Integral Health
Data Science Challenge

Chengzhen Dai

May 2019

Summary and evaluation of the raw data

As a first step, we evaluate the data we are working with

There are 7 omic datasets, with varying number of features in each dataset (Figure 1)

	samples	features
Microbiome	68	18548
PlasmaSomalogic	68	1300
CellfreeRNA	68	37101
PlasmaLuminex	68	62
ImmuneSystem	68	534
Metabolomics	68	3485
SerumLuminex	68	62

The normalization/standardization of each dataset varies.

- Microbiome data seems to be z-score normalized (mean = 0, standard deviation = 1) within sample. This is an unusual method for normalization as microbiome data is typically normalized in relative abundances, log-ratio (EdgeR), or geometric mean (DESeq2).
- Metabolomic data also seems to be z-score normalized within a sample. This is also unusual as studies typically will z-score across samples for an individual metabolite

Analysis by batch and donor

Are there differences in omic profiles between donors and between trimesters?

To answer this question, I:

1. Calculated pairwise Euclidean distances* between samples to generate a pairwise distance matrix
2. Performed a PERMANOVA analysis to test for significant differences between donors and between trimesters

For most of the omic measures:

1. Samples from the same donors are more similar than samples from different donors, except for microbiome data ($P > 0.05$).
2. With the exception of PlasmaLuminex and SerumLuminex, samples from different trimester cluster separately

	pval_donor	pval_trimester
CellfreeRNA	0.019	0.024
ImmuneSystem	0.001	0.010
Metabolomics	0.004	0.013
Microbiome	0.915	0.007
PlasmaLuminex	0.001	0.199
PlasmaSomalogic	0.003	0.001
SerumLuminex	0.001	0.313

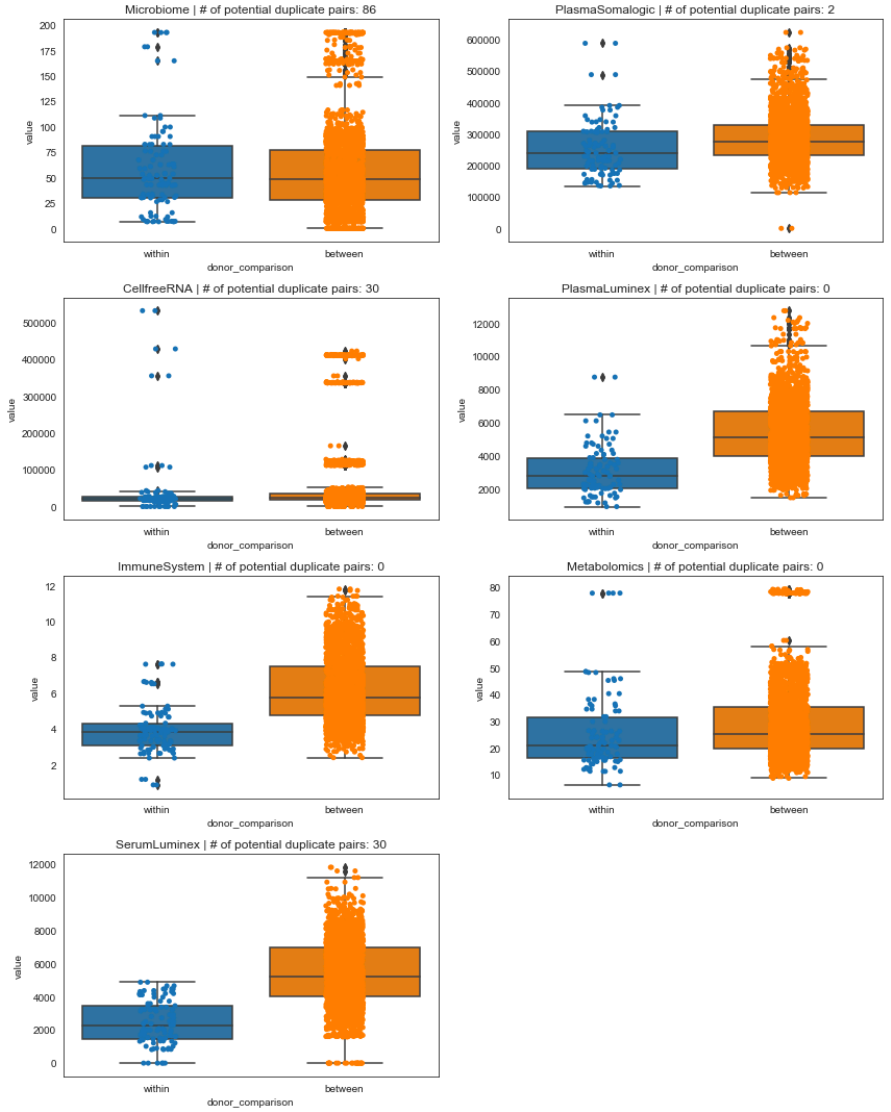
Note: The use Euclidean distances across all omic datasets was for convenience. Different omic measures often have different preferred methods for calculating distances (i.e. Bray-Curtis for microbiome)

