Which tissue is the best predictor for cancer?

Citadel CMU Datathon powered by Correlation One

Yijie Wang, Yujie Wang, Jiaqi He, Taiyan Liu

September 16, 2017

1 Topic Question

Cancer has a major impact on society in the United States and across the world. About 1 of every 4 deaths in the US is caused by cancer. Approximately 39.6% of men and women will be diagnosed with cancer at some point during their lifetimes according to the report from American Cancer Society [1]. However, another more distressing fact is that early cancer has no obvious symptoms and is often found in late stages. Therefore, it is worth investigating in predicting cancer in early stages with minimum cost and damage for human body. Right now, our best method for detecting cancer is a biopsy - cutting out a small piece of the tumour tissue for lab analysis. But biopsies are often painful and invasive, and you need to already have a tumour or at least a suspect tumour to cut something out of it. In the Genotype-Tissue Expression project, all tissue samples are extracted from donated body. On average, 15 tissue samples are extracted from every post-mortem donor. But we just cannot take 15 tissue samples from a living person! Because the process is invasive and dangerous. Can we minimize the number of tissue samples required to make useful decision? Specifically, can we use only several tissue samples, instead of 15, to decide if a person is healthy? In this project, we would like to investigate the relationships between gene expression values obtained from specific tissues and the possibility of getting the cancer. Is there any specific tissue that can efficiently predict the probability of a person getting cancer?

2 Non-Technical Summary

The goal of this project contains three perspectives: First, we would like to investigate whether abnormal gene expression values (RPKM and FPKM) could be an indicator of cancer. Second, we would like to analyze the correlation between gene expression values and the sampling position (organs/tissues), since we hypothesized that different sampling position might affect the gene expression level. Third, by building the Naive Bayes model that tells us whether a person is likely to get cancer from the gene expression value at a specific organ/tissue, we could tell which organ/tissue could provide the most useful information for cancer diagnosis, reducing the pain and cost of biopsy sampling.

After conducting the significance analysis on different gene-organ pairs, results have shown that the breast mam tissue is the most significant indicator for cancer, while the thyroid and brain also play important roles in cancer prediction. This is reasonable since the correctness of breast cancer prediction rates are much higher than other cancers right now. Also, we can notice that there are some gene-organ pairs that are way more valuable than others. Therefore, this gives us a insight that these genes might be related to the root cause of cancer. The medical researcher should pay more attention on these combinations.

3 Technical Executive Summary

1. Preprocessing
   The first step of preprocessing is to remove outliers from the dataset. We could visualize how the FPKM/RPKM ranges over different organs as Figure 1 and used the 99% quantile as a threshold to filter those abnormal values.

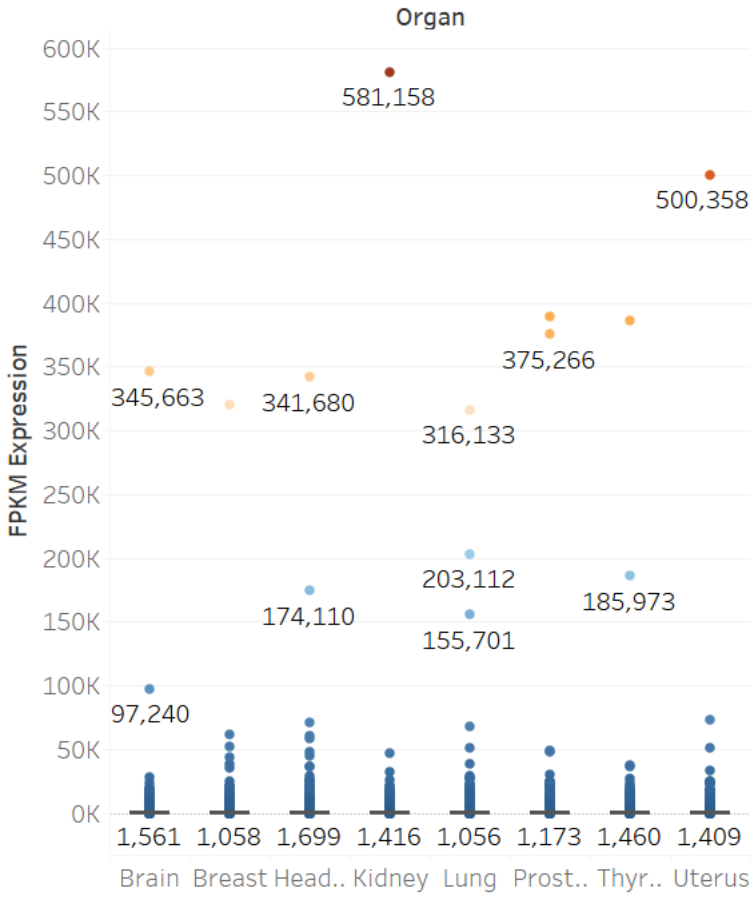
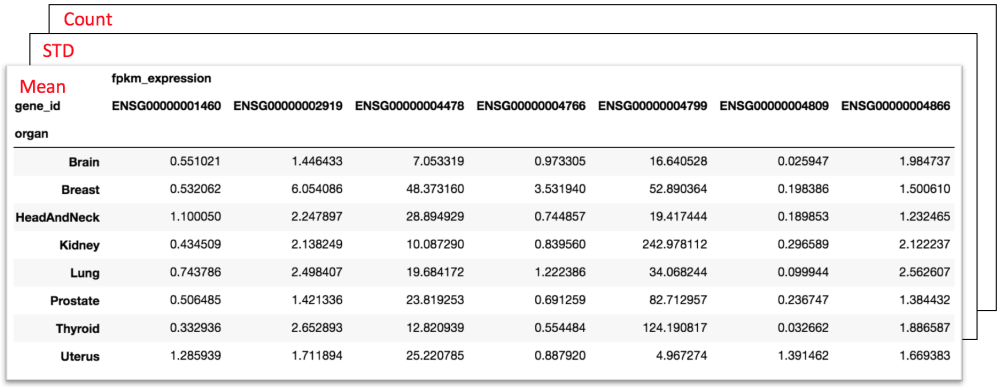


