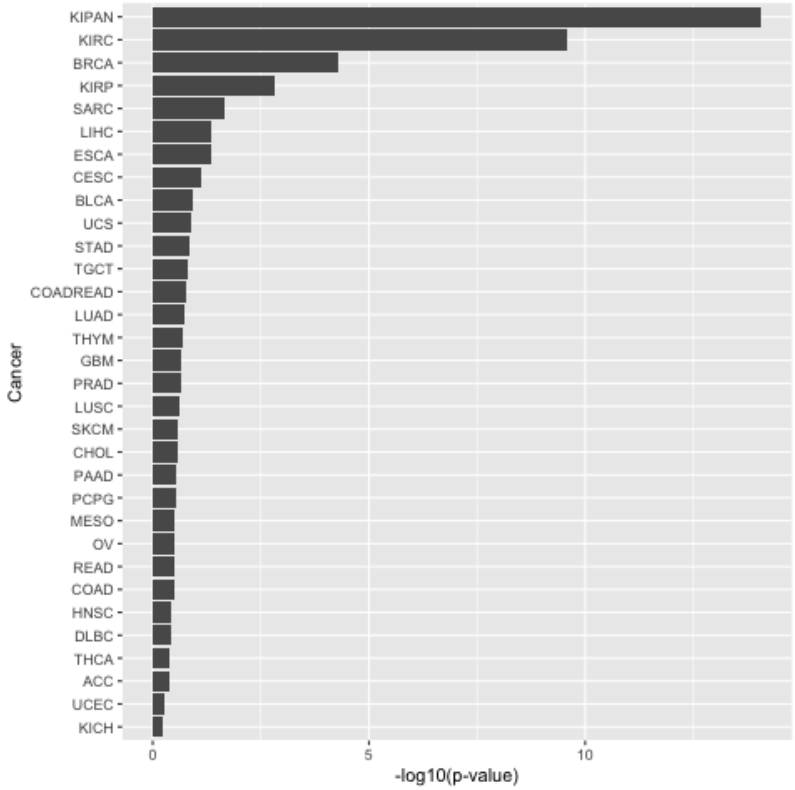
TCGA survival analysis

Mikhail Dozmorov

2018-02-23

Analysis 2: MIA in all cancers

The barplot shows the significance of MIA expression on survival in a given cancer. The wider (higher) the bar the more significant survival effect the gene has. See abbreviations of cancer types at http://www.liuzlab.org/TCGA2STAT/CancerDataChecklist.pdf



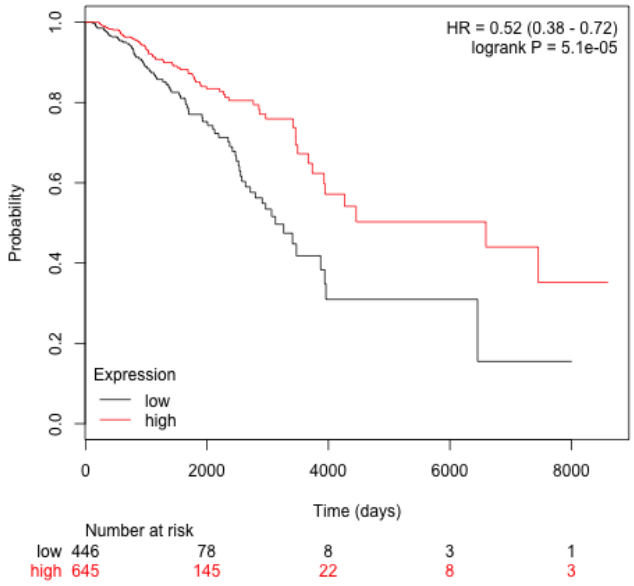
The same data in table format. Legend:

- `Cancer`, `Cancer.Name` - cancer abbreviation and description
- `Gene` - gene name for which survival analysis was run
- `p.value` - significance of the survival effect
- `HR`, `HR_left`, `HR_right` - hazard ratio, and left/right confidence interval
- `Min.`, `X1st.Qu.`, `Median`, `Mean`, `X3rd.Qu.`, `Max.` - expression level of the gene in a corresponding cancer
- `Cutoff_type`, `Cutoff_value` - gene expression cutoff best discriminating survival

