Differential expression analysis of TCGA data between groups with high/low expression of selected genes

Mikhail Dozmorov

2018-02-24

Methods

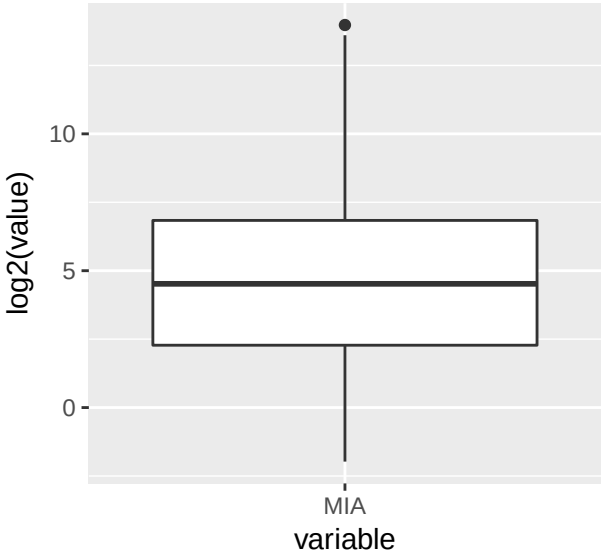
```
# Cancer type
cancer = "BRCA"
# Gene(s) of interest
selected_genes = c("MIA")  # Can be multiple
# Define quantile qutoffs
quantile_up <- 0.75  # 0.75
quantile_lo <- 0.25  # 0.25
```

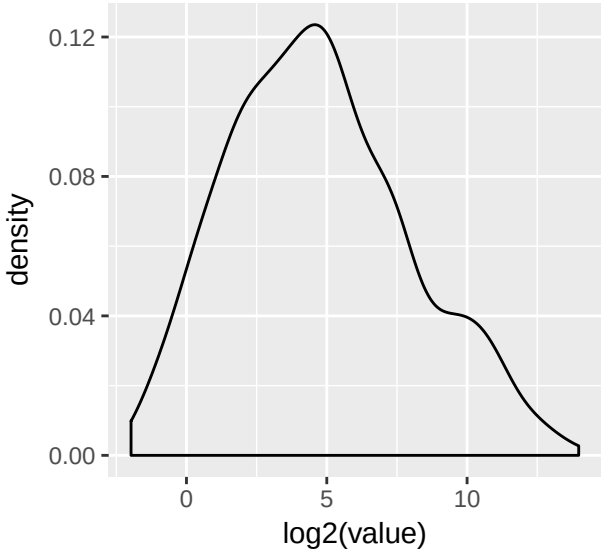
Differential expression analysis

Samples in the selected cancer cohort were sorted by expression of the selected genes. Differentially expressed genes were detected between samples in the upper 0.75 percentile of the expression gradient and samples in the lower 0.25 percentile using `limma` v 3.32.7 R package [1,2]. P-values were corrected for multiple testing using False Discovery Rate (FDR) method [3]. Genes differentially expressed at FDR < 0.01 were selected for further analysis.

Expression distribution for MIA gene

Informational plot only. Expected log2-transformed RSEM expression range distribution: $\sim 0 - 16$. Median close to 0 indicate overall low expression, should be avoided.

