Gene Network Inference from Single-Cell Omics Data and Domain Knowledge for Constructing COVID-19-Specific ICAM1-Associated Pathways

Mitsuhiro Odaka (odaka@nii.ac.jp)
The Graduate University for Advanced Studies, SOKENDAI
Morgan Magnin
École Centrale de Nantes
Katsumi Inoue
National Institute of Informatics

Research Article

Keywords:

Posted Date: February 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1300133/v1

Gene Network Inference from Single-Cell Omics Data and Domain Knowledge for Constructing COVID-19-Specific *ICAM1*-Associated Pathways

Mitsuhiro Odaka[1,2,3,4,*], **Morgan Magnin**[2,3,+], **and Katsumi Inoue**[1,2,3,+]

[1]Department of Informatics, The Graduate University for Advanced Studies, SOKENDAI, Tokyo, Japan
[2]Principles of Informatics Research Division, National Institute of Informatics, Tokyo, Japan
[3]Laboratoire des Sciences du Numérique de Nantes, École Centrale de Nantes, Nantes, France
[4]Japan Society for the Promotion of Science, Tokyo, Japan
[*]Corresponding author: Mitsuhiro Odaka; E-mail: odaka[at]nii[dot]ac[dot]jp; TEL: +818057061107
[+]These authors contributed equally to this work.

ABSTRACT

ICAM-1 is critical for interactions between cells. Previous studies have suggested that ICAM-1 triggers cell-to-cell transmission of HIV-1 or HTLV-1. SARS-CoV-2 shares several features with these viruses in interactions between cells, and SARS-CoV-2 cell-to-cell transmission is associated with COVID-19 severity. However, ICAM-1 and its associated pathways are not identified as essential factors in interactions between cells in COVID-19. For example, the current COVID-19 Disease Map has no entry for those pathways. Therefore, discovering unknown ICAM-1 pathways will be indispensable for clarifying the mechanism of COVID-19. This study builds *ICAM1*-associated pathways by gene network inference from single-cell omics data and multiple knowledge bases. First, data analyses extracted coexpressed genes with significant differences in expression levels with spurious correlations removed. Second, knowledge bases validate models. Finally, mapping the models onto existing pathways identifies new *ICAM1*-associated pathways. These pathways indicate that (1) upstream pathways include proteins in noncanonical NF-κB pathway and that (2) downstream pathways contain integrins and cytoskeleton or motor proteins for cell transformation. In this way, data-driven and knowledge-based approaches are integrated into gene network inference for *ICAM1*-associated pathway construction. The results can contribute to repairing and completing the COVID-19 Disease Map, thereby improving our understanding of the mechanisms of COVID-19.

Introduction

Deciphering the underlying mechanisms of coronavirus disease 2019 (COVID-19) is an urgent issue. Uncovering the mechanism requires a full understanding of the COVID-19-specific interactome, the complex network of interactions among different components. In previous COVID-19 studies, attempts to provide insights into interactions *within cells* have been made. However, interactions *between cells*, such as the transcriptional regulation arising from cell adhesion molecules (CAMs), have not been examined closely.

One noteworthy CAM is intercellular adhesion molecule 1 (ICAM-1; also known as CD54), which is encoded by *ICAM1*. ICAM-1 is a transmembrane glycoprotein that is expressed on leukocytes, vascular endothelial cells, and respiratory epithelial cells. Its differential expression is critical for proinflammatory immune responses and viral infection. Additionally, ICAM-1 enables interactions between cells by controlling leukocyte migration, homing, and adhesion from outside to inside the cell (outside-in) and regulation from inside to outside the cell (inside-out) [1]. These functionalities make ICAM-1 an attractive drug target and a clinically essential molecule [2].

Nevertheless, the system of signaling pathways surrounding ICAM-1, that is, the signal cascades that occur upon the functional activation of the signaling molecules, is not explicitly recognized as indispensable. Even in the case of COVID-19, there is little insight into the signaling pathway associated with ICAM-1. Previous research showed only that the gene expression level of *ICAM1* is elevated in cells infected by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus that causes COVID-19 [3].

Another important consequence of revealing *ICAM1*-associated pathways is completing and contributing to the COVID-19 Disease Map, which is a well-known graphical knowledge repository with multiscale physiological and cross-talk pathways [4]. The initial COVID-19 Disease Map was launched by highlighting the activated nodes of KEGG subpathways and inferring the cell functionalities annotated to those subpathways [5]. The enriched pathways are no more than a subset of the queried pathway models, lacking genes that have no pathway annotations [6], which makes it challenging to find unknown pathways. Moreover, the initial version of the COVID-19 Disease Map was built in an on-the-fly manner [7], so it has inherently been

incomplete and a work in progress. In fact, in the COVID-19 Disease Map, the pathways related to ICAM-1 and even ICAM-1 are absent. Therefore, it is meaningful to uncover the unknown pathways upstream and downstream of ICAM-1 (for simplicity, *ICAM1-associated pathways* for short) in COVID-19 to better understand the interactions between cells in the context of COVID-19.

Given the above, this study infers the existence of *ICAM1*-associated pathways in two ways. The first is through harnessing data and knowledge. We inferred gene networks based on data to identify unknown relations while rectifying the inferred network with domain knowledge. Integration of data-driven and knowledge-based approaches allows us to avoid biologically meaningless interpretation based on the data characteristics only [8]. The second is through analogy between SARS-CoV-2 and other retroviruses. Specifically, SARS-CoV-2 has a structurally homologous spike glycoprotein on the surface of the viral envelope that binds to a surface protein on the recipient cell during cell adhesion [9], which reflects that SARS-CoV-2 functionalization is similar to that of these other viruses [10]. Here, support is provided by the observation that human immunodeficiency virus type 1 (HIV-1) and human T cell leukemia virus type 1 (HTLV-1) have pathways for interactions between cells, called *cell-to-cell transmission*, triggered by ICAM-1 [11, 12]. After cell adhesion, HTLV-1 infection causes the formation of microtubule-organizing centers (MTOCs) and virological synapses (VSs) through the Ras-Raf-MEK-ERK pathway [13], after which cell-to-cell transmission occurs. Additionally, a previous model-driven study of COVID-19 showed an association between cell-to-cell transmission parameters and COVID-19 severity under the assumption of the presence of cell-to-cell transmission in SARS-CoV-2 infection [14].

Overall, this study indicates that *ICAM1*-associated pathways that drive cell-to-cell transmission should exist according to both data-driven gene network inference and knowledge-based rectification. Finding the unknown pathways will lead to a deeper understanding of the mechanisms of COVID-19.

Results

Figure 1 illustrates the workflow of this study. This workflow consists of the following five steps.

1. Single-cell omics data analysis
2. Undirected graphical model construction
3. Model corroboration and validation
4. Gene-to-protein conversion
5. Pathway mapping and unification

This workflow extracts beneficial information from a sparse matrix of single-cell omics data from COVID-19-positive and COVID-19-negative cases. **Steps 1** and **2** are dedicated to data-driven gene network inference. In **Step 1**, we obtained COVID-19-specific differentially expressed genes (DEGs) and a network of differentially coexpressed genes (DCGs) via single-cell omics data analysis. Here, DEGs are the genes whose expression levels are significantly different in COVID-19-positive patients and negative controls, and DCGs are coexpressed DEGs. In **Step 2**, spurious edges are removed from the correlation networks, thereby building *de novo* undirected graphical models. **Steps 3**, **4**, and **5** are rectification based on knowledge bases. In corroboration (**Step 3**), undirected graphical models are edited as dependency graphs with validated relations. In **Step 4**, genes are converted into the encoded proteins automatically through a functional annotation tool. Pathway mapping and unification (**Step 5**) refined the results as the final outputs, the *ICAM1*-associated pathways. Information beyond a subset of the existing Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways is retrieved through the workflow. Specifically, integrating single-cell omics data and multiple knowledge bases allows the inference of gene networks containing the components absent from the KEGG pathways.

