, and angiogenesis further
131 supporting the results of TraSig. Of note, we identified pairs including VEGFA/VEGFB/VEGFC
132 with FLT1/KDR, which is required for proper liver zonation, sinusoid endothelial cell specification,
133 and endothelial lipoprotein uptake [15, 16]; DLL4 with NOTCH1/NOTCH4, which is essential for
134 endothelial tip and stalk cell crosstalk and liver sinusoidal endothelial cell capillarization [17, 18];
135 CXCL12 with CXCR4, which has been shown to promote endothelial cell migration and lumen
136 formation independent of VEGF [19]; MDK with PTPRB, which is of great interest for its known
137 impact on cancer angiogenesis [20, 21]; and CYR61 with ITGAV, which represents one of the many
138 integrin interactions identified by TraSig which activate PI3K/AKT downstream signaling, and is
139 known to regulate tip cell activity and angiogenesis (Figure 4a-d) [22].

140 **Experimental validation for predicted TraSig pairs**

141 Given the success in identifying known interactions, we next experimentally validated additional
142 TraSig predictions. We first assessed if there was a correlation between the signal level of CXCL12 or
143 VEGF and vascularity via immunofluorescent staining of liver organoid cultures. As shows in Figure

144 5a-c, we found that loci with high relative expression of CXCL12 and VEGF co-localized with regions
145 of increased vessel area percentage and vessel junction density, when compared to loci with relative
146 low expression of CXCL12 and VEGF measured by AngioTool analysis of the immunofluorescent
147 staining (see also Figures S3a and S3b).

148 This motivated further investigation into the significance of predicted signaling interactions in the
149 liver organoid cultures as they pertain to vascular development. We therefore performed prolonged (5
150 days from D9-14) inhibition of several predicted signaling proteins: VEGF, NOTCH, CXCR4, MDK,
151 and PI3K (downstream of MDK and multiple integrin interactions). These experiments validated
152 several of the predictions. Specifically, we observed significant decreases in percent vessel area,
153 junction density, and average vessel length were detected in the VEGF, MDK, and PI3K conditions,
154 while NOTCH inhibition revealed an opposite effect (Figure 5d and 5e). In contrast, the local
155 correlation of increased vascular network formation with high CXCL12 expression did not carry over
156 to a negative global effect via CXCR4 inhibition, indicating opportunity for further investigation,
157 perhaps involving alternative inhibitors or assessment of the alternative CXCL12 receptor CXCR7,
158 which also plays important roles in angiogenesis and liver regeneration [23, 24].

159 Comparing TraSig with prior methods

160 We compared interactions predicted by TraSig to two popular methods for inferring cell type
161 interactinons from scRNA-Seq data: CellPhoneDB [7] and SingleCellSignalR [8]. Both methods use
162 the overall expression of genes in clusters and unlike TraSig do not use any ordering information. For
163 both methods, we tested the same cluster pairs as we did for TraSig. To make the comparisons more
164 consistent, we combined the paracrine and autocrine predicted interactions for SingleCellSignalR
165 since this is what other methods do. Figure 6a presents scores for all cluster pairs for TraSig,
166 SingCellSignalR, and CellPhoneDB. As can be seen, while some pairs score high for all methods,
167 others are only identified by one or two of the methods. Specifically, SingleCellSignalR seems to
168 assign similar scores for most pairs whereas both TraSig and CellPhoneDB assign more variable
169 scores. Figure 6c presents the Venn diagrams for the overlap between ligands and receptors identified
170 by the three methods for two example cell cluster pairs. In both cases, the receiver (receptor) cluster
171 is the day 17 endothelial edge (edge 11). While SingleCellSignalR and TraSig overlap in roughly
172 50% of the identified ligands and receptors, the overlap with CellPhoneDB is much lower. This is

173 likely a result of the database of interactions used by CellPhoneDB which is smaller than the ones
174 used by the other methods.

175 To evaluate the predicted pairs from these methods, we performed validation experiments, as
176 mentioned above, and also compared enrichment p-values for relevant GO terms using ligands
177 and receptors for several high scoring cluster pairs from each of the methods (See Supplementary
178 Notes on how we select relevant GO terms). Among the significant ligand-receptors we successfully
179 validated, DLL4-NOTCH4, MDK-PTPRB and CYR61-ITGAV are only identified by TraSig. As
180 for GO analysis, Figure 6b shows that TraSig leads to more significant relevant categories when
181 compared to the two other methods. For example, TraSig obtains a minimum p-value among cluster
182 pairs of $5.91657e - 60$ for “blood vessel morphogenesis” whereas the minimum p-values for this
183 category are higher for the other two methods ($6.40837e - 54$ and $1.10356e - 23$ for SingleCellSignalR
184 and CellPhoneDB respectfully). For “endothelial cell migration”, TraSig has a minimum p-value of
185 $6.03035e - 24$, again, lower than the minimum p-values for SingleCellSignalR ($1.64124e - 17$) and
186 CellPhoneDB ($4.90735e - 13$).

187 **TraSig identifies interactions in neocortical development**

188 To further evaluate TraSig’s performance, we applied TraSig to a mouse neocortical development
189 scRNA-seq data [25]. After preprocessing (Supplementary Notes), we obtained 18,545 cells sampled
190 at two time points: E14.5 and P0. We used the top 5000 dispersed genes to reconstruct CSHMM
191 trajectories. The CSHMM model was initialized using the cell labels from [25]. Next the model
192 was refined to improve both trajectory learning and cell assignment. The final trajectory learned
193 for this data is presented in Figure S6. The model is composed of 44 clusters (edges) of which 23
194 contain cells from the first time point and 21 from the second. Next we applied TraSig to infer
195 ligand-receptors pairs and interacting cluster pairs based on the sampling time.

196 Figure S4a presents scores for all cluster pairs. As can be seen, the method identified strongly
197 interacting cluster pairs for both time points. The highest scoring interactions identified involve
198 either endothelial cells (edge 18 from E14.5 and edge 39 from P0), radial glial cells (edge 1 from
199 E14.5), interneurons (edge 24 from P0), or astrocytes (edge 26 from P0). We performed GO analysis
200 using the significant ligands and receptors identified for radial glial cells in E14.5 or interneurons in
201 P0. Figure S4b shows the $-\log_{10}$ p-value of enriched GO terms for interactions involving either RG2

202 [14-E] cluster for the radial glial cells in E14.5 (edge 1) or Int2 [14-P] cluster for the interneurons in
203 P0 (edge 24). Radial glial cells were identified as progenitor cells for neocortical development [26]
204 and determined to function as “scaffolds” for neuronal migration [27]. GO analysis shows that the
205 signaling proteins identified by TraSig for interactions involving this cluster are indeed related to such
206 functions and include “cell migration” (p-value $1.69780e - 60$), “cell motility” (p-value $1.01291e - 56$)
207 and “regulation of cell migration” (p-value $9.23644e - 42$). Terms related to neuron development
208 are also highly enriched in the set of ligand and receptor proteins identified for the interneuron
209 cell cluster and include “neurogenesis” (p-value $1.39908e - 64$) and “neuron projection development”
210 (p-value $5.39174e - 64$).

