Supplementary Material for SUBSTRA: Supervised Bayesian Patient Stratification

Sahand Khakabimamaghani, Yogeshwar D. Kelkar, Bruno M. Grande, Ryan D. Morin, Martin Ester, Daniel Ziemek

A Simulation Studies: Settings and Results

For each type of relationship (*i.e.*, AND, OR, and XOR) the parameters used for generating the data are shown in Tables A1 to A3. The same parameters are used for simulating different noise levels. Figures A1 to A3 show the heatmaps for different relationship types and noise levels for SUBSTRA and B2PS, corresponding to the quantitative results shown in the main text in Table 2.

PC# (Size)	TC# (Size)			Phenotype
	A (10)	B (10)	Noise (1980, 380, 180)	
1 (30)	0.7	0.7	0.5	1
2 (30)	0.7	0.23	0.5	0
3 (20)	0.1	0.8	0.5	0
4 (20)	0.3	0.3	0.5	0

Table A1: Parameters and cluster sizes for AND data

PC# (Size)	TC# (Size)			Phenotype
	A (10)	B (10)	Noise (1980, 380, 180)	
1 (30)	0.17	0.7	0.5	1
2 (30)	0.76	0.17	0.5	1
3 (20)	0.7	0.8	0.5	1
4 (20)	0.3	0.3	0.5	0

Table A2: Parameters and cluster sizes for OR data

PC# (Size)	TC# (Size)			Phenotype
	A (10)	B (10)	Noise (1980, 380, 180)	
1 (40)	0.7	0.25	0.5	1
2 (20)	0.1	1.0	0.5	1
3 (30)	0.33	0.35	0.5	0
4 (10)	1.0	0.95	0.5	0

