Every which way? On predicting tumor evolution using cancer progression models

Ramon Diaz-Uriarte, Claudia Vasallo Vega
Dept. Biochemistry, Universidad Autónoma de Madrid
Instituto de Investigaciones Biomédicas "Alberto Sols" (UAM-CSIC)
Madrid, Spain*

2018-07-17 (Release: Rev: fd58149)

Abstract

Cancer progression models (CPMs) use cross-sectional samples to identify restrictions in the order of accumulation of driver mutations. CPMs implicitly encode all the possible tumor progression paths or evolutionary trajectories during cancer progression, which can be of help for diagnostic, prognostic, and treatment purposes. Here we examine whether CPMs can be used to predict the true distribution of tumor progression paths and to estimate evolutionary unpredictability. Using simulations we show that the agreement between the true and the predicted distributions of paths is generally poor, unless sample sizes are very large and fitness landscapes are single peaked (have a single global fitness maximum). Under other fitness landscapes, performance is poor and only improves slightly with increasing sample size. Detection regime can be a key determinant of performance, and evolutionary unpredictability hurts performance except under regimes with very low sample variability. Estimates of evolutionary unpredictability from CPMs tend to overestimate the true unpredictability and the bias is affected by detection regime; CPMs could be useful for estimating upper bounds to the true evolutionary unpredictability. Analysis of eleven cancer data sets supports the relevance of detection regime and shows estimates of evolutionary unpredictability in regions where useful prediction might possible for at least some data sets. But the evolutionary trajectory predictions themselves are unreliable. Our results indicate that, currently, obtaining useful predictions of tumor progression paths from CPMs is dubious and emphasize the need for methodological work that can account for the probably multi-peaked fitness landscapes in cancer.

*To whom correspondence should be addressed: ramon.diaz@iib.uam.es, rdiaz02@gmail.com, http://ligarto.org/rdiaz

1 Introduction

Cancer progression models (CPMs), such as CBN (Gerstung *et al.*, 2009), CAPRI (Ramazzotti *et al.*, 2015), or OT (Szabo and Boucher, 2008), have been developed to identify restrictions in the order of accumulation of mutations during tumor progression from cross-sectional data (Beerenwinkel *et al.*, 2015, 2016). The identification of these constraints can help find therapeutic targets and disease markers. CPMs also implicitly encode all the possible mutational paths or trajectories of tumor progression, from the initial genotype to the genotype with all driver genes mutated, and the identification of these paths or "evolutionary trajectories" is a prominent idea in recent CPM publications (e.g. Caravagna *et al.*, 2016; Ramazzotti *et al.*, 2015).

Knowing the likely paths of tumor evolution is helpful for diagnostic, prognostic, and treatment purposes as, for example, it would allow us to identify genes that block the most likely paths (Greaves, 2015; Lipinski *et al.*, 2016; McPherson *et al.*, 2018). This interest in predicting paths of progression is, of course, not exclusive to cancer (see reviews in Lässig *et al.*, 2017; Losos, 2018). For example, in some cases antibiotic resistance shows parallel evolution with mutations being acquired in a similar order (Toprak *et al.*, 2012), and here "Even a modest predictive power might improve therapeutic outcomes by informing the selection of drugs, the preference between monotherapy or combination therapy and the temporal dosing regimen (...)" (Palmer and Kishony, 2013, p. 243 i). But detailed information about the distribution of paths of tumor evolution is currently not available. Obtaining it requires large numbers of multiple within-patient samples with timing information (alternatively, detailed knowledge of the fitness landscape together with information about major determinants of dynamics such as mutation rates, population sizes, and growth models —e.g., Bank *et al.*, 2016; de Visser and Krug, 2014; Lässig *et al.*, 2017; Losos, 2018; Szendro *et al.*, 2013— might allow us to model or simulate the evolutionary process to estimate the distribution of paths). Since CPMs can capture the paths of tumor progression from cross-sectional data, they offer a promising alternative, especially given the currently available and growing number of cross-sectional data sets.

The first aim of this paper is to examine if we can use CPMs to predict the likely paths of tumor evolution. The question we will ask is how close to the truth are the predictions about the distribution of paths of tumor evolution. When addressing this question we need to take into account possible deviations from the models assumed by CPMs. In particular, CPMs assume that the acquisition of a mutation in a driver gene does not decrease the probability of gaining a mutation in another driver gene (Misra *et al.*, 2014); this implies that the fitness landscapes assumed by CPMs cannot have reciprocal sign epistasis (Diaz-Uriarte, 2018: under reciprocal sign epistasis two mutations that individually increase fitness reduce it when combined Crona *et al.*, 2013; Poelwijk *et al.*, 2007, 2011). This also means that the fitness landscapes assumed by CPMs only have a single global fitness maximum (the genotype with all drivers mutated).

But reciprocal sign epistasis is likely to be common in cancer (Chiotti *et al.*, 2014), an argument supported by how common synthetic lethality is in both cancer cells (Beijersbergen *et al.*, 2017; O'Neil *et al.*, 2017) and the human genome (Blomen *et al.*, 2015). Moreover, if there are many combinations of a small number of drivers, out of a larger pool of drivers (Tomasetti *et al.*, 2015), that result in the escape genotype, it is likely that cancer landscapes will have several local fitness maxima (i.e., be multi-peaked). As we have shown before (Diaz-Uriarte, 2018), the performance of CPMs for predicting what genotypes can and cannot exist degrades considerably when the assumption of absence of reciprocal sign epistasis is violated. Those results, however, do not provide a direct answer to the question of predictability: if our objective is predicting paths of tumor progression we want to measure directly the quality of the predictions of paths of progression. For example, getting some of the edges of the DAG of restrictions wrong, or predicting some of the genotypes incorrectly, might be of little importance if the main paths of disease are captured and this happens in evolutionary scenarios where disease progression follows only a limited set of paths. Thus, to answer the question of whether CPMs can be used to predict paths of progression we will need to look directly at the prediction of paths and do that both under scenarios where CPM's assumptions are met and under scenarios with relevant deviations from the assumptions (see Figure 1).

And relevant scenarios highlight another key consideration when attempting predictions of tumor progression. That a particular method can reconstruct well the actual distribution of paths of tumor progression might be of little importance if that happens in an evolutionary scenario where the true evolutionary unpredictability itself is very large; for practical purposes, forecasting here would be useless. Which brings us to the second and third objectives of this paper: to understand what factors, including intrinsic evolutionary unpredictability itself, affect the quality of those predictions and to asses if, regardless of the performance predicting the actual paths of tumor progression, we can use CPMs to estimate evolutionary unpredictability itself.

To address the above three questions (can we predict the paths of tumor evolution using CPMs?; how is predictability affected by evolutionary unpredictability?; can we estimate evolutionary unpredictability using CPMs?) we use evolutionary simulations on 1260 fitness landscapes that include from none to severe deviations from the assumptions of CPMs. Since the role of evolutionary unpredictability is an important focus of this paper, we simulate evolution under different population sizes and mutation rates, so as to generate varying amounts of evolutionary unpredictability. This paper does not attempt to understand the determinants of evolutionary predictability (see, e.g., Bank *et al.*, 2016; de Visser and Krug, 2014; Lässig *et al.*, 2017; Losos, 2018; Szendro *et al.*, 2013) but, instead, we focus on the effects of evolutionary unpredictability for CPMs. This is why we use variation in key determinants of predictability (e.g., variation in population sizes and mutation rates) but these factors, themselves, are only used to generate variability in unpredictability, and not themselves the focus of the study. To better assess the quality of predictions, we use sample sizes that cover the range from what is commonly used to what are much larger sample sizes than currently available. We also include variation in the cancer detection process, since it has been found before to affect the quality of inferences (Diaz-Uriarte, 2015, 2018). We find that the agreement between the true and the predicted distributions of paths is generally poor, unless sample sizes are very large and fitness landscapes conform to the restrictive assumptions of CPMs. Both detection regime and evolutionary unpredictability itself have major effects on performance. But in spite of the unreliability of the predictions of actual paths of tumor progression, CPMs can be useful for estimating upper bounds to the true evolutionary unpredictability.

We then analyze eleven cancer data sets for which the truth is not known, but using bootstrap samples allows us to examine the robustness of the inferences. Our results emphasize again the relevance of detection regime. And if the estimates of evolutionary unpredictability do indeed provide an upper bound to the true evolutionary unpredictability, the data indicate that at least some of the data reflect conditions where useful predictions could be possible. But for most data sets these results are thwarted by the unreliability of the predictions themselves. Our results question the use of CPMs for predicting paths of tumor progression.

1.1 Assumptions

CPMs assume that the different individuals in a data set constitute independent realizations of the same evolutionary process and therefore that the same constraints hold for all tumors (Beerenwinkel *et al.*, 2015, 2016; Gerstung *et al.*, 2011). Thus, a data set can be regarded as a set of replicate evolutionary experiments where all individuals are under the same genetic constraints, though they might later be exposed to different conditions. We also make other common assumptions in this field, listed in Diaz-Uriarte (2018) (section 1.1). Briefly, we use biallelic loci, back mutations are not allowed, and mutations are single-gene mutations (Beerenwinkel *et al.*, 2007; Bozic *et al.*, 2010; McFarland *et al.*, 2013). All tumors start cancer progression without any of the mutations considered, but other mutations could be present that have caused the initial tumor growth, so we absorb the cancer initiation process in the root node (Attolini *et al.*, 2010); this is necessary to simulate data consistent with cross-sectional sampling. The driver genes are known and there are no observational errors.

2 Material and methods

2.1 Overview of the simulation study

We used simulations of tumor evolution on fitness landscapes of different types (see Figure 1), for landscapes of seven and ten genes, under different initial population sizes and mutation rates. As explained above, variation in initial population size and mutation rates is used to generate variability in evolutionary predictability, but not of interest *per se*. We have used a total of 1260 fitness landscapes = 35 random fitness landscapes x two conditions of numbers of genes x three types of fitness landscapes x three initial population sizes x two mutation regimes. For each one of the 1260 fitness landscapes, we simulated 20000 evolutionary runs (with the specified parameters for initial population size and mutation rate) using a logistic-like growth model until one of the genotypes at the local fitness maxima (or the single global fitness maximum) reached fixation. Each set of 20000 simulated runs was then sampled under three detection regimes (so that each fitness landscape generated three sets of 20000 simulated genotypes). From each of these sets, we obtained five different splits of the genotypes for each of three sample sizes (50, 200, 4000); thus a total of 56700 (= 1260 x 3 x 3 x 5 combinations of 1260 fitness landscapes, 3 detection regimes, 3 sample sizes, 5 splits) data sets were produced. Each of these 56700 data sets was analyzed with each of six CPM methods.

