Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Supplementary Material: Estimating the predictability of cancer evolution

Sayed-Rzgar Hosseini [1,2], Ramon Diaz-Uriarte [3], Florian Markowetz [2] and Niko Beerenwinkel [1,4,*]

[1]Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland
[2]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, United Kingdom
[3]Department of Biochemistry, Universidad Autónoma de Madrid, Instituto de Investigaciones Biomédicas "Alberto Sols" (UAM-CSIC), Madrid, Spain and
[4]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

This Supplementary Materials file includes 10 Supplementary texts (S1-S10), 2 supplementary tables (S1-S2) and 20 Supplementary figures (S1-S20).

1 Supplementary Text S1:

Based on an evolutionary model (McFarland *et al.*, 2013) implemented in the OncoSimulR package (Diaz-Uriarte, 2017), from each fitness landscape, 20000 genotypes have been generated (Diaz-Uriarte, 2018). Each genotype represents the final outcome of an evolutionary scenario applied to an initial population of wild-type cells, in which each individual cell can divide, acquire new mutations or die. The population size is not kept constant and it affects the birth and death rates; in addition, the birth rate strongly depends on the fitness effect of the mutations (McFarland *et al.*, 2013). The population grows until the cancer is detected, and the cancer detection probability is a function of population size as: $P(N) = 1 - e^{(-\delta \cdot (N-B))}$, if $N > B$ and 0 otherwise, where N is the population size, B is the minimum population size in which cancer can be detected, and δ controls the increase in $P(N)$ with population size, and determines the average number of evolutionary steps required until the cancer is detected.

After detection of tumor in the population, as in (Diaz-Uriarte, 2015), the final genotype, which is the final outcome of the simulation, is determined based on the genotypes of all the cells in the population. More precisely, a given gene was considered mutated if it was mutated in 50% of the cells in the populations. For each fitness landscape, 20000 independent simulations have been carried out to generate 20000 genotypes.

Moreover, for each fitness landscape, simulations using four different parameter sets (combination of two different mutation rates and two different detection rates) has been done (Diaz-Uriarte, 2018). Two different δ values $7.526 \cdot 10^{-5}$ and $7.179 \cdot 10^{-6}$ that correspond to detection probabilities of 0.1 and 0.01 respectively, when the population size has doubled, and two different mutation rates (constant: 10^{-6}, constant: 10^{-5}) were used (Diaz-Uriarte, 2018).

2 Supplementary Text S2:

In this analysis, we aimed to go beyond comparison of the final predictabilities ($\phi_{\preceq}$ and ϕ_w), and we wanted to precisely check whether the mutational pathway probability distributions inferred by the two approaches ($P_w(\pi)$, (Eq. 7 in the main text)) and ($P_{\preceq}(\pi)$, (Eq. 10 in the main text)) are also in good agreement or not. To achieve this goal, first, for any of the 211 fitness landscapes and in each simulation conditions, we quantified the Jensen-Shannon divergence (Crooks, 2017; Lin, 1991), scaled between 0 and 1 —equivalent to using the logarithm of base 2— between the two probability distributions. As figure 3 in the main text indicates, in simulation conditions with low mutation rate and in slow detection regimes, the divergence between the two distributions is below 0.25 (with median 0.045 in representable and median 0.146 in non-representable fitness landscapes); however, as expected, in simulation conditions, which deviate from the SSWM assumption (i.e. high mutation rates and fast detection regimes), we observe considerably higher Jensen-Shannon divergence between the two pathway probability distributions (figure S2).

Next, we checked whether the two approaches agree well in detecting feasible mutational pathways (Π, Eq. 6 in the main text) or ($\Pi_{\preceq}$, Eq. 9 in the main text). Figure S3 shows that in all simulation conditions and in both representable and non-representable fitness landscapes, the fraction of feasible pathways based on CBN approach is significantly higher than that of the fitness landscape based approach under SSWM assumption. Thus, we conclude that CBN is milder than the SSWM assumption, as it imposes fewer restrictions on the mutational pathways than SSWM. However, figures S4 and S5 show that the two approaches are correlated

in terms of the fraction of feasible mutational pathways on both types of fitness landscapes. The correlation is highest in simulation conditions with low mutation rates and slow detection regimes, and it becomes weaker in simulation conditions, which deviate from the SSWM assumption (i.e. high mutation rate or fast detection regimes).

Finally, for each fitness landscape (and simulation conditions), we partitioned the 7! mutational pathways into four categories as follows:
i) True positive (TP): the mutational pathways, which are identified as feasible, using both CBN and fitness landscape-based approaches. ii) True negative (TN): the mutational pathways, which are identified as infeasible, using both CBN and fitness landscape-based approaches. iii) False positive (FP): the mutational pathways, which are identified as feasible using CBN, but as infeasible using the fitness landscape-based approach. iv) False negative (FN): the mutational pathways, which are identified as infeasible using CBN, but as feasible using the fitness landscape-based approach. We used these values to quantify the accuracy, precision, and recall as follows:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{2}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{3}$$

Figures S6 and S7 show that the accuracy is almost equal to one in all conditions and fitness landscapes. Moreover, recall is generally high, especially in simulation conditions under SSWM assumption (i.e. lower mutation rates, slow detection regime) and in representable fitness landscapes. This implies that CBN has a low false negative rate and so it correctly identifies infeasible mutational pathways. However, precision especially in rugged non-representable fitness landscapes is not high, because as we mentioned above (figure S3) CBN identifies higher number of feasible mutational pathways than the fitness landscape approach under the SSWM model, which results in higher false positive rate.

Thus, we conclude that CBN and fitness landscape based approaches agree well in terms of determining the pathway probability distributions. The only noticeable difference between the two methods is caused by the fact that CBN is milder and less restrictive than SSWM assumption in identifying the feasible mutational pathways, which results in high false positive rate and relatively lower precision. Nevertheless, the final pathway probability distribution ($P_{\preceq}(\pi)$) and the final quantified predictability ($\phi_{\preceq}$) are robust to this discrepancy, because $P_w(\pi)$ and $P_{\preceq}(\pi)$ show very low Jensen-Shannon divergence, and $\phi_{\preceq}$ strongly correlates with ϕ_w.

3 Supplementary Text S3:

To check the impact of sample size (N) on the variability of the inferred $\phi_{\preceq}$, we arbitrarily chose two fitness landscapes, a representable and a non-representable one. From each fitness landscape, we chose randomly i) $N = 10,000$ ii) $N = 1000$, and iii) $N = 100$ genotypes among the original $20,000$ genotypes. Next, from each of the above 3×2 new samples, we generated 100 bootstrap samples by sampling with replacement of the same number (N) of genotypes. Then, for each of the resulting 6×100 bootstrap samples we run CBN model with $n = 7$ genes to quantify $\phi_{\preceq}$. Figure S8 shows the distribution of $\phi_{\preceq}$ among the 6 bootstrap samples, each corresponding to a given fitness landscape and a given sample size N. As we observe, in smaller sample sizes, especially in $N = 100$, the variance of $\phi_{\preceq}$ is clearly higher than that of the larger sample sizes. Thus, we conclude that in lower sample sizes, the variability of the inferred $\phi_{\preceq}$ is alarmingly high.

4 Supplementary Text S4:

In order to investigate the decomposability of predictability and hence the accuracy of equation (12) in the main text, in this analysis, we aim to model the probability distribution of the mutational pathways using a single parameter, which solely determines the predictability of evolution. To this end, we assume an ideal scenario as follows:

For a sequence of n mutations arranged in ascending order of their waiting time or equivalently, in descending order of their corresponding λ, we assume the relative waiting time of consecutive mutations is constant and is equal to κ:

$$\kappa = \frac{\lambda_{i+1}}{\lambda_i} \forall i \in \{1, \cdots, n\}, \text{where} \quad 0 \leq \kappa \leq 1 \tag{4}$$

Furthermore, we assume that all $n!$ mutational pathways are feasible. However, the probability of different pathways can be widely different depending on κ. Under the special case, where $\kappa = 1$, the associated λ of all n mutations are equal, and so the pathway probability distribution will be the same as the uniform distribution and thus $\phi = 0$.
However, when $\kappa \neq 1$, the ordered sequence of λ's generates a geometric series as follows:

$$\lambda_i = \lambda_1 \kappa^{(i-1)}, \forall i \in \{1, \cdots, n\} \tag{5}$$

Thus, according to the equation (10) in the main text, the probability of a given pathway (π) is defined as follows:

$$P(\pi) = \prod_{i=1}^{n} \frac{\kappa^{(\pi_i - 1)}}{\sum_{h \in \text{Exit}(g(\pi)_i)} \kappa^{(h \setminus g(\pi)_i - 1)}} \tag{6}$$

Since in this analysis we have assumed that all mutational pathways ($\pi \in S_n$) are feasible, the denominator of the above equation can be expressed more simply as follows:

$$P(\pi) = \prod_{i=1}^{n} \frac{\kappa^{(\pi_i - 1)}}{\sum_{j \in \Omega(\pi)} \kappa^{(j-1)}} \tag{7}$$

Where: $\Omega(\pi) = \{\{1, \cdots, n\} \setminus \{\pi_1, \cdots, \pi_{i-1}\}\}$ The above equation does not need to be normalized, because the following equality always holds:

$$\sum_{\pi \in S_n} \prod_{i=1}^{n} \frac{\kappa^{(\pi_i - 1)}}{\sum_{j \in \Omega(\pi)} \kappa^{(j-1)}} = 1 \tag{8}$$

Using this probability distribution, and based on the normalized entropy (equations 3 and 4 in the main text), we quantified predictability (ϕ) as a function of κ for various n. Figure S9 shows ϕ as a function of κ for $3 \leq n \leq 10$.