Table A3: Parameters and cluster sizes for XOR data

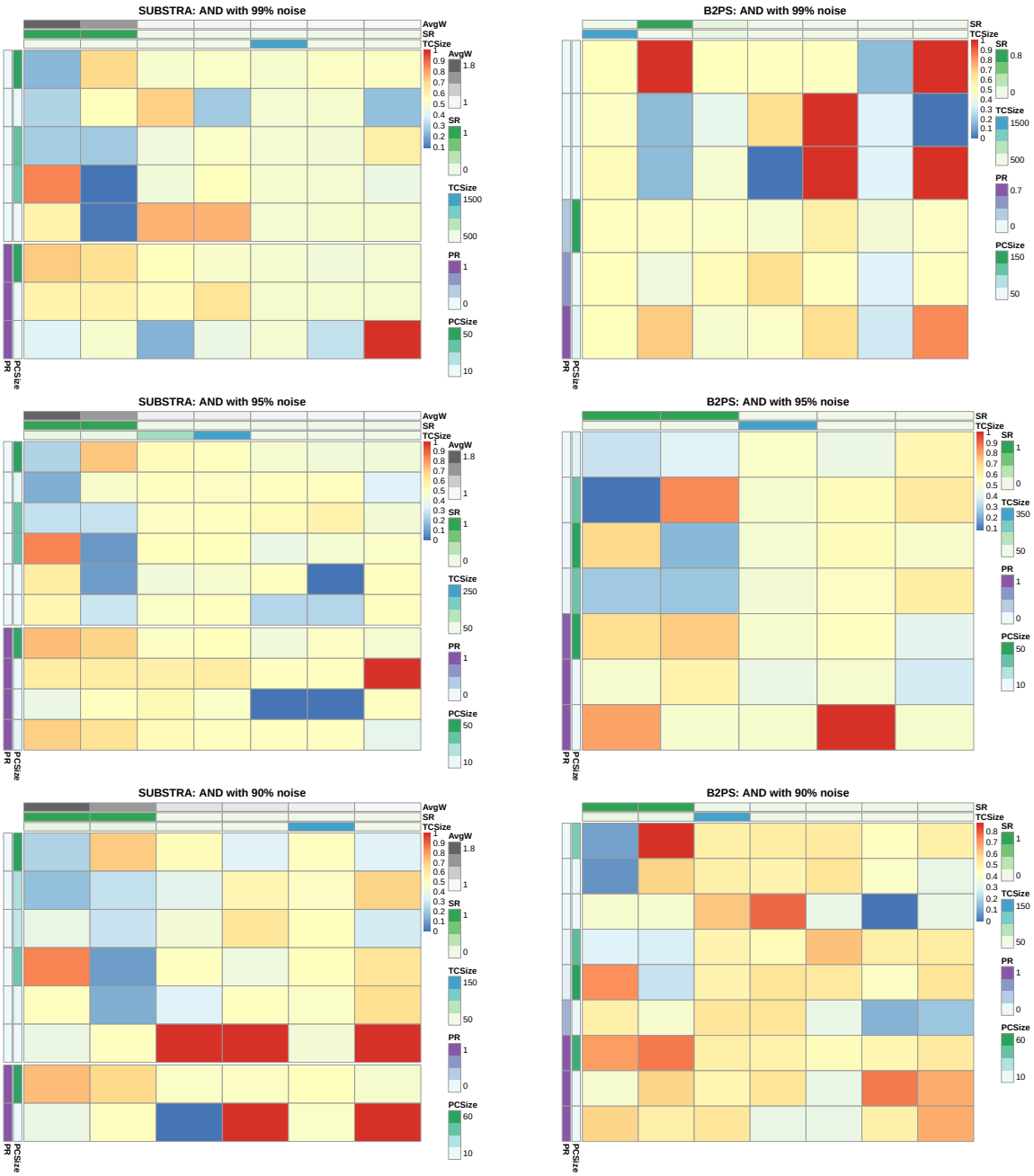

Figure A1: Simulation Results for AND

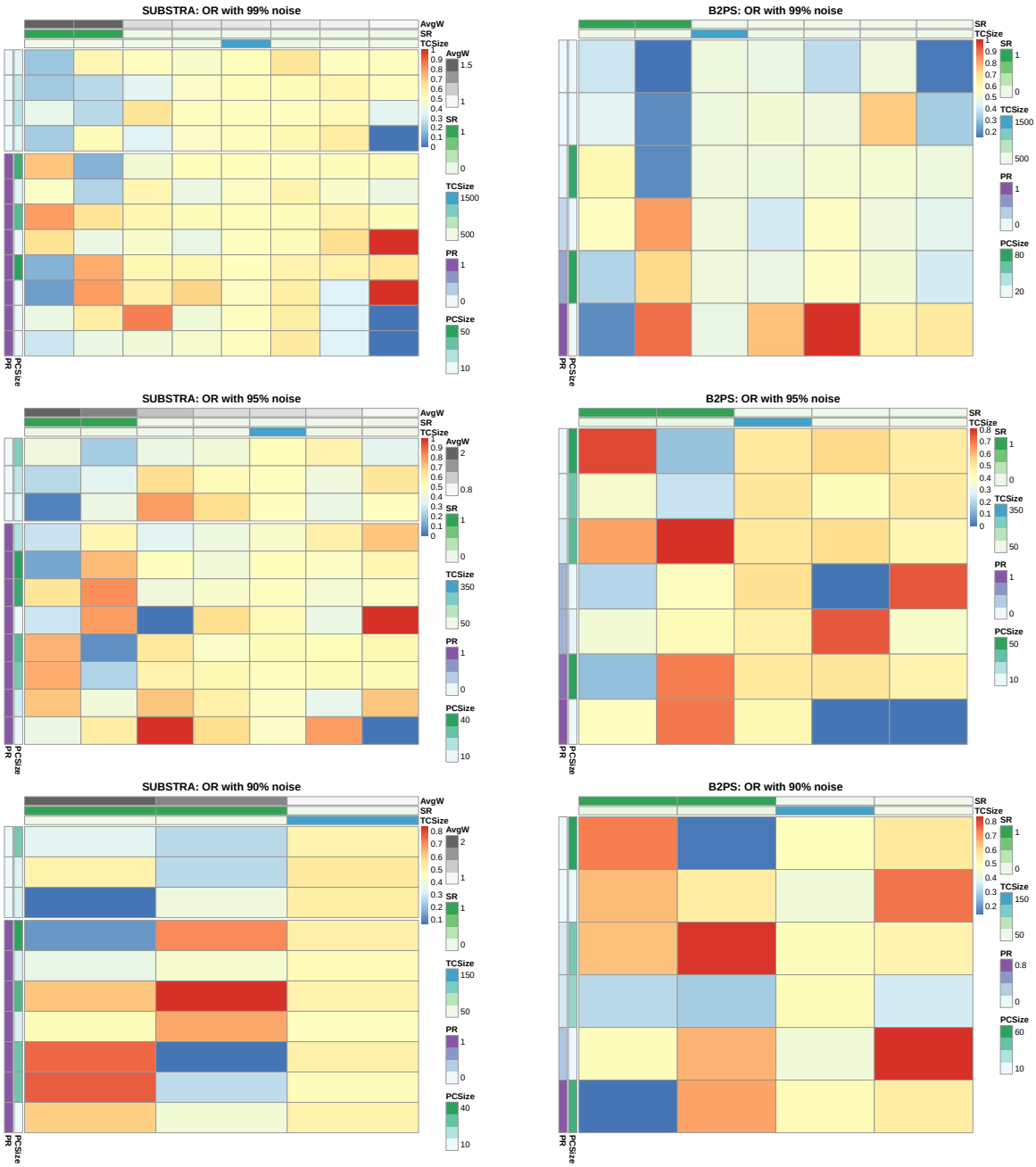

Figure A2: Simulation Results for OR

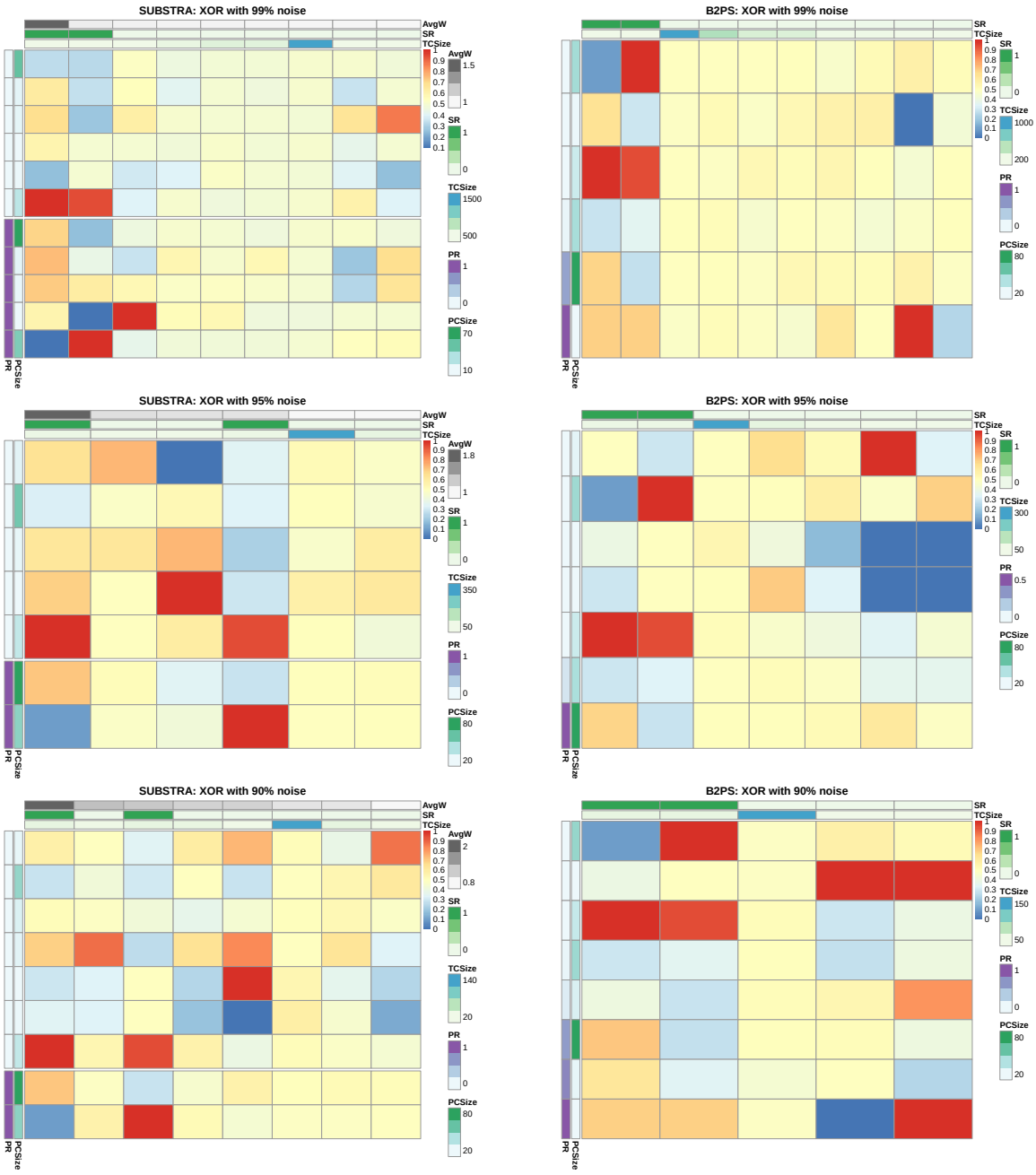

Figure A3: Simulation Results for XOR

B Results of the Experiments on Biological Data

SUBSTRA detects transcript clusters that define patient subtypes. Sorting clusters by the average of the transcript weights gives an indication of their relevance to the phenotype under consideration. We further analyzed the top 5 transcript clusters that SUBSTRA identified for each biological dataset. After mapping the transcripts to the corresponding genes, Gene Ontology (GO) and Pathway (PW) enrichment analysis based on the "tmod" R package [**?**] was performed for these transcript sets. Biological Process (BP) GO terms and KEGG pathways from the Molecu-

lar Signatures Database (MSigDB) [**?**, **?**] are used as the candidate gene modules, and all MSigDB genes are used as the background gene set. Modules with $q\text{-}value < 0.05$ are selected as significantly enriched.

The heatmaps produced by SUBSTRA and the enrichment results produced by R package "tmod" are shown in the following sections. The heatmaps depict the behavior of the transcript clusters (columns) across different patient groups (rows).The value inside each cell/bicluster indicates the average expression (a value between 0 and 1), with red being high expression and blue being low expression. "TCSize" is size of transcript cluster, "AvgW" is average weight of