Conditional prediction of consecutive tumor evolution using cancer progression models: What genotype comes next?

Juan Diaz-Colunga[1,2,3] and Ramon Diaz-Uriarte[1,2,†]

[1]*Dpt. of Biochemistry, School of Medicine, Universidad Autónoma de Madrid, Madrid, Spain*
[2]*Instituto de Investigaciones Biomédicas 'Alberto Sols' (UAM-CSIC), Madrid, Spain*
[3]*Department of Ecology & Evolutionary Biology and Microbial Sciences Institute, Yale University, New Haven, CT, USA*
[†]*To whom correspondence should be addressed:* ramon.diaz@iib.uam.es

May 10, 2021

Abstract

Accurate prediction of tumor progression is key for adaptive therapy and precision medicine. Cancer progression models (CPMs) can be used to infer dependencies in mutation accumulation from cross-sectional data and provide predictions of tumor progression paths. However, their performance when predicting complete evolutionary trajectories is limited by violations of assumptions and the size of available data sets. Instead of predicting full tumor progression paths, here we focus on short-term predictions, more relevant for diagnostic and therapeutic purposes. We examine whether five distinct CPMs can be used to answer the question "Given that a genotype with n mutations has been observed, what genotype with n + 1 mutations is next in the path of tumor progression?" or, shortly, "What genotype comes next?". Using simulated data we find that under specific combinations of genotype and fitness landscape characteristics CPMs can provide predictions of short-term evolution that closely match the true probabilities, and that some genotype characteristics can be much more relevant than global features. Application of these methods to 25 cancer data sets shows that their use is hampered by a lack of information needed to make principled decisions about method choice. Fruitful use of these methods for short-term predictions requires adapting method's use to local genotype characteristics and obtaining reliable indicators of performance; it will also be necessary to clarify the interpretation of the method's results when key assumptions do not hold.

1 Introduction

Accurately predicting tumor evolution would allow for better prognosis and treatment. Unfortunately, the stochastic components in key factors that govern tumor progression (mutation, genetic drift or clonal selection [38, 46, 67], as well as non-genetic variability [11] and the tumor microenvironment [1]) fundamentally limit predictability. This is also common to other problems [44, 47] such as the evolution of antibiotic resistance [64] or virulence [20], where even small increases in predictive power could be of critical importance for the design of treatment strategies [54].

Cross-sectional high-throughput data are widely available for many cancer types, and Cancer Progression Models (CPMs) [8, 9], originally developed to identify restrictions in the order of accumulation of mutations in tumors using cross-sectional data, could thus be key tools for predicting cancer progression. CPMs (implicitly) encode all the possible mutational trajectories that are compatible with the restrictions and, for some methods, provide probabilities of different paths (figure 1A-B). However their performance to infer the complete evolutionary trajectories, from the root genotype with no driver mutations to a final, fixated genotype, is limited [27, 40]; this has been attributed mainly to violations of the underlying assumptions of the models (fitness landscapes with a single fitness maximum in the genotype with all loci mutated, strong selection week mutation regime [36]) and the difficulty to acquire large, unbiased data sets to power them [27].

Predicting the complete evolutionary path, however, might not be the most relevant objective. Detection of a specific tumor composition at a given point in time automatically discards all mutational pathways that

do not visit the observed state, reducing the amount of possible evolutionary trajectories that can be followed from then on and thus potentially facilitating the task of predicting further tumor progression. In this work we address the question of how accurately CPMs can predict short-term cancer evolution after a genotype is detected: what is the state that comes consecutively after or "what genotype comes next". Even if the predictability of the complete path is low because of violations of method assumptions, genotype-specific features could lead to local improvements in short-term predictability [6, 21, 30]. Moreover, in contrast to the more general, non-local descriptors, these genotype-specific factors can provide more detailed insight regarding the requirements for the methods to be reliable. In fact, conditional predictions on the observed genotype are arguably much more relevant than complete evolutionary path predictions for both adaptive therapy [e.g., 39, 48, 61] and for SMART (sequential, multiple assignment, randomized trial) designs in dynamic treatment regimes for personalized medicine [17, 65].

[Fig. 1 about here]

The question of what genotype comes next, or what is the state that comes consecutively after, requires that we carefully define the meaning of "consecutive" states in this context. If cancer progresses under conditions of strong selection and weak mutation (SSWM) [36], only evolutionary trajectories of increasing fitness are possible, and the evolution of a new, higher-fitness clone results in rapid expansion and effective exclusion of previous ones before any new mutations take place. In this situation, there is a well defined sequence of genotypes with progressively increasing fitness in the path to tumor fixation, with the most abundant clone at every time being a direct descendant of a previously most abundant one itself (figure 1C). But, in cancer, large population sizes and high mutation rates can prevent successful clonal sweeps [52, 70] so that instead of SSWM, clonal interference and stochastic tunneling [22, 43, 59, 63, 68] become common: multiple clones can coexist at significant fractions for prolonged periods of time, and new higher-fitness clones can evolve from low fitness parents that represented a minority of the population (Fig 1D).

To formulate the question of "what genotype comes next" so that it does not depend on the SSWM assumption we designate an *observed* genotype as that which is most abundant in the tumor at the time of sampling. On the other hand, a Line of Descent (LOD) is defined as a sequence of parent-child genotypes from the wild-type to the (possibly local) fitness maximum where fixation occurs [63]. By construction, assuming that mutations accumulate one by one and cannot be lost, each genotype in the LOD has all the mutations of the previous one plus exactly one more (as it directly descends from it). The LOD has no more than one genotype with a given number of mutations, whereas in general two or more observable genotypes could share the same mutation count (e.g. in figure 1D both the red and green genotypes have one mutated driver and are observable). If the evolutionary process does not conform to SSWM, not all genotypes in the LOD are necessarily observable. From a clinical perspective, it is relevant to use the information that is detectable to unveil the underlying dynamics of cancer evolution, encoded in the LOD. In other words, we want to make a prediction regarding *what comes next in the LOD* based on the *observation* of a genotype. We thus formulate the following general question: *Given that a genotype with* n *mutations has been observed, what genotype with* n + 1 *mutations is in the LOD?* Formulated like this, the question is unambiguous and determined. Under SSWM conditions it can be expressed simply as: *given that a certain genotype has been observed, what genotype comes next in the path to tumor fixation?* However, it is important to emphasize that even if there is no SSWM the more general question remains well defined.

To address this question, we use a collection of evolutionary simulations from which we obtain the true paths of tumor progression. For every simulated process we keep track of the most frequent genotypes. Then for each one of them we ask what genotype has exactly one more mutation while also being a direct ancestor of the ultimately fixated clone. Averaging over all simulations we obtain a set of frequencies which represent the conditional probabilities that we are interested in. We then examine five CPMs: Mutual Hazard Networks (MHN) [58], Conjunctive Bayesian Networks (CBN) [34, 35, 50], Oncogenetic Trees (OT) [23, 62], CAncer PRogression Inference (CAPRI) [16, 57] and CAncer PRogression Extraction with Single Edges (CAPRESE) [53], some of which have more than one variant (giving a total of 13 different methods). We provide these methods with cross-sectional samples taken from the evolutionary simulations and compare their predictions with the true probabilities. We explore the factors that affect performance, focusing on those factors that depend on the properties of the observed genotype. We want to understand whether CPMs can be used to predict short-term evolution; for this question, the key response variable is the best possible performance over all CPMs considered. Although we do address method choice as a secondary question, this is not our main objective *per se*. We discuss the consequences of our results for the analysis of cancer data by analyzing 25 cancer data sets.

2 Methods

2.1 Overview

To assess if we can use CPMs for short-term predictions of evolutionary processes, we have compared predicted and true answers to the question *Given that a genotype with* n *mutations has been observed, what genotype with* n + 1 *mutations is in the LOD?* under an extensive set of simulated scenarios. First, we simulated a large number of tumor evolution trajectories under scenarios with varying departures from the assumptions of CPM methods; next, we obtained a large number of simulated tumor evolution trajectories. Next, we emulated cross-sectional sampling of the simulations, with varying sampling schemes and sample sizes. These steps are described in section Simulated evolutionary processes and sampling. The transition probabilities that provide the true answer to the question *Given that a genotype with* n *mutations has been observed, what genotype with* n + 1 *mutations is in the LOD?* is explained in section Transition probabilities from evolutionary simulations. To obtain the CPM predictions we analyzed each of the sampled data sets with a total of 13 different CPM variants and obtained the predictions as explained in section Transition probabilities from CPMs. Next, we measured how close the predictions were to the truth (section Quantification of similarity between predicted and true transition probabilities) and, finally, we analyzed the major factors that affect the similarity between predictions and truth (section Factors that affect predictive performance: linear mixed-effects model trees). To examine the implications of our results for the analyses of cancer data, we have analyzed 25 cancer data sets, described in section Cancer data sets.

2.2 Simulated evolutionary processes and sampling

Simulated data were taken from [27], where complete details are provided. In summary, simulations were conducted under evolutionary scenarios that differed in the number of genes, the type of fitness landscape, the initial population size, and the mutation rates. Landscapes were of either 7 or 10 genes. Fitness landscapes were of three types: representable, local maxima or Rough Mount Fuji (RMF). The three types of fitness landscapes incorporate increasing departures from the assumptions of most CPMs: for the representable fitness landscapes a DAG of restrictions exists with the same accessible genotypes and accessible mutational paths; for local maxima fitness landscapes the set of accessible genotypes can be represented by a DAG of restrictions, but there are local fitness maxima and the fitness graph has missing paths [27]; the genotype with all genes mutated might or might not be the genotype with largest fitness. The RMF fitness landscapes [22, 33, 51] usually have multiple local fitness maxima and considerable reciprocal sign epistasis so not even the set of accessible genotypes can be represented by a DAG of restrictions. On the above fitness landscapes evolutionary simulations were run. Three different populations sizes (2×10^3, 5×10^4 or 1×10^6) and two mutation rate regimes (mutation rate fixed at 10^{-5} or uniformly distributed in the log scale between 0.2×10^{-5} and 5×10^{-5}) were used to generate variability with respect to SSWM. For each combination of the above factors, 35 independent replicates were obtained for a total of 1260 landscapes (35 replicates $\times$ 2 number of genes $\times$ 3 types of fitness landscapes $\times$ 3 initial population sizes $\times$ 2 mutation rate regimes). Then, for each one of them, 20000 independent evolutionary processes were simulated until fixation of one of the genotypes at a fitness maximum (local or global). The complete history of the simulations was stored so that the true LOD for each simulation was known. Each set of simulations was then sampled under three detection regimes: with equal probability with respect to the size distribution (*uniform* sampling), sampling predominantly early in the tumor progression process (i.e., enriched in *small* tumors), sampling predominantly late in the tumor progression process (i.e., enriched in *large* tumors). When *sampling*, we identified the most abundant genotype and its mutated loci at the chosen time, so the sampling process is consistent with how we defined *observable* genotypes earlier. No sampling error is considered. The full sample consists of three binary matrices of observations (one per detection regime), each with 20000 rows (one observation per simulation) and as many columns as driver genes. Each element of the matrix is either 1 or 0 (true or false) depending on which genes were mutated in which observations. For each detection regime, the complete sample was used to generate five smaller, non-overlapping splits of size 50, 200 or 4000. This produced a total of 56700 data sets (1260 fitness landscapes $\times$ 3 detection regimes $\times$ 3 sample sizes $\times$ 5 splits).

