Algorithmic methods to infer the evolutionary trajectories in cancer progression[*]

Supplementary Material

Giulio Caravagna Alex Graudenzi Daniele Ramazzotti Rebeca Sanz-Pamplona
Luca De Sano Giancarlo Mauri Victor Moreno Marco Antoniotti Bud Mishra

Contents

List of Tables

List of Figures

[*] This work follows up on our earlier project initiated by BM and carried out by Milan-Bicocca and the Catalan Institute of Oncology, based on a framework discussed at the 2014 Como School of Advanced Studies on Cancer, Systems and Complexity (CSAC). PiCnIc was designed and constructed by MA's Bioinformatics lab at University of Milan-Bicocca, within a project led and supervised by GC. GC, AG and DR designed the pipeline, and GC, DR and LDS coded and executed it. Data gathering and model interpretation was done by GC, LDS, DR, AG together with BM, VM and RSP. GM, MA, VM and BM provided overall organizational guidance and discussion. GC, AG, RSP and BM wrote the original draft of the paper, which all authors reviewed and revised in the final form. BM and MA are co-senior authors.

A Reproducing this study

Code availability

The implementation of PiCnIc shown in the Main Text was performed by using, as core, the R language, and other external Java tools which we reference in this document. In R, much of the data processing and inference is done by exploiting the current version of the open-source

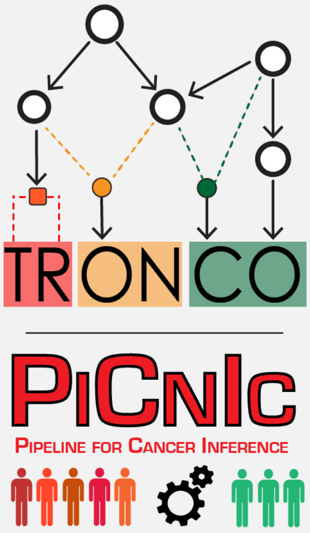
TRanslational ONCOlogy (`TRONCO`, [1], version 2.3)

package which implements up-to-date statistical algorithms to estimate cancer progression models from a list of genomic lesions (e.g., somatic mutations, copy number variations or persistent epigenetic states) in a population of independent tumors, or in a single patient.

`TRONCO`'s official webpage is reachable from the Software section of our group's webpage

`http://bimib.disco.unimib.it/`

By navigating to the Case Studies section of `TRONCO`'s official webpage one can find the source code to replicate this study (i.e., the PiCnIc's implementation) along with the documentation detailing all the implementation, as well as the data that we used. This should allow easy implementation of similar studies in different contexts.



B Glossary

This glossary of terms shall be of help to readers not familiar with the concepts mentioned in the Main Text. For clarity, terms are separate in two categories according to the fact that they are common to the statistics or the cancer biology communities. Each term which is included in this glossary appears in color.

Terms common to the statistics community

Term	Meaning
Boolean formula	In CAPRI, a formula written with standard logical operators which capture a relation among a group of alterations. In PiCnIc, these are used to detect alternative routes to selective advantage from mutually exclusive alterations. See: Fitness equivalence.
Ensemble level progression inference	Detection of the relations of selective advantage across the permanent alterations in a cohort of independent tumors (cross-sectional data). When aggregated in a graphical model, these shall picture the most common evolutionary trajectories in the population/cancer under study. See also: inter-tumor heterogeneity.
Graphical models	In this context, a direct acyclic graph with nodes (alterations) and edges (selective advantage relations), as a shorthand to represent the joint probability of observing a set of alterations in a sample (i.e., a cancer genotype). See also: Suppes-Bayes Causal Network, Model selection.
Individual level progression inference	Detection of clonal signatures and their prevalence in individual tumors by scanning multi-region or single-cell sequencing data; clones are then displayed in a phylogenetic tree structure. See also: intra-tumor heterogeneity.
Model selection	The process of selecting a model which fits data, according to some criterion. In CAPRI, this is done by balancing model likelihood (a measure of to which degree data can be explained by the model) and model complexity (the size of the graphical model). See: regularization.
Phylogenetic tree	In this context, rooted tree where each node is a clone, and edges represent ancestry relations among clones.
Regularization	Common approach to avoid overfitting (false-positives) during model selection – in CAPRI this is achieved by using the standard AIC/BIC which penalize with different severity graphical models which contain many selective advantage relations.
Simposon's paradox	A paradox in statistics, in which a trend appears in different groups of data but disappears or reverses when these groups are combined. In this context, this shall refer to genuine selective advantage relations which are not inferred unless data coming from different populations is separated before doing inference. See: heterogeneity, subtypes, formulas .
Suppes-Bayes Causal Networks	A specific type of graphical model returned by CAPRI algorithm, where each edge satisfies Suppes's conditions of probabilistic causation subsuming temporal ordering and positive statistical dependence – the statistical approach to estimate selective advantage among the alterations.

Terms common to the cancer biology community

Term	Meaning
Alterations	Somatic mutations: A change in the genome of a cell that is not inherited from a parent, but is acquired. CNVs: Structural variation of large regions of DNA segments, including deletions, insertions, duplications and complex multi-site variants.
Bulk sequencing	Genome sequencing from single tumor samples, each containing a large number of cells. The resulting genomic profiles are derived from a mixture of cells with potentially distinct evolutionary histories.
Clones; Clonal expansion	Clone: group of cells sharing an identical genome and that derive from a common ancestor. Clonal expansion: the production of descendent cells all arising originally from a single cell. In the scenario of cancer development, tumors develop through a series of clonal expansions, in which the most favorable clonal population survives and proliferate.
Cross-sectional data	Unique snapshots of data derived from samples that are collected at unknown time points. Usually derived from bulk sequencing technologies.
Driver; Passenger	Driver: (epi)genetic alteration that provides a selective advantage to a cancer clone. Passenger: alteration of a cancer cell that does not increase its fitness.
Exclusivity of alterations	Group of alterations which manifest few or no co-occurences in a cohort of different samples, and might be fitness-equivalent for tumor progression. Hard exclusivity: when co-occurrences shall be considered the result of random errors. Soft exclusivity: when few co-occurrences shall be possible. See: formulas.
Fitness	A cell's ability of surviving, proliferating and adapting to environmental changes, usually within an environment with limited and depleting resources (e.g., oxygen or nutrients).
Fitness equivalence	Groups of driver alterations, functional to the same pathway or equally dysruptive, that can independently confer a selective advantage to a cancer cell. Multiple co-occurrence of such alterations to provide no further advantage, hence leading to mutually exclusive alteration patterns across distinct samples.
Hallmark of cancer	Common traits or phenotypic properties that are supposed to drive the transformation of normal cells to cancer cells. Anti-hallmark: clonal profiles that are usually not observed, yet being theoretically possible.
Inter-tumor heterogeneity	The phenomenon according to which different patients with the same cancer type usually display a few common alterations. This is the major problem of inferring ensemble-level cancer progression models.
Intra-tumor heterogeneity	Intra-tumor heterogeneity is related to possible coexistence of different cancer clones, with different evolutionary histories and different mutational profiles, within the same tumor. This is the major problem of inferring individual-level cancer progression models.
Multiregion sequencing	Collection of genomic data obtained by processing multiple spatially separated biopsy samples from the same individual tumor.
Next Generation Sequencing (NGS)	New technologies for sequencing genomes at high speed and low cost, including, e.g., full-genome/exome sequencing, genome resequencing, transcriptome profiling (RNA-Seq), DNA-protein interactions (ChIP-Seq), and epigenome characterization.
Selective advantage relation	In successive waves of clonal expansions one or more cells of the same clone can (progressively) increase their fitness through the acquisition of additional driver alterations, leading to the emergence and development of a fitter clone. In this case a relation of selective avantage connects the earlier to the succeeding alterations.
Single-cell sequencing	Recent technology based on the retrieval and analysis of genomic information from individual cells, rather than from mixtures of cells.
Synthetic lethality	The phenomenon according to which two otherwise non-lethal alterations lead the cell death when they co-occur within the same cell. See: Anti-hallmark

C PiCnIc's implementation for COADREAD samples

Here we detail all the steps implemented to use PiCnIc for CRC progression inference.

C.1 TCGA COADREAD project data

COADREAD provides genome-scale analysis of samples with exome sequence, DNA copy number, promoter methylation, messenger RNA and microRNA expression data, which we used to define the input dataset. In particular, only samples with both mutations and CNAs profiles were used in the analysis. Supplementary Table S1 details the dataset.

Dataset used to infer models presented in the Main Text. Samples published in [2] were used as, to the best of our knowledge, these represent the highest–quality data made available by COADREAD as of today; for these samples TCGA provides somatic mutation profiles and high-resolution focal CNAs via GISTIC. These are obtained from TCGA data freeze as of 2 February 2012, downloaded on 12 March 2015, from repository:

`https://tcga-data.nci.nih.gov/docs/publications/coadread_2012/`

The following files were processed to produce the data:

- `TCGA_CRC_Suppl_Table2_Mutations_20120719.xlsx`
Somatic mutations profiles obtained via whole-exome sequencing of 224 colorectal tumors by TCGA. Data available consists of 15995 mutations in 228 samples, provided in the *Manual Annotation Format* (MAF). Samples were selected to univocally match the 224 patients as of the TCGA guidelines for aliquote disambiguation, see `https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode`. All the mutations annotated by TCGA – truncating (`De_novo_Start_OutOfFrame`, `Frame_Shift_Del`, `Frame_Shift_In`, `Nonsense_Mutation`, `Splice_Site`, `Frame_Shift_Ins`, `In_Frame_Del`), silent (`Silent`) and missense (`Missense_Mutation`) – were considered for analysis; notice that the majority of them are missense, see Figures S2 and S3.

- `crc_gistic.txt.zip`
Focal *Copy Number Alterations* (CNAs) for 564 patients derived from whole-genome sequencing using the Illumina HiSeq platform. *High-level gains* and *homozygous deletions* were considered for analysis by selecting entries with GISTIC scores ± 2;

- `crc_clinical_sheet.txt`
Clinical data summary with patient stage and *Micro Satellite Stabe/Unstable* (MSS/MSI) status being any of: MSS, MSI-high and MSI-low.

The list of patients used was first reduced to those having *both* CNAs and somatic mutation data, and then was split into two groups: MSI-HIGH and MSS. The training cohort has 152 MSS and 27 MSI-HIGH samples; samples flagged as low MSI were excluded from the study as they have not been shown to differ in their clinicopathologic features or in most molecular features from MSS tumors [3].

