Università degli Studi di Milano Bicocca

Scuola di Scienze

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di laurea in Informatica

Implementation of a pipeline to infer somatic selective advantage relations in Renal Carcinoma

Relatore: *Giancarlo Mauri*

Co-relatore: *Marco Antoniotti*

Co-relatore: *Daniele Ramazzotti*

Co-relatore: *Giulio Caravagna*

Relazione della prova finale di:

Andrea Campagner

Matricola 761976

Anno Accademico 2014-2015

Abstract

The availability of biological data, produced by modern *Next Generation Sequencing* technologies, provided researchers ways to tackle complex biological problems, like the study of cancer diseases.

An important aspect that is to be studied, in order to understand cancer, is *tumorigenesis*, the biological process resulting in the formation of a tumor, that consists in an accumulation of alterations in a set of genes that are important for cell regulatory activities.

An important problem is then to identify these relevant mutated genes, so called *drivers*, and also to infer selective advantage relations between them, in order to reconstruct progression models able to explain the progression of the disease.

As of today various algorithms and techniques to solve this progression inference problem are available and, in particular, the BIMIB group developed a novel technique called *CAPRI*.

Cancer diseases, however, usually exhibits problematic characteristics (e.g. tumor heterogeneity, presence of alterations irrelevant for the progression of the disease, etc.) that hinders the ability of existing algorithms to infer progression models.

In this document we report on the pipeline, inspired by a previous work conducted by the BIMIB group in the context of a study on Colorectal Cancer, that we implemented in order to cope with these problems, thus providing a way to reconstruct progression models from previously available expression data.

This pipeline has been implemented in order to study a specific type of cancer, known as Clear Cell Renal Cell Carcinoma, and integrates various external tools in order to solve the problems presented above so that its general structure allows to: (i) import and process raw or pre-processed expression data, (ii) extract a set of possible subtypes of patients that are likely to have a similar progression of the disease, (iii) select genes that are relevant for the progression of the disease, (iv) identify patterns of mutual exclusivity between relevant genes and (v) infer selective advantage relations and establish which are recurrently inferred across various subtypes.

In particular we applied the pipeline to two different studies: a cross-sectional study conducted by The Cancer Genome Atlas (TCGA) and a single-patient study conducted by Gerlinger et al. For the TCGA study we applied the whole pipeline, on the contrary for the single-patient study we implemented only the step corresponding to inference of progression models, because of the reduced dimensionality of the dataset.

Contents

List of Figures

Chapter 1

Introduction

Next Generation Sequencing (NGS) Technologies are producing huge amounts of data, augmenting the possibility to study and analyze complex biological phenomena like cancer diseases.

In this document we present the implentation of a pipeline, adapted from a conceptual pipeline defined for a previous study conducted by the BIMIB group in the context of Colorectal Cancer, that we applied to analyze Clear Cell Renal Cell Carinoma.

This pipeline is partially automated and integrates various pre-existing external tools in order to process and analyze NGS preprocessed data to infer relationships between genomic events that are likely to be relevant for the progression of the disease.

In this context relevant relationships means relationships of *selective advantage*, that is, a mutation of a particular gene enables clones harboring this mutation to survive and reproduce better than other cells starting a wave of clonal expansion, in which mutated clones proliferate, that will end when a mutated clone will acquire a new mutation giving further selective advantage. In this case we say that a *selective advantage relationship* occur between the two mutations, and in particular the first *selects for* the latter.

1.1 Preamble

Cancer is a *genetic disease* figuring among the leading causes of morbidity and mortality worldwide, unfortunately cancer is a *clonal disorder* that is very specific to each individual and cancer type, as suggested by the fact that clinically identical tumors have often few common genetic features. In addition to this *inter-patient heterogeneity* there is also heterogeneity at patient-level, so called *intra-tumor heterogeneity*. Heterogeneity clearly has implications

for predictive or prognostic strategies against cancer.

To understand cancer is necessary to understand *tumorigenesis* (i.e. the biological process resulting in the formation of a tumor), which consists in an accumulation of *genetic mutations* in three types of genes:

- `Oncogenes`
  Genes that regulate cell division of healthy systems; as such, mutated oncogenes may lead to cells growing out of control;

- `Tumor suppressor genes`
  Genes that prevent cells to become cancerous; as such, mutated tumor suppressors genes may have an hindered functioning thus allowing cells to progress to cancer;

- `Stability genes`
  A particular type of tumor suppressor genes, that are involved in recognizing and repairing DNA damage; as such, mutated stability genes may lead to an increased mutation rate of all genes (including oncogenes and tumor suppressor genes).

Mutations are usually ascribed to two broad categories: *drivers* (i.e. mutations that inhibit key cell regulatory processes eventually leading these cells to become cancerous) and *passengers* (i.e. mutations that have no direct effect on cancerous development), these mutations can either hit a single gene or a wide chromosomal region.

An important aspect of understanding tumorigenesis is to identify early driver mutations (i.e. driver mutations that are likely to start the progression of cancer) and also selective advantage relations between driver mutations, in order to construct and devise progression models that are able to explain cancer progression. Progression models are probabilistic observational models, over a set of mutated genes, identifying relations among these genes that capture selective advantage relations and inducing a causal structure able to explain the order in which cancer can progress to acquire increasingly higher fitness.

The most recent approaches tend to adopt *Bayesian Networks* to model these progressions, this is because they are well-suited to represent both branching and convergent evolutionary trajectories. However, although these approaches are able to infer confluent selective advantage relations involving multiple genes, they are able to do so only in the specific case in which multiple mutations *co-occur* to select for a certain event. This is a rather severe limitation in the context of cancer, because mutations disrupting a single function are usually distributed among *multiple genes of a common*

pathway, and most samples are consequently mutated in only one of those genes because additional alterations would not convey a further selective advantage to the tumor. For this reason, these confluences, usually exhibit *mutual exclusion*. In general a single gene could be selected by a group of genes related by an arbitrary relation, we call such relation a *pattern*. In order to overcome this limitation, the BIMIB group developed a novel algorithmic technique, named `CAPRI` (*Cancer Progression Inference*) [1], that we integrated in our pipeline.

1.2 Introduction to the CAPRI algorithm

CAPRI is an algorithm, developed by the BIMIB group, in order to solve the progression inference problem. The algorithm takes as input a set G of n mutational events across m samples (represented as an $m \times n$ matrix) and a set ϕ of logical formulae, representing patterns. The main idea of the algorithm is to combine a scoring function and subsequent filtering and model selection techniques (i.e. maximum likelihood estimations and bootstrap iterations) in order to filter out spurious relations.

The scoring function is based on Patrick Suppes' conditions for *probabilistic causation*, that can be stated as follows:

Definition. *Let i and j be two observables, represented as two Bernoulli random variables, and let t_i and t_j, respectively, the time of occurrence of i and j; we say that a selectivity relation hold among i and j if:*

(i) $t_i < t_j$*; that is, i occurs before j;*

(ii) $P(j|i) > P(j|\bar{i})$*; that is, observing i raises the probability of observing j.*

We integrated CAPRI in the pipeline for various reasons:

- It has the ability to test arbitrarily complex relations among events, in a supervised setting in which these relations are explicitely given as an input;
- It outperform various state-of-the-art algorithms, in particular in the presence of noise and with limited sample size.

Furthermore various convergence properties were proved for the CAPRI algorithm, in particular it has been proved that the algorithm is *sound and complete*.

1.3 Introduction to Clear Cell Renal Cell Carcinoma and data summary

As previously stated, we developed the pipeline that is the object of this document in the context of a case study, in which we analyzed Clear Cell Renal Cell Carcinoma, a specific type of cancer affecting kydneys. For our study we used two sources, namely the study conducted by The Cancer Genome Atlas (TCGA [2]) consortium and the study conducted by Gerlinger et al. in [3] from which we collected all data used in our analysis and also biologically relevant information.

The nature of these two sources is vastly different: while the study conducted in [2] is a large population study (over 400 patients) the study conducted in [3] involved only 10 patients, also, the TCGA study is a so-called *cross-sectional study* in which biopsies for multiple patients are collected at a certain time, on the contrary the study conducted by Gerlinger et al. is a Single-patient study in which for each patient multiple biopsies of different tumoral regions are considered. In what follows we refer to the study conducted by the TCGA consortium as `TCGA study` and to the study conducted in [3] as `Single-patient study`.

Introduction to Clear Cell Renal Cell Carcinoma Kidney cancers, also called Renal Cell Carcinomas, are a class of chemoterapy-resistant diseases that can be distinguished by the underlying gene mutations. Clear Cell Renal Cell Carcinoma (CCRCC), which is the most common among kidney cancers, is mainly related to mutations of the VHL gene but was also recently found to be related to alterations in the *SWI/SNF chromatin remodeling complex* which include genes like PBRM1, SETD2 and BAP1. A distinguishing caracteristic of CCRCC is a high degree of heterogeneity.

The TCGA study evidenced the importance of pathway-level alterations for the progression of the disease, in particular they highlighted importance of the *VHL/HIF*, *Chromatin remodeling* and *PI3K/AKT/MTOR* pathways , furthermore they evidenced the importance of the deletion of the 3P arm (containing all four genes mentioned above).

The Single-patient study further confirmed the heterogeneity of CCRCC also evidencing the fact that this type of cancer exhibits a branched, rather than linear, progression also confirming the importance of VHL mutations, deletions of the 3P arm and mutation of other 3P-related genes.

1.3.1 Data summary

TCGA study data summary The TCGA project for Human Clear Cell Renal Cell Carcinoma (KIRC, [2]) provides genome-scale analysis of 446 samples with exome sequence, DNA copy number, messenger RNA and microRNA expression data which we downloaded on 3 March 2015 from TCGA repository:

`https://tcga-data.nci.nih.gov/docs/publications/kirc_2013/`

We processed the following files (data freeze 19 April 2012):

- `hgsc.bcm.edu_KIRC.Mixed_DNASeq.Level_2.1.2.0.tar.gz`
  Somatic mutations profiles obtained via whole-exome sequencing for 491 samples (417 patients) with 12008 annotated mutations *Manual Annotation Format* (MAF) file). 74 patients had multiple samples associated in the MAF file; duplicated samples were resolved to have one sample per patient according to the following two-stage criteria: (*i*) we removed outliers (ie. samples with less than 10 mutations), (*ii*) we applied the TCGA guidelines for aliquote disambiguation, see `https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode`.
  For each annotated gene, Entrez Gene Id and Hugo Symbol were annotated; however for 542 genes the corresponding Entrez Id was unsolved (ie. set to 0). Since a map between Entrez Ids and corresponding Hugo Symbols is needed in subsequent stages, we solved the missing entries according to the following techniques: (*i*) we used an hand-curated map, (*ii*) we queried the web service `http://mygene.info`. All mutations annotated by the TCGA consortium were considered for our analysis.

- `all_thresholded.by_genes.kirc_120430.txt.zip`
  Copy number and *structural aberrations* - termed Copy Number Alterations (CNAs) - data was available, in GISTIC format, for 428 patients with 53419 annotated events on 21881 genes. Only *high-level gains* and *homozygous deletions* were considered for our analysis (GISTIC scores +2, -2 termed "Amplification" and "Deletion"), resulting in a reduction to 10253 events 9327 genes;

- `KIRC Clinical Data Jul-31-2012.xlsx`
  Clinical data summary (sample barcode and tumor stage) was available for 446 samples;

- `focal_data_by_genes.kirc_120430.txt.zip`
  Since the TCGA consortium states that, for CCRCC, arm-level CNAs

are more relevant than focal-level (single gene) CNAs, we useda map of genes to respective chromosome arms (at the highest possible resolution level, e.g. 3p25.1) to convert the focal-level CNA dataset, obtained as stated previously, to an arm-level CNA dataset.

The input cohort of patients with *both* CNAs and somatic mutation data consists of 411 patients with 22016 alterations on 16196 distinct genes, divided as follow:

alteration type	*count*	*source*
somatic mutations	11861	MAF file
focal amplifications	5699	GISTIC score +2
focal deletions	4456	GISTIC score -2
altered genes	16196	

The input cohort of patients with *both* arm-level CNAs and somatic mutation data consists of 411 patients with 12414 alterations on 12322 distinct sites, divided as follow:

alteration type	*count*	*source*
somatic mutations	11861	MAF file
arm-levell amplifications	334	GISTIC score +2
arm-level deletions	219	GISTIC score -2
altered sites	12322	

For subsequent analyses, described in the following sections, we selected the dataset with arm-level CNAs and somatic mutations, obtained as stated above.