2.3 Transition probabilities from evolutionary simulations

Let us consider a matrix **P** where rows correspond to observable genotypes and columns to genotypes of the LOD. The element (i, j) of the matrix, denoted as p_{ij}, represents the conditional probability that the genotype j (with $n+1$ mutations) is an element of the LOD given the genotype i (with n mutations) has been observed. In the absence of back mutations and when mutations accumulate one by one, p_{ij} is defined only if i and j are exactly one mutation away from each other. Note that the genotype j (element of the LOD and with $n+1$ mutations) need not be a direct descendant of i if SSWM does not hold. We built a transition matrix **P** for every fitness landscape that we generated by examining each of the 20000 evolutionary processes that were simulated in the corresponding landscape, listing all observable genotypes (i.e. those that represented the majority of the population at some point during the simulated evolutionary process) and extracting the LOD of every process tracking back the ancestors of the fixated genotype (definition and details in [63]; using implementation in [25]). Each element p_{ij} can thus be interpreted as the fraction of the times that j was in the LOD given that an observation of i had been made. See further details in Supplementary Material, section 1.5.

2.4 Transition probabilities from CPMs

General overviews of all the methods used here, except for MHN, are available in [8, 9, 26, 27] and detailed descriptions are given in the original references for each procedure (MHN [58], CBN [34, 35], MCCBN [50], OT [23, 62], CAPRI [16, 57], CAPRESE [53]). Briefly, CBN, MCCBN, OT, CAPRESE, and CAPRI try to identify restrictions in the order of accumulation of mutations from cross-sectional data. CAPRESE and OT code these restrictions using trees (i.e., under these two methods each event, such as a mutation, can immediately depend on at most one other event); although the output of OT and CAPRESE is often similar, the procedures are different (in OT weights along edges can be directly interpreted as probabilities of transition [62], whereas CAPRESE uses a probability raising notion of causation based on Suppes' probabilistic causation). CBN, MCCBN, and CAPRI code these restrictions using DAGs (i.e., each event can depend on multiple other events). CBN and MCCBN are very closely related methods: the CBN version we use is H-CBN as described in [34, 35], and MCCBN is an implementation of the CBN model that uses a Monte Carlo Expectation-Maximization (EM) algorithm for fitting the CBN model [50] (instead of the simulated annealing nested within EM of H-CBN); in contrast, CAPRI tries to identify probability raising in the framework of Suppes' probabilistic causation. Common to most of these methods is a model of deterministic dependencies [58], a model of accumulation of mutations where an event (a mutation) can only occur if all its dependencies are satisfied (though different methods allow for small error deviations from this requirement); CAPRI and CAPRESE try to recover "probability raising" relations and, strictly, their trees/DAGs do not encode deterministic dependencies but will be treated as doing so in this paper, since obtaining probabilities of transition from them is not possible otherwise —see details in section Transition probabilities from CPMs: supplementary details in the Supplementary Material.

For all of these methods, since a DAG of restrictions determines a fitness graph (see Fig. 1 A-B and [26, 27]), we can obtain the set of possible descendants for a genotype directly from this fitness graph. For CAPRESE and CAPRI all transitions were set as equiprobable (as probabilities of transition are not available) but for CBN, MCCBN, and OT we can compute the probabilities of transition between genotypes. CBN and MCCBN return the parameters of the waiting time to occurrence of each mutation given its restrictions are satisfied which allows us to compute transition probabilities between genotypes using competing exponentials (see [27, 40, 50]) and a similar procedure can be used with OT by using the edge weights (though, strictly, the OT model, being an untimed oncogenetic tree does not return these probabilities —see [27]). Mutual Hazard Networks (MHN) [58] differ from the previous methods in that events are modeled by a spontaneous rate of fixation and a multiplicative effect each of these events can have on other events (i.e., pairwise interactions), which allows it to model both enabling and inhibiting dependencies. The output of MHN is a transition rate matrix; as for CBN and MCCBN, we can obtain the probability of transition to each descendant genotype, given a transition, using competing exponentials from the transition rate matrix (see equation 2 and Fig. 2 in [58]). Details on software and computation of transition probabilities are provided in the Supplementary Material, sections CPMs: software and usage details and Transition probabilities from CPMs: supplementary details.

None of the CPM models considered here allow us to accommodate local maxima: under all of the models, the genotype with all genes mutated is always reached with probability 1 as time goes to infinity (this is

also the case for MHN, even if it can include inhibiting effects). To overcome this limitation without making specific additional assumptions about the sampling process, and for MHN, CBN, and MCCBN, we have used the uniformization method [37, 66] (see also [58, Fig. 6]) to approximate the continuous-time Markov chain ($\mathbf{Q}$) by a discrete time Markov chain (with transition matrix = $\mathbf{I} + \frac{1}{\gamma}\mathbf{Q}$, where $\gamma = \max(|q_{ii}|)$). We will interpret the diagonal entries of the time-discretized transition matrix as a lower bound on the probability that the given genotypes behave as local maxima. The output from CBN, MHN, and MCCBN after time discretization will be referred to as CBN_td, MHN_td, and MCCBN_td, respectively, and we will differentiate this **time-discretized** results from those obtained, as explained above, using competing exponentials. We will refer to these methods as the **TD** methods (for time-discretized) whereas their counterparts, CBN, MHN, MCCBN, obtained via competing exponentials on the rate matrix, woll be referred below as the CE methods. OT, CAPRESE, and CAPRI do not belong to any of those two groups, though none of them can accommodate local maxima either; thus, when we refer to the "non-TD" methods, that includes MHN, CBN, MCCBN (in their "non-TD" versions) as well as OT, CAPRESE, and CAPRI.

In summary, a total of 13 methods have been used: CBN, in three variants (CBN, with transition probabilities using competing exponentials conditional on transition; CBN_uw, obtained from the former by using equiprobable transitions; CBN_td, the time-discretized version of CBN); MCCBN, again in three variants (MCCBN; MCCBN_uw; MCCBN_td); MHN in two variants (MHN and MHN_td); OT in variants OT and OT_uw (equiprobable transitions); CAPRESE; CAPRI, in two variants, CAPRI_AIC, CAPRI_BIC (which differ on the penalty used, AIC or BIC). Each of these methods returns a matrix $\hat{\mathbf{P}}$ of predicted transition probabilities.

2.5 Quantification of similarity between predicted and true transition probabilities

For every observed genotype i we have compared the vector of true probabilities $\mathbf{p}_i$, from the ith row of $\mathbf{P}$ with the analogous vector of predictions from a CPM $\hat{\mathbf{p}}_i$ returned from the ith row of $\hat{\mathbf{P}}$ for each CPM. We measured the similarity between these two vectors using the Jensen-Shannon distance (JS) [19], the square root of the Jensen-Shannon divergence [45]. The Jensen-Shannon divergence is a symmetrized Kullback-Leibler divergence between two probability distributions that is 0 when the two distributions are identical and reaches its maximum value of 1 if the two distributions do not overlap (we used log of base 2). See further details in Supplementary Material, section 1.4.

2.6 Factors that affect predictive performance: linear mixed-effects model trees

To examine the relevance of the different factors on the performance of methods we used linear mixed-effects model trees. These are an extension of recursive partitioning (or tree-based) methods, where observations are first split repeatedly according to the predictor variables, which play the role of partitioning variables, so that the dependent variable becomes more homogeneous within each node; the linear mixed-effects model allows for the addition of random effects [31, 32]. In our models, the dependent variable is the minimal JS over all methods (for each genotype by replicate by id by detection regime by sample size). The possible partitioning variables include characteristics that are fitness landscape- and evolutionary dynamics-specific, sampling characteristics, genotype (within fitness landscape)-specific, and genotype (within fitness landscape) by sampling (sample size and detection regime) by replicate-specific variables. The fitness landscape- and evolutionary dynamics-specific variables are SSWM regime (`SSWM` in the figures, measured as the average frequency of the most frequent genotype), γ (`gamma` —this is not the same gamma as used in the time-discretization of transition rate matrices) a measure of the amount of epistasis in the fitness landscape, defined as "the correlation of fitness effects of the same mutation in single-mutant neighbors" [29] or, equivalently, "the correlation in fitness effects between genotypes that only differ by 1 locus, averaged across the [fitness] landscape" [12], fraction of pairs of loci with reciprocal sign epistasis (`epistRSign`), and number of observed peaks in the fitness landscape (`numObservedPeaks`). The sampling characteristics are detection regime (`detect` in the figures) and sample size (`sample_size`) —recall that each fitness landscape was sampled over all nine combinations of detection regime and sampling size. The genotype (within fitness landscape)-specific variables are number of mutations (`nMut` in the figures), fitness rank (`fitnessRank`, where the genotype of maximum fitness in a landscape has fitnessRank 1), and the proportion of times the genotype is a (possibly local) fitness maximum (`propLocalMax`) . The genotype (within fitness landscape) by sampling by replicate specific variables are proportion of times that genotype was observed in the sample (`observedProp`), and difference between the observed proportion and the true

frequency of that genotype in the samples (where the true frequency is the frequency over the 20000 samples) (`diff_obs_prop`). Random effects for the global model are ID and, nested within ID, the crossed random effects of genotype and replicate. As the meaning of the number of mutations, nMut, is different in models with 7 and 10 genes, we fitted separate models for simulations with 7 and 10 genes. Models were fitted using the R package "glmertree" [32]. The Bonferroni-corrected significance level for node-splitting was set at 0.01. We specified a minimal (weighted) size in leaves (or terminal nodes) of 1%; as the returned trees had over 50 leaves in all cases, to allow for interpretability we pruned the resulting tree by recursively merging (starting from the leaves) all children node with a fitted minimal JS > 0.0677 (which corresponds to the 20% best JS —see Table S4 in Supplementary Material). In addition, we merged all children with a fitted minimal JS that differed by less than 0.0677/4 (so as to collapse good performing nodes with minor differences). Further details about the fitting procedure are provided in the Supplementary Material section 1.7.