Cancer	Cancer.Name	Gene	p.value	HR	HR_left	HR_right	Min.	X1st.Qu.
KIPAN	Pan-kidney cohort (KICH+KIRC+KIRP)	MIA	0.00e+00	2.72	2.09	3.53	0	0.40
KIRC	Kidney renal clear cell carcinoma	MIA	0.00e+00	2.76	1.99	3.83	0	0.42
BRCA	Breast invasive carcinoma	MIA	5.12e-05	0.52	0.38	0.72	0	2.44
KIRP	Kidney renal papillary cell carcinoma	MIA	1.56e-03	2.54	1.40	4.63	0	0.43
SARC	Sarcoma	MIA	2.23e-02	1.80	1.08	3.01	0	1.10

Cancer	Cancer.Name	Gene	p.value	HR	HR_left	HR_right	Min.	X1st.Qu.
LIHC	Liver hepatocellular carcinoma	MIA	4.47e-02	1.42	1.01	2.01	0	0.62
ESCA	Esophageal carcinoma	MIA	4.55e-02	0.58	0.33	0.99	0	1.22
CESC	Cervical and endocervical cancers	MIA	7.86e-02	1.57	0.95	2.59	0	1.58
BLCA	Bladder urothelial carcinoma	MIA	1.21e-01	1.32	0.93	1.88	0	1.18
UCS	Uterine Carcinosarcoma	MIA	1.32e-01	0.58	0.29	1.19	0	2.75

Survival effect in BRCA cancer

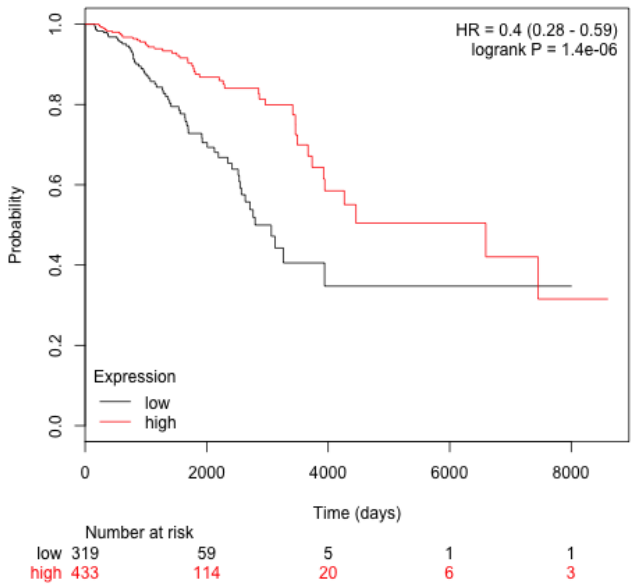


Analysis 3: MIA in BRCA, clinical subtypes

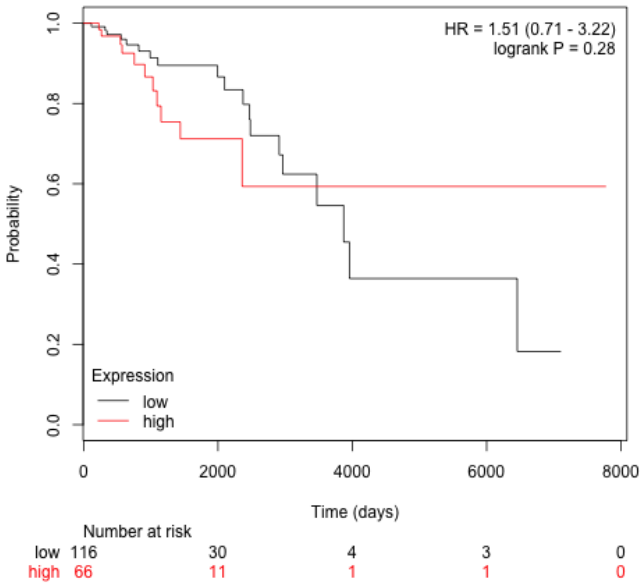
The table lists clinical subtypes where the expression of MIA gene in BRCA most significantly affects survival. The table is sorted by increasing p-values, most significant on top. Description of clinical subtypes can be found at https://gdc.cancer.gov/about-data/data-harmonization-and-generation/clinical-data-harmonization