Single-patient study data summary The single-patients study offered somatic mutation and CNAs for ten patients, named respectively `EV001`, `EV002`, `EV003`, `EV005`, `EV006`, `EV007`, `RMH002`, `RMH004`, `RMH008` and `RK26`, for each patient data was available for multiple regions. We processed the following files:

- `ng.2891-S2.xlsx`
  We manually processed the file, which contained somatic mutation data for each of the patients, to provide data that could be easily handled by our pipeline, in particular we created a separate file for each patient. We retained each type of Somatic mutation, not collapsing them into a single type as was done for the TCGA study (in which we collapsed all somatic mutation types in the Mutation type);

- `Figure 2`
  We manually created files for CNA data by extracting data from Figure 2 of the main text of [3]. CNA data was available at arm-level at the highest possible resolution (e.g. 3p25.3).

We only retained data for those genes that were found to be relevant in the previous analysis conducted by the authors of the study. The study also provided phylogenetic trees in order to predict the progression of the disease for each of the patients.

1.4 Description of the implemented pipeline

As previously stated, the main goal of this study was to devise and develop a pipeline, in the context of a specific case study, that will provide a guideline for researchers interested in performing progression inference analysis of cancer diseases.

The pipeline was implemented using the R programming language, the reason for this choice was two-fold:

- The R programming language is one of the de-facto standard programming environments in bioinformatics, providing various libraries and tools for manipulation of biological data;
- The R programming language is well suited for matrix manipulation, visualization and also naturally includes various statistical techniques.

The pipeline that we implemented is, in general, structured through 5 steps:

- `Data importation`
  In this step, raw datasets provided as input are processed in order to obtain data in a format suited to be used for the next steps of the pipeline;
- `Subtypes extraction`
  Given the heterogenity of cancer, in order to reduce confounding effects due to this heterogeneity, tumor stratification (i.e. clustering patients according to some biological criteria) is a critical step needed to define subtypes of patients that are likely to have a similar progression of the disease. For the execution of this step we interfaced the pipeline (using functionalities provided in the CAPRI tool) with an external tool. Details of this step, as applied in our case study, are found in Chapter 2;

- `Driver events selection`
  Since passengers mutations are irrelevant for the progression of the disease it's important to select only driver events (i.e. events that are likely to be important for the progression of the cancer) in order to both reduce the size of datasets and also infer progression models involving only genes likely to be relevant. In our case study we implemented this step in a supervised manner. Details are given in Chapter 2;

- `Pattern selection`
  In order to be able to test for arbitrarily complex relationg among groups of genes, the CAPRI tool, requires users to manually input *patterns* representing these relations. For this step the pipeline is interfaced with an external Java tool, named `Mutex`, but also previously known information could be employed (in a supervised fashion). This step is described in Chapter 3;

- `Reconstruction of progression models`
  The final step of the pipeline is devoted to inference of progression models from processed and filtered datasets obtained in the previous steps. For this step we use various functionalities offered by the TRONCO tool. This step is described in Chapter 4.

A graphical visualization of the pipeline, in the context of the case study we performed based on the TCGA study, is shown in Figure 1.1.

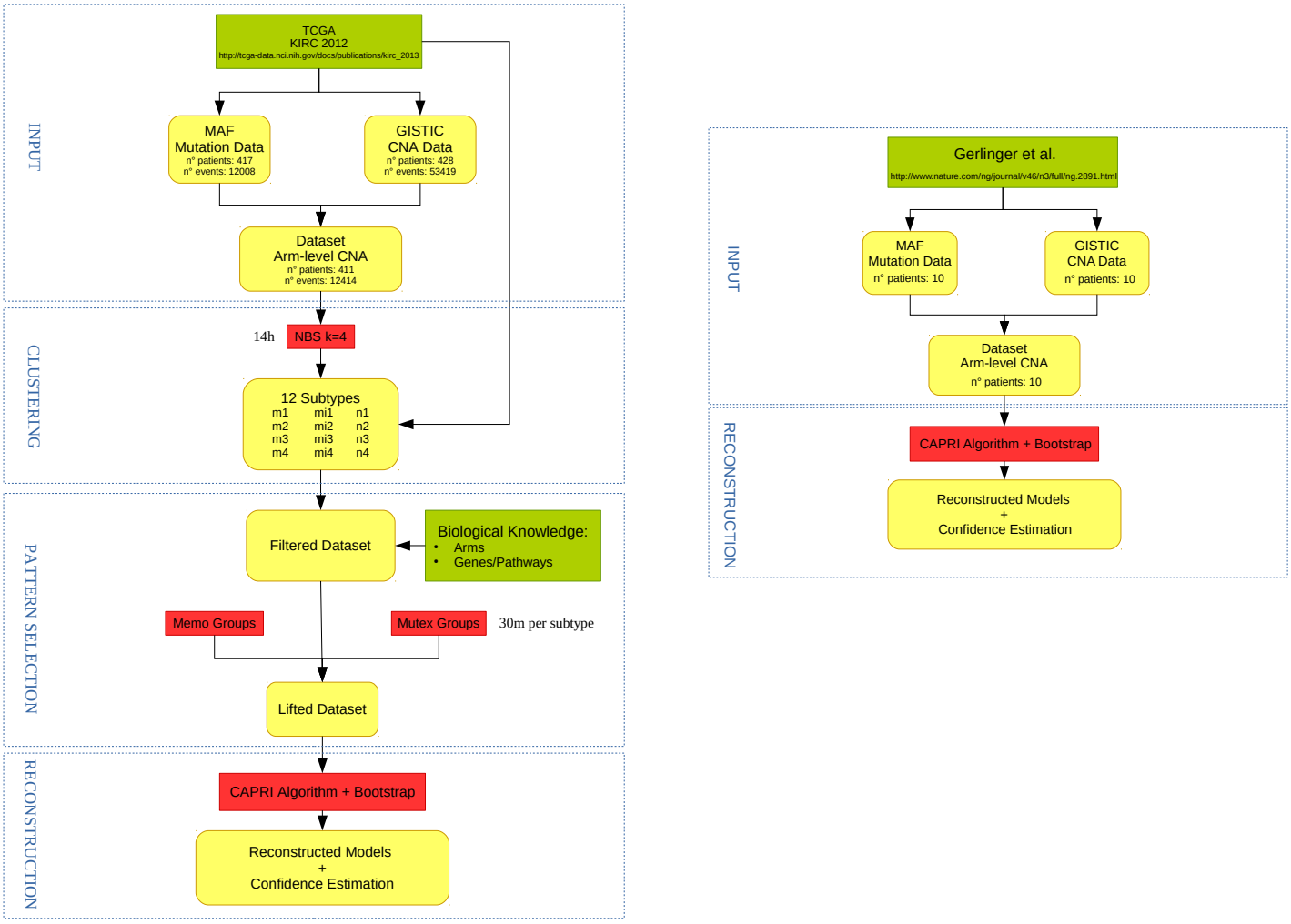


Figure 1.1: The data-analysis pipeline implemented in this work. Left panel: pipeline used for the TCGA dataset. We fetched somatic mutation and Copy Number profiles for the TCGA study KIRC (2012, [2]). The cohort of patients was split according to 3 different clustering techniques to determine potential CCRCC subtypes progressing with different patterns of somatic evolution. 12 Subtypes were determined by clustering mRNA expression, microRNA expression - provided by the TCGA consortium - or somatic mutations (3) - performed with the NBS tool [5]. Patterns to extract branched/confluent evolution trajectories over those genes were fetched from the TCGA paper and predicted with computational techniques. Genes/arms inputed to CAPRI are those either altered with frequency $> 5\%$ or part of a pattern. Confidence estimation is performed to assess the level of confidence of relations inferred in the reconstructed models. Right panel: pipeline used for the Single-patient study, in this case we directly applied the CAPRI algorithm to reconstruct progression models.

Chapter 2

Subtypes extraction and driver selection

To reduce confounding effects due to *intra-tumor heterogeneity* and augment our ability to infer CCRCC progression models we analyzed the cohort after applying different clustering techniques on the input samples, in order to identify groups of patients which are likely to progress in a similar way. In what follows and in the figures we will refer to mRNA expression, microRNA expression and mutation subtypes to indicate which TCGA data was used to compute clusters.

The result of these analyses is the extraction of the following subtypes:

	mRNA expression[†]	n	**microRNA expression**[†]	n	**Mutations**[‡]	n
1	m2	131	mi1	74	n1	64
2	m2	84	mi2	109	n2	194
3	m3	87	mi3	132	n3	52
4	m4	83	mi4	69	n4	101
total		**385**		**384**		**411**

[†]Provided as TCGA data

[‡]Computed with the NBS clustering tool

In order to reduce the dimension of the dataset, in particular the number of genes that we consider for further analyses, we performed a selection of events, selecting only events that are considered to be relevant for CCRCC progression in the literature, in particular the sources for these gene and arms are [2] and [3].

2.1 mRNA and microRNA expression subtypes

The TCGA consortium evaluated CCRCC subtypes for *mRNA* and *microRNA* expression, these clustering assignments were computed with consensus clustering via non-negative matrix factorization using the NMF [4] R package, using standard parameters: 200 primary iterations for clustering and 50 iterations for factorization rank estimation.

Non-negative matrix factorization is a clustering technique (which is also used in the tool that we employed for somatic mutation clustering) where a non-negative matrix is factorized into two smaller matrices, in an attempt to reduce dimensionality of the initial dataset, formally:

Definition. *Let V be a non-negative $n \times m$ matrix and $r > 0$ an integer. Non-Negative Matrix Factorization (NMF) consists in finding an approximation:*

$$V \simeq WH$$

where W, H are, respectively, $n \times r$ and $r \times m$ non-negative matrices.

Both mRNA and microRNA analysis identified four subgroups (termed, respectively, `m1` through `m4` and `mi1` through `mi4`). The number of samples with available mRNA expression based clustering assignments was 417, of these only 390 had somatic mutation data available. The number of samples with available microRNA expression clustering assignments was 414, of these only 390 had somatic mutation data available. We then restricted the two sets of patients, selecting only patients for which also CNA data was available, thus selecting 385 samples for mRNA expression and 384 for microRNA expression. Clustering assignments for mRNA and microRNA expression subtypes are provided as Supplementary Data in the TCGA study, see file `Data_file_S9_mRNA_miRNA_cluster_assignments.csv`.

2.2 Subtypes extracted from somatic mutations

2.2.1 Overview of the Network-Based Stratification tool

We used the *Network-Based Stratification* tool (NBS, [5]) to separate the input cohort in subtypes likely to progress through some common accumulation patterns of somatic alterations.