211 Applying TraSig to trajectories obtained by Slingshot

212 To test the ability of TraSig to generalize to pseudotime inferred by additional methods, we used
213 it to post-process trajectories inferred by Slingshot [9]. Slingshot is a trajectory inference method
214 that first infers a global lineage structure using a cluster-based minimum spanning tree (MST)
215 and then infers the cell-level pseudotimes for each lineage. We applied Slingshot and TraSig to
216 an oligodendrocyte differentiation dataset composed of 3,685 cells [28, 4]. Figure S5a presents the
217 trajectory learned by Slingshot for this data. Figure S5b presents the interactions predicted by
218 TraSig for the inferred trajectory. Cells assigned to edges 2 and 3 are more mature cells while those
219 assigned to edges 0 and 1 containing precursor cells (Figure S5a). Our results suggest that the more
220 mature oligodendrocytes are signaling to the precursors during development. As before we performed
221 GO analysis on the set of ligands and receptors predicted for strongly interacting clusters. We found
222 several relevant GO terms including “neuron projection development” (p-value $2.50804e - 24$) and
223 “neuron development” (p-value $7.129894e - 23$) (Figure S5c). Ligands in top ranking ligand-receptor
224 TraSig pairs include PDGFA, BMP4 and PTN, all of which are known to be involved in regulating
225 oligodendrocyte development [29, 30, 31].

226 Discussion

227 Initial methods for the analysis of scRNA-Seq data mainly focused on within cluster or trajectory
228 interactions. Recently, a number of methods have been developed to use these data to infer interactions

229 between different cell types or clusters [6]. These methods focus on the average expression of ligands
230 and their corresponding receptors in a pair of cell types to score and identify interacting cell types
231 pairs.

232 While the exact way in which scores are computed differs between methods developed to predict such
233 interactions, to date most methods looked at the average or sum of the expression values for ligands
234 and receptors in the two clusters or cell types. Such analysis works well when studying processes that
235 are in a steady state (for example, adult tissues) but may be less appropriate for dynamic processes.
236 For real interactions, when time or pseudotime information is available, we expect to see not just
237 average expression levels match but also trajectory matches in their expression profiles. Since many
238 methods have been developed to infer pseudotime from scRNA-Seq data, such information is readily
239 available for many studies.

240 To fully utilize information in scRNA-Seq data we developed TraSig, a new computational method
241 for inferring signaling interactions. TraSig first orders cells along a trajectory and then extracts
242 expression profiles for genes in different clusters using a sliding window approach. Matches between
243 profiles for ligands and their corresponding receptors in different clusters are then scored and their
244 significance is assessed using permutation tests. Finally, scores for individual pairs are combined to
245 obtain a cluster interaction score.

246 We applied TraSig to several different scRNA-Seq datasets and have also compared its predictions
247 to predictions by prior methods developed for this task. As we have shown, for liver organoid
248 development, TraSig was able to identify several known and novel interactions related to the
249 regulation of vascular network formation. These interactions involve endothelial, stellate, and
250 cholangiocyte cell types that have been known to reside in close proximity [13] and several ligand-
251 receptor pairs known to be involved in vascular development. While many interactions were predicted
252 by all methods we tested, there are also several interactions uniquely predicted by TraSig. We
253 validated a number of these interactions including DLL4-NOTCH4 and MDK-PTPRB which are
254 only discovered by TraSig.

255 Our experiments showed that the VEGF inhibitor Axitinib, completely ablated the vascular network
256 formation as shown previously [32, 11], and appeared to completely remove CD34 expressing cells.
257 PI3K inhibition showed similar disruption of network formation, however, in contrast to Axitinib
258 treatment, rounded CD34 expressing cells remained present and evenly spaced yet completely

259 disconnected (Figure S3b). MDK inhibition appeared to decrease branching and connectivity of
260 CD34 expressing cells significantly, however these cells still maintained a spread morphology. MDK
261 is a pleiotropic growth factor that can induce cell proliferation, migration as well as angiogenesis
262 [33, 34, 35]. It has been suggested that MDK from mesothelial cells can participate in liver
263 organogenesis [36]. While its role was suggested in cancer related angiogenesis [37, 21], less is known
264 about its function in liver development. Our combined computational and experimental analysis
265 suggests such role for MDK in vascular development in human livers.

266 Interestingly, inhibition of NOTCH resulted in increased endothelial cell numbers and vascular
267 formation. Vascularization can enable better engraftment *in vivo*. Hence modulation of notch
268 signaling might be a possible target to improve liver organoid implantation *in vivo* that warrants
269 further investigation. The mechanisms of these findings can be further investigated via cell type
270 specific genetic circuits to determine dose, timing and cell types involved. Combined, our data
271 confirms that significant signaling pathways in the liver organoids could be predicted using TraSig
272 and functionally validated.

273 We have also tested TraSig on neuron and oligodendrocyte differentiation datasets. As we have
274 shown, TraSig was able to correctly identify known and novel interacting cell types pairs for these
275 datasets as well. For the first two datasets we studied, we used CSHMM for the pseudotime inference
276 while for the oligodendrocytes, we applied TraSig to the pseudotime inferred by Slingshot [9]. This
277 demonstrates the generalizability of TraSig which can be applied to output data from any pseudotime
278 ordering method. As we have shown, the ability to identify significant interactions is independent of
279 the ordering method itself enabling the use of TraSig in post-processing of any pseudotime ordered
280 scRNA-Seq data.

281 Methods

282 To identify interacting cell types pairs, we developed TraSig (**T**rajectory based **S**ignaling genes
283 inference), which infers key genes involved in cell-cell interactions. We primarily focus on genes
284 encoding ligands and receptors at this stage but our method can accommodate other proteins likely
285 to interact. For any two groups of cells that are expected to overlap in time, TraSig takes the
286 pseudo-time ordering for each group and the expression of genes along the trajectory as input and
287 then outputs an interaction score and p-value for each possible ligand-receptor pair.

