GENETICS

Reconstructing codependent cellular cross-talk in lung adenocarcinoma using REMI

Alice Yu[1], Yuanyuan Li[1,2], Irene Li[3], Michael G. Ozawa[4], Christine Yeh[1], Aaron E. Chiou[1,2], Winston L. Trope[5], Jonathan Taylor[6], Joseph Shrager[5]†, Sylvia K. Plevritis[1,2]*

Cellular cross-talk in tissue microenvironments is fundamental to normal and pathological biological processes. Global assessment of cell-cell interactions (CCIs) is not yet technically feasible, but computational efforts to reconstruct these interactions have been proposed. Current computational approaches that identify CCI often make the simplifying assumption that pairwise interactions are independent of one another, which can lead to reduced accuracy. We present REMI (REgularized Microenvironment Interactome), a graph-based algorithm that predicts ligand-receptor (LR) interactions by accounting for LR dependencies on high-dimensional, small–sample size datasets. We apply REMI to reconstruct the human lung adenocarcinoma (LUAD) interactome from a bulk flow-sorted RNA sequencing dataset, then leverage single-cell transcriptomics data to increase the cell type resolution and identify LR prognostic signatures among tumor-stroma-immune subpopulations. We experimentally confirmed colocalization of CTGF:LRP6 among malignant cell subtypes as an interaction predicted to be associated with LUAD progression. Our work presents a computational approach to reconstruct interactomes and identify clinically relevant CCIs.

INTRODUCTION

Cell-cell interactions between and within the various cell types comprising the tissue microenvironment play a fundamental role in regulating local and systemic biological and physiological functions under normal and pathological conditions. These interactions facilitate cooperation or competition between cell types and are typically mediated between ligands and receptors. Ligands are often manifested as soluble or extracellular proteins that are expressed by the "sending" cells and bind onto a cognate receptor on the "receiving" cells (*1*, *2*). In tumor microenvironments (TMEs), cellular cross-talk between tumor, stroma, and immune cells orchestrates the establishment of preinvasive and invasive niches that enable cancer progression properties, such as tumor growth, immune evasion, and metastasis. Large-scale cellular interactions are difficult to measure using current experimental techniques, but several computational approaches have been proposed to predict these interactions using -omics data.

A majority of the current computational approaches that use high-throughput transcriptomics data to infer cell-cell interactions either calculate interaction scores on the basis of gene expression permutation tests or implement graph-based approaches (*2*). In expression permutation-based approaches, such as CellphoneDB v2.0 and NATMI, potential ligand and receptor (LR) interactions are identified by thresholding the genes on the basis of their expression level with the assumption that this predicts higher LR protein abundance (*3*, *4*). Other methods, such as CCCExplorer and NicheNet, calculate correlation metrics between the expression levels of the ligand, receptor, or downstream signaling pathway genes for each LR pair (*5*, *6*). However, a correlation between the expression of ligand, receptor, or downstream genes may be capturing an indirect association caused by another LR interaction involving the ligand or receptor of interest.

While current cross-talk inference approaches provide a valuable baseline for synthesizing hypotheses of LR interactions, they do not capture the conditional dependencies of LR pairs among multiple cell types. Current approaches make the assumption that pairwise interactions are independent, but a pairwise interaction can be influenced by other interactions via autocrine and paracrine loops (*7*, *8*). Calculating the conditional dependency of LR pairs is often an ill-defined problem for high-dimensional datasets generated across a small sample size. This challenge is common in -omics data analysis because of the large number of parameters (p) and relatively small number of samples (n). To address this challenge, we present the algorithm, called REMI (REgularized Microenvironment Interactome), to identify communities of dependent LR pairs in high-dimensional datasets of small sample size using graph-based approaches. We demonstrate the performance of REMI by simulating datasets with varying sample sizes to show how REMI outperforms existing approaches.

To compare REMI to existing interactomes, we focused on reconstructing the lung adenocarcinoma (LUAD) interactome. Various renditions of the LUAD microenvironment have been assembled using different computational approaches that have led to novel insights. Kumar *et al.* (*9*) built an interactome for LUAD using mouse single-cell RNA sequencing (RNA-seq) (scRNA-seq) data, where they used a scoring mechanism that captured highly expressed LR genes. Gentles *et al.* (*10*) created the Lung Tumor Microenvironment Interactome (LTMI) from bulk flow-sorted RNA-seq data by thresholding gene expression levels in the dataset and computing pairwise correlations between LR genes. We applied REMI to the LTMI dataset to reconstruct a rendition of LUAD interactome (REMI-LUAD) with higher specificity. We then projected an independent LUAD scRNA-seq dataset onto REMI-LUAD to increase the cell type resolution of the interactome and referred to this

[1] Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [2] Department of Radiology, Stanford University, Stanford, CA, USA. [3] Cancer Biology Interdepartmental, Program Stanford University, Stanford, CA, USA. [4] Department of Pathology, Stanford University, Stanford, CA, USA. [5] Department of Cardiothoracic Surgery, Stanford University, Stanford, CA, USA. [6] Department of Statistics, Stanford University, Stanford, CA, USA.
*Corresponding author. Email: sylvia.plevritis@stanford.edu
†Present address: Veterans Affair Palo Alto Health Care System, 3801 Miranda Ave., Palo Alto, CA 94304, USA.

interactome as the single-cell rendition of REMI-LUAD (scREMI-LUAD). Using scREMI-LUAD, we identified paracrine interactions between cell subtypes that were previously annotated as autocrine signaling interactions. To derive a signature of LUAD progression from scREMI-LUAD, we assigned a prognostic score to each cell subtype and identified prognostically associated cross-talk signatures that may lead to clinically relevant biomarkers and therapeutic targets. In summary, REMI offers a new approach to infer cell-cell interactions between many cell types by accounting for conditional dependencies in cell-cell interactions. REMI is implemented in R and is freely available on GitHub (https://github.com/plevritis-lab/REMI).

RESULTS

Impact of correlation versus partial correlation analysis on LR pair inference

Many of the current computational approaches use correlation to identify likely LR interactions. Here, we show that correlation-based analyses can provide high false-positive predictions that are removed when computing the partial correlation. Partial correlation removes potential confounding effects from correlation values that are caused by another ligand or receptor indirectly affecting the expression levels of the LR pair of interest. To demonstrate this point, we analyzed a publicly available primary LUAD bulk flow-sorted RNA-seq dataset, which contains transcriptomic data for four main cell types: malignant, fibroblast, endothelial, and pan-immune cells (GSE111907) (n = 17 patients) (*10*). As expected, hierarchical clustering performed on only the genes that expressed ligands and receptors grouped the samples by cell type (Fig. 1A). We then constructed an LR correlation network, denoted as $G_{\mathrm{LTMI-LUAD}}$, where the nodes represent ligand or receptor genes, and an edge between an LR pair exists on the basis of the threshold criteria set by Gentles *et al.* (Fig. 1B). The edge weight is set as the Pearson correlation between the gene expression of the LR.

Because many receptors have multiple potential ligand pairings and vice versa, the correlation network is composed of a large dense network with small disjoint network components, as indicated in Fig. 1B. To infer the conditional dependency of LR pairs within the network, we calculated the partial correlation of genes on a small disjoint graph, which contains one ligand node (*TNFSF10)* and two receptor nodes (*TNFRSF10B* and *TNFRSF10D)* across multiple cell types (Fig. 1B). Although simultaneous TNFSF10:TNFRSF10B and TNFSF10:TNFRSF10D interactions could occur, one interaction plays an agonistic role, while the other plays an antagonistic role (*11*). The binding of TNFSF10 to TNFRSF10B induces cell apoptosis via the TRAIL [tumor necrosis factor (TNF)–related apoptosis-inducing ligand] pathway, whereas the binding of TNFSF10 to TNFRSF10D inhibits the TRAIL pathway (*12*). If we infer LR interactions on the basis of correlation, then the ligand TNFSF10 secreted from the malignant cell interacts with the receptor TNFRSF10B on malignant and fibroblast cells (circled in green), indicating likely apoptosis of malignant cells and cancer-associated fibroblasts (Fig. 1C, left). This contradicts studies that show protumoral effects of fibroblasts and how they can express the decoy receptor to avoid apoptosis (*13*). Instead, by calculating the partial correlation of the LR pairs within the circled component, we find more biologically reasonable results. On the basis of the partial correlation values, TNFSF10 expressed by the malignant cells is more likely to interact with TNFRSF10B expressed on the immune cells, inducing protumoral effects as described in literature (*12*, *14*). This simple analysis demonstrates how partial correlation removes confounding effects in correlation analysis that may be misleading by reducing positive edges when building the interactome (Fig. 1C, right). However, calculating the partial correlation is ill-posed on larger subnetworks because of the high dimensionality of many -omics datasets. REMI extends this concept onto larger components of the network.

REMI algorithm

We introduce our novel algorithm, REMI, which identifies communities of conditionally dependent cell-cell interactions to reduce the number of false-positive edges in LR correlation–based networks. The algorithm is applicable to high-dimensional transcriptomic datasets with small sample sizes. REMI is composed of four steps: (i) build a weighted undirected LR correlation network leveraging known LR pairings, (ii) detect communities of LR groups, (iii) identify conditionally dependent LR pairs in communities, and (iv) reconstruct the global interactome from the communities (Fig. 2A) (Materials and Methods). An additional step in REMI allows for the user to measure the significance of an LR pair prediction with respect to the LR pair's REMI community.

In its first step, REMI generates a weighted bipartite LR network, denoted by G, where the nodes represent either a ligand or a receptor gene expressed in a specified cell type within the dataset. Edges of G are drawn between literature-supported LR pairings curated from the FANTOM5 database (*15*). Edge weights of G are computed as the Pearson correlation between the gene expression of the LR nodes (Fig. 2, A and I). Ideally, the conditional dependency of each LR pair would be computed with respect to all other nodes in the network. However, current high-dimensional transcriptomic datasets contain a large number of genes compared to the number of samples, which makes the conditional dependency hard to estimate with high accuracy.

To reduce the dimensionality of the problem, REMI hierarchically divides G into groups of densely connected nodes, called communities, using the Louvain community detection method (*16*). Next, REMI identifies conditionally dependent LR pairs in each community independently. A covariance matrix between all pairs of ligands and receptors is computed for each community, setting edges between ligands and edges between receptors as zero (Fig. 2A, ii). To identify these conditionally dependent LR edges within a community, we use graphical lasso (GLasso), which estimates the inverse covariance of the community using a lasso penalty that increases the graph's sparsity (Fig. 2A, iii) (*17*). The resulting edge weights represent the $l1$-regularized partial correlation metrics of each LR pair within a community, which captures its relationship with respect to other LR pairs in the community. For further downstream analyses, we analyzed a binarized version of the network, denoted as $\widehat{\Theta}$, reconstructed by aggregating the communities after filtering for edges with a positive edge weight.