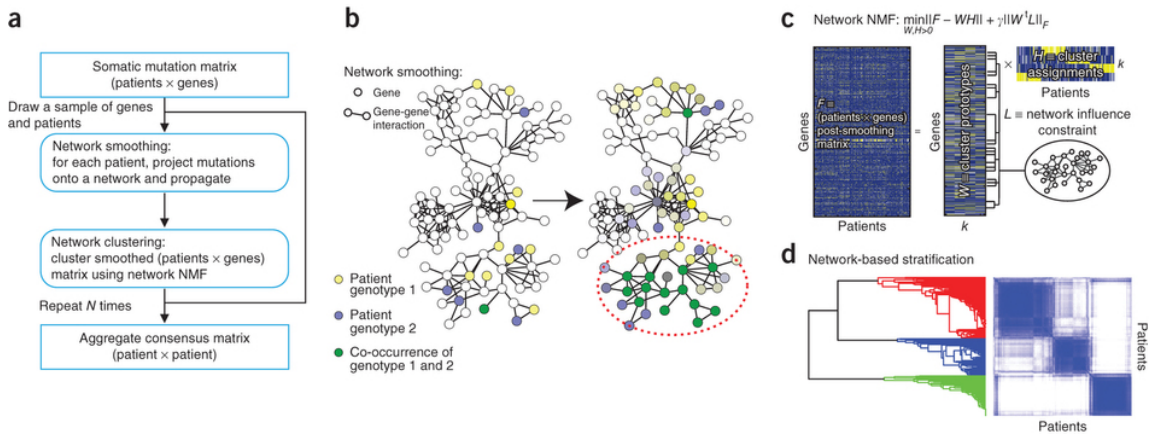


Figure 2.1: Overview of the NBS tool: NBS is a tool for clustering samples on the basis of somatic mutations. (a) Flowchart of the execution of NBS. The execution of NBS consists in 3 steps: in the first step mutation profiles are mapped onto a curated network which is then smoothed to propagate influence of mutated genes to respective neighbours, in second step mutation profiles are clustered using network NMF, in the third step the cohort is re-clustered using standard consensus clustering repeating multiple times the first two steps. (b) Example illustration of the network smoothing step of the algorithm. After smoothing mutation profiles of patients are propagated across the network. (c) Clustering of patients using non-negative matrix factorization. The decomposition attempts to minimize the function shown. (d) Final tumor subtypes are obtained from the consensus assignments of each patient after multiple applications of procedures b and c. - Picture and description taken from [5]

NBS is a tumor stratification technique designed to stratify tumor patients (i.e. cluster the patients into subgroups) on the basis of somatic mutation profiles, since somatic mutations are presumed to be causal drivers for cancer progression, using prior knowledge of the molecular network architecture of human cells. This prior knowledge is fundamental, because somatic mutation data is usually very sparse and is thus unlikely that two patients share the same mutation profiles.

The intuition behind NBS is that, although patients may not share the same mutations, they may share the same functional subnetworks affected by these mutations, thus NBS combines somatic mutation data, modeled as a matrix of 0/1 mutation profiles, with a curated gene-interaction network. Patients are separately mapped to this network that is then smoothed via network propagation, simulating a random walk on the network. This diffusion strategy is applied to propagate the influence of each mutated gene to the subnetwork of its neighbors, thus reducing the sparsity of the mutation profiles. These mutation profiles are then clustered using a variant of non-negative matrix factorization, called NetNMF. NetNMF is a variant of

standard non-negative matrix factorization that constraints standard NMF to respect the network structure underlying the matrix. Finally standard consensus clustering is applied, in which the above procedure is repeated multiple times, to promote robust cluster assignments. An overiew of the NBS tool is shown in Figure 2.1.

2.2.2 Clustering setting

NBS requires as input a set of 0/1 mutation profiles, obtained as described in the previous section, and the number $k > 0$ of clusters to extract; the input format for mutation signatures is the same we use for progression inference.

To augment the prediction capabilities of the tool - which uses solely information about somatic mutations - we used all the entries annotated in the MAF file for all the patients with exome-sequencing data available (417 patients, >12008 mutations). Results obtained with NBS were mapped to the samples with both somatic mutations and CNA data available. The tool was used with the following parameters (suggested by the authors as default):

Parameter	Value	Description
`prop-network`	`"ST90Q_adj_mat"`	*NBS propagation network*[†]
`infl-network`	`"glap_subnetwork_ST90"`	*graph influence measure derived from network*
`outDeg`	`11`	*number of nearest neighbours for Laplacian influence*
`min_mutation`	`10`	*minimum number of mutations in a sample to cluster*
`nmf_type`	`"netnmf"`	*non-negative matrix factorisation technique for consensus clustering*
`nsample`	`100`	*number of times to perform non-negative matrix factorisation technique as part of consensus clustering*
`smp_feat`	`0.8`	*proportion of samples to include in consensus clustering*
`smp_ind`	`0.8`	*proportion of genes to include in consensus clustering*
`min_mutations`	`9`	*minimum number of mutations per sample to include*
`proV`	`0.7`	*network-dependent propagation value (for STRING)*
`k`	`2, 3, 4 ,5`	*number of clusters to extract (variable)*

[†]Network prepared in [5] by including the top 10% of interactions according to the weights in STRING v.9 [6]. It contains 12233 genes and 164034 edges; it integrates evidence types including, experimental expression and literature mining approaches to derive a globally weighted network of gene interactions, it comprises of multiple types of gene interactions (e.g., protein-protein interactions, genetic and cocitation).

A visualization of the somatic mutation data for the subtypes extracted with the NBS tool, with events selected as described in Section 2.3 is shown in Figure 2.3

We also performed cluster sensitivity analysis for clustering assignments inferred via NBS, selecting the cluster assignment with k=4 as a reference, in order to assess how cluster assignments varied changing the number of clusters (i.e. parameter k). Results of this analysis are shown in Figure 2.2. For NBS-inferred clustering we selected as reference assignment the one with 4 cluster, in accordance with the number of clusters inferred in the TCGA study.

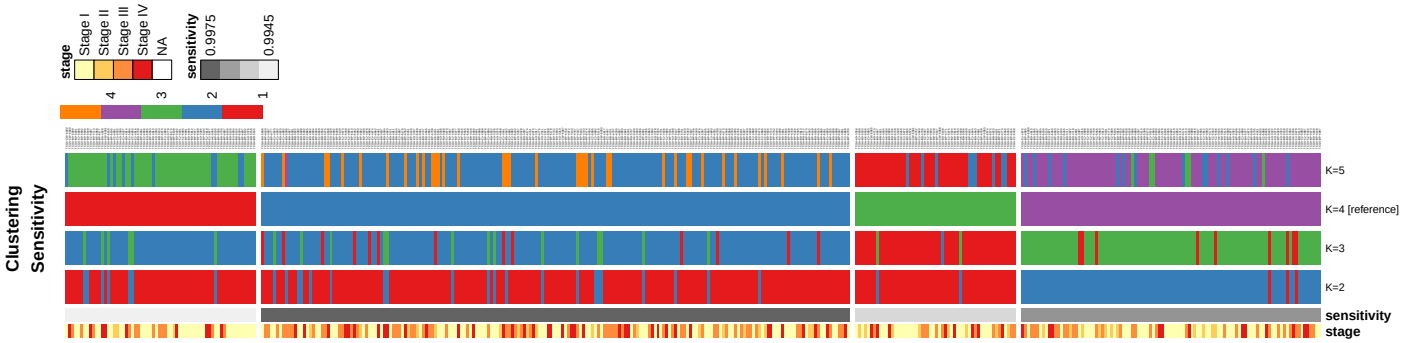


Figure 2.2: For the subtype extracted with the NBS tool we performed sensitivity analysis, selecing the cluster assignments with *k=4* clusters as a reference, in order to assess how cluster assignments varied changing the number of clusters.

2.3 Driver events selection

As previously stated, we reduced the set of CCRCC driver events to a smaller set of genes and arms - which harbour, respectively, somatic mutations and CNAs - prior to performing further analysis. We performed this selection because not all mutations and CNAs are relevant for the progression of cancer, and is thus important to select only relevant genes in order to augment the ability to infer progression models.

Various techniques can be used to select driver genes/events for cancer progression: recurrent mutations/CNAs or, more sophisticatedly, mutations in genes predicted to have a functional role. For this last category of predicted drivers a plethora of tools is available in the literature [7, 8].

For consistency with the TCGA study, in what follows we selected 50 genes that were predicted to be relevant by the TCGA consortium via MutSig [7] analysis, also for all these genes we selected the corresponding arms. In addition to this set of events we also selected a set of arms that were explicitly denoted as relevant in both the TCGA study and [3], also we selected all genes appearing in exclusivity groups identified in the TCGA study.

Not all of these genes were explicitly mapped to a pathway, so we manually annotated this information using both explicit pathway annotations and information about frequently mutated subnetworks (as obtained in the TCGA study via HotNet analysis). Follows a list of all selected genes and arms:

Chromatin Remodeling			
1	PBRM1	**4**	BAP1
2	SETD2	**5**	ARID1A
3	KDM5C		

DNA Damage	
1	TP53
2	ATM
3	CDKN2A

VHL	
1	VHL
2	TCEB1

PI3K/AKT/MTOR					
1	AKT1	**4**	EGFR	**7**	PTEN
2	AKT2	**5**	MTOR	**8**	TSC1
3	AKT3	**6**	PIK3CA	**9**	TSC2

10	GNB2L1								
11	SQSTM2								
12	RHEB								

Others									
1	MAPK9	**10**	SLC27A6	**19**	GPM6A	**28**	DIO2	**37**	TSPAN19
2	MSR1	**11**	COL6A6	**20**	MS4A12	**29**	SFXN4	**38**	DST
3	TXNIP	**12**	SPRED1	**21**	RO2L8	**30**	EMR3		
4	NFE2L2	**13**	FBN2	**22**	ZFPM2	**31**	HOXC8		
5	BTNL3	**14**	STAG2	**23**	NKAIN3	**32**	ATF7IP2		
6	SLITRK6	**15**	SECISBP2L	**24**	PGLYRP3	**33**	SCARB2		
7	NPNT	**16**	TFDP2	**25**	OR10AG1	**34**	PCNA		
8	CCNB2	**17**	HMCN1	**26**	KIAA0174	**35**	SLC17A6		
9	ZNF800	**18**	MAGEC1	**27**	FAM5B	**36**	MS4A3		

Arms									
1	14Q32.33	**14**	7Q36.1	**27**	7Q31.33	**40**	1Q21.3	**53**	12Q21.31
2	19Q13.2	**15**	3P25.3	**28**	5Q23.3	**41**	11Q11	**54**	6P12.1
3	1Q44	**16**	3P21.1	**29**	3Q22.1	**42**	16Q22.3	**55**	1Q25.1
4	7P11.2	**17**	3P21.31	**30**	15Q14	**43**	1Q25.2	**56**	2Q14.3
5	1P36.22	**18**	XP11.22	**31**	XQ25	**44**	14Q31.1	**57**	7Q22.3
6	3Q26.32	**19**	8P22	**32**	15Q21.1	**45**	10Q26.11	**58**	8Q24.21
7	10Q23.31	**20**	1Q21.1	**33**	3Q23	**46**	19P13.12	**59**	12P11.21
8	9Q34.13	**21**	8Q21.11	**34**	1Q25.3	**47**	12Q13.13	**60**	20Q13.33
9	16P13.3	**22**	2Q31.2	**35**	XQ27.2	**48**	16P13.13	**61**	4Q34.3
10	5Q35.3	**23**	13Q31.1	**36**	4Q34.2	**49**	4Q21.1	**62**	6Q22.33
11	17P13.1	**24**	1P36.11	**37**	11Q12.2	**50**	20P12.3	**63**	8P23.2
12	11Q22.3	**25**	4Q24	**38**	8Q23.1	**51**	11P14.3	**64**	3P26.1
13	9P21.3	**26**	15Q22.2	**39**	8Q12.3	**52**	11Q12.1	**65**	5Q35.2