C.2 Driver events selection

In the TCGA COADREAD study [2] integrated analysis of mutations, copy number and mRNA expression changes in 195 tumours with complete data was performed. Part of the analysis was carried out by using the MutSig tool [4], as well as manual curation. Samples were grouped by mutation status, and recurrent alterations in key CRC pathways were identified in [2] (Fig. 4, Supplementary Fig. 6 and Supplementary Table 1) as a result, we can use the consortium's list of 33 driver genes annotated to 5 pathways and use these to extract our progression models. These are well-known cancer genes, frequently reported as relevant to colorectal progression and to the major pathways involved in CRC. Driver events are alterations in the following genes (acronym resolved at `http://www.genecards.org/`):

- WNT genes (14): Adenomatous Polyposis Coli (APC), Dickkopf WNT Signaling Pathway Inhibitor 4 (DKK-4), Transcription Factor 7-Like 2 (TCF7L2), Catenin beta-1 (CTNNB1), Low Density Lipoprotein Receptor-Related Protein 5 (LRP5), F-Box And WD Repeat Domain Containing 7 (FBXW7), Dickkopf WNT Signaling Pathway Inhibitor 1(DKK-1), Frizzled Class Receptor 10 (FZD10), AT Rich Interactive Domain 1A (ARID1A), Dickkopf WNT Signaling Pathway Inhibitor 2 (DKK-2), APC Membrane Recruitment Protein 1 (FAM123B, also known as AMER1), SRY (Sex Determining Region Y)-Box 9 (SOX9), Dickkopf WNT Signaling Pathway Inhibitor 3 (DKK-3) and axin-2 (AXIN2);

- RTK/RAS genes (5): Erb-B2 Receptor Tyrosine Kinase 2 (ERBB2), Erb-B2 Receptor Tyrosine Kinase 3 (ERBB3), Neuroblastoma RAS Viral (V-Ras) Oncogene Homolog (NRAS), Kirsten Rat Sarcoma Viral Oncogene Homolog (KRAS) and B-Raf Proto-Oncogene, Serine/Threonine Kinase (BRAF);

- TGF-β genes (5): Transforming Growth Factor, Beta Receptor 1 (TGFBR1), SMAD Family Member 3 (SMAD3), Transforming Growth Factor, Beta Receptor 2 (TGFBR2), SMAD Family Member 4 (SMAD4), Activin A Receptor Type IB (ACVR1B), Activin A Receptor Type IIA (ACVR2A) and SMAD Family Member 2 (SMAD2);

- IGF2/PI3K genes (5): Insulin-Like Growth Factor 2 (IGF2), Insulin Receptor Substrate 2 (IRS2), Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha (PIK3CA), Phosphoinositide-3-Kinase Regulatory Subunit 1 Alpha (PIK3R1) and Phosphatase And Tensin Homolog (PTEN);

- P53 genes (2): Tumor Protein P53 (TP53) and ATM Serine/Threonine Kinase (ATM).

In the Main Text, RTK/RAS and IGF2/PI3K pathways are shortly denoted as RAS and PI3K.

The distinct types of mutations detected in these genes are shown in Figure S2, as well as the overall rate of COADREAD mutations. The spatial distribution (per gene) of such mutations is shown in Figure S3.

C.3 Mutual exclusivity groups of alterations.

Groups of alterations showing a trend of mutual exclusivity were scanned with MUTEX and mutations and CNA hitting any of the 33 selected genes as input. MUTEX was run independently on MSS and MSI-HIGH groups (Supplementary Table S2, running times: approximately 6 and 3.5 hours, respectively, on a standard Desktop machine).

We selected only groups with score < 0.2, where the score is derived from *p-values corrected for false discovery rate*. 3 groups are found for MSI-HIGH tumors and 6 for MSS. For MSI-HIGH tumors, the three predicted groups consists of genes ACVR1B, ACVR2A, TP53 and ERBB2, of genes BRAF, NRAS and TGFBR2, and of genes KRAS and BRAF.

Further groups of exclusive alterations were considered consistent with results reported in [2]. These include groups derived by consolidated knowledge of colorectal progression: the well-known WNT alterations in APC/CTNNB1 [5], as well as RAS alterations in KRAS, NRAS and BRAF genes [6]. Similarly, we used also a group collected by scanning non-hypermutated tumors with the MEMO tool in [2] - this group includes PIK3CA, PTEN, ERBB2 and IGF2 genes. These groups were restricted to account only for genes actually altered in a certain subtype, e.g., MSI-HIGH tumors lack CTNNB1 mutation, making the Wnt group irrelevant. Groups for MSS tumors are shown as Supplementary Figure S4, groups for MSI-HIGH tumors are in the Main Text.

C.4 CAPRI's execution

Background on the algorithm. CAPRI algorithm can be executed in two different modes, originally dubbed as "supervised" when formulas are given in input for testing, and "unsupervised" when this is not the case. This paper deals with the former; see the Main Text for the interpretation of formulas in this context and [7] for a derivation of the algorithm. CAPRI is a three-steps procedure, which we briefly recall here.

1. CAPRI starts by creating a "lifted" representation of the input data $\mathbf{M}$ which includes the input formulas; each formula – which is written in propositional logic – is so evaluated to yield a new column in the dataset. This is the input processed by the algorithm, which starts by selecting a set of candidate model edges, which are then used to constrain a score-based Bayesian model-selection problem [7].

2. The initial set of selective advantage relations $\mathcal{S}$, which determines the model edges $x \rightarrow y$, is computed by evaluating the following inequalities

$$\begin{aligned} \texttt{(Temporal priority)} \quad & p(x) > p(y) \\ \texttt{(Probability raising)} \quad & p(y \mid x) > p(y \mid \neg x)\,, \end{aligned}$$

   where $p(\cdot)$ is a marginal probability, $p(\cdot \mid \cdot)$ is a conditional, $\neg x$ is the negation of x and either x or y is not a formula. It is interesting to observe that `probability raising` implies that x and y are *positively statistically dependent*, thus imposing a minimum threshold on their association [8].

   In each of CAPRI's executions, the distribution of the observed marginals and conditionals are estimated by K non-parametric bootstrap resamples; practically, this means that we create a bootstrapped approximation for each of the four populations $\hat{p}(x)$, $\hat{p}(y)$, $\hat{p}(y \mid x)$ and $\hat{p}(y \mid \neg x)$. Then, we use a single-tail non-parametric Mann-Withney U test of the difference in mean to test the hypothesis that one of the two populations is more probable than the other, e.g., $\hat{p}(x) > \hat{p}(y)$. With this test, we can compute two p-values, one for each condition. An edge is included in $\mathcal{S}$ if at least the p-value for `probability raising` is below a significance threshold p_* and an edge is said not to be orientable if its p-value for `temporal priority` is above the same threshold. Cycles $x_1 \rightarrow x_2 \rightarrow \ldots x_k \rightarrow x_1$ that might appear in $\mathcal{S}$ are broken by deleting the edge with minimum p-value; both K and p_* are custom parameters.

3. Optimization of $\mathcal{S}$, namely detection of the subset $\mathcal{S}^*$ of $\mathcal{S}$ with the edges that we include in the final progression model is done by optimizing, via *hill climbing* or *tabu search*, the *score with regularization*

$$\mathcal{S}^* \triangleq \arg\min_{\hat{\mathcal{S}} \subset \mathcal{S}} \left\{ -2\log[\mathcal{L}(\hat{\mathcal{S}} \mid \mathbf{M})] + \theta|\hat{\mathcal{S}}| \right\} , \tag{1}$$

where $\mathcal{L}(\cdot)$ is the *model likelihood* and $\mathbf{M}$ is the input data; the estimated optimal solution is $\mathcal{S}^*$, which is displayed as a Suppes-Bayes Causal Network. The different regularization strategies mentioned in the Main Text, BIC and AIC, are obtained by the following parametrization:

$$\texttt{(Bayesian Information Criterion)} \quad \theta \triangleq \log(n)$$
$$\texttt{(Akaike Information Criterion)} \quad \theta \triangleq 2.$$

Besides edges, a model has a set of parameters $\boldsymbol{\theta}$ which define the *conditional probability table* of each edge and should be fit from data; these are necessary if one wishes to use a model as a "generator" of further data. For discrete-valued graphical models, for each parent set $\pi_x \to y$, parameter $\boldsymbol{\theta}(y) = p(y \mid \pi_x)$ can be taken either as the *maximum likelihood estimate* from the lifted input, or by using a Bayesian interpretation [9].

Usage in this context. CAPRI was run, on each group of tumors, by selecting alterations from the pool of 33 pathway genes; every alteration on a gene x is included if *any* of these apply:

- the alteration frequency of x - sum of mutation and CNA frequency - is greater than 5%;
- x it is part of an exclusivity group.

The set of selected events for MSI-HIGH training tumors is shown in the Main Text, the analogous set for MSS tumors is shown as Supplementary Figure S5.

CAPRI was executed in its supervised mode by writing formula over groups and genes with multiple alterations associated, as explained in the Main Text. For instance, for MSI-HIGH tumors with alterations in RAS pathway we grouped hard exclusivity of NRAS mutations and deletions, with soft exclusivity of KRAS and BRAF mutations. Our aim was to account for a small subset of samples with concurrent KRAS and NRAS alterations (see Figure 2, Main Text). The list of all Boolean formulas written over groups is in Table S3; this approach was adopted also when a gene harbors multiple alterations in a subtype, e.g., ERBB2 in MSS training samples which shows a trend of soft exclusivity between mutations and amplifications. We used both AIC and BIC scores to regularize inference after 100 non-parametric bootstrap iterations for estimation of the preliminary selective advantage relations – Mann-Whitney U test was performed with a minimum threshold $p_* = 0.05$. In most cases p-values are orders of magnitude below p_* - exact values reported as Dataset File S1. CAPRI's models with such p-values and non-parametric bootstrap confidence are shown in Figures S6 and S7, statistical validation of the models is discussed in the next section.

D Statistical validation of the models

P-values from hypothesis testing, as well as scores from k-fold *cross-validation* and various *bootstrapping* techniques can be used to measure the statistical consistency of models and data, each one capturing different potential errors in the inference process. Approaches such as cross-validation and bootstrap are sometimes also used in the (*ex novo*) generation and inference of models from data (see, e.g., *bootstrap consensus models* [10]), but we only use them here for the *a posteriori* evaluation of a model's confidence, and we interpret them as a quantitative measure for the *relative* assessment of each model's relation.

All the p-values and the scores that we present here are computed within TRONCO.

D.1 Edge p-values

As explained in §C.4, for each model edge $x \to y$ we get two p-values by assessing temporal priority and probability raising via Mann-Whitney U testing.

For each edge, a p-value for the *hypergeometric test* of overlap between alteration profiles x and y can be computed. More precisely, we test if there is a difference between the number of samples containing *both* x and y versus the total population of samples with x, y, or both. We would like the overlap to be significant as those samples – that determine the joint probability of x and y – are those supporting the presence of a selection trend among x and y.

An edge is fully supported if all three p-values are below a custom significance threshold, e.g., 0.05 or even better 0.01. Some edges might have the p-value for temporal priority above the threshold. If so, the selection trend might be still significant, but the temporal order of x and y – i.e., the *direction of selection* – is not supported by the data.

D.2 Bootstrap

We used non-parametric and statistical bootstrap techniques to measure the *goodness-of-fit*, as originally proposed in [11]. In this case, we distinguish two type of errors that one could make in the inference process, estimating the presence or absence of edges in the model:

- *Type I errors:* incorrect rejection of a true H_0 (null-hypthesis), i.e., a *"false positive"* edge that we wrongly include.
- *Type II errors:* incorrect acceptance (failure to reject) of a false H_0, i.e., a *"false negative"* edge that we miss.

Non-parametric bootstrap [12] computes scores to be interpreted here as follows. If a model contains an edge $x \to y$ that is a true positive, we expect its score to be high. In other words, when we sample with repetition subsets of the original data and re-run the inference process we expect to often find models which contain the edge $x \to y$. Conversely, for a node y without incoming edges, or equivalently for any edge $x \to y$ which is correctly excluded from a model – a true negative – we would expect its bootstrap score to be low. However, this reasoning can be also generalized to whole models, where we count how many times we re-infer exactly the same model. Clearly, such scores will depend also on the empirical probabilities of the nodes in our data, and their deviation from the true probabilities of the phenomenon. So, one might expect rare events to be less frequently bootstrapped, which results in a lower estimate; however, such counts can be anyway interpreted as measures of repeatability of our findings[1].

The above bootstrap approach depends on two random number generators: one to shuffle data (the a posteriori bootstrap), and one to evaluate CAPRI's inequalities via hypothesis testing (the internal CAPRI's bootstrap, see C.4). Thus, to ensure that no bias is introduced by the random number generators, we performed a *statistical bootstrap* by holding data fixed, and re-estimating CAPRI's inequalities with generators initialized with different seeds. We evaluated the robustness of our scores for all edges imputed to be genuine and hence, high-scoring.