2.2 Evolutionary simulations and data sampling

We used three **initial population sizes**, 2000, 50000, and 1×10^6 cells, for the simulations; these are ranges that have been previously used in the literature and cover a range of population sizes at tumor initiation (e.g. Beerenwinkel *et al.*, 2007; Gerstung *et al.*, 2011; McFarland *et al.*, 2013; Wodarz and Komarova, 2014). We also used two **mutation regimes**; in the first one, all genes had a common mutation rate of 1×10^{-5}; in the second, genes had different mutation rates, uniformly distributed in the log scale between $(1/5)\ 1 \times 10^{-5}$ and 5×10^{-5} (i.e., the largest ratio between largest and smallest mutation rates was 25), so that the arithmetic mean of mutation rates was 1.5×10^{-5} and the geometric mean 1×10^{-5}. These mutation rates are within ranges previously used in the literature (Bozic *et al.*, 2010; McFarland *et al.*, 2013; Nowak *et al.*, 2004), with a bias towards larger numbers (since we use only seven or ten genes relevant for population growth and we could be modeling pathways, not individual genes). Initial population size and mutation rates are not of intrinsic interest here (since our focus is not the determinants of evolutionary predictability *per se*), but are used to generate variability in evolutionary predictability; see section 3.1.

For each of the combinations of number of genes (seven and ten), initial population size (50, 200, 1×10^6), and mutation rate (constant, variable), we generated random fitness landscapes of three kinds (see Figure 1). We generated the DAG-derived **representable** fitness landscapes by generating a random DAG of restrictions and from it the fitness graph. We then assigned birth rates to genotypes using an iterative procedure on the fitness graph where, starting from the genotype without any driver mutation with a birth rate of 1, the birth rate of each descendant genotype was set equal to the maximum fitness of its parent genotypes times a random uniform variate between 1.01 and 1.19 (yielding, thus, an average multiplicate increase in fitness of 0.1, again within values previously used; Bozic *et al.*, 2010; McFarland *et al.*, 2013). Birth rate of genotypes without dependencies satisfied was set to 0. (Note that for the growth model used here —see below— birth rates determine fitness at any population size as death rates are identical for all genotypes and depend only on population size. Note also that genotypes with birth rate of 0 are never added to the population; thus, they cannot mutate before dying, so this simulation scheme strictly adheres to the assumptions about accessible and non-accessible genotypes under the CPM model). The DAG-derived **local-maxima** fitness landscapes were obtained by generating a random DAG and from it the fitness graph. Before assigning fitness to genotypes, a random selection of edges of the fitness graph were removed so that all accessible genotypes remained accessible but from a possibly much smaller set of parents. Fitness was then assigned as above (with the iterative procedure on the fitness graph, where fitness of child = max(fitness parents) $U(1.01, 1.19)$). For each DAG we repeated this procedure 50 times, and kept the one that introduced the largest number of local maxima. Creating local maxima almost always resulted in creating reciprocal sign epistasis (but see Supplementary Material "Generating random fitness landscapes"). The local maxima fitness landscapes used in this paper are representable in the weaker sense of Diaz-Uriarte (2018), as all genotypes that should be accessible under the DAG of restrictions are accessible. What the local-maxima landscapes are missing are mutational paths to the genotype with all genes mutated, because we have introduced local fitness maxima (and once we introduce local maxima there is no longer a one-to-one correspondence between DAGs of restrictions and fitness graphs and, thus, there is no longer a one-to-one correspondence between DAGs of restrictions and sets of tumor progression paths). These local maxima landscapes are "easier" than the DAG-derived fitness landscapes used in (Diaz-Uriarte, 2018), as those also missed some genotypes that should exist under the DAG of restrictions. Our local maxima are easier by design as we want to isolate the effect of multi-peaked landscapes or local maxima (or, equivalently, missing paths), without the additional burden of missing genotypes. The Rough Mount Fuji (**RMF**) fitness landscapes we obtained from an RMF model, a model that has been useful to model empirical fitness landscapes (de Visser and Krug, 2014; Franke *et al.*, 2011; Neidhart *et al.*, 2014), where the reference genotype and the decrease in birth rate of a genotype per each unit increase in Hamming distance from the reference genotype were randomly chosen (see "Random fitness landscapes" in Supplementary Material). These fitness landscape cannot be represented by DAGs of restrictions with respect to neither paths to the maximum or accessible genotypes (see also Diaz-Uriarte, 2018).

Once a fitness landscape had been generated, we simulated 20000 evolutionary processes. We used the continuous-time, logistic-like model of McFarland *et al.* (2013), in which death rate depends on total population size, as implemented in OncoSimulR (Diaz-Uriarte, 2017), with the specified parameters of initial population size and mutation rate. Each individual simulation was run until one of the genotypes at the local fitness maxima (or the single global fitness maximum) reached fixation (see details in "Simulations" in Supplementary Material). We also verified that all seven or ten genes had appeared in at least some genotypes, i.e., were part of the paths of tumor progression. If this condition was not fulfilled, a new fitness landscape was generated and the processes started again. This procedure is independent of the detection process that generates the actual samples of genotypes (next).

To obtain the actual samples of genotypes that were analyzed by the CPMs, we used three different detection regimes to emulate single-cell sampling at total tumor sizes (number of cells) that are, in the log scale, approximately uniformly distributed (**uniform** detection regime), biased towards large sizes (**large**) or biased towards small sizes (**small**). (Working on the log-scale of tumor size is appropriate as in the model of McFarland *et al.*, 2013, tumor population size increases logarithmically with number of driver mutations). We drew random deviates from beta distributions with parameters $B(1, 1)$, $B(5, 3)$, and $B(3, 5)$ (for uniform, large, and small, respectively), rescaled them to the range of observed sizes, and obtained the sample with actual population size closest to the target (see details in Supplementary Material "Detection regimes: sampling"). For each sample, the genotype returned was the single genotype with the largest frequency (so we did not introduce possible additional noise due to whole-tumor, or bulk, sequencing). Finally, for each of the three **sample sizes** of 50, 200, and 4000, we splitted the 20000 simulations into five sets of non-overlapping data sets. These are the data sets that were analyzed with the six CPMs.

2.3 Cancer Progression Models (OT, CBN, CAPRI, CAPRESE) and paths of tumor progression

We have used four different CPM methods (methods not considered here are either too slow for routine work, have no software available, or have dependencies on non-open source external libraries —see Supplementary Material "CPM software"). Two of the methods used have two with variants, yielding a total of six methods. Only a brief overview is provided here; detailed descriptions can be found in Caravagna *et al.* (2016); Desper *et al.* (1999); Gerstung *et al.* (2009, 2011); Loohuis *et al.* (2014); Montazeri *et al.* (2016); Ramazzotti *et al.* (2015); Szabo and Boucher (2008). CPM methods assume that the different individuals in a data set constitute independent realizations of the same evolutionary process —see above. These methods try to identify restrictions in the order of accumulation of mutations from cross-sectional data. The cross-sectional data is a matrix of subjects or samples by driver alteration events, where each entry in the matrix is binary coded as mutated or not-mutated. For the simulations, we will refer to these driver alteration events as "genes", but they can be individual genes, parts or states of genes, or modules or pathways made from several genes (e.g. Caravagna *et al.*, 2016; Gerstung *et al.*, 2011). When we analyze the eleven cancer data sets (see section 2.5) we will use the generic term "features" as some of those data sets use genes whereas others use pathway information. Both Oncogenetic trees (OT) (Desper *et al.*, 1999; Szabo and Boucher, 2008) and CAPRESE (Loohuis *et al.*, 2014) describe the accumulation of mutations with order constraints that can be represented as trees. A key difference between the two is that CAPRESE reconstructs these models using a probability raising notion of causation in the framework of Suppes probabilistic causation, whereas in OT weights along edges can be directly interpreted as probabilities of transition along the edges by the time of observation (Szabo and Boucher, 2008, p. 4). Both CAPRI and CBN allow modeling the dependence of an event on more than one previous event: the output of the model are graphs (DAGs) where some nodes have multiple parents, instead of a single parent (as in trees). CAPRI tries to identify events (alterations) that constitute "selective advantage relationships" again using probability raising in the framework of Suppes probabilistic causation. We have used two versions of CAPRI, that we will call CAPRI_AIC and CAPRI_BIC, that differ in the penalization used in the maximum likelihood fit (AIC or BIC, respectively). For CBN we have also used two variants, the one described in Gerstung *et al.* (2009, 2011) that uses simulated annealing with a nested EM algorithm for estimation, and MCCBN, described in Montazeri *et al.* (2016), that uses a Monte-Carlo EM algorithm that allows it to fit data sets with many more genes. See Supplementary Material, "CPM software" for further details.

Because (the transitive reduction of) a DAG of restrictions determines a fitness graph (see Figure 1 and Diaz-Uriarte, 2018), the set of paths to the maximum encoded by the output from a CPM is obtained from the fitness graph. This we did for all methods. From CBN and MCCBN we can also obtain the estimated probability of each path of tumor progression to the fitness maximum, since both CBN and MCCBN return the parameters of the transition rates between genotypes (see e.g., p. i729 in Montazeri *et al.*, 2016, section 2.2 in Gerstung *et al.*, 2009, or Hosseini, 2018). It is also possible to perform a similar operation with the output of OT, and use the edge weights from the fits of OT to obtain the probabilities of transition to each descendant genotype and, from them, the probabilities of the different paths to the global maximum. It must be noted that this is really abusing the model, since the OTs used are untimed oncogenetic trees (Desper *et al.*, 1999; Szabo and Boucher, 2008). We will refer to paths with probabilities assigned in the above way as **probability-weighted paths**. For CAPRESE and CAPRI, it is not possible to map the output to different probabilities of paths of progression (see also Supplementary Material "CAPRI, CAPRESE, and paths of tumor progression") and in all computations that require probability of paths we will assign the same probability to each path.

2.4 Measures of performance and predictability

We have characterized evolutionary unpredictability using the diversity of Lines of Descent (LOD). LODs were introduced by Szendro *et al.* (2013) and "(...) represent the lineages that arrive at the most populated genotype at the final time" (p. 572). In other words, a LOD is a sequence of parent-child genotypes, from the initial genotype to a local maximum. In the context of this paper, a LOD is the path that a tumor has taken until fixation. The final genotype in a LOD is a local fitness maximum, but there are no guarantees that any intermediate genotype in the LOD will have been the most common genotype at any time point (specially if there is clonal interference and stochastic tunneling — de Visser and Krug, 2014; Sniegowski and Gerrish, 2010). As in Szendro *et al.* (2013), we can use the entropy of these paths to measure the indeterminism with respect to the paths of evolution, or evolutionary unpredictabilty, and we will define $S_p = -\sum p_i \ln p_i$, where p_i is the observed probability of each LOD (each path) computed from the 20000 simulations, and the sum is over all paths or LODs. Evolutionary unpredictability, as estimated by the CPMs, will analogously be defined as $S_c = -\sum q_j \ln q_j$, where q_j is the probability of each path to the maximum according to the cancer progression model considered, and the sum is over all paths predicted by the CPMs . (Hosseini, 2018, normalizes predictability by dividing by the maximum entropy, similar to dividing by the prior entropy in the "information gain" statistic in Lässig *et al.*, 2017; but the maximum entropy is a constant for each number of genes, i.e., 7! or 10! for our simulations).