Single-cell omics data analysis adopts a combination of the following standard methods. Quality control (**QC**) preprocesses the single-cell omics data (scRNA-seq data). Dimensionality reduction (**DR**) embeds data points in a sparse matrix onto a two-dimensional latent space. Clustering (**CL**) classifies data points in the latent space by similarity measurement and filters out the genes that are not associated with the gene of interest. The Wilcoxon rank-sum test (**WX**) detects significant differences in expression between COVID-19-positive and COVID-19-negative patients. Genewise DEG analysis and cellwise DCG analysis handle the data sparsity through these methods, forming correlation networks.

We applied the above workflow to data from four antigen-presenting cell types in which viral particles are detectable, namely, infected alveolar type 1 & 2 cells (infected AT1 & AT2), migratory dendritic cells (migratory DCs), tissue-resident alveolar macrophages type 2 (TRAM2), and monocyte-derived alveolar macrophages type 2 (MoAM2), as well as the aggregate of these four cell types (Fig. 2).

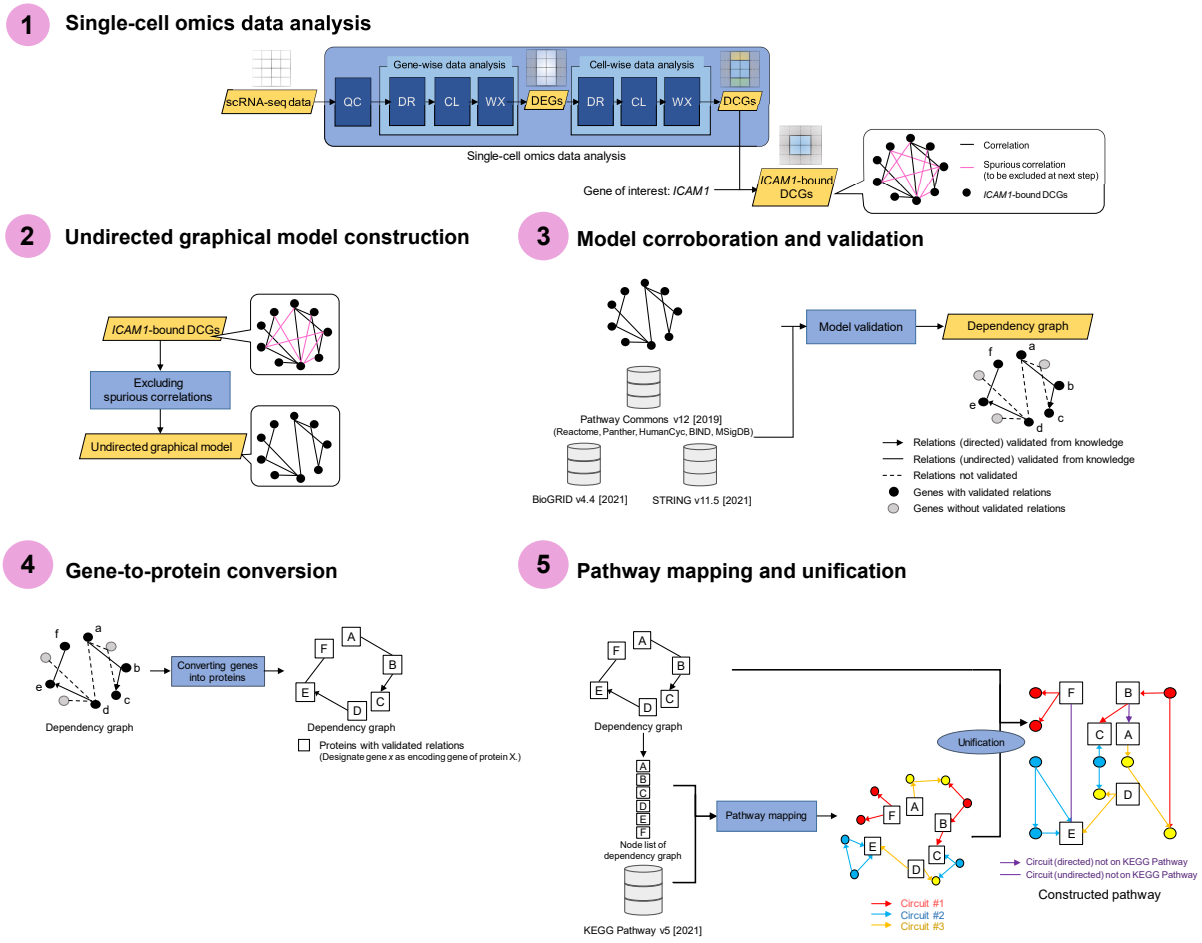
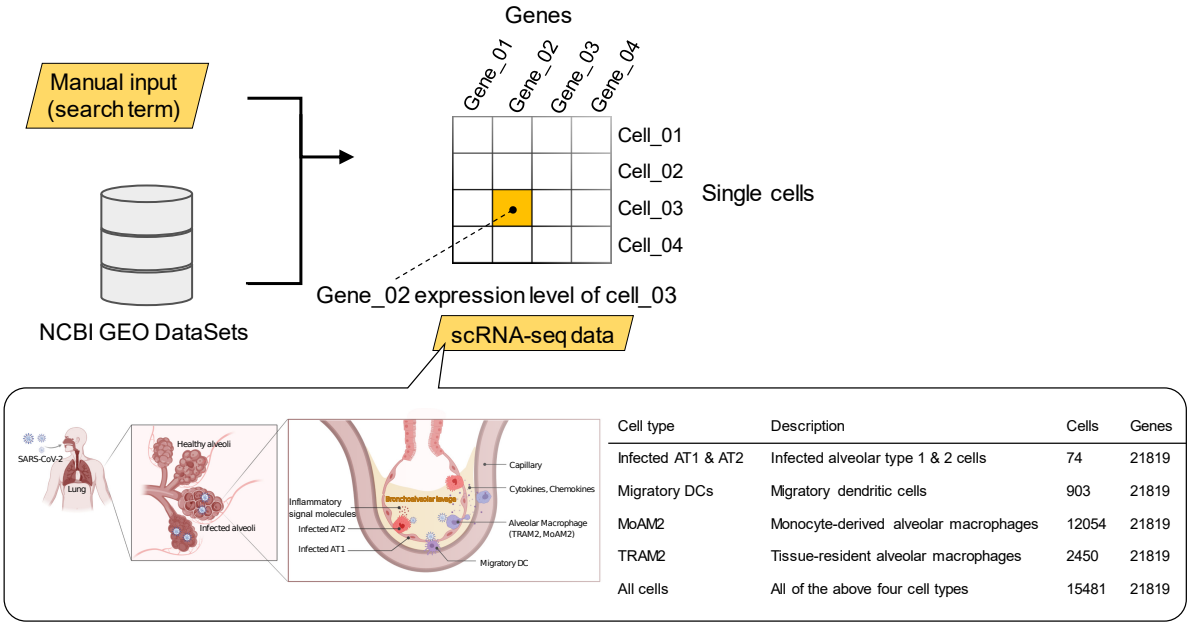


Fig. 1. The research workflow. **Step 1**: Single-cell omics data analysis. **Step 2**: Undirected graphical model construction. **Step 3**: Model corroboration and validation. **Step 4**: Gene-to-Protein conversion. **Step 5**: Pathway mapping and unification. The *circuits* are subpathways transmitting a signal from input receptor nodes to output effector nodes, where the nodes mostly represent proteins such as metabolic enzymes. See also [15].



