An instance-specific causal framework for learning intercellular communication networks that define microenvironments of individual tumors

Xueer Chen[1,2*], Lujia Chen[1,2*], Cornelius H.L. Kürten[3,4,5*], Fattaneh Jabbari[1,2], Lazar Vujanovic[3,4], Ying Ding[6], Aditi Kulkarni[3], Tracy Tabib[6], Robert Lafyatis[6], Gregory Cooper[1,2,4§], Robert Ferris[3,4§], and Xinghua Lu[1,2,4§]

[1]Department of Biomedical Informatics,
[2] Center for Causal Discovery,
[3]Department of Otolaryngology,
[4] University of Pittsburgh Hillman Cancer Center,
[5] Department of Otorhinolaryngology, Head and Neck Surgery, University Hospital Essen, University Duisburg-Essen,
[6] Department of Biostatistics,

* These authors contributed equally to the work.
§ Corresponding authors

Abstract

Cells within a tumor microenvironment (TME) dynamically communicate and influence each other's cellular states through an intercellular communication network (ICN). In cancers, intercellular communications underlie immune evasion mechanisms of individual tumors. We developed an instance-specific causal analysis framework for discovering tumor-specific ICNs. Using head and neck squamous cell carcinoma (HNSCC) tumors as a testbed, we first mined single-cell RNA-sequencing data to discover gene expression modules (GEMs) that reflect the states of transcriptomic processes within tumor and stromal single cells. By deconvoluting bulk transcriptomes of HNSCC tumors profiled by The Cancer Genome Atlas (TCGA), we estimated the activation states of these transcriptomic processes in individual tumors. Finally, we applied instance-specific causal network learning to discover an ICN within each tumor. Our results show that cellular states of cells in TMEs are coordinated through ICNs that enable multi-way communications among epithelial, fibroblast, endothelial, and immune cells. Further analyses of individual ICNs revealed structural patterns that were shared across subsets of tumors, leading to the discovery of 4 different subtypes of networks that underlie disparate TMEs of HNSCC. Patients with distinct TMEs exhibited significantly different clinical outcomes. Our results show that the capability of estimating instance-specific ICNs reveals heterogeneity of ICNs and sheds light on the importance of intercellular communication in impacting disease development and progression.

Communications among cells in a tissue influence their cellular states, leading to a homeostatic state at a given time that impacts the physiological and pathological processes of cells. As an example, cancer cells expressing neoantigens actively modulate the immune environment through intercellular communications(1-6) to evade immune surveillance, and the success of contemporary immune checkpoint inhibitors(7-9) demonstrate the potential for identifying and targeting intercellular communications in treating diseases. Importantly, heterogeneity of genomic alterations among tumors also leads to diverse TMEs(10,11) and thereby to distinct immune evasion mechanisms and heterogeneous responses to immune therapies. Thus, it is important to understand how cells communicate with each other within each tumor and gain insights into tumor-specific immune evasion mechanisms to help guide personalized immune therapies.

Studying cell-cell communication is a long-standing problem (12-14), and the advent of technologies for profiling single cells(15-20) provides unprecedented opportunities to advance the field. An increasing number of studies concentrate on modeling cell-cell communication through mining single cell transcriptome data (20-26). However, the majority of current efforts (see recent reviews by Almet et al(25) and Armingol et al (26)) concentrate on modeling potential interactions of known ligand-receptors, through studying the correlations between the expression values of ligand-receptor pairs. While informative, a knowledge-based approach cannot discover previous unknown interactions; co-expression of a pair of ligand-receptor mRNA does not entail physical interaction between their protein products; and, more importantly, correlation analysis does not provide information of biological consequence of ligand-receptor interactions. Finally, contemporary modeling of cell-cell communication is mostly performed at a population level, where a single model is developed to explain the data of

all cases. Such analysis is not suitable to investigate ICNs underlying the well-appreciated heterogeneity of TMEs observed in tumors(10,11). Thus, there is an unmet need for computational methodologies that enable studying ICNs within individual tumors to better understand how cells communicate within a tumor's TME.

An ICN is essentially a causal network, in which a change in the cellular state of a cell (or a subpopulation of cells) can causally influence the cellular states of other cells via communication channels involving physical or paracrine ligand-receptor interactions. If we can accurately detect and represent the state of cellular processes in cells, we can apply principled causal discovery methods to search for causal relationships among the cellular processes across cells, i.e., an ICN. As a discipline of machine learning and artificial intelligence, the field of computational causal discovery has witnessed significant progress in recent decades (27-29). Particularly, a new family of instance-specific causal discovery methods (30-32) has been developed, which can learn a causal network based on the data of an individual case, be it a single cell or a single tumor.

In this study, we developed and applied an instance-specific Bayesian causal discovery framework to learn tumor-specific ICNs, and we used HNSCC cancer as a testbed for our study. The framework consists of multiple steps as follows. First step, we applied a statistical model to identify GEMs from single cell data, such that expression status of a GEM would reflect the state of a transcriptomic process regulating its expression in a cell or a tumor. Identification of GEMs lays a foundation for modeling how the change of a cellular process in one type of cell (e.g., cancerous epithelial cells) causally influences the state of another process in another type of cells (e.g., CD8+ lymphocytes). Second, we inferred the activation states of the discovered transcriptomic processes in a large number of tumors, so that we could reliably learn causal

relationship of cellular processes across different types of cells. Third, we employed causal Bayesian network (CBN) learning algorithms to search for ICNs among diverse types of cells at the cohort and the individual tumor levels. Our analyses uncovered ICNs that model well the coordinated expression of GEMs across diverse types of cells. By mining the patterns of ICNs of individual tumors, we discovered distinct patterns of ICN structure that underlie disparate TMEs of HNSCCs that are associated with significantly different patient outcomes. Our methods serve as a general tool for studying intercellular communications that influence disease development and progression.

RESULTS

Identification of context-specific GEMs reflective of the states of transcriptomic processes

We collected single-cell RNA sequencing (scRNAseq) data of 109,879 cells from 18 HNSCC tumors and matched peripheral blood samples. After quality control, we retained data on 78,887 $CD45^+$ cells from tumor and peripheral blood samples of the 18 patients, and we collected 19,107 $CD45^-$ cells from 12 of the tumors (**Figure 1a, Figure S1**). Since different cells (e.g., epithelial vs lymphocytes) have significantly different transcriptomes and regulatory processes, we separated cells into 4 major categories and modeled their transcriptomic regulatory process respectively: immune-related (PTPRC/$CD45^+$; 78,887 cells), epithelial (9,852 cells), fibroblast (3,476 cells), and endothelial (5,779 cells), based on the expression of respective marker genes.

Based on the assumption that the transcriptome of each single cell is a mixture of GEMs controlled by different transcriptomic processes, we applied the nested hierarchical Dirichlet processes (NHDP)(33) model to the scRNAseq data to identify GEMs. NHDP is a non-parametric statistical model that can deconvolute a mixture of discrete variables (e.g., words in a

text documents) by searching for modules (topics) that capture co-occurrence of words in text documents. When applied to scRNAseq data from a population of cells, NHDP can discover sets of co-expressed genes (i.e., GEMs), with each GEM consisting of genes that are repeatedly co-expressed in many cells. As such, the expression status of a GEM in a cell reflects the state of a transcriptomic process that regulates the GEM. NHDP is a hierarchical model that represents the joint distribution of GEMs (i.e., co-occurrence patterns of GEMs in cells) using a tree-structured hierarchy. It can represent the differentiation (evolution) processes of cells in a cell population, such that a GEM close to the root of a tree tends to be expressed broadly among the majority of cells, representing a general transcriptomic process shared by cells of a common lineage. In contrast, a GEM close to the leaves of the tree tends to be expressed in a subpopulation of cells, reflecting the state of transcriptomic processes that lead to differentiation of a subtype of cells.