To measure how well CPMs predict actual tumor progression, we use three different statistics. To compare the overall similarity of the distribution of paths predicted by CPMs with the observed one (i.e., the distribution of LODs) we have used the Jensen-Shannon divergence (**JS**) (Crooks, 2017; Lin, 1991), scaled between 0 and 1 (equivalent to using the logarithm of base 2). JS is a symmetrized Kullback-Leibler divergence between two distributions and is defined even if the two distributions do not have the same sample space, i.e., even if $P(i) \neq 0$ and $Q(i) = 0$ (or $Q(i) \neq 0$ and $P(i) = 0$), as can often be the case for our data. A value of 0 means that the distributions are identical, and a value of 1 that they do not overlap. Therefore, predictions of CPMs are closer to the truth the smaller the value of JS. The sum of the probabilities of the paths in the LODs that are not among the paths allowed by the CPMs, $P(\neg DAG|LOD)$, is equivalent to **1 - recall**. Larger values of 1-recall mean that the CPM is not capturing a large fraction of the actual evolutionary paths to the maximum (or maxima). The sum of the predicted probabilities of paths according to the CPMs that are not used by evolution (i.e., that are not LODs), $P(\neg LOD|DAG)$, is equivalent to **1 - precision**. Larger values of 1-precision mean that the CPMs predict paths to the maximum that are not used by evolution. Some figures in the Supplementary Material also use as statistic the **probability of recovering the most common LOD**; we will rarely refer to this statistic in the main paper since it follows a pattern very similar to recall (see Supplementary Material, section "Probability of recovering the most common LOD"). Statistics 1-recall and 1-precision can, however, overestimate performance: they could both have a value of 0, even when JS is very close to 1 (see example in Supplementary Material "Example where perfect recall and precision do not guarantee Jensen-Shannon divergence of 0"). Thus, the basic overall performance measure will be JS.

2.4.1 Comparing paths from CPMs with LODs of different lengths

When all paths from the CPM and the LOD have equal length (they end in a genotype with the same number of genes mutated, K) computing the above statistics is straightforward. But paths could differ in length. In fitness landscapes with local maxima, LODs can differ in length; some could have K_i (the number of mutations at the fixated genotype, or the length of the path) larger than K_C (all paths from a CPM have the same number of mutations, since all arrive at the genotype with all K_C genes mutated). It is also possible that $K_i > K_C$ if the CPM has been built from a fitness landscape with a data set that contains fewer genes than the number of genes in the landscape (e.g., because one or more genes were absent —see Supplementary material "Preprocessing of data for CPMs"). In fact, in representable fitness landscapes, all $K_i > K_C$ if the CPM has been built from a fitness landscape with a data set that contains fewer genes than the number of genes in the landscape (all K_i will be equal to either seven or 10). We need a procedure to compute JS, 1-recall, and 1-precision that will cover all those cases. This procedure should ignore specifics of the sampling model and should reduce to the simpler procedure in the above section when all $K_i = K_C$.

Let i and i denote two paths, one from the LOD and the other from the CPM, with corresponding probabilities p_i and q_j; in contrast to the previous section, and to minimize notation, p, q could refer to a path from the LOD and a path from the CPM, xor a path from the CPM and a path from a LOD. Let K_i, K_j denote the length of paths i and j, respectively. At least one set of either K_i or K_j has all elements identical (e.g., if j refers to indices of the paths from the CPM, it is necessarily the case that $K_1 = K_2 = \ldots = K_m = K_C$, with m the total number of paths from the CPM).

The procedure has to fulfill two desiderata. a) If $K_i > K_j$, but i^k, the path i up to $K_i = k$ mutations (i.e., from the WT genotype to the genotype with k mutations) is identical to j, then path j is included in path i. All of q_j is accounted for by i. b) Path i is partially included (or accounted for) by path j, but a fraction of it, $(K_i - K_j)/K_i$, is missing or unaccounted. The above applies directly to calculations of 1-recall and 1-precision. For computing JS, there will be two entries in the vectors with the probability distributions that will be compared: $P = \left[p_i \frac{K_j}{K_i}, p_i \frac{K_i - K_j}{K_i}\right]$, $Q = \left[q_i, 0\right]$. This procedure can be applied to all elements i, j, summing all unmatched entries ($\sum p_i \frac{K_i - K_j}{K_i}$: this is the total flow in the set of is that cannot be matched by the js because they are shorter). To simplify computations, that unmatched term can include $\sum p_u$, where u denote those paths in i that do not match any j. Likewise, all paths i with $K_i > K_j$ such that the paths become indistinguishable up to K_j can be summed in a single entry: $\sum p_i \frac{K_j}{K_i}$ and $\sum p_i \frac{K_i - K_j}{K_i}$ for the matched and unmatched fractions, respectively. All computations have their corresponding counterparts for elements i, j when $K_i < K_j$. This procedure results in unique JS (remember the K are all the same for at least one of the sets of paths) as well as unique 1-precision and 1-recall, and it reduces to section 2.4 when all K_i are equal and equal to all K_j. A commented example is provided in the Supplementary Material ("Commented example for paths of unequal length")

2.4.2 Statistical modeling of performance

We have used generalized linear mixed-effects models, with a beta model for the dependent variable (Ferrari and Cribari-Neto, 2004; Grün *et al.*, 2012; Smithson and Verkuilen, 2006), to model how JS, 1-recall, and 1-precision, are affected by S_p, detection regime, sample size, number of genes, and type of fitness landscape. All models used, as response variables, the average from the five split replicates of each landscape by sample size by detection regime combination, and we have used fitness landscape as random effect . When the dependent variable had values exactly equal to 0 or 1, we have used the transformation suggested in Smithson and Verkuilen (2006). Models were fitted using sum-to-zero contrasts (McCullagh and Nelder, 1989). All regressors have been used as discrete regressors, except S_p, which has been scaled (mean 0, variance 1) for easier interpretation and so that the intercept term is interpreted as the predicted response at the average value of the regressors (see further details in Supplementary Data"Coefficients of linear models"). We have used the glmmTMB (Brooks *et al.*, 2017) and car (Fox and Weisberg, 2011) packages for R (R Core Team, 2018) for statistical model fitting and analysis.

2.5 Cancer data sets

We have used eleven cancer data sets (including glioblastoma, lung, ovarian, colorectal, and pancreatic cancer); some code mutations in terms of genes and some in terms of pathways. These data were obtained from Caravagna *et al.* (2016); Gerstung *et al.* (2011); Misra *et al.* (2014), with the original sources for the data being Bell *et al.* (2011); Ding *et al.* (2008); Jones *et al.* (2008); Network (2012); Parsons (2008); Wood *et al.* (2007). Details on sources, names, and how the data were obtained are provided in the the Supplementary Data ("Cancer data sets").

3 Results

3.1 Simulated fitness landscapes: characteristics, evolutionary predictability, clonal interference, and sampled genotypes

In the Supplementary Material (see section "Plots of fitness landscapes and inferred DAGs") we show all the fitness landscapes used; in section "Fitness landscapes: characteristics, evolutionary predictability, clonal interference, and sampled genotypes") we show the main characteristics of the fitness landscapes used, the variability in evolutionary predictability, and the main characteristics of the samples obtained under the three detection regimes. The three types of fitness landscapes had comparable numbers of accessible genotypes but differed strongly in the number of local fitness maxima and reciprocal sign epistasis, with reciprocal sign epistasis being associated to number local fitness maxima in the local maxima landscapes. Simulations resulted in varied amounts of clonal interference, as measured by the average frequency of the most common genotype (or, similarly, the inverse of the average number of clones with frequency $> 5\%$); as seen in the plots in the Supplementary Material, scenarios where clonal sweeps dominate (i.e., those characterized by the smallest clonal interference) corresponded to initial population sizes of 2000, with clonal interference being much larger at the other population sizes.

Simulations resulted in observed numbers of paths to the maximum (number of distinct LODs) that showed a wide range, from 2 to 3082 (median of 228, 95, and 55, for representable, local maxima, and RMF, respectively), with fitness landscapes with 10 genes with a disproportionately larger number (105 vs. 1340, 55 vs. 261, 33 vs. 113, for representable, local maxima, and RMF, respectively). LOD diversities (S_p) ranged from 0.3 to 8.7, with RMF models showing lower S_p, although RMF landscapes had the largest number and diversity of observed local fitness maxima. S_p was strongly associated to the number of accessible genotypes.

The number of different sampled genotypes was comparable between detection regimes, but diversity differed, with the uniform detection regime showing generally larger sampled diversity. The mean and median number of mutations of sampled genotypes differed between detection regimes in the expected direction (largest in large detection regime, and smallest in small detection regime); the standard deviation and coefficient of variation in the number of mutations were largest in the uniform detection regime (thus, the uniform detection regime showed both the largest variation in number of mutations of genotypes and the largest diversity of genotypes). The differences in sample characteristics between detection regimes often differed between fitness landscapes (in particular, the coefficient of variation in the number of mutations was largest in the RMF landscapes).

3.2 Predicting paths of evolution with CPMs: overall patterns

The six methods used can be divided into three groups: methods that return trees (OT and CAPRESE) and two families of methods that return DAGs, CAPRI (CAPRI_AIC and CAPRI_BIC) and CBN (CBN and MCCBN). As seen in the Supplementary Material "Overall patterns for the six methods", comparing within groups with respect to JS, one member of the pair consistently outperformed the other. OT was significantly better than CAPRESE (paired t-test over all 56595 pairs: $t_{56594} = -161.1.2$, $P < 0.0001$), CBN was significantly better than MCCBN ($t_{56593} = -42.6$, $P < 0.0001$), and CAPRI_AIC was significantly better than CAPRI_BIC ($t_{56594} = -41.9$, $P < 0.0001$).

This ranking within types of methods does not always apply for the other two measures of performance, most notably CAPRESE with respect to 1-recall, where its performance can be one of the best, and often better than that of OT. CAPRESE's better recall, however, is more than offset by its poor precision (often the worst or among the worst). Similar comments apply to other reversals (e.g., MCCBN's slightly better precision in some scenarios being offset by its considerably worse recall). Remember we will asses performance using mainly JS (see 2.4). In what follows, therefore, and for the sake of brevity, we will focus on OT, CBN, and CAPRI_AIC, since the overall performance of their alternatives is worse.

Figure 2 shows how the performance measures for OT, CBN, and CAPRI_AIC change with sample size for all combinations of type of landscape by detection regime by number of genes (results for probability of recovering the most common LOD are shown in the Supplementary Material and the patterns are essentially those of recall). The measures of JS and 1-precision for OT and CBN (and MCCBN) use probability-weighted paths computed as explained in 2.4, because there was strong evidence for all three methods that the probability-weighted paths led to better results (JS, paired t-test over all pairs: OT, $t_{56594} = 195.8$, $P < 0.0001$; CBN: $t_{56594} = 222.3$, $P < 0.0001$; MCCBN: $t_{56593} = 149.0$, $P < 0.0001$; 1-precision: OT: $t_{56594} = 187.6$, $P < 0.0001$; CBN: $t_{56594} = 217.6$, $P < 0.0001$; MCCBN: $t_{56593} = 130.3$, $P < 0.0001$). (See also Supplementary Material, "OT and CBN, JS, weighted vs. unweighted", "CAPRESE and OT, 1-precision, unweighted" and "CAPRI and CBN, 1-precision, unweighted" for figures that show the improvement due to weighting).