Cell type	Description	Cells	Genes
Infected AT1 & AT2	Infected alveolar type 1 & 2 cells	74	21819
Migratory DCs	Migratory dendritic cells	903	21819
MoAM2	Monocyte-derived alveolar macrophages	12054	21819
TRAM2	Tissue-resident alveolar macrophages	2450	21819
All cells	All of the above four cell types	15481	21819

Fig. 2. The five cell types for which data were collected. Pulmonary tissue illustrations: Created with BioRender.com. See also [16].

Single-cell omics data analysis extracts COVID-19-specific DEGs and *ICAM1*-associated DCGs

This section explains the pipeline of single-cell omics data analysis, while the Supplementary Information provides details of the results (see also Supplementary Figs.S1–S6). In this study, COVID-19-specific genes associated with *ICAM1* were examined via the application of standard methods refined as three subroutines (dimensionality reduction, clustering, and Wilcoxon rank-sum test) to the omics data for each gene pair and each cell pair [17]. After quality control, genewise analysis extracted the gene clusters with differential expression patterns specific to COVID-19. Each cluster includes genes with significant differences in expression from the negative control. Additionally, the genes within the same cluster share a common differential expression pattern. The entire list of DEGs can be found in Supplementary Table S1 online.

Likewise, cellwise analysis filtered the DCGs via three subroutines to classify all the cells into cell clusters based on the correlation coefficients as similarity measurements, which means that the genes within the same cell cluster are more strongly correlated with each other than with the genes in other clusters. The entire list of DCGs can be found in Supplementary Table S2 online. Constraining the DCGs with the gene of interest, *ICAM1*, provided a subset of DCGs that correlate with *ICAM1*.

Removing the spurious edges from DCG correlation networks results in undirected graphical models

Next, we obtained gene network models inferred from the data. There are several methods available for gene network inference from single-cell omics data, including regression, mutual information, correlation, Bayesian networks, and a combination of different approaches [18, 19]. Gene network inference by Bayesian networks has been well investigated; however, this approach is said to be most suitable for small datasets because Bayesian networks infer causality but incur computational costs due to scalability problems such as the variables' rapid combinatorial expansion [20]. Conversely, simple correlation networks find both direct and indirect relations at a lower cost, but they do not distinguish between causality and correlation [21]. For less time cost and direct relations, we adopt gene network inference with partial correlation graphs and *undirected graphical models*, which have been successful in displaying *de novo*-produced direct linear associations [22–24].

In undirected graphical models, nodes are the extracted DCGs, and edges are anticipated correlations with spurious correlations removed. Considering that correlated gene pairs are coexpressed with similar functions, designating any gene pairs as nodes and the correlation coefficient of gene expression levels as edges formed a correlation network of *ICAM1*-associated DCGs. Excluding the edges with zero second-order partial correlation coefficients for at least one node pair yielded undirected graphical models without spurious correlations (Table 1).

Calculation of the second-order partial correlation coefficients between all gene pairs and removal of the edges of the gene pairs with zero partial correlation coefficients for any combination as spurious correlation formed undirected graphical models with coexpressed genes as nodes [25]. The equations for the zero-order, first-order, and second-order partial correlations are

Table 1. Excluding spurious correlations.

	Nodes	Edges (full)	Edges (excluded)	Edges (output)
Infected AT1 & AT2	116	6670	6296	374
Migratory DCs	248	30628	28695	1933
TRAM2	150	11175	8690	2485
MoAM2	152	11476	7819	3657
All cells	179	15931	12529	3402

shown in equations (1), (2), and (3).

Zero-order correlation :

$$r_{xy} = \frac{\mathrm{cov}(xy)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}} \tag{1}$$

First-order partial correlation :

$$r_{xy,z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{\left(1 - r_{xz}^2\right)\left(1 - r_{yz}^2\right)}} \tag{2}$$

Second-order partial correlation :

$$r_{xy,zq} = \frac{r_{xy,z} - r_{xq,z} r_{yq,z}}{\sqrt{\left(1 - r_{xq,z}^2\right)\left(1 - r_{yq,z}^2\right)}} \tag{3}$$

The random variables denoted by x, y, z, and q represent the gene names. r_{xy} is Pearson's correlation coefficient between the gene expression-level vector running over all cells of any gene x and that of any gene y. The simple correlation network starts from connecting x and y if and only if $r_{xy} \neq 0$. Undirected graphical modeling removes the linear effect of all second-order partial correlation coefficients $r_{xy,zq}$ between two variables (x, y) conditional on all other variables. Figure 3 shows the undirected graphical model in which the nodes are the *ICAM1*-associated DCGs. The edge is weighted as $(0.5 + 0.5 \cdot r_{x,y})^{12}$ to follow the scaling law, which typically holds for a weighted gene correlation network [26].

Corroboration and validation with multiple knowledge bases edit undirected graphical models as dependency graphs

Information retrieved from any given dataset, which could have no biological meaning, requires model corroboration and validation with heuristics based on domain knowledge. In this study, we queried multiple knowledge bases, using each knowledge base's application programming interface (API) for model corroboration and validation. Fetching relations between gene pairs in the simple interaction format (SIF) enabled us to convert a subset of undirected edges to directed edges, thereby obtaining dependency graphs (Fig. 4).

The entire list of relations of dependency graphs with knowledge bases used for model validation can be found in Supplementary Table S3 online.

ICAM1-associated pathway construction by pathway mapping and unification of validated dependency graphs and KEGG pathways

There exists a gap between the inferred gene network and the KEGG pathways because the nodes of the models are genes, while the nodes of the KEGG pathways are primarily proteins. After the conversion of genes into proteins, mapping of the converted proteins onto KEGG pathways was employed to identify the activated pathways (Table 2).

Table 2. Mapping results of the dependency graphs of each cell type onto the KEGG pathways. Scores count the "matched" genes on the dependency graphs, whose encoding proteins are on any of the KEGG pathways and their proportions to total gene counts. Matched KEGG pathways exemplify how many matched genes are included in a specific pathway. For example, if gene x's encoded protein X is on KEGG pathways A and B, one is added to the score, and both A and B are represented.

Cell types	Scores	Matched KEGG pathways
Infected AT1&2 cells	0 genes (no match)	
Migratory DCs	40 genes (63.5% match)	NF-κB signaling pathway (hsa04064) (12), HTLV-1 infection (hsa05166) (9)
TRAM2	23 genes (65.7% match)	Influenza A (hsa05164) (13), HTLV-1 infection (hsa05166) (5)
MoAM2	31 genes (54.4% match)	NF-κB signaling pathway (hsa04064) (3)
All cells	18 genes (64.3% match)	TNF signaling pathway (hsa04668) (3), NF-κB signaling pathway (hsa04064) (3)

Unification of the inferred dependency graphs and the KEGG pathways in KEGG Markup Language (KGML) format by Cytoscape 3.9.0 [29] resulted in the final COVID-19-specific *ICAM1*-associated pathways for each cell type (Fig. 5).