Applying the NHDP model to pooled peripheral and TIL $CD45^+$ cells, we identified 39 GEMs (referred to as immune GEMs) and their signature genes, and we examined the expression patterns of each immune GEM among the $CD45^+$ cells (**Figure 1, Table S1** and **Figure S2**). As an example, immune GEM1 is broadly expressed in T cells, B cells, and natural killer (NK) cells, and thus, it represents the transcriptomic process that is shared by cells from the lymphoid lineage; immune GEM2 is mainly expressed in cells of myeloid (monocyte/macrophage) lineage; immune GEM3 is concentrated in tumor infiltrating lymphocytes (TILs), potentially reflecting the impact of tissue environment on T cells (34); GEM14 is mainly expressed in $CD8^+$ T cells, which contain genes commonly expressed in activated cytotoxic T cells (such as *CD8A, GZMB, CXCL13, IFNG, GZMA, GNLY, NKG7,* and *PRF1*), suggesting that this GEM is likely regulated by a T-cell activation processes. Interestingly, certain GEMs can be expressed in cells across conventionally defined immune cell lineages. For example, GEM32 is shared among T cells and

NK cells from peripheral blood (**Figure 1f, Figure S2**), indicating that our approach can discover shared transcriptomic processes among cells that otherwise have distinct characteristics.

In addition to PTPRC/CD45$^+$ cells, we identified GEMs among epithelial, fibroblast, and endothelial cells respectively as well. Similarly, some cell-type-specific GEMs represent lineage-defining transcriptomic processes, and some are unique to a small subset of cells reflecting different activation states (**Figure S2**). In summary, the NHDP model can identify GEMs that reflect the state of transcriptomic processes in individual cells, enabling us to represent each cell in the space of transcriptomic processes. This representation not only provides a more concise and robust representation of cellular state than is possible with individual genes, but it also serves as an instrument for modeling the causal relationships of transcriptomic processes across cell boundaries.

Different combinations of active transcriptomic programs define distinct cell subtypes

We set out to examine whether representing cells based on their active transcriptomic processes (expression status of GEMs) can reveal cell subtypes based their cellular states in contrast to representing cells in a gene space. We performed clustering analyses to identify subtypes within each of the following major immune-cell categories: CD8$^+$, CD4$^+$, and T regulatory cells (Treg), B cells, NK cells, monocytes and macrophages (**Figure S3**). As an example, eight subtypes (clusters) of CD8$^+$ T cells were identified (**Figure 2a**), with each exhibiting a distinct composition of GEMs. Pseudo-time analysis of CD8$^+$ T cells represented in GEM space revealed a development path of these CD8$^+$ clusters, starting from "naïve" T cells from peripheral blood, to "primed/activated" tumor infiltrating CD8$^+$ T cells, and further onto "exhausted" CD8$^+$ T cells (**Figure 2b**). Interestingly, the composition of GEMs expressed in

these subtypes can be represented as a tree (**Figure 2c**) that mimics the development path of $CD8^+$ T cells: All subtypes expressed GEM1, which represents a transcriptomic process shared by cells of the lymphoid lineage; expression of GEM3 differentiates tumor infiltrating and peripheral $CD8^+$ T cells; expression of other GEMs further leads to differentiation of cells into subtypes. Unlike conventional clustering analysis based on overall transcriptomics, our approach clearly delineated the transcriptomic processes that contributed to such development. For example, cluster CL6 constitutes a subtype of tumor infiltrating $CD8^+$ T cells that followed a unique trajectory (**Figure 2b**), and the expression of GEM7 appears to underlie the development of this subtype (**Figure 2c**). Thus, inspecting the GEM composition of each cell cluster not only reveals distinct patterns of activation states of transcriptomic processes in cells, but also illustrates how a subtype of cell develops. Using the same approach, we have identified cell subtypes (clusters) among each major PTPRC/$CD45^+$ cell category, as well as among epithelial, fibroblast, and endothelial cells (**Figure S3**).

Distinct patterns of active transcriptomic processes and cell composition revealed heterogeneous immune environments of HNSCCs

Based on the assumption that the expression status of a GEM in a tumor would reflect the presence and activation status of transcriptomic process regulating the GEM, we set out to detect the activity of transcriptomic processes by deconvoluting bulk RNA sequencing data of HNSCC tumors profiled by TCGA. Using the top 50 genes of a GEM (ranked according to the probabilities being assigned to the GEM) as a signature profile of the GEM, we employed gene set variation analysis (GSVA) (35) to determine the expression status of the GEM in each tumor. The GSVA method calculates a sample-wise enrichment score of genes within a GEM and

further estimates the variance of GEM enrichment scores across a cohort of samples. Thus, it infers the relative expression status of each immune GEM, i.e., the relative activation status of the corresponding transcriptomic process in each tumor (**Table S2a**). We performed GSVA analysis of immune GEMs using the bulk RNA data of 522 HNSCC tumors profiled by TCGA. Similarly, we used the gene expression signature of each subtype of immune cell to infer their relative enrichment scores in the tumors (**Table S2b**).

We then performed consensus clustering analyses on the tumors using as features the GSVA scores of immune GEMs, and alternatively, the cell subtypes. Both analyses revealed 4 distinct immune environments, with each cluster consisting of distinct compositions of GEMs (**Figure 3a**) or cell subtypes (**Figure 3b**), wherein the members of the clusters (defined by GEMs and cell subtypes, respectively) significantly overlapped. Survival analyses showed significant differences in patient outcomes between the clusters, supporting the impact of the modeled immune environment on disease progression.

To identify the transcriptomic processes impacting survival, we performed Cox proportional hazard analyses, using the GSVA scores of the GEMs (or the cell subtypes) as independent variables. When only considering the impact of an individual GEM (**Table S3**) and immune cell subtype (**Table S4**) on survival, 20 GEMs and 17 cell subtypes were deemed to be significantly associated with overall survival of patients ($p < 0.05$). We further performed a forward-and-backward search to investigate the impact of combinations of these GEMs (or cell types) on survival, which could detect and eliminate redundant GEMs or cell subtype, and 7 GEMs and 6 immune-cell subtypes (**Figure 3c and 3d)** were statistically significantly associated with patient survival in the multi-variate Cox proportional hazard analyses. The results of GEM-based and cell-subtype-based analyses are consistent. For example, the expression of immune

GEM7 is concentrated in cluster CL6 of $CD8^+$ T cells (**Figure 3e** and **Figure 2a**), and the enrichment of the GEM7 or cells from cluster CL6 were deemed as significantly associated with increased hazard ratio; similarly, expression of immune GEM18 is concentrated in cluster CL6 of monocytes (**Figure 3f** and **Figure S3**), and their enrichment in tumors were both significantly associated with an increased hazard ratio. The results support that HNSCC tumors exhibit distinct patterns of immune environments, which are associated with patient outcomes.