2.7 Cancer data sets

We have used a total of 25 data sets. Twenty two of them were used in [27]. These 22 data sets include six different cancer types (breast, glioblastoma, lung, ovarian, colorectal, and pancreatic cancer), use different types of features (nonsynonymous somatic mutations, copy number alterations, or both) analyzed in terms of pathways, functional modules, genes, gene events, or mutations (yielding from 3 to 192 different features), and have samples sizes from 27 to 594; the original sources are [5, 10, 13–15, 28, 41, 42, 55, 56, 69] and all have been used previously in CPM research [3, 16, 18, 27, 35, 49, 53, 57]; for further details about those 22 data sets, see the supplementary material (file S5_Text) in [27]. We have also used three CGH data sets previously used in [34, 58] that were originally obtained from the Progenetix database [7]. These are `Breast-CGH`, from 871 breast cancer patients with 10 CGH alterations, `Renal-CGH`, from 251 renal cell carcinoma with 12 CGH alterations, and `Colon-CGH`, from 570 colorectal cancers with 11 CGH alterations. These three data sets were obtained from the Supplementary data of [58]. Thus, these 25 cancer data sets constitute a wide, representative example of data to which researchers have applied or might want to apply CPMs. We have analyzed all the data sets with all the methods considered; because computing time increases nonlinearly with number of features for some methods, and in particular CBN (and, to a lesser extent, MHN), have difficulties with 16 or more features, for the data sets with more than 15 features, we have used the common procedure [24] of selecting the 15 most frequent features.

3 Results

Figure 2 shows the liner mixed-effects model tree for fitness landscapes with seven genes under the uniform detection regime and with the largest sample size (4000). In the Supplementary Material (Fig. S1, section 3.4) we show the results for 10 genes. Figures S2 and S3, in the Supplementary Material, present the models fitted to all detection regimes and sample sizes. In these figures we can see the relevance of detection regime and sample size: the smallest JS were generally achieved under the uniform detection regime and with the largest sample size (4000) (see also Supplementary material, Tables S1 to S7). Moreover, there were interactions between detection regime and sample size on the one hand and genotype-specific and fitness landscape-specific variables on the other; these interactions can make it harder to understand the relevance of fitness landscape-specific and genotype-specific variables on the quality of predictions. Figures 2 and S1 isolate the effects of fitness landscape and genotype-specific variables, which is one of the main objectives of this paper.

[Fig. 2 about here]

Best achievable performance is generally larger (i.e., JS is lower) in the representable than in the local maxima and RMF fitness landscapes, as can be seen in node 4 of Fig. 2 (see also Tables S2, S3). Moreover, gamma, reciprocal sign epistasis, and number of observed peaks, three of the variables that play important roles in the fitted trees (e.g., nodes 5, 11, 14, 15, and 23 in Fig. S2; node 5 in Fig. S1; nodes 12 and 15 in Fig. S3), allow to perfectly separate the RMF from the rest of the fitness landscape, and to separate most local maxima fitness landscapes from the representable ones (Fig. S4 in the Supplementary Material). The results indicate, though, that highly specific fitness landscape characteristics (fitness landscape itself, gamma, reciprocal sign epistasis, number of observed peaks) show complex interactions with genotype-specific characteristics. For example, in Fig. 2, fitness landscape plays a role in node 4 only for genotypes

with small fitness rank (i.e., large relative fitness of a genotype compared to the rest of the genotypes in the fitness landscape) and that are rarely local maxima. Likewise, number of observed peaks in node 19, which would separate representable from local maxima and RMF landscapes, is relevant for the one-before-last genotype. Similar interactions between genotype-specific and fitness landscape-specific characteristics can be seen in the figures in the Supplementary Material (Fig. S2: node 23, child of genotype-specific nodes 1 and 19; Fig. S3: nodes 15 and 12, mediated by genotype-specific nodes 1, 11, and 14; Fig. S1: node 5, affected and which affects ancestor and descendant nodes of genotype-specific characteristics). A general observation is that deviations from SSWM lead to poorer predictive ability; these effects can depend both on general fitness landscape characteristics (e.g., node 5 in Fig. 2) and genotype-specific characteristics (e.g., node 21 in Fig. S2 and node 19 in Fig. S3 in Supplementary Material).

Not all genotype-specific characteristics had a relevant effect. Observed proportion of a genotype and the difference between its observed proportion and true frequency in the sample are variables that do not appear in the final trees (although the former did appear in the models before prunning). Number of mutations had a non-linear effect: genotypes with either none or close to maximum number of mutations lead to better performance, whereas genotypes with intermediate number of mutations had worse performance (e.g., nodes 1 and 14 in Fig. 2, nodes 1, 3, 6 in Fig. S1, nodes 2 and 6 in Fig. S3). This indicates that prediction becomes harder when the number of potential next genotypes increases, which happens at intermediate numbers of mutations. Smaller fitnessRank (i.e., large relative fitness of a genotype) was generally associated to better predictions (e.g., nodes 9 and 15 in Fig. 2; nodes 12, 13, and 4 in Fig. S1) but, as with most other characteristics, the effect was strongly affected by other variables (node 1 in Fig. S2, nodes 11 and 21 in Fig. S3). Note that genotypes of fitness rank 1 are necessarily fitness maxima.

In fact, how often a genotype is a local fitness maximum (propLocalMax) not only affected performance but also which methods were better. This is apparent in node 3 in Fig. 2, node 2 in Fig. S1, node 19 in Fig. S2, and node 1 in Fig. S3: when a genotype was frequently (over 20 to 50%, depending on the data set) a local maximum, the TD methods lead to much better performance that the other methods. Moreover, once the frequency of local maxima was large, making it even larger lead to better performance (node 11 in Fig. 2; nodes 14, 17, and 19 in Fig. S1; node 19 in Fig. S2; node 14 in Fig. S3). This phenomenon can also be looked at from the perspective of global fitness landscape characteristics: an increase in the number of local fitness peaks (numObservedPeaks) can lead to the TD methods being better than their CE counterparts (node 19 in Fig. 2) or to the TD methods improving their performance (node 15 in Fig. S3).This is also observed with increasing reciprocal sign epistasis (episRSign —nodes 5, 11, 14 in Fig. S2), a parameter that can almost perfectly separate the RMF and local maxima landscapes from the representable ones (see Fig. S4). Note that epistRSign no longer plays a role in a switch between methods in node 23, since this node only splits genotypes in fitness landscapes that cannot be representable (genotypes that have a probability of being local maxima > 0 even when their fitness rank is > 1).

Scenarios with very good TD performance (JS less than approximately 0.1) comprise, then, between 1 and 5% of the cases (4% in nodes 10 and 14 in Fig. 2; 5% in nodes 16, 20, and 21 in Fig. S1; 1% in node 24 of Fig. S2; 2% in nodes 18 and 22 of Fig. S3; see also Tables S4 to S11). In these cases, MHN_td was generally slightly better performing than CBN_td (and MCCBN_td was generally the worse of the TD methods). Within the non-TD methods, when they could do well, most methods had roughly similar performance (except for CAPRI, that was normally the worse performer), with a slight advantage of CBN and MHN. CBN was generally among the best performers, and scenarios where MHN was the single best method were rare (but see node 20 in Fig. 2). For methods with both a weighted and an unweigthed version (CBN_uw, MCCBN_uw, OT_uw) using weights never lead to worse performance and sometimes improved it. As with the TD methods, the CE methods were consistently very good only in limited scenarios (see also Tables S4 to S11. For CE methods these were nodes 7 and 20 in Fig. 2 which comprised about 7% of the (weighted) observations, and slightly worse in nodes 6 and 16 —where MHN's performance was surprisingly poor— that comprised another 6% of the observations. Similar patterns are seen in Figures S1 (node 23, with 5% of observations), S2 (node 17, and with worse performance 6, 12, 16, comprising overall 6% of observations) and S3 (node 10 and, with worse performance, 9, both with 1% of the observations).

4 Discussion: implications for the analysis of biological data sets

Can we use cancer progression models (CPMs) to predict the short-term evolution of a tumor? Informally, we could word this question as "what genotype comes next in the path to tumor fixation?". More precisely,

we can ask whether, conditional on us observing a particular genotype with n mutations as the most abundant genotype in a tumor, we can predict what genotype with n + 1 mutations is in the Line of Descent (LOD) to the fitness maximum (where fixation occurs). We have examined the performance of 13 different CPM procedures for predicting short-term evolution, focusing mainly on the best achievable performance over all 13 methods (i.e., on how close to the truth are the predictions of short-term evolution, regardless of method) and, secondarily, on the choice of method. The analysis of simulated data indicated that it was possible to provide predictions of short-term evolution that closely matched the true probabilities of short-term evolution only under some specific combinations of genotype characteristics and fitness landscape characteristics. And when good performance was possible, no method or set of methods was consistently superior, making choice of best method also dependent on genotype and fitness landscape properties.

Best achievable performance was affected by complex interactions between fitness landscape-specific characteristics, deviations from the strong selection and weak mutation (SSWM) regime, sampling, and genotype-specific characteristics. The best performance was seen with large sample sizes, a uniform detection regime, with representable fitness landscapes and under a SSWM regime. Genotypes with maximum or close to maximum fitness (i.e., small fitness rank) with either 0 or close to the maximum number of mutations, are those for which predictions tend to be better. For scenarios where the best achievable performance was good (i.e., where it matters), these three factors also affected which methods were the best. There were sharply contrasting differences between methods, in particular, methods that can model local maxima —the TD methods MHN_td and CBN_td— tended to either both perform very well or both perform very bad. The corresponding two methods that do not model local maxima —MHN and CBN— generally present the opposite pattern, performing well when the TD methods do not and vice-versa.

Since our focus are predictions conditional on the observed genotype, we can also summarize the results by dividing genotypes into three groups: A) Genotypes for which using the TD methods lead to very good predictions (JS < 0.1); B) Genotypes for which the CE methods can provide very good predictions (JS < 0.1); C) The remaining cases, which constitute the majority of the cases, where performance was generally poor (JS > 0.25). Group A consists of genotypes that are more often than not local maxima (probability of being a local maximum > 0.5) and of large fitness (small fitnessRank) or that are often local maxima (probability of being a local maximum > 0.25) while deviations from SSWM are minor. Fitness landscape-specific characteristics that make these genotypes more likely to be observed are large number of local fitness peaks and large reciprocal sign epistasis. For these genotypes, the TD methods can correctly predict being at the end of the evolutionary process, something that the remaining methods cannot account for, which explains the very poor performance of the non-TD methods in these cases (JS > 0.75). In group B are genotypes of large relative fitness, no mutations or with all except one loci mutated, in evolutionary scenarios with SSWM and smooth fitness landscapes (little reciprocal sign epistasis, large gamma) and few local peaks (less than 3). For these genotypes, the TD methods make very poor predictions (JS > 0.75) while MHN and CBN do particularly good jobs, but CAPRESE and OT, two methods that can only model simpler patterns of dependencies, generally perform well too. In the remaining cases (group C), no method can provide good predictions. Assessing whether we can achieve good performance and, if so, deciding which method to use, requires detailed information about the genotype, the fitness landscape, and the evolutionary dynamics. For some small subset of scenarios performance could be very good but it could also be dismal if we use the incorrect method.