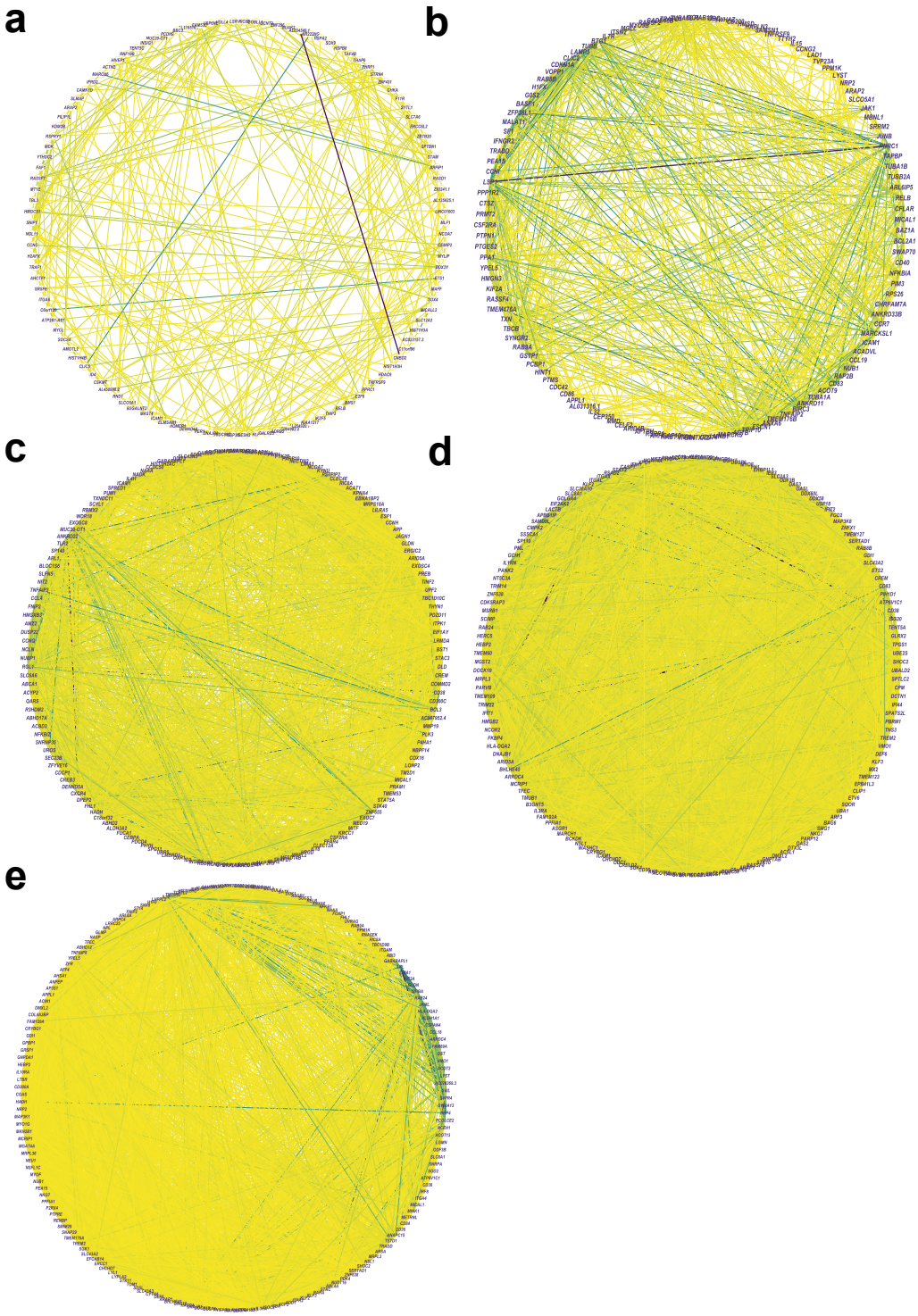


Fig. 3. Undirected graphical models. **a**: Infected alveolar type 1 & 2 cells. **b**: Migratory dendritic cells. **c**: Tissue-resident alveolar macrophages. **d**: Monocyte-derived alveolar macrophages. **e**: All cells. See also [27].

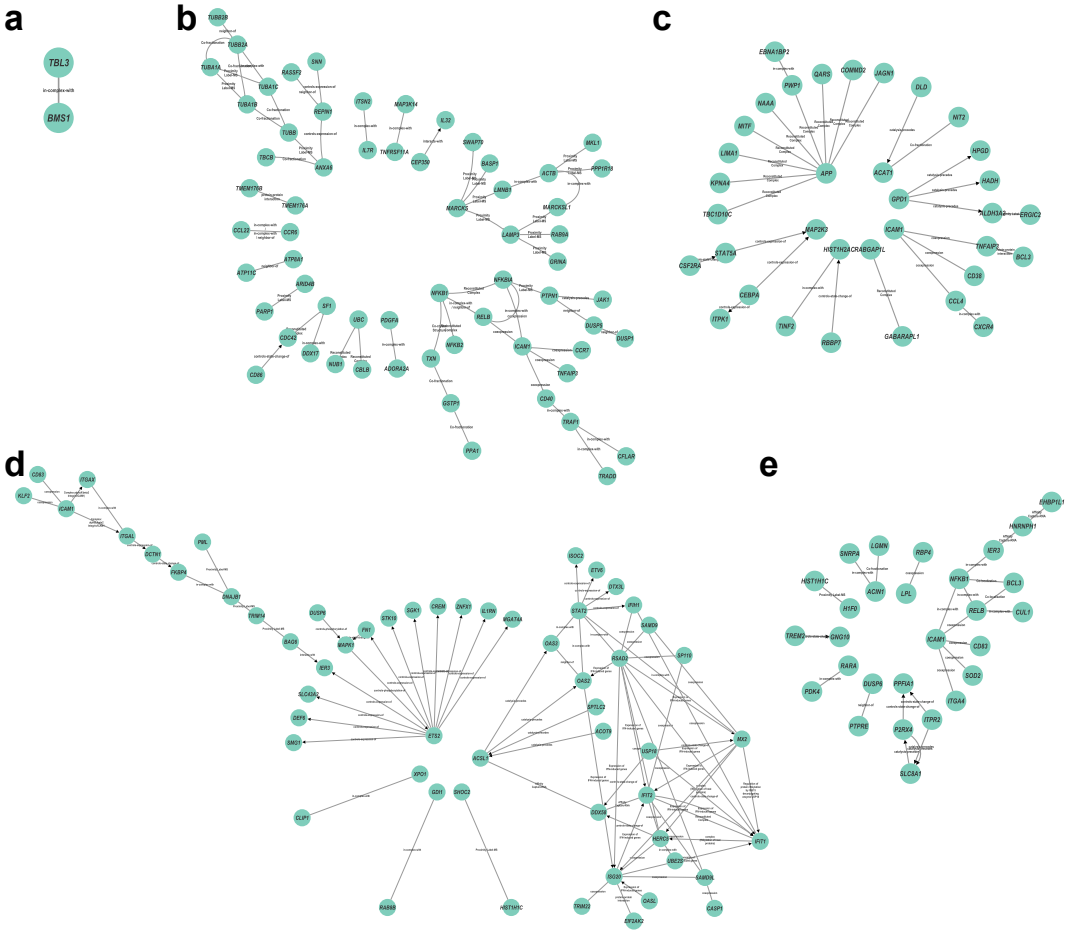


Fig. 4. Directed graphical models. **a**: Infected alveolar type 1 & 2 cells. **b**: Migratory dendritic cells. **c**: Tissue-resident alveolar macrophages. **d**: Monocyte-derived alveolar macrophages. **e**: All cells. See also [28].

