Predicting Organ Cancer Probability: An Gene Expression Level Approach

1 Topic questions (Introduction)

Researchers have been devoting years and years to understand the correlation/causality between genes' expression level and the possibility of being caught by cancer. In this data challenge, we tried to restrict our scope to the following three questions:

1. How genes expression level on a particular organ associates with organs cancer occurrence?
2. What genes contributes the most to a specific cancer (e.g. brain cancer)?
3. For the genes that we identified, how are they collaborating with each other in affecting ones chance in being caught by cancer?

2 Non-technical Executive Summary

According the TCGA dataset, we investigate eight types of cancers located in the following organs: breast, head and neck, kidney, brain, lung, and prostate, thyroid, uterus. For each type of cancer, we build a model comparing the gene expression levels on healthy individuals (GTEx) and cancer-positive individuals (TCGA).

From our predictive model (details can be found in the next section), we identify top 50 most predictive genes for cancer development in a particular organ. Our model unveils the following important results:

1. **The distribution of gene's expression are significantly different between healthy donor and cancer donor.** For all 8 organs, the samples from healthy and cancer donor are linear separable over the gene expression space, which make the classification problem a relatively easy one.
2. **A small fraction of genes contribute to most of the cancer probabilities.** Single gene's contribution to cancer probability drops sharply: Top 50 genes explains over 98.4% of the probability of being caught by cancer. And most of these genes can explain all the eight types of cancers.
3. **A lot of the gene expressions effect are joint effects.** Even though single genes contribution will be small after the first 50 most important genes, interaction term between genes can still be significant. That is to say, multiple genes together determines cancer probability.

With groups of genes that have the largest contributions to different cancers, we examine the GTEx dataset and look into the similarities and differences of the genes, such as gene type, chromosome position, and transcription characteristics etc. Mutations of genes in the future might put a person at higher risk of getting certain kinds of cancer.

Once the most predictive genes for a particular type of cancer are determined, we explore their gene expression levels on other cancer-negative organs. This generate additional information that further assist diagnosis. For example,

3 Technical Executive Summary

We build models using genes expression levels on an organ to predict whether that particular organ will develop cancer.

3.1 Modeling Data Preparation

We define a binary classification problem. The predictor x is a 1 x 5000 vector containing expression level values (FPKM) for each of the 5000 genes for the organ. For GTEx dataset containing health individuals, we divide the RPKM expression levels of cancer-positive individuals by 2, since its exactly double the value of FPKM. The response y takes value 1 if the organ is cancer-positive and 0 otherwise.

3.2 Model Building (LASSO Regression)

Since the predictor x is a sparse high-dimensional vector, we use L-1 regularized logistic regression to perform gene selection on the 5000 candidates. This regularization helps us obtain sparese parameterThis enhances both the accuracy and interpretability of our model. We choose cross entropy as our loss function because we have an inbalance modeling data.

3.3 Sensitivity Analysis and Complementary Gene Interaction on Phenotype

We rank the covariates according to the models sensitivity to the covariate. This provides information on which covariates are most important. The sensitivity of the model to a particular covariate is measured by the magnitude of the derivative of the model's predicted probability with respect to the covariate (averaged over the test data). Specifically, we calculate the sensitivity (with respect to j-th covariate) of the model's prediction as:

$$\text{Sensitivity(j)} = \frac{1}{N}\sum_{k=1}^{n}\Big|\frac{\partial p(x)}{\partial x_j}\Big|_{x=X_k} \tag{1}$$

We also rank pairs of covariates by calculate the sensitivity with respect to i-th and j-th mixed covariates of the models prediction as:

$$\text{Sensitivity(i,j)} = \frac{1}{N}\sum_{k=1}^{n}\Big|\frac{\partial^2 p(x)}{\partial x_i \partial x_j}\Big|_{x=X_k} \tag{2}$$

The partial derivative can be estimated using finite difference scheme:

$$\frac{\partial p(x)}{\partial x_j} \approx p(x+\Delta_j) - p(x) \tag{3}$$

$$\frac{\partial^2 p(x)}{\partial x_i \partial x_j} \approx p(x+\Delta_{ij}) - p(x+\Delta_i) - p(x+\Delta_j) + p(x) \tag{4}$$

Figure 1 shows the decreasing importance of genes from the most important to least important. We can see that the importance drops sharply and only about first 50 genes have importance greater than 0.00001. Figure 1 also shows the importance of interaction terms in the model. We can see that the importance drops much slower and interaction terms still have significant contribution after the first hundred gene pairs.

