Supplementary material for “Every which way? On predicting tumor evolution using cancer progression models”

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

2018-11-16 (Release: Rev: bea16f1)

Contents

*ramon.diaz@iib.uam.es, rdiaz02@gmail.com, http://ligarto.org/rdiaz

List of Figures

1. Generating random fitness landscapes

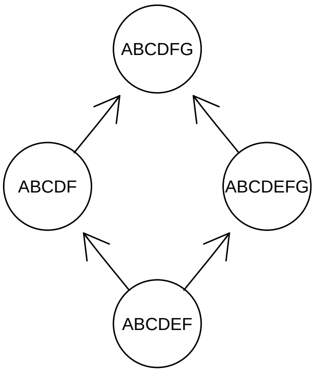
For the representable and local-maxima fitness landscapes, we started by generating random DAGs. Since no agreed upon model exists for the distribution of DAGs in CPMs, we have used two different procedures, choosing each one randomly with the same probability. One of the procedures uses the function `simOGraph`, in the OncoSimulR package. To generate random DAGs with `simOGraph` for N genes, the genes were first randomly split in a number of levels, where the number of levels used was a randomly chosen integer between 3 and $N - 1$, both included. Then, each gene from each level i was randomly connected (as descendant) to randomly chosen genes (the ancestors) from levels j, where $j < i$; the number of incoming connections of each gene is a randomly chosen integer between 1 and $maxp$ (both included), where $maxp$ is a randomly chosen integer between 2 and $N - 2$ ($maxp$ is common for all genes in a DAG, but can vary between DAGs). The DAG we use is the transitive reduction of the above generated DAG. (Note that this procedure can occasionally result in star DAGs, i.e., DAGs without any dependencies; in such a case, the DAG was discarded and a new one obtained). The other procedure uses the function `random_poset` in package MC-CBN (https://github.com/cbg-ethz/MC-CBN); this function is undocumented, but it returns the transitive reduction of a randomly-filled adjacency matrix for a DAG where the initial number of non-zero connections is equal to the number of possible connections times a constant; we used the default value of that constant (0.15).

1.1. Local maxima without reciprocal sign epistasis?

As explained in the paper, creating fitness landscapes with local maxima generally results in creating reciprocal sign epistasis and the number of local fitness maxima is associated with reciprocal sign epistasis —see Figures in section 4. There were, however, seven cases (out of 420) where introduction of local fitness maxima did not lead to introduction of reciprocal sign epistasis. These cases (which can be seen in the files referred to in section 2) are landscapes with IDs “7E10pIyu7UguIUl8I”, “8QlFQCUlUVfC10PZr”, “ bCsk2Qo5VMVm55fM”, “GedZaWDeb1029Mf88”, “ hw8kQ4g44p4XAkDa”, “WpF105HbEDoECa8vs”, “t1yUXsv5fVuo10GRi”. To look at one example in detail, we will use “GedZaWDeb1029Mf88” (it is the smallest one). The fitness of four of the relevant genotypes are

```
ABCDFG : 2.007
ABCDEFG: 1.8
ABCDF: 1.749
ABCDEF: 1.712
```

so there is no reciprocal sign epistasis (use, for example, the graphical criterion in [15] or [23]) and “ABCDFG” is a global maximum.



Of course, under an evolutionary model that assumes no back mutations (as is the case for CPMs), two of those transitions, those that involve loosing “E”

(`ABCDEF -> ABCDF` and `ABCDEFG -> ABCDFG`) are not allowed, leading to two local fitness maxima (`ABCDFG, ABCDEFG`).

Note also that here, for that set of four genotypes, mutating gene E decreases fitness. But mutating E increases fitness in genotypes "ABC" or "ABCD". Thus, this fitness landscape does not fulfill either the assumption that a mutation never decreases the probability of acquiring other mutations (even if the fraction of pairs of genotypes with reciprocal sign epistasis is 0). Regardless, one can also simply focus on the fact that this fitness landscape contains local maxima (and is missing paths relative to the corresponding fitness graph from the DAG of restrictions).

1.2. Rough Mount Fuji

In the Rough Mount Fuji fitness landscapes the reference genotype (i.e., the genotype with maximum fitness) was randomly chosen (setting `reference = 'random'` in the `rfitness` function in OncoSimulR). The standard deviation, *sd*, of the random normal variate was set to 0.2 and the decrease in fitness (strictly, birth rate) of a genotype per each unit increase in Hamming distance from the reference genotype, c, was chosen from a uniform $U(0, 0.2)$ distribution. This gives a wide variety of fitness landscapes that encompass from close to additive (large values of c) to House of Cards (c close to 0), with maximum fitness (birth rate) comparable to those of the representable and local-peaks fitness landscapes.

The generated RMF fitness landscape was checked to ensure that all 7 or 10 genes were present in at least one accessible genotype; if they were not, a new fitness landscape was generated (with, possibly, different values of c and reference genotype). Function `rfitness` from the OncoSimulR package [19] was used.

2. Plots of fitness landscapes and inferred DAGs

Files `fl-fg-7.pdf` and `fl-fg-10.pdf` show the 1260 fitness landscapes used. Each `fl-fg-x.pdf` shows the 630 fitness landscapes used for x genes. In each PDF, the first 210 fitness landscapes are fully representable fitness landscapes, the next 210 (pages 211 to 420) are the "Local maxima" fitness landscapes. The next 210 (pages 421 to 630) correspond to Rough Mount Fuji (RMF) fitness landscapes. Landscapes are not ordered according to any criterion within type of fitness landscape.

In each page, the following figures/tables are shown:

DAG of restrictions (top left) The true DAG of restrictions. This only applies properly to the representable fitness landscapes. In the local maxima fitness landscapes, not all paths between genotypes are available. (See below). For the RMF landscapes this, of course, is not available, since there is no underlying DAG of restrictions.

The true DAG of restrictions could be a tree. Among the representable fitness landscapes, 30% and 2%, of the landscapes with 7 and 10 genes, respectively, were trees (i.e., can be represented by OT and CAPRESE).

Fitness landscape (bottom left) As it says, the fitness landscape. Boxes surround the fitness maxima in the 7-genes landscapes. To minimize clutter, genotype labels are not shown in the 10-genes landscapes.

ID and landscape characteristics (center) The ID of the fitness landscape (a random string that matches the value of ID in the data tables), the number of accessible genotypes, the number of fitness maxima or peaks. Values for "removed edges" and "prop rem edges" denote, respectively, the number and proportion of edges in the fitness graph that were removed; this only applies to the "Local maxima" landscapes; thus, these values are 0 for "Representable" fitness landscapes, and NA for RMF fitness landscapes.

Fitness graph from DAG of restrictions (top right) The fitness graph implied by the DAG of restrictions. Not available for RMF (since there is no DAG of restrictions).

True fitness graph (bottom right) The actual fitness graph that corresponds to the fitness landscape. For "Representable" fitness landscapes this would be the same as the Fitness graph from DAG of restrictions, so we do not show it (to make the size of the files smaller).

3. Simulations

3.1. Runs until fixation

Simulations were run until fixation of a genotype, where the genotype was one of the genotypes among the local maxima (or the single global maximum). We used OncoSimulR, with the `fixation` option (introduced in version 2.9.8 of the program). A genotype was considered to have been fixated if it maintained a proportion ≥ 0.98 during 15000 consecutive sampling periods (this means that if after reaching a minimum frequency ≥ 0.98, at any time the proportion became smaller than 0.98 the counter of successive periods was reset to 0).

Why not require a proportion of 1.0 as evidence of fixation? Because for local maxima, if mutation rate is larger than 0 and neighboring genotypes have non-zero birth rate, the fixated genotype can occasionally generate descending genotypes that exist, with small frequencies, for short periods of time. Using much shorter number of consecutive sampling periods such as 1000 or 5000 did not produce different results over using 15000 in trial runs; however, to err on the safe side and make sure fixation had been established, we used that overly long period.

We excluded these 15000 periods from the computation of clonal interference statistics.

3.2. All genes part of lines of descent with frequency > 0.001

When the 20000 simulations were completed, we verified that the frequency of all genes in the last genotypes (i.e., the fixated genotypes or the final genotypes of the LODs) were at least 0.001. If they were not, a new fitness landscape was generated and the processes started again. In other words, we avoided fitness landscapes that have a nominal number of, say, 10 genes, but where a smaller number of genes were effectively ever part of the paths of tumor progression (this issue can affect the local maxima and RMF landscapes). In a sample of 4000 individuals, the probability that a gene with a true frequency of 0.001 is never part of a LOD is about 0.018 (= $(1 - 0.001)^{4000}$), so less than 2%. Of course, at the smallest value of the threshold, some data sets of 4000 might have at least one gene missing, and that probability is larger for data sets of 50 and 200.

3.3. Detection regimes: sampling

For each detection regime, we generated 20000 random deviates (called r, below) from the specified beta distribution ($\mathrm{B}(1,1)$, $\mathrm{B}(5,3)$, and $\mathrm{B}(3,5)$ (for uniform, large, and small, respectively).

Using those random deviates, we defined the target size of each sample as $t = \exp(r\ (\ln(M) - \ln(m)) + \ln(m))$, where M and m are the largest and smallest values, respectively, of population sizes ever attained in any of the 20000 simulations. Thus, we obtain target sizes that are uniform or biased towards large sizes or biased towards small sizes in the log scale. In the model of [32], tumor population size increases logarithmically with number of driver mutations. Therefore, uniform, small, and large biases would correspond to approximately uniform, small, or large in terms of number of driver mutations.

For each of the 20000 simulations, the actual sample was the one corresponding to the first sampling period at which the total tumor size achieved a value equal to, or larger than, t. If all values of tumor population size were $> t$, we returned the sample with the largest population size, and if all values were $< t$ the sample with smallest size.

This procedure determines at which of the sampling times we take the sample. The actual genotype returned is the single genotype with the largest frequency. Thus, we are not emulating taking a biopsy of the entire tumor or bulk sequencing but, rather, single-cell sampling, and sampling the single most common genotype.

We carried the above steps using OncoSimulR's function `samplePop`, with the values of t (thresholded as explained for $> t$ and $< t$) as arguments to `popSizeSample` and using `typeSample = 'single'`.

3.4. Other parameters of the simulations

Simulations used the implementation of the McFarland model in the OncoSimulR package [19]. In addition to the parameters specified in the main text, other parameters for the simulations on the fitness landscapes were (see specific meaning in documentation of OncoSimulR [19]): $finalTime = 10000$, $keepEvery = 1$, $sampleEvery = 0.03$, $max.wall.time = 20$, $max.num.tries = 500$.

4. Fitness landscapes: characteristics, evolutionary predictability, clonal interference, and sampled genotypes

The following figures show the main fitness landscape characteristics and the resulting variation in evolutionary predictability, clonal interference, and sampling characteristics, between types of fitness landscapes and simulation conditions (initial population size and mutation rate). Note that the "static fitness landscape" characteristics do not depend on the simulations.

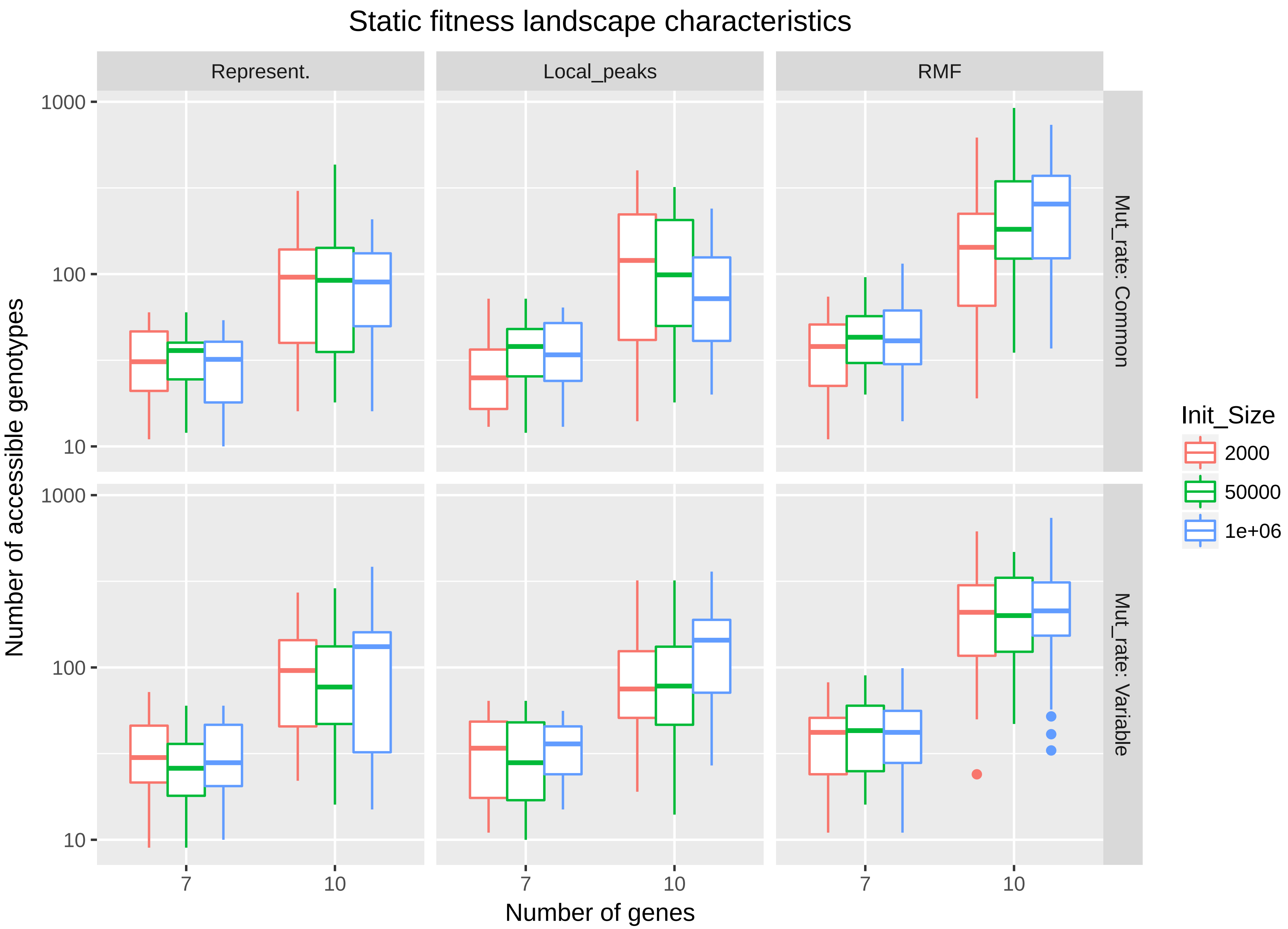


Figure S1: Simulated fitness landscapes: Number of accessible genotypes

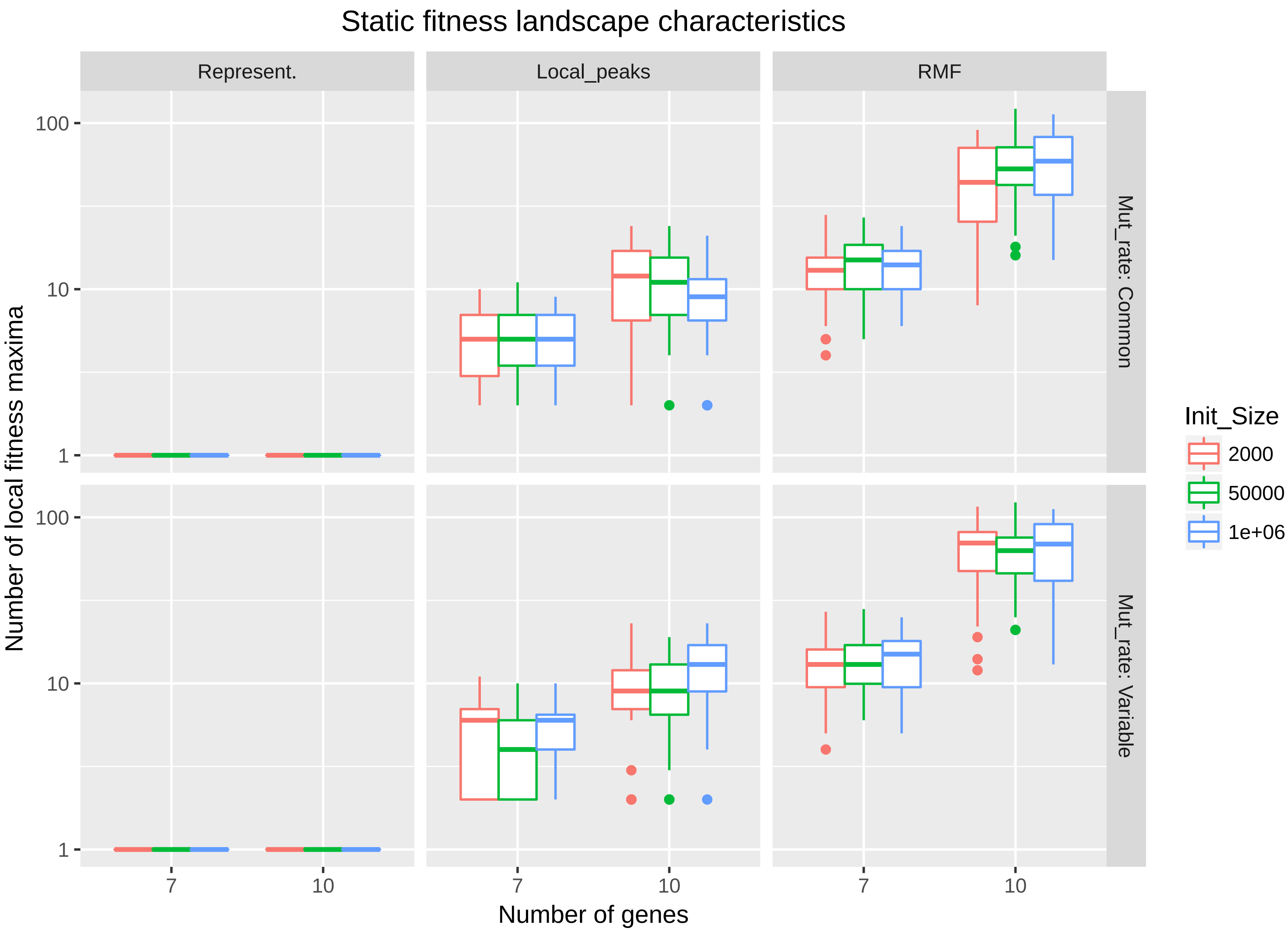


Figure S2: Simulated fitness landscapes: Number of local fitness maxima

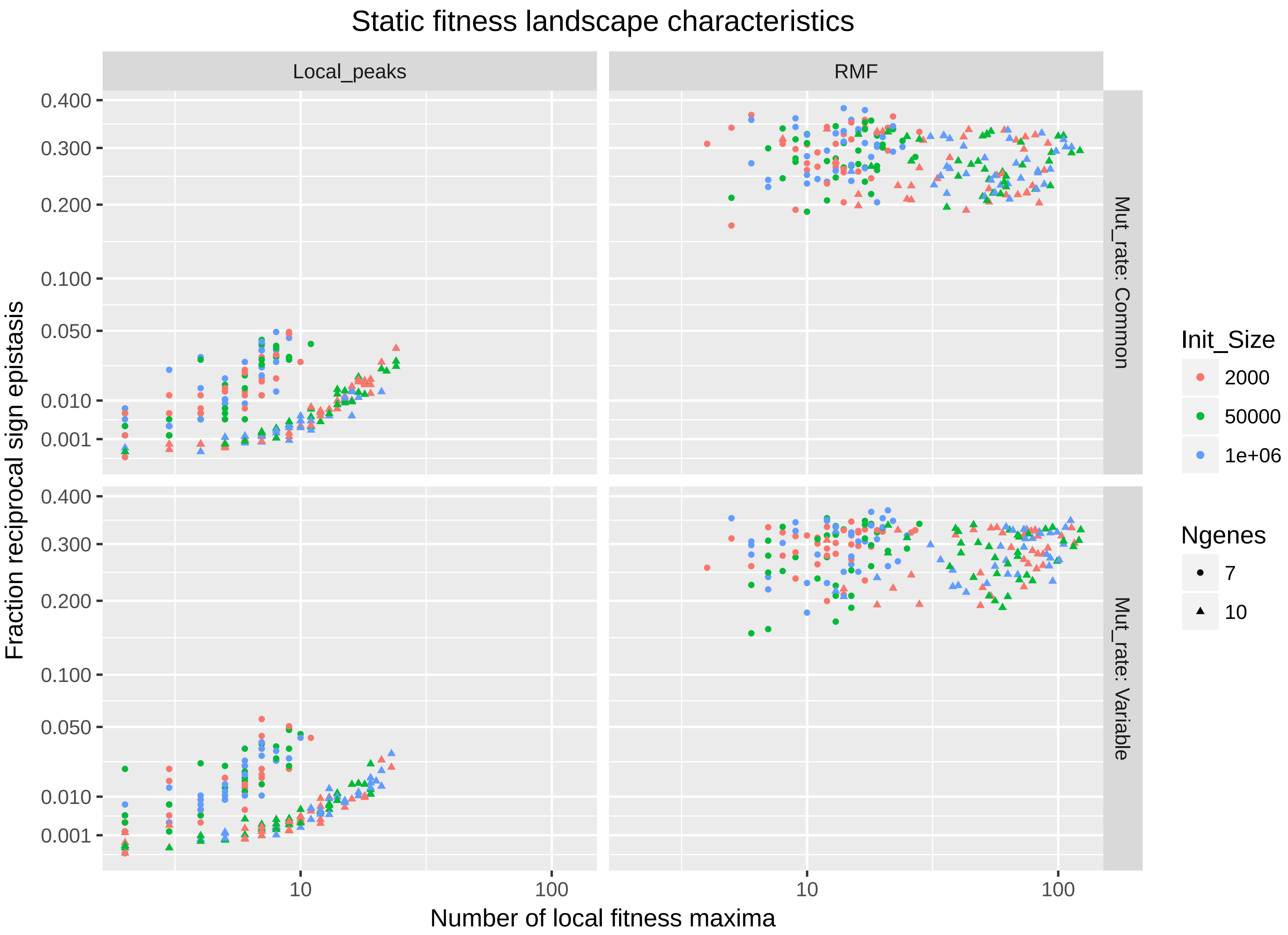


Figure S3: Simulated fitness landscapes: reciprocal sign epistasis

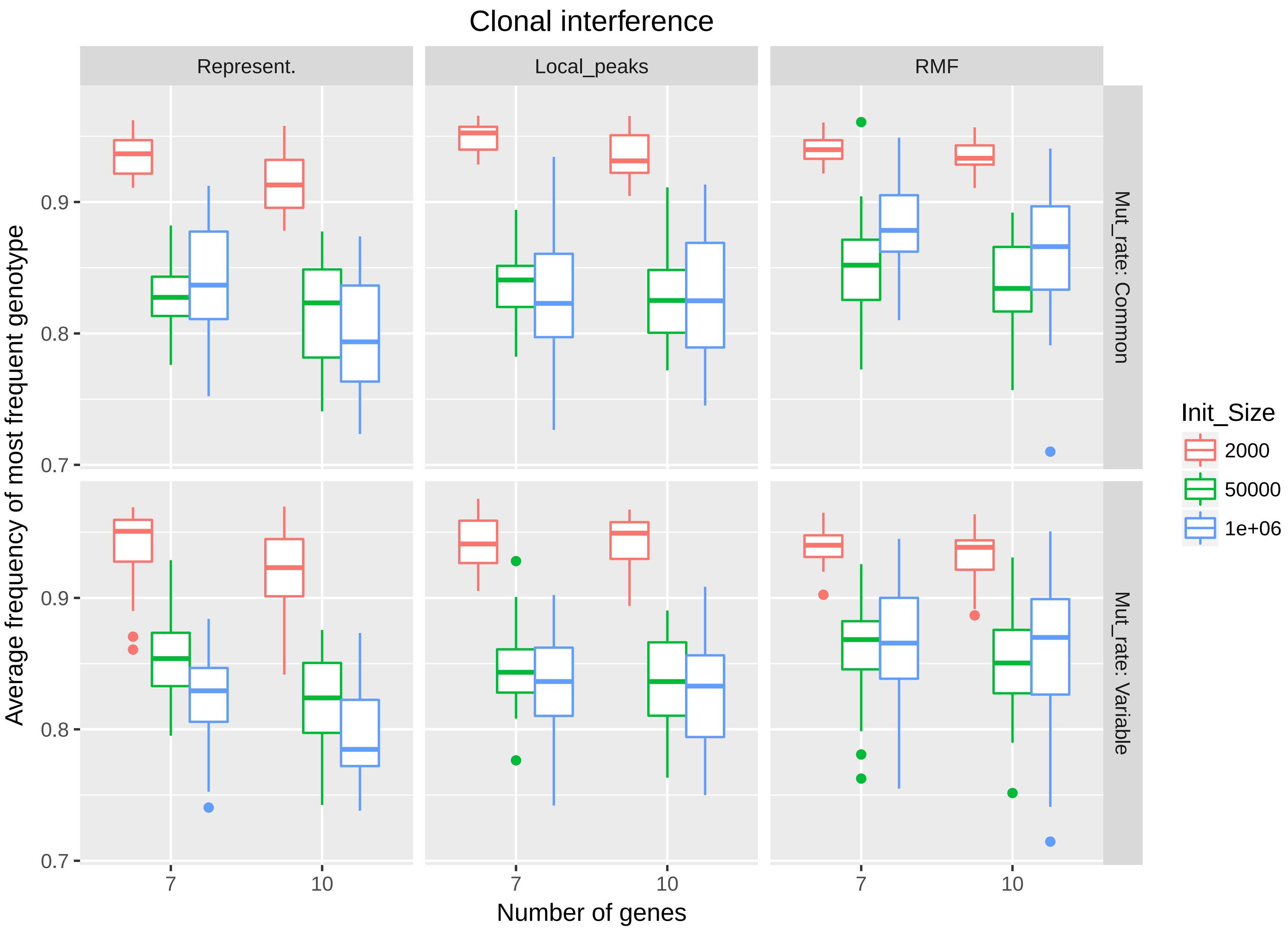


Figure S4: Simulated fitness landscapes: clonal interference (frequency of most frequent genotype)