Overall, CBN was the method with the best performance ($P < 0.0001$ from all pairwise comparisons between the six methods with Tukey's contrasts and single-step multiple testing p-value adjustment Hothorn *et al.*, 2008). It must be noted, however, that CBN was one of the most variable methods in performance, as shown in Figure 3 (also Supplementary Material, "Overall patterns for the six methods").

JS differed between type of landscape, number of genes, detection regime, and sample size, but the magnitude and even direction of effects differed between combinations of those factors, as seen in Figure 2 and 4. Generalized linear mixed-effects models fitted to the complete data set and to the different combinations of method and type of landscape (see Supplementary Material, section "Analysis of deviance tables for fitted models") also showed highly significant ($P < 0.0001$) two-, three-, and four-way interactions between most of the factors, in particular those involving type of landscape and detection regime.

Under representable fitness landscapes, performance improved with increasing sample size and with the uniform detection regime, but decreased as the number of genes increased (Figure 2, panel A; Figure 4, top row). The decrease in performance with the number of genes is related both to missing evolutionary paths (Figure 2B), and allowing paths that are not used by evolution (Figure 2C). With CAPRI, however, the effect of sample size is much weaker and increases in sample size can even lead to decreases in performance, specially under the uniform detection regime (highly significant, $P < 0.0001$, interactions of detection and sample size — see Supplementary Material); this is attributable to CAPRI excluding many paths taken during evolution (Figure 2B). This behavior of CAPRI can also be seen in Figure 6A, where the number of paths allowed under CAPRI goes from slight to very severe underestimation as sample size increases under the uniform detection regime. This is itself caused by CAPRI sometimes allowing only a few or even just one path to the maximum (Supplementary Material, section "Number of paths inferred").

Under the RMF landscape overall performance was worse. Increasing sample size for OT and CBN led to minor decreases in performance (Figure 2 and Figure 4 bottom row). CPMs failed to capture about 50% of the evolutionary paths to the local maxima (Figure 2B) and included more than 75% of paths (or fractions of paths) that were never taken by evolution (Figure 2C). The behavior under local maxima was similar to that of representable fitness landscapes in terms of the direction of most effects, but effects were generally weaker. An important exception was evolutionary unpredictability where increasing S_p was associated to a decrease in performance similar to, but of smaller magnitude than, in RMF landscapes (see next).

3.3 Predicting paths of evolution with CPMs: effects of evolutionary unpredictability

There were no marginal effects of evolutionary unpredictability on performance in representable fitness landscapes (Figure 4). But the effects of evolutionary unpredictability were, in fact, more complex than depicted in Figure 4, as there were highly significant interactions ($P < 0.0001$) between S_p, detection regime, and sample size, within representable and local maxima landscapes, as well as in the overall models (see Supplementary Material, "Analysis of deviance tables for fitted models"). In many cases, the sign of the slope was reverted from its main effect, as is shown in Figure 5 (see also Supplementary Material, "Slopes of regressions of recall and precision on LOD diversity").

In most scenarios, performance was worse with larger unpredictability (larger S_p) as seen by the positive slopes of JS on S_p. But under representable landscapes, in the small and large detection regime and for sample sizes 50 and 200, larger evolutionary unpredictability was associated with better performance; the difference in effects was itself significantly affected by the number of genes (see also Supplementary data, section "Analysis of deviance tables for fitted models"). Note, however, that despite the change in relationship between unpredictability and performance, performance was still better with a sample size of 4000 than with sample sizes of 50 or 200 (Figure 2) in representable and local maxima landscapes. Under RMF fitness landscapes, large evolutionary unpredictability (larger S_p), in contrast to what happened in representable landscapes, was associated with poorer performance. Under local maxima, the effect of evolutionary unpredictability depended strongly on sample size and detection regime, with reversal of effects from sample size of 50 compared to 4000 under the large detection regime, similar to those mentioned above for representable landscapes.

3.4 Inferring evolutionary unpredictability from CPMs

Figure 6 shows the relationship between the estimated and true numbers and diversities of paths of tumor progression.

Even under representable fitness landscapes, and for the two methods with the best behavior, CBN and OT, there was large variability in the estimates of the number of paths to the maximum relative to the true number of paths associated to differences in sample size and number of genes, as shown in Figure 6A.

Average ratios of estimated paths to the maximum over true paths to the maximum were 1.4 and 6.9 for 7 and 10 genes for CBN (and 0.4 and 1.9 for OT). But values for CBN range from 0.5 (7 genes, sample size 50, uniform detection regime), to 33.9 (10 genes, sample size 4000, large detection regime); for OT they range from 0.2 (7 genes, sample size 200, uniform, detection) to 5.6 (10 genes, sample size 4000, large detection regime). In section 3.1 (see also Supplementary Material, "Fitness landscapes: characteristics, evolutionary predictability, clonal interference") we saw that the true number of paths to the maximum increased with the number of genes; what we see here is that the inferred number of evolutionary paths to the maximum from CBN and OT often increased even faster, a consequence of worse recall under 10 genes. Detection regime and sample size had a large effect: number of paths inferred increased with sample size, specially under the large detection regime.

For both CBN and OT that disproportionate increase in the number of inferred paths carried only a small penalty in terms of estimating evolutionary unpredictability (the diversity of paths to the maximum, S_p), as can be seen from Figure 6B. For example, for CBN the ratio of inferred to observed diversities, S_c/S_p, remained close to 1 with values from 0.68× to 1.04× over all combinations of type of landscape by detection regime by number of genes by sample size (averages of 0.81× and 0.93× for seven and ten genes); the values were closest to one with sample size 4000 and under the uniform detection regime.

In contrast to CBN and OT, patterns for CAPRI seemed dominated by the tendency of CAPRI to only allow very few paths as the sample size grows large, and mainly under the uniform detection regime (see also Supplementary Material, section "Number of paths inferred"). Under representable landscapes, CAPRI underestimates, sometimes severely, the true diversity of paths to the maximum (Figure 6B).

The above results apply to representable landscapes. Under RMF, the number of paths tended to be overestimated by very large factors (averages over seven and ten genes: paths: CBN 2.9× and 55.6×; OT: 5.1× and 128×; CAPRI: 3.5× and 61×), especially with 10 genes and sample sizes of 4000 (CBN: 112×; OT: 236×; CAPRI: 61×). Diversity was also overestimated but, as above, by smaller factors (averages over seven and ten genes: CBN 1.1× and 1.6×; OT:

2.1× and 2.8×; CAPRI: 3.0× and 3.5×; values for 10 genes and sample sizes of 4000: CBN: 2.2×; OT: 3.5×; CAPRI: 4.0×).

And how does the estimated evolutionary unpredictability change with the true evolutionary unpredictability? Figure 6C shows that the slopes of regressions of estimated unpredictability from CPMs (S_c) on true unpredictability (S_p) changed depending on fitness landscape, detection regime, and sample size, including slopes over and under 1, and even inversion of signs (ranges of slopes over all combinations of type of landscape by detection regime by number of genes by sample size: CBN: 0.47 to 1.27; OT: 0.43 to 1.50; CAPRI: -1.04 to 1.19); in contrast, those slopes remained basically constant over sample size if we simply regressed sample diversity on S_p.

3.5 Cancer data sets

We will use CPMs on eleven cancer data sets to examine their usefulness for predicting tumor evolution. These data include five different cancer types, with number of patients that range from 27 to 326 and number of features from seven to 192. Three data sets code mutations in terms both of genes and pathways (colon, glioblastoma, pancreas; genes, pathways). We have analyzed all the data sets with CBN (the best performing method —see sections 3.2 and 3.4). We have run the analysis three times per data set, limiting the number of features analyzed to the seven, ten, and 13 most common ones, so as to examine how our assessments depend on the number of features; the first two thresholds use the same number of features as the simulations. (Of course, for data sets with 7 or fewer features, there are no differences in the data sets used under the 7, 10, and 13 thresholds, so the values show below reflect variability between runs; ditto for data sets with 8 to 10 features with respect to thresholds 10 and 13).

We do not know the true paths of tumor progression, but we can use the bootstrap to asses the robustness of the inferences. To do so, we repeated the process above with 20 bootstrap samples (see Supplementary Material "Bootstrapping on the cancer data sets"). We measured the JS between the distribution of paths to the maximum from the original data set and each of the bootstrapped samples ($JS_{o,b}$). Large differences in the distribution of paths between the analyses with the bootstrap samples and the analysis with the original sample suggests that the inferences cannot be trusted (but small differences do not indicate that the inferred paths match the distribution of the true ones).

The results are shown in Figure 7. Performance ($JS_{o,b}$) was generally poor for most data sets, and very poor for some of them. These results would not be unexpected, even if the true fitness landscapes were representable ones, as most of the data sets have small sample sizes, and we have seen that performance (JS) is poor for that range of sample sizes (Figure 2A). When the same data set was analyzed using pathways and genes, performance was generally better using pathways. As the number of features analyzed increased from 7 to 10 to 13, the unreliability of the inferences generally increased too.

What determines the variation in $JS_{o,b}$ in Figure 7A? The pancreas pathways data has the smallest $JS_{o,b}$ and this contrasts with the much larger $JS_{o,b}$ of the same data analyzed using genes. A similar, though not as extreme, pattern appears in glioblastoma. Determinants of performance are complex and related to the association between features and distribution of genotypes. But one possible explanation is shown in Figure 7C: the data set for pancreas pathways contains 68 subjects with the same five pathways mutated, whereas in the pancreas genes most subjects have only a few mutations. In general, data sets where most subjects had only a few mutations, relative to the number of features in the data set, had very large $JS_{o,b}$. When most individuals in the sample have very few mutations, estimations of the restrictions on the downstream mutations are probably very noisy and, thus, different between bootstrapped data sets.

Values for S_c were well within the ranges of S_c estimated by CBN for the simulated data (see Supplementary Material, "Estimated S_c by CBN"). Of course, S_c increased with numbers of features analyzed. Given the results from section 3.4, where generally $S_p < S_c$, this suggests that the true evolutionary unpredictability (when analyzing up to 13 features) for eight of the data sets should be less than that corresponding to about 100 equiprobable paths to the maximum, but only four are below the much more manageable, and useful, 20 equiprobable paths. The pancreas pathways data set is here also an extreme case: examination of the output shows that there was one single path with estimated probability > 0.97. There was a positive association between S_c and $JS_{o,b}$ between the full and bootstrapped data sets (Figure 7 D), but some data sets with small S_c had very unreliable path predictions (e.g., colon and glioblastoma genes, MSI, MSS); conversely, the data set with pathway information from three cancers (All Pathways), even if it had the largest S_c had a moderate $JS_{o,b}$.

4 Discussion

Can we predict the likely course of tumor progression using CPMs? CBN, the best performing CPM method in our study, under the representable fitness landscapes (the easiest scenario, as it fits the underlying model), returned estimates of the probability of paths of tumor evolution that were not far from the true distribution of paths of evolution (Figure 2A) when sample size was very large. But we find that performance with moderate (and more realistic) sample sizes was considerably worse and was affected by detection regimen. The analysis of the cancer data sets revealed that performance ($JS_{o,b}$) was poor or very poor for most data sets. What factors, and how, affect performance?