Cancer	Gene	p.value	HR	HR_left	I
BRCA-breast_carcinoma_estrogen_receptor_status-Positive	MIA	1.00e-07	0.35	0.23	
BRCA-race-white	MIA	1.40e-06	0.40	0.28	
BRCA-breast_carcinoma_progesterone_receptor_status-Positive	MIA	4.40e-06	0.37	0.24	
BRCA-gender-FEMALE	MIA	4.77e-05	0.52	0.38	
BRCA-PAM50Call_RNAseq-Basal	MIA	5.27e-04	0.12	0.03	
BRCA-breast_carcinoma_surgical_procedure_name-Modified Radical Mastectomy	MIA	9.79e-04	0.42	0.25	
BRCA-person_neoplasm_cancer_status-TUMOR FREE	MIA	9.83e-04	0.40	0.23	
BRCA-radiation_therapy-NO	MIA	1.50e-03	0.43	0.25	
BRCA-histological_type-Infiltrating Ductal Carcinoma	MIA	1.56e-03	0.54	0.37	
BRCA-pathologic_M-M0	MIA	1.65e-03	0.56	0.39	

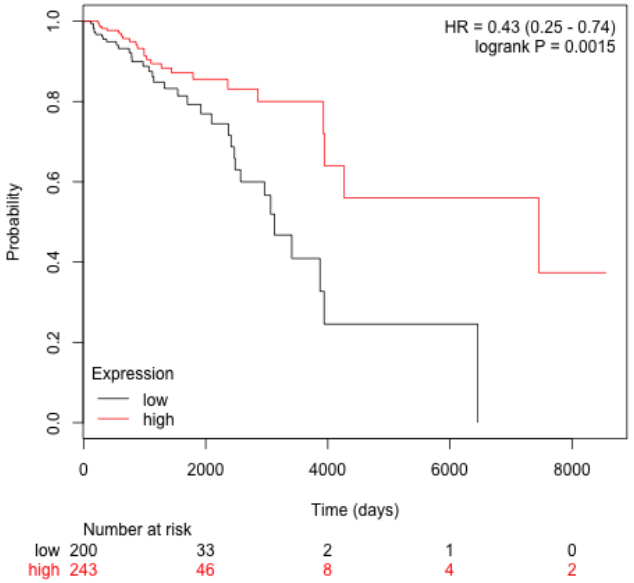
"MIA_BRCA-race-white"



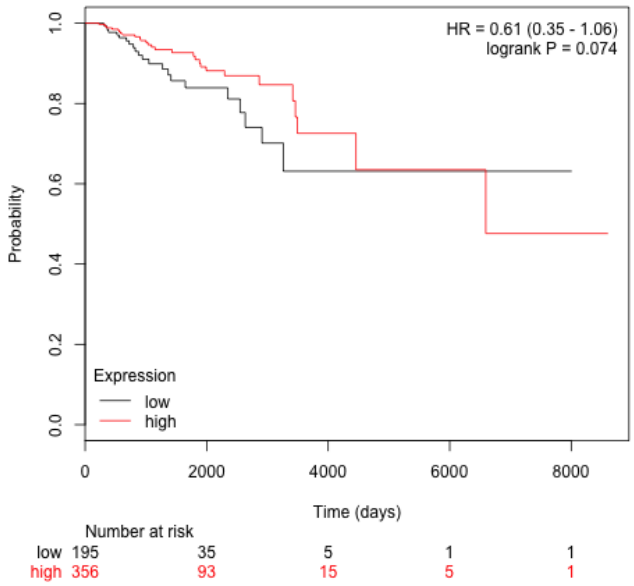
"MIA_BRCA-race-black or african american"



"MIA_BRCA-radiation_therapy-NO"

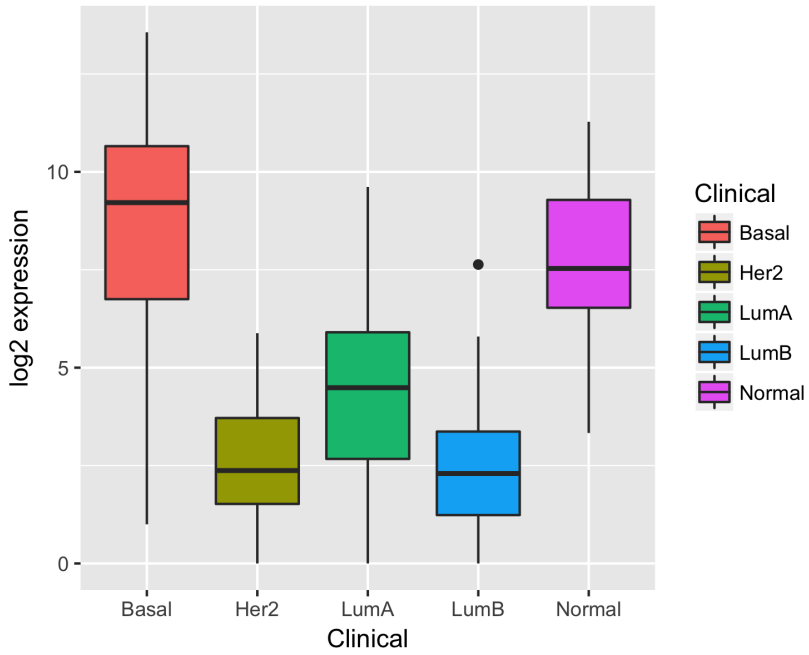


“MIA_BRCA-radiation_therapy-YES”



