Bioinformatics I (KF) VU 041035 WS2024

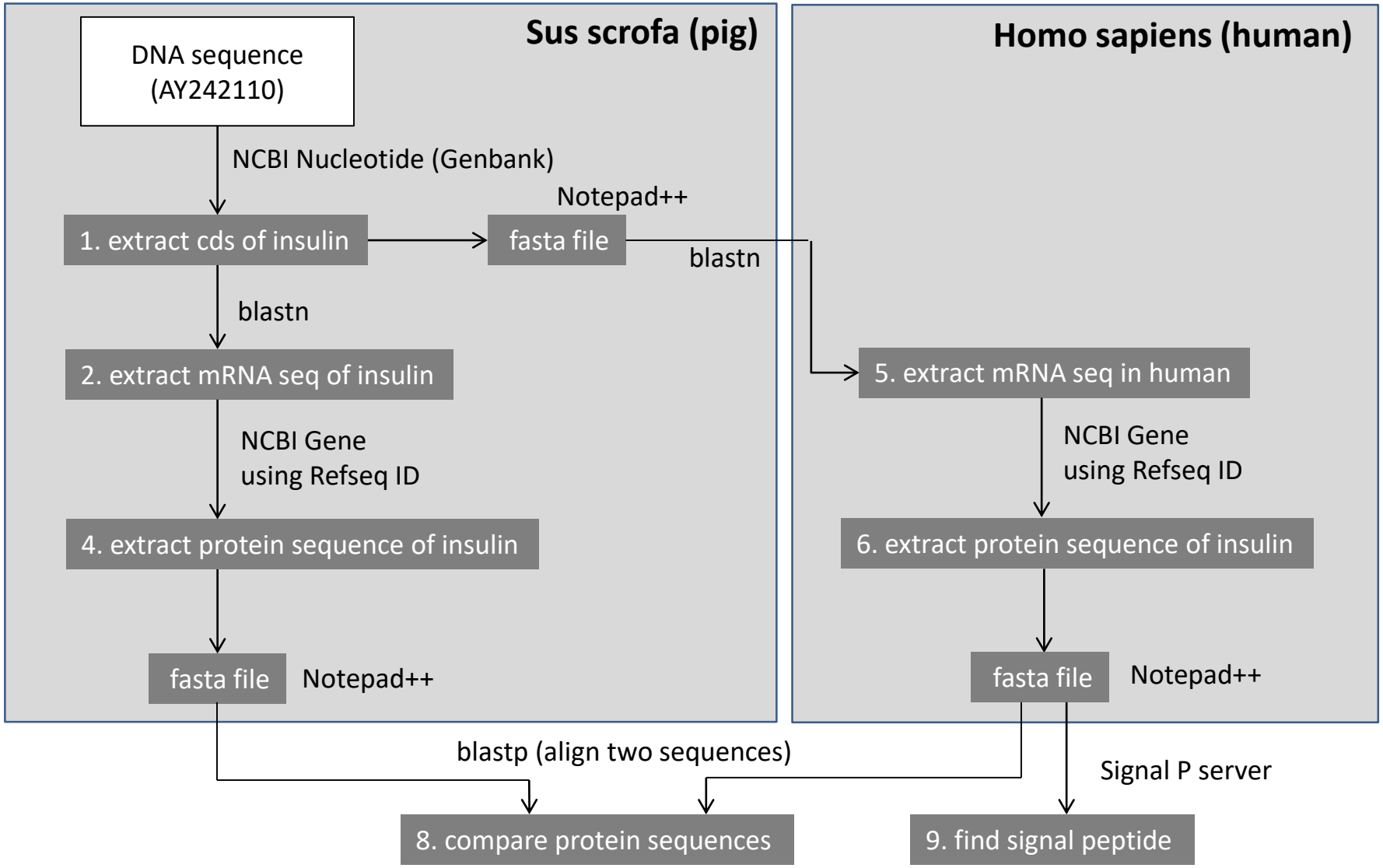
http://icbi.at/mo

Hubert Hackl
Biocenter, Institute of Bioinformatics,
Medical University of Innsbruck,
Innrain 80, 6020 Innsbruck, Austria
Tel: +43-512-9003-71403,
Email: hubert.hackl@i-med.ac.at
URL: http://icbi.at

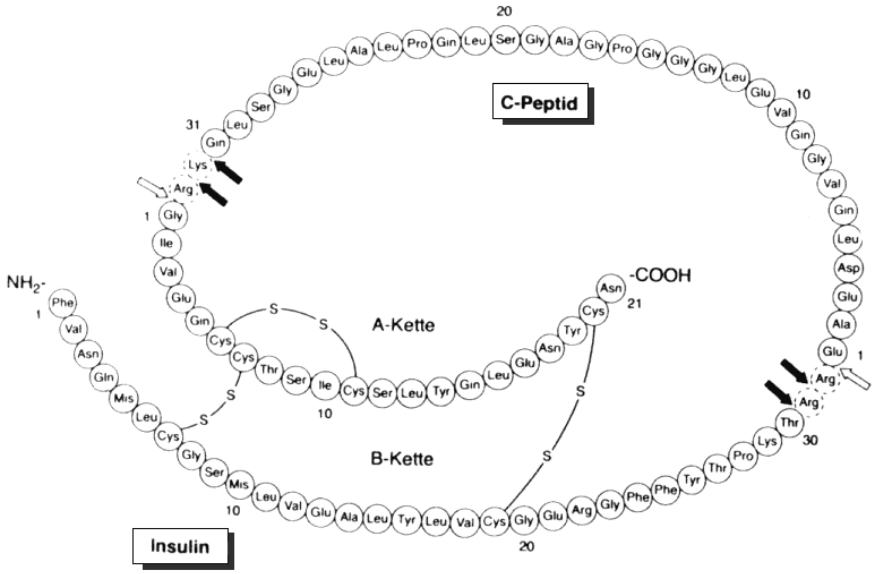
Computational exercises with online databases/tools and R scripts

(1) Functional prediction from protein sequence (BLAST, SignalP, InterPro, NetMHCpan)
(2) IntoGen
(3) cBioPortal
(4) RNAseq preprocessing
(5) R introduction
(6) Differentially expressed genes (limma, DESeq2)
(7) Functional analysis
(8) TCGA (Firebrowse) preprocessing, boxplots, KM survival analyses
(9) Heatmaps and clustering analyses (Genesis)
(10) Gene set enrichment analyses (GSEA)
(11) Predictive biomarker, logistic regression, ROC curve
(12) Single cell RNAseq analyses (Seurat Tutorial)

Difference between insulin sequence in pig and human



Difference between insulin sequence in pig and human



Difference between pig and human insulin = 1AA

Signal peptide

B-chain (30 AA)

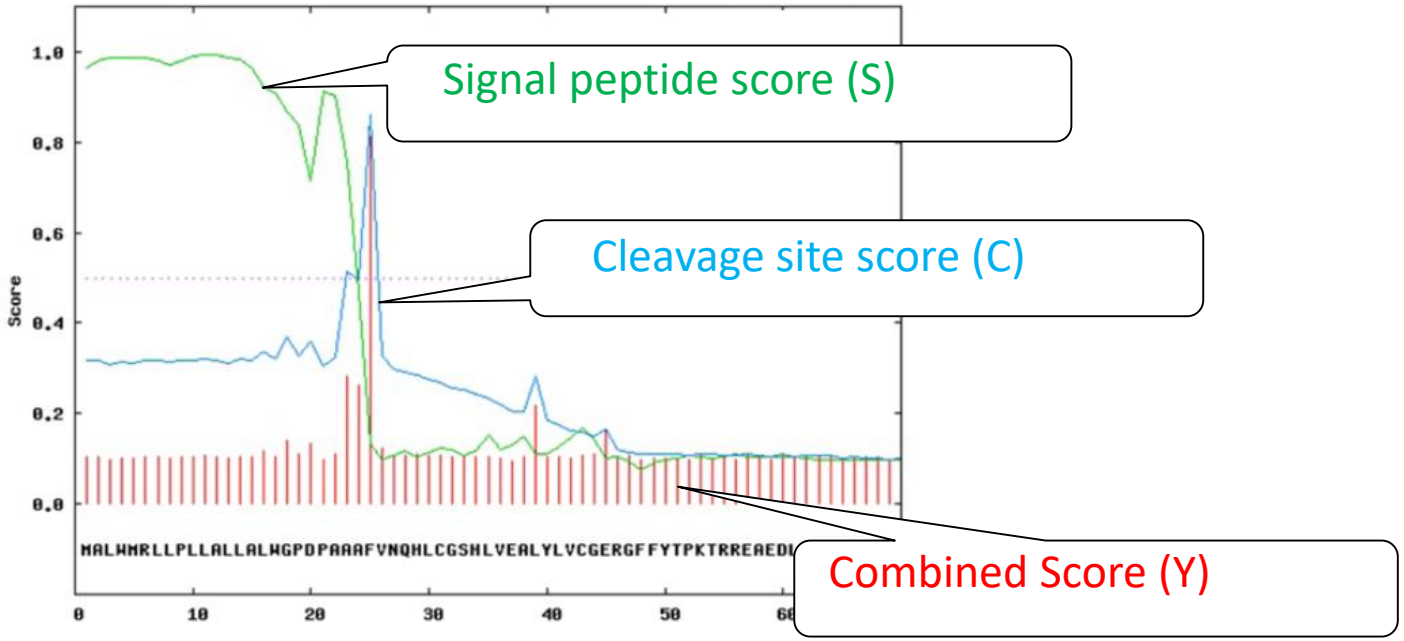
```
Query   1    MALWTRLLPLLALLALWAPAPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREAEN  60
             MALW RLLPLLALLALW P PA AFVNQHLCGSHLVEALYLVCGERGFFYTPK RREAE+
Sbjct   1    MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED  60

Query   61   PQAGAVELGG--GLGGLQALALEGPPQKRGIVEQCCTSICSLYQLENYCN  108
              Q G VELGG  G G LQ LALEG  QKRGIVEQCCTSICSLYQLENYCN
Sbjct   61   LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110
```