Thus, we discovered that a group of two or more different genes often work together to create a specific phenotype (and in our case, to determine an individuals susceptibility to certain types of cancers).

For genes that are most likely to cause brain cancer, we also calculate the correlation between gene expression level in brain and other tissues and gene expression level in other tissues. This suggests some non-related organs may also reveal risk for a particular type of cancer. For example, heart has the highest correlation with "brain": 0.89251896558686272. Thus, heart-related symptoms may foreshadow dangers for brain.

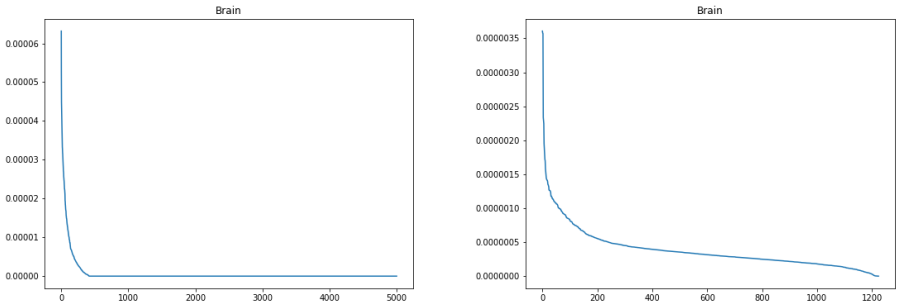
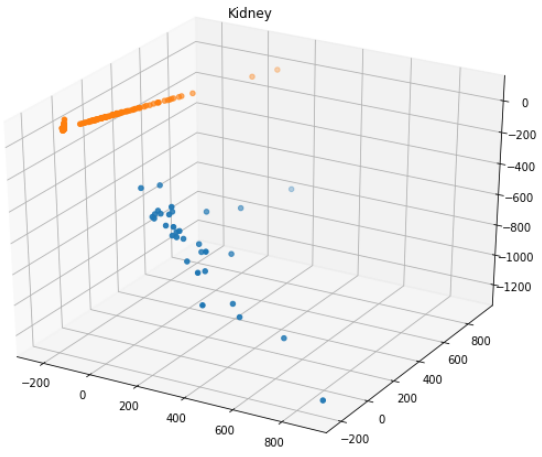


Figure 1: Marginal Effects of the largest k-th gene/gene pairs, Brain; LHS: Single gene; RHS: Gene pairs

3.4 Model Result, Interpretation and Suspect Genes Identification

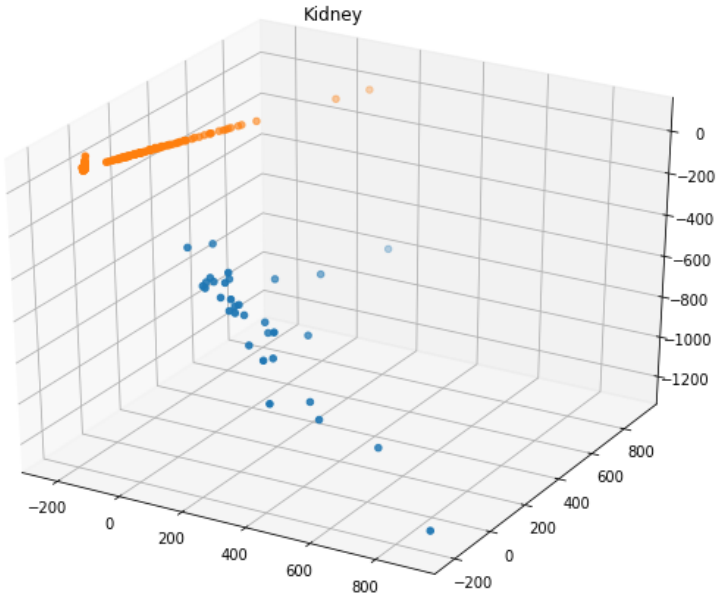
Our model achieved almost 100% accuracy. Digging into the dataset, we plot the dataset on the first three directions generated from principal component analysis (PCA) where orange dots represents cancer-positive samples and blue dots are cancer-negative samples. The linear separability of the two groups proves the reasonableness of our classification approach.