Our results for short-term predictions highlight variables that have been found to be relevant for long-term prediction of the complete evolutionary paths. Sampling regime and sample size have previously been found relevant by [27, 40] (see also [24, 26] for the effect of sampling regime in inferring the DAGs of restrictions *per se*). Likewise, evolutionary scenarios with clonal interference in rugged fitness landscapes (small gamma, large reciprocal sign epistasis, large number of local fitness maxima) lead to poorer performance of CPMs compared to SSWM regimes in single-peaked fitness landscapes [27, 40]. The role of smooth fitness landscapes for CPM's predictive ability is directly related to the CPM assumption that final fixation occurs in the genotype with all loci mutated and their inability to model reciprocal sign epistasis [26] (this problem also affects MHN, specially if there are higher-order epistatic interactions). Interestingly, these are the fitness landscapes where the TD methods (which allow us to capture local maxima) perform better than the CE methods. The role of SSWM for CPMs has been discussed before by [27, 60] and [40] used it explicitly for predicting tumor evolution. In fact, all of the methods considered here implicitly assume SSWM mainly for two reasons. First, CPMs are generally applied to cross sectional data from bulk sequencing (not single-cell sequencing) without undergoing any deconvolution step (i.e. without attempting to infer the clonal composition of the sequenced tumor), which reflects the implicit assumption

that an actual existing genotype can be obtained from bulk sequencing (see [2] for discussion of how this assumption can fail with bulk samples as we collapse over existing genotypes); if this assumption does not hold, the question "what genotype comes next" is undefined because we are conditioning on a non-existing genotype. Second, the observed genotypes are considered true steps in the path to tumor fixation, thus neglecting the possibility of tunneling effects [22, 43, 59, 63]. These two assumptions are true if SSWM holds (see Fig. 1, and note that the reverse is not necessarily true), in which case the most fit genotype at any given time is effectively the only one in the population, thus the only one observed, and an ancestor of the ultimately fixated one. Even when CPMs specifically include error estimations (e.g. [34]), these typically refer to sequencing error but not necessarily to deviations from SSWM. And our reformulation of the question "what genotype comes next" as we did in section Introduction allows us to deal with the second violation, but of course not with the first. In our analysis of the simulated data the first assumption holds (we sample true, existing genotypes); yet, our results show that using CPMs to predict short-term evolution when the second assumption does not hold resulted in poorer performance. It must be emphasized, though, that the above comments refer to the effects of deviations from SSWM on the performance of the methods considered, not on predictability itself (the role of SSWM on predictability itself is more nuanced: see, e.g., [4, 6, 22, 30, 43, 63, 70]).

In addition to the above factors, which affect both short- and long-term predictions, our results highlight that when we want to make conditional, short-term predictions, features of the genotypes for which we make the predictions can be as relevant as, or more relevant than, fitness landscape-specific features (see also [6, 21, 30] for a comparison of short-term vs. long-term predictability). We have already mentioned the combined importance of fitness rank and probability of being a local maximum, and the latter, together with reciprocal sign epistasis, illustrate factors that can increase short-term predictability while decreasing global-path predictability. The role of number of mutations, where predictions are worse from genotypes with intermediate numbers of mutations, can be understood as a consequence of the inverted U-shape relationship between the number of entries that can be non-zero in $\mathbf{P}$ and the number of mutations in a genotype ($\binom{N}{n+1}$: see section Transition probabilities from evolutionary simulations): the problem of prediction becomes harder at intermediate numbers of mutations. The role of genotype-specific factors is a noteworthy feature because it suggests that CPMs might be leveraged to provide short-term predictions even when global, long-term predictions are difficult because fitness landscape- and evolutionary model-specific assumptions are violated.

What are the implications of the above results for the analysis of cancer data? We have compared the predictions from the best performing methods on 25 biological data sets: even if we do not know the true transition probabilities, we can examine the consequences of assuming that we choose one method when another method is the one that gives the best (or even perfect) predictions. This is what we do in Fig. 3. The left panel of Fig. 3A, for example, shows the difference in predictions (measured using JS) if the method that gave perfect predictions were CBN but we had used MHN (or viceversa). If we found that, for a given genotype, two or more methods gave similar predictions (i.e. comparing their predicted probabilities yielded a small JS) we could argue that they could all be potentially providing good results. On the other hand, very different predictions would indicate that at least one of them (if not all) is failing, and in that case there are no solid grounds to choose one or the other since we lack access to the underlying factors that affect performance. As can be seen, for most data sets there is wide variability in the consequences of using the wrong method, and often the difference in predictions is very large, specially if we use a TD vs. a non-TD method (MHN vs. MHN_td, right-most panel). The high variability between method predictions shown in Fig. 3A —even across genotypes within a same dataset— highlights the importance of conditioning predictions to the observed genotype, as different methods can make very similar or substantially different predictions depending on it. Particularly in the latter case, if any of the methods was indeed providing accurate predictions, determining which one is the correct one would depend on the properties of the observed genotype.

[Figs. 3 and 4 about here]

However, in general we do not have access to the underlying parameters that would allow us to make an informed decision on method (or method family) choice. So how do the above results contribute to the applied analysis of actual cancer data? Fig. 3B shows the dissimilarity between the predictions of two specific CE methods (CBN and MHN), two specific TD methods (CBN_td and MHN_td), and two methods from different groups (MHN and MHN_td). Many genotypes accumulate in the area of the plot where the predictions of the two CE methods, as well as those of the two TD methods, are very different, with variable similarity between CE and TD predictions. But a few genotypes escape this trend. Looking back

at the simulations, we have seen that there is a sharp difference in behavior of TD and CE methods when the best achievable performance is good; and necessarily CBN and MHN (or CBN_td and MHN_td) can only both make correct predictions if both are making similar predictions. This same pattern —predictions of methods within the same group are similar and predictions of methods from different groups are not— is observed in the biological data for a small subset of genotypes. What about limiting our interpretation to those genotypes? This approach by itself will most likely yield only modest improvements in performance (see Tables **S8** to **S11** showing that, for TD, the proportion of cases with very good TD performance could at most increase from an unconditional 0.04 to a conditional 0.06, and the proportion of cases with very good performance of CE from an unconditional 0.04 to a conditional 0.16). Fig. 4 shows the fraction of genotypes, for each data set, that meet fairly stringent conditions of similarity either within the CE methods, i.e. between CBN and MHN, or within the TD methods, i.e. between CBN_td and MHN_td (see Supplementary Material, figures in Section 3.6, for similarity for each individual genotype for each data set). For 11 of the data sets, there are more than 10% of the genotypes that fall in either of these categories: which among the predictions that fulfill those similarity conditions should we trust? The difficulty of this choice is even more dramatic when the same genotype simultaneosuly shows large similarity within TD methods, large similarity within CE methods, and large dissimilarity between the CE and TD methods, as indicated by the gray points in Fig. 4. In these cases, predictions between families of methods are very different and yet method (or method family) choice would require fitness landscape-specific, evolutionary model-specific and genotype-specific information.

Even if this information is generally unknown in practice, our results provide a foundation to narrow down the search for genotypes for which good performance is possible. Furthermore, we have demonstrated that we do not need complete information on the general properties of the fitness landscape; instead, even partial information on those specific genotypes could serve to make informed decisions on method choice. Gathering information on the properties of individual genotypes is, in principle, a more reasonable objective than obtaining a full description of the complete fitness landscape for every cancer type.

4.1 Conclusion

CPMs could, under very specific combinations of genotype characteristics and fitness landscape characteristics, be used to obtain good predictions of the short-term evolution of a tumor, even when long-term predictions are not possible because of violations of assumptions of CPM procedures. But method choice and assessing if the predictions obtained are to be trusted, for example to guide therapeutic decisions, requires detailed knowledge which is not available with empirical data. More generally, our work shows the promises and practical difficulties of using cross-sectional data in an evolutionary context to guide individualized therapies, even when cross-sectional data recapitulate the equivalent of many repeated evolutionary processes. Exploiting the potential of these methods will require further work to, first, examine the interpretation of their output and the consequences of their usage when key assumptions do not hold; second, identify if the methods' output, by itself or in combination with data set features, can provide indicators of performance; and third, develop strategies to determine the characteristics of individual genotypes in real tumors and adapt method use consequently. For all three tasks, it will be important to make explicit the relationship between assumptions with respect to the evolutionary model (e.g., SSWM), fitness landscape characteristics (e.g., local maxima and reciprocal sign epistasis), sampling regimes (late vs. early tumors), and predictions, all in the context of well defined concepts such as lines of descent and for specified tasks such as interventions.

5 Acknowledgments

We thank A. Zeileis for answers about the use of the glmertree package and R. Schill for answers about the use of MHN, and R. Guantes and C. Lazaro-Perea for comments on the ms.

6 Funding

Supported by BFU2015-67302-R (MINECO/FEDER, EU) and PID2019-111256RB-I00/AEI/10.13039/501100011033 to RDU. JDC supported by PEJD-2018-POST/BMD-8960 from Comunidad de Madrid.

7 Data & code availability

Code for the analyses in this article is available at https://github.com/rdiaz02/what_genotype_next.

References

1. Albini, A., Sporn, M. B., 2007. The tumour microenvironment as a target for chemoprevention. *Nature Reviews Cancer*, **7(2)**:139–147. doi:10.1038/nrc2067. URL https://www.nature.com/articles/nrc2067.

2. Alves, J. M., Prieto, T., Posada, D., 2017. Multiregional Tumor Trees Are Not Phylogenies. *Trends in Cancer*, **3(8)**:546–550. doi:10.1016/j.trecan.2017.06.004. URL http://www.cell.com/trends/cancer/abstract/S2405-8033(17)30121-8.

3. Attolini, C., Cheng, Y., Beroukhim, R., Getz, G., Abdel-Wahab, O., Levine, R. L., Mellinghoff, I. K., Michor, F., 2010. A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proceedings of the National Academy of Sciences*, **107(41)**:17604–17609. doi:10.1073/pnas.1009117107/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1009117107. URL http://www.pnas.org/content/107/41/17604.short.

4. Bailey, S. F., Blanquart, F., Bataillon, T., Kassen, R., 2017. What drives parallel evolution? *BioEssays*, **39(1)**:e201600176. doi:10.1002/bies.201600176. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201600176.

5. Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P. A., Stratton, M. R., Wooster, R., 2004. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer*, **91(2)**:355–358. doi:10.1038/sj.bjc.6601894. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2409828/.

