iProMix: A decomposition model for studying the function of ACE2 based on bulk proteogenomic data for coronavirus pathogenesis

Xiaoyu Song[1], Jiayi Ji[1], Pei Wang[2]

[1] Tisch Cancer Institute and Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY

[2] Department of Genetics and Genomic Sciences and Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY

Abstract

Both SARS-CoV and SARS-CoV-2 use ACE2 receptors to enter epithelial cells in lung and many other tissues to cause human diseases. Genes and pathways that regulate ACE2 may facilitate/inhibit viral entry and replication, and genes and pathways that are controlled by ACE2 may be perturbed during infection, both affecting disease severity and outcomes. It is critical to understand how genes and pathways are associated with ACE2 in epithelial cells by leveraging proteomic data, but an accurate large-scale proteomic profiling at cellular resolution is not feasible at current stage. Therefore, we propose iProMix, a novel framework that decomposes bulk tissue proteomic data to identify epithelial cell component specific associations between ACE2 and other proteins. Unlike existing decomposition based association analyses, iProMix allows both predictors and outcomes to be impacted by cell type composition of the tissue and accounts for the impacts of decomposition variations and errors on hypothesis tests. It also builds in the functions to improve cell type estimation if estimates from existing literature are unsatisfactory. Simulations demonstrated that iProMix has well-controlled false discovery rate and large power in non-asymptotic settings with both correctly and mis-specified cell-type composition. We applied iProMix to the 110 adjacent normal tissue samples of patients with lung adenocarcinoma from Clinical Proteomic Tumor Analysis Consortium, and identified that interferon α and γ pathways were most significantly associated with ACE2 protein abundances in epithelial cells. Interestingly, the associations were sex-specific that the positive associations were only observed in men, while in women the associations were negative.

1 Introduction

In the past two decades, three types of coronaviruses emerged to cause serious and widespread human illness and death, including SARS-CoV (2002), MERS-CoV (2004) and SARS-CoV-2 (2019). While MERS-CoV used DPP4 receptor to enter human cells, both SARS-CoV and SARS-CoV-2 used the same ACE2 receptor for cell entry [1]. The abundance of ACE2 receptor on the cell surface has been identified as the limiting factor of viral attachment, fusion and entry [2], and thus affects the viral replication rate and disease severity [3]. The upstream genes and pathways that regulate the abundance of ACE2 receptor may facilitate or inhibit the virus infection to host cells, and the downstream genes and pathways that are regulated by ACE2 receptors may be perturbed when the receptor is bounded with the virus and cannot perform its regulatory functions. Therefore, identifying genes and pathways associated with ACE2 receptor paves the way for understanding the viral pathogenesis and the host defense mechanisms to prevent and treat the severe illness. Tools that identify these associations are not only for useful understanding current COVID-19 pandemic but also prepare us for new types of coronavirus potentially emerging in the future.

The ACE2 was primarily expressed in epithelial cells [1, 2, 4]. A recent study that leveraged the single cell RNA sequencing (scRNAseq) data of airway epithelial cells identified the interferon (IFN)α and INFγ pathways may stimulate ACE2 expression levels [4]. This result brought concerns over the complex and contradictory role of IFNs and the general immune system in combating coronaviruses: on one side it induces interferon-stimulated genes to promote host antiviral defense, but on the other side it may also induce ACE2 which makes the virus easier to enter the cells and replicate. Validating the result is critical to generate insights for clinical prescription of IFN-based therapies and development of novel therapeutics. As a result, a validation experiment was performed by perturbing IFN in cell lines, which although resulted in differential expression of ACE2, the effects on ACE2 protein abundances were not identified [4]. Proteins are critical molecules that carry out almost every single cellular function. They are poorly represented by gene expressions as proteins employee complex, multi-level post translational modification that are not reflected at the mRNA level. Studies that quantified both mRNA gene expressions and proteomic abundances identified a median correlation of <0.5 in lung [5] and many other tissues [6, 7]. Therefore, identifying and validating genes and pathways at protein level that are associated with ACE2 is crucial for further understanding the roles of this gene in coronavirus infection.

Unlike scRNAseq, the proteomic profiling technology at the cellular resolution is not mature enough for large-scale studies. Analyses at tissue level have been confounded by other cell types [5]. On the other hand, statistical methods for decomposing bulk -omic data have been intensively studied these years [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24] and shed light on developing statistical methods for studying cell-type specific associations with ACE2. The existing decomposition methods can be broadly categorized into two groups based on their goals. The first group focuses on estimation, such as estimating the cell-type composition [14, 17, 18] and the cell-type specific transcriptomic levels averaged across samples or for each individual sample [25, 20]. The second group focuses on associations by leveraging pre-estimated cell-type composition for cell-type controlled or specific effects. For example, cell-type composition estimates are often treated as covariates in simple association analysis for "unconfounded" effects [6, 7]. Cell-type specific associations are also considered by modeling interactive effects between the estimated cell-type composition and covariates [26] or by considering hierarchical mixture models to decompose the observed tissue-level traits into their values in different cell types for their cell-type specific associations [27]. There are a few difficulties for applying existing methods to our study to identify genes and pathways associated with ACE2 abundances in epithelial cells: First, the protein abundance (or gene expression) levels of both ACE2 and other genes are impacted by the cell type composition. Existing methods decompose the outcomes and regress them on the known predictors. In our setting both predictor (e.g. ACE2) and outcome (e.g. other gene) need to be decomposed. Secondly, existing cell-type specific association analyses assumed the cell-type composition as known and did not account for its estimation biases and variations in hypothesis tests. However, the cell-type compositions are estimated and how the estimation would affect the test statistics is not well studied and accounted, and thus the results may suffer from uncontrolled false positives and reduced study power. Lastly, the estimation for cell-type composition is an ongoing effort that is far from achieving perfection. For example, most well-established methods estimate cell type proportions for a limited number of cell types and epithelial cells are not among the most popular cell types. The xCell [17] is a method that considers 64 cell types including epithelial cells, but it provides only an estimation for relative abundances instead of proportions of a cell type. In other words, it describes how the abundance of epithelial cells in one sample relative to other samples, instead of proportion that tells how the abundance of epithelial cells compared to other cell types within the same sample.

In this study, we propose a new statistical method *iProMix* to analyze the association between ACE2 and other proteins in epithelial cells by leveraging proteogenomic data in bulk tissue. iProMix is building upon a mixture model which treats the observed bulk profiles of both ACE2 and other proteins as a weighted average from epithelial and non-epithelial cell types. It models the strength of cell-specific associations via their conditional joint distributions, and accounts for the uncertainty of decomposition in hypothesis testing. It leverages existing estimation on cell-type composition but builds in an option to allow an updated estimation from the data. We applied iProMix to adjacent normal tissue of lung adenocarcinoma (LUAD) from Clinical Proteomic Tumor Analysis Consortium (CPTAC), and confirmed the role of proteins in IFNα and IFNγ

2

pathways on the ACE2 expression and protein abundance levels in epithelial cells. These associations were not observed by looking at the tissue level data without decomposition. Strikingly, we observed sex-specific effects that were not observed in previous studies.

The article is structured as follows: Section 2 describes the iProMix model that identifies cell specific associations between ACE2 and other proteins in epithelial cells with bulk tissue data. Section 3 presents various simulation results for the performance of our approach in comparison with traditional tissue-level models for error control and study power in correctly and mis-specified models. Section 4 presents the application to CPTAC LUAD data with a focus on IFNα and IFNγ pathways.

2 Method

2.1 Model

In this section, we introduce the proposed iProMix, an *i*ntegrative *Pro*teogenomic model for bulk tissues that are a *Mix*ture of different cell types. Suppose a data set has proteomic profiling of N tissues. In tissue i $(i = 1, ..., N)$, we let x_i be the ACE2 protein abundance (or expression level), y_i be the protein abundance of a different gene, and z_i be a set of covariates such as age. iProMix decomposes the observed protein abundance (or gene expression) levels into their unobserved levels in epithelial-cell component (x_{ui}, y_{ui}) and non-epithelial cell components (x_{vi}, y_{vi}). Similar to many models [24, 27], we assume that the observed tissue-level value is a linear combination of their values in the two major cell types. We have

$$x_i = \pi_i x_{ui} + (1 - \pi_i) x_{vi} \text{ and } y_i = \pi_i y_{ui} + (1 - \pi_i) y_{vi},$$

with π_i denotes the proportion of epithelial cells in the tissue.

As both x_{ui} and y_{ui} can not be directly observed, regression-based framework cannot work for identifying their associations. We propose to leverage a joint modeling idea to identify their conditional correlation given z_i. In details, we model their dependence structures via multivariate Gaussian distribution as follows: $u_i = (x_{ui}, y_{ui})^T \sim N(\mu_{ui}, \Sigma_u)$ and $v_i = (x_{vi}, y_{vi})^T \sim N(\mu_{vi}, \Sigma_v)$, where the mean functions depend on covariates z_i as such $\mu_{ui} = z_i^T \beta_u$ and $\mu_{vi} = z_i^T \beta_v$, and the variance-covariance functions $\Sigma_u = \begin{pmatrix} \sigma_{ux}^2 & \rho_u \sigma_{ux} \sigma_{ux} \\ \rho_u \sigma_{ux} \sigma_{ux} & \sigma_{uy}^2 \end{pmatrix}$ and $\Sigma_v = \begin{pmatrix} \sigma_{vx}^2 & \rho_v \sigma_{vx} \sigma_{vx} \\ \rho_v \sigma_{vx} \sigma_{vx} & \sigma_{vy}^2 \end{pmatrix}$ quantify their dependence structure in the epithelial- and non-epithelial components, respectively. Collectively, $\Theta = (\beta_u, \beta_v, \rho_u, \rho_v, \sigma_{ux}^2, \sigma_{uy}^2, \sigma_{vx}^2, \sigma_{vy}^2)$ and $\pi = \{\pi_1, ..., \pi_n\}$ are the model parameters. The hypothesis test for the association between ACE2 and protein abundance given z_i in epithelial cell type becomes to test $H_0 : \rho_u = 0$ v.s. $H_a : \rho_u \neq 0$.

