Integrating airway microbiome and plasma proteomics data to identify multi-omic networks associated with cystic fibrosis pulmonary exacerbation treatment response

Brenton Graham

Department of Biostatistics and Informatics
University of Colorado Anschutz Medical Campus

Committee: Laura Saba (PhD), Brandie Wagner (PhD), Jonathan K Harris (PhD)

7/3/23

Cystic Fibrosis

- Cystic fibrosis (CF) is a chronic, genetic disease that causes the body to produce abnormally thick mucus
- People with CF are at high risk of chronic bacterial infections, inflammation, and progressive respiratory complications
- ~40,000 children and adults have been diagnosed with CF in the United States

Pulmonary Exacerbations

- Pulmonary exacerbations (PExs) are the leading cause of morbidity in CF
- PExs are significant life events associated with...
  - Acute decrease in lung function
  - Reduced quality of life (QOL)
  - Shortened survival
- Lung function is often not fully recovered despite seemingly appropriate therapies (e.g., targeted IV antibiotic treatment)

Airway Microbiome & Blood Proteome in CF

- Inflammatory biomarkers in both the airway and blood have been shown to decrease after treatment of a PEx
- Evidence suggests that airway infection in CF results in a robust host immune response
- Identifying associations between specific airway bacteria or bacterial communities and host-response may be critical to understanding the pathogenicity of CF bacteria

The Thesis

- **Goal:** To identify multiomic (taxon—protein) networks at PEx onset that are indicative of PEx recovery
- We use an extension of canonical correlation analysis (CCA) called sparse multiple canonical correlation network (SmCCNet) for data integration (Shi et al., Kechris Lab)
- We hope to provide insights into the **variability observed in PEx recovery**

Study Design & Population

- 33 PEx events from a cohort of 29 subjects aged 10 to 22
- Participants could be reenrolled if PEx events were separated by ≥ 6 months
  - 25 subjects with one PEx event
  - 4 subjects with two PEx events
- Participants were recruited prospectively and enrolled at the time of hospital admission for IV antibiotic therapy of a clinically diagnosed PEx
- IV antibiotics were targeted for specific CF pathogens as determined through microbial culture

Study Design & Population

- The study focuses on two time points
  - Hospital admission (i.e., PEx onset, day 0-2)
  - Hospital discharge (i.e., After IV treatment, day 4-21)

- Study procedures at each visit included
  - A physical
  - A spirometry test
  - **A standardized PEx score**
  - A validated QOL measure
  - Specimen collection (**blood** and **sputum** samples)

The Phenotype: $\%\Delta\text{PExS}$

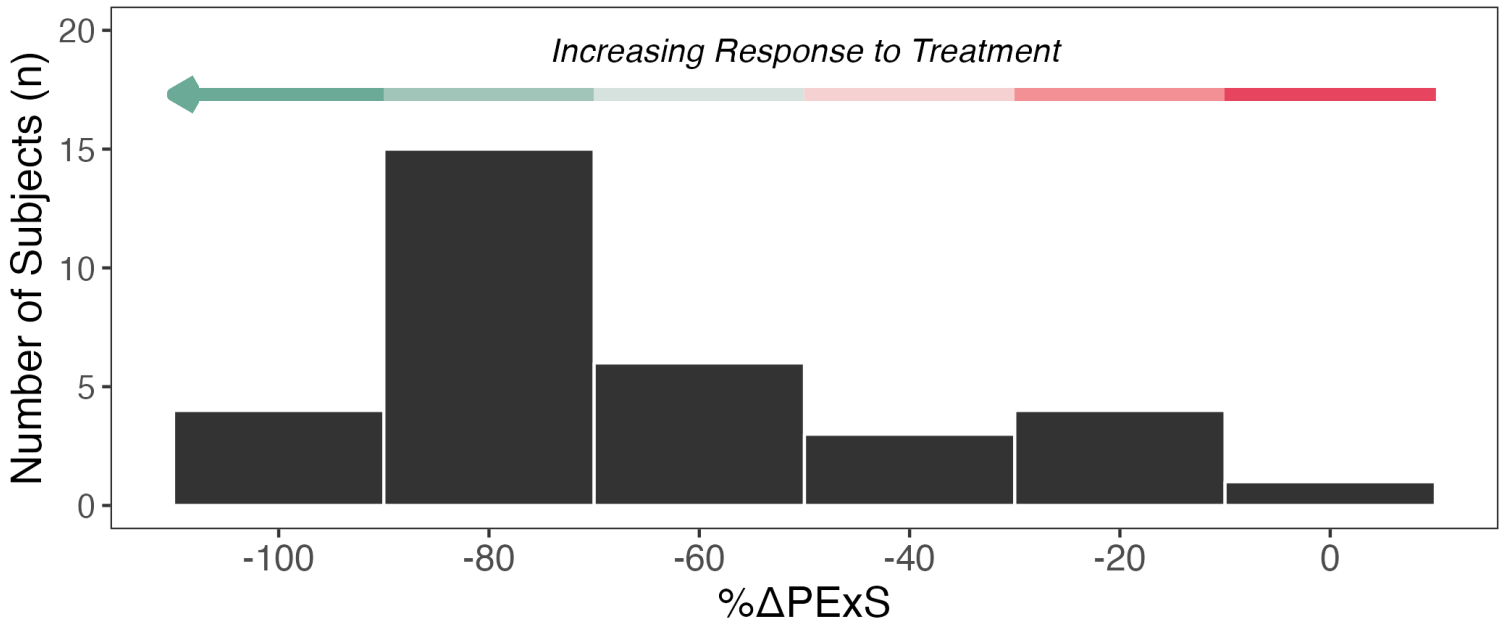
- The phenotype of interest is % change in PEx score (PExS) between hospital admission (t_1) and discharge (t_2), $\%\Delta\text{PExS}$

$$\%\Delta\text{PExS} = \frac{\text{PExS}_{t_2} - \text{PExS}_{t_1}}{\text{PExS}_{t_1}} \times 100\%$$

