Development of graph-based artificial intelligence techniques for knowledge discovery from gene networks

2021

Yoshihisa Tanaka

Contents

Preface

Living organism is a system composed of a wide variety of molecules. Their homeostasis is maintained by the concerted formation of networks in which various biomolecules are intricately interacted. For example, within the cell, there is a web of networks at various hierarchical levels, such as genome, RNA, protein, and metabolism. Therefore, in order to elucidate the mechanisms of biological phenomena, it is essential to understand these molecular networks that operate in the complex biological system.

In recent years, the development of high-performance measurement technologies have facilitated the comprehensive acquisition of a variety of biomolecular data. Among them, transcriptome analysis can provide information on the expression levels of a very large number of genes at once. The most widely used method is differential gene expression analysis, which extracts genes individually whose expression levels change significantly between samples under different conditions. While this method is useful for the characterization of samples and the search for biomarkers, it does not describe the relationships between genes. Another method is to map the gene set extracted as differentially expressed genes (DEGs) to known pathways in order to determine what biological functions the gene set is associated with. However, since this analysis relies on existing knowledge like pathways, it is difficult to find new relationships such as correlation and regulation between genes.

On the other hand, various methods have been proposed in order to discover novel relationships between genes from their expression data [1]. These methods represent latent regulatory gene networks, contributing to systematic understanding of biological phenomena and hypothesis generation. However, the gene networks that have been inferred are static, and therefore fail to represent the multiple states of the cell. Cells intrinsically take different states depending on time, environment, and biological function. As the transcriptional network changes drastically during cell differentiation [2], cells in different states are regulated by different networks. If a comprehensive gene network that can estimate the state changes under different conditions of each sample can be constructed, it is expected to obtain biological knowledge at the network level unlike the conventional analyses.

In this study, I developed a novel method to quantitatively analyze changes in the state of the network by constructing a basal gene network from a set of gene expression data and mapping each sample onto it. To construct the basal gene network, I employed a Bayesian network, which is a method of representing causal relationships between variables (genes) as a network based on data. The regulatory relationship between genes is represented by a nonlinear regression model on the expression levels of each gene. Here, as an extension of Bayesian network theory, I proposed a method to quantify the network for each sample by giving a sample-specific quantitative value to the estimated relationships. This enables us to characterize condition-specific networks for each sample, which was previously impossible. I described the development and establishment of the proposed method using cancer cell line data in Chapter 1. Next, I carried out evaluations of this method by applying it to coronavirus disease 2019 in Chapter 2. Furthermore, I attempted a unique approach to comprehensively analyze a large number of networks in different states using Graph Neural Networks for drug-induced liver injury in Chapter 3. The results of these studies were discussed in the following three chapters.

Chapter 1

System-based differential gene network analysis for characterizing a sample-specific subnetwork

1 Introduction

The use of high throughput technologies in molecular biology has led to the generation of large volumes of data, and thus, the development of precise methods for handling such large data is required. Understanding the cellular mechanisms at a system level forms a fundamental goal of these methods. Elucidation of the cellular mechanisms is indispensable for discovering biologically significant events, especially for clinical applications like finding new drug targets [3].

Researchers have developed several approaches to elucidate cellular mechanisms. One of the most popular methods is the conventional mRNA expression analysis to identify differentially expressed genes (DEG) [4, 5]. This method extracts independent genes using differences of expression levels between different conditions, e.g., control and perturbated samples. Likewise, pathway analysis maps genes onto known pathways to classify genes based on their enrichment in the maps of well-known pathways, such as the EGFR signaling and DNA damage repair pathways [6, 7, 8]. These traditional approaches, however, are unable to discover *de novo* pathways and relationships.

Network analysis constitutes a promising method to interpret the big omics data and researchers have applied it to many biomolecular investigations [9, 10]. However, since the gene network estimation brings in a huge number of putative gene regulatory relationships often expressed as a hairball [11], establishing a method for identifying biologically significant subnetworks or inter-gene relationships from such a large and complicated network

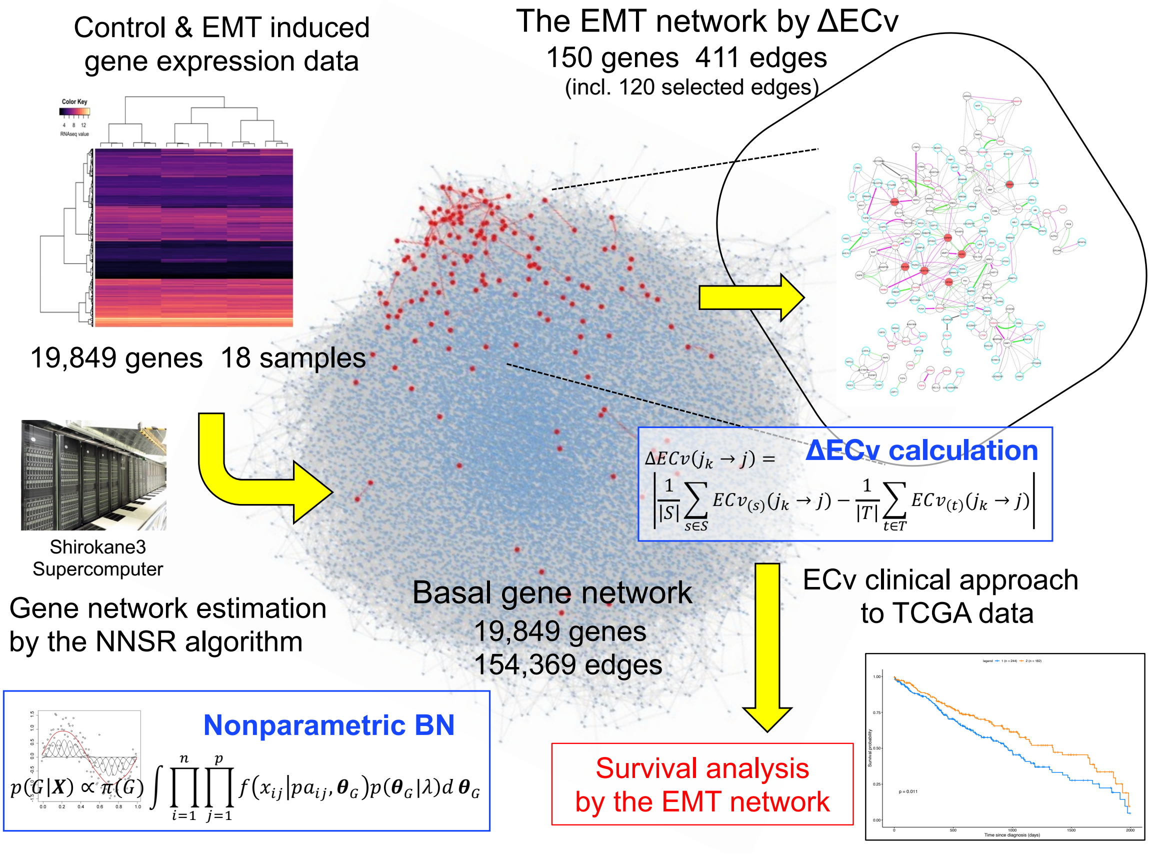

Control & EMT induced
gene expression data
Color Key
RNAseq value
19,849 genes 18 samples
The EMT network by ΔECv
150 genes 411 edges
(incl. 120 selected edges)
ΔECv calculation
$\Delta ECv(j_k \to j) = \left| \frac{1}{|S|} \sum_{s \in S} ECv_{(s)}(j_k \to j) - \frac{1}{|T|} \sum_{t \in T} ECv_{(t)}(j_k \to j) \right|$
Shirokane3
Supercomputer
Gene network estimation
by the NNSR algorithm
Basal gene network
19,849 genes
154,369 edges
ECv clinical approach
to TCGA data
Survival analysis
by the EMT network
Nonparametric BN
$p(G|\boldsymbol{X}) \propto \pi(G) \int \prod_{i=1}^{n} \prod_{j=1}^{p} f(x_{ij}|pa_{ij}, \boldsymbol{\theta}_G) p(\boldsymbol{\theta}_G|\lambda) d\,\boldsymbol{\theta}_G$

Figure 1.1: Overview of the proposed method. The center hairball (blue) is the basal network. The red nodes in the basal network represent the ΔECv-extracted EMT network.

has been challenging. Established methods that identify regulatory relationships in large gene networks mainly adopt two approaches: one focuses on hub genes that have a large number of connections [12, 13, 14, 15], whereas the other, known as differential network analysis [16], compares network structures derived from different conditions. While the former approach fails to extract edges regardless of the gene network analysis, the latter requires individual networks reflecting different conditions. Nonetheless, it remains difficult to satisfy sufficient sampling to generate data-driven and condition-dependent networks. For instance, my study presented in this Chapter 1 uses a dataset that consists of only three samples per condition, which makes a Bayesian network (BN) and structure-based approach unfeasible. In general, machine learning algorithms including BN require a large number of samples.

Herein, I propose a novel method to extract a biomedically differentiated subnetwork from a huge number of edges in a gene network estimated using a relatively small number of samples. Compared to the structure-based network analysis, my approach aims to extract edges, which represent the system-level differences between the samples in cellular networks without comparing network structures. Therefore, even with a limited number of samples for conditions of interest, this method will extract system-based differential networks. The outline of the proposed method is shown in Figure 1.1. Following the reported method [17], a gene network from the available gene expression data was estimated using nonparametric BN. This generates a basal network with more than a hundred thousand edges between approximately 20,000 genes. Most importantly, this method calculates an *Edge Contribution value* (ECv) to every single edge in a gene network with respect to each sample. The proposed ECv can quantify a particular edge for each sample using the estimated model. Therefore, based on the differences in ECv's between different samples, which imitate biologically or clinically-specific phenomena, this method highlights the differentiated subnetwork. This method defines the edges out of a possible hundred thousand links between genes, and thus demonstrates major differences in regulating the basal gene network potentially associated with target diseases or phenomena. Since the extracted subnetwork highlights specific differences between samples, it may portray expression datasets that are not used in network estimation or subnetwork extraction. I proved that the network ECv pattern with respect to a patient sample for certain types of diseases can be adopted for clinical classifications by validating real data.

The idea of estimating sample specific gene networks has already been addressed. For instance, Shimamura *et al.* (2011) [18] proposed a structural equation-based model that estimated sample specific regulator-modulator-target relationships. Their method assumes the gradual effect of modulators to parent-child relationships throughout the collected samples. Thus, this method generally requires sufficient samples to detect such relationships. Yu *et al.* (2015) [19] tried to extract personalized gene networks based on their differential network model. Their model assumes an existing network, such as the protein-protein interaction network, and combines it with genes extracted by traditional methods such as DEG and gene pairs that are independently evaluated by their novel differential expression covariance method on that network. Kuijjer *et al.* (2019) [20] proposed a method to estimate sample-specific regulatory networks where the linear combination of edge weights in particular networks forms the aggregated network. Unlike these existing methods, the proposed method estimates a globally optimized structure as a cellular model with the BN model and correlates edge differences to the samples using the estimated model parameters. None of the existing methods realize extraction of the edge from a limited number of samples and a large network except for the proposed method.

I examined this method on epithelial-mesenchymal transition (EMT) [21] to understand its process through the representative EMT subnetwork; I then applied it to The Cancer

Genome Atlas project (TCGA) clinical data [22]. In this analysis, the basal network was estimated from the small number of samples mimicking EMT in lung cancer cell lines. By comparing ECv's between EMT-induced and control samples, I succeeded in extracting the EMT-characterized network. In addition, I applied this EMT network to TCGA clinical data to test if the network can generate a prognosis profile, as EMT is a major factor involved in the prognosis of cancer patients. These results show that the prognosis for lung cancer patients was partially associated with the ECv patterns in the EMT network, according to the survival analysis. This indicates that the proposed method correctly extracts a subnetwork determining patients' EMT characteristics from a large number of nodes and edges in the network.

2 Methods

2.1 Nonparametric Bayesian network

As described in the Introduction, I used nonparametric BN to estimate a gene-to-gene regulatory system from gene expression data [23]. The BN estimation is an unsupervised machine learning algorithm that is able to capture cause-and-effect relationships among variables from its observations by optimizing the global structure of the network. Therefore, it constitutes an ideal method to estimate gene regulatory systems and has been successfully applied to many gene network analyses [14, 24, 25]. In this section, I provide a brief explanation of the model.

Let $\boldsymbol{X}$ be an n-by-p data matrix whose element x_{ij} corresponds to the observation of the j-th variable, that is, the expression value of the j-th gene, at the i-th sample, where n represents the total number of samples and p the number of variables. In the nonparametric BN model, I consider the joint density of all the variables and assume that it can be decomposed as the product of the local conditional densities such as

$$f(x_{i1}, \ldots, x_{ip}; \boldsymbol{\theta}_G) = \prod_{j=1}^{p} f(x_{ij} | \boldsymbol{pa}_{ij}^{G}(x_{ij}); \boldsymbol{\theta}_j), \tag{1.1}$$

where $\boldsymbol{pa}_{ij}^{G}(x_{ij}) = (pa_{i1}^{(j)}, \ldots, pa_{iq_j}^{(j)})$ is the set of observations in the i-th sample of q_j dependent variables of the j-th variable and $\boldsymbol{\theta}_G = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$ is the parameters of the conditional densities. This decomposition can be represented by a directed acyclic graph consisting of nodes representing variables and edges connecting the nodes called "children," to their dependent variables called "parents." To model parent-child relationships, B-spline nonparametric regression is employed. In the model, the gene expression is represented by

$$x_{ij} = m_1^{(j)}(pa_{i1}^{(j)}) + \cdots + m_{q_j}^{(j)}(pa_{iq_j}^{(j)}) + \varepsilon_j, \tag{1.2}$$

where $m_k^{(j)}(pa_{ik}) = \sum_{l=1}^{M} \gamma_{lk}^{(j)} b_{lk}^{(j)}(pa_{ik}^{(j)})$ for $1 \le k \le q_j$ and $\varepsilon_j \sim N(0, \sigma_j)$. Here, $b_{lk}^{(j)}(\cdot)$ is a third-order B-spline function which is determined by the range of the observations and $\gamma_{lk}^{(j)}$ is its coefficient. M is the number of the B-splines and $M = 20$ is used as shown in Imoto *et al.* (2002) [23]. The local density in Eq. (1.1) can be written as

$$f(x_{ij}|\boldsymbol{pa}_{ij}^{G}(x_{ij}); \boldsymbol{\theta}_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{\left(x_{ij} - \sum_{k=1}^{q_j} \sum_{l=1}^{M} \gamma_{lk}^{(j)} b_{lk}^{(j)}(pa_{ik}^{(j)})\right)^2}{2\sigma_j^2} \right),$$

where $\boldsymbol{\theta}_j = (\gamma_{1,1}^{(j)}, \ldots, \gamma_{M,q_j}^{(j)}, \sigma_j^2)$ is the parameter vector for the local density to be estimated from the observations. The network structure can be determined based on the maximization of the marginal posterior

$$p(G|\boldsymbol{X}) \propto \pi(G) \int \prod_{i=1}^{n} f(x_{i1}, \ldots, x_{ip}; \boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})d\boldsymbol{\theta}_G,$$

where G represents the network structure, $\pi(G)$ the prior probability of G, and $\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})$ the prior distribution of $\boldsymbol{\theta}_G$. Here $\boldsymbol{\lambda}$ is the hyperparameter vector determined also by the maximization of the posterior. Searching for the optimal structure of the BN is known to be an NP-hard problem. Thus, I use the Neighbor Node Sampling and Repeat (NNSR) algorithm [17] that is applicable to more than twenty thousand genes for obtaining an approximated structure of the huge BN. This network is called *the basal network* in the later steps.

2.2 Proposed method for evaluating sample specific edge contribution values

The basic idea is to evaluate a single edge value of an observed sample through the mathematical model estimated as a BN. In the model, the expression value of a child gene is represented by a linear combination of their parent values transformed by the functions denoted as $m_k^{(j)}(\cdot)$ in Eq. (1.2). Considering this nonparametric regression function, $m_k^{(j)}(pa_{ik}^{(j)})$ can be regarded as a contribution of the k-th parent of the j-th gene to the expression value of x_{ij}, because these values of q_j parents constitute the value of their child. According to this, I define *Edge Contribution value* (ECv) of edge $j_k \to j$ with respect to the i-th sample as

$$\mathrm{ECv}_{(i)}(j_k \to j) = m_k^{(j)}(pa_{ik}^{(j)}), \tag{1.3}$$

where j_k represents the index of the k-th parent of the j-th gene.

Since ECv is calculated from the model parameters to every single edge in the estimated network, the ECv of an edge can be considered as its edge weight, representing how it contributes to the link between a node pair for a certain sample. However, this calculation of ECv alone does not represent an evaluation of the absolute importance of the edge, because it is impossible to determine the size of calculated ECv. This is similar to gene expression values where differential expression is considered. Therefore, two ECv's of an edge should be compared between samples. For instance, let us assume that there are control and drug-perturbated samples in an *in vitro* cell assay. To obtain the differential edge in terms of ECv, I define a variation of ECv, ΔECv, as

$$\Delta\mathrm{ECv}(j_k \to j) = \left| \frac{1}{|S|} \sum_{s \in S} \mathrm{ECv}_{(s)}(j_k \to j) - \frac{1}{|T|} \sum_{t \in T} \mathrm{ECv}_{(t)}(j_k \to j) \right|, \tag{1.4}$$

where S and T are sets of indices of samples observed in particular conditions, respectively. The graphical representation of ΔECv is shown in Figure 1.2.

The ΔECv of an edge is the absolute difference that stands for the difference in edge contribution between samples of different conditions. Thus, an edge that shows significant ECv difference between certain samples represents a distinctive edge. In this experiment, ΔECv implementation did not result in a large number of candidate edges that are needed

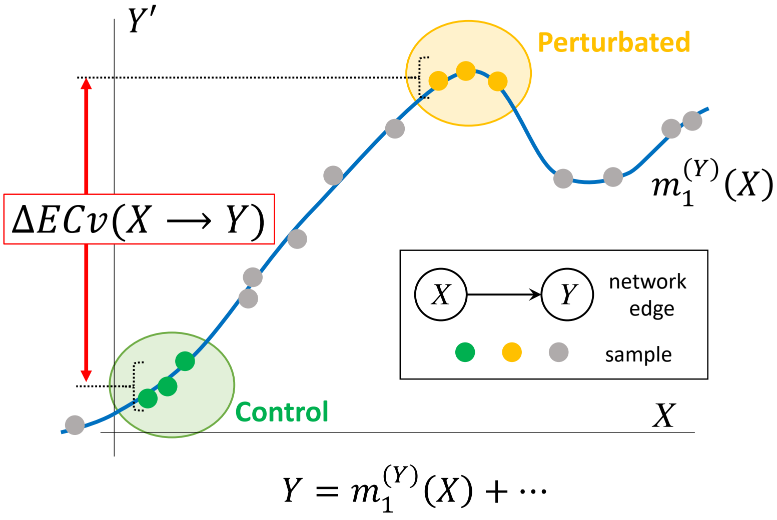


Figure 1.2: Graphical representation of ΔECv. Line (blue) is a nonparametric regression curve $m_1^{(Y)}(X)$ for edge $X \longrightarrow Y$ estimated with Bayesian Network. Plots for X axis correspond to actual mRNA signal values and Y' axis partial residuals where the effects of the other parents are subtracted from the children's mRNA signal values. Plots (green) and plots (yellow) represent, for instance, control and perturbated samples, respectively. Plots (gray) represent other values used for determining the regression curve. Values in Y' axis also stand for output through the regression function with parents' mRNA signals. By the definition, these correspond to ECv's. ΔECv is defined as the difference between two conditions. In this example, the difference of ECv's between perturbated and control samples, i.e., ΔECv of these two conditions is depicted.

for following the investigation. This suggests that ΔECv functions as a screening method for candidate networks from the estimated basal networks that are important under certain conditions. The network estimation with a small number of samples generally increases the number of false positive edges due to spurious correlations between variables. It is expected that the ΔECv can overcome this problem, because ΔECv might highlight a sample difference which is determined by not only the estimated model but also the sample-specific differences through the estimated model. Note that the number of samples in S and T here can be 1 because I focus only on differences of ECv so that only a single pair of samples is required to extract edges.

As a consequence of ECv calculation, a matrix consisting of ECv's for all of the edges and samples is obtained. This matrix is called an *ECv matrix*. More precisely, the ECv matrix $\boldsymbol{E}$ is an n-by-m matrix whose element e_{iv} corresponds to ECv of the v-th edge at the i-th sample, where m represents the total number of estimated edges and n the number of samples. The ECv matrix thus can be considered as a set of each sample's ECv's. Since ECv originally represents sample-specific profiling for each estimated edge, clustering of the ECv matrix highlights differences according to ECv's for each sample, which is capable of grouping samples based on their system-level similarities.

2.3 Data preparation

The microarray data for EMT analysis were acquired from Gene Expression Omnibus (GSE49644) [21]. The dataset is composed of three human non-small cell lung cancer (NSCLC) cell lines: A549, HCC827, and NCI-H358. The microarray experiments were replicated 3 times for both control and TGFβ-treated cells. Thus, the data consists of $3 \times 3 \times 2 = 18$ samples in total and 19,849 genes. The $\log_2$-transformed values of preprocessed data were applied to BN estimation and ECv calculation. The clinical and RNA-seq data of lung cancer patients [22] were acquired from the Genomic Data Commons Data Portal at TCGA and UCSC Xena [26]. NSCLC patients with either lung squamous cell carcinoma (LUSC) or lung adenocarcinoma (LUAD) were selected. The patients were first screened to obtain tumor specimens. RNA-seq data (ver 2017-10-13) was filtered to remove genes with a mean percentile lower than 15, resulting in 17,450 genes. In the clinical data (downloaded at 9 Dec 2019), entries for patients whose follow-up or decease data was more than 2000 days were removed. Further to these preprocesses, the patient data that were not common in the RNA-seq and clinical data were deleted. The number of the final patients for analyses was 426 (alive: 238, deceased: 188) for LUSC and 457 (alive: 285, deceased: 172) for LUAD.

2.4 Differential expression gene analysis

The differential expression gene analysis for GSE49644 microarray data was performed using R package limma [27]. Benjamini-Hochberg method was applied for calculation of false discovery rate [28].

2.5 Molecular function analysis

The functional analysis was generated through the use of Ingenuity Pathway Analysis [29].

2.6 ECv matrix clustering and survival analysis

Unsupervised-clustering for the ECv matrix was performed using the ‘ward.D2’ clustering method with Euclidean distance for both edges and samples in R. The following survival analysis was performed with log-rank test using R package surveminer and TCGAbiolinks [30, 31].

2.7 Network analysis and visualization

The network visualization and a part of network analysis were performed using Cytoscape [32].

2.8 Computation environment

All the computation for the network estimation and ECv calculation in this study was performed by the SHIROKANE supercomputer system (Shirokane3) at Human Genome Center, the Institute of Medical Science, the University of Tokyo, where the computation nodes were equipped with dual Intel Xeon E5-2670 v3 2.3GHz CPUs and 128GB memory per node.

3 Results

3.1 Basal gene network estimation

For the basal gene network estimation using the BN, I adopted the NNSR algorithm [17] as described in Materials and Methods. The algorithm repeatedly iterates subnetwork estimations in parallel many times for gene sets extracted by the neighbor node sampling method, and determines a final network structure by incorporating edges whose estimated frequencies are greater than the cutoff threshold. The frequencies correspond to strengths of edges in terms of stability or confidence of them. To begin with, I tuned these parameter settings for this analysis. As for the cutoff threshold, since the number of extracted edges is fixed with another quantitative measurement after the network estimation, I employed a threshold of 0.1, which is slightly relaxed comparing the default setting (threshold of 0.2), to include weak putative edges in this step. Although the algorithm can estimate networks for more than ten thousand genes, it supposedly requires more than a hundred samples for an input data. Therefore, I need to confirm if the algorithm works with an extremely small number of samples, i.e., 18 samples in this case. For this purpose, I defined the degree of concordance between two networks as the ratio of edges that are estimated in both networks to the total number of edges, and then I tested if the algorithm produces stable results using this degree. Following the recommendations of Tamada *et al.* (2011) [17], I performed the network estimation three times with 10,000 times for the number of iterations (denoted as "T") of the subnetwork estimation, and calculated the averaged degree of concordance for every estimated-network pair. As a result, the degree of concordance was 72.7%, suggesting that the algorithm could not produce stable gene network structures (Table 1.1). Therefore, I assessed whether the increased T results in stable network structures with the EMT dataset. The identical evaluations were performed as above for T=100,000, 500,000 and 1,000,000. I found that T=1,000,000 is sufficient for this network analysis owing to reduced error rate below 5% and the stable networks with just 18 samples. The results are summarized in Table 1.1. The final network structure consisted of 154,369 edges with a threshold of 0.1 and the node degree average of 15.55. This basal network is shown in Figure 1.1. When we used the dataset consisting of 19,849 genes and 18 samples, the computation time required for this network estimation was 7h 55m 42s at T=1,000,000 using 64 CPU cores.

Table 1.1: The list of concordance following different iterations.

T	concordance
10,000	72.7%
100,000	89.0%
500,000	94.3%
1,000,000	95.6%

3.2 ΔECv highlights the EMT-characterized edges

To examine if the proposed method is applicable to real data, I applied an ECv calculation to the basal network. The microarray data cells were treated with TGFβ to induce EMT. EMT is a cellular process in which metastasis and invasion are involved [33]. Although EMT has been well-investigated and many EMT-related genes were identified, its molecular mechanism is not fully understood. Once cancer cells invade tissues, this is critical for treatment of cancer and even more so for patient prognosis. Therefore, I aimed to extract an EMT-characterized subnetwork which represents a putative core mechanism of EMT as a cellular system modeled by the nonparametric regression. Compared to control cells, TGFβ-treated cells represented the EMT profile, which was confirmed by the alteration in cellular morphology and expression levels of EMT markers in the previous study [21]. Following ECv calculation, a 18-by-154,369 ECv matrix was obtained. Since the samples were replicated 3 times for both *control* and *EMT-induced* conditions over three cell lines, I calculated ΔECv with respect to each cell line as

$$\Delta\mathrm{ECv}^{(u)}(j_k \to j) = \left| \frac{1}{3} \sum_{s \in \{\mathit{EMT\text{-}induced}\}} \mathrm{ECv}^{(u)}_{(s)}(j_k \to j) - \frac{1}{3} \sum_{t \in \{\mathit{control}\}} \mathrm{ECv}^{(u)}_{(t)}(j_k \to j) \right|,$$

where u = A549, HCC827 and NCI-H358. The edges that were assigned high ΔECv can be considered as significant differences in the cellular system between control and EMT-induced samples in each cell line. The distribution of ΔECv values for each cell line is displayed in Figure 1.3A. As shown, most edges do not show significant differences between the two conditions. The number of edges with ΔECv more than 1.0 for A549, HCC827 and NCI-H358 were only 946, 1420 and 1041, respectively, out of 154,369. Given that the total number of edges in the estimated network is 154,369, ΔECv filters out edges which are assumed to be significantly different in EMT. To compare the $\log_2$ fold change (FC) distribution of mRNAs, which is a standard indicator in DEGs, their histograms were overlapped (Figure 1.3A). This showed that the distribution of ΔECv is much steeper than that of $\log_2$FC throughout the thresholds, suggesting that ΔECv can be considered as a better indicator for extracting condition-dependent edges. I have set ΔECv threshold as 1.0 because it corresponds to approximately top 0.1% of edges out of the total number of edges. To gain reliable EMT-distinctive edges, I extracted 120 edges which exceeded the threshold in all of the three cell lines that are composed of 150 genes (Figure 1.3B, C). Because there are 9 samples both for TGFβ-treated and control cells, I can evaluate statistical significance of the extracted edges. I performed t-test for ECv between TGFβ-treated and control samples, and found that 108 edges out of ΔECv-extracted 120 satisfied a criteria of FDR-corrected p-value < 0.01 (Figure 1.3D). This supports that ΔECv extracted edges are statistically significant even though the basal network was estimated from a small number of samples. Furthermore, considering that $\log_2$-transformed expression data was used, threshold 1.0 for ΔECv was generally supposed to be a 2-FC cutoff for the estimated system. Therefore, I hypothesized that 1.0 was a moderate threshold for the EMT dataset and used this in the following analyses. The ΔECv heat map reflects the samples' distinctive ECv matrix for the selected 120 edges (Figure 1.3E). This shows that the EMT-induced and control samples were clearly separated into two clusters, which further suggests that the ECv method captures the EMT-induced *pattern* of cellular network differences.