2.2 Estimation of π

We propose to estimate epithelial cell proportion π separately from Θ, as π is a feature of the tissue that remains the same regardless of the genes, -omic data types and inclusion of other tissues. As a result, we can use -omic data from other technologies (e.g. RNAseq) that are more established and have more prior information than proteomic profiling to improve the estimation accuracy. Among many methods that are available for estimating cell-type composition, most are based on cell-type signature genes, which are not specific to the tissue types and sample conditions (e.g. healthy vs. diseased). Therefore, we employ a hybrid approach to both take advantages of existing knowledge and extract additional signals from the data for cell-type composition estimation. We assume there exists a prior estimate of epithelial cell proportion h_i for tissue i from independent sources, and use *Beta* distribution to link it to the true π_i as done in previous studies [24]. Specifically, we assume $h_i \sim Beta(\alpha_i, \beta_i)$, where $\alpha_i = \pi_i \delta$ and $\beta_i = (1 - \pi_i)\delta$ for some positive parameter δ. By assumption, $E(h_i) = \pi_i$ and $var(h_i) = \pi_i(1 - \pi_i)/(\delta + 1)$ that h_i is an unbiased estimator of the true level of π_i. Suppose we use G individual genes collected on a -omic data type of M tissue samples to estimate $\pi = \{\pi_i\}_{i=1}^M$ and δ. These M sample should include N samples used in downstream analysis but can also include additional samples (e.g. sample with with RNAseq but not proteomic data) to improve

estimation accuracy. We let $W_{M \times G} = \{w_i^g\}$ be the observed profiling of M samples and G genes in the tissue, $S_{M \times G} = \{s_i^g\}$ be the unobserved levels in epithelial cells and $T_{M \times G} = \{t_i^g\}$ be the unobserved levels in non-epithelial cells. Under the assumption that the observed tissue level is as a linear combination of levels in the epithelial- and non-epithelial- cells, we have

$$w_i^g = \pi_i s_i^g + (1 - \pi_i) t_i^g, \quad g = 1, ..., G, i = 1, ..., M.$$

As the primary goal is to estimate $\boldsymbol{\pi}$ instead of capturing the interrelationship among the genes, we consider the marginal distribution of the genes ignoring their correlations to allow a computationally efficient estimation. We have

$$(s_i^1, ..., s_i^G) \sim N(\boldsymbol{\mu}_S, \sigma_S I); (t_i^1, ..., t_i^G) \sim N(\boldsymbol{\mu}_T, \sigma_T I)$$

with I being the $G \times G$ dimensional identity matrix. Assuming independence between $\{h_i\}$ and $\{w_i^g\}$, the estimation process is the solution of the following maximization problem:

$$\max_{\boldsymbol{\Gamma}, \boldsymbol{\pi}, \delta} \sum_{g=1}^{G} \sum_{i=1}^{M} [l(\boldsymbol{\Gamma}, \pi_i \mid w_i^g) + l(\pi_i, \delta \mid h_i)],$$

where $\boldsymbol{\Gamma} = (\boldsymbol{\mu}_S, \boldsymbol{\mu}_T, \sigma_S, \sigma_T)$, δ and $\boldsymbol{\pi}$ are the model parameters. The $l(\boldsymbol{\Gamma}, \boldsymbol{\pi}_i \mid w_i^g)$ and $l(\boldsymbol{\pi}_i, \delta \mid h_i)$ are the log likelihood of the observed gene expression profile and epithelial cell proportion estimate, respectively. Given this likelihood function, we carry out a generalized Expectation-Maximization (EM) algorithm to estimate $\boldsymbol{\pi}$ and δ which can be summarized in the following two steps:

1. E-step: Given the current estimates of the model parameters, i.e. $\boldsymbol{\pi}^{(t)}, \delta^{(t)}, \boldsymbol{\Gamma}^{(t)}$, we calculate $Q^{(t)}(\boldsymbol{\pi}, \delta, \boldsymbol{\Gamma}) = E_{(S|W, \boldsymbol{\pi}^{(t)}, \delta^{(t)}, \boldsymbol{\Gamma}^{(t)})}[l(\boldsymbol{\pi}, \delta, \boldsymbol{\Gamma}; s_i, o_i)]$.

2. M-step: We sample each parameter $\boldsymbol{\pi}, \delta, \boldsymbol{\Gamma}$ conditioning to other parameters via the Expectation Conditional Maximization (ECM) algorithm [28] to calculate $\max_{(\boldsymbol{\pi}^{(t)}, \delta^{(t)}, \boldsymbol{\Gamma}^{(t)})} Q^{(t)}(\boldsymbol{\pi}, \delta, \boldsymbol{\Gamma})$.

As the estimated parameters depend on prior, input genes and samples, we propose bootstrap procedure to use different priors, genes and samples for estimation, and aggregate the multiple estimates for a robust estimation. We propose to calculate a tail truncated mean for aggregation to minimize the impact of outliers. It is worthy noticing that as $E(h_i) = \pi_i$, the averaged value of final estimates $\widehat{\boldsymbol{\pi}}$ depends on the mean of prior, and thus the downstream analysis should be robust against a mis-specification of the averaged prior.

2.3 Estimation of $\boldsymbol{\Theta}$

We propose a likelihood based method to estimate the parameter $\boldsymbol{\Theta}$ in iProMix assuming $\widehat{\boldsymbol{\pi}}$ is known. The complete data log-likelihood function, l, of the observed data $\boldsymbol{o}_i = (x_i, y_i)$ and unobserved data $\boldsymbol{u}_i$ and $\boldsymbol{v}_i = \frac{\boldsymbol{o}_i - \widehat{\pi}_i \boldsymbol{u}_i}{1 - \widehat{\pi}_i}$ with respect to the parameter $\boldsymbol{\Theta}$ can be expressed as follows:

$$l(\boldsymbol{\Theta}; \{\boldsymbol{o}_i\}_{i=1}^n, \{\boldsymbol{u}_i\}_{i=1}^N) = \sum_{i=1}^{n} (-\frac{1}{2} \log |\Sigma_u| - \frac{1}{2} \log |\Sigma_v| - \frac{1}{2} (\boldsymbol{u}_i - \boldsymbol{z}_i^T \boldsymbol{\beta}_u)^T \Sigma_u^{-1} (\boldsymbol{u}_i - \boldsymbol{z}_i^T \boldsymbol{\beta}_u)$$
$$- \frac{1}{2} (\frac{\boldsymbol{o}_i - \widehat{\pi}_i \boldsymbol{u}_i}{1 - \widehat{\pi}_i} - \boldsymbol{z}_i^T \boldsymbol{\beta}_v)^T \Sigma_v^{-1} (\frac{\boldsymbol{o}_i - \widehat{\pi}_i \boldsymbol{u}_i}{1 - \widehat{\pi}_i} - \boldsymbol{z}_i^T \boldsymbol{\beta}_v)^T) + \text{constant}.$$

The identifiability of the parameter $\boldsymbol{\Theta}$ is demonstrated in Supplementary Materials. If $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are observed, maximum likelihood estimates of the parameters $\boldsymbol{\Theta}$ could be directly obtained. The latent (unobserved) nature of $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ requires the adoption of the EM algorithm. Specifically, the EM algorithm summarizes into the following steps:

- E-step: Given the current estimates of the model parameters, i.e. $\boldsymbol{\Theta}^{(t)}$, we calculate $Q^{(t)}(\boldsymbol{\Theta}) = E_{(\boldsymbol{u}|\boldsymbol{o}, \boldsymbol{\Theta}^{(t)})}[l(\boldsymbol{\Theta}; \boldsymbol{u}_i, \boldsymbol{o}_i)]$.

- M-step: We find $\boldsymbol{\Theta}^{(t+1)}$ which is the solution to the following $\max_{\boldsymbol{\Theta}} Q^{(t)}(\boldsymbol{\Theta})$.

In the EM algorithm, a good initialization can lead to faster convergence than random starts. We assumed ρ_u and ρ_v to be zero and used the the mean and variance from regressing $\boldsymbol{o}_i$ on $\boldsymbol{z}_i$ as the initial values.

2.4 Two hypothesis test strategies

Building upon parameter estimation, the goal of iProMix is to identify proteins associated with ACE2 in epithelial cells. If the cell-type level data is observed, the hypothesis test of $\rho_u = 0$ can be solved by likelihood ratio test (LRT), which approximately follows a χ_1^2 distribution under the null hypothesis. iProMix, however, considers a decomposition model that includes unobserved data and leverages quantities such as $\hat{\boldsymbol{\pi}}$ that are estimated with errors and variations. Therefore, in addition to iProMix (χ^2), we consider two empirical distribution based strategies, iProMix (Permutation) and iProMix (Knockoff), to leverage the parallel nature of the genes in -omics data types to identify individual genes with controlled false discovery rate (FDR). This idea borrows information across genes to assist inference about each gene individually, which has been extensively employed in -omics studies, such as in FDR [29] and limma [30].