288 Learning trajectories for time series scRNA-Seq data

289 There have been several methods developed to infer trajectories from time series scRNA-Seq data [4].
290 Several of these methods first reduce the dimension of the data and then infer trajectory structures
291 by using minimum spanning trees in the reduced dimension space [4]. While such methods work
292 well for obtaining global ordering and for groupings cells, they may not be as accurate for the exact
293 ordering of cells in the same edge (cluster), especially for clusters with small number of cells. Since
294 the ordering is only based on the low dimension representation, genes that are only active in a small
295 number of cells may have little impact on the representation of the cell in the lower dimension [10].
296 Since such ordering is critical for the ability to infer the activation or repression of individual genes
297 along the pseudotime, we instead use another method for trajectory inference which works in the
298 original gene space. This method, termed CSHMM, uses probabilistic graphical models to learn
299 trajectories and to assign cells to specific points along the trajectories. CSHMM (Continuous-state
300 Hidden Markov Model) [10] learns a generative model on the expression data using transition states
301 and emission probabilities. CSHMM assumes a tree structure for the trajectory and assigns cells to
302 specific locations on its edges. This enables both, the inference of the gene expression trajectories
303 for each edge and the determination of overlapping edges (in time) which are potential interacting
304 groups. In CSHMM, the expression of a gene j in cell i assigned to state $s_{p,t}$ is modeled as

$$x_j^i \sim \mathcal{N}(\mu_{s_{p,t}}, \sigma_j^2)$$

305 , where $s_{p,t}$ is determined by both the edge p and the specific location t on the edge the cell is
306 assigned to, and

$$\mu_{s_{p,t}} = g_{aj} \exp(-K_{p,j}t) + g_{bj}(1 - \exp(-K_{p,j}t)).$$

307 g_{aj} and g_{bj} are the mean expressions for gene j at branching node a and b (the beginning and the
308 end of edge p , respectively) and $K_{p,j}$ is the rate of change for gene j on edge p . σ_j^2 is the variance of
309 gene j . CSHMM is learned by using an initial assignment based on clustering single cells and then
310 iteratively refining the model and assignment using an EM algorithm [10].

311 Selecting paired clusters

312 While most current methods look at all possible cluster pairs when searching for interactions, when
313 using time series data we can constrain the search space and reduce false positives. Specifically, cells
314 can only interact if both are active at the same time. For example, predicting interactions between
315 clusters representing cells in day 1 and day 30 in a developmental study is unlikely to lead to real
316 signaling interactions. TraSig can either use the time in which cells were profiled for this or it can
317 use the tree structure provided by CSHMM to match edges based on their predicted pseudotime.
318 Interactions are only predicted for pairs of edges (clusters) representing overlapping time.

319 Ordering cells and inferring expression profiles

320 Given two groups of cells (cells assigned to two edges in the model) selected as discussed above, we
321 first obtain a smooth expression profile for each gene along each of the edges. For this we first divide
322 each edge into 101 equal size bins. We then use a sliding window approach that summarizes expression
323 levels for genes along overlapping windows of equal size. We tested window sizes comprising of
324 $L = \{5, 10, 20, \text{ and } 30\}$ bins and found that window size of 20 works best (Supplementary Notes).
325 Windows overlap by $L - 1$ bins so the first $L - 1$ bins of a window are the last $L - 1$ bins of its
326 predecessor. Since most cells are usually assigned to locations that are near the branching nodes
327 (start and end of the edges, Figure 3a), we use $L/2$ as the length of the first sliding window and
328 then increase to L when we reach the first L bins (Figure 2). We next generate an expression profile
329 for each gene using its mean expression within each window. Using overlapping intervals allows
330 us to overcome issues related to dropout and noise while still obtaining an accurate profile of the
331 expression of the gene along the edge.

332 Computing interaction scores for ligands and receptors

333 We used genes determined to be ligands or receptors from Ramilowski et al [38]. This database
334 consists of 708 ligands, and 691 receptors with 2,557 known ligand-receptor interactions. To calculate
335 an interaction score between a ligand in group A (sender) and its corresponding receptor in group
336 B (receiver), we use the expression profile for each edge calculated as discussed above. Denote the
337 expression values of the ligand in group A as $\mathbf{x} = (x_1, x_2, \dots, x_M)$ and those for the receptor in group
338 B as $\mathbf{y} = (y_1, y_2, \dots, y_M)$, where M is the total number of overlapping intervals. We use the dot
339 product function to compute the score by calculating $\mathbf{x}^T \mathbf{y} = \sum_i^M x_i y_i$. The advantage of using dot
340 product for such analysis is that it enables the use of both the magnitude and the similarity of
341 expression's change over time to rank the top pairs.

342 To compute a p-value for the score, we use randomization analysis. Specifically, we permute the
343 assignment of cells to edges and pseudotime in the model and re-compute the score as discussed
344 above for the same pair of genes along the two clusters. Such permutation allows the method to
345 identify both time dependent interactions and cluster (or cell type) specific interactions since genes
346 that are active in most of the clusters will likely be also ranked high when permuting assignments
347 between the clusters. We perform 100,000 permutations leading to a minimum p-value of 0.00001.
348 We use Benjamini-Hochberg to control the false discovery rate (FDR) at 0.05 for multiple testing
349 correction. For each pair of clusters, we also provide a summary score over all ligand-receptor pairs
350 by counting how many ligand-receptor pairs are significant for this cluster pair.

351 Using trajectories inferred by other methods

352 While we mainly discuss the use of TraSig with CSHMM, as we show in Results, it can be used with
353 the output of any other trajectory inference tool. For this TraSig uses dynverse [4], which provides
354 an R package that transforms the output of several popular trajectory inference and pseudotime
355 ordering methods to a common output. Specifically, TraSig uses the “milestone_progression” output
356 from dynverse which represents the location of a cell on an edge. This is a value in [0, 1] which we
357 use to determine the pseudo-time assignment for each cell on an edge. All other steps are the same
358 as when using CSHMM’s trajectory output.

359 Assessment of cell-cell interaction to probe vascular formation in liver organoids

360 For evaluation of whole culture vascular network formation, liver organoids were cultured on 8
361 mm glass coverslips in a 48 well plate [11]. On day 9 of culture, indicated inhibitors 50 ng/mL
362 Axitinib (Sigma, Cat PZ0193-5MG), 15 uM WZ811 (Cayman, Cat 13639), 10 uM DAPT (Stem Cell
363 Technologies, Cat 082), 10 uM LY294002 (Stem Cell Technologies, Cat 72152), 1 uM iMDK (Millipore,
364 Cat 5.08052.0001), or vehicle control (DMSO, Sigma, Cat D2650-100mL) were supplemented to the
365 culture medium daily for 5 days. After fixation with 4% PFA for 20 minutes at room temperature
366 on day 14, the cultures were washed 3x in PBS and stained as explained previously [11] with
367 CD34 antibody (Abcam, Cat ab81289) and the whole coverslip was imaged using an EVOS M7000.
368 Raw images were exported to ImageJ and applied a threshold to generate binary images of the
369 CD34+ vasculature networks. Four 1200 pixel (2-3 mm) diameter circular areas were selected per
370 coverslip for assessment in AngioTool (<https://ccrod.cancer.gov/confluence/display/ROB2>) [39]. For
371 evaluation of CXCL12 and VEGF localized vascular network formation, liver organoid cultures were
372 fixed on day 14 and stained for CD34 along with either CXCL12 or VEGF. Loci (with diameter of
373 300 pixels) with high and low relative CXCL12 or VEGF expression, determined by fluorescence,
374 were selected and vascular network was analyzed using AngioTool.