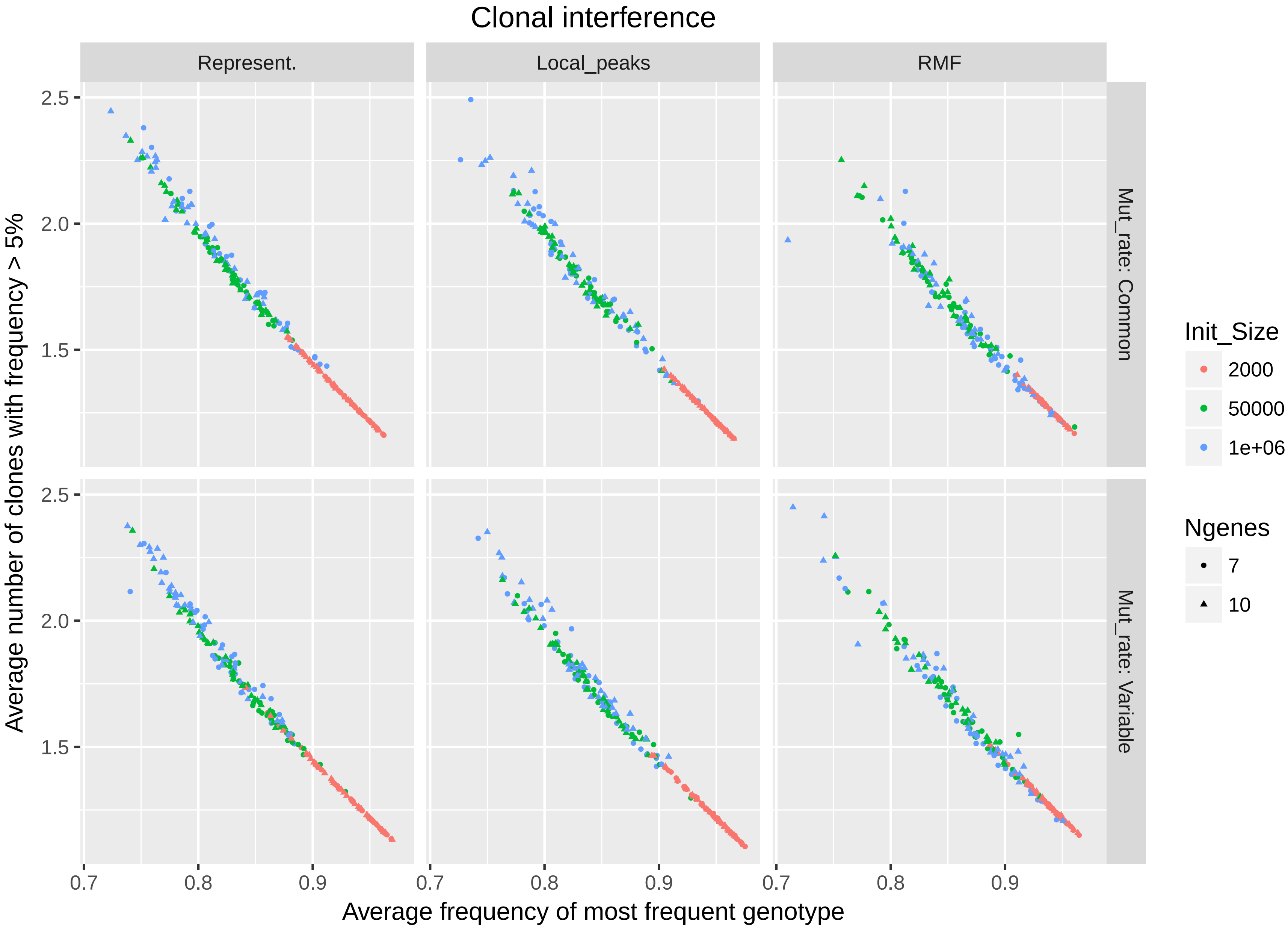


Figure S5: Simulated fitness landscapes: clonal interference (average number of clones with frequency > 5%)