Suppose we consider a total of P other proteins $\mathbf{y}_i = (y_i^1, ..., y_i^P)$ for their association with ACE2 x_i. In both approaches, we permute the label ACE2 to generate $\tilde{x}_i$ that preserves the distribution of ACE2 but breaks its correlation with $\mathbf{y}_i$'s. The $\tilde{x}_i$ copy behaves in the same way as the original variables but, unlike them, is known to be null. How we handle $\tilde{x}_i$ in the two strategies are, however, different. The results for iProMix (Permutation) and iProMix (Knockoff) are non-asymptotic.

iProMix (Permutation) We calculate the LRT statistics in original and permuted separately, and compare their values for FDR controlled discoveries. In original data for protein p, we estimate the full log likelihood function of (x_i, y_i^p) as l_{full}^p, and the reduced log likelihood function $l_{reduced}^p$ by forcing $\rho_u^p = 0$. The corresponding LRT statistic for ρ_u^p is defined as $LR_x^p = -2(l_{reduced}^p - l_{full}^p)$. Similarly, in permuted data, we model $(\tilde{x}_i, y_i^p)$ and get its LRT statistic $LR_{\tilde{x}}^p$. Across all P proteins, we consider a pre-specified LRT cut-off value c and a protein was counted if its LRT statistics is $> c$. We calculated empirical FDR (eFDR) as

$$T = \min\left\{c : \frac{1 + \#\{p : LR_{\tilde{x}}^p > c\}}{\#\{p : LR_x^p > c\}} \leqslant \alpha\right\}, \tag{1}$$

where α is a pre-determined nominal FDR level. By the definition of false discovery proportion (FDP), we have $FDP(c) = \frac{\#\{p : LR_{\tilde{x}}^p > c\}}{\#\{p : LR_x^p > c\}}$, and incrementing the number of numerator by one the slightly more conservative procedure in Equation (1) controls the FDR.

iProMix (Knockoff) The null distribution for $\rho_u^p = 0$ includes two cases: (1) both $\rho_u^p = 0$ and $\rho_v^p = 0$, and (2) $\rho_u^p = 0$ but $\rho_v^p \neq 0$. While iProMix (Permutation) provides a negative control for ACE2 and protein association in epithelial cells, it also breaks the ACE2 and protein association in non-epithelial cells, and thus its null distribution does not include case (2). To overcome this issue, we consider an alternative strategy iProMix (knockoff) to model the joint distribution of $(x_i, \tilde{x}_i, y_i^p)$ in iProMix framework. Assume the observed data can be decomposed into the epithelial component $\tilde{u}_i^p = (x_{ui}, \tilde{x}_{ui}, y_{ui}^p)$ and non-epithelial component $\tilde{v}_i^p = (x_{vi}, \tilde{x}_{vi}, y_{vi}^p)$, and their cell-type specific dependence can be modeled by multivariate Gaussian distribution similar as described above. Then ρ_{ux}^p and $\rho_{u\tilde{x}}^p$ denote the (x_{ui}, y_{ui}^p) and $(\tilde{x}_{ui}, y_{ui}^p)$ correlation conditional on other variables in the model, respectively. A reduced model for $\rho_{ux}^p = 0$ gives a log-likelihood function $\breve{l}_{reduced,x}^p$ and a reduced model for $\rho_{u\tilde{x}}^p = 0$ gives a log-likelihood function $\breve{l}_{reduced,\tilde{x}}^p$. Then, we define the iProMix (Knockoff) test statistic as the difference of two LRTs that

$$W_p = -2(\breve{l}_{reduced,x}^p - \breve{l}_{reduced,\tilde{x}}^p).$$

When protein p is under the null, $\breve{l}_{reduced,x}^p$ is close to $\breve{l}_{reduced,\tilde{x}}^p$ and W_p is close to 0. $W_p > 0$ indicates that the association of protein p with ACE2 is more important than with its knockoff copy. Hence a large positive value of W_p is an indication that this protein is a genuine signal and associated with ACE2 in the model. Finally, we leverage the empirical distribution of from all P proteins to control the FDR. For a pre-specified cut-off value $c > 0$, protein p was considered positive if its W_p is $> c$. The FDR controlled cutoff of W_p is defined as

$$T = \min \left\{ c : \frac{1 + \#\{p : W_p \leqslant -c\}}{\#\{p : W_p > c\}} \leqslant \alpha \right\}, \tag{2}$$

where α is the pre-determined nominal FDR level. This equation employs the property of W_p that it is symmetric under the null, and for any fixed threshold $c > 0$, we have $\#\{p : W_p \leqslant -c\} = \#\{$null $p : W_p \leqslant -c\} = \#\{$null $p : W_p \geqslant c\}$. By incrementing the negatives by one, the slightly more conservative procedure controls the FDR.

2.5 Pathway enrichment analysis

One can perform pathway enrichment analysis with significant genes, but this analysis is sensitive to the power and error control of gene identification. Alternatively, we propose to compare the relative rank of LRT statistics for all proteins, which is robust against to the distribution assumptions of LRTs that are perturbed by decomposition. For protein p, we define Q^p as the rank of its signed LRT score across all proteins as

$$Q^p = \text{rank}(\text{signed}(LR_x^p)),$$

where the sign comes from the estimated correlation $\hat{\rho}_u^p$ of the data, and the LR_x^p is the likelihood ratio test statistic of contrasting full and reduced models of (x_i, y_i^p). For a pathway, we let $\boldsymbol{Q}^{in}$ indicate the scores for all proteins inside of the pathway, and $\boldsymbol{Q}^{out}$ indicate scores for all proteins outside of the pathway. Then, we compared $\boldsymbol{Q}^{in}$ versus $\boldsymbol{Q}^{out}$ such as using non-parametric Wilcoxon rank score test to identify ACE2 enriched pathways.

Suppose a pathway is not associated with ACE2, then $\boldsymbol{Q}^{in}$ and $\boldsymbol{Q}^{out}$ present similar distributions. If $\boldsymbol{Q}^{in}$ is significantly greater than $\boldsymbol{Q}^{out}$, it indicates this pathway is positively associated with ACE2. Similarly, if $\boldsymbol{Q}^{in}$ is significantly smaller than $\boldsymbol{Q}^{out}$, it indicates this pathway is negatively associated with ACE2. We considered 50 pathways known from Hallmark database [31], including INFα and IFNγ, and considered Benjamini–Yekutieli (BY) procedure for FDR control due to highly overlapped genes across pathways.

3 Simulation results

3.1 Finite sample performance under correctly specified π

In this section, we evaluated the FDR and power of iProMix methods for identifying cell-specific associations under correctly specified cell-type composition $\boldsymbol{\pi}$, and compared the results with methods based on tissue-level data. For iProMix, we considered three hypothesis test strategies: (1) iProMix (χ^2) which uses the theoretical distribution of χ_1^2 for hypothesis test, (2) iProMix (Permutation) which calculates eFDR using null distribution generated by permuting the label of ACE2, and (3) iProMix (Knockoff) which jointly models the distribution of protein, ACE2 and permuted ACE2. For tissue-level analyses, we considered three additional strategies: (1) regressing ACE2 on other proteins ignoring cell-type composition, (2) regressing ACE2 on other proteins adjusting cell-type composition as a covariate, and (3) regressing main and interactive effects of cell-type composition and ACE2 on other proteins.

In data synthesization, we mimic the real data of CPTAC LUAD to generate 100 tissues. For each tissue i, we simulated two major cell types to mimic epithelial and non-epithelial cell components with cell type 1 (epithelial) proportion π_i follows a rescaled Beta distribution as $\pi_i \sim Beta(\alpha = 3, \beta = 2)/1.4 + 0.07$. The rescale guarantees cell type 1 is a major cell type with a minimal 7% and mean 50% cell proportion. For ACE2 synthesization, we let the mean level of ACE2 in cell type 1 to be more abundant than in cell type 2, such that $x_{1i} = 2 + e_{1i}$ and $x_{2i} = -2 + e_{2i}$, where $e_{2i}, e_{2i} \sim N(0, 1)$. When synthesizing protein data for two cell types ($\boldsymbol{y}_{1i}$ and $\boldsymbol{y}_{2i}$), we let the mean and standard deviation of P proteins in cell type 1 and cell type 2 to be randomly generated from $N(0, 1)$ and $\max(0.05, N(1, 0.25))$. As multiple proteins are simultaneously tested, we further considered the correlation of proteins in data generation, assuming a block-wise correlation structure with 50 proteins in each block. The correlation of two proteins j and l within the same block follows a autocorrelation structure of $\rho = 0.6^{|j-l|}$. Finally, as the performance of hypothesis tests for cell type 1

may be impacted by the other cell type, we considered five association patterns across the two cell types to comprehensively evaluate the methods:

1. Association in both cell types with the same direction : $\text{cor}(x_{1i}, y_{1i}) = \text{cor}(x_{2i}, y_{2i}) = 0.5$,

2. Association in cell type 1 only: $\text{cor}(x_{1i}, y_{1i}) = 0.5$, $\text{cor}(x_{2i}, y_{2i}) = 0$,

3. Association in cell type 2 only: $\text{cor}(x_{1i}, y_{1i}) = 0$, $\text{cor}((x_{2i}, y_{2i}) = 0.5$,

4. Association in both cell types with different directions: $\text{cor}(x_{1i}, y_{1i}) = 0.5$, $\text{cor}(x_{2i}, y_{2i}) = -0.5$, and