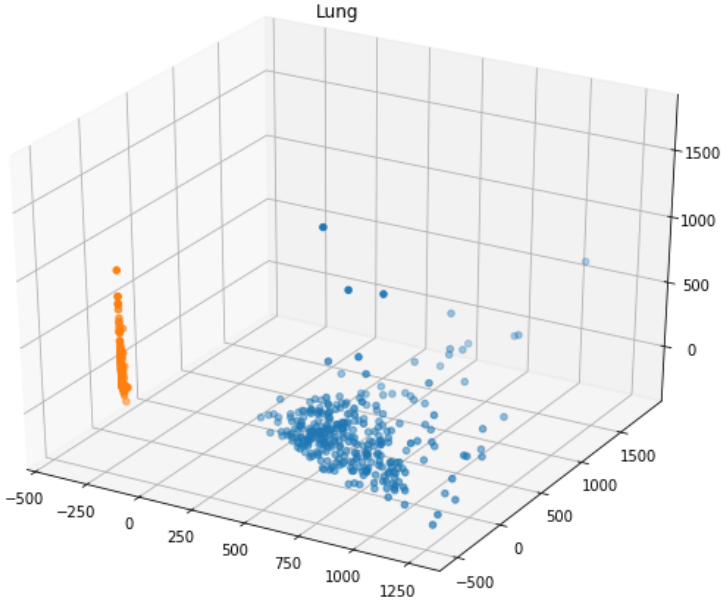


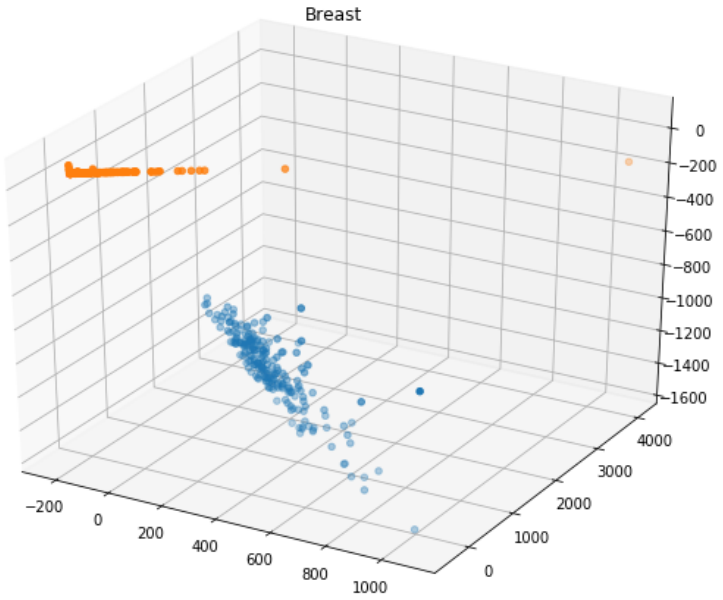
A Appendix

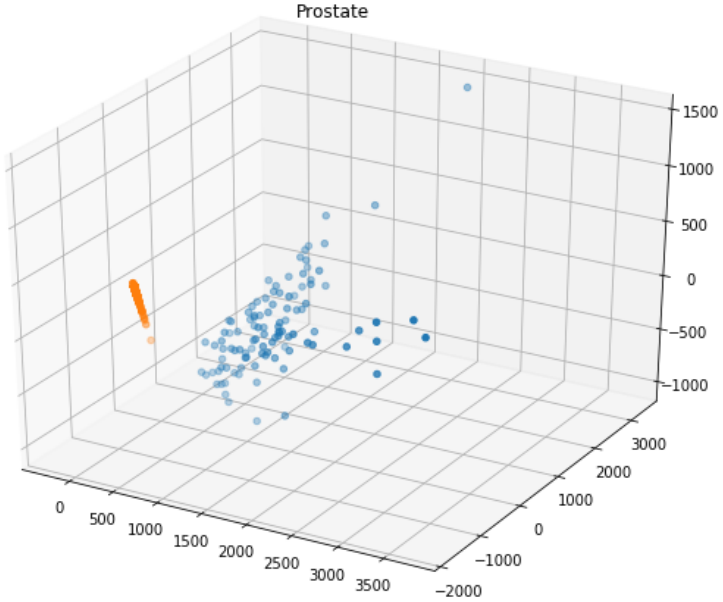
A.1 PCA Decomposition of Gene's Expression