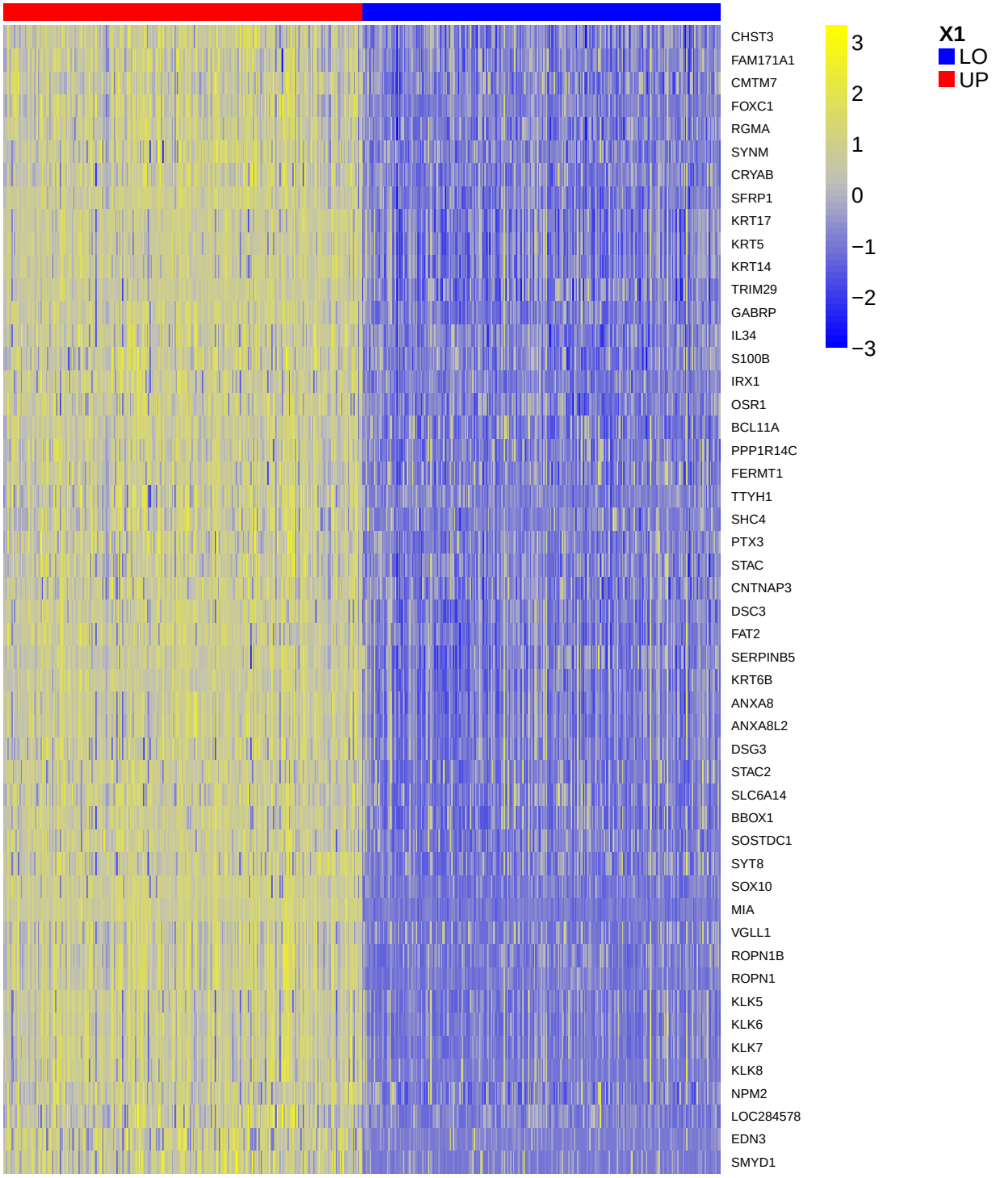




Differential expression analysis

We split BRCA cohort into 273 x 273 patients having MIA in lower 0.75 and upper 0.25 expression quantiles. We have a total of 12802 differentially expressed genes at FDR corrected p-value 0.01. 7155 are upregulated, 5647 are downregulated.

Top 50 the most differentially expressed genes are shown



Results are stored in the Excel file results/BRCA_MIA_UP75LO25_DEGs_0.01.xlsx

- Legend for gene lists: “Gene” - gene annotations; “logFC” - log fold change; “AveExpr” - average expression, log2; “t” - t-statistics; “P.Val”/“adj.P.Val” - non-/FDR-adjusted p-value, “B” - another statistics.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
MIA	8.013	0.7394	68.46	4.066e-270	8.335e-266	605.4
SOX10	8.587	1.23	43.98	1.453e-181	1.49e-177	404.1
ROPN1	6.469	-0.8354	37.59	1.42e-153	9.706e-150	340.2
SFRP1	6.042	5.387	36.21	2.809e-147	1.44e-143	325.8
GABRP	7.838	3.692	36.11	8.938e-147	3.665e-143	324.6
FAT2	4.77	1.681	30.9	5.836e-122	1.994e-118	267.8
KRT6B	7.318	2.458	30.78	2.149e-121	6.293e-118	266.5
IRX1	4.837	1.404	30.46	8.305e-120	2.128e-116	262.9
DSC3	5.884	2.057	30.32	3.919e-119	8.928e-116	261.3
ROPN1B	5.583	-0.4379	30	1.501e-117	3.047e-114	257.7

Functional enrichment analysis

KEGG canonical pathway enrichment analysis

Up- and downregulated genes are tested for pathway enrichment separately. Each table has enrichment results for both up-/downregulated genes. The “direction” column indicate which pathways are enriched in “UP”- or “DN”-regulated genes.. FDR cutoff of the significant enrichments - 0.3. Top 10 genes shown.

Legend: “database” - source of functional annotations, “category” - name of functional annotation, “pval” - unadjusted enrichment p-value, “qval” - FDR-adjusted p-value, “genes” - comma-separated differentially expressed genes enriched in a corresponding functional category. “direction” - UP/DN, an indicator whether genes are up- or downregulated.