5. Association in neither cell type: $\text{cor}(x_{1i}, y_{1i}) = \text{cor}(x_2, y_2) = 0$.

We considered a grid of the signal levels to allow the proportion of alternatives ranging 5% to 50% out of a maximum of 3000 simulated proteins. The resulting cell-type specific levels of ACE2 and other proteins cannot be observed, but their weighted averages $x_i = \pi_i x_{1i} + (1-\pi_i) x_{2i}$ for ACE2 and $\boldsymbol{y}_i = \pi_i \boldsymbol{y}_{1i} + (1-\pi_i) \boldsymbol{y}_{2i}$ for protein abundance are observed for analysis.
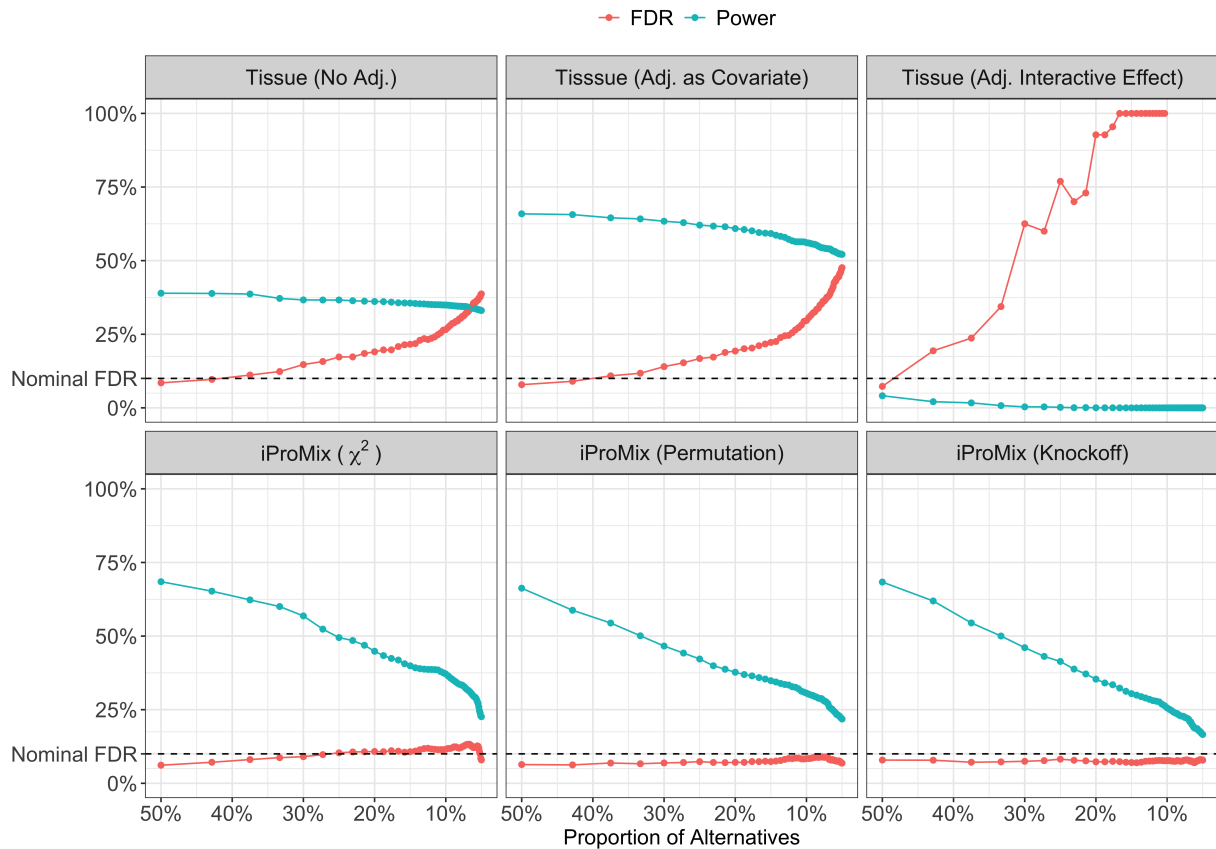

Figure 1: Simulation results for the performance of six comparison methods under correctly specified $\boldsymbol{\pi}$ (cell-type composition) across a wide range of signal levels.

Figure 1 demonstrated the FDR and power of six comparison methods under correctly specified cell-type composition $\boldsymbol{\pi}$ across a wide range of signal levels (alternative proportions). When the signals are very abundant that approximately 50% of proteins are associated with ACE2, all methods hold valid FDRs. As the signals become more and more sparse, the FDRs blow up in all tissue-level methods regardless of

adjustment for cell-type composition or not. The iProMix (χ^2) performs better than tissue-level methods, but still has inflated FDR when the signal level is $<25\%$ of the tests. As the number of proteins associated with ACE2 in real data is believed the be small, iProMix (χ^2) fails to provide valid results. Finally, both iProMix (Permutation) and iProMix (Knockoff) provide valid error control throughout the simulation, and their power are in general comparable.

3.2   Evaluation under misspecification

$\widehat{\pi}$	iProMix (Permutation)		iProMix (Knockoff)	
	FDR	Power	FDR	Power
True π	0.01	0.44	0.07	0.47
Scale change:				
$\pi * 0.8$	0.03	0.32	0.00	0.02
$\pi * 1.2$	0.04	0.63	0.09	0.71
Mean shift (with scale change):				
$\pi - 0.05$	0.02	0.43	0.06	0.39
$\pi + 0.1$	0.03	0.59	0.07	0.59
$\pi + 0.2$	0.05	0.65	0.12	0.74
$\pi * 0.7 + 0.45$	0.08	0.67	0.15	0.75
Reduced correlation:				
$cor(\widehat{\pi}, \pi) = 0.9$	0.04	0.43	0.08	0.38
$cor(\widehat{\pi}, \pi) = 0.6$	0.04	0.25	0.09	0.30
Non-linear transformation:				
$\exp(\pi - 1) - 0.1$	0.03	0.32	0.00	0.02
$\log(\pi + 1) + 0.1$	0.02	0.38	0.02	0.08
More than two cell types:				
$\pi + (1 - \pi) * \pi_{cell3}$	0.05	0.48	0.10	0.58

Table 1: The FDR and power of iProMix (Permutation) and iProMix (Knockoff) methods under mis-specified cell type composition.

As in real data cell-type composition is estimated with errors and noises, we further evaluate the performance of iProMix (Permutation) and iProMix (Knockoff) under various scenarios of the misspecified $\widehat{\pi}$. In this simulation, we consider 1000 proteins with five blocks of equal sizes, and each block corresponds to one of the five association patterns mentioned above. We consider various scenarios where $\widehat{\pi}$ deviates from the true π, such as having scale change, mean shift, reduced correlation, non-linear transformation and including of a third cell type with small cell counts. In details, scale change and mean shift were evaluated by linear transformation of the true π; reduced correlation was achieved by adding random uniformly distributed noises and rescaled to have mean 0.5 and range $(0, 1)$; non-linear transformation included the logarithm and exponential transformation of π; and finally, we generated three cell types by splitting cell type 2 into two components with the newly generated cell type 3 taking 10% of the cell counts, and considered the misspecified estimates as a sum of cell type 1 and 3.

The resulting FDR and power of iProMix (Permutation) and iProMix (Knockoff) under misspecified models are listed in Table 1. We found that while both methods have well-controlled FDR rates and comparable power in correctly specified models, iProMix (Permutation) is robust against various model misspecifications but iProMix (Knockoff) is not. iProMix (Knockoff) has inflated FDRs when $\widehat{\pi}$ is much greater than true π (e.g. mean shift from 0.5 to 0.7) and highly impacted power under non-linear transformation and reduced scale. On the contrary, while iProMix (Permutation) is overconservative at true π, it controls FDR and have reasonable power across all mis-specified scenarios under consideration. Therefore, we will applied iProMix (Permutation) to real data where we do not have a perfect knowledge on the cell type composition.

4 Application to CPTAC LUAD tissue

To identify genes and pathways associated with ACE2 in lung epithelial cells, we applied iProMix to analyze the proteogenomic data sets of 110 normal adjacent tumor (NAT) lung tissue samples from the CPTAC LUAD study [5]. Specifically, the 110 NAT samples were collected from 72 male and 38 female LUAD patients. The transcriptomic and global proteomics data sets contain 18,099 genes and 10,699 proteins respectively. Data prepossessing and quality control were describe in [5]. The gene expression levels and protein abundances of ACE2 in the 110 samples showed no correlation ($p = 0.254$), as illustrated in Figure 2.
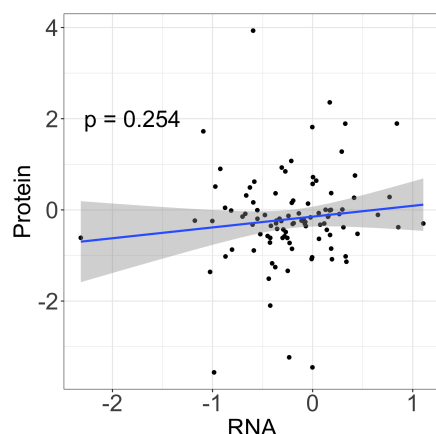
Figure 2: The correlation of the gene expression and protein abundance levels of ACE2 in adjacent normal tissues of lung adenocarcinoma using data from Clinical Proteomic Tumor Analysis Consortium.

4.1 Epithelial cell proportion estimation

Prior for π. First, to obtain a prior estimates of π, we derived epithelial cell scores using xCell [17], a commonly used tool for cell type deconvolution analysis based on RNAseq data. Using cell type specific gene signatures curated from the literature, xCell employs the idea of the single sample gene set enrichment analysis (ssGSEA) scores [32] to capture the relative abundance of each cell type in each sample. The cell type scores from xCell are not direct estimates of cell proportions, and we re-scaled the xCell epithelial scores to have five different mean levels and used re-scaled results as five different sets of priors of π for iProMix analysis. Specifically, the re-scaling was done by (1) fitting a *Beta* distribution to the xCell epithelial scores; (2) altering the parameter α of the *Beta* distribution to generate a new *Beta* distribution with similar shape but different mean, and (3) quantile normalizing the original scores to this new distribution, such that the mean values of the 5 resulting sets of epithelial cell "proportion" vectors evenly spaced from 0.3 to 0.5.

