Section 1:

This data is of Host-Microbial interactions in Idiopathic Pulmonary Fibrosis patients, and how they influence the progression of IPF disease in these patients. The observations are each patient gene sample, since there is a baseline sample and 4 samples (1 month, 3 months, 6 months, and a year) for 60 patients with IPF and 20 control patients. Gene expression profiles were generated using Affymetrix Human Gene 1.1 ST arrays. Comes out to 174 samples.

The features are each gene of the person in question, and how that gene is associated with an IPF diagnosis. 21661 features.

Some relevant baseline data is included in the metadata, such as FVC, age, and sex.

Research question: How do certain IPF gene modules correlate with IPF disease progression with respect to host-microbial interactions? I think that the researchers were considering that certain host-microbial interactions had a strong effect on IPF disease progression already, and it seems that hypothesis was confirmed from the data.

Another research question could be which host-microbial interactions positively affect IPF disease progression? This data already shows positive or negative correlation between interaction and disease progression, but it doesn't clarify if the correlation being positive or negative is significant. I think the positive interactions could be interesting if there are some.

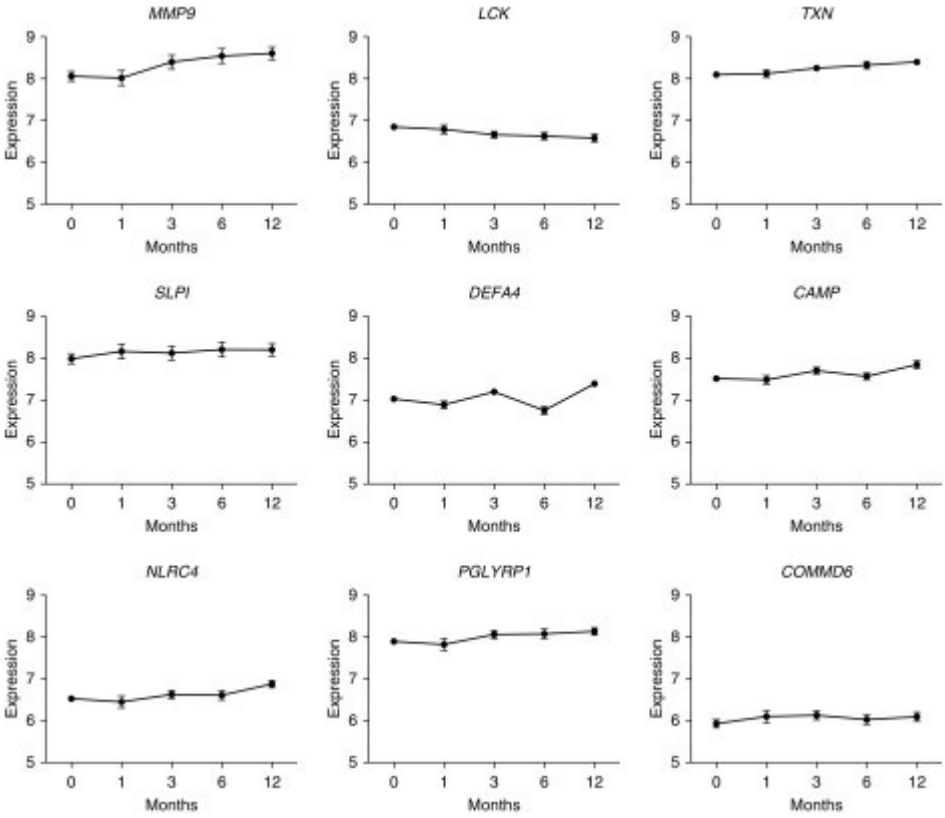
Section 2:

Module	Turquoise		Yellow		Brown		Blue		Green	
	Corr Coef	*p*	Corr Coef	*p*	Corr Coef	*p*	Corr Coef	*p*	Corr Coef	*p*
IPF	-0.53	***	-0.59	***	0.49	***	0.47	***	0.45	***
Survival	0.41		0.36	***	-0.31	**	-0.39	***	-0.31	**
MUC5B	-0.03		0.02		0.056		0.00		-0.06	
TOLLIP1	-0.01		-0.15		0.00		0.01		-0.02	
TOLLIP2	0.16		0.00		-0.01		-0.15		-0.19	
Blood Monocytes	0.14		0.13		-0.07		-0.17		-0.22	*
Blood Lympocytes	0.40	***	0.21		-0.12		-0.49	***	-0.36	***
Blood Neutrophils	-0.48	***	-0.32	**	0.22	*	0.56	***	0.43	***
Blood Eosinophils	0.24	*	0.20		-0.10		-0.22		-0.23	*
BAL Macrophages	0.13		-0.04		0.08		-0.17		-0.24	
BAL Lymphocytes	0.22		0.17		-0.08		-0.22		-0.04	
BAL Neutrophils	-0.37	***	-0.19		0.12		0.37	***	0.34	**
BAL Eosinophils	-0.19		0.07		-0.06		0.17		0.06	
Bacterial Burden	-0.24	*	0.09		-0.05		0.24	*	0.17	
Neiserria	0.18		-0.10		0.09		-0.26	*	-0.16	
Haemophilus	-0.05		0.01		0.03		0.02		-0.05	
Veillonella	-0.08		0.01		0.10		0.13		0.36	***
Streptococcus	-0.18		-0.04		-0.02		0.10		0.10	