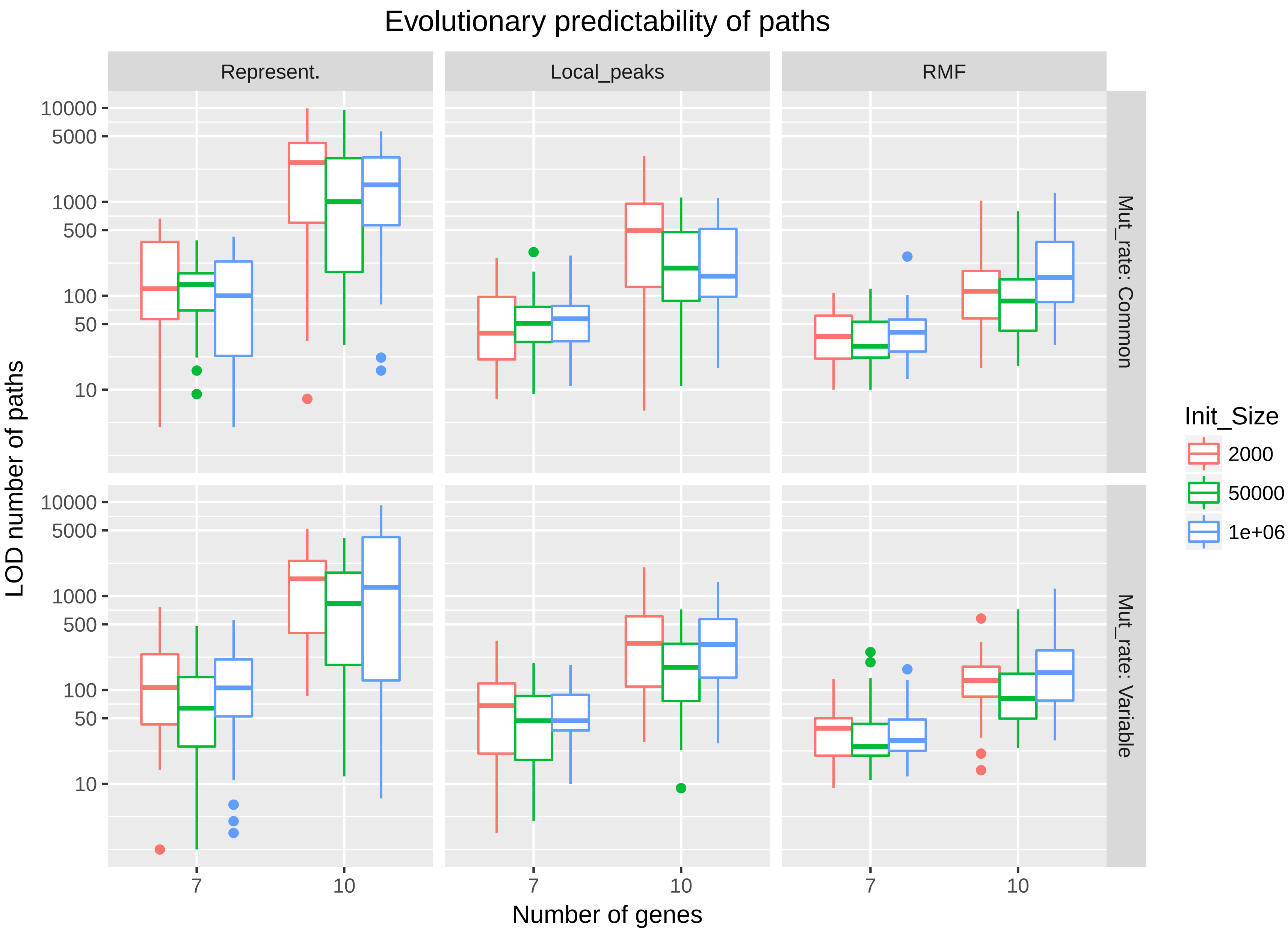


Figure S6: Simulated fitness landscapes: LOD

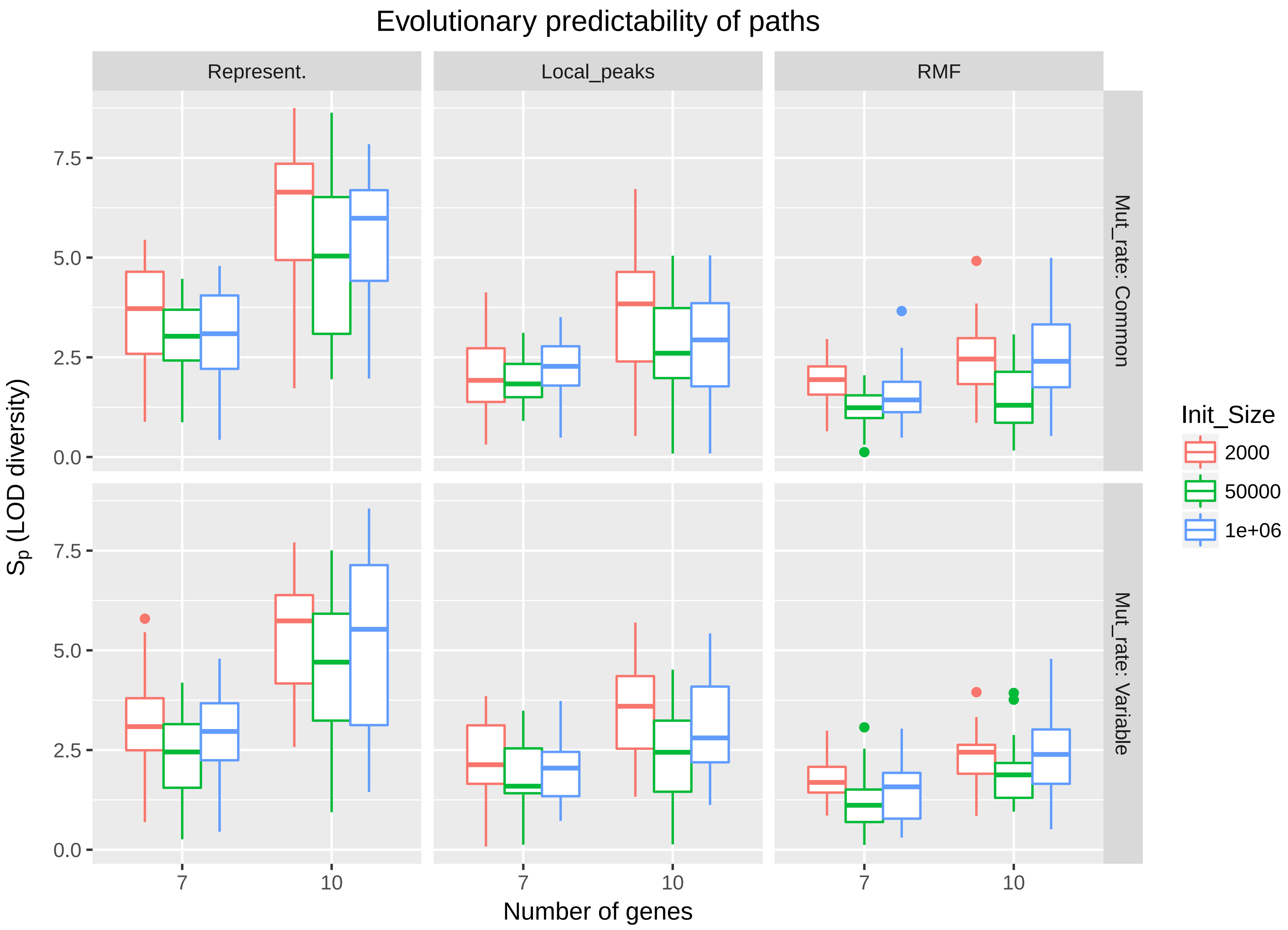


Figure S7: Simulated fitness landscapes: S_p

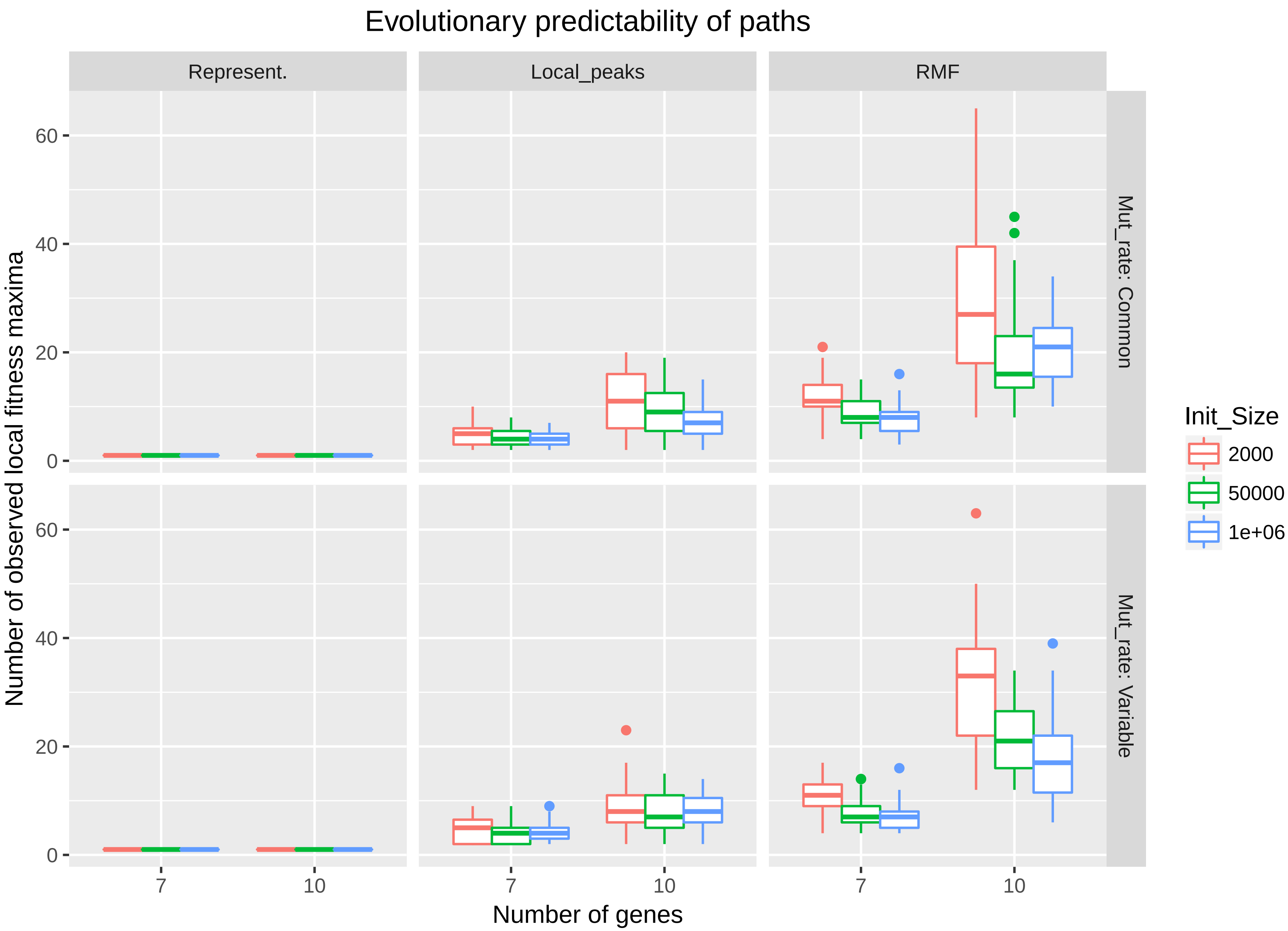


Figure S8: Simulated fitness landscapes: number of observed local fitness maxima

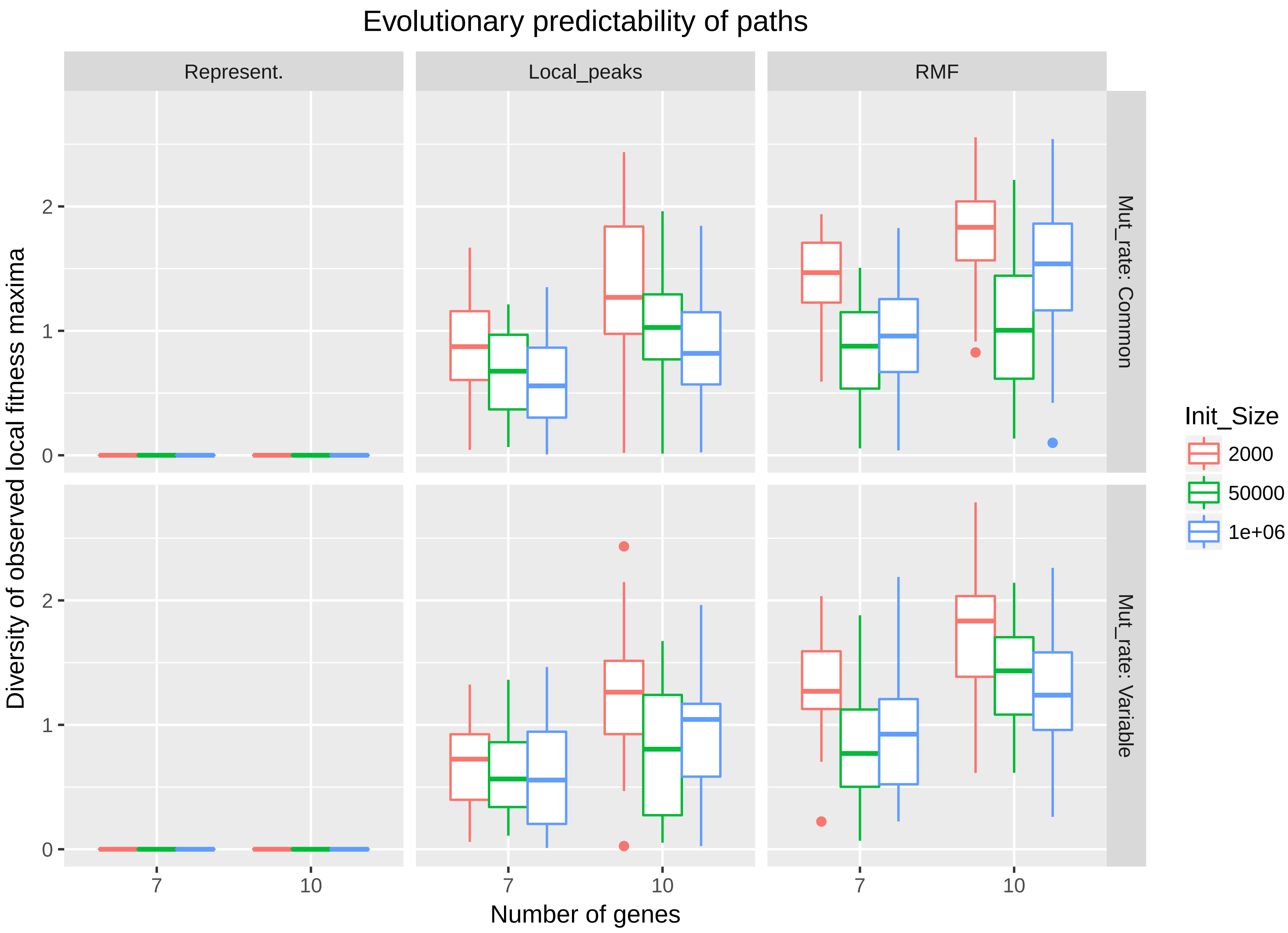


Figure S9: Simulated fitness landscapes: diversity of observed fitness maxima

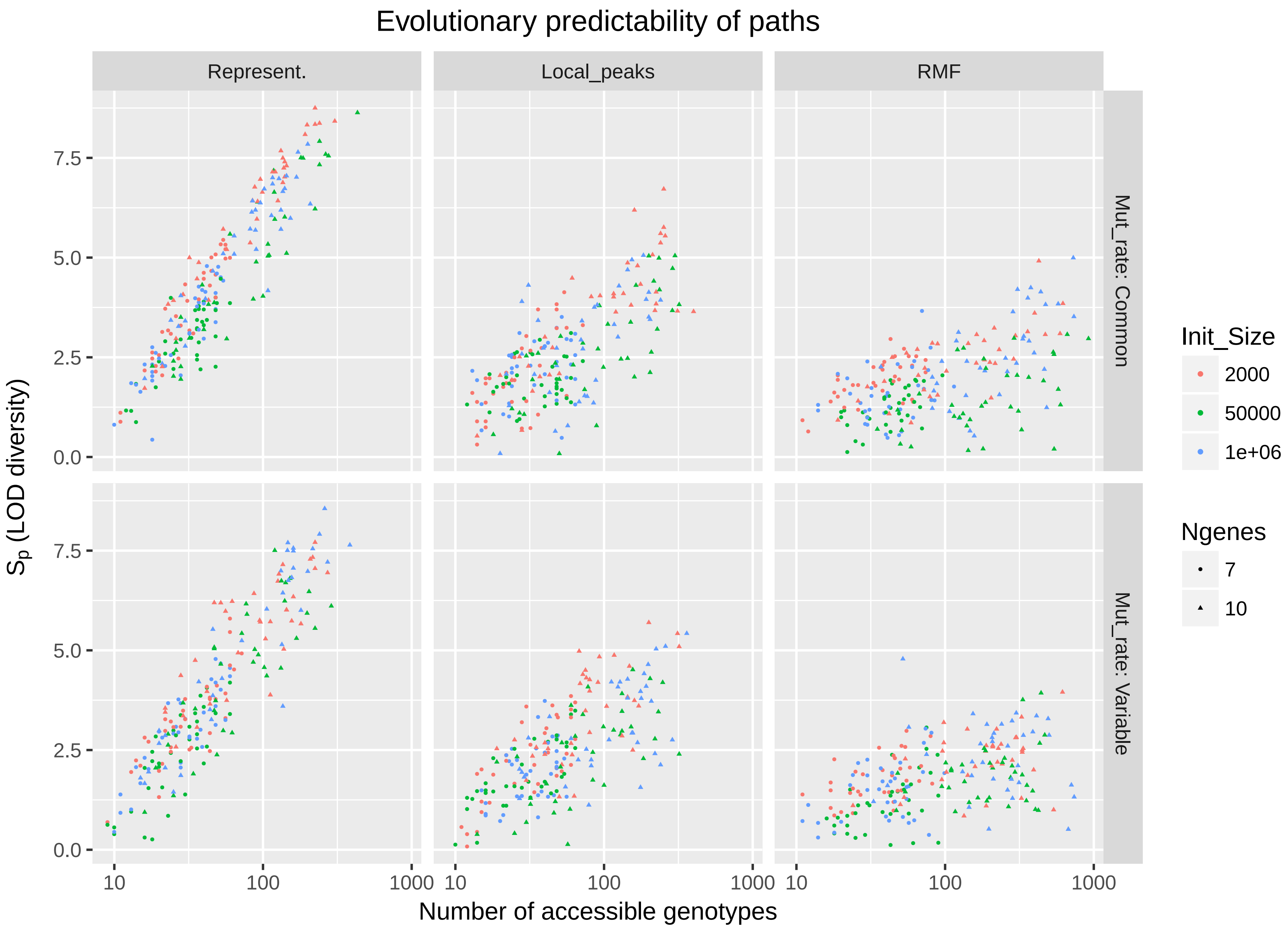


Figure S10: Simulated fitness landscapes: S_p vs. accessible genotypes

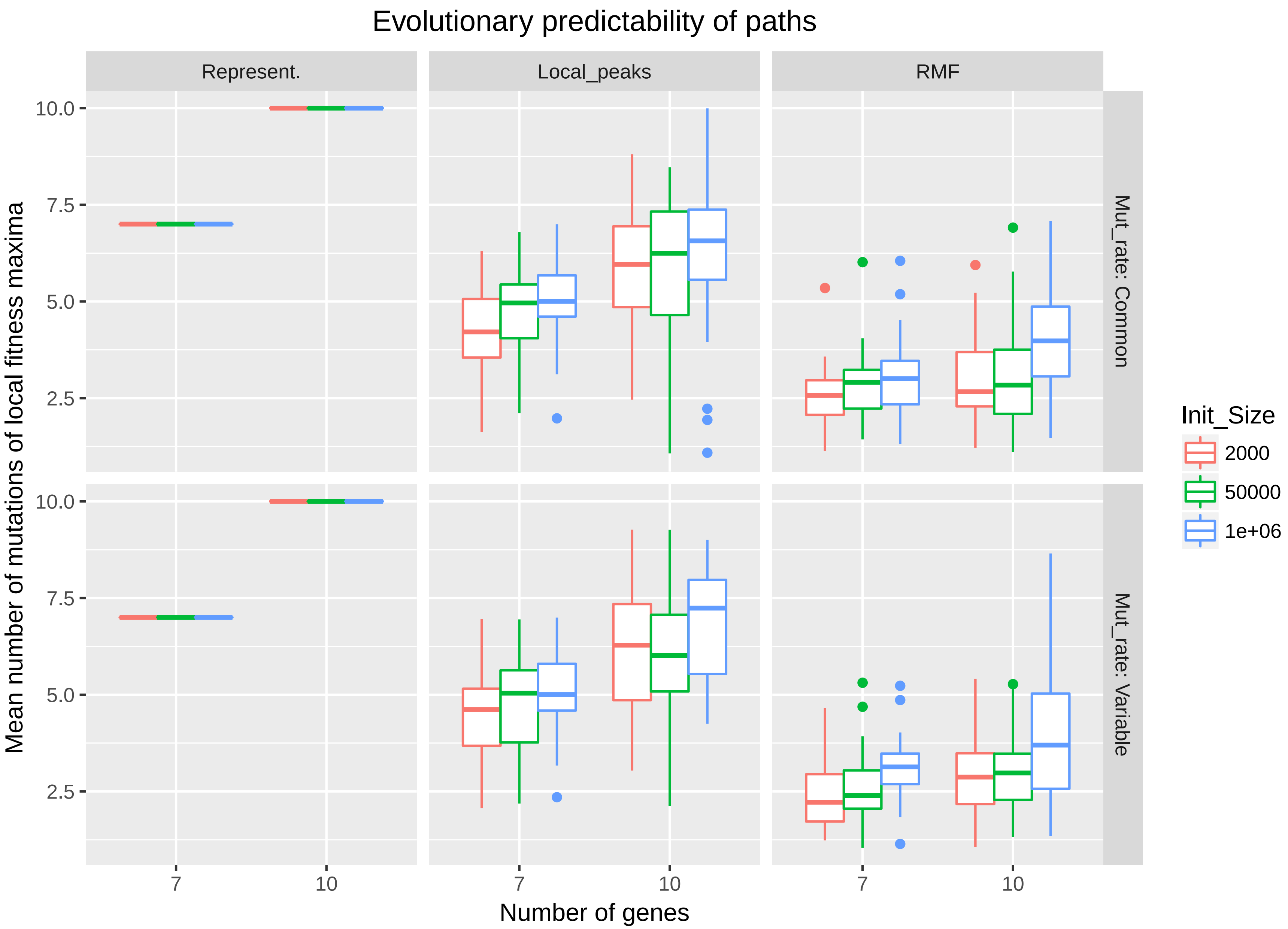


Figure S11: Simulated fitness landscapes: number of mutations of fitness maxima

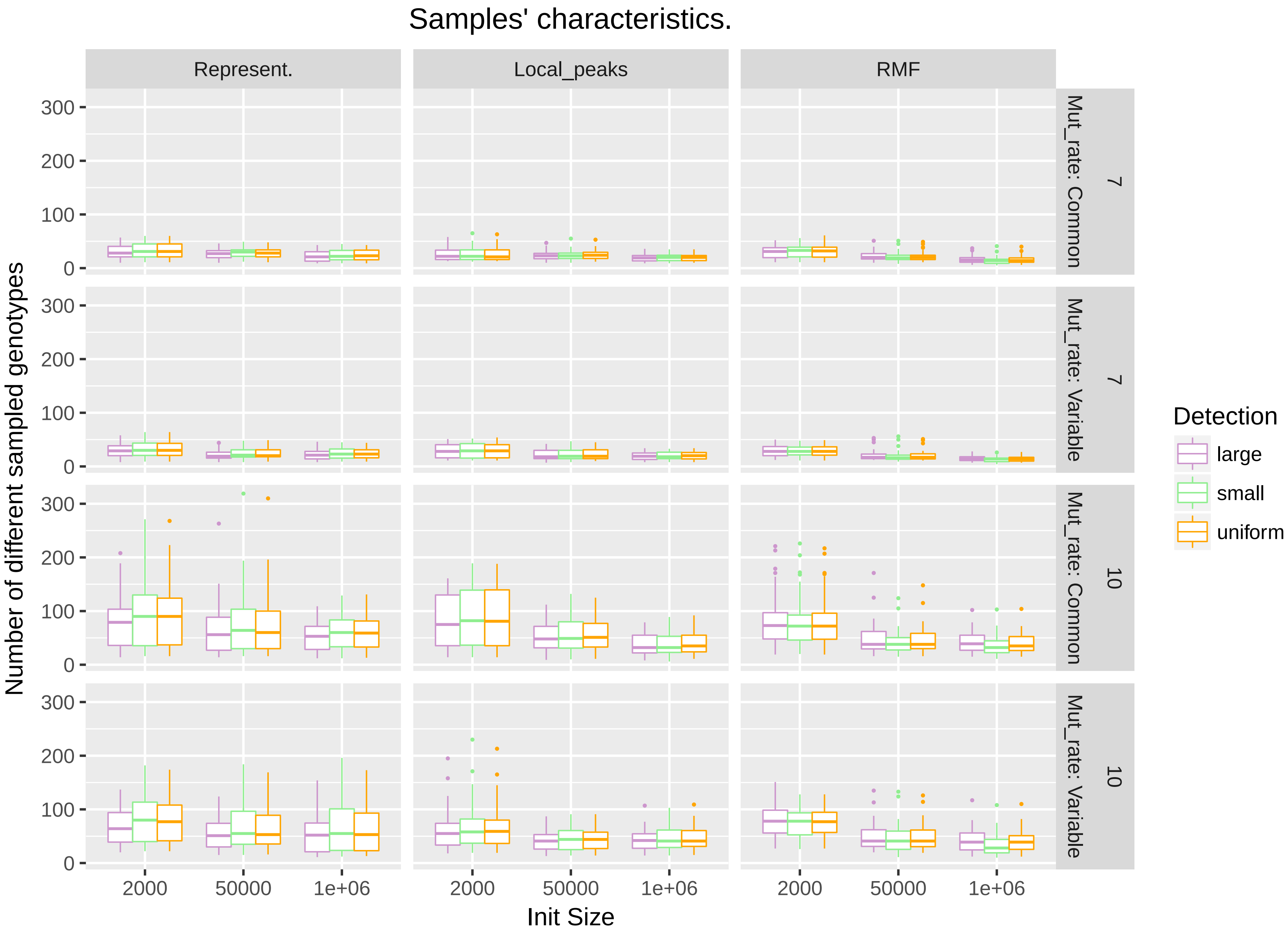


Figure S12: Simulated samples' characteristics: number of genotypes

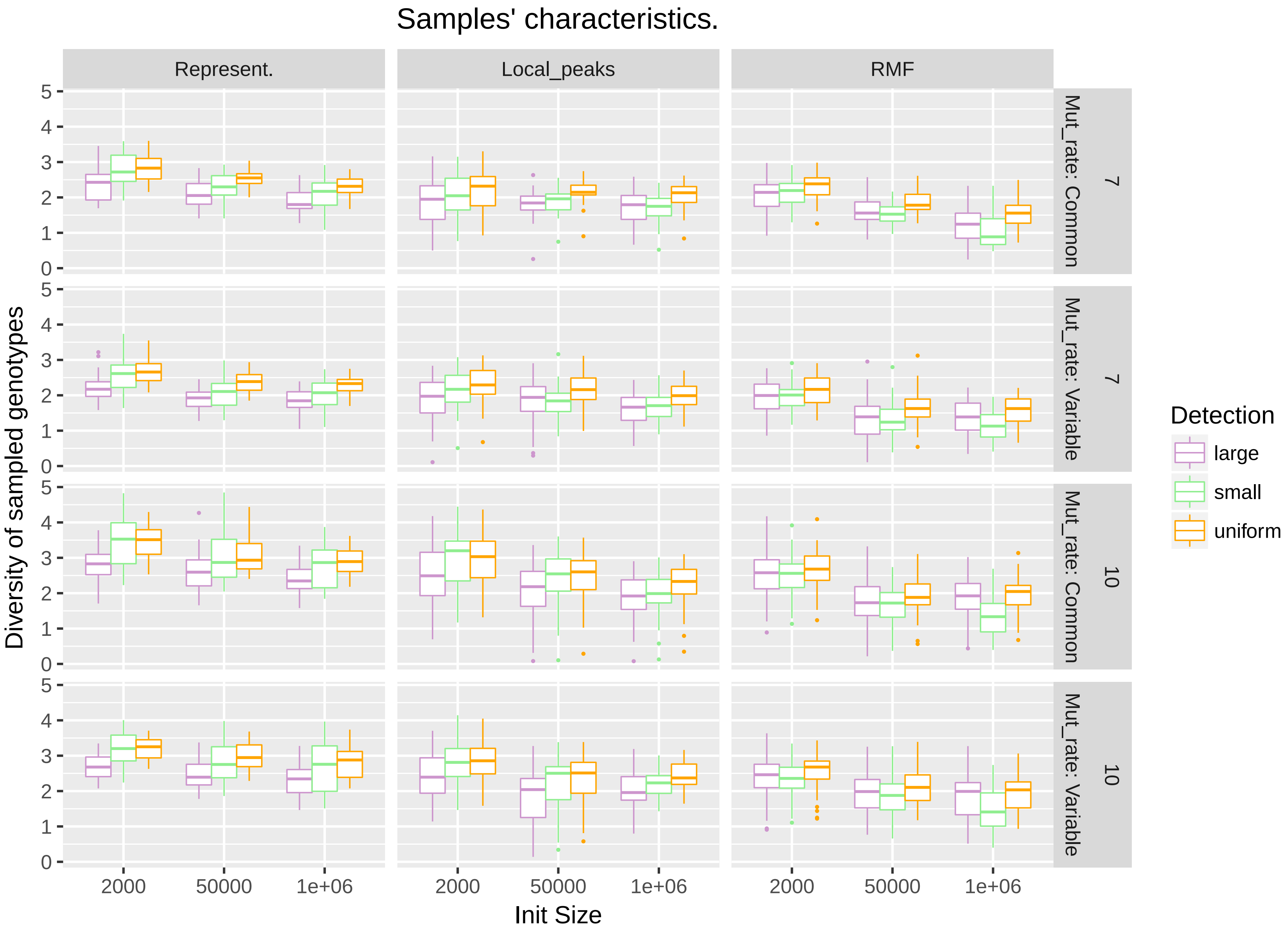


Figure S13: Simulated samples' characteristics: diversity of genotypes

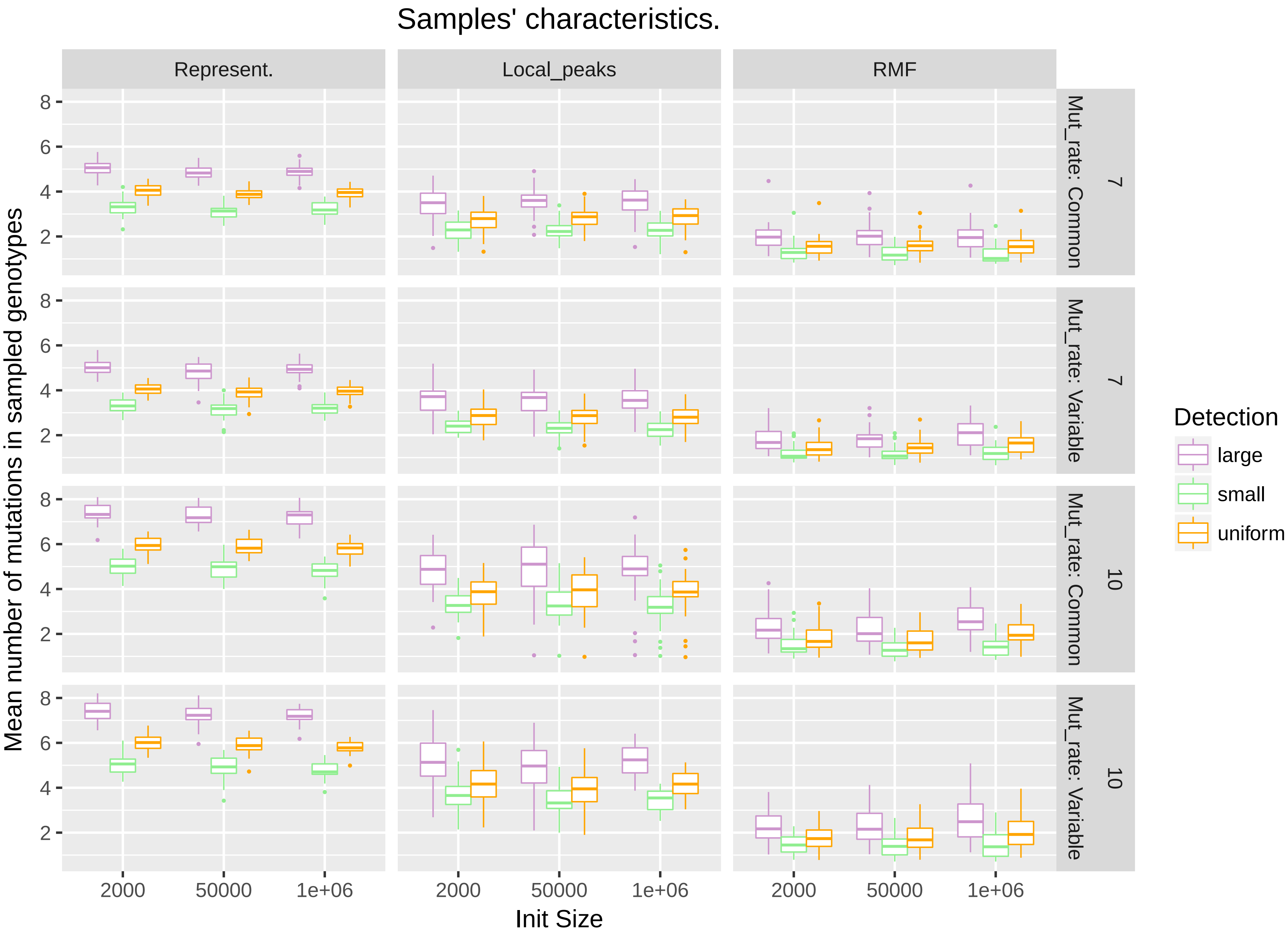


Figure S14: Simulated samples' characteristics: mean number of mutations in genotypes

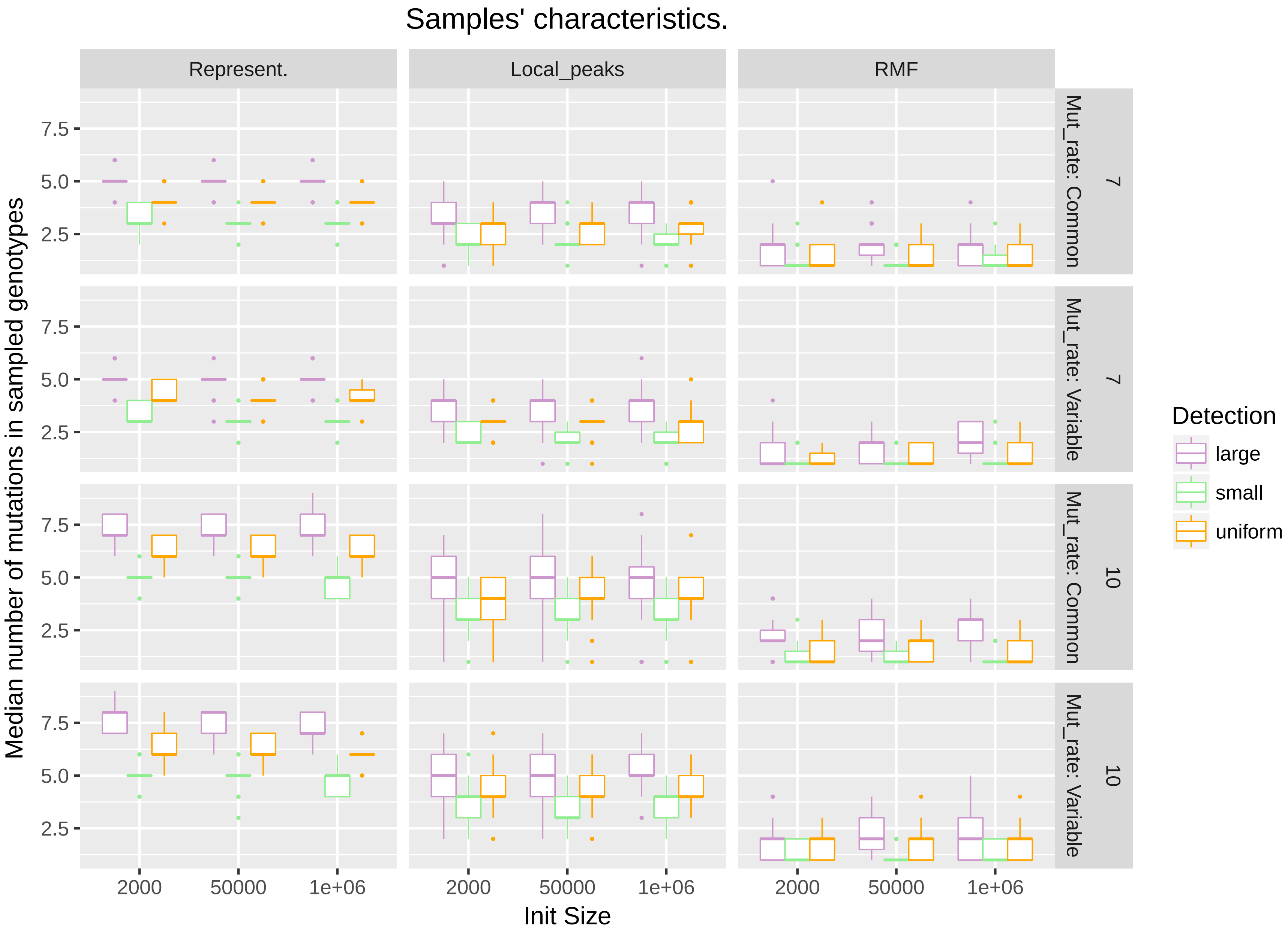


Figure S15: Simulated samples' characteristics: median number of mutations in genotypes

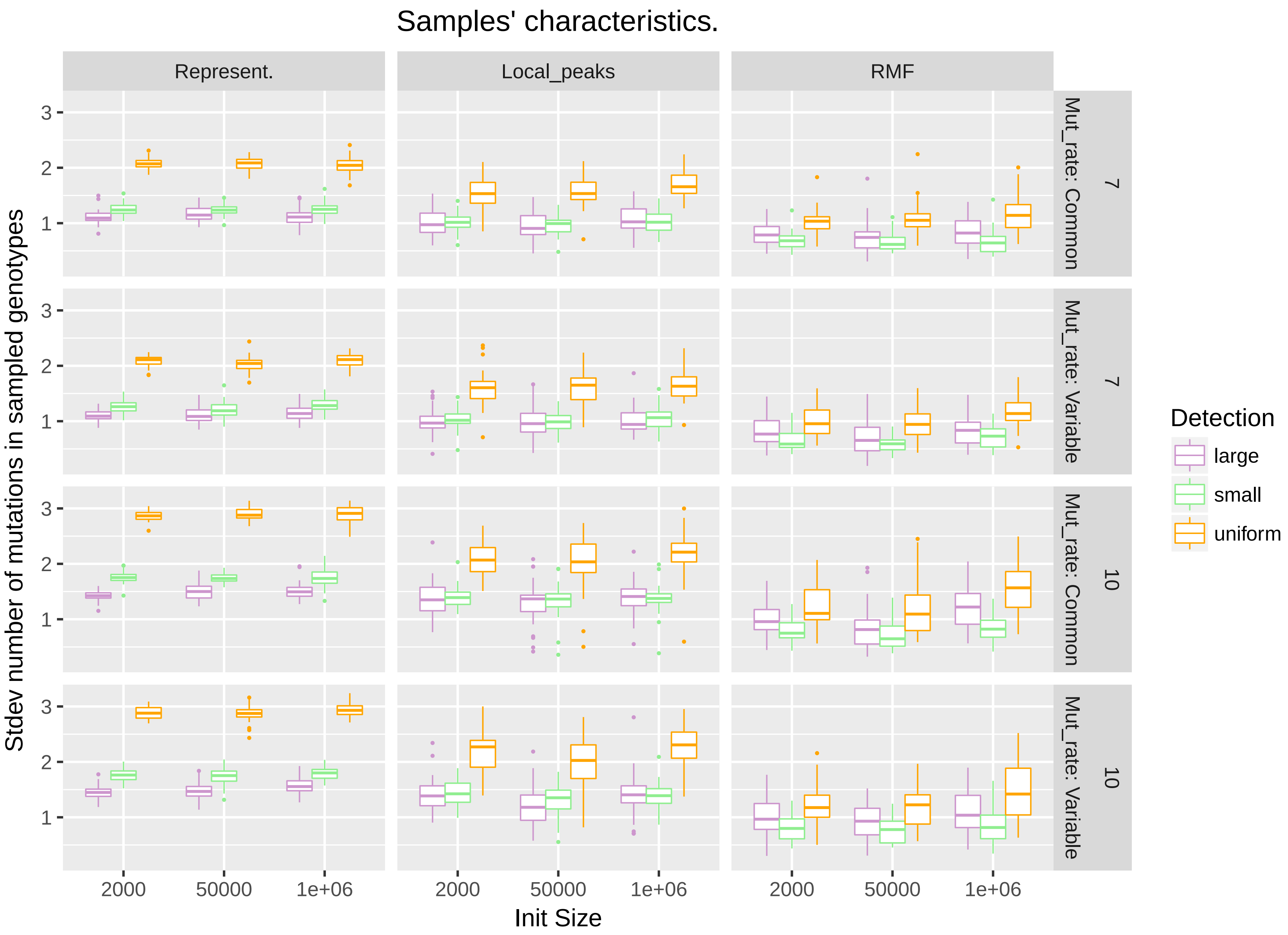


Figure S16: Simulated samples’ characteristics: standard deviation number of mutations in genotypes

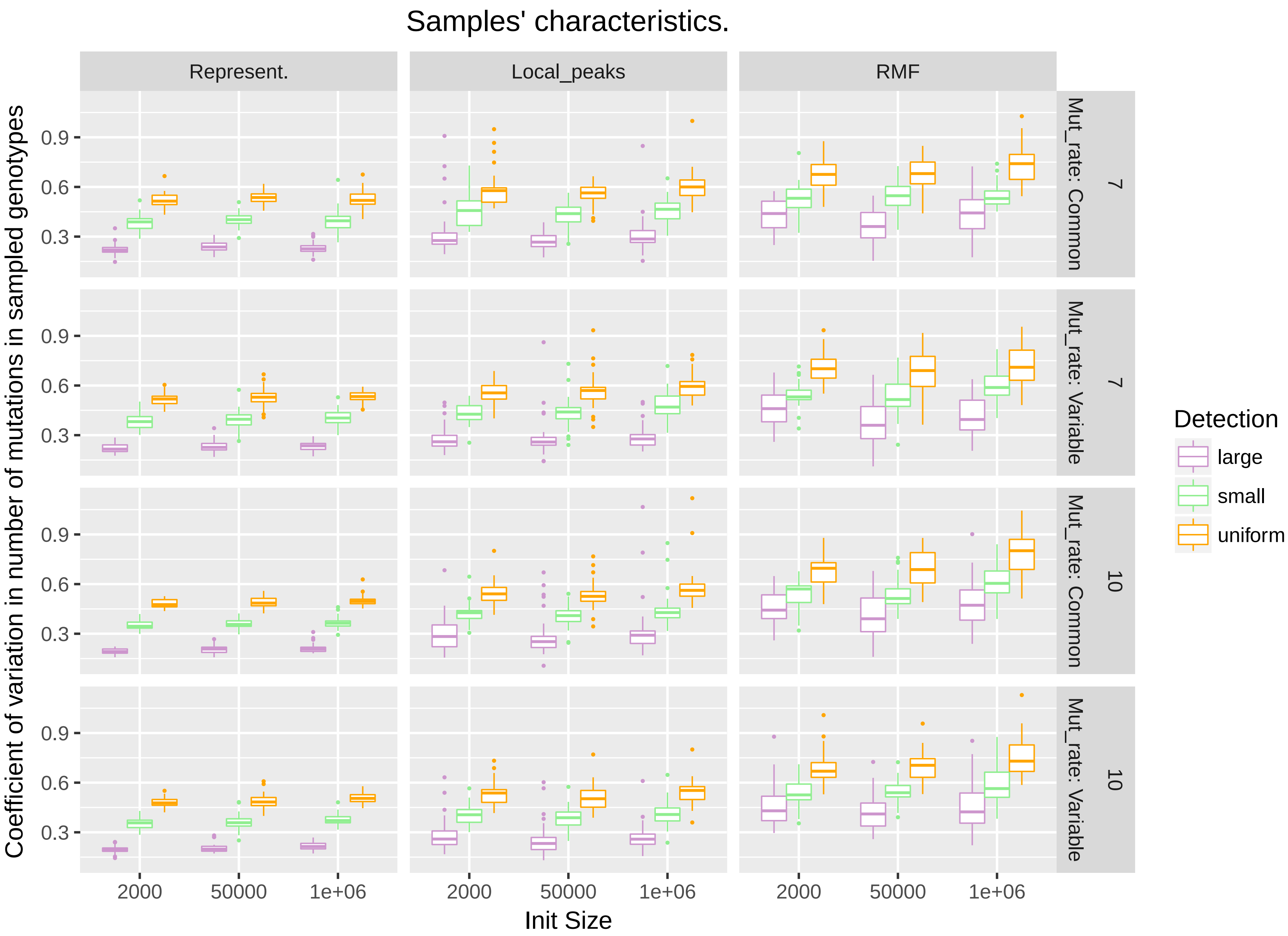


Figure S17: Simulated samples' characteristics: coefficient of variation in number of mutations in genotypes

5. Material and methods: others

5.1. Terminology

The **number of local (fitness) maxima** is a static feature of the landscape (number of genotypes such that all genotypes within a distance of one mutation have lower fitness). The number of **observed local (fitness) maxima** can be smaller, since some peaks (local maxima) might never be visited. For representable fitness landscapes, both numbers are 1. For the other two landscapes, those numbers were ≥ 2.