Notice that, in principle, even *parametric bootstrap* scores could be computed if we used the model to generate bootstrapped data [12]. However, as the support of the distribution subsumed by a model with n nodes consists of 2^n possible outcomes, sampling uniformly from large models might be computationally hard. For this reason, and because such scores are overestimates of the non-parametric ones, we did not include them in our computation.

The MSS and MSI progression models are annotated with the non-parametric bootstrap scores in Figures S6 and S7. Non-parametric and statistical bootstrap scores for a set of selected edges are shown and commented in Figure S8. For the same set of edges, we also report the p-values for CAPRI's inequalities assessed in the MSS and MSI progression models (temporal priority and probability raising, as a measure of the selectivity among the alterations, and hypergeometric, as a measure of the randomness in the overlap of two alteration profiles). Additional comments are in the caption.

D.3 Cross-validation

Next we study the sufficiency of the data sizes for model inference and its ability to characterize the underlying progression (*goodness-of-data*). Thus, we focus on the Type III errors, which occur when the sample size is inadequate or the sample is a poor descriptor for the reference phenomenon[2], and thus failing to represent the progression. For this purpose, we used cross-validation with the data used for the models built (see the Main Text), and followed the best practices developed by the Bayesian Networks community [9].

`TRONCO` exploits the cross-validation routines implemented in the `bnlearn` package [16]. The approach that we adopt is a k-fold non-exhaustive cross-validation, which we repeat 10 times to average its results. Exhaustive strategies might be used for datasets of small sample size. Each run of cross-validation consists in computing a loss function for a model; its steps are the followings:

- split randomly the data in $k = 10$ groups, and then repeat the following two steps, for each group in turn:
    1. set one of the groups to be the "training" $\mathbf{M}_{tr}$;
    2. merge the others $k-1$ to be the "test" $\mathbf{M}_{te}$;
    3. by holding fixed the model structure (i.e., the edges in $\mathcal{S}^*$), fit the model parameters $\boldsymbol{\theta}$ over the training data via *maximum likelihood estimates*, compute a score over the test $\mathbf{M}_{te}$ (see below) and the corresponding loss;

[1] Bootstrapping techniques have been widely used to gauge uncertainty in estimates, but also subjects of philosophical debate about their precise interpretation, especially when coupled with various significance thresholds – unlike the situation with a p-value for a null hypothesis. An exhaustive review on the topic is provided by Soltis in [13]. We follow the ideas originally developed in the area of phylogenetic analysis [14], suggesting that the scores can be alternatively interpreted as a measure of accuracy of the method or of the robustness of the data [15].

[2] Consider the case where the samples are from two different unknown and heterogeneous groups, with random chance making low values to be sampled from a group that actually has a majority of high values, and vice versa. In this case, the samples will not be the best descriptor of the two groups.

- combine the k loss estimates to give an overall loss for data.

Let $\boldsymbol{\theta}_{tr}$ be the parameters fit from the training set $\mathbf{M}_{tr}$, and $\mathcal{S}^*$ the edges in the model, three scores are computed with cross-validation:

1. the *negative entropy* of a model – i.e., the negated expected log-likelihood of the test set for the Bayesian network fitted from the training set, that is

$$\texttt{eloss}(\mathcal{S}^*, \boldsymbol{\theta}_{tr}) = -\mathbb{E}[\mathcal{L}(\mathcal{S}^*, \boldsymbol{\theta}_{tr} \mid \mathbf{M}_{te})] \,.$$

2. the *prediction error* for a single node x and its parents set X, i.e., we measure how precisely we can predict the values of x by using only the information present in its local distribution $\boldsymbol{\theta}_{tr}(x)$. This parameter corresponds to computing the misclassification rates from $p_{te}(x)$, the empirical marginal probability of x estimated from the test.

3. the *posterior classification error* for a single node x and one of its parent node $y \in X$ – i.e., the values of x are predicted using only the information present in y by likelihood weighting and Bayesian posterior estimates.

See [17] for a discussion on these loss functions. The first statistics measures the *log-likelihood loss* when we "forget" some of the samples used to infer a model (indeed, the test samples); see Figure S9. Roughly, we are measuring how the model's predictive power changes as we look at the data from different viewpoints. This is a score for a whole model, it has no scale – so cannot be used to say how good the models/data are, in any absolute sense. Nonetheless, it can be used to evaluate how "stable" a model is, for a certain dataset.

The second and third statistics measure the accuracy of the parent-set, X, for a child x; the second statistics dealing with the whole parent set as predictor, and the third, the individual contribution of each of the parents. For these two statistics, we desire the prediction error to be low, as a measure of goodness. These are shown in Figure S8 (selected edges), S10 and S11 (all edges, prediction error).

E Supplementary Tables and Figures

Datasets (CNAs and mutations provided by TCGA)

	statistics			**alteration type**		
cancer[†]	n	m	$\lvert G\rvert$	*mutations*	*amplifications*	*deletions*
MSI-HIGH	27	16100	13798	11556	2888	1656
MSS	152	21317	16371	12417	6925	1975

[†] Samples were classified as MSI-HIGH/LOW and MSS by TCGA; see flag `MSI_status` in clinical data available for the COADREAD project.

Table S1: **COADREAD Data.** Data used in this study, derived from the TCGA COADREAD project [2].

MUTEX parameters

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 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; available for download at `https://code.google.com/p/mutex/`.

MUTEX groups with score $< .2$

	MSI-HIGH Groups	*score*	*q-value*
1	KRAS, BRAF,	0.095	0.48
2	NRAS, BRAF, TGFBR1	0.1677	0.45
3	ERBB2, TP53, ACVR1B, ACVR2A	0.1703	0.355

	MSS Groups	*score*	*q-value*
1	TP53, ATM,	0.051	0.34
2	ARID1A, TP53	0.075	0.193
3	KRAS, NRAS, BRAF,	0.0864	0.1975
4	CTNNB1, APC, DKK2,	0.098	0.144
5	DKK1, TP53, ATM, DKK2	0.1387	0.176
6	PIK3CA, TP53, ATM	0.164	0.207

Table S2: **MUTEX: parameters and results.** Top: Parameters used to run MUTEX on the original TCGA MSS/MSI-HIGH datasets with input CNA and somatic mutations in the pathway genes described in text. Bottom: MUTEX identified 3 and 6 groups of alterations showing a trend of mutual exclusivity in these groups with score below the suggested cutoff of 0.2.

Formulas input for testing to CAPRI[†]

	MSI-HIGH tumors	*description*
1	(NRAS:m ⊕ NRAS:d) ∨ KRAS:m ∨ BRAF:m	RAF exclusivity
2	PIK3CA:m ∨ ERBB2:m ∨ PTEN:m ∨ IGF2:d	MEMO group
3	(ACVR1B:m ⊕ ACVR1B:d) ∨ ACVR2A:m ∨ TP53:m ∨ ERBB2:m	MUTEX group
4	(NRAS:m ⊕ NRAS:d) ∨ TGFBR1:m ∨ BRAF:m	MUTEX group
5	KRAS:m ∨ BRAF:m	MUTEX group
6	ACVR1B:m ⊕ ACVR1B:a	multiple alterations
7	NRAS:m ⊕ NRAS:a	multiple alterations
8	FBXW7:m ∨ FBXW7:a	multiple alterations [‡]
	MSS tumors	*description*
1	(APC:m ⊕ APC:d) ∨ CTNNB1:m	WNT exclusivity
2	(KRAS:m ∨ KRAS:a) ∨ (NRAS:m ⊕ NRAS:a) ∨ (BRAF:m ⊕ BRAF:a)	RAF exclusivity and MEMO group
3	PIK3CA:m ∨ (ERBB2:m ∨ ERBB2:a) ∨ (PTEN:m ⊕ PTEN:d) ∨ IGF2:a	MEMO group
4	(TP53:m ⊕ TP53:d) ∨ (ATM:m ⊕ ATM:d)	MUTEX group
5	(TP53:m ⊕ TP53:d) ∨ ARID1A:m	MUTEX group
6	(TP53:m ⊕ TP53:d) ∨ ARID1A:m	MUTEX group
7	(APC:m ⊕ APC:d) ∨ CTNNB1:m ∨ DKK2:m	MUTEX group
8	(TP53:m ⊕ TP53:d) ∨ (ATM:m ⊕ ATM:d) ∨ DKK2:m ∨ DKK1:m	MUTEX group
9	(TP53:m ⊕ TP53:d) ∨ (ATM:m ⊕ ATM:d) ∨ PIK3CA:m	MUTEX group
10	(APC:m ⊕ APC:d)	multiple alterations
11	(TP53:m ⊕ TP53:d)	multiple alterations
12	(SMAD4:m ⊕ SMAD4:d)	multiple alterations
13	(TCF7L2:m ⊕ TCF7L2:d)	multiple alterations
14	(ATM:m ⊕ ATM:d)	multiple alterations
15	(NRAS:m ⊕ NRAS:d)	multiple alterations
16	(ERBB2:m ∨ ERBB2:a)	multiple alterations
17	(PTEN:m ⊕ PTEN:d)	multiple alterations
18	(SMAD2:m ⊕ SMAD2:a)	multiple alterations
19	(DKK4:m ⊕ DKK4:a)	multiple alterations
20	(SOX9:m ⊕ SOX9:d)	multiple alterations
21	(BRAF:m ⊕ BRAF:a)	multiple alterations

[†] Events type: mutation (m), deletion (d), amplification (a). Hard (⊕) and soft (∨) exclusivity.

[‡] Formula not included as it creates a duplicated signature in the dataset.

Table S3: **CAPRI formulas from exclusivity groups.** Formulas created for the groups, and input to CAPRI for testing. These are either derived from exclusivity groups or from genes involved in different types of alterations.

