Project 1, Phase 3

Agnieszka Sitko, Annanina Koster, Piotr Obarski

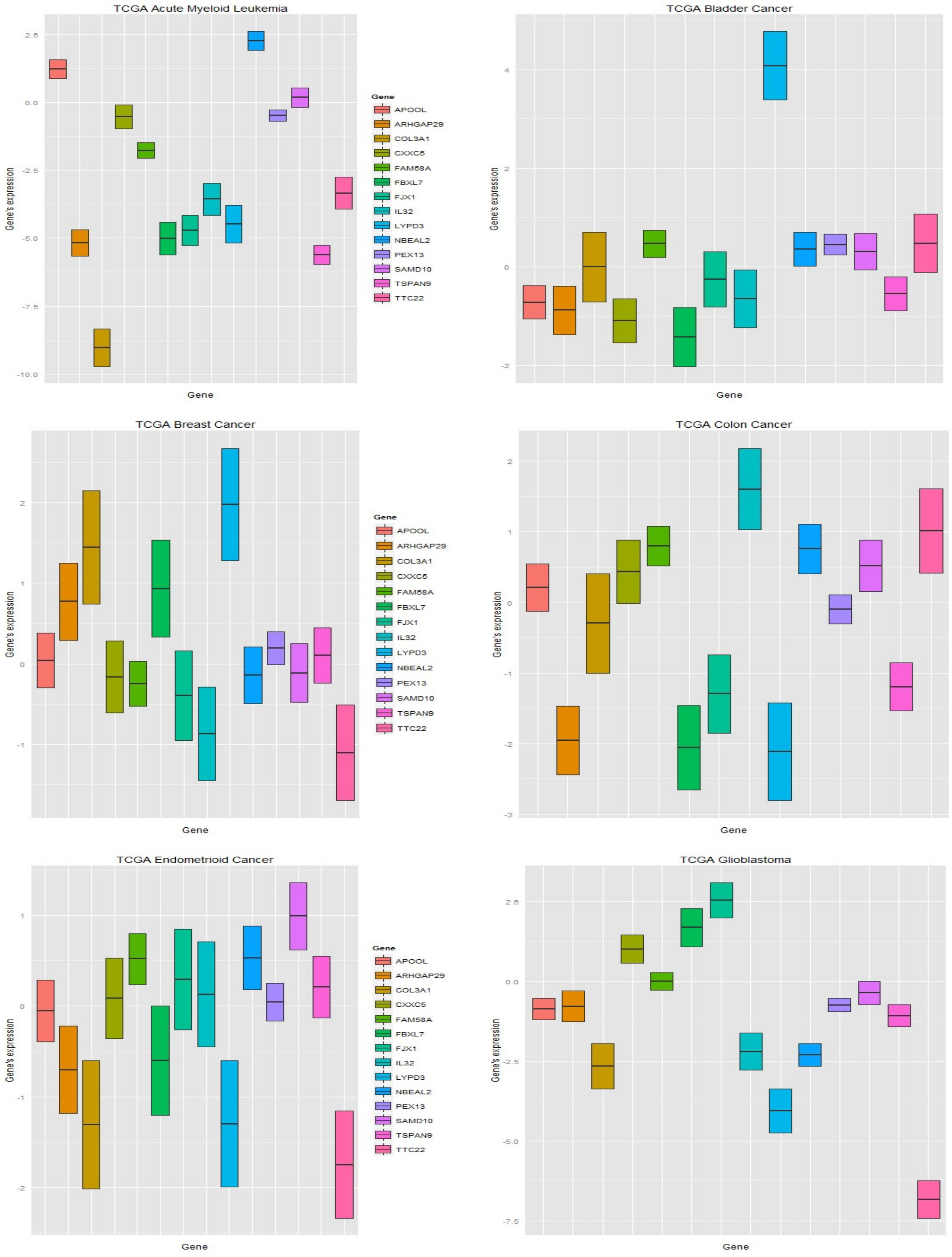
Strategy of choosing the genes and summary of pervious phases