Detection regime and LOD diversity (S_p) affected individually and jointly all performance measures (Figures 2, 4, 5), and increasing sample size improved all performance measures. CBN had better performance under the uniform detection regime (where more genotypes, with larger diversity, and larger dispersion in the number of mutations, are represented in the samples). But all performance measures (Figure 5, Supplementary Material "Slopes of regressions of recall and precision on LOD diversity"), improved with increasing unpredictability in the large and small detection regimes, specially with smaller samples sizes. The large and small detection regimes differ from the uniform in number of mutations, each in a different direction. What is common to both the large and small detection regimes is that increased evolutionary unpredictability leads to a larger range of observed genotypes. Increased unpredictability is thus associated to improved performance probably because it provides a surplus variability (but the improved performance does not rise up to the levels of the uniform detection regime). The summary messages from the simulations are, thus: a) increased evolutionary unpredictability hurts performance, unless the sampled genotypes (because of sample size or detection regime) have too low variability; b) detection regime can be a key determinant of performance, as already found in previous work (Diaz-Uriarte, 2015, 2018). The analysis of the cancer data sets reinforces the last conclusion: the distribution of mutations per sample is likely a major determinant of performance.

Performance was also affected by the number of features analyzed, the dimension of the fitness landscape. Performance in the simulated data sets (JS) was worse with 10 than with 7 seven genes (Figure 2) and, of course, S_p itself was larger under 10 genes (see Supplementary Material). In the eleven cancer data sets, both estimated unpredictability (S_c) and deviations of JS between the full and bootstrapped data sets ($JS_{o,b}$) increased with number of features. This result is not surprising, but brings forth the problem of the selection of the relevant features for analysis (Caravagna *et al.*, 2016; Cristea *et al.*, 2016; Gerstung *et al.*, 2011). We have shown that feature selection can have a very detrimental impact on the performance of CPM methods (Diaz-Uriarte, 2015). Using pathways instead of genes in the analyses (see, e.g., Cristea *et al.*, 2016; Raphael and Vandin, 2015) can alleviate some of the problems of feature selection. Pathways can also improve predictability and how close the estimates of path distributions are to the truth because they are more similar to heritable phenotypes, which often have smoother phenotype-fitness maps and tend to show more repeatable evolution (Lässig *et al.*, 2017; see also Wang *et al.*, 2015, but also Chebib and Guillaume, 2017; Sailer and Harms, 2017). Gerstung *et al.* (2011) found that analysis using pathways gave stronger evidence for order constraints than analysis using genes and we also see in Figure 7 that both S_c and $JS_{o,b}$ tend to decrease if we use pathways; “All Pathways” constitutes a promising case because it has large S_c but moderate $JS_{o,b}$.

Hosseini (2018) has reanalized the DAG-derived representable and a subset (those where the fully mutated genotype has largest fitness) of the DAG-derived non-representable fitness landscapes in Diaz-Uriarte (2018). He finds good agreement (small Kullback-Leibler divergence) between the distributions of paths to the maximum from CBN and the fitness landscape-based probability distribution of paths to the maximum. Our results for CBN under the best conditions are not as optimistic. Two differences in the studies explain the differences. First, Hosseini (2018) computes the fitness landscape-based probability of paths assuming a strong selection weak mutation regime, not by directly examining the actual distribution of the paths to the maximum in each simulation (i.e., he does not use the LODs) and, second, he uses CBN with the very large sample size of 20000 (the full data sets in Diaz-Uriarte, 2018).

An important caveat of the discussion so far (which also applies to Hosseini, 2018,’s results) is that very good performance simply tells us that the true and estimated probability distributions of the paths to the maximum agree closely. If the true evolutionary unpredictability is large, then for practical purposes our capacity to predict what will happen (in the sense of providing a small set of likely outcomes) is very limited. To give a feeling for these values, 25 equiprobable paths have a diversity of 3.2, and 400 equiprobable paths a diversity of 6.0 (see also Figure 7), values of diversity comparable to those seen for representable fitness landscapes of seven and ten genes in our data (see Supplementary Material), and similar to those of several of the eleven cancer data sets. The inability to narrow down the likely paths to a small set of paths in these cases is, of course, not a limitation of the methods, but a problem inherent to the unpredictability of the evolutionary process in many scenarios, which could severely limit the usefulness of even perfect predictions.

We have focused on three families of methods. The behavior of CAPRI contrasted with that of OT and CBN in terms of recall: it improved little or even degraded with increasing sample size. That is because CAPRI allowed few paths to the maximum (it encodes too many restrictions in the DAGs of restrictions), and their number decreased with sample size (Supplementary Material “Number of paths inferred” and Figure 6A). CAPRI’s precision is not surprising: the few paths to the maximum that it allowed were actually used by evolution. This is related to the increased false negative discoveries for CAPRI discussed in Diaz-Uriarte (2018). Briefly, because of both the objective of CAPRI (identification of “probability raising” relationships) and its workings (fitting a Bayesian Network using penalized likelihood after filtering relationships to fit probability raising and temporal priority), it is not possible to obtain probabilities of paths (section 2.4) but, moreover, the results of CAPRI cannot be mapped unambiguously to sets of possible and non-possible paths of tumor evolution, in contrast to OT and CBN (see details in Supplementary Material “CAPRI, CAPRESE, and paths of tumor progression”).

The above discussion has centered on representable fitness landscapes. As argued above, it is reasonable to suspect fitness landscapes with local fitness maxima are common in cancer. Interestingly, for small sample sizes, recall was sometimes better in DAG-derived and RMF than under representable landscapes (Figure 2): with local fitness maxima, achieving good recall involves the relatively easier task of getting right the first part of short paths to the maximum (see Supplementary Material,“Number of mutations of local maxima and performance”, where 1-recall increases with the average number of mutations of local fitness maxima). But good recall was more than offset by the low precision: overall predictability was very poor. The decrease in precision is the consequence of local fitness maxima: CPMs are fitting models with paths of tumor progression that extend beyond the true end point of the progression. In addition, RMF fitness landscapes strongly violate the CPM assumption that acquiring a mutation in one gene does not decrease the probability of acquiring a mutation in another gene (see Diaz-Uriarte, 2018).

Returning to our third original question, even if achieving good performance in predicting the actual paths of tumor progression is unlikely, inferring evolutionary unpredictability could be an easier task. Can we use inferences of evolutionary unpredictability from CPMs as estimates of the true evolutionary unpredictability? Under representable fitness landscapes, CBN, the best performing method also for this task (Figure 6B), returned values of S_c very similar to S_p, the evolutionary unpredictability estimated from the diversity of paths, and this held over detection regimes and sample sizes. Hosseini (2018) also finds that the estimates of predictability from CBN correlate well with the true evolutionary predictability, and his slopes of the regression of CPM-based on landscape-based predictability are generally slightly below 1, which agrees with our Figure 6C (left-most column). These good results do not hold under the other two fitness landscapes: evolutionary unpredictability is overestimated, and increasing sample sizes makes the problems worse. Could we do better by trying to infer the true evolutionary unpredictability by using an inverse regression procedure of a regression of CPM path diversity (estimated unpredictability, S_c) on LOD diversity (true unpredictability, S_p)? As shown in Figure 6C this is also unlikely to succeed: different scenarios, sample sizes, and detection regimes have different relationships of estimated predictability regressed on true unpredictability, and marginalizing over all those factors leads to a very large spread around a single regression line (see Supplementary Material "Regression of individual CBN unpredictability estimates on LOD diversity"). Trying the inverse regression approach properly with experimental data would require that many other details of the process (e.g., type of landscape, detection regime) were known, and this are currently unknown (and unlikely to be known in the future). But our results indicate that we can use CBN to set upper bounds on the true S_p; obtaining tighter estimates is an issue for further research to explore. And here our analysis of eleven cancer data sets suggests that the true evolutionary unpredictability of at least some cancer scenarios might be reasonably small, specially if S_c is overestimating the true unpredictability.

4.1 Conclusion

The answer to the question "can we predict the likely course of tumor progression using CPMs?" is, unfortunately, "only with moderate success and only under representable fitness landscapes and with very large sample sizes; but even perfect predictions might be of little use if evolutionary unpredictability is large". Estimating upper bounds to evolutionary unpredictability is a more modest, though more likely to succeed, use of CPMs. There are three key difficulties for successful prediction: the sheer size of the problem even for moderate numbers of genes, the intrinsic evolutionary unpredictability in many scenarios, and the deviations from the assumptions of CPMs that are likely to hold in most cancer data. In addition to the caveat about using methods under scenarios where performance is very poor, this paper raises the general question of what can we really predict about likely paths of tumor progression from cross-sectional data, for instance to guide therapeutic interventions.

5 Acknowledgements

S. Alvarez-Tolcheff for help writing some of the code to generate to local maxima fitness landscapes. N. Beerenwinkel, S. Posada-Céspedes, and G. Caravagna for discussion about progression models or software, and S.-R. Hosseini for providing a preprint of his MSc. thesis.

Supported by BFU2015-67302-R (MINECO/FEDER, EU). CVV supported by PEJD-2016/MED-2116 from Comunidad de Madrid.

References

Attolini, C. et al (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.

Bamford, S. et al (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer*, **91**(2), 355–358.

Bank, C. et al (2016). On the (un)predictability of a large intragenic fitness landscape. *PNAS*, **113**(49), 14085–14090.

Beerenwinkel, N. et al (2007). Genetic progression and the waiting time to cancer. *PLoS computational biology*, **3**(11), e225.

Beerenwinkel, N. et al (2015). Cancer evolution: Mathematical models and computational inference. *Systematic Biology*, **64**(1), e1–e25.

Beerenwinkel, N., Greenman, C.D. and Lagergren, J. (2016). Computational Cancer Biology: An Evolutionary Perspective. *PLoS Comput. Biol.*, **12**(2), e1004717.

Beijersbergen, R.L., Wessels, L.F.A. and Bernards, R. (2017). Synthetic Lethality in Cancer Therapeutics. *Annual Review of Cancer Biology*, **1**(1), 141–161.

Bell, D. et al (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**(7353), 609–615.

Blomen, V.A. et al (2015). Gene essentiality and synthetic lethality in haploid human cells. *Science*, **350**(6264), 1092–1096.

Bozic, I. et al (2010). Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 18545–18550.

Brooks, M.E. et al (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, **9**(2), 378–400.

Brouillet, S. et al (2015). MAGELLAN: A tool to explore small fitness landscapes. *bioRxiv*, page 031583.

Caravagna, G. et al (2016). Algorithmic methods to infer the evolutionary trajectories in cancer progression. *PNAS*, **113**(28), E4025–E4034.

Chebib, J. and Guillaume, F. (2017). What affects the predictability of evolutionary constraints using a G-matrix? The relative effects of modular pleiotropy and mutational correlation. *Evolution*, **71**(10), 2298–2312.

Chiotti, K.E. et al (2014). The Valley-of-Death: Reciprocal sign epistasis constrains adaptive trajectories in a constant, nutrient limiting environment. *Genomics*, **104**(6, Part A), 431–437.

Cristea, S., Kuipers, J. and Beerenwinkel, N. (2016). pathTiMEx: Joint Inference of Mutually Exclusive Cancer Pathways and Their Progression Dynamics. *Journal of Computational Biology*.