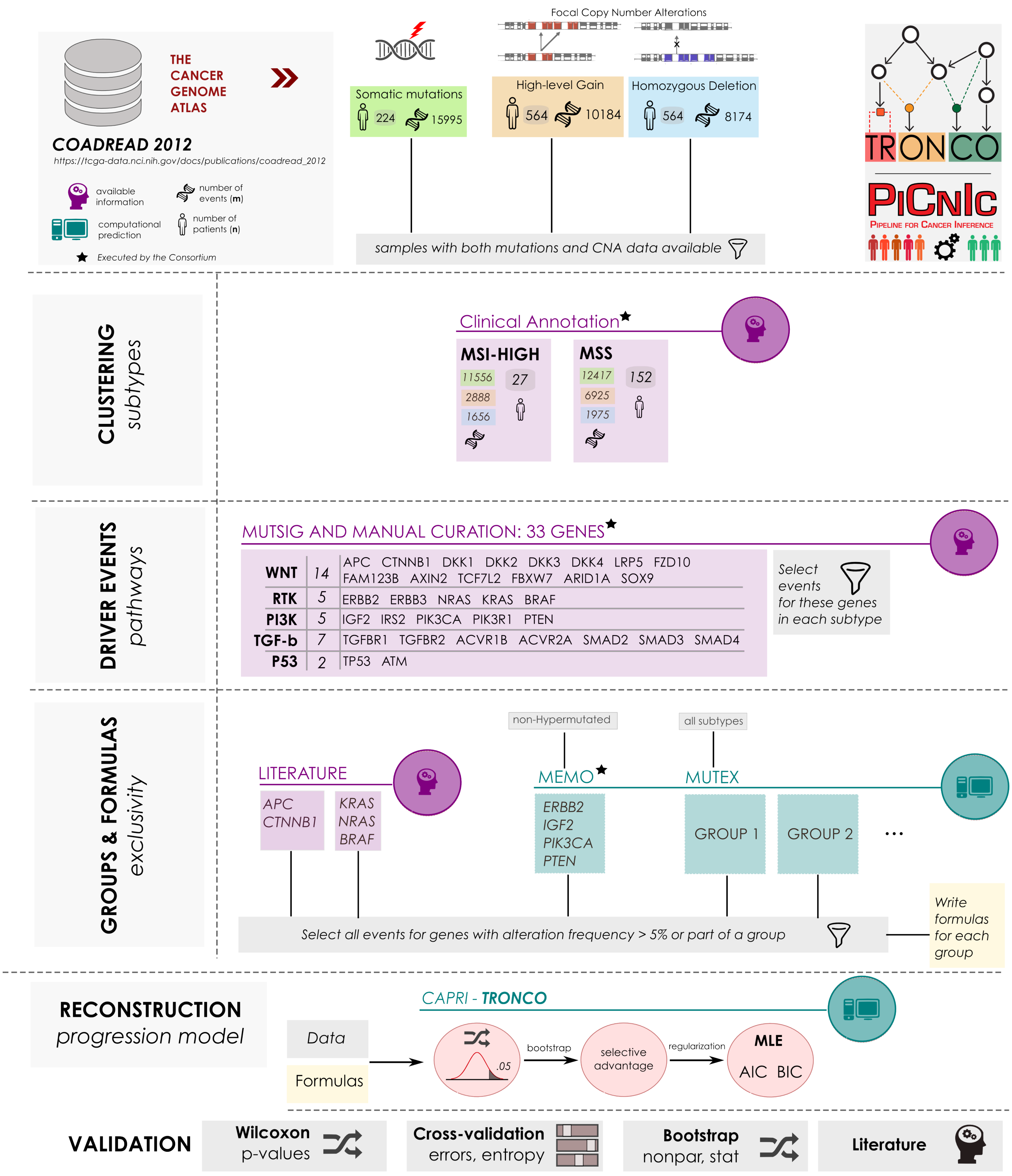


Figure S1: **PiCnIc pipeline processing MSI-HIGH/MSS tumors.** We process with PiCnIc Microsatellite Stable and highly unstable tumors collected from the The Cancer Genome Atlas project "Human Colon and Rectal Cancer" (subtypes annotations provided as clinical data). We implement a study on selected somatic mutations and focal CNAs in 33 driver genes manually annotated with 5 pathways in the COADREAD project. We scan groups of exclusive alterations with computational tools run by us and by TCGA, and we exploit previous knowledge on CRC; we select which alterations we input to CAPRI. Next, inference is performed with various settings of regularization and confidence. Statistical confidence of the models is assessed with standard techniques from the literature (p-values from statistical testing, bootstrap scores and cross-validation statistics).

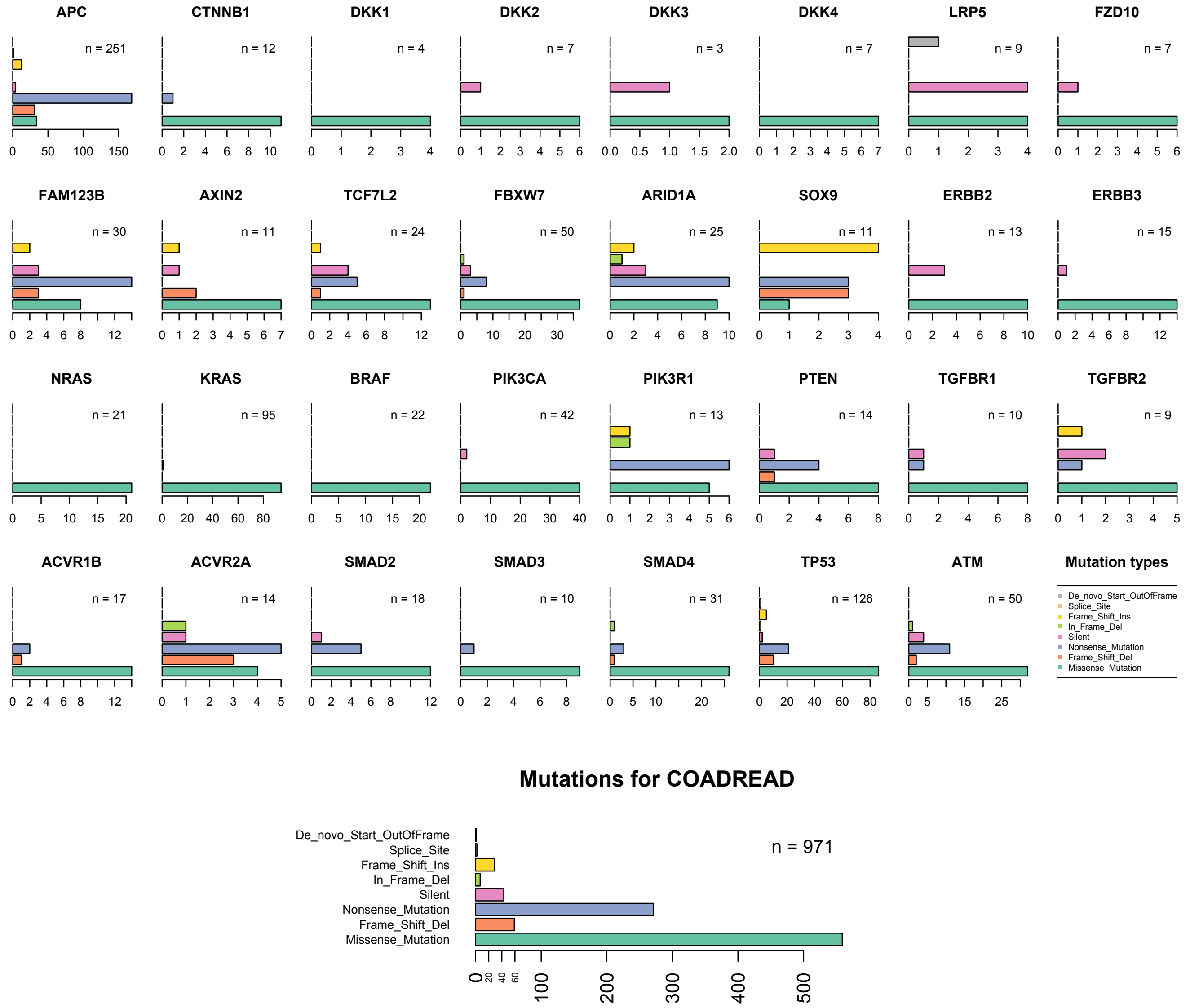


Figure S2: **Mutations annotated by TCGA for the driver genes.** Top: the majority of the mutations annotated by TCGA for the driver genes that we consider are *missense* – this in almost all genes and in all the cohort. Bottom: overall summation of the frequencies determine the mutations across all driver genes.

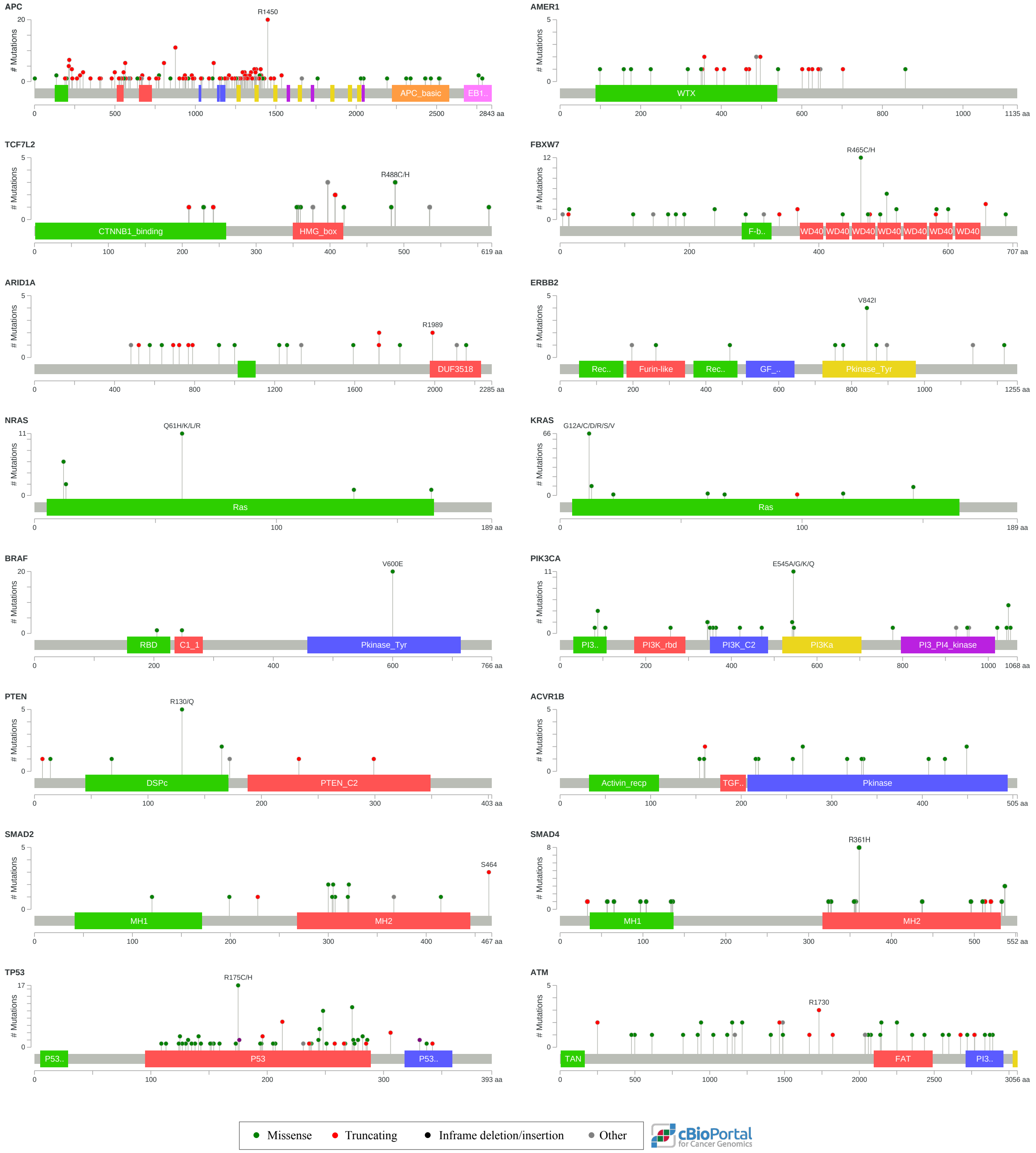


Figure S3: **Lolliplot diagrams of TCGA mutations.** Diagrams generated from the cBio portal for the COADREAD project (see `http://www.cbioportal.org/`). These display the physical distribution of the annotated mutations for each gene. Here we shown only genes with a total mutation count greater than 15; FAM123B is called with its synonym AMER1, as in the portal.