Figure 1: Gene expression value distribution over different organs

Notice that the data from GTEX and the data from TCGA have different formats. The first thing we need to do is to find a bridge to connect them together. We first selected the shared $gene_id$ and $organ$ between two tables, which has 4691 genes and 7 organs in total. Then for each table, the $(gene\_id, organ) : gene\_expression\_value$ key-value pair was chosen as a feature and we calculated the mean, standard deviation and count for each combination of gene id and organ to form the feature matrix of the Naive-Bayes model. A typical feature matrix is shown as figure 2:



	fpkm_expression						
gene_id	ENSG00000001460	ENSG00000002919	ENSG00000004478	ENSG00000004766	ENSG00000004799	ENSG00000004809	ENSG00000004866
organ							
Brain	0.551021	1.446433	7.053319	0.973305	16.640528	0.025947	1.984737
Breast	0.532062	6.054086	48.373160	3.531940	52.890364	0.198386	1.500610
HeadAndNeck	1.100050	2.247897	28.894929	0.744857	19.417444	0.189853	1.232465
Kidney	0.434509	2.138249	10.087290	0.839560	242.978112	0.296589	2.122237
Lung	0.743786	2.498407	19.684172	1.222386	34.068244	0.099944	2.562607
Prostate	0.506485	1.421336	23.819253	0.691259	82.712957	0.236747	1.384432
Thyroid	0.332936	2.652893	12.820939	0.554484	124.190817	0.032662	1.886587
Uterus	1.285939	1.711894	25.220785	0.887920	4.967274	1.391462	1.669383

Figure 2: Feature matrix for Naive Bayes Model

Since RPKM and FPKM are two different gene expression values, to compare the magnitude of them, we need to normalize the expression values. Here we use a simple assumption which might not reflect the actual situation: RPKM = FPKM/2. To validate whether the assumption is reasonable, we compared the average value between RPKM(5.02) and FPKM(8.58) which are really close to the assumption.

2. Correlation Investigation
   To validate our hypothesis, we first evaluated whether different sampling positions affects the gene expression values. Therefore, we randomly sampled 50 gene ids and checked their expression values at different organs. The figure below showed that the variation of FPKM/RPKM at different organs are significant. Therefore, it's reasonable for us to use the organ-gene expression value pair as a feature to predict the probability of having a cancer.

   We then validated whether there was a significant difference of gene expression value between positive samples(have cancer) and negative samples(healthy).

3. Model Design Here we proposed our significance analysis module. In this project, we used the concept from Naive Bayes Model to show that given the gene expression value $x_{i,j}$ of the gene i from a specific tissue j, we can tell the log-likelihood ratio between whether the person is healthy or the person has

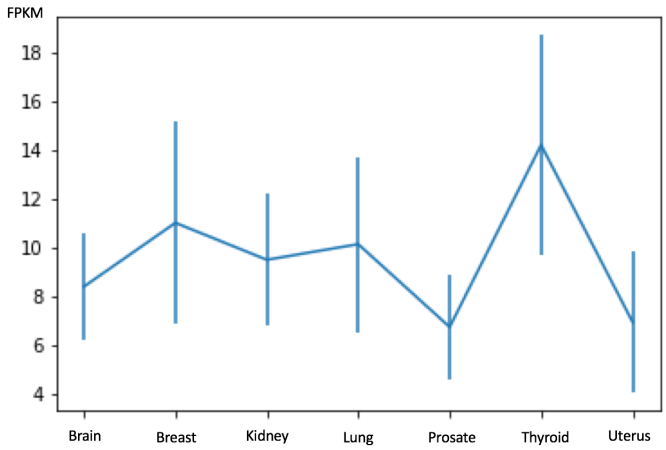


Figure 3: FPKM of Gene ENSG00000004897 over different organs

cancer. The significance ratio of a tissue β is shown below:

$$\beta_j = \log(\frac{P(\mathbf{x_j}|y = \text{healthy})}{P(\mathbf{x_j}|y = \text{cancer})})$$