the gene/transcript cluster, "PCSize" is patient cluster size, and "PR" is the Phenotype Ratio (the proportion of '1' phenotypes inside the patient cluster). The transcript clusters are sorted based on the average transcript weight ("AvgW") from left to right in descending order.

In the enrichment plots, rows correspond to GO terms or KEGG pathways (depending on the type of the plot) and columns are the top 5 transcript clusters with the same order as in the heatmaps. The intensity of the colour of the red circles indicate the hyper-geometric test *q-value*. In the following sections we provide the highlights of the descriptive results based on the gene clusters identified in SUBSTRA's output.

B.1 Kidney 1

The 'Kidney 1' dataset was obtained from biopsies extracted more than a year after the kidney transplants einec. The authors of this study developed a classifier for transplant failure versus acceptance, and identified 886 genes whose expression was significantly associated with graft failure. Of the 30 top genes most frequently used by the classifier, five (HAVCR1, ITGB3, LTF, PLK2 and SERPINA3) were clustered in the second top cluster (C2) identified by SUBSTRA. C2 was enriched in pathways associated with cellular death and differentiation, and extracellular matrix organization circulatory system development (see Figures B2 and B3). C1 was broadly enriched in immune signalling pathways, C3 was enriched in transmembrane transport, and C5 in tissue development. SUBSTRA suggests that all these processes are possibly critical to determining the success of allograft.