- PEx score (PExS) is a standardized score that considers
  - Patient symptoms (2 week change in exercise tolerance, cough, sputum production, chest congestion, school/work attendance, appetite)
  - Physical examination findings (increased adventitial sounds on auscultation of the chest, change in FEV_1)

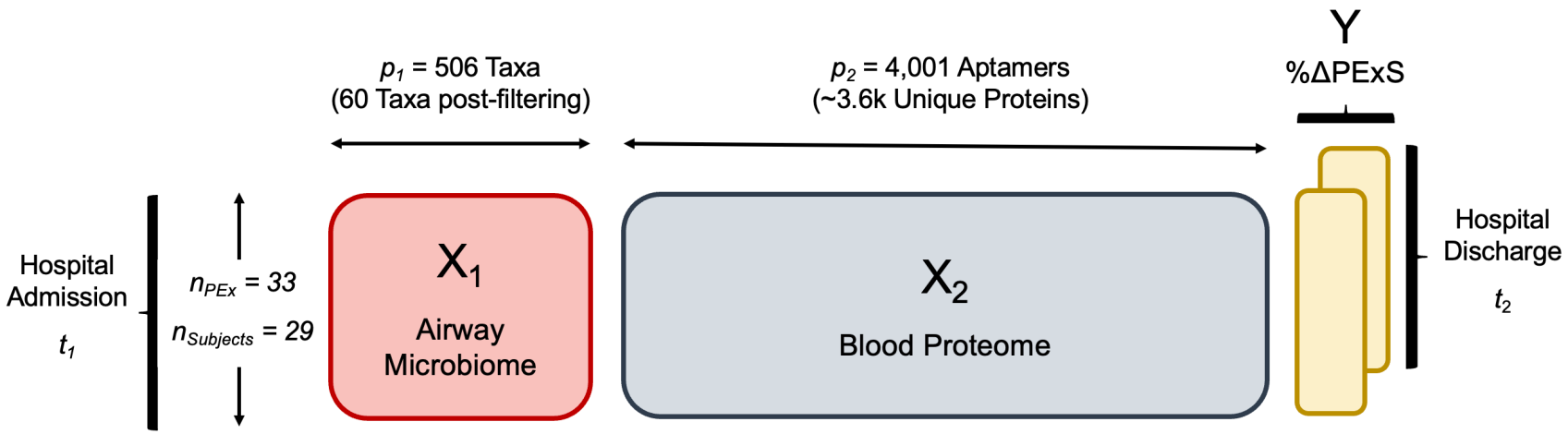
The Phenotype: %ΔPExS

- %ΔPExS as a clinical measure of PEx recovery



Study Design

- The predictors (X_1 and X_2) are measured at PEx onset
- The outcome, Y, is a longitudinal measure



Study Demographics

Cohort Subjects (n = 29)	
Age (Years)	15.9 [10.5, 22.1]
Female	15 (51.7%)
Body Mass Index (BMI)	19.4 [15.6, 26.1]
Genotype (CF Mutations)	
0 F508del	2 (6.9%)
1 F508del	9 (31.0%)
2 F508del	18 (62.1%)
All PEx Events (n = 33)	
FEV-1 % Predicted at Admission	81.0 [30.0, 119.0]
PEx Score at Admission	12.0 [8.0, 16.0]
%ΔPExS	-72.7 [-100.0, 0.0]
CF Bacteria Culture Detection	
P. aeruginosa	11 (33.3%)
S. aureus	19 (57.6%)
Haemophilus	1 (3.0%)
Stenotrophomonas	5 (15.2%)
Burkholderia	5 (15.2%)

Airway Samples, Sequencing & Sequence Analysis

- Spontaneously expectorated **sputum** was used for airway microbiome analysis; sputum induction was performed for participants unable to spontaneously expectorate
- Amplicons were generated using primers targeting approximately 300 base pairs of the V1/V2 variable region of the 16S rRNA gene
- Illumina paired-end sequencing was performed on the MiSeq platform using a 500 cycle v2 reagent kit
- Assembled sequences were aligned and classified with SINA (1.2.11) using the Silva 111 database as reference

Airway Microbiome Data Preprocessing

- Microbiome data were filtered to include only prevalent taxa
  - **Detection Threshold**: 0.1% Relative Abundance (RA)
  - **Prevalence Threshold**: 10% of Samples
  - *Taxa must exceed 0.1% RA in $\geq$ 10% of samples*
- Count data were transformed using the centered log-ratio (CLR) transformation given by

$$clr(x) = ln\, x_i - \frac{1}{D}\sum_{j=1}^{D} ln\, x_j$$

  where D represents the number of components (or taxa)
- A pseudocount of $\mathrm{RA}_{min}/2$ was applied to exact zero RA entries before CLR transformation

Blood Proteomics Assay & Data Preprocessing

- Blood samples were sent to SomaLogic for proteomics analysis
- Proteomics data were measured using the SomaScan multiplex proteomics assay, an aptamer-based assay measuring ~3.6k unique proteins
- RFU values were transformed using a $\log_2$-transformation
- All data were standardized prior to statistical analysis