5.2. CPM software

Other methods for cancer progression models have been described but either are too slow for routine use such as [40], or have dependencies on external libraries that are not open source such as DiP [22], or have no software available (e.g., [3, 14]). See further details in [18].

For CBN, we used version 0.1.04b from March 2016, and still current as of April 2018, downloaded from https://www.bsse.ethz.ch/cbg/software/ct-cbn.html. We wrote a wrapper to call CBN from R, and we used the default settings for temp ($-T = 1$) and steps ($-N =$ number of nodes^2) —tough we ensure a minimum of 25 steps are used, even if number of nodes is less than five; the simulated annealing search started for the best poset from an initial poset built using Oncogenetic Trees [42], as preliminary runs suggested this initial poset is as good as, or better than, the default linear poset in [26]. The parameters for the transition rates between genotypes (λs) were obtained doing an additional run on the fitted model from the previous step, as in [26].

MCCBN was run using version 1.1.9 of the MC-CBN package, downloaded from github (https://github.com/cbg-ethz/MC-CBN).

OT was run using version 0.3.3 of the Oncotree package [42].

For CAPRI and CAPRESE we used version 2.11.0 of the TRONCO BioConductor package, current as of April 2018, downloaded from the official BioConductor site. All options were left at the recommended defaults (e.g., 100 bootstrap samples for the estimation of the selective advantage scores with p-value of 0.05, and heuristic search using Hill Climbing).

We wrote wrapper code for all methods to obtain the fitness graphs and, for OT, CBN, and MCCBN, the weighted predicted paths. For OT, we use `ot.fit$parent$est.weight` to obtain the probabilities of transition to each descendant genotype; if the OT fit, however, cannot return an error estimate, that operation fails and thus we use the `ot.fit$parent$obs.weight` component.

5.3. Preprocessing of data for CPMs

Before analyzing data with CPMs, data were preprocessed as follows:

- All columns that had all 0s (i.e., genes that were absent in all samples) were removed. Since these are never present in the data given to the methods, no inference can be made about the removed genes, and this necessarily decreases the dimension of the fitness landscape implied by the CPM and the length of the paths to the maximum.

- If two or more columns (genes) were identical over all individuals (i.e., were indistinguishable), all the identical replicate columns, except one, were removed from the analyses[1]. This, of course, will preclude the matching of some (or all) of the true paths since we are constructing the CPM from a data set of smaller dimensionality and the CPM's paths are shorter than they ought to be (see also details in section 5.6).

[1] In more detail, the identical columns were flagged as such by "fusing" the names of the genes, so as to be able to identify them. Then, the paths were post-processed before the analysis to remove the combined name, leaving one of the two, or more, identical names. For example, if we have fused B and C, a path could appear as $WT \rightarrow A \rightarrow A, B\_C \rightarrow A, B\_C, D$. We would then write as $WT \rightarrow A \rightarrow A, B \rightarrow A, B, D$. Here "C" has been dropped (it was indistinguishable from "B").

Indistinguishable events will unavoidably create problems. Alternative ways of handling them are not better. If the indistinguishable columns are left in the data, some methods (e.g., CAPRI and CAPRESE) complaint about it, whereas others (CBN, MCCBN) make the indistinguishable events depend on one another, with a very large λ, with the order in the DAG depending on the order on the column of data (leftmost events placed as ancestors). OT also places them as independent events (as CAPRI and CAPRESE do).

The consequence of leaving the events as in CBN would be similar to expanding the paths of progression, post-analysis, and placing the indistinguishable events one right after the other in the path. The order would, of course, have to be arbitrary and, in most cases, this would actually make matching any true LOD harder. If only one of the replicates is left in the path, the LOD needs to match, by chance, one particular order. If two or more are placed, the probability of matching decreases.

The proportion of data sets with one or more identical columns is about .16, .16, and .11 for the representable, local maxima, and RMF landscapes. They decrease with sample sizes, generally being under 0.05 for sample size 4000.

- Whenever one of more genes were present in all samples, to prevent the removal of these genes present in all cases, we added one case (one "pseudosample") with no mutations to the data set (this is not unlike [18], but we add only one sample, not a fixed percentage, to minimize altering any estimates of probabilities of paths). This allows us to use exactly the same data for all methods (CAPRI cannot deal with data where one or more columns are present in all subjects, OT removes them, whereas CBN can use this data).

  Even more importantly, this procedure does not decrease the dimensionality of the data set and, thus, does not decrease the length of the CPM's paths to the maximum

  The event that has a frequency of 1 is placed at the top of the DAG of restrictions (it is the first mutation after WT in all the paths to the maximum). The (very minor) inconvenience is that it has a minor effect on CBN's λs estimates, but that should be inconsequential for practical purposes.

  The proportion of data sets with pseudosamples added was .30, .21, and .01, for representable, local maxima, and RMF fitness landscapes.

After the data pre-processing, a total of 105 data sets, out of the original 56700, led to data sets in which all methods failed to fit a model (e.g., because the final data set contained only a single column). MCCBN, in addition, failed to fit a model to one additional case (to which all other methods were able to fit a model).

5.4. Computing probabilities of paths

The procedure we used is as follows (it might be simpler to understand it by referring to p. i729 of Montazeri et al., 2016 [35]):

1. Obtain the set of genotypes that can exist under the poset (as we will use it in step 3 below).

2. Obtain the set of paths that can exist under the poset (to be used in step 5, below). This itself is obtained from 1.

3. Obtain the transition rate matrix between genotypes from the lambdas (e.g., what is shown in matrix S in Montazeri et al.). As explained in Montazeri et al., "the non-zero off-diagonal elements of the transition matrix are the transition rates from each genotype to its successive genotypes in the genotype lattice, also shown in Figure 1(b)." See also legend of Figure 1: "(b). Directed transition rates among neighboring genotypes are shown on the edges of the lattice".

4. Set the diagonal of the previous matrix to 0[2] and for each row of the transition rate matrix, divide by $\sum \lambda$. Now the entries are probabilities of transition to each descendant genotype given a transition.

5. Go over the list of paths (from step 2) and for each path, obtain its probability by multiplying the probabilties of the transitions (from step 4) between the genotypes in a path.

6. (Check: verify sum of probabilities of all paths equals 1, within numerical margin of error of machine.)

In the code, steps 3 and 4 above were carried out by creating, from the set of paths to the maximum and the output from CBN, what we called a weighted fitness graph: the fitness graph of paths to the maximum with the λs on the edges (or the weighted adjacency matrix corresponding to paths to the maximum where weights are λs). This would be Figure 1b in [35]. Dividing by the row sum gives us the transition matrix between genotypes in step 4.

As an example, suppose we obtain from CBN the following DAG of restrictions and estimated lambdas:

From	To	λ
Root	A	2
Root	B	3
A	C	4
C	D	5

The paths to the maximum, with their probabilities, are:

path	probability
WT → A → A, B → A, B, C → A, B, C, D	2/5 * 3/7 * 1 * 1
WT → A → A, C → A, B, C → A, B, C, D	2/5 * 4/7 * 3/8 * 1
WT → A → A, C → A, C, D → A, B, C, D	2/5 * 4/7 * 5/8 * 1
WT → B → A, B → A, B, C → A, B, C, D	3/5 * 1 * 1 * 1

For example, from WT with a probability of 2/5 we take the path to A and with 3/5 to B. Once in genotype A, we can either add a B mutation (probability = 3/7) or a C mutation (probability = 4/7). If we add a B mutation, from genotype AB we can only move to ABC (probability 1). Etc. This procedure is equivalent to the one used by Hosseini [27].

An analogous procedure was used with OT.

5.5. Example where perfect recall and precision do not guarantee Jensen-Shannon divergence of 0

Suppose the following set of two paths to the maximum, with predicted and observed probabilities as shown:

Path	CPM predicted probability	True LOD probability
WT → A → AB	0.99	0.01
WT → B → AB	0.01	0.99

The JS divergence (on a scale 0 to 1) is 0.9192 (remember 1 is the maximum value of divergence), even when 1-recall and 1-precision are both 0 (none of the CPM's paths are missing

[2] In fact, the diagonal entries are never computed explicitly and are always 0.

from the LODs, and none of the LODs are missing from the CPM's paths).

5.6. Measuring predictability: comparing paths from CPMs and LODs of different lengths

Let i and j denote two paths, one from the LOD and the other from the CPM, with corresponding probabilities p_i and q_j. Here, the i index refers to paths from the LOD, and the j to paths from the CPM (this is in contrast to the paper, where we did not make this specification, to keep the description general).

Let K_i, K_j denote the length of paths i and j, respectively. Note that all K_j are equal (as all go up to the genotype with all genes mutated), and we will refer to that unique value of K_j as K^C. When there is ambiguity about which K we are referring to, we will use K_j^C for the K_j from the CPM (again, $K_1^C = K_2^C = \ldots = K_C$) and K_i^L for the K_i from the LOD.

The vectors P, Q, for the computation of JS will have the following types of matching pairs:

1. p_i, q_j when $K_i = K_j$
2. $p_i \frac{K_j}{K_i}, q_j$, when i is partially included in j ($K_i > K_j$), and this accounts for the part of i included in j,
3. $p_i, q_j \frac{K_i}{K_j}$ when j is partially included in i ($K_j > K_i$), and this accounts for the part of j included in i,
4. $\sum p_i \frac{K_i - K_j}{K_i}, 0$ when i is partially included in j ($K_i > K_j$), and this accounts for the part of i not included in j,
5. $0, \sum q_j \frac{K_j - K_i}{K_j}$ when j with is partially included in i ($K_j > K_i$), and this accounts for the part of j not included in i,
6. $\sum p_u, 0$, for all paths u among the paths from the LOD that do not match any j (any path from the CPM),
7. $0, \sum q_v$, for all paths v among the paths from the CPM that do not match any i (any path from the LOD).

(Where some notation above, again, can be simplified by noting that all $K_j = K_C$).

We can sum, as appropriate, the relevant entries to simplify computations as the JS is the same for the pair of vectors $P = [p_1, p_2, 0, 0, p_5, p_6]$, $Q = [q_1, q_2, q_3, q_4, 0, 0]$ and the pair of vectors
$P' = [p_1, p_2, 0, p_5 + p_6]$, $Q' = [q_1, q_2, q_3 + q_4, 0]$ (this follows from the definition of JS).

In other words, we have the 0 entry in Q correspond to $\sum p_i \frac{K_i - K_j}{K_i} + \sum p_u$, where the i are the paths in the LOD with some partial match among the paths in the CPM, and the u denote those paths in the LOD that do not match any path in the CPM.

The relevant entries above can be used to compute 1-recall and 1-precision. For example, for 1-recall, $P(\neg DAG|LOD)$, the sum of the probabilities of the paths in the LODs that are not among the paths allowed by the CPMs, we will use $\sum p_i \frac{K_i - K_j}{K_i} + \sum p_u$.

5.6.1. Commented example for paths of unequal length

To give a specific example, suppose the following paths from a LOD and a CPM. In this example, the CPM has been constructed from a data set that had only mutations A and B (C and D were missing). And we have the three possible cases: some paths from the LOD are shorter than the paths from the CPM, some paths from the LOD are the same length as those from the CPM, and some paths from the LOD are larger than those from the CPM.

Path (i)	LOD frequency	p_i	K_i
1	WT → A	0.1	1
2	WT → B → AB	0.3	2
3	WT → B → AB → ABC	0.1	3
4	WT → B → BC → ABC	0.2	3
5	WT → B → BC → ABC → ABCD	0.3	4

Path (j)	CPM predicted probability	q_j	$K_j (= K^C)$
1	WT → A → AB	0.6	2
2	WT → B → AB	0.4	2

Where $i = 1, 2, 3, 4, 5$ are the four LODs and $j = 1, 2$ the two paths to the maximum from the CPM. $K_j = K^C = 2$. To compute JS, 1-recall and 1-precision, it is much simpler to use an algorithm that splits the cases to be considered into three:

1. $K_i < K_j$ (i.e., $K_i < K^C$), those paths from the LOD that are shorter than the paths from the CPM;

2. $K_i = K_j$ (i.e., $K_i = K^C$), those paths from the LOD that have the same length as the paths from the CPM;

3. $K_i > K_j$ (i.e., $K_i > K_C$), those paths from the LOD that are larger than the paths from the CPM;

We can iterate over all distinct k for $K_i < K_j$, and weight the output by w_k, the sum of all paths from the LOD that end at k mutations; computations for $K_i = K_j$ can be subsumed into those for $K_i < K_j$. Thus, one part of the algorithm iterates over all $k \leq K_j = K_C$ (remember all K_j are identical and equal to K_C, the single, unique K at which the paths from the CPM stop). Computations for $K_i > K_j$ can be done in one iteration.

It might also be helpful to think about a cut operation on a path. For instance, for each k where $K_i < K_j$, we can cut the paths from the CPM at k mutations, leaving only the paths from WT up to k mutations (and collapsing, as appropriate, any collection of now indistinguishable subsets of paths, summing their probabilities).

In the exposition below, some computations could be further simplified; they are left as they are for clarity (e.g., we multiply by total frequencies of mutations for the weights when we have previously scaled the total probability by it, so it is 1 in each k, etc).

JS

1. LOD $i = 1$ finishes at one mutated gene. Cutting the CPM path at $k = 1$, the JS for $k = 1$ is obtained from the vectors of probabilities $P = \left[1, 0, 0\right]$ (from the LOD) and $Q = \left[0.6\,0.5, 0.4\,0.5, 0.5\right]$ from the CPM. The last entry in Q is the sum of all the flow through the paths of the CPM that cannot be matched because the length of the CPM paths is $K_C = 2$. And the 1 in P comes from $p_1 / \sum p_i$ for all i that end in $k = 1$ which is only p_1.

The first and second entries are q_1^1 and q_2^1 multiplied by k/K_C, i.e., the probabilities of the fractions of paths from the CPM cut at $k = 1$ mutations.

The weight, w_k, for this value is 0.1 (the frequency of LODs that finish at $k = 1$).

2. LOD $i = 2$ finishes at $k = 2$. Here the comparison is the immediate one for equal length paths and the vectors used for JS are: $P = \left[1, 0\right]$ and $Q = \left[0.4, 0.6\right]$. $w_2 = 0.3$.

3. Paths $i = 3, 4, 5$ are longer than the CPM paths. The flow $AB \rightarrow ABC$ for $i = 3$, the flow $BC \rightarrow ABC$ for $i = 4$, etc, cannot be matched by the CPM.

   Here the total amount of evolutionary flow through the LOD that cannot be captured by the CPM, because the CPM ends prematurely, is $\sum_i p_i(K_i - K^C)/K_i = (0.1 * (1/3) + 0.2 * (1/3) + 0.3 * (1/2)) * (1/0.6) = 0.25/0.6$, for $i = 3, 4, 5$, where the $1/0.6$ scales relative to the total probability in paths $i = 3, 4, 5$. Then, to compute JS, the two vectors of probabilities would be $P = \left[0, (2/3)\ (0.1/0.6), (2/3)\ (0.2/0.6), (1/2)\ (0.3/0.6), 0.25/0.6\right]$, from the LOD, and $Q = \left[0.6, 0.4, 0, 0, 0\right]$, for the CPM.

   But that is equivalent to using the two vectors $P = \left[0, (2/3)\ (0.1/0.6), (0.5/0.6)\ +\ (1/3)\ (0.1/0.6)\right]$, $Q = \left[0.6, 0.4, 0\right]$. The last entry in P might be easier to see from adding LODs $i = 4, 5$ and the unmatched portion of $i = 3$. Here $w_i = 0.6$

4. We can now add all the JS with their corresponding weights.

1-recall For 1-recall, the sum of the probabilities of the paths in the LODs that are not among the paths allowed by the CPMs, we have: $p_5 + p_4 + p_3\ \frac{(K_3 - K_C)}{K_3} = 0.3 + 0.2 + (1/3)\ 0.1$, where $K_C = 2$ (all CPM paths end at $k = 2$). Note that these same values can be obtained by iterating over k as above, and doing a weighted sum, but in this example it is much simpler to use the computation directly. Had we used weighted sums, we would have got: $0\ w_1 + 0\ w_2 + (p_5 + p_4 + p_3\ \frac{K_3 - K_C}{K_3})\ \frac{1}{0.6}\ w_{\geq 3} = 0.8889\ 0.6 = 0.533 = 0.3 + 0.2 + (1/3)\ 0.1$ (where the $\frac{1}{0.6}$ scales so that the probabilities considered when $k \geq 3$ add to 1 —and, sure, we are dividing by 0.6 only to multiply by it because the scaling factor is the weight of the $k_{\geq 3}$ stratum).

1-precision For 1-precision, the sum of the probabilities of the paths in the CPM's paths that are not among the LOD paths (the paths followed by evolution), it is simpler to use a weighted sum:

$((q_1\ \frac{K^C - K_1^L}{K^C}) + q_2)\ w_1 + q_1\ w_2 + q_1\ w_{\geq 3}$.

When $k = 1$, $i = 1$ is $WT \rightarrow A$, and thus all of q_2 is not captured, and half of q_1 is captured; at $k = 2$, all of q_2 but none of q_1 is captured; for $k \geq 3$ again path $j = 1$, with $q_1 = 0.6$ is not captured. Thus, we have $(0.6 * (1/2) + 0.4) * 0.1 + 0.6 * 0.3 + 0.6 * 0.6$.

Note that for 1-precision we do not need to re-scale so that probabilities always add up to 1 because they already do (we consider all the j). This was not the case for 1-recall (where only some of the i might be considered in turn when we iterate over k).

To recap, as stated in the paper, we want to compute JS, 1-recall, and 1-precision taking into account that:

1. Any LOD that finishes at k mutations and matches a CPM path up to k mutations is a perfect match, up to k; this is the reason we match each LOD with the fitness graph from CPMs cut at the number of mutations of the final genotype of the LOD.

2. Any set of LODs that finishes at k mutations when the CPM goes to K with $K > k$ necessarily misses $(K - k)/K$ of the total evolutionary flow, all that which goes from $k + 1$ to K. This is why we use a category unmatchable by construction.

Without this, it would be possible to obtain perfect JS from LODs that missed most of the evolutionary flow, for instance very short LODs that finished at one mutation.

This part of the procedure, thus, accounts for that part of the evolutionary process that the CPM predicts and is not matched by stopping evolution at a local fitness maximum; remember, again, that by construction the CPMs predict that the evolutionary process should go all the way to a global maximum with all genes mutated.

3. A similar reasoning applies when the paths from the CPM are shorter than the paths from the LOD.

5.7. Coefficients of linear models

Coefficients from the generalized linear mixed-effects models shown in the manuscript (Figure 4) are from overparameterized models, and those are not the models fitted. What we have done is fit the models several times, always with sum-to-zero contrasts, but changing the level of the factor set to $-\Sigma$ (rest of levels) so as to explicitly obtain the coefficients and standard errors for all levels of all terms (e.g., the coefficients that correspond to “Detection, Uniform”, “Detection, Small” and “Detection, Large”).

6. Cancer data sets

The cancer data sets used here are a representative example of data sets to which researchers have applied CPMs or data sets to which researchers might want to apply CPMs. All of these data sets (in at least one of their variants) have been used previously in studies with CPMs except for the BRCA data sets, that were obtained *de novo* for this paper.

These data sets vary in:

- sample size (27 to 594 samples);
- data types (nonsynomymous somatic mutations, and copy number aberrations, or both);
- levels of analysis: altered/non-altered pathways —e.g., Pan_pa, GBM_pa, Col_pa, all _pa—, functional modules —GBM_mo–, exclusivity groups [11] —Col_msi_co, Col_mss_co, ACML_co—, genes —e.g., BRCA_ba_s, BRCA_he_s, Pan_ge, GBM_ge, Col_ge, Ov, Lu—, and different types of gene-level events as insertion/deletions, missense point mutations, nonsense point mutations —ACML and ACML_co.
- different procedures for driver selection, from simple frequency-based selection of features (e.g., GBM_ge, Pan_ge, Col_ge) to state-of-the-art methods for the identification of significantly altered genes [43] (e.g., BRCA_he_ s, BRCA_ba_s, Col_msi, Col_mss, GBM_CNA);
- restriction of patient subtypes (with the purpose of achieving sample homogeneity —e.g., BRCA_ba_s, BRCA_he_s, Col_msi, Col_mss);

Thus, in several cases the same source data set has been processed in different ways to produce two different versions. For three of the data sets, two versions, one coded in terms of mutations of genes and one in terms of pathway alterations, were available (Col_ge and Col_pa, GBM_ge and GBM_pa, Pan_ge and Pan_pa). For three other data sets, we have analyzed both the original data (Col_msi, Col_mss, ACML), and the same data after accounting for so-called "exclusivity relations" (see 11; Col_msi_co, Col_mss_co, ACML_co). Another data set, GBM_CNA, was also analyzed in terms of "functional modules" (GBM_mo).

Other data sets have been obtained from a single source and split to increase subject homogeneity (e.g., BRCA_ba_s and BRCA_he_s; Col_mis, Col_mss).