Analysis 5: Clinical-centric analysis. Selected cancer, selected clinical subcategory, gene expression differences across categories

Expression of MIA in selected clinical subcategories



Cancer	Gene	p.value	HR
BRCA	PAM50Call_RNAseq-LumA-Her2	0.00536	0.44
BRCA	PAM50Call_RNAseq-LumA-LumB	0.0149	1.72
BRCA	PAM50Call_RNAseq-Basal-Her2	0.0314	2.14
BRCA	PAM50Call_RNAseq-LumB-Basal	0.0584	1.74
BRCA	PAM50Call_RNAseq-Normal-Basal	0.427	1.49
BRCA	PAM50Call_RNAseq-Normal-LumA	0.448	1.42
BRCA	PAM50Call_RNAseq-Normal-LumB	0.521	0.73
BRCA	PAM50Call_RNAseq-LumB-Her2	0.542	0.82
BRCA	PAM50Call_RNAseq-Normal-Her2	0.565	0.74
BRCA	PAM50Call_RNAseq-LumA-Basal	0.722	1.1

ANOVA and Tukey's test

What are the means of log2-expression per clinical subgroup""

```
   Basal     Her2     LumA     LumB   Normal
8.389736 2.540233 4.317911 2.335421 7.872572
```

ANOVA

Is the expression of gene MIA significantly different across clinical subgroups? Significant "Pr(>F)" suggests "Yes"

```
Analysis of Variance Table
```