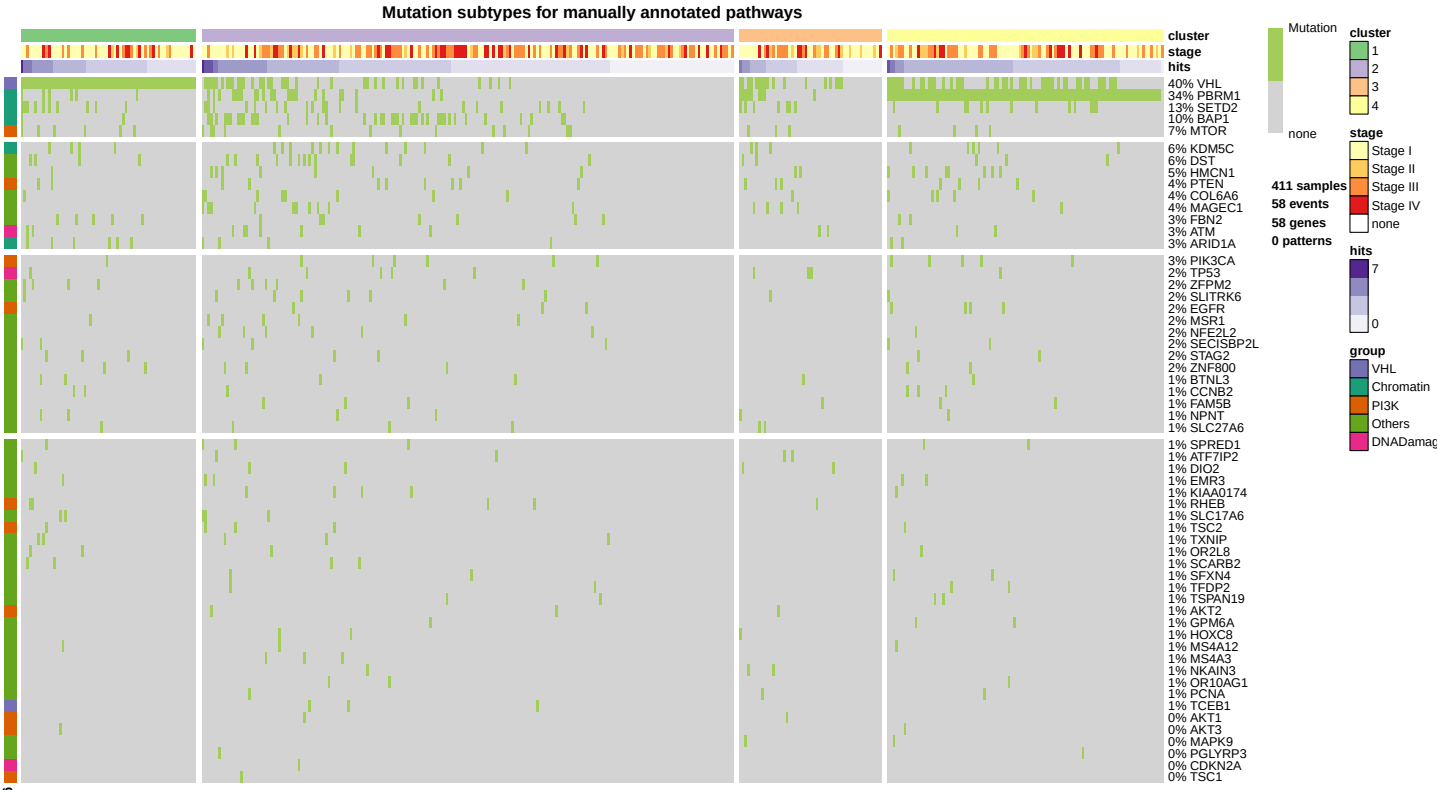


Figure 2.3: We used the NBS tool to perform somatic mutation-based clustering and we identified 4 subtypes. We then performed a selection of event, in order to retain only those genes and arms that are relevant for the progression of CCRCC. In the Figure we represented somatic mutation events, selected as described in Section 2.3, for subtypes extracted with NBS. Somatic mutation data, obtained as described in Chapter 1, was given as input to the NBS tool which extracted 4 subgroups, we then added also CNA events to these subgroups. Mutational profiles of subtypes are represented as a heatmap in which a colored cell denotes a mutated gene for a certain patient (with different colors identifying different types of alterations, as shown in the legend). Each gene is assigned to a pathway/functional group, as defined in section 2.3.

Chapter 3

Patterns selection for CAPRI's hypotheses generation

To exploit CAPRI's ability to infer complex selective advantage relations - via hypotheses testing in its supervised mode - we looked for patterns of *soft/hard mutually exclusive* alterations functional to cancer progression[1].

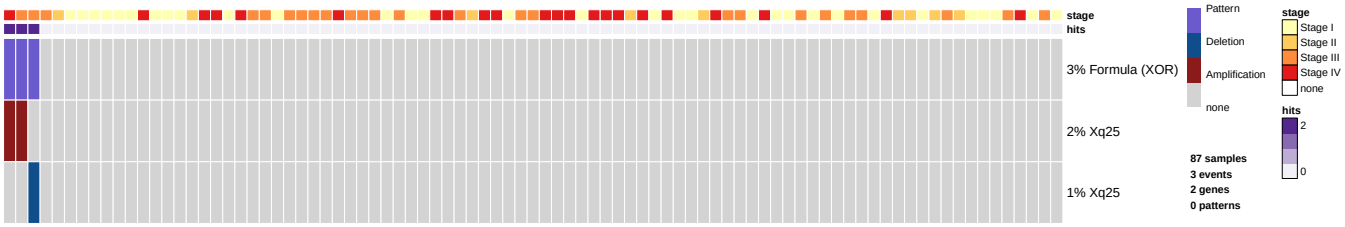
Patterns were either assessed in a computational fashion or by fetching the literature. It is to note that, by inputting a pattern to CAPRI, a selective advantage relation for the pattern to be inferred is not forced, rather the pattern is tested and competes with all other relations for the inference of a model maximizing data likelihood.

3.1 Preamble

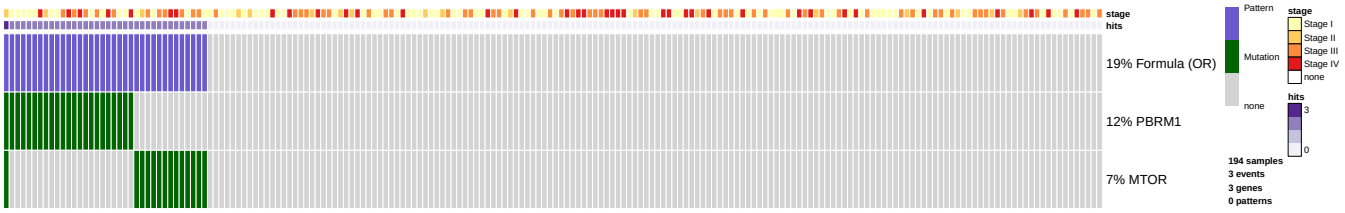
A pattern is a formula over the somatic mutations and CNAs which we include to infer a progression model. For instance, MTOR*:mutation* $\vee$ RHEB*:mutation* is a soft exclusivity pattern selecting samples with mutations in MTOR *and* germline RHEB, together with samples with mutation in RHEB *and* germline MTOR, together with samples with mutations in *both* MTOR and RHEB. In order to exclude the latter the pattern should be written as an *hard exclusivity* pattern, using $\oplus$ connective.

In general, a pattern suggests an "observational trend" which captures a certain relation between a group of events, e.g., a strong form of exclusivity

[1]We adopt the terminology introduced in [9]. Hard mutual exclusivity refers to *strictly mutually exclusive* (in testing the null hypothesis being that overlaps between them can be explained by random errors); this will be in what follows denoted via the *logical exclusivity operator* $\oplus$. Soft mutual exclusivity weakens the overlap constraint for independent events overlapping less than expected by chance because of a statistical interaction. This will be denoted with the *logical disjunction operator* $\vee$.



(a)



(b)

Figure 3.1: (a) Example of hard-exclusivity pattern, as can be seen there is no overlap for the alteration of XQ25, the corresponding formula is XQ25:*amplification* $\oplus$ XQ25:*deletion*. (b) Example of soft-exclusivity pattern, as can be seen there is a patient harboring a mutation of both PBRM1 and MTOR, the corresponding formula is PBRM1:*mutation* $\vee$ MTOR:*mutation*.

possibly subsuming *functional equivalence* and equivalent selective advantage. For instance, one might infer an evolutionary trajectory with VHL mutations selecting for the above pattern and RHEB, in turn, selecting for TP53 mutations. This would suggest the following *clonal-expansion interpretation*: VHL-mutated clones will enjoy a clonal expansion and selection and progress by acquiring mutations in MTOR or RHEB, or both. In turn, clones with RHEB mutated will be selected by acquiring a TP53 mutations.

A pattern - usually in exclusivity or co-occurrence form - can be either fetched by scanning the literature or predicted by computational techniques. In the latter case purely categorical tests such as Fisher can be used, possibly in combination with biological priors [9, 10]. In this study we use exclusivity patterns predicted by constraining statistical approaches with biological information[2]. The computational techniques that we used predict gene-level patterns, e.g., exclusivity between MTOR and RHEB alterations. For this reason we added sub-patterns of: (*i*) soft exclusivity between genes and their

[2]This is an attempt to diminish the rate of false positives/negatives for statistical tests such as Fisher which might be biased by sparse input alteration profiles. Our strategy is more conservative, it might capture less novel relations, but should diminish the amount of false selective advantage relations that we infer.

respective arm, (*ii*) hard exclusivity between different alterations (across different samples) for each given arm.

Examples of pattern of soft and hard mutual exclusivity patterns are shown in Figure 3.1.

3.2 MUTEX groups (computational priors)

We used the MUTEX tool [9] to identify mutual exclusivity groups of gene-level alterations.

3.2.1 Overview of MUTEX

MUTEX is a method for the identification of sets of mutually exclusive gene alterations in a set of genomic profiles. The strategy implemented by the tool combines detailed prior pathway information, via a signaling network, with a statistical metric in order to assign a score to groups of mutated genes exhibiting a mutual exclusivity pattern and also validate the results. MUTEX uses both somatic mutation and CNA data. The signaling network is used to search for groups of mutually exclusive genes which share a common downstream effect, this search is initialized by setting an altered gene as the seed of the group and then greedily expanding the group with the next best candidate gene (ie. the gene that best improves the group score). MUTEX introduces also a novel statistical metric to measure mutual exclusion of a group of genes, by testing each gene against the union of all other alterations in the group (correcting for multiple hypothesis testing). Thus by using a biological prior (the signalling network) the tool restricts statistical testing to identify biologically relevant groups of alterations. Currently , MUTEX is the state-of-the art solution for the identification of exclusivity patterns from somatic mutations and CNAs.

3.2.2 Patterns selection setting

MUTEX was run on every subtype, with the alterations of the pathway genes given as input (running time: approximately 3 hours for each set of clusters) and the following parameters (suggested by the authors as default):

Parameter	Value	Description
`signalling-network`	-	*MUTEX network*[†]
`max-group-size`	5	*maximum size of a result group*
`first-level-random-iteration`	10000	*number of randomisation to estimate null distribution of member p-values in mutex groups*
`second-level-random-iteration`	100	*number of runs to estimate the null distribution of final scores*
`fdr-cutoff`	-	*false-discovery-rate cutoff maximising the expected value of true positives - false positives is estimated from data*
`search-on-signaling-network`	`TRUE`	*reduce the search space using the signalling network*

[†] Manually curated from Pathway Commons, SPIKE and SignaLink databases. Provided with the tool.

The tool returns, for each group, a *score derived from p-values corrected for false discovery rate*; we selected only groups with score < 0.2. In the majority of subtypes no group was selected, in particular we selected one group for subtype n2, one group for subtype mi2 and two groups for subtype m3. Follows the list of selected group scores:

Subtype	Group	Score	q-val
m3	TP53, VHL, CDKN2A	0.0471	0.015
m3	CDKN2A, PBRM1	0.15109	0.01275
mi2	CDKN2A, PBRM1	0.1441	0.12
n2	PBRM1, MAPK9, TP53, MTOR, TCEB1	0.1535	0.34

The added groups, as selected by MUTEX (thus, with focal-level somatic mutation and CNAs) are shown in Figure 3.3.

3.3 MEMO groups (computational priors)

The TCGA consortium ran the MEMO tool (*Mutual Exclusivity Modules in cancer*), on the whole dataset, to extract groups of mutually exclusive alterations from the whole cohort. MEMO [10] searches for functional modules (ie. groups) whose member genes are: (*i*) recurrently altered across a set of samples, (*ii*) known to or likely to participate in the same biological process (pathway-level mapping) and show a trend towards mutual exclusivity. The tool integrates multiple data types, maps genomic alterations to pathways and uses a statistical model that preserves the number of alterations per gene and per sample. In particular they identified eight groups, mainly related to genes in the PI3K/AKT/MTOR and DNA Damage pathways, that were not

specified to be soft or hard exclusive; we created soft exclusivity patterns for each of these groups:

	MEMO groups
1	AKT1, AKT2, AKT3, EGFR, MTOR, PIK3CA, PTEN
2	AKT1, AKT2, AKT3, MTOR, PIK3CA, PTEN, TSC1, TSC2
3	AKT1, AKT2, AKT3, EGFR, GNB2L1, PIK3CA
4	AKT1, AKT2, AKT3, MAPK9, MTOR
5	AKT1, AKT2, AKT3, EGFR, TP53, PTEN
6	ATM, CDKN2A, TP53
7	AKT1, AKT2, AKT3, PIK3CA, SQSTM2
8	AKT1, AKT2, AKT3, MTOR, RHEB, TSC1, TSC2

Genes involved in MEMO groups, for mutation subtype `n2`, are shown in Figure 3.4.

