PiCnIc pipeline for Lung Adenocarcinoma

Davide Cozzi, 829827 - Mattia Sgrò, 829474

Dipartimento di Informatica, Sistemistica e Comunicazione (DISCo)
Università degli Studi di Milano Bicocca

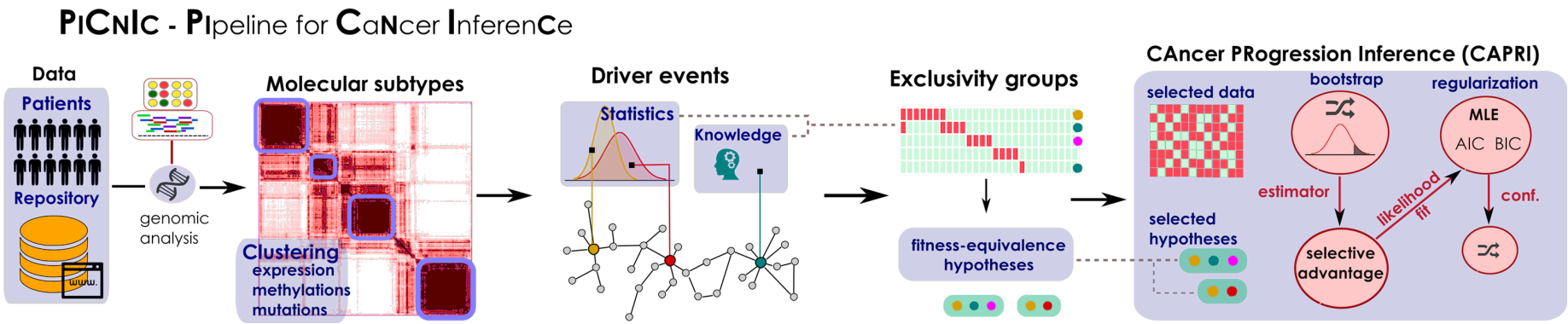
Outline

Outline

The pipeline for LUAD analysis

Aim of the project

With this project we intend to built a cancer progression model from cross-sectional ***lung adenocarcinoma*** data, using the **PiCnIc pipeline** [1] via the **R TRONCO library** [2].



Lung Cancer

Histological classification of lung cancer [3]

Classification is based on:

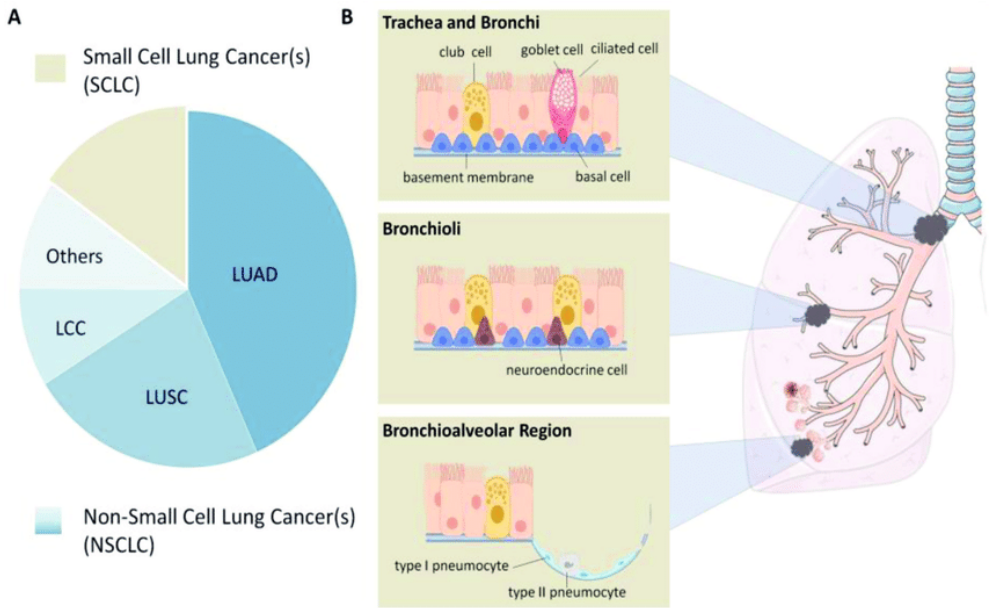
- Histological type of lung cancer
- Location of the tumors and cell origins

Lung Cancer

Histological classification of lung cancer [3]

Classification is based on:

- Histological type of lung cancer
- Location of the tumors and cell origins



Lung Adenocarcinoma I

Lung adenocarcinoma (*LUAD*) is the most common histological type of non-small cell lung cancer (*NSCLC*).
Due to the nonspecific early symptoms (that are similar across other forms of lung cancer, i.e. persistent cough and shortness of breath), the majority of the diagnosed *LUAD* patients are in the middle and late stages, with multiple metastases, and have missed the optimal period for treatment [4].

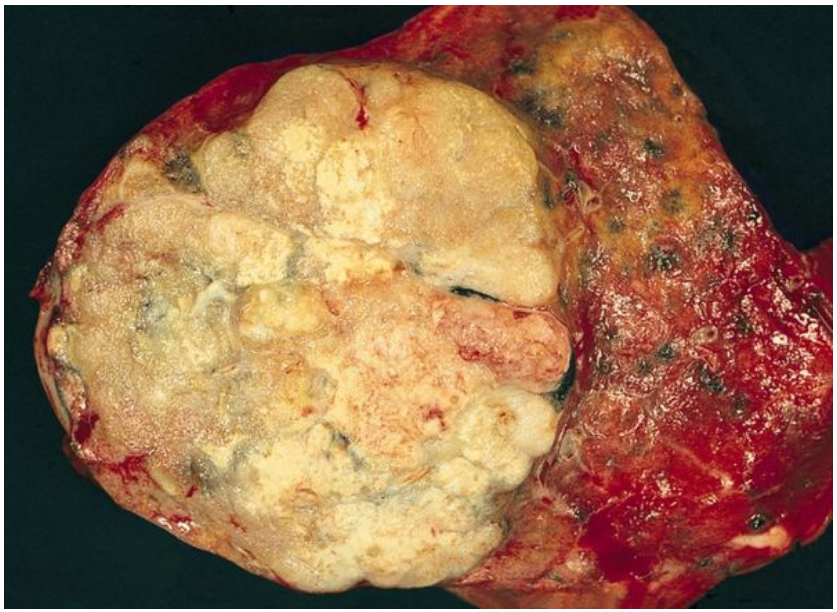
Lung Adenocarcinoma I

Lung adenocarcinoma (*LUAD*) is the most common histological type of non-small cell lung cancer (*NSCLC*).
Due to the nonspecific early symptoms (that are similar across other forms of lung cancer, i.e. persistent cough and shortness of breath), the majority of the diagnosed *LUAD* patients are in the middle and late stages, with multiple metastases, and have missed the optimal period for treatment [4].

Adenocarcinoma is more common in patients with a history of cigarette smoking, and is the most common form of lung cancer in younger women and Asian populations. The pathophysiology of adenocarcinoma is complicated, but generally follows a histologic progression from cells found in healthy lungs to distinctly dysmorphic, or irregular cells [5].

Lung Adenocarcinoma II

Adenocarcinoma of the lung is the leading cause of cancer death worldwide [6].



Outline

Genes Drivers and Pathways I

A first attempt

At first we tried to use the results of some software present in the literature for the choice of genes drivers [7].

Genes Drivers and Pathways I

A first attempt

At first we tried to use the results of some software present in the literature for the choice of genes drivers [7].

IMC Driver	*SCS*	*ACtive Driver*	MutSigCV	*Dawn Rank*	*Driver ML*	*Driver Net*
TPS3	SRGAP3	NUP155	C12ORF5	MAPK3	EGFR	GRIN2B
LRP1B	PCDHGC5	LCLAT1	C16ORF3	PRKCB	SETD2	MET
KRAS	ZNFI17	NEK10	C19ORF70	PRKACB	STK11	RYR2
TRN	MST1R	SIK31	C8ORF59	PRKX	TP53	PIK3CA
KEAP1	ZWINT	SOS1	EGFR	HRAS	RB1	ADCY8
...	...	...	...	...	...	...

Genes Drivers and Pathways II

Final selection

At the end we decided to use the genes drivers and pathways noted in the marker paper [6].

Published in final edited form as:
Nature. 2014 July 31; 511(7511): 543–550. doi:10.1038/nature13385.

Comprehensive molecular profiling of lung adenocarcinoma

The Cancer Genome Atlas Research Network

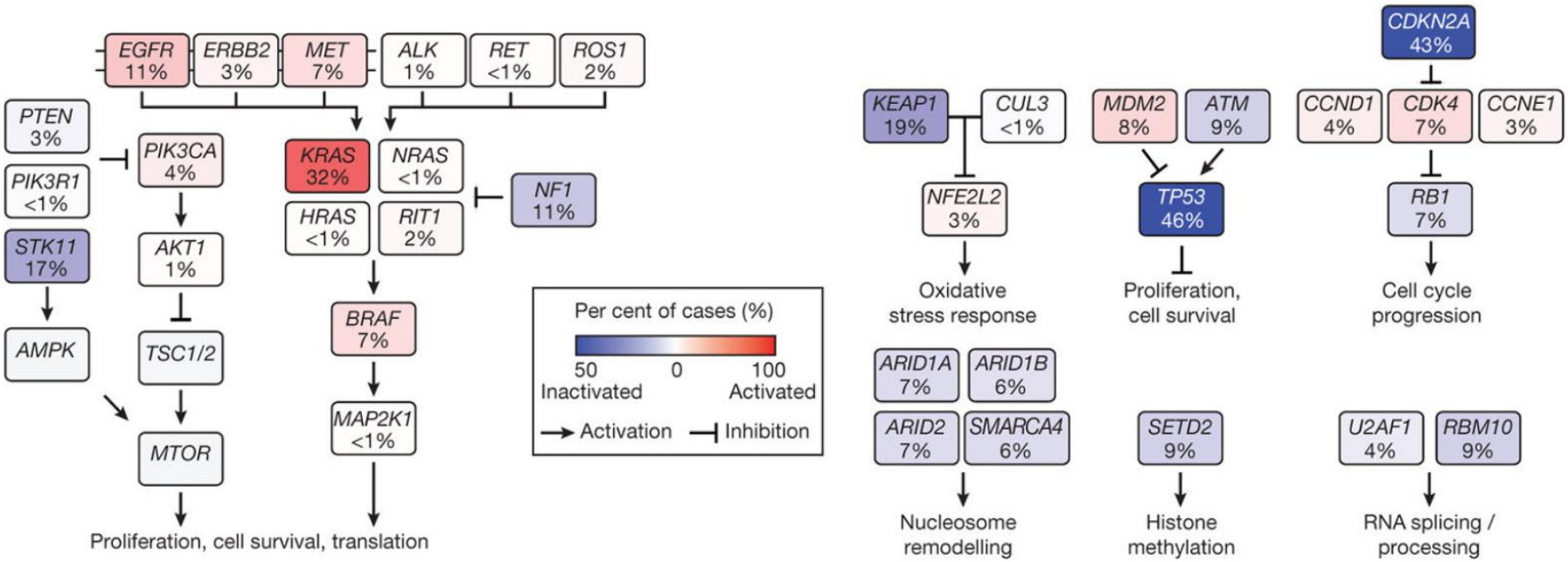
Abstract

Adenocarcinoma of the lung is the leading cause of cancer death worldwide. Here we report molecular profiling of 230 resected lung adenocarcinomas using messenger RNA, microRNA and DNA sequencing integrated with copy number, methylation and proteomic analyses. High rates of somatic mutation were seen (mean 8.9 mutations per megabase). Eighteen genes were statistically significantly mutated, including *RIT1* activating mutations and newly described loss-of-function *MGA* mutations which are mutually exclusive with focal *MYC* amplification. *EGFR* mutations were more frequent in female patients, whereas mutations in *RBM10* were more common in males. Aberrations in *NF1*, *MET*, *ERBB2* and *RIT1* occurred in 13% of cases and were enriched in samples otherwise lacking an activated oncogene, suggesting a driver role for these events in certain tumours. DNA and mRNA sequence from the same tumour highlighted splicing alterations driven by somatic genomic changes, including exon 14 skipping in *MET* mRNA in 4% of cases. MAPK and PI(3)K pathway activity, when measured at the protein level, was explained by known mutations in only a fraction of cases, suggesting additional, unexplained mechanisms of pathway activation. These data establish a foundation for classification and further investigations of lung adenocarcinoma molecular pathogenesis.

Genes Drivers and Pathways II

Final selection

At the end we decided to use the genes drivers and pathways noted in the marker paper [6].



Genes Drivers and Pathways III

Pathway details using **WikiPathways R library** [8] (although some results are not directly related to *LAUD*)

Short Name	*Genes*
P53	*TP53 ATM MDM2*
MAPK	*KRAS NRAS HRAS RIT1 NF1 BRAF MAP2K1 EGFR ERBB2 MET ALK RET ROS1*
MTOR	*PTEN PIK3CA PIK3R1 STK11 AKT1 AMPK TSC1 TSC2 MTOR*
OXI	*KEAP1 CUL3 NFE2L2*
PROG	*CDKN2A CCND1 CDK4 CCNE1 RB1*
REMO	*ARID1A ARID1B ARID2 SMARCA4*
HIME	*SETD2*
RNASPL	*RBM10 U2AF1*

Genes Drivers and Pathways III

Pathway details using **WikiPathways R library** [8] (although some results are not directly related to *LAUD*)

Short Name	*First title by WikiPathways*
P53	ATM signaling pathway
MAPK	EGFR tyrosine kinase inhibitor resistance
MTOR	Energy dependent regulation of mTOR by LKB1-AMPK
OXI	Photodynamic therapy-induced NFE2L2 (NRF2) survival signaling
PROG	G1 to S cell cycle control
REMO	Tumor suppressor activity of SMARCB1
HIME	Histone modifications
RNASPL	Processing of Capped Intron-Containing Pre-mRNA

Genes Drivers and Pathways III

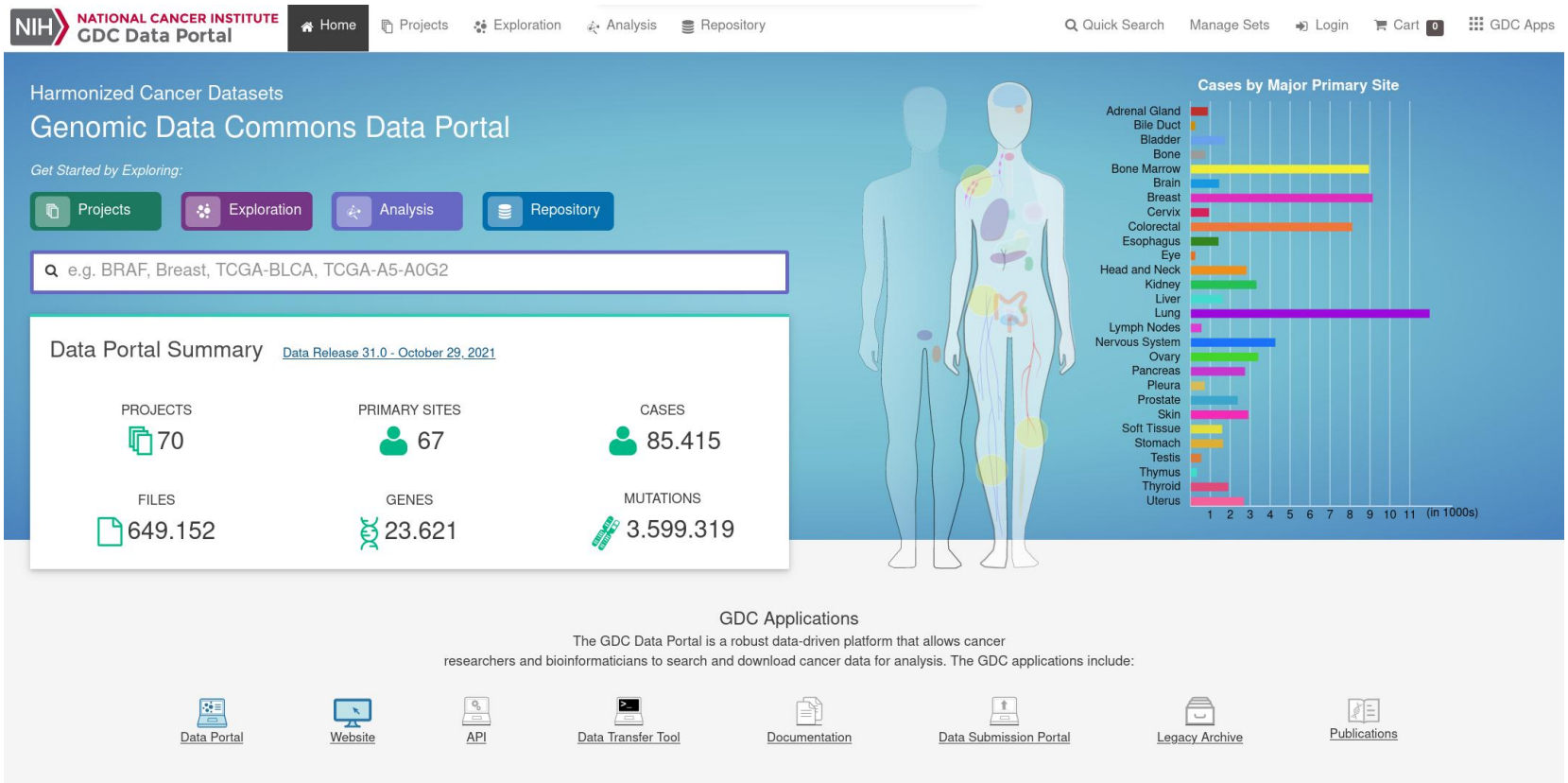
Pathway details using **WikiPathways R library** [8] (although some results are not directly related to *LAUD*)

Short Name	*Second title by WikiPathways*
P53	TP53 network
MAPK	NRF2-ARE regulation
MTOR	Focal adhesion: PI3K-Akt-mTOR-signaling pathway
OXI	NRF2-ARE regulation
PROG	Cell cycle
REMO	Transcriptional regulation by RUNX1
HIME	
RNASPL	mRNA processing

Outline

Data Download

Regarding the data (MAF files, GISTIC files, clinical files and subtypes files) we choose to use *GDC portal* and *Firehose*, using the **TCGAbiolinks R library** [9] to automate the download directly into the source code.



MAF File I

First the MAF file is downloaded, containing the data of the various somatic mutations. Important are the column called *Hugo Symbol*, containing the code name of the gene, and the *Variant classification* column, indicating the type of mutation.