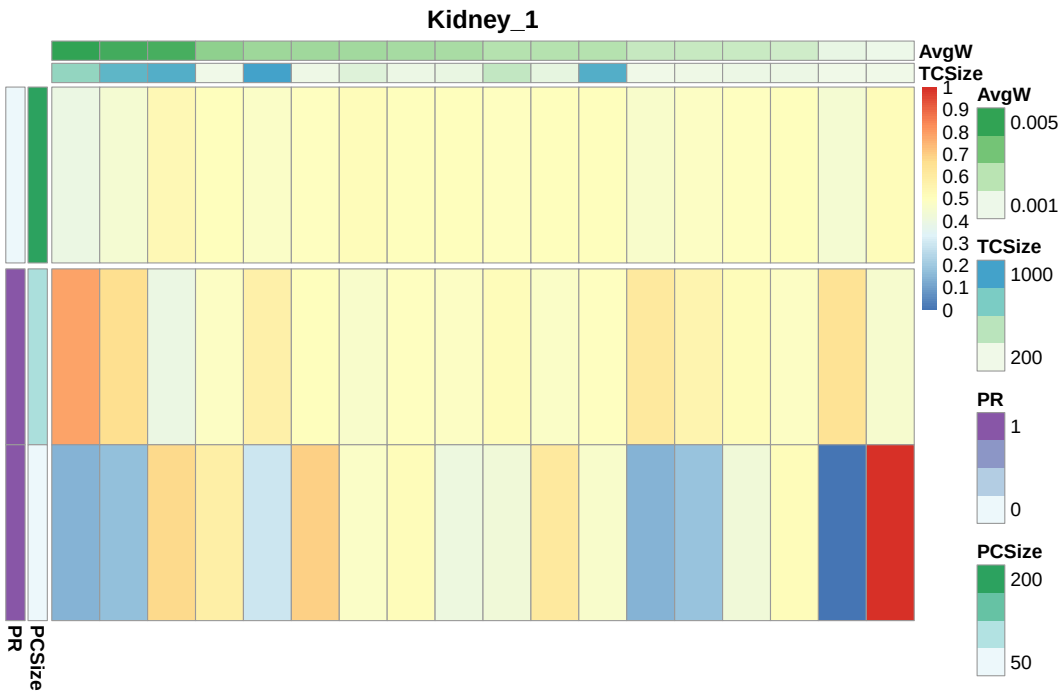

Figure B1: Heatmap for Kidney 1 Dataset

Figure B2: GO Enrichment for Kidney 1 Dataset

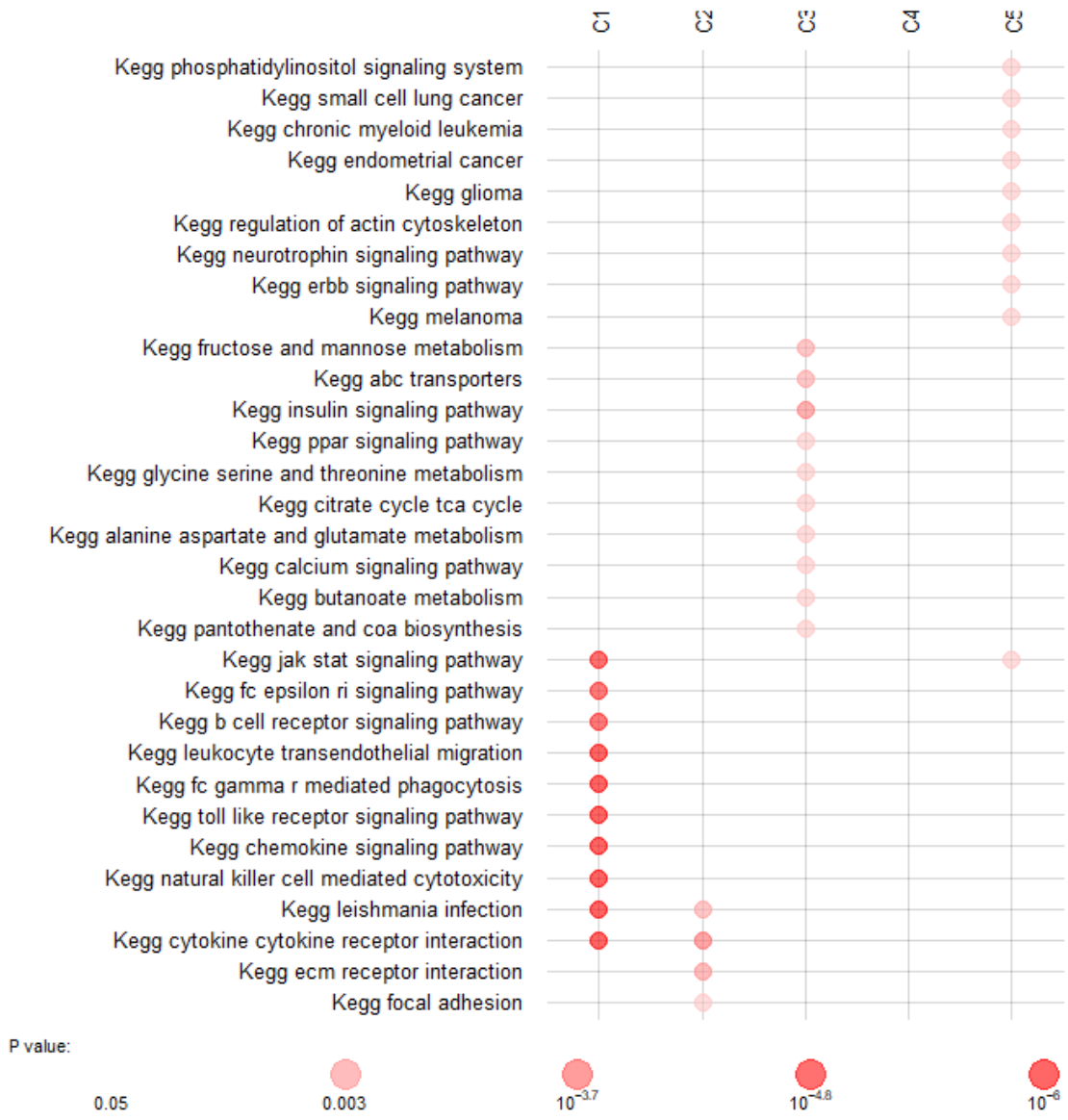

Figure B3: Pathway Enrichment for Kidney 1 Dataset

B.2 Kidney 2

In the study associated with the 'Kidney 2' dataset, khatr identified a 'common rejection module' consisting of 11 genes that were differentially expressed in rejection of transplanted organs : BASP1, CD6, CD7, CXCL9, CXCL10, INPP5D, ISG20, LCK, NKG7, PSMB9, RUNX3 and TAP1. SUBSTRA placed six of these genes – CXCL9, CXCL10, LCK, NKG7, PSMB9 and RUNX3, in the fourth gene cluster, supporting the conclusions of Khatri *et al.*, that these genes form a distinct module that differentiates graft rejection from non-rejection. The second top cluster shows enrichment of 'graft versus host disease', allograft rejection, immune signaling pathways, as well as related pathways such as cell, leukocyte, and lymphocyte activation (see Figures B5 and B6). The remaining gene clusters (except C3) exhibit similar enrichment of immune response pathways.