375 Data availability

376 Single cell data for the liver organoid is available from the Gene Expression Omnibus (GEO)
377 under accession number GSE159491. Single cell data for neocortical development [25] is available
378 from the Gene Expression Omnibus (GEO) under accession number GSE123335. Single cell data
379 for oligodendrocyte differentiation [28, 4] is downloaded from <https://doi.org/10.5281/zenodo.1443566>.
380

381 Code availability

382 TraSig is implemented in Python and is available at <https://github.com/doraadong/TraSig>.

383 Acknowledgements

384 Work was partially supported by NIH grants 1R01GM122096 and OT2OD026682 and by a C3.ai
385 DTI Research Award to ZB-J. M.R.E. is supported by NIH grants EB028532, HL141805 and
386 P30DK120531. J.H. is supported by the CATER Predoctoral Fellowship (NIBIB T32 EB001026).
387 Figure 4A was created with Biorender.com.

388 Author contributions

389 D.L., J.D., Z.B.-J. designed the research; D.L., J.D., Z.B.-J. developed the method; D.L. implemented
390 the software; All authors analyzed the method outputs to select validation experiments. J.J.V., J.H.
391 and M.R.E. designed and performed the validation experiments; D.L. and J.J.V. performed the
392 analysis of validation data; All authors wrote the manuscript.

393 Conflict of interest

394 M.R.E and J.J.V. have a patent (WO2019237124) for the organoid technology used in this publication.

395 References

- 396 [1] Lin, C., Ding, J. & Bar-Joseph, Z. Inferring tf activation order in time series scRNA-seq studies.
397 *PLoS Comput. Biol.* **16**, e1007644 (2020).
- 398 [2] Hurley, K. *et al.* Reconstructed single-cell fate trajectories define lineage plasticity windows
399 during differentiation of human pSC-derived distal lung progenitors. *Cell Stem Cell* **26**, 593–608
400 (2020).
- 401 [3] Abdelaal, T. *et al.* A comparison of automatic cell identification methods for single-cell RNA
402 sequencing data. *Genome Biol.* **20**, 1–19 (2019).
- 403 [4] Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory
404 inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
- 405 [5] Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. Benchmarking algorithms
406 for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* **17**,
407 147–154 (2020).
- 408 [6] Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell–cell interactions and
409 communication from gene expression. *Nat. Rev. Genet.* 1–18 (2020).
- 410 [7] Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. Cellphonedb: inferring
411 cell–cell communication from combined expression of multi-subunit ligand–receptor complexes.
412 *Nat. Protoc.* **15**, 1484–1506 (2020).
- 413 [8] Cabello-Aguilar, S. *et al.* Singlecellsignalr: inference of intercellular networks from single-cell
414 transcriptomics. *Nucleic Acids Res.* **48**, e55–e55 (2020).
- 415 [9] Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics.
416 *BMC Genomics* **19**, 477 (2018).
- 417 [10] Lin, C. & Bar-Joseph, Z. Continuous-state hmms for modeling time-series single-cell RNA-seq
418 data. *Bioinformatics* **35**, 4707–4715 (2019).
- 419 [11] Velazquez, J. J. *et al.* Gene regulatory network analysis and engineering directs development
420 and vascularization of multilineage human liver organoids. *Cell Syst.* **12**, 41–55 (2021).
- 421 [12] Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
- 422 [13] Si-Tayeb, K., Lemaigre, F. P. & Duncan, S. A. Organogenesis and development of the liver.
423 *Dev. Cell* **18**, 175–189 (2010).
- 424 [14] Gouysse, G. *et al.* Relationship between vascular development and vascular differentiation
425 during liver organogenesis in humans. *J. Hepatol.* **37**, 730–740 (2002).
- 426 [15] Walter, T. J., Cast, A. E., Huppert, K. A. & Huppert, S. S. Epithelial vegf signaling is required
427 in the mouse liver for proper sinusoid endothelial cell identity and hepatocyte zonation in vivo.
428 *Am. J. Physiol. Gastrointest. Liver Physiol.* **306**, G849–G862 (2014).
- 429 [16] Carpenter, B. *et al.* Vegf is crucial for the hepatic vascular development required for lipoprotein
430 uptake. *Development* **132**, 3293–3303 (2005).

- 431 [17] Blanco, R. & Gerhardt, H. Vegf and notch in tip and stalk cell selection. *Cold Spring Harb.*
432 *Perspect. Med.* **3**, a006569 (2013).
- 433 [18] Chen, L. *et al.* Delta-like ligand 4/dll4 regulates the capillarization of liver sinusoidal endothelial
434 cell and liver fibrogenesis. *Biochim. Biophys. Acta, Mol. Cell Res.* **1866**, 1663–1675 (2019).
- 435 [19] Kanda, S., Mochizuki, Y. & Kanetake, H. Stromal cell-derived factor-1alpha induces tube-like
436 structure formation of endothelial cells through phosphoinositide 3-kinase. *J. Biol. Chem.* **278**,
437 257–262 (2003). URL <https://doi.org/10.1074/jbc.M204771200>.
- 438 [20] Maeda, N. *et al.* A receptor-like protein-tyrosine phosphatase ptpzeta/rptpbeta binds a heparin-
439 binding growth factor midkine. involvement of arginine 78 of midkine in the high affinity binding
440 to ptpzeta. *J. Biol. Chem.* **274**, 12474–12479 (1999). URL <https://doi.org/10.1074/jbc.274.18.12474>.
- 442 [21] Filippou, P. S., Karagiannis, G. S. & Constantinidou, A. Midkine (mdk) growth factor: a key
443 player in cancer progression and a promising therapeutic target. *Oncogene* **39**, 2040–2054
444 (2020). URL <https://doi.org/10.1038/s41388-019-1124-8>.
- 445 [22] Park, M.-H. *et al.* Ccn1 interlinks integrin and hippo pathway to autoregulate tip cell activity.
446 *eLife* **8** (2019). URL <https://europepmc.org/articles/PMC6726423>.
- 447 [23] Zhang, M. *et al.* Cxcl12 enhances angiogenesis through cxcr7 activation in human umbilical
448 vein endothelial cells. *Sci. Rep.* **7**, 8289 (2017). URL <https://europepmc.org/articles/PMC5557870>.
- 450 [24] Ding, B.-S. *et al.* Divergent angiocrine signals from vascular niche balance liver regeneration and
451 fibrosis. *Nature* **505**, 97–102 (2014). URL <https://europepmc.org/articles/PMC4142699>.
- 452 [25] Loo, L. *et al.* Single-cell transcriptomic analysis of mouse neocortical development. *Nat.*
453 *Commun.* **10**, 1–11 (2019).
- 454 [26] Barry, D. S., Pakan, J. M. & McDermott, K. W. Radial glial cells: key organisers in cns
455 development. *Int. J. Biochem. Cell Biol.* **46**, 76–79 (2014).
- 456 [27] Sild, M. & Ruthazer, E. S. Radial glia: Progenitor, pathway, and partner. *Neuroscientist*
457 **17**, 288–302 (2011). URL <https://doi.org/10.1177/1073858410385870>. PMID: 21558559,
458 <https://doi.org/10.1177/1073858410385870>.
- 459 [28] Marques, S. *et al.* Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous
460 system. *Science* **352**, 1326–1329 (2016).
- 461 [29] Fruttiger, M. *et al.* Defective oligodendrocyte development and severe hypomyelination in
462 pdgf-a knockout mice. *Development* **126**, 457–467 (1999).
- 463 [30] See, J. *et al.* Oligodendrocyte maturation is inhibited by bone morphogenetic protein. *Mol.*
464 *Cell. Neurosci.* **26**, 481–492 (2004).
- 465 [31] Tanga, N. *et al.* The ptn-ptprz signal activates the afap1l2-dependent pi3k-akt pathway for
466 oligodendrocyte differentiation: Targeted inactivation of ptprz activity in mice. *Glia* **67**, 967–984
467 (2019).