Canonical Correlation Analysis (CCA)

- CCA aims to find the linear combination of variables that maximizes the correlation (i.e., canonical correlation) between two multivariate data sets (e.g., X_1, X_2)
- Canonical weights w_1 and w_2 are defined as

$$(w_1, w_2) = \arg\ \max_{\tilde{w}_1, \tilde{w}_2} \mathrm{Cor}(X_1\tilde{w}_1, X_2\tilde{w}_2)$$

where $\mathrm{Cor}(X_1\tilde{w}_1, X_2\tilde{w}_2)$ denotes the canonical correlation between X_1 and X_2 and $\mathrm{Cor}(X_1\tilde{w}_1, X_2\tilde{w}_2) = \tilde{w}_1^T X_1^T X_2 \tilde{w}_2$, subject to $\tilde{w}_1^T X_1^T X_1 \tilde{w}_1 = \tilde{w}_2^T X_2^T X_2 \tilde{w}_2 = 1$

Sparse Multiple CCA (SmCCA)

- SmCCA incorporates a third data type (i.e, the phenotype Y) into the integration task by accounting for phenotype—omic correlation within the canonical weight objective function
- The definition of (w_1, w_2) becomes

$$(w_1, w_2) = \arg\ \max_{\tilde{w}_1, \tilde{w}_2} (a\tilde{w}_1^T X_1^T X_2 \tilde{w}_2 + b\tilde{w}_1^T X_1^T Y + c\tilde{w}_2^T X_2^T Y)$$

where a, b, and c are scaling constants that can be used to used to prioritize correlations with the phenotype (i.e., taxon—%ΔPExS or protein—%ΔPExS correlation)

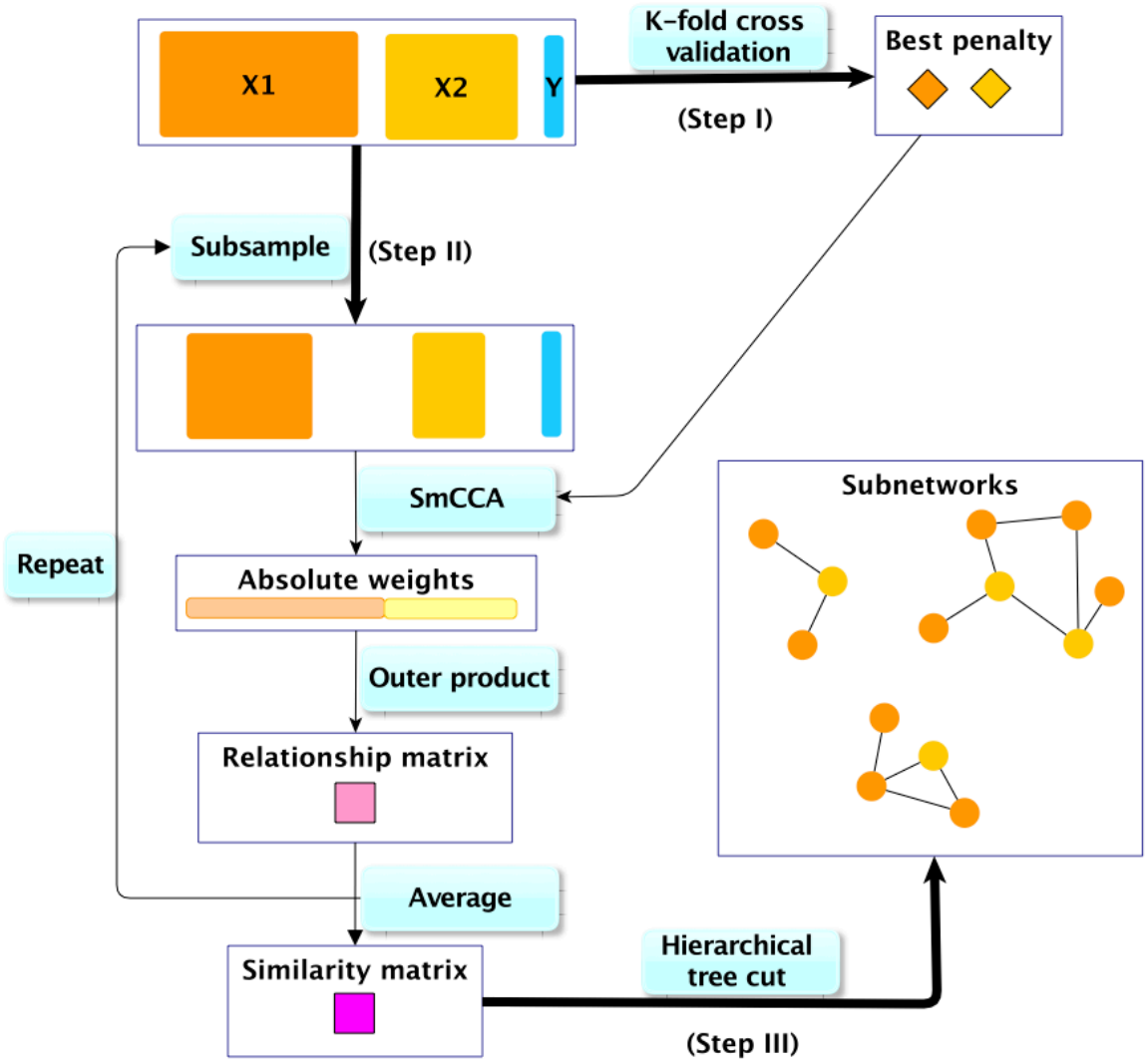
Sparse Multiple CCA (SmCCA)