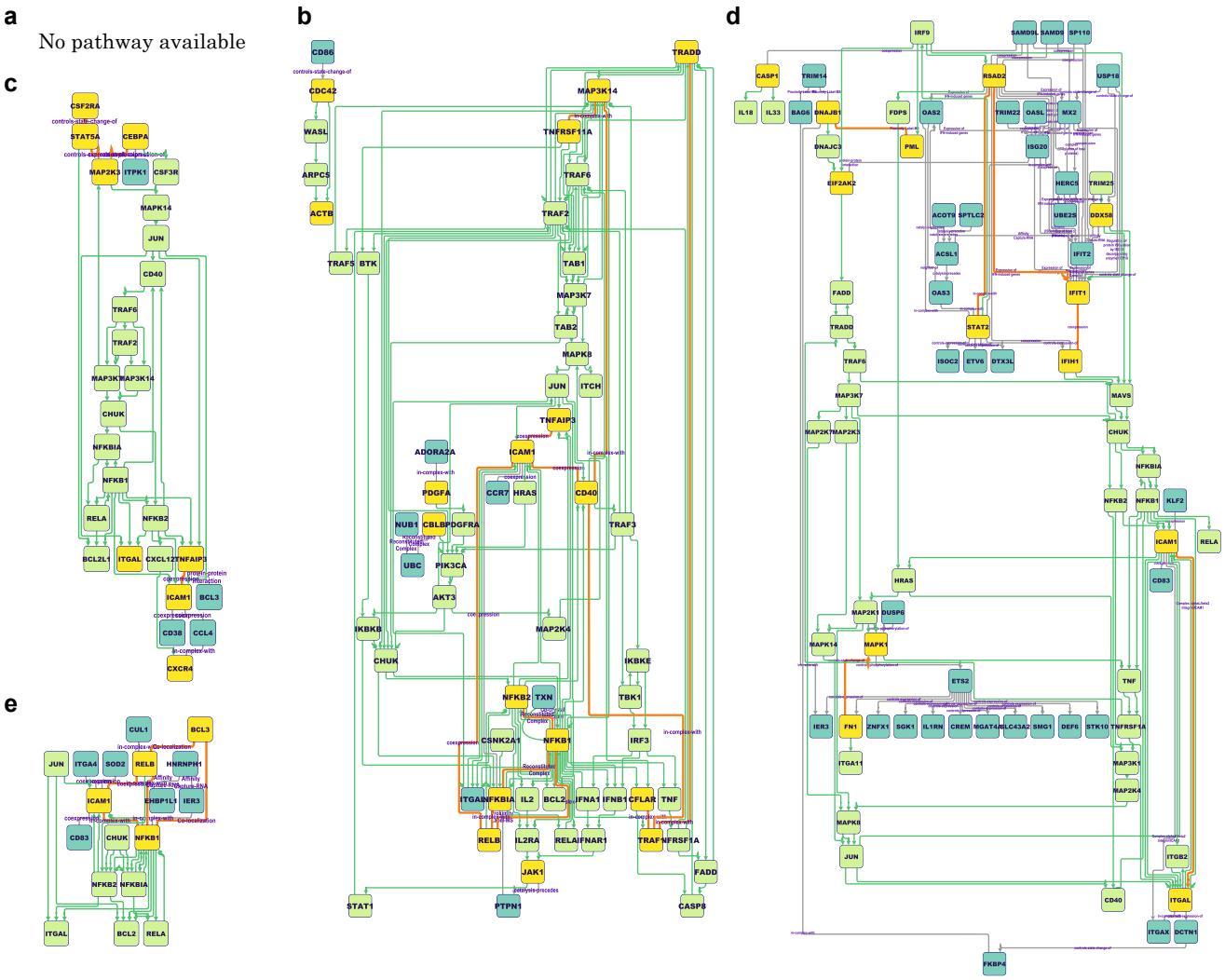


Fig. 5. *ICAM1*-associated pathways. **a**: Infected alveolar type 1 & 2 cells (no pathway available); **b**: Migratory dendritic cells; **c**: Tissue-resident alveolar macrophages; **d**: Monocyte-derived alveolar macrophages; **e**: All cells. The rectangular nodes colored blue, yellow, and lime green reflect the proteins on dependency graphs only, the proteins common to both the dependency graphs and KEGG pathways, and the proteins on KEGG pathways only, respectively. Gray lines are the directed or undirected edges on the dependency graphs only; orange lines are the directed or undirected edges between yellow nodes on the dependency graphs; and green lines are the directed edges on the KEGG pathways only. Orange edges do not have a direction if the KEGG pathways connect its yellow node pair indirectly. See also [30].

The pathway for migratory DCs comprises *NFKB1*, *NFKB2*, and *RELB*. *NFKB2* and *RELB* lie in the noncanonical NF-κB pathway, which is a known pathway upstream of *ICAM1* [31]. *MAP2K3*, *MAPK14*, *JUN*, *FOS*, *ITGA2*, *ITGB1*, *RSAD2*, *OAS*, and *STAT2* have already been mapped onto the COVID-19 Disease Map, while *RELB*, *ITGAL*, *CDC42*, *ACTB*, *CD40*, *DCTN1*, *BCL3*, and *CD83* in the pathway in this study were still absent. All cell types also contain some integrins that interact with ICAM-1 to stabilize cell adhesion, including *ITGAL* in macrophages (gene encoding CD11a also known as LFA1A), *ITGAX* (gene encoding CD11c), and *ITGB2* (gene encoding CD18). Some of them are downstream, such as actin in migratory DCs and dynein in MoAM2 (*DCTN1* recruiting and tethering dynein to microtubules).

Discussion

Our results imply that COVID-19 involves (1) ICAM-1-upstream pathways with proteins on the noncanonical NF-κB pathway and (2) downstream pathways with integrins and cytoskeletal elements associated with actin and the motor protein dynein for cell transformation. The noncanonical NF-κB pathway is reasonable because it is relevant to the proinflammatory response in viral infections such as COVID-19. The involvement of downstream pathways leading to the cytoskeleton (the internal filaments of eukaryotic cells), including actin filaments and microtubules, in COVID-19 is also plausible. After cell adhesion is regulated by the interaction between ICAM-1 and integrin, the motor protein myosin would move on actin filaments, inducing cell transformation and movement, while the moter protein dynein would move on microtubules transporting molecules in the cytoplasm to the MTOCs.

Pathways such as the Ras-Raf-MEK-ERK pathway in HTLV-1 cells were inactive in this study. Meanwhile, *RAC1* and *CDC42* were found to be conserved. Ras-related C3 botulinus toxin substrate 1 (Rac1, encoded by *RAC1*) and cell division control protein 42 homolog (Cdc42, encoded by *CDC42*) are essential for VS formation in HTLV-1 cells [32]. Therefore, although it is unclear whether SARS-CoV-2 has a VS formation mechanism analogous to those of HIV-1 or HTLV-1, we cannot rule out the possibility that MTOC formation and VS formation never occur. Further phenomenological verification of MTOC and VS formation through infection experiments or molecular dynamics tracking using high-end live-cell imaging techniques [33] would be desirable.

The need to identify unknown pathways has accelerated work related to gene network inference. For example, Hasankhani *et al.* obtained signaling pathways associated with the main hallmarks of COVID-19 by differential coexpression network analysis [34]. Tanaka *et al.* revealed host cellular gene networks by Bayesian network [35].