Data-driven estimates. For each prior set, we estimated epithelial cell proportions using 126 xCell epithelial signature genes on random bootstrap samples with the iProMix pipeline. We considered a total of 100 bootstrap samples with each being a random drawn of 80% samples, and calculated a truncated mean as our final estimation, which truncated at 5% of tails in both sides and took an average of the remaining 90% estimates. For a total of five priors with mean 0.3-0.5, we have five estimates $\widehat{\pi}(0.3)$, $\widehat{\pi}(0.35)$, $\widehat{\pi}(0.4)$, $\widehat{\pi}(0.45)$ and $\widehat{\pi}(0.5)$ for downstream analysis, and in later models we aggregate identifications with all priors.

The performance of estimated epithelial cell proportions is presented in Figure 3, and compared to xCell scores (prior information). The correlations of 126 epithelial signature genes with estimated epithelial proportions are higher than the correlation with xCell scores, with a median correlation improved from 0.246 in xCell to $0.292 - 0.296$ in iProMix. A total of 82% epithelial signatures were positively correlated with
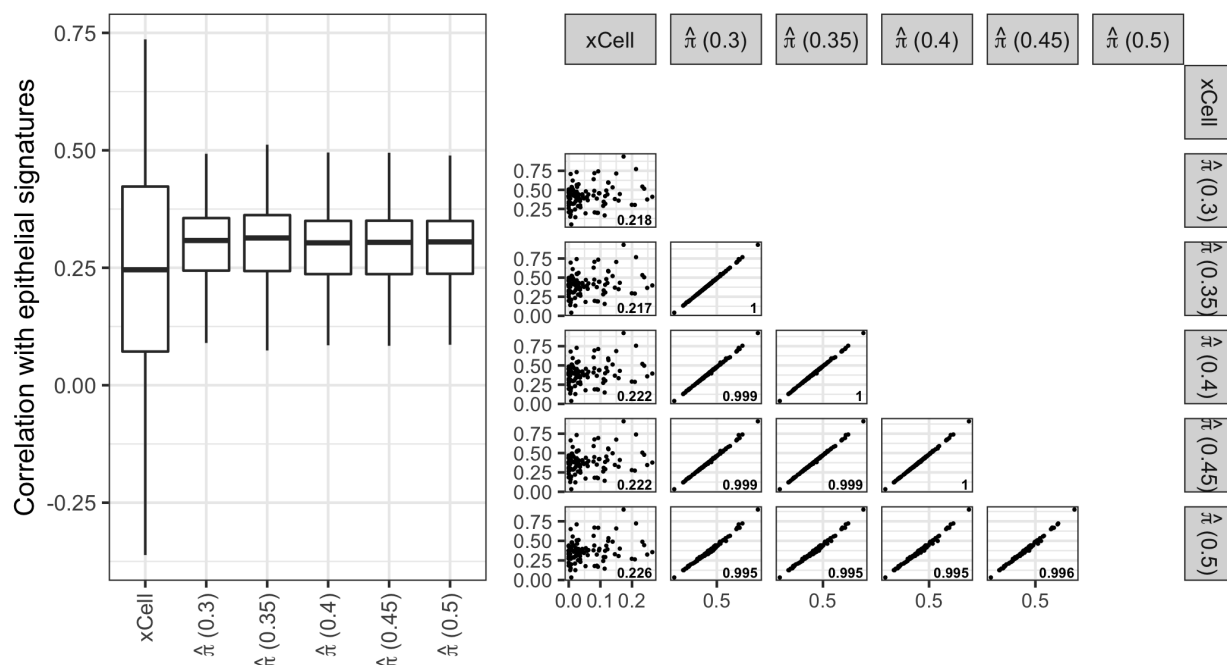
Figure 3: Evaluation of epithelial cell proportion estimates for CPTAC NAT lung tissue samples. (a). Boxplot of the correlation between xCell epithelial score, estimated epithelial cell proportions from the iProMix pipeline using different prior parameters, and the expression levels of 126 individual epithelial cell signature genes from xCell [17]. (b). Correlation between xCell epithelial score and estimated epithelial cell proportions from the iProMix pipeline.

the xCell scores, while all epithelial signatures were positively correlated with the iProMix estimates. This indicated the $\hat{\boldsymbol{\pi}}$'s were more relevant to epithelial cells than xCell scores. In Figure 3(b), the correlations among iProMix estimates and with xCell scores were contrasted. A high correlation among iProMix estimates were observed (r = 0.995-1; p-value < 2e-16), indicating a highly consistent estimation with different priors. Their correlation with xCell score is low (r= 0.217-0.226; p-value = 0.023 -0.030).

4.2   Genes identification

With estimated epithelial cell proportions, we applied iProMix to the proteomics data of CPTAC LUAD adjacent normal tissues to identify epithelial-specific associations. We first considered the protein abundance of ACE2 with each of the 10,698 remaining proteins quantified in the data. The analysis included age and smoking status as covariates, and stratified by sex to allow sex-specific effects as ACE2 is a gene located on X chromosome. Then, we repeated the same analysis for the expression level of ACE2 with all 10,699 proteins in order to further understand these genes for their signals at expression levels.

We employed iProMix (Permutation) procedure to estimate the cell-type specific correlations and identify proteins at a 10% eFDR for each of the five $\hat{\boldsymbol{\pi}}$'s. We then considered the consensus voting across the five $\hat{\boldsymbol{\pi}}$'s to declare significant proteins that were significant in at least 50% of the tests. For comparison purposes, we also modeled the association between ACE2 and other proteins at tissue level with three strategies (1) regression with no adjustment of cell type composition, (2) regression adjusting cell type composition as a covariate, and regressing the interactive effect of ACE2 and cell type composition on other proteins.

Table 2 presented the number of genes identified by each approach, as well as names of significant genes. While no genes were identified to be associated with ACE2 protein levels at tissue-level models, iProMix

	Tissue (No Adj.)	Tissue (Adj. as Cov)	Tissue ($\pi\times$ ACE2)	iProMix (Epithelial)
ACE2 Protein with all other proteins:				
Male	0	0	0	1 (RNASE1)
Female	0	0	0	20*
ACE2 RNA with all proteins:				
Male	0	0	1 (JADE1)	0
Female	0	0	0	0

Table 2: Number of genes identified at FDR 0.1 using tissue-level methods and iProMix, as well as name of identified genes. * The 20 genes include PBXIP1, SGPL1, FAM50B, OTUB1, TMCC1, DDX11, RE-TREG3, HACD3, HPS1, RNF167, PROX1, IKZF2, SOX4, FAM3A, SRD5A3, CFAP157, METAP1D, FGD1, SPOCD1 and NDUFAF8.

identified RNASE1 in men and 20 genes in women. However, none of the identified genes were replicated at the expression level of ACE2 at FDR 10%. The expression level of ACE2 were only identified to be associated with the interactive effect of JADE1 and epithelial cell proportion in men.

4.3 Pathway enrichment for LUAD proteomics

We considered 50 pathways from Hallmark database [31] including IFNα and IFNγ to identify pathways associated with ACE2 in epithelial cell component. iProMix was applied with each of the five pre-estimated $\hat{\pi}$, and the resulting p-values were aggregated using Cauchy combination [33] to get the final p-values across all $\hat{\pi}$. The BY procedure was then applied to the combined p-values to identify pathways associated with ACE2 in epithelial cells. For comparison purposes, we also performed tissue-level analysis and compare for BY controlled pathway discoveries, and if $\hat{\pi}$'s were used, the same Cauchy combination was applied to combine p-values for comparison.

The most significant pathways showing positive association with ACE2 protein abundances in epithelial cells were IFNα/γ pathways (Figure S1). These associations were also observed by looking at the expression levels of ACE2 (Figure S2). Figure 4 compared the association p-values of iProMix with tissue-level methods for IFNγ and INFα pathways. When the cell type composition was not adjusted, in both men and women, at both gene expression and protein abundance levels, the IFNα response pathway was not identified. After adjusting for estimated epithelial cell composition as a covariate, the IFNα response was positively associated with ACE2 protein abundance level in men. The interaction model identified the effects of ACE2 on IFNα response pathway differ by epithelial cell composition in men at its protein level and in women at both RNA and protein levels. Finally, iProMix identified that both ACE2 protein and expression levels in epithelial cell component were positively associated with IFNα response pathway in men, and both were negatively associated with IFNα response pathway in women. The results for the IFNγ response pathway was very similar to INFα that we were able to identify positive associations between ACE2 and INFγ in men and negative associations in women at both protein and RNA levels.

More intriguingly, in the iProMix results, the positive associations between ACE2 and IFN-α/γ were observed only in the tissue samples from male patients, while the association direction is negative in samples from female patients. The sex difference for the association between $ACE2$ and IFN pathways was uniquely identified in our study and may provide potential explanations for the sex difference in immune response to SARS-Cov-2. To provider further understanding of the key players that contribute to sex difference, we further contrasted men and women for the rank of signed LRT scores of the genes within these two pathways . The one with the most striking sex difference was TRIM14 (Figure 5).