REMI's iterative community detection algorithm component allows for community sizes to change for downstream prediction. We ran several simulations to test the sensitivity and specificity of varying community sizes. REMI's best performance occurs when the number of nodes in each community is equal to the sample size (fig. S2B). Therefore, REMI hierarchically breaks down large communities until the size of the communities is approximately equivalent to sample size. Communities identified by Louvain are referred to as "within communities" (WC). This step prioritizes assigning a node

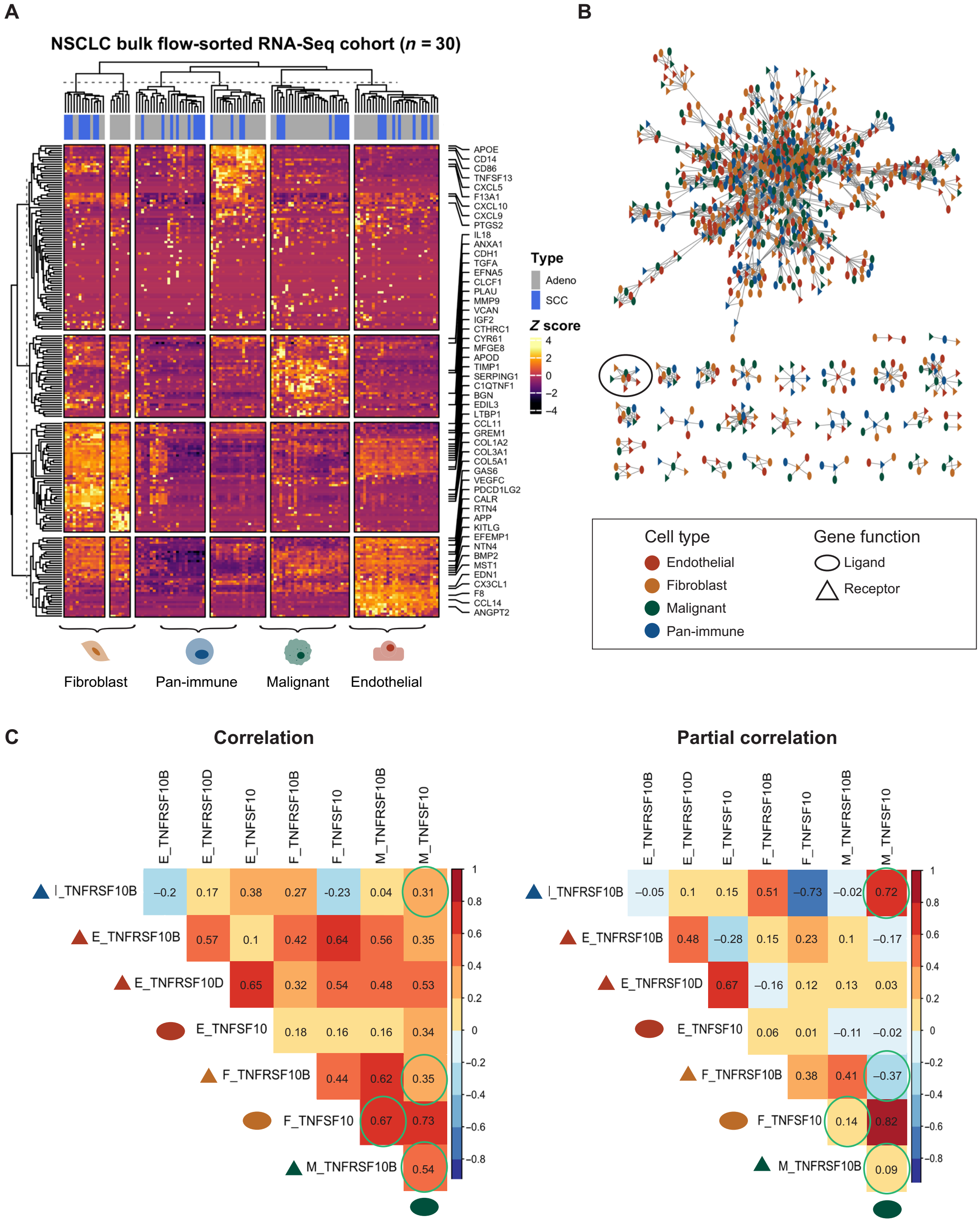
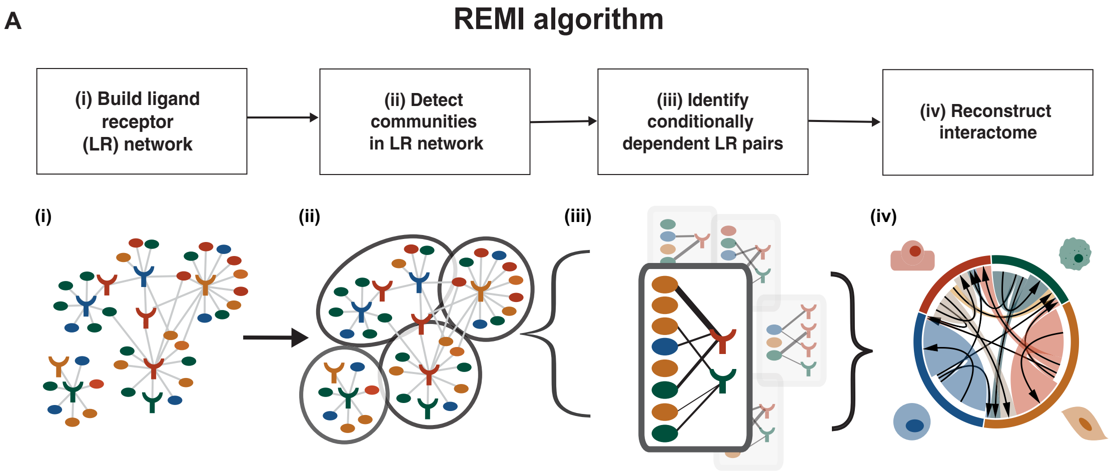


Fig. 1. Network representation of LTMI shows dense components of cognate LR pairs. (**A**) All LR genes in the LUAD flow-sorted RNA-seq dataset hierarchically clustered. Genes labeled are the most highly expressed genes in the top five percentile. (**B**) Correlation network of expressed LR pairs from the Gentles *et al.* dataset. Color of nodes represent cell type of gene. Shape of node represents type of gene. Black circle indicates network component used for correlation analysis in panel below. (**C**) Correlation and partial correlation values of the nodes that are found within the black circled component in (B). Correlation matrix values are on the left, and partial correlation matrix values are on the right. The text labels of the correlation matrices correspond to the gene's cell type added in front of the gene name (celltype_name). In addition, a colored (cell type) shape (gene function) is placed in front of the text label. Colors correspond to the legend in (B). Bottom dark green circle corresponds to M_TNFSF10 labeled at the top of the column. Green circles highlight examples of correlation values that are changed after accounting for confounding variables.


A
REMI algorithm
(i) Build ligand receptor (LR) network
(ii) Detect communities in LR network
(iii) Identify conditionally dependent LR pairs
(iv) Reconstruct interactome
(i)
(ii)
(iii)
(iv)

B

	REMI		GLasso	
n	Sensitivity	Specificity	Sensitivity	Specificity
15	0.69 (0.15)	0.52 (0.15)	0.82 (0.02)	0.29 (0.03)
25	0.73 (0.14)	0.55 (0.14)	0.82 (0.02)	0.37 (0.02)
50	0.81 (0.11)	0.54 (0.12)	0.83 (0.02)	0.49 (0.02)
100	0.87 (0.07)	0.57 (0.09)	0.86 (0.02)	0.62 (0.02)
200	0.93 (0.04)	0.65 (0.07)	0.90 (0.01)	0.74 (0.02)

C

	REMI	CellphoneDB	NicheNET
n	AUC	AUC	AUC
15	0.61 (0.06)	0.53 (0)	0.5 (3.5×10^{-3})
25	0.63 (0.075)	0.53 (0)	0.5 (3.2×10^{-3})
50	0.66 (0.1)	0.53 (0)	0.5 (3.1×10^{-3})
100	0.69 (0.12)	0.53 (0)	0.5 (2.8×10^{-3})
200	0.69 (0.13)	0.53 (0)	0.5 (2.6×10^{-3})

D

	CellphoneDB	NicheNET
n	Jaccard index	Jaccard index
15	0.05 (0.03)	0.2 (0.11)
25	0.06 (0.009)	0.24 (0.09)
50	0.04 (0.03)	0.23 (0.11)
100	0.09 (0.004)	0.29 (0.15)
200	0.09 (0.004)	0.33 (0.16)

Fig. 2. Schematic overview of REMI and simulation studies. (**A**) Diagram of REMI: (i) A network of potential LR interactions between cell types of interest is built given a reference LR database. Nodes represent either ligand or receptor genes. Edges represent known LR pairs. (ii) Network represented as density-based communities. (iii) Conditionally independent LR pairs removed: Each community is converted into a graphical Gaussian model, and conditionally independent LR pairs are removed. (iv) Conditionally dependent LR pairs across communities are reassembled into global interactome. (**B**) Simulation studies: Performance of REMI versus GLasso for sample sizes $n = 15$ to 200 when compared to a GLasso interactome derived from the entire LUAD TCGA RNA-seq dataset ($n = 1031$). Cohorts of various sample sizes were generated by downsampling TCGA data and then used to compare REMI's and GLasso's performance. At each sample size, 50 cohorts were generated. Mean AUC and SD (in parentheses) values are generated across the 50 cohorts per sample size. (**C**) Test performance of REMI compared with other computational cross-talk prediction algorithms. At each sample size, 50 cohorts were generated by randomly sampling TCGA data. Mean AUC and SD (in parentheses) are provided. (**D**) Average Jaccard indices of 50 cohorts per sample size calculated for each method compared to REMI.

to a unique community but leaves edges that represent potential LR pairs unassigned. REMI creates additional communities to account for the newly unassigned edges by aggregating unassigned edges that exist between two communities to create "between communities" (BCs) (fig. S1A). The conditionally dependent WC and BC edges are aggregated to reconstruct the interactome (Fig. 2A, iv).

To reconstruct the global interactome from the independent community calculations, we developed a method to measure the significance of each computed edge with respect to its community. The *P* value is calculated by randomly permuting the edge weight of the LR pair with randomly sampled correlation values. Because multiple correlation values can predict a binary edge in REMI, we build a decision tree to sort out which correlation values predict the edge of interest. We then sample from our null Wishart distribution, which represents a previous covariance matrix distribution, and predict edges using the decision tree. Edges predicted using values sampled from the null represent our REMI null distribution. We then calculate a one-sided *P* value by measuring the number of null edges that had a greater correlation value than the original edge weight. The one-sided *P* value is then converted to a two-sided *P* value (fig. S1B). Although the statistical test is computationally intensive, it can be used for determining the presence or absence of the edges when aggregating the communities to reconstruct the global interactome.

Testing robustness of REMI parameters via simulations

To assess REMI's performance with respect to implementing GLasso alone, we generated a population-level regularized interactome by leveraging the large publicly available Cancer Genome Atlas [The Cancer Genome Atlas (TCGA)] LUAD bulk RNA-seq dataset. We made the assumption that the 1013 patients in this dataset represent an entire population and regularized a network of LR pairs within the dataset using GLasso. Because of the large sample size, confounding effects for all possible interactions within the microenvironment are captured, and the resulting network, denoted as $\widehat{\Theta}_{\text{TCGA}}$ (Supplementary Dataset), contains these conditionally dependent LR pairs.