First Figure: Figure 1: From Polyneaux et al. 2017

It appears that the module on the left refers to a set of genes, called a gene module, where BAL is associated with genes NLRC4, PGLYRP1, MMP9, and DEFA4. The color at the top making several columns likely splits up the sample sets, with it starting from baseline, to 1 month, to 3 months, 6 months, and 1 year if still alive. The correlation coefficient is likely calculated by averaging the correlation of IPF disease progression for each gene module.

This is likely the centerpiece of the paper, that says that the gene modules related to IPF, IPF survival, BAL neutrophils and Blood neutrophils are very significantly correlated with IPF progression.

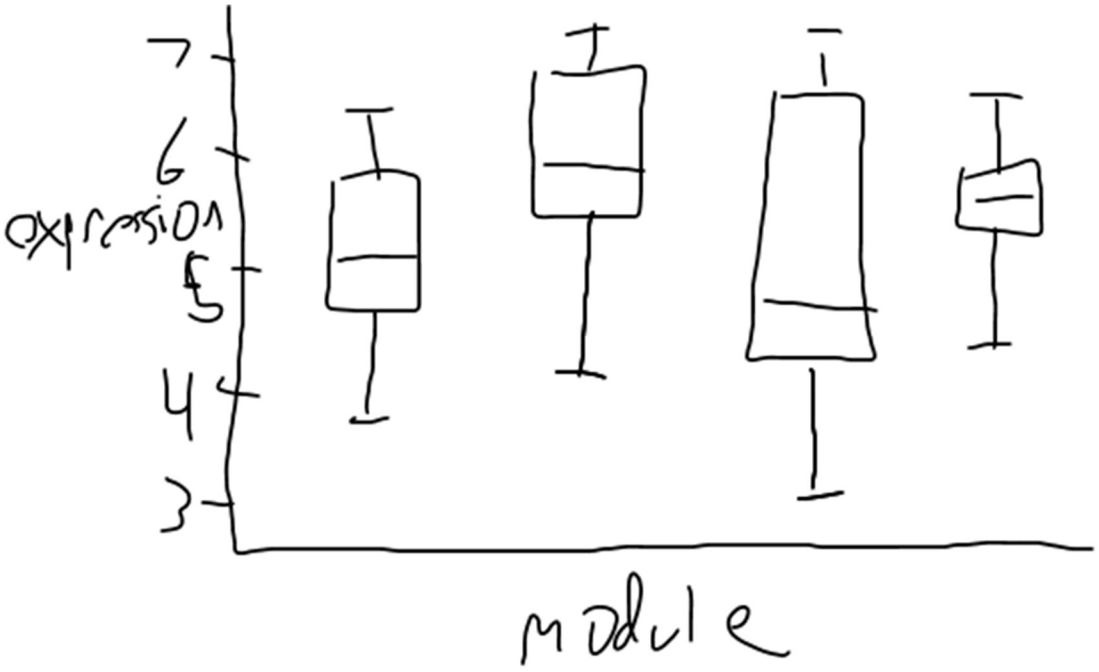


Second Figure: Figure 3: From Polyneaux et al. 2017

This figure shows the expression of certain genes over time from 0 months to 1 year across all patients in the IPF group. The x axis is the time period, and the y axis is gene expression level. Essentially this figure shows how gene expression changes over time with disease progression, and certain genes like LCK are negatively correlated while other genes, like MMP9 are positively correlated. The authors likely chose this figure to demonstrate how most genes are not significant in IPF disease progression, but certain genes show a noticeable correlation. If these genes are associated with the host-microbial gene module from the above table, then those host-microbial interactions have a strong correlation with

IPF disease progression, and from this figure, it is likely that MMP9 and LCK are both strongly correlated.

Another possible figure that could be made from this data could be a box plot for each gene module that shows how spread out and variable data from each gene could possibly be. The x axis would be gene module, and the y axis would be gene expression. Based on the above table, it would look something like this:



This model can tell us which genes have closely related and predictable expression, and which genes are very variable and inconsistent.

Section 3:

174 columns, 21661 rows. The file is a tsv so the delimiter is tab.