The limitations of this study includes the gap between gene and protein layers in the systems. The gene network constructed from the transcriptome is a transcriptional regulatory network, where the transcripts directly determine the gene expression levels. Although the links between the gene pairs represent direct interactions between a transcription factor and controlled gene, there are interactions between the upper layer components other than transcripts or genes, including proteins and metabolites [36]. Filling this gap between the different layers requires the integration of the observed data for each layer. For example, we can not guarantee that our method would produce the same results using datasets measuring variations in protein expression levels directly, which requires further analysis, such as reversed-phase protein array (RPPA).

Given the above limitation, future work will include the following two tasks. One is calibration experiments on multilayered models extended from viral dynamics models [37] with multiomics data incorporating distinct layer datasets [38]. The current results rely on the transcriptome data only, making it challenging to check the upper layer of the biological hierarchy. Consequently, we can consider a multilayered model, such as a combination of a metabolic flux model and Hoffmann's NF-κB model [39]. The other is to produce a simulation for the explicit use of direct transcription factor perturbations (knockout or overexpression) [40]; for example, one such study significantly improved prediction accuracy for downstream targets [41].

Overall, the contributions of this study were as follows: first, we discovered novel pathways currently absent from the COVID-19 Disease Map; second, we demonstrated the use of a new methodology. While previous analysis or curation work found the canonical NF-κB pathway [42], the noncanonical pathways were not known to be involved in the COVID-19 Disease Map. The discovered pathways suggested the existence of unknown pathways in the map, upstream noncanonical NF-κB pathway, and a downstream pathway that may lead to MTOC formation subject to further experiments. Furthermore, if *ICAM1* can be replaced with other targets such as *ACE2*, then the same workflow may lead to other findings. Although further experiments remain, the data-driven and knowledge-based gene network inference and the *ICAM1*-associated pathways constructed in this study will expedite the repair and completion of the COVID-19 Disease Map for a deeper understanding of SARS-CoV-2 pathogenesis.

Methods

Machine configuration

The machine configuration was as follows: Python 3.8 Google Compute Engine back end TPUv2, CPU Xeon @ 2.30 GHz x2, and 35.25 GB of RAM.

Data

Omics data are comprehensive bioinformatics data that can be generated through high-throughput technologies with the advantage of illustrating multilevel systems ranging from genes and proteins to metabolites. For example, the transcriptome represents the entire body of transcripts. While there are two primary sequencing techniques for obtaining the transcriptome, bulk sequencing and single-cell sequencing, the data from single-cell sequencing offer the advantage of addressing the heterogeneity of cell populations and thus providing different interpretations than bulk sequencing.

In this study, transcriptome data were adopted, given their three benefits. First, it is more easily generated than other omics data. Second, given that biological systems exist as an interactome with multiple layers, the gene network, consisting of genes, transcripts, and gene regulation by transcripts, is the foundational layer projected to the upper layers, i.e., signaling pathways, metabolic pathways, or protein-protein interactions [43]. Third, even though the correspondence between genes and transcripts is not one-to-one due to exon gaps or splicing invariance, 94% of functionality can be determined from transcriptome data [44].

The single-cell omics data (GSE155249_supplement.h5ad) used in this study originate from a primary source [45]. Manually inputting the search term *((COVID-19 OR SARS-CoV-2) AND gse[entry type]) AND "Homo sapiens" AND h5ad* to the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) DataSet [46] provided the data. The data included the gene expression profiles of bronchoalveolar lavage fluid samples isolated from 10 patients with severe COVID-19 and two negative controls via high-throughput single-cell RNA sequencing. The gene expression level is a signal magnitude of its transcripts, such as messenger ribonucleotide acid (mRNA), microRNA (miRNA), noncoding RNA, and small interfering RNA (siRNA). Among the transcripts, some mRNAs undergo translation into proteins.

Single-cell omics data analysis

Quality control

For further analysis, the data were processed through quality control, including filtering, scaling, and normalization by Scanpy [47]. The data originally contained 77,650 cells × 24,714 genes. Given cell quality, one can regard the cells with overexpressed mitochondrial RNA per data count tagged by a unique molecular identifier (UMI) [48] as dead or broken cells. Likewise, cells with many genes per data count tagged by UMI could be identified as doublets. The filtering process excluded 8,916 cells with over 5,000 expressed genes, 700 genes that were detected in fewer than three cells, mitochondrially encoded genes, and cells with a low percentage (< 10%) of mitochondrial genes. After filtering, the dataset contained 68,734 cells × 24,001 genes. The count data were scaled with regression on total UMI counts and normalization per feature based on standard deviation.

Single-cell omics data were isolated from bronchoalveolar lavage fluid samples collected from 10 patients with severe COVID-19 and two negative controls. The data were processed and annotated in a sparse matrix of mRNA expression levels in a single-cell population. The primary article classified the cell types of clusters based on cell-type-specific marker genes. The input data underwent dimensionality reduction, clustering, and statistical tests for each cell type. This analysis resulted in a hash table of DEGs, including the cluster number corresponding to the cell type, gene name string, and log fold change.

Then, the DEGs were extracted from 21,819 genes among the SARS-CoV-2-infected 15,481 cells (cells with SARS-CoV-2 transcripts detected); infected alveolar type 1 and 2 cells, *CCR7*+ migratory dendritic cells, *CCL2*+ monocyte-derived alveolar macrophages, and *CCL4*+, *CCL20*+, *CXCL10*+, *CXCL11*+, *IL1B*+, *TNFSF10*+, *DEFB1*+ tissue-resident alveolar macrophages. In addition to the preprocessing mentioned above to the whole sparse matrix, the first step was quality control to obtain preprocessed data suitable for further analysis. Removing the cells with a high proportion of mitochondrial RNA resulted in 15,220 cells. The doublet discrimination provided 14,723 cells. Cells with less than one gene count were filtered out, leaving 9,050 cells. Subsequently, to ensure gene quality, the genes detected in fewer than three cells were filtered out, leaving 17,644 genes. Subsequently, normalization of the gene expression data adjusted for the RNA composition bias and allowed comparison of the values among the cells. Finally, log-transformation prepared the data for calculating the log-fold changes reflecting the gene expression difference. The quality control ultimately yielded log-transformed normalized gene expression data for 17,644 genes in 9,050 single cells.

Dimensionality reduction and clustering

The step after quality control was the imputation of zeros, representing either technically missing data or biologically absent genes, within a sparse matrix of single-cell omics data [49]. One zero imputation method is matrix factorization, which decomposes a sparse matrix into components in cell-by-factor weighted low-rank matrices and gene-by-factor weighted matrices. We adopted two matrix factorization methods: principal component analysis (PCA) and uniform manifold approximation and projection (UMAP). These methods detected possible batch effects and embedded the matrix in the latent space, reducing its dimensionality. Computing 50 PCA coordinates on the sparse matrix for mean centering [50], one obtains eigenvalues and eigenvectors with the singular value decomposition solver ARPACK (ARnoldi PACKage) [51]. PCA reduced the dimension to 100 by a Gaussian kernel, and UMAP embedded the data points onto the two-dimensional latent space. Given the 50 decomposed coordinates, the connectivities (weighted adjacency matrix) of the k-nearest neighborhood graph are computed and thresholded at the 15 closest neighbors defined for data points of the manifold in Euclidean space. Initializing with spectral

embedding, UMAP [52] embedded the neighborhood graph onto a two-dimensional latent space. By implementing clustering with the Louvain algorithm, greedy optimization of local modularity to detect the groups [53] classified the cell population into 18 subgroups with similar gene expression profiles.