3.4 Arm-level groups and genes with both mutations and arm-level CNAs

We observed that, in the datasets, some arms harboured different events (on different samples). For this reason we added hard-exclusivity patterns accounting for all such events, this is because these events obviously showed mutual exclusivity. An example for this kind of pattern, included in mutation subtype `n2`, is shown in panel `(a)` of Figure 3.2. For this subtype we added pattern:

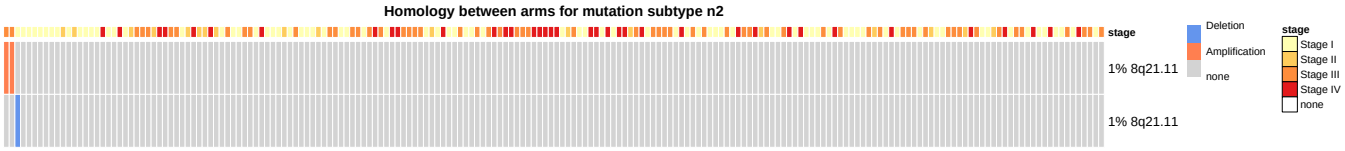
$$8\text{Q}21.11\textit{:deletion} \oplus 8\text{Q}21.11\textit{:amplification}$$

Also, in order to account for mutual exclusivity between somatic mutations of genes and CNAs on the respective arms, we added soft-exclusivity patterns between genes and respective arms. An example for this kind of pattern is shown in panel `(b)` of Figure 3.2. For this subtype we added pattern:

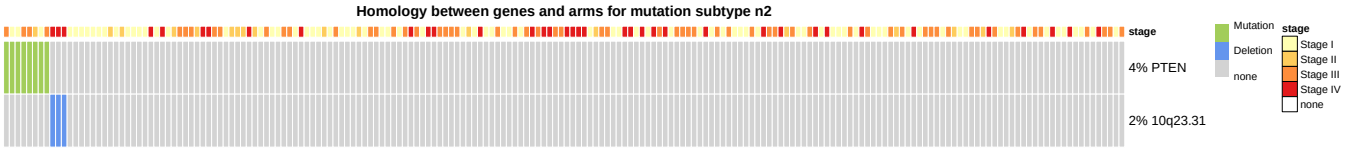
$$\text{PTEN}\textit{:mutation} \vee 10\text{Q}23.31\textit{:deletion}$$

In every subtype, patterns are instantied by including only genes and arms with at least an alteration in the set of considered samples. An example of a complete pattern, with both somatic mutations and CNAs, is shown in panel `(c)` of Figure 3.2. For this subtype we included pattern:

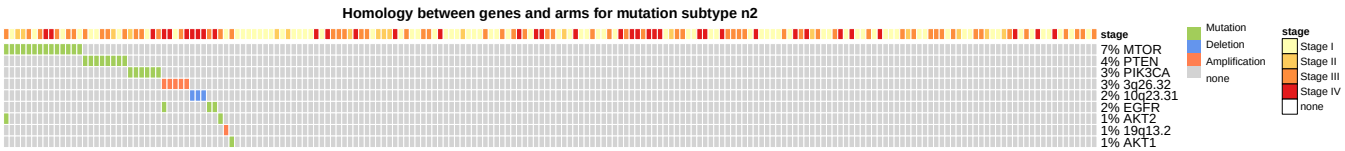
$$\text{MTOR}\textit{:mutation} \vee \text{PTEN}\textit{:mutation} \vee \text{PIK3CA}\textit{:mutation} \vee \text{EGFR}\textit{:mutation} \vee \text{AKT2}\textit{:mutation} \vee \text{AKT}\textit{:mutation} \vee 3\text{Q}26.32\textit{:amplification} \vee 10\text{Q}23.31\textit{:deletion} \vee 19\text{Q}13.2\text{em:amplification} \vee 1\text{Q}44\textit{:amplification}$$



(a)

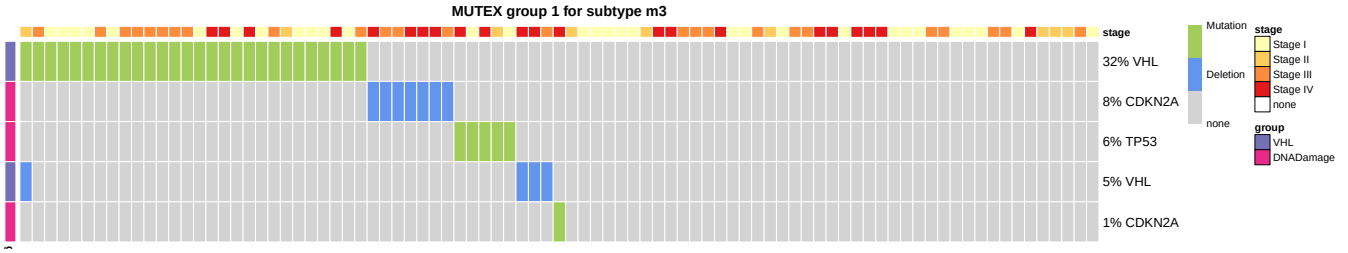


(b)

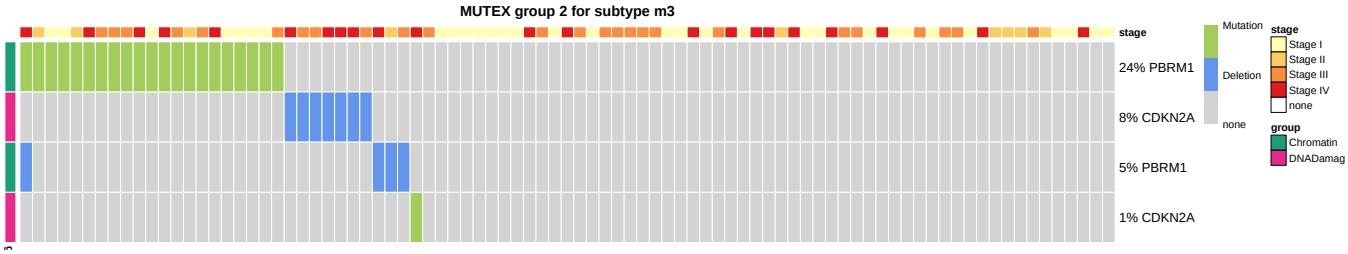


(c)

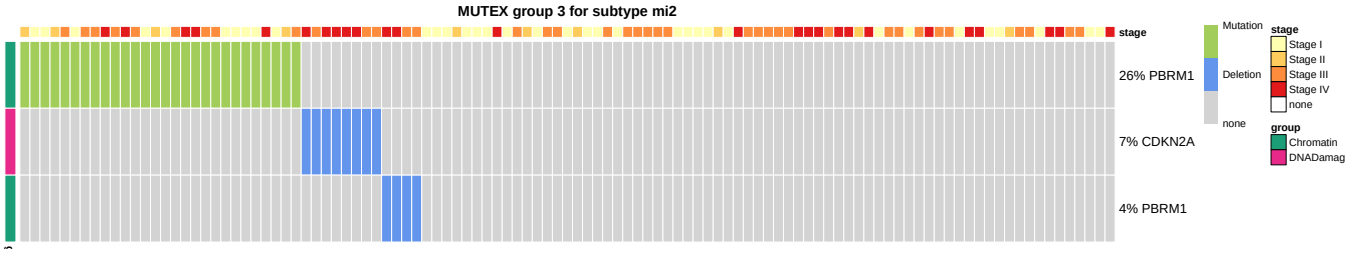
Figure 3.2: `(a)` Example of hard-exclusivity pattern between multiple events of a single arms, included for mutation subtype `n2`. `(b)` Example of soft-exclusivity pattern between events of a gene and events of the respective arm, included for mutation subtype `n2`. `(c)` Example of a complete pattern, including both focal-level somatic mutations and arm-level CNAs, included for mutation subtype `n2`.



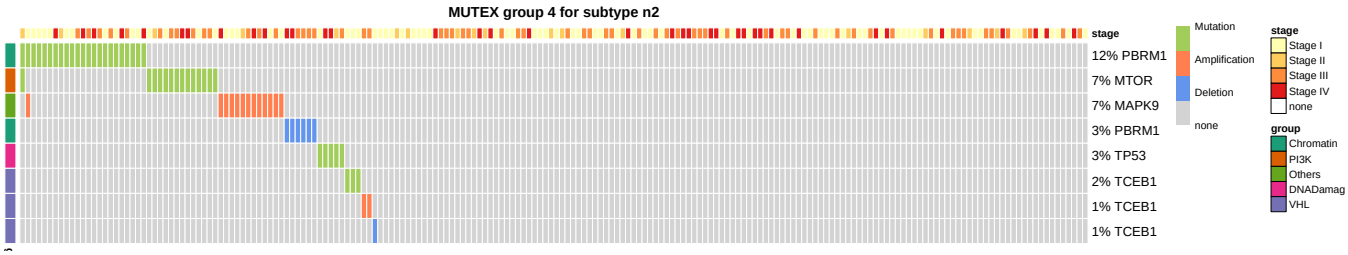
(a)



(b)



(c)



(d)

Figure 3.3: Oncoprint of mutual exclusivity group, as inferred with the MUTEX [9] tool (thus with focal-level somatic mutations and CNAs). For each of these groups we then remapped focal-level CNAs to arm-level CNAs and created patterns to be included for furthere analyses.

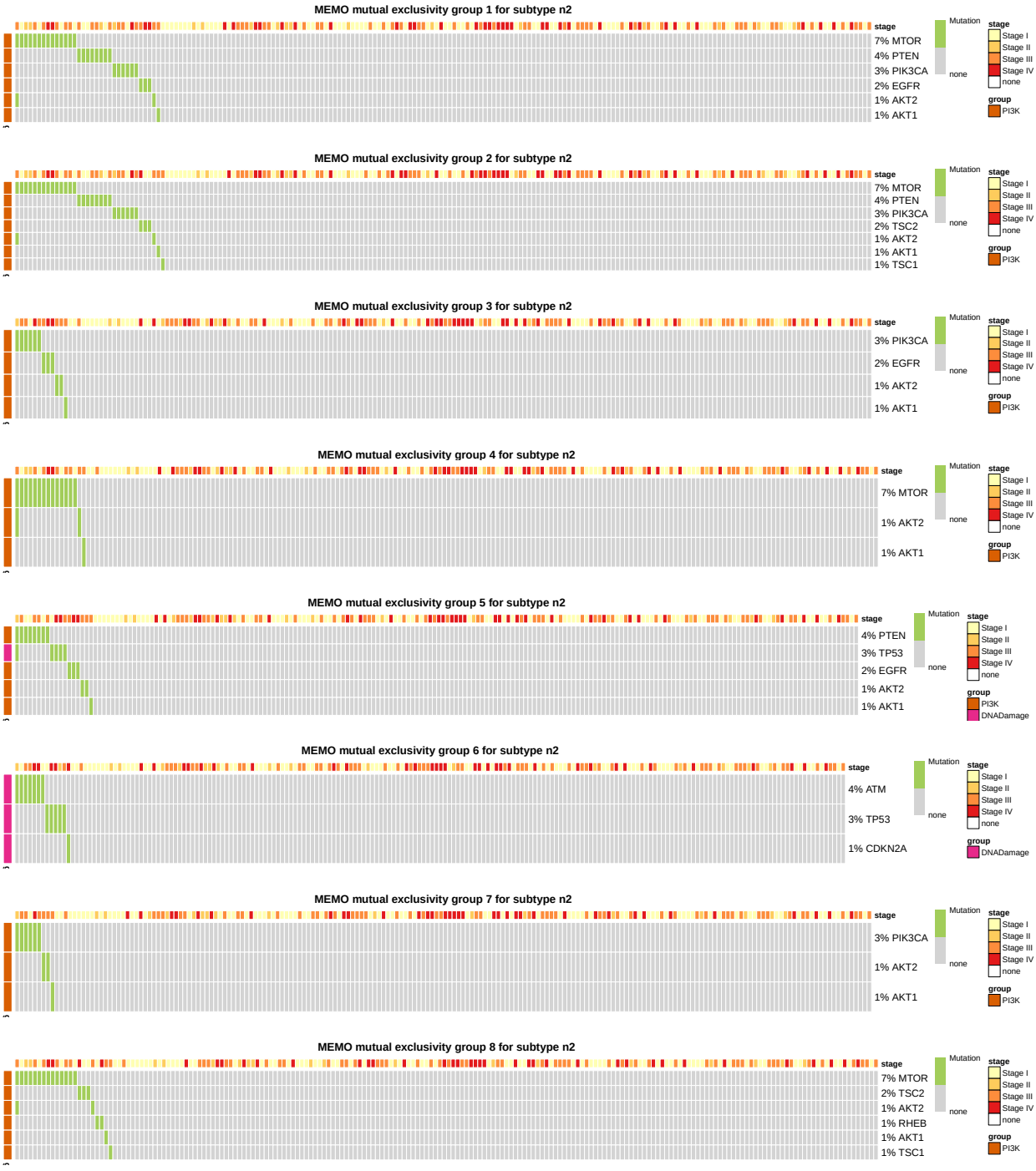


Figure 3.4: Oncoprint of mutual exclusivity groups, computed with MEMO in the TCGA consortium study, for the mutation-based subtype n2. As can be seen, all genes involved in mutual exclusivity groups are related to the PI3K/AKT/MTOR pathway or the DNADamage functional group (as defined in section 2.3). Genes involved in these groups are mutated with relatively low frequency but exhibit, evidently, hard exclusivity. As an example, for group 7 we added pattern PIK3CA*:mutation* $\vee$ AKT1*:mutation* $\vee$ AKT2*:mutation*