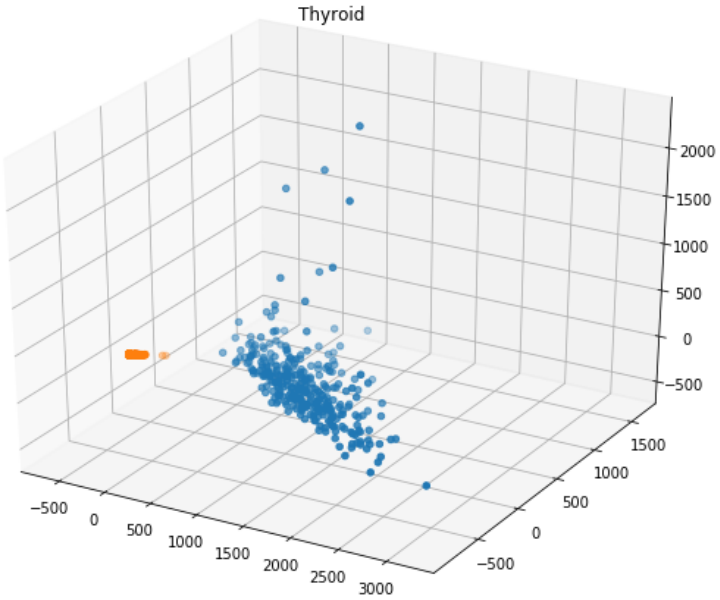
We use orange points to denote cancer-positive samples and blue points to denote healthy samples. Each points are projected onto the first three PCAs.

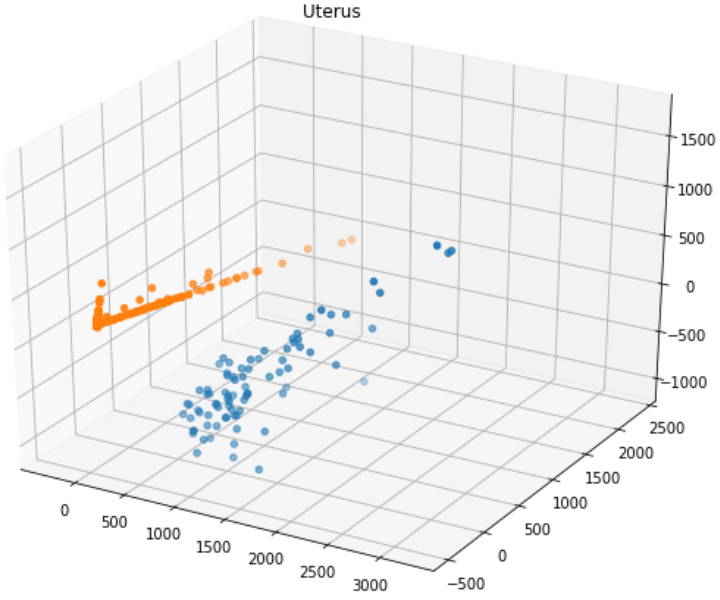





Breast
0
−200
−400
−600
−800
−1000
−1200
−1400
−1600
4000
3000
2000
1000
0
−200
0
200
400
600
800
1000


Prostate
1500
1000
500
0
−500
−1000
3000
2000
1000
0
−1000
−2000
0
500
1000
1500
2000
2500
3000
3500


Thyroid
2000
1500
1000
500
0
−500
1500
1000
500
0
−500
−500
0
500
1000
1500
2000
2500
3000


Uterus
1500
1000
500
0
−500
−1000
2500
2000
1500
1000
500
0
−500
0
500
1000
1500
2000
2500
3000