Crona, K., Greene, D. and Barlow, M. (2013). The peaks and geometry of fitness landscapes. *Journal of Theoretical Biology*, **317**, 1–10.

Crooks, G.E. (2017). On measures of entropy and information. Technical report.

de Visser, J.A.G.M. and Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. *Nat Rev Genet*, **15**(7), 480–490.

Desper, R. et al (1999). Inferring tree models for oncogenesis from comparative genome hybridization data. *J Comput Biol*, **6**(1), 37–51.

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), 0–36.

Diaz-Uriarte, R. (2017). OncoSimulR: Genetic simulation with arbitrary epistasis and mutator genes in asexual populations. *Bioinformatics*, **33**(12), 1898–1899.

Diaz-Uriarte, R. (2018). Cancer progression models and fitness landscapes: A many-to-many relationship. *Bioinformatics*, **34**(5), 836–844.

Ding, L. et al (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, **455**(7216), 1069–1075.

Ferrari, S. and Cribari-Neto, F. (2004). Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, **31**(7), 799–815.

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression, 2nd Ed*. Sage, Thousand Oaks, CA.

Franke, J. et al (2011). Evolutionary Accessibility of Mutational Pathways. *PLoS Comput Biol*, **7**(8), e1002134.

Gerstung, M. et al (2009). Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics (Oxford, England)*, **25**(21), 2809–2815.

Gerstung, M. et al (2011). The Temporal Order of Genetic and Pathway Alterations in Tumorigenesis. *PLoS ONE*, **6**(11), e27136.

Greaves, M. (2015). Evolutionary Determinants of Cancer. *Cancer Discovery*, **5**(8), 806–820.

Grün, B., Kosmidis, I. and Zeileis, A. (2012). Extended Beta Regression in *R* : Shaken, Stirred, Mixed, and Partitioned. *Journal of Statistical Software*, **48**(11).

Hosseini, S.R. (2018). *Quantifying the Predictability of Cancer Progression Using Conjunctive Bayeisan Networks*. Ph.D. thesis, ETH, Zurich.

Hothorn, T., Bretz, F. and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biom J*, **50**(3), 346–363.

Jones, S. et al (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science (New York, N.Y.)*, **321**(5897), 1801–6.

Lässig, M., Mustonen, V. and Walczak, A.M. (2017). Predicting evolution. *Nature Ecology & Evolution*, **1**(3), s41559–017–0077–017.

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

Lipinski, K.A. et al (2016). Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. *Trends in Cancer*, **2**(1), 49–63.

Loohuis, L.O. et al (2014). Inferring Tree Causal Models of Cancer Progression with Probability Raising. *PLoS ONE*, **9**(10), e108358.

Losos, J.B. (2018). *Improbable Destinies: Fate, Chance, and the Future of Evolution.* Riverhead Books, S.l.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, 2nd Ed*. Chapman and Hall/CRC, London.

McFarland, C.D. et al (2013). Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(8), 2910–5.

McPherson, A.W., Chan, F.C. and Shah, S.P. (2018). Observing Clonal Dynamics across Spatiotemporal Axes: A Prelude to Quantitative Fitness Models for Cancer. *Cold Spring Harb Perspect Med*, **8**(2).

Misra, N., Szczurek, E. and Vingron, M. (2014). Inferring the paths of somatic evolution in cancer. *Bioinformatics (Oxford, England)*, **30**(17), 2456–2463.

Montazeri, H. et al (2016). Large-scale inference of conjunctive Bayesian networks. *Bioinformatics*, **32**(17), i727–i735.

Neidhart, J., Szendro, I.G. and Krug, J. (2014). Adaptation in Tunably Rugged Fitness Landscapes: The Rough Mount Fuji Model. *Genetics*, **198**(2), 699–721.

Network, T.C.G.A. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**(7407), 330–337.

Nowak, M.A. et al (2004). Evolutionary dynamics of tumor suppressor gene inactivation. *PNAS*, **101**(29), 10635–10638.

O'Neil, N.J., Bailey, M.L. and Hieter, P. (2017). Synthetic lethality and cancer. *Nat Rev Genet*, **18**(10), 613–623.

Palmer, A.C. and Kishony, R. (2013). Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance. *Nature Reviews Genetics*, **14**(4), 243–248.

Parsons, B.L. (2008). Many different tumor types have polyclonal tumor origin: Evidence and implications. *Mutation research*, **659**(3), 232–47.

Poelwijk, F.J. et al (2007). Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, **445**(7126), 383–6.

Poelwijk, F.J. et al (2011). Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *Journal of Theoretical Biology*, **272**(1), 141–144.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Ramazzotti, D. et al (2015). CAPRI: Efficient inference of cancer progression models from cross-sectional data. *Bioinformatics*, **31**(18), 3016–3026.

Raphael, B.J. and Vandin, F. (2015). Simultaneous Inference of Cancer Pathways and Tumor Progression from CrossSectional Mutation Data. *Journal of Computational Biology*, **22**(00), 250–264.

Sailer, Z.R. and Harms, M.J. (2017). Molecular ensembles make evolution unpredictable. *PNAS*, **114**(45), 11938–11943.

Sjoblom, T. et al (2006). The Consensus Coding Sequences of Human Breast and Colorectal Cancers. *Science*, **314**(5797), 268–274.

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, **11**(1), 54–71.

Sniegowski, P.D. and 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.

Szabo, A. and Boucher, K.M. (2008). Oncogenetic trees. In W.-Y. Tan and L. Hanin, editors, *Handbook of Cancer Models with Applications*, pages 1–24. World Scientific.

Szendro, I.G. et al (2013). Predictability of evolution depends nonmonotonically on population size. *PNAS*, **110**(2), 571–576.

Tomasetti, C. et al (2015). Only three driver gene mutations are required for the development of lung and colorectal cancers. *PNAS*, **112**(1), 118–123.

Toprak, E. et al (2012). Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nature Genetics*, **44**(1), 101–105.

Wang, E. et al (2015). Predictive genomics: A cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. *Seminars in Cancer Biology*, **30**, 4–12.

Wodarz, D. and Komarova, N.L. (2014). *Dynamics of Cancer: Mathematical Foundations of Oncology*.

Wood, L.D. et al (2007). The Genomic Landscapes of Human Breast and Colorectal Cancers. *Science*, **318**(5853), 1108–1113.

6 Figures

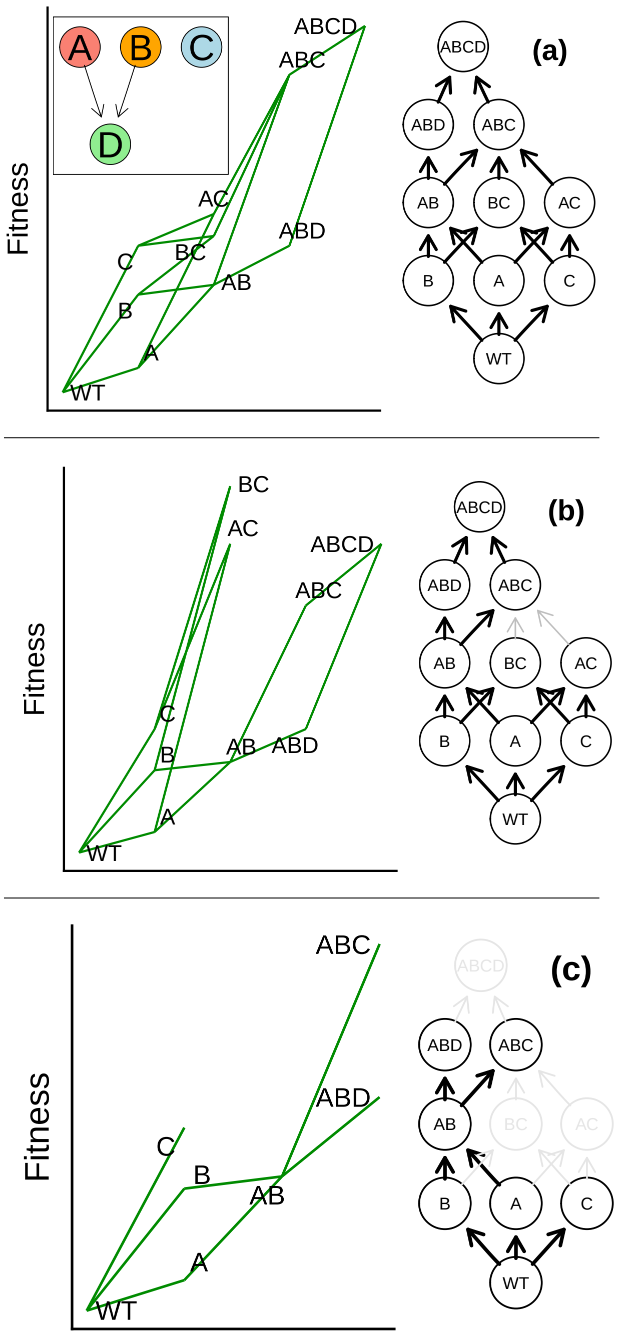

Figure 1: Fitness landscapes, paths of tumor progression, and DAGs of restrictions in the order of accumulation of mutations for the three types of landscapes used. (a) representable; (b) local maxima; (c) RMF. In each row, on the left, the fitness landscape (representation based on Brouillet *et al.*, 2015) that shows the accessible genotypes (where the notation "AB" means a genotype with both genes A and B mutated) and on the right the fitness graphs or graphs of mutational paths (Crona *et al.*, 2013; de Visser and Krug, 2014; Franke *et al.*, 2011), where nodes are genotypes and arrows point toward mutational neighbors of higher fitness. Fitness graphs show all the paths of tumor progression, the set of accessible mutational paths and adaptive walks that, under the restriction that there can be no back mutations, start from the "wild type" (WT) genotype —where we absorb all cancer initiation events— and end in the local fitness maxima (or single global fitness maximum). Each path from WT to a maximum corresponds to a different Line of Descent (LOD). For (b) and (c), gray edges and nodes denote those that are present in (a) but missing in (b) or (c). The inset in the first row shows the DAG of restrictions in the order of accumulation of mutations that applies to (a) and (b). A DAG of restrictions shows genes in the nodes; an arrow (directed edge) from gene *i* to gene *j* indicates a direct dependency of a mutation in *j* on a mutation on *i*; a mutation in *j* cannot be observed unless *i* is mutated. In the example, a mutation in gene D can only be observed if both A and B are mutated; the absence of an arrow between two genes indicates a lack of direct dependencies between the two genes. The set of genotypes that can exist under both (a) and (b) is the same, and all of them satisfy the restrictions in the DAG of restrictions. But the fitness landscape in (b) has three maxima; there are fewer paths to "ABCD" and several paths end in the other two maxima ("AC", "BC"). Thus, the fitness graph of (b) does not fulfil the assumptions of CPMs. The defining features of (b) are that the set of accessible genotypes can be represented by a DAG of restrictions, but there are missing paths. The fitness landscape in (c) cannot be represented by any DAG of restrictions; e.g., no DAG of restrictions can account at the same time for the presence of genotypes "A", "B", "C", and the absence of every double mutant with "C". Relative to (a), (c) is missing both paths and genotypes (relative to other DAGs of restrictions it could either be missing and/or adding genotypes and paths).