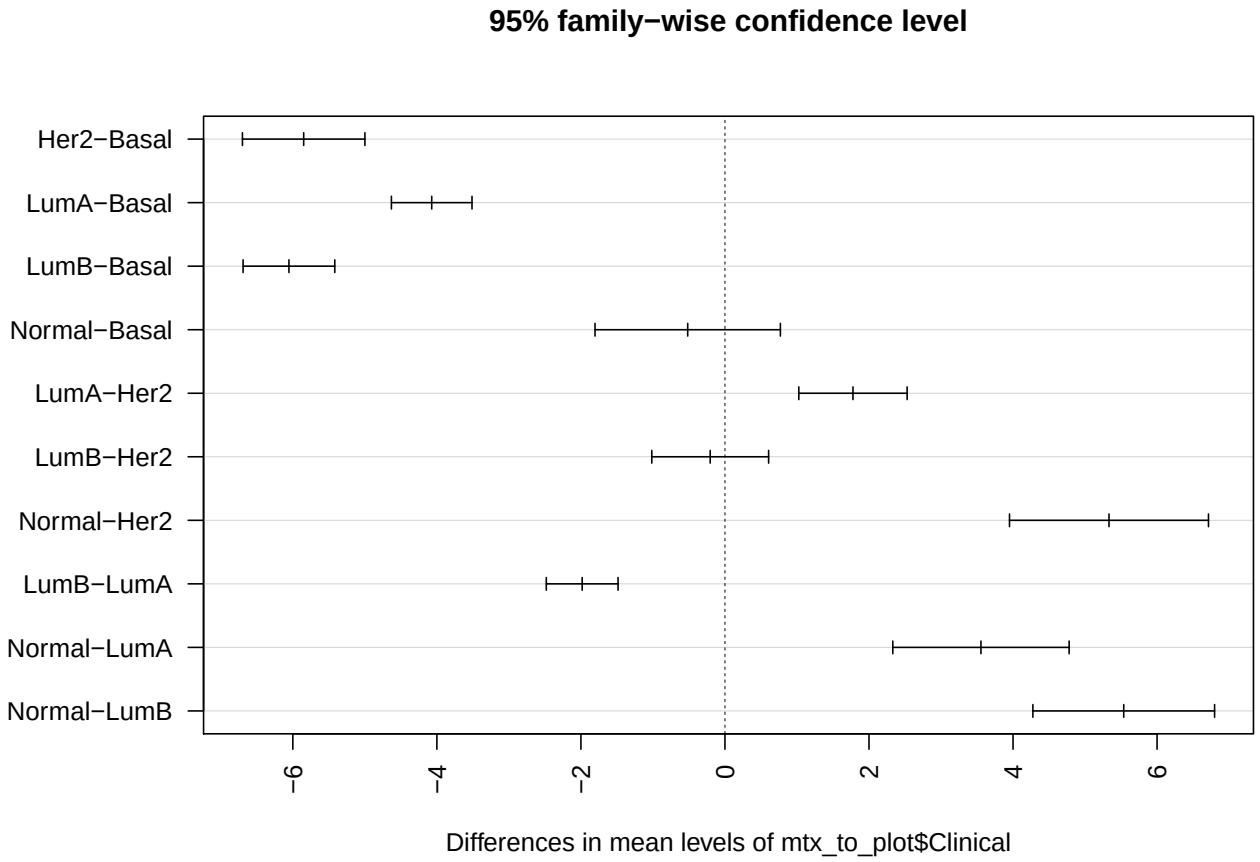
```
Response: mtx_to_plot$Gene
                      Df Sum Sq Mean Sq F value    Pr(>F)
mtx_to_plot$Clinical   4 3535.5  883.87  202.13 < 2.2e-16 ***
Residuals            835 3651.2    4.37
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey HSD (Honest Significant Difference) test

Which pair of clinical categories has significant differences? "p.adj" and confidence intervals that do not cross 0 suggest significant differences in gene expression between the subgroups in the corresponding pairwise comparison.

- **mtx_to_plot$Clinical**:

	diff	lwr	upr	p adj
Her2-Basal	-5.85	-6.7	-4.999	3.164e-13
LumA-Basal	-4.072	-4.631	-3.512	3.164e-13
LumB-Basal	-6.054	-6.691	-5.418	3.164e-13
Normal-Basal	-0.5172	-1.804	0.7697	0.8073
LumA-Her2	1.778	1.026	2.53	1.764e-09
LumB-Her2	-0.2048	-1.016	0.6063	0.9586
Normal-Her2	5.332	3.951	6.714	3.624e-13
LumB-LumA	-1.982	-2.481	-1.484	3.537e-13
Normal-LumA	3.555	2.33	4.779	4.368e-13
Normal-LumB	5.537	4.276	6.799	3.165e-13


95% family−wise confidence level
Her2−Basal
LumA−Basal
LumB−Basal
Normal−Basal
LumA−Her2
LumB−Her2
Normal−Her2
LumB−LumA
Normal−LumA
Normal−LumB
−6
−4
−2
0
2
4
6
Differences in mean levels of mtx_to_plot$Clinical

Methods

Survival analysis of gene expression data from TCGA

Level 3 gene expression data summarized as RSEM values was obtained using the `TCGA2STAT` R package v 1.2, along with the corresponding clinical annotations. Data for each of the 34 cancers was obtained separately. The data was log2-transformed and analyzed using Kaplan-Meier curves and Cox proportional hazard model. Each gene of interest was analyzed for its effect on survival by separating patients into high/low expression subgroups. A modified approach from [1] was used to estimate the best gene expression cutoff that separates high/low expression subgroups with differential survival.

We took the advantage of the availability of clinical annotations. To identify if expression of a gene of interest affects survival in any specific clinical subgroup, subsets of patients annotated with specific clinical annotations were selected (e.g., "males" or "females" in the "gender" clinical annotation). Subgroups with $<$ 40 patients were not considered.

Differential expression analysis

Samples in the selected cancer cohort were sorted by expression of the selected genes. Differentially expressed genes were detected between samples in the upper 75 percentile of the expression gradient and samples in the lower 25 percentile using `limma` v 3.32.6 R package [2,3]. P-values were corrected for multiple testing using False Discovery Rate (FDR) method [4]. Genes differentially expressed at FDR < 0.01 were selected for further analysis.

References

1. Mihály Z, Kormos M, Lánczky A, Dank M, Budczies J, Szász MA, Győrffy B: **A meta-analysis of gene expression-based biomarkers predicting outcome after tamoxifen treatment in breast cancer**. *Breast Cancer Res Treat* 2013, **140**:219–3210.1007/s10549-013-2622-y.

2. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **Limma powers differential expression analyses for rna-sequencing and microarray studies**. *Nucleic Acids Res* 2015, **43**:e4710.1093/nar/gkv007.

3. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Stat Appl Genet Mol Biol* 2004, **3**:Article310.2202/1544-6115.1027.

4. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995, **57**:289–300Available: http://www.jstor.org/stable/2346101.