- Sparsity is imposed on the canonical weights (w_1, w_2) since not all features contribute to the true canonical correlation
- w_1 and w_2 in SmCCA are subject to

$$||\tilde{w}_s||^2 = 1,\ P_s(\tilde{w}_s) \leq c_s,\ s = 1, 2$$

where $P(\cdot)$ represent penalty functions (e.g., the LASSO) and c_s represent pre-specified sparse penalty constants

SmCCNet Workflow

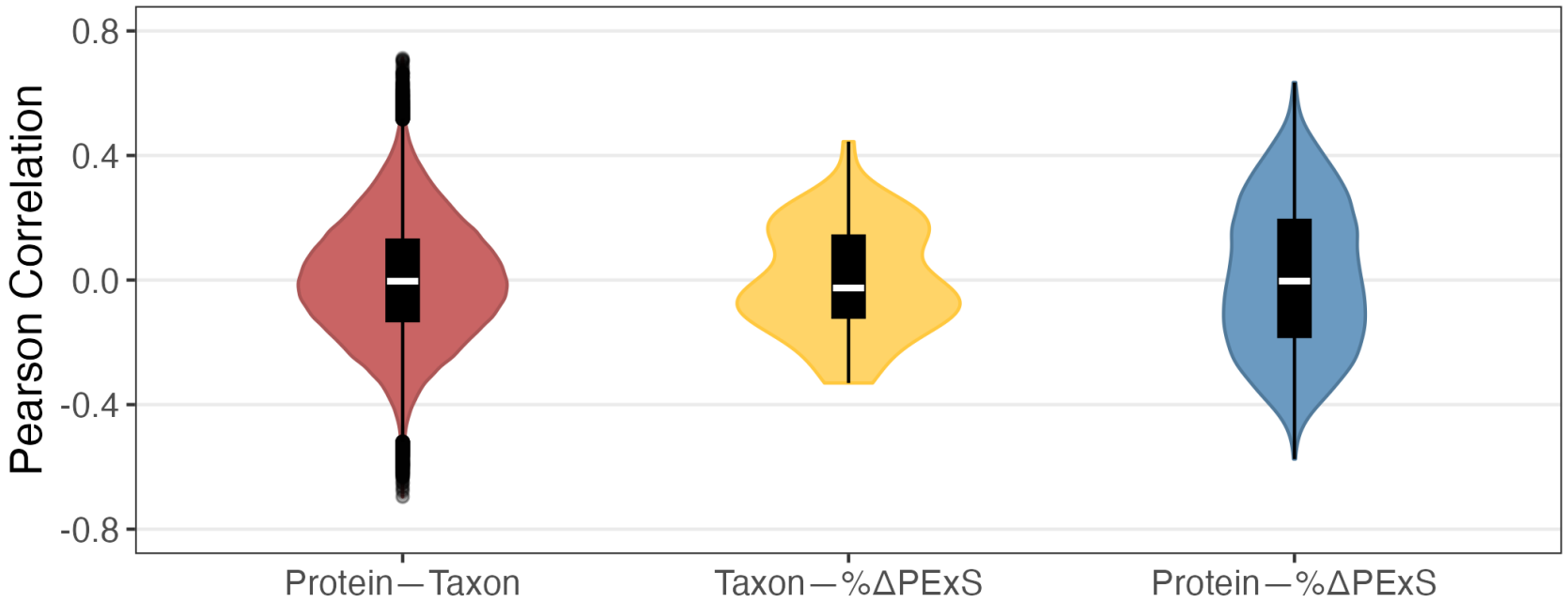


Hyperparameter Tuning: Sparse Penalty Selection

- We used 5-fold cross-validation (CV) and a randomized grid search approach to select the optimal penalty pair
- The selected penalty pair corresponds to the pair that minimizes the prediction error between training and test sets
- Counterintuitively, increasing the value of a penalty parameter weakens the strength of regularization
- We searched a range of larger values for the X_1 penalty parameter (0.4 to 0.6) and a range of smaller values for the X_2 penalty parameter (0.1 to 0.3) due to feature imbalance between X_1 (60 taxa) and X_2 (4,001 aptamers)
- We used feature subsampling proportions of 0.90 and 0.70 for X_1 and X_2 to further account for dimensionality imbalance

Hyperparameter Tuning: SmCCA Weighting Scheme

- We explored the *weighted* version of SmCCNet (the case where a, b, and c are not equal) since correlations between taxon—protein were stronger than correlations between taxon—%ΔPExS and protein—%ΔPExS



Hyperparameter Tuning: SmCCA Weighting Scheme

- Remember the SmCCA canonical weight objective function

$$(w_1, w_2) = \arg\ \max_{\tilde{w}_1, \tilde{w}_2} (a\tilde{w}_1^T X_1^T X_2 \tilde{w}_2 + b\tilde{w}_1^T X_1^T Y + c\tilde{w}_2^T X_2^T Y)$$

- We tried various (a, b, c) weighting schemes to test the effect of increasing taxon—phenotype correlation importance (b)
- Tested (a, b, c): (1, 1, 1), (1, 2, 1), (1, 5, 1)
- Optimal weighting scheme was determined by considering:
  - Subnetwork—phenotype correlation strength
  - Subnetwork size (i.e., number of nodes)
  - Taxon—protein balance