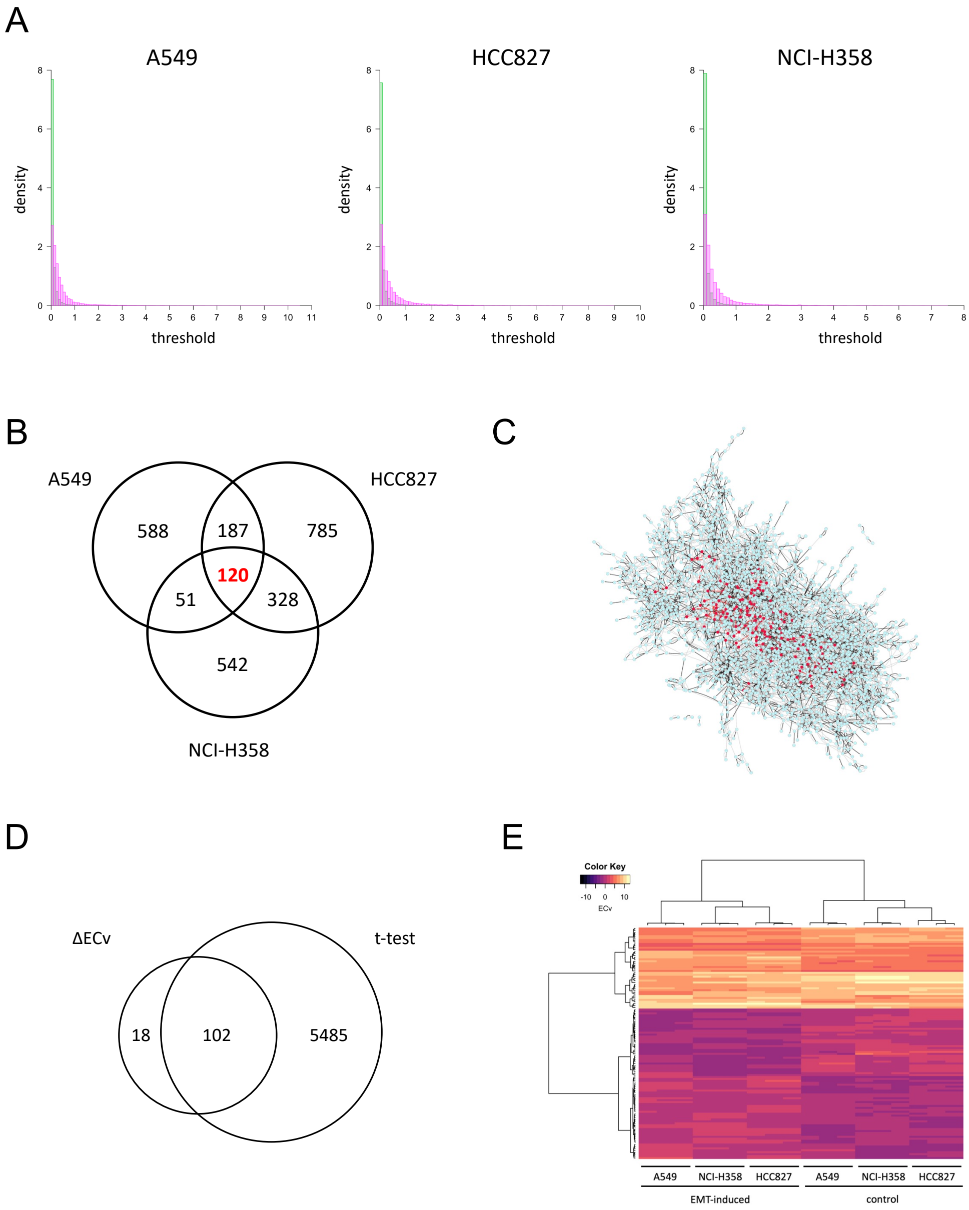


Figure 1.3: ΔECv analysis. (A) The histograms of absolute ΔECv (green) and absolute $\log_2$FC (magenta) for each cell line. FC is defined as TGFβ-treated/control. The number of total edges is 154,369 for ΔECv. The total number of genes for $\log_2$FC is 19,849. Y axis stands for density. X axis corresponds to the threshold for each ΔECv and $\log_2$FC. (B) The Venn diagram represents the numbers of ΔECv-extracted edges for all the cell lines with threshold 1.0. (C) The subnetwork of all the 2601 edges which were extracted from each cell line by ΔECv. The nodes (150) and edges (120) with red highlight the common edges for each cell line. (D) Venn diagram for ΔECv-extracted edges and t-test-significant edges. The t-test for ECv between TGFβ-treated (9 samples) and control (9 samples) were performed. The t-test-significant edges were selected with a criteria of FDR-corrected p-value < 0.01. The number of the obtained edges was displayed in Venn diagram. (E) Heat map and the result of hierarchical clustering for the ECv matrix of ΔECv-extracted 120 edges with 18 samples.

3.3 ΔECv unveils the EMT networks

Using these ΔECv-extracted edges, I aimed to build and visualize a network. I mapped 120 edges and linked the edges that are absent in ΔECv using their connections in the basal network, resulting in the establishment of the EMT-characterized subnetwork (hereafter referred as *the EMT network*) (Figure 1.4). Eventually, this network comprised 150 nodes and 411 edges. Remarkably, the EMT network shared many nodes of ΔECv-extracted edges, i.e., these edges were linked to each other. The biggest connected component in the 120 ΔECv-extracted edges consisted of 54 edges with 50 nodes. Many other connected components or independent edges were also connected to the biggest one within a single edge of the basal network, whereas only few edges were completely isolated. The inclusion of the basal network edges resulted in the connected component consisting of 371 edges with 127 nodes in the EMT network. Additionally, if I highlighted the ΔECv-extracted nodes and edges in the basal network whose node layout was arranged only using its topological structure, I observed that these were closely located and seemed to constitute a module in the basal network (Figure 1.1). Furthermore, I mapped top 5% rank hub genes for the basal network and the EMT network to compare their topological localization in the EMT network and identified 28 of the 1156 genes as the ECv-extracted genes (Figure 1.4). Interestingly, these basal hub genes did not function as hub genes in the EMT network.

3.4 Biological validation for the EMT network with the comparison between ΔECv and DEG

To investigate the extent to which ΔECv-extracted genes explain the EMT features biologically, I conducted a method-based comparison between ΔECv and DEG. The DEG analysis was performed by a criteria of absolute $\log_2$FC > 2 and FDR-corrected p-value < 0.00001, approximately following a previous report [21], resulting in 125 genes (Figure 1.5A). This DEG-extracted gene set principally reflects a difference between control and TGFβ-treated samples. The number of shared genes obtained by ΔECv and DEG was 71 (Figure 1.5B), suggesting that some population of the ΔECv genes belongs to the DEG-extracted gene set. Considering that 2-FC is a standard lowest cutoff for making a decision for potential DEGs, the remaining 79 genes for ΔECv and 54 genes for DEG might be exclusive for each method (Figure 1.5B). This implies that network-driven ΔECv can extract genes that the conventional DEG method never does. To get more of an insight into the biology involved, I examined whether biological functions are different between the gene sets obtained by these two methods. The molecular functions in the top 6 ranks out of 10 are exactly the same

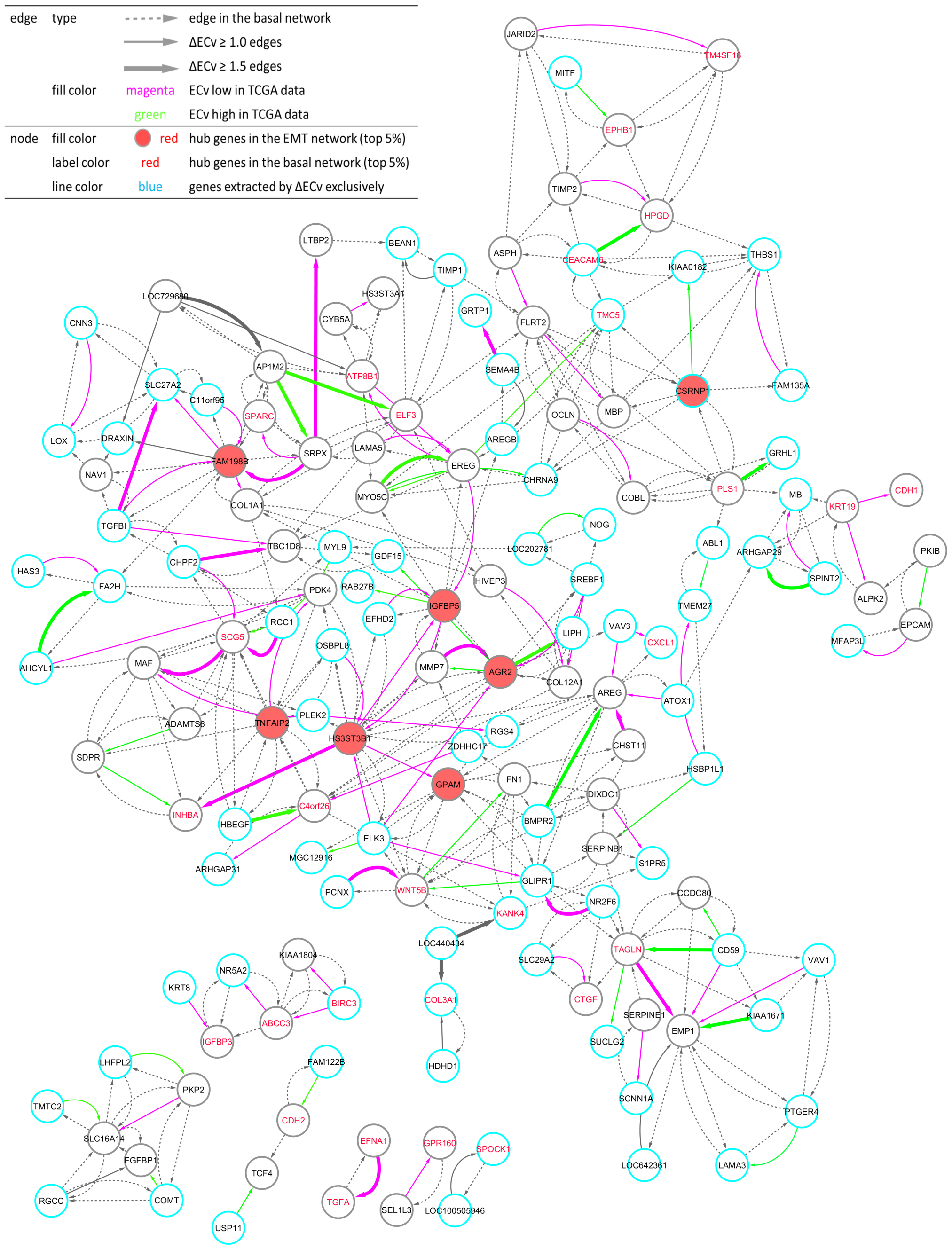


Figure 1.4: Visualization of the EMT network. 150 nodes and 411 edges constitute the network. The total number of connected components is 7. Node: Top 5% hub genes (filled with red) in the EMT network, and top 5% (labeled with red) hub genes in the basal network are displayed. Nodes (blue line) represent genes extracted by ΔECv exclusively. Edge: Bold edges (gray) and standard edges (gray) represent ones with absolute ΔECv more than 1.5 and 1.0, respectively. ECv high 38 (green) and low (magenta) 70 edges in TCGA data-fitting experiment (Figure 1.6A, B) are labeled. Dotted edges (gray) originally belong to the basal network.

between them (Figure 1.5C, D), and the top 3 functions of "cellular movement", "cellular development" and "cellular growth and proliferation" might represent the EMT features. This further supports that at least the major population of ΔECv-extracted genes consists of EMT-related genes. Moreover, I observed that representative EMT markers, CDH1 and CDH2 were included in the EMT network (Figure 1.4). On the other hand, the lower four molecular functions are different between ΔECv and DEG, which might reflect their characteristics (Figure 1.5C, D). These results show that the proposed method enables us to identify genes that are not done through DEG along with a biological validity, indicating the advantage of the ΔECv method.

3.5 A clinical approach using the EMT network

Finally, considering that ΔECv enables us to emphasize network differences between normal and clinically relevant samples, I attempted to investigate if the pattern of ΔECv-extracted edges over patients' samples identifies their properties regarding their EMT network suitability. I directly applied the EMT network on the RNA-seq data of the two types of lung cancer (LUSC and LUAD) patients by calculating their own RNA-seq specific ECv as

$$\mathrm{ECv}'_{(i)}(j_k \to j) = m_k^{(j)}(pa_{ik}'^{(j)})$$

where $pa_{ik}'^{(j)}$ represents the RNA-seq expression of the k-th parent of the j-th node in the basal network, and i the patient index in the TCGA dataset. Note that $m_k^{(j)}(\cdot)$ here is the same as the one estimated for the 18-sample microarray data and I calculated them for only the ΔECv-extracted edges. This experiment tried to reassign the EMT network, which was obtained by the microarray data, to a new RNA-seq data. I hypothesized that it can also highlight the EMT network in the RNA-seq data. Due to the differences in the total number of genes between the microarray and RNA-seq data, this was performed on the 136 genes shared with both EMT network and RNA-seq to gain an RNA-seq-derived ECv, resulting in 108 edges out of 120 edges. Therefore, this ECv calculation produces the 426-by-108 and 457-by-108 ECv matrix for both LUSC and LUAD. The ECv patterns of these matrices were shown as a heat map along with the result of the unsupervised clustering analysis (Figure 1.6A, B). The 108 edges were classified into ECv high and low clusters, indicating that these are involved in network gene up- or down- regulation. Since the patients were clearly clustered into two groups, I considered that these groups highlight a system difference in the process of EMT. Given that the metastasis and invasion that is a

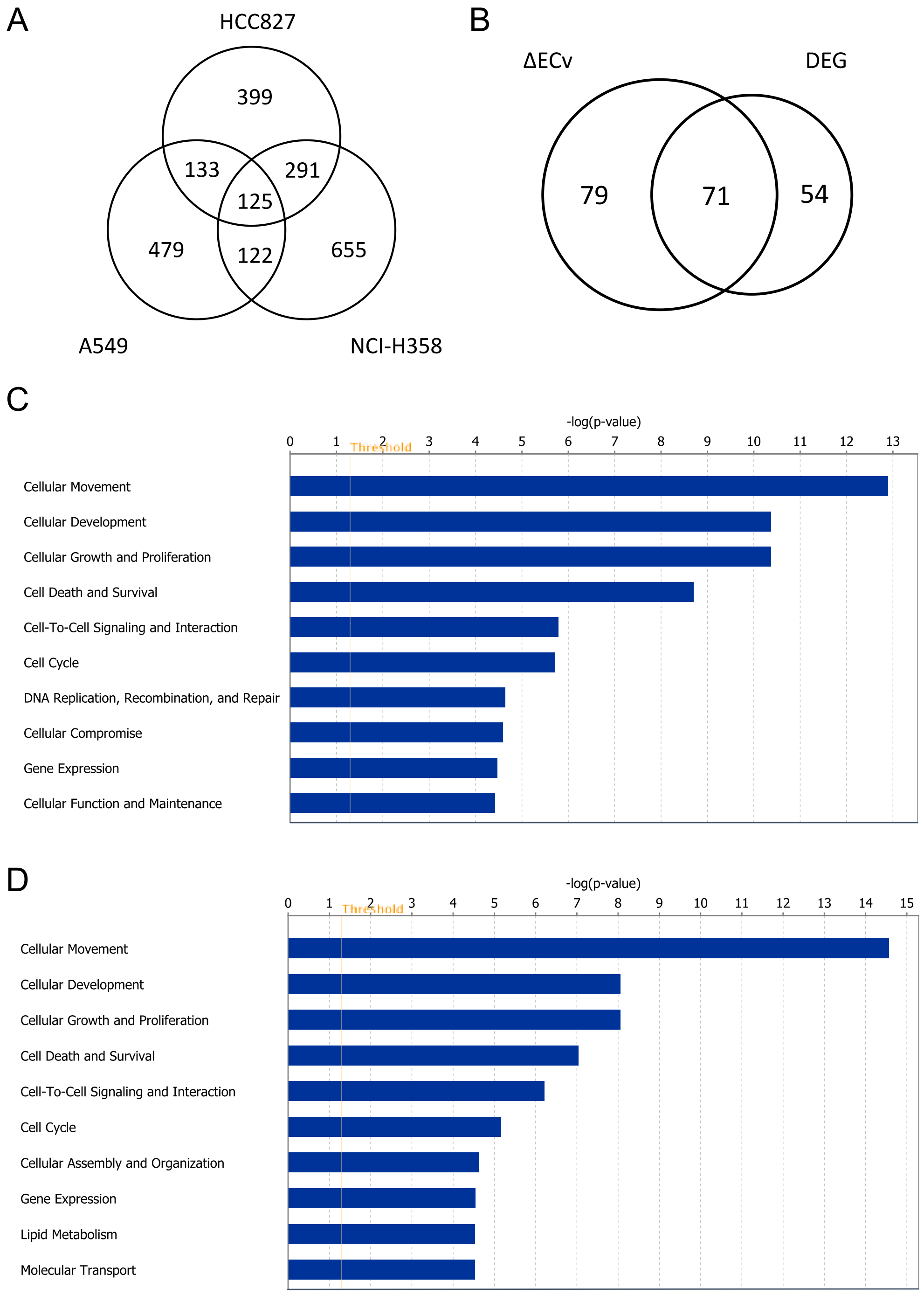


Figure 1.5: Comparison of ΔECv and DEG. (A) DEG analysis was performed with the criteria of absolute $\log_2$FC > 2 and FDR-corrected p-value < 0.00001 for each cell line; A549, HCC827 and NCIH-358. The number of the obtained genes was displayed in Venn diagram. (B) Venn diagram for the genes extracted by ΔECv and DEG. (C) Top 10 terms of the molecular function analysis for the ΔECv-extracted 150 genes. (D) Top 10 terms of the molecular function analysis for the 125 genes through DEG.

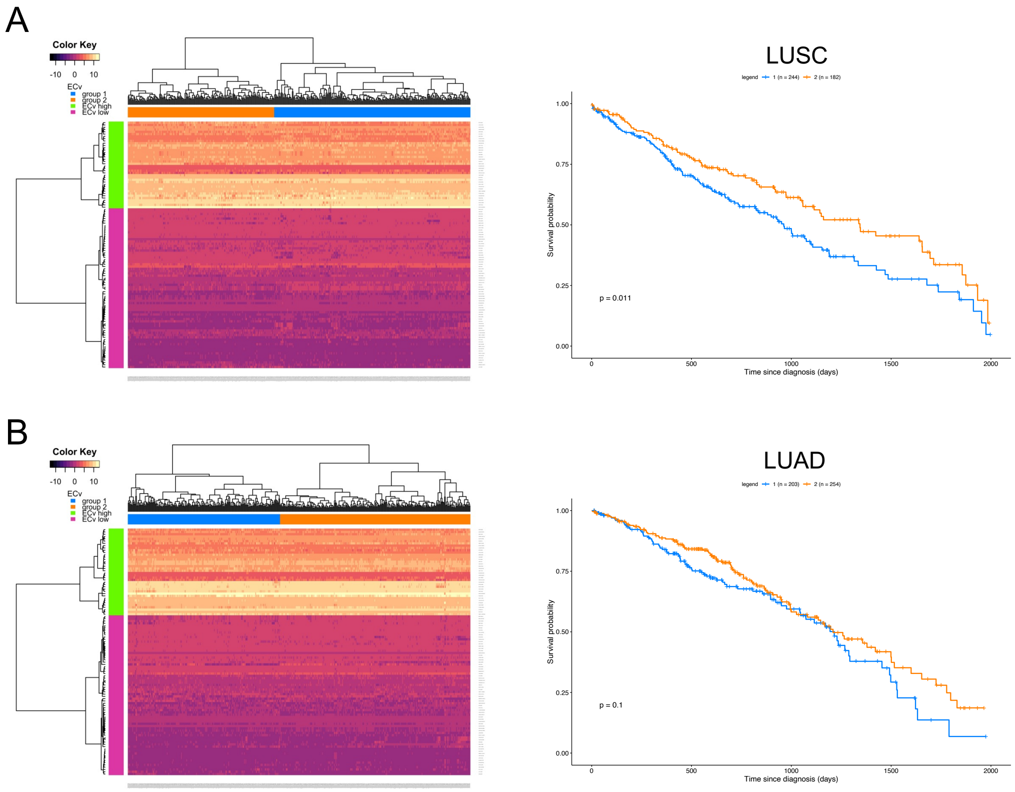


Figure 1.6: Unsupervised clustering and survival analysis for LUSC and LUAD. (A) Heat map with hierarchical clustering for the ECv matrix of 108 edges with 426 samples in LUSC RNA-Seq data. Kaplan-Meier curves for the two patient groups; group 1 (blue, n: 244) and group 2 (orange, n: 182), corresponding the patient clusters in the heat map in A. (B) Heat map with hierarchical clustering for the ECv matrix of 108 edges with 457 samples in LUAD RNA-Seq data. Kaplan-Meier curves for the two patient groups; group 1 (blue, n: 203) and group 2 (orange, n: 254), corresponding the patient clusters in the heat map in B. The survival analysis was performed using log rank-test for p value calculation.

consequence of EMT is fatal for patients' cancer prognosis, in order to ask whether these two groups are clinically different, I performed a survival analysis. The result showed that, although this failed (log-rank test p-value=0.1) for LUAD (Figure 1.6B), these two groups were significantly different (p-value=0.011 < 0.05) in terms of their prognosis for LUSC (Figure 1.6A), suggesting that the LUSC patients were distinguished into groups of better and worse prognosis by their ECv patterns based on their EMT network differences.

4 Discussion

Here, I report a novel method to obtain a sample-specific subnetwork using ECv derived from the estimated parameters of the BN. I investigated the application of this method in EMT biology and clinical data analysis. This method shows a potential ability of capturing the subnetwork and the patient-specific ECv patterns, and these patterns further characterized the prognosis. The prognosis in cancer patients depends on more than its metastasis. This is probably one reason why the combination of clustering and survival analysis works on LUSC, but not LUAD. On the other hand, the results of conventional clustering analysis on ECv matrices showed clear discriminations into two groups both on cell line microarray data and patient tumor sample RNA-seq data. Therefore, these results imply that individuals are distinguished through network differentiation using the proposed quantification of the edges.

Although translating laboratory experiments into the clinical context remains a big hurdle, this study indicates that my approach could be a powerful tool for bridging the divide between them. However, if we use it for this purpose, *in vitro* experiment design such as using cell lines should at least relate directly to practical clinical realities. In addition, because some datasets in replicate experiments involving *in vitro* assay are markedly homogeneous, I calculate ΔECv using mean ECv difference for each condition-specific sample set in this study. This, however, would not be appropriate for a heterogeneous dataset because mean ECv is not supposed to reflect a representative ECv for a particular set of samples. These issues are current limitations of the proposed method and can be improved in the future work.

As discussed previously, existing gene network analyses generally focus on hub genes that are supposed to catalog important regulatory genes in a network. In the EMT network, most of the basal network hub genes were either located at the corners of the EMT network or completely isolated from the biggest connected component (Figure 1.4). Only a few of these hubs are at the network's center, suggesting that the basal network hub genes are the master regulators responsible for various cellular processes in the basal network, but not in specific functional modules in the EMT network as discussed below. This may imply that the major difficulty in hub gene analysis lies in the acquisition of significant genes, especially for samples relating to specific conditions. However, this may also suggest that the network estimation involves appropriate consideration of all the system-level cellular features even when the number of samples are very small and are measured for a specific condition.

Within the biggest connected component in Figure 1.4, HS3ST3B1, FAM198B and IGFBP5 are the three top hub genes of the EMT network. Genes farthest from the hub genes were closely linked, suggesting that the subnetwork centered on the hub genes essentially represents EMT profiles depicted in the EMT dataset. In particular, I found that these components are more enriched in extracellular matrix (ECM) genes. ECM constructs a multi-layer scaffold structure located outside of the cell membrane and functions as an attachment between cells and tissues, assisting in cell growth, movement, development and differentiation [34]. Given that collagen, proteoglycan, and glycosaminoglycans (GAGs) constitute ECM, it is reasonable that ECM-functional genes, HAS3, FN1 and MMP7 are located proximal to the top three hub genes. HS3ST3B1 encodes an enzyme that controls the ECM environment following organization of heparin sulfate in GAGs and reports show its expression level regulates EMT [35]. A low level of FAM198B attenuates tumor growth and metastasis [36]. Overexpression of GFBP5 reduces EMT [37]. Although these reports support the notion that hub genes identified in this study actually engage in the EMT process, elucidation of the EMT mechanism remained incomplete as clear interactions have not been defined. Given that dysregulation of ECM was found to be a profound connection with EMT [38], the EMT network indicates a possible regulatory system of interactions between ECM and EMT. Moreover, PDK4—identified biologically as a novel EMT-associated gene [21]—has a network location close to the hub genes, supporting the finding of including this previously validated gene and the interaction of PDK4 with EMT-related genes.

Recently, the emergence of deep learning technology has allowed us to apply it to many scientific fields, including biology and medicine. However, such a situation depends on a large number of data. More importantly, the explainability of deep learning technology remains elusive. This disadvantage is a key issue, especially for biological or clinical situations, because the predictive process must also have responsibility for interpretation and ultimate outcome. In contrast, BN has been conventionally developed as an explainable model, and thus it is considered to be more appropriate to these fields. However, BN was unable to explain the individuality of samples as well as additional statistical models intended to interpret the population of certain data. In this study, by overcoming this weakness of BN, ECv has proved to be a more powerful tool in terms of explainability in machine learning. Therefore, the application possibilities of this BN model to analyses of biomedical data beyond existing bioinformatics methods makes it superior to other models. Although multi-omics analysis is becoming mainstream in biology, single transcriptome analysis still possesses a broad scope. Integration of other types of omics data such as genome, proteome, and epigenome with transcriptome using BN would lead us to analyze these big data more precisely than in existing studies, which in turn may result in a new approach in precision medicine.

Chapter 2

Dynamic changes in gene-to-gene regulatory networks in response to SARS-CoV-2 infection

1 Introduction

The newly emerging coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has spread rapidly worldwide [39, 40], with more than 21,000,000 cases of coronavirus disease 2019 (COVID-19) and 770,000 deaths as of August 18, 2020 [41]. This pandemic outbreak has drastically changed our society and has compelled us to be vigilant of the continuous risk of SARS-CoV-2 infection [42]. To overcome this dire situation, the development of novel drugs or vaccines continues to be an urgent global challenge. During the therapeutic development process, the elucidation of cellular mechanisms is essential for the discovery of potential targets; the fundamental question to be solved is how the SARS-CoV-2 influences host cells and causes COVID-19 at the molecular level. However, the cellular mechanisms underlying COVID-19 are poorly understood.

High-throughput technologies have contributed to the acquisition of a large amount of "omics" data, which has provided comprehensive information on cellular systems. These technologies have also been utilized during the current research into SARS-CoV-2. Several reports have provided various clues to understanding the global cellular signatures in response to SARS-CoV-2 infection at both the proteome and transcriptome levels [43, 44, 45]. Recently, network-based approaches have attracted great interest in the use of emerging omics data for drug discovery and systems biological analysis in the current field of SARS-CoV-2 research [46, 47, 48, 49, 50, 51]. Their major approaches combine publicly available sources, including knowledge of the already established pathways and drugs with these omics data to reconstruct molecular networks. However, these networks do not sufficiently

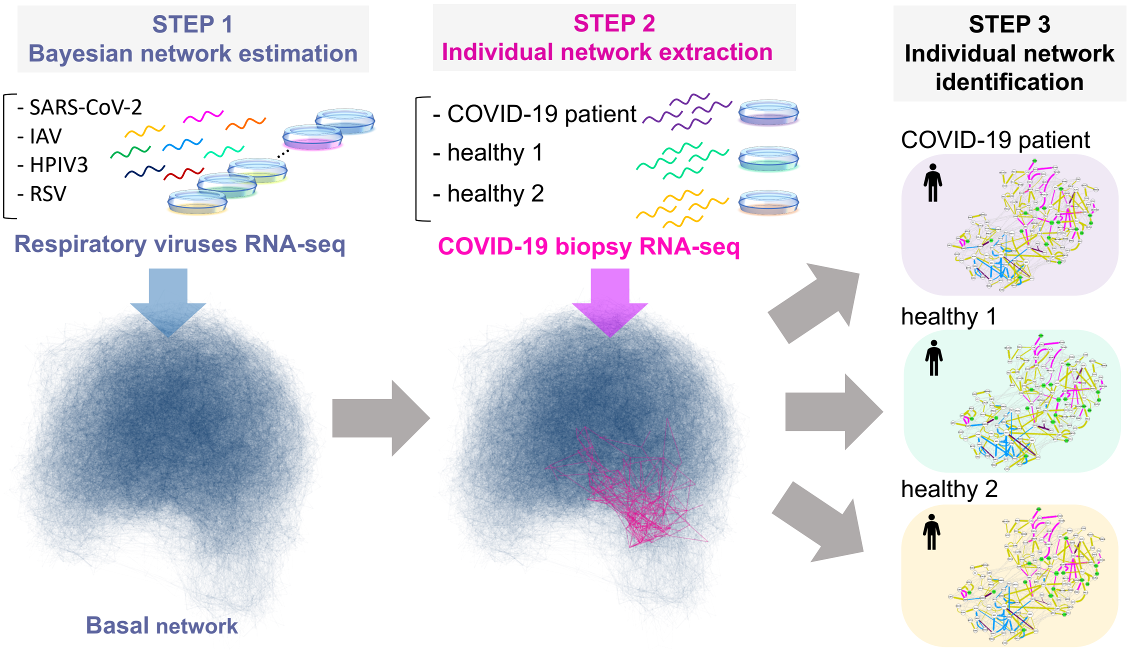


Figure 2.1: Illustration of overview. The hairball (blue) is the basal network consisting of 127,126 edges and 15,258 nodes established using the respiratory viruses RNA-seq including SARS-CoV-2. The highlighted-network (magenta) in the basal network represents the COVID-19-perturbated network extracted by using the biopsy RNA-seq.

represent a real cellular system for the following two main reasons: 1) public data consist of heterogeneous knowledge that has been accumulated throughout longstanding biological research; and 2) previous studies use mixed networks that combine data from various samples, but cannot reflect an individual cell-/patient-specific network.