C-peptide

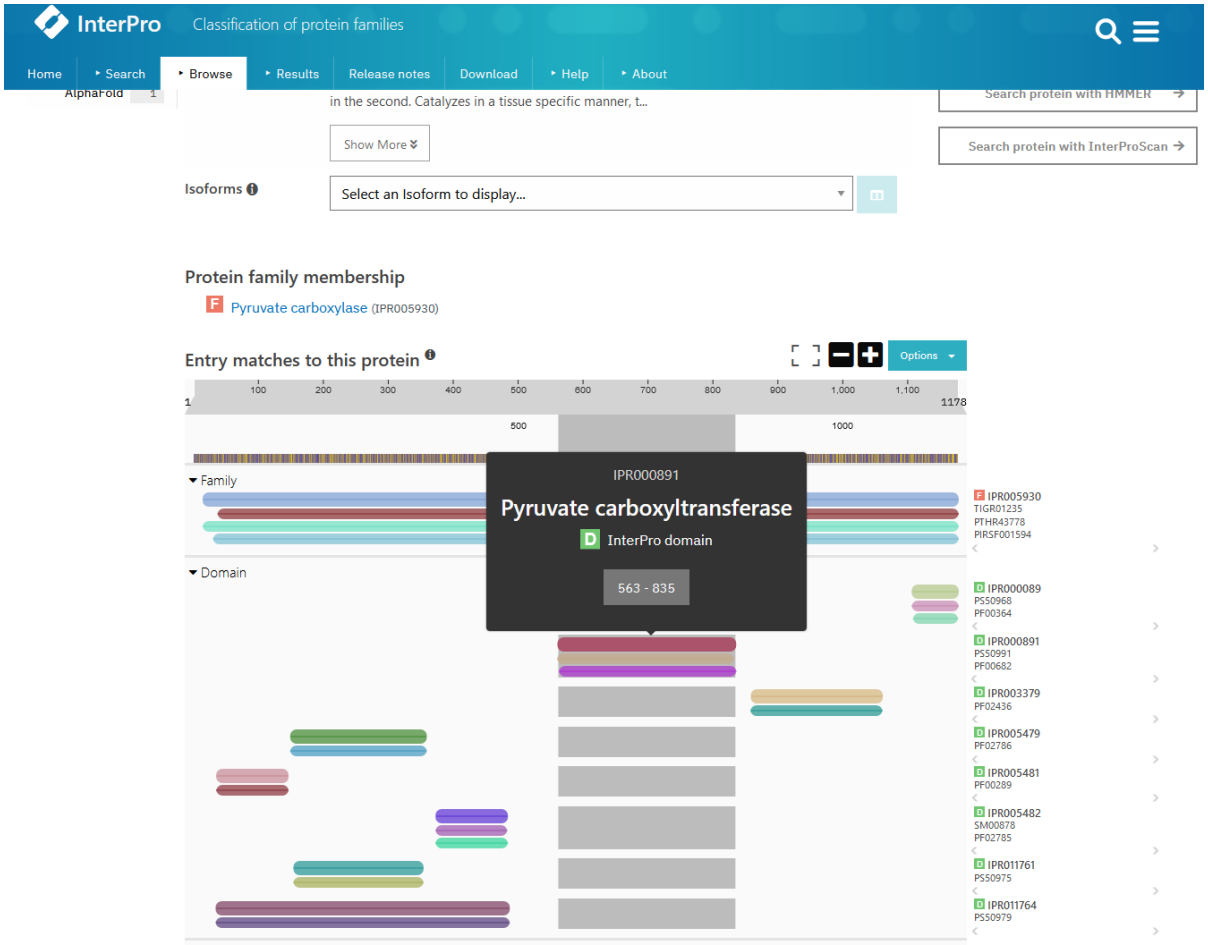
A-chain (21 AA)

SignalP

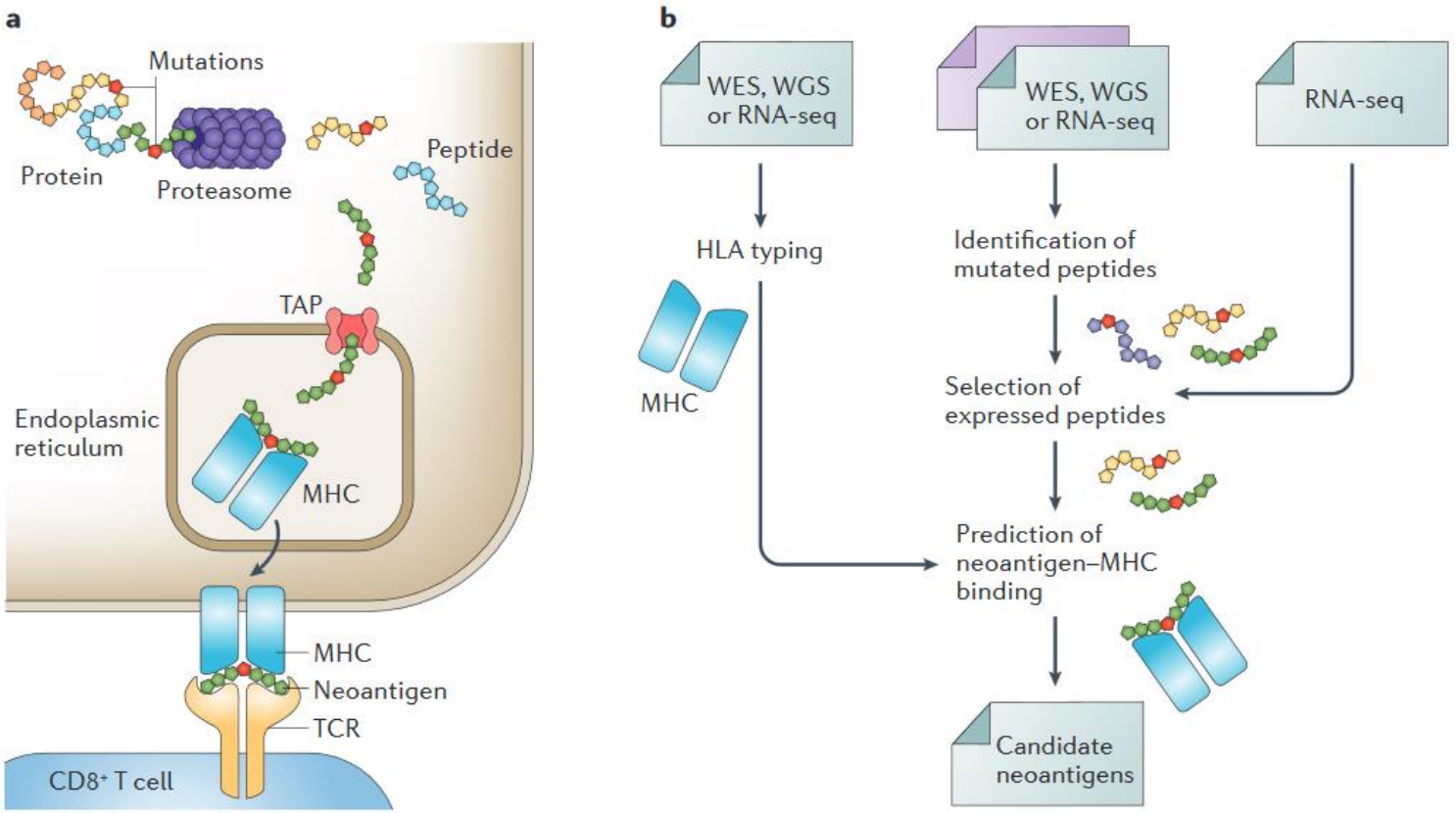
- Neural network trained based on phylogeny
  - Gram-negative prokaryotic
  - Gram-positive prokaryotic
  - Eukaryotic
- Predicts secretory signal peptides
- http://www.cbs.dtu.dk/services/SignalP/