From the TCGA LUAD cohort, we created randomly subsampled cohorts of different sample sizes (n = 15, 25, 50, 100, and 200), generating 50 cohorts at each sample size. We ran REMI and GLasso on each downsampled cohort and compared their performance in terms of sensitivity and specificity (Fig. 2, B and C). We found that in larger sample sizes (n = 200), REMI and GLasso performed with comparable sensitivity (0.93 and 0.90, respectively) and specificity (0.65 and

0.74, respectively) (Fig. 2B). As the sample size decreased ($n = 15$), GLasso retained high sensitivity values compared with REMI. GLasso's sensitivity dropped from 0.90 to 0.83, whereas REMI's sensitivity dropped from 0.93 to 0.69. However, REMI retained higher specificity than GLasso as the sample size decreased: GLasso's specificity dropped from 0.74 to 0.29 (45% decrease), whereas REMI's specificity dropped from 0.65 to 0.52 (13% decrease). In summary, REMI has higher specificity compared with GLasso for datasets with smaller sample sizes, which we regard as a preferred property when selecting candidate-relevant interactions for experimental validation.

To further evaluate the performance of REMI, we ran the same simulations and measured the sensitivity and specificity of WC and BC REMI predictions as a function of sample size. WCs have a higher density compared with BCs and yielded relatively higher sensitivity and lower specificity as sample size decreased from $n = 200$ to $n = 15$. In WCs, sensitivity dropped from 0.93 to 0.75 (18% decrease), and specificity dropped from 0.63 to 0.45 (18% decrease). On the other hand, BCs performed with relatively lower sensitivity and higher specificity as sample size decreased from $n = 200$ to $n = 15$. In BCs, sensitivity dropped from 0.88 to 0.45 (43%), and specificity stayed relatively the same from 0.74 to 0.76 (2% increase) (fig. S2A). Next, we considered the effect of the community size on REMI's performance. REMI's community sizes are set to be equal to the sample size of the dataset. When the number of nodes within a community was increased from $2n$ to $50n$ times, the density of the communities increased and specificity decreased from 0.52 to 0.33 (19%), whereas sensitivity increased from 0.66 to 0.78 (12%) (fig. S2B). Community sizes closer to the sample size of the dataset have increased specificity as opposed to larger community sizes, thereby reinforcing REMI's optimization for specificity.

Next, we compared REMI to current cell-cell interaction computational inference approaches designed for bulk and scRNA-seq. We selected methods that represent the expression permutation-based approaches (CellphoneDB v2.0) and network-based approaches (NicheNet) and ran the methods on the same simulation (Fig. 2C). REMI's performance ranged from an area under the curve (AUC) of 0.61 to 0.69 as sample size increased. CellphoneDB v2.0 performed with an AUC of 0.53, and NicheNet performed with an AUC of 0.5 across all cohort sizes. Both methods' performance did not vary as size increased, suggesting that these approaches are more biased toward the reference LR network and do not readily adapt to the variations seen in the data. We also calculated the Jaccard index of NicheNet and CellphoneDB v2.0's results compared to REMI and found both to be low because of the high number of predicted interactions in NicheNet and CellphoneDB v2.0. NicheNet's regression-based approach had higher Jaccard indices (0.20 to 0.33) as opposed to CellphoneDB v2.0 (0.05 to 0.09), showing more similarity between NicheNet and REMI rather than CellphoneDB v2.0 and REMI.

Reconstructing LUAD interactome using REMI

We applied REMI to the bulk flow-sorted LUAD RNA-seq cohort to reconstruct the REMI-LUAD interactome, with the graph denoted as $G_{\mathrm{REMI-LUAD}}$. Compared with $\mathrm{G}_{\mathrm{LTMI-LUAD}}$, $G_{\mathrm{REMI-LUAD}}$ had 16% fewer nodes ($V = 868$ in LTMI and $V = 727$ in REMI) and 45% fewer edges ($E = 2652$ in LTMI and $E = 1435$ in REMI), highlighting the regularization underlying REMI's approach (Fig. 3A and Supplementary

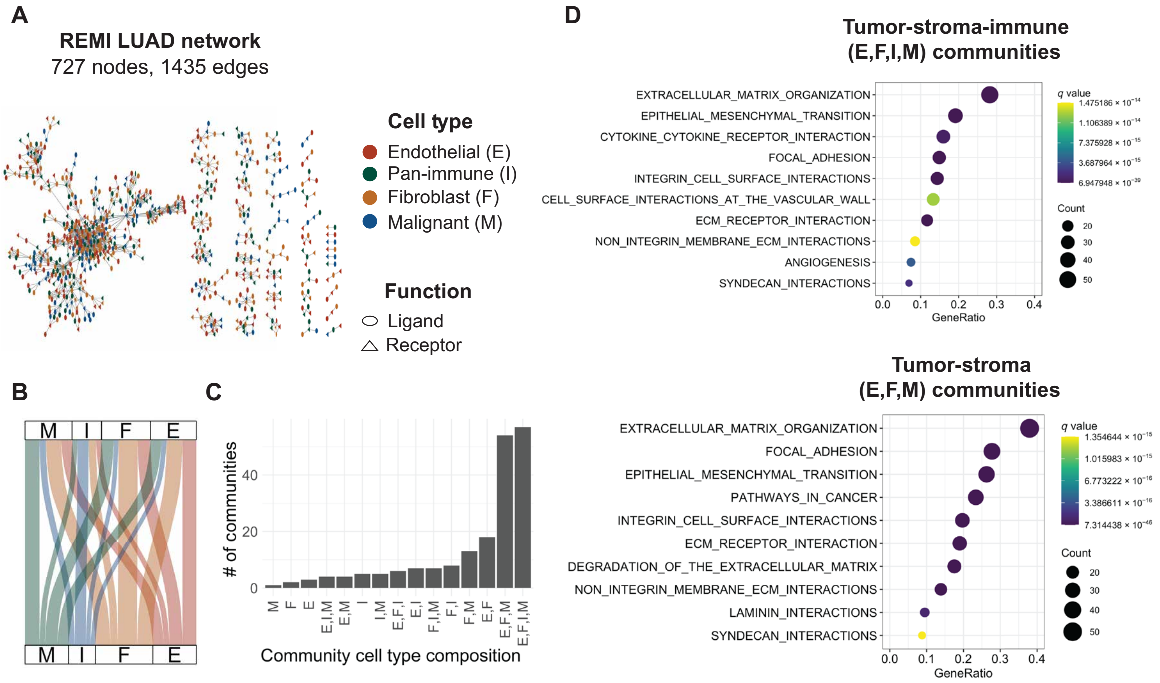


Fig. 3. Reconstructed LUAD TME interactome reveals tumor-stroma–specific communities. (**A**) Force-directed layout of the REMI-LUAD interactome network. Shape of the nodes represent gene function. Color of the node represents cell type as denoted by the legend. (**B**) Alluvial plot showing the ratio of paracrine and autocrine signaling interactions occurring between cell types. Thickness of lines represent the number of LR pairs. Full table is available in the Supplementary Materials. (**C**) Distribution of number of communities with certain cell type compositions. (**D**) Top 10 most significant enrichments for genes within the tumor-immune-stroma–specific communities and tumor-stroma–specific communities. GeneRatio is the ratio of the number of community-specific genes that overlapped with the gene set to the number of community-specific genes that overlapped with all the gene sets in the collection. Count represents the number of genes.

Dataset). REMI-LUAD removed LR edges that were densely connected (average node degree of edges removed = 12) and also identified a few new LR pairs (fig. S1C). In REMI-LUAD, known lung cancer–specific LR pairs were identified, including *GREM1:KDR*, which was experimentally validated between fibroblast and malignant cells (*10*). We observed fewer immune interactions within REMI-LUAD than within LTMI-LUAD. We suspect that this is a consequence of the greater heterogeneity among the pan-immune cells relative to the other three tumor-stroma cell types in the bulk-sorted samples, as demonstrated in published scRNA-seq studies (*18*). Grouping the diverse immune-suppressive and immune-reactive TME immune cells into a pan-immune group reduced their overall effect in REMI. We are reassured by this finding because it would have been challenging to interpret interactions of pan-immune cells and other cell types that included both tumor-promoting and tumor-suppressing features.

To understand which cell types are interacting in REMI-LUAD, we analyzed the cell type composition within each community. Majority of the communities were either represented by all cell types (E, F, I, and M) (36%) or tumor-stroma cell types (E, F, and M) (25%) (Fig. 3C). To assess the phenotypic properties enriched in these immune-stroma and stroma communities, we performed Gene Set Enrichment Analysis (GSEA) on the LR genes in each community separately. Communities with tumor-immune-stroma interactions were enriched for cytokines, angiogenesis, extracellular matrix (ECM) organization, and leukocyte extravasation. Angiogenesis and ECM remodeling are crucial for building a niche architecture for tumor growth. Communities with tumor-stroma–specific interactions were uniquely enriched for ECM degradation (Fig. 3D). ECM stiffness is associated with increased immunosuppression and found to stimulate epithelial-to-mesenchymal transformation in the tumor, whereas ECM degradation aids in creating paths for cancer cell migration (*19*). Integrins found within REMI-LUAD, in particular, can detect ECM mechanical stiffness and assist with cell migration through degraded areas (*20*). Together, our results suggest that REMI-LUAD captures cell-cell interactions associated with processes ranging from tumor growth to invasion.

Increasing cell type resolution of REMI-LUAD using a scRNA-seq LUAD dataset

To increase the cell type resolution of the REMI-LUAD, we analyzed REMI-LUAD using a publicly available independent scRNA-seq LUAD dataset from Lambrecht *et al.* ($n = 2$ patients; 22,681 total cells) (E-MTAB-6149) (*21*). We reclustered them individually using Louvain clustering and applied the annotated cell type labels from the manuscript (Fig. 4A). We labeled the immune clusters with broad immune cell subtypes as denoted from the manuscript (myeloid, T cells, and B cells). For the remaining cell types, we identified five malignant [denoted as M(1), M(2), M(3), M(4), and M(5)], three endothelial [denoted as E(1), E(2), and E(3)], and four fibroblast subtypes [denoted as F(1), F(2), F(3), and F(4)] (Fig. 4, B to D, and fig. S3) (*22*).