To address these problems, I developed a method to extract a core sample-specific network from a massive gene network generated from a Bayesian network in Chapter 1 [52]. Gene regulatory network estimation has been developed as a prospective method to model the cellular system using omics data [12, 13, 14, 15, 16]. Although the Bayesian network-based approach can infer the cause-and-effect relationships between genes with transcriptome data, the key issue is the extraction of biologically significant information from a huge, complicated network, which is often sarcastically referred to as a hairball [11]. My unique framework consists of the following three steps: 1) estimation of a global gene network; 2) extraction of context-specific core networks based on differences in molecular systems from the global network; and 3) identification of a sample or patient-specific network (Figure 2.1). The prominent advantage is that it enables us to identify putative context-specific or sample-specific potential sets of edges in the form of a network, that is, gene-to-gene relationships with directions as well as nodes.

In this study, by using the developed framework for gene network analysis, I have presented the core host cellular systems involved in SARS-CoV-2 over several *in vitro* experiments, including different viral loads, cell lines, and respiratory viruses. No studies have been performed on the computational data-driven gene regulatory network approach for SARS-CoV-2. I characterized interferon signaling and subsequent inflammatory signaling cascades as significantly changed networks in human host cells, which represent the innate antiviral immune system in response to SARS-CoV-2 infection. Additionally, recent studies have reported that patients with COVID-19 exhibit various clinical outcomes depending on each patient, and that a certain proportion of patients will experience a severe disease [53, 54, 55]. Therefore, it is much more important to reveal the cellular mechanisms causing these clinical symptoms at the individual level. To this end, I have further identified the gene networks specifically for patients with COVID-19. I believe that the landscape of gene networks characterized in this study is beneficial for understanding the mechanisms by which cellular systems respond to SARS-CoV-2 and to further drug development.

2 Methods

2.1 Global gene network estimation and core network extraction

The network estimation and network extraction were performed using the method developed in Chapter 1. Briefly, this method first estimates a global gene network, called the *basal network*, which includes all the genes in a dataset using a Bayesian network with B-spline nonparametric regression with the NNSR algorithm [23, 17]. After the basal network estimation, I then quantified every single edge with respect to a certain sample in terms of the system-level usage of the edge with the estimated mathematical model using the ECv method developed in Chapter 1 [52]. The ΔECv definition in this study followed Eq. (1.4) in Chapter 1, where S and T were sets of infected and control (mock) replicated samples, respectively. As the target dataset includes control samples for a particular series of experiments, I can extract certain core networks from them by calculating ΔECv for the series of experiments using their corresponding control samples. For example, a SARS-CoV-2-perturbated core network was extracted by calculating ΔECvs for the SARS-CoV-2-infected and their corresponding mock-triplicate samples. As shown in Chapter 1 [52], we generally employ $\Delta\text{ECv} \geq 1.0$ for the threshold and carry out statistical t-tests to extract a core network extraction. This threshold approximately corresponds to 2-fold changes in differentially expressed genes for the extracted genes. Thus, I considered that the extracted networks, including edges and nodes, were significantly activated by the infection in cellular regulatory systems. In this study, statistical tests were not performed due to the small number of samples.

2.2 Proposed relative contribution of edges for characterization of individual networks

This ECv development allowed for a new solution for gene network analyses. In Chapter 1 [52], I succeeded in characterizing network profiles by calculating ECvs for edges in a ΔECv-extracted core network with respect to many samples from patients with cancer. Conventional clustering onto these calculated ECvs led to the identification of prognosis-related subgroups. Thus, I demonstrated that the differences and similarities in the edge profiles of the network could be captured as patterns of ECvs. Despite the high availability of ECvs, it is impossible to directly compare ECvs between individual samples because ECvs have different sizes depending on the estimated pairwise edge and the sample. The normalization of ECvs across samples is inappropriate for this purpose due to the mutual dependency of the individual network on each sample. Thus, it is not possible to highlight the differences in regulatory systems at an individual level.

For these reasons, ECvs are not appropriate for the analysis of individual networks. To overcome these drawbacks, I have proposed a novel method, *relative contribution* (RC), to quantify edges with respect to individual samples using the estimated gene network model. I hypothesized that the differences in individual samples in terms of the cellular system could be attributed to the differences in the ratios of the contributions of edges connecting to a certain node in the network. Edges with different samples need to be described as differently weighted edges according to the ratios of effects between parents that regulate or are connected to a certain gene. Additionally, the quantification of a network with a single sample needs to be independent from other samples and their distributions. To achieve this, I define the relative contribution of an edge with respect to a sample as

$$\mathrm{RC}_{(i)}(j_k \to j) := \frac{|\mathrm{ECv}_{(i)}(j_k \to j)|}{\max_{1 \le k' \le q_j} |\mathrm{ECv}_{(i)}(j_{k'} \to j)|},$$

where i represents a certain sample ($0 < \mathrm{RC} \le 1$). That is, an RC of the edge is the relative strength of the contribution of the edge to the maximum strength among the parents connecting to the same child node. The reason why an RC is not divided by the sum of the ECvs is that the range of RCs does not shrink depending on the number of parents of the child node. One drawback of RCs is that if the ratio of ECvs of the parents is not changed, the changes in parent values do not affect the RCs. However, the RCs of their downstream edges will be affected by such changes. Therefore, this drawback is not problematic in terms of the specification of differences in individual networks. Note that, similar to ECvs, sample i does not necessarily need to be a single sample used for the network estimation. As illustrated in the Results section, I have shown that RCs can be used to analyze individual networks, even if we have a single sample, or only a few samples, of gene expression data, as long as a basal network can be estimated from other datasets. RC, therefore, offers a significant enhancement to the framework for gene network analysis. The results have demonstrated that the framework, through an integration of the three key pieces—Bayesian network estimation, ECv, and RC—provides a powerful data-driven solution to seek biological phenomena through cellular systems ranging from a global level to an individual level.

2.3 Dataset

The transcriptome dataset GSE147507 was downloaded from the NCBI Gene Expression Omnibus [45]. The samples were infected with respiratory viruses, including SARS-CoV-2, and biological replicates were performed. The samples exclusive for human RNA-seq with 78 samples were selected. The detailed descriptions of samples are listed in Table 2.1, which was created according to the source paper [45]. Among the samples, four samples of the *in vivo* experiment (biopsy) data were pre-eliminated. The $\log_2$-transformed dataset was filtered to remove genes with a mean percentile lower than 30%, resulting in 74 samples and 15,258 genes. This preprocessed dataset of the 74 × 15,258 matrix was used as input for the basal network estimation. The biopsy dataset eliminated above, prior to global network estimation, consisted of four samples (two healthy samples and two COVID-19-positive samples). The RPM (reads per million)-normalized biopsy dataset was $\log_2$-transformed, and genes with at least one zero value were removed to obtain more reliable data. The two technical replicate samples for COVID-19 were averaged for the RC calculation. Following this preprocessing, the input dataset for the RC calculation finally comprised a 3 × 4,516 matrix. The RNA-seq samples used for ΔECv calculations in this study were: SARS-CoV-2 in A549 cells (MOI of 0.2/2 for 24 hr, n = 3) and the corresponding mock (n = 3); SARS-CoV-2 in normal human bronchial epithelial (NHBE) cells (MOI of 2 for 24 hr, n = 3) and the corresponding mock (n = 3); SARS-CoV-2 in Calu-3 cells (MOI of 2 for 24 hr, n = 3) and the corresponding mock (n = 3); human respiratory syncytial virus (RSV) in A549 cells (MOI of 2 for 24 hr, n = 3) and the corresponding mock (n = 3); human parainfluenza virus

3 (HPIV3) in A549 cells (MOI of 2 for 24 hr, n = 3) and the corresponding mock (n = 3); influenza A virus (IAV) in A549 cells (MOI of 5 for 9 hr, n = 2) and the corresponding mock (n = 2); COVID-19 (n = 2) and healthy (n = 2).

Table 2.1: The detailed list of sample descriptions. hACE2, human ACE2; IAVdNS1, a mutant IAV lacking its antiviral antagonist.

Series	cell	treatment	time	replicates
1	NHBE	mock	24 hr	3
1	NHBE	SARS-CoV-2 (MOI 2)	24 hr	3
2	A549	mock	24 hr	3
2	A549	SARS-CoV-2 (MOI 0.2)	24 hr	3
3	A549	mock	24 hr	2
3	A549	RSV (MOI 15)	24 hr	2
4	A549	mock	24 hr	2
4	A549	IAV (MOI 5)	24 hr	2
5	A549	mock	24 hr	3
5	A549	SARS-CoV-2 (MOI 2)	24 hr	3
6	A549	mock with hACE2 vector	24 hr	3
6	A549	SARS-CoV-2 (MOI 0.2) with hACE2 vector	24 hr	3
7	Calu3	mock	24 hr	3
7	Calu3	SARS-CoV-2 (MOI 2)	24 hr	3
8	A549	mock	24 hr	3
8	A549	RSV (MOI 2)	24 hr	3
8	A549	HPIV3 (MOI 2)	24 hr	3
9	NHBE	mock	12 hr	4
9	NHBE	IAV (MOI 3)	12 hr	4
9	NHBE	IAVdNS1 (MOI 3)	12 hr	4
9	NHBE	human IFNβ	4 hr	2
9	NHBE	human IFNβ	6 hr	2
9	NHBE	human IFNβ	12 hr	2
16	A549	mock with hACE2 vector	24 hr	3
16	A549	SARS-CoV-2 (MOI 0.2) with hACE2 vector	24 hr	3
16	A549	SARS-CoV-2 (MOI 0.2) with hACE2 vector and Ruxolitinib	24 hr	3

2.4 Pathway analysis

The canonical pathway analysis was performed through the use of Ingenuity Pathway Analysis software [29].

2.5 Network analysis and visualization

The network visualization and the network analysis were performed using Cytoscape (version 3.7.2 and 3.8.0) [32]. The genes for known drug targets were acquired from IPA knowledge database [29] and the representative drugs were listed in Table 2.2.

2.6 Computational environments

All the computations for the network estimation and the ECv calculations in this study were performed by the SHIROKANE supercomputer system (Shirokane5) at Human Genome Center, the Institute of Medical Science, the University of Tokyo, where the computation nodes were equipped with dual Intel Xeon Gold 6154 3.0GHz CPUs and 192GB memory per node.

3 Results

3.1 Estimation of the basal gene network in the involvement of respiratory virus infection using a Bayesian network

I first characterized a global gene network (hereafter referred to as the *basal network*) using a Bayesian network with a transcriptome dataset involved in the engagement of respiratory virus infection, including SARS-CoV-2, in several human cell lines [45] (Table 2.1). Since the outstanding characteristic of my approach is to capture sample-specific signatures from the basal network, it is preferable that various reactions are included to model complex gene regulatory systems [52]. To determine the basal network structure, I performed a network estimation using the neighbor node sampling and repeat algorithm [17], and screened the best algorithm parameters for the target dataset, as described in Chapter 1 [52]. Briefly, the network estimation was run three times independently, and the subsequent concordance test was performed to ensure the robustness and stability of the estimated basal network. I confirmed that the iteration number $T = 500,000$ satisfied less than 5% error (Error=4.0% for $T = 500,000$; error=5.3% for $T = 300,000$). The final basal network comprised 127,126 edges and 15,258 nodes, with a threshold of 0.05 and an average degree of 16.7. This final basal network was used for subsequent analyses.

3.2 Dynamics of host cellular network profiles at different viral loads of SARS-CoV-2

To examine the transition of host cellular system dynamics during the increase of SARS-CoV-2 viral loads, I characterized the networks perturbated by SARS-CoV-2 with two viral loads, namely a low multiplicity of infection (MOI) of 0.2 and a high MOI of 2 in A549 cells. I expected that cells exposed to different viral loads would present a unique cellular system, and that my approach could capture the fluctuation of system dynamics in whole cellular systems. To obtain differential core gene networks for each viral load, I followed multiple steps using an edge quantification technique, called the *edge contribution value* (ECv), established in Chapter 1 [52]. I first calculated ΔECvs following Eq. (1.4) , where S is SARS-CoV-2 infected and T is mock samples for each MOI condition (see Methods). The distributions of ΔECv showed that the innate cellular system was extensively more perturbated in the cells exposed to the high MOI than those exposed to low MOI (Figure 2.2A). I next set a threshold of 1 for ΔECv and obtained *differentially regulated edges* (DREs) from the basal network. The Venn diagram analysis for the ΔECv-extracted DREs showed that the number of DREs in the high MOI was larger than that in the low MOI (Figure 2.2B). Interestingly, the number of shared DREs between high- and low- MOIs was 42, which was only 6% of the total number of DREs in both conditions, thereby indicating that the underlying regulatory system between them was not similar. To confirm the biological involvement of the DREs, I performed canonical pathway analysis for the genes contained in the ΔECv-obtained DREs, which showed that these genes were associated with some cellular antiviral systems (Figure 2.2C). These results support that the components of the DREs are biologically relevant to viral infection.

To gain a greater insight into the profiles of the DREs from the perspective of network topology, I next generated networks using a set of all the DREs in the Venn diagram (Figure 2.2B). These DREs connected mutually and, in turn, generated subnetwork fragments of various sizes (Figure 2.2D). I reasoned that if these fragments had biological significance, these features should be reflected as modular, as biologically close functions in cellular systems link together and shape modules [56]. Hence, small-sized fragments were likely to be less informative, and I focused on the largest connected component among the various fragments. The largest connected component was extracted and the basal edges were additionally mapped on this network, which established the SARS-CoV-2-perturbated network with 130 nodes and 305 edges (Figure 2.3). I found that this network clearly consisted of three modules linked to each other. One module (module 1, yellow-marked region) was mainly composed of a set of DREs under low-MOI conditions, and its constituent

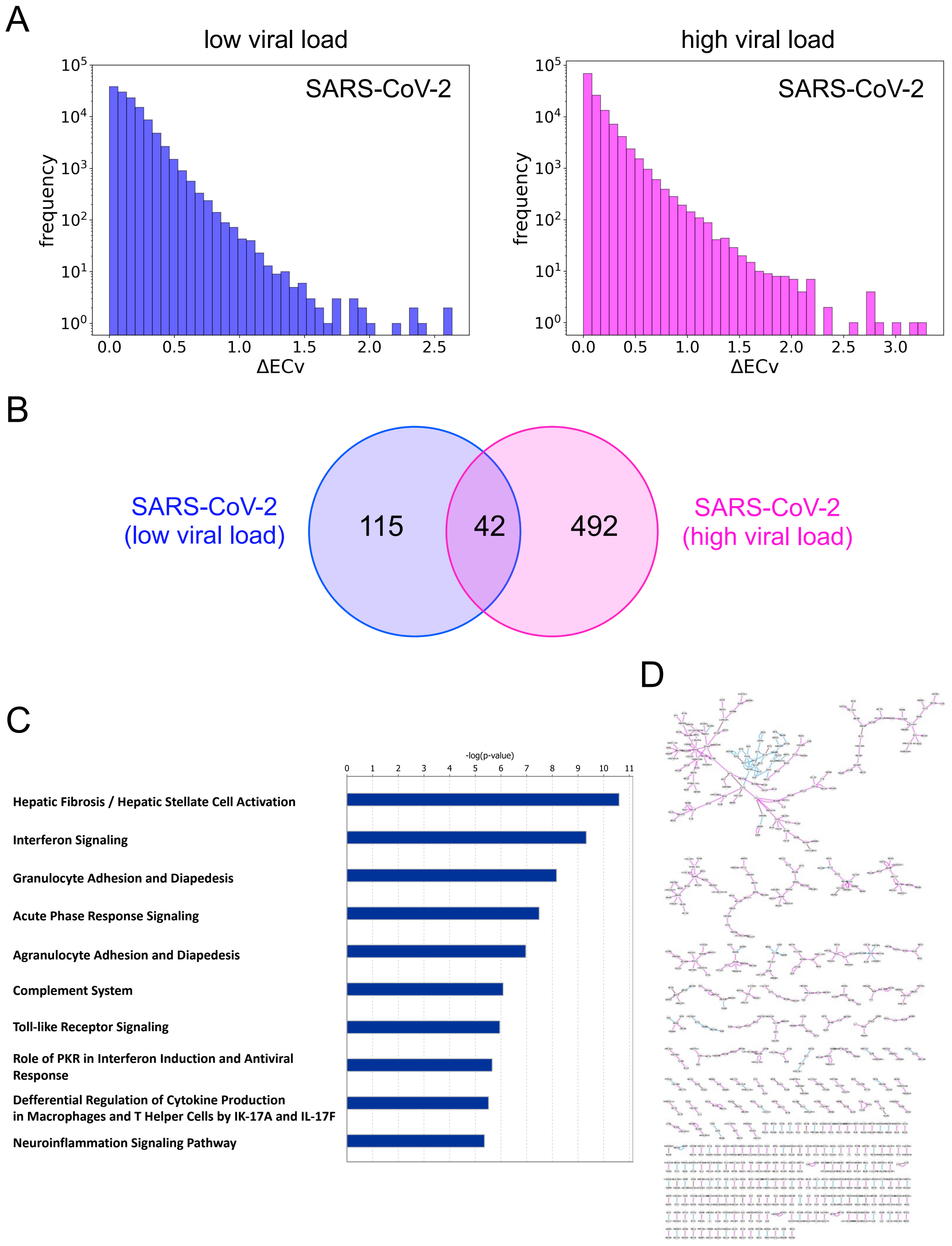


Figure 2.2: Dynamics of the SARS-CoV-2-perturbated network for different viral loads in host cells. (A) The histograms of ΔECv for different SARS-CoV-2 viral loads; a low MOI of 0.2 (blue) and a high MOI of 2 (magenta). The X-axis corresponds to the threshold for each ΔECv. The Y-axis shows the number of edges on a log scale. (B) The Venn diagram represents the numbers of differentially regulated edges (DREs) for two SARS-CoV-2 viral loads (blue: low MOI, magenta: high MOI) with a threshold of 1.0 for ΔECv in Figure 2.2A. (C) The top 10 terms of canonical pathway analysis for the genes comprising a union set of ΔECv-extracted DREs in the Venn diagram analysis (Figure 2.2B). (D) The whole illustration for the subnetwork fragments of various sizes is shown. Image of how the ΔECv-extracted DREs mutually connected and generated the subnetworks.

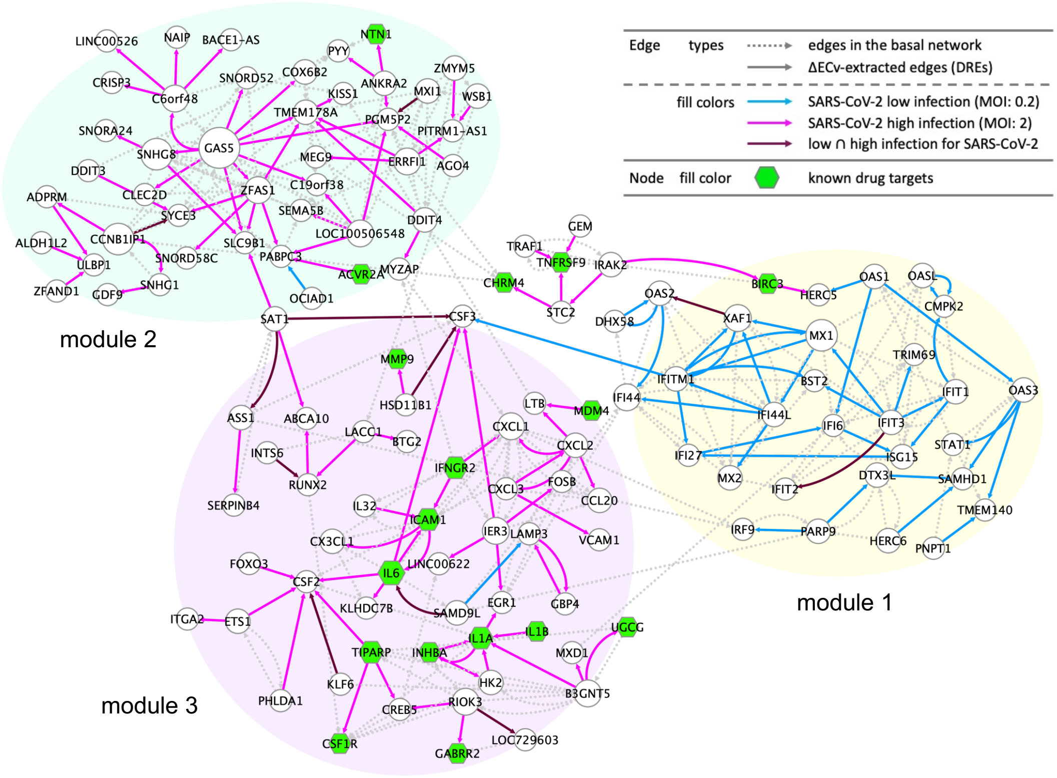


Figure 2.3: The SARS-CoV-2-perturbated network. The SARS-CoV-2-perturbated host cellular network in response to different viral loads in A549 cells (the SARS-CoV-2-perturbated network). The network comprises 130 nodes and 305 edges (including 155 basal edges). The colored solid edges represent the SARS-CoV-2-perturbated DREs; high MOI of 2 (magenta), low MOI of 0.2 (blue), and high MOI ∩ low MOI (purple). Dotted edges represent basal edges (gray). The nodes (green) represent the known drug target genes (Table 2.2). The node size represents the extent of outdegree.

elements were interferon (IFN)-stimulated genes (ISGs), namely IFIs, MXs, OASs, TRIMs, IFTMs, IRFs, and STATs. These highly orchestrated webs of various ISGs are induced by transductions of both IFN signaling and subsequent JAK/STAT signaling [57]. This evidence strongly suggests that module 1 represents the consequences of activation of both these signaling pathways by the acute antiviral response. Contrary to module 1, the other two modules (module 2, green-marked region; and module 3, purple-marked region) are mainly shaped by a set of DREs in the high-MOI condition. Modules 2 and 3 were found to comprise fewer IFN-related genes. While module 2 appeared to be a GAS5-centralized module, module 3 was composed of chemokines (CXCL1, CXCL2, CXCL3, CX3CL1, and CCL20), interleukins (IL6, IL1A, IL1B, and IL32), and colony-stimulating factors (CSF2 and CSF3), which have been implicated in inflammatory-related cytokine signaling followed by the acute activation of IFN and JAK/STAT signaling represented in module 1. In particular, the cluster of modules 1 and 3 likely represents the transition of the gene regulatory system in response to SARS-CoV-2 infection. Specifically, the cellular system perturbated by SARS-CoV-2 gradually switched to inflammatory signaling (module 3) via IFN and JAK/STAT signaling (module 1) as the viral load increased. This was consistent with the clinical observations of COVID-19, and may thus partially explain the process of cytokine storm syndromes, which is a severe clinical feature of COVID-19 [53, 55]. I also performed the same analyses among the four respiratory viruses (HPIV3, IAV, RSV, and SARS-CoV-2) and found that module 3 was exclusive for SARS-CoV-2 (Figure 2.4A-D). Collectively, I identified the SARS-CoV-2-perturbated network and its three modules, which reflected distinctive host cellular functions in response to SARS-CoV-2 infection.

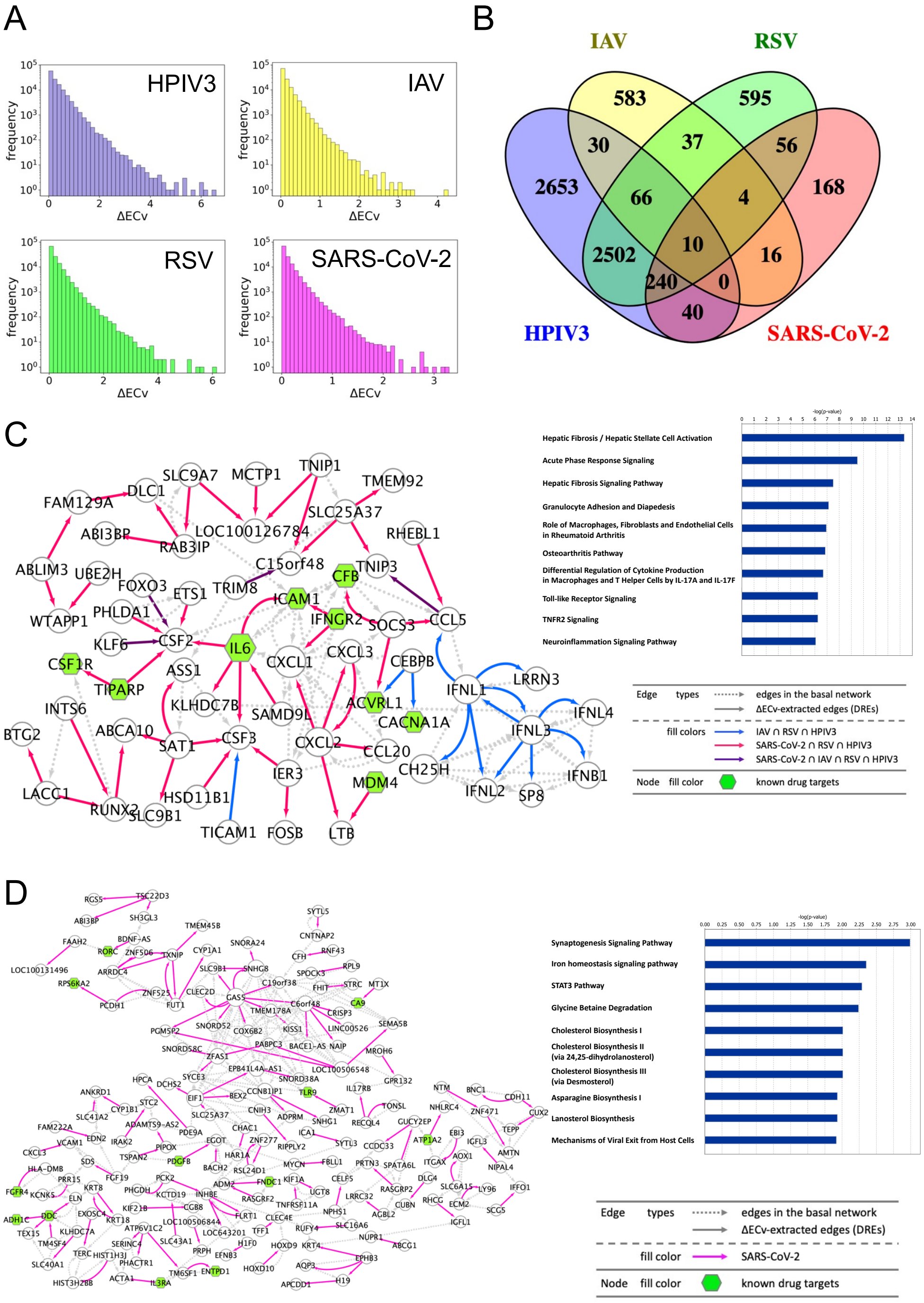
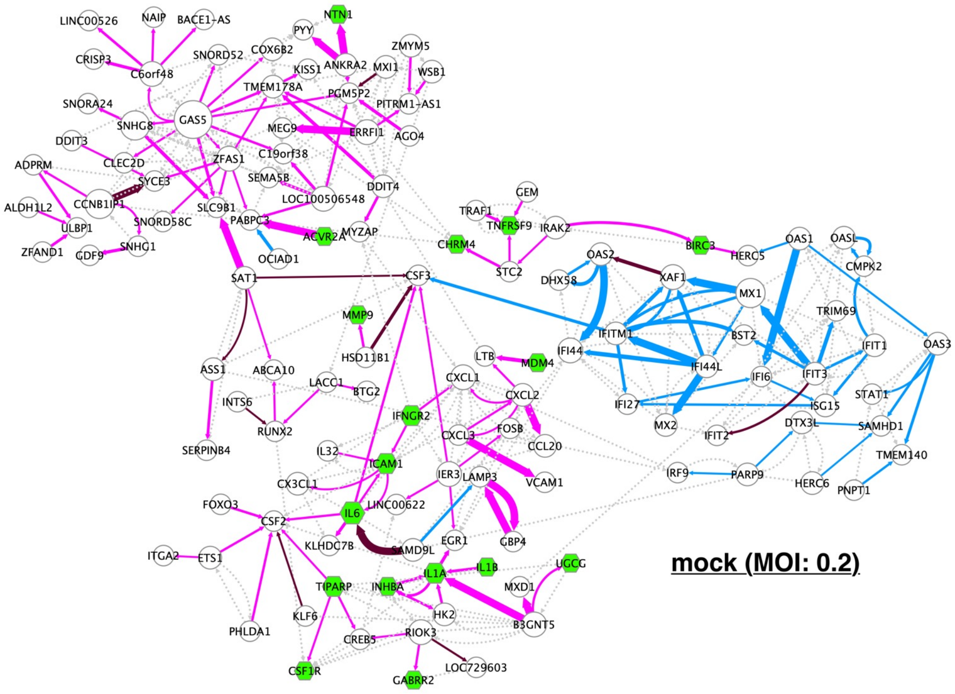


Figure 2.4: Network comparison analyses across four respiratory viruses. (A) The histograms of ΔECv for each respiratory virus as indicated: SARS-CoV-2 (MOI: 2), HPIV3, IAV, and RSV. The X-axis corresponds to the threshold for each ΔECv. The Y-axis stands for the number of edges on a log scale. (B) The Venn diagram represents the numbers of ΔECv-extracted edges for all respiratory viruses. (C) The respiratory viruses-shared network comprised 62 nodes and 116 edges (including 53 basal edges). The colored solid edges represent DREs; SARS-CoV-2 ∩ IAV ∩ HPIV3 ∩ RSV (purple), SARS-CoV-2 ∩ RSV ∩ HPIV3 (red), IAV ∩ RSV ∩ HPIV3 (blue). The top 10 terms of canonical pathway analysis for the genes of ΔECv-extracted DREs shared by at least three viruses in the Venn diagram (Figure 2.4B). (D) The SARS-CoV-2 specific network comprising 182 nodes and 295 edges (including 171 basal edges). The solid edges (magenta) represent DREs for SARS-CoV-2 (MOI: 2). The dotted edges represent the basal edges (gray). The size of the node represents the extent of outdegree. The nodes (green) are target genes for existing drugs (Table 2.2). The top 10 terms of canonical pathway analysis for the genes of ΔECv-extracted DREs exclusive for the SARS-CoV-2 in the Venn diagram (Figure 2.4B).