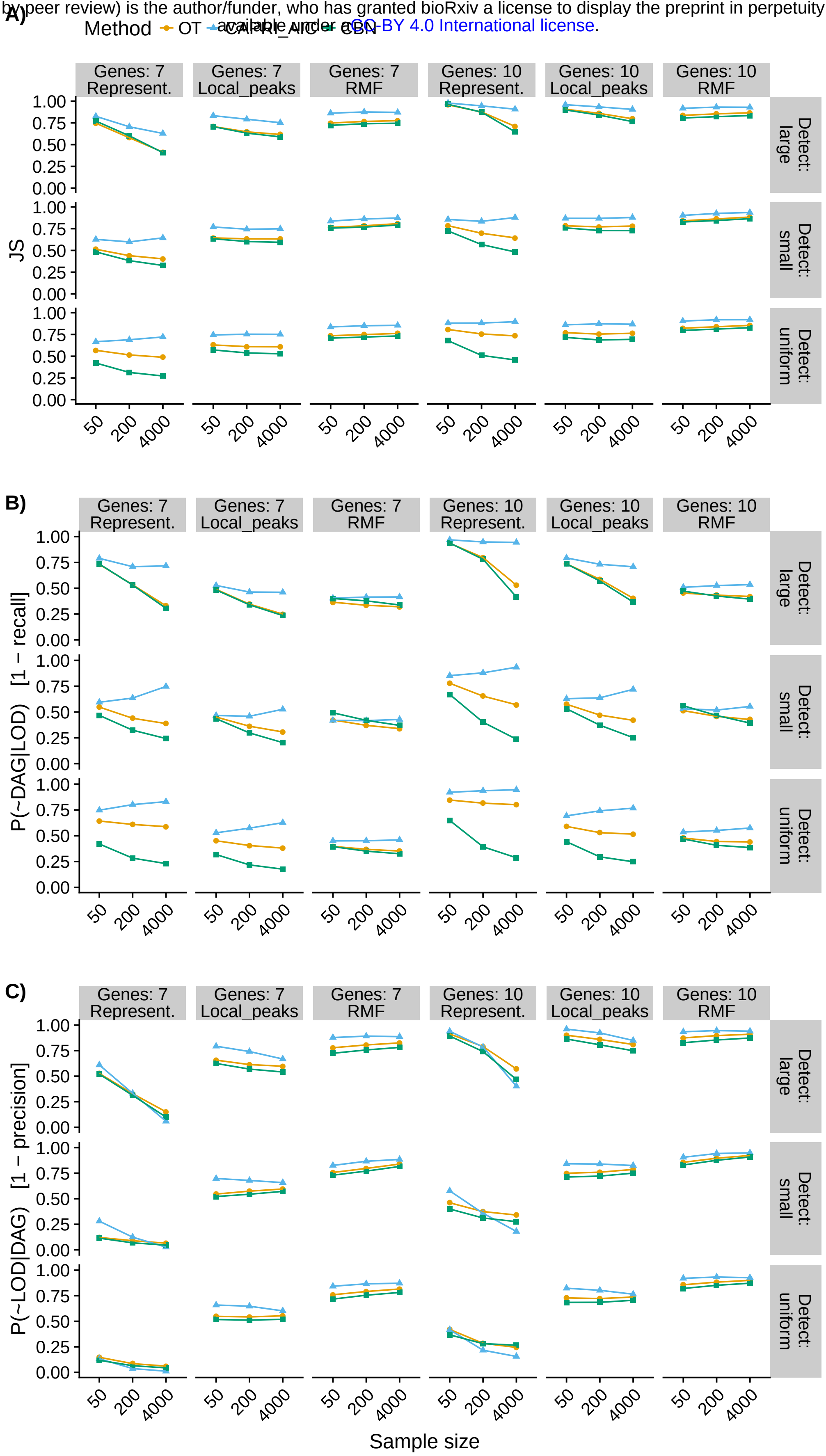

Figure 2: Summary performance measures (see definitions in 2.4) for OT, CAPRI (with AIC penalty) and CBN for all combinations of sample size by type of landscape by detection regime by number of genes. For all measures, smaller is better. For OT and CBN, Jensen-Shannon divergence (JS) and 1-precision use probability-weighted predicted paths (see text). Each point represented is the average of 210 points (35 replicates of each one of the six combinations of 3 initial size by 2 mutation rate regimes —see 2.1); we are thus marginalizing over mutation rate by initial simulation size combinations. Each one of the 210 points is, itself, the average of five runs on different partitions of the simulated data. See Supplementary Material, "Overall patterns for the six methods" for results for all six methods used.

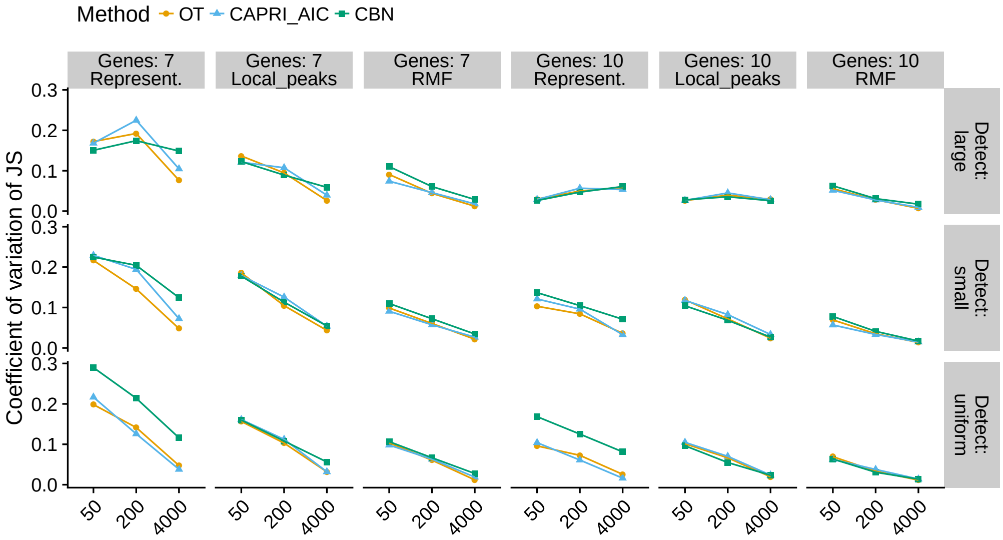

Figure 3: Coefficient of variation (standard deviation/mean) of JS for each method for all combinations of sample size by type of landscape by detection regime by number of genes. The coefficient of variation has been computed from the five runs for each landscape on each combination of sample size and detection regime. For OT and CBN, JS is computed using the probability-weighted predicted paths (see text). Each point plotted is the average of 210 points (see Figure 2).

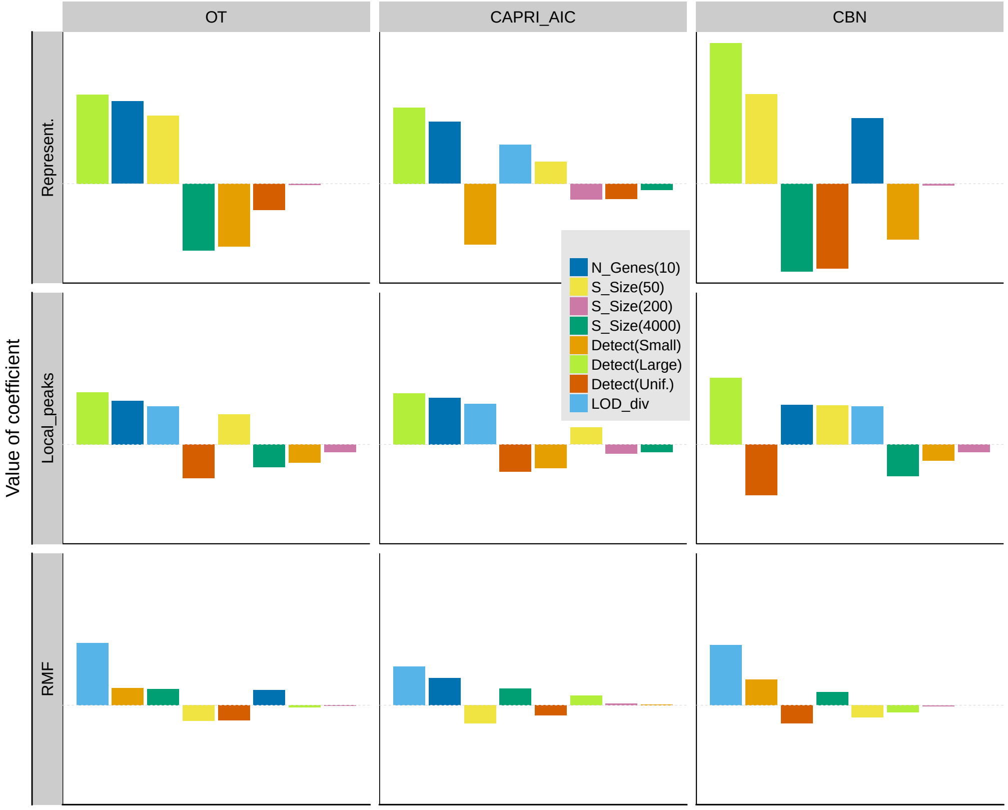

Figure 4: Coefficients from generalized linear mixed-effects models, with separate models fitted for each combination of method and type of fitness landscape. Coefficients are from models with sum-to-zero contrasts (see text and Supplementary Material). Within each panel, coefficients have been ordered from left to right according to decreasing absolute value of coefficient. The dotted horizontal gray line indicates 0 (i.e. no effect). Coefficients with a large positive value indicate factors that lead to a large decrease in performance. Only coefficients that correspond to a term with a P-value $<$ 0.05 in Type II Wald chi-square tests are shown. The coefficient that corresponds to Number of genes 7 is not shown (as it is minus the coefficient for 10 genes —from using sum-to-zero contrasts). "N_Genes": number of genes; "S_Size": sample size; "Detect": detection regime; "LOD_div": LOD diversity (S_p).