Within-between sample comparisons

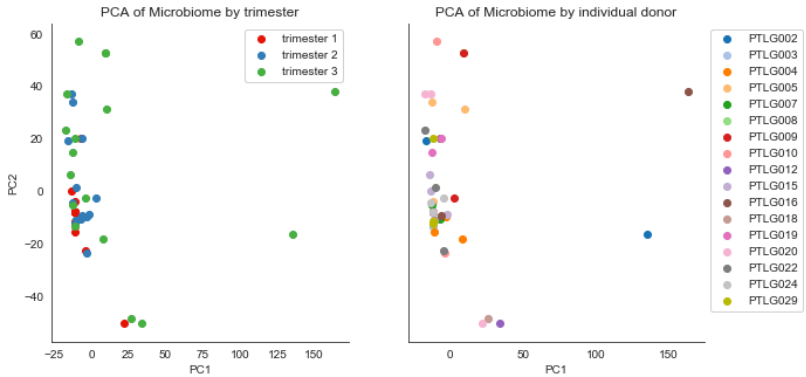
To visualize the within-sample and between-sample comparison, I generated boxplots and strip-plots for each omic measure

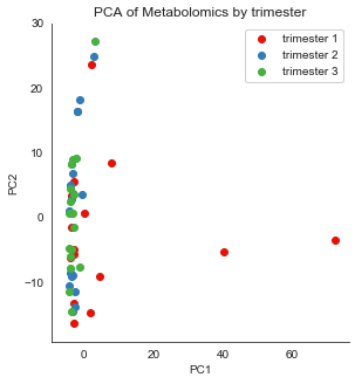
Some observations:

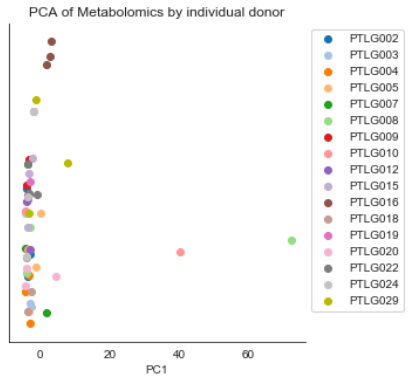
- There exists a number of sample pairs (from different samples) in which the Euclidean distance between samples are 0. This suggests potential duplicates as it's rare for samples to be this perfectly similar. The number of pairs are indicated in the title of each subplot
- There also seems to be samples that are outlier. These are particularly evident in the within-sample comparisons, as there are groups of sample pairs that have high Euclidean distances.
- Thus, we identify potential duplicates and outliers, but further analysis may be needed to confirm such inferences.