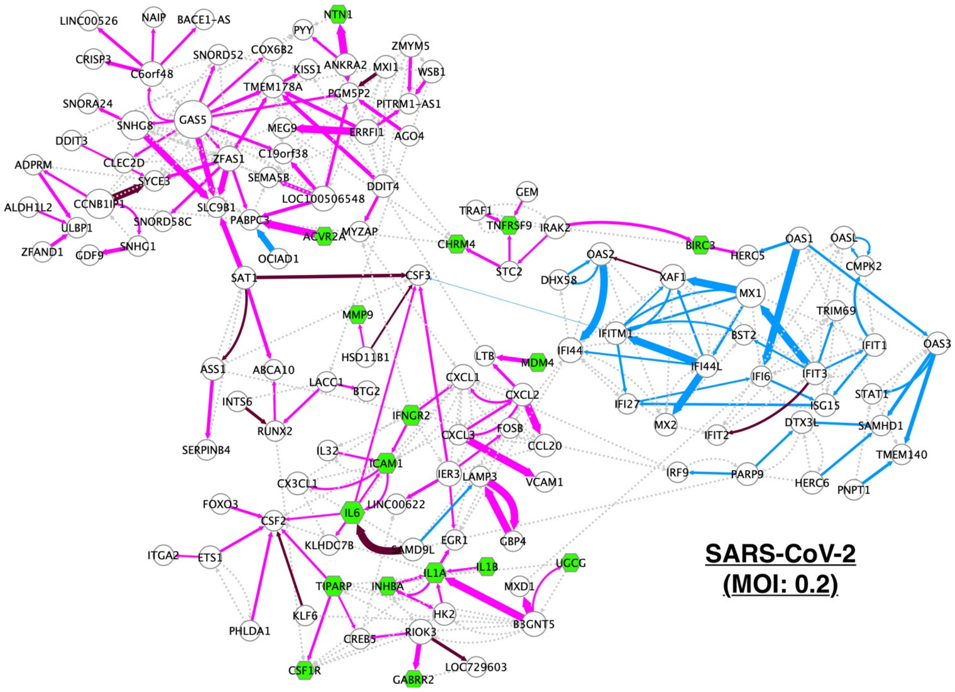
3.3 Characterization of the SARS-CoV-2-perturbated network at the individual sample level

Next, I determined how the signaling represented in the SARS-CoV-2-perturbated network (Figure 2.3) changed across the samples. To this end, I developed a novel quantitative method, called *relative contribution* (RC), to measure the edge contribution at an individual level. The mathematical definition of RC is described in the Methods section. Within a set of pairwise parent-child relations for a certain child, the RC captures how parent genes influence a child gene in response to the pairwise parent's mRNA expression, and it can therefore reveal local regulatory changes in response to SARS-CoV-2 infection at an individual sample level. To characterize the individual networks, I calculated RCs for 12 samples within four groups (mock × 3 for SARS-CoV-2-infected (MOI: 0.2), SARS-CoV-2-infected × 3 (MOI: 0.2), mock × 3 for SARS-CoV-2-infected (MOI: 2), SARS-CoV-2-infected × 3 (MOI: 2)) involved in the network generation process, as shown in Figure 2.3. Since I confirmed that the RC profiles exhibit almost the same between the replicates, four representative samples were selected from each group. By representing RCs as the sizes of edge widths, I depicted these four sample-specific individual networks (Figure 2.5), and found that the vicinity of the GAS5-centralized module (module 2) drastically changed at an RC level (Figure 2.6). Interestingly, this module included GAS5, SNHG8, ZFAS1, SNORD52, SNORD58C, SNORA24, and LOC100506548, which encode non-coding RNA (ncRNA) genes. Given that GAS5 appears to function as a hub gene, these results suggest that the genes downstream of GAS5 are regulated by different cellular systems in the mock and SARS-CoV-2 infections at a local system level. In particular, GAS5, ZFAS1, and SNHG8 were found to be dominant for SLC9B1 in SARS-CoV-2-infected samples compared with the mock samples, suggesting that the regulatory system used was significantly different between them (Figure 2.6). GAS5 is a single-stranded lncRNA, and one study demonstrated that the mRNA expression of GAS5 was elevated in response to hepatitis C virus infection and that GAS5 impaired virus replication by the interaction between truncated-GAS5 and HCV NS3 protein in human cells [58]. Combined with this evidence, these results suggest the possibility that this module 2 related to ncRNA may play a novel clear role in SARS-CoV-2 infection.

Conversely, of the four individual networks, the two networks for mocks exhibited no significant change in RC (Figure 2.5 and Figure 2.6). This is consistent with the prerequisite experimental design, as the mock samples are supposed to exhibit the same behavior, which further supports the validity of the developed method. Moreover, the RC-highlighted edges displaying little to no changes showed that their local regulatory system, presented as a set



mock (MOI: 0.2)



SARS-CoV-2 (MOI: 0.2)

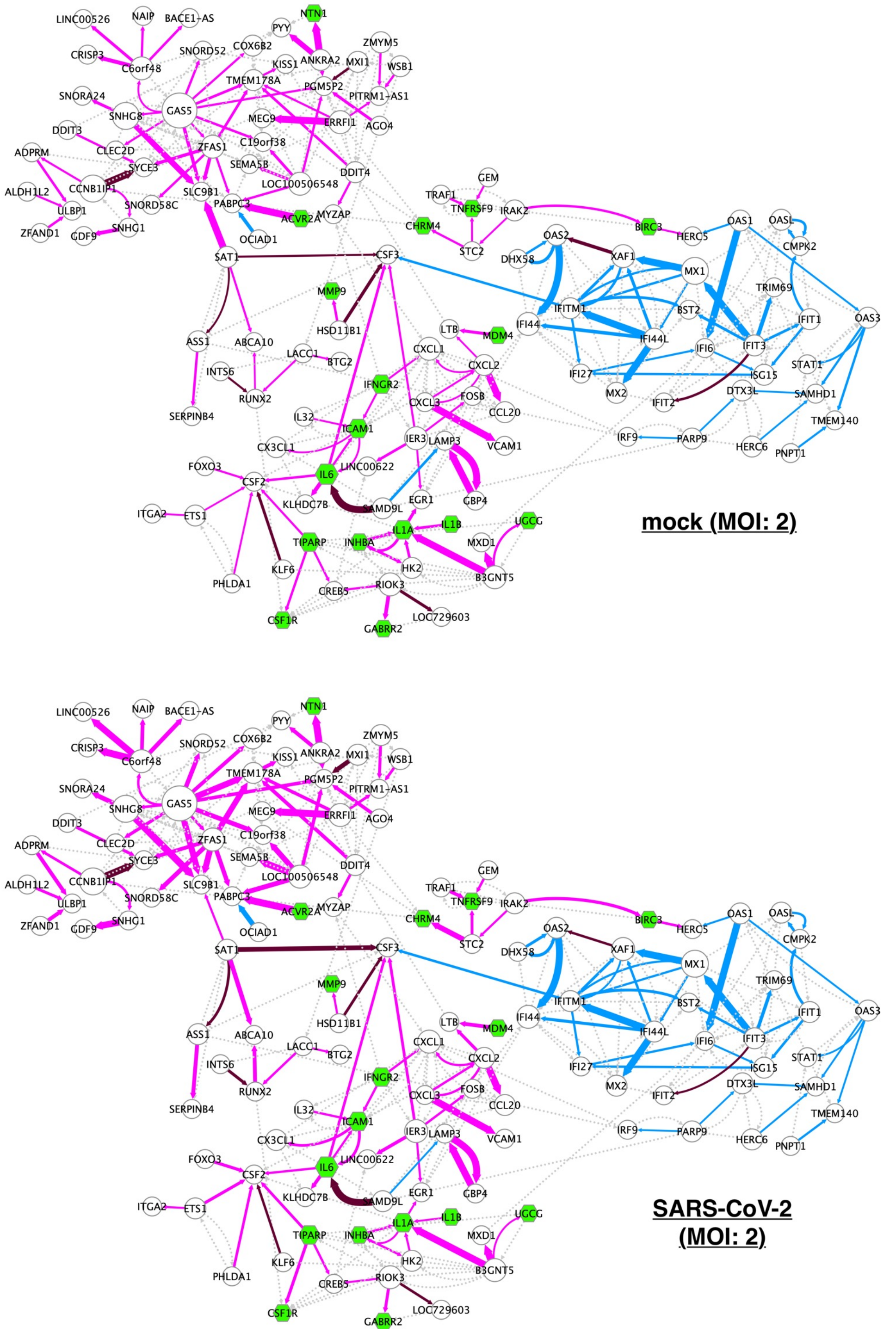


Figure 2.5: Sample-specific individual networks in the SARS-CoV-2-perturbated network. The RC-introduced individual networks are shown for the representative four samples from each group (mock for SARS-CoV-2-infection (low MOI: 0.2), SARS-CoV-2-infection (low MOI: 0.2), mock for SARS-CoV-2-infection (high MOI: 2), SARS-CoV-2-infection (high MOI: 2)). The depicted network is established in Figure 2.3 (the SARS-CoV-2-perturbated network). The region of module 3 is shown in Figure 2.6. RCs are represented as edge sizes to show individual differences.

of pairwise parent-child relationships for one child, did not change between the individual samples. Collectively, these data demonstrate that the RC method can capture the local system differences in network signaling at an individual level.

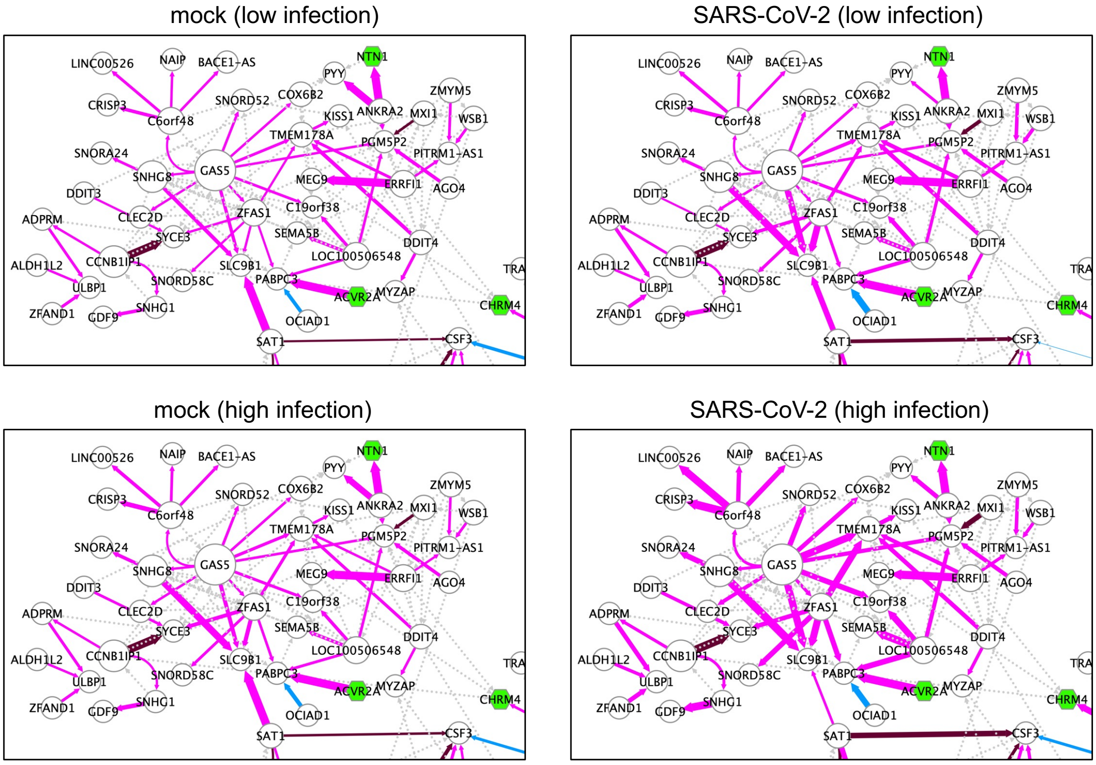


Figure 2.6: Sample-specific individual networks around the GAS5-centralized module. The GAS5-centralized module (module 3) in the SARS-CoV-2-perturbated network (presented in Figure 2.3) is displayed for four representative samples from each group (mock for SARS-CoV-2-infection (low MOI: 0.2), SARS-CoV-2-infected (low MOI: 0.2), mock for SARS-CoV-2-infection (high MOI: 2), and SARS-CoV-2-infected (high MOI: 2)). RCs are represented as edge sizes to show individual differences. The node size represents the extent of outdegree.

3.4 Identification of specific individual networks in patients with COVID-19

Finally, I aimed to establish COVID-19 individual networks with a human biopsy dataset (healthy: two samples; COVID-19-positive: two samples) on the basis of the estimated basal network model. I expect that the *in vivo* biopsy dataset would provide a more clinically relevant perspective compared with the *in vitro* experiments. Usually, network estimation is impossible with such a small number of samples due to the difficulty in acquisition of a robust network structure, yet my approach using the basal network model was capable of generating a context-specific network, even with a few samples of a different dataset (Figure 2.1). By using the B-spline regression model of the Bayesian network acquired by the estimation of the basal network, I first computed the ECv for the preprocessed biopsy dataset, despite the absence of some genes compared with the dataset used for the basal network estimation. To obtain DREs, I calculated ΔECv between healthy (regarded as control) and COVID-19-positive samples according to Eq. (1.4), where S is healthy ($|S| = 2$) and T is COVID-19-positive ($|T| = 2$) (see Methods). The ΔECvs were distributed over a broad range, and 4,242 DREs were observed at a threshold of 1 for ΔECv (Figure 2.7A). To extract more reliable DREs induced by COVID-19, I set a threshold of 2.3, corresponding approximately to $\log_2$FC where FC=5, which resulted in 638 DREs. These DREs were mapped as networks and the largest connected component (167 DREs) was depicted with inclusion of the basal edges, generating the COVID-19-perturbated network, which comprised 127 nodes and 412 edges (Figure 2.7B). This network is supposedly a representation of the distinctive cellular system in patients with COVID-19. The pathway analysis of genes contained in this network showed that they were involved in the immune and inflammatory response (Figure 2.7C), thereby supporting the consistency of the established network with biological observations in COVID-19.

To determine the signatures of the acquired DREs in the COVID-19-perturbated network, I measured the ECv similarity for a set of the 167 DREs across the other experimental samples. This result showed that the ECv profiles in COVID-19 were similar to the sample with HPIV3 rather than SARS-CoV-2 in the *in vitro* experiments (Figure 2.8A), thus suggesting that there is a physiological gap between *in vitro* and *in vivo*. I further explored the extent to which the 167 COVID-19-related DREs overlapped with the Venn diagram illustrated in Figure 2.2B. I observed that a moderate number of DREs were shared by the cell models of SARS-CoV-2 perturbation (Figure 2.8B), then these overlapped edges were mapped onto the COVID-19-perturbated network (Figure 2.8C). Unlike the network observations shown in Figure 2.3, I found that both the ISG-related webs (module 1) and subsequent cytokine signaling (module 3) involved in inflammatory cascades were concurrently present in the COVID-19-perturbated network, indicating that these two modules continued to be mutually activated in COVID-19.

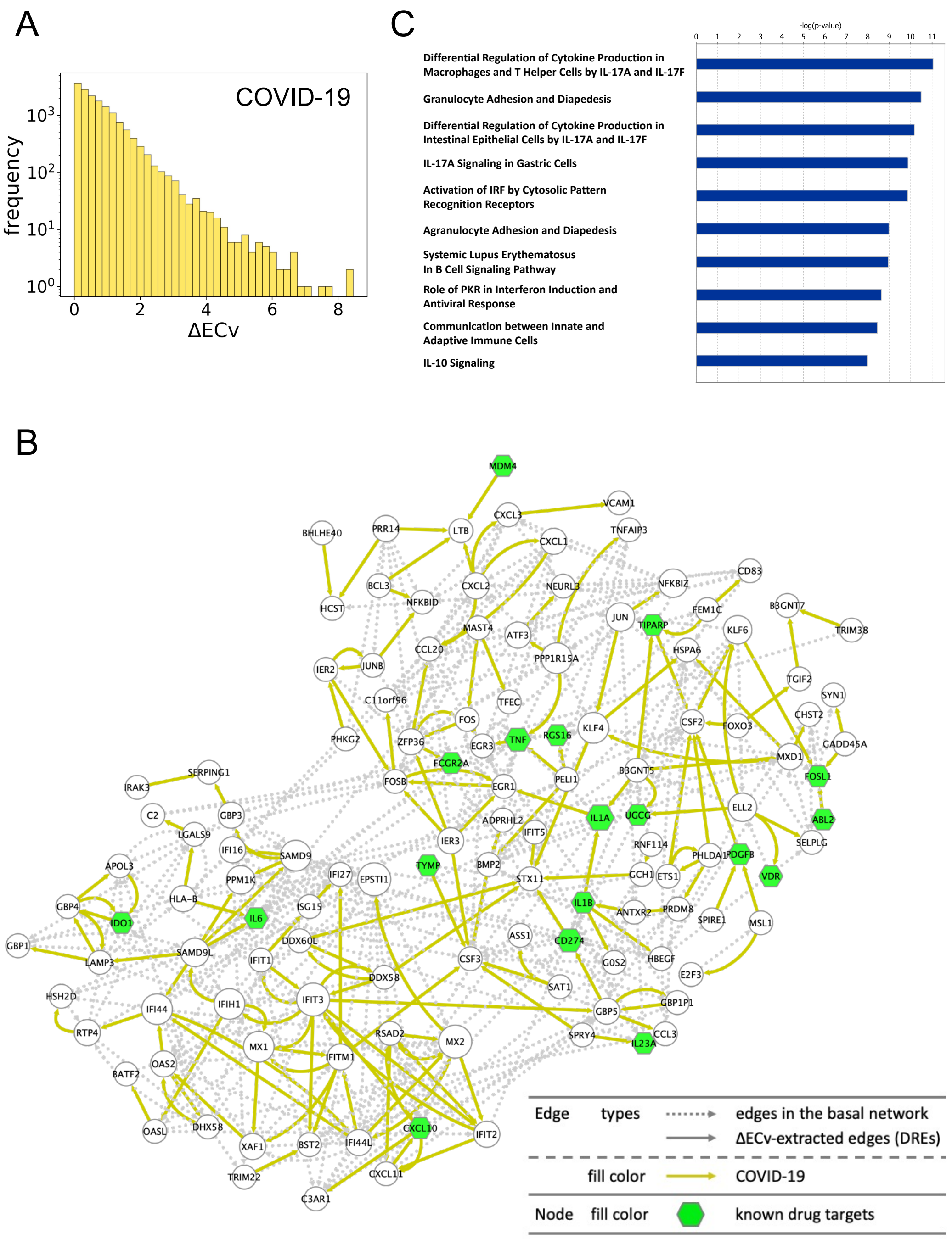


Figure 2.7: The COVID-19-perturbated network analysis. (A) The histograms of ΔECv for the biopsy dataset. The X-axis corresponds to the threshold for the ΔECv. The Y-axis stands for the number of edges with log scale. (B) The COVID-19-perturbated network is shown. The network is composed of 127 nodes and 412 edges (including 245 basal edges). The colored solid edges represent DREs perturbated by COVID-19 (yellow). The dotted edges represent the basal edges (gray). The nodes (green) represent the known drug target genes (Table 2.2). The node size represents the extent of outdegree. (C) The top 10 terms of canonical pathway analysis for the genes in the COVID-19-perturbated network.

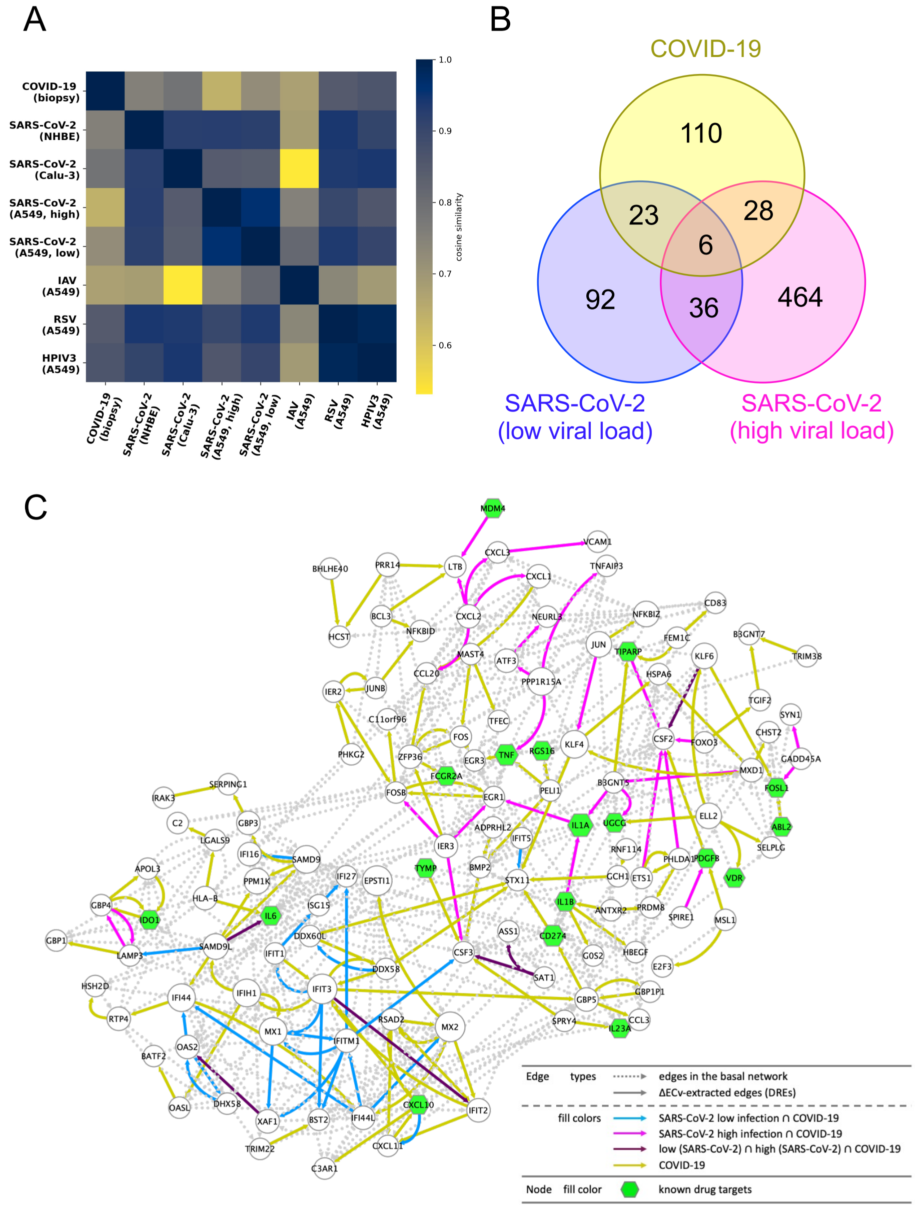
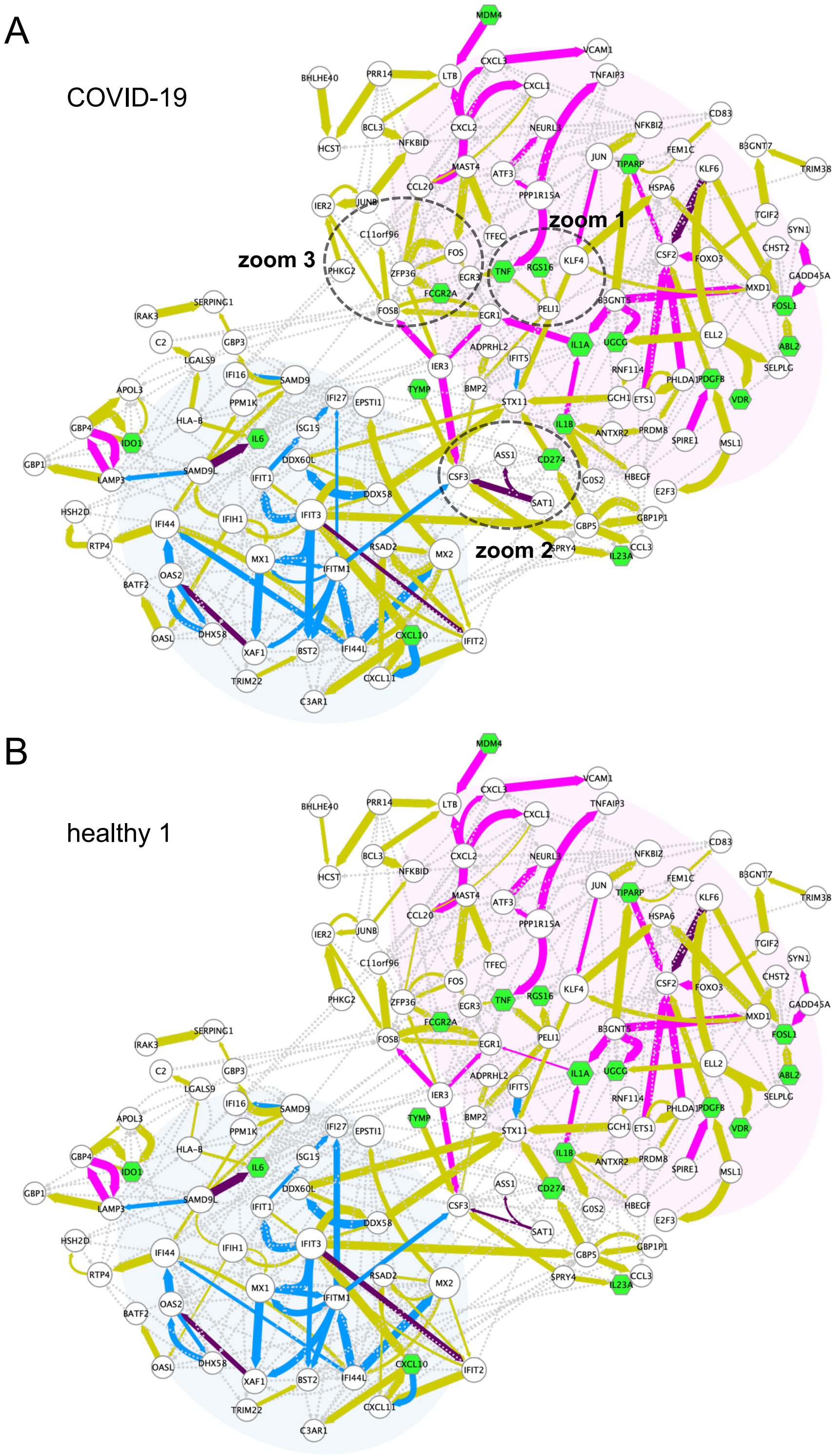


Figure 2.8: Analyses for generating the COVID-19-perturbated network. (A) The similarity heatmap is shown and samples for comparisons are labeled as indicated. Similarity is calculated with cosine distance method for ECvs of the 167 DREs. (B) The Venn diagram shows the numbers of the ΔECv-extracted DREs (ΔECv threshold 2.3) induced by COVID-19 perturbation for the biopsy dataset (yellow) overlapped with the two DREs through the Venn diagram analysis in Figure 2.2B. (C) The COVID-19 patient-specific network in combination with the Venn diagram analysis (Figure 2.8B). The network is composed of 127 nodes and 412 edges (including 245 basal edges). The colored solid edges represent DREs; SARS-CoV-2 (high MOI: 2) ∩ COVID-19-perturbated (magenta), SARS-CoV-2 (low MOI: 0.2) ∩ COVID-19-perturbated (blue), SARS-CoV-2 (high MOI: 2) ∩ SARS-CoV-2 (low MOI: 0.2) ∩ COVID-19-perturbated (purple), COVID-19-perturbated exclusive edges (yellow). The dotted edges represent the basal edges (gray). The nodes (green) represent the known drug target genes (Table 2.2). The node size represents the extent of outdegree.