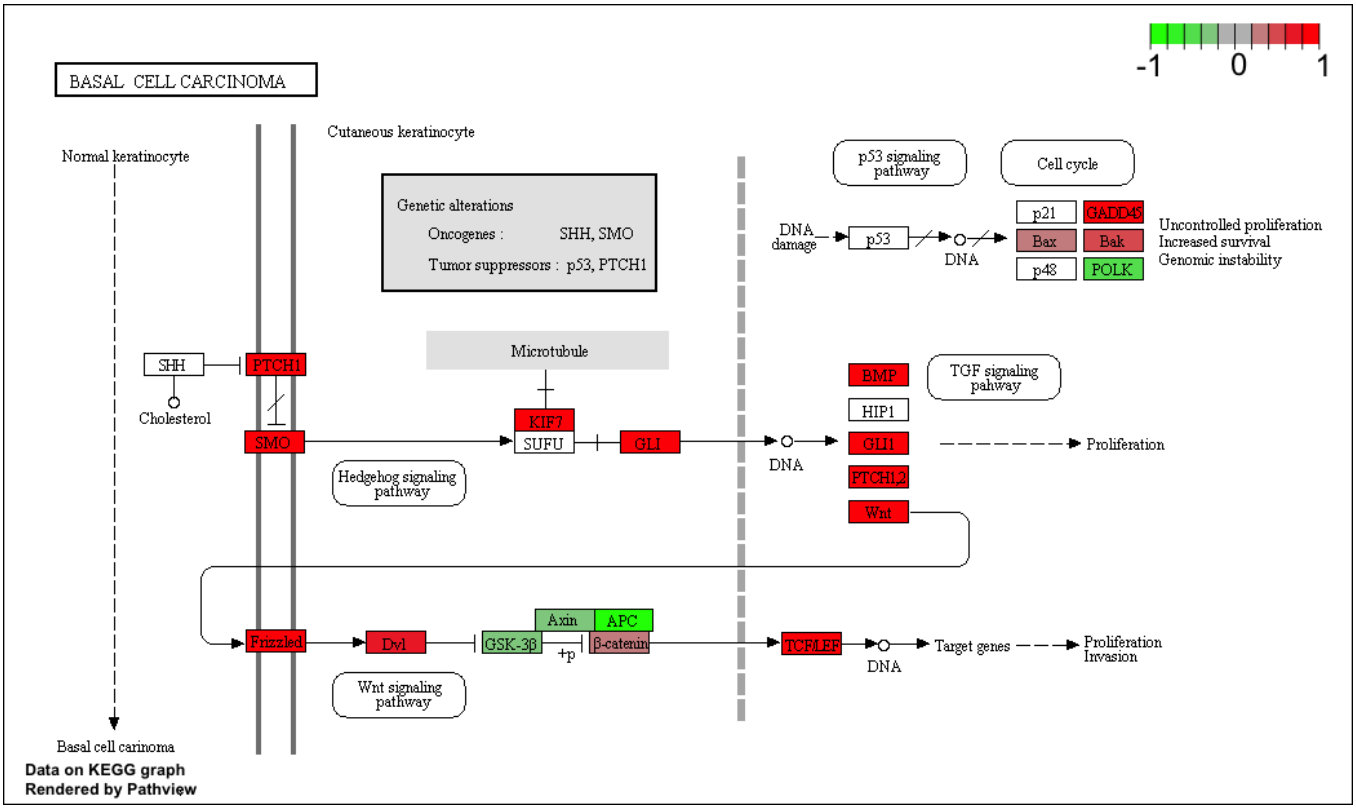
```
[1] "KEGG pathway run on 3726 upregulated and 2274 downregulated genes."
[1] "Running KEGG_2016 analysis"
```

A total of 569 KEGG pathways were detected as significantly affected at FDR 0.3. Top 10 shown.

database	category	pval	qval
KEGG_2016	Ribosome_Homo sapiens_hsa03010	1.786e-16	5.125e-14
KEGG_2016	Cytokine-cytokine receptor interaction_Homo sapiens_hsa04060	8.048e-15	1.155e-12
KEGG_2016	TNF signaling pathway_Homo sapiens_hsa04668	4.680e-08	4.477e-06
KEGG_2016	Hematopoietic cell lineage_Homo sapiens_hsa04640	3.136e-07	2.250e-05
KEGG_2016	NF-kappa B signaling pathway_Homo sapiens_hsa04064	1.301e-05	5.332e-04
KEGG_2016	Primary immunodeficiency_Homo sapiens_hsa05340	7.198e-06	3.443e-04
KEGG_2016	Glycosphingolipid biosynthesis - lacto and neolacto series_Homo sapiens_hsa00601	1.608e-06	9.229e-05
KEGG_2016	Cell adhesion molecules (CAMs)_Homo sapiens_hsa04514	5.823e-05	2.089e-03
KEGG_2016	HTLV-I infection_Homo sapiens_hsa05166	2.811e-04	7.333e-03
KEGG_2016	Basal cell carcinoma_Homo sapiens_hsa05217	1.824e-04	5.818e-03

Selected pathway

Red/Green - up/downregulated genes in upper vs. lower MIA expressing samples. Gray - marginal fold change, yet significant. White - gene is not differentially expressed



References

1. 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.

2. 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.

3. 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.