- 468 [32] Guye, P. *et al.* Genetically engineering self-organization of human pluripotent stem cells into a
469 liver bud-like tissue using gata6. *Nat. Commun.* **7**, 10243 (2016). URL <https://europepmc.org/articles/PMC4729822>.
- 470
- 471 [33] Ang, N. B. *et al.* Midkine-a functions as a universal regulator of proliferation during epimorphic
472 regeneration in adult zebrafish. *PLoS One* **15**, e0232308 (2020). URL <https://europepmc.org/articles/PMC7292404>.
- 473
- 474 [34] Qi, M. *et al.* Haptotactic migration induced by midkine. involvement of protein-tyrosine
475 phosphatase zeta. mitogen-activated protein kinase, and phosphatidylinositol 3-kinase. *J. Biol.*
476 *Chem.* **276**, 15868—15875 (2001). URL <https://doi.org/10.1074/jbc.m005911200>.
- 477
- 478 [35] Weckbach, L. T. *et al.* Midkine acts as proangiogenic cytokine in hypoxia-induced angiogenesis.
479 *Am. J. Physiol. Heart Circ. Physiol.* **303**, H429—38 (2012). URL <https://doi.org/10.1152/ajpheart.00934.2011>.
- 480
- 481 [36] Onitsuka, I., Tanaka, M. & Miyajima, A. Characterization and functional analyses of hepatic
482 mesothelial cells in mouse liver development. *Gastroenterology* **138**, 1525—35, 1535.e1—6
483 (2010). URL <https://doi.org/10.1053/j.gastro.2009.12.059>.
- 484
- 485 [37] Shin, D. H. *et al.* Midkine is a potential therapeutic target of tumorigenesis, angiogenesis, and
486 metastasis in non-small cell lung cancer. *Cancers* **12** (2020). URL <https://europepmc.org/articles/PMC7563676>.
- 487
- 488 [38] Ramilowski, J. A. *et al.* A draft network of ligand–receptor-mediated multicellular signalling in
489 human. *Nat. Commun.* **6**, 1–12 (2015).
- 490
- 491 [39] Zudaire, E., Gambardella, L., Kurcz, C. & Vermeren, S. A computational tool for quantitative
492 analysis of vascular networks. *PLoS One* **6**, e27385 (2011).
- 493

490 **Figures**

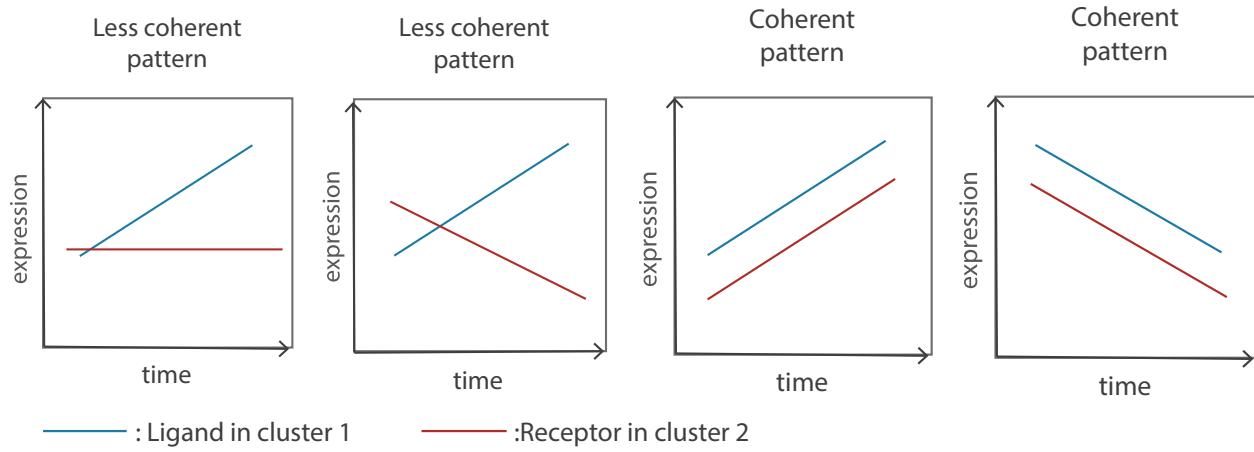


Figure 1: Example cases where the *average* expressions of the ligand and receptor that are known to interact are the same. Of these four figures only the last two represent correlated activation and repression of these proteins. Methods that only use the average expression of genes in clusters cannot differentiate between these 4 profiles and so will score all of them the same.