Wilcoxon rank-sum test

Nonparametric pairwise statistical hypothesis tests between pairs of subgroups identified a significant difference. The Wilcoxon rank-sum test ranks the genes characteristic of each subgroup, comparing the signal values between each subgroup and the union of the other subgroups with the Benjamini–Hochberg method for adjusting the false discovery rate and correcting the *p value* [54]. Wilcoxon rank-sum tests between pairs of groups of cells identified significant differences. Excluding the duplicated genes extracted 1,434 DEGs in 9,050 single cells.

The DEGs were filtered to distinguish those whose gene expression levels were correlated. Here, given that functionally related genes are coexpressed in the same clusters, the identified gene clusters can be considered coexpressed genes [55]. PCA and UMAP were conducted to achieve the above. The notable difference was the similarity metrics used for embedding. Louvain clustering yielded 11 clusters based on the correlations between gene pairs in the embedded space. The Louvain method classifies the cells by kernel *k*-means clustering, where the kernel function is Gaussian. One of the 11 clusters included *ICAM1* and 178 coexpressed genes.

Model corroboration and validation

Until the previous data analysis, gene networks consisting of the *ICAM1* gene of interest were inferred without guaranteeing the validity of each edge. To corroborate and validate each relation of gene networks, we queried multiple knowledge bases, including Pathway Commons Web Service 12 [56], BioGRID REST Web Service [57], and STRING version 11.5 [58]. Pathway Commons API provided access to the significant pathway databases Reactome, Panther, HumanCyc, BIND, and MSigDB. BioGRID served as a complementary source of the latest knowledge since Pathway Commons was not up-to-date [59]. Additionally, HumanCyc, which has richer information on biochemical reactions and regulatory relationships than the KEGG pathways alone [60], enabled the obtained model to include more information than a subset of the KEGG pathways. STRING returned annotations of functional or physical interactions between the queried proteins.

Gene-to-protein conversion

To overlay the coexpressed genes in inferred gene networks onto the KEGG pathways, DAVID Bioinformatics Resources 6.8 [61, 62] converted genes into proteins.

Pathway mapping and unification

Simultaneously, DAVID was used to map the converted proteins onto KEGG pathways to identify the activated pathways (Supplementary Fig. S7).

Reporting summary

Further information on the research design is available in the Nature Research Reporting Summary linked to this article.

Data availability statement

Supplementary figures are provided in the Supplementary Information. All the datasets derived from this study, including Supplementary figures, have been deposited in the *figshare* repository under the URL: https://doi.org/10.6084/m9.figshare.c.5756363. Other data are available from the corresponding author (odaka[at]nii[dot]ac[dot]jp) on reasonable request. Single-cell RNA sequencing data supporting the findings of this study were obtained from the NCBI Gene Expression Omnibus (GEO) with the accession code GSE155249.

Code availability

All the source codes to reproduce the results in this study are available in the GitHub repository, Mitsuhiro-ODAKA/COVID19-ICAM1.

References

1. Luo, B., Carman, C. V. & Springer, T. A. Structural basis of integrin regulation and signaling. *Annu. Rev. Immunol.* **25**, 619–647, DOI: 10.1146/ANNUREV.IMMUNOL.25.022106.141618 (2007).

2. Wilson, R. W. *et al.* Gene targeting yields a CD18-mutant mouse for study of inflammation. *J. Immunol.* **151**, 1571–8 (1993).

3. Tong, M. *et al.* Elevated Expression of Serum Endothelial Cell Adhesion Molecules in COVID-19 Patients. *The J. Infect. Dis.* **222**, 894–898, DOI: 10.1093/infdis/jiaa349 (2020).

4. Ostaszewski, M. *et al.* Community-driven roadmap for integrated disease maps. *Brief. Bioinform.* **20**, 659–670, DOI: 10.1093/bib/bby024 (2019).

5. Mitsos, A. *et al.* Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data. *PLoS Comput. Biol.* **5**, e1000591, DOI: 10.1371/journal.pcbi.1000591 (2009).

6. Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **14**, 482–517, DOI: 10.1038/s41596-018-0103-9 (2019).

7. Ostaszewski, M. *et al.* Author Correction: COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. *Sci. Data* **7**, DOI: 10.1038/s41597-020-00589-w (2020).

8. Zuo, Y., Cui, Y., Yu, G., Li, R. & Ressom, H. W. Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical lasso. *BMC Bioinform.* **18**, DOI: 10.1186/s12859-017-1515-1 (2017).

9. Weissenhorn, W. *et al.* Structural basis for membrane fusion by enveloped viruses. *Mol. Membr. Biol.* **16**, 3–9, DOI: 10.1080/096876899294706 (1999).

10. Walls, A. C. *et al.* Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **183**, 1735, DOI: 10.1016/j.cell.2020.11.032 (2020).

11. Bracq, L., Xie, M., Benichou, S. & Bouchet, J. Mechanisms for Cell-to-Cell Transmission of HIV-1. *Front. Immunol.* **9**, 260, DOI: 10.3389/fimmu.2018.00260 (2018).

12. Igakura, T. *et al.* Spread of HTLV-I between lymphocytes by virus-induced polarization of the cytoskeleton. *Science* **299**, 1713–1716, DOI: 10.1126/science.1080115 (2003).

13. Nejmeddine, M., Negi, V. S., Mukherjee, S. *et al.* HTLV-1-Tax and ICAM-1 act on T-cell signal pathways to polarize the microtubule-organizing center at the virological synapse. *Blood* **114**, 1016–1025, DOI: 10.1182/blood-2008-03-136770 (2009).

14. Odaka, M. & Inoue, K. Modeling viral dynamics in SARS-CoV-2 infection based on differential equations and numerical analysis. *Heliyon* **7**, e08207, DOI: 10.1016/j.heliyon.2021.e08207 (2021).

15. Odaka, M., Magnin, M. & Inoue, K. The research workflow. *figshare*. https://doi.org/10.6084/m9.figshare.18095717 (2022).

16. Odaka, M., Magnin, M. & Inoue, K. The five cell types for which data were collected. *figshare*. https://doi.org/10.6084/m9.figshare.18095717 (2022).

17. Li, X. *et al.* Network embedding-based representation learning for single cell rna-seq data. *Nucleic Acids Res.* **45**, e166–e166, DOI: 10.1093/nar/gkx750 (2017).

18. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804, DOI: 10.1038/nmeth.2016 (2012).

19. Mochida, K., Koda, S., Inoue, K. & Nishii, R. Statistical and Machine Learning Approaches to Predict Gene Regulatory Networks From Transcriptome Datasets. *Front. Plant Sci.* **9**, 1770, DOI: 10.3389/fpls.2018.01770 (2018).

20. Schäfer, J. & Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4**, DOI: 10.2202/1544-6115.1175 (2005).

21. Opgen-Rhein, R. & Strimmer, K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst. Biol.* **1**, 37, DOI: 10.1186/1752-0509-1-37 (2007).

22. Butte, A. J. & Kohane, I. S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.* 418–429, DOI: 10.1142/9789814447331_0040 (2000).

23. de la Fuente, A., Bing, N., Hoeschele, I. & Mendes, P. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinform.* **20**, 3565–3574, DOI: 10.1093/bioinformatics/bth445 (2004).