Next, we aimed to check the accuracy of the equation (12) in the main text, which claims that the predictability based on a set of n genes is approximated by the mean of the associated predictability of its $\binom{n}{n'}$ subsets, each including n' genes. Figure S10 shows that in all pairs of n and n', where $n' \leq n$, the deviation of the approximated ϕ from the exact ϕ is negligible, as all the scatter plots of $\phi_{\preceq}$(approximated) versus $\phi_{\preceq}$(exact) are dramatically close to the identity line. Thus, we conclude that in an ideal scenario, where the ratio of the λ of consecutive mutations stays constant along a sequence of mutations arranged in ascending order of the waiting time of their arrival (i.e. $\frac{\lambda_{i+1}}{\lambda_i} = \kappa, \forall i \in \{1, \cdots, n\}$), we observe that the approximated ϕ based on the equation (12) in the main text, is almost exactly the same as the exact ϕ.

However, in real scenarios, $\frac{\lambda_{i+1}}{\lambda_i}$ may not stay constant. Moreover, the CBN-based estimated λ using a subset including n' genes may not be exactly the same as the original CBN-based estimated λ using all n genes. Furthermore, in the ideal scenario above, all $n!$ pathways are assumed to be

feasible, but in real scenarios, majority of the pathways are unfeasible and have $P(\pi) = 0$, which might cause further deviation of the approximated ϕ from the exact ϕ. Therefore, to systematically check the validity of the above approximation, we need to use the CBN-based estimated pathway probability distributions (equation (10) in the main text) both for simulated and real datasets (See texts S5 and S6 below).

5 Supplementary Text S5:

In this analysis, we aimed to further check the validity of the approximation that we proposed in equation (12) of the main text using our simulated data. For this purpose, for each of the 100 representable and 111 non-representable fitness landscapes, we re-quantified the predictability of cancer progression based on the average predictability (see equation (12) in the main text) among i) all $\binom{7}{2}$ possible doublets (pairs) of genes ($n' = 2$), ii) all $\binom{7}{3}$ possible triplets ($n' = 3$), iii) all $\binom{7}{4}$ possible quartets ($n' = 4$), and iv) all $\binom{7}{5}$ possible quintets ($n' = 5$). In this analysis, we used the simulation data with mutation rate of 10^{-6} and slow detection rate.

Figure S11 shows that the average predictability among the quartets ($n' = 4$) and quintets ($n' = 5$) approximates the predictability using the full CBN (with $n = 7$ genes) very well. Moreover, even the average predictability of triplets ($n' = 3$) strongly correlates with the predictability of full CBN. Thus, we conclude that the CBN-based predictability is approximately decomposable, meaning that we can approximate the predictability of a given CBN with a given set ($\mathcal{E}$) of n genes based on the predictability of smaller CBNs including only a subset ($\mathcal{E}'$) including $n' < n$ genes.

6 Supplementary Text S6:

In this analysis, we aimed to further check the validity of the approximation that we proposed in equation (12) of the main text using real genomic data. For this purpose, for each of the 15 cancer types in both TCGA and MSK datasets, we first quantified $\phi_{\preceq}$ based on the $n = 7$ most frequently mutated driver genes in each cancer type (full CBN). Then, we aimed to approximate $\phi_{\preceq}$ based on equation (12) in the main text (subsets CBN), by quantifying the average $\phi_{\preceq}$ among, i) all $\binom{7}{3}$ possible triplets ($n' = 3$), ii) all $\binom{7}{4}$ possible quartets ($n' = 4$), and iii) all $\binom{7}{5}$ possible quintets ($n' = 5$).

Figure S12 shows that the approximation strongly holds in both datasets and in all three values of n'. Thus, we conclude that even in real data, where sample size is substantially smaller than the simulated data, the predictability of a given CBN with a given set of n genes ($\mathcal{E}$) can still be reliably approximated based on the predictability of smaller CBNs including only a subset ($\mathcal{E}'$) of $n' < n$ genes.

However, so far we have only checked this approximation, when the full CBN includes seven genes ($n = 7$). This raises the concern that the approximation might be valid only when n' and n are close enough to each other. To check whether the approximation is valid also for larger n, we examined this approximation in $n = 10$. For this purpose, for each of the 15 cancer types in both TCGA and MSK datasets, we calculated the average $\phi_{\preceq}$ among i) all $\binom{10}{3}$ possible triplets ($n' = 3$), ii) all $\binom{10}{4}$ possible quartets ($n' = 4$), iii) all $\binom{10}{5}$ possible quintets ($n' = 5$), iv) all $\binom{10}{6}$ possible sextets ($n' = 6$), and v) all $\binom{10}{7}$ possible septets ($n' = 7$).

Figure S13 shows that in both datasets and in all cancer types, the approximated $\phi_{\preceq}$ converges to almost the same value from $n' = 4$ until $n' = 7$. Therefore, convergence of the approximated $\phi_{\preceq}$ in $n' \geq 4$ is not peculiarity of $n = 7$ but is also valid in larger $n = 10$, which is considerably far away from $n' = 4$. In other words, the approximation holds even if n' and n are not very close to each other. This feature makes the inference step of the CBN model remarkably more efficient, because to find maximum likelihood DAG of restrictions it would suffice to explore a dramatically smaller space of potential DAGs.

7 Supplementary Text S7:

In this analysis, we aimed to compare the stability of $\phi_{\preceq}$ estimated based on the approximation approach (equation (12) in the main text) with the exact method. For this purpose, we used real datasets and for each of the 15 cancer types in both datasets, we defined genotypes using the $n = 7$ most frequently mutated driver genes. While in the exact method, we used all $n = 7$ genes to quantify $\phi_{\preceq}$, in the approximation method we used average of $\phi_{\preceq}$ over all $\binom{7}{4}$ possible quartets ($n' = 4$) of genes. To measure the stability of $\phi_{\preceq}$ in both methods, for each cancer type in each dataset, we generated 100 bootstrap samples by sampling with replacement the same number of genotypes from the corresponding original pool of genotypes.

Figure S14 shows that in all cancer types and in both datasets, the 95% bootstrap confidence interval for the exact method is considerably wider than that of the approximate approach. This excess instability of the exact method stems from the fact that the structure-learning step of the CBN model relies on simulated annealing, which is a stochastic approach. While in the approximate method, the space of possible structures to be explored contains only $3^{\binom{4}{2}} = 729$ distinct DAGs of restrictions, in the exact method a vast space including $3^{\binom{10}{2}} > 10^{21}$ possible DAGs needs to be explored to find the maximum likelihood DAG of restrictions, and hence the chance to deviate from the optimal DAG is much higher in the exact method, which culminates in a higher variability and a much wider confidence interval for the estimated $\phi_{\preceq}$.

8 Supplementary Text S8:

In this analysis, we used the approximate approach (equation (12) in the main text) with $n = 10$ and $n' = 4$ to quantify $\phi_{\preceq}$ for the 15 different cancer types in both datasets. Then, we subjected each cancer type to 10 leave-one-out events in each of which one of the $n = 10$ genes are removed and then the $\phi_{\preceq}$ is re-estimated based on $n = 9$ and $n' = 4$. Finally, we calculated the difference ($\Delta\phi_{\preceq}$) between the estimated predictability before and after removal of the i^{th} frequently mutated driver gene. Figure S17 shows that except for the most frequently mutated gene, whose removal slightly reduces $\phi_{\preceq}$, removal of the other nine genes has a negligible effect on the estimated predictability ($\Delta\phi_{\preceq}$<0.05). Thus, we conclude that at $n = 10$, the estimated ($\Delta\phi_{\preceq}$) is robust to the removal of a driver gene.