Protein domains

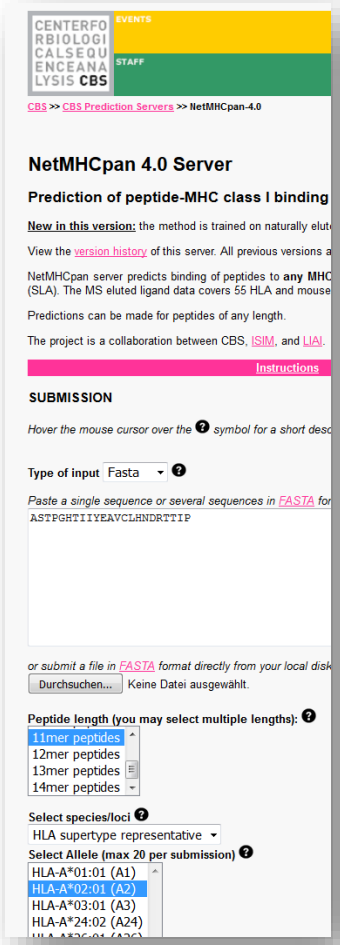


Neoantigen prediction



Hackl et al. Nat Rev Genet 17: 441-458 (2016)

NetMHCpan



Fasta input:

>Gag_180_209
TPQDLNTMLNTVGGHQAAMQMLKETINEEA

Peptide length: **8, 9, 10, 11, 12**
Allele: **HLA-A*0301**
Toggle **Sort by prediction score**

will return the following predictions:

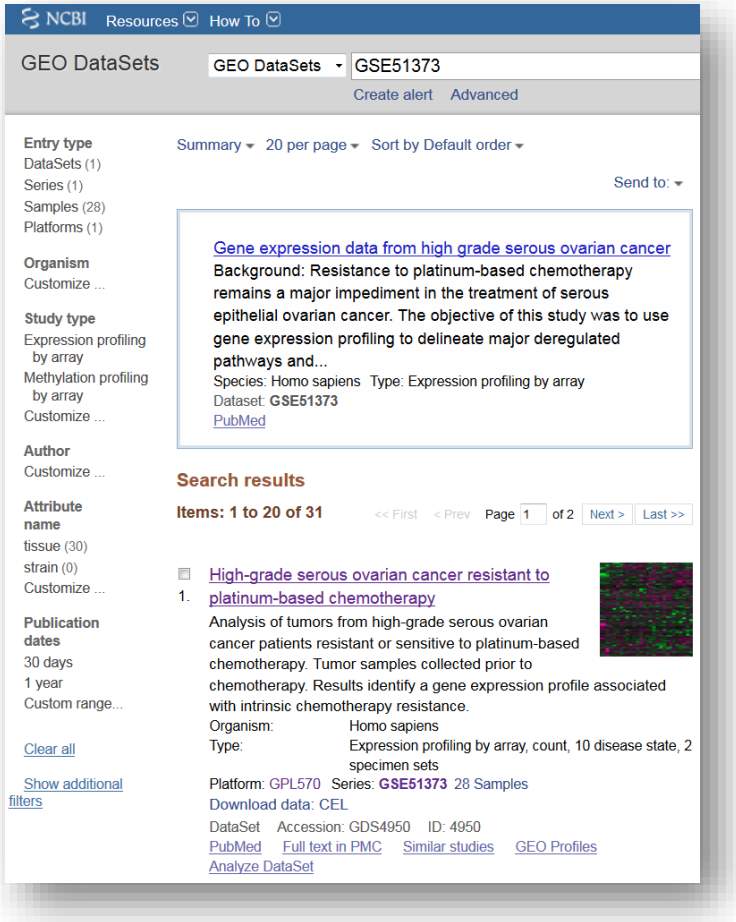
```
# NetMHCpan version 4.0

# Tmpdir made /usr/opt/www/webface/tmp/server/netmhcpan/59DBCCFF00005A84DAFF1311/netMHCpanVszuD8
# Input is in FSA format

# Peptide length 8,9,10,11,12

# Make Eluted ligand likelihood predictions

HLA-A03:01 : Distance to training data  0.000 (using nearest neighbor HLA-A03:01)

# Rank Threshold for Strong binding peptides   0.500
# Rank Threshold for Weak binding peptides   2.000
-----------------------------------------------------------------------------------
  Pos          HLA         Peptide       Core Of Gp Gl Ip Il        Icore        Identity     Score   %Rank  BindLevel
-----------------------------------------------------------------------------------
   15  HLA-A*03:01       HQAAMQMLK  HQAAMQMLK  0  0  0  0  0    HQAAMQMLK     Gag_180_209 0.5697290  0.2857 <= SB
   14  HLA-A*03:01      GHQAAMQMLK  GQAAMQMLK  0  1  1  0  0   GHQAAMQMLK     Gag_180_209 0.2137130  1.1582 <= WB
    7  HLA-A*03:01       TMLNTVGGH  TMLNTVGGH  0  0  0  0  0    TMLNTVGGH     Gag_180_209 0.0487720  3.0466
    8  HLA-A*03:01       MLNTVGGHQ  MLNTVGGHQ  0  0  0  0  0    MLNTVGGHQ     Gag_180_209 0.0319510  3.7842
   13  HLA-A*03:01     GGHQAAMQMLK  GQAAMQMLK  0  1  2  0  0  GGHQAAMQMLK     Gag_180_209 0.0313010  3.8215
   12  HLA-A*03:01    VGGHQAAMQMLK  VQAAMQMLK  0  1  3  0  0 VGGHQAAMQMLK     Gag_180_209 0.0166440  5.2079
   15  HLA-A*03:01      HQAAMQMLKE  HQAAMQMLK  0  0  0  0  0    HQAAMQMLK     Gag_180_209 0.0124970  5.9719
   16  HLA-A*03:01        QAAMQMLK  QAA-MQMLK  0  0  0  3  1     QAAMQMLK     Gag_180_209 0.0086270  7.1279
   21  HLA-A*03:01       MLKETINEE  MLKETINEE  0  0  0  0  0    MLKETINEE     Gag_180_209 0.0079270  7.4157
--
--
-----------------------------------------------------------------------------------

Protein Gag_180_209. Allele HLA-A*03:01. Number of high binders 1. Number of weak binders 1. Number of peptides 105

Link to Allele Frequencies in Worldwide Populations HLA-A03:01
-----------------------------------------------------------------------------------
```

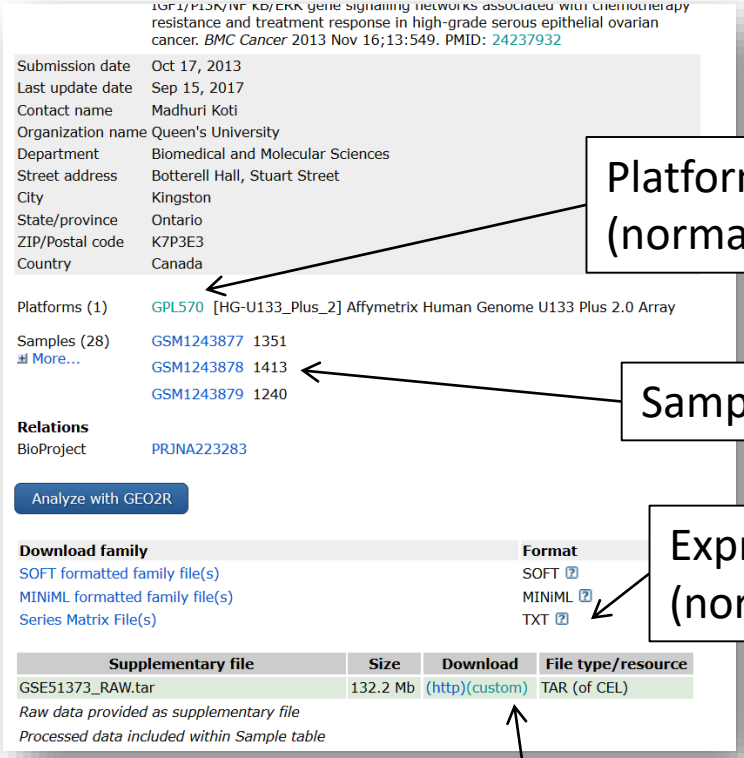
http://www.cbs.dtu.dk/services/NetMHCpan/

Intogen

- What is the most common BRAF mutation
- In which cancer types IDH1 is a cancer driver and in which cancer type mutation of IDH1 is most frequent
- Most common drivers in breast carcinoma
- Mutation frequency of VHL

Gene Expression Omnibus (GEO)