6.1. Cancer data sets: sources and characteristics

Name	Source	Original source	Number of features	Number of subjects	Type of event	Abbreviation
All Pathways	[26]	(From sources for colon, glioblastoma, and pancreas genes data sets)	12	268	candidate mut	all_pa
Colon Genes	[26]	[44]	8	95	candidate mut	Col_ge
Colon Pathways	[26]	[44]	10	95	candidate mut	Col_pa
Glioblastoma Genes	[26]	[37]	8	78	candidate mut	GBM_ge
Glioblastoma Pathways	[26]	[37]	10	78	candidate mut	GBM_pa
Pancreas Genes	[26]	[29]	7	90	candidate mut	Pan_ge
Pancreas Pathways	[26]	[29]	7	90	candidate mut	Pan_pa
Lung	[34]	[21]	51	161	recurrent mut	Lu
Ovarian	[34]	[8]	192	326	recurrent mut	Ov
Ovarian driver	[34]	[8]	9	326	significant mut	Ov_drv
Colon MSI	[11]	[9]	30	27	significant mut and CNA	Col_msi
Colon MSS	[11]	[9]	34	152	significant mut and CNA	Col_mss
Colon MSI mutual exclusivity groups collapsed	[11]	[9]	20	27	significant mut and CNA	Col_msi_co
Colon MSS mutual exclusivity groups collapsed	[11]	[9]	13	152	significant mut and CNA	Col_mss_co
ACML	[16, 39]	[38]	16	64	recurrent mut	ACML
ACML mutual exclusivity groups collapsed	[16, 39]	[38]	11	64	recurrent mut	ACML_co
GBM CNA	[13, 25]	[5]	48	563	significant CNA	GBM_CNA
GBM CNA modules	[13, 25]	[5]	9	563	significant CNA	GBM_mo
GBM co-occurrent	[3]	[4, 7]	3	594	significant co-occurrent CNA	GBM_coo
Ovarian CNV	[42]	[30]	7	87	recurrent arm-level CNA	Ov_CNV
BRCA HER2, subtypes	[13, 25]	[10]	4	57	significant mut	BRCA_he_s
BRCA basal-like, subtypes	[13, 25]	[10]	6	81	significant mut	BRCA_ba_s

Table S1: Cancer data sets used. Source refers to where the data have been obtained from, generally also the first reference where data set has been used with CPMs. Data sets BRCA_he_s and BRCA_ba_s have been obtained from original sources for this paper. A data set very similar to GBM_CNA was used in [14], but we obtained it from [13, 25], as explained in the text. Type of event: mut: nonsynonymous somatic mutations; CNA: copy number alterations; candidate mut: nonsynonymous mutations on candidate genes [41, 43, 44]; significant mut: nonsynonymous mutations on significant mutated genes, as defined by state-of-the-art algorithms [43] MuSiC [17] or MutSigCV [31]; significant CNA: significant copy number alterations, as defined by GISTIC2.0 [33].

These are further details about how the data were obtained and the rationale for the data processing:

All Pathways and Colon, Glioblastoma, Pancreas pathways Data sets Colon Genes, Glioblastoma genes, and Pancreas genes are from [26], with original sources [44], [37], and [29], respectively.

For the corresponding data sets in terms of pathways, the mapping from genes to pathways was done by [26], from the original papers with data sets. Our scripts to reproduce the analysis are provided with the code. Note that for Pancreas pathways we eliminate the four pathways that were present in all subjects (see also [26] and notes in the code for details). For Glioblastoma pathways, two pathways had identical patterns (Apoptosis and Small GTPase-dependent signaling (other than KRAS)) and only one was used.

What we call "All Pathways" here, for brevity, is called "All cancer types" in [26].

Lung Original data from [21]. They were obtained from text file `Lung_SM4` from the supplementary material of [34] (file "BMLv1.tar.gz").

Ovarian Original data from [8]. They were obtained from text file `OV_SM5` from the supplementary material of [34] (file "BMLv1.tar.gz").

Ovarian driver Data come from dataset Ovarian, restricting events to 9 significantly mutated genes described in Table 2 of source paper [8].

Colon MSI Colorectal cancer, microsatellite unstable tumors. The original data (as well as Colon MSS) come from COADREAD [9]; we obtained them from [11], where the original data were split by tumour subtype into MSI and MSS (see also comments about patient stratification under "BRCA basal-like, subtypes, BRCA HER2, subtypes").

We used GIMP to open the pdf file page (Figure 3 on page 6 of [11]) where the figure was and cropped the grid of the figure and exported it as JPEG with high resolution. Then we imported it in ImageJ(Fiji) (`https://fiji.sc/`), converted it to 8-bit, applied threshold option and set it to B/W, then exported it as text image (matrix as txt). Then we imported the text image in R and used the code in `fig_to_matrix_capri_pnas.R` to convert the text image into a matrix of genotypes. The data were checked against the original figures.

Colon MSS Colorectal cancer, microsatellite stable tumors. The original data come from COADREAD [9] and we obtained them from [11], where the original data were split by tumour subtype into MSI and MSS (see above).

From [11]. Same process as for Colon MSI; the figure is Figure S5 from page 16 of the supplementary material to [11]. The authors explain that "Events selected for reconstruction are those involving genes altered in at least 5% of the cases, or part of group of alterations showing an exclusivity trend (see Figure S4)."

Colon MSI mutual exclusivity groups collapsed, Colon MSS mutual exclusivity groups collapsed Data sets were obtained from data sets **Colon MSI** and **Colon MSS**, respectively, processed so events showing mutual exclusivity patterns described in [11] were collapsed in a single event representing an exclusivity group.

Mutual exclusivity patterns, as explained in [11], could decrease the performance of CPMs (as CPMs assume no events show exclusivity or otherwise reduce the probability of another event occurring [34]). How to deal with these exclusivity patterns with CPMs such as CBN, OT, CAPRESE, or CAPRI without additional formulas, is not clear, however. For the Colon MSI and Colon MSS data sets, the exclusivity groups identified in Caravagna *et al.* (2016), [11], are supposed to represent fitness-equivalent exclusive sets of alterations. Some of these exclusivity groups share events, some represent "hard" exclusivity relations whereas others represent "soft" exclusivity relations [11]. What we have done is consider each one of the exclusivity groups as analogous to a pathway in Gerstung *et al.* (2011) [26] or a module in Cheng *et al.* (2012) [14] (where the same gene can be part of different pathways/modules or, in this case, different exclusivity groups). This amounts to considering each exclusivity group as a "fitness equivalent" (*sensu* [11]) set of alterations for some "phenotype" shared by the exclusivity group, similar to what [11] did, and should not introduce any additional difficulties for the inference of downstream dependencies.

We removed from the data set any alteration that was a member of one or more exclusivity groups as an individual alteration. Thus, the data sets with exclusivity groups differ from the original ones by adding exclusivity groups and removing alterations that belong to those exclusivity groups. We used the Table S3 of the Supplementary Material of [11] as the canonical source of exclusivity groups. However, notice that there must be the following mistakes in Table S3: in row 6 it says ACVR1B:a, but there is no amplification event that affects ACVR1B in data set **Colon MSI**, according to Figure 3 in the paper (and Figure 5 and Figure S11), but mutation and deletion; in row 7 it says NRAS:a but there is no amplification event that affects NRAS in data set **Colon MSI**, according to Figure 3 in the paper (and Figure 5 and Figure S11), but mutation and deletion; in row 15 it says NRAS:d but there is no deletion event that affects NRAS in data set **Colon MSS**, according to Figure S5 (and Figure S4 and Figure S10), but mutation and amplification. So we assumed it should say ACVR1B:d in row 6, NRAS:d in row 7 and NRAS:a in row 15. Additionally, when a mutual exclusivity group in Table S3 was contained in another (for instance, group in row 5 is part of group of row 1) we used only the bigger one. This procedure results, for **Colon MSI**, in a new data set of 27 subjects and 20 columns (five from the exclusivity groups, and 15 from the 30 alterations in the original data set minus the 15 removed as they are in one or more exclusivity groups). For **Colon MSS**, this results in a new data set of 152 subjects and 13 columns (11 from the exclusivity groups, and 2 from the 34 alterations in the original data set minus the 32 removed as they are in one or more exclusivity groups).

ACML Data are originally from [38], and were obtained from the `aCML` data set in R package "TRONCO" [16] and processed to keep the 16 events used in [39]. The data includes alterations with a frequency above 5 % in original data set from [38] and additional selected alterations hypothesized to be part of a functional ACML progression path in the literature and are shown in Figure 5 of [39]. As explained in [39], events are categorized as insertion/deletions, missense point mutations, and nonsense point mutations. This data set shows mutual exclusivity patterns described in Section 4.2 of [39].

ACML mutual exclusivity groups collapsed Data come from data set **ACML**, processed so events showing mutual exclusivity patterns described in [39] were collapsed in a single event. As with data sets **Colon MSI** and **Colon MSS**, here we dealt with mutual exclusivity patterns in the data by collapsing individual events in exclusivity groups considered as "fitness equivalent" groups. The two exclusivity groups were obtained from Section 4.2 of [39]: one involves all types of alterations of genes ASXL1 and SF3B1 (ASXL1 nonsense point, ASXL1 ins/del, SF3B1 missense point) and the other involves all types of alterations of genes TET2 and IDH2 (TET2 nonsense point, TET2 missense point, TET2 ins/del, IDH2 missense point); thus, there are 11 columns, 2 from the exclusivity groups, and 9 from the 16 alterations minus the 7 removed as they belong into exclusivity groups.

GBM CNA Data come from TCGA GBM PUB CNA data set [5] and were obtained from cBioPortal [13, 25] using the R package "cgdsr" [28], selecting only CNA data from 51 driver genes used by Cheng *et al.* (2012) and detailed in Table 1 of [14], with a GISTIC score of 2 or -2. By doing so, we intended to follow the author's indications to obtain the same data set Cheng *et al.* (2012), [14], used to infer a cancer progression model. Although [14] cite [7] as the source of their data, we understand that the original data set used in [14] should have been the "Provisional" TCGA data set at that time since they got more patients (462) than the total number of patients in the only published TCGA glioblastoma study at the date [7] (206). So, we used the data from the TCGA glioblastoma study published in 2013 [5] on the belief that this is the closest we can get to reproduce the data set used in [14] (the study of 2013, [5], contains 198 patients in common with the study of 2008, [7]). Note that 3 of the 51 genes were not altered in any subject and then were removed from the data set. Also note that, contrary to [14], to avoid mutual exclusivity patterns we only analyze one level of gain/loss per gene; so, here we used high-level amplification (GISTIC score of 2) and homozygous deletion (GISTIC score of - 2).

GBM CNA modules Data come from **GBM CNA**, processed so individual alterations were grouped in modules at phenotype level of cancer-related pathways.

We reproduced the procedure used in Cheng *et al.* (2012) [14] to map alterations to functional modules of positive or negative effects within different cancer-related pathways as originally described in [12] and as detailed in Table 1 of [14].

GBM co-ocurrent Data come originally from [7] and [4] and were obtained from Supplementary Material file `ST01.xls` (GBM_copy_number tab) of Attollini *et al.* (2010) and processed to keep only three highly correlated events described in [3], namely PTEN homozygous deletion, P16 homozygous deletion and EFGR low-level amplification.

Ovarian CNV Obtained from data set `ov.cgh` in the R package "Oncotree" [42]. Data are originally from [30]; `ov.cgh` from the "Oncotree" R package has also been used in [36].

BRCA basal-like, subtypes, BRCA HER2, subtypes Original data come from TCGA BRCA PUB mutation data set [10] and were obtained from cBioPortal [13, 25] using the R package "cgdsr" [28], restricting the data to subtype-specific significantly mutated genes within patients subtypes.

The data were then split in two, restricting the subjects to those classified as basal-like and HER2-enriched subtypes, respectively. To split the dataset according to cancer subtypes we used the patient's classification by the gene expression-based PAM50 technique as detailed in Supplementary Table 1 of [10]. Then, for each subtype, we restricted the features to subtype-specific significantly mutated genes identified by MuSiC algorithm [17] and detailed in Supplementary Table 2 of [10] (Supplementary Tables 1-4.xls file in supplementary file nature11412-s2.zip).

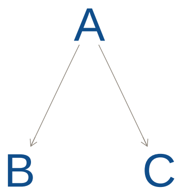
The reason for splitting the data into two subsets is that cancer subtypes are believed to follow distinct evolutionary trajectories, and hence rely on at least some different drivers and/or pathways, and show differences in the chronology of accumulation of alterations [1, 2, 14]. Thus, sample heterogeneity in terms of different cancer subtypes (intertumor heterogeneity) can be confounding and hamper the identification of existing relations in the data. Sample stratification can alleviate this to some extent and should allow to focus on relevant events for specific subsets of subjects [11].

6.2. Bootstrapping on the cancer data sets

If the bootstrapping process resulted in a feature becoming absent from the data, or two or more features having identical patterns (i.e., one feature being identical to another) we discarded the bootstrap sample and obtained a new one; this is done to ensure that all bootstrapped data sets have paths of identical length (see also section 5.3). This, therefore, leads to JS values that are more optimistic (smaller).

7. CAPRI, CAPRESE, and paths of tumor progression

With both OT and CBN if we see a DAG such as



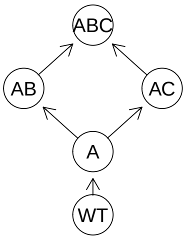
this is saying that, except for errors (errors in the model and observational errors) the genotypes that can exist under the model are only (with our usual notation of using a capital letter to denote that the gene is mutated, and no letter to denote absence of mutation)

```
WT
A
AB
AC
ABC
```

and, consequently, there are only two paths to the maximum:

```
WT -> A -> AB -> ABC
WT -> A -> AC -> ABC
```

or



Which means that, for example, under OT and CBN the following paths to the maximum are not allowed under the model:

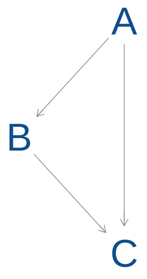
```
WT -> B -> BC -> ABC
WT -> B -> AB -> ABC
```

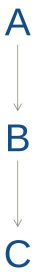
This interpretation, however, does not necessarily follow with CAPRI. CAPRI returns, as output, a DAG and a set of conditional probability tables (CPTs) associated to each node. What CAPRI seems to say, as can be checked from non-zero entries in the CPTs for B without A or C without A in a DAG as above, is that the DAG says that the most likely mutational paths are

```
WT -> A -> AB -> ABC
WT -> A -> AC -> ABC
```

but other mutational paths could also take place under the model. (This follows directly from seeing CAPRI return a CPT where, say, $P(B|\neg A) = 0.4$, i.e., mutating B without A is certainly not a rare event, even when an arrow exists from A to B). In fact, any path might happen (with some conditional probability tables), and one would seem to need to look at the CPTs to understand which ones can or cannot happen (but see below). And we are given no clear definition of what most likely paths really mean (e.g., "trends of selective advantage among genomic alterations" or "most common evolutionary trajectories" in the wording of 11) in terms of what probabilities are considered or not.

We see a similar issue with the following two DAGs:





because, of course, the transitive reduction of the first DAG is the second DAG. So from the point of view of paths from the non-mutated to the fully mutated genotype, both DAGs have the same meaning, as both imply that the same order of events needs to happen (or the same restrictions in the accumulation of mutations hold). Yet CAPRI seems to make a distinction between them. A distinction that could only be disentangled, presumably, by looking at the CPTs.

However, no information is available on how to go from CPTs to the conditional probabilities of genotypes implied by the model. In fact, directly using the CPT information seems discouraged and not something that users are supposed to do: access to the CPT has to be done via `TRONCO:::as.bnlearn.network(some_tronco_model)`, i.e., using the `:::` which denotes that we are accessing a non-exported function from the software. (This was the case with v. 2.11.0 and is the case as of July-2018 with version v. 2.13.0).

This contrasts with, say, CBN (and MCCBN) and OT which provide, respectively, estimated λs and `edge weights` (`object$parent$est.weight`). These conditional probabilities are what we use to obtain the probability-weighted paths implied by the model.

This is a key difference between OT and CBN on the one hand and CAPRI (and CAPRESE) on the other: OT and CBN incorporate errors in their models, but they return the estimates of the parameters of their models, i.e., the λs or the `est.weights`, that map directly into what can happen under the model, what genotypes can arise and from which other genotypes. (This is analogous to a simple linear regression: there is an error component in the model $y = \alpha + \beta x + \epsilon$, the ϵ, often assumed to be independent and identically distributed from a normal distribution with mean 0 and variance σ^2, etc, but the method, and the software, return the α and β which allow us to predict the expected value of y given x, under the model).

In contrast, with CAPRI we can (again, using a non-exported function) obtain the CPTs that correspond to the DAG returned. Remember that essentially what CAPRI is doing is fitting a Bayesian Network to the observational data, with the DAG built so that arrows respect the temporal priority and probability raising restrictions. But the CPTs themselves are not estimated parameters of a model that could be mapped into probabilities of paths. The CPTs of CAPRI seem to be the conditional probabilities of observing what we observe under the DAG and they incorporate errors (model errors and noise in the data), and can contain non-zero entries for child nodes when their parents are absent. Obtaining probabilities of paths from these CPTs is, thus, impossible.

An additional issue that has been noted with CAPRI in the paper is its behavior with increasing sample sizes. That was also noted in [20] and seems related to the penalization in the fits: CAPRI does not seem suited to deal with large data sets as it tends to allow only one, or a few, paths to the maximum when N is 4000. (The transitive reduction of) the DAGs returned from CAPRI is often a linear sequence. This behavior did not change considerably whether we used AIC or BIC.

CAPRESE does not return DAGs, but trees, and in that sense it is simpler to deal with. But still, as with CAPRI, there is no information available on how to go from CPTs to the conditional probabilities of genotypes implied by the model (and accessing the CPT does not seem to be encouraged) and, as for CAPRI, the CPTs seem to mix the underlying model with the error model, and obtaining probabilities of paths from these CPTs is, thus, impossible.

8. Overall patterns for the six methods

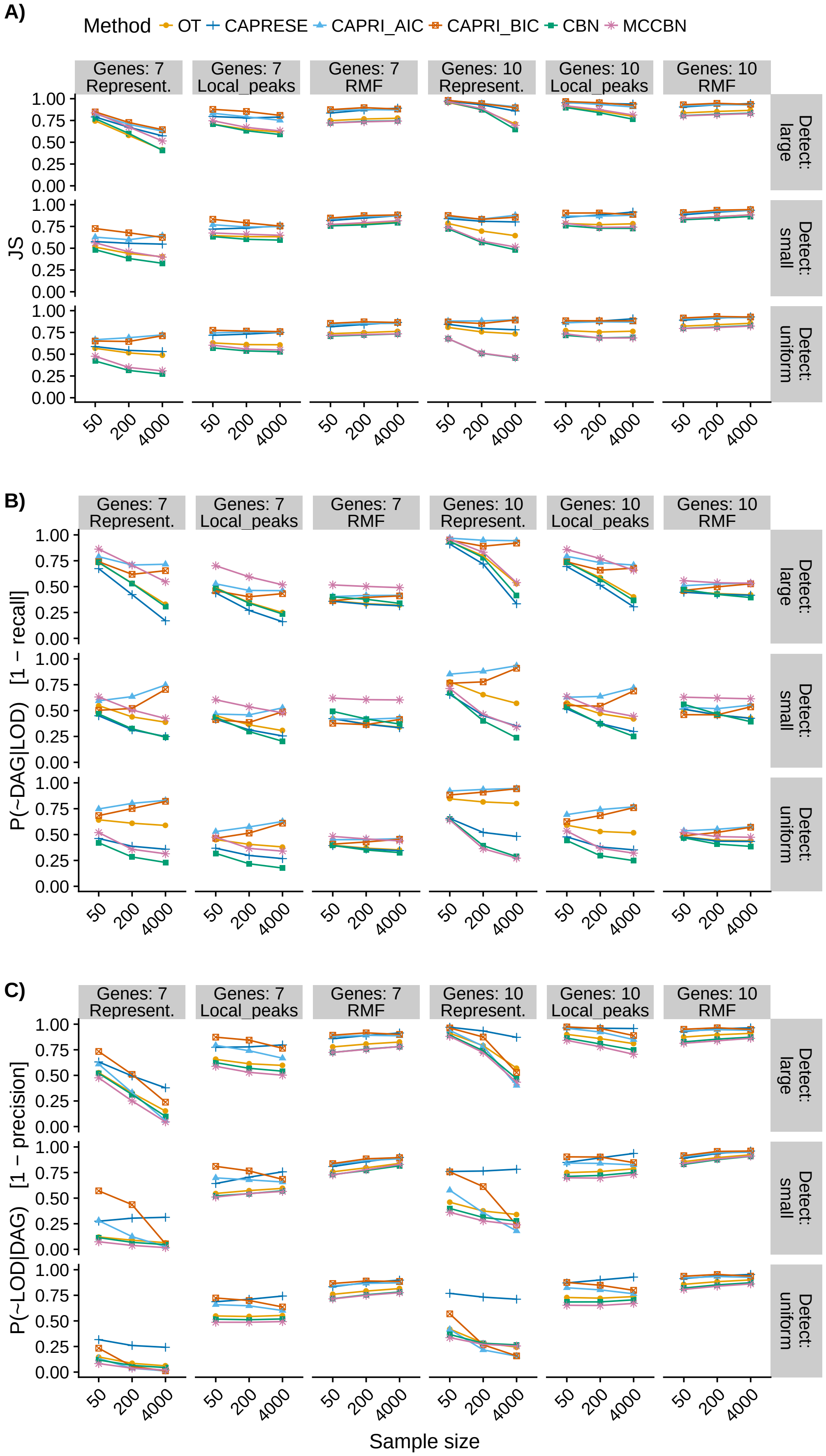


Figure S18: Summary performance measures for all six methods for all combinations of sample size, type of landscape, detection regime, and number of genes. For all measures, smaller is better. For OT, CBN, and MCCBN, Jensen-Shannon entropy and 1-precision use probability-weighted predicted paths (see 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; we are thus marginalizing over initial size by mutation rate; each one of the 210 points is, itself, the average of five runs on different partitions of the simulated data.

9. Probability of recovering the most common LOD

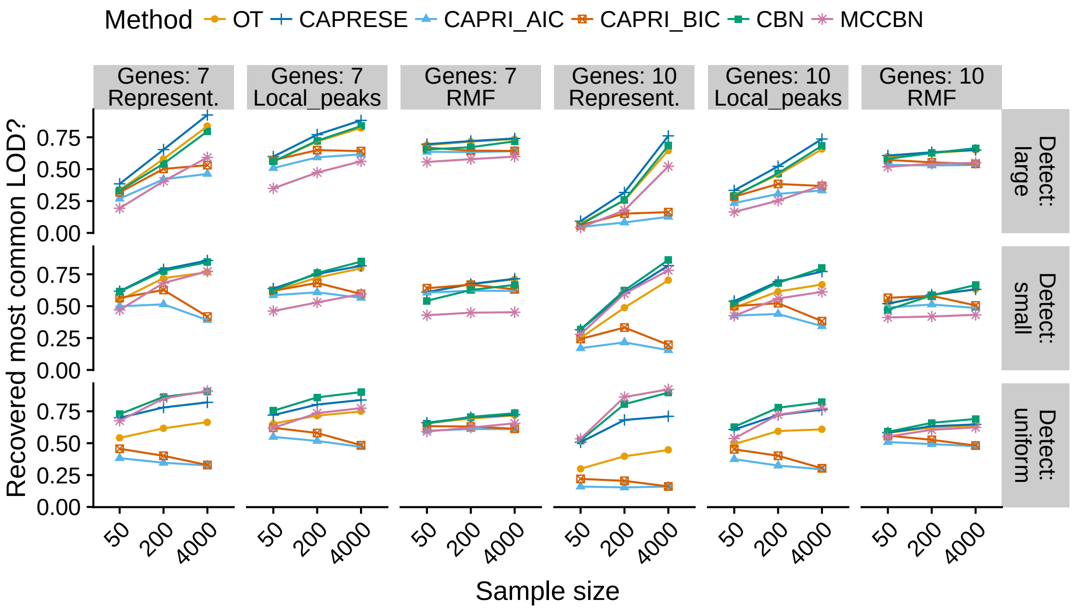


Figure S19: Probability of recovering the most common LOD: probability that the most common observed path to the maximum is among the paths allowed by the CPMs.