9 Supplementary Text S9:

In this analysis, we aim to gain an intuitive interpretation of ϕ by attempting to relate it to the fraction of feasible mutational pathways (α). In other words, by quantifying α, we aim to gain insight into the extent of the evolutionary constraint on the progression of cancer.

However, the exact inference of mutational pathways depends on the inferred DAG of restrictions (equation (9) in the main text). As we discussed in section 3.2 in the main text and supplementary text S7, the exact inference for large n is inefficient and non-robust. Moreover, the approximate solution that we offered in equation (12) in the main text is based on sub-pathways of length n' and not on the original pathways of length n. Therefore, we need to derive an alternative approximate relationship such that we can directly map a given ϕ to its corresponding α.

For this purpose, we need to make simplifying assumptions to derive such a one-to-one map. As we mentioned earlier, two factors directly affect ϕ: i) the fraction of feasible pathways and ii) the non-uniformity of

the distribution of the probability among the feasible pathways. Thus, to model ϕ by a single parameter, we need to take one of the following approximate scenarios:
i) We assume that all pathways are feasible and differ only in their probability (non-uniform pathway probability distribution) as we discussed in text S4 by using single parameter κ and ii) we assume that only a fraction (α) of pathways are feasible, but the feasible pathways are all equally probable.

We take the second assumption in this analysis, and hence define the probability of a given pathway as a function of α:

$$P(\pi) = \begin{cases} \frac{1}{\alpha n!}, & \text{if } \pi \in \Pi_{\preceq} \\ 0, & \text{if } \pi \notin \Pi_{\preceq} \end{cases} \tag{9}$$

Where $\Pi_{\preceq}$ includes the set of feasible mutational pathways as defined in equation (9) in the main text.

Under the second scenario mentioned above, and according to equations 3 and 4 in the main text, the predictability ($\phi_{\preceq}$) can be expressed as a function of α and n as follows:

$$\phi_{\preceq} = 1 - \frac{\sum_{\pi \in \Pi_{\preceq}} \frac{1}{(\alpha n!)} \log(\alpha n!)}{\log(n!)} \tag{10}$$

Since the number of feasible mutational pathways under this scenario equals $\alpha n!$, the above equation is further simplified as follows (see figure S18):

$$\phi_{\preceq} = 1 - \frac{\log(\alpha \cdot n!)}{\log(n!)} \tag{11}$$

10 Supplementary Text S10:

In this analysis, we aimed to dissect the relative contributions of the mere frequency of mutations versus the restrictions (i.e., evolutionary constraints), which originate from the interactions between mutations. For this purpose, we quantified $\phi_{\preceq}$ with $n = 10$ on an empty CBN, where there is no edge in the DAG of restrictions. In the empty CBN, mutations are considered as independent of each other and the interactions between mutated genes are neglected. In this scenario, the only factor contributing to the predictability of evolution is the frequency of mutations and how non-uniform they are. Figure S20 compares the $\phi_{\preceq}$ based on i) empty CBN and ii) (normal) CBN, for all 15 cancer types and both TCGA and MSK datasets. It reveals that $\phi_{\preceq}$ based on empty CBN is substantially lower than that of the normal CBN-based approach. Empty CBN-based $\phi_{\preceq}$ is consistently below 0.25 in majority of the cancer types and thus it explains only a minor fraction of the total predictability. Instead, the restriction in the ordering of mutations accounts for the majority of the predictability. Thus, using a simple frequency-based approach without taking the evolutionary constraints into account considerably under-estimates the evolutionary predictability of cancer. This highlights the power of CBN as a statistical graphical model in identifying the interactions between mutations, which is manifested as the edges in the inferred DAG of restrictions.

11 Supplementary Tables:

This section contains 2 tables (page 5).

12 Supplementary Figures:

This section contains 20 figures (pages 6-19).

Bibliography

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

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**(1), 41.

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

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

Lawrence, M. S. *et al.* (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**(7457), 214–218.

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

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

Raynaud, F. *et al.* (2018). Pan-cancer inference of intra-tumor heterogeneity reveals associations with different forms of genomic instability. *PLOS Genetics*, **14**(9), e1007669.

index	Cancer type	MSK	TCGA
1	Lung Adenocarcinoma	1357	588
2	Breast Invasive Ductal Carcinoma	927	836
3	Colon Adenocarcinoma	724	439
4	Prostate Adenocarcinoma	698	499
5	Pancreatic Adenocarcinoma	384	186
6	Bladder Urothelial Carcinoma	312	413
7	Glioblastoma Multiforme	286	607
8	Renal Clear Cell Carcinoma	202	538
9	Cutaneous Melanoma	195	480
10	Breast Invasive Lobular Carcinoma	190	219
11	Lung Squamous Cell Carcinoma	170	511
12	Stomach Adenocarcinoma	151	478
13	Esophageal Adenocarcinoma	112	186
14	Uterine Endometrioid Carcinoma	95	549
15	Papillary Thyroid Cancer	93	507

Table 1. (S1): Number of samples per cancer type. Each row corresponds to a given cancer type specified in the second column based on MSK data (the third column) or TCGA data (fourth column).

Cancer type index (Table S1)	Top 20 most frequently mutated driver genes
1	{TP53,KRAS,KEAP1,STK11,EGFR,NF1,ATM,BRAF,MGA,SETD2,RBM10,SMARCA4,ARID1A,ZEB1,MET,PIK3CA,ARID2,RB1,APC,CDKN2A}
2	{TP53,PIK3CA,GATA3,MAP3K1,NCOR1,PTEN,MAP2K4,SPEN,NF1,ARID1A,HRNR,RB1,CBFB,ERBB2,RUNX1,EYS,CTCF,CDH1,TBX3,SF3B1}
3	{APC,TP53,KRAS,PIK3CA,SYNE1,RYR2,CSMD3,USH2A,FBXW7,NEB,BRAF,ATM,SOX9,HMCN1,MDN1,VPS13B,ARID1A,NBEA,SMAD4,DNAH17}
4	{SPOP,TP53,FOXA1,PTEN,NBPF1,CTNNB1,KDM6A,PIK3CA,MED12,CNTNAP1,MED15,CDKN1B,IDH1,TNRC18,EHHADH,SMG7,LMOD2,NUDT11,AGAP6,EOMES}
5	{KRAS,TP53,SMAD4,CDKN2A,GNAS,RNF43,TGFBR2,RREB1,RYR2,KDM6A,MED12L,AFF2,PCLO,FCGBP,OTOF,DNAH5,KCNA4,LAMA1,KLK15,ZFHX3}
6	{TP53,ARID1A,KDM6A,PIK3CA,RB1,EP300,FGFR3,CREBBP,STAG2,ATM,FAT1,SPTAN1,CDKN1A,ELF3,ERBB2,ERBB3,ERCC2,FBXW7,TSC1,ARID2}
7	{PTEN,TP53,EGFR,NF1,PIK3R1,PIK3CA,RB1,ATRX,IDH1,STAG2,PDGFRA,CHD8,KDR,SEMG1,IL4R,TMPRSS6,BRAF,SLC26A3,RPL5,TP63}
8	{VHL,PBRM1,SETD2,MTOR,KDM5C,PTEN,ATM,TP53,PIK3CA,NEFH,NF2,NUDT11,GPR50}
9	{BRAF,NRAS,COL3A1,ALPK2,DSG3,TP53,ARID2,NF1,RPTN,KEL,CDKN2A,SLC38A4,TRERF1,PTEN,NBPF1,DDX3X,PPP6C,MAP2K1,RAC1,IDH1}
10	{CDH1,PIK3CA,RUNX1,TBX3,PTEN,TP53,FOXA1,MAP3K1,GATA3,NCOR1,SPEN,ERBB2,NF1,ARID1A,CASZ1,SF3B1,KDM6A,TGS1,CDKN1B,HRNR}
11	{TP53,CDKN2A,NFE2L2,PIK3CA,CPS1,KEAP1,PTEN,NOTCH1,RB1,CYP11B1,EP300,ASB5,IRF6}
12	{TP53,ARID1A,OBSCN,ZFHX4,PIK3CA,CUBN,COL12A1,DNAH7,ANK3,APC,VPS13B,ERBB4,HSPG2,MUC6,ATM,ZFHX3,COL11A1,SCN10A,CDKN2A,CELSR1}
13	{TP53,PIK3CA,CDKN2A,NFE2L2,DNAH10,NOTCH1,IVL,ARID1A,ERBB2,SMAD4,FBXW7,TGFBR2,ZNF750,PTCH1,PAXIP1,CORO7,RB1,LIMA1,ITGA6,IPP}
14	{PTEN,PIK3CA,ARID1A,TP53,PIK3R1,CTNNB1,CTCF,ZFHX3,KRAS,FAT1,ARHGAP35,FBXW7,ATM,BCOR,MKI67,PPP2R1A,FGFR2,SACS,ARID5B,POLE}
15	{BRAF,NRAS,HRAS,TG,EIF1AX,PPM1D,COL5A3,NUP93,KRAS,AKT1,NLRP6,ABL1,RPTN}

Table 2. (S2): Frequently mutated driver genes used for genotyping of the patients and quantifying evolutionary predictability of cancer types. Each row corresponds to a given cancer specified by the index according to table S1. The second column lists up to 20 most frequently mutated driver genes for a given cancer type. Driver genes are ordered from left to right in descending order of their mutation frequency among the samples. Note that for some cancer types the list of driver genes contains fewer than 20 genes, because either the corresponding number of significantly mutated driver genes is less than 20.