Network Summarization

- Principal component analysis (PCA) was used for subnetwork summarization
- We used the correlation between subnetwork-specific PC1s and $\%\Delta\mathrm{PExS}$ to measure subnetwork-phenotype association
- Absolute subnetwork—phenotype correlations are reported as the use of PC1 obscures the interpretability of $+/-$ relationships
- Defining *strong* associations as $|\rho| > 0.3$

Subnetwork Pruning

- We aimed to incorporate a rational/systematic process to limit subnetwork sizes to 325 nodes (<10% of the feature space)
- The thought was to limit the number of proteins in a large subnetwork to ~300 proteins
- The pruning process aims to trim the least important nodes in a given subnetwork

The Pruning Process for Large Subnetworks (>325 Nodes)

1. Rank nodes by *importance* using the PageRank algorithm
2. Select the top 325 ranked nodes

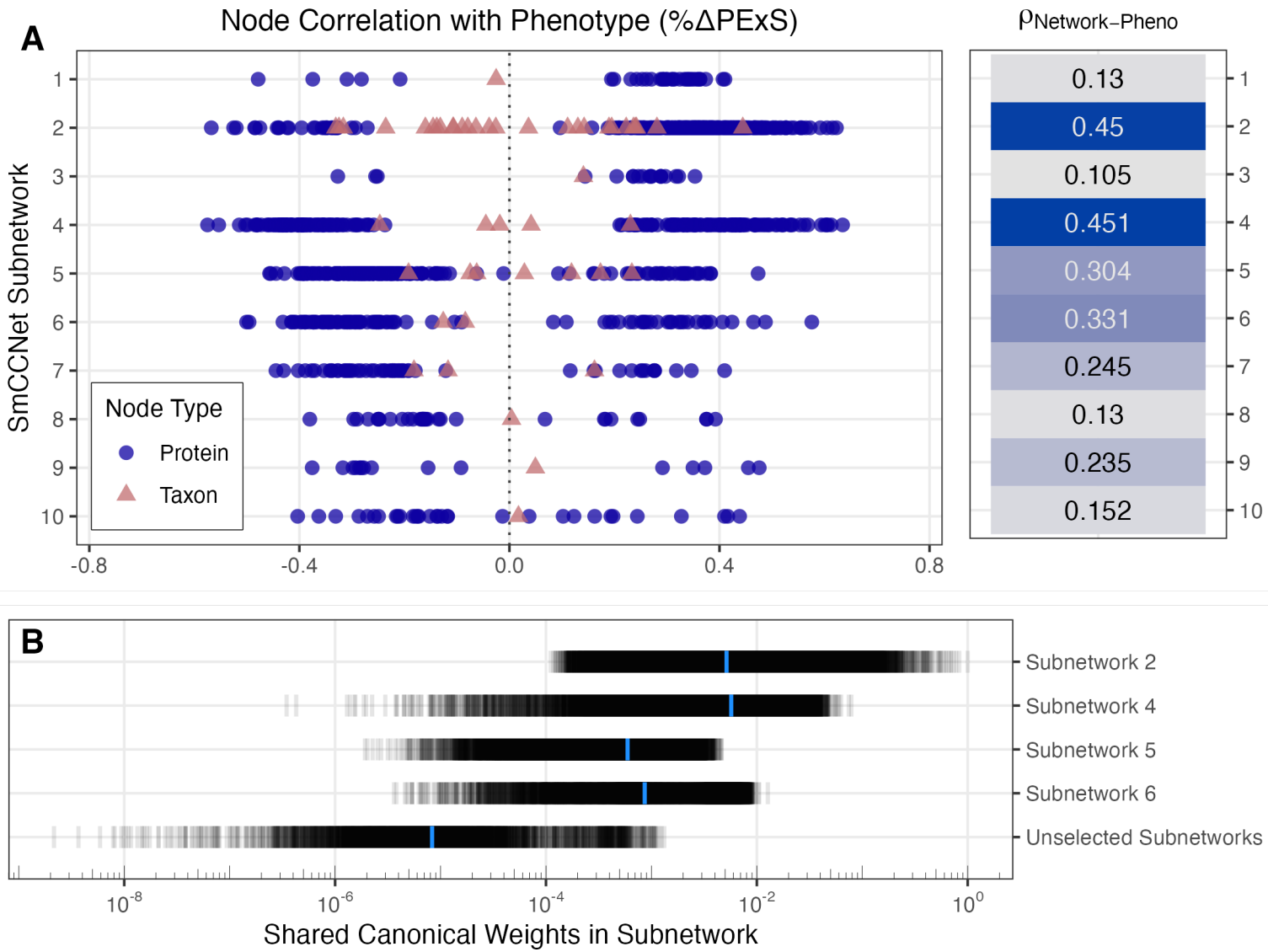
Subnetwork Visualization

- **Edge Pruning**
  - Weak node-to-node connections (edges) can blur biologically relevant relationships in subnetwork visualizations
  - Edges were removed (i.e., set to 0) if between node correlations were weak ($\rho < 0.2$)
- **GO-Specific Subnetwork Visualization**
  - Large subnetwork visualization is difficult due to the number of nodes, edges, and subnetwork attributes
  - We selected one GO pathway per subnetwork to visualize
  - Network visualizations include the proteins contained within the selected GO pathway and subnetwork-specific taxa