24. Schäfer, J. & Strimmer, K. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinform.* **21**, 754–764, DOI: 10.1093/bioinformatics/bti062 (2005).

25. Zuo, Y., Yu, G., Tadesse, M. G. & Ressom, H. W. Biological network inference using low order partial correlation. *Methods* **69**, 266–273, DOI: 10.1016/j.ymeth.2014.06.010 (2014).

26. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559–559, DOI: 10.1186/1471-2105-9-559 (2008).

27. Odaka, M., Magnin, M. & Inoue, K. Undirected graphical models. *figshare*. https://doi.org/10.6084/m9.figshare.17261825 (2022).

28. Odaka, M., Magnin, M. & Inoue, K. Directed graphical models. *figshare*. https://doi.org/10.6084/m9.figshare.17261780 (2022).

29. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504, DOI: 10.1101/gr.1239303 (2003).

30. Odaka, M., Magnin, M. & Inoue, K. *ICAM1*-associated pathways. *figshare*. https://doi.org/10.6084/m9.figshare.17261540 (2022).

31. Liu, Z. *et al.* A NIK–SIX signalling axis controls inflammation by targeted silencing of non-canonical NF-κB. *Nature* **568**, 249–253, DOI: 10.1038/s41586-019-1041-6 (2019).

32. Gross, C. & Thoma-Kress, A. K. Molecular Mechanisms of HTLV-1 Cell-to-Cell Transmission. *Viruses* **8**, 74, DOI: 10.3390/v8030074 (2016).

33. Real, F., Sennepin, A., Ganor, Y., Schmitt, A. & Bomsel, M. Live Imaging of HIV-1 Transfer across T Cell Virological Synapse to Epithelial Cells that Promotes Stromal Macrophage Infection. *Cell Rep.* **23**, 1794–1805, DOI: 10.1016/j.celrep.2018.04.028 (2018).

34. Hasankhani, A. *et al.* Differential Co-Expression Network Analysis Reveals Key Hub-High Traffic Genes as Potential Therapeutic Targets for COVID-19 Pandemic. *Front. Immunol.* **12**, 5371, DOI: 10.3389/fimmu.2021.789317 (2021).

35. Tanaka, Y. *et al.* Dynamic changes in gene-to-gene regulatory networks in response to SARS-CoV-2 infection. *Sci. Rep.* **11**, 11241, DOI: 10.1038/s41598-021-90556-1 (2021).

36. Zenobi, R. Single-cell metabolomics: analytical and biological perspectives. *Science* **342**, DOI: 10.1126/science.1243259 (2013).

37. Odaka, M. & Inoue, K. Computational Modeling and Simulation of Viral Load Kinetics in SARS-CoV-2 Replication. In *CSBio'20: Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics*, 75–82, DOI: 10.1145/3429210.3429214 (2020).

38. Stephenson, E., Reynolds, G., Botting, R. A. *et al.* Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Med.* **27**, 904–916, DOI: 10.1038/s41591-021-01329-2 (2021).

39. Hoffmann, A., Levchenko, A., Scott, M. L. & Baltimore, D. The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation. *Science* **298**, 1241–1245, DOI: 10.1126/science.1071914 (2002).

40. Loucera, C. *et al.* Drug repurposing for COVID-19 using machine learning and mechanistic models of signal transduction circuits related to SARS-CoV-2 infection. *Sig. Transduct. Target Ther.* **5**, DOI: 10.1038/s41392-020-00417-y (2020).

41. Noh, H., Shoemaker, J. E. & Gunawan, R. Network perturbation analysis of gene transcriptional profiles reveals protein targets and mechanism of action of drugs and influenza A viral infection. *Nucleic Acids Res.* **46**, e34, DOI: 10.1093/nar/gkx1314 (2018).

42. Fujisawa, K., Shimo, M., Taguchi, Y. H. *et al.* PCA-based unsupervised feature extraction for gene expression analysis of COVID-19 patients. *Sci. Rep.* **11**, DOI: 10.1038/s41598-021-95698-w (2021).

43. Brazhnik, P., de la Fuente, A. & Mendes, P. Gene networks: how to put the function in genomics. *Trends Biotechnol.* **20**, 467–472, DOI: 10.1016/s0167-7799(02)02053-x (2002).

44. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476, DOI: 10.1038/nature07509 (2008).

45. Grant, R. A. *et al.* Circuits between infected macrophages and T cells in SARS-CoV-2 pneumonia. *Nature* **590**, 635–641, DOI: 10.1038/s41586-020-03148-w (2021).

46. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Res.* **41**, D991–D995, DOI: 10.1093/nar/gks1193 (2013).

47. Wolf, F., Angerer, P. & Theis, F. SCANPY: large-scale single-cell gene expression data analysis. *Genome. Biol.* **19**, 15, DOI: 10.1186/s13059-017-1382-0 (2018).

48. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74, DOI: 10.1038/nmeth.1778 (2012).

49. Hicks, S. C., Townes, F. W. T., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562–578, DOI: 10.1093/biostatistics/kxx053 (2018).

50. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830, DOI: 10.5555/1953048.2078195 (2011).

51. Lehoucq, R. B., Sorensen, D. C. & Yang, C. ARPACK users'guide — solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods. In *Software, environments, tools*, vol. 6, DOI: 10.1137/1.9780898719628 (SIAM, 1998).

52. McInnes, L. & Healy, J. UMAP: uniform manifold approximation and projection for dimension reduction. *ArXiv* **abs/1802.03426** (2018).

53. Blondel, V. D., Guillaume, J., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.* **2008**, P10008, DOI: 10.1088/1742-5468/2008/10/p10008 (2008).

54. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300, DOI: 10.1111/j.2517-6161.1995.tb02031.x (1995).

55. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **96**, 10943–10943, DOI: 10.1073/pnas.96.19.10943-c (1999).

56. Rodchenkov, I. *et al.* Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Res.* **48**, D489—-D497, DOI: 10.1093/nar/gkz946 (2020).

57. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–D539, DOI: 10.1093/nar/gkj109 (2006).

58. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613, DOI: 10.1093/nar/gky1131 (2019).

59. Wadi, L., Meyer, M., Weiser, J., Stein, L. D. & Reimand, J. Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods* **13**, 705–706, DOI: 10.1038/nmeth.3963 (2016).

60. Altman, T., Travers, M., Kothari, A., Caspi, R. & D., K. P. A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinform.* **14**, DOI: 10.1186/1471-2105-14-112 (2013).

61. Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57, DOI: 10.1038/nprot.2008.211 (2009).

62. Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13, DOI: 10.1093/nar/gkn923 (2009).

Acknowledgements

This research was supported in part as a COVID-19 Research Project by Research Organization of Information and Systems (ROIS). K.I. and M.O. have been supported by JSPS KAKENHI Grant Numbers JP21H04905 and JP21J22938, respectively.

Author contributions

M.O., M.M., and K.I. designed the study; M.O. performed the experiments and analyzed the data; M.O., M.M., and K.I. wrote the manuscript.

Ethics declarations

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 220126SciRepSuppl.pdf
- 220203SciRepSuppl2.pdf
- SupplementaryFigure01.eps
- SupplementaryFigure02.eps
- SupplementaryFigure03.eps
- SupplementaryFigure04.eps
- SupplementaryFigure05.eps
- SupplementaryFigure06.eps
- SupplementaryFigure07.eps
- SupplementaryTable01.xlsx
- SupplementaryTable02.xlsx
- SupplementaryTable03.xlsx