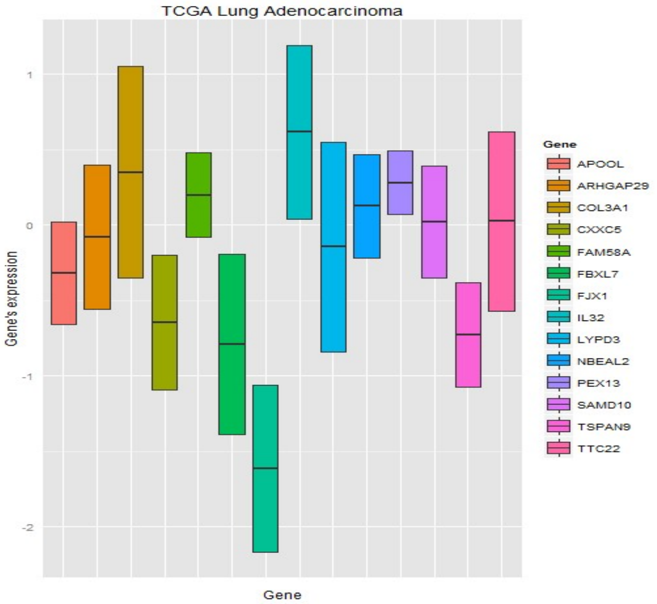
We were given data set with around 3600 patients and 16 thousnds genes. Every patient had some cohort of cancer, which subsequently we will call cancer type. We decided to divide patients according to cancer type (so the patients form groups) and for each gene tested whether there is any difference in means of gene's expression in the groups. We have done it with ANOVA. Afterwards we performed so-called post-hoc test (we used Scheffe test) to check what means in groups are and to get Critical Difference, i.e. value of maximal difference of two means to consider those means to be statistically equal. Then groups were created in the following way: two cancer types are in the same group if the means are statistically equal. Finally, we chosen, and consider as important, genes which have the highest number of groups. Why? Because we see each group as an information, and taking such a genes we maximaze informations while minimizing number of genes. To explain it better, consider an example. Assume we have 5 (we are going to denote it by roman numbers) cancer types. We are given a patient with 3 genes - A, B, C. Firstly, we consider gene A. Let's say that expression of gene A has a value such that according to our post hoc test, patient is likely to have cancer type I, III or IV. So then we consider gene B and we are interested only in cancer types I, III or IV. So it may turn out that on the basis of gene B, patient may have cancer type I or V but since we don't consider V, we infer that patient is likely to have cancer type I. Then we can check if according to gene C, gene's expression is in group of cancer type I. One can say that we treat each genes as a kind of filter, and if patient successfully passes oall filters he is arguably ill. In other words, we take a subset of possible cancers according to each gene and then intersect them. We can also modify it, by computing absolute value of distance in each gene, between gene expression and cancer type, then add those value and see for which cancer type the value is the least. This modification is resistant to the situation where patient for a few genes can be very close to be in a specific cancer type, but he is not, and in the rest of genes he fits perfectly. Above approach has many advantages. Firstly, as it was said, it minimizes number of genes while maximizing amount of informations. Secondly, Each gene is much or less important for each cancer type, so we infer on the basis of 14 genes, while normal approach would be to find a gene (or genes) which are specific for a certain cancer type and is irrelevant for others, so we would have to either infer on one gene or drastically increase number of genes. Thirdly, in our approach it is very probable that if we have a patient with some gene's expression, he will be classified to some group, because if we have many groups, there are many groups, which don't overlap, so they cover more space. This approach has also some drawbacks. The most important one is that there are known diseases for which only one gene is responsible. In our case, we won't detect it. Also it may turn out that one gene has many different group in post hoc test, and other genes are correlated with them, so they have the same high number. But we check correlation, and fortunately they are not, moreover we looked at each post hoc test at confirmed that genes which were chosen explain different things. The last disadvantage is that despite high number of groups we can have a pathological situation that for a one gene's expression it can say us nothing about cancer type. But as well this situation we excluded by checking post hoc tests one by one. So we chose genes:

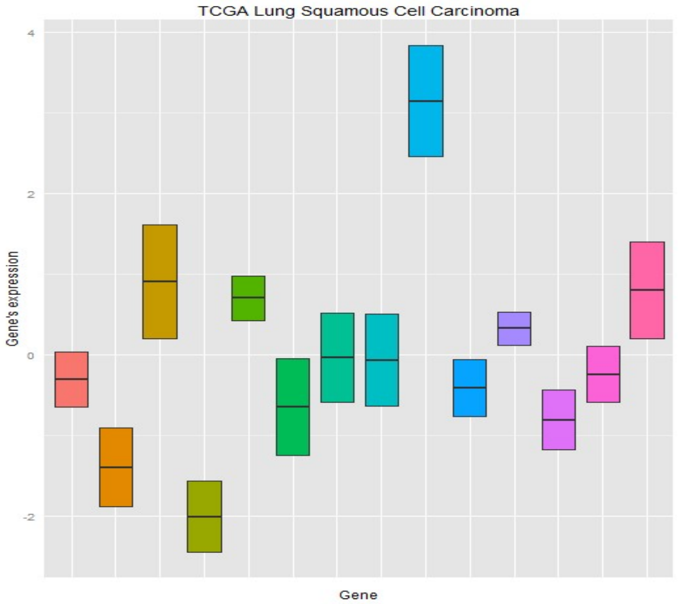
```
##  [1] "APOOL"    "ARHGAP29" "COL3A1"   "CXXC5"    "FAM58A"   "FBXL7"
##  [7] "FJX1"     "IL32"     "LYPD3"    "NBEAL2"   "PEX13"    "SAMD10"
## [13] "TSPAN9"   "TTC22"
```