To uncover the differences in the local regulatory system, I next examined the profiles of the COVID-19-perturbated network at an individual level using the RC method (Figure 2.1). As the two COVID-19 samples were originally derived from a single patient who tested positive for COVID-19, RCs were calculated for three individuals (healthy 1, healthy 2, and COVID-19 patient). The depiction of the RC as the edge sizes eventually led to the establishment of the COVID-19 patient-specific network, which was likely to show how the cellular system changed in the patients with COVID-19 compared with the healthy controls (Figure 2.9A). The panel of three individual networks dramatically exhibited a great magnitude of differences, showing that the cellular regulatory systems were quite distinctive among individuals (Figure 2.9A-C). As this approach captures differences in the system between COVID-19 and healthy individuals as a network, genes that are not normally considered to be up/down-regulated in healthy people will also be included in the network. In comparison with the SARS-CoV-2-perturbated network established by the well-organized *in vitro* experiments using cell lines (Figure 2.5), this broad range of RC fluctuation for each *in vitro* sample likely reflects further differences among individuals. The representative regions where local regulatory systems are different among individuals are illustrated in Figure 2.10. In the zoom 1 region, PELI1 is a parent gene for both TNF and RGS16; these two signals were dominant in the healthy individuals, but not in the patient with COVID-19. In contrast, the zoom 2 and 3 regions showed that local signals were clearly different, not only between the healthy patients and the patient with COVID-19, but also even between two healthy individuals.


A
COVID-19
zoom 1
zoom 2
zoom 3
MDM4
VCAM1
CXCL3
BHLHE40
PRR14
LTB
TNFAIP3
CXCL1
CD83
BCL3
CXCL2
NEURL3
NFKBIZ
HCST
NFKBID
JUN
FEM1C
B3GNT7
TRIM38
MAST4
ATF3
TIPARP
KLF6
HSPA6
CCL20
PPP1R15A
IER2
JUNB
TGIF2
SYN1
C11orf96
TFEC
FOS
CSF2
FOXO3
CHST2
GADD45A
PHKG2
ZFP36
EGR3
TNF
RGS16
KLF4
MXD1
FCGR2A
PELI1
B3GNT5
FOSL1
IRAK3
SERPING1
FOSB
EGR1
ELL2
C2
GBP3
IL1A
UGCG
ABL2
LGALS9
ADPRHL2
IFIT5
RNF114
IER3
SELPLG
IFI16
SAMD9
APOL3
TYMP
BMP2
PHLDA1
PDGFB
PPM1K
IFI27
EPSTI1
STX11
GCH1
ETS1
VDR
HLA-B
IL1B
GBP4
ISG15
ANTXR2
PRDM8
IDO1
IL6
SPIRE1
MSL1
ASS1
CD274
GBP1
SAMD9L
DDX60L
LAMP3
IFIT1
CSF3
G0S2
HBEGF
DDX58
E2F3
SAT1
HSH2D
IFIH1
IFIT3
GBP1P1
IFI44
GBP5
RTP4
RSAD2
SPRY4
CCL3
MX2
IL23A
MX1
IFITM1
OAS2
BATF2
OASL
DHX58
XAF1
BST2
IFI44L
CXCL10
IFIT2
CXCL11
TRIM22
C3AR1
B
healthy 1
MDM4
VCAM1
CXCL3
BHLHE40
PRR14
LTB
TNFAIP3
CXCL1
CD83
BCL3
CXCL2
NEURL3
NFKBIZ
HCST
NFKBID
JUN
FEM1C
B3GNT7
TRIM38
MAST4
ATF3
TIPARP
KLF6
HSPA6
CCL20
PPP1R15A
IER2
JUNB
TGIF2
SYN1
C11orf96
TFEC
FOS
CSF2
FOXO3
CHST2
GADD45A
PHKG2
ZFP36
EGR3
TNF
RGS16
KLF4
MXD1
FCGR2A
PELI1
B3GNT5
FOSL1
IRAK3
SERPING1
FOSB
EGR1
ELL2
C2
GBP3
IL1A
UGCG
ABL2
LGALS9
ADPRHL2
IFIT5
RNF114
IER3
SELPLG
IFI16
SAMD9
APOL3
TYMP
BMP2
PHLDA1
PDGFB
PPM1K
IFI27
EPSTI1
STX11
GCH1
ETS1
VDR
HLA-B
IL1B
GBP4
ISG15
ANTXR2
PRDM8
IDO1
IL6
SPIRE1
MSL1
ASS1
CD274
GBP1
SAMD9L
DDX60L
LAMP3
IFIT1
CSF3
G0S2
HBEGF
DDX58
E2F3
SAT1
HSH2D
IFIH1
IFIT3
GBP1P1
IFI44
GBP5
RTP4
RSAD2
SPRY4
CCL3
MX2
IL23A
MX1
IFITM1
OAS2
BATF2
OASL
DHX58
XAF1
BST2
IFI44L
CXCL10
IFIT2
CXCL11
TRIM22
C3AR1

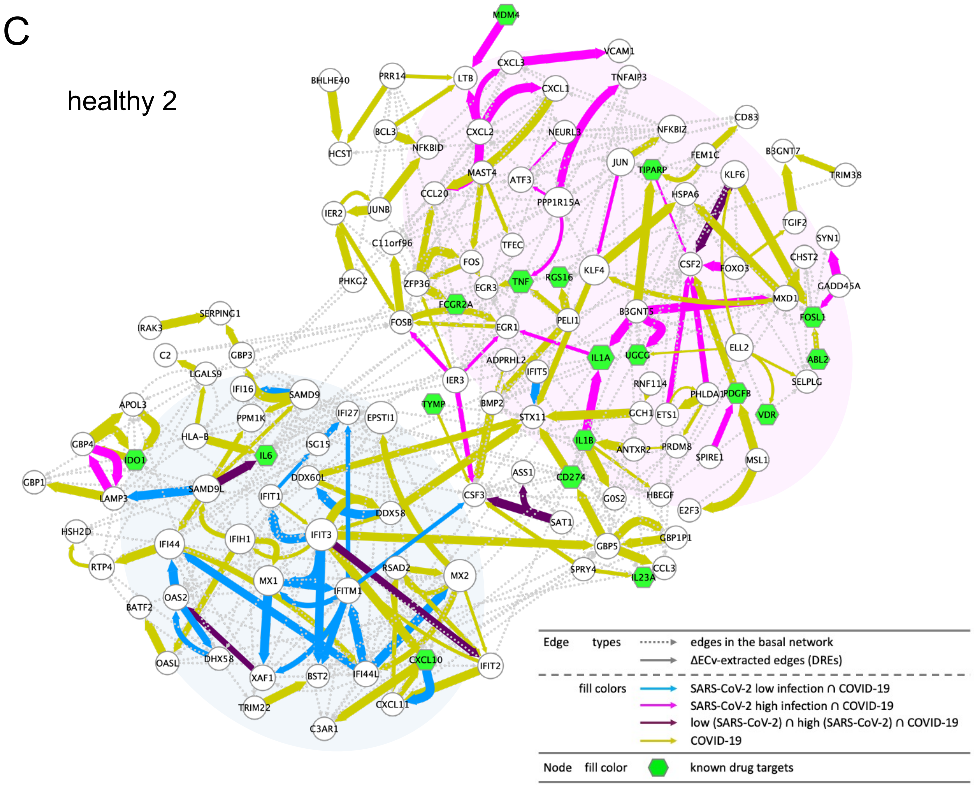


Figure 2.9: Establishment of the individual-specific networks in the COVID-19-perturbated network. The RC-introduced individual networks are shown for three individuals (patient with COVID-19, healthy 1, and healthy 2). The depicted network is established in Figure 2.7 and Figure 2.8 (the COVID-19-perturbated network). (A) The COVID-19 patient-specific network. (B, C) The healthy-specific networks. The network comprises 127 nodes and 412 edges (including 245 basal edges). The colored solid edges represent DREs; SARS-CoV-2 (high MOI: 2) ∩ COVID-19-perturbated (magenta), SARS-CoV-2 (low MOI: 0.2) ∩ COVID-19-perturbated (blue), SARS-CoV-2 (high MOI: 2) ∩ SARS-CoV-2 (low MOI: 0.2) ∩ COVID-19-perturbated (purple), COVID-19-perturbated exclusive edges (yellow). The dotted edges represent the basal edges (gray). The nodes (green) represent the known drug target genes (Table 2.2). The node size stands for the extent of outdegree. RCs are represented as edge sizes to show individual differences.

4 Discussion

Here, I have presented the host cellular gene networks perturbated by SARS-CoV-2 both *in vitro* and *in vivo* by using the proposed framework for gene network analysis. As the networks I established to be associated with SARS-CoV-2 were generated through RNA-seq data, these networks explained how genes were systematically regulated at the transcriptome level. Although this approach depends on the initial network estimation with an experimental dataset and may therefore risk the inclusion of false relationships or the exclusion of true relationships, I have succeeded in capturing the biologically explainable immune response systems in human cells induced by SARS-CoV-2 at the level of signaling

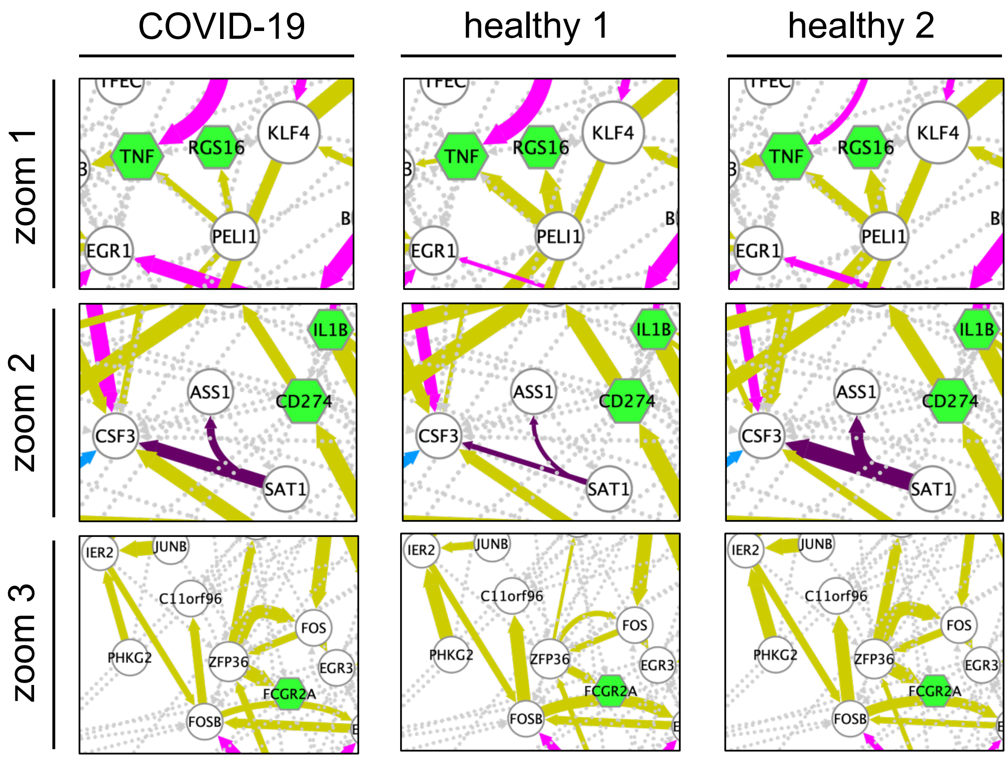


Figure 2.10: Differences in local regulatory systems among the three individuals. Zoomed regions indicated in Figure 2.9A for three individuals (the patient with COVID-19, healthy 1, and healthy 2). RCs are represented as edge sizes to show individual differences.

networks.

Sensing of viruses causes an immune defense system in host cells, which induces acute IFN signaling activation followed by the expression of IFNs. These IFNs amplify the JAK/STAT signaling to promote the expression of various ISGs and accelerate subsequent cytokine signaling [57]. As illustrated in Figure 2.3, the mutually interacting module of ISGs (module 1) followed by the IFN and JAK/STAT signaling was shown to be an early response to SARS-CoV-2 infection. During the process of cells exposed to high SARS-CoV-2 viral loads, the signaling appears to move to the next stage, represented by inflammatory signaling, including the involvement of various cytokines (Figure 2.3). The recently reported drug, dexamethasone, could be effective for patients with severe COVID-19 via suppression of these orchestrated inflammatory signaling cascades [59]. In the network (Figure 2.3), IL6 was located as a hub gene to regulate downstream cascades, including chemokines and colony-stimulating factors, which have been reported to be increased in patients with COVID-19 [53]. The web of chemokines, such as CXCL1, CXCL2, and CXCL3, may represent how the SARS-CoV-2-infected cells present a signal to induce

Table 2.2: The gene list for known drug targets. The representative drug for each gene is listed.

Gene	Drug	Gene	Drug
ABL2	dasatinib	IDO1	epacadostat
ACVR2A	bimagrumab	IFNGR2	interferon gamma-1b
ACVRL1	panulisib	IL1A	MABp1
ADH1C	fomepizole	IL1B	canakinumab
ATP1A2	digoxin	IL23A	guselkumab
BIRC3	birinapant	IL3RA	DT388IL3
CA9	acetazolamide	IL6	tocilizumab
CACNA1A	bepridil	INHBA	STM 343
CD274	atezolizumab	MDM4	ALRN-6924
CFB	IONIS-FB-LRx	MMP9	marimastat
CHRM4	acetylcholine	NTN1	NP137
CSF1R	bosutinib	PDGFB	pegpleranib
CXCL10	MDX-1100	RGS16	AGS-16M18
DDC	levodopa	RORC	AZD0284
ENTPD1	TTX-030	RPS6KA2	PMD-026
FCGR2A	IgG	TIPARP	RBN-2397
FGFR4	erdafitinib	TLR9	agatolimod
FNDC1	AGS-8M4	TNF	adalimumab
FOSL1	MORAb-202	TNFRSF9	ADG106
GABRR2	diazepam	TYMP	tipiracil
ICAM1	tyroserleutide	UGCG	eliglustat
		VDR	alfacalcidol

leukocyte chemotaxis and infiltration. The localization of ICAM1 in the vicinity of IL6 and chemokines is supportive of this, as ICAM1 is known to be a scaffold for the accumulation of leukocytes at inflammatory sites and its expression is regulated by cytokines, including IL6 [60, 61]. This tendency was also observed in the network comparison analyses across the four respiratory viruses, including SARS-CoV-2 (Figure 2.4C). These data showed that IL6 was not exclusive to SARS-CoV-2, but a universal factor in response to respiratory viral infection, with the exception of the influenza A virus. Given that several studies have reported that tocilizumab, an inhibitor of the IL6 receptor, is a potential drug that can suppress the cytokine storm observed in many critical patients with COVID-19 [62, 63], the accumulated evidence strongly suggests that IL6 is a central regulator of the inflammatory cascade, even from a network perspective. Additionally, the networks showed that CSF2 was regulated by various factors, including IL6, which strengthened previous reports suggesting that CSF2 might be a promising therapeutic target in combination with IL6 [64, 65]. Moreover, other immune defense signaling pathways such as complement and macrophage were identified (Figure 2.2C). In contrast, coagulation cascade reported abnormalities in patients with COVID-19 [66] was not identified, probably because this aberration of coagulation may have occurred at the physiological system level rather than at the cellular level. This developed method only captures the system at the cellular level, and it is not yet possible to see the response of the entire biological system (e.g. between organs). This would be a limitation of the current approach.

Several recent studies have shown that ACE2 plays a key role in the process of SARS-CoV-2 infection. SARS-CoV-2 enters into host cells via ACE2 [67], and ACE2 was found to be an ISG in human airway epithelial cells [68]. Considering that the SARS-CoV-2-perturbated network includes several ISGs (Figure 2.3), it can be reasoned that some clues regarding ACE2 may be present in this network. In this context, I found that ACE2 was closely located to this network and was downstream of TNFRSF9, ATF3, and ARRDC3 via ACHE (Figure 2.11); these are potential candidates for further investigation of the relationship between ACE2 and ISGs. Among them, ATF3 would be the most promising as it was found to be a direct transcriptional target for ACE2 [69]. Thus, the established networks provide promising information to elucidate SARS-CoV-2 profiles from a broad biological perspective.

The second noteworthy outcome in this study was that I succeeded in the characterization of sample-specific individual networks by introducing the new edge-quantitative technique of RC. In particular, although it is impossible to estimate a network with a small number of samples, such as the four biopsy samples used in this case, the basal network model that was already obtained through the analysis of the *in vitro* dataset with both RC and ECv methods led to establishment of the COVID-19 patient-specific individual network. This process represents the method of extrapolation between *in vitro* and *in vivo* experiments. Each sample exhibits a unique regulatory profile, especially in actual individuals (such as those obtained from biopsy) rather than well-controlled *in vitro* samples (Figure 2.5 and Figure 2.9). These results probably reflect a more realistic clinical situation and increase the importance of the most effective utilization of a biopsy dataset. In the current outbreak of COVID-19, we need to look into both biological and clinical aspects to explore COVID-19 therapy. The individual networks regarding COVID-19 show the extent to which individuals possess their own network, which ultimately links to the necessity of a personalized treatment. Therefore, my efforts are a potential contribution to the emerging field of personalized medicine. The biopsy dataset used was not sufficient to allow interpretation of the comprehensive information through individual networks in patients with COVID-19, as it contained fewer COVID-19 samples. In addition, since the lung samples were obtained postmortem patients with COVID-19, I could not determine at what time point in disease progression the identified regulatory system is in this study

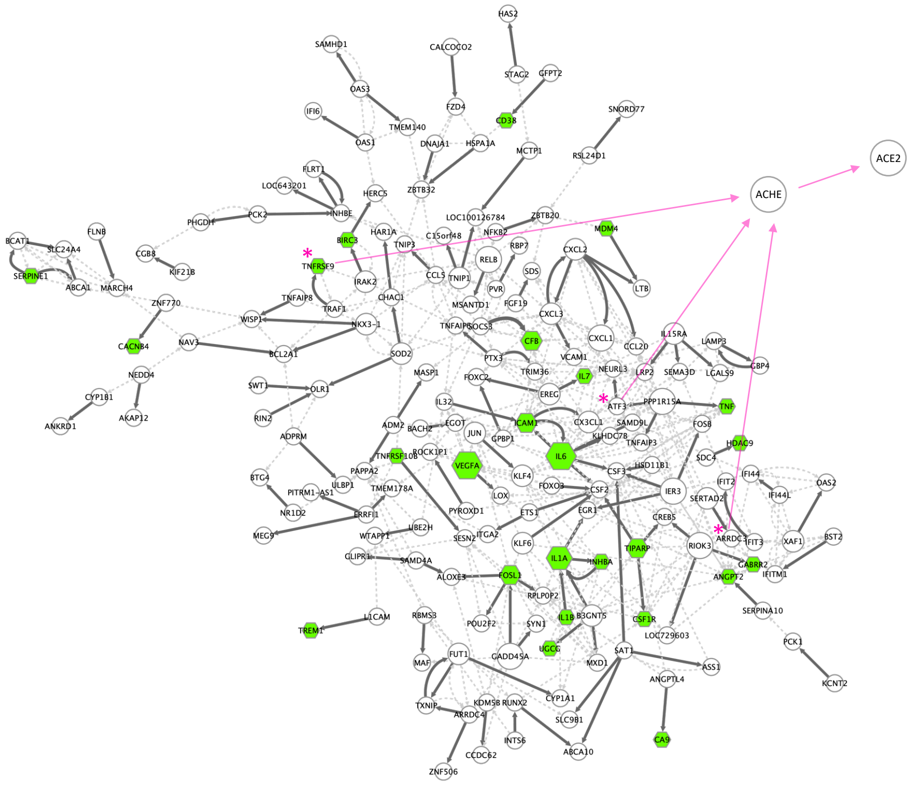


Figure 2.11: The schematic network illustration of ACE2 involvement in immune defense system of host cells. The network (188 nodes and 420 edges) was generated by the ΔECv-extracted edges (black) shared with A549, NHBE and Calu-3 cells in response to SARS-CoV-2 infection. The nodes (green) represent the known drug target genes. The edges (magenta), and nodes (marked magenta) represent potential regulatory signalings to ACE2 via ACHE. The dotted edges represent the basal edges (gray).

(Figure 2.9 and Figure 2.10). More clinical samples with time series and disease severity information from patients with COVID-19 can lead to the determination of key regulatory systems at a clinical level. As data regarding SARS-CoV-2 has been currently accumulated by the efforts of researchers, I hope that this panel of network analyses will be of help to the SARS-CoV-2 research field and to establishment further treatments for COVID-19.

Chapter 3

Large-scale gene network analysis for identification of drug-induced liver injury signature in human hepatocyte

1 Introduction

Liver has the ability to metabolize and detoxify many xenobiotics, but liver injury occurs when the threshold of its processing capacity is exceeded or when toxic intermediate metabolites are produced [70]. In clinical situations, various drugs have been reported to induce liver injury, which is one of the side effects of drugs called drug-induced liver injury (DILI) [71]. While some drugs, such as acetaminophen, cause DILI in a dose-dependent manner, many of DILI-concerned drugs infrequently cause severe liver injury in a small percentage of patients, which sometimes lead to the need for liver transplantation or even death [72]. Since this side effect is very difficult to detect in clinical trials due to the small number of incidences, it is one of the major causes of withdrawal of drugs from the market and is also a critical risk for the pharmaceutical industry [73, 74]. Thus, DILI is an important issue in both the medical and drug discovery fields. However, a solution to avoid this risk has not yet been fully established.

A major reason for this is that the mechanisms of DILI are largely unknown, and hence there are no specific predictive markers for DILI [75, 76]. To address this problem, large-scale toxicogenomics data [77, 78, 79] and adverse reaction data [80, 81] has been accumulated in recent years, and many studies were carried out from various perspectives including chemical structure properties [82], genetics [83], and gene expression [84, 85, 86]. Among them, the network-based approach from the perspective of systems biology has

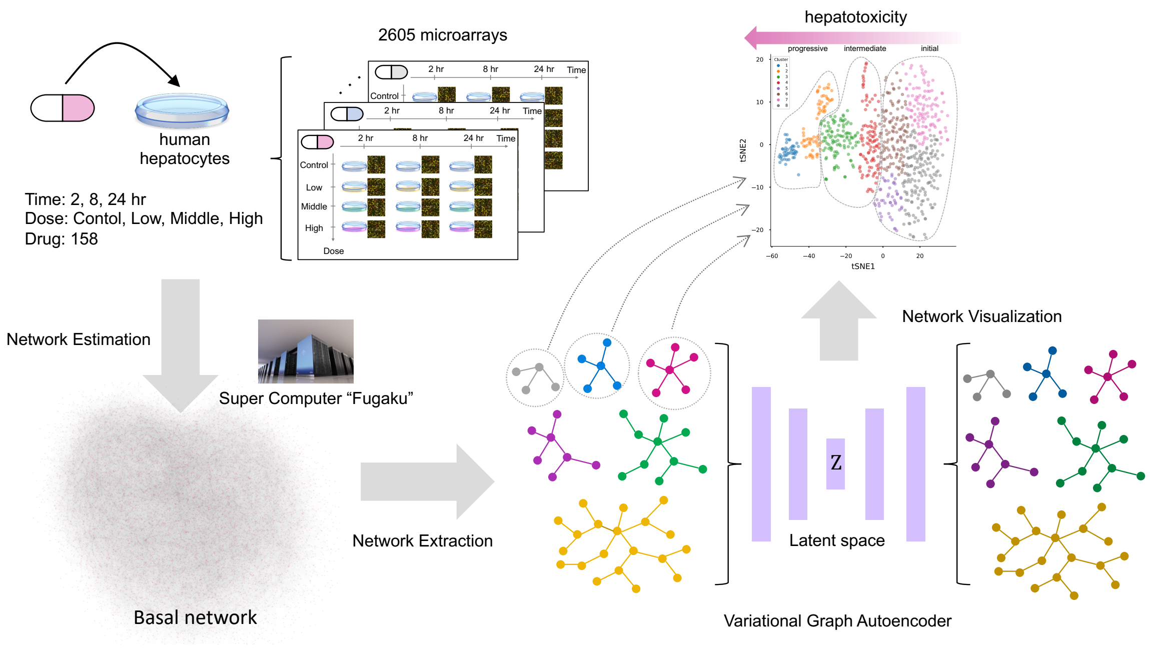


Figure 3.1: Overview of the study. The basal network (bottom left, gray) consists of 10,554 nodes and 113,054 edges. The extracted subnetworks were used as input data for the VGAE (bottom right). The plots in latent variable visualization correspond to the subnetworks (top right).

recently attracted great interest as a promising method for exploring molecular mechanisms involved in DILI [87]. The mainstream approach is to integrate gene expression profiles and network analysis, which can be divided into two types: one is to use the weighted correlation network analysis (WGCNA) [88] to construct networks based on the correlation of expression between genes after identifying differentially expressed genes (DEGs) [89, 90, 91], and the other is to combine gene expression data with existing signaling pathway information [92, 93]. However, these methods do not consider the whole cellular system. This is because these methods require a predetermined selection of a certain number of genes or rely on existing pathway knowledge such as Reactome and KEGG [94, 95], which might miss relations that are not included in the prior knowledge but are important for cellular system regulations. Furthermore, for example, different types of drugs are considered to trigger different intracellular networks, but these existing methods have not been able to lead these diverse states of the network.

In order to overcome these drawbacks of existing studies, this study explored condition-specific networks involved in hepatotoxicity by establishing a unique gene network analysis approach using large-scale gene expression data in a network information-driven manner. This approach could be expected to make a contribution to the elucidation of condition-specific mechanisms through the gene networks. The overall picture of the study is illustrated in Figure 3.1. Firstly, a Bayesian network estimation method was used on gene expression data from 2605 samples of human hepatocytes exposed to 158 drugs at multiple time points and doses to reflect the whole cellular system as a gene network, which systematically responds to more than 10,000 available genes. From this huge network, 762 condition-specific subnetworks that are systematically perturbated by the exposure of DILI-concerned drugs were extracted according to the combinations of drug, exposure time, and dose using the developed method in Chapter 1 [52]. Then, in order to comprehensively characterize these subnetworks, a novel network classification method was developed using graph neural networks (GNNs), which is a type of deep learning technology for graph-structured data [96]. Eventually, I identified the subnetworks involved in hepatotoxicity, and further showed the connection between the subnetworks and human DILI. These results reveal for the first time that networks play a key role as biomarkers of DILI.

2 Methods

2.1 The gene network estimation

The network estimation was performed using the NNSR algorithm [17] following previously established protocols described in Chapter 1 and 2 [52, 97]. The NNSR algorithm allows us to estimate a network with all the available number of genes over 10000, which generates the universal gene regulatory network (*the basal network*) underlying cellular systems. The difference from these previous studies is that this one had a much larger sample size. The evaluation of the network estimation protocol in the latest supercomputer system Fugaku was performed using the multiple simulation datasets following the same protocol as in the previous study [17]. The four datasets used in the simulation experiments consist of 10540 nodes, with 500, 1000, 2000, and 3000 samples, respectively. As these simulation datasets were generated from the original network, to evaluate the network estimation performance, the networks estimated from them were subjected to structural comparisons with the true network. In the precision recall curve, recall is defined as TP/(TP + FN), and precision as TP/(TP + FP), where TP is the number of correctly estimated edges (true positive), FN is the number of edges that were not estimated but included in the true network (false negative), and FP is the number of incorrectly estimated edges (false positive). The network estimation was performed three times independently for each dataset. The average of recall and precision over each run was used in the precision recall curve since the results of the three runs were almost identical. The area under the precision-recall curve was calculated with the network threshold between 0.2 and 0.7. The iteration number T for the network estimation was fixed at 100,000 for every dataset.

2.2 Extraction of sample-specific subnetworks from a basal network

The subnetwork extraction was performed using the ΔECv method described in Chapter 1 [52]. Briefly, edge contribution value (ECv), defined in Eq. (1.3) in Chapter 1, is a quantitative measure of how strongly the estimated edge was used in the network for all estimated edges in each sample. This method enables us to extract the changes in the cellular system between arbitrary samples by calculating the differences in their ECvs as a subnetwork from the basal network. In this study, I defined ΔECv as follows,

$$\Delta\mathrm{ECv}_d^{c,t}(u \to v) = \left|\mathrm{ECv}_d^{c,t}(u \to v) - \mathrm{ECv}_{\mathrm{control}}^{c,t}(u \to v)\right|,$$

where $\mathrm{ECv}_d^{c,t}(u \to v)$ is the ECv of the estimated edge from u to v with respect to drug $c \in \{\mathit{158\ drugs}\}$, time $t \in \{\mathrm{2hr}, \mathrm{8hr}, \mathrm{24hr}\}$, and dose $d \in \{\mathrm{Control}, \mathrm{Low}, \mathrm{Middle}, \mathrm{High}\}$. Here, u and v are genes in the network, and the ECv was averaged when the replicate for the same condition is available. The ΔECv threshold was set to extract differentially regulated edges (DREs) from the basal network [97]. Here, a single subnetwork was defined as a set of DREs for a certain combination of drug, time and dose following the ΔECv definition. In this study, $\Delta\mathrm{ECv} \geq 0.5$ was employed to capture system changes even at weak perturbation levels, such as low dose or short exposure time of 2 hr.