5 Discussion

In this work, we develop a new method *iProMix* to test for cell-type specific associations among genes/proteins based on bulk tissue proteogenomic profiles. Besides providing association inferences, iProMix is able to take
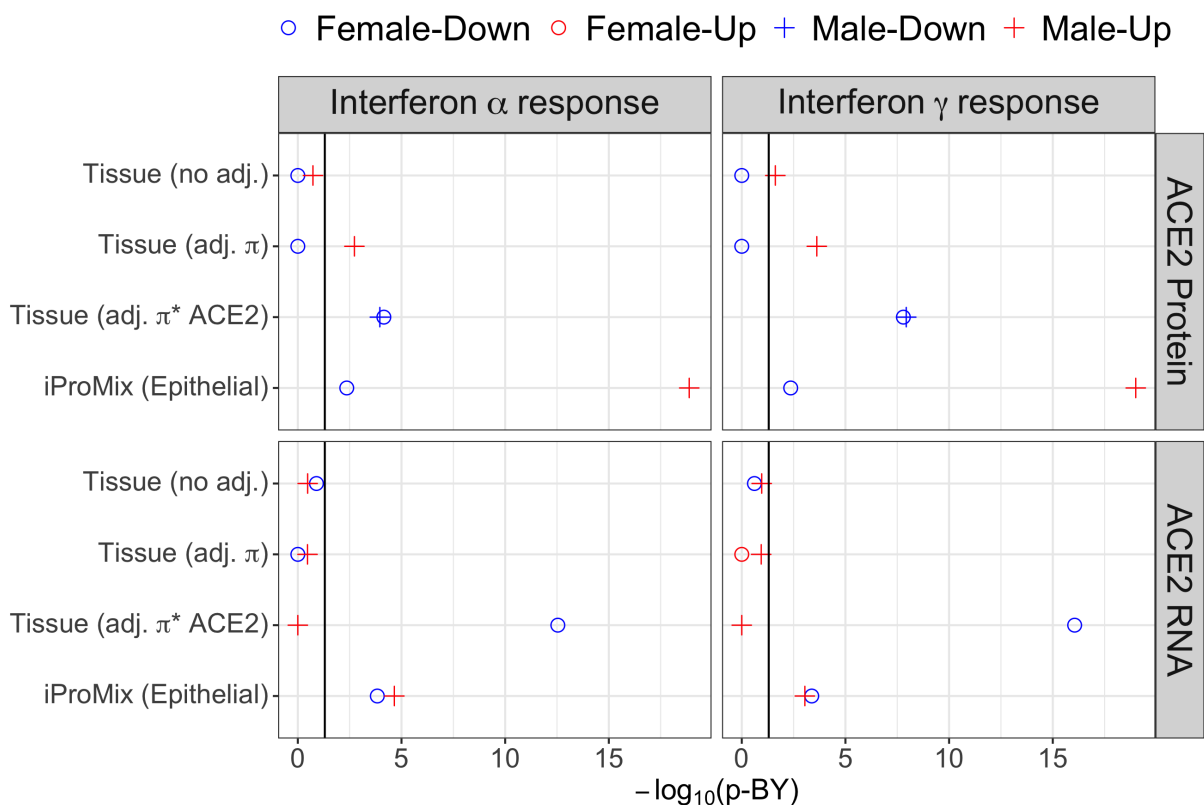
11

Figure 4: Pathway enrichment results for IFNγ and INFα pathways using tissue-level models and iProMix for men and women. Sex differences in immune responses was observed. Both IFNγ and INFα pathways were significantly positively associated with both ACE2 expression and protein levels in epithelial cell component in men. These two pathways were negatively associated with ACE2 expression and protein levels in epithelial cell component in women.

(noisy) cell type percentage estimates as priors and deliver improved percentage estimates in the final results. In the simulation study, we show that *iProMix* properly controls the FDRs and is more powerful than tissue-level analysis for detecting the cell-type specific associations. When iProMix is applied to the CPTAC LUAD proteogenomic data to study ACE2 related genes/proteins, it not only validates the associations between ACE2 and IFN pathways in epithelial cells of human lung tissues, but also revealed novel sex-specific effects.

ACE2 has been suggested to be an IFN stimulated gene in human airway epithelial cells by a recent study based on single cell RNAseq data and *in vitro* experiments [4]. Thus, the fact that iProMix detected IFNα and INFγ as the most significant pathways associated with ACE2 suggests the proposed method is able to capture real biological signals in the bulk tissue data. Moreover, intriguingly, in the iProMix results, the positive associations between ACE2 and IFNα and INFγ were observed only in male patients, while the association was negative in female patients. It has been reported that COVID-19 produced more severe symptoms and higher mortality among men than women [34, 35]. Studies examining the antibody titres and plasma cytokines in COVID-19 patients revealed a sex difference in immune response to COVID-19 virus [36]. The result in our analysis further suggested that the sex difference in immune response to COVID-19 could be partly due to the different regulation patterns between IFN and ACE2 in lung epithelial cells in males and females. For genes in the INFα and INFγ pathways, TRIM14 was identified to show the most striking sex difference. TRIM14 is a mitochondrial adaptor that facilitates retinoic acid-inducible gene-I–like
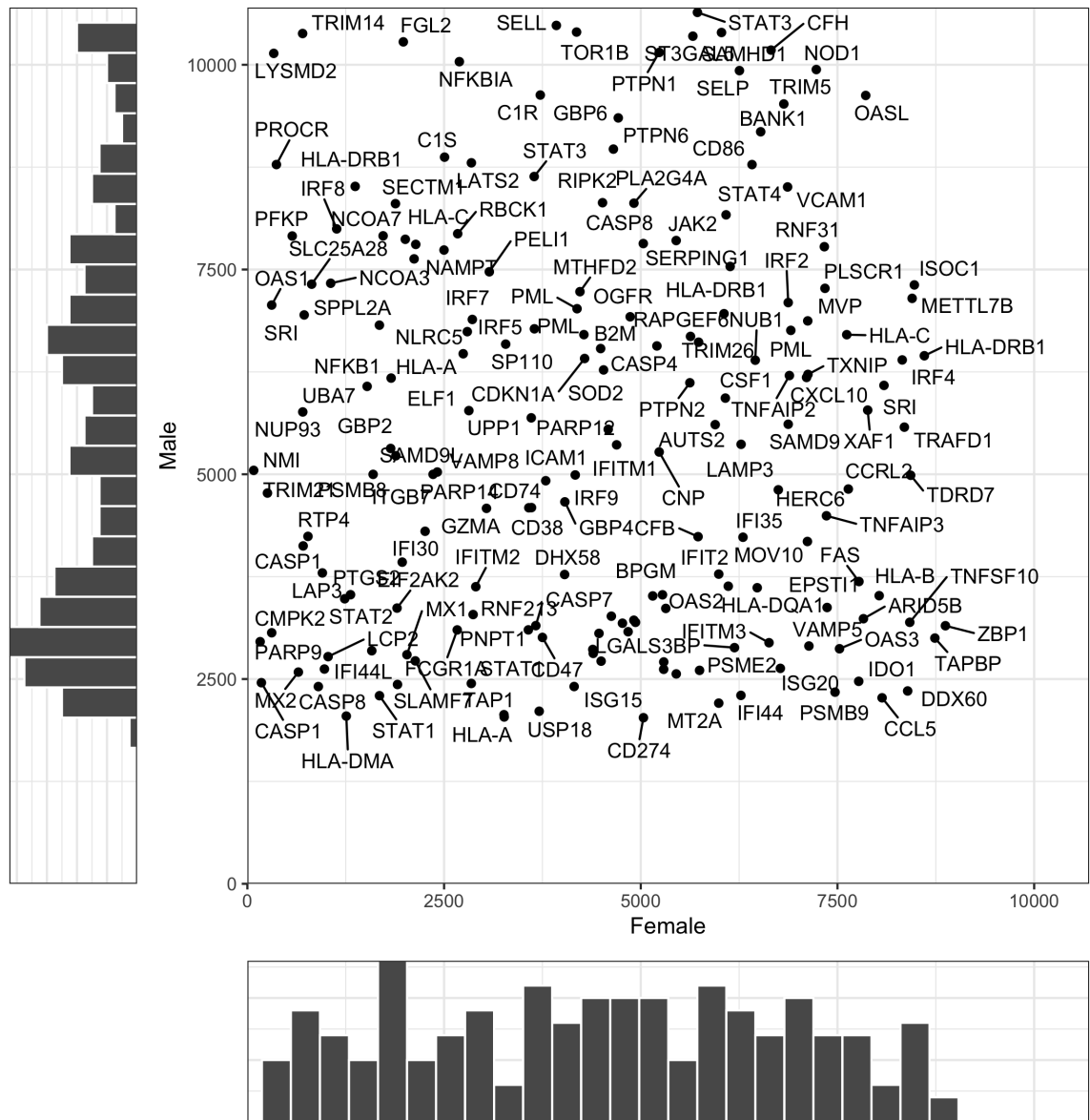
Figure 5: A contrast of ranks of signed LRT scores for genes in INFα and INFγ pathways in male and female patients. The upper left corner (TRIM14) presents that genes that contribute the most sex differences in their associations with ACE2 protein levels.

receptor-mediated innate immune response [37]. It has been reported as a key regulator of the Type I IFN response in other virus infections [38]. Therefore, by applying the new tool iProMix to the real data, we are able to detect promising biological signals in the proteomics that are otherwise missed by using tissue level methods. Nevertheless, cautions should be given in scientific interpretation the results. Analyses were performed on adjacent normal tissue of lung cancer, which may have different protein levels and associations from normal tissues of non-cancer patients. Also, A comparison of men and women should be aware that they have different sample sizes in our data which may affect the power and accuracy in detecting their signals.

The method is applicable to a wide range of studies where the interest is on the dependence of two genes in single or multiple -omic data types that are impacted by cell type composition. For example, analysis on the interrelationship of genes within or across epigenomics, transcriptomics, proteomics and metabolisms can employ iProMix framework for cell-type specific associations. Extension of the tool to more than two cell types is theoretical feasible but the analytical performance for rare cell types needs careful evaluation. Joint consideration of multiple genes, such as via Gaussian graphic models, for their cell-type specific conditional associations is another direction for extension of the method, in which regularization should be further investigated for a good performance. Finally, gene identification is currently achieved by consensus voting of results from multiple priors, and an exploration of more strategies for aggregating results may further improve the study power.

Software implementing the proposed iProMix is available on R CRAN at https://cran.r-project.org/iProMix [To be announced upon acceptance], the data containing genes and pathways identified by iProMix is available on Github at [To be announced upon acceptance].

Acknowledgements

References

[1] Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S Schiergens, Georg Herrler, Nai-Huei Wu, Andreas Nitsche, et al. Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *Cell*, 2020.