Chapter 4

Reconstructed models

4.1 Events selection and statistically indistinguishable events

For each subtype we ran the *Cancer Progression Inference* (CAPRI, [1]) algorithm with pathway events selected according to the following criterion:

(*i*) we selected Mutation events for genes with alteration frequency (i.e., sum of all events frequency) greater than 5%, for all such genes we also selected all events for respective arms;

(*ii*) we selected Mutation events for any gene part of a MEMO/MUTEX or biological pattern, regardless its frequency, for all such genes we also selected all events for respective arms;

(*iii*) we selected all events for arms in the list of relevant arms having alteration frequency greater than 5%.

First condition imposes a minimal cutoff to the genes with alterations likely to be relevant, the second one mimics the fact that the frequency of a functional pattern is more important than the individual frequency of its constituting events, the third one imposes a minimal cutoff to the arms with alterations likely to be relevant.

An example of the result of this selection, for somatic mutation-based subtype `n2` is shown in Figure 4.1. Even though the current CAPRI version does not require previous deletion, or merging, of indistinguishable events (in terms of their respective occurrences, ie. they occur exactly in the same samples), thus consolidating the respective subtype, for 3 subtypes we decided to merge into a single event groups of indistinguishable events.

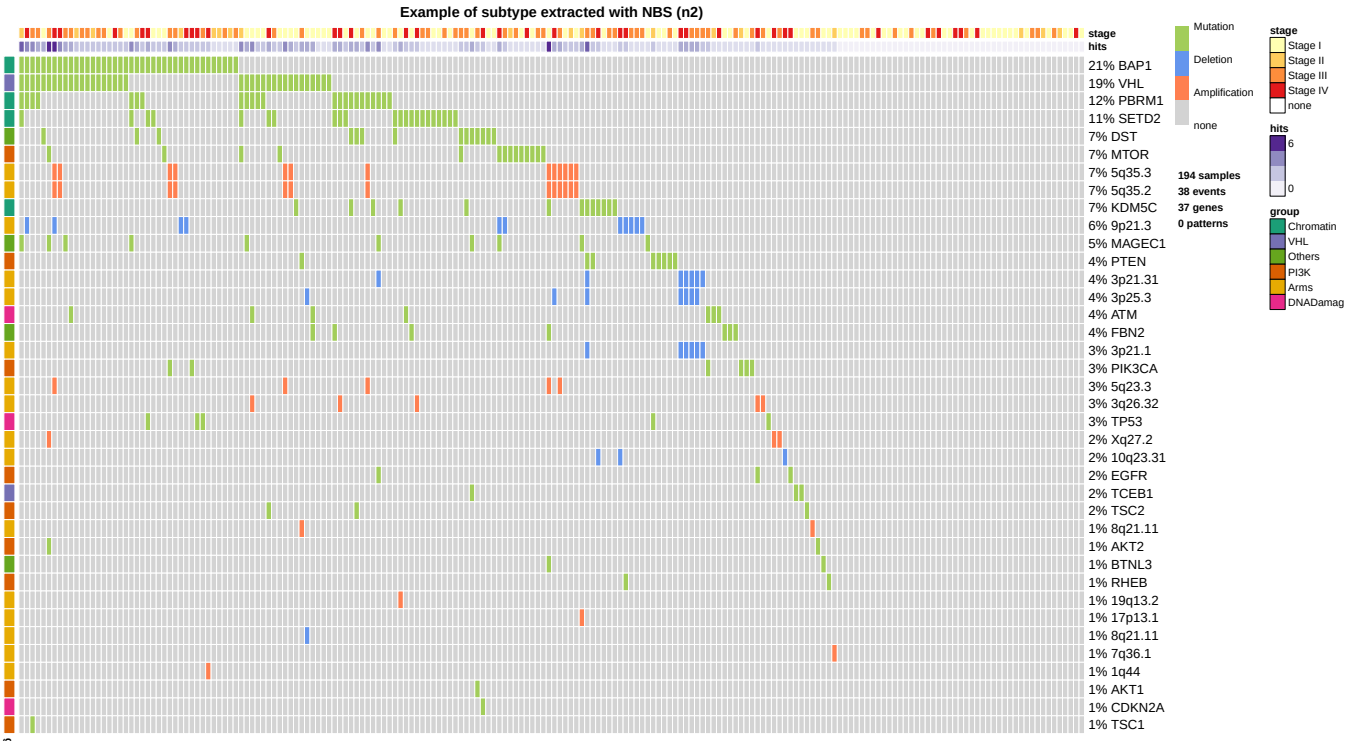


Figure 4.1: Oncoprint of Subtype 2 obtained with mutation-based clustering using the NBS [5] tool after performing Events selection. The selection was performed according to the following criterion: (*i*) we selected Mutation events for genes with alteration frequency greater than 5%, for all such genes we also selected all events for respective arms, (*ii*) we selected Mutation events for any gene part of a MEMO/MUTEX or biological pattern, regardless its frequency, for all such genes we also selected all events for respective arms, (*iii*) we selected all events for arms in the list of relevant arms having alteration frequency greater than 5%.

We performed this preprocessing in order to reduce complexity of the inferred models, in particular because in the case of statistically indistinguishable events, these events are selected *randomly* in the likelihood fit step of the reconstruction algorithm (i.e. since there is no theoretical way to choose one event over another, which events is chosen depends on how the likelihood fit algorithm explores the search space). In these cases we deleted all events for the involved genes or arms from the dataset and created a merged event with type MERGED. This process of merging events is supervised:

- mRNA expression subtypes: In subtype m3 we merged events 3P21.1*:deletion*, 3P25.3*:deletion* and 3P21.31*:deletion*, and created event 3P21.1-/3P25.3-/3P21.31-*:merged* (the minuses in the name of the merged event are inserted in order to account for the original

events types, in this case deletion). The indistinguishability of the deleted events could be explained as a loss affecting a large portion of the 3p arm. For this subtype we manually added hypothesis between the created merged event and genes mapped on the deleted arms;

- microRNA expression subtypes: In subtype mi1 we merged events 3p26.1*:deletion*, 3p25.3*:deletion* and 3p21.31*:deletion*, and created event 3p26.1-/3p25.3-/3p21.31-*:merged*;

- Mutation subtypes: In subtype n3 we merged events 9p21.3*:deletion*, 1p36.22*: deletion*,10q23.31*:deletion* and 11q22.3*:deletion*, and created event 9p21.3-/1p36.22-/10q23.31-/11q22.3-*:merged*.

4.2 Reconstruction setting

CAPRI algorithm was executed with the following parameters:

Parameter	**Value**	**Description**
`nboot`	`100`	*bootstrap iterations for Wilcoxon testing of selective advantage scores (temporal priority and probability raising)*
`regularization`	`aic`, `bic`†	*regularization techniques for likelihood fit*

†`aic`: Akaike Information Criterion, `bic`: Bayesian Information Criterion

The first parameter determines the number of times the input dataset is bootstrapped to estimate, via Wilcoxon testing, p-values for the following inequalities

condition†	**formula**‡	**interpretation**
temporal priority	$p_i > p_j$	*event i is earlier than j*
probability raising	$p_{j\,\vert\,i} > p_{j\,\vert\,\bar{i}}$	*event i selects for j*

† j can be either an event (gene mutation or CNA) or a pattern, i just an event.

‡ p_i (resp. p_j) is the probability of event i (resp. j), $p_{j\,\vert\,i} = p_{j,i}/p_i$ is the conditional probability of j given i, $p_{j,i}$ is the joint and $p_{\bar{i}} = 1 - p_i$. All the probabilities are estimated from input data.

Bootstrap iterations are repeated until `nboot` observations for each marginal and joint probability value are available. When these inequalities hold, CAPRI considers this as a potential selective advantage relation among i and j, termed *prima facie*. To select only those relations maximising the likelihood of data, given the model, CAPRI runs a likelihood-fit algorithm with these relations as prior constraints.

The second parameter specifies which regularization technique is used to achieve a model with a minimal set of prima facie relations. We used two distinct score-based approaches to *penalise complex models*, i.e. models with many relations, and favour smaller ones. The two penalty scores, where x is input data (i.e., the alteration signatures) and R is the number of prima facie relations in the model have general form $\ell(x)$

$$\ell(x) = -2\log(\mathcal{L}(x)) + \theta R\,. \tag{4.1}$$

Akaike Information Criterion (AIC, `aic`) is when $\theta = 2$, *Bayesian Information Criterion* (BIC, `bic`) when $\theta = \log(n)$ (n sample size) thus imposing a stronger penalty term; both scores are approximately correct according to a different goal and a different set of asymptotic assumptions. These scores have generally two different aims: AIC being more prone to overfitting errors, but likely to provide also good future predictions from data, BIC being more prone to underfitting errors but providing a parsimonious model. Thus, AIC is better in situations when a false negative finding would be considered more misleading than a false positive, and BIC is better in situations where a false positive is as misleading as, or more misleading than, a false negative. For this reason, we used - as is commonly done - both approaches in model selection, distinguishing which relations are selected by BIC and which by AIC.

4.3 Graphical interpretation of models

Inferred models are pictured as acyclic graphs, with node size proportional to event frequency. Edges in grey represent relations inferred with AIC regularization, black one those inferred with BIC regularization - which are a strict subset of those inferred with AIC. Two simple models (reconstructed, respectively, from mRNA expression subtypes m1 and m2) are shown in 4.3 and 4.4.