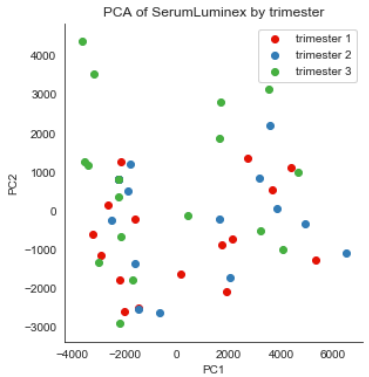


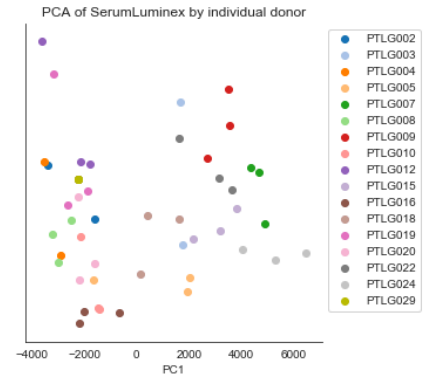
PCA analysis by batch and donor

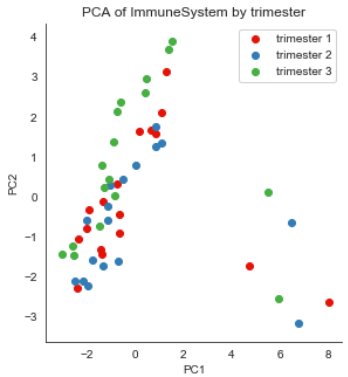


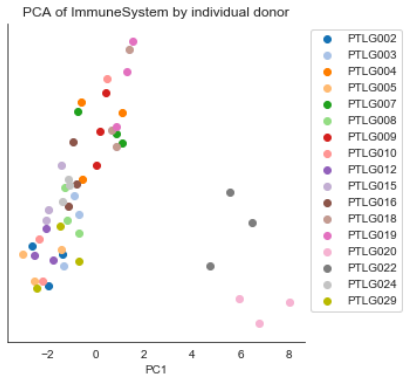


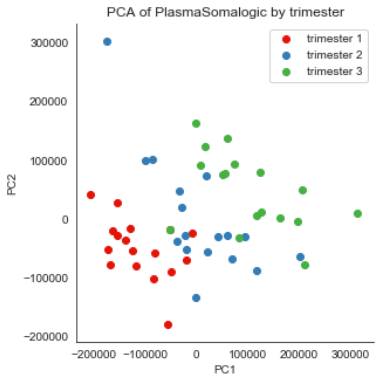


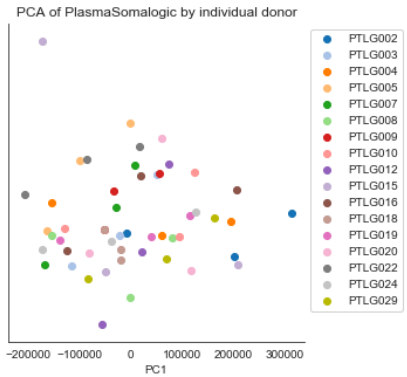


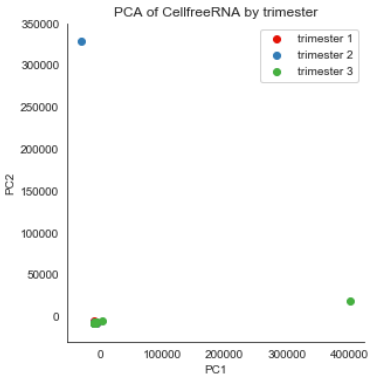


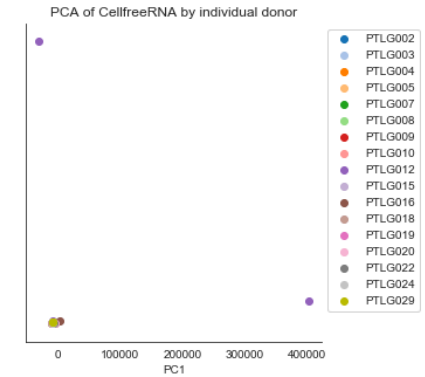


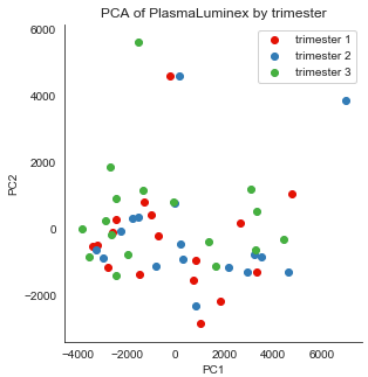


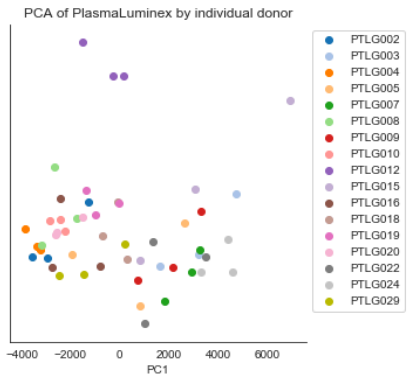












While PC1 and PC2 do not always explain the majority of variance, samples from the same donor does appear to be clustering together. Visually, the clustering of samples from the same trimester is strongest in PlasmaSomalogic. These PCA plots are similar to those from the original paper, suggesting replicability.