	Hugo_Symbol	Entrez_Gene_Id	Center	NCBI_Build	Chromosome	Start_Position	End_Position	Strand	Variant_Classificatio
1	AC215219.2	0	BI	GRCh38	chr12	14878	14878	+	3'Flank
2	AC215219.2	0	BI	GRCh38	chr12	15938	15938	+	3'Flank
3	WASH1	100287171	BI	GRCh38	chr9	15952	15952	+	Silent
4	MIR6859-4	103504738	BI	GRCh38	chr16	16943	16943	+	3'Flank
5	MIR6859-4	103504738	BI	GRCh38	chr16	16989	16989	+	3'Flank
6	WASH1	100287171	BI	GRCh38	chr9	17145	17145	+	Missense_Mutation
7	FAM110C	642273	BI	GRCh38	chr2	45522	45522	+	Silent
8	FAM110C	642273	BI	GRCh38	chr2	45600	45600	+	Silent
9	FAM110C	642273	BI	GRCh38	chr2	45614	45614	+	Missense_Mutation
10	TUBB8	347688	BI	GRCh38	chr10	47273	47273	+	Silent
11	TUBB8	347688	BI	GRCh38	chr10	47380	47380	+	Missense_Mutation
12	TUBB8	347688	BI	GRCh38	chr10	47453	47453	+	Silent
13	TUBB8	347688	BI	GRCh38	chr10	47468	47468	+	Silent
14	TUBB8	347688	BI	GRCh38	chr10	47513	47513	+	Missense_Mutation
15	TUBB8	347688	BI	GRCh38	chr10	47523	47523	+	Missense_Mutation
16	TUBB8	347688	BI	GRCh38	chr10	47632	47632	+	Missense_Mutation
17	TUBB8	347688	BI	GRCh38	chr10	47633	47633	+	Silent
18	TUBB8	347688	BI	GRCh38	chr10	47913	47913	+	Missense_Mutation
19	TUBB8	347688	BI	GRCh38	chr10	47923	47923	+	Nonsense_Mutation

MAF file II

We can plot information about the complete MAF file, using **Maftools R library** [10], such as:

- *variant classification and type*
- *single-nucleotide variant class*
- *genes mutated more frequently*

Some passages of the marker paper can be found in these data

Past or present smoking associated with cytosine to adenine ($C > A$) nucleotide transversions as previously described both in individual genes and genome-wide [6].
We will talk more about smokers.

MAF file II

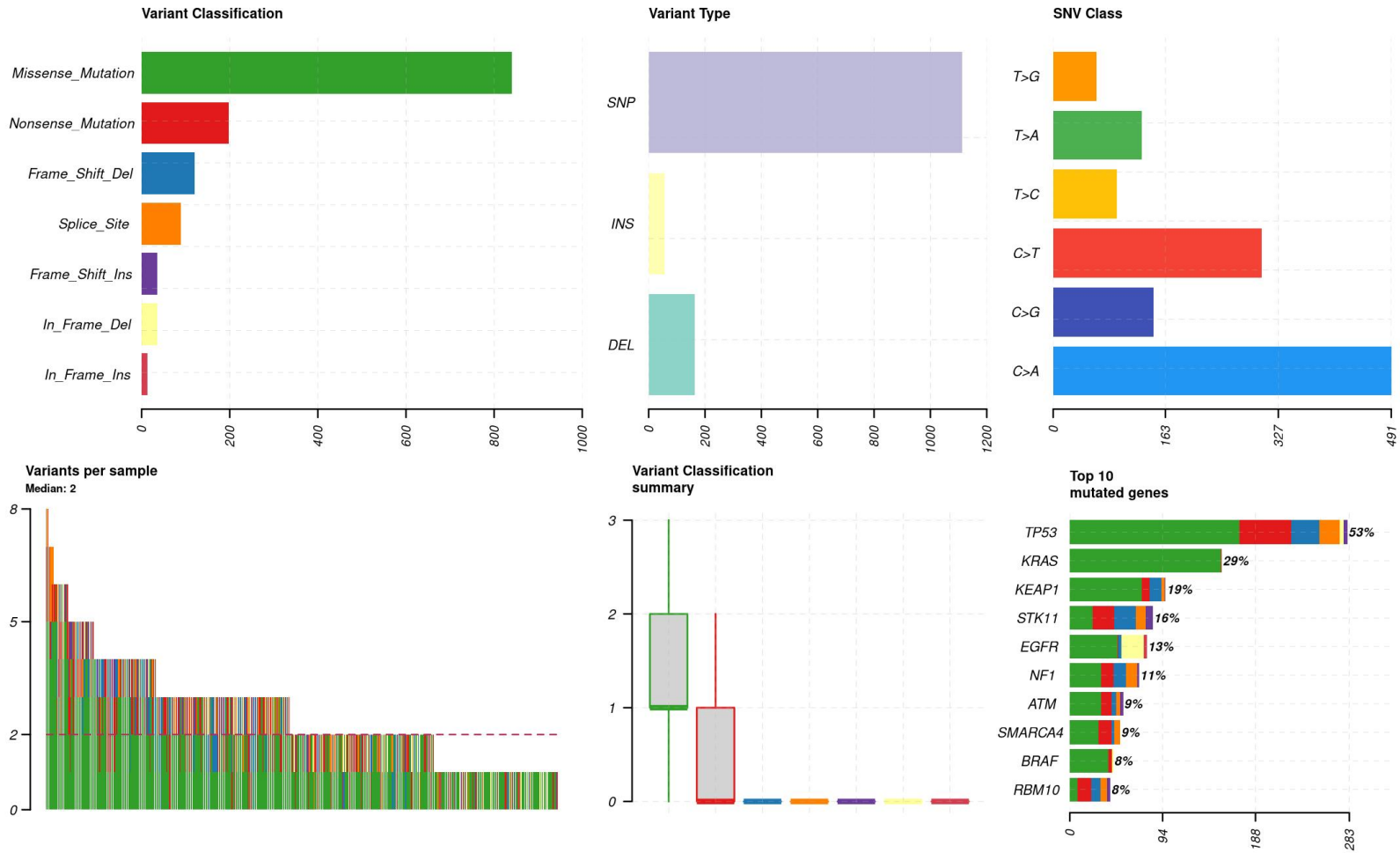


Figure: **Maftools** analysis after the selection of the desired genes.

MAF file III

MAF analysis with TRONCO library

By studying the file we have:

- 522 samples
- 38 genes

We decided to group each type of somatic mutation (such as *missense mutations, nonsense mutations*, etc...) into one type, simply noted as **Mutation**, therefore each gene is related to a single event.

MAF file III

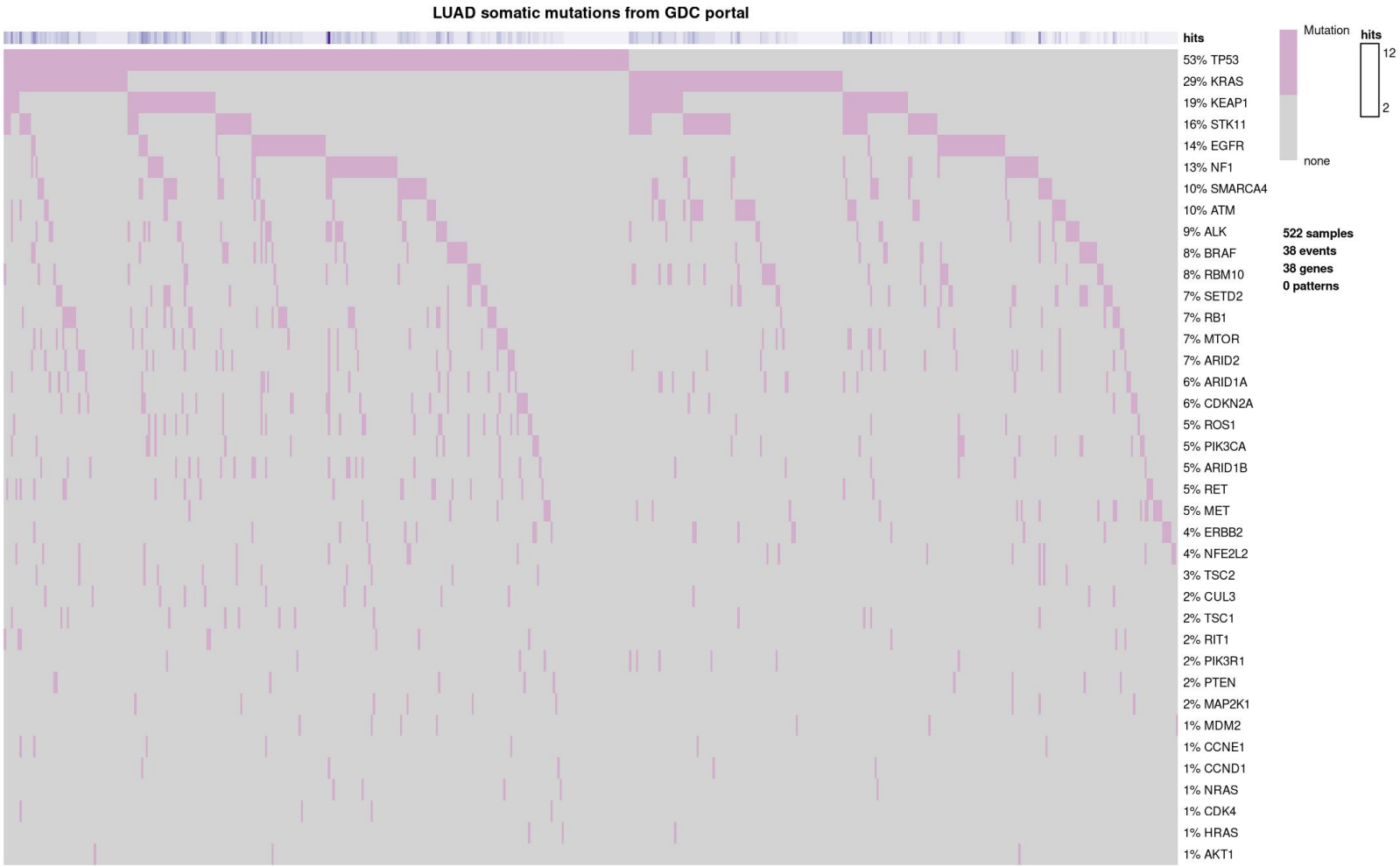


Figure: Oncoprint of somatic mutations from MAF file.

GISTIC File I

Then we download data regarding *CNAs*, annotated in the **GISTIC file**, where we have four mutation type [11], each indicated with a value in the matrix:

1. *Homozygous loss (-2)*
2. *Heterozygous loss (-1)*
3. *Low-level gain (1)*
4. *High-level gain (2)*

There is also 0 to indicate the absence of a mutation (*Diploid*)

	MTOR	ARID1A	RIT1	NFE2L2	SETD2
TCGA-05-4244-01A-01D-1877-01	0	0	1	0	0
TCGA-05-4249-01A-01D-1877-01	-1	-1	2	0	-1
TCGA-05-4250-01A-01D-1877-01	-1	-1	1	1	0
TCGA-05-4382-01A-01D-1204-01	0	0	1	1	-1
TCGA-05-4384-01A-01D-1752-01	0	0	1	0	0
TCGA-05-4389-01A-01D-1204-01	-1	-1	2	1	-1
TCGA-05-4390-01A-02D-1752-01	-1	-1	1	-1	-1

GISTIC File II

We decided to focus only on *high-confidence scores* [11]:

1. *Heterozygous loss (-1)*, as **Deletion**
2. *Low-level gain (1)*, as **Amplification**

GISTIC File II

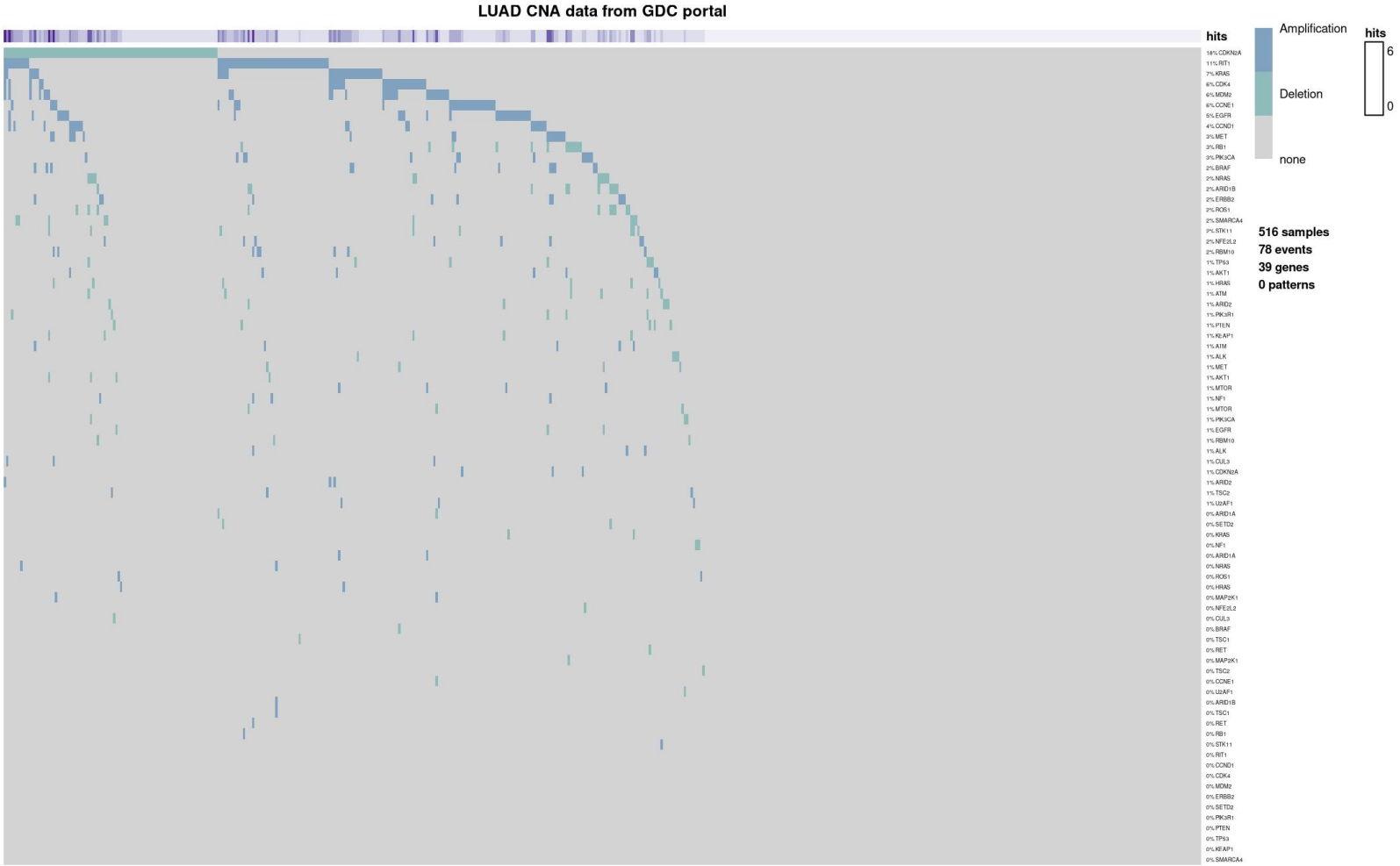


Figure: Oncoprint of selected CNAs from GISTIC file.

Clinical data I

The next step was to download the clinical data, in order to have information, for most of the samples, regarding stages of tumor progression.

Clinical data information

In *TCGA* we have clinical information on each sample such as:

- personal data (such as *age*, *ethnicity* etc...)
- temporal information about the tumor (such as *date of initial diagnosis* etc...)
- oncological data (such as **pathologic stage** etc...)
- information on smoking

Clinical data I

The next step was to download the clinical data, in order to have information, for most of the samples, regarding stages of tumor progression.

	Composite Element REF	years_to_birth	vital_status	days_to_death	days_to_last_followup	tumor_tissue_site	pathologic_stage
tcga.05.4245	value	81	0	*NA*	730	lung	stage iiia
tcga.05.4382	value	68	0	*NA*	607	lung	stage ib
tcga.05.4384	value	66	0	*NA*	426	lung	stage iiia
tcga.05.4396	value	76	1	303	*NA*	lung	stage iiib
tcga.05.4402	value	57	1	244	*NA*	lung	stage iv
tcga.05.4405	value	74	0	*NA*	610	lung	stage ib

Clinical data II

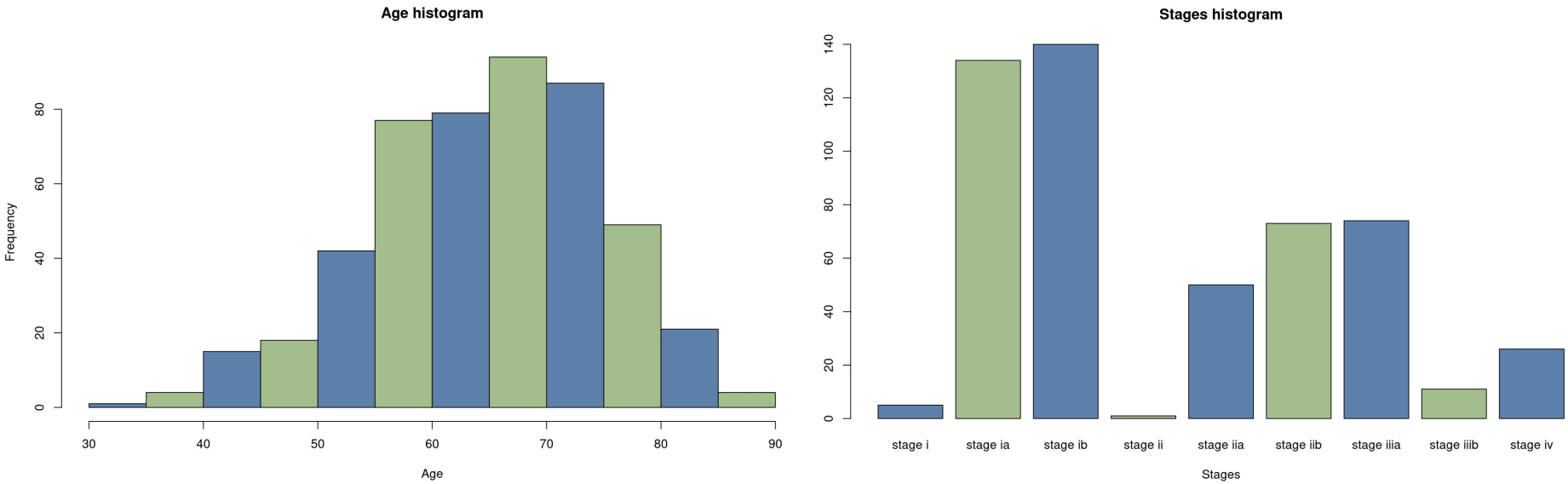


Figure: Distribution of ages and tumor stages on samples.