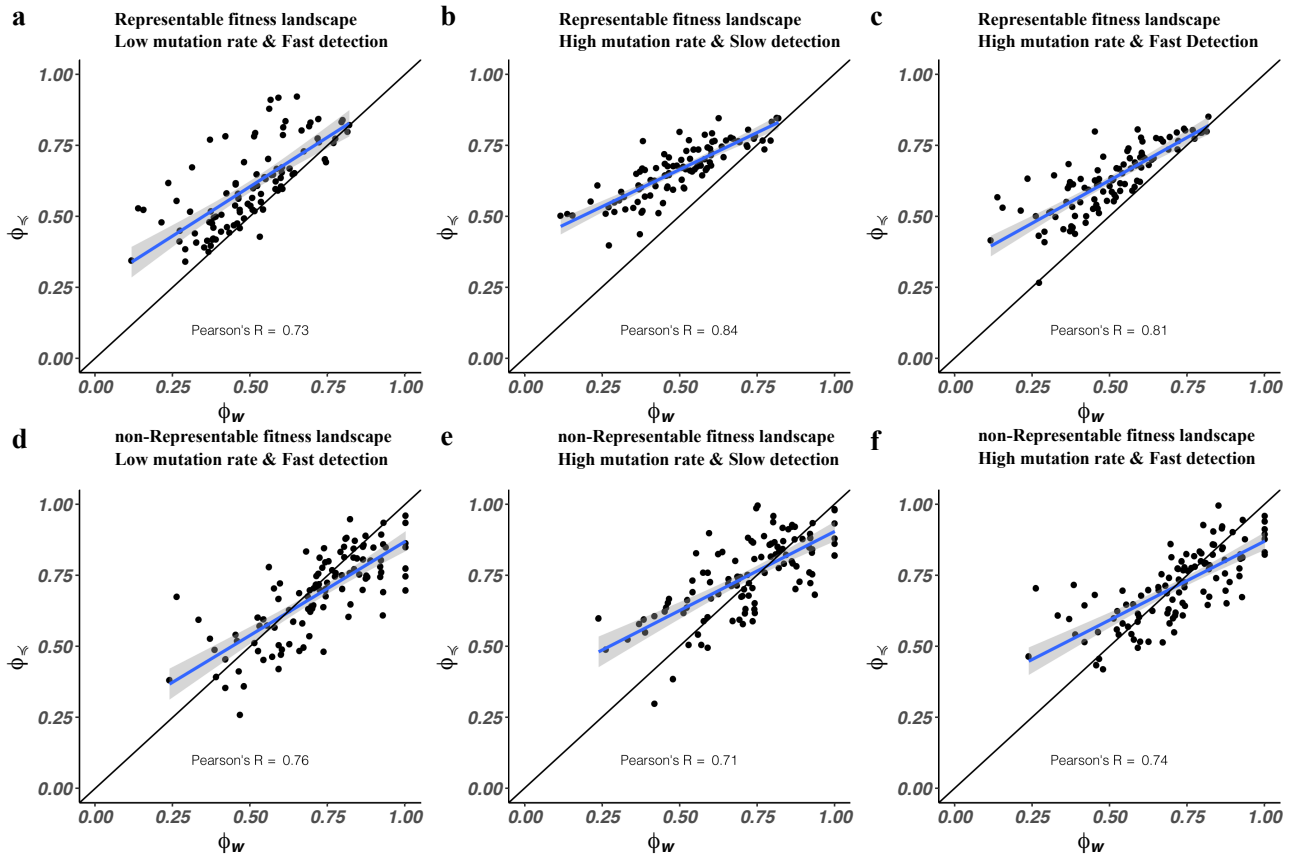

Figure S 1. Effects of mutation and detection rate on the correlation between ϕ_w and $\phi_{\preceq}$. Upper and lower panels respectively correspond to representable and non-representable fitness landscapes. Each point in the panels corresponds to a given fitness landscape and its x-axis and y-axis respectively show its fitness landscape based (ϕ_w) and CBN-based ($\phi_{\preceq}$) predictabilities. The black lines are the identity lines, and the blue lines are the regression lines surrounded by a shaded confidence interval region. The genotypes are the final outcome of evolutionary simulations (See text S1) with low mutation rate (10^{-6}) and fast detection (in panels a and d), high mutation rate (10^{-5}) and slow detection (in panels b and e), and high mutation rate (10^{-5}) and fast detection (in panels c and f).

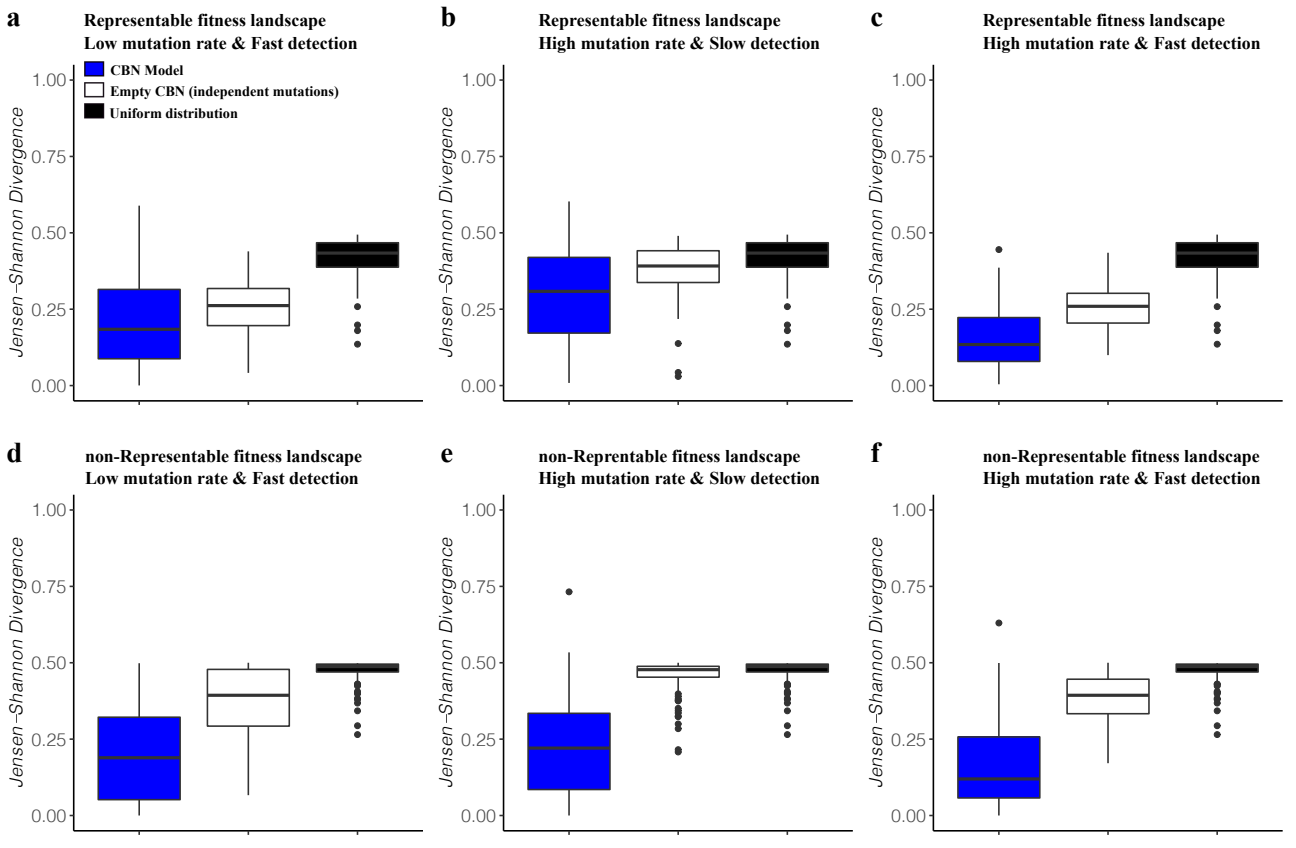

Figure S 2. CBN model-based versus fitness landscape-based pathway probability distributions. In each panel, the vertical axis shows the Jensen-Shannon divergence between the pathway probability distributions of fitness landscape approach ($P(\pi_w)$, Eq. 7 in the main text) and that of the CBN-based approach ($P(\pi_{\preceq})$, Eq. 10 in the main text, blue boxes), the empty CBN model (white boxes) and the uniform pathway probability distribution (black boxes). Upper and lower panels respectively correspond to representable and non-representable fitness landscapes. The genotypes are the final outcome of evolutionary simulations (See text S1) with low mutation rate (10^{-6}) and fast detection (in panels a and d), high mutation rate (10^{-5}) and slow detection (in panels b and e), and high mutation rate (10^{-5}) and fast detection (in panels c and f). Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima.

