Automated analysis of epigenetic and transcriptomic biomarkers to assist individualised treatment success predictions

February 12th 2023, Bled

Gábor Balogh

Introduction - *p*ersonalised/*p*recision medicine

– **PM** $\rightarrow$ customization of healthcare
– treatments tailored to a subgroup of patients (instead of a one-drug-fits-all model)

Introduction - *p*ersonalised/*p*recision medicine

– **PM** $\rightarrow$ customization of healthcare
– treatments tailored to a subgroup of patients (instead of a one-drug-fits-all model)
– appropriate and optimal therapies based on:
  - ▶ sex
  - ▶ medical history
  - ▶ life habits
  - ▶ molecular or cellular traits
  - ▶ patient's genetics ($\rightarrow$ needs approved biomarkers in prior)

Introduction - Database

– Using the TCGA-Projects (The Cancer Genome Atlas):

- ► cancer genomics program, molecularly characterized over 20,000 primary cancer
- ► spanning over 33 cancer types
- ► metadata available (treatment regime, survival data etc.)

Introduction - Pipeline

– With help of publicly available data we aim to:
  ▶ link epigenomic and transcriptomic data with treatment and survival data
  ▶ identify possible biomarkers that indicate treatment success

Introduction - Pipeline

– With help of publicly available data we aim to:
  ▶ link epigenomic and transcriptomic data with treatment and survival data
  ▶ identify possible biomarkers that indicate treatment success

– focus on:
  ▶ flexible query
  ▶ automatization
  ▶ reproducibility
  ▶ parallelisation

Usage of the pipeline in command line

Where to save the
downloaded data and the
results?:

Choose from any
TCGA-Project:

Choose any drug/
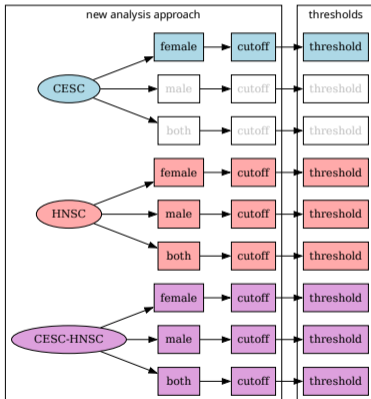drugcombination:

-o OUTPUT_PATH

-p TCGA-CESC
-p TCGA-HNSC

-d carboplatin
-d carboplatin,paclitaxel
-d cisplatin

TreSucMs -o OUTPUT_PATH -p TCGA-CESC -p TCGA-HNSC -d carboplatin -d carboplatin,paclitaxel -d cisplatin

Set combinations

Command line call:

```
TreSucMs -o OUTPUT_PATH -p TCGA-CESC -p TCGA-HNSC
-d carboplatin -d carboplatin,paclitaxel -d cisplatin:
```
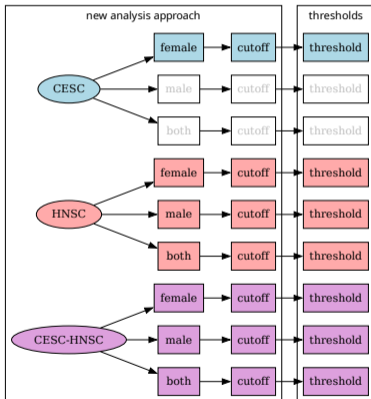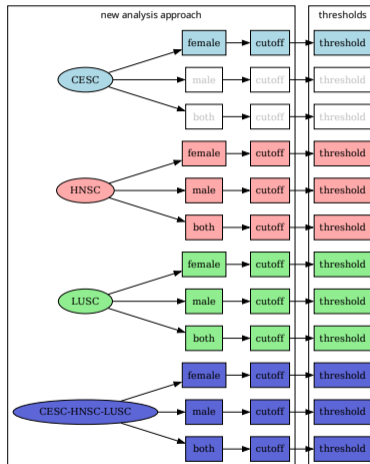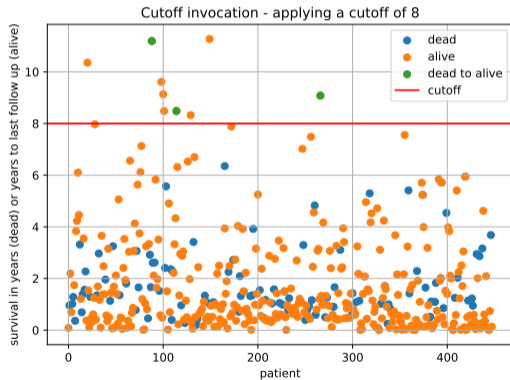
Set combinations