Clinical data III

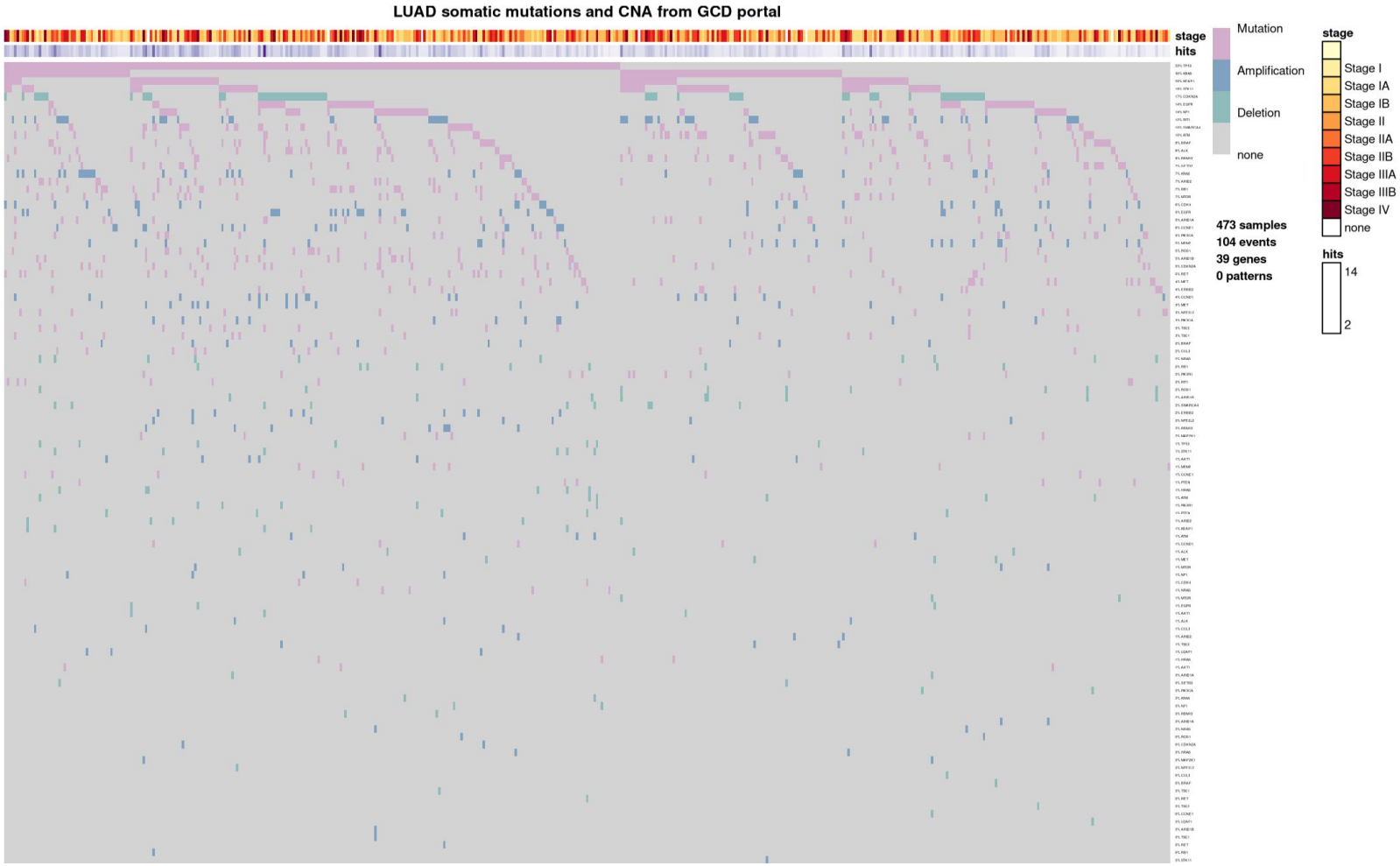


Figure: Oncoprint of both somatic mutations and CNAs with tumor stages annotated.

Smoke

Smoking is one of the main causes of *LUAD*

Speaking of lung adenocarcinoma, it is interesting to talk about smoking. The clinical data records, for some patients, the number of packs smoked per year, as well as the year they started smoking.

Smoke

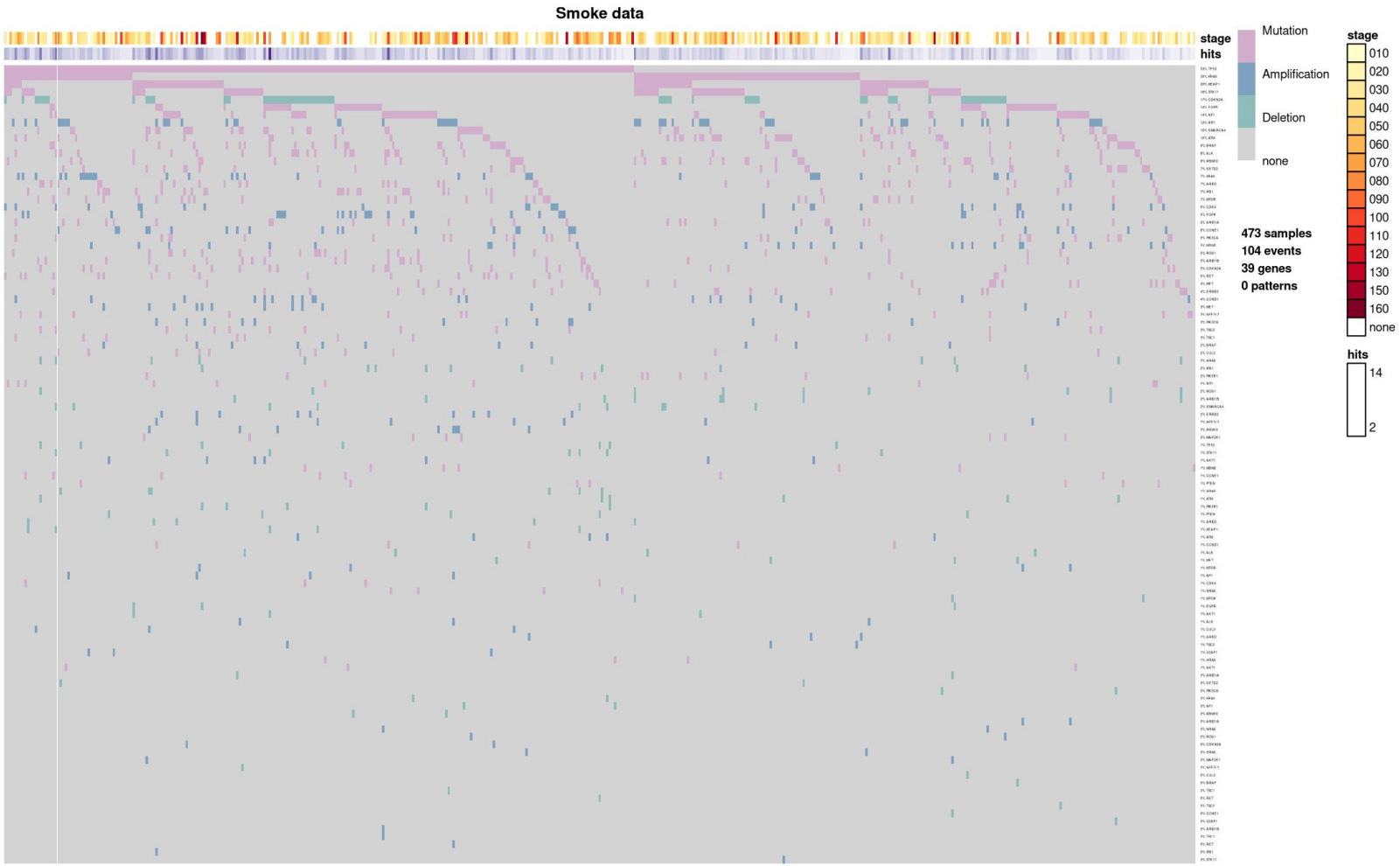


Figure: A little "hack" to use *oncoprint* function to plot the annotation of smoker in clinical data, using the number of packs smoked per year, grouped by tens, as "stages".

Final TRONCO dataset for LUAD

As final preprocessing of the data we filter the events on the basis of a minimum frequency: 3% [12].

Final TRONCO dataset for LUAD

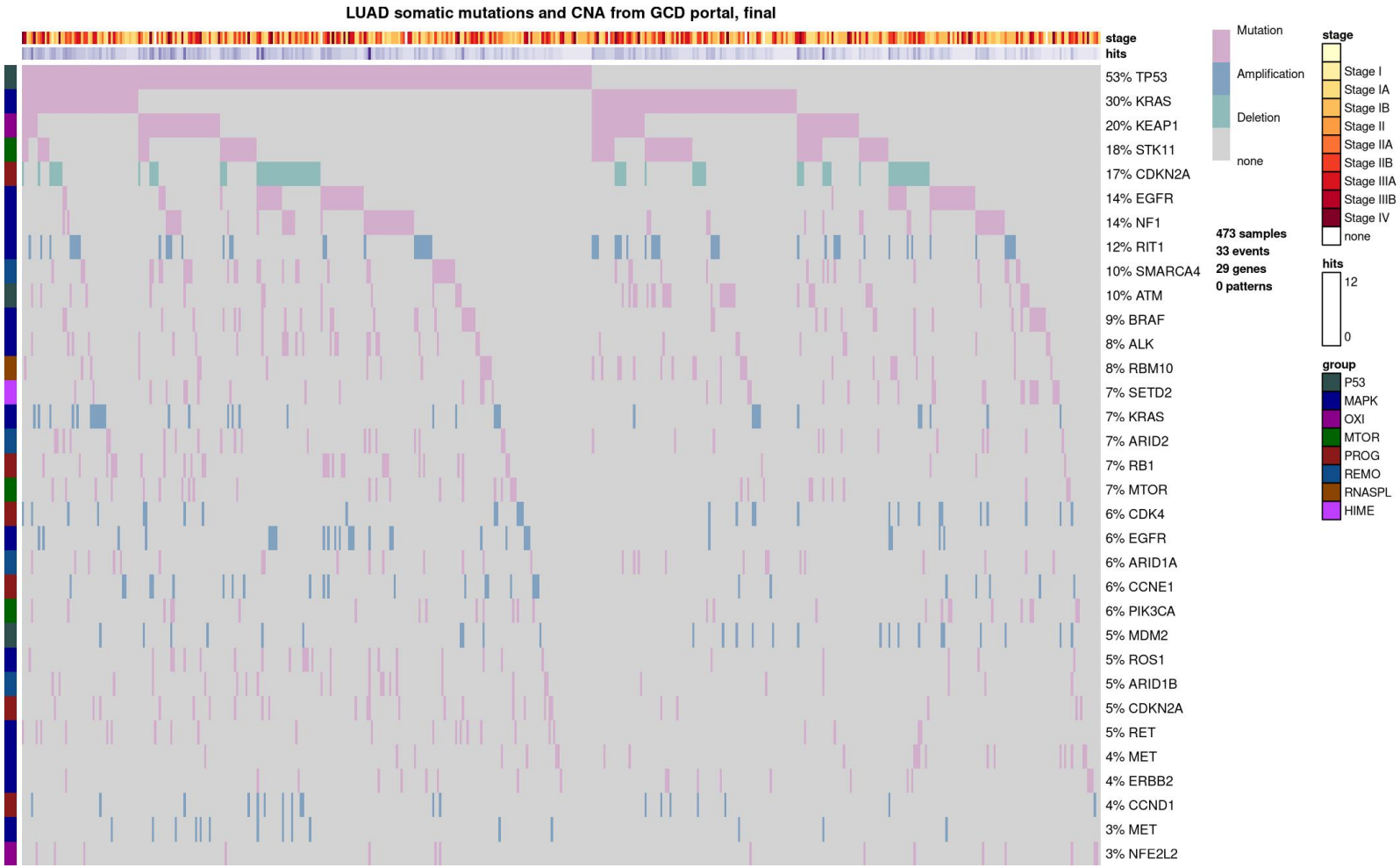


Figure: Oncoprint of complete dataset, after frequency selection, with stages and pathways annotated.

Outline

Subtyping I

Marker paper subtyping

"*To coordinate naming of the transcriptional subtypes with the histopathological, anatomic and mutational classifications of lung adenocarcinoma, we propose an updated nomenclature: the* ***terminal respiratory unit*** *(TRU, formerly bronchioid), the* ***proximal-inflammatory*** *(PI, formerly squamoid), and the* ***proximal-proliferative*** *(PP, formerly magnoid) transcriptional subtypes*" [6].

Subtyping I

Marker paper subtyping

"*To coordinate naming of the transcriptional subtypes with the histopathological, anatomic and mutational classifications of lung adenocarcinoma, we propose an updated nomenclature: the* ***terminal respiratory unit*** *(TRU, formerly bronchioid), the* ***proximal-inflammatory*** *(PI, formerly squamoid), and the* ***proximal-proliferative*** *(PP, formerly magnoid) transcriptional subtypes*" [6].

Given these premises, we decided to study the three subtypes listed in the marker paper.
Furthermore, for the sake of completeness, we have also decided to study the complete dataset without division into subtypes.

Subtyping II

In order to associate the respective subtype to a certain sample we again relied on TCGA, again through *TCGAbiolinks*.

	patient	Sex	Age.at.diagnosis	expression_subtype	T.stage	N.stage
1	TCGA-05-4249	MALE	67	TRU	T2	N0
2	TCGA-05-4382	MALE	68	prox.-inflam	T2	N0
3	TCGA-05-4384	MALE	66	TRU	T2	N2
4	TCGA-05-4389	MALE	70	prox.-prolif.	T1	N0
5	TCGA-05-4390	FEMALE	58	prox.-prolif.	T2	N0
6	TCGA-05-4395	MALE	76	prox.-inflam	T4	N2
7	TCGA-05-4396	MALE	76	prox.-prolif.	T4	N1

Subtyping III

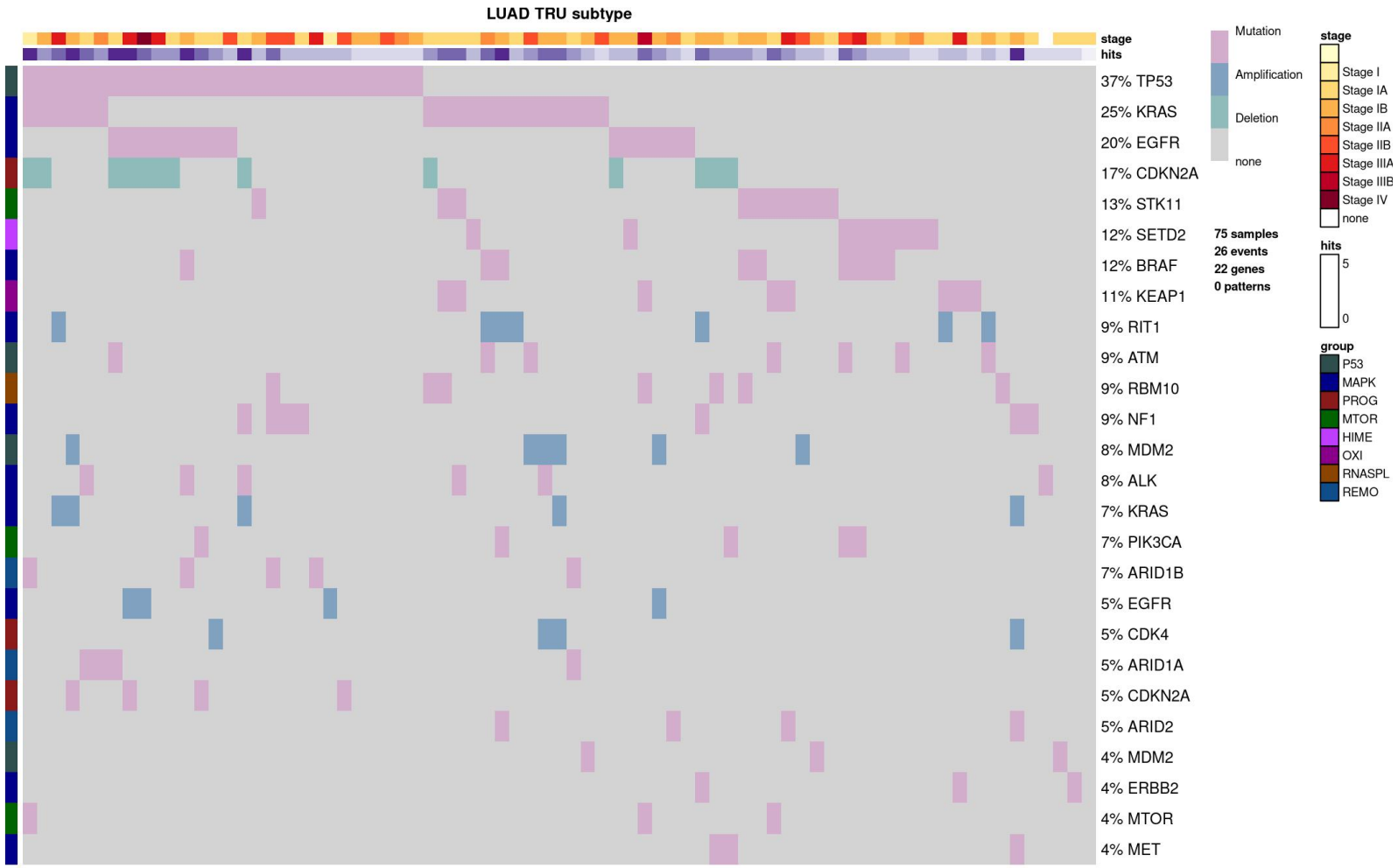


Figure: Oncoprint of *TRU* subtype, with stages and pathways annotated.

Subtyping III

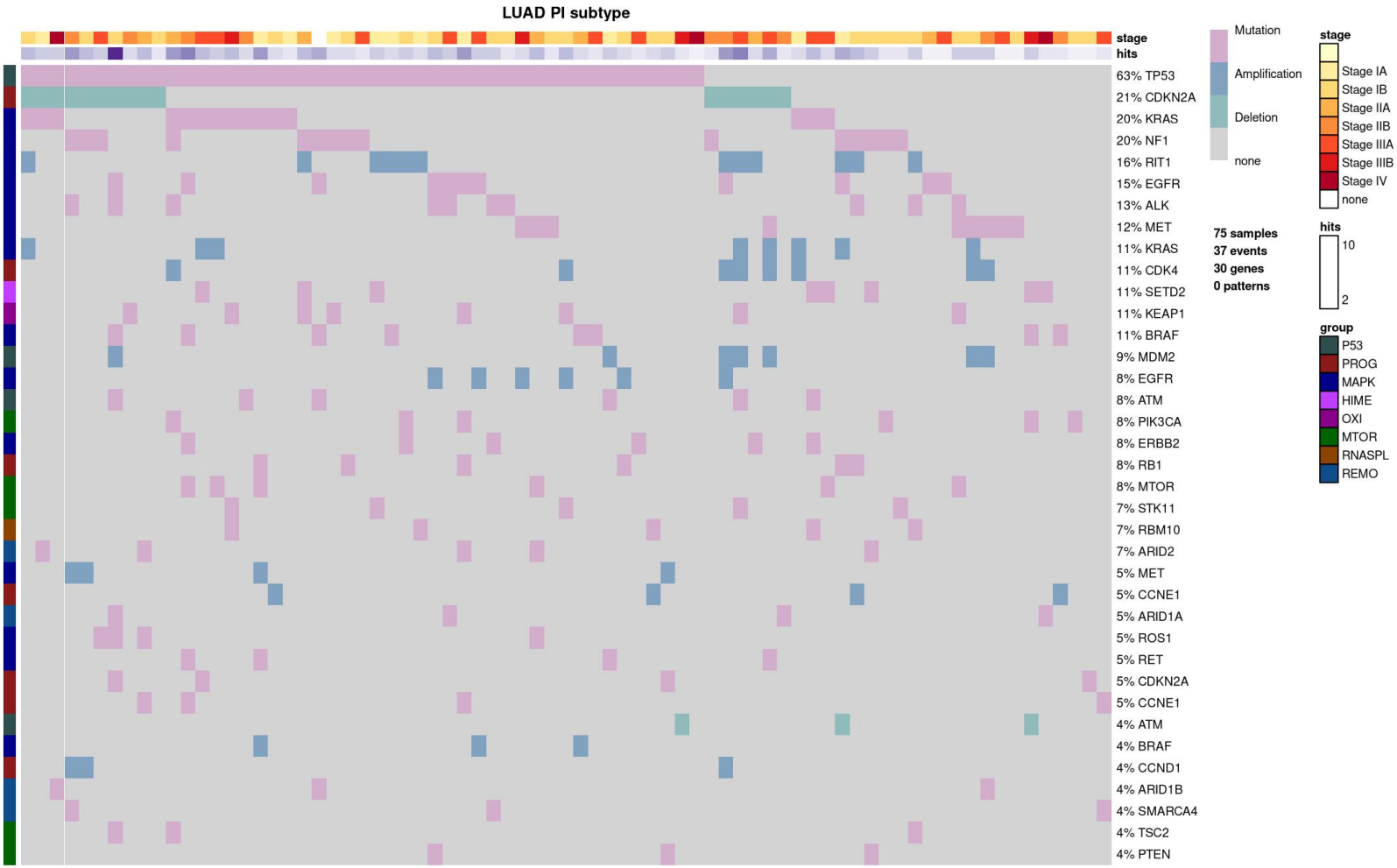


Figure: Oncoprint of *PI* subtype, with stages and pathways annotated.