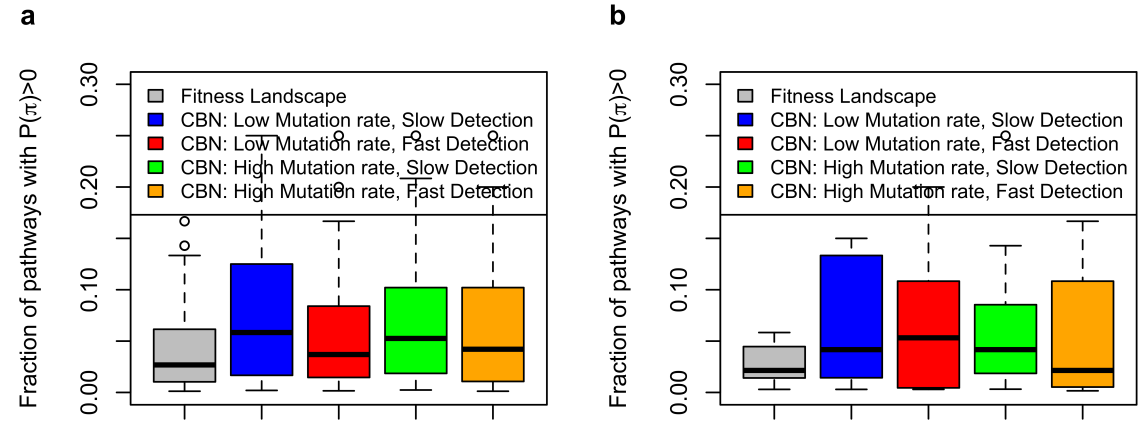

Figure S 3. CBN-based approach predicts higher number of mutational pathways with non-zero probabilities than the fitness landscape based approach. In each panel, the vertical axis shows the fraction of mutational pathways with non-zero probability based on fitness landscape approach (red boxes) versus the CBN approach in four different simulation conditions, whose corresponding boxes are color-coded according to the legend. Panel a) corresponds to representable fitness landscapes, while panel b) corresponds to non-representable ones. Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima.

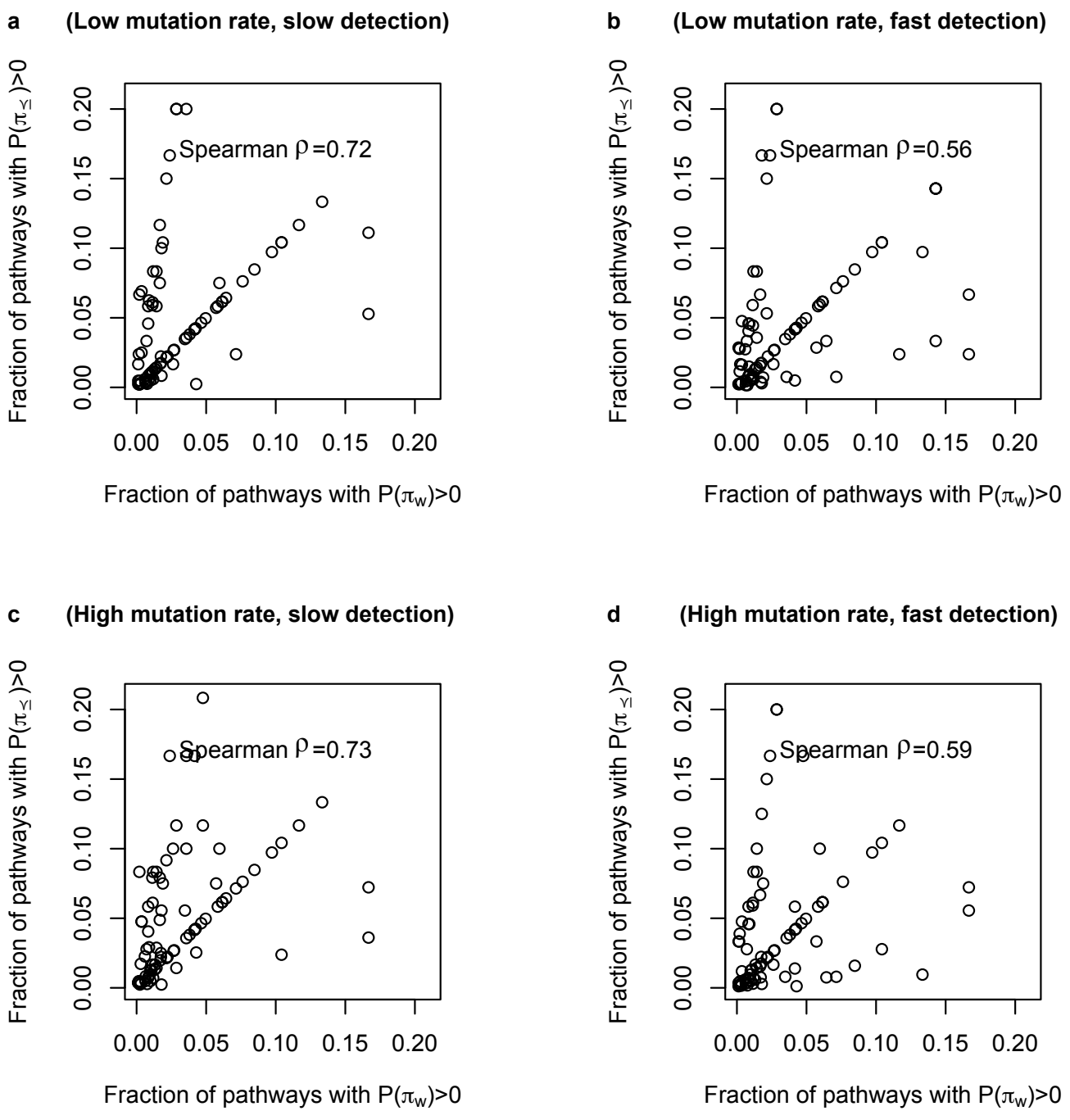

Figure S 4. Fraction of pathways with nonzero probabilities in the CBN based versus the fitness landscape based approaches (representable fitness landscapes). Each panel corresponds to a given simulation condition: a) low mutation rate (10^{-6}) and slow detection, b) low mutation rate (10^{-6}) and fast detection, c) high mutation rate (10^{-5}) and slow detection, and d) high mutation rate (10^{-5}) and fast detection (the mutation rate and the cancer detection rate). In each panel, each point corresponds to one of the 100 representable fitness landscapes and the x-axis and y-axis respectively show the fraction of non-zero mutational pathways using fitness landscape based and CBN based approaches respectively. In each panel, the majority of the points are aligned with the identity line, showing that the two approaches assign non-zero probabilities to the same number of mutational pathways. However, in some of the landscapes the CBN based approach overestimates the number of pathways with non-zero probability, and only in a few of the landscapes, the CBN based approach underestimates the fraction of pathways with non-zero probability. The two approaches correlate well in terms of the fraction of the pathways with non-zero probabilities. Note that the points, which are aligned along the identity line, correspond to the fitness landscapes for which CBN and SSWM predict exactly the same fraction of feasible pathways.