Pipeline schema

Data preprocessing:
Acquiring, filtering, revising, linking

Analysis step,
either metilene or DESeq2

Postprocessing:
validation steps, ranking of outcomes

gathering raw data according to users made options

according to the filtering and grouping of the given metadata, the DESeq2 or metilene analysis are started

most significant differential features are observed furher

base plots

Kaplan Meier plots based on the mean of medians of the features

threshold parameter

differential analysis between dead and alive vital status

Feature names (either ENSG or position in DMR

cutoff parameter

threshold parameter

validation plots

confirming the suggestion of the base plot with the complete patients set, separated by the mean of medians in up and down regulated:

Up-group

Cutoff parameter

UNIVERSITÄT
LEIPZIG

Pipeline schema

Data preprocessing:
Acquiring , filtering, revising, linking

Analysis step, either metilene or DESeq2

Postprocessing: validation steps, ranking of outcomes

gathering raw data according to users made options

according to the filtering and grouping of the given metadata, the DESeq2 or metilene analysis are started

most significant differential features are observed furher

differential analysis between dead and alive vital status

Feature names (either ENSG or position in DMR

cutoff parameter

threshold parameter

base plots

Kaplan Meier plots based on the mean of medians of the features

threshold parameter

validation plots

confirming the suggestion of the base plot with the complete patients set, separated by the mean of medians in up and down regulated:

Up-group
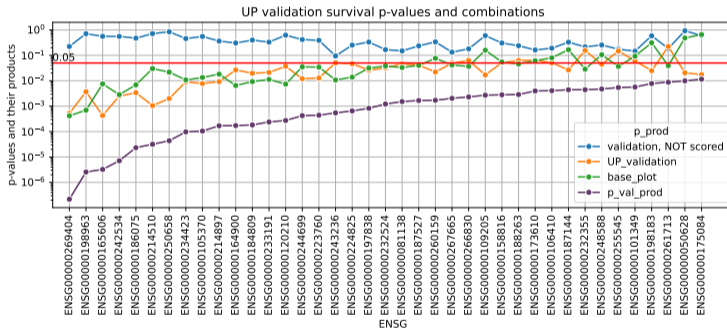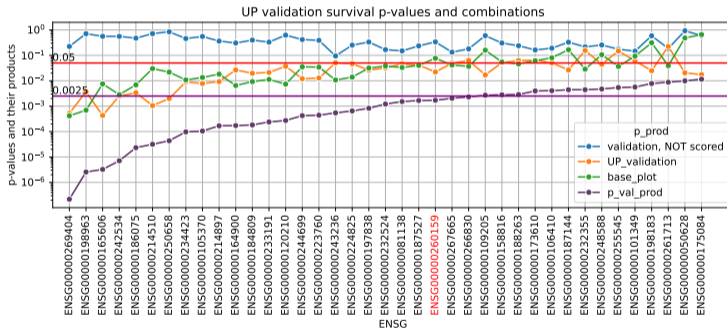
Aggregating all results of a patient cohort

aggregation over base and validation plot, sorted over product of p values, ENSGs (norm_count)
TCGA-CESC, TCGA-HNSC, TCGA-LUSC, female, male,
carboplatin; carboplatin,paclitaxel; cisplatin, cutoff=0

UP validation survival p-values and combinations
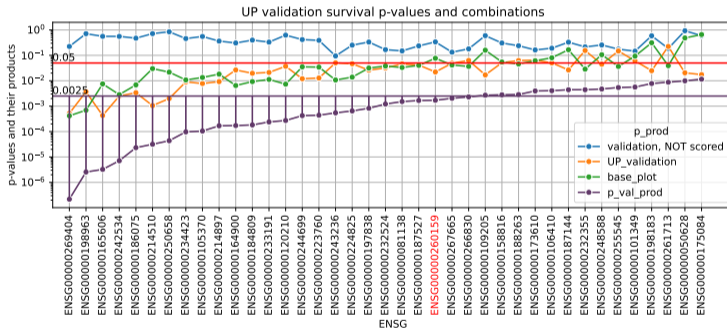
IZBI Leipzig