Subtyping III

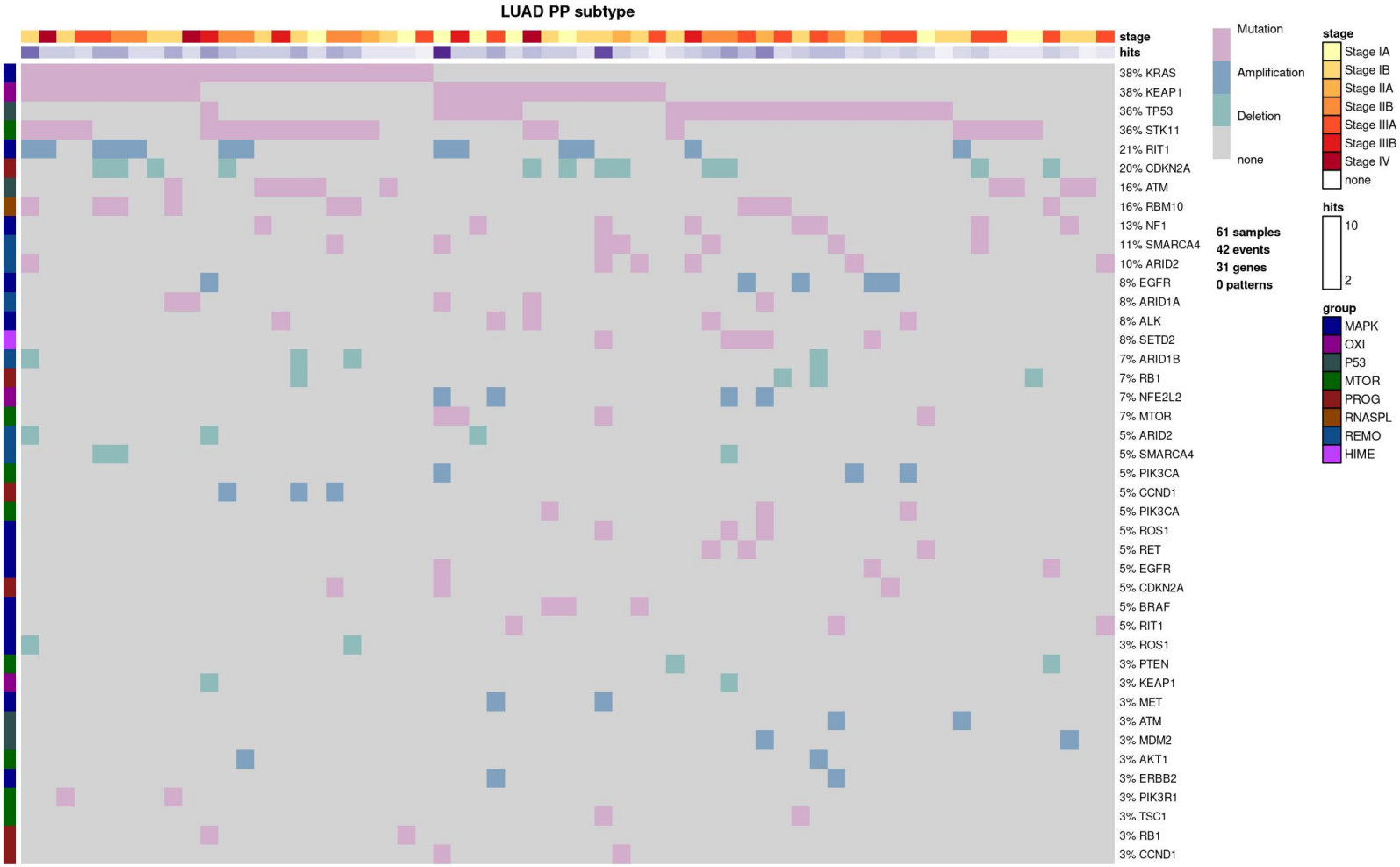


Figure: Oncoprint of *PP* subtype, with stages and pathways annotated.

Outline

Mutex I

As recommended by the authors of the PiCnIc pipeline, one of the main ways for the study of mutually exclusive groups is the use of a tool called **Mutex** [12].

In any case, since the execution of this tool was very expensive, we used the results for LUAD already present in "additional files" of the paper.

The file with all the results is imported through the *TRONCO* utilities and then used to create hypotheses of mutually exclusive groups, making hypotheses with **OR**.

Mutex I

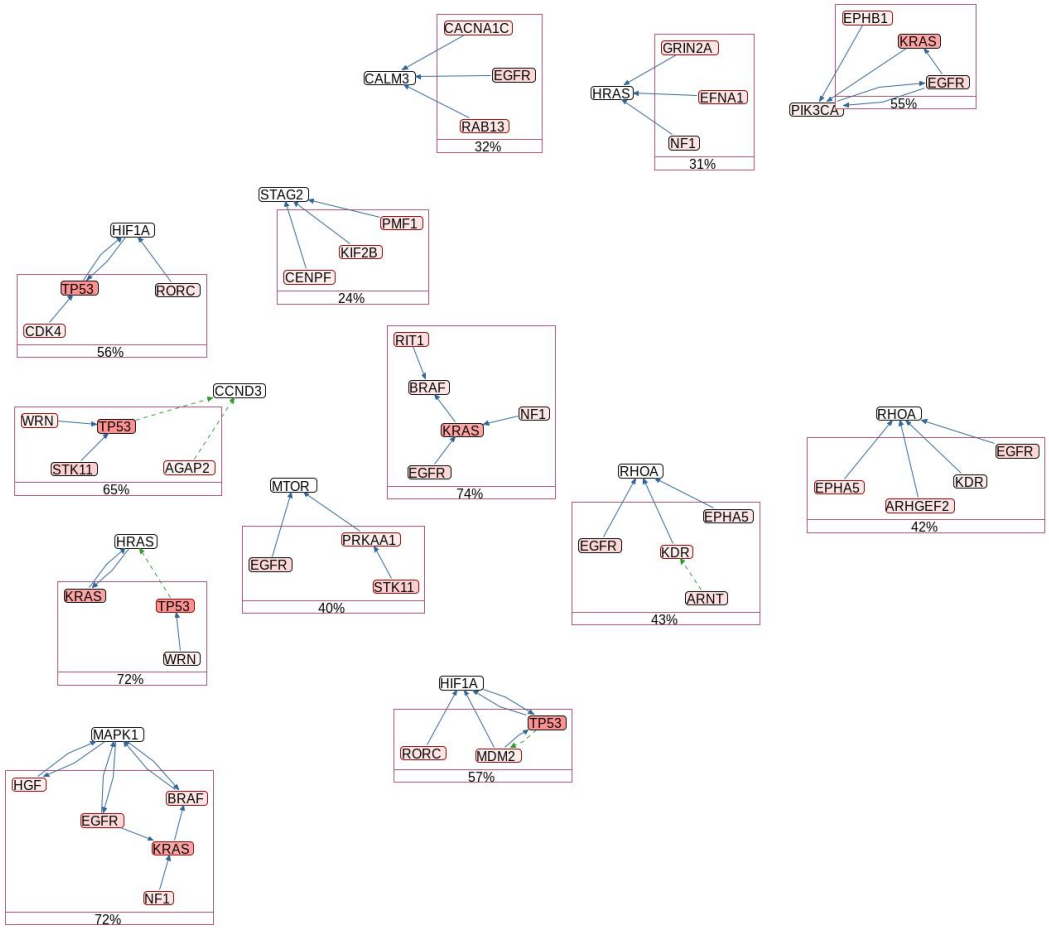


Figure: Visualization of the exclusivity groups using **Chisio BioPAX Editor (*ChiBE*)** [13] of the *Mutex* results for *LUAD*.

Mutex II

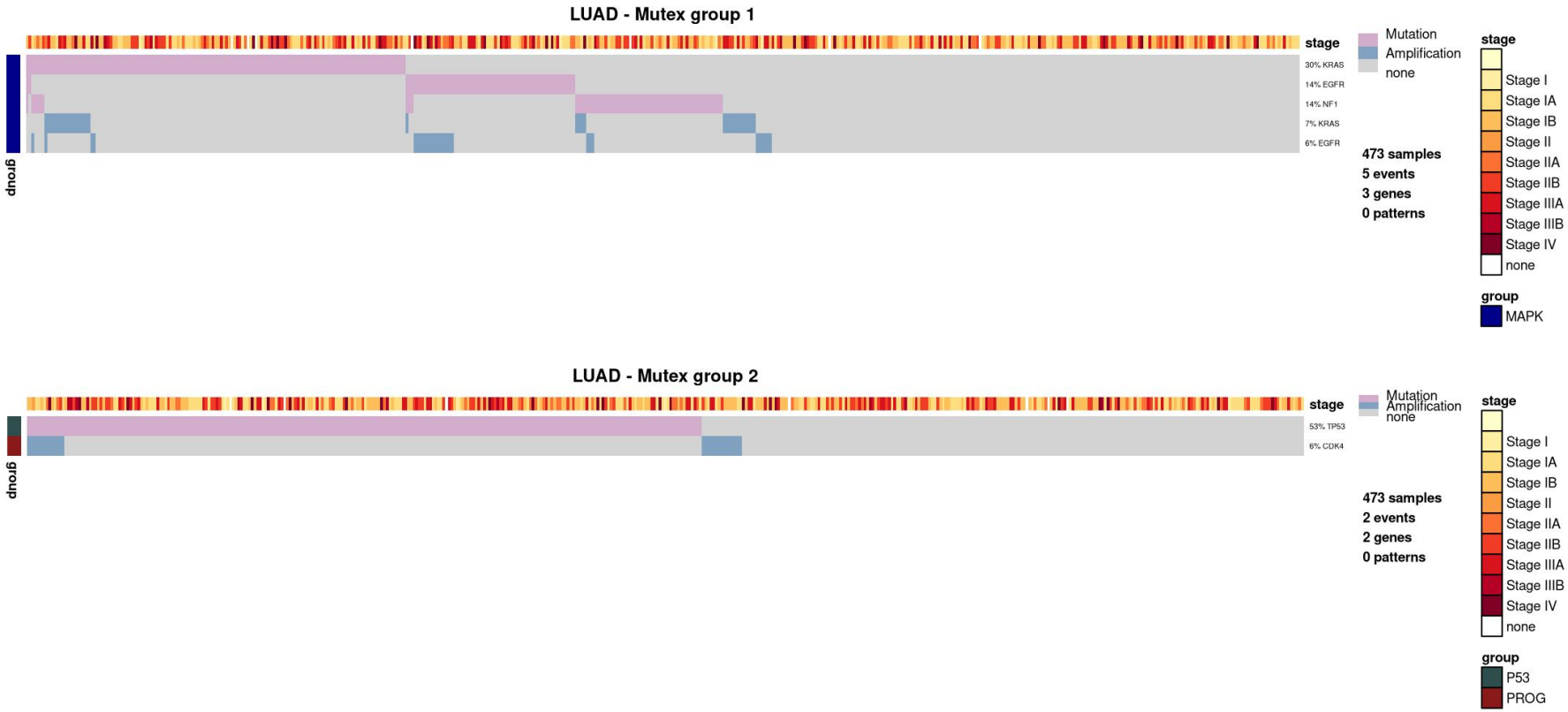
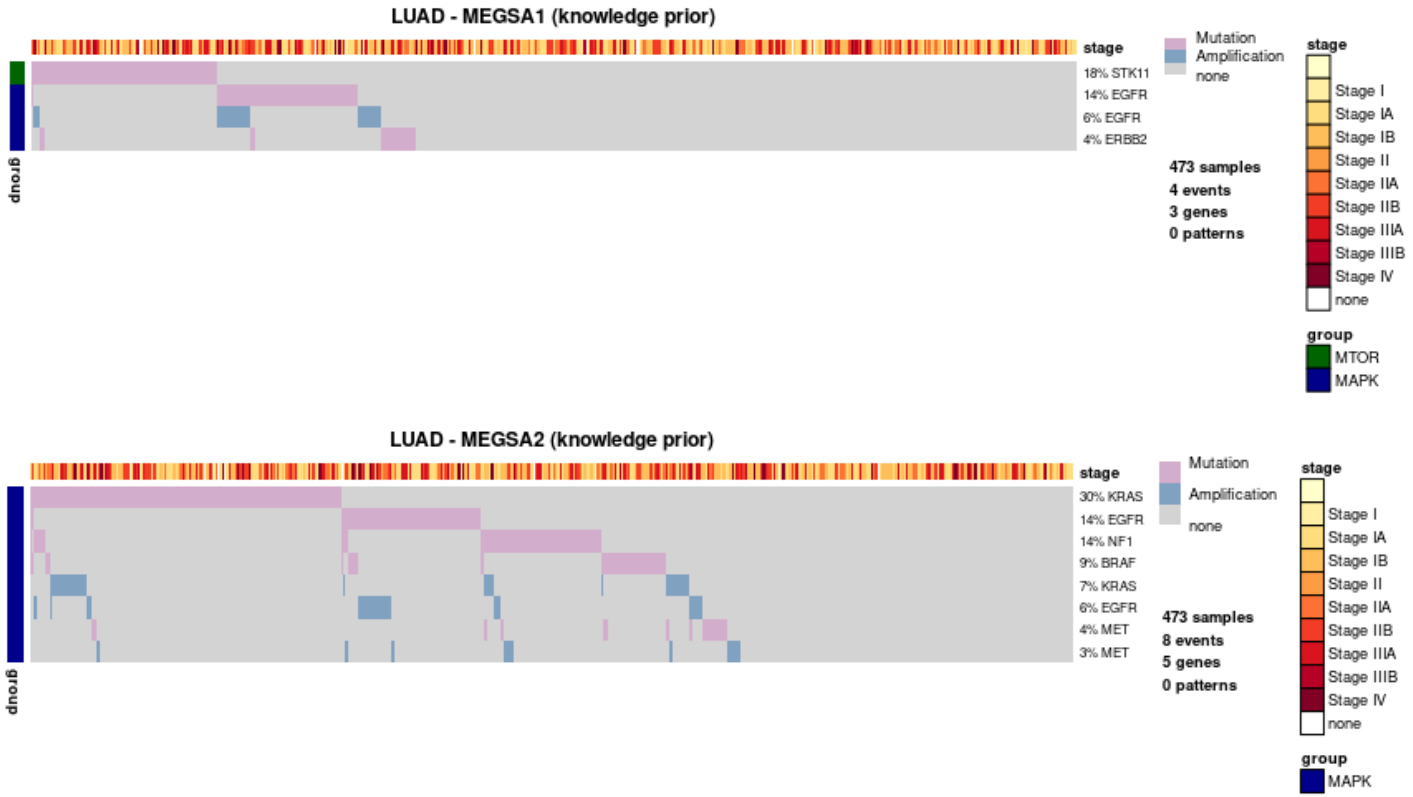


Figure: Oncoprint of the first two groups in *Mutex* results for *LUAD*.

MEGSA

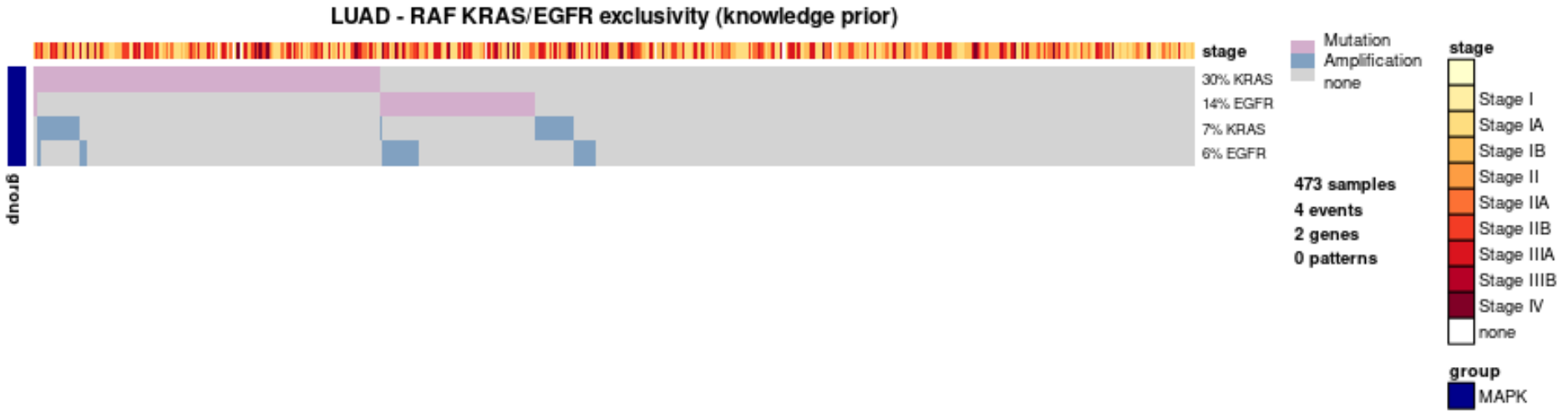
In order to have other hypotheses we did a search in the literature, finding a paper [14] on *LUAD* analysis where a tool called **MEGSA** [15] was used. We therefore decided to include the two annotated hypotheses, as **OR** hypotheses.



Marker Paper Hypotheses I

Marker paper Hypothesis

Finally we included a hypothesis from the marker paper. Infact we have: "*Mutations in KRAS (33%) were mutually exclusive with those in EGFR (14%).*" so we add this as a **XOR** hypotheses.



Marker Paper Hypotheses II

Subtyping Hypotheses

For the subtypes we some extra hypotheses based on the marker paper:

- for *PP* subtype we can find in the paper: "*The PP subtype was enriched for mutation of KRAS, along with inactivation of the STK11 tumour suppressor gene by chromosomal loss, inactivating mutation, and reduced gene expression*" so we add this as an **AND** hypothesis
- for *PI* subtype we can find in the paper: "*the PI subtype was characterized by solid histopathology and co-mutation of NF1 and TP53*" so we add this as an **AND** hypothesis

Marker Paper Hypotheses II

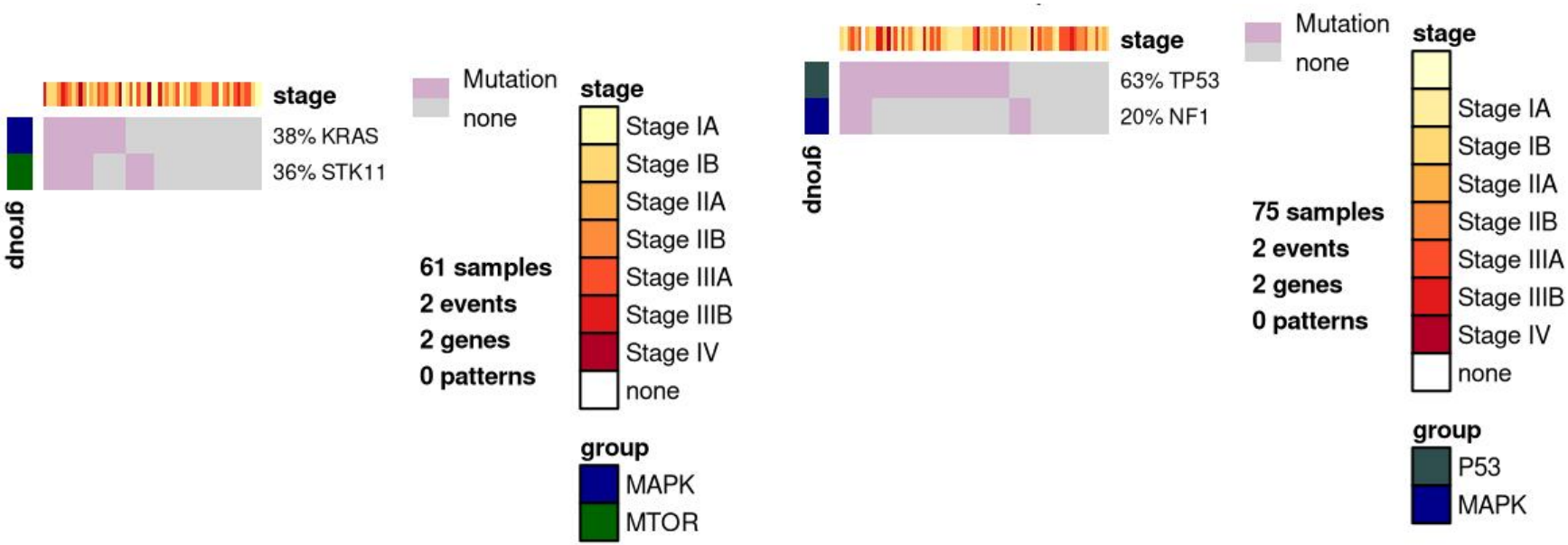


Figure: Oncoprint of the hypotheses for subtypes from the marker paper.