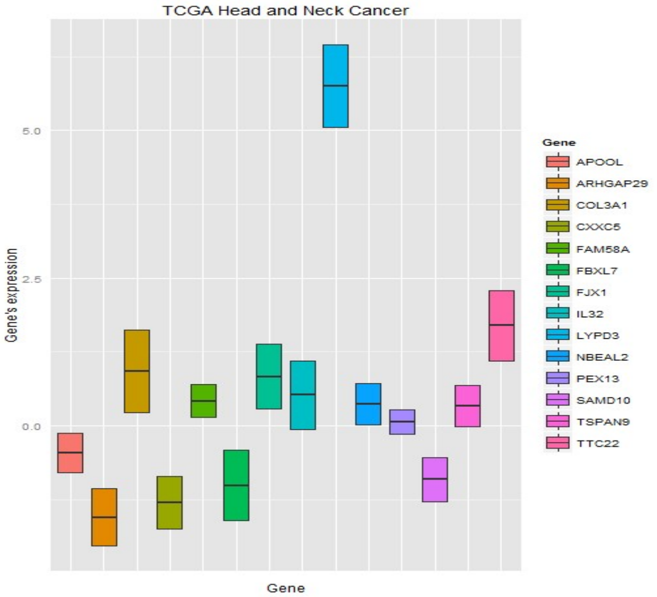
Characteristic genes’ expressions

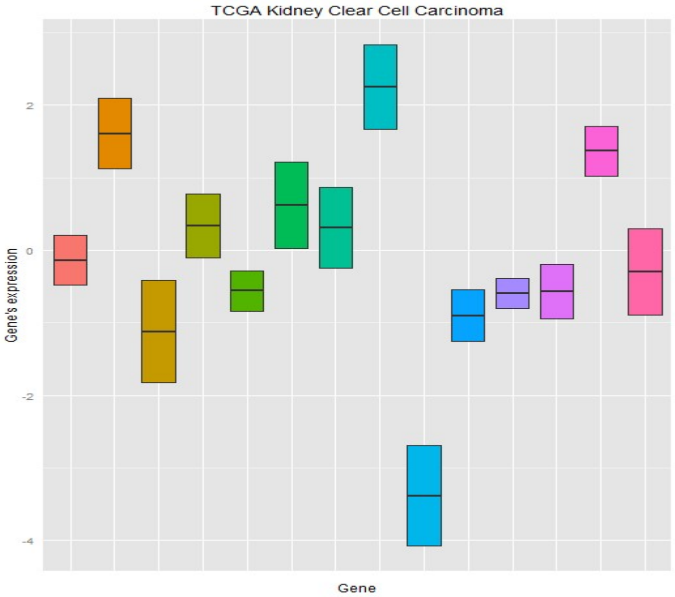
Now we present characteristic genes' expressions for each cancer type. One plot concerns one cancer type. Each rectangle is one gene, thick line is a mean of gene's expression and top and bottom borders corresponds to mean +/- critical difference. So if patient has value of gene which is in the rectangle it means, on the basis of this gene, he is prone to have this cancer type.