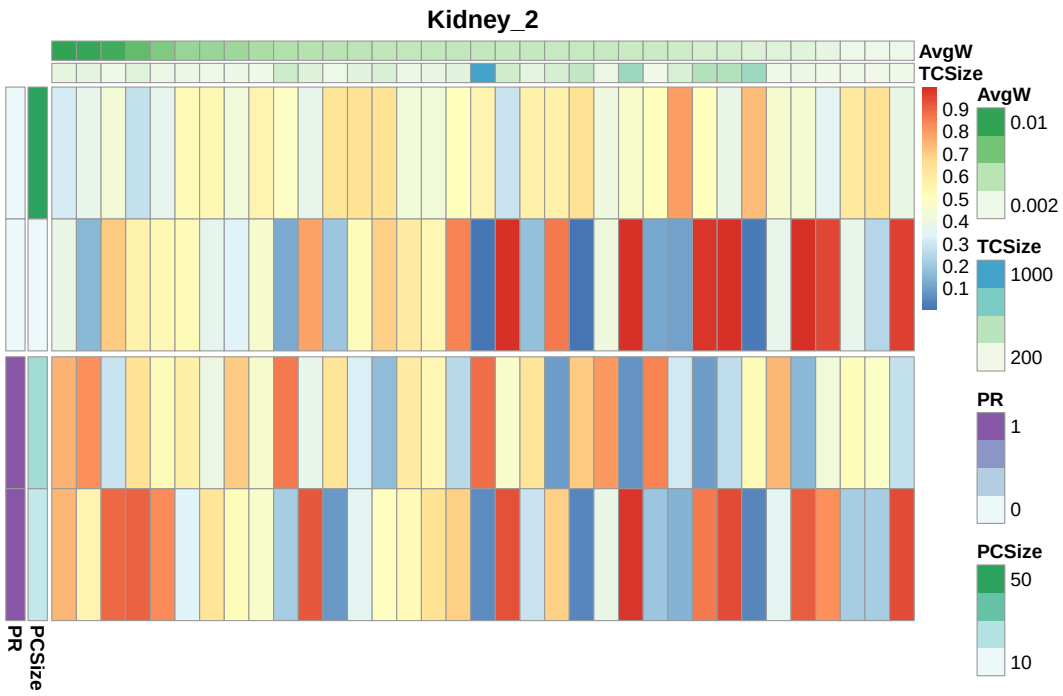

Figure B4: Heatmap for Kidney 2 Dataset

Figure B5: GO Enrichment for Kidney 2 Dataset

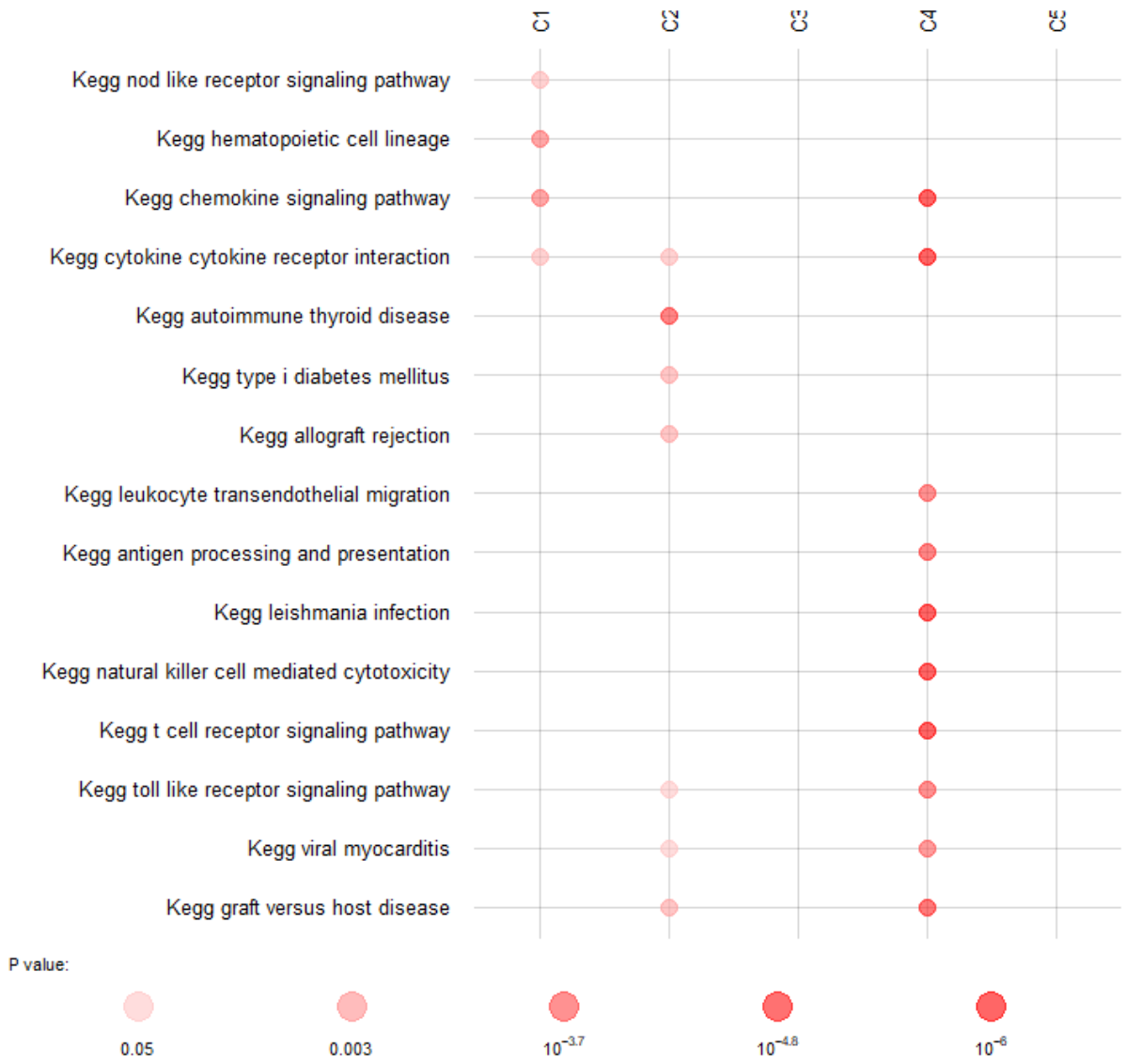

Figure B6: Pathway Enrichment for Kidney 2 Dataset

B.3 Drug Response

'Drug Response' dataset barre contains gene expression information from cancer cell lines treated with AZD6244, also known as selumetinib. Selumetinib's target, MEK, is implicated in the epithelial-mesenchymal transition (EMT), which is an important step in the initiation of metastasis barth. Among many other physiological changes, EMT involves the loss of cell-cell junctions such as tight junctions that are characteristic of epithelial cells. Our method identifies a transcript cluster related to EMT, which is involved in cell-substrate adhesion, as key pathways that respond to selumetinib (see Figure B8).

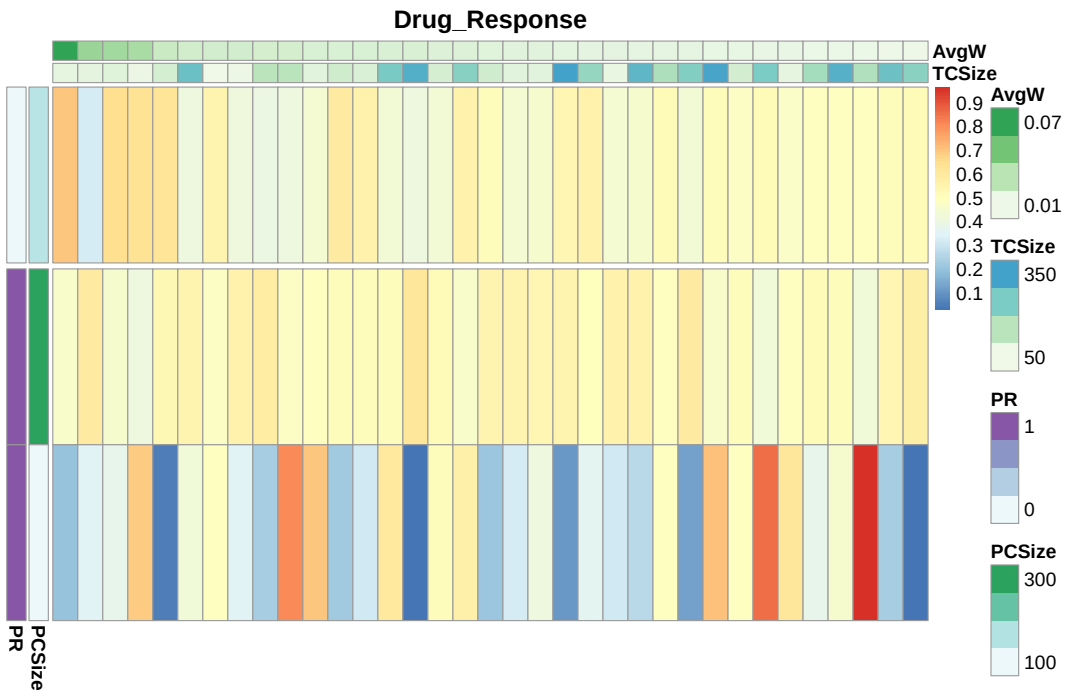

Figure B7: Heatmap for Drug Response Dataset

Figure B8: GO Enrichment for Drug Response Dataset

B.4 Multiple Myeloma

In 'Multiple Myeloma', tian identified DKK1 as an important gene involved in the formation of focal bone lesions. As an inhibitor of the Wnt signaling pathway, DKK1's exact role in modulating this phenotype can be related to any of the pathway's many downstream effects, such as cell fate determination, cell motility, body axis formation, cell proliferation and stem cell renewal komiy. SUBSTRA recapitulated the original analysis by assigning the greatest weight to DKK1 within the third relevant cluster. Interestingly, this cluster also harbors some of the most significantly enriched pathways. Gene set enrichment analysis identified the cell cycle and MAPK signalling as pathways enriched in genes of this cluster (C3 in Figures B10 and B11). This result suggests that DKK1 might be modulating cell proliferation as opposed to other cellular processes associated with the Wnt signaling pathway. Furthermore, previous work has shown an interplay between the Wnt and MAPK signalling pathways in skeletal development zhang. MAPK signalling may be playing an

important role in the formation of osteolytic lesions, a potential discovery that is not described in the original study. This shows that SUBSTRA biclustering and weight assignment can complement other methods such as differential gene expression analysis to provide additional biological context.