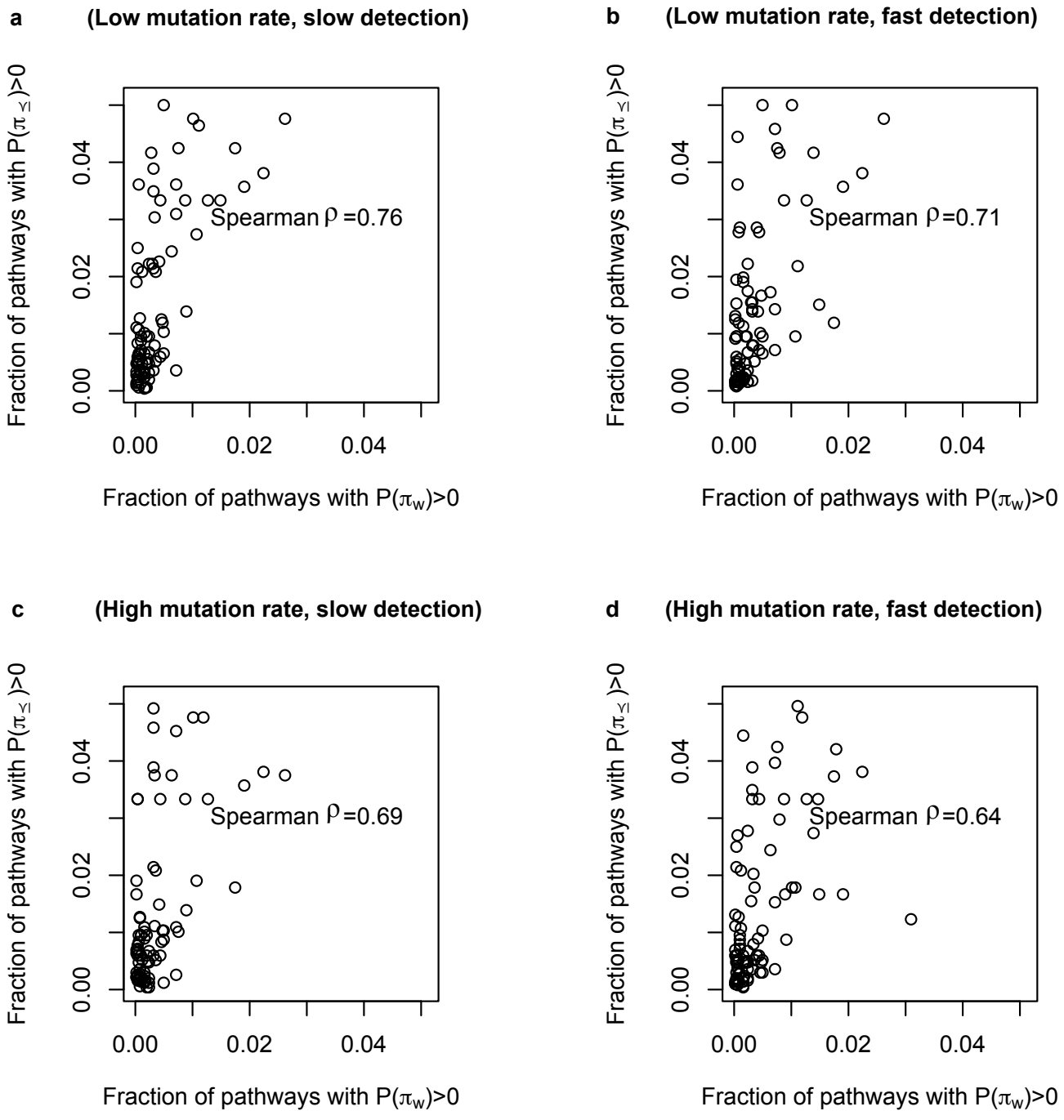

Figure S 5. Fraction of pathways with nonzero probabilities in the CBN based versus the fitness landscape based approaches (non-representable). Each panel corresponds to a given simulation condition: a) low mutation rate (10^{-6}) and slow detection, b) low mutation arte (10^{-6}) and fast detection, c) high mutation rate (10^{-5}) and slow detection, and d) high mutation rate (10^{-5}) and fast detection (the mutation rate and the cancer detection rate). In each panel, each point corresponds to one of the 111 non-representable fitness landscapes and the x-axis and y-axis respectively show the fraction of non-zero mutational pathways using fitness landscape based and CBN based approaches respectively. The two approaches correlate well in terms of the fraction of the pathways with non-zero probabilities.

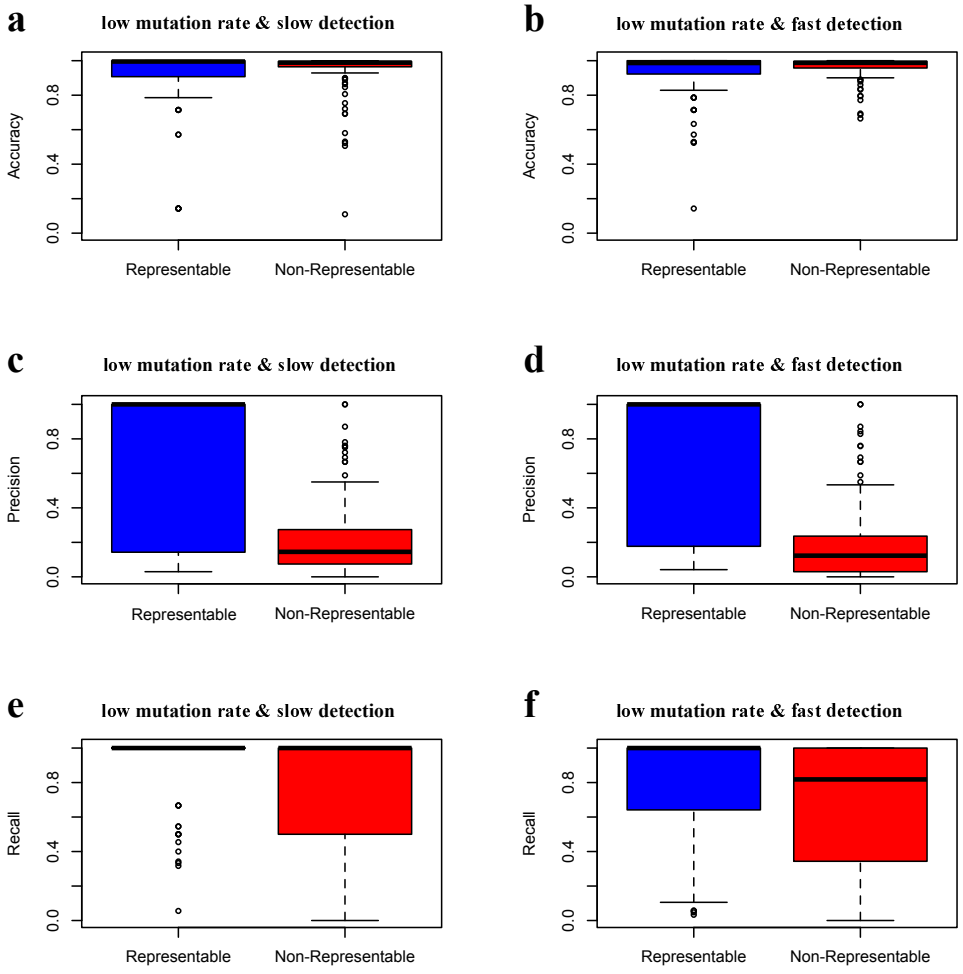

Figure S 6. Identification of mutational pathways with non-zero probability in CBN-based approach as compared to those of the fitness landscape based approach (mutation rate: 10^{-6}). The vertical axes show accuracy (panels a and b), precision (panels c and d), and recall (panels e and f) measured for each fitness landscapes based on the set of mutational pathways with non-zero probability according to the CBN-based approach versus that of the corresponding fitness landscape based approach. The blue and the red box plots correspond respectively to the representable and non-representable fitness landscapes. Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima. The left and right panels correspond to simulations based on slow and fast cancer detection rates respectively. However, in all panels mutation rate in the evolutionary simulations is the same as 10^{-6}.