Transcriptomic programs of cells are coordinated in TMEs

The capability of detecting the states of transcriptomic processes in cells and individual tumors enabled us to further investigate whether the transcriptomic programs in different cell types are coordinated. We first performed pairwise correlation analyses between GEMs (**Table S5)** across the four major cell types: epithelial, immune, endothelial, and fibroblast cells (**Figure 4a** and **Figure S4)**. The results showed that expression of GEMs from different types of cells are significantly correlated, supporting the notion that intercellular communication may underlie coordinated activation/inactivation of transcriptomic processes across major cell types. To investigate the impact of cancer and stromal cells on the states of immune cells, we further trained regression models using GSVA scores of the non-immune cell GEMs as independent variables to predict expression of each of the immune GEMs in 10-fold cross-validation experiments. The results show that the expression status (reflected as GSVA scores) of almost all immune GEMs could be accurately predicted (**Figure 4b**) based on those of non-immune cells. The mean and standard deviation (SD) of R^2 of predicted and observed values of 39 immune GEMs were 0.75 and 0.11 respectively, and p-values for all regression R^2 (except GEM33) were less than 0.001. The results provide evidence that transcriptomic programs of cells in HNSCC

TMEs are coordinated, and such coordination may be the result of causal intercellular communications, may confounded by unknown factors, or both, which can be resolved by future causal analyses.

A global ICN of HNSCC identified by causal Bayesian network learning

To go beyond the correlation analysis, we next set out to find direct support for general causal relationships between the transcriptomic processes, i.e., discovering an ICN. We employed the greedy fast causal inference (GFCI) (36) algorithm, which is capable of learning causal relationships from observational data, including the possibility of latent (hidden) confounding when two or more variables are caused by some hidden process. The GFCI algorithm takes a set of measured variables (e.g., GSVA scores) as input data and returns a partial ancestor graph (PAG) (37), which represents the causal relationships among the measured variables. In our case, a direct causal edge from *GEM_A* to *GEM_B* across cell types indicates that the state of the transcriptomic process regulating *GEM_A* in a subpopulation of cells is inferred to be causally influencing the state of the transcriptomic process regulating *GEM_B* in another subpopulations of cells.

Using the GSVA scores of GEMs from TCGA HNSCC tumors as inputs to GFCI algorithm, we first learned a consensus PAG (see Methods). During the process, we constrained the GFCI algorithm to exclude the candidate edges between the GEMs that were known to be co-expressed in same subpopulation of cells, and thus, the returned CBN captured the causal of relationships between GEMs expressed across cell subpopulations. **Figure 4c** shows a subgraph of the consensus PAG, and the complete causal network is shown in **Figure S5** and **Table S6.** **Figure 4c** uses the immune GEM16 to illustrate causal relationships with other GEMs, where it

is shown at the center surrounded by a set of neighboring GEMs (parents, children, and parents of children, which are commonly referred to as Markov blanket). We examined the directions of the causal edges between major cell categories to provide an overview of how different cells communicate with each other in tumor microenvironment of HNSCCs (**Table 1**), according to our model. The results indicate dense and multi-way intercellular communications among cells in TMEs.

Since experimentally validating the discovered causal edges is beyond the scope of this study, we set out to mathematically validate that the learned ICN precisely captured statistical relationships among the transcriptomic processes in diverse types of cells. Conditional independence plays a central role in discovering causal relationships(27-29). In a causal Bayesian network, the value of a variable is independent of the remainder of the variables in the network conditioned on the values of its Markov boundary, or in other words, the Markov boundary variables serve as the smallest sufficient set of variables for optimally predicting the state of the GEM of interest. The better we can predict the expression status of a GEM based on the expression status of GEMs in its Markov boundary, the more support (although not proof) it provides for the causal edges within the Markov boundary (as shown in **Figure 4c**) being valid(38). For each immune GEM, we identified its Markov boundary GEMs and used their expression status (GSVA scores) as independent variables to train a regression model to predict the expression status of the immune GEM of interest. Most immune GEMs could be predicted well based on the expression of a few neighboring GEMs discovered by GFCI, with R^2 ranging from 0.36 to 0.97 and the mean and standard deviation of the R^2 values being 0.67 and 0.16 respectively (**Figure 4d**), and the p-values of regressions (except GEM23) are less than 0.001. Thus, the PAG resulting from our analysis provides a parsimonious representation of an ICN that

serves as a supported hypothesis of the communications between the cells expressing specific GEMs.

Tumor-specific ICNs of individual tumors underlie heterogeneous immune environment of HNSCC

While an ICN trained at the tumor cohort level depicts well the general causal relationships among transcriptomic processes, it does not capture the idiosyncratic characteristics of intercellular communications of individual tumors, which we conjecture underlie the diverse TMEs as shown in Figure 3. We then applied an instance-specific causal discovery algorithm previously developed by us, which we call the instance-specific GFCI (iGFCI) algorithm (31) (see Methods) and discovered a tumor-specific ICN for each of 522 HNSCC tumors.

Inspecting the individual ICNs, we note that while each ICN is unique, certain patterns of network structure were conserved among subpopulations of tumors. We represented each tumor-specific network with a vector spanning the union of discovered causal edges from tumors, in which element i =1 if and only if that network contained edge i. We then performed consensus clustering analyses of these networks. Four distinct clusters of networks emerged (**Figure 5a**), with certain edges being distinguishably present or absent from some clusters of tumors (**Figure 5b**). The patients from the 4 subtypes exhibited significantly different survival outcomes (**Figure 5c**), and presence/absence of the certain edges were significantly associated with patient outcomes (**Figure 5d**).

We further investigated whether the distinct ICN patterns were associated with distinct immune environments, to explore the hypothesis that an immune environment is defined by a subtype type of ICN. We compared the overlap of cluster membership of tumors when they

were clustered based on immune GEMs (Figure 3a) and causal network edges (Figure 5a), respectively; we found that the tumor clusters derived from two the clustering analyses share a significant number of tumors (**Figure S6**) (chi-square test *p*-value < 2.2E-16). Particularly, the Network Cluster # 1 closely aligns with the GEM-based Cluster # 4, and patients assigned to these clusters have the best outcomes compared to patients assigned to the other clusters; similarly, Network Cluster #2 aligns best with the GEM-based Cluster #2, and the patients assigned to these clusters exhibited the worst outcomes compared to patients assigned to the other clusters. Thus, the results supports the notion that distinct patterns of tumor-specific ICNs underlie different immune environments detected from transcriptomic profiling.

Discussion

In this study, we have developed a framework for inferring tumor-specific ICNs, which involves two major conceptual steps: Firstly, mining single cell transcriptomes to identify GEMs to serve as surrogate representations of cellular signaling processes. Secondly, modeling intercellular communication as a causal network between these processes to infer intercellular communication at the level of individual tumors. We anticipate that the capability of inferring ICN, particularly tumor-specific ICNs, will lay a foundation for future mechanistic investigations of how intercellular communications define the state of cells in a TME, and in turn, how tumors evade the immune surveillance of hosts.

While it is well-appreciated that cells within a TME actively communicate and influence the functional states of each other (19-21), few computational methods have been developed to analyze the fundamental nature of an ICN, especially the causal effects of communication. Our work represents an effort in this direction. Our results show that the inferred ICNs serve as

concise statistical models that explain well the coordinated transcriptomic programs across different types of cells and that tumor-specific ICNs provide a new approach to investigate the heterogeneity of immune environments in cancer. Based on the assumptions of Bayesian causal inference, we believe that certain communication channels (directed edges) of learned ICNs capture causal effects of intercellular communications, beyond accurately modeling the joint distribution of cellular events (controlled expression of GEMs) in different cells. However, we realize that those assumptions may not always hold in a complex system as the ones we are studying, and thus, we view the learned ICNs as hypotheses, which can be experimentally tested in future.