The probability above indicated how important a tissue sample is for predicting cancers. The reason why we designed a model like this was because though we had many negative samples from the GTEX dataset, there was no information about which sample came from whom in the TCGA dataset. Due to the lack of

After traversing the 40 million data, the result was shown as figure 4.

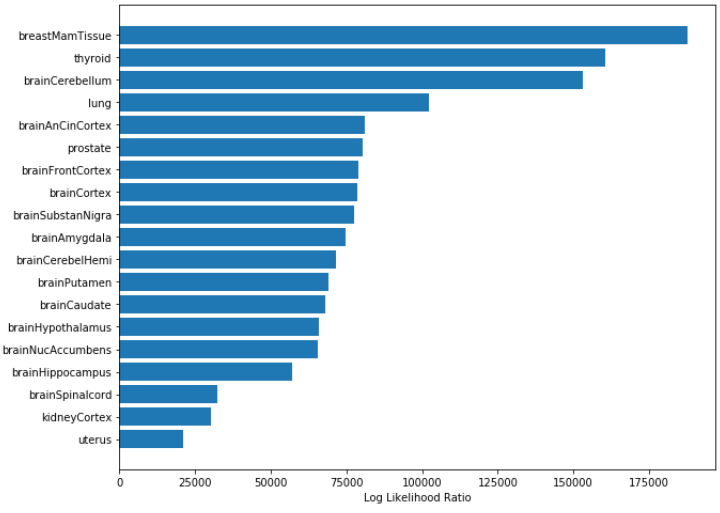


Figure 4: The significance of tissue when predicting the cancer

From figure 3 we observed that the breast mam tissue is the most informative for predicting cancer, which means, if you'd like to check whether yourself is

likely to have a cancer by using the RNA sequencing technique, the best choice is to obtain a tissue sample from breast. Thyroid, brain cerebellum, and lung are also good places for sampling as well. The rest tissues won't make much difference.

Next we were also interested in knowing which gene-organ pair was the best cancer indicator in our feature. Therefore, we conducted a symmetric K-L divergence analysis to select the most important gene-organ pair. From Figure 5 we can tell that the there are several gene-organ pairs are of great importance for cancer prediction. The result also matched the tissue analysis above since the most informative genes lies in breast, brain, and thyroid respectively. We should pay more attention to the expression value change of these genes in the further study.

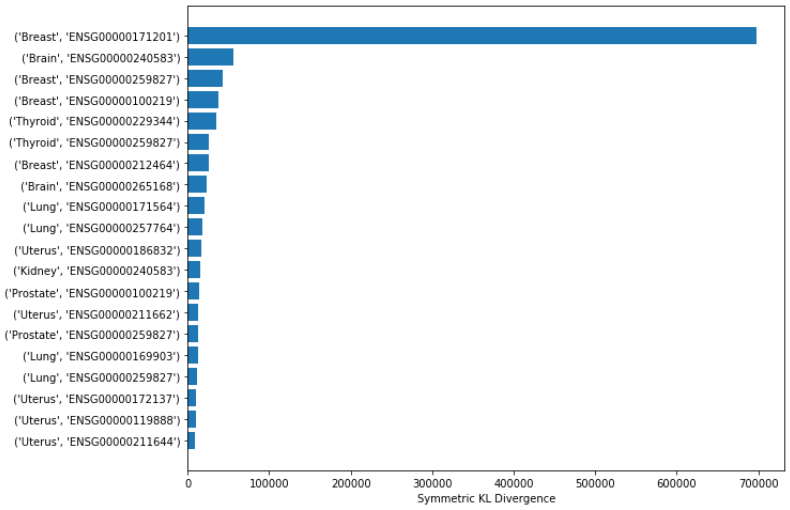


Figure 5: The significance of tissue when predicting the cancer

References

[1] American Cancer Society, *Cancer Facts & Figures*, Atlanta, Georgia, 2016.

[2] Grossman, Robert L., Heath, Alllison P., Ferretti, Vincent, Varmus, Harold E. Lowy, Dougla R., Kibbe, Warren A., Staudt, Louis M., *Toward a Shared Vision for Cancer Genomic Data. New England Journal of Medicine 375:12, 1109-1112*, 2016.