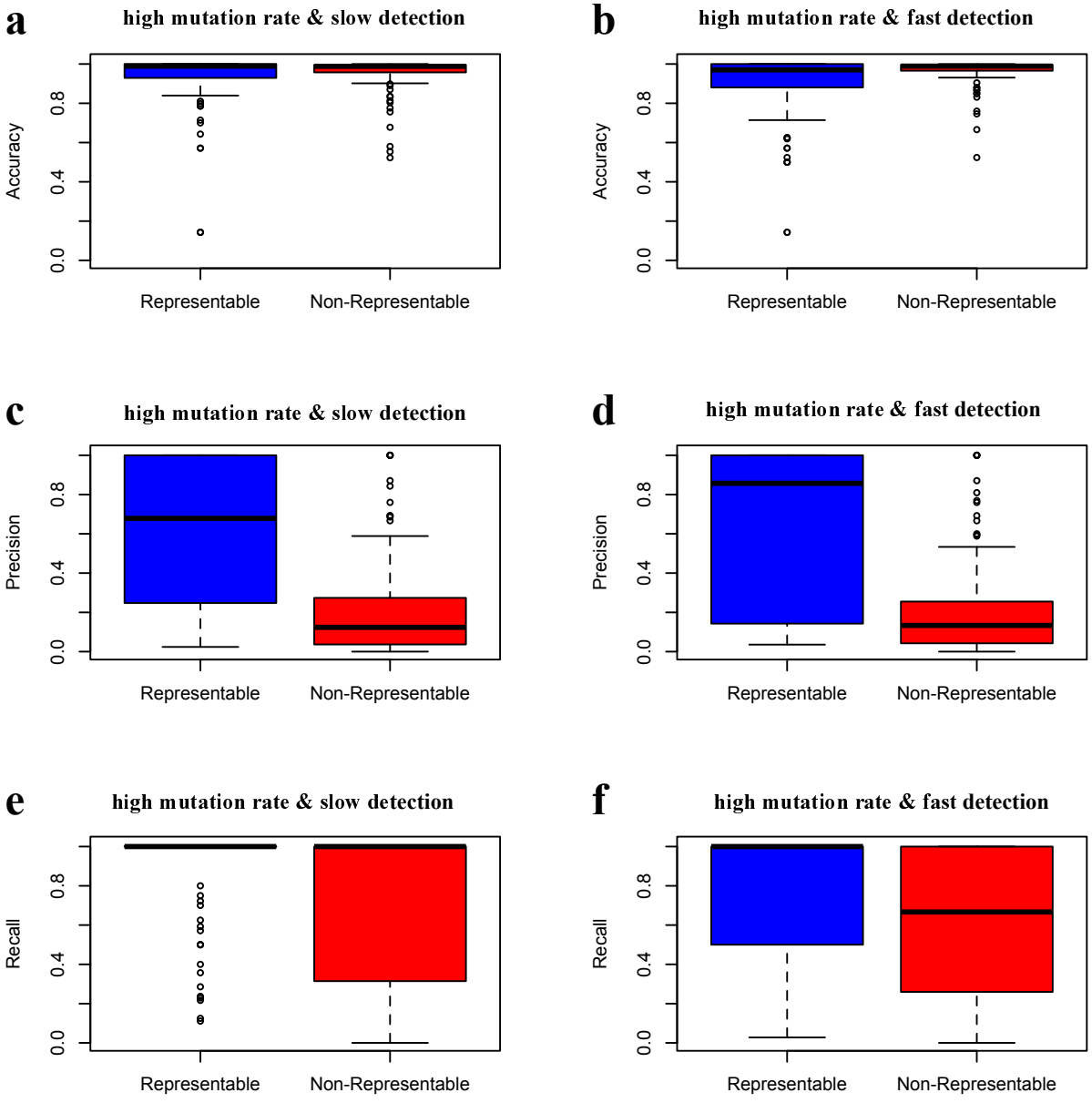

Figure S 7. Identification of mutational pathways with non-zero probability in CBN-based approach as compared to those of the fitness landscape based approach (mutation rate: 10^{-5}). The vertical axes show accuracy (panels a and b), precision (panels c and d), and recall (panels e and f) measured for each fitness landscapes based on the set of mutational pathways with non-zero probability according to the CBN-based approach versus that of the corresponding fitness landscape based approach. The blue and the red box plots correspond respectively to the representable and non-representable fitness landscapes. Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima. The left and right panels correspond to simulations based on slow and fast cancer detection rates respectively. However, in all panels mutation rate in the evolutionary simulations is the same as 10^{-5}.

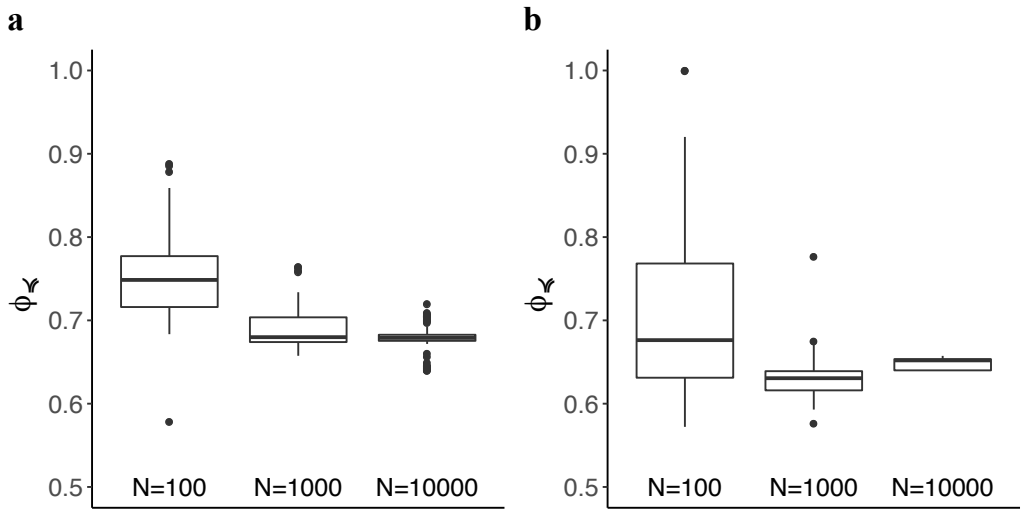

Figure S 8. Effect of sample size on the stability of the CBN-based estimation of the predictability. The analyses correspond to an arbitrary chosen a) representable and b) non-representable fitness landscapes. Each box plot indicates the distribution of $\phi_{\preceq}$ for 100 bootstrap samples from an initial sample with a given number of genotypes ($N = 100$, 1000 and 10000). In lower sample sizes, the variability of the estimated $\phi_{\preceq}$ is higher.

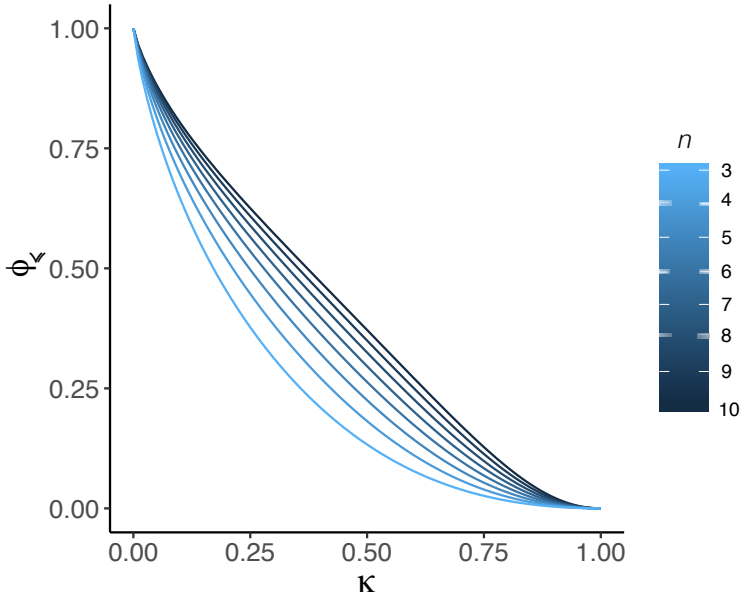

Figure S 9. Predictability (ϕ) as a function of κ. Each curve represents predictability (vertical axis) as a function of κ (horizontal axis) for a given n, which is color coded according to the color-bar (see text S4 for more details).

Figure S 10. Approximated predictability versus exact predictability. In each panel, the black curve represents the approximated predictability (vertical axis), which is calculated according to the Eq. 12 in the main text, versus the exact predictability (horizontal axis), and the red line is the identity line. Each panel corresponds to a given n and n'. It is noticeable, that in all values of n and n', the deviation of the black curve from the identity red line is negligible, confirming the accuracy of the approximated predictability. However, this level of accuracy is achievable based on the assumption that the λ of the consecutive mutations remains constant and equal to κ (see text S4 for more details).