In this study, we illustrated the capability and utility of our methods in studying ICNs in the cancer (HNSCC) domain. The principled framework introduced here is quite general and can be applied in wide range of biomedical domains to study communications of cells in a tissue or organ. With the advent of more single-cell omics data, particularly in situ spatial omics studies of tissues (15,39,40), the framework we present can be used to study intercellular communications under a wide range of physiological (e.g., embryo development) and pathological (e.g., rheumatism or aplastic anemia) settings besides cancer, which provides many opportunities for future research.

Methods

Single cell RNA sequencing data processing

FASTQ files were processed with the CellRanger software (10X Inc, version 2.0). We aggregated the 46 samples of all the 18 patients, including samples from CD45+/- tumor specimen and peripheral blood. After the aggregation and default depth normalization using the CellRanger, we collected the count matrix of 103,006 cells in total, with a post-normalization mean reads per cell of 31,303 and median genes per cell of 1,005. We then processed the aggregated count matrix in Scanpy (41). We filtered cells and genes with cells expressing at least 200 and at most 5000 genes. Cells with more than 90 percent of genes belonging to mitochondria genes were removed. Genes were then filtered with dispersion range (minimum mean of 0.0125, maximum mean of 5 and minimum dispersion of 0.1). Scanpy also provides the implementation of feature reduction and visualization called UMAP (42,43). We visualized all the cells using in UMAP in which cells are pseudo-colored to mark cell types based canonical gene signatures of general cell types or manual labels.

NHDP modeling on single cell RNAseq data

We applied the NHDP on scRNAseq data to identify the gene modules which potentially indicate different biological processes. The NHDP method allows users to define a hierarchical structure and the algorithm will allocate modules along the tree. In this study, we designed a three-layer hierarchical tree structure, with branching factors of 5, 4, and 3 from the root to the 2nd and 3rd layers. That is, starting from the root, the 1st layer had 5 nodes; each node in the 1st layer is connected to 4 children nodes in the 2nd layer; each node in 2nd layer is connected to 3 children nodes in the 3rd layer. Each node defines a distribution over the space of genes reflecting the

information on which genes are commonly assigned to the module. Input to the NHDP model is a binarized expression matrix of scRNAseq data, where rows are individual cells and columns are genes. An element of the matrix was set to "1" if the corresponding gene is expressed in a given cell, and "0" otherwise. The algorithm generates two outputs: 1) For each node in the hierarchical tree, a distribution over the space of genes is returned. From the distribution, one can extract a ranked list of genes most commonly assigned to the node (GEM). 2) A cell-by-GEM count matrix is returned. An element of the matrix reflects the number of genes expressed in a given cell that are assigned to a specific GEM. This is equivalent to projecting a cell into the space of GEMs, which enabled us to search cell subtypes through clustering analysis.

Enrichment of GEMs in bulk RNAseq and tumor subtype identification

We analyzed the enrichment of all the immune and non-immune GEMs in three bulk RNAseq datasets, namely, TCGA HNSCC tumors and two independent validation sets from the GEO database (GSE39366, GSE27020) (44). Genes within a GEM are ranked according to the probability mass assigned to them by the NHDP model, and we used the top 50 genes of a GEM to represent its signature co-expression profile. We applied the gene set variation analysis (GSVA) method (35) to estimate the enrichment of a GEM in a tumor relative to a cohort of tumor samples. The distribution of the GSVA scores among the input tumor samples generally followed a normal distribution in the range of -3 to 3.

With the estimated GSVA scores for all the immune GEMs and cell subtypes in TCGA HNSCC tumor samples, we applied consensus clustering (implemented in the R package as ConsensusClusterPlus (45)) and identified subtypes of tumor samples with distinct enrichment

patterns of immune GEMs. The optimal number of clusters were chosen using the relative change in area under the empirical cumulative density function (CDF) curve (46), which shows the relative increase in consensus and cluster number at which there is no appreciable increase.

In addition to GEMs, we also calculated the GSVA enrichment scores of immune cell subtypes, which were identified in **Figure S3**. The signatures of each cell subtype were curated based on the GEMs that were enriched in the corresponding cell subtypes; in particular, we used the union set of top30 genes of the GEMs enriched in individual cell subtypes. Following the same procedure, we estimated the cell-subtype enrichment GSVA scores in bulk RNA data and performed consensus clustering to identify tumor clusters with distinct composition of cell subtypes.

Evaluate association between patient survival and GEMs or cell subtypes. We performed Cox Proportional-Hazards regression (*coxph* function of R language) using both the immune GEMs and cell subtypes enrichment scores as input to evaluate which GEMs or cell subtypes were significantly associated with patient survival outcomes. We conducted both individual Coxph regression on individual immune GEMs and alternatively on cell types. To do so, we used a stepwise selection of multiple variates to obtain a minimum and optimized set of immune GEMs or cell types that could fit the survival curves well.

Cross-cell-type GEM correlations analysis

We examined the correlations among the enrichment scores of GEMs with both Pearson correlation and multiple regression models. Across all cell categories (immune, epithelial, fibroblast, and endothelial), we performed Pearson correlation analysis of the expression levels between all pairs of GEMs. We excluded the intra-cell-category correlation analysis to avoid

potential confounding effects of coordinated cell-type-specific expression programs. We also applied the GLMNET (47) package from the R language to train regression models, which predict each immune GEM using non-immune GEMs from epithelial, endothelial, and fibroblast cells as independent variables. We performed 10-fold cross-validation experiments, in which we used 9 folds of data to train a regression model and used the trained model to predict immune GEMs for the tumors in the held-out fold. We then evaluated the performance of the regression models by performing a correlation analysis using predicted GEM-expression values versus observed GEM expression values and recorded R^2 values.

Discovering population-wide causal relations among GEMs with GFCI

We employed the GFCI algorithm to infer from data whether one GEM is supported by the data (under assumptions) as causally influencing the expression of another GEM, particularly across cell types. For estimating a population-wide ICN, we applied the GFCI algorithm implemented in the Tetrad Package (https://cloud.ccd.pitt.edu/nexus/content/repositories/releases/edu/cmu/tetrad-gui/), which is available from the Center for Causal Discovery (www.ccd.pitt.edu) at the University of Pittsburgh and Carnegie Mellon University. GFCI (36) is a general causal inference algorithm which can model the possibility that there are latent variables that confound the statistical relationships among the measured variables. This capability is of particular importance when we model the causal relationships among all GEMs, because certain GEMs may exhibit cell-type-specific expression patterns, and thus, the cell type may be a confounding factor that leads to GEMs being correlated. On the other hand, there may be patterns of association among GEMs that strongly support that one GEM causes the expression of another without any latent confounding; we are most interested in analyzing these results for insights. The output of the GFCI algorithm is

a partial ancestral graph (PAG) (37,48), which contains multiple types of causal edges including those that indicate causation and others that indicate latent confounding.

The Tetrad implementation of the GFCI algorithm allows a user to input prior knowledge regarding the network, such as edges to be included, as well as those to be excluded. To avoid confounding effects of cell-type-specific co-expression of GEMs, we first examined the distribution of GEMs in cells based on the results shown in **Figure S2** to identify a list of pairs of GEMs that were co-expressed in a set of cells. This list was input to GFCI as prior knowledge, so that the algorithm would exclude edges between a pair of GEMs in the list. Using the GEM data of TCGA HNSCC tumors as inputs, we produced 50 bootstrap samples by drawing 90% of the tumors each time with replacement; we applied the GFCI algorithm to search for a causal network using each bootstrap sample. GFCI scored the causal relationships among GEMs by applying the Bayesian information criterion (see "sem-bic" in the Tetrad manual) and the Fisher-*z*-test. We created a consensus causal PAG by keeping the edges that were conserved more than 20% of the time in the bootstrap results.