A.2 Marginal Effects of Single Genes and Gene pairs

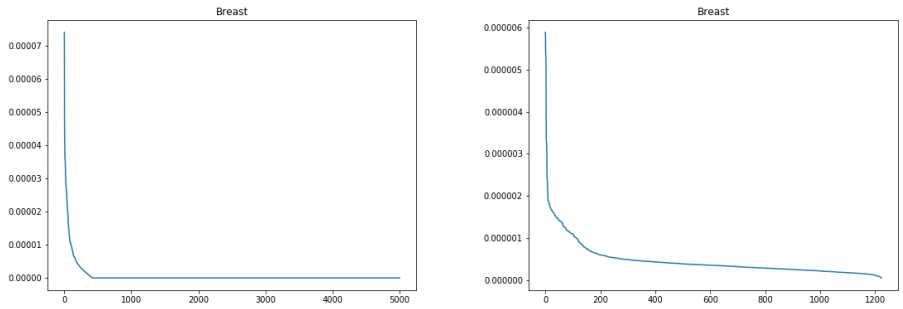


Figure 2: Marginal Effects of the largest k-th gene/gene pairs, Breast; LHS: Single gene; RHS: Gene pairs

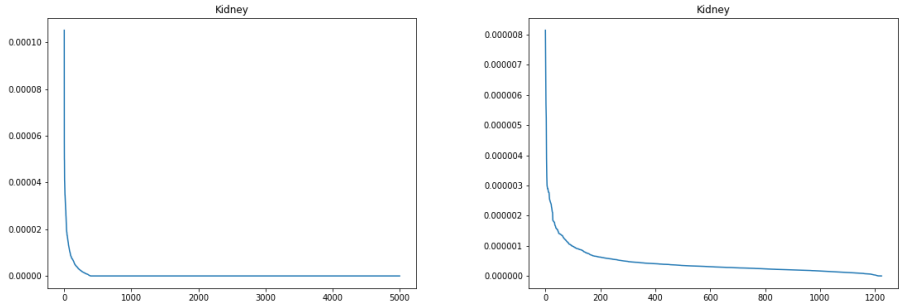


Figure 3: Marginal Effects of the largest k-th gene/gene pairs, Kidney; LHS: Single gene; RHS: Gene pairs

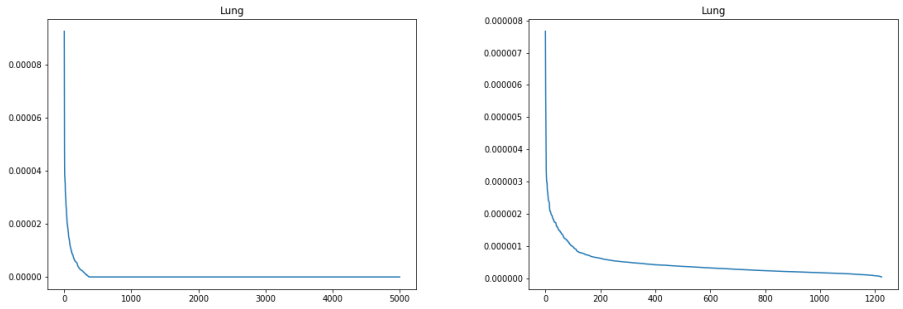


Figure 4: Marginal Effects of the largest k-th gene/gene pairs, Lung; LHS: Single gene; RHS: Gene pairs

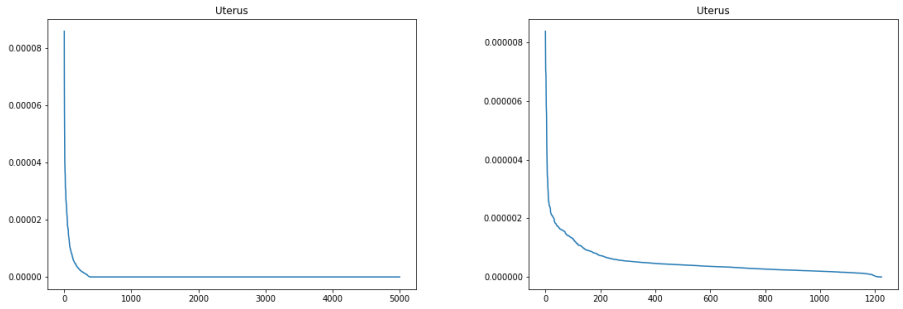


Figure 5: Marginal Effects of the largest k-th gene/gene pairs, Uterus; LHS: Single gene; RHS: Gene pairs

B Reference

[1] Chu, Minjie, Ruyang Zhang, Yang Zhao, Chen Wu, Huan Guo, Baosen Zhou, Jiachun Lu et al. ”A genome-wide genegene interaction analysis identifies an epistatic gene pair for lung cancer susceptibility in Han Chinese.” Carcinogenesis 35, no. 3 (2013): 572-577