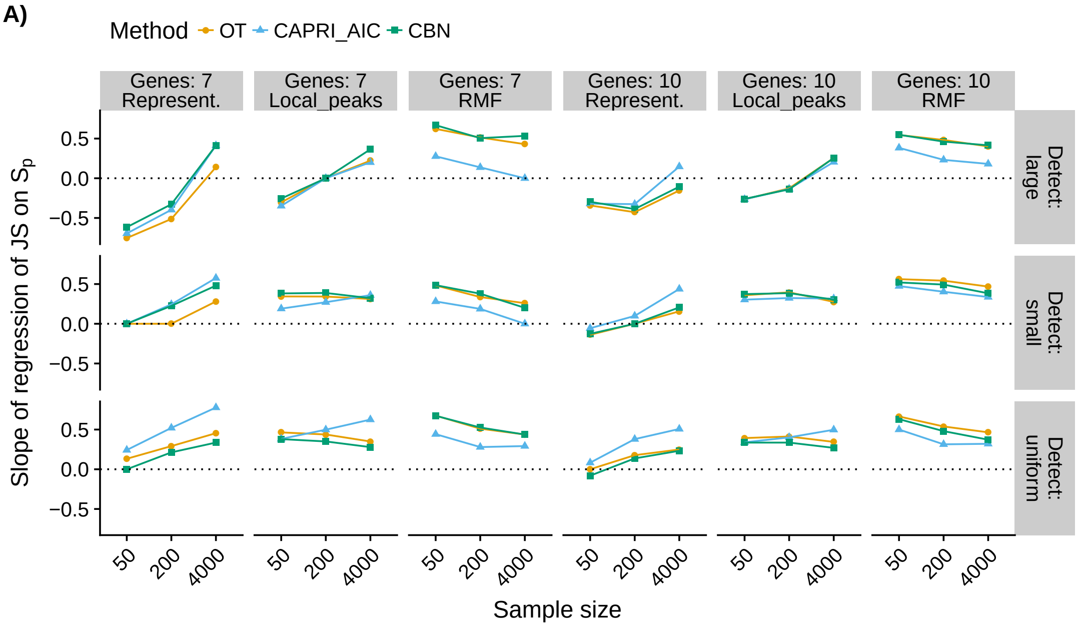

Figure 5: Estimated slopes of the regression of Jensen-Shannon divergence (JS) on LOD diversity (S_p) for all combinations of type of landscape by detection regime by number of genes by sample size. A beta regression was fitted to each subset of data. Slopes not significantly different from 0 ($P > 0.05$) shown as 0. Each regression was fitted to 210 points, each of which is itself the average of five replicates, one for each of the five runs on different partitions of the simulated data.

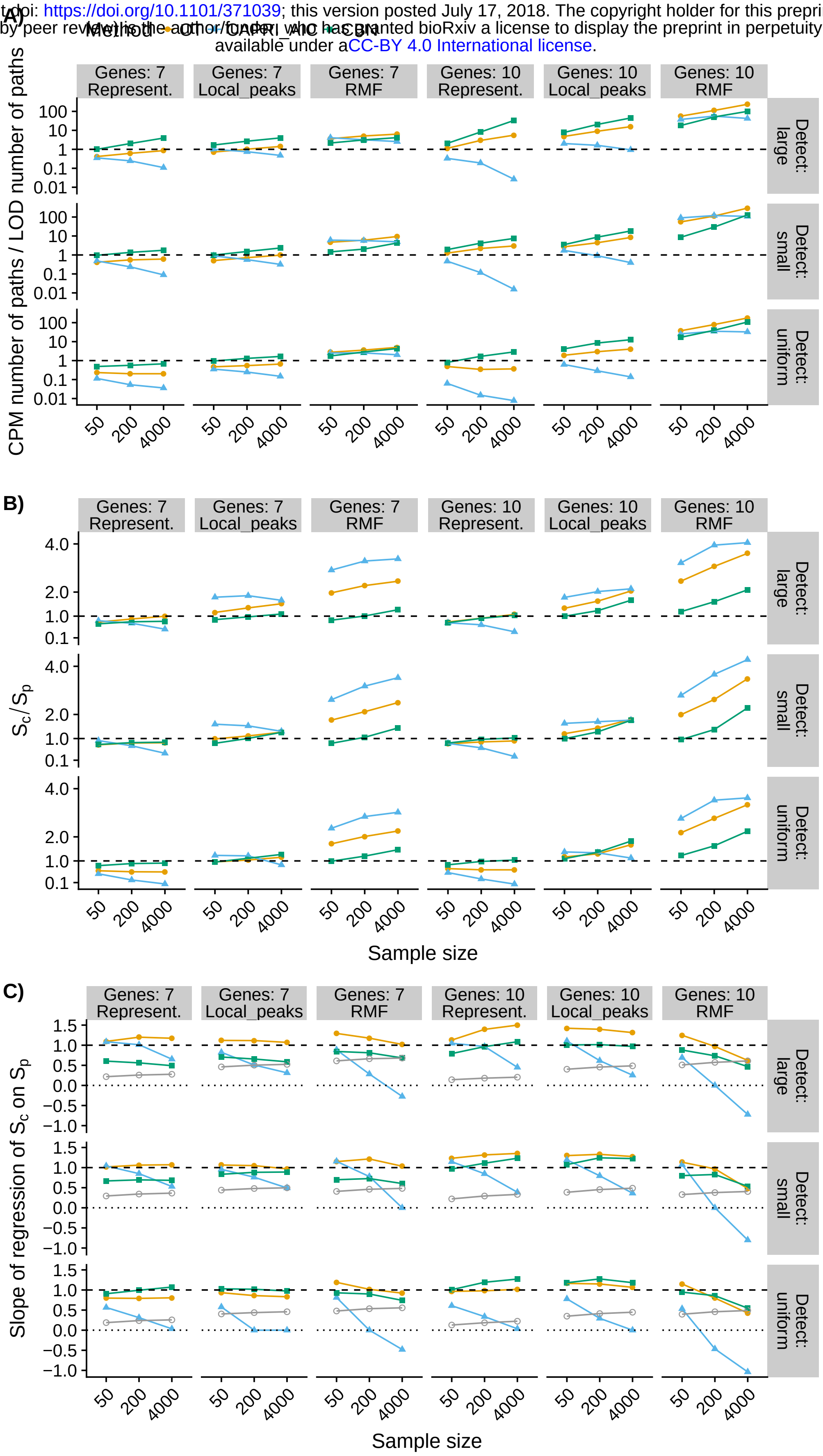

Figure 6: Number of paths and path diversities inferred from CPMs relative to the true values from LODs. Panel A: average of the ratio of number of paths to the maximum from the CPMs relative to the observed number of distinct LODs for all combinations of type of landscape by detection regime by number of genes by sample size. Panel B like panel A, but for diversities of paths to the maxima. As in Figure 2, each point represented is the average of 210 points. Panel C shows the slope of the regression S_c on S_p; each point is thus a slope from a regression of 210 points, each of which is itself the average of 5 replicates (see Figure 5). For comparison, panel C shows also the regression of diversity of the observed genotype samples on S_p (gray line). Panels B and C show different features of the data: panel B shows whether evolutionary unpredictability (S_p) tends to be over- or under-estimated by S_c; panel C shows how S_c changes with S_p —see Supplementary Material, "LOD and CPM diversity: ratios and slopes" for an example of positive ratios with negative slopes.

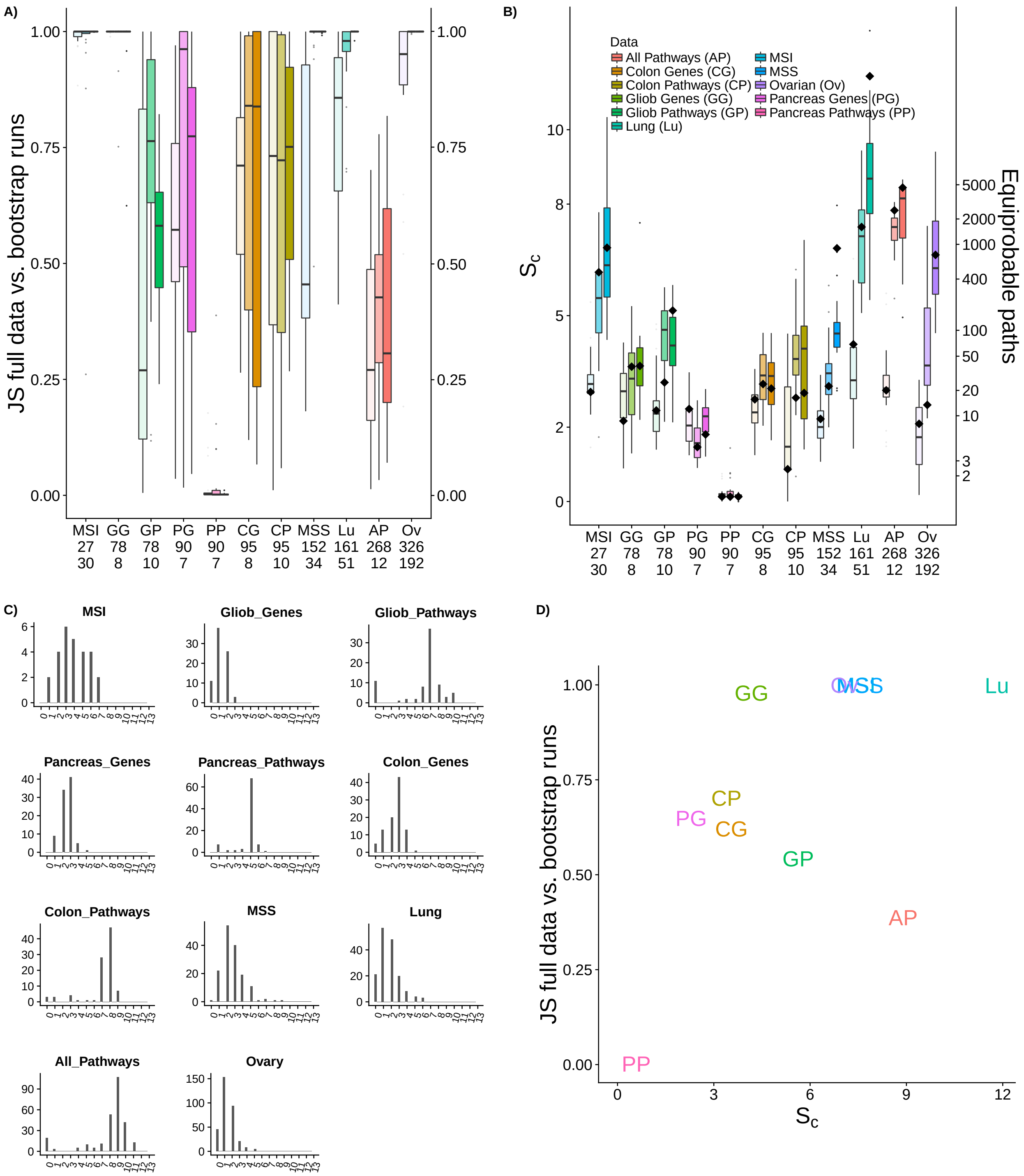

Figure 7: Results from the cancer data sets analyzed with CBN. In A) and B), data sets have been ordered by increasing sample size, and the x-axis labels provide the acronym (shown in full in the inset legend). Below the data set acronym are the sample size and the total number of features, respectively. We used CBN for all analyses. Analysis were run three times, limiting the number of features analyzed to the seven, ten, and 13 most common ones; the boxplots for each data set are shown in increasing order of number of features. For data sets such as, say, Pancreas genes, using 7, 10, or 13 maximum features makes no difference in the actual number of features analyzed; the three replicate runs show run-to-run variability. A) $JS_{o,b}$: JS statistic for the comparison of the distribution of paths from running CBN on the original data set against the distribution of paths from running CBN on each one of the bootstrap runs. B) Diamonds show the S_c from the full data, and boxplots the S_c from the boostrap runs. Right axis labeled by number of equiprobable paths equivalent to the S_c. C) Histograms of number of mutations per individual in the data set. D) Scatterplot of the mean of the JS statistic (panel A) vs. S_c (both from analyses with up to 13 features).