Discovering tumor-specific intercellular communication networks

To investigate the mechanisms underlying the heterogeneous immune environments among tumors, we applied an instance-specific version of GFCI (iGFCI) method developed by (31) to identify causal graphs for individual tumors. The inputs for iGFCI analysis are GEM enrichment scores in bulk TCGA HNSCC tumors.

Instance-specific GFCI (iGFCI) is a causal learning algorithm that is based on the GFCI algorithm. GFCI [1] learns from observational data a partial ancestral graph (PAG), which is a

causal graphical model with the edge types shown in Figure S5. iGFCI is an instance-specific version of GFCI. An *instance* is a sample (e.g., a tumor) for which a set of variables have a specific set of values, which are called *features T*. An instance-specific PAG captures context-specific independence (CSI) [2] of the causal relationships among the measured variables. In particular, the parents of a node X can change, depending on the values of the other measured variables in the instance, as given by T.

Figure S7a shows an example of a causal Bayesian network (CBN). Suppose that $T = \{X_1 = 1, X_2 = 1, X_3 = 0\}$. For this instance T, Figure 1a indicates that X_4 is independent of X_3 given $X_1 = 1$ and $X_2 = 1$; this independence implies that for this instance, the parents of X_4 are just X_1 and X_2, as shown in Figure 1b. Figure S7 shows the PAG that can be learned in the large sample limit from data generated by the CBN shown in Figure 1b; with additional measured variables or background knowledge, it may be possible to further resolve the causal relationships among these variables.

iGFCI takes as input a set D of training instances and a test instance T, and it returns as output an instance-specific PAG P_{IS}. The iGFCI algorithm operates in two steps. In the first step (line 1 in Algorithm 1 below), it applies the population-wide GFCI algorithm [1] using dataset D. By performing GES search using the BDeu score [3], GFCI initially learns a population-wide CBN, which we denote as G_{PW}. GFCI then uses G_{PW} as the initial adjacency graph and performs additional conditional independence tests to further prune the adjacency structure using a Bayesian independence test called BSC [4]. Finally, GFCI applies at set of orientation rules to obtain a population-wide PAG P_{PW}. In the second step (line 2 in Algorithm 1), iGFCI applies an instance-specific version of GFCI. For the score-based part, it applies the GFCI algorithm with an instance-specific Bayesian score called IS-Score [5] to learn an instance-specific CBN given D, T, and the

population-wide CBN G_{PW} from the first step. For the conditional independence testing part, it applies an instance-specific BSC test called IS-BSC [6] to find an instance-specific PAG P_{PW} given D, T, and G_{PW}; we use GFCI2 to denote this instance-specific application of GFCI.

Algorithm 1 iGFCI ($\boldsymbol{D}, \boldsymbol{T}$)

Input: a dataset $\boldsymbol{D}$ and a test case $\boldsymbol{T}$

Output: an instance-specific PAG $\boldsymbol{P_{IS}}$

1: $\boldsymbol{G_{PW}}, \boldsymbol{P_{PW}} \leftarrow$ GFCI ($\boldsymbol{D}$)

2: $\boldsymbol{G_{IS}}, \boldsymbol{P_{IS}} \leftarrow$ GFCI2 ($\boldsymbol{D}, \boldsymbol{T}, \ \boldsymbol{G_{PW}}$)

3. return $\boldsymbol{P_{IS}}$

Finding patterns among tumor-specific ICNs

We used bootstrap sampling to enhance the robustness of the results generated by iGFCI. In particular, for each tumor, we applied iGFCI to 20 bootstrap datasets of the original data, where each dataset was obtained by 90% sampling of the original dataset with replacement. For a given tumor T, we then identified the conserved edges among the 20 bootstrapping results. In particular, we kept the edges that were present in more than two bootstrap-derived causal graphs for each of the population causal graph and the causal graph for tumor T. The union of the retained causal edges yielded a "consensus" intercellular communication network for each tumor T.

We then clustered the tumor samples based on their individual ICNs learned using iGFCI to obtain novel subtypes of HNSCC with different cellular communication patterns. Each tumor is represented as a vector over the space of the union of causal edges of the tumor-specific intercellular communication networks. An element of the vector of a tumor T is set to "1" if the causal edge is present in the tumor-specific intercellular communication network for T, and to "0"

otherwise. We applied the ConsensusClusterPlus package from the R Bioconductor Consortium to cluster these vectors.

References

1. Spranger S, Bao R, Gajewski TF. Melanoma-intrinsic beta-catenin signalling prevents anti-tumour immunity. Nature **2015**;523:231-5
2. Collins DC, Chenard-Poirier M, Lopez JS. The PI3K Pathway at the Crossroads of Cancer and the Immune System: Strategies for Next Generation Immunotherapy Combinations. Curr Cancer Drug Targets **2018**;18:355-64
3. Concha-Benavente F, Ferris RL. Oncogenic growth factor signaling mediating tumor escape from cellular immunity. Curr Opin Immunol **2017**;45:52-9
4. Seton-Rogers S. Oncogenes: Driving immune evasion. Nat Rev Cancer **2018**;18:67
5. Spranger S, Gajewski TF. Mechanisms of tumor cell–intrinsic immune evasion. Annual Review of Cancer Biology **2018**;2:213-8
6. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC*, et al.* A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. Cell **2018**;175:984-97 e24
7. Robert C. A decade of immune-checkpoint inhibitors in cancer therapy. Nat Commun **2020**;11:3801
8. Hodi FS, O'Day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB*, et al.* Improved survival with ipilimumab in patients with metastatic melanoma. N Engl J Med **2010**;363:711-23
9. Robert C, Ribas A, Wolchok JD, Hodi FS, Hamid O, Kefford R*, et al.* Anti-programmed-death-receptor-1 treatment with pembrolizumab in ipilimumab-refractory advanced melanoma: a randomised dose-comparison cohort of a phase 1 trial. Lancet **2014**;384:1109-17
10. Chen DS, Mellman I. Elements of cancer immunity and the cancer-immune set point. Nature **2017**;541:321-30
11. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH*, et al.* The Immune Landscape of Cancer. Immunity **2018**;48:812-30 e14
12. Nishida-Aoki N, Gujral TS. Emerging approaches to study cell-cell interactions in tumor microenvironment. Oncotarget **2019**;10:785-97
13. Schwager SC, Taufalele PV, Reinhart-King CA. Cell-Cell Mechanical Communication in Cancer. Cell Mol Bioeng **2019**;12:1-14
14. Song D, Yang D, Powell CA, Wang X. Cell-cell communication: old mystery and new opportunity. Cell Biol Toxicol **2019**;35:89-93
15. Burgess DJ. Spatial transcriptomics coming of age. Nature reviews Genetics **2019**;20:317
16. Eng CL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y*, et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. Nature **2019**;568:235-9
17. Lasken RS. Single-cell sequencing in its prime. Nature biotechnology **2013**;31:211-2
18. Ray LB. Single-cell biology. Cells go solo. Introduction. Science **2013**;342:1187
19. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S*, et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell **2017**;171:1611-24 e24
20. Cillo AR, Kurten CHL, Tabib T, Qi Z, Onkar S, Wang T*, et al.* Immune Landscape of Viral- and Carcinogen-Driven Head and Neck Cancer. Immunity **2020**;52:183-99 e9