2.3 The VGAE architecture

I present a unique approach to predict the features of a dimension-free graph using Variational Autoencoder (VAE) framework [98]. I design an autoencoder that encodes a graph into a latent variable $z \in \mathbb{R}^m$ and decodes the graph from z. Therefore, the goal of this autoencoder is to optimize the parameters that can reconstruct the input graph via latent variables by end-to-end learning (Figure 3.2). An input graph is denoted as $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$ where $\boldsymbol{V}$ is a set of nodes and $\boldsymbol{E}$ is a set of edges. A graph $\boldsymbol{G}$ having n nodes are represented in a matrix form as $n \times n$ adjacency matrix $\boldsymbol{A}$ and the element A_{ij} of $\boldsymbol{A}$ at i-th row and j-th column is defined as 1 if edge exists between i-th node and j-th node, otherwise 0. A node feature matrix is denoted as $\boldsymbol{X} \in \mathbb{R}^{n \times f}$ where f is a dimension of features. The input node features are given arbitrarily.

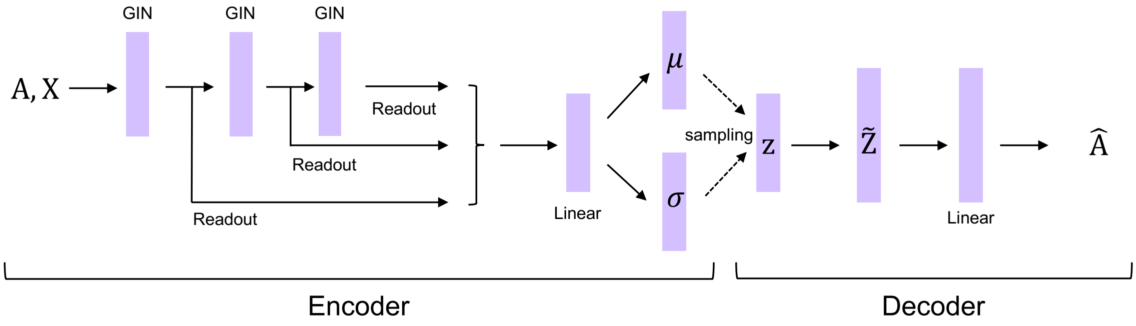


Figure 3.2: Illustration of the VGAE architecture.

I first consider an encoder of this VGAE using Graph Isomorphism Network (GIN). The GIN is one of the architectures in GNNs to learn graph-structured data using adjacency matrix $\boldsymbol{A}$ and node feature matrix $\boldsymbol{X}$ [99]. Here, the encoder is defined as,

$$\begin{aligned} &q(z|\boldsymbol{X},\boldsymbol{A}) := \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2)), \\ &\boldsymbol{\mu} = \mathrm{GIN}_{\mu}^{(K)}(\boldsymbol{X},\boldsymbol{A}), \quad \log\boldsymbol{\sigma} = \mathrm{GIN}_{\sigma}^{(K)}(\boldsymbol{X},\boldsymbol{A}). \end{aligned} \tag{3.1}$$

The function $\mathrm{GIN}_{*}^{(K)}(\cdot)$ consists of two layers, the K-th layer GIN and a linear transformation. Let $\boldsymbol{N}_i$ denote a set of adjacent nodes of the i-th node, and $\boldsymbol{x}_i$ is a i-th row vector of $\boldsymbol{X}$. The K-th layer GIN is defined using a multi-layer perceptron (MLP) as follow,

$$\begin{aligned} \boldsymbol{x}_i^{(k+1)} &= \mathrm{MLP}^{(k+1)}\left(\boldsymbol{x}_i^{(k)} + \sum_{u\in N_i} \boldsymbol{x}_u^{(k)}\right), k = 0,\ldots,K-1, \\ \boldsymbol{x}_i^{(0)} &= \boldsymbol{x}_i \end{aligned}$$

where $\boldsymbol{x}_i^{(k)}$ is a i-th node feature vector in $\boldsymbol{X}^{(k)}$ and K is the maximum number of GIN layers. This node feature matrix $\boldsymbol{X}$ is converted into a graph feature vector using a readout function defined as $\mathrm{Readout}(\boldsymbol{X}) := \sum_i^n \boldsymbol{x}_i$ by a row-wise aggregation of all node vector for each GIN layer. The graph feature vector $\boldsymbol{h}$ is computed with a concatenation across all GIN layers as followed,

$$\boldsymbol{h} = [\mathrm{Readout}(\boldsymbol{X}^{(1)}), \mathrm{Readout}(\boldsymbol{X}^{(2)}), \ldots, \mathrm{Readout}(\boldsymbol{X}^{(K)})].$$

The $\boldsymbol{h}$ is applied to a linear transformation defined as $(\boldsymbol{W}_*\boldsymbol{h}+\boldsymbol{b}_*)$ to change feature dimension, where the $\boldsymbol{W}_*$ is a weight matrix and the $\boldsymbol{b}$ is a bias vector. Therefore, the map to estimate averages and variances of the encoder (3.1) is given by $\mathrm{GIN}_*(\boldsymbol{X},\boldsymbol{A}) = \boldsymbol{W}_*\boldsymbol{h} + \boldsymbol{b}_*$.

I next consider a decoder as follow,

$$p(\boldsymbol{A}|z) := \prod_{i,j=1}^{n} \text{Bernoulli}(A_{ij}|\pi_{ij}),$$

where $\pi_{ij} = \text{sigmoid}(\tilde{z}_i^{\mathrm{T}}\tilde{z}_j)$. Here, $\tilde{z}_i$ are the i-th row vector in a matrix $\tilde{\boldsymbol{Z}}$, which is the $n \times m$ matrix generated with row-wise replication of the vector z to reconstruct the node feature matrix from the graph feature vector.

For learning, the variational lower bound L is optimized by computing the reconstruction loss of the adjacency matrix and Kullback-Leibler (KL) divergence as follow:

$$L := \mathbb{E}_{q(z|\boldsymbol{X},\boldsymbol{A})}[\log p(\boldsymbol{A}|z)] - \beta \text{KL}[q(z|\boldsymbol{X},\boldsymbol{A})||p(z)].$$

The β is a hyper parameter to control learning performance following the previous study [100]. The function of $\text{KL}[q(\cdot)||p(\cdot)]$ represents the KL divergence, which measures the distribution difference of $q(\cdot)$ from $p(\cdot)$. This KL divergence can be calculated when $p(z)$ is supposed to the follow Gaussian distribution with $\mathcal{N}(z|\boldsymbol{0},\boldsymbol{I})$ as follow,

$$\text{KL}[q(\cdot)||p(\cdot)] = -\frac{1}{2}\sum_{r=1}^{m}\left(1 + 2\log\sigma_r - \mu_r^2 - \sigma_r^2\right),$$

where m is the dimension of z [98]. For the reconstruction loss, since the distribution of each element of the adjacency matrix is supposed to follow Bernoulli distribution, this can be calculated by the binary cross entropy. Considering the case where a graph is sparse, the number of edges present in the graph (*positive*) is much less than the number of edges in the complete graph. Here, edges of a complete graph that is not connected in the sparse graph is called *negative*. Due to this imbalance in the number of *positive* and *negative* edges, it is not realistic to consider all these edges for the reconstruction loss calculation in terms of computer memory resources. To speed up the learning process, the same number of *negative* edges were randomly sampled as the number of *positive* edges from the set of negative elements when computing the reconstruction loss [101]. Therefore, the reconstruction loss for both $A_{ij} = 1$ and $A_{ij} = 0$ is calculated as follow,

$$\mathbb{E}_{q(z|\boldsymbol{X},\boldsymbol{A})}\big[\log p(\boldsymbol{A}|z)\big] = \sum_{(i,j)\in\boldsymbol{E}_+\cup\boldsymbol{E}_-}\left(A_{ij}\log\hat{A}_{ij} + (1-A_{ij})\log(1-\hat{A}_{ij})\right),$$

where $\boldsymbol{E}_-$ is a set of sampled *negative* edges and $\boldsymbol{E}_+$ is a set of *positive* edges $\hat{\boldsymbol{A}}$ is a reconstructed adjacency matrix. For training, the z is sampled using the reparameterization trick defined as $z = \boldsymbol{\mu} + \boldsymbol{\sigma} \cdot \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is a m dimensional standard normal distribution [102].

Thus, the VGAE constructed by the encoder and decoder estimates the latent variables that lie behind the graph structure data. The basic framework of this VGAE is based on the previous study [102]. However, unlike the original model, the outstanding characteristics of this VGAE is to be designed to acquire the latent features of a graph at the graph level rather than the node level. In the experiments, three GIN layers were used. The MLP in one GIN layer consisted of two linear layers with a rectified linear unit (ReLU) as the activation function. The dimension size of vector $\boldsymbol{x}$ was 32 for the first and second GIN layers, and was 16 for the third GIN layer. The batch normalization was used at the first MLP layer for each GIN layer [103]. During the reconstruction of the adjacency matrix, an additional linear layer was applied before the sigmoid function in order to change the dimension size m to be the same as the input dimension size f. The adam optimizer was applied with the learning rate at 0.001 [104]. For the both datasets, batch size was 64; latent dimension size m was 16; β was 0.001. The epoch size was 10 and 5 for the model network dataset and the toxicogenomics network dataset, respectively. For the model network dataset, the node degree encoded as a one hot vector was used for the input node features, and the input dimension was 59. For the toxicogenomics network dataset, gene expression levels normalized as minimum 0 and maximum 1 across all the samples were used for the input node features, and the input dimension was 1304. The 5-fold cross validation was performed and one of the models was used for the following analyses. The VGAE is implemented using the python package PyTorch Geometric [105].

2.4 Toxicogenomics microarray dataset

The toxicogenomics microarray data were downloaded from the Open TG-GATEs (Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System) website (`https://toxico.nibiohn.go.jp/`) ("Toxicogenomics Project and Toxicogenomics Informatics Project under CC Attribution-Share Alike 2.1 Japan") [78]. In human hepatocyte experiments, there are four levels of dose (Control, Low, Middle, and High) and three time points (2 hr, 8 hr, and 24 hr) for one drug, with a maximum of two replicates for each. There are 158 drugs in total. Thus, there are at most 24 microarray samples per drug. Here, not all drugs have all these 24 combinations, which means that missing microarray samples exist for some drugs. All the available 2605 microarray samples were processed by MAS5 [106] and concatenated for all the samples, which generated a gene-by-sample expression data matrix. The probe name was converted into the gene name and the expression values of probes with the same gene name were averaged. The genes including the missing expression value more than 25% of the total samples were removed, and the missing expression values for the remaining genes were complemented with 0, eventually resulting in the expression dataset with 10554 genes and 2605 samples. This dataset was transformed with log (base is 2), which was used as the input dataset for the basal network estimation.

2.5 Model network dataset

The Barabási-Albert (BA) model [107], the Watt-Strogatz (WS) model [108] and Erdös-Rényi (ER) model [109] were used as the model networks for the VGAE evaluation. All the networks for the BA model contain 100 nodes and 475 edges. All the networks for the WS model contain 100 nodes and 500 edges. All the networks for the ER model contain 100 nodes and approximately 500 edges. Three model networks (BA, WS and ER) were generated using the python package networkx [110].

2.6 Graph visualization and classification analysis

The latent variables acquired by the VGAE were transformed into two dimensions to visualize by using t-SNE [111]. The hierarchical clustering was performed with Euclidean distance and Ward method.

2.7 Cytotoxicity analysis

Cytotoxicity data was also downloaded from the Open TG-GATEs website [78]. The cytotoxicity was measured as the relative DNA content at a certain dose (Low, Middle and High) sample compared to the control sample corresponding to the exposure time (2 hr, 8 hr and 24 hr) for each drug. There were samples with missing values. The relative cell viability was calculated as follow,

$$\text{Relative viability}(\%) = \text{DNA}_d^c(\%) - 100(\%),$$

where dose $d \in \{\text{Low}, \text{Middle}, \text{High}\}$ and drug c.

2.8 Network visualization

The network visualization was performed using Cytoscape (version 3.8.2) [32].

2.9 Pathway analysis

Pathway analyses were performed using R packages ReactomePA and clusterProfiler [112, 113] based on the Reactome knowledge database [94]. The significant top 10 pathway terms were listed for each cluster. The set of gene nodes that comprise the subnetworks included in the cluster was used for input.

2.10 Computational environments

All the computations for the network estimation and the ECv calculations in this study were performed by the Fugaku supercomputer system at RIKEN Center for Computational Science, where the computation nodes were equipped with Fujitsu A64FX CPUs and 32GiB memory per node.

3 Results

3.1 Establishment of the basal gene network using the large-scale toxicogenomics data

In order to establish the basal network, I performed the Bayesian network estimation using a large-scale toxicogenomics microarray dataset including 2605 samples, which was measured by exposing hepatocytes to many drugs at different doses and times *in vitro*. The Bayesian network estimation method is supposed to generate a more robust and expressive model by using a larger number of samples [17]. However, existing studies have never evaluated a performance of the network estimation with more than approximately 1000 samples [114]. Prior to the network estimation using the toxicogenomics dataset, to ask whether the network estimation is feasible on a large-scale sample size, I first performed simulation experiments using test datasets following the previous study (see Methods) [17].

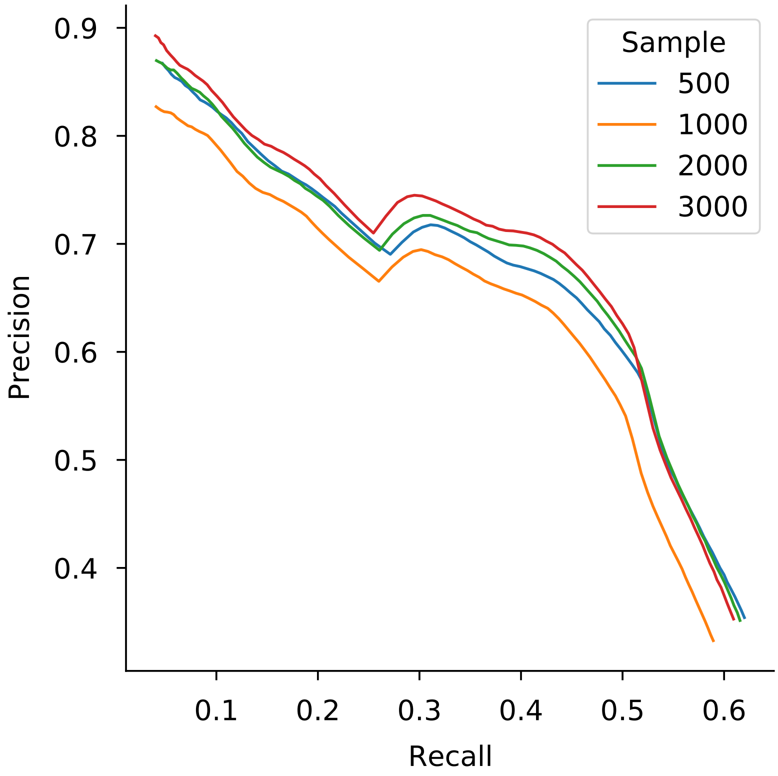
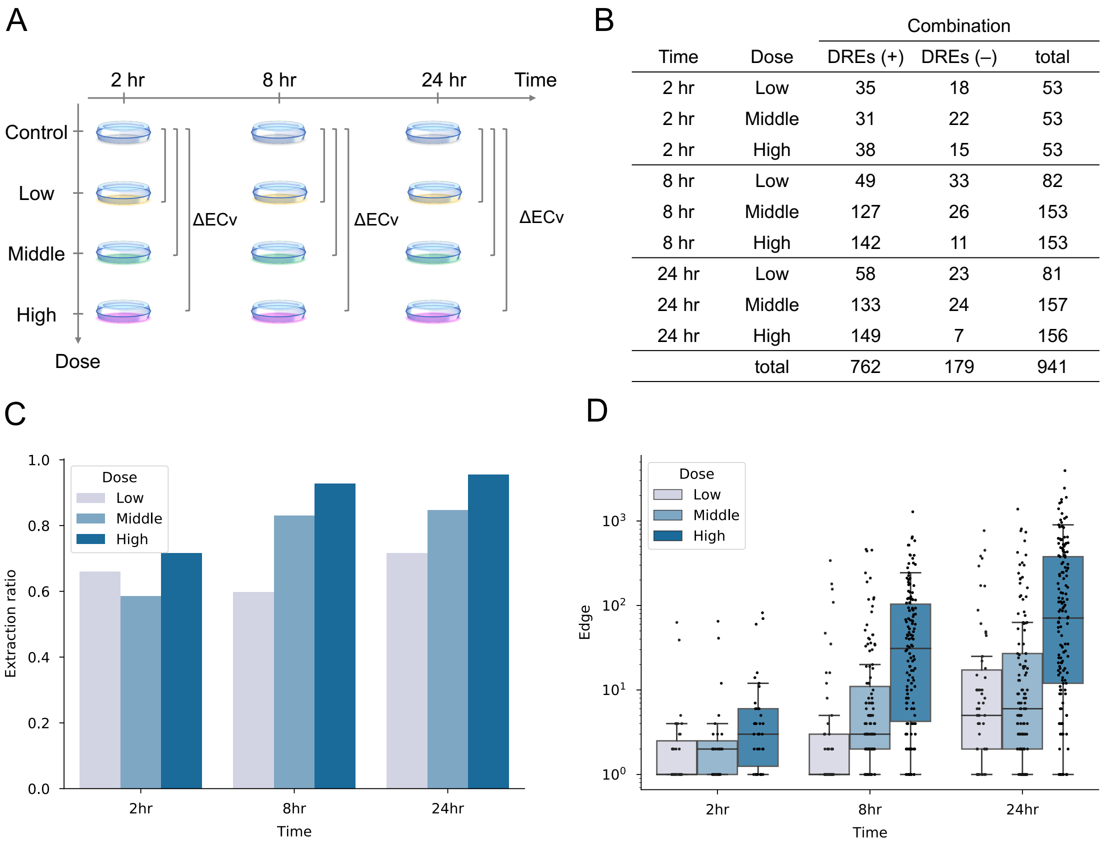


Figure 3.3: The precision-recall curve for different sample sizes of simulation datasets. The sample sizes are 500 (blue), 1000 (orange), 2000 (green) and 3000 (red).

Briefly, the four types of datasets were generated with 500, 1000, 2000 and 3000 sample sizes, and then the network estimation was performed on each dataset. The performance of the network estimation was evaluated by checking whether the structure of the original network can be reproduced. The area under the precision-recall curve (PRAUC) was 0.387, 0.360, 0.395, 0.399 for sample sizes 500, 1000, 2000, and 3000, respectively (Figure 3.3). The PRAUC tended to become larger as the sample size increases, which is consistent with the previous study [17]. This result suggests that the estimated network structure approaches the true structure as the number of samples increases. Furthermore, this shows that it is feasible to estimate a network up to 3000 sample sizes in a realistic time using the supercomputer. To construct the basal gene network for the toxicogenomics dataset, the network estimation was performed with 384 CPU nodes along with 8 processes per node, resulting in a network of 10,554 nodes and 113,054 edges in 17 hours 18 minutes.

3.2 Identification of condition-specific subnetworks from the basal network according to the combination of drug, time, and dose



		Combination		
Time	Dose	DREs (+)	DREs (–)	total
2 hr	Low	35	18	53
2 hr	Middle	31	22	53
2 hr	High	38	15	53
8 hr	Low	49	33	82
8 hr	Middle	127	26	153
8 hr	High	142	11	153
24 hr	Low	58	23	81
24 hr	Middle	133	24	157
24 hr	High	149	7	156
	total	762	179	941

Figure 3.4: Subnetwork extraction analysis. (A) Illustration of ΔECv calculation. (B) The list of the number of combinations obtained from ΔECv calculation for all the possible combinations of drugs, doses and times. DREs(+) indicates that at least one edge was extracted, DREs(-) that not even one edge was extracted. (C) The extraction rate of subnetworks calculated as #(DREs(+) samples)/#(total samples) from Figure 3.4B result. (D) Edge distribution of the subnetworks extracted by ΔECv calculation.

To examine changes in the network state, I extracted a subnetwork defined as a set of differentially regulated edges (DREs) from the basal network for each combination of drug, time and dose using ΔECv method [52, 97]. Since primary hepatocytes were exposed to a drug with different doses until given three time points in this dataset, the ΔECv calculation was defined as the difference between certain dose samples and the control samples at the corresponding single time point (Figure 3.4A). This calculation was performed for every available combination. As a result, there were 941 sample combinations, of which 762 combinations were able to extract at least one DREs, while 179 combinations were unable to extract a single DRE (Figure 3.4B). Namely, in these 179 combinations, no observable perturbations were captured at the ΔECv threshold used. The subnetwork extraction rate was proportional to the dose and time, which shows that the cellular perturbations are more likely caused under the severe conditions, such as high dose or long exposure time (Figure 3.4C). In contrast, there were some populations in which no perturbation was observed even under these conditions (Figure 3.4B). To grasp the magnitude of the extracted 762 subnetworks, the distribution of the number of edges was examined for each subnetwork. The network size was found to increase in proportion to the dose and time, indicating that the level of the perturbated intracellular system induced by the exposure of DILI-concerned drugs also increases depending on the dose and time (Figure 3.4D).

3.3 Establishment of the Variational Graph Autoencoder

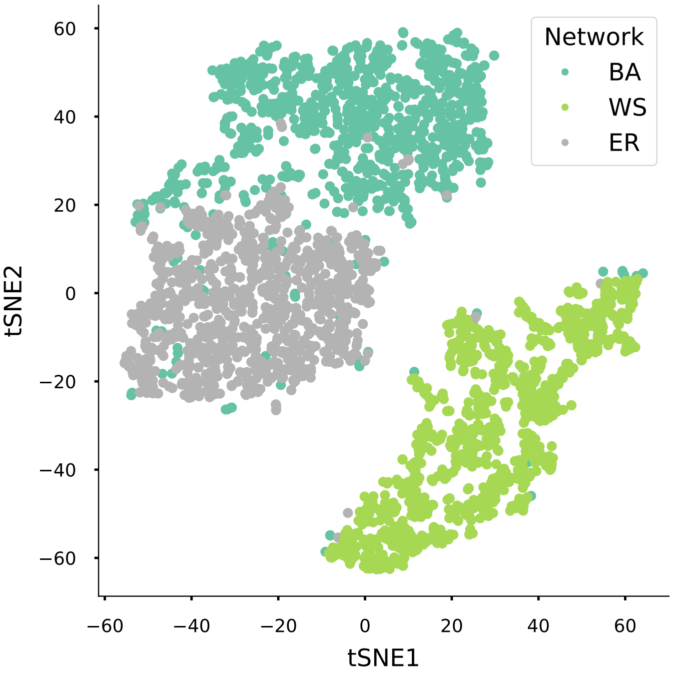
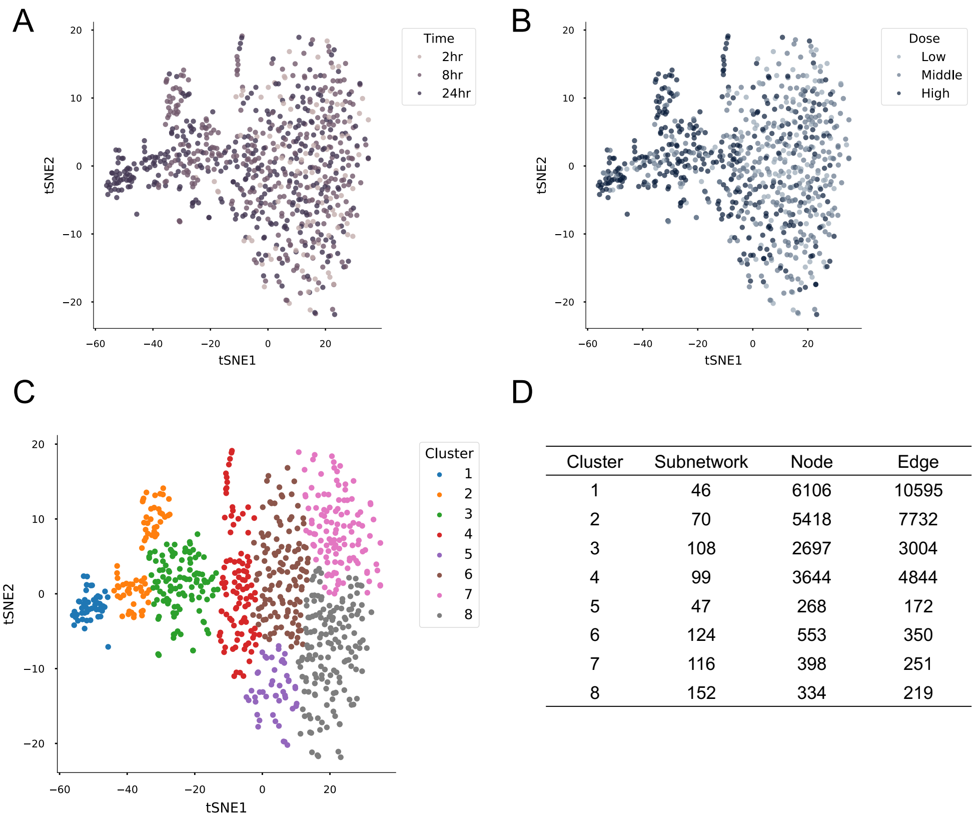


Figure 3.5: The visualization of latent features for the three model networks by the VGAE. The number of networks is 1,000 each for BA (green), WS (yellow), and ER (gray).

This wide variety of 762 subnetworks suggested that they contain essential clues for interpreting the drug-induced cellular perturbations. To comprehensively analyze these subnetworks based on the network information, I developed the multiple graph-level Variational Graph Autoencoder (VGAE), which was designed to acquire the latent features of a network as a vector representation (Figure 3.2). To begin with, in order to test the validity of the established VGAE, the VGAE performance was evaluated using three types of artificially-generated networks: the BA model (scale-free network) [107], the WS model (small-world network) [108] and ER model (random network) [109]. Since these three types of networks are known to have different network structures, I expected that if the VGAE worked, it would be able to distinguish between each network. The networks were prepared in 1000 pieces for each model network. Then, these networks acquired the latent features through the VGAE, which were visualized in two dimensional space using t-SNE [111]. The result showed that these networks are categorized into three groups for each model network (Figure 3.5). This demonstrates that the latent features acquired by the VGAE represent the characteristic network information.

3.4 The VGAE identifies multiple clusters characterized by network size, level of exogenous perturbations, and cytotoxicity.



Cluster	Subnetwork	Node	Edge
1	46	6106	10595
2	70	5418	7732
3	108	2697	3004
4	99	3644	4844
5	47	268	172
6	124	553	350
7	116	398	251
8	152	334	219

Figure 3.6: Characterization of the subnetworks in latent space with the VGAE. (A, B) The visualization of latent features in subnetworks mapped with doses (A) and times (B). (C) The clustering of the 762 subnetworks based on their latent features. (D) The list of the subnetwork information for clusters. The nodes and edges represent the set of nodes and edges that comprise the subnetworks contained in each cluster.