Outline

Models Preparation

Before proceeding with the reconstruction of the cancer progression model with the **CAPRI algorithm** we cleaned up the data:

- we only kept the genes drivers that are present in the hypotheses
- we collapse multiple events per gene in one unique event
- we *consolidate* the *TRONCO* object, deleting indistinguishable events, having alterations which have the same signature across the samples, "programmatically"

Models Preparation

Before proceeding with the reconstruction of the cancer progression model with the **CAPRI algorithm** we cleaned up the data:

- we only kept the genes drivers that are present in the hypotheses
- we collapse multiple events per gene in one unique event
- we *consolidate* the *TRONCO* object, deleting indistinguishable events, having alterations which have the same signature across the samples, "programmatically"

Then we added all the hypotheses using *TRONCO* methods, iterating over the *Mutex object* and adding all the hardcoded hypotheses stated before. Finally we added automatically all the hypotheses related to homologous events, using the homonymous function.

CAPRI I

We executed the **CAPRI algorithm** [16] almost with default settings:

- **regularizers** for the likelihood estimation [1]:
  - **Akaike Information Criterion (*AIC*)**, more prone to overfitting but also likely to provide good predictions from data and is better when false negatives are more misleading than positive one
  - **Bayesian Information Criterion (*BIC*)**, more prone to underfitting errors, thus more parsimonious and better in the opposite direction.
- **p-value $<$ 0.05** to accept the valid selective advantage relations
- **heuristic search** is performed with *Hill Climbing*
- **seed** for bootstrap iterations has been set to 42 - *the Answer to the Ultimate Question of Life, the Universe, and Everything*

In addition the **number of boostrap iterations** is choosen at the begin of the pipeline and for this project we use 100 iterations.
The details on the labels of the models will be explored further on.

CAPRI II

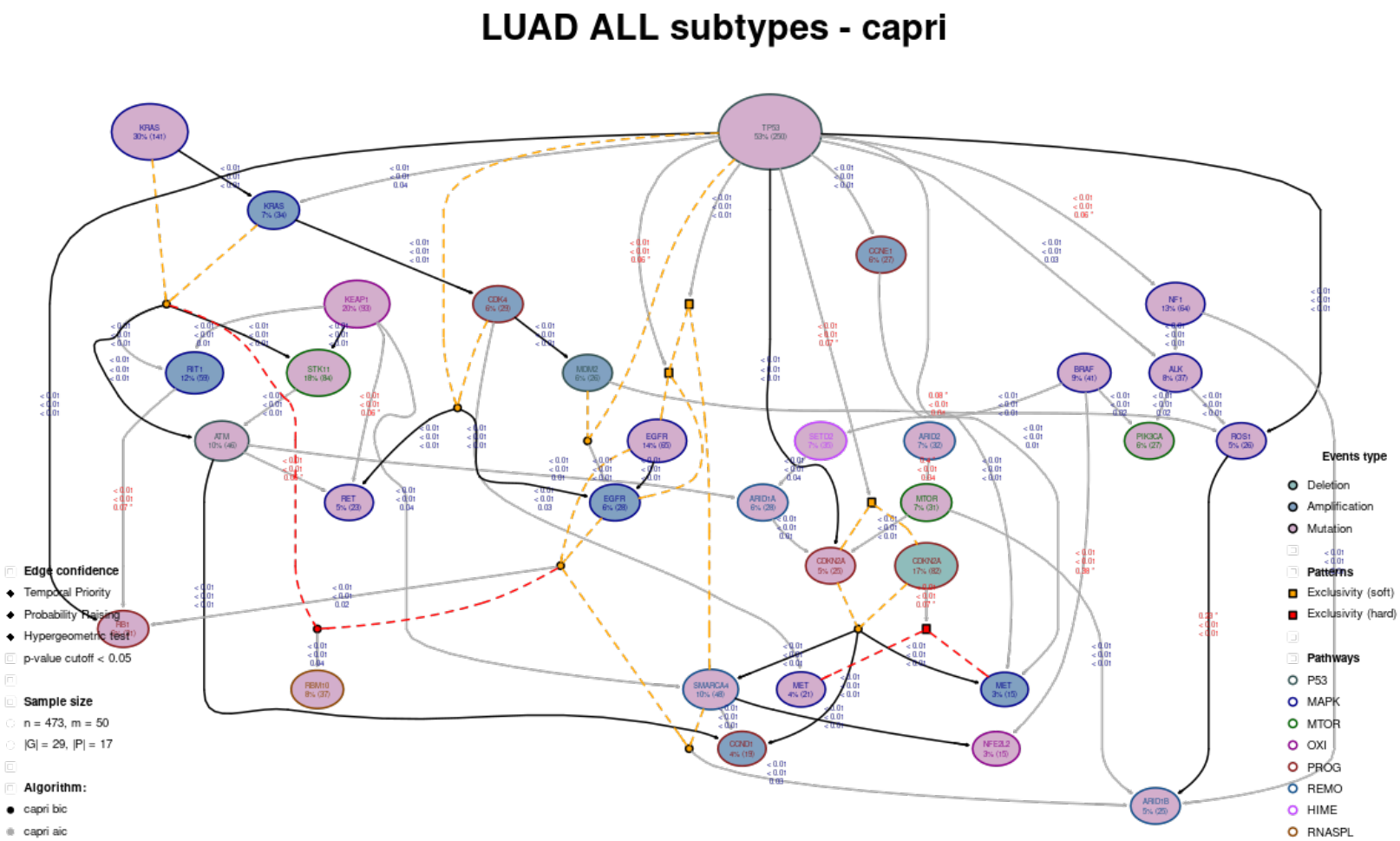


Figure: *TRONCO.plot* of the reconstructed model from the data including all subtypes.

CAPRI II

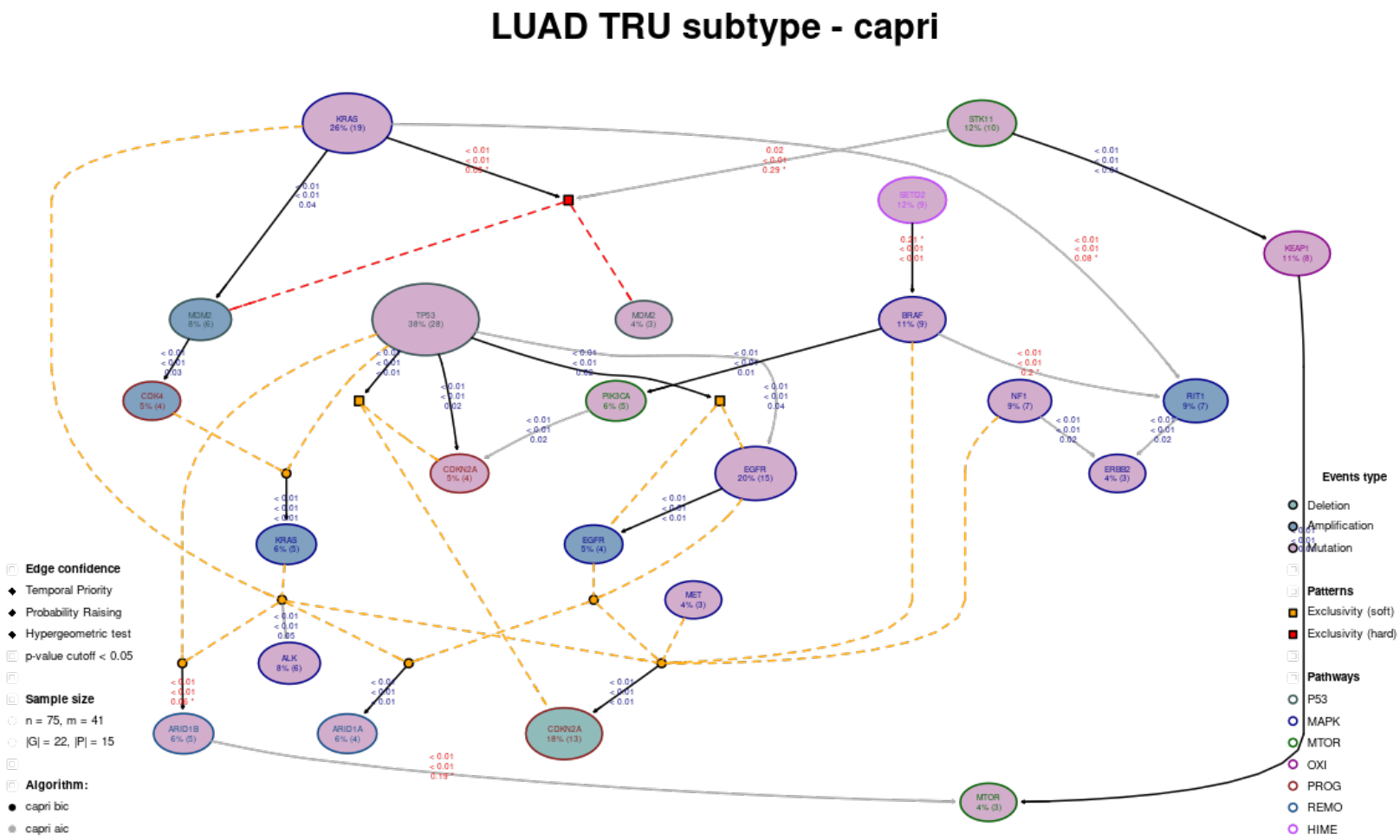


Figure: *TRONCO.plot* of the reconstructed model from the data of TRU subtype.

CAPRI II

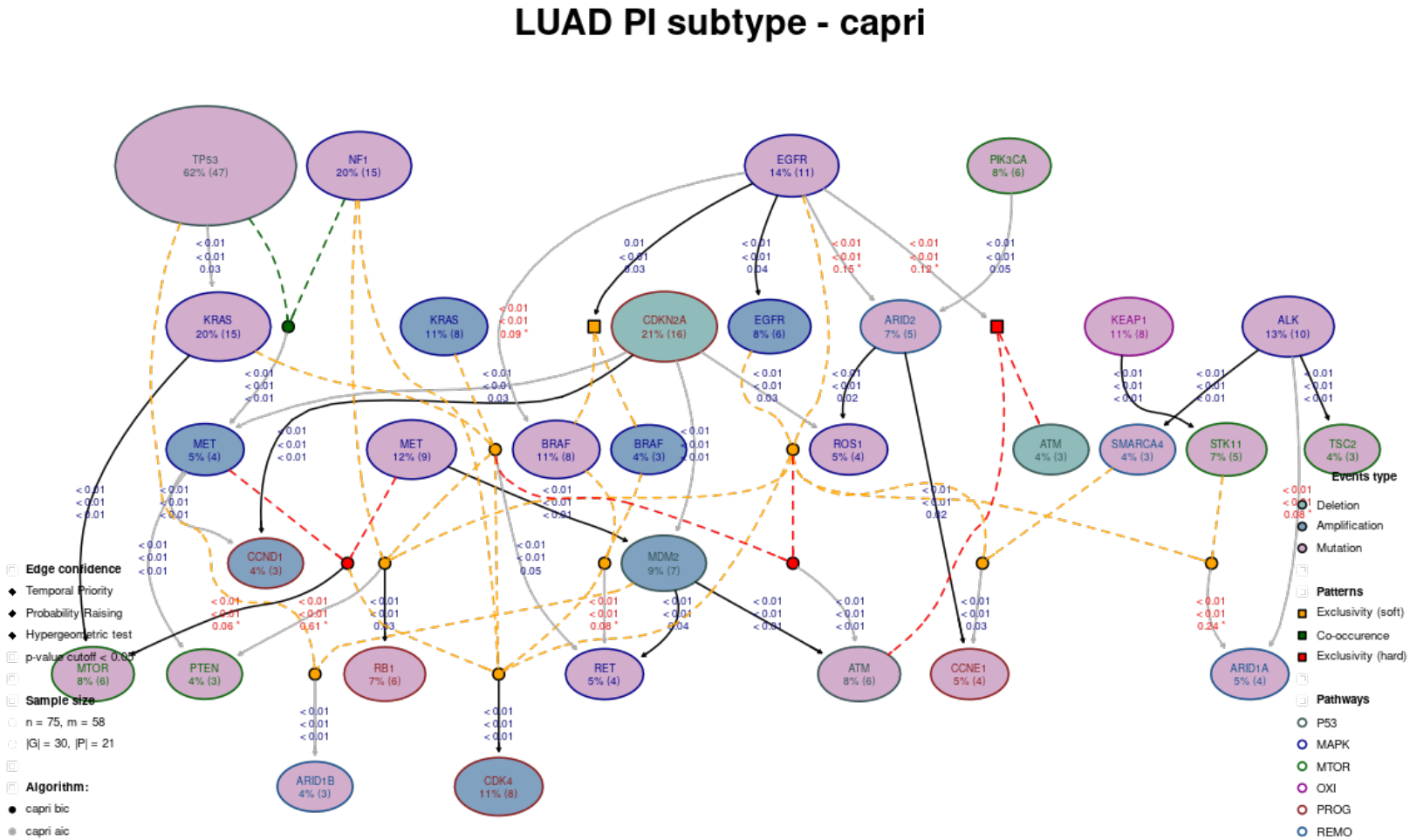


Figure: *TRONCO.plot* of the reconstructed model from the data of PI subtype.

CAPRI II

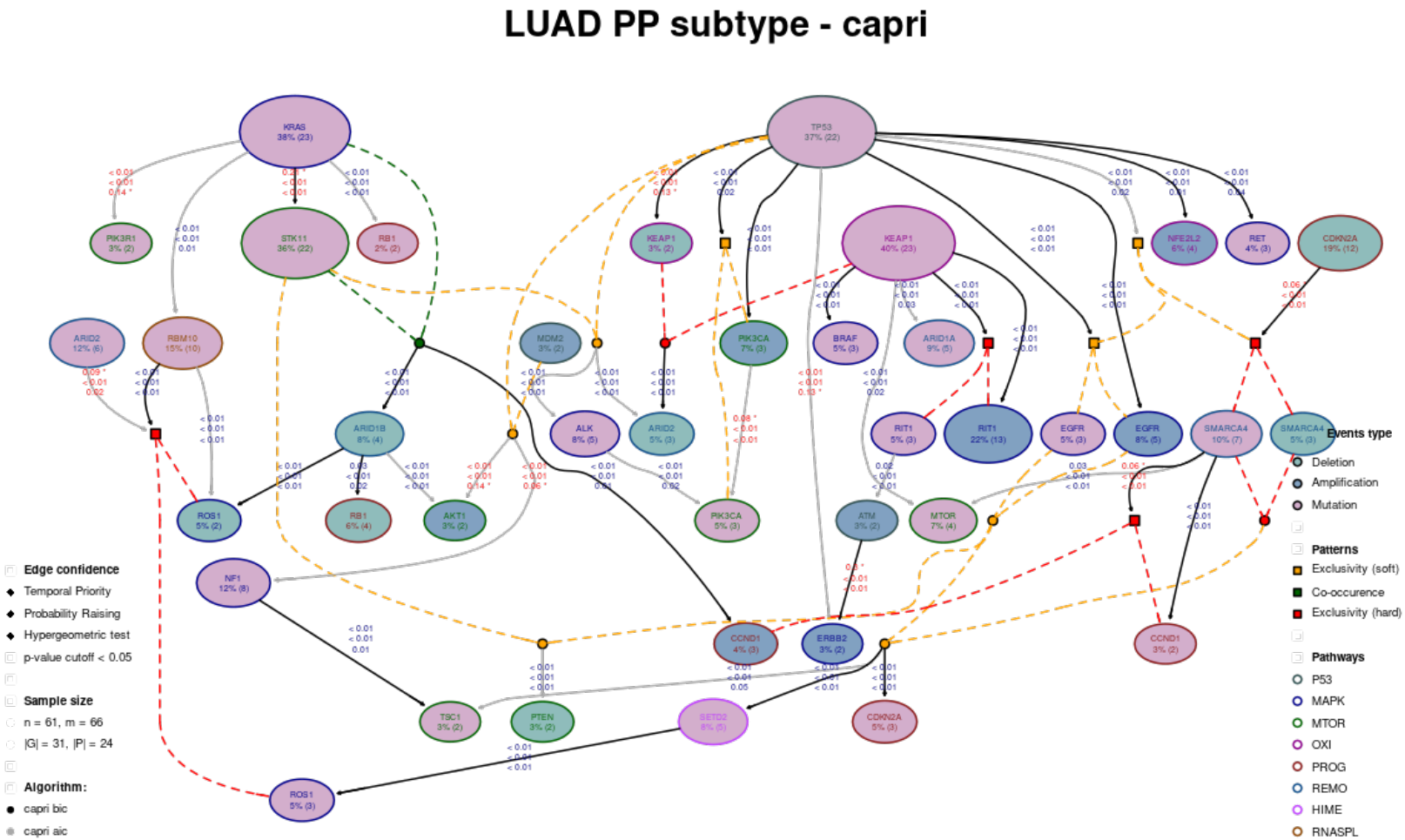
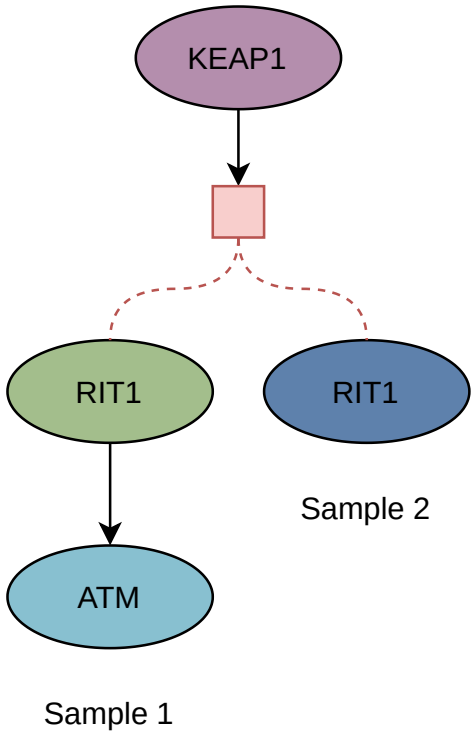


Figure: *TRONCO.plot* of the reconstructed model from the data of PP subtype

Evolutionary Trajectories I

Branching trajectory

In this case we have that from one mutation we could have two different and independent mutation in different samples (and eventually other mutations to follow).
In the model we have, for example in PP subtype model, something like this:



Evolutionary Trajectories I

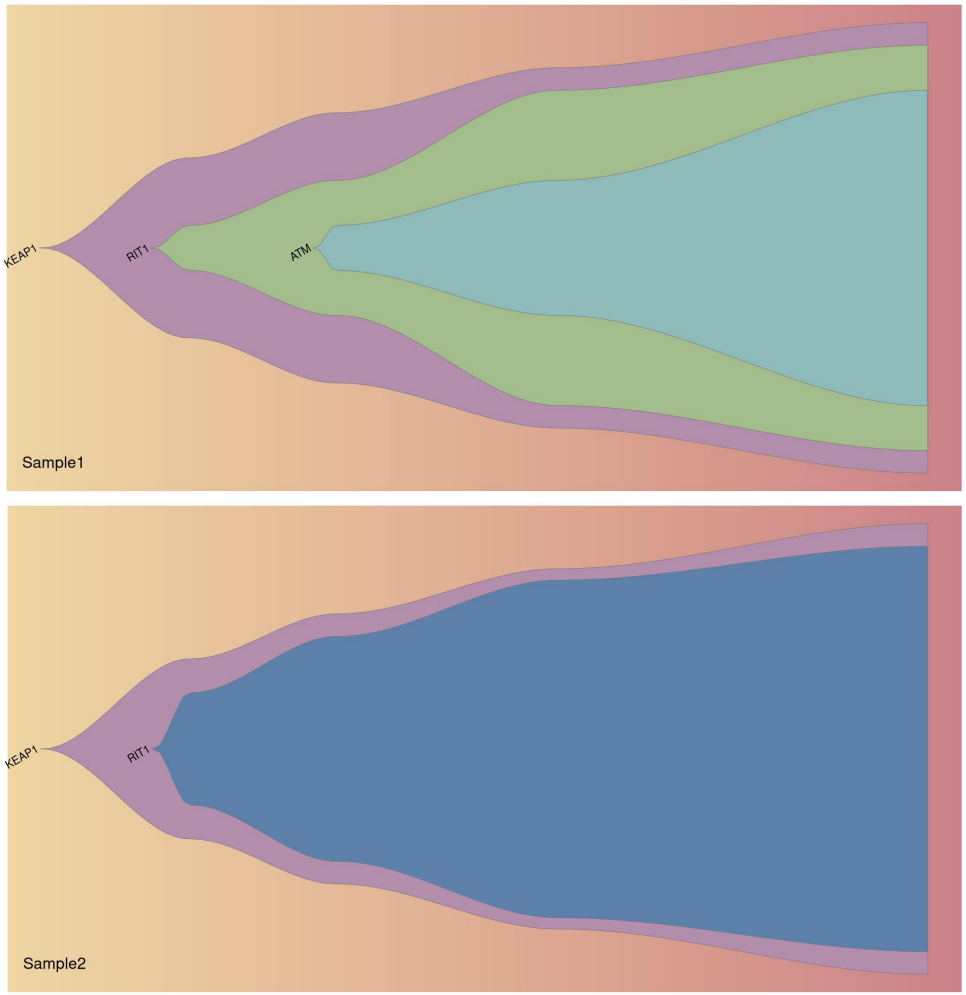
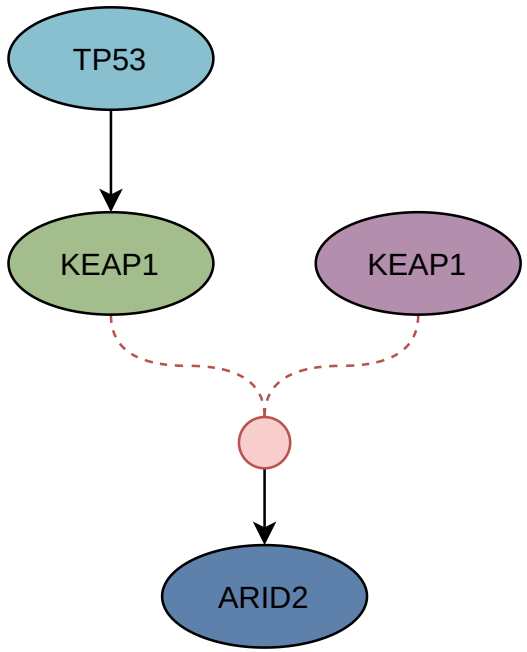


Figure: Manually curated branching trajectories from PP subtype, plotted using *Fishplot library* [17].

Evolutionary Trajectories II

Branching trajectory

In this case we have that from two different mutation paths we converge to the same mutation.
In the model we have, for example in PP subtype model, something like this:



Evolutionary Trajectories II

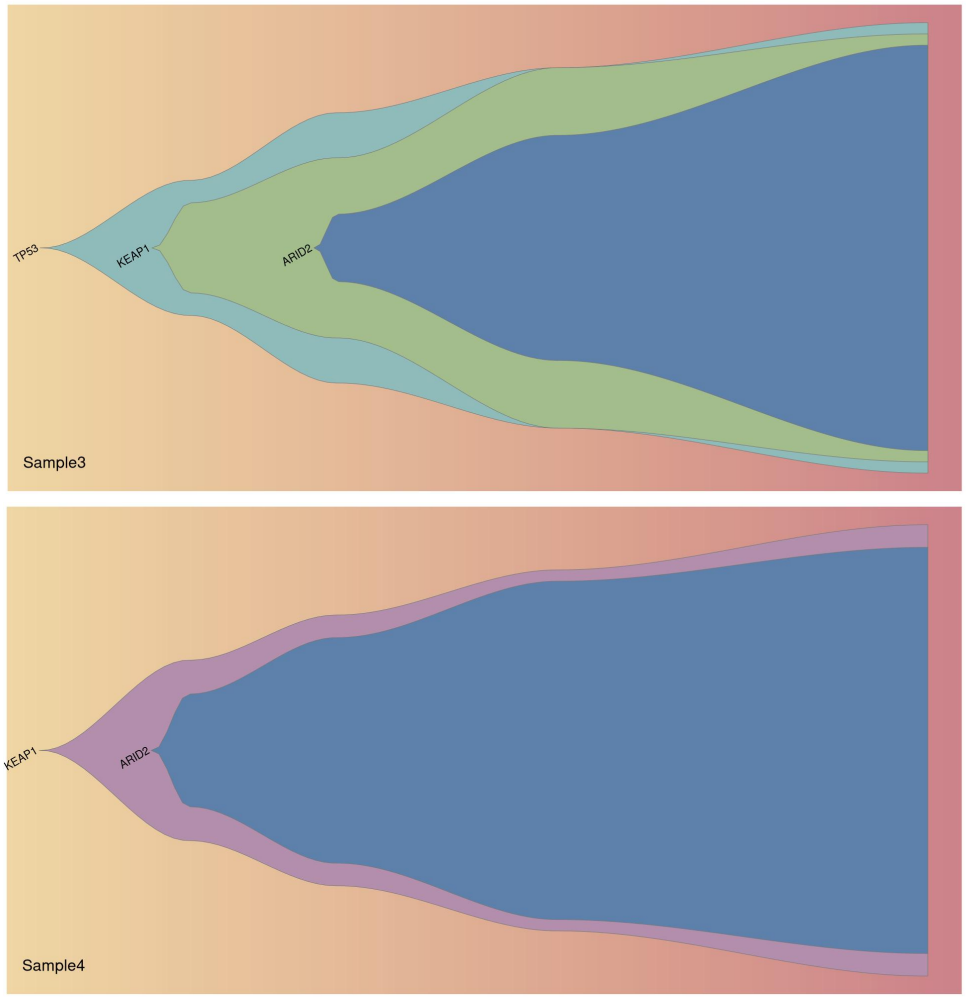


Figure: Manually confluencing trajectories from PP subtype, plotted using *Fishplot library* [17].

Outline

Post-Reconstruction I

We calculated some scores:

- **non-parametric boostrap scores** resampling the dataset, re-running *CAPRI algorithm* and computing the scores based on how many times we re-infer the same edge
- **statistical bootstrap scores**, re-running the statistical test for temporal priority and probability raising by initializing a random number generator with different seeds

Post-Reconstruction I

We calculated some scores:

- **non-parametric boostrap scores** resampling the dataset, re-running *CAPRI algorithm* and computing the scores based on how many times we re-infer the same edge
- **statistical bootstrap scores**, re-running the statistical test for temporal priority and probability raising by initializing a random number generator with different seeds

As we'll see later, the scores obtained in our analysis seem to validate the goodness of our models.

Post-Reconstruction II

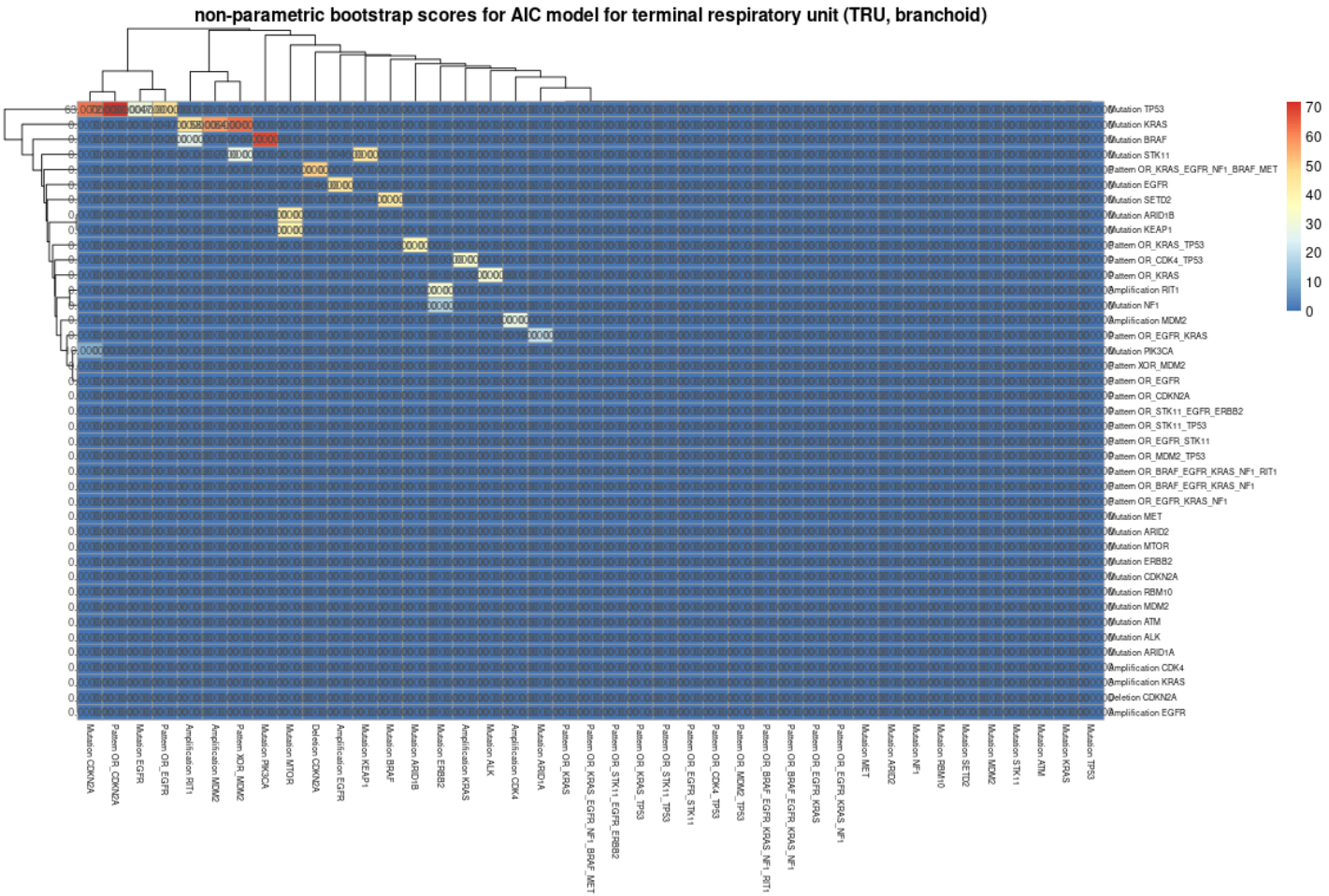


Figure: Example of Non Parametric and Statistical results shown as heatmap for AIC bootstrap scores for TRU subtype model.

Post-Reconstruction III

Edge labels

Given an event c and an event e, at time t_c and t_e, with $0 < Pr(c), Pr(e) < 1$:

- **temporal priority**, *p-value* of temporal priority, $t_c < t_e$, via *Mann-Whitney U testing*,
- **probability raising**, *p-value* of probability raising, $Pr(e|\neg c) < Pr(e|c)$, via *Mann-Whitney U testing*
- **hypergeometric test**, is used to underline if there is a difference between samples containing $c \wedge e$ versus the total population of samples containing $c \vee e$. We would like the overlap (joint probability of c and e) to be significant as this supports the presence of a selection trend among c and e.

An edge is fully supported if all those three p-values are below a significance threshold set at 0.05. Some edges might have p-value for the *temporal priority* above the threshold, meaning there's still a significant selection trend but with a direction (i.e. the temporal order of c and e) not supported by the data [1].

Post-Reconstruction III

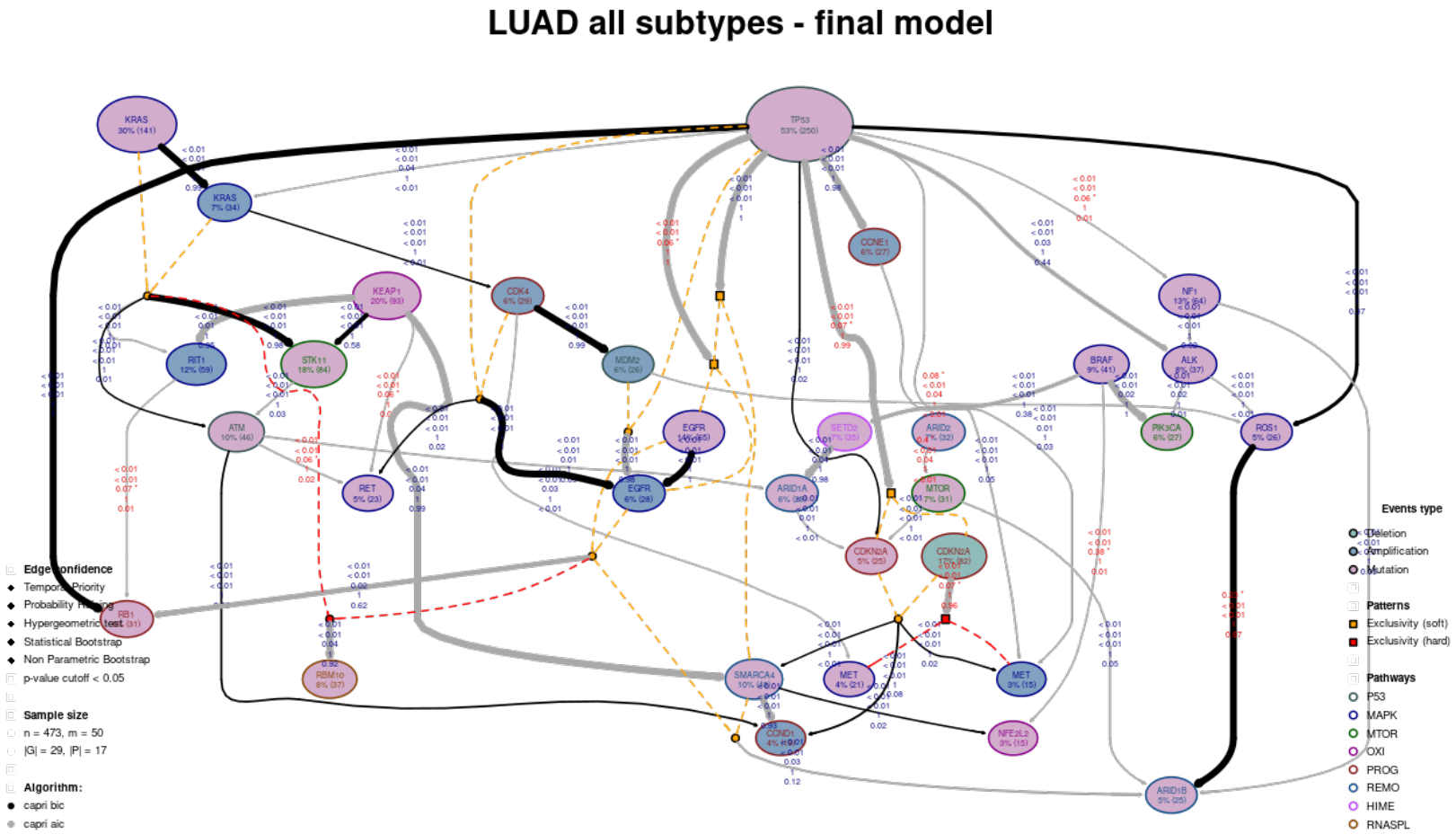


Figure: *TRONCO.plot* with bootstrap scores of the reconstructed model from the data including all subtypes. Edge size represents non-parametric bootstrap scores.

Post-Reconstruction III

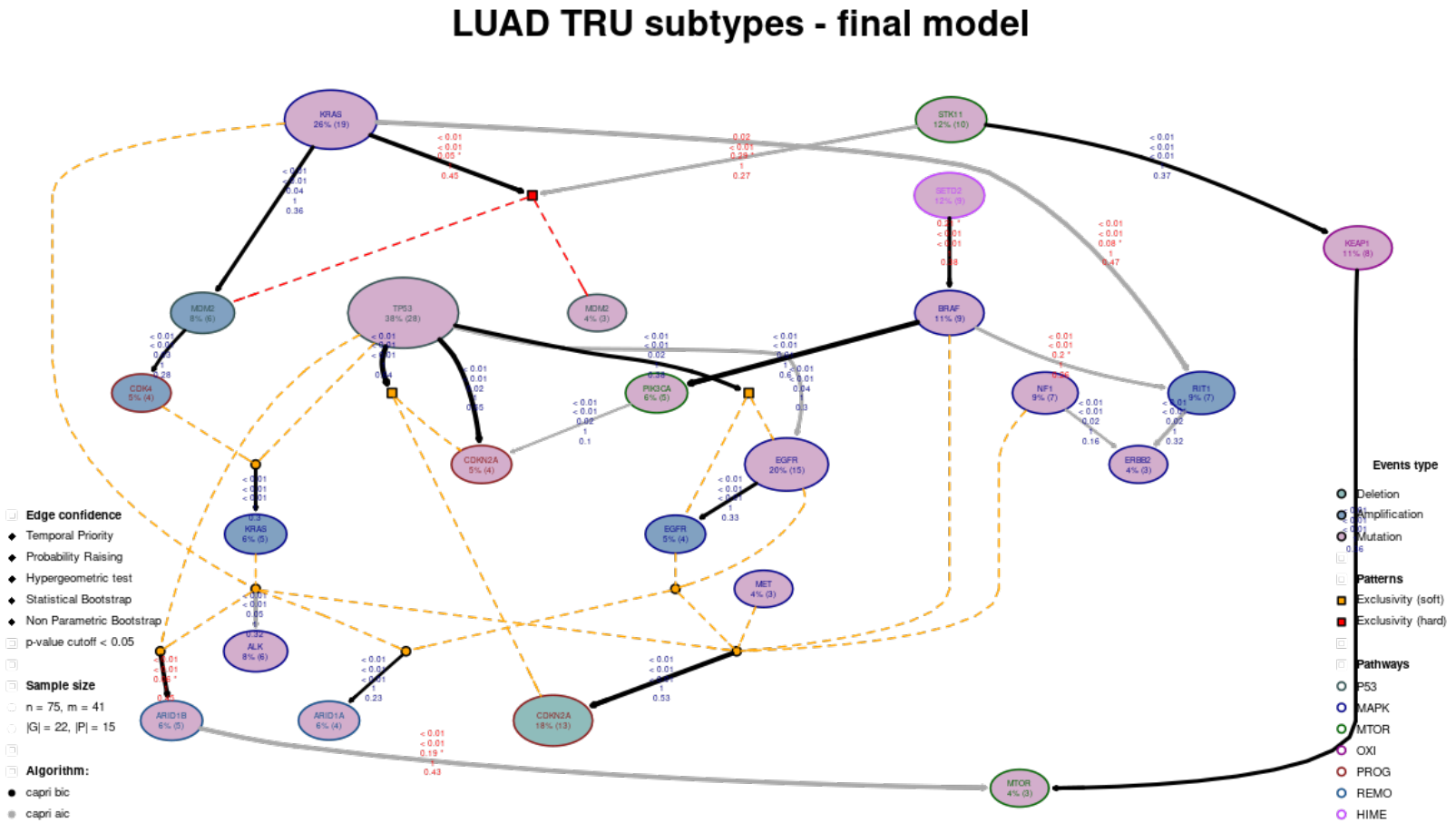


Figure: *TRONCO.plot* with bootstrap scores of the reconstructed model from the data of TRU subtype. Edge size represents non-parametric bootstrap scores.