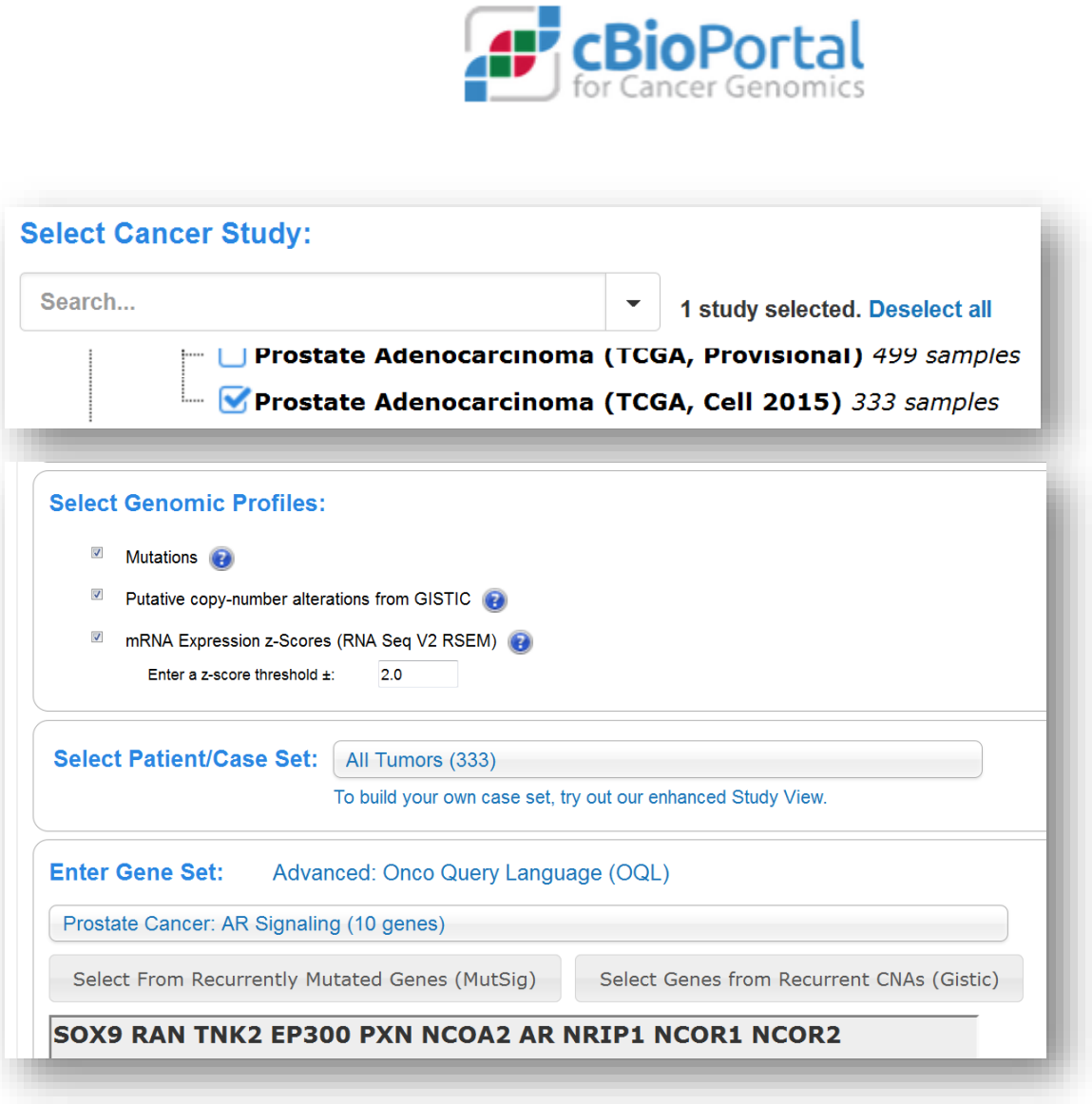


Platform (microarray) (normalized data)

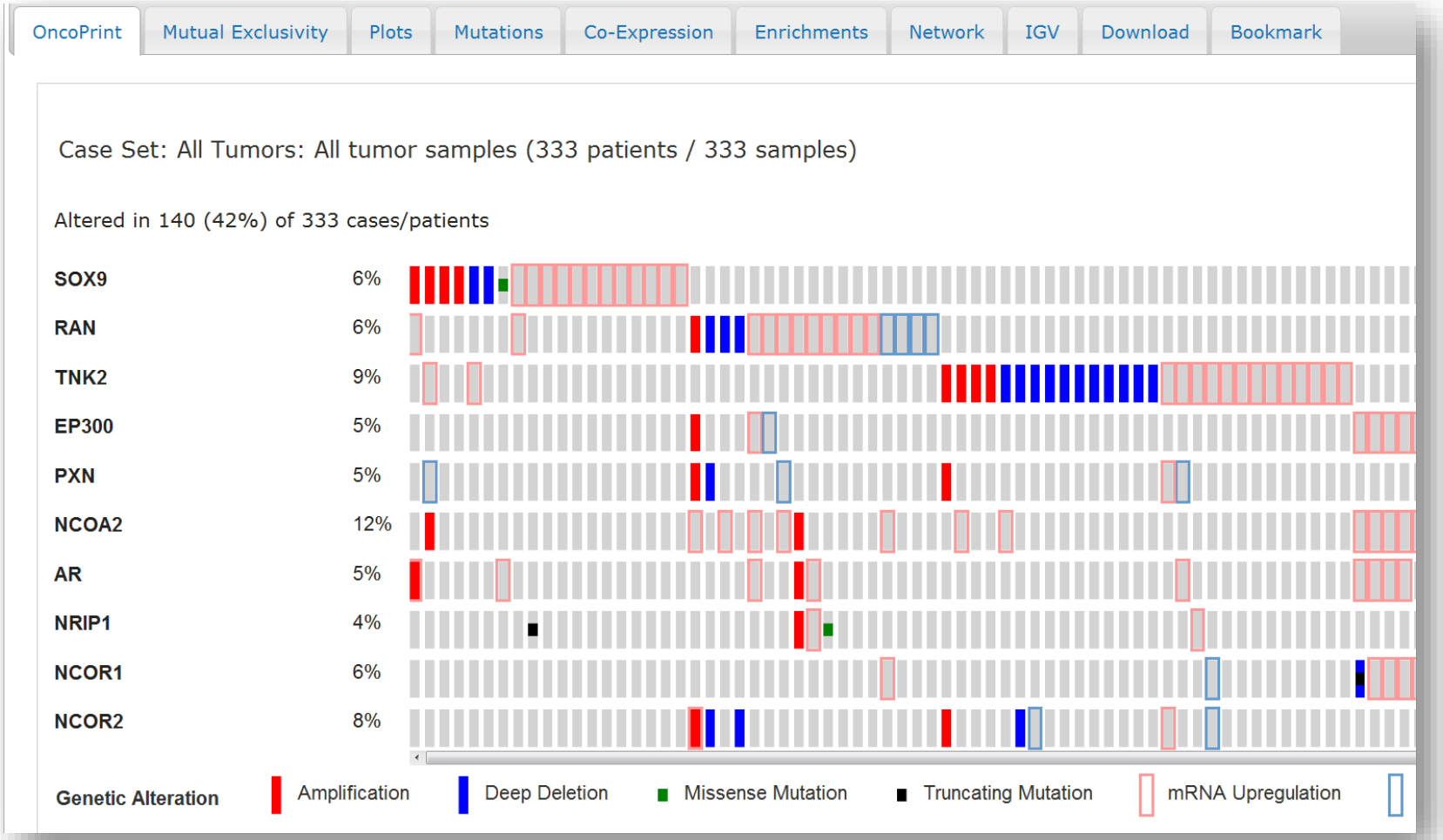
Sample data

Expression matrix (normalized data)

Raw data (cel files)


cBioPortal
for Cancer Genomics
Select Cancer Study:
Search...
1 study selected. Deselect all
Prostate Adenocarcinoma (TCGA, Provisional) 499 samples
Prostate Adenocarcinoma (TCGA, Cell 2015) 333 samples
Select Genomic Profiles:
Mutations
Putative copy-number alterations from GISTIC
mRNA Expression z-Scores (RNA Seq V2 RSEM)
Enter a z-score threshold ±:
2.0
Select Patient/Case Set:
All Tumors (333)
To build your own case set, try out our enhanced Study View.
Enter Gene Set:
Advanced: Onco Query Language (OQL)
Prostate Cancer: AR Signaling (10 genes)
Select From Recurrently Mutated Genes (MutSig)
Select Genes from Recurrent CNAs (Gistic)
SOX9 RAN TNK2 EP300 PXN NCOA2 AR NRIP1 NCOR1 NCOR2