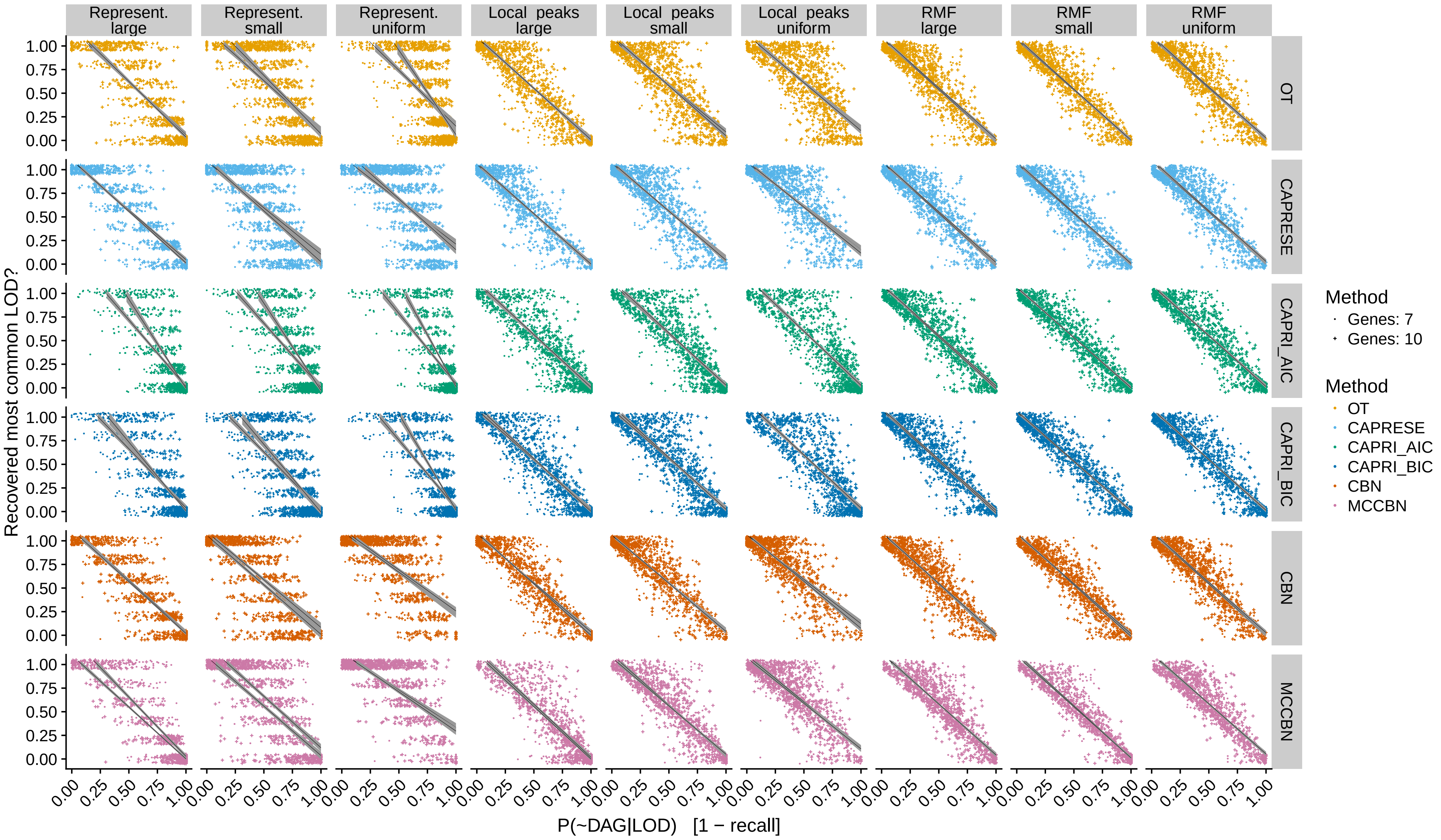


Figure S20: Probability of recovering the most common LOD and 1-recall: relationship.

10. OT and CBN, JS, weighted vs. unweighted

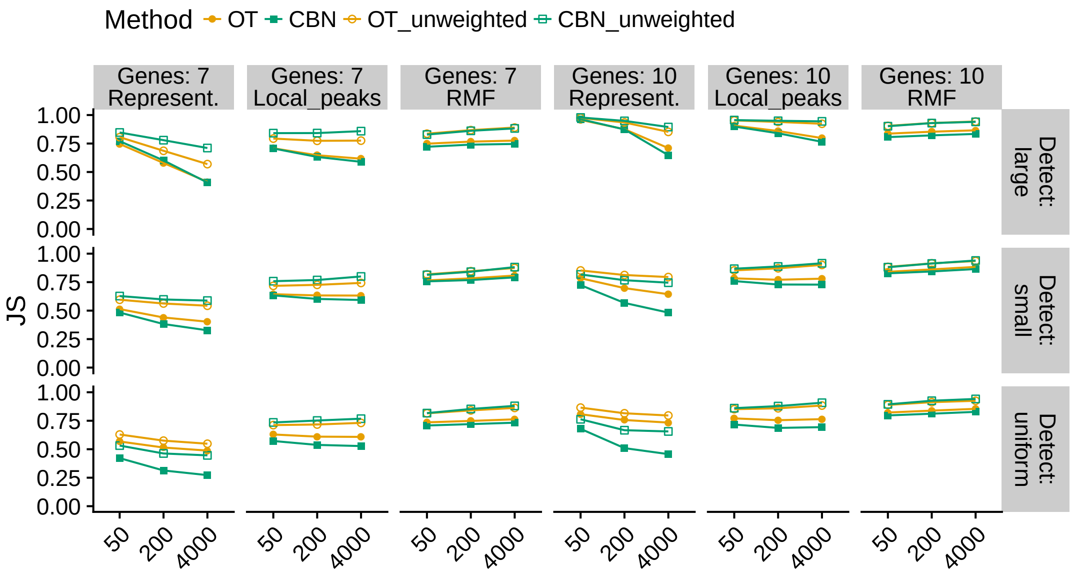


Figure S21: Comparison of the performance of OT and CBN using weighted and unweighted probabilities of paths to the maximum.

11. CAPRI and CBN, 1-precision, unweighted

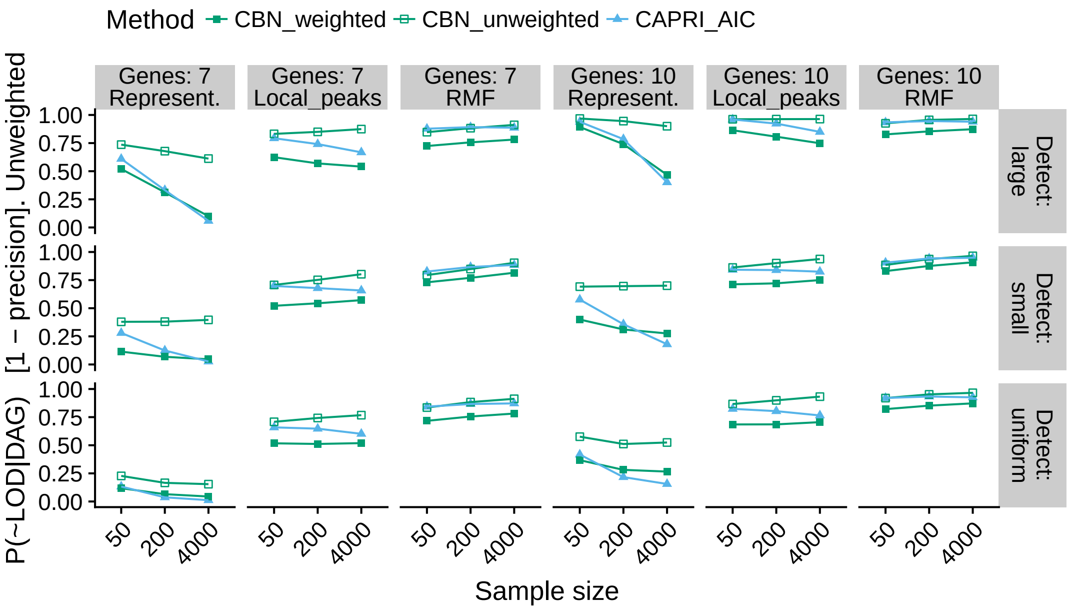


Figure S22: Comparison of the performance of CAPRI with CBN using weighted and unweighted probabilities of paths to the maximum.

12. CAPRESE and OT, 1-precision, unweighted

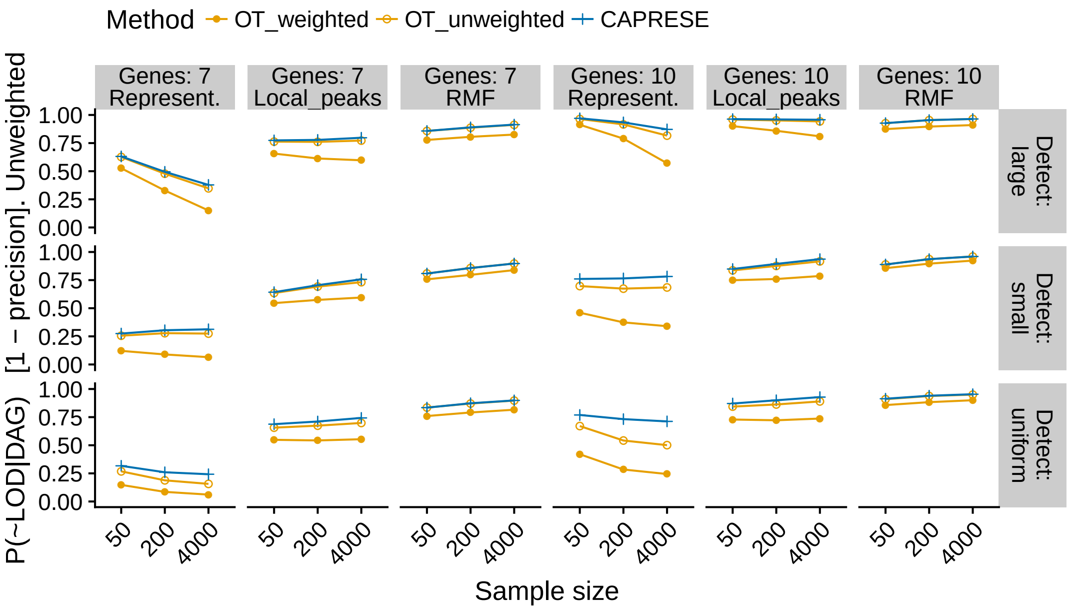


Figure S23: Comparison of the performance of CAPRESE with OT using weighted and unweighted probabilities of paths to the maximum.

The results of OT here are remarkable because OT can only build trees, and therefore cannot reflect the dependency of a mutation on two or more upstream mutations so it is prone to allow more paths to the maximum. The results of OT contrasts with those of CAPRESE, the other method that only builds trees. CAPRESE is building DAGs of restrictions that have too few restrictions and, therefore, allow for too many paths to the maximum. One notable difference between the two methods is that with OT it is relatively simple to use a measure of 1-precision that weights by the probability of each path. The performance of OT, even if we use unweighted probabilities of paths, is much better than that of CAPRESE but improves even further when using weighted paths, again highlighting the usefulness of weighting paths to obtain more accurate predictions.

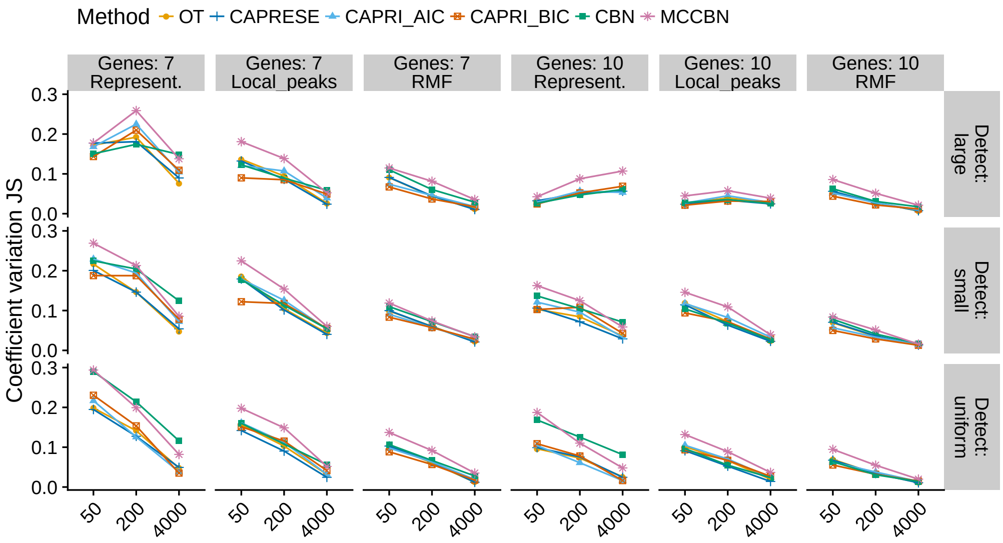


Figure S24: Coefficient of variation (standard deviation/mean) of JS for each combination of method and type of fitness landscape. The coefficient of variation has been computed from the five runs for each landscapes on each combination of sample size and detection regime.

13. Number of paths inferred

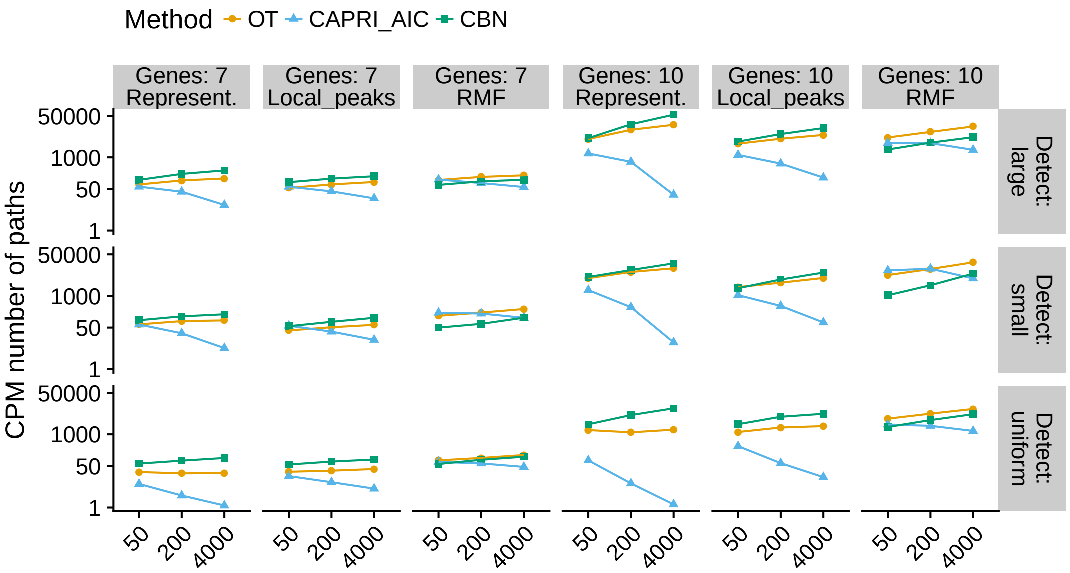


Figure S25: Number of paths to the maximum according to the CPMs.

14. Slopes of regressions of 1-recall and 1-precision on LOD diversity, S_p

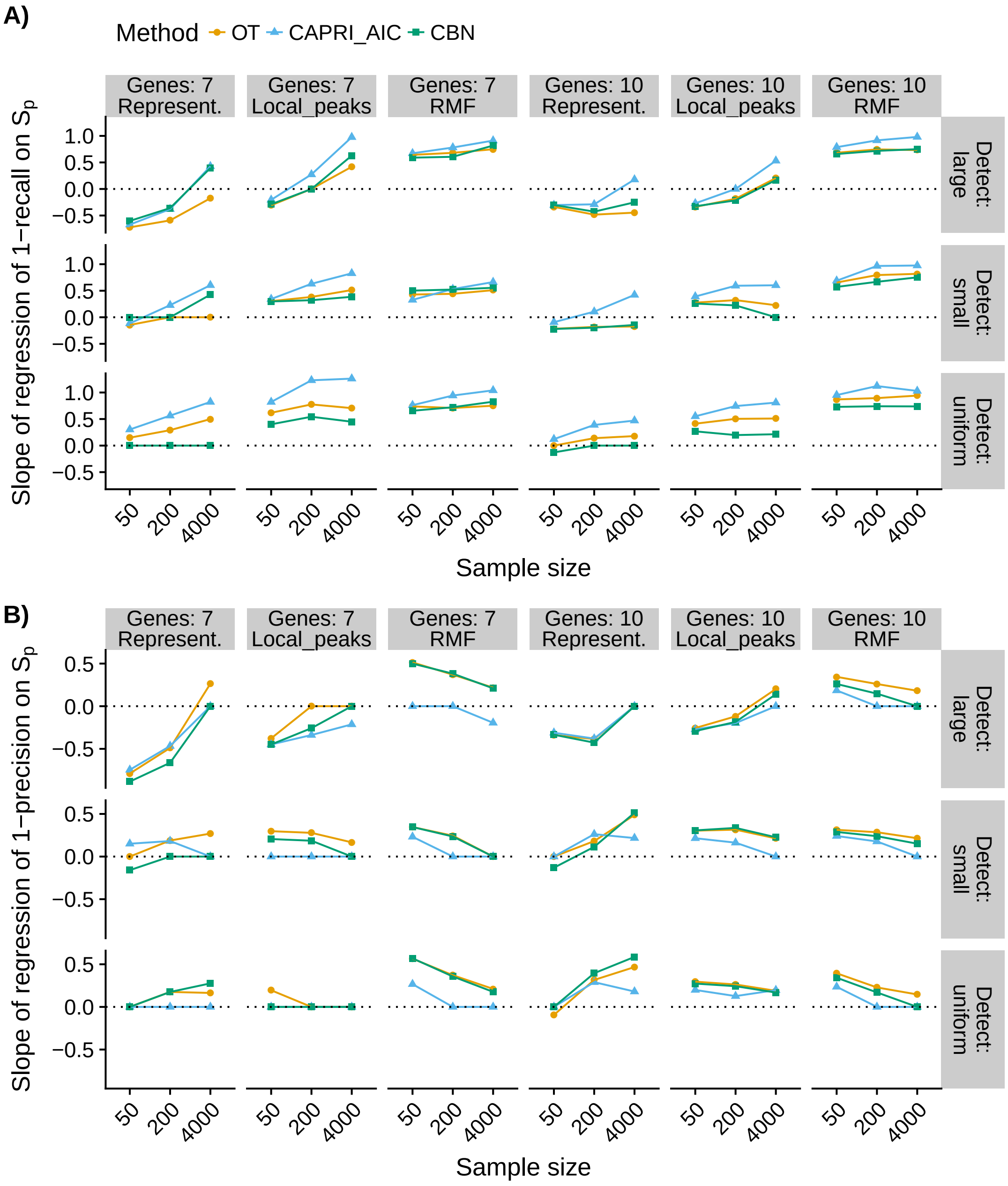


Figure S26: Slopes of regressions of 1-recall and 1-precision on LOD diversity, S_p

15. Coefficient of variation of S_c

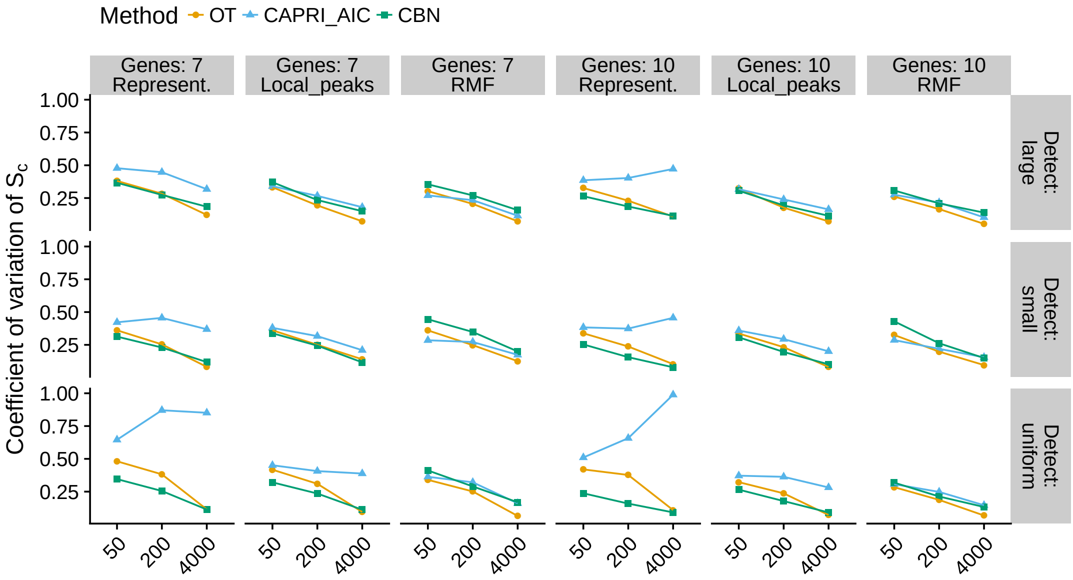


Figure S27: Coefficient of variation (standard deviation/mean) of S_c for each combination of method and type of fitness landscape. The coefficient of variation has been computed from the five runs for each landscapes on each combination of sample size and detection regime. For OT and CBN, it is computed using the probability-weighted predicted paths (see text). Each point plotted is the average of 210 points.

16. Estimated S_c by CBN

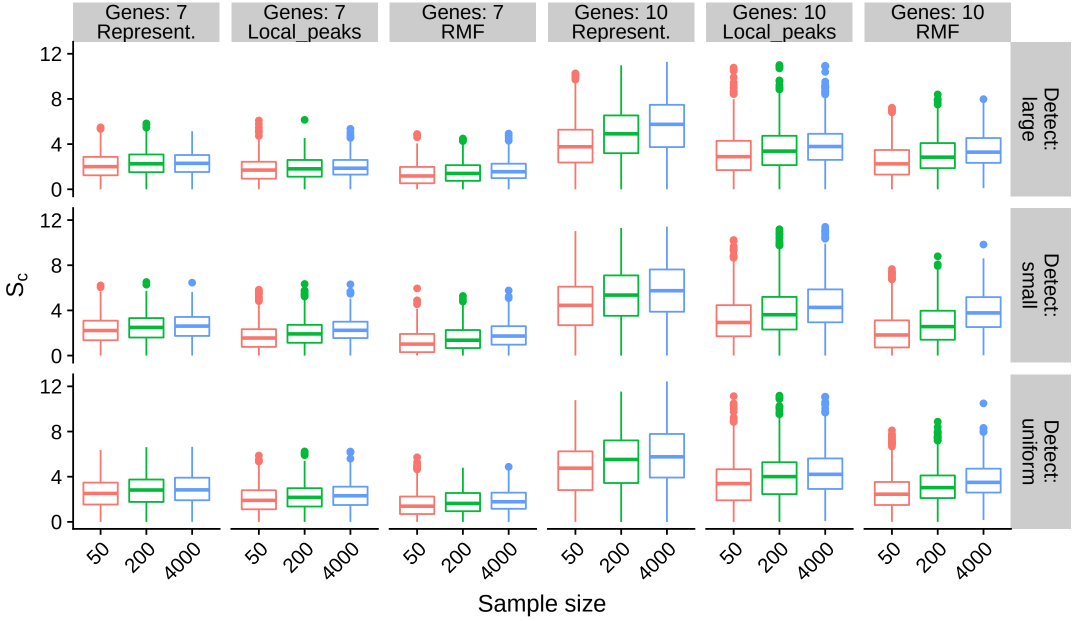


Figure S28: Estimated S_c by CBN for all combinations of sample size by type of landscape by detection regime by number of genes. Each box plot shows 1050 points.

17. Analysis of deviance tables for fitted models

The tables below show the analysis of deviance tables for the generalized (beta regressions) linear mixed effects models. Models where fitted using the R package glmmTMB [6]. Analysis of deviance tables are from package car [24]. All analysis of deviance tables use Type II Wald chi-square tests.

Analysis have been run on the complete data set (section 17.1), and after splitting for the different combinations of method and fitness landscape type (section 17.2).

17.1. Models fitted to the complete data set

(Notice the strong evidence we see for three and four and even five way interactions.)