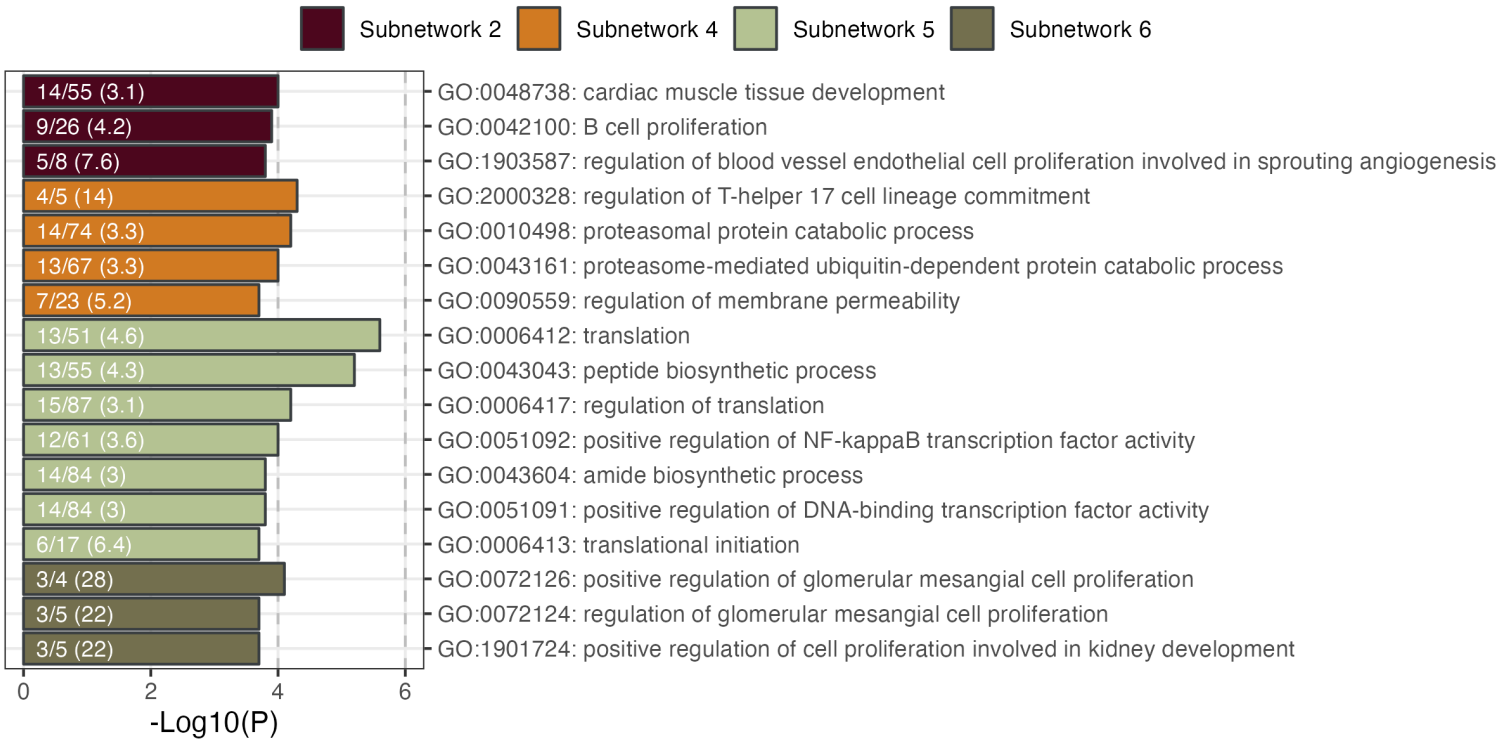
GO Enrichment Analysis

- **Metascape** was used for GO enrichment analysis using subnetwork-specific protein sets
  - P-value threshold: 0.001
  - Enrichment threshold: 3
  - Minimum protein threshold: 3
  - The full set of unique proteins targeted by the assay was used as the background list