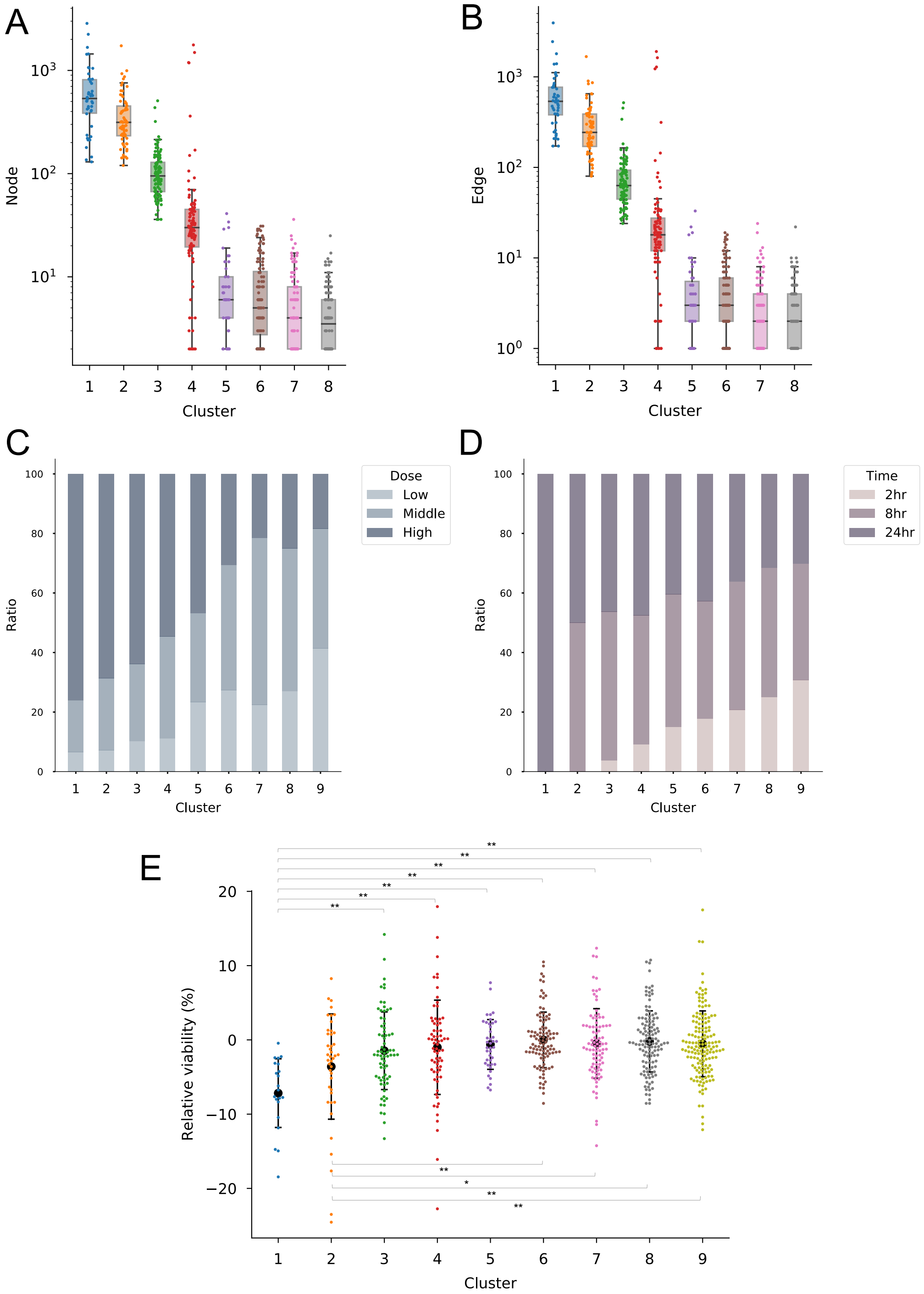


Figure 3.7: The cluster characterization analyses. (A, B) The distribution of nodes (A) and edges (B) for clusters. (C, D) The ratio of the subnetworks comprising each cluster for doses (C) and times (D). (E) The cytotoxicity assay for cluster 1 (n=21), cluster 2 (n=42), cluster 3 (n=72), cluster 4 (n=70), cluster 5 (n=40), cluster 6 (n=103), cluster 7 (n=89), cluster 8 (n=124), and cluster 9 (n=155, negative control). Turkey test was performed for statistical test; **p<0.01; *p<0.05. Error bar, mean ± standard deviation.

Given that the VGAE result on the artificially-generated networks, I reasoned that the extracted 762 subnetworks (Figure 3.4) could also be analyzed using the VGAE based on their network features. These subnetworks were subjected to the VGAE and the acquired latent features were visualized. The labels for three doses and times were mapped onto it, showing that the subnetworks labeled with high dose or long exposure are likely to be enriched especially in the left area compared to other subnetworks (Figure 3.6A, B). To further investigate the subnetwork distribution in this two dimensional space, clustering was performed on the subnetworks, which resulted in eight clusters (Figure 3.6C). The summary of each cluster is shown in Figure 3.6D. Here, cluster 9 was additionally defined for the population from which no subnetworks could be extracted in the ΔECv calculation (Figure 3.4B), and thus this cluster 9 comprised of the 179 combinations was intended as a negative control group.

To identify the characteristics of the clusters, I examined the network size of the subnetworks included in each cluster. The clusters were then found to be separated by the difference in the size of nodes and edges (Figure 3.7A, B). The network size increased from cluster 8 to cluster 1. The composition ratios of the labels for each of three doses and times were also examined (Figure 3.7C, D). This analysis showed that the percentage of subnetworks subjected to strong levels of exogenous perturbation, such as high dose or long exposure, gradually rises from cluster 9 to cluster 1. By contrast, that of subnetworks subjected to weak levels of exogenous perturbation, such as low dose or short exposure, decreases from cluster 9 to cluster 1. Next, to determine whether there is a link between cluster and cell viability, I assessed the extent to which cytotoxicity occurs across clusters. The cytotoxicity tended to enhance from cluster 9 to cluster 1, of which clusters 1 and 2 exhibited significantly higher cytotoxicity than the other clusters (Figure 3.7E). Taken together, these analyses revealed that the identified clusters are associated with gradual changes in the network size, the level of the exogenous perturbation, and the cytotoxicity.

3.5 The clusters represent a sequential cascade involved in drug metabolism

To confirm the biological significance of the clusters, pathway analysis was performed (Figure 3.8). The pathways related to early drug metabolism, such as cytochrome P450 reaction, glucuronidation, and phase I and II reactions, were strongly enriched in cluster 5, 6, 7, and 8. In parallel with the weakening of these initial signaling pathways, immune response pathways, such as interferon signalings and complement signaling, gradually appeared in

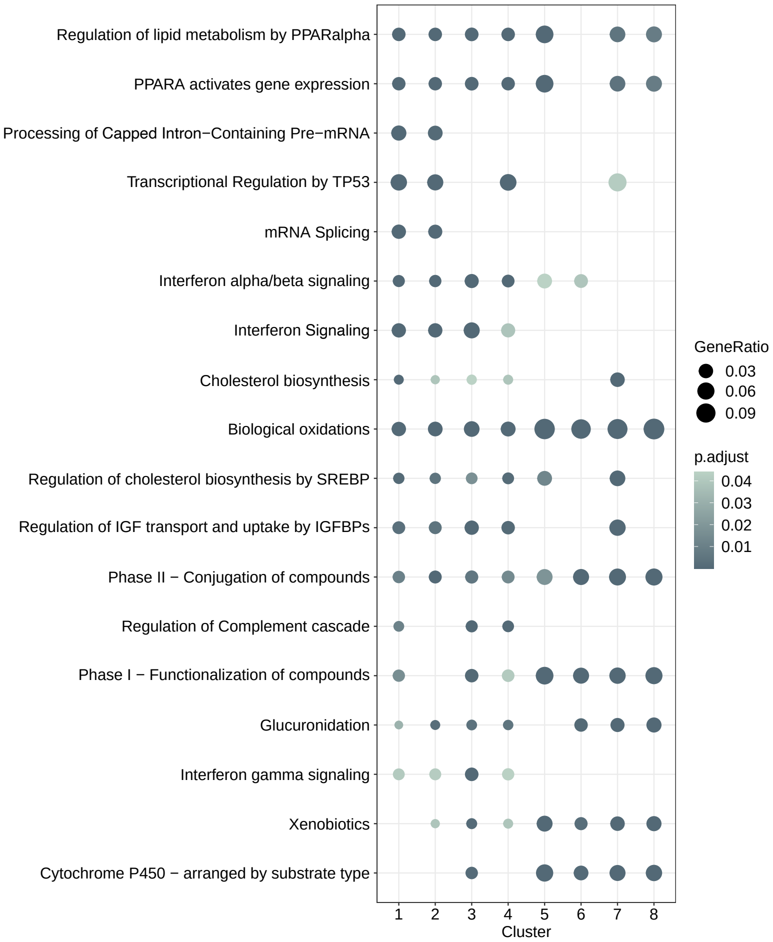


Figure 3.8: Pathway analysis for each cluster. The top 10 terms are shown for each cluster.

clusters 1, 2, 3, and 4. Moreover, p53-related signaling pathway appeared especially in clusters 1 and 2. Since p53 pathway and immune response pathways are known to be involved in apoptosis and cell phagocytosis induced by immune cells, respectively, these factors imply the signs of cytotoxicity [70, 115]. This could support the cytotoxicity observed in cluster 1 and 2 (Figure 3.7E). Collectively, these results suggest that the transition of cluster 8 to 1 corresponds to a sequential cascade from early drug metabolism to hepatotoxicity [70, 116].

3.6 Drugs with a high DILI risk show characteristic distribution patterns across the clusters.

The cluster characterization analyses suggested that clusters 5, 6, 7, and 8 represent networks in which the initial metabolic process has begun, clusters 3 and 4 represent networks in

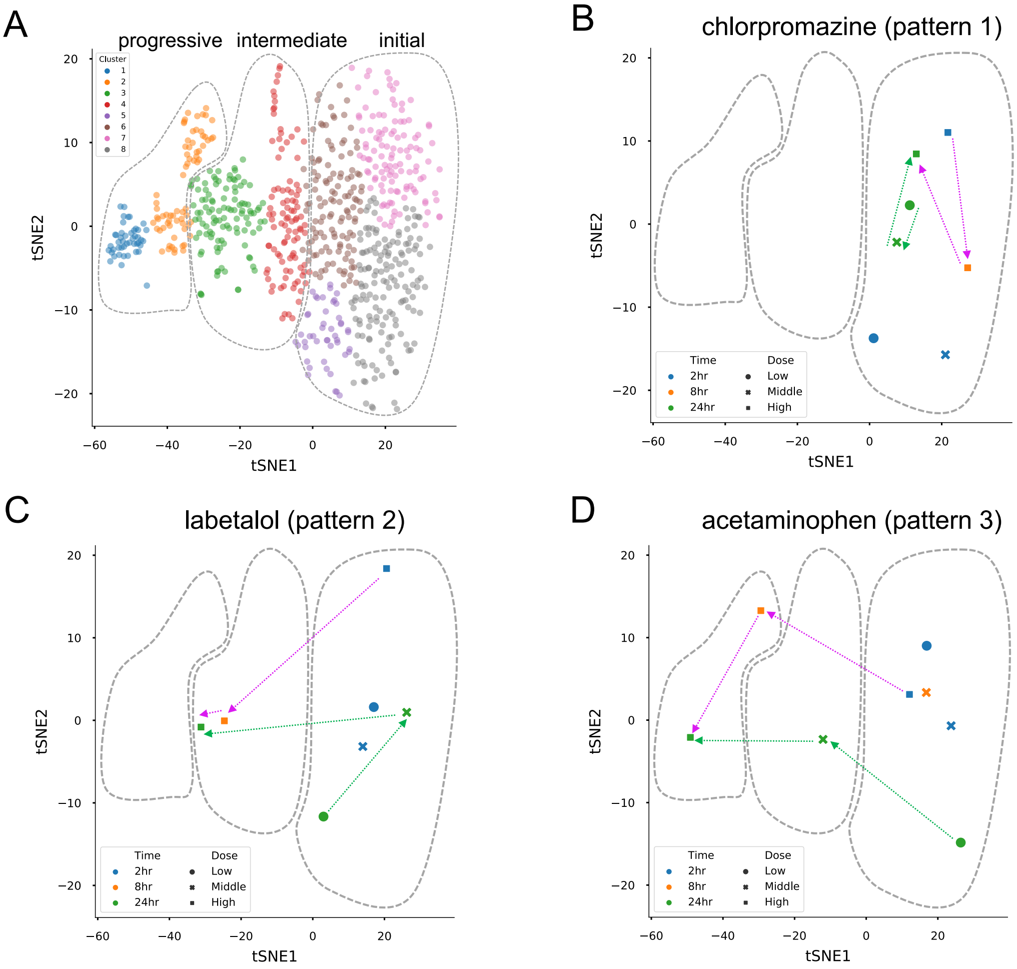


Figure 3.9: (A) The clusters were divided into the three zones of progressive (cluster 1 and 2), intermediate (cluster 3 and 4) and initial (cluster 5, 6, 7 and 8). (B-D) The network transition patterns with representative drugs are shown. Chlorpromazine for pattern 1 (B), labetalol for pattern 2 (C), acetaminophen for pattern 3 (D). The green dotted line shows the transition of the three doses at 24 hr time point, and the magenta dotted line shows the transition of the three time points at high dose.

which the drug metabolism response has gradually progressed from the state of clusters 5, 6, 7, and 8 but has not yet reached cytotoxicity, and clusters 1 and 2 represent networks in which the state of clusters 3 and 4 has been further progressed to cytotoxicity. To simplify the classification of clusters based on these observations, clusters 5, 6, 7, and 8 were assigned as the “initial” zone, clusters 3 and 4 as the “intermediate” zone, and clusters 1 and 2 as the “progressive” zone (Figure 3.9A). Considering that the subnetworks were located somewhere in these three zones according to the level of drug-induced perturbation, the transition profiles for each drug were examined. In order to compare the transitions in response to the dose- and time-series perturbation, 52 drugs were selected for which all microarray samples were available under all the three conditions of doses and times.

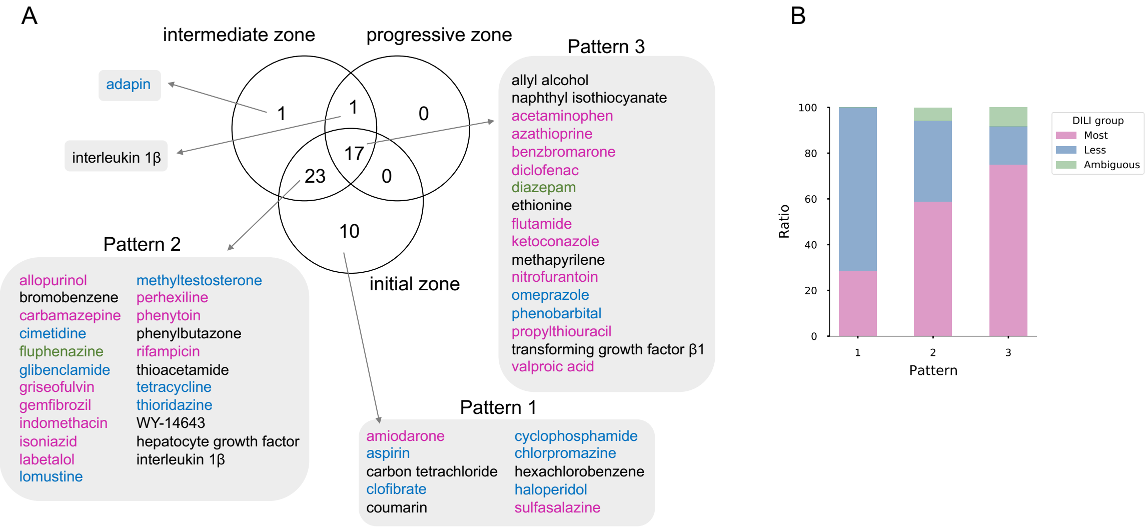


Figure 3.10: (A) The Venn diagram shows the distribution of 52 drugs according to the three zones. Drugs labeled with Most-concern DILI, Less-concern DILI, Ambiguous-concern DILI in the benchmark dataset are colored with magenta, blue, and green, respectively. Drugs that are not listed in the benchmark dataset are colored with black. (B) The stack bar represents the percentage of drugs labeled with DILI risk in the indicated zones. The DILI risk-unlabeled drugs (black) were excluded from the calculation.

The network transitions of the drugs were found to show three patterns: drugs that stay in the initial zone exclusively (pattern 1) (Figure 3.9B), drugs that transition through the two initial and intermediate zones (pattern 2) (Figure 3.9C), and drugs that transition through all three zones (pattern 3) (Figure 3.9D). The drugs in patterns 2 and 3 tended to shift from initial to progressive zone in a dose- and time-dependent manner, while the drugs in pattern 1 did not shift significantly regardless of doses and times (Figure 3.9B-D). The Venn diagram analysis showed that the number of drugs in patterns 1, 2, and 3 were 10, 23, and 17, respectively (Figure 3.10A). To examine the association between these characterized drugs and human DILI, they were inquired against a benchmark dataset of DILI risk, which categorizes the FDA-approved drugs into several groups (Most-, Less-, Ambiguous-, and No-DILI concern) according to the grade of DILI [117]. The drugs with high DILI risk were found to be enriched in patterns 2 and 3 compared to pattern 1 (Figure 3.10B).

4 Discussion

By analyzing the status of many networks according to the combination of drug, dose and time, this study revealed the relationship between those networks and hepatotoxicity. Networks are believed to be essential for understanding living organisms, but analyzing biological networks is challenging due to their huge size and complexity [56]. With the recent development of GNNs technology, which aims to learn network features from network structures, there has been increasing attention to the application of this technology in biological networks [118]. However, the interpretation of GNNs' results usually remains elusive. This study shows that it is possible to interpret what GNNs have learned with respect to the hepatotoxicity, the network sizes and the perturbation levels, which further characterizes drugs. These data indicate that the VGAE is interpretable and could be used to analyze biological networks, which is also supported by another recent study [119].

In the application of GNNs to biological networks, a single network has been largely used so far. The main reason for this is that there is not a large amount of network data representing different biological states. Therefore, the concept of using GNNs to analyze many networks has not yet been realized in biology. This study has presented one possible solution for this, which can be brought by the emergence of the following three things: 1) the development of methods for extracting sample-specific networks; 2) the advance of artificial intelligence techniques with GNNs; 3) the progress of computational resources. On the other hand, the issue with this approach is that it does not directly deal with a huge scale network such as the basal network due to the limitations of current GNNs technology. The solution to this issue would provide further network information for understanding the cellular mechanisms.

Furthermore, this study shows that the human DILI risk could be extrapolated from *in vitro* hepatotoxicity using network latent features, because the drugs showing pattern 2 or 3 are strongly associated with DILI (Figure 3.10B). While human DILI is generally known to occur after several weeks of taking drugs [120, 71], the data used in this study were derived from a time series of *in vitro* hepatocyte experiments up to 24 hours [78]. This indicates that the observations in this study represent relatively early cellular responses compared to the occurrence of human DILI. Nevertheless, the results exhibit the potential to identify drugs with the DILI risk by capturing dose- and time-dependent network transitions in a short experimental scheme. Given that early prediction of DILI is needed for long-term drug administration [121, 122], this approach might be useful to assess the DILI risk in advance. For example, although diazepam is still considered an Ambiguous-DILI concern in the DILI benchmark dataset [117], this study indicates that it might have a high DILI risk (Figure 3.10A). The possibility of misclassification remains, but the use of diazepam would require careful attention. Additionally, if data are obtained for a new drug under the same conditions as those used in this study, it will be possible to evaluate the DILI risk using this model. Since DILI is caused not only by internal factors in the liver but also by complex factors in the whole body, it is not sufficient to investigate the intracellular mechanisms *in* *vitro* hepatocytes. Further analysis and establishment of methods at the organ and body level are necessary. In this study, the link between condition-specific gene networks and hepatotoxicity was confirmed, allowing us to take a new approach in terms of individual networks based on toxicogenomics data. A comprehensive integration of conventional methods, network-based methods, and wet experiments would further drive the elucidation of the DILI mechanism.

Summary

In this study, by integrating gene expression information from multiple samples with different conditions and constructing a single network representing the cellular system, I proposed a new data-driven approach to extract and classify networks that are perturbated specifically among arbitrary samples, which realizes a sample specific network analysis. The approach was then applied to cancer, infectious disease, and hepatotoxicity data in the three chapters.

In Chapter 1, I developed a novel method to analyze differences between samples based on networks representing underlying relationships between genes using Bayesian Network. This method quantifies sample specific networks using the proposed *Edge Contribution value* (ECv) based on the estimated system, which realizes condition-specific subnetwork extraction. To validate the method, I applied it to a dataset of TGFβ-treated lung cancer cells that are related to the process of metastasis and thus prognosis in cancer biology. I successfully extracted and established the network implicated in the epithelial-mesenchymal transition process, which is consistent with the previous biological findings on TGFβ. Furthermore, I found that the sample specific ECv patterns of this network can characterize the survival of lung cancer patients. These results show that this method allows us to analyze sample specific networks based on cellular system changes and to discover new relationships between genes.

In Chapter 2, the network analysis method developed in Chapter 1 was applied to the COVID-19. To elucidate how SARS-CoV-2 behaves in human host cells, I examined the dynamic changes in gene-to-gene regulatory networks in response to SARS-CoV-2 infection, revealing that interferon signaling gradually switched to the subsequent inflammatory cytokine signaling cascades as the intracellular virus increased. Furthermore, I succeeded in capturing a COVID-19 patient-specific network in which transduction of these signals was concurrently induced, which is consistent with the pathological progression of COVID-19 to some extent. This enabled us to explore the local regulatory systems influenced by SARS-CoV-2 in host cells more precisely at an individual level. This panel of network analyses has provided new insights into SARS-CoV-2 research from the perspective of

cellular systems.

In Chapter 3, a large-scale gene network analysis was carried out for DILI using a large gene expression dataset of 2605 samples from exposure of human hepatocytes to 158 drugs at multiple time points and dose levels with a hybrid approach of the method developed in Chapter 1 and GNN techniques. Firstly, by employing the latest supercomputer, I succeeded in estimating a gene network with a largest number of sample sizes ever from these microarray data. Secondly, a multitude of condition-specific networks were determined from the estimated network according to the combination of drugs, times, and doses using the method presented in Chapter 1. Lastly, to comprehensively analyze the state of these 762 condition-specific networks, I newly developed a network classification method using GNN techniques, which reveals that the networks can be classified into several groups according to the cellular perturbation levels. The characterization of these groups led to identify features of the network that could be potentially responsible for hepatotoxicity. These results show that networks play an important role in the prediction and mechanism elucidation of DILI.

In conclusion, from the viewpoint of understanding living organisms as systems through molecular networks, I developed and evaluated a new network analysis approach and proved the validity of it. This study has solved the bottleneck that made it impossible to profile sample-specific networks with a new method, thus providing a new avenue for network analysis. Furthermore, by combining this method with the emerging GNN techniques, I have presented a comprehensive analysis strategy for multiple networks that can be obtained from the developed method, which further extends the scope of network analysis. The methods and findings obtained in this study suggest the possibility of data-driven approaches in biology, medicine, and pharmacology in the future, and will lead to new discoveries of biomolecular mechanisms, which are expected to contribute to the mutual integration and bridging of wet and dry research.

Acknowledgments

I would like to express my sincere gratitude to my advisors Prof. Yasushi Okuno and Prof. Fumiyoshi Yamashita. My sincere thanks also go to Prof. Yoshinori Tamada. I thank all the members who contributed to my published works. I also thank to members of the Okuno lab and the Yamashita lab for their helpful discussions and intellectual contributions, especially many thanks to Prof. Mitsuru Hashida, Prof. Masahiko Nakatsui, Mayumi Kamada, Ryosuke Kojima, Yuji Okamoto, Eiichiro Uchino and Shoichi Ishida. I wish to sincerely thank Peter Carlton, Georgia Kafer, Aya Sato, Xuan Li and members of the Carlton lab for guiding me into the world of science and supporting me over the years. Finally, I would like to thank my family for their long support and encouragement.

List of Publications

Journal articles related to this thesis

- Y. Tanaka, Y. Tamada, M. Ikeguchi, F. Yamashita, and Y. Okuno, “System-based differential gene network analysis for characterizing a sample-specific subnetwork,” *Biomolecules*, Volume 10(2), 306, 2020.

- Y. Tanaka, K. Higashihara, M. A. Nakazawa, F. Yamashita, Y. Tamada, Y. Okuno, “Dynamic changes in gene-to-gene regulatory networks in response to SARS-CoV-2 infection,” *Scientific Reports*, Volume 11, 11241, 2021.

References

[1] F. M. Delgado and F. Gómez-Vela, "Computational methods for gene regulatory networks reconstruction and analysis: A review," *Artif. Intell. Med.*, vol. 95, pp. 133–145, Apr. 2019.

[2] H. Niwa, "The pluripotency transcription factor network at work in reprogramming," *Curr. Opin. Genet. Dev.*, vol. 28, pp. 25–31, Oct. 2014.

[3] P. Lecca and C. Priami, "Biological network inference for drug discovery," *Drug Discov. Today*, vol. 18, pp. 256–264, Mar. 2013.

[4] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis, "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 93, pp. 10614–10619, Oct. 1996.

[5] F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci, and D. Betel, "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data," *Genome Biol.*, vol. 14, no. 9, p. R95, 2013.

[6] V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop, "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nat. Genet.*, vol. 34, pp. 267–273, June 2003.

[7] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, pp. 15545–15550, Oct. 2005.

[8] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources,” *Nat. Protoc.*, vol. 4, no. 1, pp. 44–57, 2009.

[9] L. E. Chai, S. K. Loh, S. T. Low, M. S. Mohamad, S. Deris, and Z. Zakaria, “A review on the computational approaches for gene regulatory network construction,” *Comput. Biol. Med.*, vol. 48, pp. 55–65, May 2014.

[10] P. Creixell, J. Reimand, S. Haider, G. Wu, T. Shibata, M. Vazquez, V. Mustonen, A. Gonzalez-Perez, J. Pearson, C. Sander, B. J. Raphael, D. S. Marks, B. F. F. Ouellette, A. Valencia, G. D. Bader, P. C. Boutros, J. M. Stuart, R. Linding, N. Lopez-Bigas, L. D. Stein, and Mutation Consequences and Pathway Analysis Working Group of the International Cancer Genome Consortium, “Pathway and network analysis of cancer genomes,” *Nat. Methods*, vol. 12, pp. 615–621, July 2015.

[11] K.-K. Yan, D. Wang, A. Sethi, P. Muir, R. Kitchen, C. Cheng, and M. Gerstein, “Cross-Disciplinary network comparison: Matchmaking between hairballs,” *Cell Syst*, vol. 2, pp. 147–157, Mar. 2016.

[12] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano, “ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context,” *BMC Bioinformatics*, vol. 7 Suppl 1, p. S7, Mar. 2006.

[13] H. Araki, Y. Tamada, S. Imoto, B. Dunmore, D. Sanders, S. Humphrey, M. Nagasaki, A. Doi, Y. Nakanishi, K. Yasuda, Y. Tomiyasu, K. Tashiro, C. Print, D. Stephen Charnock-Jones, S. Kuhara, and S. Miyano, “Analysis of PPARα-dependent and PPARα-independent transcript regulation following fenofibrate treatment of human endothelial cells,” *Angiogenesis*, vol. 12, pp. 221–229, Sept. 2009.

[14] L. Wang, D. G. Hurley, W. Watkins, H. Araki, Y. Tamada, A. Muthukaruppan, L. Ranjard, E. Derkac, S. Imoto, S. Miyano, E. J. Crampin, and C. G. Print, “Cell cycle gene networks are associated with melanoma prognosis,” *PLoS One*, vol. 7, p. e34247, Apr. 2012.

[15] M. Affara, D. Sanders, H. Araki, Y. Tamada, B. J. Dunmore, S. Humphreys, S. Imoto, C. Savoie, S. Miyano, S. Kuhara, D. Jeffries, C. Print, and D. S. Charnock-Jones, “Vasohibin-1 is identified as a master-regulator of endothelial cell apoptosis using gene network analysis,” *BMC Genomics*, vol. 14, p. 23, Jan. 2013.

[16] A. J. Singh, S. A. Ramsey, T. M. Filtz, and C. Kioussi, "Differential gene regulatory networks in development and disease," *Cell. Mol. Life Sci.*, vol. 75, pp. 1013–1025, Mar. 2018.

[17] Y. Tamada, S. Imoto, H. Araki, M. Nagasaki, C. Print, D. S. Charnock-Jones, and S. Miyano, "Estimating genome-wide gene networks using nonparametric bayesian network models on massively parallel computers," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 8, pp. 683–697, May 2011.

[18] T. Shimamura, S. Imoto, Y. Shimada, Y. Hosono, A. Niida, M. Nagasaki, R. Yamaguchi, T. Takahashi, and S. Miyano, "A novel network profiling analysis reveals system changes in epithelial-mesenchymal transition," *PLoS One*, vol. 6, p. e20804, June 2011.

[19] X. Yu, T. Zeng, X. Wang, G. Li, and L. Chen, "Unravelling personalized dysfunctional gene network of complex diseases based on differential network model," *J. Transl. Med.*, vol. 13, p. 189, June 2015.

[20] M. L. Kuijjer, M. G. Tung, G. Yuan, J. Quackenbush, and K. Glass, "Estimating Sample-Specific regulatory networks," *iScience*, vol. 14, pp. 226–240, Apr. 2019.

[21] Y. Sun, A. Daemen, G. Hatzivassiliou, D. Arnott, C. Wilson, G. Zhuang, M. Gao, P. Liu, A. Boudreau, L. Johnson, and J. Settleman, "Metabolic and transcriptional profiling reveals pyruvate dehydrogenase kinase 4 as a mediator of epithelial-mesenchymal transition and drug resistance in tumor cells," *Cancer Metab*, vol. 2, p. 20, Nov. 2014.

[22] Cancer Genome Atlas Research Network, "Comprehensive molecular profiling of lung adenocarcinoma," *Nature*, vol. 511, pp. 543–550, July 2014.

[23] S. Imoto, T. Goto, and S. Miyano, "Estimation of genetic networks and functional structures between genes by using bayesian networks and nonparametric regression," *Pac. Symp. Biocomput.*, vol. 7, pp. 175–186, 2002.