21. Kumar MP, Du J, Lagoudas G, Jiao Y, Sawyer A, Drummond DC*, et al.* Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. Cell Rep **2018**;25:1458-68 e4
22. Hou R, Denisenko E, Ong HT, Ramilowski JA, Forrest ARR. Predicting cell-to-cell communication networks using NATMI. Nat Commun **2020**;11:5011
23. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat Protoc **2020**;15:1484-506
24. Camp JG, Sekine K, Gerber T, Loeffler-Wirth H, Binder H, Gac M*, et al.* Multilineage communication regulates human liver bud development from pluripotency. Nature **2017**;546:533-8
25. Almet AA, Cang Z, Jin S, Nie Q. The landscape of cell–cell communication through single-cell transcriptomics. Current Opinion in Systems Biology **2021**;26:12-23
26. Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell-cell interactions and communication from gene expression. Nature reviews Genetics **2021**;22:71-88
27. Pearl J, Glymour M, Jewell N. Causal Inference in Statistics: A Primer Wiley; 2016.
28. Zhang K, Scholkopf B, Spirtes P, Glymour C. Learning causality and causality-related learning: some recent progress. Natl Sci Rev **2018**;5:26-9
29. Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. . Frontiers in Genetics **2019**;10:524
30. Cai C, Cooper GF, Lu KN, Ma X, Xu S, Zhao Z*, et al.* Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference. PLoS computational biology **2019**;15:e1007088
31. Jabbari F, Visweswaran S, Cooper GF. Instance-specific Bayesian network structure learning. Proceedings of Machine Learning Research **2018**;72:169-80
32. Johnson A, Cooper GF, Visweswaran S. Patient-Specific Modeling with Personalized Decision Paths. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium **2020**;2020:602-11
33. Paisley J, Wang C, Blei D, Jordan. M. Nested hierarchical Dirichlet processes. IEEE transactions on pattern analysis and machine intelligence **2015**;37:256-70
34. Szabo PA, Levitin HM, Miron M, Snyder ME, Senda T, Yuan J*, et al.* Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. Nat Commun **2019**;10:4706
35. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics **2013**;14:7
36. Ogarrio J, Spirtes P, Ramsey J. A hybrid causal search algorithm for latent variable models. JMLR Workshop and Conference Proceedings **2016**;52:368-79
37. Zhang J. Causal reasoning with ancestral graphs. Journal of Machine Learning Research **2008**;9:1437-74
38. Ramsey JD, Zhang J, Spirtes P. Adjacency-Faithfulness and Conservative Causal Inference. 22nd Uncertain in AI. Cambridge, MA2006.
39. Morf J, Wingett SW, Farabella I, Cairns J, Furlan-Magaril M, Jimenez-Garcia LF*, et al.* RNA proximity sequencing reveals the spatial organization of the transcriptome in the nucleus. Nature biotechnology **2019**;37:793-802

40. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR*, et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science **2019**;363:1463-7
41. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome biology **2018**;19:15
42. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426 **2018**
43. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IW, Ng LG*, et al.* Dimensionality reduction for visualizing single-cell data using UMAP. Nature biotechnology **2019**;37:38
44. Walter V, Yin X, Wilkerson MD, Cabanski CR, Zhao N, Du Y*, et al.* Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes. PloS one **2013**;8:e56823
45. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics **2010**;26:1572-3
46. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine learning **2003**;52:91-118
47. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw **2010**;33:1-22
48. Spirtes P, Glymour C, Scheines R, Heckerman D. Causation, prediction, and search.: MIT Press; 2000.

Tables

From \ To	ENDO	EPI	FIB	IMM
ENDO	25	63	24	38
EPI	57	96	35	63
FIB	29	34	20	26
IMM	28	41	9	90

Table 1. Statistics of directed causal edges between different cell types. Rows indicate cell types (endothelial, epithelial, fibroblast, and immune cells) from which edges emit and columns represent cell types to which edges point. A number in the table represents the count of edges between two types of GEMs.

Figures and legends

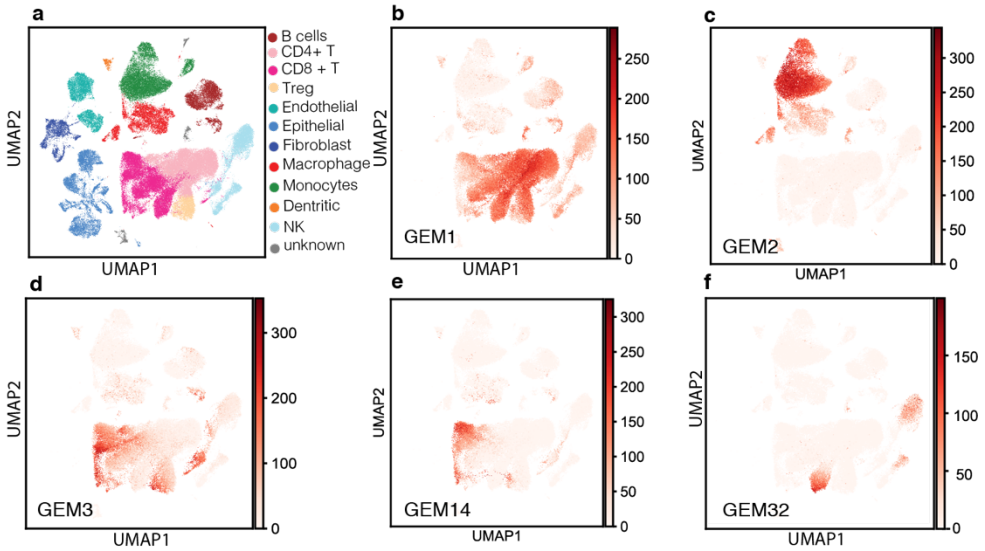

Figure 1. a. UMAP plot of all cells collected from 17 HNSCC tumors. Cells were represented in the original transcriptome and were projected to 2-D space using the UMAP method. Pseudocolors illustrate the cell types. **b – f. Distribution of example immune-cell GEMs among CD45^{+} cells**. The sum of the probabilities of the genes expressed in a cell being a member of an indicated GEM is represented as pseudocolors.

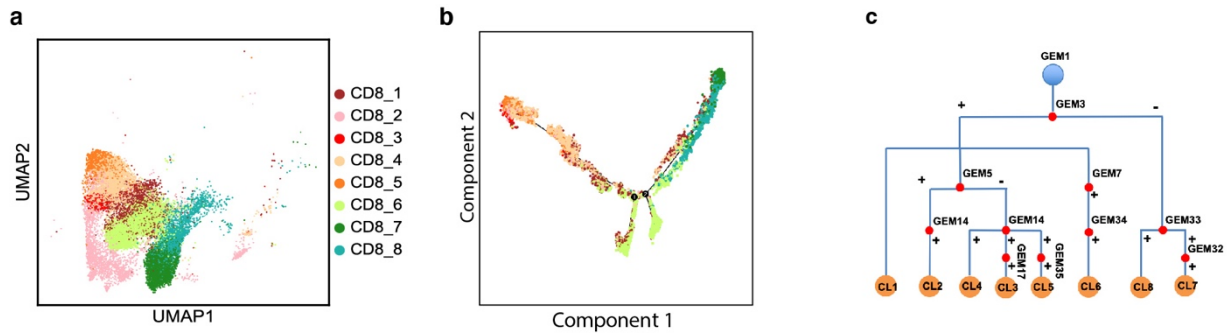

Figure 2. CD8$^+$ T cell subtyping based on GEM compositions. a. UMAP plot of CD8$^+$ T cells. Eight subtypes were identified, which are indicated by pseudo-colors. **b. Pseudo-time plot reflecting evolution of subtypes of CD8$^+$ T cells** (cell subtype indicated with the same pseudo color scheme as in **a**). **c. A hierarchical tree reflecting GEM composition of cell subtypes.**