17.1.1. Two-way interactions

	Chisq	Df	Pr(>Chisq)
Num. Genes	223.34	1.00	$< .0001$
Method	9459.99	2.00	$< .0001$
Landscape	322.37	2.00	$< .0001$
Detection	4768.56	2.00	$< .0001$
Sample Size	1599.17	2.00	$< .0001$
LOD diversity	139.98	1.00	$< .0001$
Num. Genes:Method	22.98	2.00	$< .0001$
Num. Genes:Landscape	91.35	2.00	$< .0001$
Num. Genes:Detection	2318.60	2.00	$< .0001$
Num. Genes:Sample Size	381.80	2.00	$< .0001$
Num. Genes:LOD diversity	8.72	1.00	0.0032
Method:Landscape	192.88	4.00	$< .0001$
Method:Detection	678.41	4.00	$< .0001$
Method:Sample Size	818.96	4.00	$< .0001$
Method:LOD diversity	170.20	2.00	$< .0001$
Landscape:Detection	6534.14	4.00	$< .0001$
Landscape:Sample Size	3110.94	4.00	$< .0001$
Landscape:LOD diversity	78.21	2.00	$< .0001$
Detection:Sample Size	2420.56	4.00	$< .0001$
Detection:LOD diversity	3303.42	2.00	$< .0001$
Sample Size:LOD diversity	939.31	2.00	$< .0001$

Table S2: Full model, 2-way interactions

17.1.2. Three-way interactions

	Chisq	Df	Pr(>Chisq)
Num. Genes	197.19	1.00	$< .0001$
Method	11331.58	2.00	$< .0001$
Landscape	411.00	2.00	$< .0001$
Detection	4217.33	2.00	$< .0001$
Sample Size	1536.56	2.00	$< .0001$
LOD diversity	146.46	1.00	$< .0001$
Num. Genes:Method	52.74	2.00	$< .0001$
Num. Genes:Landscape	80.98	2.00	$< .0001$
Num. Genes:Detection	2394.91	2.00	$< .0001$
Num. Genes:Sample Size	354.72	2.00	$< .0001$
Num. Genes:LOD diversity	8.15	1.00	0.0043
Method:Landscape	285.47	4.00	$< .0001$
Method:Detection	707.50	4.00	$< .0001$
Method:Sample Size	842.82	4.00	$< .0001$
Method:LOD diversity	117.61	2.00	$< .0001$
Landscape:Detection	6924.74	4.00	$< .0001$
Landscape:Sample Size	3407.98	4.00	$< .0001$
Landscape:LOD diversity	77.13	2.00	$< .0001$
Detection:Sample Size	2507.80	4.00	$< .0001$
Detection:LOD diversity	4229.09	2.00	$< .0001$
Sample Size:LOD diversity	1120.46	2.00	$< .0001$
Num. Genes:Method:Landscape	90.47	4.00	$< .0001$
Num. Genes:Method:Detection	36.64	4.00	$< .0001$
Num. Genes:Method:Sample Size	10.43	4.00	0.0338
Num. Genes:Method:LOD diversity	41.35	2.00	$< .0001$
Num. Genes:Landscape:Detection	1302.06	4.00	$< .0001$
Num. Genes:Landscape:Sample Size	347.91	4.00	$< .0001$
Num. Genes:Landscape:LOD diversity	3.70	2.00	0.1572
Num. Genes:Detection:Sample Size	553.21	4.00	$< .0001$
Num. Genes:Detection:LOD diversity	3.91	2.00	0.1417
Num. Genes:Sample Size:LOD diversity	56.88	2.00	$< .0001$
Method:Landscape:Detection	250.27	8.00	$< .0001$
Method:Landscape:Sample Size	192.75	8.00	$< .0001$
Method:Landscape:LOD diversity	605.91	4.00	$< .0001$
Method:Detection:Sample Size	19.44	8.00	0.0127
Method:Detection:LOD diversity	126.85	4.00	$< .0001$
Method:Sample Size:LOD diversity	117.71	4.00	$< .0001$
Landscape:Detection:Sample Size	2084.94	8.00	$< .0001$
Landscape:Detection:LOD diversity	867.54	4.00	$< .0001$
Landscape:Sample Size:LOD diversity	1163.94	4.00	$< .0001$
Detection:Sample Size:LOD diversity	736.44	4.00	$< .0001$

Table S3: Full model, 3-way interactions

17.1.3. Four-way interactions

	Chisq	Df	Pr(>Chisq)
Num. Genes	193.68	1.00	< .0001
Method	11619.48	2.00	< .0001
Landscape	426.49	2.00	< .0001
Detection	4059.78	2.00	< .0001
Sample Size	1500.59	2.00	< .0001
LOD diversity	150.27	1.00	< .0001
Num. Genes:Method	54.25	2.00	< .0001
Num. Genes:Landscape	77.19	2.00	< .0001
Num. Genes:Detection	2370.70	2.00	< .0001
Num. Genes:Sample Size	357.14	2.00	< .0001
Num. Genes:LOD diversity	8.47	1.00	0.0036
Method:Landscape	298.90	4.00	< .0001
Method:Detection	716.32	4.00	< .0001
Method:Sample Size	841.73	4.00	< .0001
Method:LOD diversity	119.30	2.00	< .0001
Landscape:Detection	6697.76	4.00	< .0001
Landscape:Sample Size	3459.44	4.00	< .0001
Landscape:LOD diversity	77.85	2.00	< .0001
Detection:Sample Size	2466.93	4.00	< .0001
Detection:LOD diversity	4132.54	2.00	< .0001
Sample Size:LOD diversity	1103.02	2.00	< .0001
Num. Genes:Method:Landscape	101.44	4.00	< .0001
Num. Genes:Method:Detection	35.22	4.00	< .0001
Num. Genes:Method:Sample Size	11.99	4.00	0.0174
Num. Genes:Method:LOD diversity	43.71	2.00	< .0001
Num. Genes:Landscape:Detection	1262.35	4.00	< .0001
Num. Genes:Landscape:Sample Size	350.64	4.00	< .0001
Num. Genes:Landscape:LOD diversity	3.55	2.00	0.1699
Num. Genes:Detection:Sample Size	532.95	4.00	< .0001
Num. Genes:Detection:LOD diversity	2.93	2.00	0.231
Num. Genes:Sample Size:LOD diversity	50.25	2.00	< .0001
Method:Landscape:Detection	229.42	8.00	< .0001
Method:Landscape:Sample Size	215.46	8.00	< .0001
Method:Landscape:LOD diversity	627.29	4.00	< .0001
Method:Detection:Sample Size	21.27	8.00	0.0065
Method:Detection:LOD diversity	131.95	4.00	< .0001
Method:Sample Size:LOD diversity	88.97	4.00	< .0001
Landscape:Detection:Sample Size	2216.86	8.00	< .0001
Landscape:Detection:LOD diversity	867.13	4.00	< .0001
Landscape:Sample Size:LOD diversity	1175.97	4.00	< .0001
Detection:Sample Size:LOD diversity	819.51	4.00	< .0001
Num. Genes:Method:Landscape:Detection	12.73	8.00	0.1214
Num. Genes:Method:Landscape:Sample Size	7.44	8.00	0.4898
Num. Genes:Method:Landscape:LOD diversity	14.84	4.00	0.0051
Num. Genes:Method:Detection:Sample Size	25.87	8.00	0.0011
Num. Genes:Method:Detection:LOD diversity	53.99	4.00	< .0001
Num. Genes:Method:Sample Size:LOD diversity	2.51	4.00	0.6425
Num. Genes:Landscape:Detection:Sample Size	321.87	8.00	< .0001
Num. Genes:Landscape:Detection:LOD diversity	74.72	4.00	< .0001
Num. Genes:Landscape:Sample Size:LOD diversity	101.05	4.00	< .0001

Num. Genes:Detection:Sample Size:LOD diversity	115.78	4.00	< .0001
Method:Landscape:Detection:Sample Size	24.77	16.00	0.0739
Method:Landscape:Detection:LOD diversity	19.36	8.00	0.013
Method:Landscape:Sample Size:LOD diversity	23.93	8.00	0.0024
Method:Detection:Sample Size:LOD diversity	10.02	8.00	0.2634
Landscape:Detection:Sample Size:LOD diversity	237.96	8.00	< .0001

Table S4: Full model, 4-way interactions

17.2. Models fitted to each combination of fitness landscape by method

Remember each model uses 3780 observations: 35 replicates, 3 mutation rates, 2 variance settings, 2 number of genes, 3 detection regimes, 3 sample sizes. These correspond to 420 different fitness landscapes: 35 by 3 by 2 by 2. Each observation is itself the average of five different splits of the set of 20000 simulations.

17.2.1. Main effects

	Chisq	Df	Pr(>Chisq)
Num. Genes	325.32	1.00	< .0001
Sample Size	797.56	2.00	< .0001
Detection	1101.14	2.00	< .0001
LOD diversity	2.41	1.00	0.1206

Table S5: Represent..OT

	Chisq	Df	Pr(>Chisq)
Num. Genes	84.38	1.00	< .0001
Sample Size	179.66	2.00	< .0001
Detection	470.52	2.00	< .0001
LOD diversity	63.71	1.00	< .0001

Table S6: Local Peaks.OT

	Chisq	Df	Pr(>Chisq)
Num. Genes	9.65	1.00	0.0019
Sample Size	199.97	2.00	< .0001
Detection	206.52	2.00	< .0001
LOD diversity	162.60	1.00	< .0001

Table S7: RMF.OT

	Chisq	Df	Pr(>Chisq)
Num. Genes	206.70	1.00	< .0001
Sample Size	62.82	2.00	< .0001
Detection	692.92	2.00	< .0001
LOD diversity	79.61	1.00	< .0001

Table S8: Represent..CAPRI_AIC

	Chisq	Df	Pr(>Chisq)
Num. Genes	110.80	1.00	< .0001
Sample Size	48.42	2.00	< .0001
Detection	383.72	2.00	< .0001
LOD diversity	81.86	1.00	< .0001

Table S9: Local Peaks.CAPRI_AIC

	Chisq	Df	Pr(>Chisq)
Num. Genes	35.67	1.00	$< .0001$
Sample Size	144.12	2.00	$< .0001$
Detection	48.52	2.00	$< .0001$
LOD diversity	70.86	1.00	$< .0001$

Table S10: RMF.CAPRI_AIC

	Chisq	Df	Pr(>Chisq)
Num. Genes	157.75	1.00	$< .0001$
Sample Size	1331.50	2.00	$< .0001$
Detection	2493.82	2.00	$< .0001$
LOD diversity	0.35	1.00	0.5538

Table S11: Represent..CBN

	Chisq	Df	Pr(>Chisq)
Num. Genes	71.83	1.00	$< .0001$
Sample Size	315.26	2.00	$< .0001$
Detection	823.90	2.00	$< .0001$
LOD diversity	64.31	1.00	$< .0001$

Table S12: Local Peaks.CBN

	Chisq	Df	Pr(>Chisq)
Num. Genes	3.82	1.00	0.0507
Sample Size	100.38	2.00	$< .0001$
Detection	320.43	2.00	$< .0001$
LOD diversity	151.14	1.00	$< .0001$

Table S13: RMF.CBN

17.2.2. Two-way interactions

	Chisq	Df	Pr(>Chisq)
Num. Genes	229.87	1.00	< .0001
Sample Size	1084.79	2.00	< .0001
Detection	1145.03	2.00	< .0001
LOD diversity	0.02	1.00	0.8958
Num. Genes:Sample Size	168.67	2.00	< .0001
Num. Genes:Detection	705.00	2.00	< .0001
Num. Genes:LOD diversity	4.46	1.00	0.0347
Sample Size:Detection	726.35	4.00	< .0001
Sample Size:LOD diversity	308.31	2.00	< .0001
Detection:LOD diversity	1019.12	2.00	< .0001

Table S14: Represent..OT

	Chisq	Df	Pr(>Chisq)
Num. Genes	66.47	1.00	< .0001
Sample Size	163.41	2.00	< .0001
Detection	449.46	2.00	< .0001
LOD diversity	74.54	1.00	< .0001
Num. Genes:Sample Size	34.28	2.00	< .0001
Num. Genes:Detection	443.26	2.00	< .0001
Num. Genes:LOD diversity	0.02	1.00	0.895
Sample Size:Detection	307.15	4.00	< .0001
Sample Size:LOD diversity	21.38	2.00	< .0001
Detection:LOD diversity	403.27	2.00	< .0001

Table S15: Local Peaks.OT

	Chisq	Df	Pr(>Chisq)
Num. Genes	9.65	1.00	0.0019
Sample Size	213.84	2.00	< .0001
Detection	215.42	2.00	< .0001
LOD diversity	155.94	1.00	< .0001
Num. Genes:Sample Size	31.75	2.00	< .0001
Num. Genes:Detection	5.39	2.00	0.0676
Num. Genes:LOD diversity	0.05	1.00	0.8191
Sample Size:Detection	7.68	4.00	0.1041
Sample Size:LOD diversity	190.09	2.00	< .0001
Detection:LOD diversity	9.29	2.00	0.0096

Table S16: RMF.OT

	Chisq	Df	Pr(>Chisq)
Num. Genes	144.41	1.00	< .0001
Sample Size	70.06	2.00	< .0001
Detection	720.04	2.00	< .0001
LOD diversity	89.03	1.00	< .0001
Num. Genes:Sample Size	154.58	2.00	< .0001
Num. Genes:Detection	574.17	2.00	< .0001
Num. Genes:LOD diversity	5.16	1.00	0.0231
Sample Size:Detection	691.60	4.00	< .0001
Sample Size:LOD diversity	734.28	2.00	< .0001
Detection:LOD diversity	958.49	2.00	< .0001

Table S17: Represent..CAPRI_AIC

	Chisq	Df	Pr(>Chisq)
Num. Genes	90.86	1.00	< .0001
Sample Size	36.29	2.00	< .0001
Detection	351.07	2.00	< .0001
LOD diversity	89.21	1.00	< .0001
Num. Genes:Sample Size	8.98	2.00	0.0112
Num. Genes:Detection	309.60	2.00	< .0001
Num. Genes:LOD diversity	0.01	1.00	0.91
Sample Size:Detection	262.51	4.00	< .0001
Sample Size:LOD diversity	80.39	2.00	< .0001
Detection:LOD diversity	429.26	2.00	< .0001

Table S18: Local Peaks.CAPRI_AIC

	Chisq	Df	Pr(>Chisq)
Num. Genes	33.69	1.00	< .0001
Sample Size	164.54	2.00	< .0001
Detection	50.28	2.00	< .0001
LOD diversity	65.36	1.00	< .0001
Num. Genes:Sample Size	34.37	2.00	< .0001
Num. Genes:Detection	2.23	2.00	0.3273
Num. Genes:LOD diversity	1.24	1.00	0.2663
Sample Size:Detection	15.51	4.00	0.0037
Sample Size:LOD diversity	154.86	2.00	< .0001
Detection:LOD diversity	40.59	2.00	< .0001

Table S19: RMF.CAPRI_AIC

	Chisq	Df	Pr(>Chisq)
Num. Genes	111.35	1.00	< .0001
Sample Size	1727.19	2.00	< .0001
Detection	2829.04	2.00	< .0001
LOD diversity	2.47	1.00	0.116
Num. Genes:Sample Size	252.55	2.00	< .0001
Num. Genes:Detection	583.85	2.00	< .0001
Num. Genes:LOD diversity	11.81	1.00	6e-04
Sample Size:Detection	583.96	4.00	< .0001
Sample Size:LOD diversity	367.88	2.00	< .0001
Detection:LOD diversity	634.67	2.00	< .0001

Table S20: Represent..CBN

	Chisq	Df	Pr(>Chisq)
Num. Genes	56.06	1.00	$< .0001$
Sample Size	313.03	2.00	$< .0001$
Detection	836.24	2.00	$< .0001$
LOD diversity	76.98	1.00	$< .0001$
Num. Genes:Sample Size	35.78	2.00	$< .0001$
Num. Genes:Detection	413.32	2.00	$< .0001$
Num. Genes:LOD diversity	0.30	1.00	0.5868
Sample Size:Detection	296.90	4.00	$< .0001$
Sample Size:LOD diversity	19.84	2.00	$< .0001$
Detection:LOD diversity	292.03	2.00	$< .0001$

Table S21: Local Peaks.CBN

	Chisq	Df	Pr(>Chisq)
Num. Genes	3.82	1.00	0.0506
Sample Size	107.26	2.00	$< .0001$
Detection	331.99	2.00	$< .0001$
LOD diversity	144.95	1.00	$< .0001$
Num. Genes:Sample Size	21.97	2.00	$< .0001$
Num. Genes:Detection	4.24	2.00	0.12
Num. Genes:LOD diversity	1.16	1.00	0.2823
Sample Size:Detection	4.15	4.00	0.3856
Sample Size:LOD diversity	181.70	2.00	$< .0001$
Detection:LOD diversity	1.36	2.00	0.5071

Table S22: RMF.CBN

17.2.3. Four-way interactions

	Chisq	Df	Pr(>Chisq)
Num. Genes	236.38	1.00	$<$ .0001
Sample Size	1105.42	2.00	$<$ .0001
Detection	1046.08	2.00	$<$ .0001
LOD diversity	0.84	1.00	0.3593
Num. Genes:Sample Size	193.98	2.00	$<$ .0001
Num. Genes:Detection	778.62	2.00	$<$ .0001
Num. Genes:LOD diversity	6.38	1.00	0.0116
Sample Size:Detection	722.33	4.00	$<$ .0001
Sample Size:LOD diversity	322.04	2.00	$<$ .0001
Detection:LOD diversity	1027.32	2.00	$<$ .0001
Num. Genes:Sample Size:Detection	96.81	4.00	$<$ .0001
Num. Genes:Sample Size:LOD diversity	28.18	2.00	$<$ .0001
Num. Genes:Detection:LOD diversity	26.39	2.00	$<$ .0001
Sample Size:Detection:LOD diversity	116.00	4.00	$<$ .0001
Num. Genes:Sample Size:Detection:LOD diversity	76.54	4.00	$<$ .0001

Table S23: Represent..OT

	Chisq	Df	Pr(>Chisq)
Num. Genes	57.63	1.00	$<$ .0001
Sample Size	166.26	2.00	$<$ .0001
Detection	380.88	2.00	$<$ .0001
LOD diversity	72.98	1.00	$<$ .0001
Num. Genes:Sample Size	35.33	2.00	$<$ .0001
Num. Genes:Detection	409.75	2.00	$<$ .0001
Num. Genes:LOD diversity	0.04	1.00	0.8394
Sample Size:Detection	326.02	4.00	$<$ .0001
Sample Size:LOD diversity	24.07	2.00	$<$ .0001
Detection:LOD diversity	412.53	2.00	$<$ .0001
Num. Genes:Sample Size:Detection	198.32	4.00	$<$ .0001
Num. Genes:Sample Size:LOD diversity	1.85	2.00	0.3967
Num. Genes:Detection:LOD diversity	5.28	2.00	0.0715
Sample Size:Detection:LOD diversity	219.97	4.00	$<$ .0001
Num. Genes:Sample Size:Detection:LOD diversity	8.77	4.00	0.0672

Table S24: Local Peaks.OT

	Chisq	Df	Pr(>Chisq)
Num. Genes	9.69	1.00	0.0019
Sample Size	215.06	2.00	$< .0001$
Detection	215.20	2.00	$< .0001$
LOD diversity	155.03	1.00	$< .0001$
Num. Genes:Sample Size	32.40	2.00	$< .0001$
Num. Genes:Detection	5.40	2.00	0.0673
Num. Genes:LOD diversity	0.05	1.00	0.8191
Sample Size:Detection	7.55	4.00	0.1096
Sample Size:LOD diversity	191.25	2.00	$< .0001$
Detection:LOD diversity	9.34	2.00	0.0094
Num. Genes:Sample Size:Detection	0.26	4.00	0.9925
Num. Genes:Sample Size:LOD diversity	1.73	2.00	0.4219
Num. Genes:Detection:LOD diversity	32.05	2.00	$< .0001$
Sample Size:Detection:LOD diversity	1.65	4.00	0.799
Num. Genes:Sample Size:Detection:LOD diversity	1.63	4.00	0.8035

Table S25: RMF.OT

	Chisq	Df	Pr(>Chisq)
Num. Genes	163.81	1.00	$< .0001$
Sample Size	53.20	2.00	$< .0001$
Detection	661.33	2.00	$< .0001$
LOD diversity	100.78	1.00	$< .0001$
Num. Genes:Sample Size	170.47	2.00	$< .0001$
Num. Genes:Detection	654.22	2.00	$< .0001$
Num. Genes:LOD diversity	5.37	1.00	0.0205
Sample Size:Detection	691.21	4.00	$< .0001$
Sample Size:LOD diversity	750.44	2.00	$< .0001$
Detection:LOD diversity	972.35	2.00	$< .0001$
Num. Genes:Sample Size:Detection	48.16	4.00	$< .0001$
Num. Genes:Sample Size:LOD diversity	24.57	2.00	$< .0001$
Num. Genes:Detection:LOD diversity	42.41	2.00	$< .0001$
Sample Size:Detection:LOD diversity	116.19	4.00	$< .0001$
Num. Genes:Sample Size:Detection:LOD diversity	53.68	4.00	$< .0001$

Table S26: Represent..CAPRI_AIC

	Chisq	Df	Pr(>Chisq)
Num. Genes	87.57	1.00	$< .0001$
Sample Size	35.06	2.00	$< .0001$
Detection	329.14	2.00	$< .0001$
LOD diversity	88.97	1.00	$< .0001$
Num. Genes:Sample Size	9.80	2.00	0.0074
Num. Genes:Detection	309.31	2.00	$< .0001$
Num. Genes:LOD diversity	0.03	1.00	0.8725
Sample Size:Detection	272.21	4.00	$< .0001$
Sample Size:LOD diversity	87.34	2.00	$< .0001$
Detection:LOD diversity	437.56	2.00	$< .0001$
Num. Genes:Sample Size:Detection	63.36	4.00	$< .0001$
Num. Genes:Sample Size:LOD diversity	1.44	2.00	0.4879
Num. Genes:Detection:LOD diversity	4.16	2.00	0.1248
Sample Size:Detection:LOD diversity	73.43	4.00	$< .0001$
Num. Genes:Sample Size:Detection:LOD diversity	2.63	4.00	0.6213

Table S27: Local Peaks.CAPRI_AIC

	Chisq	Df	Pr(>Chisq)
Num. Genes	33.91	1.00	$< .0001$
Sample Size	165.45	2.00	$< .0001$
Detection	50.85	2.00	$< .0001$
LOD diversity	65.50	1.00	$< .0001$
Num. Genes:Sample Size	34.42	2.00	$< .0001$
Num. Genes:Detection	2.30	2.00	0.3163
Num. Genes:LOD diversity	1.25	1.00	0.2641
Sample Size:Detection	15.66	4.00	0.0035
Sample Size:LOD diversity	155.70	2.00	$< .0001$
Detection:LOD diversity	41.21	2.00	$< .0001$
Num. Genes:Sample Size:Detection	1.76	4.00	0.7805
Num. Genes:Sample Size:LOD diversity	9.35	2.00	0.0093
Num. Genes:Detection:LOD diversity	3.06	2.00	0.2169
Sample Size:Detection:LOD diversity	2.52	4.00	0.6414
Num. Genes:Sample Size:Detection:LOD diversity	0.83	4.00	0.9342

Table S28: RMF.CAPRI_AIC

	Chisq	Df	Pr(>Chisq)
Num. Genes	109.99	1.00	< .0001
Sample Size	1798.63	2.00	< .0001
Detection	2781.42	2.00	< .0001
LOD diversity	4.29	1.00	0.0384
Num. Genes:Sample Size	292.35	2.00	< .0001
Num. Genes:Detection	633.77	2.00	< .0001
Num. Genes:LOD diversity	13.30	1.00	3e-04
Sample Size:Detection	577.47	4.00	< .0001
Sample Size:LOD diversity	383.39	2.00	< .0001
Detection:LOD diversity	647.75	2.00	< .0001
Num. Genes:Sample Size:Detection	91.32	4.00	< .0001
Num. Genes:Sample Size:LOD diversity	48.42	2.00	< .0001
Num. Genes:Detection:LOD diversity	16.31	2.00	3e-04
Sample Size:Detection:LOD diversity	97.29	4.00	< .0001
Num. Genes:Sample Size:Detection:LOD diversity	98.90	4.00	< .0001

Table S29: Represent..CBN

	Chisq	Df	Pr(>Chisq)
Num. Genes	46.47	1.00	< .0001
Sample Size	326.11	2.00	< .0001
Detection	762.69	2.00	< .0001
LOD diversity	74.96	1.00	< .0001
Num. Genes:Sample Size	35.59	2.00	< .0001
Num. Genes:Detection	377.15	2.00	< .0001
Num. Genes:LOD diversity	0.43	1.00	0.5135
Sample Size:Detection	315.03	4.00	< .0001
Sample Size:LOD diversity	21.87	2.00	< .0001
Detection:LOD diversity	296.12	2.00	< .0001
Num. Genes:Sample Size:Detection	213.53	4.00	< .0001
Num. Genes:Sample Size:LOD diversity	2.25	2.00	0.3248
Num. Genes:Detection:LOD diversity	20.79	2.00	< .0001
Sample Size:Detection:LOD diversity	233.91	4.00	< .0001
Num. Genes:Sample Size:Detection:LOD diversity	4.47	4.00	0.3461