[2] Waradon Sungnak, Ni Huang, Christophe Bécavin, Marijn Berg, Rachel Queen, Monika Litvinukova, Carlos Talavera-López, Henrike Maatz, Daniel Reichart, Fotios Sampaziotis, et al. Sars-cov-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nature medicine*, 26(5):681–687, 2020.

[3] Keiji Kuba, Yumiko Imai, Shuan Rao, Hong Gao, Feng Guo, Bin Guan, Yi Huan, Peng Yang, Yanli Zhang, Wei Deng, et al. A crucial role of angiotensin converting enzyme 2 (ace2) in sars coronavirus–induced lung injury. *Nature medicine*, 11(8):875–879, 2005.

[4] Carly GK Ziegler, Samuel J Allon, Sarah K Nyquist, Ian M Mbano, Vincent N Miao, Constantine N Tzouanas, Yuming Cao, Ashraf S Yousif, Julia Bals, Blake M Hauser, et al. 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*, 2020.

[5] Michael A Gillette, Shankha Satpathy, Song Cao, Saravana M Dhanasekaran, Suhas V Vasaikar, Karsten Krug, Francesca Petralia, Yize Li, Wen-Wei Liang, Boris Reva, et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell*, 182(1):200–225, 2020.

[6] David J Clark, Saravana M Dhanasekaran, Francesca Petralia, Jianbo Pan, Xiaoyu Song, Yingwei Hu, Felipe da Veiga Leprevost, Boris Reva, Tung-Shing M Lih, Hui-Yin Chang, et al. Integrated proteogenomic characterization of clear cell renal cell carcinoma. *Cell*, 179(4):964–983, 2019.

[7] Francesca Petralia, Nicole Tignor, Boris Reva, Pichai Raman, Shrabanti Chowdhury, Dmitry Rykunov, Azra Krek, Weiping Ma, Jiayi Ji, Xiaoyu Song, et al. Integrated proteogenomic characterization across seven histological types of pediatric brain tumors, 2020.

[8] Xiaoping Su, Li Zhang, Jianping Zhang, Funda Meric-Bernstam, and John N Weinstein. Purityest: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics*, 28(17):2265–2266, 2012.

[9] David Venet, F Pecasse, Carine Maenhaut, and Hugues Bersini. Separation of samples into their constituents using gene expression data. *Bioinformatics*, 17(suppl_1):S279–S287, 2001.

[10] Timo Erkkilä, Saara Lehmusvaara, Pekka Ruusuvuori, Tapio Visakorpi, Ilya Shmulevich, and Harri Lähdesmäki. Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics*, 26(20):2571–2577, 2010.

[11] Christopher R Bolen, Mohamed Uduman, and Steven H Kleinstein. Cell subset prediction for blood genomic studies. *BMC bioinformatics*, 12(1):258, 2011.

[12] Shai S Shen-Orr, Robert Tibshirani, Purvesh Khatri, Dale L Bodian, Frank Staedtler, Nicholas M Perry, Trevor Hastie, Minnie M Sarwal, Mark M Davis, and Atul J Butte. Cell type–specific gene expression differences in complex tissues. *Nature methods*, 7(4):287, 2010.

[13] Jason E Shoemaker, Tiago JS Lopes, Samik Ghosh, Yukiko Matsuoka, Yoshihiro Kawaoka, and Hiroaki Kitano. Cten: a web-based platform for identifying enriched cell types from heterogeneous microarray data. *BMC genomics*, 13(1):460, 2012.

[14] Kosuke Yoshihara, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, Hui Shen, Peter W Laird, Douglas A Levine, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications*, 4:2612, 2013.

[15] Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, et al. Absolute quantification of somatic dna alterations in human cancer. *Nature biotechnology*, 30(5):413, 2012.

[16] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457, 2015.

[17] Dvir Aran, Zicheng Hu, and Atul J Butte. xcell: digitally portraying the tissue cellular heterogeneity landscape. *Genome biology*, 18(1):220, 2017.

[18] Elior Rahmani, Regev Schweiger, Liat Shenhav, Theodora Wingert, Ira Hofer, Eilon Gabel, Eleazar Eskin, and Eran Halperin. Bayescce: a bayesian framework for estimating cell-type composition from dna methylation without the need for methylation reference. *Genome biology*, 19(1):1–18, 2018.

[19] Ankur Chakravarthy, Andrew Furness, Kroopa Joshi, Ehsan Ghorani, Kirsty Ford, Matthew J Ward, Emma V King, Matt Lechner, Teresa Marafioti, Sergio A Quezada, et al. Pan-cancer deconvolution of tumour composition using dna methylation. *Nature communications*, 9(1):1–13, 2018.

[20] Zeya Wang, Shaolong Cao, Jeffrey S Morris, Jaeil Ahn, Rongjie Liu, Svitlana Tyekucheva, Fan Gao, Bo Li, Wei Lu, Ximing Tang, et al. Transcriptome deconvolution of heterogeneous tumor samples with immune infiltration. *iScience*, 9:451–460, 2018.

[21] Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, 10(1):1–9, 2019.

[22] Gregory J Hunt, Saskia Freytag, Melanie Bahlo, and Johann A Gagnon-Bartsch. dtangle: accurate and robust cell type deconvolution. *Bioinformatics*, 35(12):2093–2099, 2019.

[23] Aaron M Newman, Chloé B Steen, Chih Long Liu, Andrew J Gentles, Aadel A Chaudhuri, Florian Scherer, Michael S Khodadoust, Mohammad S Esfahani, Bogdan A Luca, David Steiner, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology*, 37(7):773–782, 2019.

[24] Francesca Petralia, Li Wang, Jie Peng, Arthur Yan, Jun Zhu, and Pei Wang. A new method for constructing tumor specific gene co-expression networks based on samples with tumor purity heterogeneity. *Bioinformatics*, 34(13):i528–i536, 2018.

[25] Jaeil Ahn, Ying Yuan, Giovanni Parmigiani, Milind B Suraokar, Lixia Diao, Ignacio I Wistuba, and Wenyi Wang. Demix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics*, 29(15):1865–1871, 2013.

[26] Margaret KR Donovan, Agnieszka D'Antonio-Chronowska, Matteo D'Antonio, and Kelly A Frazer. Cellular deconvolution of gtex tissues powers discovery of disease and cell-type associated regulatory variants. *Nature communications*, 11(1):1–14, 2020.

[27] Xiangyu Luo, Can Yang, and Yingying Wei. Detection of cell-type-specific risk-cpg sites in epigenome-wide association studies. *Nature communications*, 10(1):1–12, 2019.

[28] Xiao-Li Meng and Donald B Rubin. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

[29] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[30] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[31] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.

[32] David A Barbie, Pablo Tamayo, Jesse S Boehm, So Young Kim, Susan E Moody, Ian F Dunn, Anna C Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, et al. Systematic rna interference reveals that oncogenic kras-driven cancers require tbk1. *Nature*, 462(7269):108–112, 2009.

[33] Yaowu Liu and Jun Xie. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020.

[34] Nanshan Chen, Min Zhou, Xuan Dong, Jieming Qu, Fengyun Gong, Yang Han, Yang Qiu, Jingli Wang, Ying Liu, Yuan Wei, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. *The lancet*, 395(10223):507–513, 2020.

[35] Catherine Gebhard, Vera Regitz-Zagrosek, Hannelore K Neuhauser, Rosemary Morgan, and Sabra L Klein. Impact of sex and gender on covid-19 outcomes in europe. *Biology of sex differences*, 11:1–13, 2020.

[36] Takehiro Takahashi, Mallory K Ellingson, Patrick Wong, Benjamin Israelow, Carolina Lucas, Jon Klein, Julio Silva, Tianyang Mao, Ji Eun Oh, Maria Tokuyama, et al. Sex differences in immune responses that underlie covid-19 disease outcomes. *Nature*, 588(7837):315–320, 2020.

[37] Zhuo Zhou, Xue Jia, Qinghua Xue, Zhixun Dou, Yijie Ma, Zhendong Zhao, Zhengfan Jiang, Bin He, Qi Jin, and Jianwei Wang. Trim14 is a mitochondrial adaptor that facilitates retinoic acid-inducible gene-i–like receptor-mediated innate immune response. *Proceedings of the National Academy of Sciences*, 111(2):E245–E254, 2014.