cBioPortal for Cancer Genomics



cBioPortal for Cancer Genomics

Horizontal Axis

Genetic Profile Clinical Attribute

Clinical Attribute

Subtype

Vertical Axis

Genetic Profile Clinical Attribute

Gene EP300

Profile Type mRNA

Profile Name mRNA expression (RNA Sec

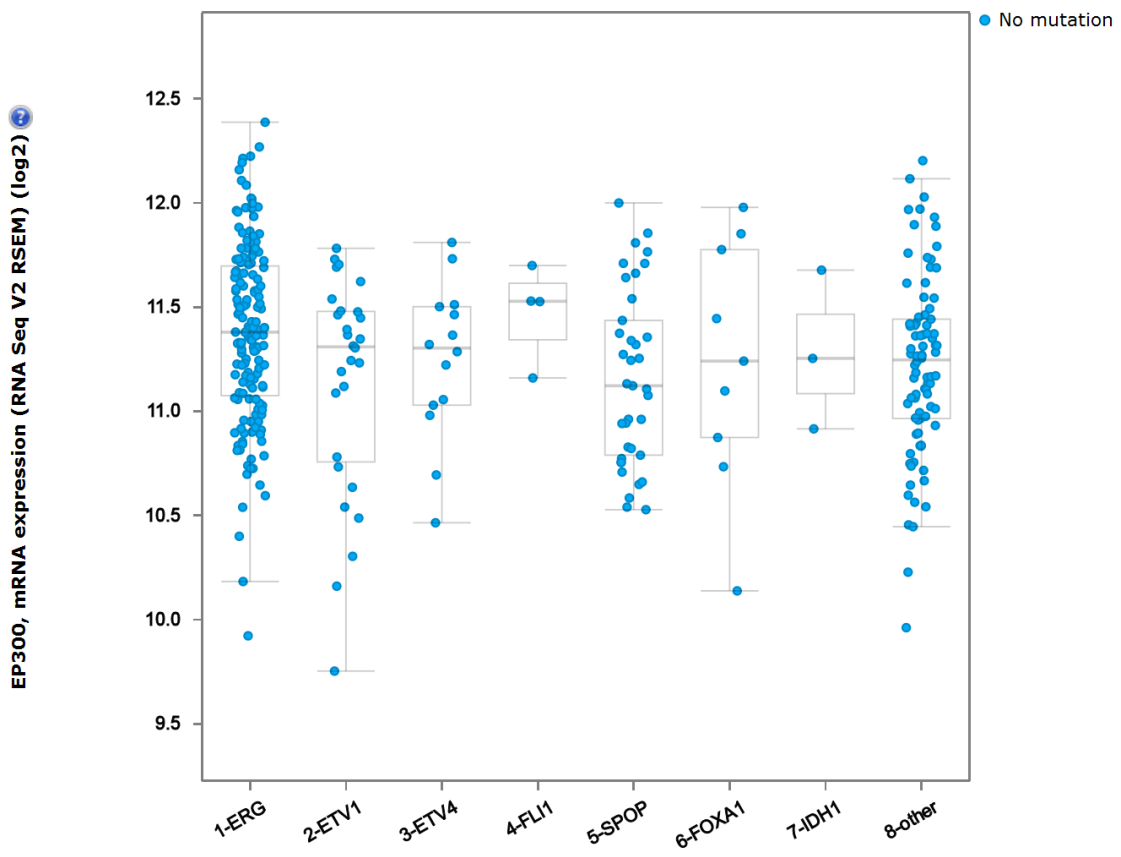
Apply Log Scale

Utilities

Search Case(s) Case ID..

Search Mutation(s) Protein Change..

Download SVG PDF Data

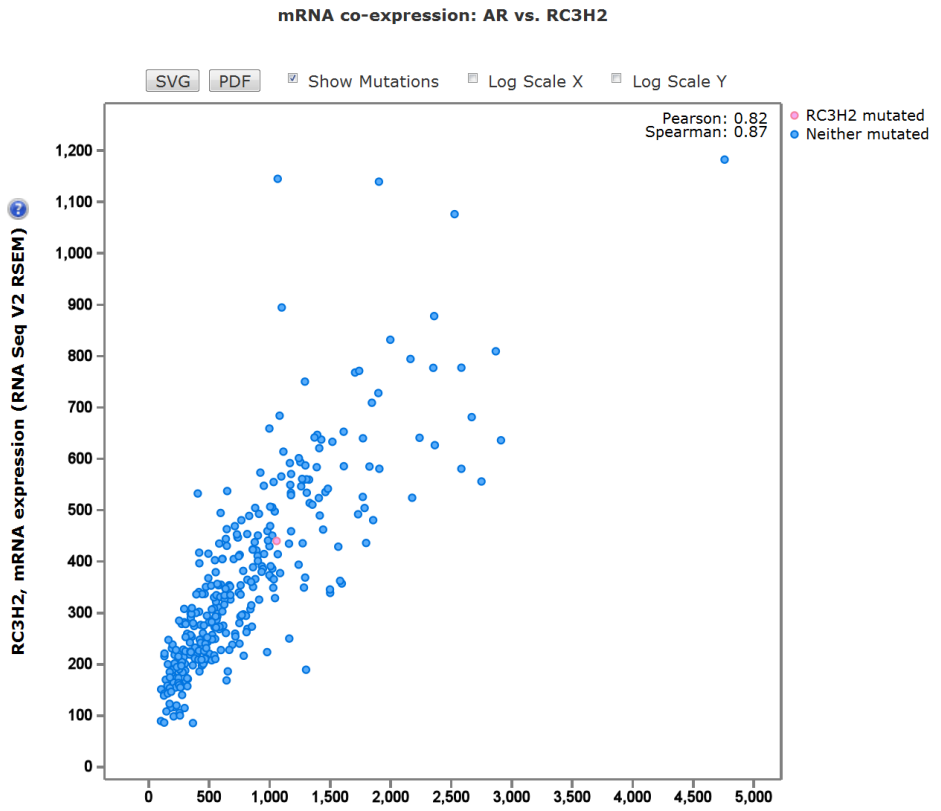


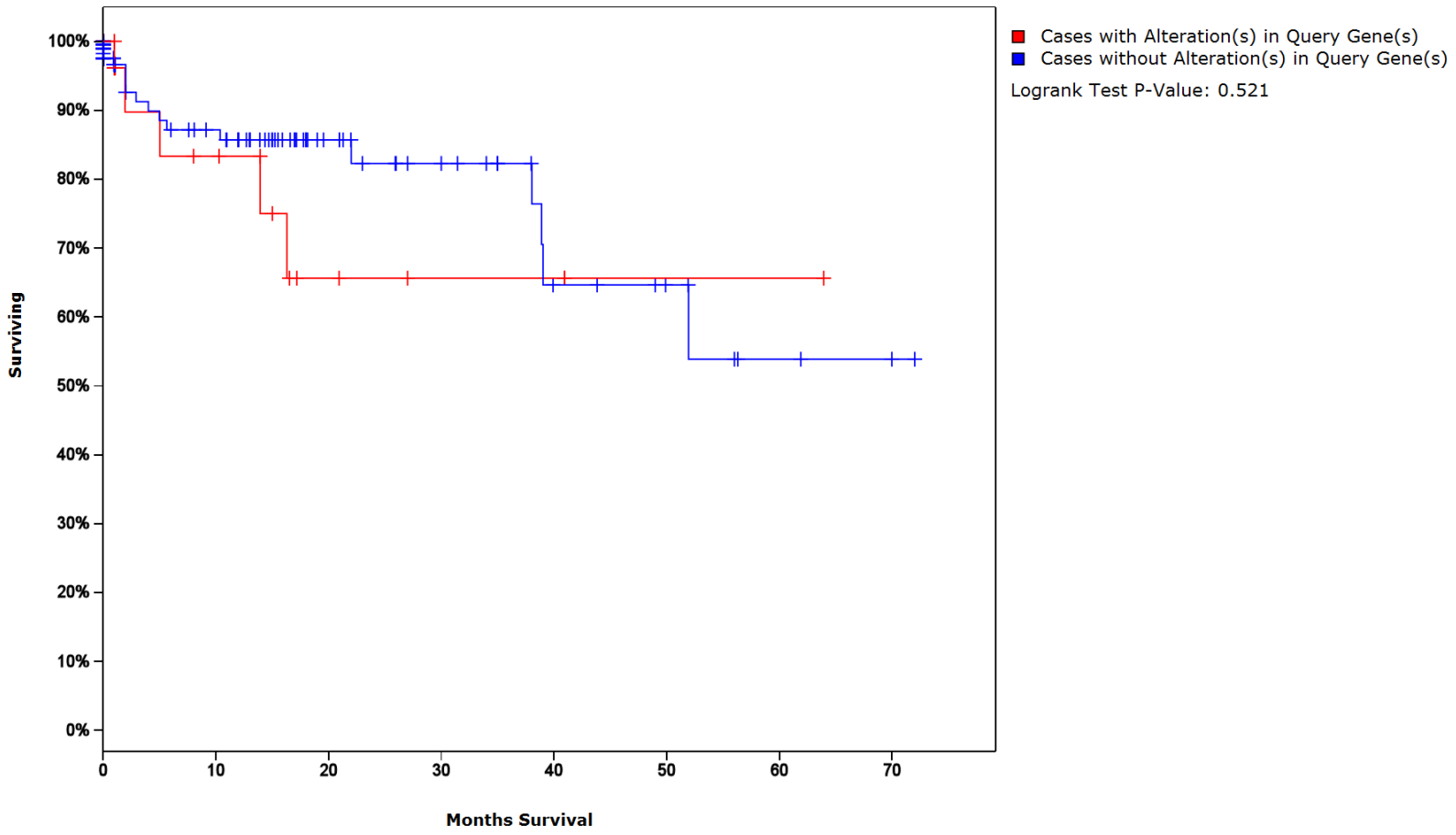
cBioPortal for Cancer Genomics

Search Gene

Show All

Correlated Gene	Pearson's Correlation	Spearman's Correlation
RC3H2	0.82	0.87
ERCC6L2	0.80	0.81
C9ORF129	0.79	0.85
BRWD3	0.79	0.79
UHMK1	0.79	0.86
CCNT1	0.78	0.86
GTF2A1	0.78	0.87
FAM168A	0.78	0.89
UBXN7	0.78	0.85
TOR1AIP2	0.77	0.79
TAOK1	0.77	0.85
BPTF	0.76	0.80
NCOA2	0.76	0.88
KLHL11	0.76	0.84
APOOL	0.75	0.86
ARID1A	0.75	0.80
HUWE1	0.75	0.83
GTF3C4	0.75	0.83
CLOCK	0.75	0.87
CEP97	0.75	0.82
UHRF1BP1	0.75	0.81
WDFY3	0.75	0.84
DDI2	0.75	0.88
BIRC6	0.75	0.84
LMBRD2	0.75	0.85
REST	0.74	0.84
ZNF426	0.74	0.83
NUP155	0.74	0.79