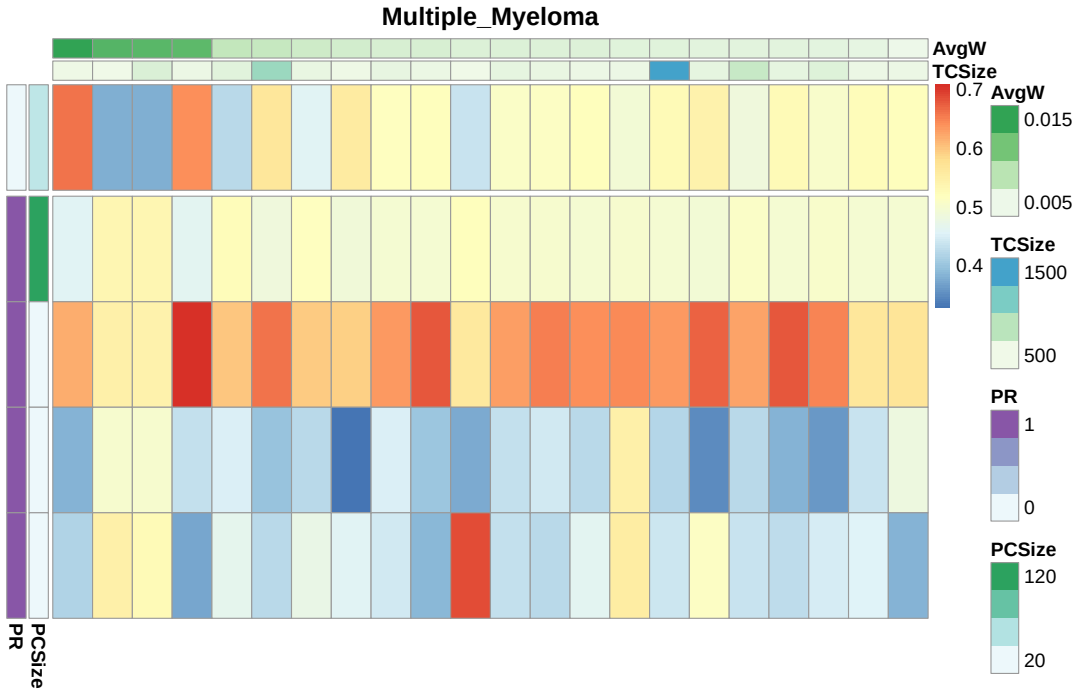

Figure B9: Heatmap for Multiple Myeloma Dataset

Figure B10: GO Enrichment for Multiple Myeloma Dataset

Figure B11: Pathway Enrichment for Multiple Myeloma Dataset

B.5 Lung Cancer

For the 'Lung Cancer' dataset, gordo originally identified eight genes differentially expressed between adenocarcinoma of the lung (ADCA) and malignant pleural mesothelioma (MPM): calretinin (CALB2), VAC-β (ANXA8), TACSTD1 (EPCAM), claudin-7 (CLDN7), TITF-1 (NKX2-1), MRC OX-2 antigen (CD200), PTGIS, and KIAA0977 (COBLL1). SUBSTRA reported all but one gene (CLDN7) in the top 3 transcript clusters, although other claudin genes, namely CLDN3 and CLDN4, were included in the top cluster. Cell and focal adhesion are among the enriched GO terms and KEGG pathways in top 5 transcript clusters (see Figures B13 and B14). These biological processes are consistent with some of the features reported in the original study. Moreover, these top transcript cluster are particularly enriched in several additional pathways that differentiate ADCA and MPM, including extracellular receptor interaction, MAPK signalling, and cytokine receptor

interactions.

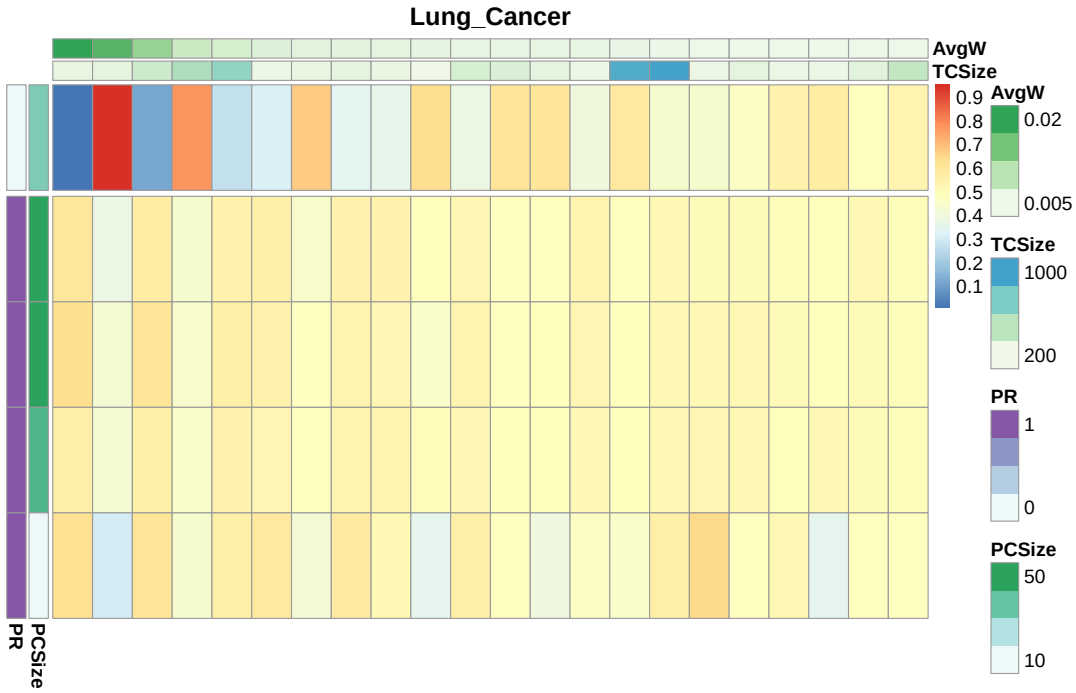

Figure B12: Heatmap for Lung Cancer Dataset

Figure B13: GO Enrichment for Lung Cancer Dataset

Figure B14: Pathway Enrichment for Lung Cancer Dataset

C Details of the Experimental Design

To determine unbiased parameter values, 3x7-fold nested cross-validation (CV) and 5x7-fold nested CV are used for all methods, respectively, for the synthetic and real data predictive analysis. For the descriptive experiments, 3-fold CV is used for tuning the parameters for both of the synthetic and real data. AUC is used as the metric for CV. For SVM, the *radial basis function* kernel is used and the model is tuned through grid search over $\gamma \in \{10^i | -8 \leq i \leq -1\}$ and $C \in \{1..5\}$. For SUBSTRA, the weight magnitude variable μ (see section 2.3.1 in the main manuscript) is tuned over values $\{0.0001, 0.001, 0.01, 0.1, 1, 10\}$. To accommodate for random initialization, the descriptive experiments are repeated 5 times for each dataset and the weights are averaged and the clusters are identified through consensus clustering monti. For HARP, the user should provide

a lower bound for the number of clusters, for which we used the true number of clusters 4 for the AND, OR and XOR datasets and 6 for the UNCLES dataset.