Post-Reconstruction III

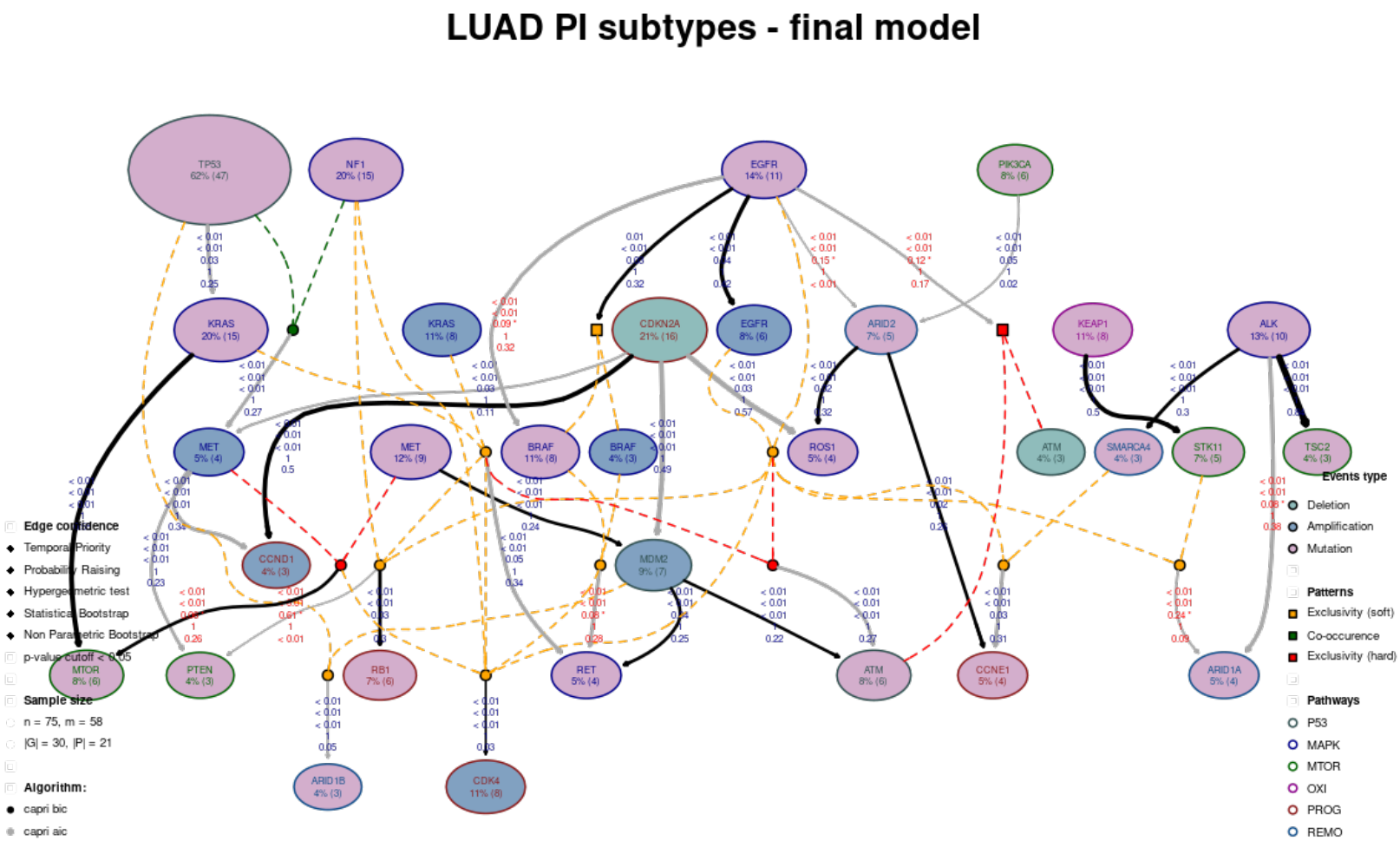


Figure: *TRONCO.plot* with bootstrap scores of the reconstructed model from the data of PI subtype. Edge size represents non-parametric bootstrap scores.

Post-Reconstruction III

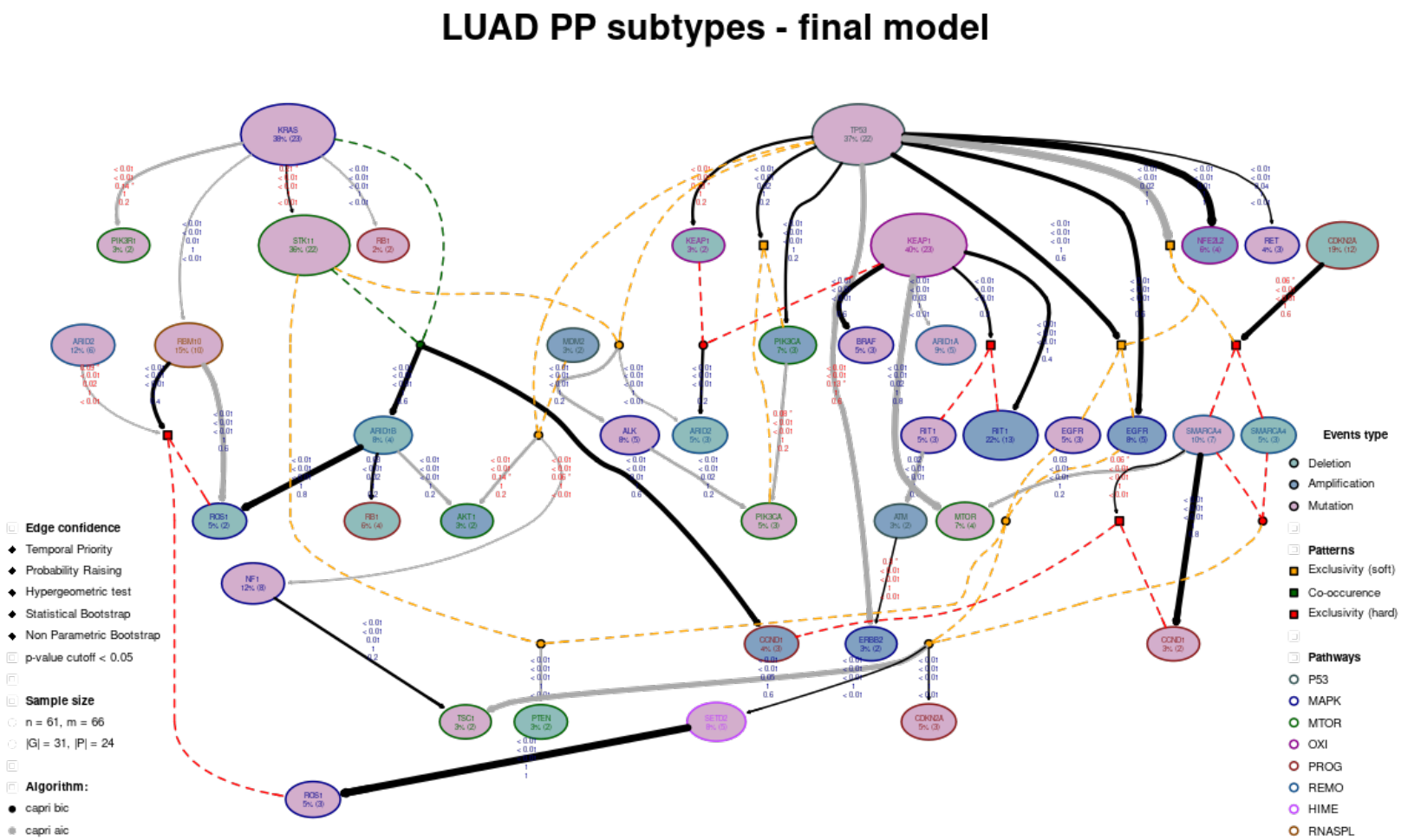


Figure: *TRONCO.plot* with bootstrap scores of the reconstructed model from the data of PP subtype. Edge size represents non-parametric bootstrap scores.

10-Fold Cross Validation

Then we perform a **10-fold cross validation** (with 10 iterations) studying using the methods included in *TRONCO library*:

- the **entropy-loss** for each model computed, that's the negated expected log-likelihood of the test set for the Bayesian network fitted from the training set
- the **prediction error** for each parent set of a certain node
- the **posterior classification error** for each edge that connect a node to a child

These studies have been performed both for *CAPRI AIC models* and *CAPRI BIC models*, producing good results.

10-Fold Cross Validation

capri_bic.SELECTS	capri_bic.SELECTED	c	c	c	c	c	capr	capri_	capri_bic.MEAN.PREDERR	capri_bic.SD.PREDERR	capri_bic.MEAN.POSTERR	capri_bic.SD.POSTERR
Mutation KEAP1	Mutation MTOR						80	100	0.04	0	0.04	0
Pattern OR_CDK4_TP53	Amplification KRAS						60	100	0.066666667	0	0.066666667	0
Mutation TP53	Pattern OR_CDKN2A						60	100	0.216	0.00843274	0.214666667	0.009838197
Mutation TP53	Pattern OR_EGFR						60	100	0.213333333	0	0.213333333	0
Pattern OR_EGFR_KRAS	Mutation ARID1A						40	100	0.053333333	0	0.053333333	0
Mutation KRAS	Amplification MDM2						20	100	0.08	0	0.08	0
Pattern OR_KRAS_TP53	Mutation ARID1B						20	100	0.066666667	0	0.066666667	0
Mutation TP53	Mutation CDKN2A						20	100	0.053333333	0	0.053333333	0
Mutation KRAS	Pattern XOR_MDM2						20	100	0.12	0	0.12	0
Mutation SETD2	Mutation BRAF						0	100	0.157333333	0.017554149	0.149333333	0.018645491
Pattern OR_CDKN2A	Mutation EGFR						0	100	0.224	0.021591036	0.236	0.014124665
Mutation STK11	Mutation KEAP1						0	100	0.129333333	0.020893616	0.126666667	0.01692394

Figure: Example of data extracted from 10-fold cross validation.

Entropy Loss

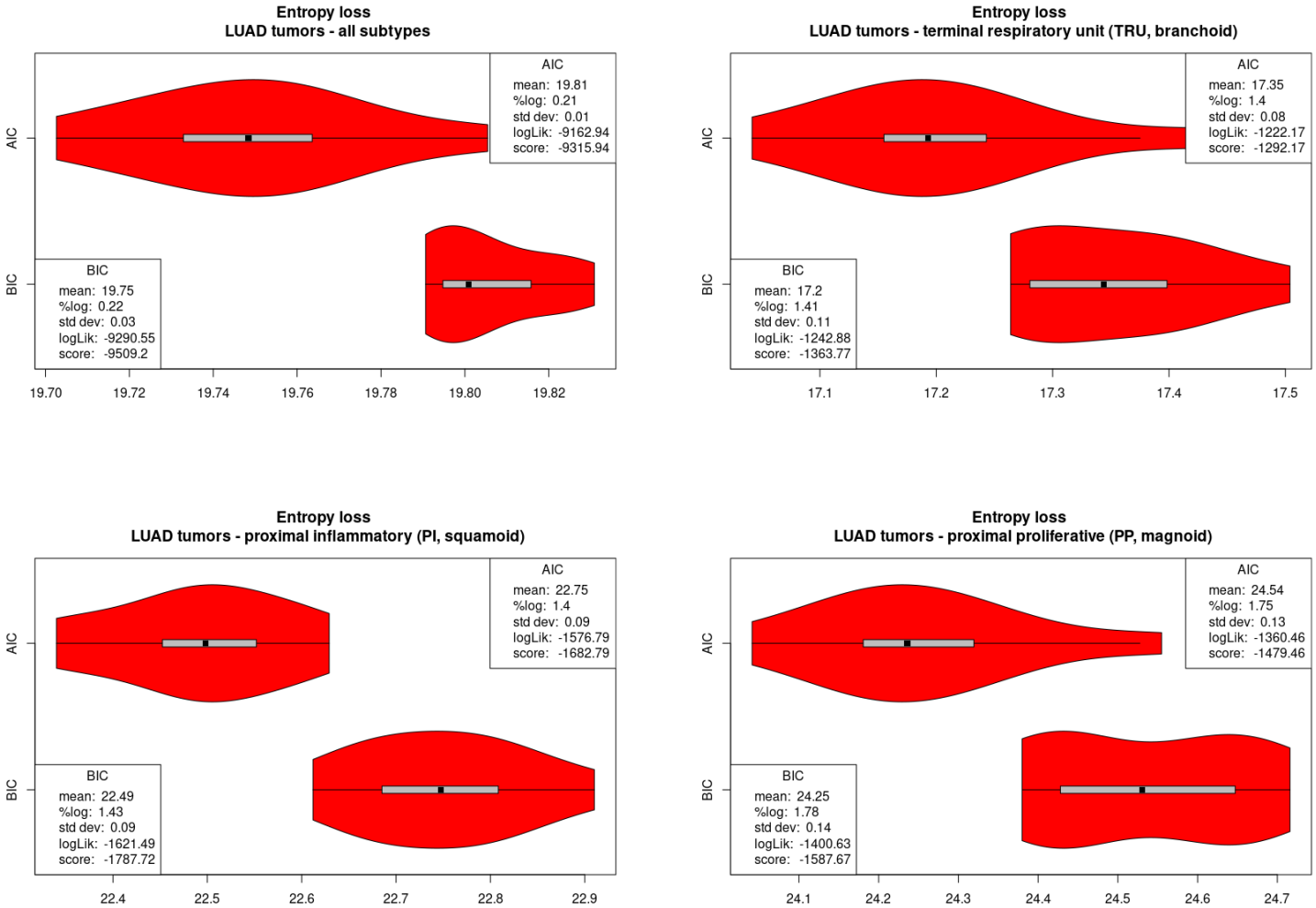


Figure: Violin plot, obtained by Vioplot [18], for entropy loss of all models.

Prediction Error

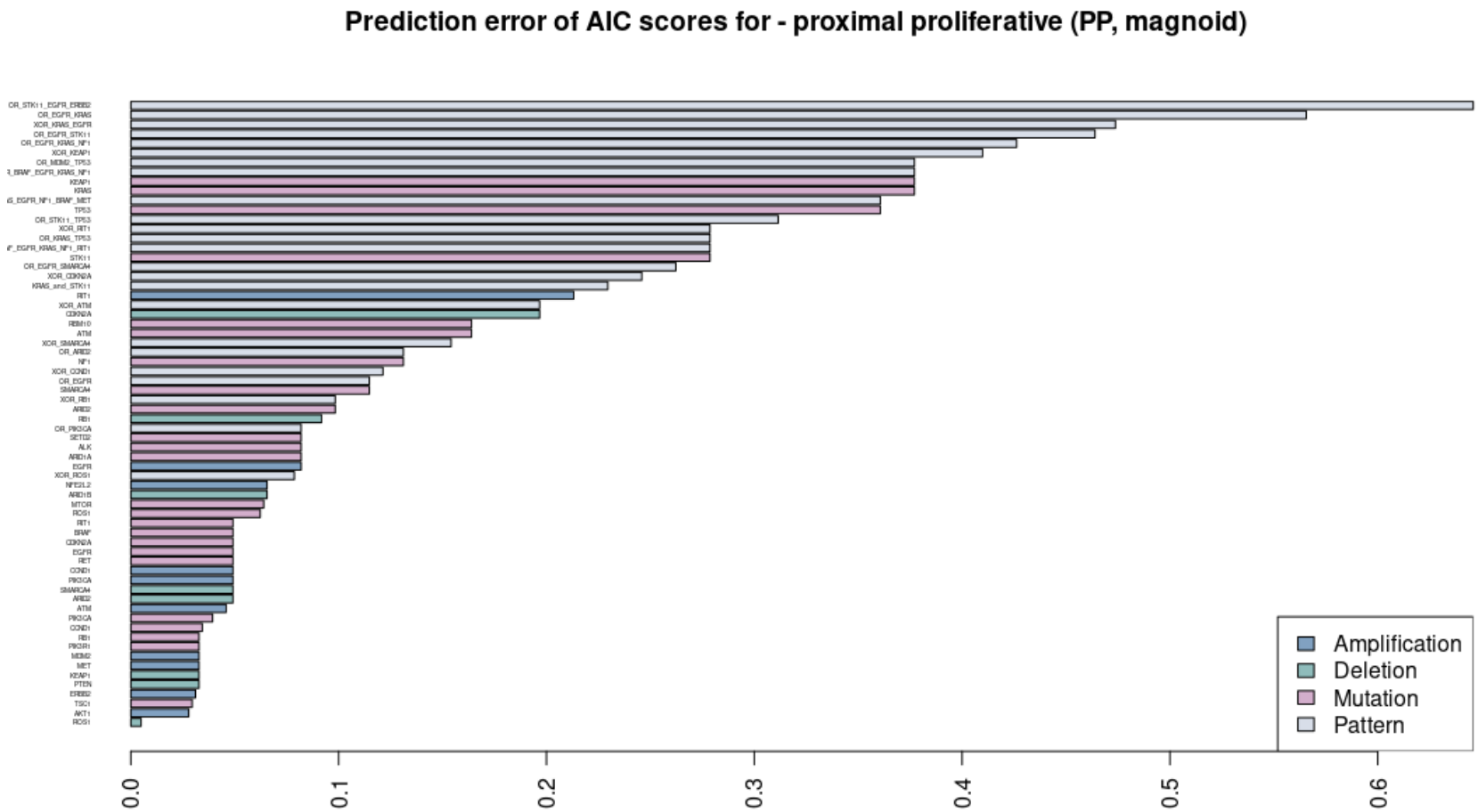


Figure: Barplot of prediction error of AIC model for PP subtype reconstruction.