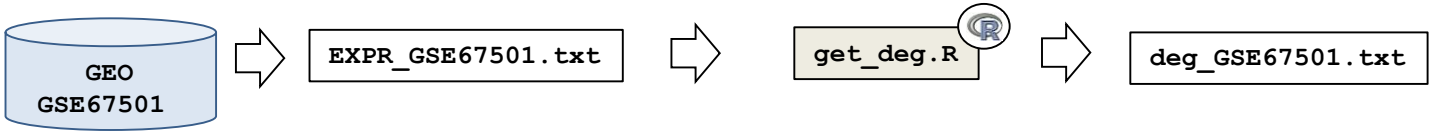

Surviving
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
10
20
30
40
50
60
70
Months Survival
Cases with Alteration(s) in Query Gene(s)
Cases without Alteration(s) in Query Gene(s)
Logrank Test P-Value: 0.521

Differentially expressed genes

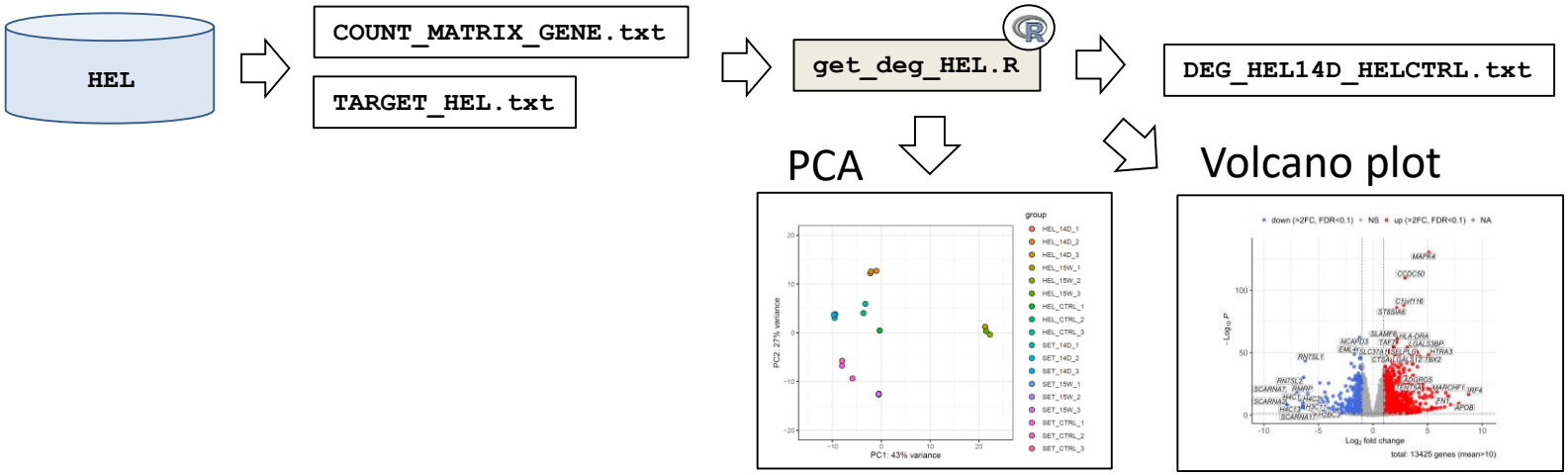
RNA sequencing preprocessing

R introduction

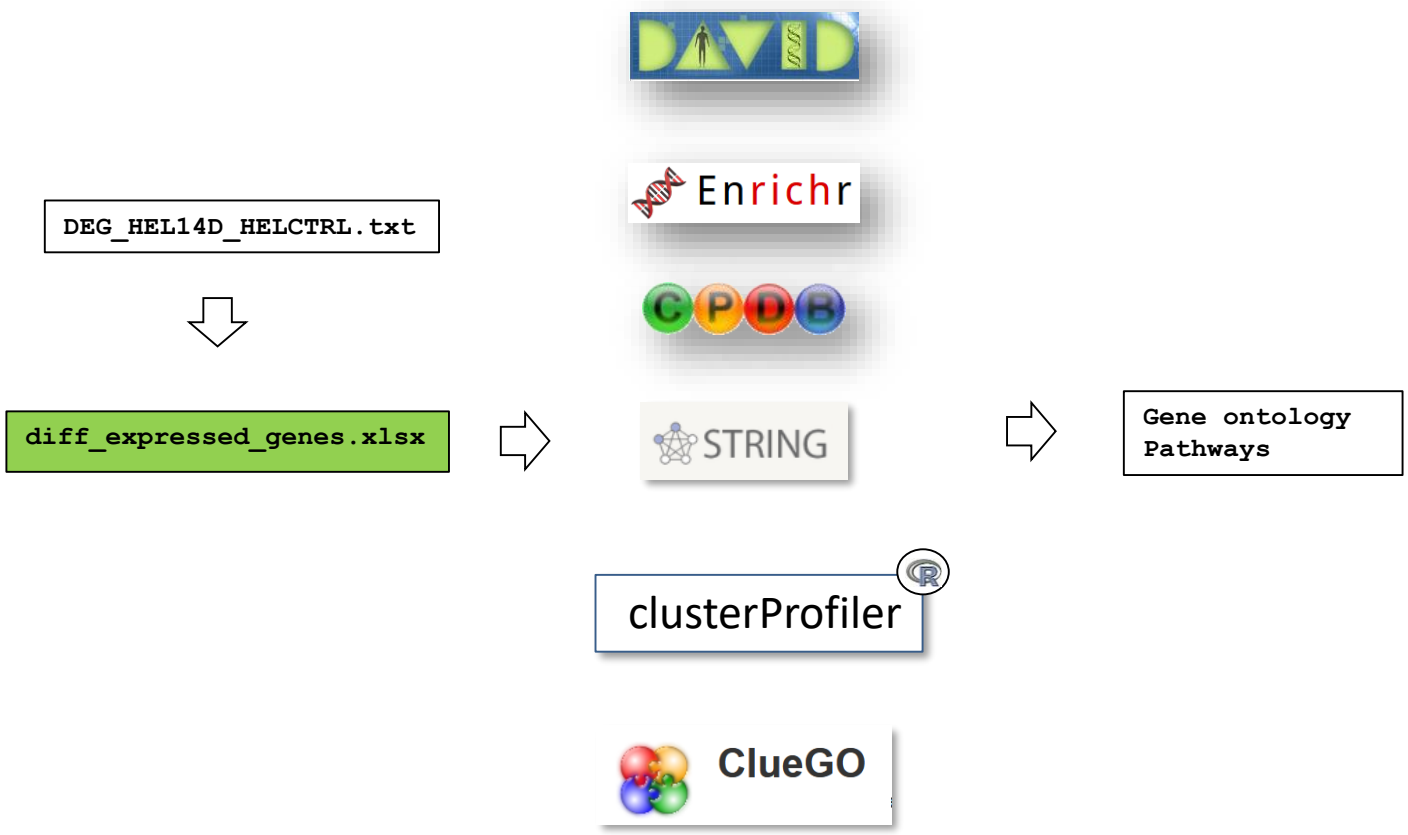
Microarray data (limma)