To adapt REMI for scRNA-seq data, we averaged each gene's expression across each cell subtype per patient. The Lambrecht *et al.* dataset consists of only two patients with LUAD, which does not provide enough power for REMI's inverse covariance–based calculations. For this reason, we projected the scRNA-seq cell types onto $G_{\text{REMI–LUAD}}$ and relabeled the nodes' cell types on the basis of the expression levels in the single-cell dataset. This network is referred to as single-cell REMI-LUAD (scREMI-LUAD). For each subpopulation, we filtered for differentially expressed (DE) LR genes that had an averaged expression level greater than 0.4 (Fig. 4, A to C). Sixty-five percent of the interactions in REMI-LUAD were present in scREMI-LUAD, confirming the high expression of these LRs in two independent patients. Because many LR genes appear in multiple cell subtypes, the number of interactions in scREMI-LUAD is greater than in REMI-LUAD, providing potential insight into the extent of heterogeneity of the interactome. For example, REMI-LUAD inferred that fibroblasts secrete *CYR61*, which interacts with *ITGB5* on fibroblasts, suggestive of autocrine signaling. In scREMI-LUAD, *CYR61* is expressed in F(2) and *ITGB5* in F(4). Hence, the REMI-LUAD *CYR61:ITGB5* autocrine fibroblast interaction is now redefined as a paracrine signaling between different fibroblast cell subtypes in scREMI-LUAD (Fig. 4E). Notably, 86% autocrine signaling interactions were converted into paracrine interactions between cell subtypes (Fig. 4F). This increased number of paracrine interactions highlights the extent of the complex cross-talk in the TME.

Inferring disease progression in scREMI-LUAD based on subpopulation prognostic significance

To determine which cell-cell interactions may be involved with disease progression, we computed a prognostic score for each cell subtype and a cross-talk score between pairs of cell subtypes within scREMI-LUAD. The prognostic score represents each cell subtype's inferred prognosis based on which ligands are expressed. First, we downloaded LUAD-specific prognostic gene scores calculated from bulk survival meta-analysis from the PRECOG (PREdiction of Clinical Outcomes from Genomic profiles) database (*23*). Positive and negative PRECOG *z* scores are associated with poor and good prognosis, respectively. Then, we multiplied each DE ligand's average cell type expression by its PRECOG score within each subpopulation. The weighted ligand scores were normalized by the sum of all the ligand's expression levels per subpopulation (Fig. 4G). Poor-prognostic ligand scoring subpopulations included granulocytes, F(4), M(3), M(4), and M(5), and good-prognostic ligand scoring subpopulations included F(1) and F(2). The malignant subpopulations have increasingly poor-prognostic scores, suggesting that different malignant cell subpopulations may be associated with different phases of tumor progression (Fig. 4H). From this perspective, we filtered scREMI-LUAD by selecting the ligand from the cell subtype that had the closest weighted PRECOG score to its receptor to focus on interactions potentially associated with disease progression (Supplementary Dataset).

Between every cell subtype, the cross-talk score was calculated using the number of interactions occurring between each subpopulation, standardized with respect to the number of interactions involving the sending cell. This measures the variability of the number of interacting ligands for each sending cell subpopulation. Cross-talk scores were clustered using *k*-means consensus clustering, and three groups of sending (S1, S2, and S3) subpopulations and four groups of receiving (R1, R2, R3, and R4) subpopulations were identified (Fig 4I). S2 and R2 are enriched for subpopulations associated with relatively good prognosis, such as F(1), F(2), E(1), E(2), natural killer cells, and M(2). S3 and R3 are enriched for subpopulations associated with poor prognosis, including the remaining malignant subpopulations, E(3) and F(4). We highlight three clusters of scores: (i) good-prognostic interactions between S2 and R2, (ii) mixed-prognostic interactions between S2 and R3, and (iii) poor-prognostic interactions between S3 and R3. To understand what types of interactions are occurring within each group, we performed GSEA on the good-,

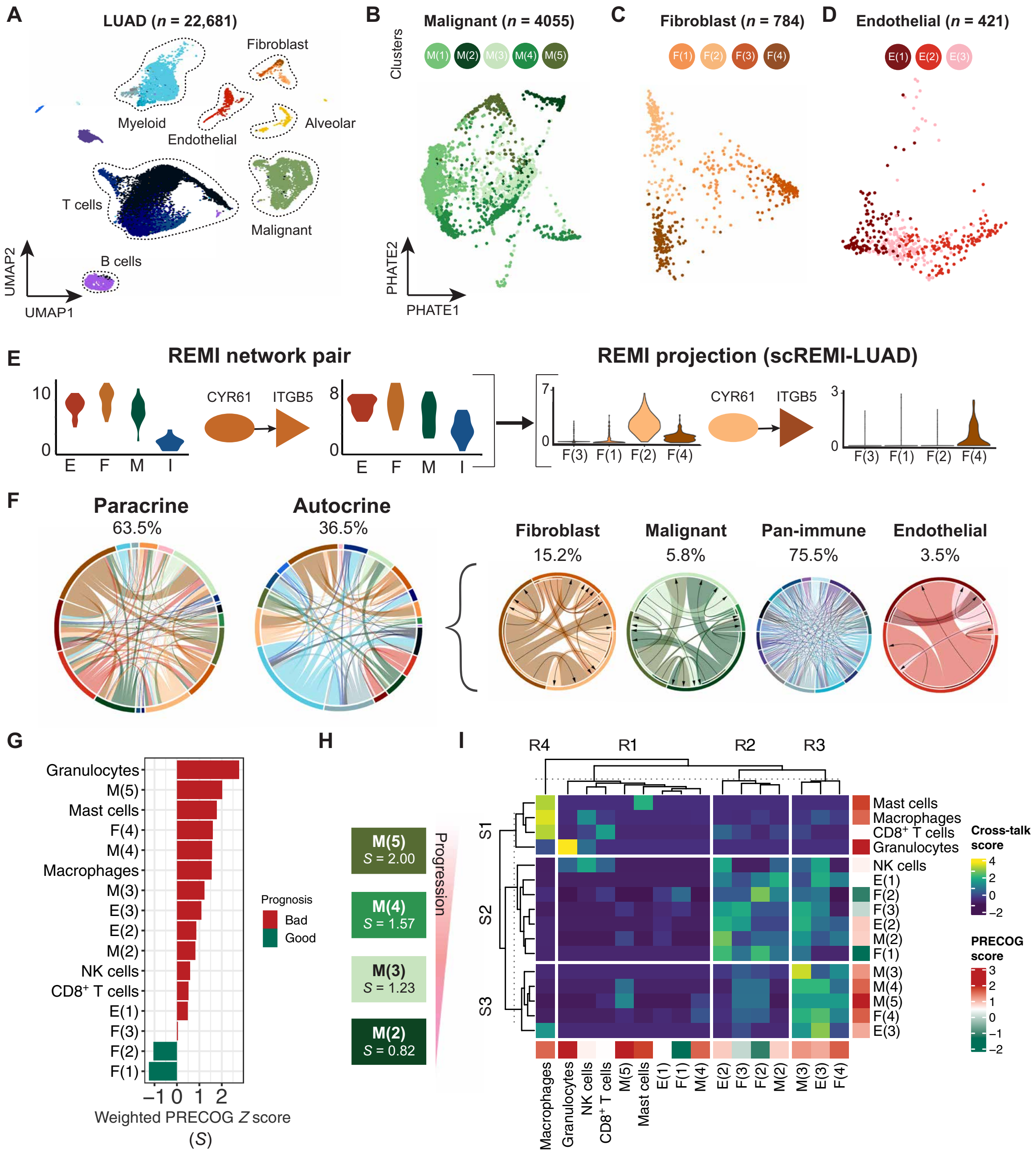


Fig. 4. Projecting scRNA-seq information onto LUAD interactome reveals prognostic progression in malignant subpopulations. (**A**) Uniform manifold approximation and projection (UMAP) of Lambrecht *et al.* lung carcinoma scRNA-seq dataset (*n* = 22,681 cells, two patients). (**B**) PHATE (potential of heat diffusion for affinity-based transition embedding) representation of malignant cell population and its five subpopulations (*n* = 4055 cells). (**C**) PHATE representation of fibroblast population and its four subpopulations (*n* = 784 cells). (**D**) PHATE representation of endothelial population and its three subpopulations (*n* = 421 cells). (**E**) Deconvolution pipeline of REMI interactome using scRNA-seq. The example shows one LR pair where its cell type node label is relabeled by the single-cell subpopulation labels. (**F**) Proportion of interactions involving each cell type in scREMI-LUAD. Two chords on the left show interactions that were labeled as paracrine and autocrine in REMI-LUAD. Four chords on the right show that the interactions that were previously labeled as autocrine are now labeled as paracrine signaling interactions between subpopulations. (**G**) Weighted prognosis scores for each subpopulation. (**H**) Progression of malignant subpopulations ranked from poor to good prognosis. *S* indicates prognosis score value. (**I**) Clustered cell-cell interaction cross-talk scores between cell subpopulations. Cross-talk score represents the variability of the number of LR interactions between two cell types with respect to the sending cell type. Cell types are labeled by weighted prognosis score.

mixed-, and poor-prognostic interactions (fig. S4). The good-prognostic interactions are enriched for coagulation and elastic fiber formation. One of the interactions occurs between SLIT2:ROBO1 as an autocrine interaction among F(2). SLIT-ROBO signaling has been shown to inhibit lung cancer migration in murine and in vitro models (*24*, *25*), and we suspect that it may play a similar role in primary lung cancer. The mixed-prognostic interactions are enriched for angiogenesis, ECM degradation, and platelet adhesion to collagen. One such interaction that occurs between M(2) and M(3), MMP7:CD151, has been confirmed to interact at the edge of LUAD nests. This suggests that mixed-prognostic interactions may occur on the invasive front (*26*). The poor-prognostic interactions were enriched for ECM degradation, angiogenesis, and Notch signaling activation. These interactions involve the ligands TNC (Tenascin-C) and VCAN (Versican), which are associated with metastatic phenotypes, such as the mesenchymal-to-epithelial transition (*27*, *28*). On the basis of the enrichments of interactions within each group, good-prognostic interactions may be associated with less invasive TMEs, poor-prognostic interactions may be associated with more invasive TMEs, and mixed-prognostic interactions may capture the transition from less to more invasive TMEs.

In situ experimental validation of CTGF and LRP6 colocalization among malignant subpopulations in LUAD

We focused on the mixed-prognostic interactions because these interactions may be involved with priming the TME for invasion. Proportionally, many mixed-prognostic interactions occur between the malignant subpopulations M(2) and M(3), suggesting potential subclonal cooperation between these subpopulations (Fig. 5a). Among the M(2)-to-M(3) interactions, we highlight the LR interaction, CTGF:LRP6, because of the increasing interest in anti-CTGF therapy (Fig. 5B) (*29*, *30*). CTGF expressed by F(2) is also inferred to interact with LRP6 expressed on M(3) within the same community ($P < 0.1$) (Fig. 5C and fig. S5, A to D, H, and I). LRP6 is a coreceptor for the WNT signaling pathway and regulates tissue homeostasis (*31*). CTGF is a complex gene involved in many different processes, such as wound healing, angiogenesis, cell adhesion, migration, fibrosis, and ECM deposition (*32*). It activates both transforming growth factor–β and WNT/β-catenin downstream signaling pathways and has been found to mediate signaling during proliferative invasion, but its involvement in LUAD remains unclear (*28*).