6. Bank, C., Matuszewski, S., Hietpas, R. T., Jensen, J. D., 2016. On the (un)predictability of a large intragenic fitness landscape. *PNAS*, **113(49)**:14085–14090. doi:10.1073/pnas.1612676113. URL http://www.pnas.org/content/113/49/14085.

7. Baudis, M., Cleary, M. L., 2001. Progenetix.net: An online repository for molecular cytogenetic aberration data. *Bioinformatics*, **17(12)**:1228–1229. doi:10.1093/bioinformatics/17.12.1228. URL https://academic.oup.com/bioinformatics/article/17/12/1228/225653.

8. Beerenwinkel, N., Greenman, C. D., Lagergren, J., 2016. Computational Cancer Biology: An Evolutionary Perspective. *PLoS Comput. Biol.*, **12(2)**:e1004717. doi:10.1371/journal.pcbi.1004717. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004717.

9. Beerenwinkel, N., Schwarz, R. F., Gerstung, M., Markowetz, F., 2015. Cancer evolution: Mathematical models and computational inference. *Systematic Biology*, **64(1)**:e1–e25. doi:10.1093/sysbio/syu081. URL http://sysbio.oxfordjournals.org/content/64/1/e1.

10. Brennan, C. W., Verhaak, R. G. W., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., Zheng, S., Chakravarty, D., Sanborn, J. Z., Berman, S. H., Beroukhim, R., Bernard, B., Wu, C.-J., Genovese, G., Shmulevich, I., Barnholtz-Sloan, J., Zou, L., Vegesna, R., Shukla, S. A., Ciriello, G., Yung, W. K., Zhang, W., Sougnez, C., Mikkelsen, T., Aldape, K., Bigner, D. D., Van Meir, E. G., Prados, M., Sloan, A., Black, K. L., Eschbacher, J., Finocchiaro, G., Friedman, W., Andrews, D. W., Guha, A., Iacocca, M., OtextquoterightNeill, B. P., Foltz, G., Myers, J., Weisenberger, D. J., Penny, R., Kucherlapati, R., Perou, C. M., Hayes, D. N., Gibbs, R., Marra, M., Mills, G. B., Lander, E., Spellman, P., Wilson, R., Sander, C., Weinstein, J., Meyerson, M., Gabriel, S., Laird, P. W., Haussler, D., Getz, G., Chin, L., Benz, C., Barrett, W., Ostrom, Q., Wolinsky, Y., Bose, B., Boulos, P. T., Boulos, M., Brown, J., Czerinski, C., Eppley, M., Kempista, T., Kitko, T., Koyfman, Y., Rabeno, B., Rastogi, P., Sugarman, M., Swanson, P., Yalamanchii, K., Otey, I. P., Liu, Y. S., Xiao, Y., Auman, J. T., Chen, P.-C., Hadjipanayis, A., Lee, E., Lee, S., Park, P. J., Seidman, J., Yang, L., Kalkanis, S., Poisson, L. M., Raghunathan, A., Scarpace, L., Bressler, R., Eakin, A., Iype, L., Kreisberg, R. B., Leinonen, K., Reynolds, S., Rovira, H., Thorsson, V., Annala, M. J., Paulauskis, J., Curley, E., Hatfield, M., Mallery, D., Morris, S., Shelton, T., Shelton, C., Sherman, M., Yena, P., Cuppini, L., DiMeco, F., Eoli, M., Maderna, E., Pollo, B., Saini, M., Balu, S., Hoadley, K. A., Li, L., Miller, C. R., Shi, Y., Topal, M. D., Wu, J., Dunn, G., Giannini, C., Aksoy, B. A., Antipin, Y., Borsu, L., Cerami, E., Gao, J., Gross, B., Jacobsen, A., Ladanyi, M., Lash, A., Liang, Y., Reva, B., Schultz, N., Shen, R., Socci, N. D., Viale, A., Ferguson, M. L., Chen, Q.-R., A, D., Dillon, L. A. L., Shaw, K. R. M., Sheth, M., Tarnuzzer, R., Wang, Z., Yang, L., Davidsen, T., Guyer, M. S., Ozenberger, B. A., Sofia, H. J., Bergsten, J., Eckman, J., Harr, J., Smith, C., Tucker,

K., Winemiller, C., Zach, L. A., Ljubimova, J. Y., Eley, G., Ayala, B., Jensen, M. A., Kahn, A., Pihl, T. D., Pot, D. A., Wan, Y., Hansen, N., Hothi, P., Lin, B., Shah, N., Yoon, J.-g., Lau, C., Berens, M., Ardlie, K., Carter, S. L., Cherniack, A. D., Noble, M., Cho, J., Cibulskis, K., DiCara, D., Frazer, S., Gabriel, S. B., Gehlenborg, N., Gentry, J., Heiman, D., Kim, J., Jing, R., Lander, E. S., Lawrence, M., Lin, P., Mallard, W., Onofrio, R. C., Saksena, G., Schumacher, S., Stojanov, P., Tabak, B., Voet, D., Zhang, H., Dees, N. N., Ding, L., Fulton, L. L., Fulton, R. S., Kanchi, K.-L., Mardis, E. R., Wilson, R. K., Baylin, S. B., Harshyne, L., Cohen, M. L., Devine, K., Sloan, A. E., VandenBerg, S. R., Berger, M. S., Carlin, D., Craft, B., Ellrott, K., Goldman, M., Goldstein, T., Grifford, M., Ma, S., Ng, S., Stuart, J., Swatloski, T., Waltman, P., Zhu, J., Foss, R., Frentzen, B., McTiernan, R., Yachnis, A., Mao, Y., Akbani, R., Bogler, O., Fuller, G. N., Liu, W., Liu, Y., Lu, Y., Mills, G., Protopopov, A., Ren, X., Sun, Y., Yung, W. K. A., Zhang, J., Chen, K., Weinstein, J. N., Bootwalla, ., 2013. The Somatic Genomic Landscape of Glioblastoma. *Cell*, **155(2)**:462–477.

11. Brock, A., Chang, H., Huang, S., 2009. Non-genetic heterogeneity — a mutation-independent driving force for the somatic evolution of tumours. *Nature Reviews Genetics*, **10(5)**:336–342. doi:10.1038/nrg2556.

12. Brouillet, S., Annoni, H., Ferretti, L., Achaz, G., 2015. MAGELLAN: A tool to explore small fitness landscapes. *bioRxiv*, p. 031583. doi:10.1101/031583. URL http://biorxiv.org/content/early/2015/11/13/031583.

13. Cancer Genome Atlas Research Network, 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455(7216)**:1061–1068.

14. Cancer Genome Atlas Research Network, 2011. Integrated genomic analyses of ovarian carcinoma. *Nature*, **474(7353)**:609–615. doi:10.1038/nature10166.

15. Cancer Genome Atlas Research Network, 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487(7407)**:330–337. doi:10.1038/nature11252. URL https://www.nature.com/articles/nature11252.

16. Caravagna, G., Graudenzi, A., Ramazzotti, D., Sanz-Pamplona, R., Sano, L. D., Mauri, G., Moreno, V., Antoniotti, M., Mishra, B., 2016. Algorithmic methods to infer the evolutionary trajectories in cancer progression. *PNAS*, **113(28)**:E4025–E4034. doi:10.1073/pnas.1520213113. URL http://www.pnas.org/content/113/28/E4025.

17. Chakraborty, B., Moodie, E. E. M., 2013. *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. Springer, edición: 2013 edition.

18. Cheng, Y.-K., Beroukhim, R., Levine, R. L., Mellinghoff, I. K., Holland, E. C., Michor, F., 2012. A mathematical methodology for determining the temporal order of pathway alterations arising during gliomagenesis. *PLoS computational biology*, **8(1)**:e1002337. doi:10.1371/journal.pcbi.1002337. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3252265&tool=pmcentrez&rendertype=abstract.

19. Crooks, G. E., 2017. On measures of entropy and information. Technical report. URL http://threeplusone.com/on_information.pdf.

20. Day, T., Proulx, S. R., 2004. A general theory for the evolutionary dynamics of virulence. *The American Naturalist*, **163(4)**:E40–E63. doi:10.1086/382548.

21. de Visser, J. A. G. M., Elena, S. F., Fragata, I., Matuszewski, S., 2018. The utility of fitness landscapes and big data for predicting evolution. *Heredity*, **121(5)**:401. doi:10.1038/s41437-018-0128-4. URL https://www.nature.com/articles/s41437-018-0128-4.

22. de Visser, J. A. G. M., Krug, J., 2014. Empirical fitness landscapes and the predictability of evolution. *Nat Rev Genet*, **15(7)**:480–490. doi:10.1038/nrg3744. URL http://www.nature.com/nrg/journal/v15/n7/full/nrg3744.html.

23. Desper, R., Jiang, F., Kallioniemi, O. P., Moch, H., Papadimitriou, C. H., Schäffer, A. A., 1999. Inferring tree models for oncogenesis from comparative genome hybridization data. *J Comput Biol*, **6(1)**:37–51. URL http://view.ncbi.nlm.nih.gov/pubmed/10223663.

24. Diaz-Uriarte, R., 2015. Identifying restrictions in the order of accumulation of mutations during tumor progression: Effects of passengers, evolutionary models, and sampling. *BMC Bioinformatics*, **16(41)**. doi:doi:10.1186/s12859-015-0466-7. URL http://www.biomedcentral.com/1471-2105/16/41/abstract.

25. Diaz-Uriarte, R., 2017. OncoSimulR: Genetic simulation with arbitrary epistasis and mutator genes in asexual populations. *Bioinformatics*, **33(12)**:1898–1899. doi:10.1093/bioinformatics/btx077. URL https://academic.oup.com/bioinformatics/article/33/12/1898/2982052/OncoSimulR-genetic-simulation-with-arbitrary.

26. Diaz-Uriarte, R., 2018. Cancer progression models and fitness landscapes: A many-to-many relationship. *Bioinformatics*, **34(5)**:836–844. doi:10.1093/bioinformatics/btx663. URL https://academic.oup.com/bioinformatics/article/34/5/836/4557185.

27. Diaz-Uriarte, R., Vasallo, C., 2019. Every which way? On predicting tumor evolution using cancer progression models. *PLOS Computational Biology*, **15(8)**:e1007246. doi:10.1371/journal.pcbi.1007246. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007246.

28. Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D. M., Morgan, M. B., Fulton, L., Fulton, R. S., Zhang, Q., Wendl, M. C., Lawrence, M. S., Larson, D. E., Chen, K., Dooling, D. J., Sabo, A., Hawes, A. C., Shen, H., Jhangiani, S. N., Lewis, L. R., Hall, O., Zhu, Y., Mathew, T., Ren, Y., Yao, J., Scherer, S. E., Clerc, K., Metcalf, G. A., Ng, B., Milosavljevic, A., Gonzalez-Garay, M. L., Osborne, J. R., Meyer, R., Shi, X., Tang, Y., Koboldt, D. C., Lin, L., Abbott, R., Miner, T. L., Pohl, C., Fewell, G., Haipek, C., Schmidt, H., Dunford-Shore, B. H., Kraja, A., Crosby, S. D., Sawyer, C. S., Vickery, T., Sander, S., Robinson, J., Winckler, W., Baldwin, J., Chirieac, L. R., Dutt, A., Fennell, T., Hanna, M., Johnson, B. E., Onofrio, R. C., Thomas, R. K., Tonon, G., Weir, B. A., Zhao, X., Ziaugra, L., Zody, M. C., Giordano, T., Orringer, M. B., Roth, J. A., Spitz, M. R., Wistuba, I. I., Ozenberger, B., Good, P. J., Chang, A. C., Beer, D. G., Watson, M. A., Ladanyi, M., Broderick, S., Yoshizawa, A., Travis, W. D., Pao, W., Province, M. A., Weinstock, G. M., Varmus, H. E., Gabriel, S. B., Lander, E. S., Gibbs, R. A., Meyerson, M., Wilson, R. K., 2008. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, **455(7216)**:1069–1075. doi:10.1038/nature07423. URL https://www.nature.com/articles/nature07423.

29. Ferretti, L., Schmiegelt, B., Weinreich, D., Yamauchi, A., Kobayashi, Y., Tajima, F., Achaz, G., 2016. Measuring epistasis in fitness landscapes: The correlation of fitness effects of mutations. *Journal of Theoretical Biology*, **396**:132–143. doi:10.1016/j.jtbi.2016.01.037. URL http://www.sciencedirect.com/science/article/pii/S0022519316000771.

30. Ferretti, L., Weinreich, D., Tajima, F., Achaz, G., 2018. Evolutionary constraints in fitness landscapes. *Heredity*, **121(5)**:466. doi:10.1038/s41437-018-0110-1. URL https://www.nature.com/articles/s41437-018-0110-1.

31. Fokkema, M., Edbrooke-Childs, J., Wolpert, M., 2020. Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data. *Psychotherapy Research*, **0(0)**:1–13. doi:10.1080/10503307.2020.1785037. URL https://doi.org/10.1080/10503307.2020.1785037.

32. Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., Kelderman, H., 2018. Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behav Res*, **50(5)**:2016–2034. doi:10.3758/s13428-017-0971-x. URL https://doi.org/10.3758/s13428-017-0971-x.

33. Fragata, I., Blanckaert, A., Dias Louro, M. A., Liberles, D. A., Bank, C., 2019. Evolution in the light of fitness landscape theory. *Trends in Ecology & Evolution*, **34(1)**:69–82. doi:10.1016/j.tree.2018.10.009. URL https://linkinghub.elsevier.com/retrieve/pii/S0169534718302660.

34. Gerstung, M., Baudis, M., Moch, H., Beerenwinkel, N., 2009. Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics*, **25(21)**:2809–2815. doi:10.1093/bioinformatics/btp505. URL http://dx.doi.org/10.1093/bioinformatics/btp505http://www.bsse.ethz.ch/cbg/software/ct-cbn.

35. Gerstung, M., Eriksson, N., Lin, J., Vogelstein, B., Beerenwinkel, N., 2011. The Temporal Order of Genetic and Pathway Alterations in Tumorigenesis. *PLoS ONE*, **6(11)**:e27136. doi:10.1371/journal.pone.0027136. URL http://dx.plos.org/10.1371/journal.pone.0027136http://www.bsse.ethz.ch/cbg/software/ct-cbn.

36. Gillespie, J. H., 1983. A simple stochastic gene substitution model. *Theor Popul Biol*, **23(2)**:202–215. doi:10.1016/0040-5809(83)90014-x. URL https://www.sciencedirect.com/science/article/pii/004058098390014X.

37. Grassmann, W. K., 1977. Transient solutions in markovian queueing systems. *Computers & Operations Research*, **4(1)**:47–53. doi:10.1016/0305-0548(77)90007-7. URL http://www.sciencedirect.com/science/article/pii/0305054877900077.

38. Greaves, M., 2015. Evolutionary Determinants of Cancer. *Cancer Discovery*, **5(8)**:806–820. doi:10.1158/2159-8290.CD-15-0439. URL http://cancerdiscovery.aacrjournals.org/content/5/8/806.

39. Hansen, E., Read, A. F., 2020. Modifying adaptive therapy to enhance competitive suppression. *bioRxiv*, p. 2020.10.26.355701. doi:10.1101/2020.10.26.355701. URL https://www.biorxiv.org/content/10.1101/2020.10.26.355701v1.

40. Hosseini, S.-R., Diaz-Uriarte, R., Markowetz, F., Beerenwinkel, N., 2019. Estimating the predictability of cancer evolution. *Bioinformatics*, **35(14)**:i389–i397. doi:10.1093/bioinformatics/btz332. URL https://academic.oup.com/bioinformatics/article/35/14/i389/5529151.

41. Jones, S., Zhang, X., Parsons, D. W., Lin, J. C.-H., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A., Hong, S.-M., Fu, B., Lin, M.-T., Calhoun, E. S., Kamiyama, M., Walter, K., Nikolskaya, T., Nikolsky, Y., Hartigan, J., Smith, D. R., Hidalgo, M., Leach, S. D., Klein, A. P., Jaffee, E. M., Goggins, M., Maitra, A., Iacobuzio-Donahue, C., Eshleman, J. R., Kern, S. E., Hruban, R. H., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V. E., Kinzler, K. W., 2008. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science (New York, N.Y.)*, **321(5897)**:1801–6. doi:10.1126/science.1164368. URL http://www.ncbi.nlm.nih.gov/pubmed/18772397.

42. Knutsen, T., Gobu, V., Knaus, R., Padilla-Nash, H., Augustud, M., Strausberg, R. L., Kirsch, I. R., Sirotkin, K., Ried, T., 2005. The Interactive Online SKY/M-FISH & CGH Database and the Entrez Cancer Chromosomes Search Database: Linkage of Chromosomal Aberrations with the Genome Sequence. *Genes, Chromosomes and Cancer*, **44(1)**:52–64.

43. Krug, J., 2019. Accessibility percolation in random fitness landscapes. *arXiv:1903.11913 [math, q-bio]*. URL http://arxiv.org/abs/1903.11913.

44. Lässig, M., Mustonen, V., Walczak, A. M., 2017. Predicting evolution. *Nature Ecology & Evolution*, **1(3)**:s41559–017–0077–017. doi:10.1038/s41559-017-0077. URL https://www.nature.com/articles/s41559-017-0077.

45. Lin, J., 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, **37(1)**:145–151.

46. Lipinski, K. A., Barber, L. J., Davies, M. N., Ashenden, M., Sottoriva, A., Gerlinger, M., 2016. Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. *Trends in Cancer*, **2(1)**:49–63. doi:10.1016/j.trecan.2015.11.003. URL http://linkinghub.elsevier.com/retrieve/pii/S2405803315000692.

47. Losos, J. B., 2018. *Improbable Destinies: Fate, Chance, and the Future of Evolution.* Riverhead Books, S.l. ISBN 978-0-525-53413-6.

48. Melnikov, S. V., Stevens, D. L., Fu, X., Kwok, H. S., Zhang, J.-T., Shen, Y., Sabina, J., Lee, K., Lee, H., Söll, D., 2020. Exploiting evolutionary trade-offs for posttreatment management of drug-resistant populations. *Proc Natl Acad Sci USA*, **117(30)**:17924–17931. doi:10.1073/pnas.2003132117. URL http://www.pnas.org/lookup/doi/10.1073/pnas.2003132117.

49. Misra, N., Szczurek, E., Vingron, M., 2014. Inferring the paths of somatic evolution in cancer. *Bioinformatics (Oxford, England)*, **30(17)**:2456–2463. doi:10.1093/bioinformatics/btu319. URL http://www.ncbi.nlm.nih.gov/pubmed/24812340.

50. Montazeri, H., Kuipers, J., Kouyos, R., Böni, J., Yerly, S., Klimkait, T., Aubert, V., Günthard, H. F., Beerenwinkel, N., Study, T. S. H. C., 2016. Large-scale inference of conjunctive Bayesian networks. *Bioinformatics*, **32(17)**:i727–i735. doi:10.1093/bioinformatics/btw459. URL http://bioinformatics.oxfordjournals.org/content/32/17/i727.

51. Neidhart, J., Szendro, I. G., Krug, J., 2014. Adaptation in Tunably Rugged Fitness Landscapes: The Rough Mount Fuji Model. *Genetics*, **198(2)**:699–721. doi:10.1534/genetics.114.167668. URL http://www.genetics.org/content/198/2/699.

52. Nichol, D., Robertson-Tessi, M., Anderson, A. R. A., Jeavons, P., 2019. Model genotype–phenotype mappings and the algorithmic structure of evolution. *Journal of The Royal Society Interface*, **16(160)**:20190332. doi:10.1098/rsif.2019.0332. URL https://royalsocietypublishing.org/doi/full/10.1098/rsif.2019.0332.

53. Olde Loohuis, L., Caravagna, G., Graudenzi, A., Ramazzotti, D., Mauri, G., Antoniotti, M., Mishra, B., 2014. Inferring Tree Causal Models of Cancer Progression with Probability Raising. *PLOS ONE*, **9(10)**:e108358. doi:10.1371/journal.pone.0108358. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0108358.

54. Palmer, A. C., Kishony, R., 2013. Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance. *Nature Reviews Genetics*, **14(4)**:243–248. doi:10.1038/nrg3351. URL https://www.nature.com/articles/nrg3351.

55. Parsons, D. W., Jones, S., Zhang, X., Lin, J. C.-H., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.-M., Gallia, G. L., Olivi, A., McLendon, R., Rasheed, B. A., Keir, S., Nikolskaya, T., Nikolsky, Y., Busam, D. A., Tekleab, H., Diaz, L. A., Hartigan, J., Smith, D. R., Strausberg, R. L., Marie, S. K. N., Shinjo, S. M. O., Yan, H., Riggins, G. J., Bigner, D. D., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V. E., Kinzler, K. W., 2008. An Integrated Genomic Analysis of Human Glioblastoma Multiforme. *Science*, **321(5897)**:1807–1812. doi:10.1126/science.1164382. URL http://science.sciencemag.org/content/321/5897/1807.