[24] C. Arima, T. Kajino, Y. Tamada, S. Imoto, Y. Shimada, M. Nakatochi, M. Suzuki, H. Isomura, Y. Yatabe, T. Yamaguchi, K. Yanagisawa, S. Miyano, and T. Takahashi, "Lung adenocarcinoma subtypes definable by lung development-related miRNA expression profiles in association with clinicopathologic features," *Carcinogenesis*, vol. 35, pp. 2224–2231, Oct. 2014.

[25] R. Gendelman, H. Xing, O. K. Mirzoeva, P. Sarde, C. Curtis, H. S. Feiler, P. McDonagh, J. W. Gray, I. Khalil, and W. M. Korn, "Bayesian network inference modeling

identifies TRIB1 as a novel regulator of Cell-Cycle progression and survival in cancer cells," *Cancer Res.*, vol. 77, pp. 1575–1585, Apr. 2017.

[26] M. J. Goldman, B. Craft, M. Hastie, K. Repečka, F. McDade, A. Kamath, A. Banerjee, Y. Luo, D. Rogers, A. N. Brooks, J. Zhu, and D. Haussler, "Visualizing and interpreting cancer genomics data via the xena platform," *Nat. Biotechnol.*, vol. 38, pp. 675–678, May 2020.

[27] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res.*, vol. 43, p. e47, Apr. 2015.

[28] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. R. Stat. Soc.*, vol. 57, pp. 289–300, Jan. 1995.

[29] A. Krämer, J. Green, J. Pollard, Jr, and S. Tugendreich, "Causal analysis approaches in ingenuity pathway analysis," *Bioinformatics*, vol. 30, pp. 523–530, Feb. 2014.

[30] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, M. Ceccarelli, G. Bontempi, and H. Noushmehr, "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data," *Nucleic Acids Res.*, vol. 44, p. e71, May 2016.

[31] M. Mounir, M. Lucchetta, T. C. Silva, C. Olsen, G. Bontempi, X. Chen, H. Noushmehr, A. Colaprico, and E. Papaleo, "New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx," *PLoS Comput. Biol.*, vol. 15, p. e1006701, Mar. 2019.

[32] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, pp. 2498–2504, Nov. 2003.

[33] S. Heerboth, G. Housman, M. Leary, M. Longacre, S. Byler, K. Lapinska, A. Willbanks, and S. Sarkar, "EMT and tumor metastasis," *Clin. Transl. Med.*, vol. 4, p. 6, Feb. 2015.

[34] R. O. Hynes, "The extracellular matrix: not just pretty fibrils," *Science*, vol. 326, pp. 1216–1219, Nov. 2009.

[35] K. Song, Q. Li, Z.-Z. Jiang, C.-W. Guo, and P. Li, “Heparan sulfate d-glucosaminyl 3-O-sulfotransferase-3B1, a novel epithelial-mesenchymal transition inducer in pancreatic cancer,” *Cancer Biol. Ther.*, vol. 12, pp. 388–398, Sept. 2011.

[36] C.-Y. Hsu, G.-C. Chang, Y.-J. Chen, Y.-C. Hsu, Y.-J. Hsiao, K.-Y. Su, H.-Y. Chen, C.-Y. Lin, J.-S. Chen, Y.-J. Chen, Q.-S. Hong, W.-H. Ku, C.-Y. Wu, B.-C. Ho, C.-C. Chiang, P.-C. Yang, and S.-L. Yu, “FAM198B is associated with prolonged survival and inhibits metastasis in lung adenocarcinoma via blockage of ERK-Mediated MMP-1 expression,” *Clin. Cancer Res.*, vol. 24, pp. 916–926, Feb. 2018.

[37] J. Wang, N. Ding, Y. Li, H. Cheng, D. Wang, Q. Yang, Y. Deng, Y. Yang, Y. Li, X. Ruan, F. Xie, H. Zhao, and X. Fang, “Insulin-like growth factor binding protein 5 (IGFBP5) functions as a tumor suppressor in human melanoma cells,” *Oncotarget*, vol. 6, pp. 20636–20649, Aug. 2015.

[38] G. Tzanakakis, R.-M. Kavasi, K. Voudouri, A. Berdiaki, I. Spyridaki, A. Tsatsakis, and D. Nikitovic, “Role of the extracellular matrix in cancer-associated epithelial to mesenchymal transition phenomenon,” *Dev. Dyn.*, vol. 247, no. 3, pp. 368–381, 2018.

[39] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. F. Gao, W. Tan, and China Novel Coronavirus Investigating and Research Team, “A novel coronavirus from patients with pneumonia in china, 2019,” *N. Engl. J. Med.*, vol. 382, pp. 727–733, Feb. 2020.

[40] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, and Y.-Z. Zhang, “A new coronavirus associated with human respiratory disease in china,” *Nature*, vol. 579, pp. 265–269, Mar. 2020.

[41] E. Dong, H. Du, and L. Gardner, “An interactive web-based dashboard to track COVID-19 in real time,” *Lancet Infect. Dis.*, vol. 20, pp. 533–534, May 2020.

[42] J. Hellewell, S. Abbott, A. Gimma, N. I. Bosse, C. I. Jarvis, T. W. Russell, J. D. Munday, A. J. Kucharski, W. J. Edmunds, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, S. Funk, and R. M. Eggo, “Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts,” *Lancet Glob Health*, vol. 8, pp. e488–e496, Apr. 2020.

[43] D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, K. M. White, M. J. O'Meara, V. V. Rezelj, J. Z. Guo, D. L. Swaney, T. A. Tummino, R. Hüttenhain, R. M. Kaake, A. L. Richards, B. Tutuncuoglu, H. Foussard, J. Batra, K. Haas, M. Modak, M. Kim, P. Haas, B. J. Polacco, H. Braberg, J. M. Fabius, M. Eckhardt, M. Soucheray, M. J. Bennett, M. Cakir, M. J. McGregor, Q. Li, B. Meyer, F. Roesch, T. Vallet, A. Mac Kain, L. Miorin, E. Moreno, Z. Z. C. Naing, Y. Zhou, S. Peng, Y. Shi, Z. Zhang, W. Shen, I. T. Kirby, J. E. Melnyk, J. S. Chorba, K. Lou, S. A. Dai, I. Barrio-Hernandez, D. Memon, C. Hernandez-Armenta, J. Lyu, C. J. P. Mathy, T. Perica, K. B. Pilla, S. J. Ganesan, D. J. Saltzberg, R. Rakesh, X. Liu, S. B. Rosenthal, L. Calviello, S. Venkataramanan, J. Liboy-Lugo, Y. Lin, X.-P. Huang, Y. Liu, S. A. Wankowicz, M. Bohn, M. Safari, F. S. Ugur, C. Koh, N. S. Savar, Q. D. Tran, D. Shengjuler, S. J. Fletcher, M. C. O'Neal, Y. Cai, J. C. J. Chang, D. J. Broadhurst, S. Klippsten, P. P. Sharp, N. A. Wenzell, D. Kuzuoglu-Ozturk, H.-Y. Wang, R. Trenker, J. M. Young, D. A. Cavero, J. Hiatt, T. L. Roth, U. Rathore, A. Subramanian, J. Noack, M. Hubert, R. M. Stroud, A. D. Frankel, O. S. Rosenberg, K. A. Verba, D. A. Agard, M. Ott, M. Emerman, N. Jura, M. von Zastrow, E. Verdin, A. Ashworth, O. Schwartz, C. d'Enfert, S. Mukherjee, M. Jacobson, H. S. Malik, D. G. Fujimori, T. Ideker, C. S. Craik, S. N. Floor, J. S. Fraser, J. D. Gross, A. Sali, B. L. Roth, D. Ruggero, J. Taunton, T. Kortemme, P. Beltrao, M. Vignuzzi, A. García-Sastre, K. M. Shokat, B. K. Shoichet, and N. J. Krogan, "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing," *Nature*, vol. 583, pp. 459–468, July 2020.

[44] D. Bojkova, K. Klann, B. Koch, M. Widera, D. Krause, S. Ciesek, J. Cinatl, and C. Münch, "Proteomics of SARS-CoV-2-infected host cells reveals therapy targets," *Nature*, vol. 583, pp. 469–472, July 2020.

[45] D. Blanco-Melo, B. E. Nilsson-Payant, W.-C. Liu, S. Uhl, D. Hoagland, R. Møller, T. X. Jordan, K. Oishi, M. Panis, D. Sachs, T. T. Wang, R. E. Schwartz, J. K. Lim, R. A. Albrecht, and B. R. tenOever, "Imbalanced host response to SARS-CoV-2 drives development of COVID-19," *Cell*, vol. 181, pp. 1036–1045.e9, May 2020.

[46] J. Yan, S. L. Risacher, L. Shen, and A. J. Saykin, "Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data," *Brief. Bioinform.*, vol. 19, pp. 1370–1381, Nov. 2018.

[47] M. Recanatini and C. Cabrelle, "Drug research meets network science: Where are we?," *J. Med. Chem.*, May 2020.

[48] P. H. Guzzi, D. Mercatelli, C. Ceraolo, and F. M. Giorgi, “Master regulator analysis of the SARS-CoV-2/Human interactome,” *J. Clin. Med. Res.*, vol. 9, Apr. 2020.

[49] Y. Zhou, Y. Hou, J. Shen, Y. Huang, W. Martin, and F. Cheng, “Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2,” *Cell Discov*, vol. 6, p. 14, Mar. 2020.

[50] P. Fagone, R. Ciurleo, S. D. Lombardo, C. Iacobello, C. I. Palermo, Y. Shoenfeld, K. Bendtzen, P. Bramanti, and F. Nicoletti, “Transcriptional landscape of SARS-CoV-2 infection dismantles pathogenic pathways activated by the virus, proposes unique sex-specific differences and predicts tailored therapeutic strategies,” *Autoimmun. Rev.*, p. 102571, May 2020.

[51] D. M. Gysi, Í. do Valle, M. Zitnik, A. Ameli, X. Gan, O. Varol, S. D. Ghiassian, J. J. Patten, R. A. Davey, J. Loscalzo, and A.-L. Barabási, “Network medicine framework for identifying drug-repurposing opportunities for COVID-19,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 118, May 2021.

[52] Y. Tanaka, Y. Tamada, M. Ikeguchi, F. Yamashita, and Y. Okuno, “System-Based differential gene network analysis for characterizing a Sample-Specific subnetwork,” *Biomolecules*, vol. 10, Feb. 2020.

[53] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao, L. Guo, J. Xie, G. Wang, R. Jiang, Z. Gao, Q. Jin, J. Wang, and B. Cao, “Clinical features of patients infected with 2019 novel coronavirus in wuhan, china,” *Lancet*, vol. 395, pp. 497–506, Feb. 2020.

[54] W.-J. Guan, Z.-Y. Ni, Y. Hu, W.-H. Liang, C.-Q. Ou, J.-X. He, L. Liu, H. Shan, C.-L. Lei, D. S. C. Hui, B. Du, L.-J. Li, G. Zeng, K.-Y. Yuen, R.-C. Chen, C.-L. Tang, T. Wang, P.-Y. Chen, J. Xiang, S.-Y. Li, J.-L. Wang, Z.-J. Liang, Y.-X. Peng, L. Wei, Y. Liu, Y.-H. Hu, P. Peng, J.-M. Wang, J.-Y. Liu, Z. Chen, G. Li, Z.-J. Zheng, S.-Q. Qiu, J. Luo, C.-J. Ye, S.-Y. Zhu, N.-S. Zhong, and China Medical Treatment Expert Group for Covid-19, “Clinical characteristics of coronavirus disease 2019 in china,” *N. Engl. J. Med.*, vol. 382, pp. 1708–1720, Apr. 2020.

[55] P. Mehta, D. F. McAuley, M. Brown, E. Sanchez, R. S. Tattersall, J. J. Manson, and HLH Across Speciality Collaboration, UK, “COVID-19: consider cytokine storm syndromes and immunosuppression,” *Lancet*, vol. 395, pp. 1033–1034, Mar. 2020.

[56] A.-L. Barabási and Z. N. Oltvai, “Network biology: understanding the cell’s functional organization,” *Nat. Rev. Genet.*, vol. 5, pp. 101–113, Feb. 2004.

[57] E. V. Mesev, R. A. LeDesma, and A. Ploss, “Decoding type I and III interferon signalling during viral infection,” *Nat Microbiol*, vol. 4, pp. 914–924, June 2019.

[58] X. Qian, C. Xu, P. Zhao, and Z. Qi, “Long non-coding RNA GAS5 inhibited hepatitis C virus replication by binding viral NS3 protein,” *Virology*, vol. 492, pp. 155–165, May 2016.

[59] RECOVERY Collaborative Group, P. Horby, W. S. Lim, J. R. Emberson, M. Mafham, J. L. Bell, L. Linsell, N. Staplin, C. Brightling, A. Ustianowski, E. Elmahi, B. Prudon, C. Green, T. Felton, D. Chadwick, K. Rege, C. Fegan, L. C. Chappell, S. N. Faust, T. Jaki, K. Jeffery, A. Montgomery, K. Rowan, E. Juszczak, J. K. Baillie, R. Haynes, and M. J. Landray, “Dexamethasone in hospitalized patients with covid-19 - preliminary report,” *N. Engl. J. Med.*, vol. 384, pp. 693–704, Feb. 2021.

[60] A. Meager, “Cytokine regulation of cellular adhesion molecule expression in inflammation,” *Cytokine Growth Factor Rev.*, vol. 10, pp. 27–39, Mar. 1999.

[61] L. Velazquez-Salinas, A. Verdugo-Rodriguez, L. L. Rodriguez, and M. V. Borca, “The role of interleukin 6 during viral infections,” *Front. Microbiol.*, vol. 10, p. 1057, May 2019.

[62] B. Fu, X. Xu, and H. Wei, “Why tocilizumab could be an effective treatment for severe COVID-19?,” *J. Transl. Med.*, vol. 18, p. 164, Apr. 2020.

[63] X. Xu, M. Han, T. Li, W. Sun, D. Wang, B. Fu, Y. Zhou, X. Zheng, Y. Yang, X. Li, X. Zhang, A. Pan, and H. Wei, “Effective treatment of severe COVID-19 patients with tocilizumab,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, pp. 10970–10975, May 2020.

[64] Y. Zhou, B. Fu, X. Zheng, D. Wang, C. Zhao, Y. Qi, R. Sun, Z. Tian, X. Xu, and H. Wei, “Pathogenic t-cells and inflammatory monocytes incite inflammatory storms in severe COVID-19 patients,” *Natl Sci Rev*, vol. 7, pp. 998–1002, June 2020.

[65] F. M. Lang, K. M.-C. Lee, J. R. Teijaro, B. Becher, and J. A. Hamilton, “GM-CSF-based treatments in COVID-19: reconciling opposing therapeutic approaches,” *Nat. Rev. Immunol.*, June 2020.

[66] M. Levi, J. Thachil, T. Iba, and J. H. Levy, “Coagulation abnormalities and thrombosis in patients with COVID-19,” *Lancet Haematol.*, vol. 7, pp. e438–e440, June 2020.

[67] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, M. A. Müller, C. Drosten, and S. Pöhlmann, "SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor," *Cell*, vol. 181, pp. 271–280.e8, Apr. 2020.

[68] C. G. K. Ziegler, S. J. Allon, S. K. Nyquist, I. M. Mbano, V. N. Miao, C. N. Tzouanas, Y. Cao, A. S. Yousif, J. Bals, B. M. Hauser, J. Feldman, C. Muus, M. H. Wadsworth, 2nd, S. W. Kazer, T. K. Hughes, B. Doran, G. J. Gatter, M. Vukovic, F. Taliaferro, B. E. Mead, Z. Guo, J. P. Wang, D. Gras, M. Plaisant, M. Ansari, I. Angelidis, H. Adler, J. M. S. Sucre, C. J. Taylor, B. Lin, A. Waghray, V. Mitsialis, D. F. Dwyer, K. M. Buchheit, J. A. Boyce, N. A. Barrett, T. M. Laidlaw, S. L. Carroll, L. Colonna, V. Tkachev, C. W. Peterson, A. Yu, H. B. Zheng, H. P. Gideon, C. G. Winchell, P. L. Lin, C. D. Bingle, S. B. Snapper, J. A. Kropski, F. J. Theis, H. B. Schiller, L.-E. Zaragosi, P. Barbry, A. Leslie, H.-P. Kiem, J. L. Flynn, S. M. Fortune, B. Berger, R. W. Finberg, L. S. Kean, M. Garber, A. G. Schmidt, D. Lingwood, A. K. Shalek, J. Ordovas-Montanes, and HCA Lung Biological Network, "SARS-CoV-2 receptor ACE2 is an Interferon-Stimulated gene in human airway epithelial cells and is detected in specific cell subsets across tissues," *Cell*, vol. 181, pp. 1016–1035.e19, May 2020.

[69] X. Liu, X. Liu, M. Li, Y. Zhang, W. Chen, M. Zhang, C. Zhang, and M. Zhang, "Mechanical stretch induces smooth muscle cell dysfunction by regulating ACE2 via P38/ATF3 and post-transcriptional regulation by mir-421," *Front. Physiol.*, vol. 11, p. 540591, 2021.

[70] R. J. Andrade, N. Chalasani, E. S. Björnsson, A. Suzuki, G. A. Kullak-Ublick, P. B. Watkins, H. Devarbhavi, M. Merz, M. I. Lucena, N. Kaplowitz, and G. P. Aithal, "Drug-induced liver injury," *Nature Reviews Disease Primers*, vol. 5, pp. 1–22, Aug. 2019.

[71] J. H. Hoofnagle and E. S. Björnsson, "Drug-Induced liver injury — types and phenotypes," *N. Engl. J. Med.*, vol. 381, pp. 264–273, July 2019.

[72] R. Teschke, "Idiosyncratic DILI: Analysis of 46,266 cases assessed for causality by RUCAM and published from 2014 to early 2019," *Front. Pharmacol.*, vol. 10, p. 730, July 2019.

[73] S. Babai, L. Auclert, and H. Le-Louët, "Safety data and withdrawal of hepatotoxic drugs," *Therapie*, vol. 76, pp. 715–723, Nov. 2021.

[74] R. J. Weaver, E. A. Blomme, A. E. Chadwick, I. M. Copple, H. H. J. Gerets, C. E. Goldring, A. Guillouzo, P. G. Hewitt, M. Ingelman-Sundberg, K. G. Jensen, S. Juhila, U. Klingmüller, G. Labbe, M. J. Liguori, C. A. Lovatt, P. Morgan, D. J. Naisbitt, R. H. H. Pieters, J. Snoeys, B. van de Water, D. P. Williams, and B. K. Park, "Managing the challenge of drug-induced liver injury: a roadmap for the development and deployment of preclinical predictive models," *Nat. Rev. Drug Discov.*, vol. 19, pp. 131–148, Nov. 2019.

[75] L. Meunier and D. Larrey, "Drug-Induced liver injury: Biomarkers, requirements, candidates, and validation," *Front. Pharmacol.*, vol. 10, p. 1482, Dec. 2019.

[76] S. Fu, D. Wu, W. Jiang, J. Li, J. Long, C. Jia, and T. Zhou, "Molecular biomarkers in Drug-Induced liver injury: Challenges and future perspectives," *Front. Pharmacol.*, vol. 10, p. 1667, 2019.

[77] C. Hardt, M. E. Beber, A. Rasche, A. Kamburov, D. G. Hebels, J. C. Kleinjans, and R. Herwig, "ToxDB: pathway-level interpretation of drug-treatment data," *Database*, vol. 2016, Apr. 2016.

[78] Y. Igarashi, N. Nakatsu, T. Yamashita, A. Ono, Y. Ohno, T. Urushidani, and H. Yamada, "Open TG-GATEs: a large-scale toxicogenomics database," *Nucleic Acids Res.*, vol. 43, pp. D921–7, Jan. 2015.

[79] B. Ganter, S. Tugendreich, C. I. Pearson, E. Ayanoglu, S. Baumhueter, K. A. Bostian, L. Brady, L. J. Browne, J. T. Calvin, G.-J. Day, N. Breckenridge, S. Dunlea, B. P. Eynon, L. M. Furness, J. Ferng, M. R. Fielden, S. Y. Fujimoto, L. Gong, C. Hu, R. Idury, M. S. B. Judo, K. L. Kolaja, M. D. Lee, C. McSorley, J. M. Minor, R. V. Nair, G. Natsoulis, P. Nguyen, S. M. Nicholson, H. Pham, A. H. Roter, D. Sun, S. Tan, S. Thode, A. M. Tolley, A. Vladimirova, J. Yang, Z. Zhou, and K. Jarnagin, "Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action," *J. Biotechnol.*, vol. 119, pp. 219–244, Sept. 2005.

[80] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The SIDER database of drugs and side effects," *Nucleic Acids Res.*, vol. 44, pp. D1075–9, Jan. 2016.

[81] M. Chen, V. Vijay, Q. Shi, Z. Liu, H. Fang, and W. Tong, "FDA-approved drug labeling for the study of drug-induced liver injury," *Drug Discov. Today*, vol. 16, pp. 697–703, Aug. 2011.

[82] A. Liu, M. Walter, P. Wright, A. Bartosik, D. Dolciami, A. Elbasir, H. Yang, and A. Bender, "Prediction and mechanistic analysis of drug-induced liver injury (DILI) based on chemical structure," *Biol. Direct*, vol. 16, p. 6, Jan. 2021.

[83] M. Koido, E. Kawakami, J. Fukumura, Y. Noguchi, M. Ohori, Y. Nio, P. Nicoletti, G. P. Aithal, A. K. Daly, P. B. Watkins, H. Anayama, Y. Dragan, T. Shinozawa, and T. Takebe, "Polygenic architecture informs potential vulnerability to drug-induced liver injury," *Nat. Med.*, vol. 26, pp. 1541–1548, Sept. 2020.

[84] J. D. Zhang, N. Berntenis, A. Roth, and M. Ebeling, "Data mining reveals a network of early-response genes as a consensus signature of drug-induced in vitro and in vivo toxicity," *Pharmacogenomics J.*, vol. 14, pp. 208–216, June 2014.

[85] P. Kohonen, J. A. Parkkinen, E. L. Willighagen, R. Ceder, K. Wennerberg, S. Kaski, and R. C. Grafström, "A transcriptomics data-driven gene space accurately predicts liver cytopathology and drug-induced liver injury," *Nat. Commun.*, vol. 8, pp. 1–15, July 2017.

[86] A. Roth, F. Boess, C. Landes, G. Steiner, C. Freichel, J.-M. Plancher, S. Raab, C. de Vera Mudry, T. Weiser, and L. Suter, "Gene expression-based in vivo and in vitro prediction of liver toxicity allows compound selection at an early stage of drug development," *J. Biochem. Mol. Toxicol.*, vol. 25, pp. 183–194, May 2011.

[87] J. Jiang, C. D. Pieterman, G. Ertaylan, R. L. M. Peeters, and T. M. C. M. de Kok, "The application of omics-based human liver platforms for investigating the mechanism of drug-induced hepatotoxicity in vitro," *Arch. Toxicol.*, vol. 93, pp. 3067–3098, Nov. 2019.

[88] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, p. 559, Dec. 2008.

[89] G. Callegaro, S. J. Kunnen, P. Trairatphisan, S. Grosdidier, M. Niemeijer, W. den Hollander, E. Guney, J. Piñero Gonzalez, L. Furlong, Y. W. Webster, J. Saez-Rodriguez, J. J. Sutherland, J. Mollon, J. L. Stevens, and B. van de Water, "The human hepatocyte TXG-MAPr: gene co-expression network modules to support mechanism-based risk assessment," *Arch. Toxicol.*, vol. 95, pp. 3745–3775, Dec. 2021.

[90] J. J. Sutherland, Y. W. Webster, J. A. Willy, G. H. Searfoss, K. M. Goldstein, A. R. Irizarry, D. G. Hall, and J. L. Stevens, "Toxicogenomic module associations with pathogenesis: a network-based approach to understanding drug toxicity," *Pharmacogenomics J.*, vol. 18, pp. 377–390, May 2018.

[91] I. M. Copple, W. den Hollander, G. Callegaro, F. E. Mutter, J. L. Maggs, A. L. Schofield, L. Rainbow, Y. Fang, J. J. Sutherland, E. C. Ellis, M. Ingelman-Sundberg, S. W. Fenwick, C. E. Goldring, B. van de Water, J. L. Stevens, and B. K. Park, "Characterisation of the NRF2 transcriptional network and its response to chemical insult in primary human hepatocytes: implications for prediction of drug-induced liver injury," *Arch. Toxicol.*, vol. 93, pp. 385–399, Feb. 2019.

[92] P. Trairatphisan, T. M. de Souza, J. Kleinjans, D. Jennen, and J. Saez-Rodriguez, "Contextualization of causal regulatory networks from toxicogenomics data applied to drug-induced liver injury," *Toxicol. Lett.*, vol. 350, pp. 40–51, Oct. 2021.

[93] G. Barel and R. Herwig, "Network and pathway analysis of toxicogenomics data," *Front. Genet.*, vol. 9, p. 484, Oct. 2018.

[94] B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, F. Loney, B. May, M. Milacic, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio, "The reactome pathway knowledgebase," *Nucleic Acids Res.*, vol. 48, pp. D498–D503, Jan. 2020.

[95] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, pp. 27–30, Jan. 2000.

[96] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," *arXiv preprint arXiv:1709.05584*, Sept. 2017.

[97] Y. Tanaka, K. Higashihara, M. A. Nakazawa, F. Yamashita, Y. Tamada, and Y. Okuno, "Dynamic changes in gene-to-gene regulatory networks in response to SARS-CoV-2 infection," *Sci. Rep.*, vol. 11, p. 11241, May 2021.

[98] D. P. Kingma and M. Welling, "Auto-Encoding variational bayes," *arXiv preprint arXiv:1312.6114*, Dec. 2013.

[99] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," *arXiv preprint arXiv:1810.00826*, Oct. 2018.

[100] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework." Nov. 2016.

[101] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural*

Information Processing Systems (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds.), vol. 26, Curran Associates, Inc., 2013.

[102] T. N. Kipf and M. Welling, “Variational graph Auto-Encoders,” *arXiv preprint arXiv:1611.07308*, Nov. 2016.

[103] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 448–456, PMLR, 2015.

[104] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, Dec. 2014.

[105] M. Fey and J. E. Lenssen, “Fast graph representation learning with PyTorch geometric,” *arXiv preprint arXiv:1903.02428*, Mar. 2019.

[106] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, “affy—analysis of affymetrix GeneChip data at the probe level,” *Bioinformatics*, vol. 20, pp. 307–315, Feb. 2004.

[107] A. L. Barabasi and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, pp. 509–512, Oct. 1999.

[108] D. J. Watts and S. H. Strogatz, “Collective dynamics of ’small-world’ networks,” *Nature*, vol. 393, pp. 440–442, June 1998.

[109] P. Erdős and A. Rényi, “On the evolution of random graphs,” *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, no. 1, pp. 17–60, 1960.

[110] A. Hagberg, P. Swart, and D. S Chult, “Exploring network structure, dynamics, and function using NetworkX,” tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[111] L. v. d. Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[112] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, “clusterprofiler: an R package for comparing biological themes among gene clusters,” *OMICS*, vol. 16, pp. 284–287, May 2012.

[113] G. Yu and Q.-Y. He, “ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization,” *Mol. Biosyst.*, vol. 12, pp. 477–479, Feb. 2016.

[114] M. A. Nakazawa, Y. Tamada, Y. Tanaka, M. Ikeguchi, K. Higashihara, and Y. Okuno, “Novel cancer subtyping method based on patient-specific gene regulatory network,” *Sci. Rep.*, vol. 11, pp. 1–11, Dec. 2021.

[115] K. Wang, “Molecular mechanisms of hepatic apoptosis,” *Cell Death Dis.*, vol. 5, p. e996, Jan. 2014.

[116] C. Xu, C. Y.-T. Li, and A.-N. T. Kong, “Induction of phase i, II and III drug metabolism/transport by xenobiotics,” *Arch. Pharm. Res.*, vol. 28, pp. 249–268, Mar. 2005.

[117] M. Chen, A. Suzuki, S. Thakkar, K. Yu, C. Hu, and W. Tong, “DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans,” *Drug Discov. Today*, vol. 21, pp. 648–653, Apr. 2016.

[118] G. Muzio, L. O’Bray, and K. Borgwardt, “Biological network analysis with deep learning,” *Brief. Bioinform.*, Nov. 2020.

[119] L. Seninge, I. Anastopoulos, H. Ding, and J. Stuart, “VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics,” *Nat. Commun.*, vol. 12, pp. 1–9, Sept. 2021.

[120] N. Kaplowitz, “Idiosyncratic drug hepatotoxicity,” *Nat. Rev. Drug Discov.*, vol. 4, pp. 489–499, June 2005.

[121] P. A. Walker, S. Ryder, A. Lavado, C. Dilworth, and R. J. Riley, “The evolution of strategies to minimise the risk of human drug-induced liver injury (DILI) in drug discovery and development,” *Arch. Toxicol.*, vol. 94, pp. 2559–2585, Aug. 2020.

[122] V. M. Lauschke, “Toxicogenomics of drug induced liver injury - from mechanistic understanding to early prediction,” *Drug Metab. Rev.*, vol. 53, pp. 245–252, May 2021.