To determine the potential phenotypic role of CTGF:LRP6, we built a downstream signaling network for malignant cells using signaling pathway genes from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and protein-protein interactions from the publicly available database BioGRID (*33*, *34*). We set the edge weight as the correlation between two signaling genes calculated using the bulk flow-sorted RNA-seq dataset and measured the log-transformed eigenvector centrality score (ranging from 0 to 1) of each receptor. A low versus high score indicates that the receptor has low or high correlation, respectively, with its downstream pathway genes. This centrality score represents how influential a node is in the network in terms of how correlated it is with each of its nodes to the downstream network (fig. S5E). *LRP6*'s downstream nodes were *GSK3B* and *CTNNB1* (β-catenin) with a centrality score of 0.27 and 0.31, respectively, which is high with respect to the mean centrality score of 0.05 in the network (fig. S5, G and F). High centrality scores indicate that the genes were correlated with their signaling neighbors within the downstream network. Within the scRNA-seq dataset, M(3) cells expressed *GSK3B* and *CTNNB1*, and M(2) cells expressed *CTNNB1* (fig. S5I). The expression patterns suggest that the paracrine interaction CTGF:LRP6 may activate WNT signaling downstream within M(3) to induce an invasive phenotype. The Lrp6–Gsk3b–β-catenin–Tcf–Ctgf autocrine axis was implicated in a study of lung cancer progression in an in vivo preclinical model (*35*).

To confirm whether the colocalization of CTGF:LRP6 occurs in primary LUAD, we performed immunofluorescence (IF) and paired hematoxylin and eosin (H&E) imaging of CTGF and LRP6 in whole-tissue specimens from four patients, composed of two patients with low-grade LUAD and two with high-grade LUAD (fig. S6). Low-grade versus high-grade samples were selected because they are typically indicative of better or worse prognosis, respectively. In both low-grade samples, we see CTGF$^+$ and LRP6$^+$ cell subtypes, but we generally do not observe CTGF$^+$LRP6$^+$ colocalization [Fig. 5D (i) and fig. S6, A and B]. In both high-grade samples, we observe CTGF$^+$LRP6$^+$ colocalization in malignant cells (fig. S6, C and D). The colocalization was observed in well-to-moderately differentiated regions of these high-grade tumors. These observations are consistent with our inference that CTGF:LRP6 interactions occur in regions of the tumor that may be becoming primed for more aggressive tumor growth. We observed in one of the high-grade samples distinctive CTGF$^+$LRP6$^-$ malignant cells (arrow 4 in Fig. 5D, ii) suggestive of M(2) subtype from the scRNA-seq analysis.

DISCUSSION

REMI is a novel algorithm that uses graph-based approaches to generate a global cell-cell cross-talk network by estimating the conditional dependency of LR pairs on high-dimensional data of small sample size. As opposed to current methods, REMI captures the complexity of multicellular interactions by accounting for the potential confounding effects that LR pairs may have upon one another. REMI offers the main advantage of identifying codependent interactions on a global scale with small sample size by leveraging previous knowledge, network analysis, and sparsity principles. We demonstrated REMI's performance on a large transcriptomic dataset from TCGA, showing that estimating the inverse covariance matrix using REMI is robust. We measured REMI's performance across varying sample sizes and found that REMI's specificity decreased less than its sensitivity relative to GLasso. REMI also outperformed existing cell-cell interaction inference approaches. The difference in performance metrics can be explained by the difference in the assumptions underlying each algorithm. CellphoneDB v2.0 predicts LR pairs on the basis of the mean expression values, as higher LR genes are more likely to be expressed and interacting. NicheNet relies on both literature-derived models and user-defined inputs to specify downstream-activated genes. Instead, REMI focuses on capturing the linear transcriptional relationship between a ligand and a receptor, which is not as well captured in the other methods. Overall, REMI's higher specificity is favorable when choosing interactions for experimental validation to reduce the risk of false positives.

REMI is a modularized algorithm that allows investigators to adapt each module to best fit their biological question. In step 1, we used the FANTOM5 database as our reference for literature-supported LR pairs, but the database can be substituted or complemented with any other LR databases. In step 2, we used the Louvain community detection algorithm, but this can be substituted for the investigator's clustering algorithm of choice. REMI's community

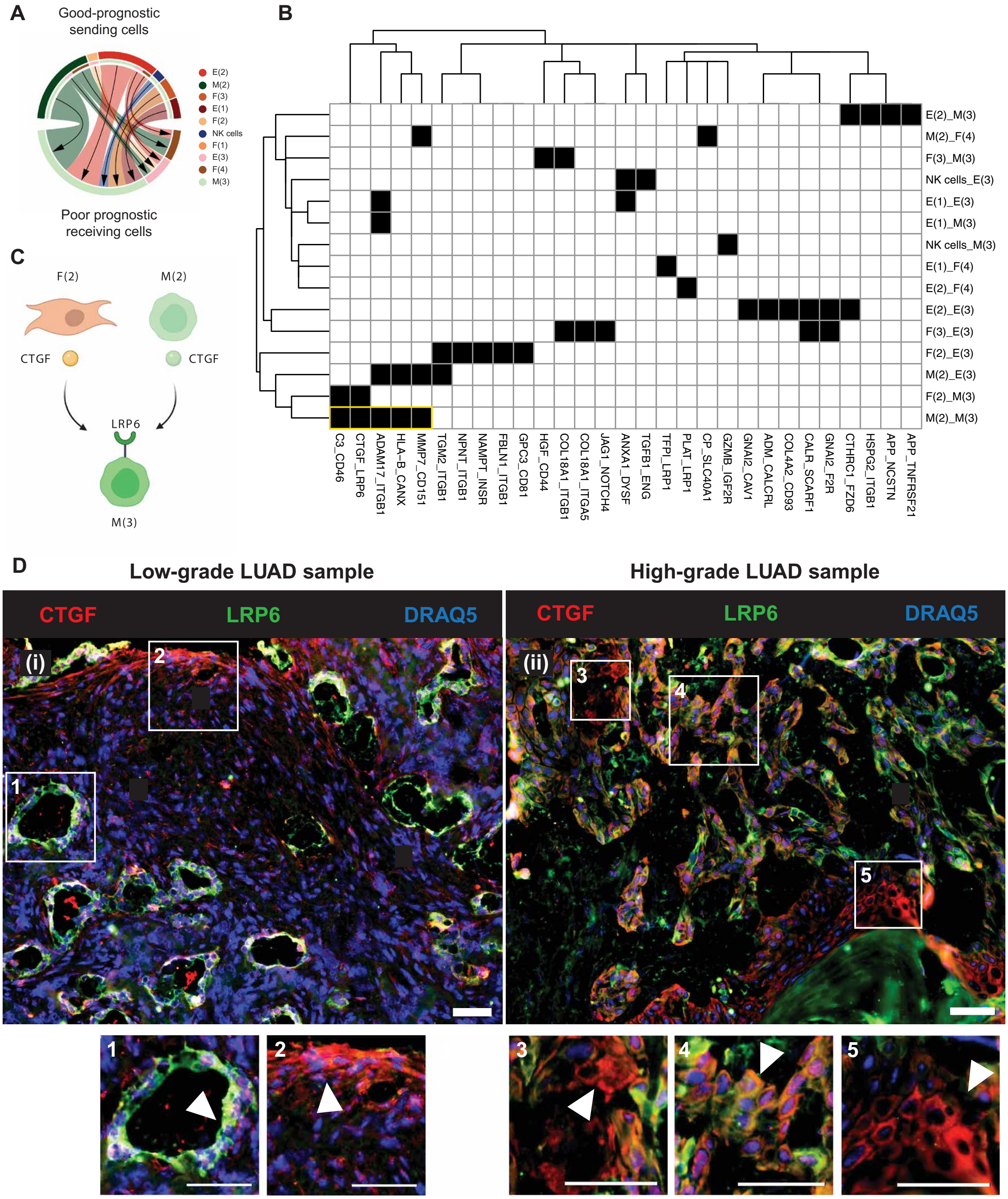


Fig. 5. Cell-cell interaction associated with progression in LUAD. (**A**) Circle plot indicates ratio of interactions occurring in mixed-prognostic subpopulations. The color of the chords corresponds to the legend on the right. Arrows point from sending to receiving cell types. NK, natural killer. (**B**) Binarized heatmap of cell-cell interactions between good-prognostic sending subpopulations and poor-prognostic receiving subpopulations. Interactions were filtered by average receptor expression >0.5. Black indicates a predicted cell-cell interaction between LR. White indicates no LR interaction prediction. Yellow box highlights interactions occurring between M(2) and M(3). (**C**) Diagram of CTGF:LRP6 interactions predicted by scREMI-LUAD. (**D**) Immunofluorescence (IF) staining analysis of CTGF (red), LRP6 (green), and DRAQ5 (blue) in fresh-frozen LUAD samples from patients with low-grade (i) versus high-grade (ii) disease. In (i), box nos. 1 and 2 contain CTGF$^-$LRP6$^+$ malignant cells and CTGF$^+$LRP6$^-$ fibroblast cells, respectively (see H&E in fig. S6). In (ii), box nos. 3, 4, and 5 contain CTGF$^+$LRP6$^-$ malignant cells, CTGF$^+$LRP6$^+$ colocalized malignant cells, and CTGF$^+$LRP6$^-$ bronchial cells with potential malignancy, respectively. Scale bars, 50 μm.

detection step is hierarchical, which allows it to accommodate high-dimensional datasets while retaining high specificity. Thus, it can be applied to the many recent single-cell studies that have revealed substantial intratumoral heterogeneity. Last, REMI was applied to transcriptomics data, but it can be generalized when applied to other -omics data as well.

We applied REMI to an LUAD bulk flow-sorted RNA-seq dataset composed of four broad cell types (malignant, immune, fibroblast, and endothelial) to assemble the REMI-LUAD microenvironment interactome. We showed that REMI-LUAD captured tumor-stroma– and tumor-stroma-immune–specific communities that were enriched for ECM remodeling interactions and highlighted the prevalence of tumor-stroma–specific interactions. To increase the cell type resolution of REMI-reconstructed networks, we projected a publicly available scRNA-seq data onto REMI-LUAD to construct scREMI-LUAD and uncovered potential interactions associated with LUAD progression. Because of the increased granularity of the interactions, we observed that previously labeled autocrine signaling interactions within REMI-LUAD became paracrine signaling interactions between cell subtypes in scREMI-LUAD. This observation suggests that the currently documented autocrine signaling interactions may be involved in paracrine signaling between different cell states for a given cell type.

To infer a cancer progression signature within scREMI-LUAD, we calculated the prognosis of each cell subtype according to its ligand expression. We then applied unsupervised analysis of prognostic-specific cellular cross-talk signatures and ordered interactions on the basis of prognosis as a proxy to tumor progression. Interactions among good-prognostic cell subtypes consisted of antitumoral or preinvasive activities, whereas interactions among poor-prognostic cell subtypes were involved in more invasive properties. We reasoned that mixed-prognostic interactions may be associated with progression from a less to more invasive TME.