56. Piazza, R., Valletta, S., Winkelmann, N., Redaelli, S., Spinelli, R., Pirola, A., Antolini, L., Mologni, L., Donadoni, C., Papaemmanuil, E., Schnittger, S., Kim, D.-W., Boultwood, J., Rossi, F., Gaipa, G., De Martini, G. P., di Celle, P. F., Jang, H. G., Fantin, V., Bignell, G. R., Magistroni, V., Haferlach, T., Pogliani, E. M., Campbell, P. J., Chase, A. J., Tapper, W. J., Cross, N. C. P., Gambacorti-Passerini, C., 2013. Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. *Nature Genetics*, **45(1)**:18–24. doi:doi.org/10.1038/ng.2495. URL https://www.nature.com/articles/ng.2495.

57. Ramazzotti, D., Caravagna, G., Olde Loohuis, L., Graudenzi, A., Korsunsky, I., Mauri, G., Antoniotti, M., Mishra, B., 2015. CAPRI: Efficient inference of cancer progression models from cross-sectional data. *Bioinformatics*, **31(18)**:3016–3026. doi:10.1093/bioinformatics/btv296. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv296.

58. Schill, R., Solbrig, S., Wettig, T., Spang, R., 2020. Modelling cancer progression using Mutual Hazard Networks. *Bioinformatics*, **36(1)**:241–249. doi:10.1093/bioinformatics/btz513. URL https://academic.oup.com/bioinformatics/article/36/1/241/5524604.

59. Sniegowski, P. D., Gerrish, P. J., 2010. Beneficial mutations and the dynamics of adaptation in asexual populations. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **365(1544)**:1255–1263. doi:10.1098/rstb.2009.0290. URL http://rstb.royalsocietypublishing.org/content/365/1544/1255.

60. Sprouffske, K., Pepper, J. W., Maley, C. C., 2011. Accurate reconstruction of the temporal order of mutations in neoplastic progression. *Cancer prevention research (Philadelphia, Pa.)*, **4(7)**:1135–44. doi:10.1158/1940-6207.CAPR-10-0374. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3131446&tool=pmcentrez&rendertype=abstract.

61. Staňková, K., Brown, J. S., Dalton, W. S., Gatenby, R. A., 2019. Optimizing Cancer Treatment Using Game Theory: A Review. *JAMA Oncol*, **5(1)**:96–103. doi:10.1001/jamaoncol.2018.3395. URL https://jamanetwork.com/journals/jamaoncology/fullarticle/2696342.

62. Szabo, A., Boucher, K. M., 2008. Oncogenetic trees. In Tan, W.-Y., Hanin, L., editors, *Handbook of Cancer Models with Applications*, pp. 1–24. World Scientific. URL http://www.worldscibooks.com/lifesci/6677.html.

63. Szendro, I. G., Franke, J., de Visser, J. A. G. M., Krug, J., 2013. Predictability of evolution depends nonmonotonically on population size. *PNAS*, **110(2)**:571–576. doi:10.1073/pnas.1213613110. URL http://www.pnas.org/content/110/2/571.

64. Toprak, E., Veres, A., Michel, J.-B., Chait, R., Hartl, D. L., Kishony, R., 2012. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nature Genetics*, **44(1)**:101–105. doi:10.1038/ng.1034. URL https://www.nature.com/articles/ng.1034.

65. Tsiatis, A. A., Davidian, M., Holloway, S. T., Laber, E. B., 2020. *Dynamic Treatment Regimes Statistical Methods for Precision Medicine*. ISBN 978-1-4987-6977-8.

66. van Dijk, N. M., van Brummelen, S. P. J., Boucherie, R. J., 2018. Uniformization: Basics, extensions and applications. *Performance Evaluation*, **118**:8–32. doi:10.1016/j.peva.2017.09.008. URL http://www.sciencedirect.com/science/article/pii/S0166531617301451.

67. Williams, M. J., Werner, B., Heide, T., Curtis, C., Barnes, C. P., Sottoriva, A., Graham, T. A., 2018. Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics*, **50(6)**:895–903. doi:10.1038/s41588-018-0128-6. URL https://www.nature.com/articles/s41588-018-0128-6.

68. Wodarz, D., Komarova, N. L., 2014. *Dynamics of Cancer: Mathematical Foundations of Oncology*. ISBN 978-981-4566-36-0. doi:10.1142/8973. URL https://www.worldscientific.com/doi/abs/10.1142/8973.

69. Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P. A., Kaminker, J. S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J. K. V., Sukumar, S., Polyak, K., Park, B. H., Pethiyagoda, C. L., Pant, P. V. K., Ballinger, D. G., Sparks, A. B., Hartigan, J., Smith, D. R.,

Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S. D., Parmigiani, G., Kinzler, K. W., Velculescu, V. E., Vogelstein, B., 2007. The Genomic Landscapes of Human Breast and Colorectal Cancers. *Science*, **318(5853)**:1108–1113. doi:10.1126/science.1145720. URL http://dx.doi.org/10.1126/science.1145720.

70. Zhao, B., Hemann, M. T., Lauffenburger, D. A., 2016. Modeling Tumor Clonal Evolution for Drug Combinations Design. *Trends in Cancer*, **2(3)**:144–158. doi:10.1016/j.trecan.2016.02.001. URL https://linkinghub.elsevier.com/retrieve/pii/S2405803316000212.

8 Figures

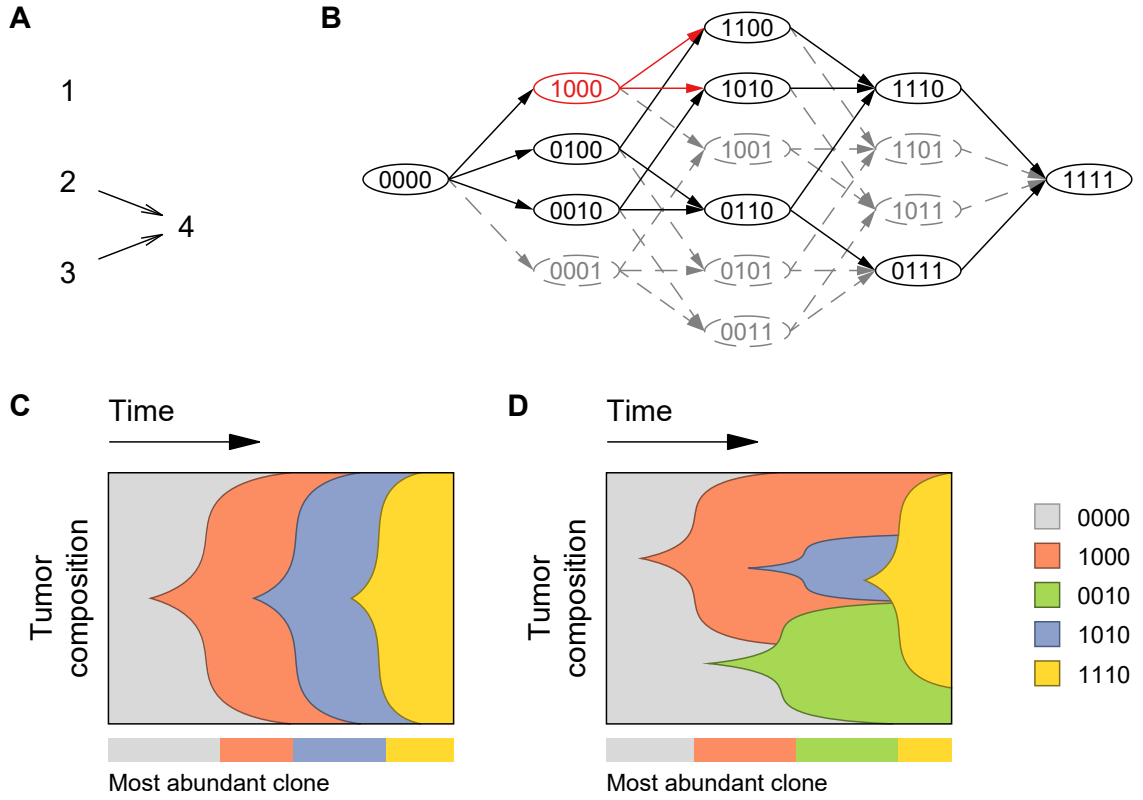

Fig. 1. Inferring evolutionary trajectories using Cancer Progression Models. A. Example of a directed acyclic graph (DAG) of restrictions in the order of accumulation of mutations, analogous to the ones produced by most CPMs, such as CBN, MCCBN, OT, CAPRESE, and CAPRI. Given four driver genes (numbered 1 to 4), the DAG indicates the constraints in the order of accumulation of mutations (arrows). Here, a mutation in the 4th loci is conditioned to the 2nd and 3rd ones having been previously mutated. MHN uses a network to encode both promoting and inhibiting effects between pairs of events; the network could be identical to the one in A (meaning that mutations 2 and 3 have a promoting effect on mutation 4), but it could also include inhibiting effects; in contrast to DAGs of restrictions, MHNs do not denote deterministic dependencies. **B.** Restrictions in the order of accumulation of mutations limit the amount of available evolutionary trajectories and stochastic dependencies change the probabilities of trajectories. In this network, nodes represent different genotypes of four driver genes and edges connect those that are a single mutation away. Genotype names are assigned according to which genes are mutated (1) or not mutated (0). Without restrictions many mutational pathways are possible (any route through both solid and/or dashed edges, from left to right). However, only a subset of genotypes satisfy the constraints in panel A (solid nodes); thus, only trajectories connecting these are predicted to be accessible pathways (solid edges). For example, given a specific genotype (1000, red node), restrictions in the mutational order limit the possible clones that can evolve from it consecutively after (red edges). With CBN, MCCBN, and OT, in addition to the set of accessible pathways, it is possible to obtain transition probabilities along the black lines. Under the MHN model, as there are no deterministic dependencies, all transitions (to genotypes with one additional mutation) are possible, but transitions affected by inhibitory stochastic dependencies will have very low probability (in this simplified depiction they would be the gray lines) and those affected by promoting stochastic dependencies will have higher probability (black lines); MHN provides transition probabilities for all the lines in the figure. **C.** Example of cancer progression with four driver genes under SSWM conditions. Progression towards the fitness maximum occurs through a defined sequence of parent-child genotypes, each one of them establishing as the most abundant rapidly after evolving. **D.** Example of cancer progression with four driver genes under conditions of no SSWM. Here, the sequence of parent-child genotypes from the wild-type with no driver mutations to the final one (i.e. the Line Of Descent or LOD) is: 0000 → 1000 → 1010 → 1110. Notice that in the most general case, the ultimately fixated genotype need not be the one with all driver mutations (1111). Some genotypes can be most abundant in the tumor during prolonged periods of time despite not being part of the LOD (0010, in green), and genotypes that do belong to the LOD may never be most abundant (1010, in blue). In all models and figures we do assume that there are no back mutations and that crossing valleys in the fitness landscape using a single multi-mutation step is not possible.