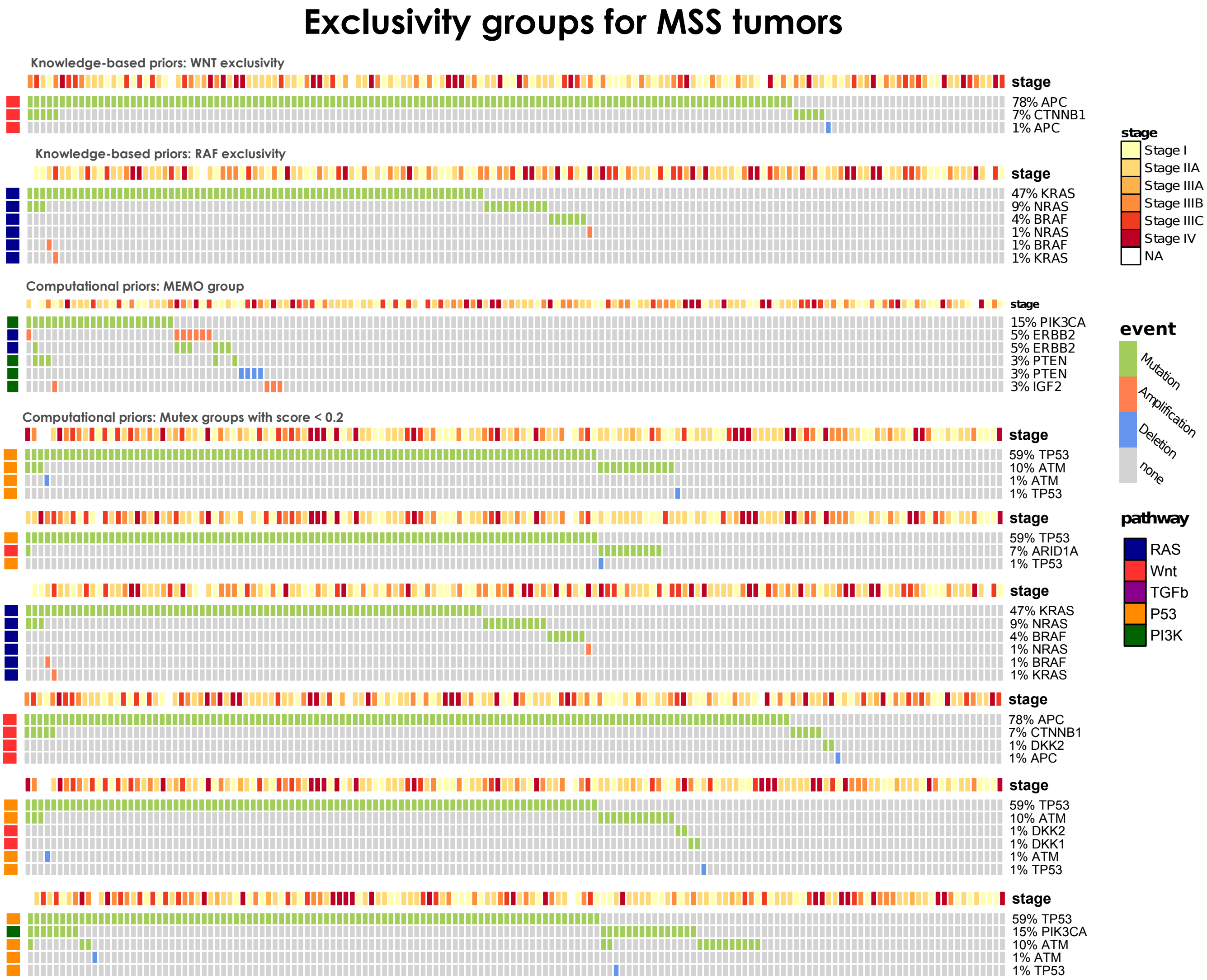


Figure S4: **Groups of exclusive alterations for MSS tumors.** Knowledge-based groups of exclusive alterations consist of: KRAS, NRAS and BRAF genes (RAF pathway) and APC and CTNNB1 genes (WNT pathway). The MEMO [18] group identified in [2] in this cohort consists of genes PIK3CA, ERBB2, IGF2 and PTEN. Finally, 6 groups are predicted by MUTEX [19] with score below .2, one of these is equivalent to the known exclusive alterations in RAF pathway.

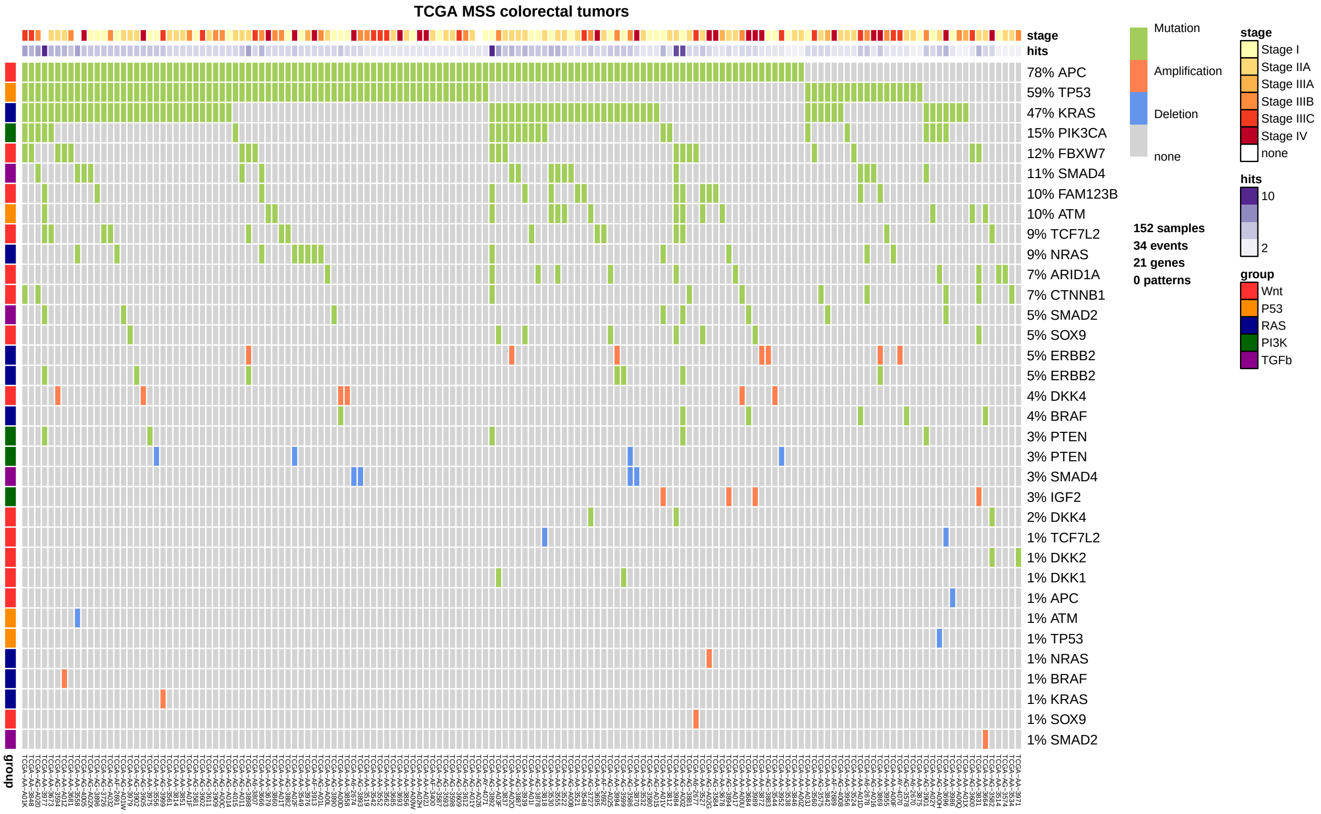


Figure S5: **Selected data for MSS tumors.** Colorectal tumors with Microsatellite Stable clinical status in the TCGA COADREAD project, restricted to 152 samples with both somatic mutations and CNA data available. 33 driver genes annotated with 5 pathways are selected from the list published in [2] to automatically detect groups of mutually exclusive alterations. Events selected for reconstruction are those involving genes altered in at least 5% of the cases, or part of group of alterations showing an exclusivity trend (see Figure S4). This dataset is used to infer the set of selective advantage relations which constitute the MSS progression model presented in the Main Text.

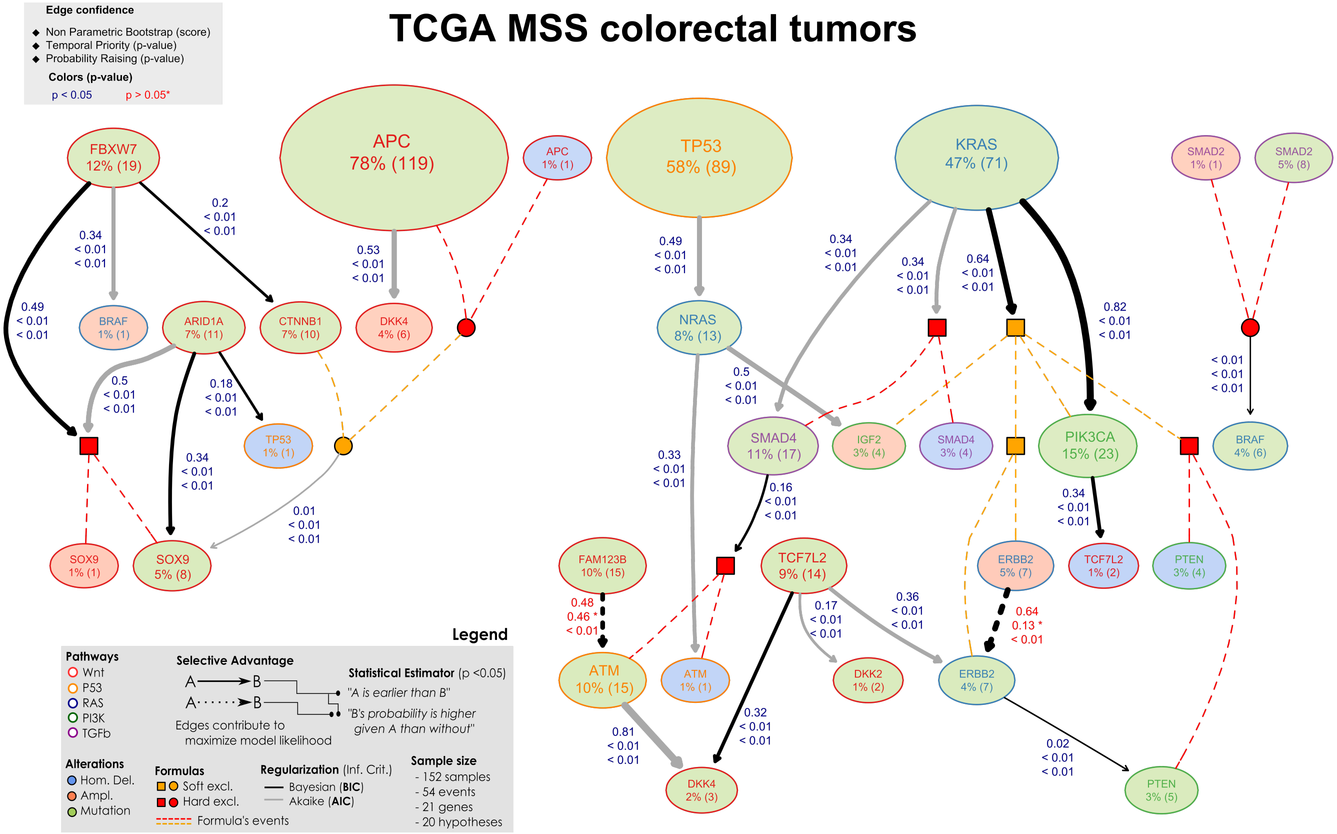


Figure S6: **Non-parametric bootstrap scores for MSS progression.** Progression model for MSS tumors with confidence shown as edge labels. The first label represents the relation confidence estimated with 100 non-parametric bootstrap iterations, the second and third are p-values for temporal priority and probability raising. Red p-values are above the minimum significane threshold of $.05$. See Figure 4 in the Main Text for an interpretation of this model.