We noted one mixed-prognostic interaction, namely, between CTGF and LRP6 among malignant subpopulations, and experimentally verified the presence versus the absence of colocalization of CTGF and LRP6 in high- versus low-grade primary LUAD samples. More specifically, we observed CTGF and LRP6 colocalization in well-to-moderately differentiated malignant regions of high-grade primary LUAD samples and did not observe their colocalization in low-grade tumors. These findings, while preliminary, are consistent with our inferences that CTGF and LRP6 interactions may be priming the tumor toward a more aggressive phenotype. This work suggests that anti-CTGF (*29*, *30*) therapy to inhibit progression in LUAD warrants consideration. We did not observe evidence of CTGF expression by fibroblasts interacting with LRP6 on malignant cells, which was another predicted interaction in scREMI-LUAD. This may be attributed to the low amounts of fibroblasts seen in our tissue specimens. We also did not have access to tissue blocks that captured more poorly differentiated malignant regions of the high-grade cases for further depiction of the CTGF$^+$ and LRP6$^+$ colocalization. Therefore, both scenarios warrant further experimental investigation. More generally, this work suggests that cell subtypes with prognostically relevant cross-talk warrant further investigation as they may be associated with disease progression.

While REMI can identify LR pairs with high specificity, it has limitations. First, the utility of REMI may be limited by the assumption that interacting ligands and receptors are correlated. By leveraging these correlations between LR genes representative of LR pairs, interactions that involve secretory molecules, such as chemokines that are typically secreted by immune cells, may not be captured well by REMI because chemokines can affect distant tissue sites. Second, REMI does not yet account for large protein complexes. One way to potentially address this issue is to regularize edges in the resulting interactome between receptors in the same protein complex along with the LR pairs in the community. Third, REMI does not account for potential external influences on LR pairs, such as spatial or temporal considerations. This may be remedied by implementing a weighted graphical Gaussian model instead. A weighted approach may also allow users to capture LR affinity or downstream signaling effects. Last, the significance test is designed for each community because the edge permutations change the communities identified by REMI. This stochasticity is difficult to capture in REMI's current statistical test. Calculating *P* values for all edges within the network is also time intensive. Further work needs to be done to implement a fast interactome-wide hypothesis test.

In summary, we developed REMI and demonstrated the benefit of using its networks as a discovery tool. REMI provides a novel way to infer global cellular cross-talk for subsequent functional validation. We anticipate that REMI can be applied to many more tissue microenvironments to better understand normal and diseased processes and identify more precise therapeutics.

MATERIALS AND METHODS

Bulk flow-sorted RNA-seq data

We used publicly available bulk flow-sorted RNA-seq data (GSE111907) that had 17 primary LUAD tumor samples that, using flow cytometry, were sorted into the four cell types of interest: immune, endothelial, fibroblast, and malignant cancerous cells using the markers CD45$^+$/EpCAM$^-$, CD31$^+$/CD45$^-$/EpCAM$^-$, CD10$^+$/EpCAM$^-$/CD45$^-$/CD31$^-$, and epithelial cellular adhesion molecule (EpCAM)/CD45$^-$, respectively. Patients without data from all four cell types were removed. The data included bulk RNA-seq measurements for each cell type defined as $X = \{X_{c1}, X_{c2}, X_{c3}, X_{cn}\}$. The TPM (transcripts per million) values were log transformed [$\log_2(\text{TPM} + 1)$] and standardized to 0 mean and 1 SD.

Publicly available databases

FANTOM5 database

We obtained a total of 1904 validated LR pairs (650 ligands and 594 receptors) from the FANTOM5 database (*15*). The database extracted known LR pairs from the Database of Ligand-Receptor Partners (DLRP), IUPHAR (International Union of Basic and Clinical Pharmacology), and the Human Plasma Membrane Receptome. We filtered for LR pairs that had literature-supported evidence and were experimentally validated. We included 12 more literature-supported experimentally validated LR pairs not found in the database: PDL1/PD-1, PD-2/PDL1, CD80/CTLA4, CD80/CD28, CD86/CD28, GREM1/KDR, PDL2/PD-1, NECTIN2/CD226, NECTIN2/TIGHT, PVR/TIGHT, SIGLEC1/SPN, and CTGF/TNFRSF1A.

KEGG database

From the KEGG database, we downloaded all the pathway genes from the Environmental Information Processing category. This includes signaling pathways associated with membrane transport, signal transduction, and signaling molecules and interactions. Genes from the phosphotransferase system and bacterial secretion system were excluded.

BioGRID database

We downloaded the *Homo sapiens* proteins from BioGRID and their interaction network. Only protein-protein interactions that were experimentally validated were used. We removed proteins that were found in the publicly available database CRAPome (*36*). The remaining BioGRID network has 18,074 proteins and 176,120 interactions.

LTMI reconstruction

For the LTMI reconstruction, the log-transformed bulk flow-sorted RNA-seq data from Gentles *et al.* were split into patients with LUAD and those with lung squamous cell carcinoma. Using only the patients with LUAD ($n = 17$), LR genes were filtered by TPM > 10. A graph was constructed by creating a node for every LR gene that passed the filtering criteria, and an edge was drawn if it was a known pair according to the FANTOM5 database. This was cross-referenced with the LTMI built on all samples. The igraph R package was used to construct and illustrate the graph. We also used igraph to identify components in the network. To calculate partial correlation for the LTMI subgraph, we used the following equation

$$r_{\mathrm{ABC}} = \frac{r_{\mathrm{AB}} - r_{\mathrm{AC}} r_{\mathrm{BC}}}{\sqrt{(1 - r_{\mathrm{AC}}^2)(1 - r_{\mathrm{BC}}^2)}}$$

This equation was extended to capture all combinations of gene interactions within the subgraph. Correlation was measured using the gene expression dataset.

REMI algorithm

(i) Build weighted bipartite LR network. On the basis of a user-defined LR list, an LR network is created. The default LR list is obtained from the FANTOM5 database. Cell types from the input dataset are denoted as $c_1, c_2, \ldots, c_n \in X$. x_c represents all ligands found in cell type c ($x_1, x_2, \ldots, x_a \in X_c$). y_c represents all receptors found in cell type c ($y_1, y_2, \ldots, y_b \in X_c$). Given dataset X, we represent all potential LR pairs interacting between all cell types in the undirected network $G_{\mathrm{LR}} = (V_1 = V_x \cap V_y, E_1)$, where $V_x = \{x : x_{c1}, x_{c2}, \ldots, x_{cn}\}$ and $V_y = \{y : y_{c1}, y_{c2}, \ldots, y_{cn}\}$. An edge represents a relationship between ligand and its cognate receptor denoted as $E_1 = \{(x, y) : x \in V_x, y \in V_y, (x, y) \in N_{\mathrm{FANTOM5}}\}$. Edge weight w_{ij} in G_{LR} is the Pearson correlation between x_i and y_j calculated using the gene expression data.

(ii) Detect cognate LR communities. Communities (C_k) are identified in G_{LR} using Louvain community detection algorithm. The algorithm identifies clusters of nodes, called communities, that are optimized for maximum modularity (Q) within the network

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w)$$

A_{vw} is the Spearman correlation between nodes v and w. m is the number of edges. k is the degree of node. δ is the Kronecker delta piecewise function indicating whether node v or w is in the same community. c is an indicator function for whether a node is in a community. The measurement ensures that the communities detected are data driven and not based on chance. G_{LR} can contain disjoint components because some nodes were removed because of their low expression within the dataset. Disconnected small components are referred to as a community. If the number of nodes within C_k is greater than the sample size, then we perform hierarchical community detection using Louvain to further reduce the density of the communities until the number of nodes in every community is less than or equal to the sample size. We then iterate between every pairwise combination of communities and create a pseudocommunity or BC. The BCs contain nodes that have an edge between two communities.

(iii) Remove conditionally independent LR pairs. According to the Hammersley-Clifford theorem, an inverse covariance matrix, or partial correlation, can be used to indicate relationships between nodes. A zero in an inverse covariance matrix indicates conditional independence between the two variables in the model given all others. The algorithm GLasso uses this concept and estimates the inverse covariance matrix that best fits the observations in a multivariate Gaussian distributed dataset (*17*). A *l*1 penalty is added to the log-likelihood of a Gaussian Markov field to add data-driven sparsity to the model. Here is the optimization function

$$\log(\det\Theta - \mathrm{tr}(S\Theta) - \lambda \| w^* \Theta \|_1)$$

In this equation, $\Theta = \sigma^{-1}$ and S are the empirical covariance matrix. w represents the weighted adjacency matrix. tr is the trace of the matrix or the sum of the diagonal. $\|\Theta\|_1$ is the *l*1 penalization of the inverse covariance matrix. When λ increases, the network becomes sparser. When λ is zero, the resulting network is equivalent to a partial correlation network (*17*). We use GLasso from the glasso R package to measure the conditional dependence between two connected nodes while controlling for the effect of all other nodes for each community detected. Because the communities identified within the largest network component are not disjoint clusters, we also created graphical models using the nodes and edges BCs. These BCs only include nodes that have edges BCs and excludes nodes that do not. Communities with one node were not included.

The GLasso similarity matrix can be represented as a fully connected LR community graph. This includes all potential edges between LRs, ligand and ligand (LL), and receptor and receptor (RR). We simplify the LR prediction problem by focusing on only LR interactions and make the assumption that the underlying network that represents the data only contains edges between LRs. LL and RR edges within the G_{LR} communities are penalized and set to zero in the weighted adjacency matrix, indicating conditional independence. The penalized weighted matrix is defined as

$$w_{ij} = \begin{cases} \lambda, & if\,(i,j) \in C_k \\ 0, & if(i) \in C_k, (j) \in C_{k'}, k != k' \end{cases}$$

The λ tuning parameter in the penalized log-likelihood equation is chosen by iterating through 20 values logarithmically spaced between lambda values 0.01 to 0.9. The optimal λ is the one that best minimizes the Bayesian information criterion. The accuracy of the GLasso method is dependent on the number of features in the network.

(iv) Reconstruct communities into predicted interactome. The final reconstructed LR network includes edges between nodes that have a weight larger than zero.

(v) Significance test for P value. We provide an optional additional tool to calculate P values for an edge in a community. P values are calculated by measuring the occurrence of a test statistic compared with its null. In the REMI algorithm, an LR pair is a subset of a test statistic, the regularized covariance matrix. Therefore, our significance test measures the conditional probability that an LR pair is conditionally dependent within its REMI community given added randomization. First, we sample from a uniform distribution of correlation values and replace the correlation value of the LR pair of interest. Then, REMI is rerun 1000 times to generate a distribution of predicted scores for the LR pair. A regression tree is fit on the sampled correlation values to predict binary REMI scores. A null distribution is lastly generated by simulating a null training dataset by creating a Gaussian distribution (G) and training it on the regression tree to predict a null set of outcomes (w). w is multiplied by the Wishart distribution to create a null distribution that represents the conditional test statistic (*37*). The P value is then calculated by using the LR pair's data-derived correlation value ($R[i, j]$). The first equation represents the one-sided P value, and the second equation represents the two-sided P value

$$p = \frac{\sum_n w * (G > R[i, j])}{\sum_n w}$$

$$p = 2 * \min(p, 1 - p)$$

Validation dataset

We tested the method on a bulk RNA-seq dataset downloaded from the publicly available database, TCGA. The non–small cell lung carcinoma (NSCLC) RNA-seq dataset contains bulk RNA-seq data for 1013 patients with NSCLC. This includes 584 patients with LUAD and 429 patients with lung squamous cell carcinoma. The data were log transformed and scaled to a mean of 0 and a variance of 1.