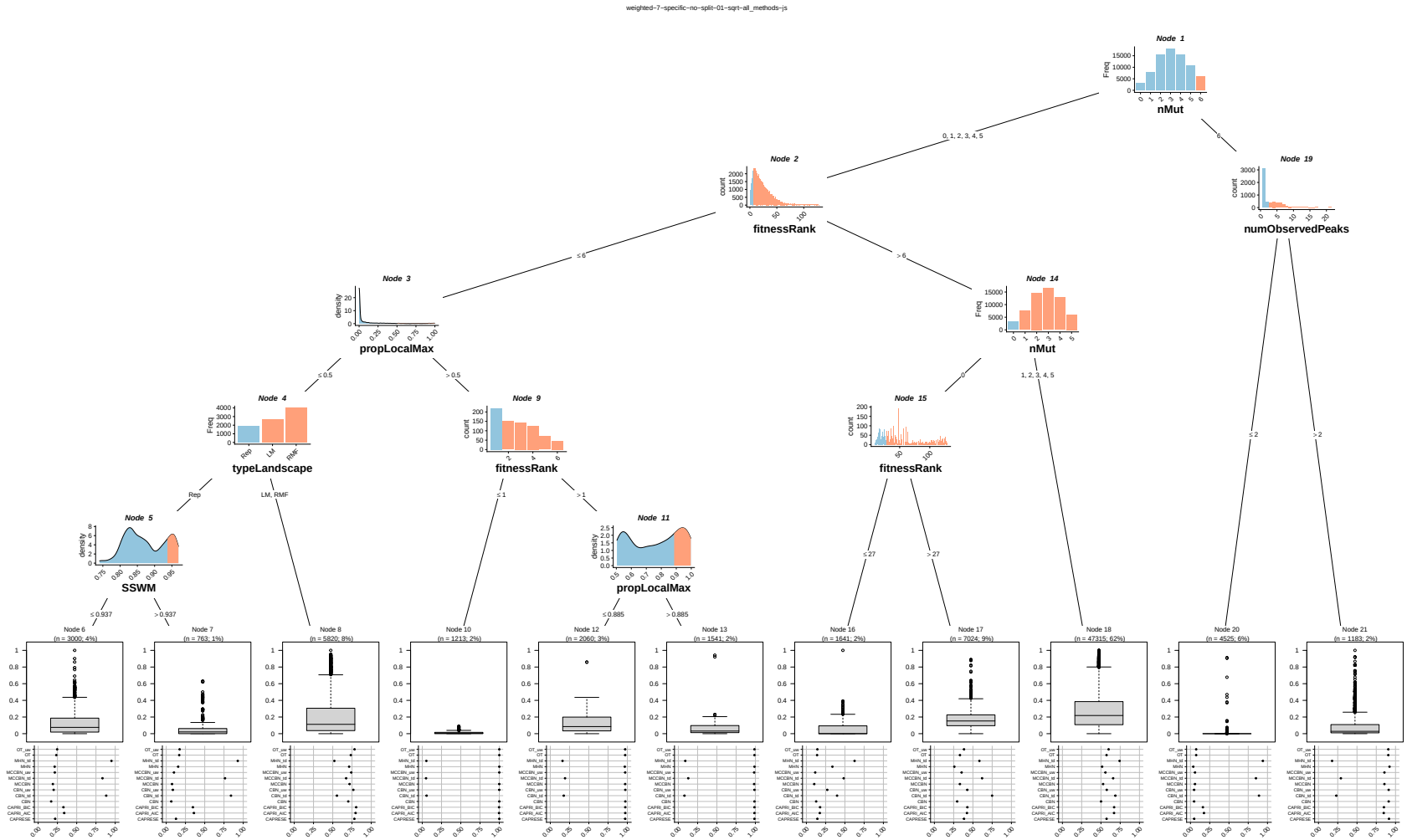

Fig. 2. Linear mixed-effects model tree fitted to data from fitness landscapes with 7 loci, sample size of 4000 and uniform sampling. Internal nodes show the partitioning variables and the values that lead to each partition; light blue denotes the values or categories that lead down the left split and salmon the values or categories that lead down the right split (partitioning criteria are also shown in the edges leading down a node). Leaves or terminal nodes show (weighted) boxplots of the dependent variable for the observations that fall in that leave and, as label, the name of the node and the sample size (sum of weighted observations and % of the total data set for the fit). The dot plots underneath the boxplots show the (weighted) mean JS for each method for that leave.

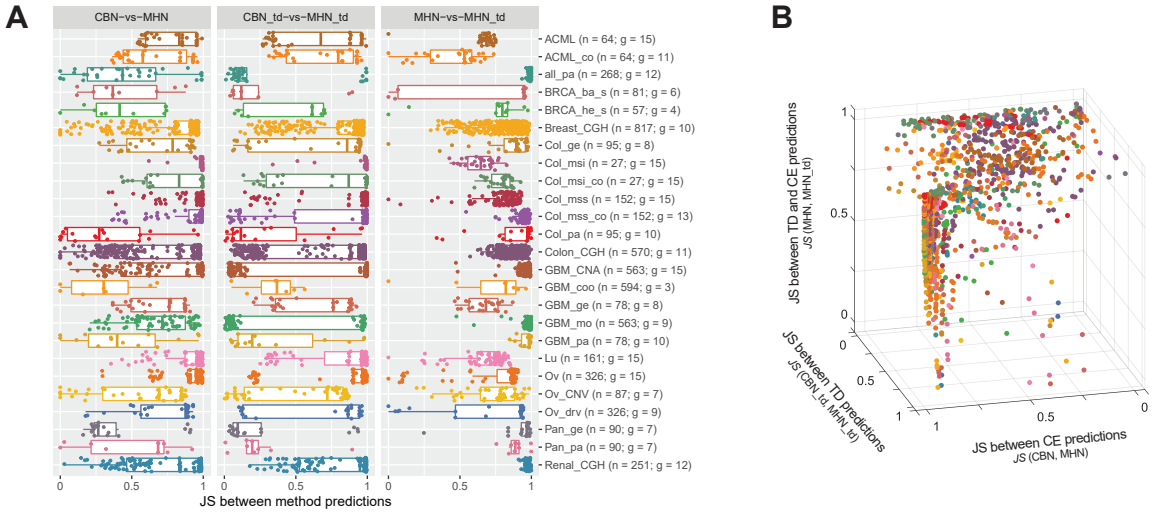

Fig. 3. Similarity between predictions of three pairs of methods for the 25 biological data sets. For each data set, the JS divergence between the predictions of a pair of methods is represented in **A.** each of the three panels or **B.** each of the three axes. Dots correspond to all genotypes for which we computed predictions, which are the genotypes observed in the samples, except for the genotype with all genes mutated if it was observed. In parentheses, after the data set, we indicate the total sample size (n) and number of genes (g).

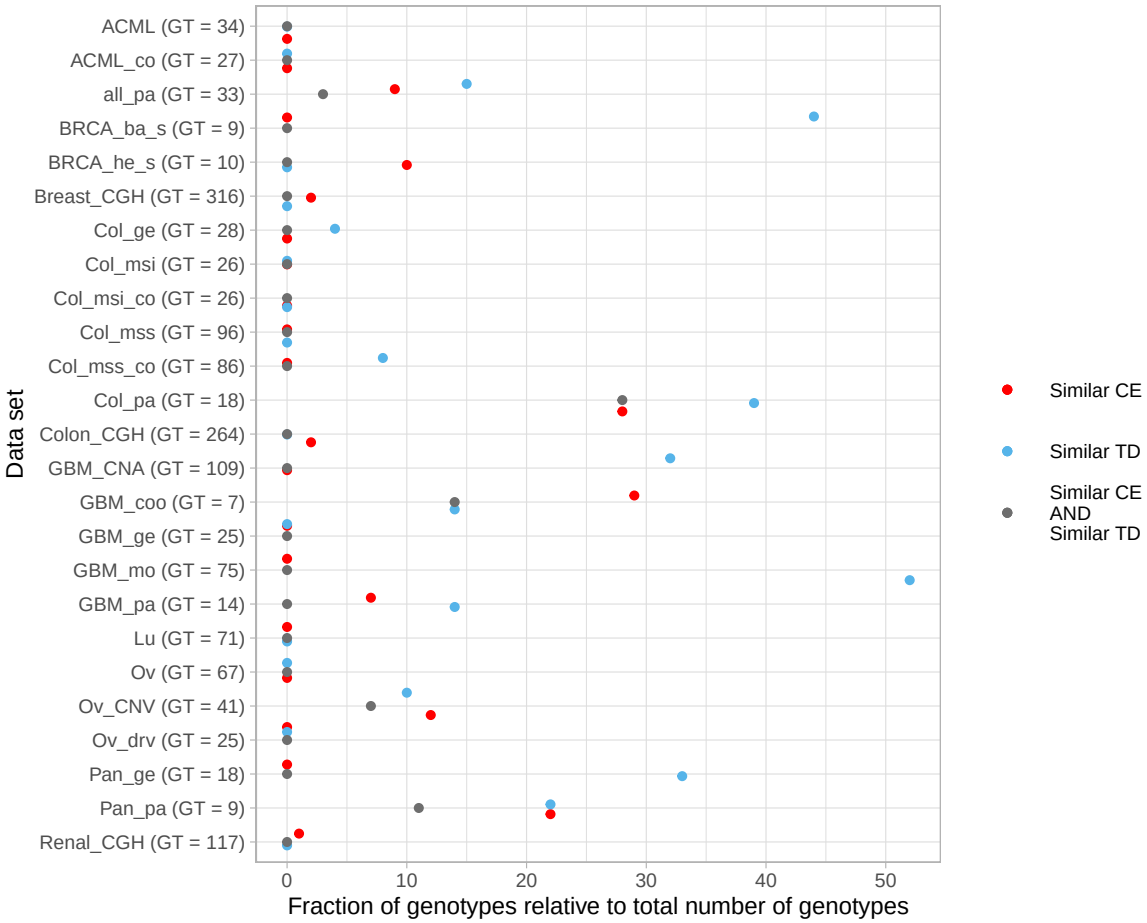

Fig. 4. Percentage of genotypes with similar TD and similar CE predictions for the 25 biological data sets. For each data set, the orange points show the percentage of genotypes in which predictions had both similar CE ($JS_{CBN,MHN} \leq 0.0677$) and different CE-TD ($(1/2)\ (JS_{CBN,CBN_td} + JS_{MHN,MHN_td}) \geq 0.7$), the blue dots those with similar TD ($JS_{CBN_td,MHN_td} \leq 0.0677$) and different CE-TD, and the gray points the percentage of genotypes that had similar CE, similar TD, and different CE-TD. Values shown are the percentage relative to the total number of genotypes for which we computed predictions, which are the genotypes observed in the samples, except for the genotype with all genes mutated, if it was observed. In parentheses, after the data set, we indicate the total number of genotypes observed (GT).