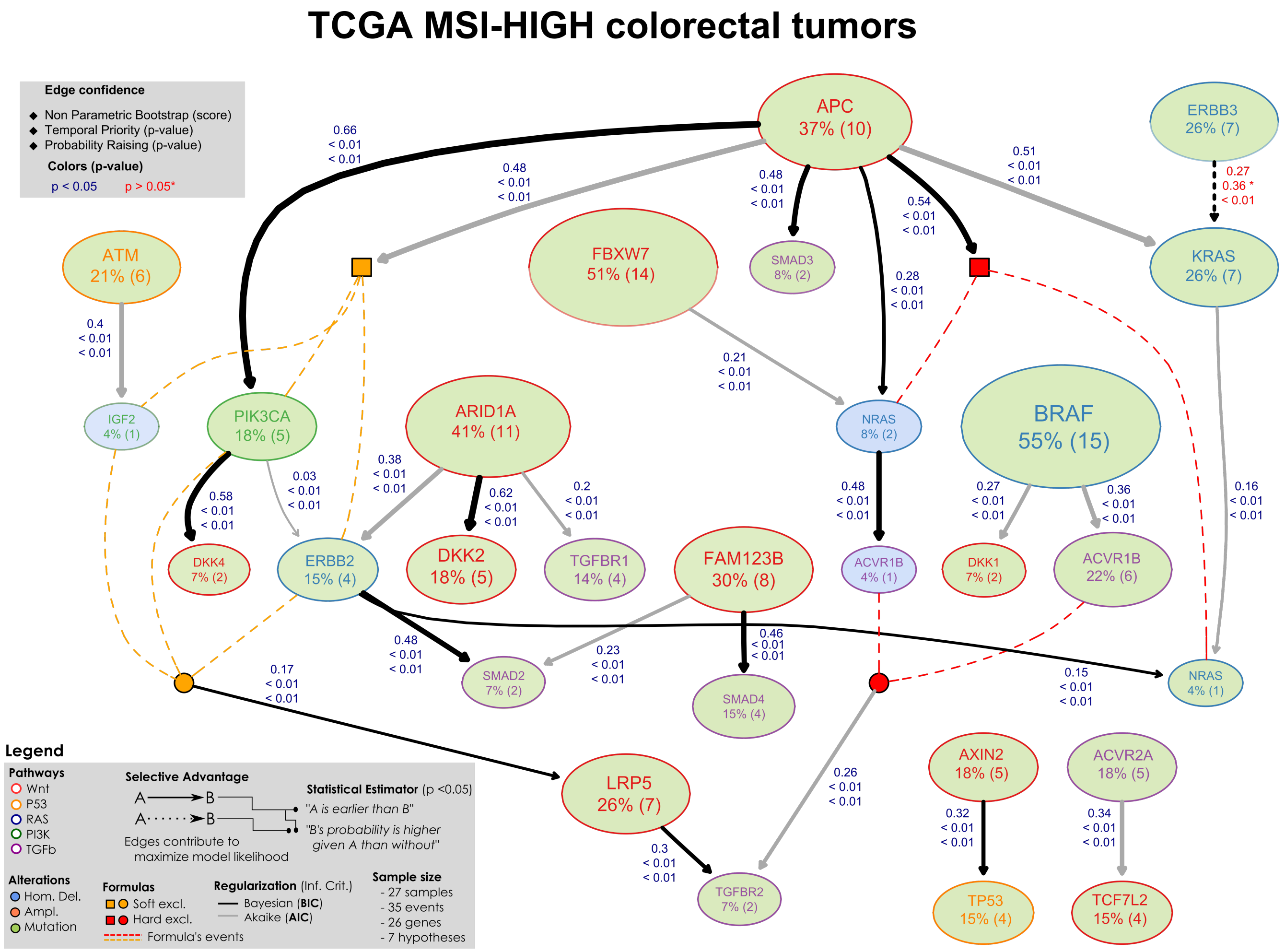
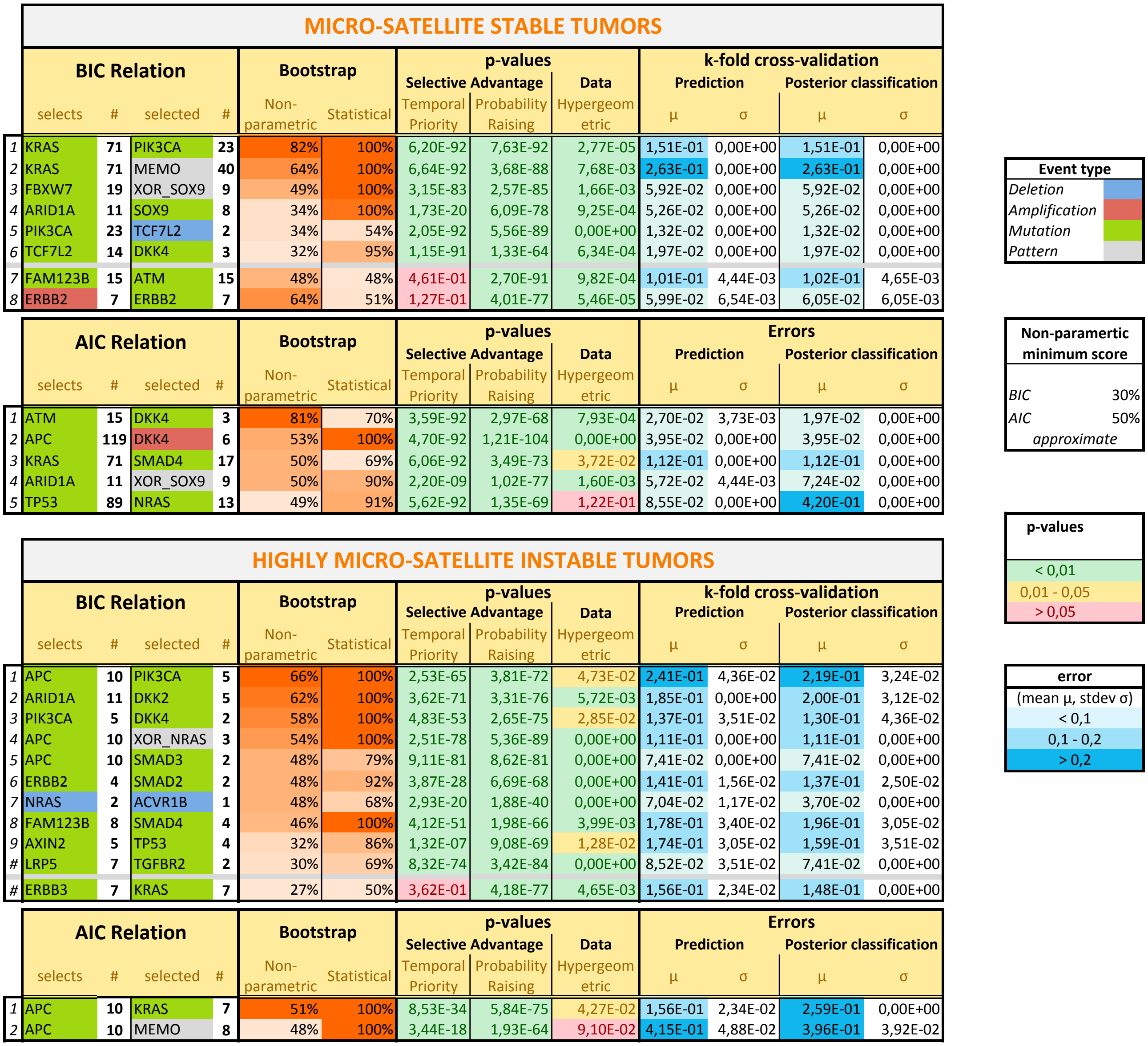


Figure S7: **Non-parametric bootstrap scores for MSI-HIGH progression.** Progression model for MSI-HIGH tumors with confidence shown as edge labels. The first label represents the relation confidence estimated with 100 non-parametric bootstrap iterations, the second and third are p-values for temporal priority and probability raising. Red p-values are above the minimum significane threshold of $.05$. See Figure 5 in the Main Text for an interpretation of this model.



MICRO-SATELLITE STABLE TUMORS

	BIC Relation				Bootstrap		p-values			k-fold cross-validation			
							Selective Advantage		Data	Prediction		Posterior classification	
	selects	#	selected	#	Non-parametric	Statistical	Temporal Priority	Probability Raising	Hypergeometric	μ	σ	μ	σ
1	KRAS	71	PIK3CA	23	82%	100%	6,20E-92	7,63E-92	2,77E-05	1,51E-01	0,00E+00	1,51E-01	0,00E+00
2	KRAS	71	MEMO	40	64%	100%	6,64E-92	3,68E-88	7,68E-03	2,63E-01	0,00E+00	2,63E-01	0,00E+00
3	FBXW7	19	XOR_SOX9	9	49%	100%	3,15E-83	2,57E-85	1,66E-03	5,92E-02	0,00E+00	5,92E-02	0,00E+00
4	ARID1A	11	SOX9	8	34%	100%	1,73E-20	6,09E-78	9,25E-04	5,26E-02	0,00E+00	5,26E-02	0,00E+00
5	PIK3CA	23	TCF7L2	2	34%	54%	2,05E-92	5,56E-89	0,00E+00	1,32E-02	0,00E+00	1,32E-02	0,00E+00
6	TCF7L2	14	DKK4	3	32%	95%	1,15E-91	1,33E-64	6,34E-04	1,97E-02	0,00E+00	1,97E-02	0,00E+00
7	FAM123B	15	ATM	15	48%	48%	4,61E-01	2,70E-91	9,82E-04	1,01E-01	4,44E-03	1,02E-01	4,65E-03
8	ERBB2	7	ERBB2	7	64%	51%	1,27E-01	4,01E-77	5,46E-05	5,99E-02	6,54E-03	6,05E-02	6,05E-03

	AIC Relation				Bootstrap		p-values			Errors			
							Selective Advantage		Data	Prediction		Posterior classification	
	selects	#	selected	#	Non-parametric	Statistical	Temporal Priority	Probability Raising	Hypergeometric	μ	σ	μ	σ
1	ATM	15	DKK4	3	81%	70%	3,59E-92	2,97E-68	7,93E-04	2,70E-02	3,73E-03	1,97E-02	0,00E+00
2	APC	119	DKK4	6	53%	100%	4,70E-92	1,21E-104	0,00E+00	3,95E-02	0,00E+00	3,95E-02	0,00E+00
3	KRAS	71	SMAD4	17	50%	69%	6,06E-92	3,49E-73	3,72E-02	1,12E-01	0,00E+00	1,12E-01	0,00E+00
4	ARID1A	11	XOR_SOX9	9	50%	90%	2,20E-09	1,02E-77	1,60E-03	5,72E-02	4,44E-03	7,24E-02	0,00E+00
5	TP53	89	NRAS	13	49%	91%	5,62E-92	1,35E-69	1,22E-01	8,55E-02	0,00E+00	4,20E-01	0,00E+00