Feature selection via Elastic Net

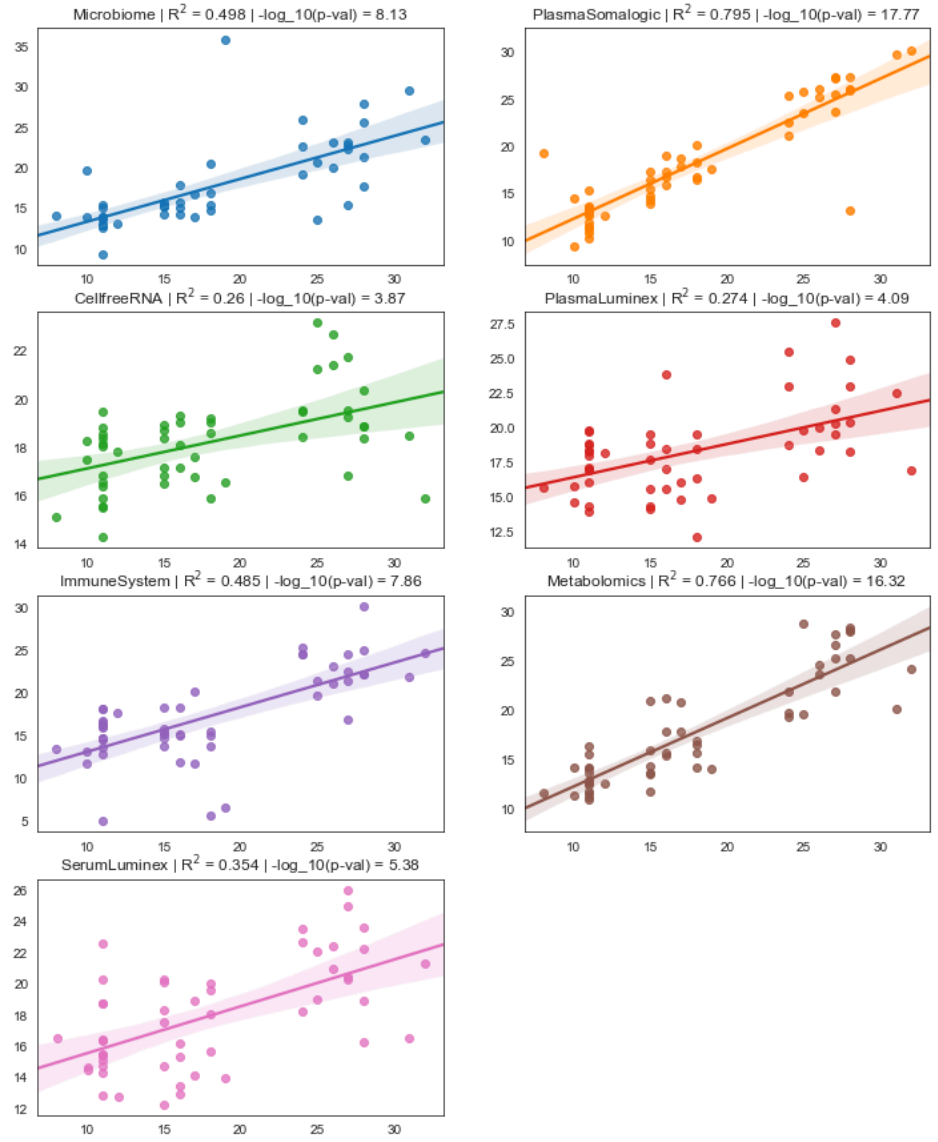
To understand how features were weighted in the Elastic Net models, I replicated the Elastic Net aspect of the study

I took a nested cross-validation approach in which:

- In the outer loop, all 3 samples from a single patient was withheld (leave-one-patient-out)
- In the inner loop, hyperparameter tuning was done using samples from the remaining 16 patients with 5-fold cross-validation
- The selected model was then used to predict gestation age on the withheld sample

The results were similar to the original study:

- Certain modalities were better (e.g. Metabolomics, SerumLuminex) while others were worse (e.g. ImmuneSystem, Microbiome, PlasmaLuminex)
- I found the use of –log(pvalue) in the original study as a method to evaluate the performance of the model. An R^2 value is more interpretable



Feature selection via Elastic Net

Across all models 1691 features were used.