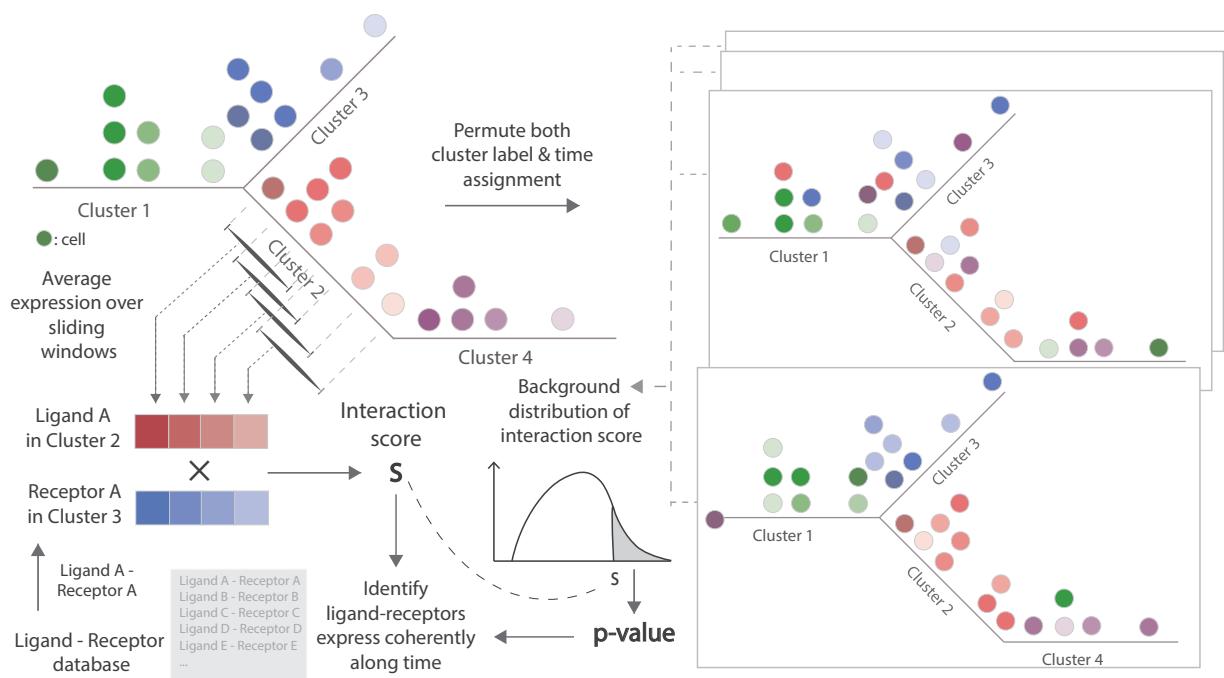


Figure 2: TraSig workflow. Top Left: For a time series scRNA-seq dataset, we use the reconstructed pseudotime, trajectory and the expression data as inputs. Bottom Left: We next determine expression profiles for genes along each of the edges (clusters) using sliding windows and compute dot product scores for pairs of genes in edges. Right: Finally, we use permutation tests to assign significance levels to the scores we computed.

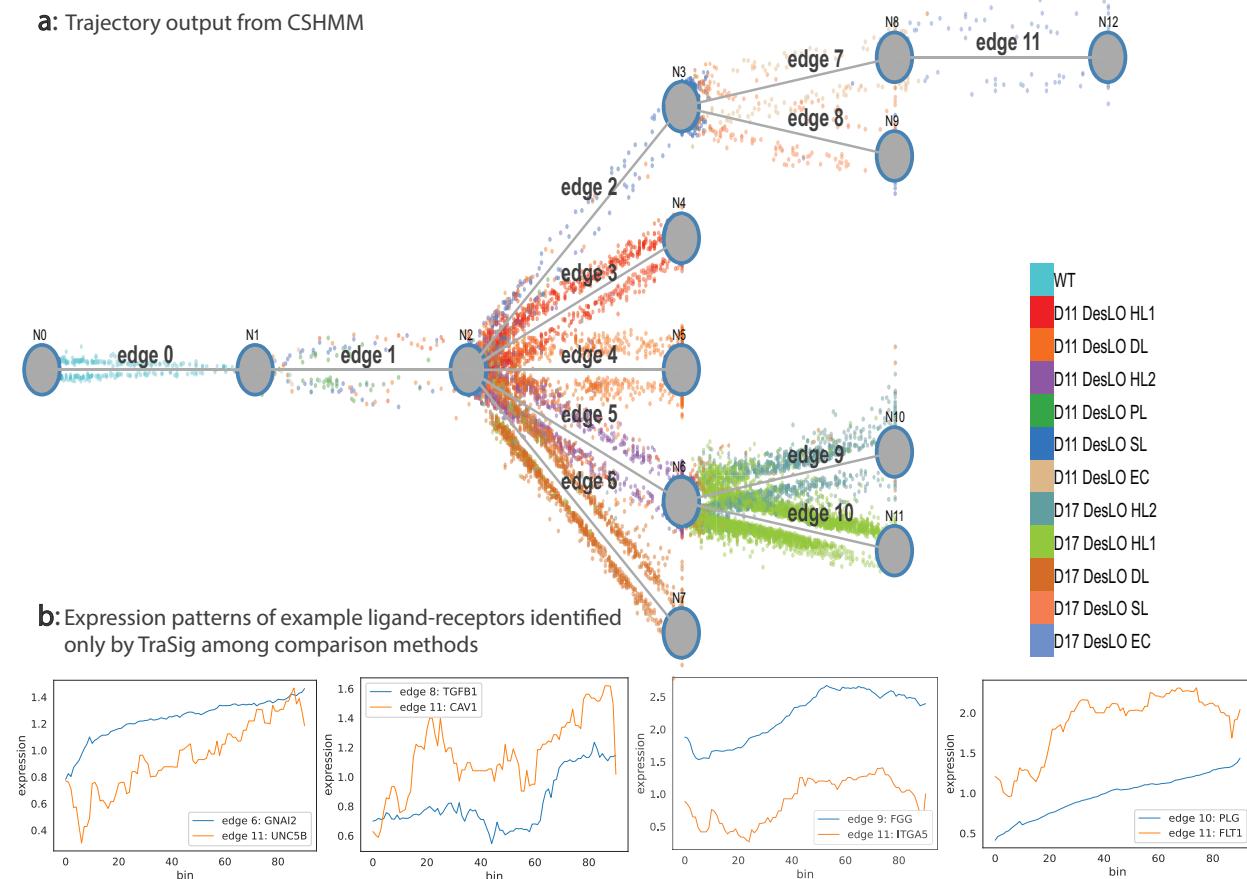


Figure 3: CSHMM results and expression patterns of identified ligand-receptor pairs. Top: Reconstructed trajectory for liver organoid differentiation. CSHMM identifies a tree-structured trajectory that clusters cells to edges based on their expression pattern and relationship to the expression patterns of prior edges (Methods). Cells are colored by their cell type labels and are shown as dots ordered by their pseudo-time assignment. DesLO - designer liver organoid; HL - hepatocyte-like cells; DL - ductal/cholangiocyte-like cells; SL - stellate-like cells; EC - endothelial-like cells; PL - progenitor-like cells; WT - wild type. Bottom: Sliding window expression for four example ligand-receptor pairs predicted to interact by TraSig.