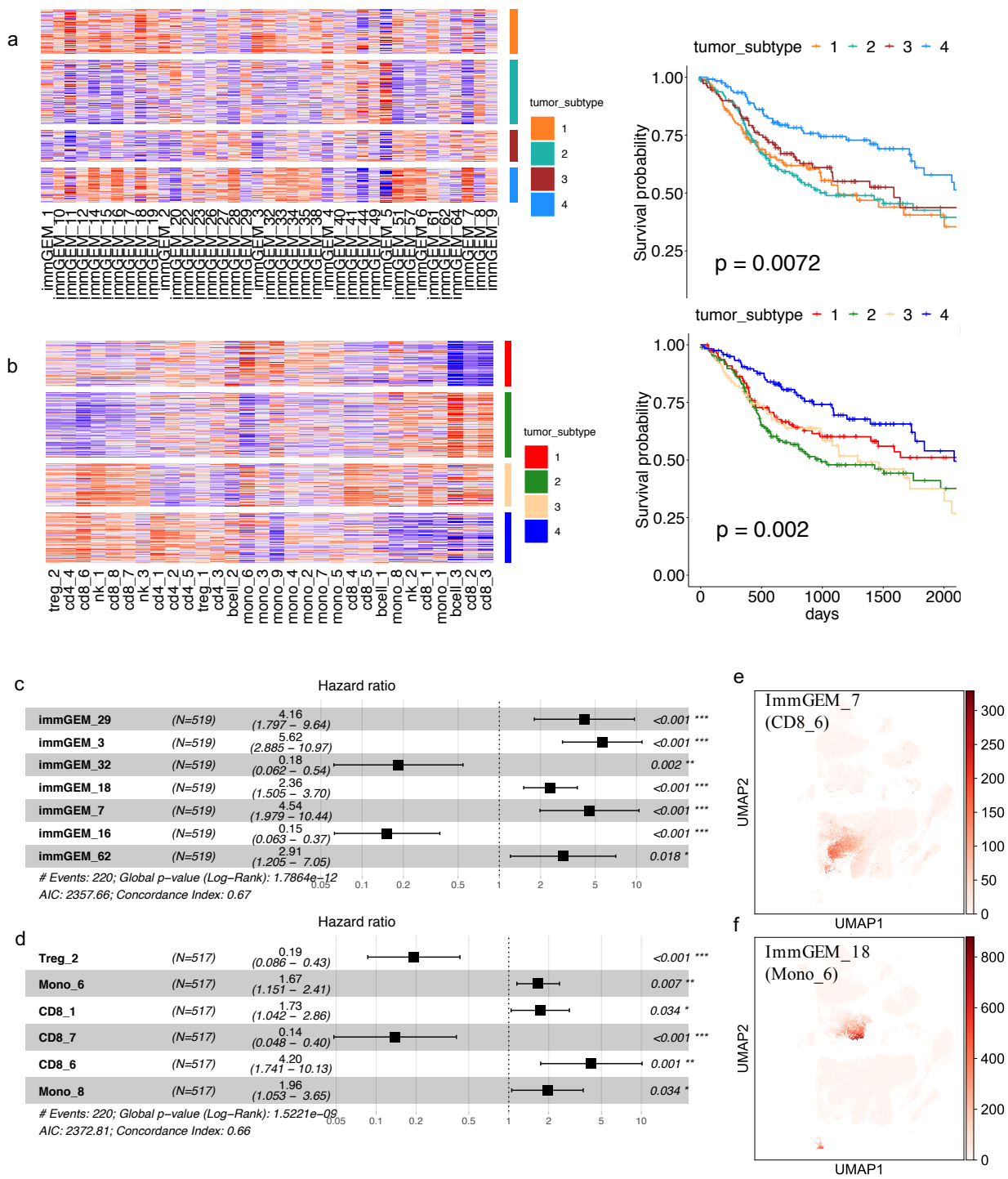

Figure 3. Distinct immune environment of HNSCCs defined by GEMs and cell subtypes. a. Heatmap reflecting relative expression levels of GEMs (columns) among TCGA tumors (rows). Tumors are organized according to the clustering results, with color bars that indicate the

tumor subtype cluster id. Kaplan-Meier curves of patients belonging to different clusters are shown on the right, using the same color scheme ids. **b.** Heatmap reflecting relative enrichment of immune cell subtypes among TCGA HNSCC tumors. Tumors are organized according to clustering results, and Kaplan-Meier curves of patients belonging to different clusters are shown on the right. **c.** Immune GEMs (with id indicated by the number) that are associated with statistically significant hazard ratios. **d.** Immune cell subtypes (indicated by major cell type, followed by subtype cluster id) that are associated with statistically significant hazard ratios. **e.** UMAP plot of immune cells illustrating that immune GEM7 is dominantly expressed in the cluster CL6 of $CD8^+$ T cells (also see Figure 2). **f.** UMAP plot of immune cells illustrating enriched expression of immune GEM18 in the cluster CL6 of monocytes.