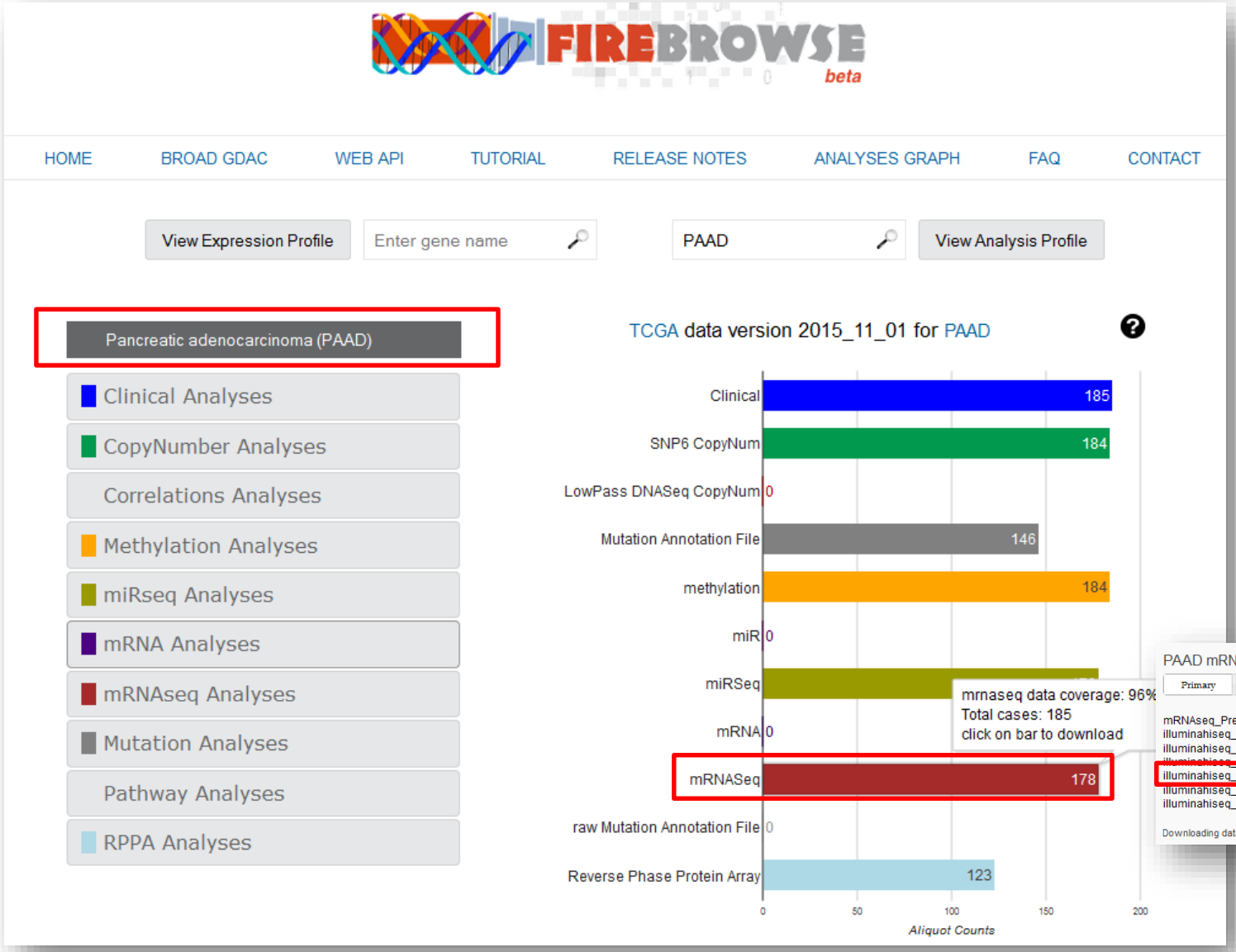
RNAseq data (matrix with raw counts) (DESeq2)



Functional analyses



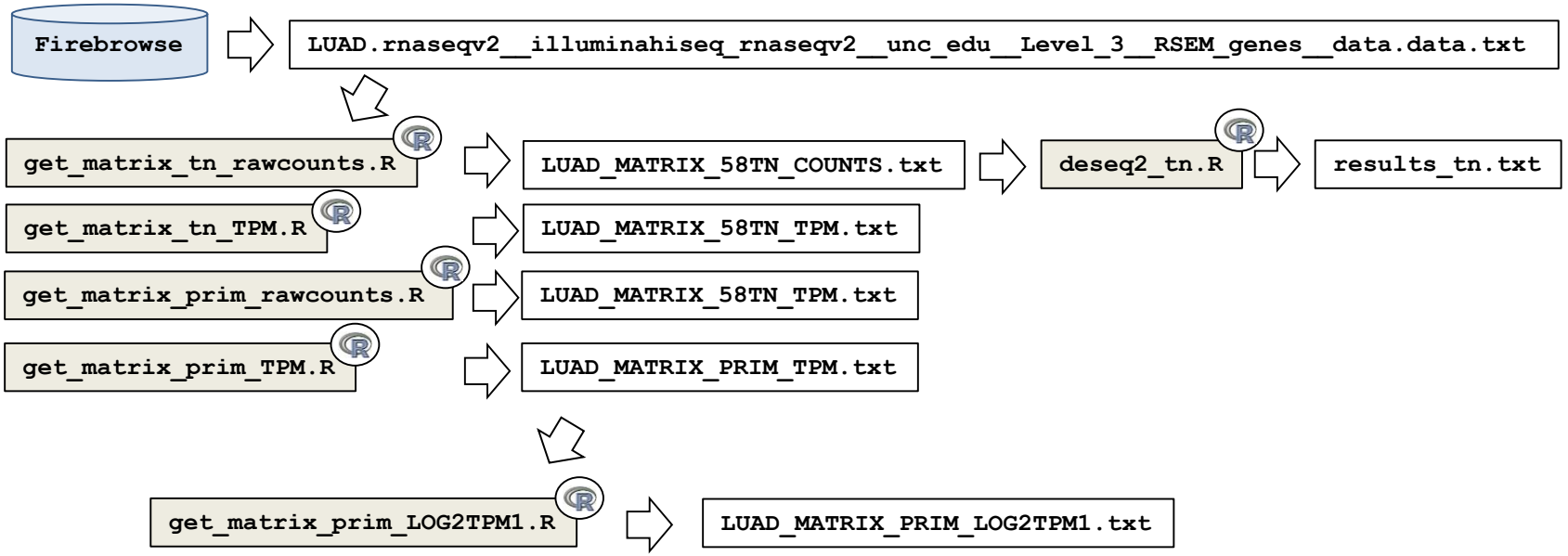
Download RNAseqV2 with Firebrowse



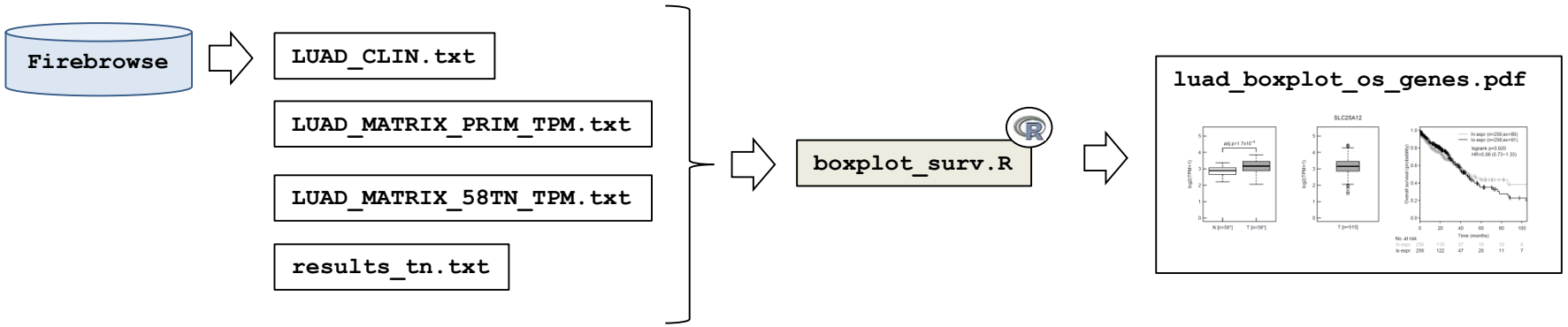
~_Level_3__RSEM_genes__data.data.txt

Hybridization REF	TCGA-2J-AAB1-01A-11R-A41B-07	TCGA-2J-AAB1-01A-11R-A41B-07	TCGA-2J-AAB1-01A-11R-A41B-07	TCGA-2J-AAB4-01A-12R-A41B-07	TCGA-2J-AAB4-01A-12R-A41B-07	TCGA-2J-AAB4-01A-12R-A41B-07	TCGA-2J-AAB6-01A-11R-A41B-07
gene_id	raw_count	scaled_estimate	transcript_id	raw_count	scaled_estin	transcript_id	raw_count
A1BG|1	167.92	3.43E-06	2qsd.3,uc002	134.85	2.46E-06	uc002qsd.3,u	141.16
A1CF|29974	52	9.63E-07	uc001jjk.1,uc0	127	2.03E-06	uc001jjh.2,uc	14
A2BP1|54715	1	8.82E-09	2cyx.2,uc002c	5	4.07E-08	uc002cyr.1,u	0
A2LD1|87769	370.02	8.87E-06	1,uc001vor.2,u	263.92	6.07E-06	uc001voq.1,u	278.94
A2ML1|144568	176	1.49E-06	lqva.1,uc001c	0	0	uc001quz.3,u	3105
A2M|2	40392.8	0.000548528	,uc001qvk.1,u	37630.67	0.00050451	uc001qvj.1,u	14564.83
A4GALT|53947	3160	5.56E-05	3bdb.2,uc010	2744	4.47E-05	uc003bdb.2,u	1917
A4GNT|51146	893	1.89E-05	uc003ers.2	113	2.21E-06	uc003ers.2	2
AAA1|404744	4	1.44E-07	uc010kwp.1,u	1	5.74E-08	uc003tdz.2,u	2
AAAS|8086	1402	2.85E-05	01scr.3,uc001s	1268	2.38E-05	uc001scr.3,uc	1427
AACSL|729522	1	1.69E-08	2,uc011dgk.1,	2	3.13E-08	uc003mjk.2,u	0
AACS|65985	2445	2.89E-05	2,uc009zyg.2,	2915	3.28E-05	uc001uhc.2,u	994

Differentially expressed genes (TCGA, Firebrowse data)