[38] Caitlyn T Hoffpauir, Samantha L Bell, Kelsi O West, Tao Jing, Allison R Wagner, Sylvia Torres-Odio, Jeffery S Cox, A Phillip West, Pingwei Li, Kristin L Patrick, et al. Trim14 is a key regulator of the type i ifn response during mycobacterium tuberculosis infection. *The Journal of Immunology*, 205(1):153–167, 2020.
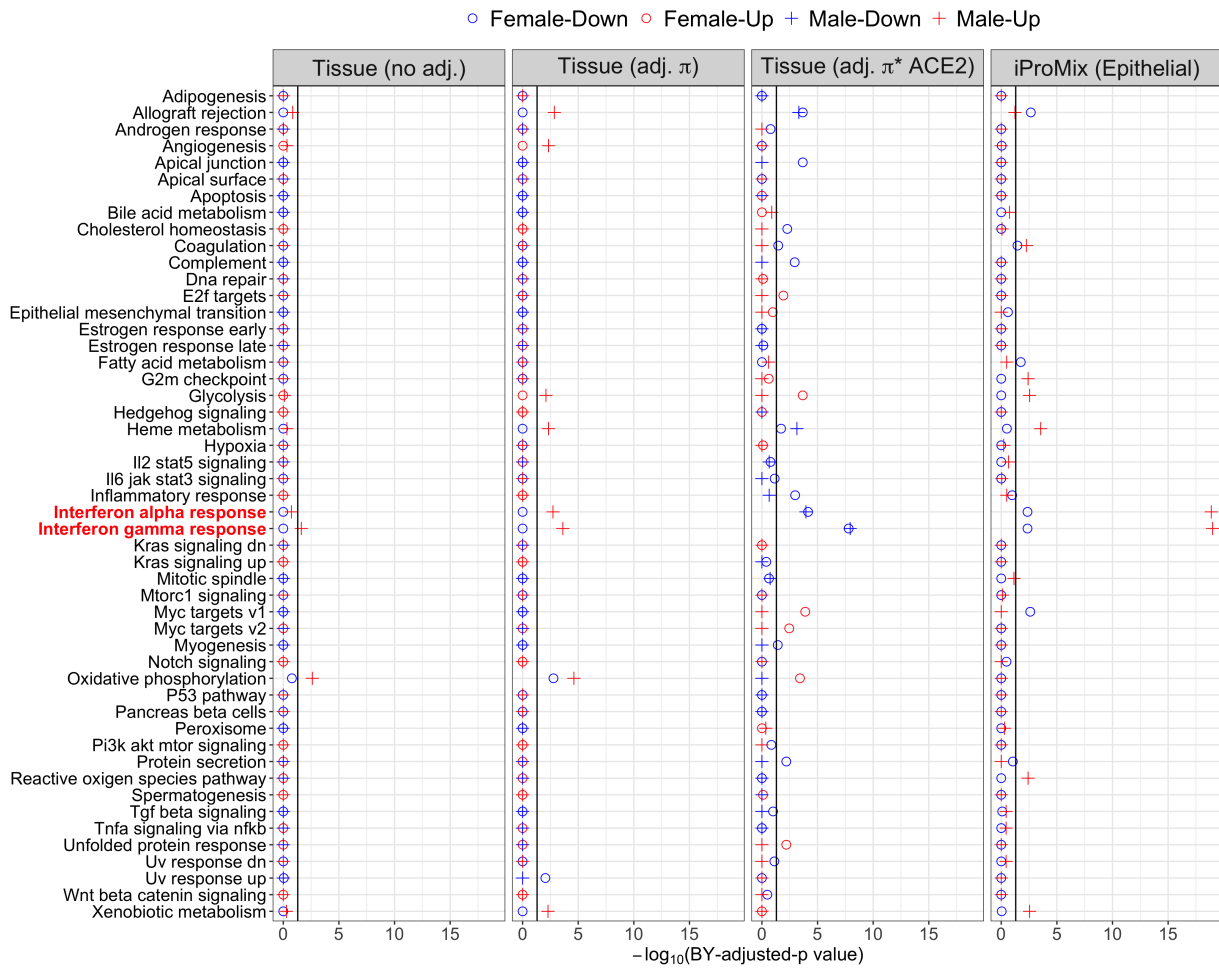
Supplementary Materials

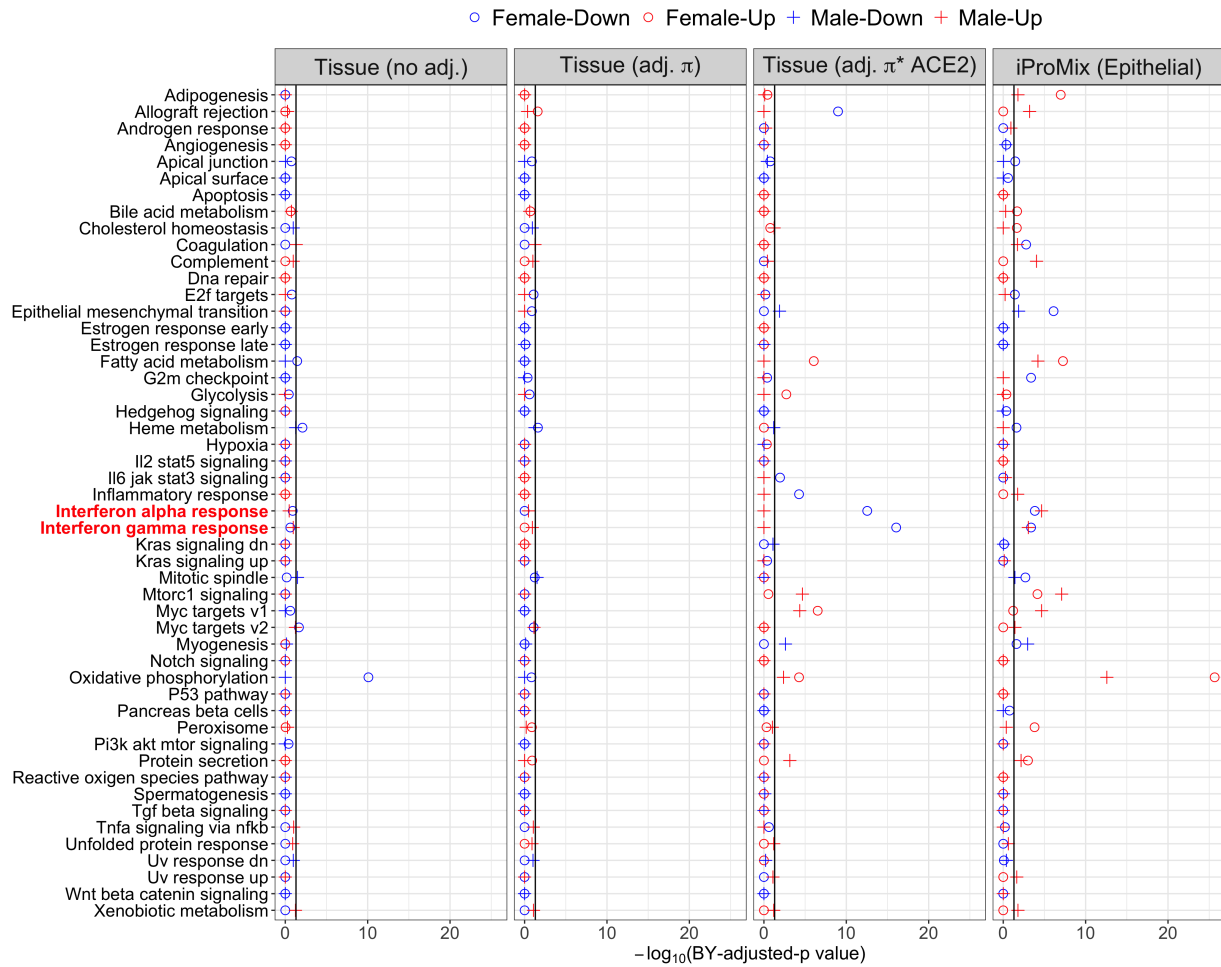Figure S1: Pathway enrichment analysis for 50 Hallmark pathways associated with ACE2 protein level.

Figure S2: Pathway enrichment analysis for 50 Hallmark pathways associated with ACE2 RNA level.

Identifiability of Θ.

Suppose we consider an alternative sets of parameter $\widetilde{\Theta}$, and then the iProMix model is identifiable if $l(\Theta \mid \{o_i\}_{i=1}^n) = l(\widetilde{\Theta} \mid \{o_i\}_{i=1}^n)$ for any observed data o_i's implies $\Theta = \widetilde{\Theta}$. Suppose the $\breve{z}_i = (1, z_i^T)^T$ includes the intercept and P dimensional covariates. As the normal distribution is identifiable, the mean function of the ACE2 can be written as

$$
\pi_i \sum_{p=1}^{P+1} \breve{z}_{ip} \beta_{ux}^{(p)} + (1 - \pi_i) \sum_{p=1}^{P+1} \breve{z}_{ip} \beta_{vx}^{(p)} = \pi_i \sum_{p=1}^{P+1} \breve{z}_{ip} \widetilde{\beta}_{ux}^{(p)} + (1 - \pi_i) \sum_{p=1}^{P+1} \breve{z}_{ip} \widetilde{\beta}_{vx}^{(p)}.
$$

This equation can be rewritten into a matrix form as follows:

$$
\left(\pi_i \breve{z}_{i1}, ..., \pi_i \breve{z}_{i(P+1)}, (1 - \pi_i) \breve{z}_{i1}, ..., (1 - \pi_i) \breve{z}_{i(P+1)} \right)
\begin{pmatrix}
\beta_{ux}^{(1)} \\
\vdots \\
\beta_{ux}^{(P+1)} \\
\beta_{vx}^{(1)} \\
\vdots \\
\beta_{vx}^{(P+1)}
\end{pmatrix}
$$

$$
= \left(\pi_i \breve{z}_{i1}, ..., \pi_i \breve{z}_{i(P+1)}, (1 - \pi_i) \breve{z}_{i1}, ..., (1 - \pi_i) \breve{z}_{i(P+1)} \right)
\begin{pmatrix}
\widetilde{\beta}_{ux}^{(1)} \\
\vdots \\
\widetilde{\beta}_{ux}^{(P+1)} \\
\widetilde{\beta}_{vx}^{(1)} \\
\vdots \\
\widetilde{\beta}_{vx}^{(P+1)}
\end{pmatrix}.
$$

By combining these n equations, the first component $A = \left(\pi_i z_{i1}, ..., (1 - \pi_i) z_{i1}, ..., (1 - \pi_i) z_{i(P+1)} \right)_{i=1}^n$ is a $n \times 2(P+1)$ matrix with a rank of $2(P+1)$ (full column rank). A has a left inverse A^{-1}, and multiplying by A^{-1} from the left on both sides, we obtain $\beta_{ux} = \widetilde{\beta}_{ux}$ and $\beta_{vx} = \widetilde{\beta}_{vx}$. Similarly, we were able to prove $\Sigma_{ux} = \widetilde{\Sigma}_{ux}$ and $\Sigma_{vx} = \widetilde{\Sigma}_{vx}$, and the rest of the paramters for the protein. As a result, $\Theta = \widetilde{\Theta}$ and iProMix is identifiabile.