HIGHLY MICRO-SATELLITE INSTABLE TUMORS

	BIC Relation				Bootstrap		p-values			k-fold cross-validation			
							Selective Advantage		Data	Prediction		Posterior classification	
	selects	#	selected	#	Non-parametric	Statistical	Temporal Priority	Probability Raising	Hypergeometric	μ	σ	μ	σ
1	APC	10	PIK3CA	5	66%	100%	2,53E-65	3,81E-72	4,73E-02	2,41E-01	4,36E-02	2,19E-01	3,24E-02
2	ARID1A	11	DKK2	5	62%	100%	3,62E-71	3,31E-76	5,72E-03	1,85E-01	0,00E+00	2,00E-01	3,12E-02
3	PIK3CA	5	DKK4	2	58%	100%	4,83E-53	2,65E-75	2,85E-02	1,37E-01	3,51E-02	1,30E-01	4,36E-02
4	APC	10	XOR_NRAS	3	54%	100%	2,51E-78	5,36E-89	0,00E+00	1,11E-01	0,00E+00	1,11E-01	0,00E+00
5	APC	10	SMAD3	2	48%	79%	9,11E-81	8,62E-81	0,00E+00	7,41E-02	0,00E+00	7,41E-02	0,00E+00
6	ERBB2	4	SMAD2	2	48%	92%	3,87E-28	6,69E-68	0,00E+00	1,41E-01	1,56E-02	1,37E-01	2,50E-02
7	NRAS	2	ACVR1B	1	48%	68%	2,93E-20	1,88E-40	0,00E+00	7,04E-02	1,17E-02	3,70E-02	0,00E+00
8	FAM123B	8	SMAD4	4	46%	100%	4,12E-51	1,98E-66	3,99E-03	1,78E-01	3,40E-02	1,96E-01	3,05E-02
9	AXIN2	5	TP53	4	32%	86%	1,32E-07	9,08E-69	1,28E-02	1,74E-01	3,05E-02	1,59E-01	3,51E-02
#	LRP5	7	TGFBR2	2	30%	69%	8,32E-74	3,42E-84	0,00E+00	8,52E-02	3,51E-02	7,41E-02	0,00E+00
#	ERBB3	7	KRAS	7	27%	50%	3,62E-01	4,18E-77	4,65E-03	1,56E-01	2,34E-02	1,48E-01	0,00E+00

	AIC Relation				Bootstrap		p-values			Errors			
							Selective Advantage		Data	Prediction		Posterior classification	
	selects	#	selected	#	Non-parametric	Statistical	Temporal Priority	Probability Raising	Hypergeometric	μ	σ	μ	σ
1	APC	10	KRAS	7	51%	100%	8,53E-34	5,84E-75	4,27E-02	1,56E-01	2,34E-02	2,59E-01	0,00E+00
2	APC	10	MEMO	8	48%	100%	3,44E-18	1,93E-64	9,10E-02	4,15E-01	4,88E-02	3,96E-01	3,92E-02

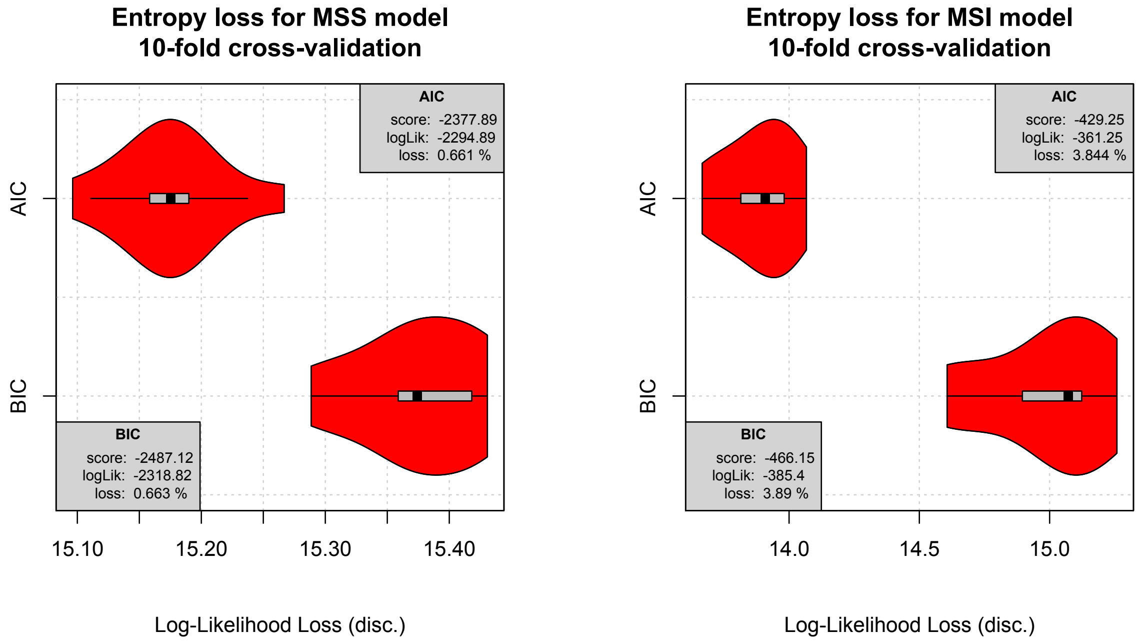
Event type	
Deletion	🟦
Amplification	🟥
Mutation	🟩
Pattern	⬜

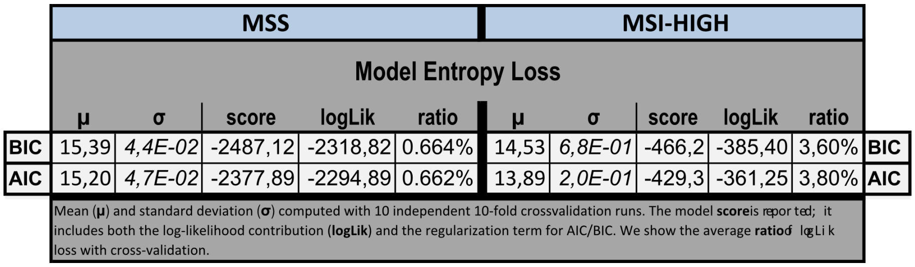
Non-parametric minimum score	
BIC	30%
AIC	50%
approximate	

p-values
< 0,01
0,01 - 0,05
> 0,05

error
(mean μ, stdev σ)
< 0,1
0,1 - 0,2
> 0,2

Figure S8: **COADREAD statistics for models confidence.** For BIC models we show statistics for edges with non-parametric bootstrap score approximately greater than 30%, for AIC models those greater than 50%s. *i) p-values* (100 repetition of non-parametric bootstrap, prior to Wilcoxon testing) for each edge statistics of selective advantage (*direction and statistical dependence, and hypergeometric*). In general, the edges that we selected show very strong support ($p \ll 10^{-10}$), but for those edges connecting events with the same marginal frequencies, where we can not be confident in the edge direction ($p > 0.05$) but still we find strong statistical dependence. (ii) *A posteriori* model confidence against Type I and II errors estimated with *non-parametric and statistical bootstraps* (100 repetitions) – edges annotated in Figures S6 and S7. (iii) Values of *posterior classification and prediction errors* are estimated from 10 repetitions of 10-fold *cross-validation*. The former reports how much error is due to predicting, for each set of edges $X = \{x_1, \ldots, x_n\} \to y$, the value of y according to the value of each $x_i \in X$. The latter reports the same statistics when we predict y from the whole set of parents X.





	MSS					MSI-HIGH					
	Model Entropy Loss										
	μ	σ	score	logLik	ratio	μ	σ	score	logLik	ratio	
BIC	15,39	*4,4E-02*	-2487,12	-2318,82	0.664%	14,53	*6,8E-01*	-466,2	-385,40	3,60%	**BIC**
AIC	15,20	*4,7E-02*	-2377,89	-2294,89	0.662%	13,89	*2,0E-01*	-429,3	-361,25	3,80%	**AIC**

Mean (**μ**) and standard deviation (**σ**) computed with 10 independent 10-fold crossvalidation runs. The model **score** is reported; it includes both the log-likelihood contribution (**logLik**) and the regularization term for AIC/BIC. We show the average **ratio** of logLik loss with cross-validation.

Figure S9: **Entropy loss for MSI-HIGH/MSS models.** Violin plot computed from 10 runs of k-fold cross-validation with $k = 10$, where we compute the "loss of log-likelihood" at each fold. In the plot and in the table we report also the overall log-likelihood, as well as the BIC and AIC scores for the models. We present the ratio of log-likelihood loss as a measure of stability of these models for these two datasets – we can observe that the MSS models lose $< 1\%$ of their likelihood, while the MSI lose slightly more (still, $< 4\%$), possibly because of the smaller sample size. From a statistical point of view, the greater (despite small) loss of likelihood by the models regularized via BIC confirms its tendency to underestimate the true model (i.e., the model should have false negatives, which could be AIC's edges).

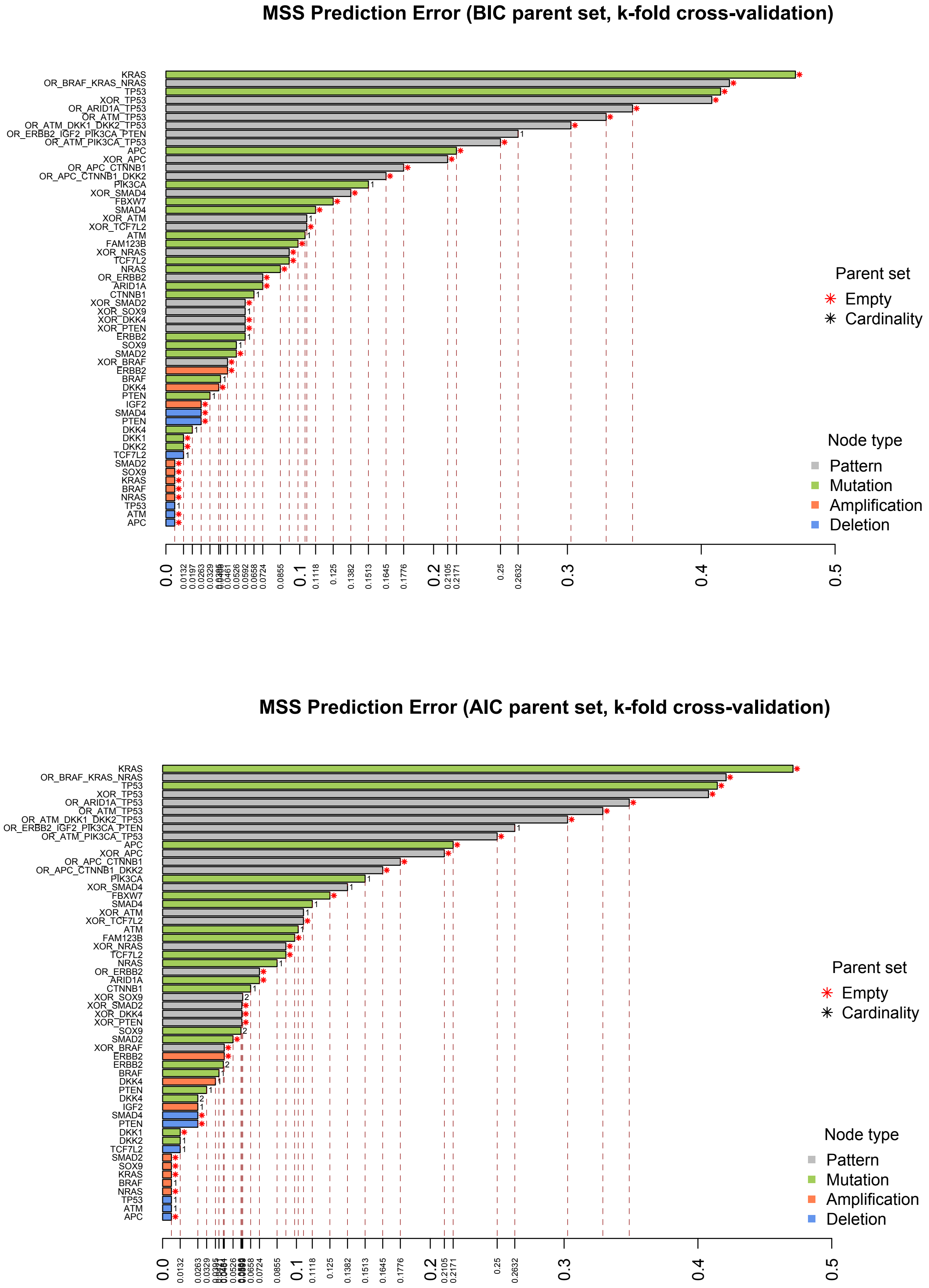


Figure S10: **Prediction error for each parent set of BIC and AIC models of MSS tumors.**