Prediction Error

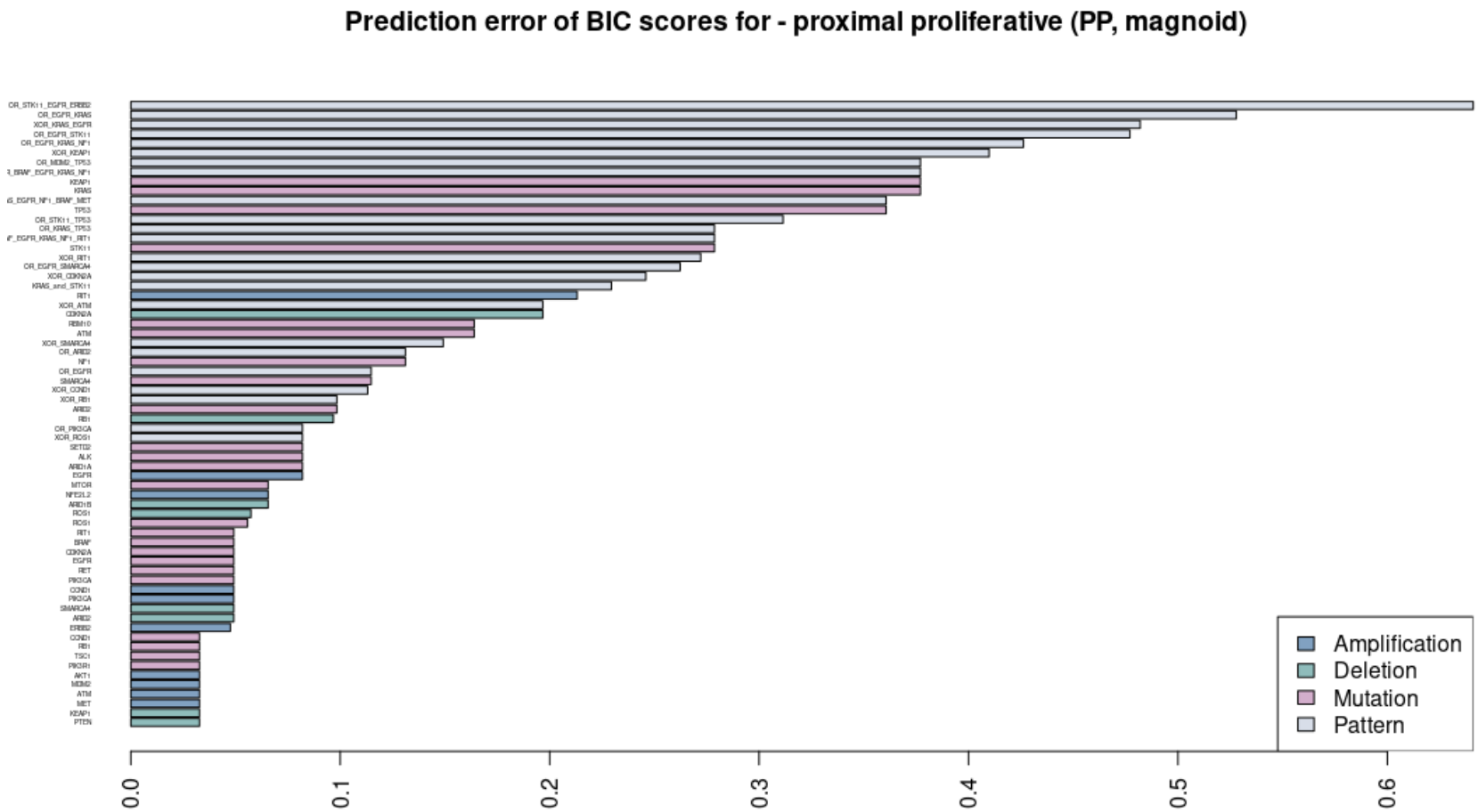


Figure: Barplot of prediction error of BIC model for PP subtype reconstruction.

Posterior Classification

Selects	#	Selected	#	Posterior Cl. (mi)	Posterior Cl. (sigma)
KEAP1	23	XOR_RIT1	16	0.301639344	0.026992544
KRAS	23	STK11	22	0.28852459	0.017622421
TP53	22	OR_EGFR_SMARCA4	16	0.278688525	0.02677038
KEAP1	23	RIT1	13	0.213114754	0
KRAS	23	RBM10	10	0.163934426	0
CDKN2A	12	XOR_SMARCA4	10	0.160655738	0.02158297
OR_MDM2_TP53	23	NF1	8	0.131147541	0
TP53	22	OR_EGFR	7	0.114754098	0
SMARCA4	7	XOR_CCND1	5	0.109836066	0.017366392
ARID1B	4	RB1	4	0.096721311	0.005184062
TP53	22	EGFR	5	0.081967213	0
RBM10	10	XOR_ROS1	5	0.081967213	0
TP53	22	OR_PIK3CA	5	0.081967213	0
OR_STK11_TP53	42	ALK	5	0.081967213	0
KEAP1	23	ARID1A	5	0.081967213	0
OR_EGFR_SMARCA4	16	SETD2	5	0.081967213	0
ARID2	6	XOR_ROS1	5	0.081967213	0
TP53	22	NFE2L2	4	0.06557377	0
KRAS_and_STK11	14	ARID1B	4	0.06557377	0
KEAP1	23	MTOR	4	0.06557377	0
SMARCA4	7	MTOR	4	0.06557377	0
SETD2	5	ROS1	3	0.054098361	0.019008226
PIK3CA	3	PIK3CA	3	0.054098361	0.01106473
RIT1	3	ATM	2	0.052459016	0.010368123
KRAS_and_STK11	14	CCND1	3	0.049180328	0

KEAP1	23	BRAF	3	0.049180328	0
TP53	22	PIK3CA	3	0.049180328	0
XOR_KEAP1	25	ARID2	3	0.049180328	0
ATM	2	ERBB2	2	0.049180328	0
OR_STK11_TP53	42	ARID2	3	0.049180328	0
OR_EGFR_SMARCA4	16	CDKN2A	3	0.049180328	0
ALK	5	PIK3CA	3	0.049180328	0
TP53	22	RET	3	0.049180328	0
ARID1B	4	ROS1	2	0.036065574	0.010368123
ARID1B	4	AKT1	2	0.03442623	0.005184062
SMARCA4	7	CCND1	2	0.032786885	0
TP53	22	KEAP1	2	0.032786885	0
NF1	8	TSC1	2	0.032786885	0
OR_MDM2_TP53	23	AKT1	2	0.032786885	0
TP53	22	ERBB2	2	0.032786885	0
OR_EGFR_STK11	28	PTEN	2	0.032786885	0
RBM10	10	ROS1	2	0.032786885	0
KRAS	23	PIK3R1	2	0.032786885	0
KRAS	23	RB1	2	0.032786885	0
OR_EGFR_SMARCA4	16	TSC1	2	0.032786885	0

del
amp
mut
pat

Figure: Posterior classification (μ and σ) of AIC model for PP subtype reconstruction.

Posterior Classification

Selects	#	Selected	#	Posterior Cl. (mi)	Posterior Cl. (sigma)
KEAP1	23	XOR_RIT1	16	0.306557377	0.037102175
KRAS	23	STK11	22	0.280327869	0.005184062
KEAP1	23	RIT1	13	0.214754098	0.005184062
CDKN2A	12	XOR_SMARCA4	10	0.147540984	0.025630687
TP53	22	OR_EGFR	7	0.114754098	0
SMARCA4	7	XOR_CCND1	5	0.113114754	0.023756355
ARID1B	4	RB1	4	0.086885246	0.013496272
OR_EGFR_SMARCA4	16	SETD2	5	0.081967213	0
RBM10	10	XOR_ROS1	5	0.081967213	0
TP53	22	EGFR	5	0.081967213	0
TP53	22	OR_PIK3CA	5	0.081967213	0
KRAS_and_STK11	14	ARID1B	4	0.06557377	0
TP53	22	NFE2L2	4	0.06557377	0
ARID1B	4	ROS1	2	0.062295082	0.006912082
SETD2	5	ROS1	3	0.060655738	0.019008226
KEAP1	23	BRAF	3	0.049180328	0
KRAS_and_STK11	14	CCND1	3	0.049180328	0
OR_EGFR_SMARCA4	16	CDKN2A	3	0.049180328	0
TP53	22	PIK3CA	3	0.049180328	0
TP53	22	RET	3	0.049180328	0
XOR_KEAP1	25	ARID2	3	0.049180328	0
ATM	2	ERBB2	2	0.047540984	0.005184062
SMARCA4	7	CCND1	2	0.040983607	0.017706942
NF1	8	TSC1	2	0.032786885	0
TP53	22	KEAP1	2	0.032786885	0

del
amp
mut
pat

Figure: Posterior classification (μ and σ) of BIC model for PP subtype reconstruction.

Outline

Result Analysis I

From a broad analysis of the results we can see in each model the presence of **mutations for *TP53***, primarily responsible for apoptosis.

Result Analysis I

From a broad analysis of the results we can see in each model the presence of **mutations for *TP53***, primarily responsible for apoptosis.

Another recurrent result is ***KEAP1* mutation** which concerns the response to oxidative stress.

Result Analysis I

From a broad analysis of the results we can see in each model the presence of **mutations for *TP53***, primarily responsible for apoptosis.

Another recurrent result is ***KEAP1* mutation** which concerns the response to oxidative stress.

Finally, there are **frequent mutations** concerning the genes of the *Ras family*, which also have regulatory functions of apoptosis, and genes of the *Raf family* that are related to oncogenes and participate in ***RAS-RAF-MEK-ERK signal transduction cascade***.

Result Analysis I

From a broad analysis of the results we can see in each model the presence of **mutations for *TP53***, primarily responsible for apoptosis.

Another recurrent result is ***KEAP1* mutation** which concerns the response to oxidative stress.

Finally, there are **frequent mutations** concerning the genes of the *Ras family*, which also have regulatory functions of apoptosis, and genes of the *Raf family* that are related to oncogenes and participate in ***RAS-RAF-MEK-ERK signal transduction cascade***.

These results seem to be consistent with what is present in the literature [19].

Result Analysis II

Molecular subtypes

Regarding the three subtypes we can observe in the models that the hypotheses extracted from the marker paper are represented [6]:

- "the PI subtype was characterized by solid histopathology and co-mutation of *NF1* and *TP53*"
- "the PP subtype was enriched for mutation of *KRAS*, along with inactivation of the *STK11* tumour suppressor gene"

On the other hand, for TRU subtype, in the marker paper we have: "the TRU subtype harboured the majority of the *EGFR-mutated* tumours as well as the kinase fusion expressing tumours". This could be confirmed in the *DAG* due to the presence of genes mutations from *RAF/RAS families*, in addition to *EGFR* mutations.

OncoBN Comparison I

Trying to validate our results we searched the literature for further studies on *LUAD* and one of them is another *R* library called **OncoBN** [20].
In the paper there are also some comparisons with *CAPRI algorithm*

In this study the authors recovered three highconfident roots: *KRAS*, *KEAP1* and *EGFR* plus a high confident edge *TP53*$\rightarrow$*RB1*.

From a quick visual comparison, the model obtained with *OncoBN* seems consistent with our model obtained from the *PiCnIc pipeline* (with all subtypes), even if the latter takes into account a much more complex input which involves an equally complex final model.

OncoBN Comparison I

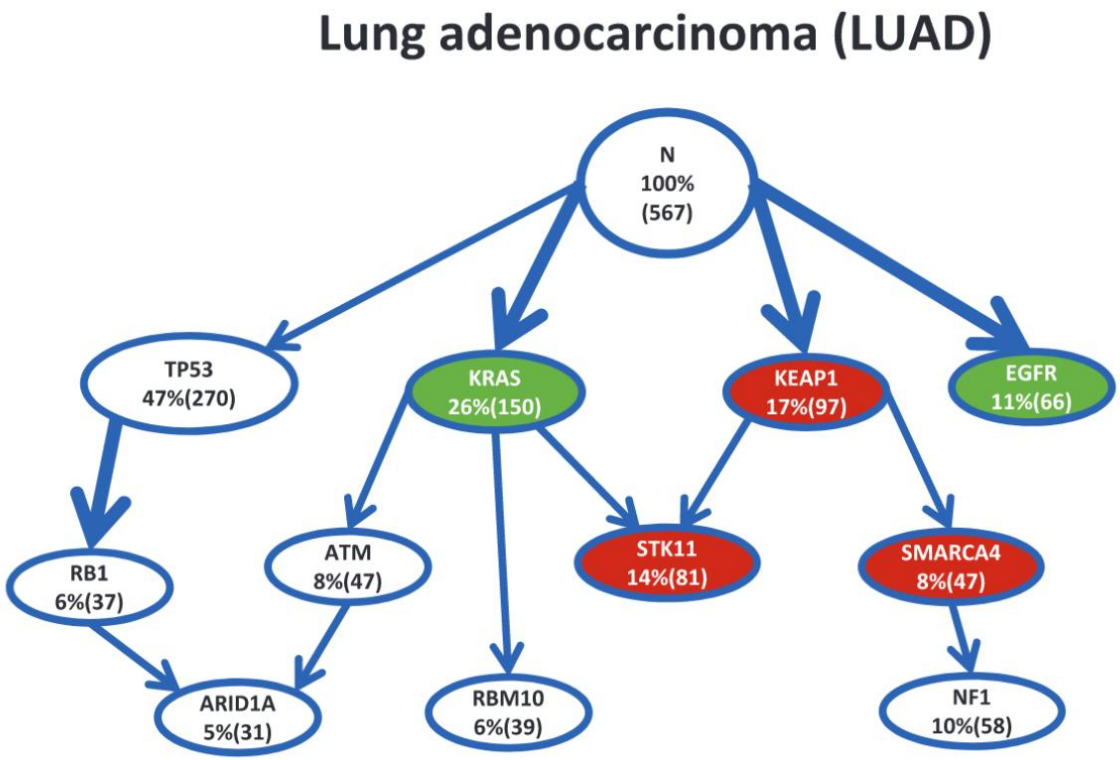


Figure: Synthetically lethal mutations of **LUAD**, *KRAS* and *EGFR* (green nodes), appear in disjoint branches. Frequently co-occurred mutations *STK11*, *KEAP1* and *SMARCA4* occupy a branch of the inferred network (red nodes). Subtype defining mutations *TP53* and *RB1* are ordered with high confident [20].

OncoBN Comparison II

Comparison with *PiCnIc models*

- **All subtypes**: all the edges and nodes can be found in *CAPRI* model
- **TRU subtype**: we can't find any edge but we have some consistent nodes
- **PI subtype**: we can't find any edge, except *KEAP1*→*STK11*, but we have some consistent nodes
- **PP subtype**: we have some consistent edges but it has all nodes. It's interesting to report that *TP53* is strongly related with some mutations, with few samples, that are not shown in *OncoBN* plot.

These can be considered expected results due to the fact that *OncoBN* does not study mutational subtypes.

Outline

References I

[1] Giulio Caravagna, Alex Graudenzi, Daniele Ramazzotti, Rebeca Sanz-Pamplona, Luca De Sano, Giancarlo Mauri, Victor Moreno, Marco Antoniotti, and Bud Mishra.
Algorithmic methods to infer the evolutionary trajectories in cancer progression.
Proceedings of the National Academy of Sciences, 113(28):E4025–E4034, 2016.

[2] Luca De Sano, Giulio Caravagna, Daniele Ramazzotti, Alex Graudenzi, Giancarlo Mauri, Bud Mishra, and Marco Antoniotti.
Tronco: an r package for the inference of cancer progression models from heterogeneous genomic data.
Bioinformatics, 32(12):1911–1913, 02 2016.

[3] Mateus Camargo Barros-Filho, Florian Guisier, Leigha D Rock, Daiana D Becker-Santos, Adam P Sage, Erin A Marshall, and Wan L Lam.
Tumour suppressor genes with oncogenic roles in lung cancer.
In *Genes and Cancer*. IntechOpen, 2019.

[4] Yangyang Liu, Lu Liang, Liang Ji, Fuquan Zhang, Donglai Chen, Shanzhou Duan, Hao Shen, Yao Liang, and Yongbing Chen.
Potentiated lung adenocarcinoma (luad) cell growth, migration and invasion by lncrna dars-as1 via mir-188-5p/klf12 axis.
Aging (Albany NY), 13(19):23376, 2021.

[5] `https://en.wikipedia.org/wiki/Adenocarcinoma_of_the_lung`.

References II

[6] Cancer Genome Atlas Research Network.
Comprehensive molecular profiling of lung adenocarcinoma.
Nature, 511(7511):543–550, Jul 2014.
25079552[pmid].

[7] Tong Zhang, Shao-Wu Zhang, and Yan Li.
Identifying driver genes for individual patients through inductive matrix completion.
Bioinformatics, 37(23):4477–4484, 06 2021.

[8] Denise N. Slenter, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, Nuno Nunes, Jonathan Mélius, Elisa Cirillo, Susan L. Coort, Daniela Digles, Friederike Ehrhart, Pieter Giesbertz, Marianthi Kalafati, Marvin Martens, Ryan Miller, Kozo Nishida, Linda Rieswijk, Andra Waagmeester, Lars M. T. Eijssen, Chris T. Evelo, Alexander R. Pico, and Egon L. Willighagen.
Wikipathways: a multifaceted pathway database bridging metabolomics to other omics research.
Nucleic acids research, 46(D1):D661–D667, Jan 2018.
29136241[pmid].

[9] Antonio Colaprico, Tiago C. Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S. Sabedot, Tathiane M. Malta, Stefano M. Pagnotta, Isabella Castiglioni, Michele Ceccarelli, Gianluca Bontempi, and Houtan Noushmehr.
TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data.
Nucleic Acids Research, 44(8):e71–e71, 12 2015.

References III

[10] Anand Mayakonda, De-Chen Lin, Yassen Assenov, Christoph Plass, and H Phillip Koeffler.
Maftools: efficient and comprehensive analysis of somatic variants in cancer.
Genome research, 28(11):1747–1756, 2018.

[11] `https://docs.cbioportal.org/1.-general/faq`.

[12] Özgün Babur, Mithat Gönen, Bülent Arman Aksoy, Nikolaus Schultz, Giovanni Ciriello, Chris Sander, and Emek Demir.
Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations.
Genome biology, 16(1):1–10, 2015.

[13] Ozgun Babur, Ugur Dogrusoz, Emek Demir, and Chris Sander.
Chibe: interactive visualization and manipulation of biopax pathway models.
Bioinformatics, 26(3):429–431, 12 2009.

[14] Jianxin Shi, Xing Hua, Bin Zhu, Sarangan Ravichandran, Mingyi Wang, Cu Nguyen, Seth A Brodie, Alessandro Palleschi, Marco Alloisio, Gianluca Pariscenti, et al.
Somatic genomics and clinical features of lung adenocarcinoma: a retrospective study.
PLoS medicine, 13(12):e1002162, 2016.

[15] Xing Hua, Paula L. Hyland, Jing Huang, Lei Song, Bin Zhu, Neil E. Caporaso, Maria Teresa Landi, Nilanjan Chatterjee, and Jianxin Shi.
Megsa: A powerful and flexible framework for analyzing mutual exclusivity of tumor mutations.
American journal of human genetics, 98(3):442–455, Mar 2016.
26899600[pmid].

References IV

[16] Daniele Ramazzotti, Giulio Caravagna, Loes Olde Loohuis, Alex Graudenzi, Ilya Korsunsky, Giancarlo Mauri, Marco Antoniotti, and Bud Mishra.
Capri: efficient inference of cancer progression models from cross-sectional data.
Bioinformatics, 31(18):3016–3026, 05 2015.

[17] Christopher A Miller, Joshua McMichael, Ha X Dang, Christopher A Maher, Li Ding, Timothy J Ley, Elaine R Mardis, and Richard K Wilson.
Visualizing tumor evolution with the fishplot package for r.
BMC genomics, 17(1):1–3, 2016.

[18] Daniel Adler and S. Thomas Kelly.
vioplot: violin plot, 2020.
R package version 0.3.6.

[19] Jian Carrot-Zhang, Xiaotong Yao, Siddhartha Devarakonda, Aditya Deshpande, Jeffrey S Damrauer, Tiago Chedraoui Silva, Christopher K Wong, Hyo Young Choi, Ina Felau, A Gordon Robertson, et al.
Whole-genome characterization of lung adenocarcinomas lacking the rtk/ras/raf pathway.
Cell reports, 34(5):108707, 2021.

[20] Phillip B. Nicol, Kevin R. Coombes, Courtney Deaver, Oksana Chkrebtii, Subhadeep Paul, Amanda E. Toland, and Amir Asiaee.
Oncogenetic network estimation with disjunctive bayesian networks.
Computational and Systems Oncology, 1(2):e1027, 2021.

Thanks for your attention!
Questions?

https://github.com/dlcgold/DCB-project