Simulations

The TCGA-LUAD GLasso interactome was generated by running GLasso interactome (GLasso on G_{LR}) using LR genes found in the TCGA NSCLC bulk RNA-seq dataset. In the resulting network ($V = 893$ and $E = 1331$), a true-positive edge is an edge with weight greater than zero. All other edges are true-negative edges. Five cohort sizes ($n = 10$, 25, 50, 100, and 2000) were sampled 50 times each from the whole TCGA NSCLC cohort to simulate various cohort sizes. REMI was run on each sampled cohort dataset using default parameters and unfiltered list of LR genes. GLasso was also performed on each sampled cohort dataset with restrictions on the LL and RR edges. Performance metrics were calculated by setting the reference "gold standard" interactome as the TCGA-LUAD GLasso interactome. Correlation values of the LR pairs were used as the cutoff for REMI's AUC calculations. The same analysis was performed by separating WCs' and BCs' predictions and calculating their performance metrics.

Comparison against other methods

NicheNet

We ran NicheNet on TCGA NSCLC bulk RNA-seq dataset using the same filtered genelist as the ones used in TCGA simulations described above. The background gene set used included the genes found in the ligand target matrix of the receiving cell. The LR list was replaced with the FANTOM5 list we generated. For TCGA data, we treated the bulk dataset as an autocrine signaling. We ran NicheNet on every autocrine and paracrine combination of cell types. We used the best ligand filter of Pearson >0.

NATMI

We used both log-transformed and scaled log-transformed TCGA RNA-seq dataset as input and set all samples as one cell type. Default parameters were used, and edges were filtered on the basis of specificity score.

CellphoneDB

We used the log-transformed TCGA RNA-seq dataset as input for the method. We used the database provided and filtered for significant LR pairs based on the significant means score.

CCCExplorer

We ran the CCCExplorer on TCGA NSCLC bulk RNA-seq dataset and bulk flow-sorted RNA-seq dataset. To identify differentially expressed gene (DEG) ligands for the input, we used samr to identify DEGs (*38*). For TCGA, we identified DEG between healthy patients and patients with lung cancer. For the bulk flow-sorted cohort, we identified DEG between the sending cell and receiving cell. The DEG cutoffs used were logFC >1.2 and $q < 0.05$. The same DEG input from NicheNet was used. We calculated cell-cell interactions for every pairwise paracrine and autocrine combination.

REMI analysis in the LUAD dataset

For the REMI interactome construction, the log-normalized gene expression measurements were binned into three evenly split discretized groups for each cell type. Genes with expression levels falling within the first group, representing low expressed genes, were removed. The dispersion ratio (variance/mean) was calculated for each gene and normalized within each bin. Genes in the first bin, which includes genes with negative and fairly low expression levels, were removed. The data were scaled after filtering, and the filtered dataset was used to identify LR genes. The scaled unfiltered gene expression dataset was used for all other analyses. All other parameters in REMI were set to default.

Gene enrichment analysis

The clusterProfiler R package was used to perform gene enrichment analysis. The gene lists were filtered for the databases: Hallmark, BIOCARTA, and REACTOME. Enrichments with adjusted $P < 0.05$ were used for analysis.

scRNA-seq analysis

Data

We used publicly available scRNA-seq datasets E-MTAB-6149 and E-MTAB-6653 generated by Lambrecht *et al.* (*21*). The droplet-based scRNA-seq (10X genomics) datasets had two patients with squamous cell carcinoma, two patients with adenocarcinoma, and one patient with NSCLC. We used the Seurat V3 R package to process the patients with LUAD and used predefined cell type labels defined by the authors (*39*).

Preprocessing

In cells with more than 10% mitochondria content, over 6000 and under 100 features were filtered out. We used the top 30 principal components and the IntegrateData function in Seurat to correct for batch effect between the patients. Malignant, fibroblast, and endothelial cells were reclustered using Louvain clustering on a Shared Nearest Neighbor graph with 10 principal components using 0.2, 0.2, and 0.1 resolutions, respectively. DEGs across all cell types and within clusters were calculated using model-based analysis of single cell transcriptomics (MAST) with default parameters and normalized counts (averaged $\log_2$FC > 0.25; Bonferroni-corrected $P < 0.05$) (*40*). Visualizations are shown using uniform manifold approximation and projection (UMAP) (30 principal components). PHATE (potential of heat diffusion for affinity-based transition embedding) was used to infer the underlying hierarchical manifold in each cell type using default parameters (*41*). Each cell type population was imputed using adaptively-thresholded low rank approximation (ALRA) after identifying subpopulations (*42*).

Deconvolving interactome

LR genes in the REMI-LUAD interactome were filtered on the basis of whether they were expressed in their respective cell type (average expression, >0.4). The cell type metadata associated with each node in the interactome was relabeled using the subpopulation in which the ligand or receptor was found in. The cell type must be the same between bulk and single-cell data. Nodes with genes that were found across multiple subpopulations within the same cell type were duplicated, and an edge drawn to its respective predicted cognate pair.

Measuring prognosis of subpopulations

Bulk LUAD-specific prognosis scores were downloaded from the PRECOG metaZ database. Then, the normalized log-transformed average expression level for the ligands within each subpopulation was calculated using the ALRA imputed gene expression data. Each subpopulation was imputed independently. For each subpopulation, ligands that were not DE, as calculated by MAST in a prior step, were removed. DE ligand cognate receptors with an average scaled expression less than 0 were removed. Subpopulations with less than three DE ligands predicted to be secreted were also removed from the calculations to increase the power of the calculation. In the remaining list of LR pairs, we then multiplied the DE ligand's averaged expression level by its PRECOG score [$a = \sum_n e_i^L * \text{PRECOG}(L)$]. The total gene expression values of ligands ($b = \sum_n e_i^L$) were aggregated, and the prognosis of the subpopulation by a/b was calculated. n is the total number of cells in a subtype. i is the gene index. L is the ligand of interest. Additional filtering of the LR pairs was then applied to duplicate LR pairs that have the same genes but different sending and receiving cells. The pair with the closest PRECOG score between sending and receiving cells was selected.

IF and H&E staining

Clinical aspects of this study were approved by the Stanford Institutional Review Board (IRB) in accordance with the Declaration of Helsinki guidelines for the ethical conduct of research. All patients involved provided a written informed consent. Collection and use of human tissues were approved and in compliance with data protection regulations regarding patient confidentiality (IRB protocol no. 15166). Following surgical resection of primary tumors from patients at Stanford Hospital, LUAD specimens were immediately embedded in optimal cutting temperature (OCT) compound. Fresh-frozen human LUAD tissues were serially sectioned at a thickness of 8 μm and mounted on poly-L-lysine–coated coverslips. One section was successively fixed in 95% ethanol and 10% formalin for 10 min each and subjected to progressive H&E staining following the standard protocol described (*43*) and reviewed by a board-certified pathologist (M.E.O.). For IF analysis, sections were fixed in acetone for 10 min at room temperature (RT) and then washed twice with phosphate-buffered saline (PBS). Fluorophore bleaching was performed at RT for 90 min to reduce autofluorescence by immersing the fixed coverslip in freshly prepared bleaching solution [4.5% (w/v) H_2O_2 and 20 mM NaOH in PBS] and illuminating the container between two light-emitting diode (LED) light panels, followed by four washes with PBS to remove bleaching solution (*44*). The pretreated coverslips were blocked with 1% horse serum in PBS for 30 min and incubated with primary antibodies in incubation solution (1% bovine serum albumin, 1% normal donkey serum, 0.3% Triton X-100, and 0.01% sodium azide in PBS) in a humidified chamber overnight at 4°C. After three washes in PBS for 15 min per wash, the coverslips were incubated with secondary antibodies in incubation solution for 1 hour at RT in dark. The coverslips were intensively washed with PBS and then stained with DRAQ5 (Thermo Fisher Scientific) at 1:1000 in PBS for 10 min. The sections were observed under the BZ-X800 fluorescence microscope (Keyence, IL, USA). Antibodies used in the study include anti-CTGF (ab6992, Abcam), anti-LRP6 (MAB1505, R&D System), anti-EpCAM (no. 2929, Cell Signaling Technology), anti-pan Cytokeratin (ab86734, Invitrogen), Cy3-conjugated goat anti-rabbit immunoglobulin G (IgG) (ab6939, Abcam), goat anti-mIgG1 Alexa Fluor 488 (A-21121, Invitrogen), and donkey anti-mouse IgG H&L Alexa Fluor 488 (ab150105, Abcam) antibodies.

Downstream influence score

An undirected downstream signaling pathway network $G_{\text{ds}} = (V_{\text{ds}}, E_{\text{ds}})$ is created for each cell type in the dataset. Each G_{ds} consists of approximately 1000 nodes representing signal transduction pathway genes (N_{KEGG}) downloaded from the publicly available database, KEGG ($V_{\text{ds}} = \{x : x \in N_{\text{KEGG}}\}$) (*34*). An edge between two nodes represents a physical protein-protein interaction between the nodes, as indicated in the database BioGRID ($E_{\text{ds}} = \{(x_i, x_j) \in N_{\text{BioGRID}}\}$) (*33*). The edge weight in the network is set as the Pearson correlation between the two gene nodes calculated from the gene expression data from the cell type of interest. For each G_{ds} per cell type, we calculate the eigenvector centrality for each node. Eigenvector centrality is a centrality measurement of a node x_i that is relational to the sum of importance of its neighbor nodes: $x_i = \frac{1}{\lambda} \sum_{j \in G} w_{i,j} x_j$, where $w_{i,\,j}$ is the edge weight between nodes x_i and x_j. The centrality measurement for each receptor within G_{ds} is the receptor's influence score.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abi4757

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. S. K. Wculek, I. Malanchi, Neutrophils support lung colonization of metastasis-initiating breast cancer cells. *Nature* **528**, 413–417 (2015).
2. E. Armingol, A. Officer, O. Harismendy, N. E. Lewis, Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22**, 71–88 (2021).
3. M. Efremova, M. Vento-Tormo, S. A. Teichmann, R. Vento-Tormo, CellPhoneDB: Inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).
4. R. Hou, E. Denisenko, H. T. Ong, J. A. Ramilowski, A. R. R. Forrest, Predicting cell-to-cell communication networks using NATMI. *Nat. Commun.* **11**, 5011 (2020).
5. H. Choi, J. Sheng, D. Gao, F. Li, A. Durrans, S. Ryu, S. B. Lee, N. Narula, S. Rafii, O. Elemento, N. K. Altorki, S. T. C. Wong, V. Mittal, Transcriptome analysis of individual stromal cell populations identifies stroma-tumor crosstalk in mouse lung cancer model. *Cell Rep.* **10**, 1187–1201 (2015).
6. R. Browaeys, W. Saelens, Y. Saeys, NicheNet: Modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162 (2020).
7. A. Patsialou, J. Wyckoff, Y. Wang, S. Goswami, E. R. Stanley, J. S. Condeelis, Invasion of human breast cancer cells in vivo requires both paracrine and autocrine loops involving the colony-stimulating factor-1 receptor. *Cancer Res.* **69**, 9498–9506 (2009).
8. L. Hernandez, T. Smirnova, D. Kedrin, J. Wyckoff, L. Zhu, E. R. Stanley, D. Cox, W. J. Muller, J. W. Pollard, N. van Rooijen, J. E. Segall, The EGF/CSF-1 paracrine invasion loop can be triggered by heregulin beta1 and CXCL12. *Cancer Res.* **69**, 3221–3227 (2009).
9. M. P. Kumar, J. Du, G. Lagoudas, Y. Jiao, A. Sawyer, D. C. Drummond, D. A. Lauffenburger, A. Raue, Analysis of single-cell RNA-seq identifies cell-cell communication associated with tumor characteristics. *Cell Rep.* **25**, 1458–1468.e4 (2018).