Table S30: Local Peaks.CBN

	Chisq	Df	Pr(>Chisq)
Num. Genes	3.85	1.00	0.0499
Sample Size	107.17	2.00	$< .0001$
Detection	332.57	2.00	$< .0001$
LOD diversity	143.99	1.00	$< .0001$
Num. Genes:Sample Size	22.27	2.00	$< .0001$
Num. Genes:Detection	4.25	2.00	0.1195
Num. Genes:LOD diversity	1.16	1.00	0.2811
Sample Size:Detection	4.19	4.00	0.3803
Sample Size:LOD diversity	182.19	2.00	$< .0001$
Detection:LOD diversity	1.29	2.00	0.5253
Num. Genes:Sample Size:Detection	0.42	4.00	0.9807
Num. Genes:Sample Size:LOD diversity	0.44	2.00	0.8014
Num. Genes:Detection:LOD diversity	24.00	2.00	$< .0001$
Sample Size:Detection:LOD diversity	8.52	4.00	0.0744
Num. Genes:Sample Size:Detection:LOD diversity	1.49	4.00	0.8284

Table S31: RMF.CBN

18. Number of mutations of local maxima and performance

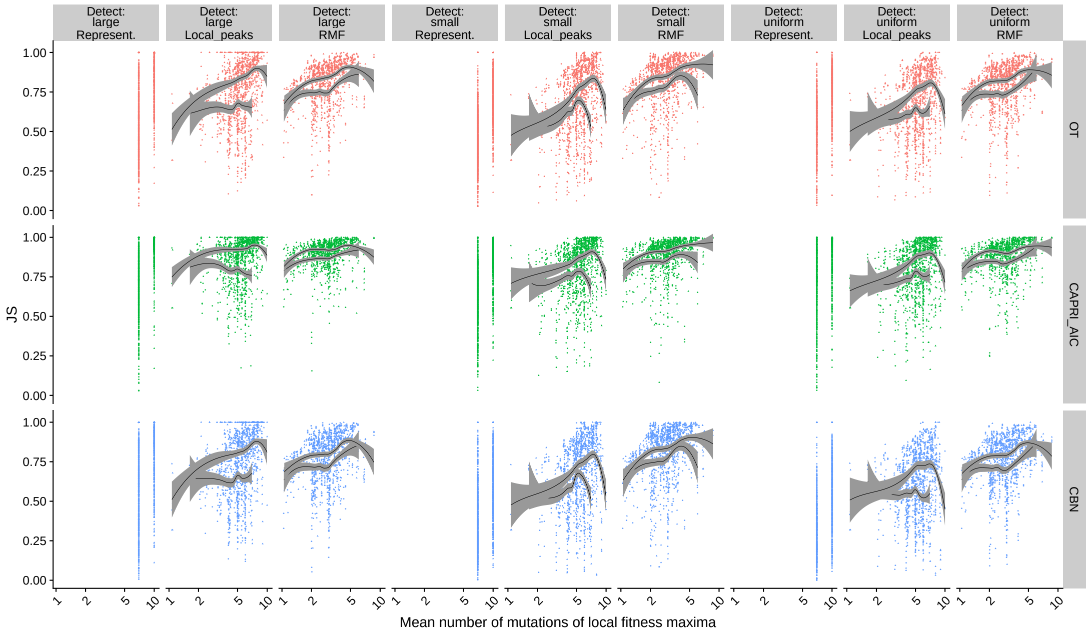


Figure S29: Mean number of mutations of local maxima and JS

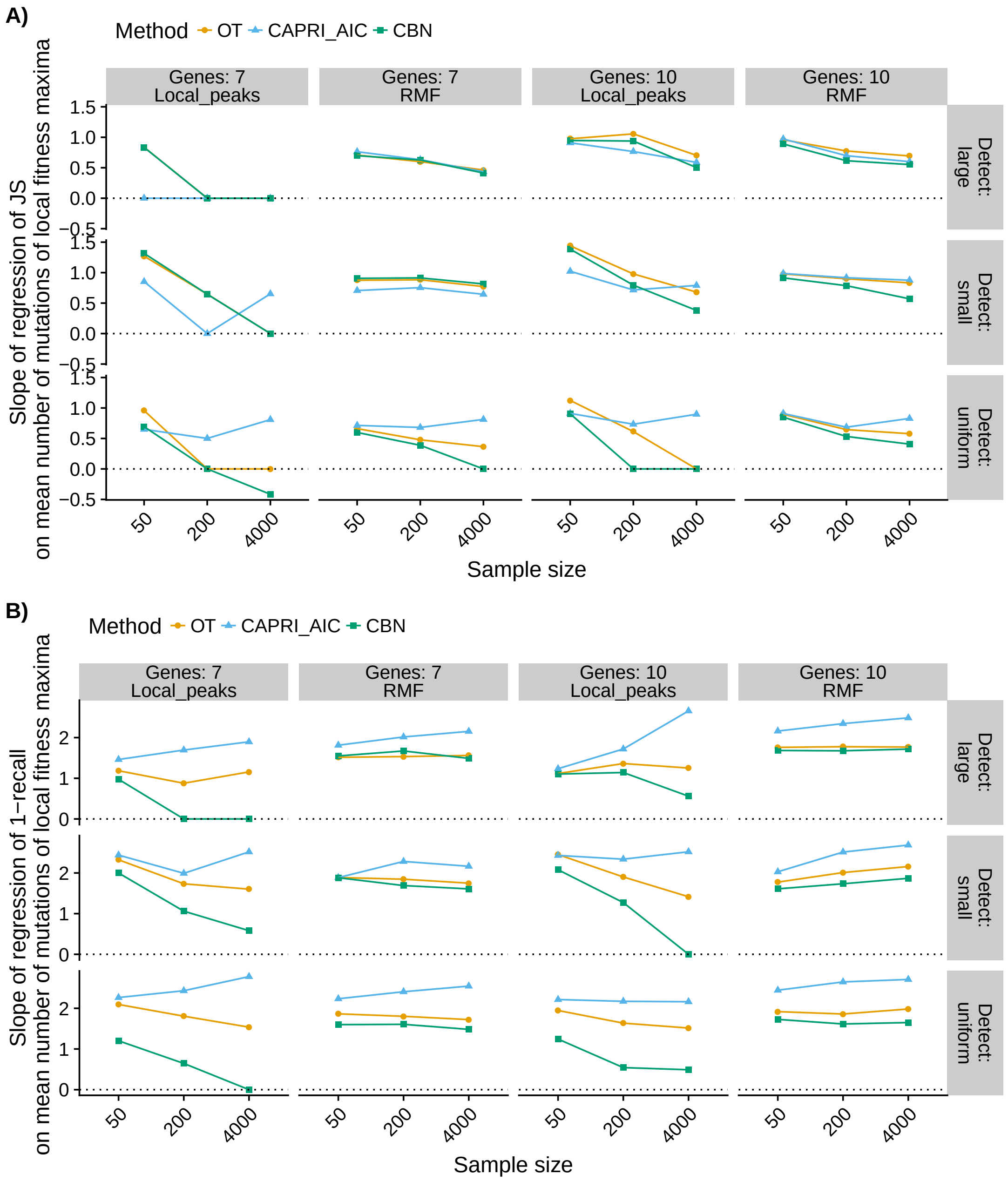


Figure S30: Slopes of regression of JS and 1-recall on mean number of mutations of local maxima.

19. LOD and CPM diversity: ratios and slopes

The following R code will, via a simple example, show that it is easy to have data where the average of the ratios is larger than one whereas the slope of the regression is negative:

```
a <- 10
n <- 100
sd <- 0.5
x <- runif(n, min = 1, max = 5)
y <- -1 * x + a + rnorm(n, mean = 0, sd = sd)
plot(y ~ x)
summary(lm(y ~ x))
mean(y/x)
```

20. Cancer data sets: additional results, figures

20.0.1. Cancer data sets: distribution of number of mutations per subject

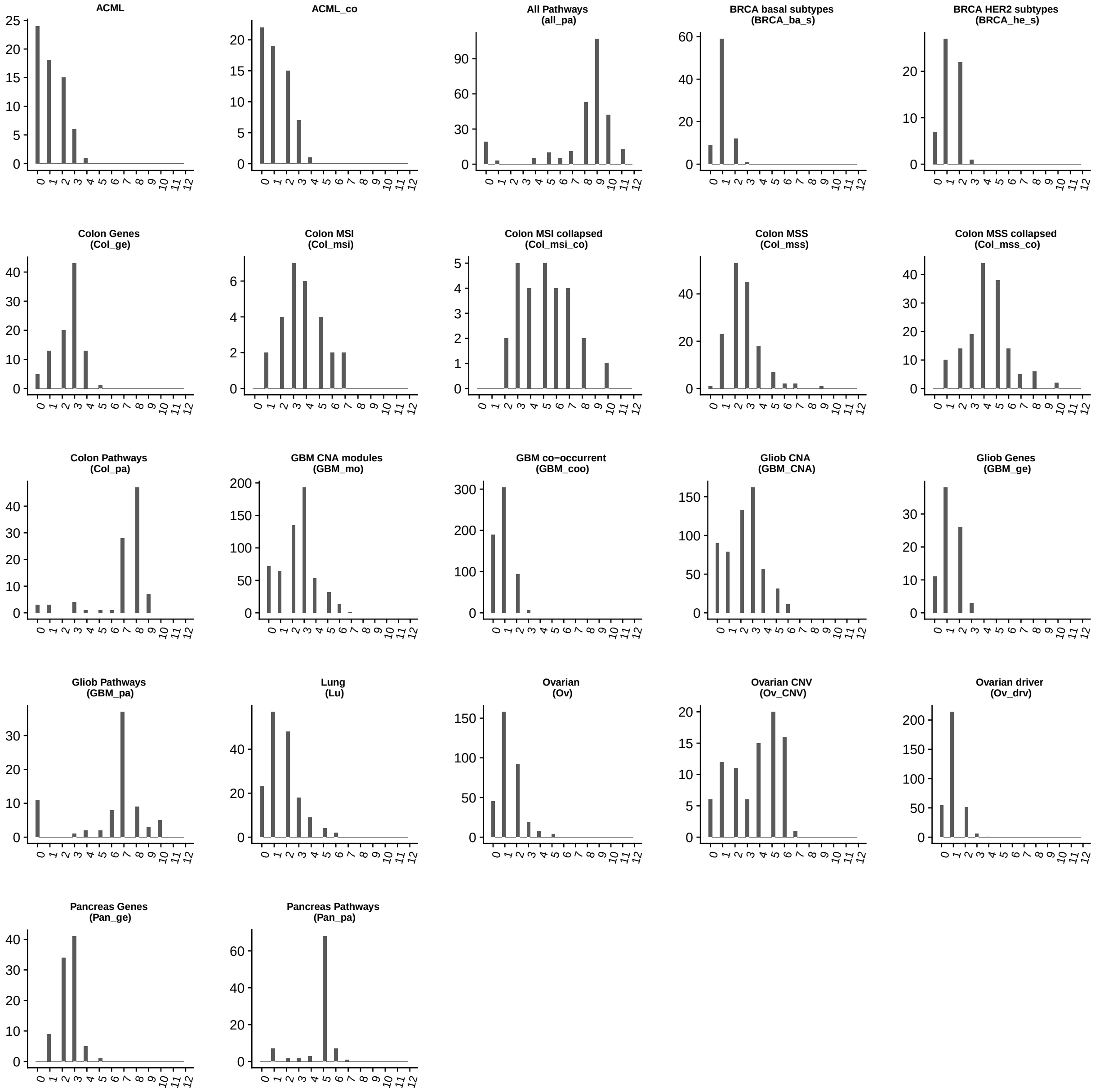


Figure S31: Cancer data sets: Histograms of number of mutations per subject in the data sets.

20.0.2. Cancer data sets: proportion of individuals in which a mutation is present

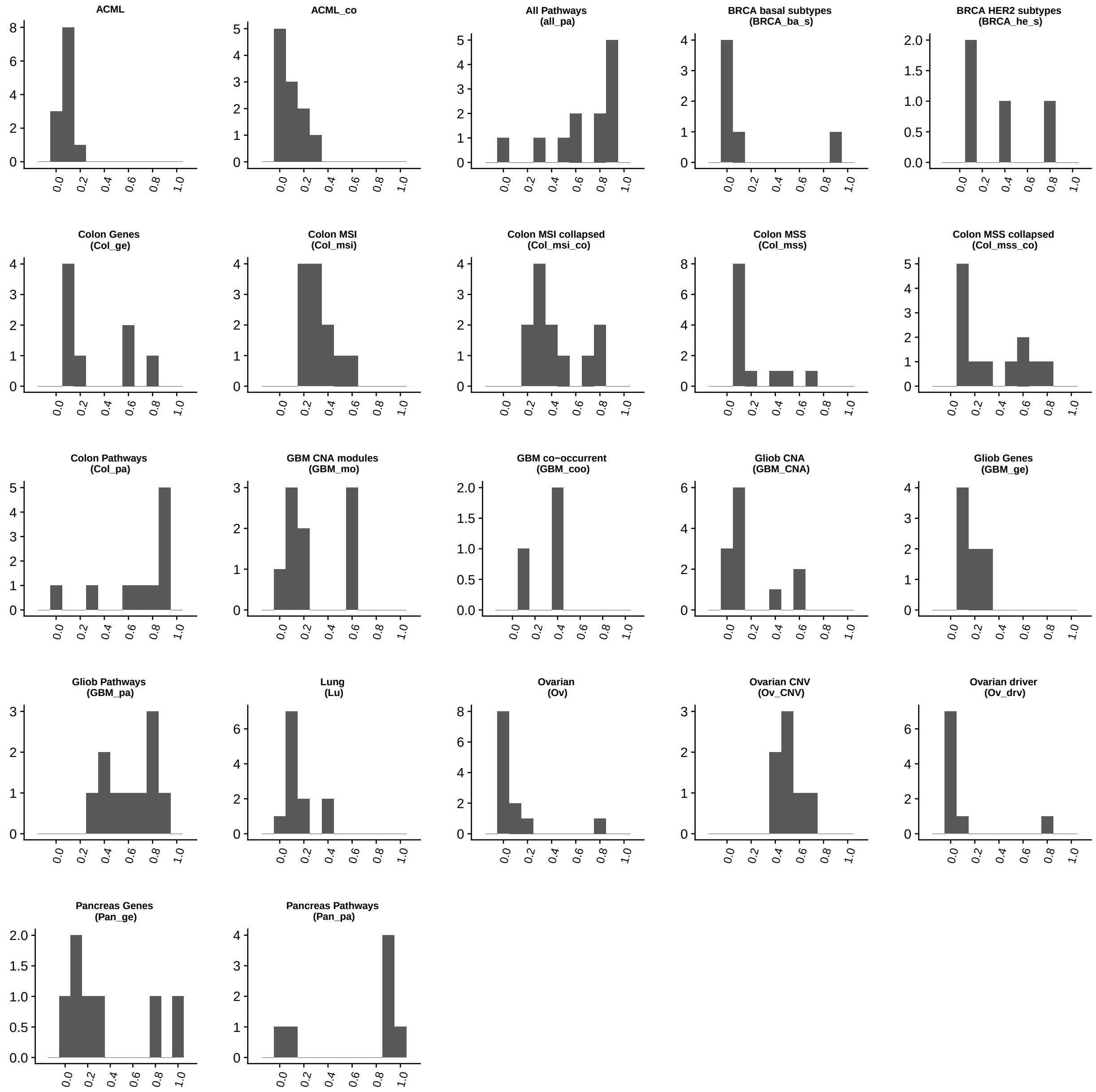


Figure S32: Cancer data sets: Histograms of proportion of individuals in which each mutation is present. For example, in the PP data set, there are four mutations that are present in 80% to 90% of the individuals in the data set, 1 mutation present in 90% to 100% of the individuals, 1 mutation in between 0 and 10% of the individuals, and 1 in between 10% and 15%.

20.0.3. Cancer data sets: scatterplots of $JS_{o,b}$, S_c, and number of paths to the maximum

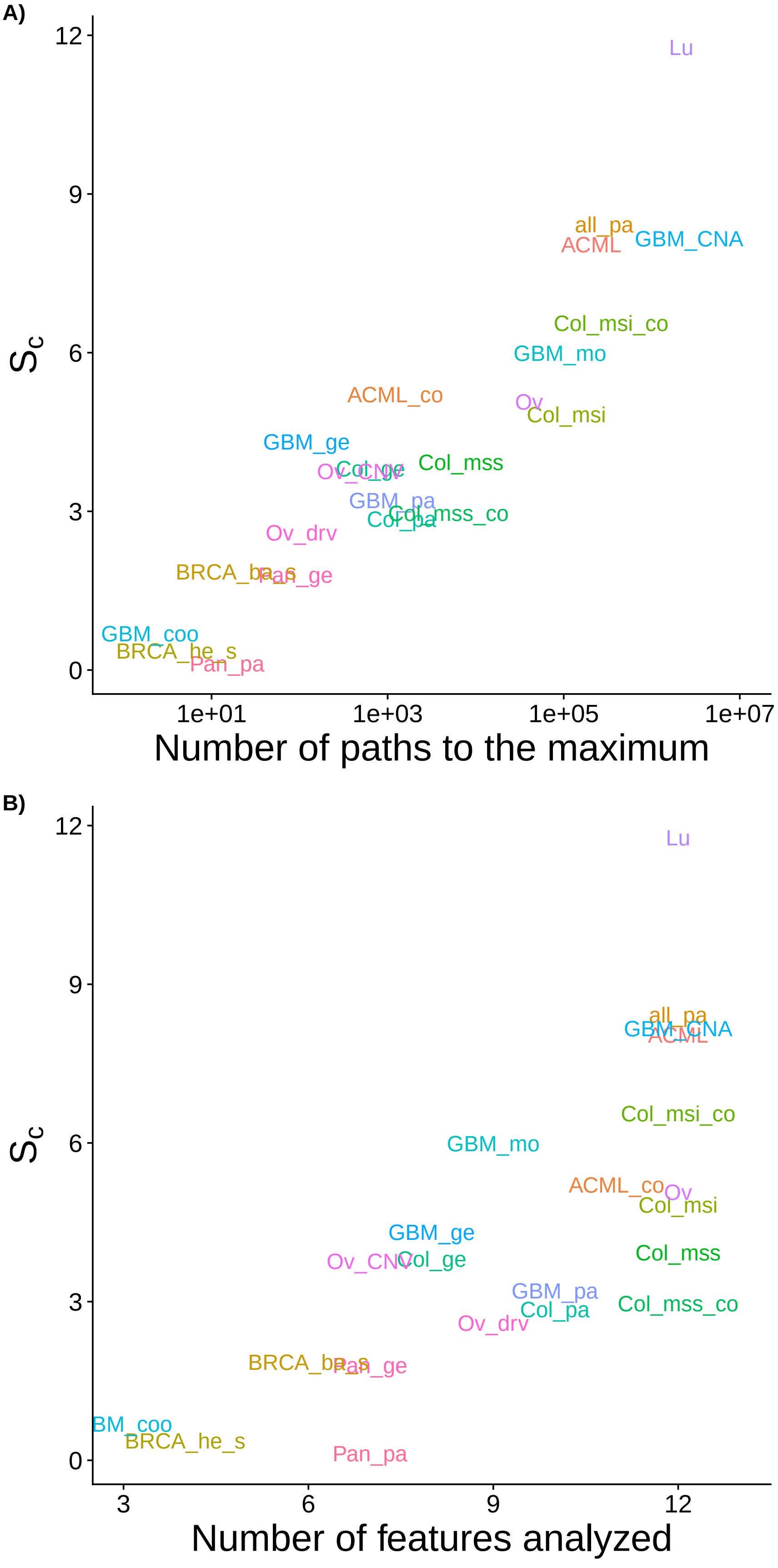


Figure S33: Cancer data sets: catterplots of the relationship between $JS_{o,b}$, S_c, and number of paths to the maximum, using the data labels, using the statistics from analyses with 12 features.

21. Data and code availability

All data for this article, along with source code, is available from file SupplMat_Code_and_Data.zip.

22. References

[1] a. Burrell, R., McGranahan, N., Bartek, J., Swanton, C., 2013. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, **501(7467)**:338–345. doi: 10.1038/nature12625. URL http://www.nature.com/doifinder/10.1038/nature12625.

[2] Anderson, W. F., Rosenberg, P. S., Prat, A., Perou, C. M., Sherman, M. E., 2014. How many etiological Subtypes of Breast Cancer:two,three, Four, or more? *JNCI: Journal of the National Cancer Institute*, **106(8)**:1–11.

[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] 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/.

[5] 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.

[6] Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., Bolker, B. M., 2017. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, **9(2)**:378–400. URL https://journal.r-project.org/archive/2017/RJ-2017-066/index.html.

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

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

[9] 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.

[10] Cancer Genome Atlas Research Network, 2012. Comprehensive molecular portraits of human breast tumours. *Nature*, **490(7418)**:61–70.

[11] 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.

[12] Cerami, E., Demir, E., Schultz, N., Taylor, B. S., Sander, C., 2010. Automated Network Analysis Identifies Core Pathways in Glioblastoma. *PLoS ONE*, **5(2)**:e8918.

[13] Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., Schultz, N., 2012. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1. *Cancer Discovery*, **2(5)**:401–404.

[14] 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.

[15] Crona, K., Greene, D., Barlow, M., 2013. The peaks and geometry of fitness landscapes. *Journal of Theoretical Biology*, **317**:1–10. doi:10.1016/j.jtbi.2012.09.028. URL http://www.sciencedirect.com/science/article/pii/S0022519312005061.

[16] De Sano, L., Caravagna, G., Ramazzotti, D., Graudenzi, A., Mauri, G., Mishra, B., Antoniotti, M., 2016. TRONCO: An R package for the inference of cancer progression models from heterogeneous genomic data. *Bioinformatics*, **32(12)**:1911–1913. doi:10.1093/bioinformatics/btw035.

[17] Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., Mooney, T. B., Callaway, M. B., Dooling, D., Mardis, E. R., Wilson, R. K., Ding, L., 2012. MuSiC: Identifying mutational significance in cancer genomes. *Genome Research*, **22(8)**:1589–1598.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Farahani, H. S., Lagergren, J., 2013. Learning oncogenetic networks by reducing to mixed integer linear programming. *PloS ONE*, **8(6)**:e65773. doi:10.1371/journal.pone.0065773. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3683041&tool=pmcentrez&rendertype=abstract.

[23] 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.

[24] Fox, J., Weisberg, S., 2011. *An R Companion to Applied Regression, 2nd Ed*. Sage, Thousand Oaks, CA.

[25] Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., Schultz, N., 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, **6(269)**:pl1. doi:10.1126/scisignal.2004088.

[26] 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.0027136%0020http://www.bsse.ethz.ch/cbg/software/ct-cbn.

[27] Hosseini, S.-R., 2018. Quantifying the predictability of cancer progression using Conjunctive Bayesian Networks. M.Sc. Thesis, Swiss Federal Institute of Technology, Zürich.

[28] Jacobsen, A., Questions, c., 2018. Cgdsr: R-Based API for Accessing the MSKCC Cancer Genomics Data Server (CGDS). URL https://CRAN.R-project.org/package=cgdsr.

[29] 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.

[30] 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.

[31] Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., Carey, V. J., 2013. Software for computing and annotating genomic ranges. *PLoS computational biology*, **9(8)**:e1003118. doi:10.1371/journal.pcbi.1003118. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3738458&tool=pmcentrez&rendertype=abstract.

[32] McFarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R., Mirny, L. A., 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. doi:10.1073/pnas.1213968110. URL http://www.ncbi.nlm.nih.gov/pubmed/23388632.

[33] Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., Getz, G., 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, **12(4)**:R41.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] Sakoparnig, T., Beerenwinkel, N., 2012. Efficient sampling for Bayesian inference of conjunctive Bayesian networks. *Bioinformatics (Oxford, England)*, **28(18)**:2318–24. doi:10.1093/bioinformatics/bts433. URL http://www.ncbi.nlm.nih.gov/pubmed/22782551%0020http://www.bsse.ethz.ch/cbg/software/bayes-cbn.

[41] Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S. D., Willis, J., Dawson, D., Willson, J. K. V., Gazdar, A. F., Hartigan, J., Wu, L., Liu, C., Parmigiani, G., Park, B. H., Bachman, K. E., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., Velculescu, V. E., 2006. The Consensus Coding Sequences of Human Breast and Colorectal Cancers. *Science*, **314(5797)**:268–274. doi:10.1126/science.1133427. URL http://dx.doi.org/10.1126/science.1133427.

[42] Szabo, A., Pappas, L., 2013. Oncotree: Estimating oncogenetic trees. R package version 0.3.3. URL http://cran.r-project.org/package=Oncotree.

[43] Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., Karchin, R., 2016. Evaluating the evaluation of cancer driver genes. *PNAS*, **113(50)**:14330–14335. doi:10.1073/pnas.1616440113. URL http://www.pnas.org/content/113/50/14330.

[44] 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.