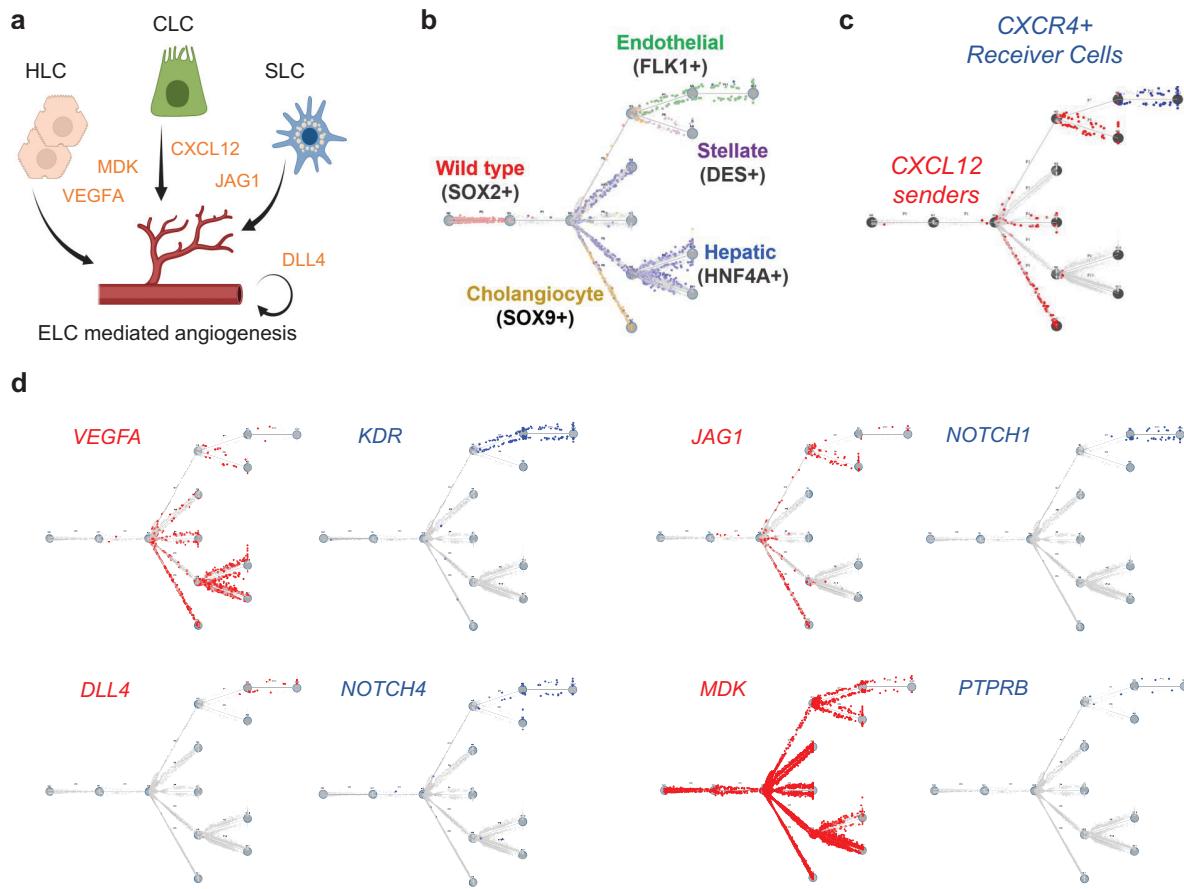


Figure 4: Ligand-receptor interaction predictions from TraSig of interest for functional studies. (a) Cartoon of cell signaling interaction between different DesLO cell types (HLC, hepatocyte-like cells; CLC, cholangiocyte-like cells; SLC, stellate-like cells; ELC, endothelial-like cells) (b) Trajectory plot showing cell type assignments with key identifying genes highlighted by different colors (Red = SOX2+ non induced cells, Yellow = SOX9 cholangiocyte-like cells, Blue = Hepatocyte-like cells, Purple = Stellate-like cells, Green = Endothelial-like cells). (c) Sender CXCL12 cells from the Cholangiocyte and Stellate populations in red shown with the receiver CXCR4 expressing endothelial cell population in blue. (d) Sender and receiver signaling populations (red = senders/ligands; blue = receivers/receptors)