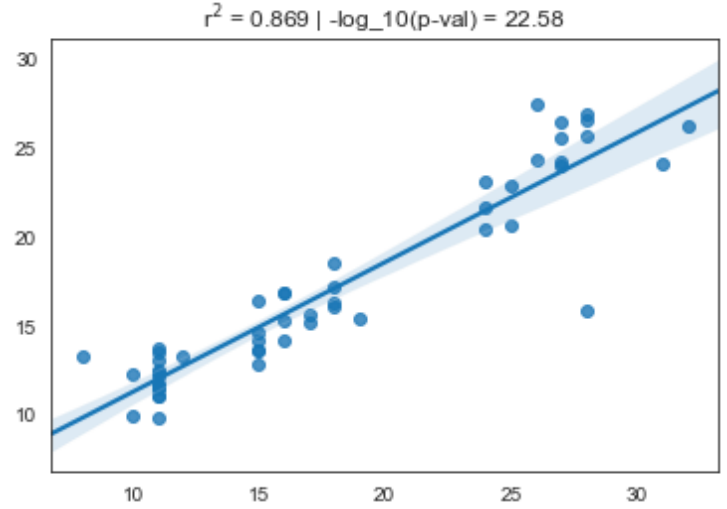
	r_squared	num_features_used	p-value
CellfreeRNA	0.259693	694	1.337641e-04
ImmuneSystem	0.484916	272	1.393723e-08
Metabolomics	0.765576	236	4.732843e-17
Microbiome	0.497787	233	7.407106e-09
PlasmaLuminex	0.273615	29	8.209887e-05
PlasmaSomalogic	0.795202	204	1.697154e-18
SerumLuminex	0.353744	23	4.176300e-06

The number of features in the Elastic Net model varies by modality

- EN models for CellfreeRNA and Microbiome used proportionally less features than other omics

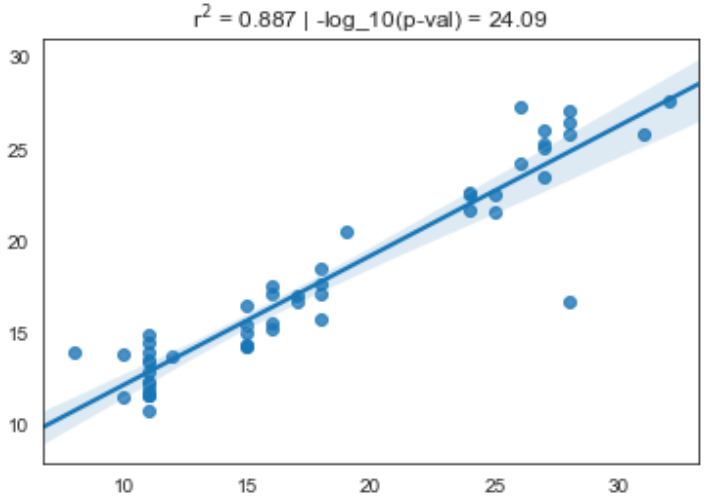
I next tried to see if using just these 1691 features in a single EN model would allow for greater predictability (i.e. combine all feature tables and subset to these 1691 features):

- The thought is that fitting a model between features from different omic datasets, one can achieve greater predictability than using one dataset at a time and then combining it.
- R^2 = 0.869; -log(p-value) = 22.58
- The -log(p-value) suggest that this approach is comparable to the stacked generalization method



Prediction of gestation age using Elastic Net model with all omic data at once

As an alternative to stack generalization, I explored the application of training using Elastic Net models on all of the omic data together (rather than separate datasets).



All feature tables were joined into one table and Elastic Net models were built using the same nested CV approach as before.

The prediction results suggest:

- R^2 = 0.887; -log(p-value) = 24.09; 1543 features were used
- This approach performs equally, if not better, than the stack generalization approach (based on p-value comparisons) and is likely less computationally intensive