Nodes are circled with a color univocally determining their pathway. We graphically distinguish when we infer that an event selects for a pattern, and viceversa; and we color with red hard exclusivity patterns (not shown here), and orange soft ones. In the former case we picture the pattern symbol with a squared box connecting the pattern elements, in the latter we use a circular notation. Somatic mutation events are represented by green-colored nodes, amplification events are represented by orange-colored nodes and deletion events are represented by blue-colored nodes. An example of a soft-exclusivity relation is the selection of pattern TP53:*:mutation* $\vee$ ATM:*:mutation* by BAP1:*:mutation*, which should be interpreted as: "BAP1

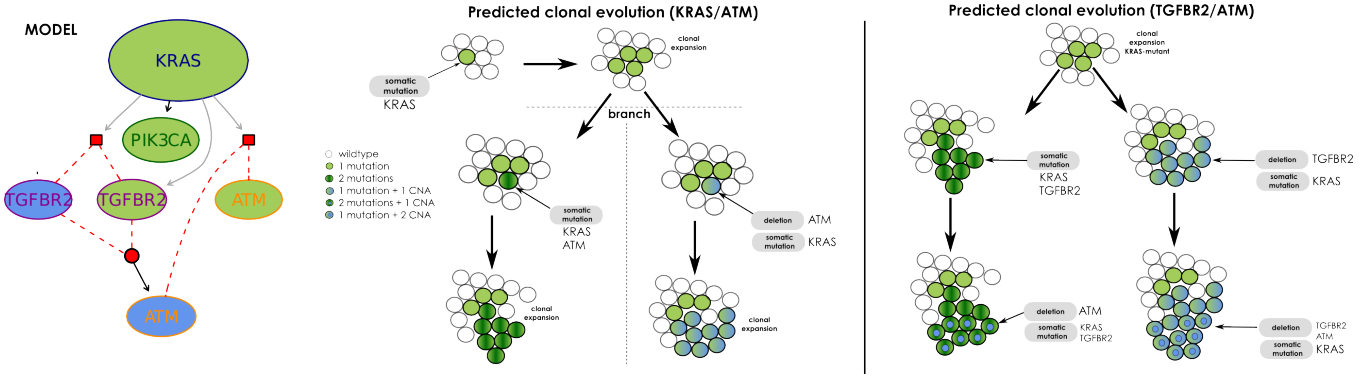


Figure 4.2: Example of progression involving mutual exclusivity patterns, in particular hard-exclusivity patterns. Shown in the left panel is a simple model picturing relations among 4 genes. Shown in the center panel an explanation of a "branching" progression (i.e. progression involving a relation of the type gene to pattern): KRAS is predicted to be the early event, KRAS mutated clones are subsequently selected for either (but bot both) a mutation or a deletion of ATM. Shown in the right panel an example of a "branching" progression followed by a "confluency" (i.e. progression involving a relation of the type pattern to gene): KRAS is predicted to be the early event, KRAS mutated clones are subsequently selected for either (but not both) a mutation or a deletion of TGFBR2, subsequently both clones with a mutation in TGFBR2 or a deletion of TGFBR2 will be selected for deletion of gene ATM. These model is taken from a progression inference study for Colorectal Cancer.

mutant clones select for a clone harbouring one, or both, of mutations in ATM or in TP53". A simple model, with an explanation of the meaning of the two types of relations involving patterns (i.e. pattern to gene and gene to pattern) is shown in Figure 4.2

Annotated on edges are represented confidence levels representing the confidences of such edges, these confidence levels are colored red when at least one of the three scores is over 0.01, colored blue otherwise.

4.4 Confidence estimation

For each reconstructed model we also performed confidence estimation via bootstrap. In particular for each reconstructed model we performed two types of bootstrap, namely *non-parametric bootstrap* and *statistical bootstrap*.

Non-parametric bootstrap is a bootstrapping procedure which performs confidence estimation by sampling with replacement of the dataset, assuming an uniform distribution of the samples.

Statistical bootstrap is a confidence estimation procedure which estimates how CAPRI is sensitive to seed choice. In the reconstruction of progression

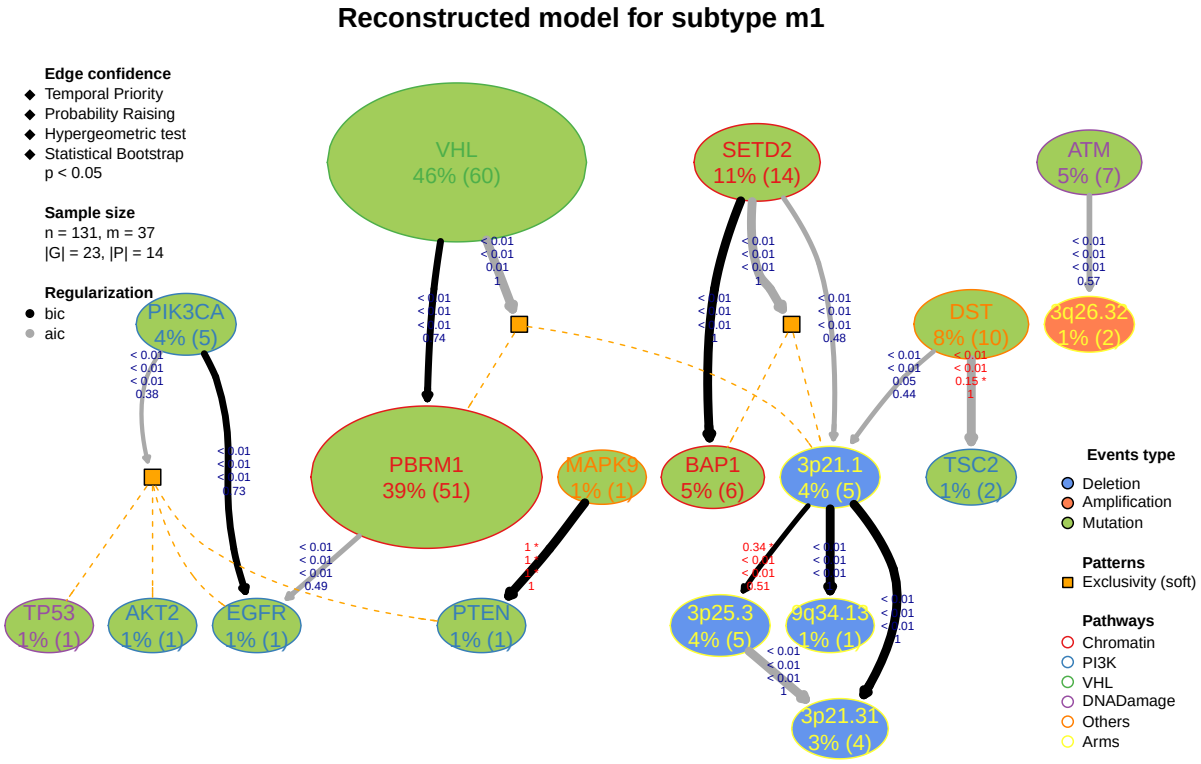


Figure 4.3: Reconstructed model for mRNA expression subtype `m1`.

models, random numbers based on a given initial seed are generated to evaluate the prima facie conditions; the statistical bootstrap technique assess the sensitivity of the reconstructed models in relation to the choice of the initial seed choice given a fixed input dataset.

This analysis was performed using libraries in the TRONCO tool with these parameters:

Parameter	Value	Description
`nboot`	100	*number of bootstrap iterations)*
`confidence`	`npb`, `sb`†	*bootstrap techniques*

†`npb`: Non-Parametric Bootstrap, `sb`: Statistical Bootstrap

Graphically, confidence is represented by thickness of edges, in particular more confident relations are represented by thicker edges.

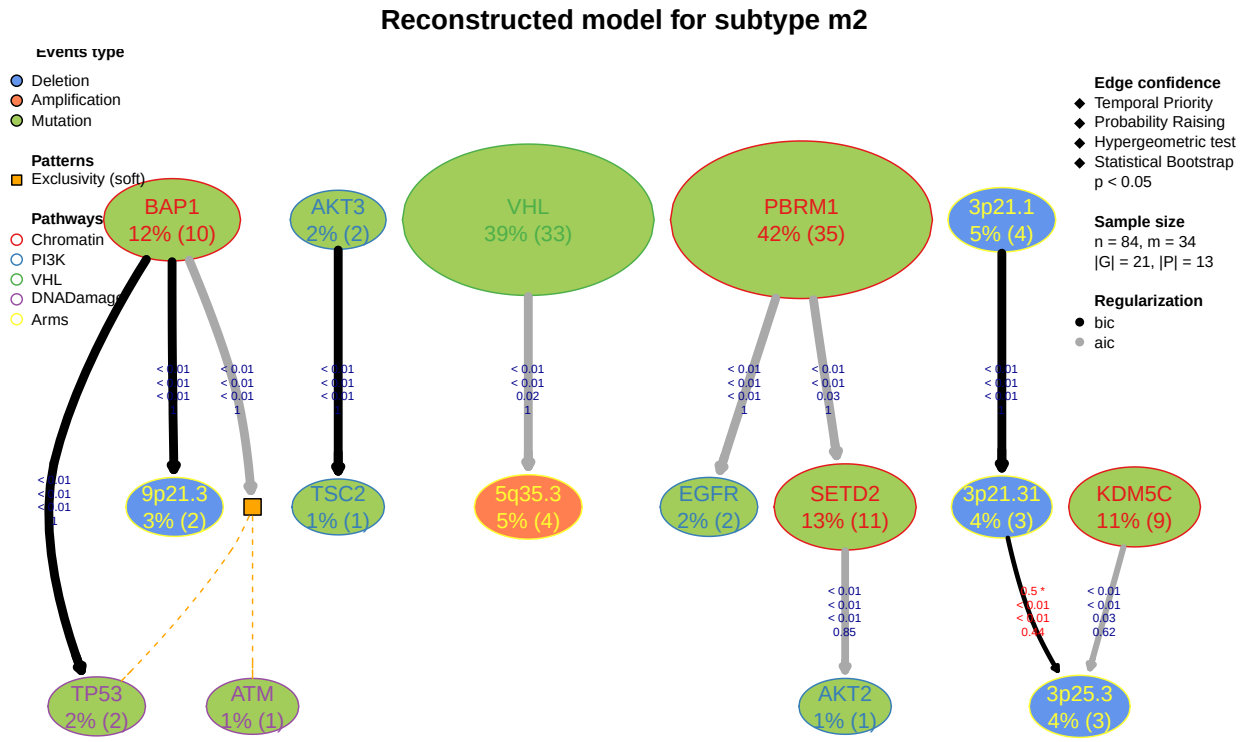


Figure 4.4: Reconstructed model for mRNA expression subtype `m2`. Mutational events are represented by nodes (with different colors denoting different types of alterations), encircled with a color denoting the pathway/functional subgroup. Soft-exclusivity patterns are represented with an orange symbol (box or circle, depending the type of selection: event for pattern or pattern for events). The model could be interpreted to predict that: (*i*) no general early event could be inferred, however 5 different subprogression were identified having as first event, respectively, BAP1, AKT3, VHL, PBRM1 and 3P21.1(*ii*) BAP1 mutant clones are selected to acquire 9P21.3 deletions or *any combination of* TP53 mutations or ATM mutations, (*iii*) AKT3 altered clones will select for a TSC2 mutation, (*iii*) VHL mutant clones will select for amplification of 5Q35.3, (*iv*) PBRM1 altered clones will select for EGFR mutations or SETD2 mutations, the latter will in turn select for acquisition of AKT2 mutations, (*v*) clones with a deletion of 3P21.1 will select also for deletion of 3P21.31, in turn clones harboring both KDM5C mutations and a deletion of 3P21.31 will select for deletion of 3P25.3. This last branch is particularly interesting because it shows a sequence of events all related to arm 3P, which is considered important for CCRCC

Chapter 5

Single-patient study

5.1 Somatic mutations and structural variants

In addition to the progression inference study based on the TCGA study, as described in previous sections, we also performed a study based on multiregion sequencing data as given by [3].

For this study somatic mutation and arm-level CNAs data was available for 10 patients, and for all these patients previous analysis was performed to identify relevant events and also progression models. For each patient, data for multiple regions was available, for this reason we considered each sample as a separate dataset on which to infer progression models considering single regions as samples.

Groups of somatic mutation events that were found to be indistinguishable were manually merged into a single event.

An example of extracted dataset is shown in Figure 5.1 Since we considered each patient separately and the number of regions for patient was quite small (ie. about 10) we didn't perform clustering and subtypes extraction. We didn't perform any event selection, since relevant genes and arms were already identified in [3]. Finally we didn't perform pattern selection because of the reduced dimension of the datasets.

5.2 Reconstructed models and confidence analysis

Since all and only relevant driver genes were identified in the main text we performed no further selection, thus retaining the whole dataset.