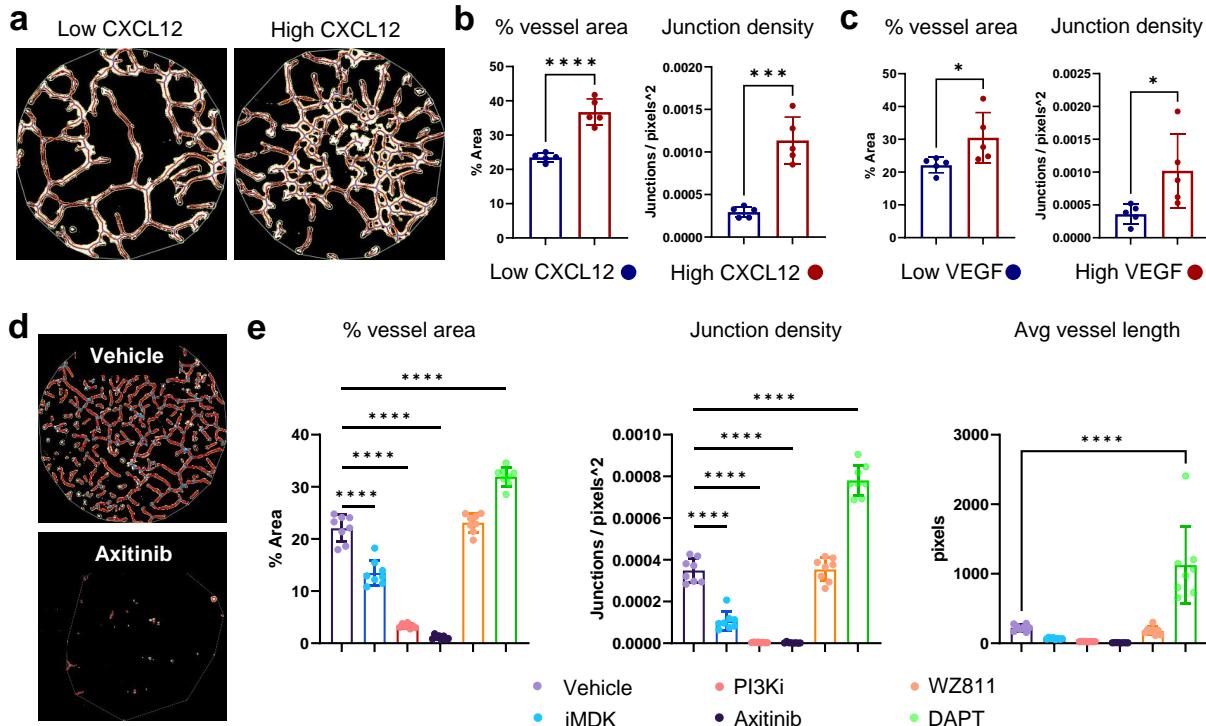
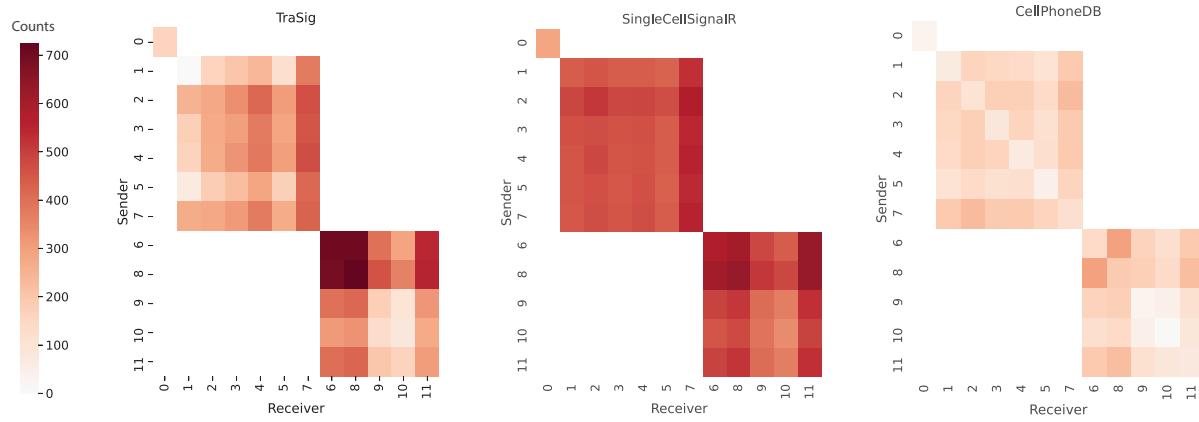
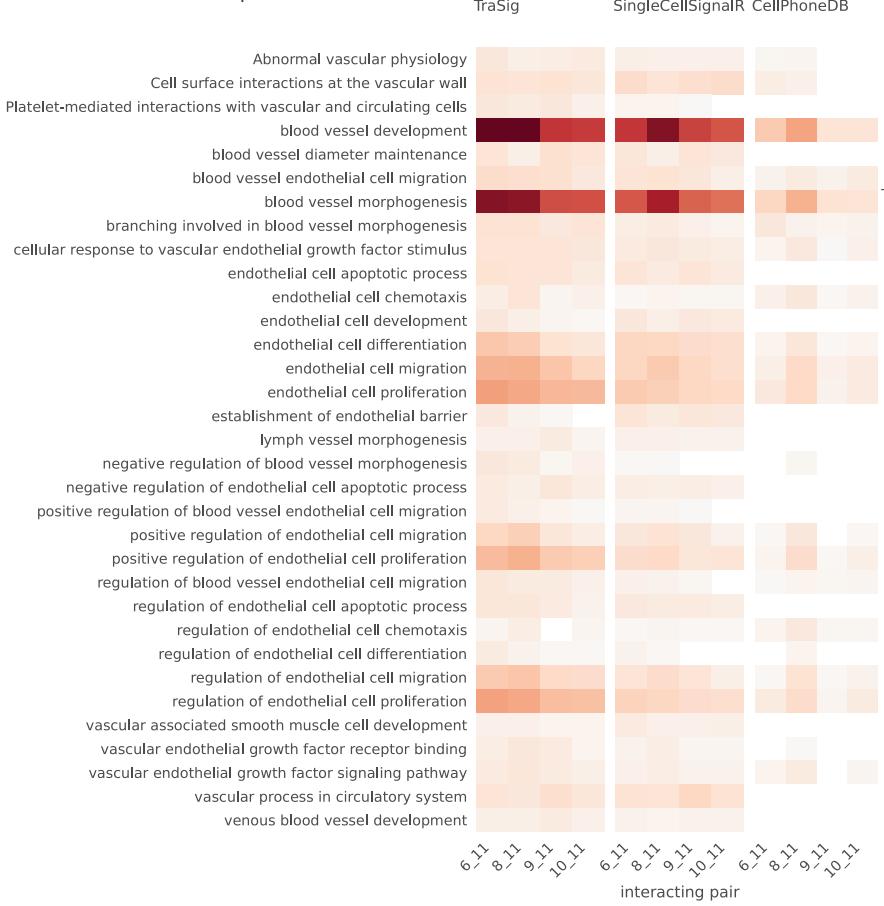


Figure 5: Functional validation of TraSig ligand-receptor signaling predictions. (a) Example of AngioTool analysis of CD34 vascular network at low vs high CXCL12 loci. (b) Percent vessel area and junction density measured at CXCL12 and (c) VEGF low vs high loci from day 14 liver organoid cultures using AngioTool. n=4 loci for high CXCL12/VEGF expression and n=4 loci for low CXCL12/VEGF on one coverslip per staining combination. (d) Example of AngioTool evaluation of CD34 stained liver organoid cultures from the vehicle control (top) and Axitinib (bottom) conditions. (e) Percent vessel area, junction density, and average vessel length vascular metrics determined by AngioTool analysis results of CD34 stained liver organoid cultures with different inhibitor conditions. n=2 biological replicates with 4 sampled areas per coverslip. For b and c, Unpaired two tailed t test was used, * p<0.05, **** p<0.0001. For e, ANOVA with Tukey post comparison test was used, **** p<0.0001. Data are represented as mean ± SE for b, c, and e.

a: Number of identified ligand-receptor pairs for each interacting cluster pair



b: GO terms enrichment comparison



c: overlap in identified ligands and receptors from different methods

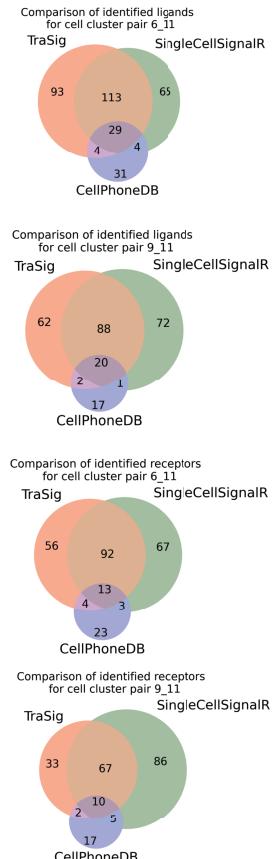


Figure 6: Results from comparing TraSig with SingleCellSignalR and CellPhoneDB. Top: Heatmaps for scores assigned by the three different methods for all cluster pairs representing cells sampled at the same time. TraSig and SingleCellSignalR identified more ligand-receptors pairs leading to higher scores. Bottom left: $-\log_{10} p$ -value for enriched GO terms related to endothelial cells and vascular development. Bottom right: Venn diagrams for the overlap in identified ligands and receptors among the three methods. The overlap between TraSig and SingleCellSignalR is high though roughly 50% of the identified proteins by each method are not identified by the other.