Identified SmCCNet Networks

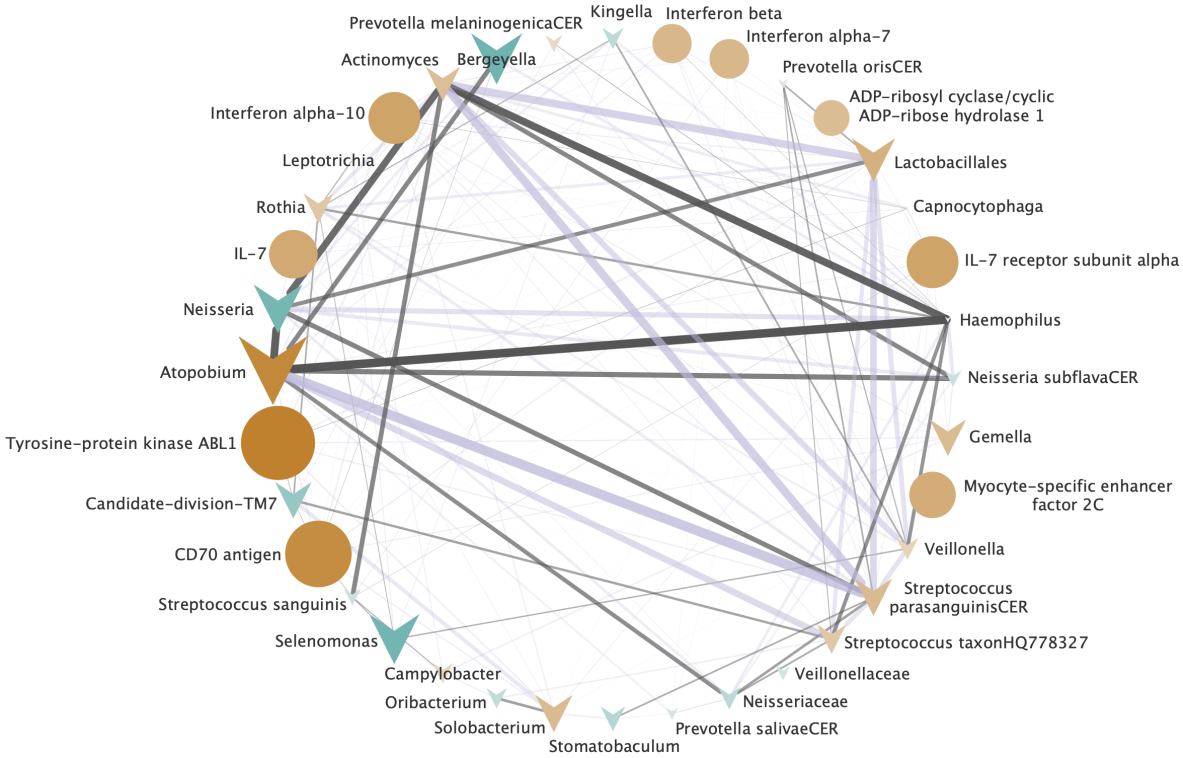


GO Enrichment Results



Subnetwork 2: GO:0042100

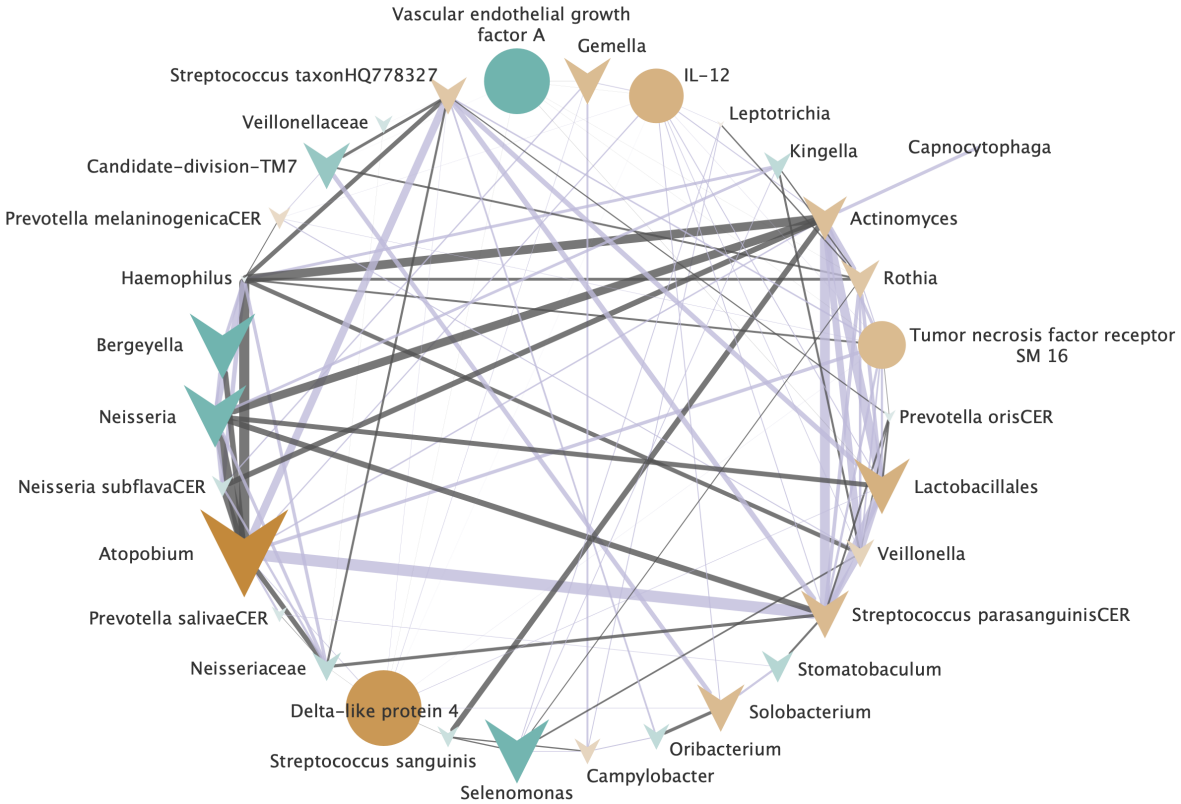
Node Count	Protein Count	Taxon Count	Network—%ΔPExS Corr	Node—%ΔPExS Corr Range
325	298	27	0.45	(-0.568, 0.623)



