Biomarker Discovery for Kidney Cancer

Mehrdad Bakhtiari

1 Introduction

Kidney cancer is a type of cancer that starts from a cell in kidney. In united states, 2.5% of the cancer deaths is caused by different types of kidney cancer. [3]. In this study, we search for the biomarkers in cancer patients that predict the survival of the patient. Here, we study three types of kidney cancers: a) kidney renal cell carcinoma (KIRC) that is responsible for 80% of the kidney cancers, b) kidney renal papillary cell carcinoma (KIRP), and c) Chromophobe Renal Cell Carcinoma (KICH).

2 Method

Data sources: We used Pan-kidney cohort (KICH+KIRC+KIRP) data downloaded from Broad GDAC Firehose for our analysis. We used the gene expression data for 20533 genes on 973 individuals. For assessing the survival durations, we used days_to_death when available, otherwise we used days_to_last_followup which could be seen as equivalent of survival since it is the last alive timepoint that we are aware of.

Filtering: We eliminate the genes where at least half of them have zero expression. In those cases, lack of expression could be due to error in the data and also anyone with minimum expression will be grouped in to high expression since half of the individuals don’t have expression.

Gene finding: For each filtered gene, we divide the individuals based on their expressions on that gene. If expression is higher than median it belongs to group 1 and otherwise it belongs to group 2. Then, we used Kaplan-Meier estimator to find the effect of expression of gene on the survival prediction. In this graph, x-axis shows the time and y-axis shows the survived individuals of each group and we search for the genes that the difference of survived individuals is significant.

Reducing the noise: When we study the gene expression data of two groups, we expect to see one Gaussian distribution for individuals with high expression and one Gaussian distribution for individuals with low expression such that one of this distributions is predictor of survival. However, this two distributions are not completely separated and we don’t know the underlying distribution for each data point (expression). So, dividing the data points from the middle will lead to having many false assignment of points to the group which affects our analysis. To overcome this problem, we divide the expressions into three percentiles (0%-33%, 33%-66%, and 66%-100%) and eliminate the middle group which cannot be assigned to a group confidently. Then, we use two other groups for our analysis.
Figure 1: Biomarker Genes. (A) Shows high expression of IMPA2 gene predicts longer survival. (B) Shows low expression of INTS8 gene predicts longer survival.

Figure 2: Biological relevance of genes. (A) Shows expression of IMPA2 in different tissues. (B) Shows INTS8 is highly expressed in different cancer tissues including three types of kidney cancer that we study.

Boolean Analysis of Genes: Previous works has shown the boolean analysis of gene expression data is useful to find biomarkers in prostate cancers. Here, we use a similar idea to consider the joint expression of genes to study the survival time analysis instead of considering single gene at a time.

3 Results

Gene finding: By looking at top genes found by Kaplan-Meier and computed p-value, we found two genes that are biologically relevant. The Kaplan-Meier estimate is shown in Figure 1 for these two genes. IMPA2 is highly expressed gene in kidney tissues as shown in Figure 2A. Transcriptome analysis shows INTS8 is highly mutated in specific cancers. In addition, it is known cancer marker in many cancers including kidney renal clear cell carcinoma with the sign of over-expression which is shown in Figure 2B.

Reducing the noise: By eliminating the mid-range expression data, we find ZIC2 (Figure 3A) which is a known biomarker for kidney cancer and is over-expressed in other types of kidney tumors.

Boolean Analysis of Genes: We evaluated every possible pairs of top 20 genes (190 pairs of genes), and for each pair, we tested the separation for 4 possible cases (low expression in both genes, high expression in both genes, and low/high expression combinations). Figure 3B shows the boolean analysis on best pair of genes in terms of separation. The best case happens for combination of IMPA2 and INTS8 which previously shown as significant genes, and here we show they could be found easier without evaluating many top genes to find biologically relevant genes.
Figure 3: Improved Gene Finding. (A) Shows expression of ZIC gene can predict the survival more accurately. We found this gene by eliminating the data with mid-range expression. (B) Shows the joint expression of INTS8 and IMPA2 genes is stronger predictor of survival compare to predictions based on single genes.

References