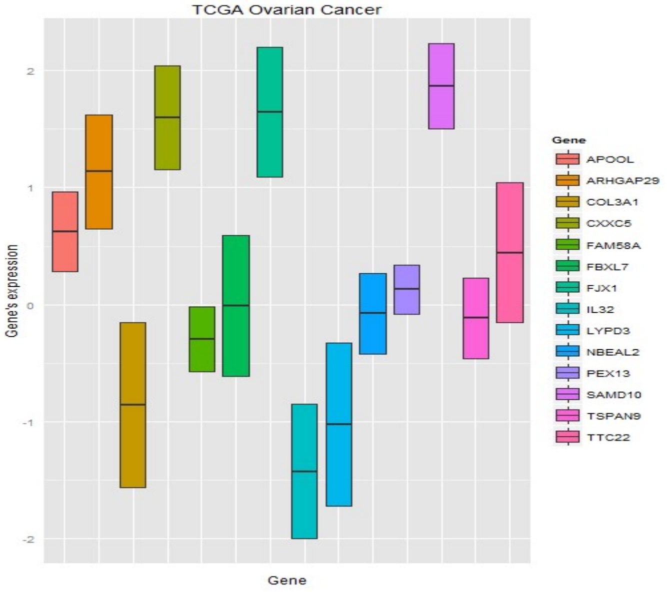
TCGA Acute Myeloid Leukemia
2.5
0.0
-2.5
-5.0
-7.5
-10.0
Gene's expression
Gene
TCGA Bladder Cancer
4
2
0
-2
Gene's expression
Gene
Gene
APOOL
ARHGAP29
COL3A1
CXXC5
FAM58A
FBXL7
FJX1
IL32
LYPD3
NBEAL2
PEX13
SAMD10
TSPAN9
TTC22
TCGA Breast Cancer
2
1
0
-1
Gene's expression
Gene
TCGA Colon Cancer
2
1
0
-1
-2
-3
Gene's expression
Gene
Gene
APOOL
ARHGAP29
COL3A1
CXXC5
FAM58A
FBXL7
FJX1
IL32
LYPD3
NBEAL2
PEX13
SAMD10
TSPAN9
TTC22
TCGA Endometrioid Cancer
1
0
-1
-2
Gene's expression
Gene
TCGA Glioblastoma
2.5
0.0
-2.5
-5.0
-7.5
Gene's expression
Gene
Gene
APOOL
ARHGAP29
COL3A1
CXXC5
FAM58A
FBXL7
FJX1
IL32
LYPD3
NBEAL2
PEX13
SAMD10
TSPAN9
TTC22

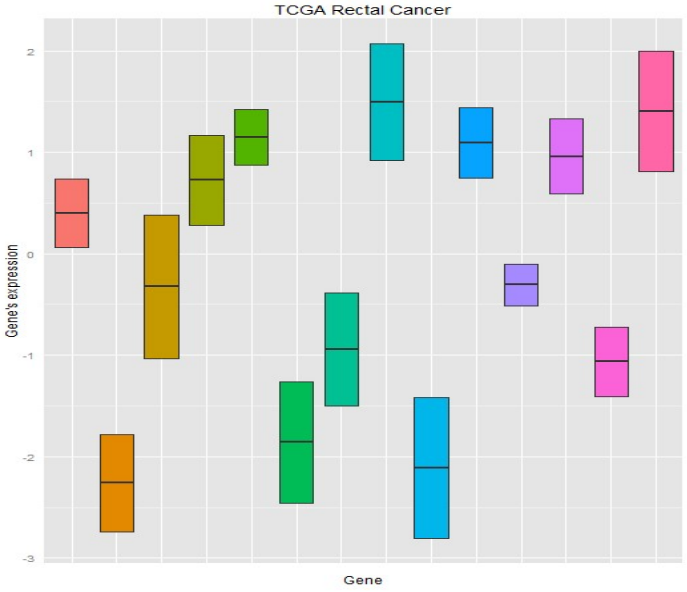

TCGA Lung Adenocarcinoma
1
0
-1
-2
Gene's expression
Gene
Gene
APOOL
ARHGAP29
COL3A1
CXXC5
FAM58A
FBXL7
FJX1
IL32
LYPD3
NBEAL2
PEX13
SAMD10
TSPAN9
TTC22


TCGA Lung Squamous Cell Carcinoma
4
2
0
-2
Gene's expression
Gene


TCGA Head and Neck Cancer
5.0
2.5
0.0
Gene's expression
Gene
Gene
APOOL
ARHGAP29
COL3A1
CXXC5
FAM58A
FBXL7
FJX1
IL32
LYPD3
NBEAL2
PEX13
SAMD10
TSPAN9
TTC22


TCGA Kidney Clear Cell Carcinoma
2
0
-2
-4
Gene's expression
Gene


TCGA Ovarian Cancer
2
1
0
-1
-2
Gene's expression
Gene
Gene
APOOL
ARHGAP29
COL3A1
CXXC5
FAM58A
FBXL7
FJX1
IL32
LYPD3
NBEAL2
PEX13
SAMD10
TSPAN9
TTC22


TCGA Rectal Cancer
2
1
0
-1
-2
-3
Gene's expression
Gene