- Include canonical weight distribution

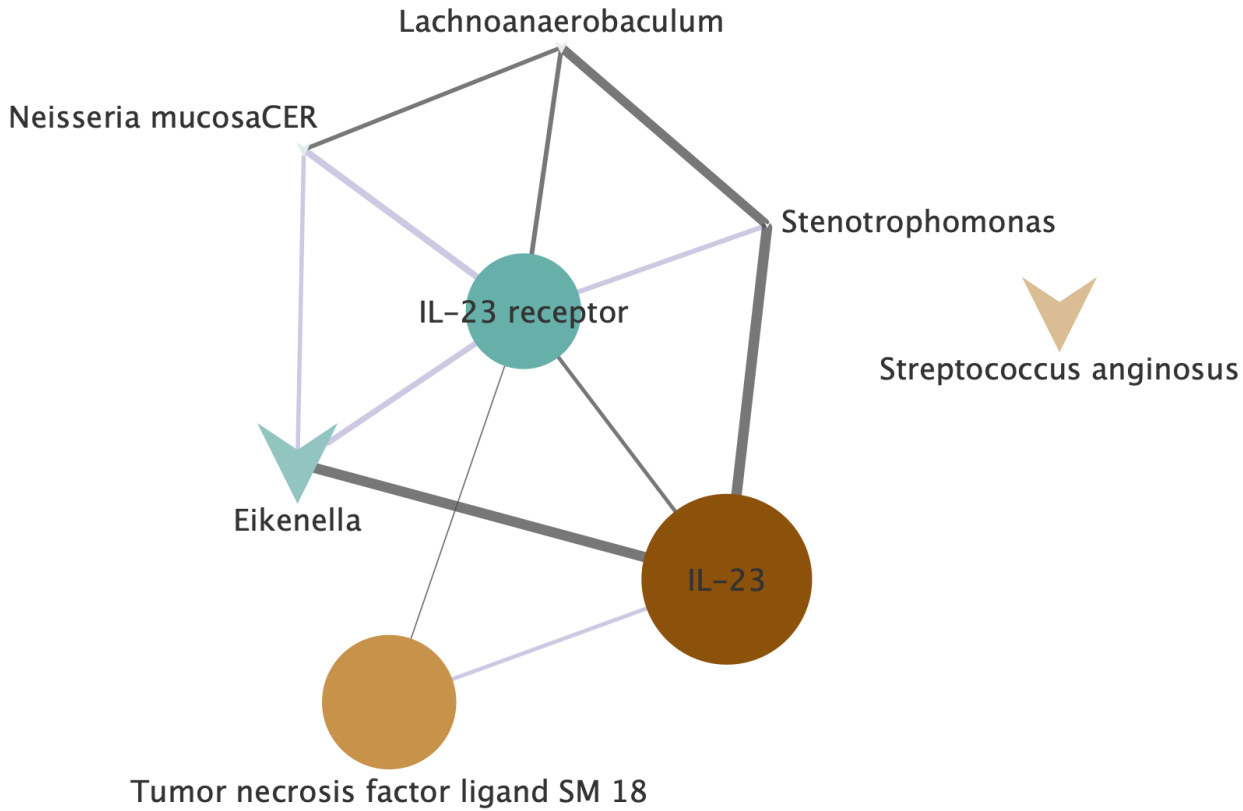
Subnetwork 2: GO:1903587

Node Count	Protein Count	Taxon Count	Network—%ΔPExS Corr	Node—%ΔPExS Corr Range
325	298	27	0.45	(-0.568, 0.623)



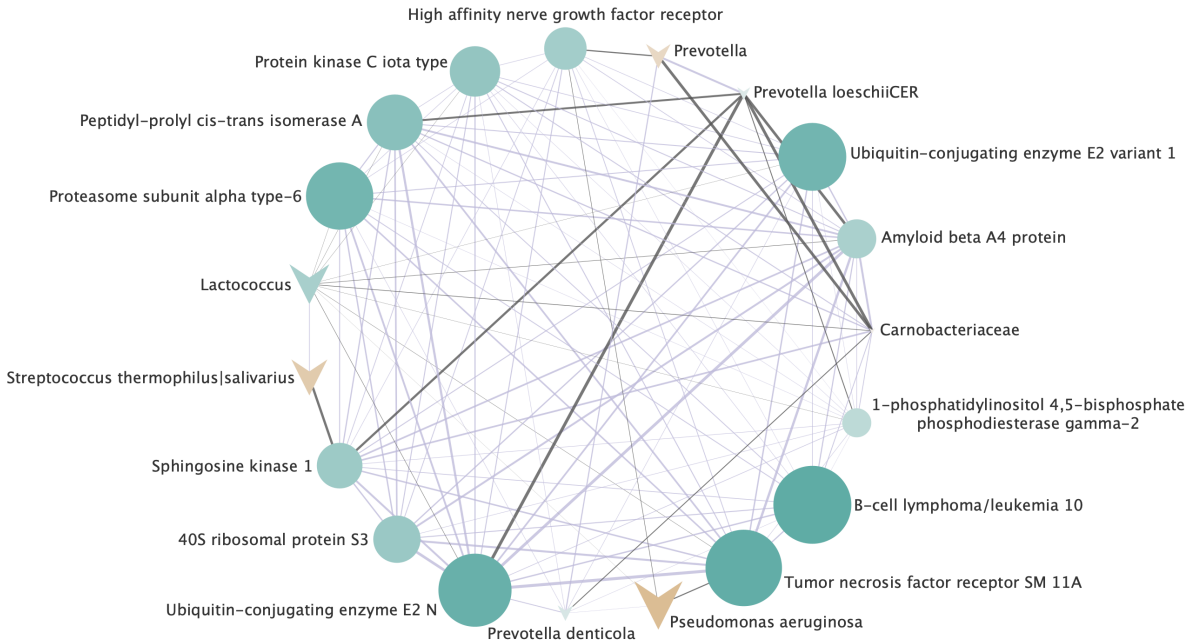
Subnetwork 4: GO:2000328

Node Count	Protein Count	Taxon Count	Network—%ΔPExS Corr	Node—%ΔPExS Corr Range
209	204	5	0.451	(-0.575, 0.634)



Subnetwork 5: GO:0051092

Node Count	Protein Count	Taxon Count	Network—%ΔPExS Corr	Node—%ΔPExS Corr Range
208	201	7	0.304	(-0.457, 0.474)



Subnetwork 6

Node Count	Protein Count	Taxon Count	Network—%ΔPExS Corr	Node—%ΔPExS Corr Range
98	96	2	0.331	(-0.501, 0.576)

Discussion Point 1

Discussion Point 2