UNIVERSITÄT
LEIPZIG

Aggregating all results of a patient cohort

aggregation over base and validation plot, sorted over product of p values, ENSGs (norm_count)
TCGA-CESC, TCGA-HNSC, TCGA-LUSC, female, male,
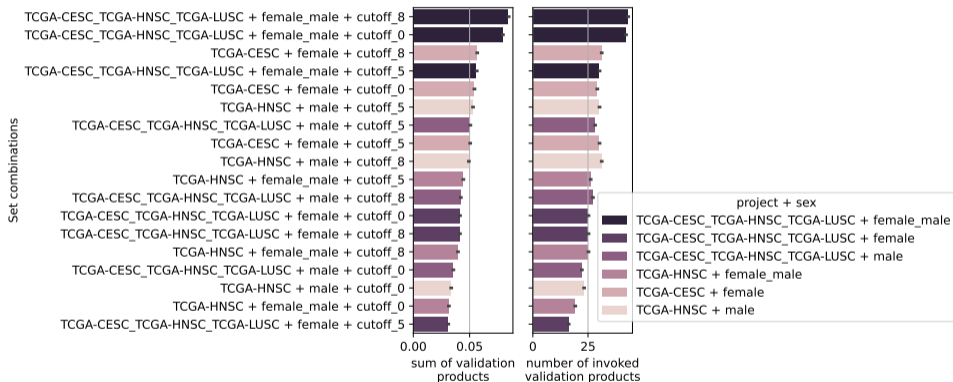carboplatin; carboplatin,paclitaxel; cisplatin, cutoff=0

Aggregating all results of a patient cohort

aggregation over base and validation plot, sorted over product of p values, ENSGs (norm_count)
TCGA-CESC, TCGA-HNSC, TCGA-LUSC, female, male,
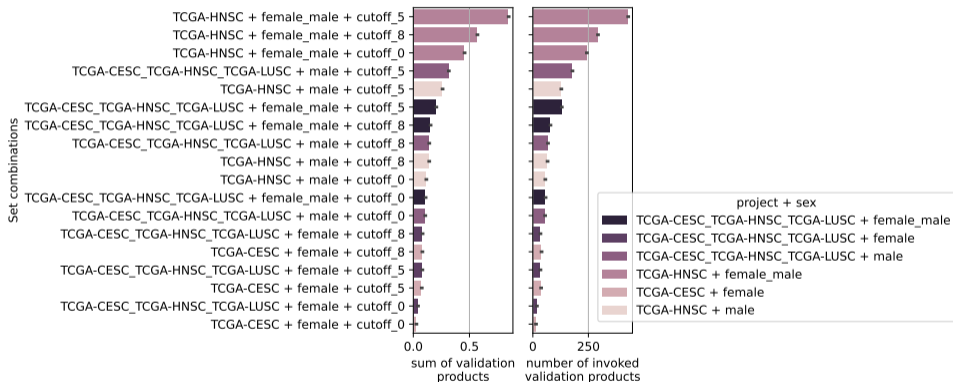carboplatin; carboplatin,paclitaxel; cisplatin, cutoff=0