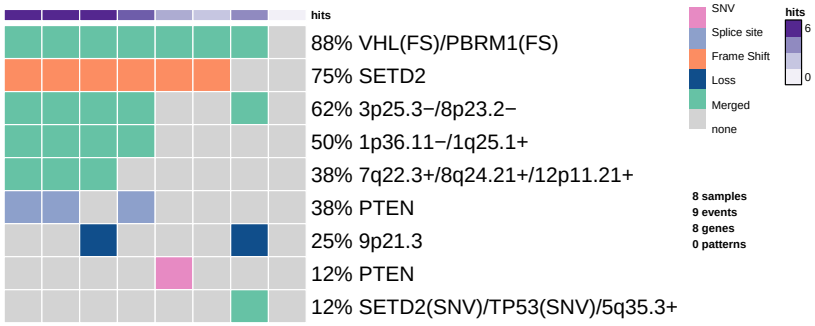


Figure 5.1: Oncoprint for patient EV002 of the Single-Patient Study conducted in [3], which analysed different regions of 10 patients. The dataset includes both somatic mutations, for example patient EV002 harbour a Frame Shift event for gene SETD2, and arm-level CNAs. Since the size of the datasets is small we observed a large number of statistically undistinguishable events that we manually merged, for example for this dataset we merged PBRM1*:Frame Shift* and VHL*:Frame Shift*.

Since a large number of genes were found to be indistinguishable, we merged into single events groups of indistinguishable events in order to reduce complexity of inferred models:

- EV001: For patient EV001 we merged 1p36.11*:Loss* and 14q31.1*:Loss* into 1p36.11-/14q31.1-*:Merged*; 3p25.3*:Loss* and VHL*:Frame Shift* into VHL(FS)/3p25.3-*:Merged*; 2q14.3*:Gain* and SETD2*:Splice site* into 2q14.3+/SETD2(SS)*:Merged*;
- EV002: For patient EV002 we merged 1p36.11*:Loss* and 1q25.1*:Gain* into 1p36.11-/1q25.1+*:Merged*; SETD2(SNV)/TP53(SNV)*:Merged* and 5q35.3*:Gain* into SETD2(SNV)/TP53(SNV)/5q35.3+*:Merged*;
- EV003 For patient EV003 we merged 3p25.3*:Loss* and 5q35.3*:Gain* into 3p25.3-/5q35.3+*:Merged*; 9p21.3*:Loss* and 1q25.1*:Gain* into 9p21.3-/1q25.1+*:Merged*;
- EV005: For patient EV005 we merged 3p25.3-/4q34.3-*:Merged* and VHL(DEL)/PBRM1(FS)*:Merged* into 3p25.3-/4q34.3-/VHL(DEL)/-PBRM1(FS)*:Merged*; 6q22.33*:Loss* and 5q35.3*:Gain* into 6q22.33-/5q35.3+*:Merged*; 14q31.1*:Loss* and SF3B1*:SNV* into 14q31.1-/SF3B1-(SNV)*:Merged*;
- EV006: For patient EV006 we merged 3p25.3-/8p23.2-/14q31.1-*:Merged* and 5q35.3*:Gain* into 3p25.3-/8p23.2-/14q31.1-/5q35.3+*:Merged*;
- RMH002: For patient RMH002 we merged 3p25.3*:Loss* and BAP1*:Frame Shift* into 3p25.3-/BAP1(FS)*:Merged*; 7q22.3+/20q13.33+*:Merged* and TP53*:Splice site* into 7q22.3+/20q13.33+/TP53(SS)*:Merged*

- RMH004: For patient RMH004 we merged 3p25.3*:Loss*, 5q35.3*:Gain* and VHL*:Frame Shift* into 3p25.3-/5q35.3+/VHL(FS)*:Merged*; 4q34.3-/8p23.2-/9p21.3-/14q31.1-*:Merged*, 1q25.1*:Gain*, 7q22.3*:Gain*, 20q13.33*:Gain* and PBRM1(FS)/ATM(SC)*:Merged* into 4q34.3-/8p23.2-/9p21.3-/14q31.1-/1q25.1+/7q22.3+/20q13.33+/PBRM1(FS)/ATM(SC)*:Merged* (in the dataset termed as LARGE-GROUP*:Merged*); 2q14.3*:Gain* and 12p11.21*:Gain* into 2q14.3+/12p11.21+*:Merged*;

- RMH008: For patient RMH008 we merged 3p25.3*:Loss* and 5q35.3*:Gain* into 3p25.3-/5q35.3+*:Merged*;

- RK26: For patient RK26 we merged PBRM1*:SNV*, 1q25.1*:Gain* and 1p36.11*:Loss* into PBRM1(SNV)/1q25.1+/1p36.11-*:Merged*; 3p25.3*:Loss* and 5q35.3*:Gain* into 3p25.3-/5q35.3+*:Merged*

We then performed reconstruction of progression models for each patient with the same parameters used for the TCGA study:

Parameter	**Value**	**Description**
`nboot`	`100`	*bootstrap iterations for Wilcoxon testing of selective advantage scores (temporal priority and probability raising)*
`regularization`	`aic`, `bic`†	*regularization techniques for likelihood fit*

†`aic`: Akaike Information Criterion, `bic`: Bayesian Information Criterion

We also performed confidence analysis via bootstrap with the same parameters used for the TCGA study:

Parameter	**Value**	**Description**
`nboot`	`100`	*number of bootstrap iterations)*
`confidence`	`npb`, `sb`†	*bootstrap techniques*

†`npb`: Non-Parametric Bootstrap, `sb`: Statistical Bootstrap

An example of reconstructed model is shown in Figure 5.2. We found that models reconstructed with our pipeline, using CAPRI, were in compliance with models predicted in the study.

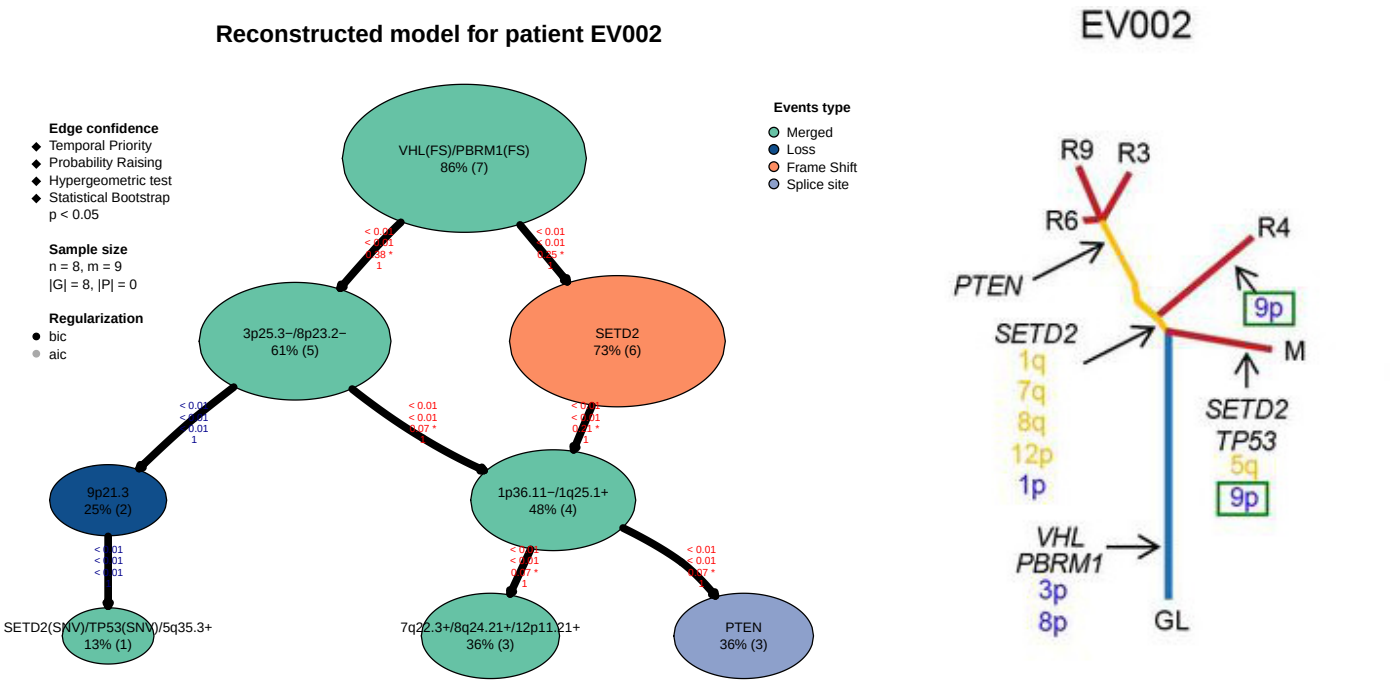


Figure 5.2: Left panel: Reconstructed model for patient EV002. The model predicts that: (*i*) VHL and/or PBRM1 (they are statistically indistinguishable) are the earliest event for the progression of cancer, (*ii*) VHL/PBRM1 (both of type Frame Shift) mutated clones will undergo selective pressure and will be selected for deletion of 3p25.3/8p23.2 and SETD2 mutations (Frame Shift), (*iii*) clones with a deletion of 3p25.3/8p23.2 will be selected also for a deletion of 9p21.3 and in turn an amplification of 5q35.3 and/or a mutation of SETD2/TP53 (both of type Single Nucleotide Variant), (*iv*) clones harboring both a mutation of SETD2 and a deletion of 3p25.3/8p23.2 will be selected for a deletion of 1p36.11 and/or an amplification of 1q25.1, which in turn will select for an amplification of 7q22.3/8q24.21/12p11.21 or a mutation of PTEN (Splice site). The reconstructed model has good compliance with the model for patient EV002, predicted in [3], shown in the Right panel.

Chapter 6

Conclusions

The advent of Next Generation Sequencing Technologies provided researchers huge amounts of data, allowing them to tackle complex biologial phenomena like cancer diseases.

In the context of the study of cancer is of fundamental importance the study of *tumorigenesis*, the process of accumulation of mutations through which cancer arises and progresses. The leading aspect to understand tumorigenesis is then to identify relevant mutations, so-called drivers, and selective advantage relationships between them in order to construct progression models able to explain the order in which cancer progresses.

Various algorithms and techniques have been proposed to tackle this progression model inference problem, among which the CAPRI algorithm, developed by the BIMIB group.

However certain characteristics of cancer hinders the ability of these techniques to reconstruct progression models, namely tumor heterogeneity and the presence of mutations irrelevant to the progression of the disease.

In this document we presented the implementation of a pipeline, designed in the context of a real case study in which we applied the pipeline to study Clear Cell Renal Cell Carcinoma, to conduct progression inference studies by integrating various pre-existing tools in order to solve the many problems intrinsic in this type of study (e.g. tumor stratification, selection of driver events, etc.). A secondary goal of this pipeline was also to provide a guideline for researchers interested in conducting similar studies.

6.1 Future works

Although our pipeline is, at the time of the writing of this document, already partially automated, a direction for future development could be an engineering of the pipeline, in order to ease researchers attempts to adapt it to their respective necessities. However, for reasons intrinsic to this type of study a completely black-box

implementation of the pipeline would be difficult to implement: different studies usually have different data availability and relevance, some steps of the study may need to be executed in a supervised fashion, etc.

Since our pipeline, as previously said, was specifically developed in the context of a study of Clear Cell Renal Cell Carcinoma, a biological expert feedback would be needed in order to validate the obtained results.

Bibliography

[1] Ramazzotti, D. et al. CAPRI: Efficient Inference of Cancer Progression Models from Cross-sectional Data. Bioinformatics (2015)

[2] The Cancer Genome Atlas Research Network Comprehensive molecular characterization of clear cell renal cell carcinoma (2013)

[3] Gerlinger, M. et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nature Genetics 46, 225233 (2014)

[4] Renaud Gaujoux, Cathal Seoighe A flexible R package for nonnegative matrix factorization. Bioinformatics. BMC Bioinformatics (2010)

[5] Hofree, M. et al. Network-based stratification of tumor mutations (2013)

[6] Szklarczyk, D. et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 39, D561?D568 (2011).

[7] Lawrence, M. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499, 214-218 (2013)

[8] Gundem G, Perez-Llamas C, Jene-Sanz A, Kedzierska A, Islam A, Deu-Pons J, Furney S and Lopez-Bigas N. IntOGen: Integration and data-mining of multidimensional oncogenomic data. Nature Methods 7, 92-93 (2010).

[9] Babur, O. et al. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations (2015)

[10] Ciriello, G. et al. Mutual exclusivity analysis identifies oncogenic network modules (2012)