10. A. J. Gentles, A. B. Y. Hui, W. Feng, A. Azizi, R. V. Nair, G. Bouchard, D. A. Knowles, A. Yu, Y. Jeong, A. Bejnood, E. Forgó, S. Varma, Y. Xu, A. Kuong, V. S. Nair, R. West, M. van de Rijn, C. D. Hoang, M. Diehn, S. K. Plevritis, A human lung tumor microenvironment interactome identifies clinically relevant cell-type cross-talk. *Genome Biol.* **21**, 107 (2020).
11. J. E. Allen, W. S. El-Deiry, Regulation of the human TRAIL gene. *Cancer Biol. Ther.* **13**, 1143–1151 (2012).
12. M. de Looff, S. de Jong, F. A. E. Kruyt, Multiple interactions between cancer cells and the tumor microenvironment modulate TRAIL signaling: implications for TRAIL receptor targeted therapy. *Front. Immunol.* **10**, 1530 (2019).
13. D. Deng, K. Shah, TRAIL of hope meeting resistance in cancer. *Trends Cancer* **6**, 989–1001 (2020).
14. T. Hartwig, A. Montinaro, S. von Karstedt, A. Sevko, S. Surinova, A. Chakravarthy, L. Taraborrelli, P. Draber, E. Lafont, F. A. Vargas, M. A. El-Bahrawy, S. A. Quezada, H. Walczak, The TRAIL-induced cancer secretome promotes a tumor-supportive immune microenvironment via CCR2. *Mol. Cell.* **65**, 730–742.e5 (2017).
15. J. A. Ramilowski, T. Goldberg, J. Harshbarger, E. Kloppmann, M. Lizio, V. P. Satagopam, M. Itoh, H. Kawaji, P. Carninci, B. Rost, A. R. R. Forrest, A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat. Commun.* **6**, 7866 (2015).
16. K. Komurov, Modeling community-wide molecular networks of multicellular systems. *Bioinformatics* **28**, 694–700 (2012).
17. J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441 (2008).
18. X. Guo, Y. Zhang, L. Zheng, C. Zheng, J. Song, Q. Zhang, B. Kang, Z. Liu, L. Jin, R. Xing, R. Gao, L. Zhang, M. Dong, X. Hu, X. Ren, D. Kirchhoff, H. G. Roider, T. Yan, Z. Zhang, Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat. Med.* **24**, 978–985 (2018).
19. J. Winkler, A. Abisoye-Ogunniyan, K. J. Metcalf, Z. Werb, Concepts of extracellular matrix remodelling in tumour progression and metastasis. *Nat. Commun.* **11**, 5120 (2020).
20. V. Gkretsi, T. Stylianopoulos, Cell adhesion and matrix stiffness: coordinating cancer cell invasion and metastasis. *Front. Oncol.* **8**, 145 (2018).
21. D. Lambrechts, E. Wauters, B. Boeckx, S. Aibar, D. Nittner, O. Burton, A. Bassez, H. Decaluwé, A. Pircher, K. van den Eynde, B. Weynand, E. Verbeken, P. de Leyn, A. Liston, J. Vansteenkiste, P. Carmeliet, S. Aerts, B. Thienpont, Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat. Med.* **24**, 1277–1289 (2018).
22. R. Kalluri, The biology and function of fibroblasts in cancer. *Nat. Rev. Cancer* **16**, 582–598 (2016).
23. A. J. Gentles, A. M. Newman, C. L. Liu, S. V. Bratman, W. Feng, D. Kim, V. S. Nair, Y. Xu, A. Khuong, C. D. Hoang, M. Diehn, R. B. West, S. K. Plevritis, A. A. Alizadeh, The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* **21**, 938–945 (2015).
24. R. Kong, F. Yi, P. Wen, J. Liu, X. Chen, J. Ren, X. Li, Y. Shang, Y. Nie, K. Wu, D. Fan, L. Zhu, W. Feng, J. Y. Wu, Myo9b is a key player in SLIT/ROBO-mediated lung tumor suppression. *J. Clin. Invest.* **125**, 4407–4420 (2015).
25. R. K. Gara, S. Kumari, A. Ganju, M. M. Yallapu, M. Jaggi, S. C. Chauhan, Slit/Robo pathway: A promising therapeutic target for cancer. *Drug Discov. Today* **20**, 156–164 (2015).
26. R. Sadej, A. Grudowska, L. Turczyk, R. Kordek, H. M. Romanska, CD151 in cancer progression and metastasis: A complex scenario. *Lab. Invest.* **94**, 41–51 (2014).
27. N. K. Altorki, G. J. Markowitz, D. Gao, J. L. Port, A. Saxena, B. Stiles, T. McGraw, V. Mittal, The lung microenvironment: An important regulator of tumour growth and metastasis. *Nat. Rev. Cancer* **19**, 9–31 (2019).
28. A. L. Parker, T. R. Cox, The role of the ECM in lung cancer dormancy and outgrowth. *Front. Oncol.* **10**, 1766 (2020).
29. Z. Chen, N. Zhang, H. Y. Chu, Y. Yu, Z. K. Zhang, G. Zhang, B. T. Zhang, Connective tissue growth factor: From molecular understandings to drug discovery. *Front. Cell Dev. Biol.* **8**, 593269 (2020).
30. A. Resovi, P. Borsotti, T. Ceruti, A. Passoni, M. Zucchetti, A. Berndt, B. L. Riser, G. Taraboletti, D. Belotti, CCN-based therapeutic peptides modify pancreatic ductal adenocarcinoma microenvironment and decrease tumor growth in combination with chemotherapy. *Cells* **9**, 952 (2020).
31. J. Raisch, A. Côté-Biron, N. Rivard, A role for the WNT co-receptor LRP6 in pathogenesis and therapy of epithelial cancers. *Cancers* **11**, 1162 (2019).
32. K. E. Lipson, C. Wong, Y. Teng, S. Spong, CTGF is a central mediator of tissue remodeling and fibrosis and its inhibition can reverse the process of fibrosis. *Fibrogenesis Tissue Repair* **5**, S24 (2012).
33. C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, M. Tyers, BioGRID: A general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–D539 (2006).
34. M. Kanehisa, Y. Sato, M. Furumichi, K. Morishima, M. Tanabe, New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595 (2019).
35. L. Jiang, Y. Yamashita, S. H. Chew, S. Akatsuka, S. Ukai, S. Wang, H. Nagai, Y. Okazaki, T. Takahashi, S. Toyokuni, Connective tissue growth factor and β-catenin constitute an autocrine loop for activation in rat sarcomatoid mesothelioma. *J. Pathol.* **233**, 402–414 (2014).
36. D. Mellacheruvu, Z. Wright, A. L. Couzens, J. P. Lambert, N. A. St-Denis, T. Li, Y. V. Miteva, S. Hauri, M. E. Sardiu, T. Y. Low, V. A. Halim, R. D. Bagshaw, N. C. Hubner, A. al-Hakim, A. Bouchard, D. Faubert, D. Fermin, W. H. Dunham, M. Goudreault, Z. Y. Lin, B. G. Badillo, T. Pawson, D. Durocher, B. Coulombe, R. Aebersold, G. Superti-Furga, J. Colinge, A. J. R. Heck, H. Choi, M. Gstaiger, S. Mohammed, I. M. Cristea, K. L. Bennett, M. P. Washburn, B. Raught, R. M. Ewing, A. C. Gingras, A. I. Nesvizhskii, The CRAPome: A contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods* **10**, 730–736 (2013).
37. J. Markovic, J. Taylor, J. Taylor, Inference after black box selection. arXiv:1901.09973 (2019).
38. V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 5116–5121 (2001).
39. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).
40. G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, P. S. Linsley, R. Gottardo, MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
41. K. R. Moon, D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. van den Elzen, M. J. Hirn, R. R. Coifman, N. B. Ivanova, G. Wolf, S. Krishnaswamy, Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* **37**, 1482–1492 (2019).
42. G. C. Linderman, J. Zhao, Y. Kluger, Zero-preserving imputation of scRNA-seq data using low-rank approximation. *bioRxiv* (2018).
43. R. D. Cardiff, C. H. Miller, R. J. Munn, Manual hematoxylin and eosin staining of mouse tissue sections. *Cold Spring Harb. Protoc.* **2014**, 655–658 (2014).
44. Z. Du, J.-R. Lin, R. Rashid, Z. Maliga, S. Wang, J. C. Aster, B. Izar, P. K. Sorger, S. Santagata, Qualifying antibodies for image-based immune profiling and multiplexed tissue imaging. *Nat. Protoc.* **14**, 2900–2930 (2019).

Acknowledgments: We acknowledge valuable discussions with L. Karacosta on cell states in lung cancer; G. Bouchard on experimental validation; S. Gonzales on autofluorescence in the IF images; and B. Anchang, D. Knowles, W. Zhang, and A. Bejnood for the discussions on the methodology. We also thank A. Gentles, P. Mallick, and T. Desai for discussions about the biological properties of LR signaling and method development. **Funding:** We acknowledge funding from the National Cancer Institute grant U54CA209971, the National Library of Medicine grant TL15LM007033, and the Tobacco-Related Disease Research Program predoctoral fellowship grant T29DT0760. **Author contributions:** A.Y. and S.K.P. conceptualized the study and methods and wrote the manuscript. A.Y. implemented and tested the method and conducted the computational analysis. Y.L. and I.L. contributed to experimental validation design. Y.L. conducted experimental validation. C.Y. and A.E.C. contributed to computational analysis. M.G.O. provided pathological analyses of the tissue images. J.T. designed and contributed to the method's statistical significance test. W.L.T. received and processed the clinical samples; J.S. collected the clinical samples. **Competing interests:** A.Y. and S.P. are inventors on a pending patent related to this work filed by Stanford OTL no. S21-038, filed June 30. The authors declare that they have no other competing interests. **Data and materials availability:** REMI is available at https://github.com/plevritis-lab/REMI. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 25 March 2021
Accepted 27 January 2022
Published 18 March 2022
10.1126/sciadv.abi4757