UP validation survival p-values and combinations

UNIVERSITÄT
LEIPZIG

Aggregating all patient cohorts, DESeq2

DESeq2: Summary of the p-value validation products for all sets

UNIVERSITÄT
LEIPZIG

Aggregating all patient cohorts, metilene

metilene: Summary of the p-value validation products for all sets

UNIVERSITÄT
LEIPZIG

Best outcomes metilene

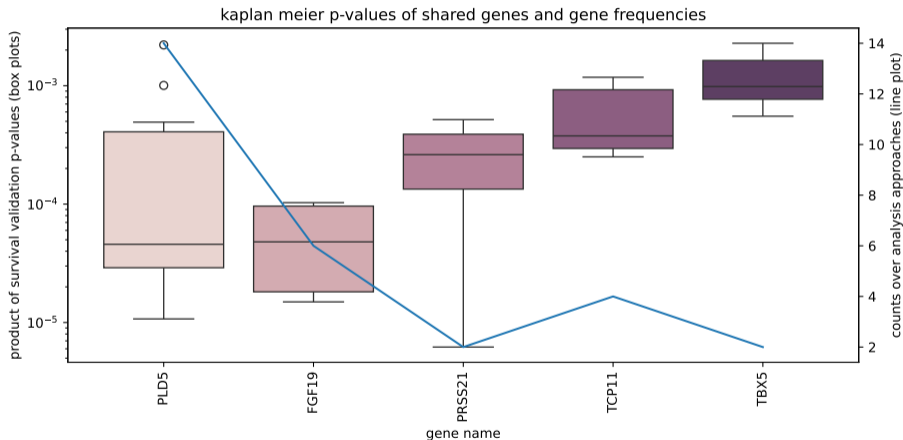| project	sex	cutoff	gene name	CMP	validation products	mmd
CESC-HNSC-LUSC	female	cutoff-8	SALL1	UP	2.51e-09	0.063997
CESC-HNSC-LUSC	female-male	cutoff-0	MSX2	UP	5.38e-09	0.064413
CESC-HNSC-LUSC	female-male	cutoff-8	TRAPPC9	UP	1.31e-08	0.164241
CESC-HNSC-LUSC	female-male	cutoff-8	FLRT2	UP	1.98e-08	0.077226
CESC-HNSC-LUSC	female-male	cutoff-8	RP11-497E19.1	UP	1.98e-08	0.077226
CESC-HNSC-LUSC	female-male	cutoff-8	RP11-497E19.2	UP	1.98e-08	0.077226
CESC-HNSC-LUSC	female-male	cutoff-8	CAMTA2	UP	5.75e-08	0.149052
CESC-HNSC-LUSC	female-male	cutoff-8	RBP1	UP	1.04e-07	0.134231
CESC-HNSC-LUSC	female-male	cutoff-8	RP11-319G6.1	UP	1.04e-07	0.134231
CESC-HNSC-LUSC	female	cutoff-8	WNT16	UP	1.57e-07	0.137616

Table: Ordered after the validation products. 10 best biomarker candidates found on the bases of metilene. CMP refers to the group with longer survival expectation.

Best outcomes DESeq2

project	sex	cutoff	gene name	regulation	validation products	log2FCh
CESC-HNSC-LUSC	female-male	cutoff-8	SPIB	UP	3.91e-08	-2.604771
CESC	female	cutoff-0	EREG	DOWN	5.02e-07	2.598291
HNSC	female-male	cutoff-0	FADD	DOWN	6.79e-07	1.201673
HNSC	male	cutoff-5	RP11-20B7.1	DOWN	8.46e-07	2.665860
CESC-HNSC-LUSC	female-male	cutoff-8	RORB	UP	9.10e-07	-2.925838
CESC	female	cutoff-8	RAET1E-AS1	DOWN	9.16e-07	2.544801
CESC-HNSC-LUSC	female	cutoff-0	SLN	DOWN	1.55e-06	3.524010
CESC	female	cutoff-0	SCG2	DOWN	1.97e-06	2.142930
CESC-HNSC-LUSC	male	cutoff-5	RP11-94A24.1	DOWN	2.96e-06	2.105692
CESC	female	cutoff-0	CXCL2	DOWN	3.11e-06	2.062064

Table: Ordered after the validation products. 10 best biomarker candidates found on the bases of DEseq2. CMP refers to the group with longer survival expectation.

Analysis comprehensive results

kaplan meier p-values of shared genes and gene frequencies

Summary

– metilene:
  ► found 497 genes which passed the p-value product boundary
– DEseq2:
  ► found 182 genes which passed the p-value product boundary
– 5 genes shared across both pipelines
– 674 unique genes all together
– 321 genes unique to cutoff =! 0
– 35 genes unique to cutoff == 0
– 318 in common

Summary

patients from TCGA-CESC, TCGA-HNSC, TCGA-LUSC:
total cases: 448 patients

no cutoff		cutoff 8		cutoff 5	
vital status	count	vital status	count	vital status	count
alive	338	alive	341	alive	348
dead	110	dead	107	dead	100

Summary

patients from TCGA-CESC, TCGA-HNSC, TCGA-LUSC:
total cases: 448 patients

no cutoff			cutoff 8			cutoff 5		
vital status	in therapy	count	vital status	in therapy	count	vital status	in therapy	count
alive	yes	174	alive	yes	177	alive	yes	179
alive	no	164	alive	no	164	alive	no	169
dead	no	59	dead	no	59	dead	no	54
dead	yes	51	dead	yes	48	dead	yes	46

Thank You!

Gábor Balogh

IZBI Leipzig
Department for Computer Science

`gabor@bioinf.uni-leipzig.de`

`https://www.bioinf.uni-leipzig.de`