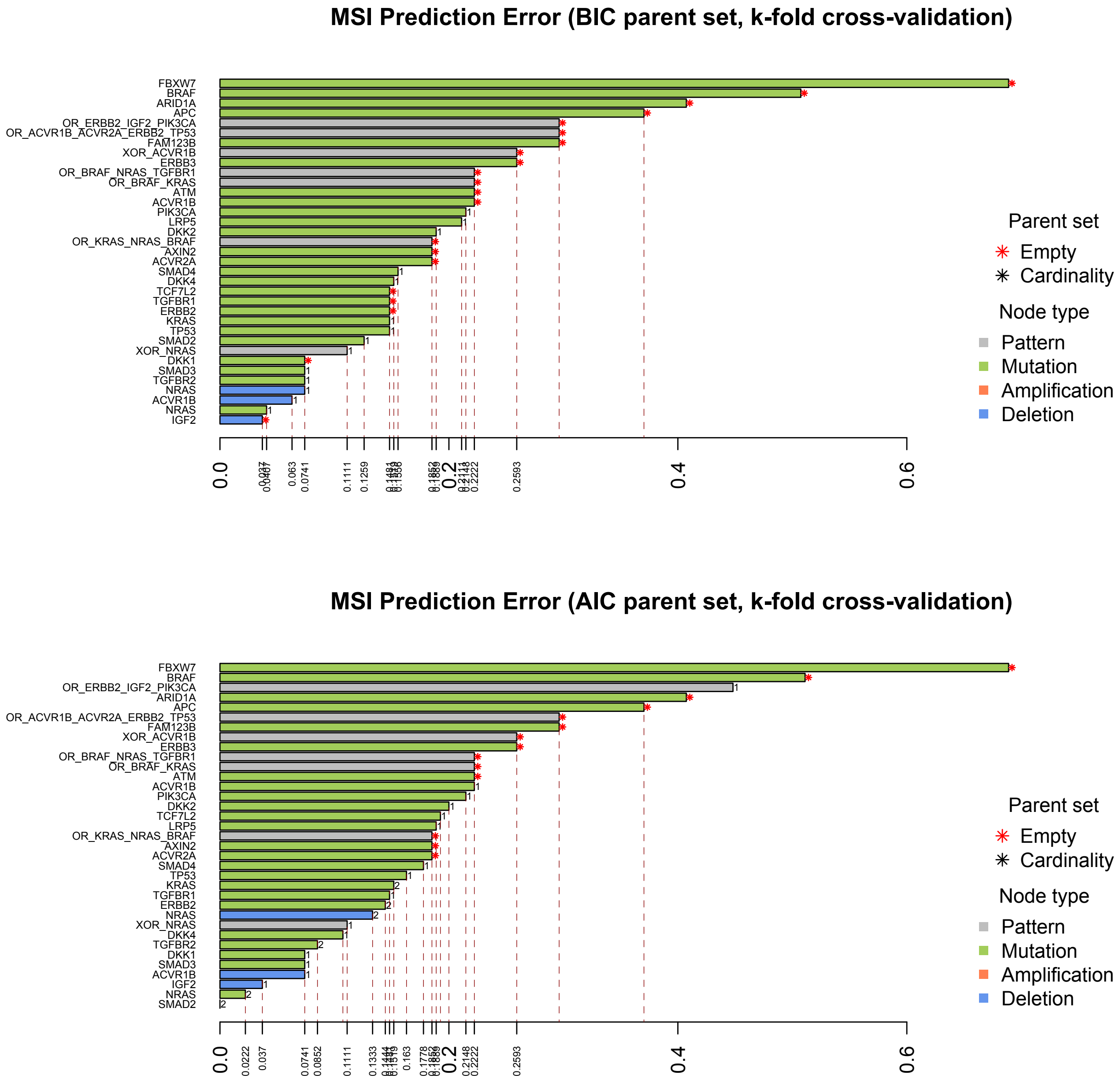
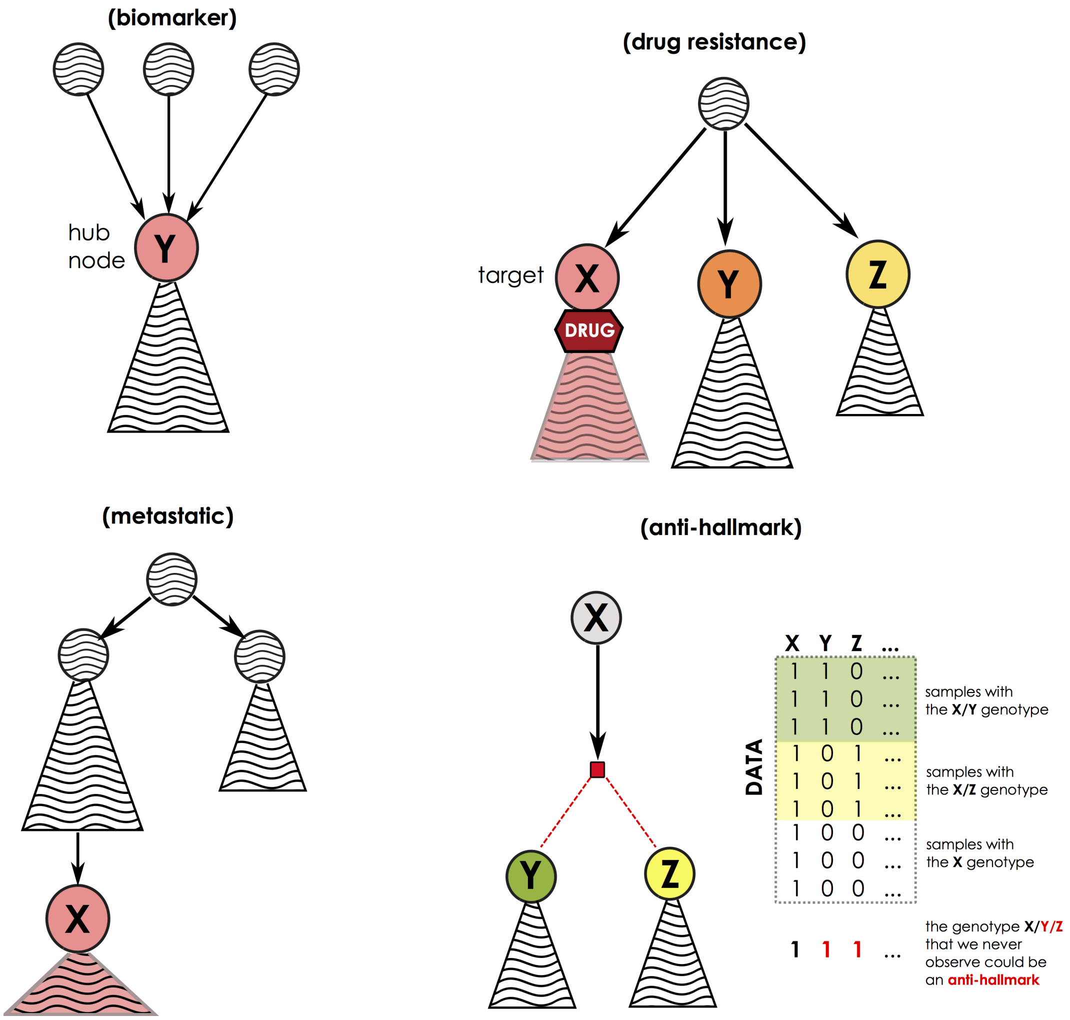


Figure S11: **Prediction error for each parent set of BIC and AIC models of MSI-HIGH tumors.**


(biomarker)
hub node
Y
(drug resistance)
target
X
DRUG
Y
Z
(metastatic)
X
(anti-hallmark)
X
Y
Z
DATA
samples with the X/Y genotype
samples with the X/Z genotype
samples with the X genotype
the genotype X/Y/Z that we never observe could be an anti-hallmark

X	Y	Z	...
1	1	0	...
1	1	0	...
1	1	0	...
1	0	1	...
1	0	1	...
1	0	1	...
1	0	0	...
1	0	0	...
1	0	0	...
1	1	1	...

Figure S12: **Models and the phenotype that they might explain. (biomarker)** Independent evolutionary trajectories depicted by a model might share common routes through a certain alteration Y; that could point to a new biomarker harbored by most of the tumors under study. **(drug resistance)** When a progression model branches in many independent sub-progressions, each one identified by alterations X, Y and Z, if a certain drug is known to target only a certain type of such clones (e.g., those where biomarker X is present), we might get insights on which are the biomarkers which make the drug ineffective for certain patients (e.g., those were cancer evolves through Y and Z). **(metastatic)** When a model is extracted from data representative of various tumor stages, we might discover which "late events" are those conferring a metastatic phenotype to a tumor – X in the figures. **(anti-hallmarks)** Relation between anti-hallmarks and formulas. Exclusivity formulas allow to capture fitness-equivalent events (Y and Z in the figure), and the presence of alternative routes – here those identified by the genotypes X/Y or X/Z. These could point us to genotype/phenotype that we do no observe in our cohort – here the X/Y/Z – which could be exploited for targeted therapy if a synthetic lethality is screened among Y and Z, the anti-hallmark.

References

[1] 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, 10.1093/bioinformatics/btw035*, 2016.

[2] The Cancer Genome Atlas Network et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, 2012.

[3] Timothy M Pawlik, Chandrajit P Raut, and Miguel A Rodriguez-Bigas. Colorectal carcinogenesis: MSI-H versus MSI-L. *Disease Markers*, 20(4-5):199–206, 2004.

[4] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013.

[5] Amy V Gerstein, Teresa Acosta Almeida, Guojing Zhao, Eric Chess, Ie-Ming Shih, Kent Buhler, Kenneth Pienta, Mark A Rubin, Robert Vessella, and Nickolas Papadopoulos. APC/CTNNB1 (β-catenin) pathway alterations in human prostate cancers. *Genes, Chromosomes and Cancer*, 34(1):9–16, 2002.

[6] Chen-Hsiang Yeang, Frank McCormick, and Arnold Levine. Combinatorial patterns of somatic gene mutations in cancer. *The FASEB Journal*, 22(8):2605–2622, 2008.

[7] 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, 2015.

[8] Loes Olde Loohuis, Giulio Caravagna, Alex Graudenzi, Daniele Ramazzotti, Giancarlo Mauri, Marco Antoniotti, and Bud Mishra. Inferring tree causal models of cancer progression with probability raising. *PLOS ONE*, 9(12):e115570, 2014.

[9] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[10] Joseph Felsenstein. *Inferring phylogenies*, volume 2. Sinauer Associates Sunderland, 2004.

[11] Aniko Szabo and Kenneth Boucher. Estimating an oncogenetic tree when false negatives and positives are present. *Mathematical biosciences*, 176(2):219–236, 2002.

[12] Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. CRC press, 1994.

[13] Pamela S Soltis and Douglas E Soltis. Applying the bootstrap in phylogeny reconstruction. *Statistical Science*, 18(2):256–267, 2003.

[14] Joseph Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, pages 783–791, 1985.

[15] Bradley Efron, Elizabeth Halloran, and Susan Holmes. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences*, 93(23):13429–13429, 1996.

[16] Marco Scutari. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software*, 35(i03), 2010.

[17] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[18] Giovanni Ciriello, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research*, 22(2):398–406, 2012.

[19] 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), 2015.