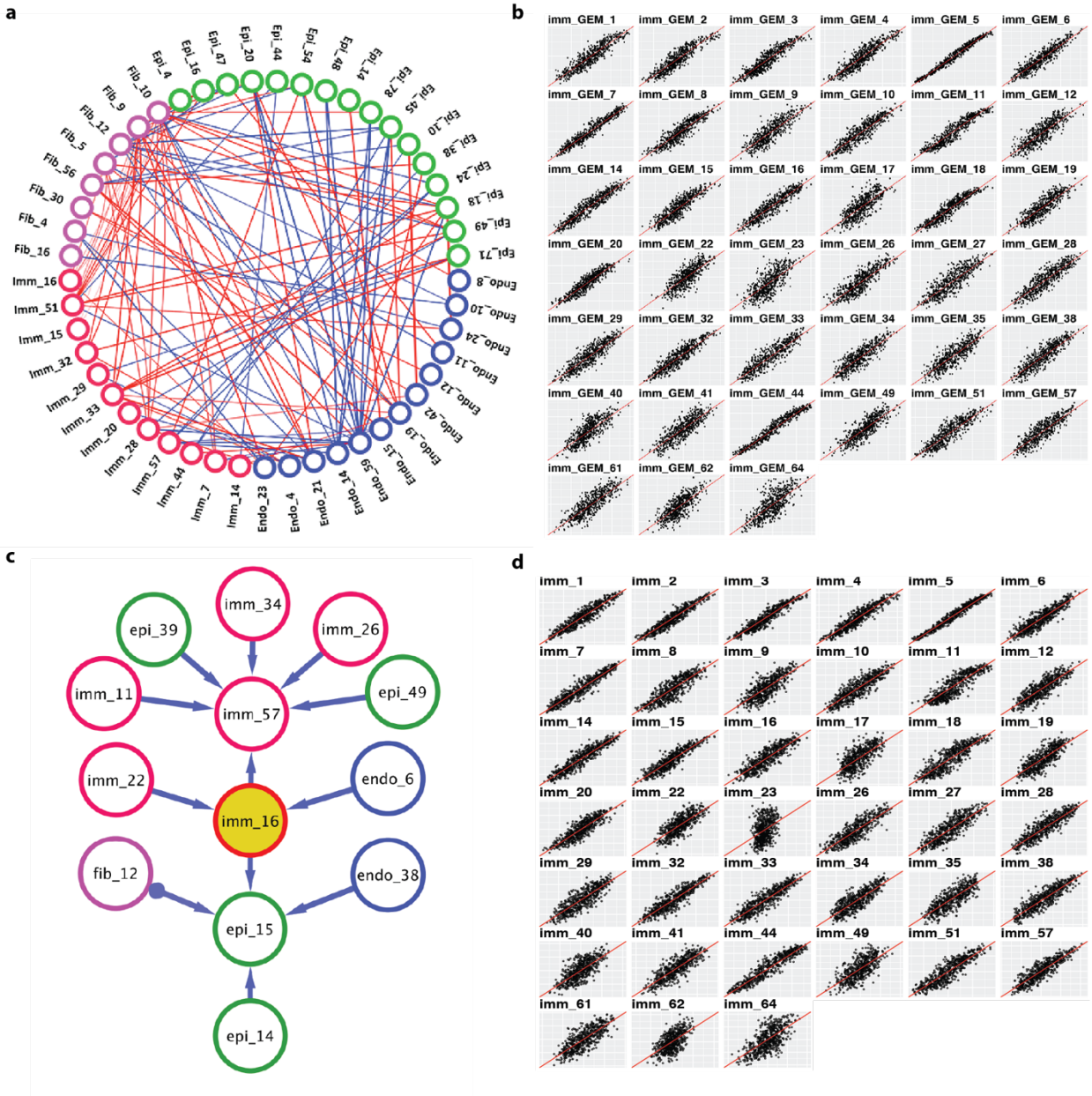

Figure 4. Intercellular communication among cells in HNSCC TMEs. a. Pair-wise correlation between GEMs from different type of cells. GEMs from major cell categories are color-coded and the top 10 strongest correlated GEMs for each cell type are shown as edges connecting the GEMs. Positive and negative correlated pairs are indicated as red and blue edges, respectively. **b. Scatter plot of predicted immune GEM enrichment values using non-immune GEMs as**

independent variables in a regression model vs observed GEMs values. The average R^2 is 0.71 . **c. A subgraph of the PAG learned by GFCI algorithm.** GEMs from different cell types are color-coded. A direct edge (A→B) represents a direct causal relationship, i.e., A causes B, and blunt tailed edge (A ⟔ B) indicates either A causes B, or there is a latent variable causes both A and B, or both situations. **d. Scatter plot of predicted immune GEM enrichment values using the Markov boundary of a GEM as independent variables in a regression model vs observed GEM values.** The mean of R^2 is 0.67, and the standard deviation is 0.16.

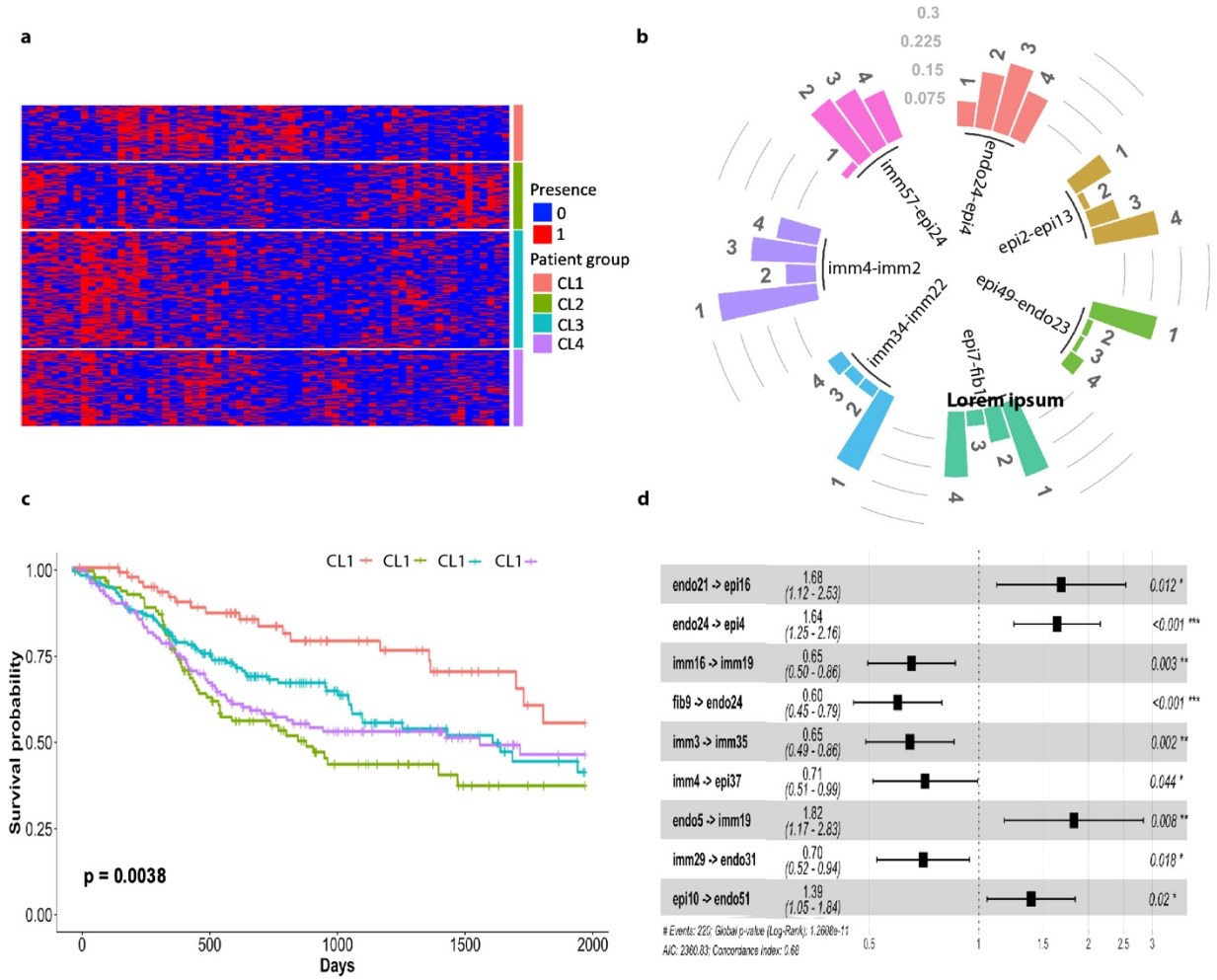

Figure 5. Distinct intercellular communication networks revealed by instance-specific CBN discovery. a. Clustering of tumor-specific intercellular communication networks. Tumor-specific causal networks returned by the iGFCI algorithm were represented in the space spanned by the union of edges, where the presence/absence of an edge is represented by a binary variable ("1" / "0" respectively). Tumors were divided into 4 clusters, and the presence/absence of edges that exhibited high variance is shown as a heatmap. **b. Examples of causal edges with high variance among tumor clusters.** Eight edges are shown, and the directions of the edges are indexed by the cause-effect GEMs. The height of a bar reflects the ratio (or percentage) of tumors in a cluster (indicated by a number from 1 to 4) that have the

edge in their tumor-specific network. **c. Kaplan-Meier survival curves of patients assigned to different clusters. d. A list of the causal edges that are associated with a statistically significant proportional hazard ratios according to a Cox proportional hazard analysis**.