For NMF, we used a method based on mean squared error (MSE) of NMF-based missing value imputations for detecting the rank k of the data matrices. This method is provided in R package NNLM. First, 20% of the matrix entries are set to missing values. Then, using the information from the remaining elements and for different values of $2 \leq k \leq 22$, missing values are imputed based on the latent factors learned by NMF and the MSE is measured. The rank k resulting in the smallest MSE is selected. For the synthetic AND, OR, and XOR datasets, this approach was not successful (i.e., returned $k = 2$ which was meaningless according to the structure of the datasets). Accordingly, we used the true rank $k = 4$ for these experiments.

D Cluster and Class Purity Metrics

Charu Aggarwal defines the cluster and class purities in the "Data Mining" book as follows:

- Purity of clusters: percentage of cluster elements belonging to dominant class.
- Purity of classes: percentage of class elements belonging to dominant cluster.

In our experiments, we refer to the ground-truth clusters as the classes and the method output clusters are considered as the clusters. Because, the true clustering of the noise transcripts is unknown, only the purity of signal transcript clusters and their corresponding clusters in the method outputs is considered. For cluster purity, we identify the method clusters that correspond to the signals by based on the highest purity with regard to the signal classes (i.e., clusters with the highest proportion of either of the signal transcript classes are selected). Then we take the weighted average of the purities of those two clusters using the cluster sizes as the weight. For class purity, we only consider the clustering of transcripts that belong to either of the classes. Similar to cluster purity, we take the weighted average of the two purities (each for one of the two signals) with class sizes as the weights.

E Runtime Analysis

A series of experiments were performed to investigate the effect of the problem size on the execution time. The studied problem size factors included the number of patients/samples, the number of transcripts, the numbers of sample clusters, and the number of transcript clusters. When examining the effect of each of the four factors, the other three factors were kept constant. For each setting of the factors, first a corresponding synthetic dataset was generated. Next, 20 iterations of SUBSTRA consisting of 10 Phase I and 10 Phase II iterations were performed. The procedure was executed 10 times for each setting and the runtimes were averaged.

Figure E1 shows the results. Based on these results, the runtime scales linearly with respect to the number of patients, the number of transcripts, and the number of patient clusters. However, the number of transcript clusters did not have any effects on the runtime. The last observation is due to the fact that the dominant more expensive computations of the algorithm, which involve the feature weights, include only the other three factors.

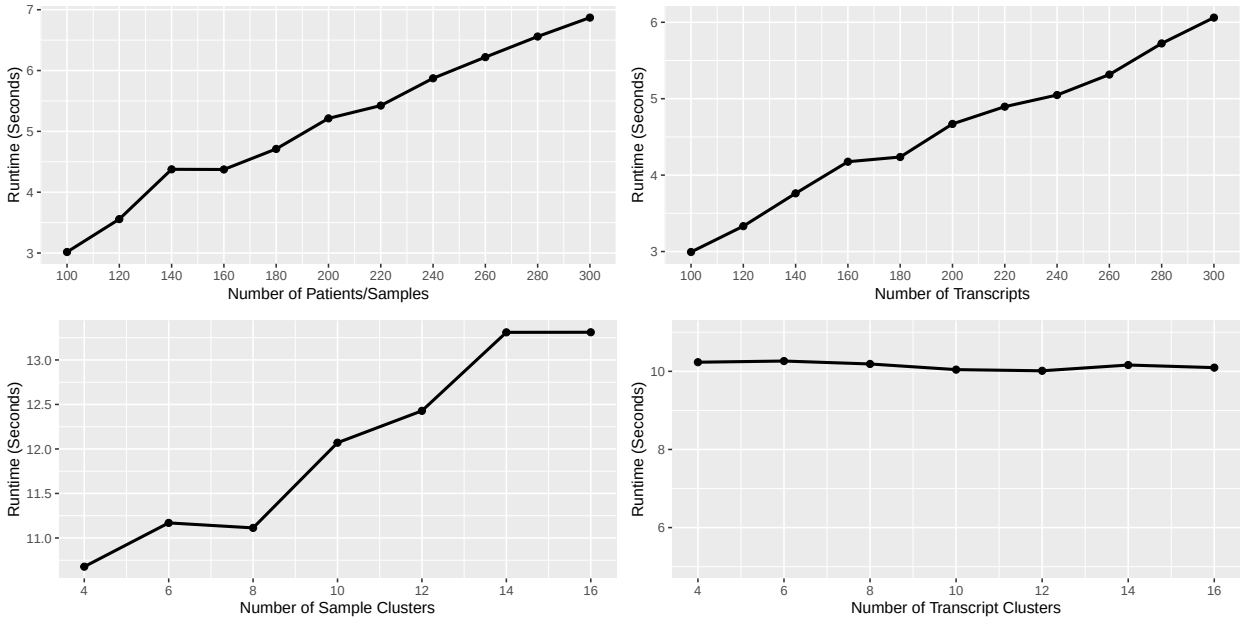

Figure E1: Results of the Runtime Analysis Experiments: For the top left curve, the numbers of sample and transcript clusters were set to 4, the number of transcripts was fixed at 100, and the number of samples varied from 100 to 300. For the top right curve, the numbers of sample and transcript clusters were set to 4, the number of samples was fixed at 100, and the number of transcripts varied from 100 to 300. For the bottom left curve, the numbers of samples and transcripts were fixed at 240, the number of transcript clusters were set to 4, and the number of sample clusters varied between 4 and 16. For the bottom right curve, the numbers of samples and transcripts were fixed at 240, the number of sample clusters were set to 4, and the number of transcript clusters varied between 4 and 16.