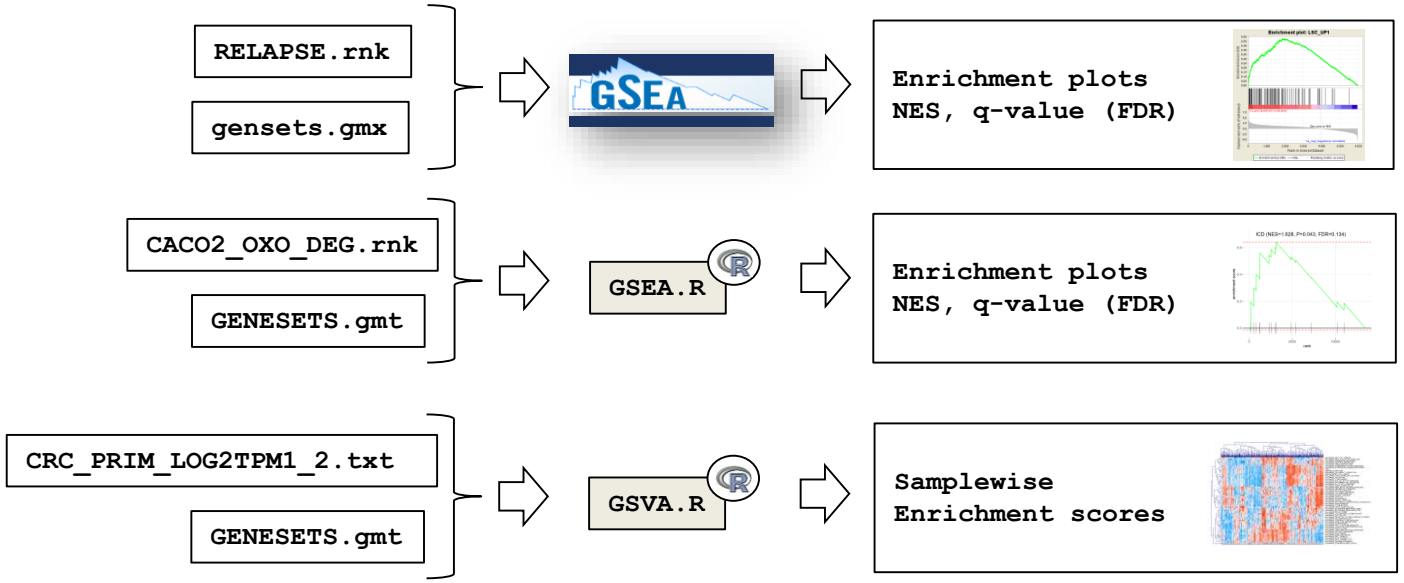


Boxplots and survival analyses



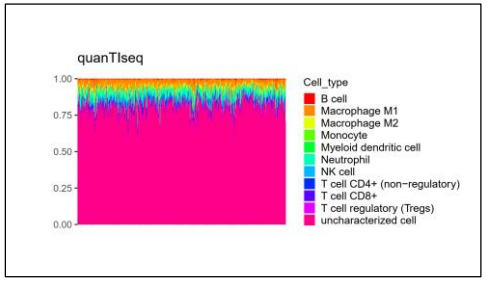
Heatmap and cluster analyses

Gene Set Enrichment Analyses (GSEA)



Deconvolution

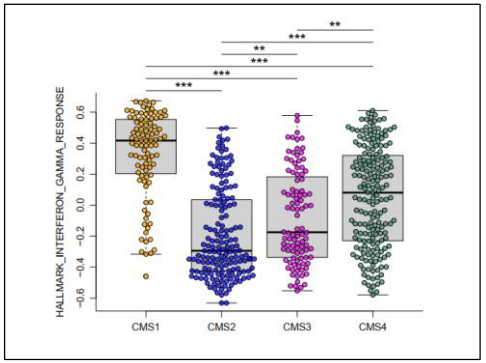
CRC_PRIM_TPM_2.txt → deconv.R →



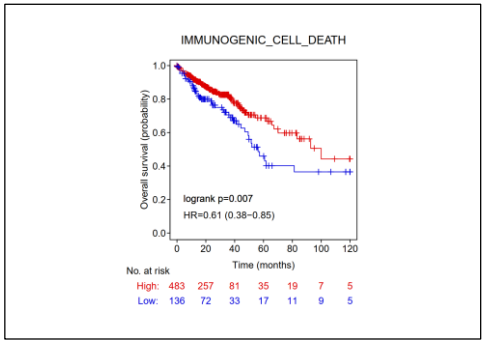
Molecular subtypes

CRC_PRIM_LOG2TPM1_2.txt
GSVA_CRC.txt
CRC_CLIN_2_CMS_MSI_TMB.txt

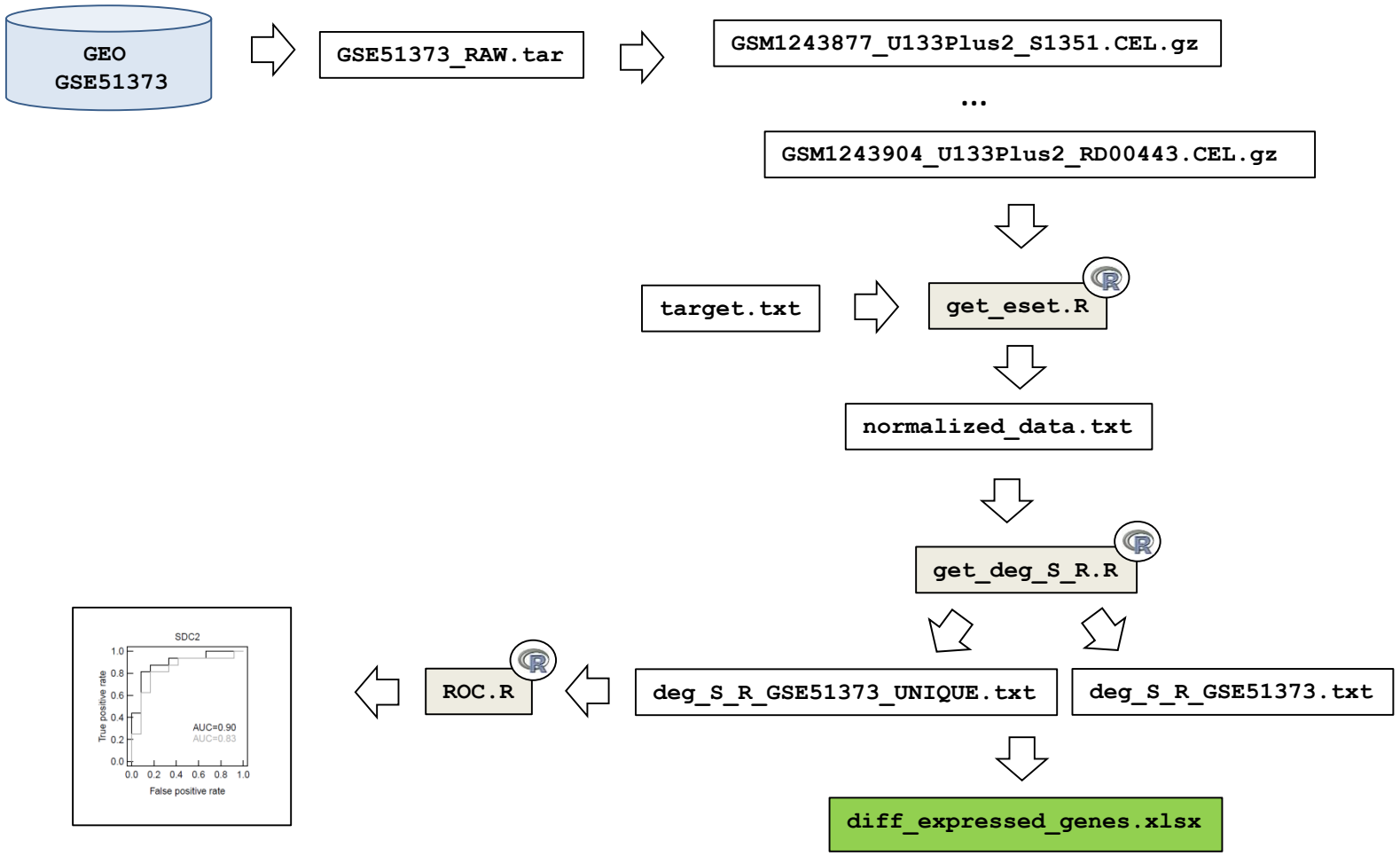
→ BOXPLOT.R →



→ OS.R →

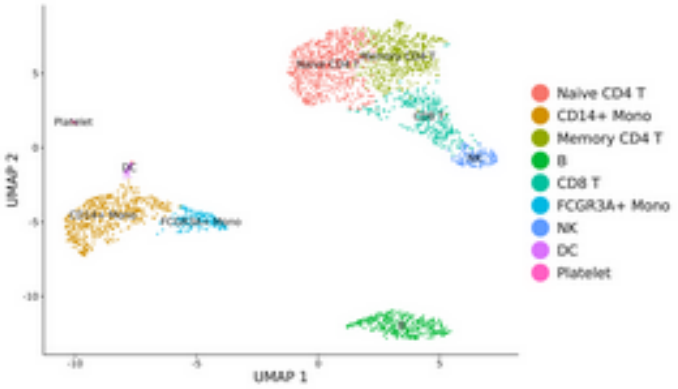


Predictive biomarker



Single cell RNAseq analysis (Seurat)

Guided tutorial — 2,700 PBMCs



A basic overview of Seurat that includes an introduction to common analytical workflows.

GO