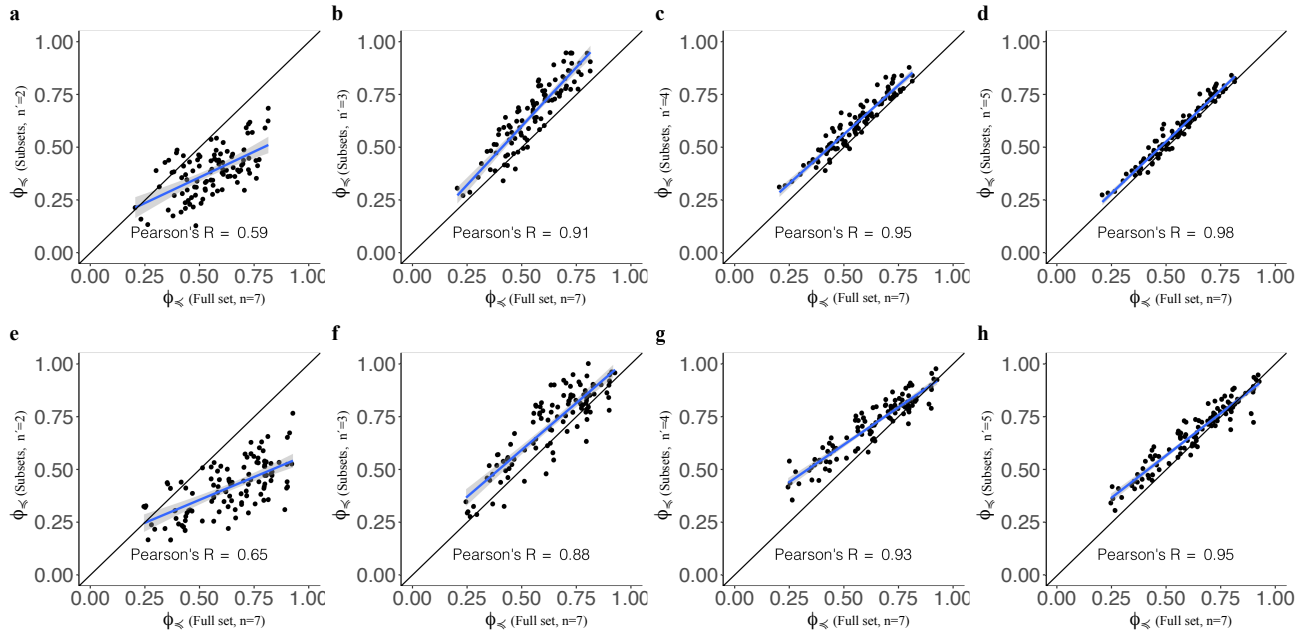

Figure S 11. The approximated subsets-based predictability ($\phi_{\preceq} subsets$) versus the full CBN based ($\phi_{\preceq} fullset$) predictability (Simulations). In all panels, each point corresponds to a given representable (upper panels) and non-representable (lower panels) fitness landscape. The x-axis shows the predictability based on CBN including all $n = 7$ genes, while the y-axis is the approximated predictability ($\phi_{\preceq} subsets$); see Eq. 12 in the main text) estimated based on the average of subsets including: i) all $\binom{7}{2}$ subsets of $n' = 2$ genes (doublets; panels a and e), ii) all $\binom{7}{3}$ subsets of $n' = 3$ genes (triplets; panels b and f), iii) all $\binom{7}{4}$ subsets of $n' = 4$ genes (quartets; panels c and g) iv) all $\binom{7}{5}$ subsets of $n' = 5$ genes (quintets; panels d and h). The black lines are the identity lines, and the blue lines are the regression lines surrounded by a shaded confidence interval region. In this analysis, mutation rate is 10^{-6} and simulations are based on slow cancer detection.

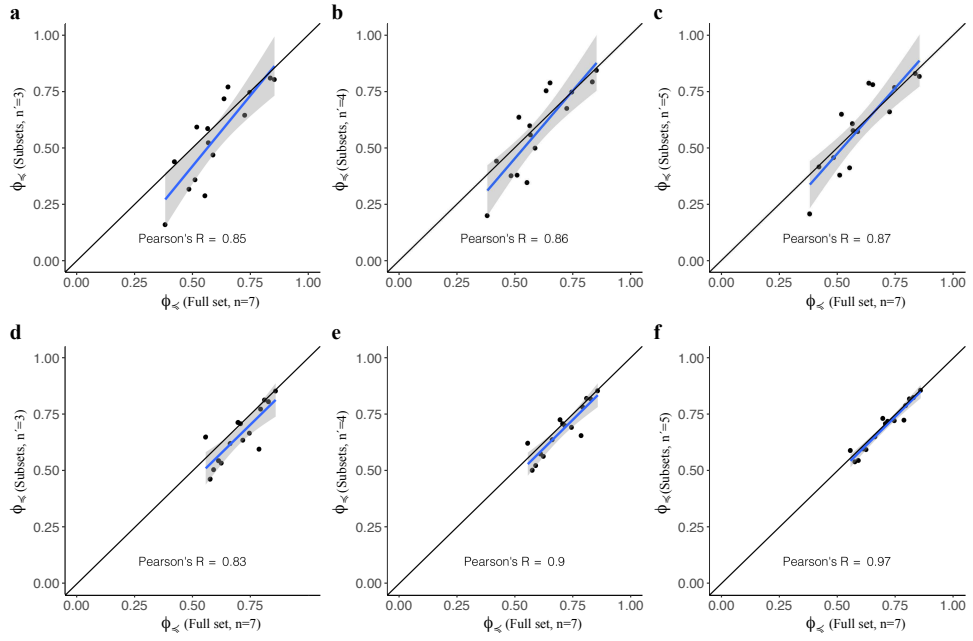

Figure S 12. The approximated subsets-based predictability ($\phi_{\preceq} subsets$) versus the full CBN based ($\phi_{\preceq} fullset$) predictability (Real data). In all panels, each point corresponds to one of the 15 cancer types, for which the genotypes are defined based on the corresponding 7 most frequently mutated driver genes using TCGA dataset (upper panels) or MSK dataset (lower panels). The x-axis shows the predictability based on CBN including all $n = 7$ genes, while the y-axis is the approximated predictability ($\phi_{\preceq} subsets$; see Eq. 12 in the main text) estimated based on the average of subsets including: i) all $\binom{7}{3}$ subsets of $n' = 3$ genes (triplets; panels a and d), ii) all $\binom{7}{4}$ subsets of $n' = 4$ genes (quartets; panels b and e), and iii) all $\binom{7}{5}$ subsets of $n' = 5$ genes (quintets; panels c and f). The black lines are the identity lines, and the blue lines are the regression lines surrounded by a shaded confidence interval region.

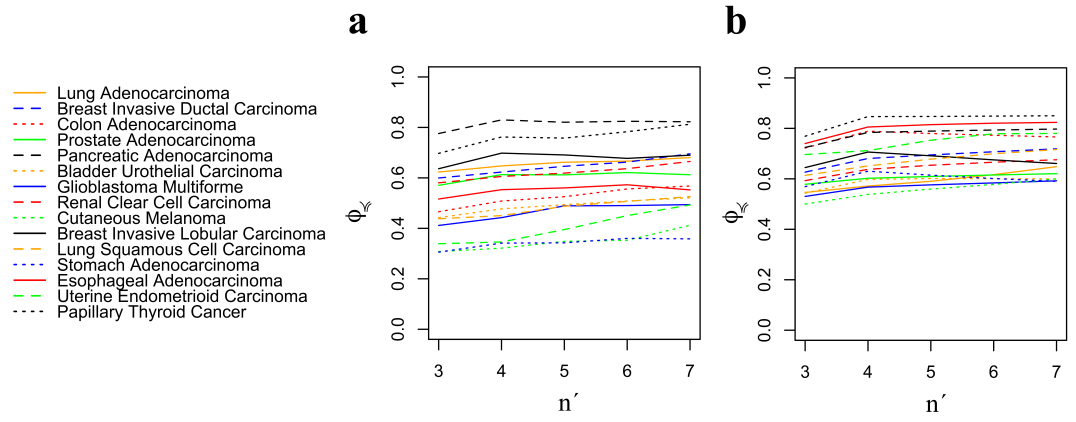

Figure S 13. Convergence of the subsets-based approximation of the CBN-based predictability ($\phi_{\preceq} subsets$). In this analysis, for each of the 15 cancer types, we define the genotypes based on the corresponding 10 most frequently mutated driver genes using TCGA dataset (panel a) or MSK dataset (panel b). Each line in the panels correspond to a given cancer type specified according to the legend. The y-axes show the quantified predictability and the x-axes indicate the number of genes (n') included in the subsets. According to the Eq. 12 in the main text, the predictability for a given n' is calculated as the average predictability among all $\binom{10}{n'}$ CBNs with n' genes. Thus, even for $n > 7$ (here $n = 10$), the ($\phi_{\preceq} subsets$) converges when $n' = 4$ in both datasets and all cancer types.

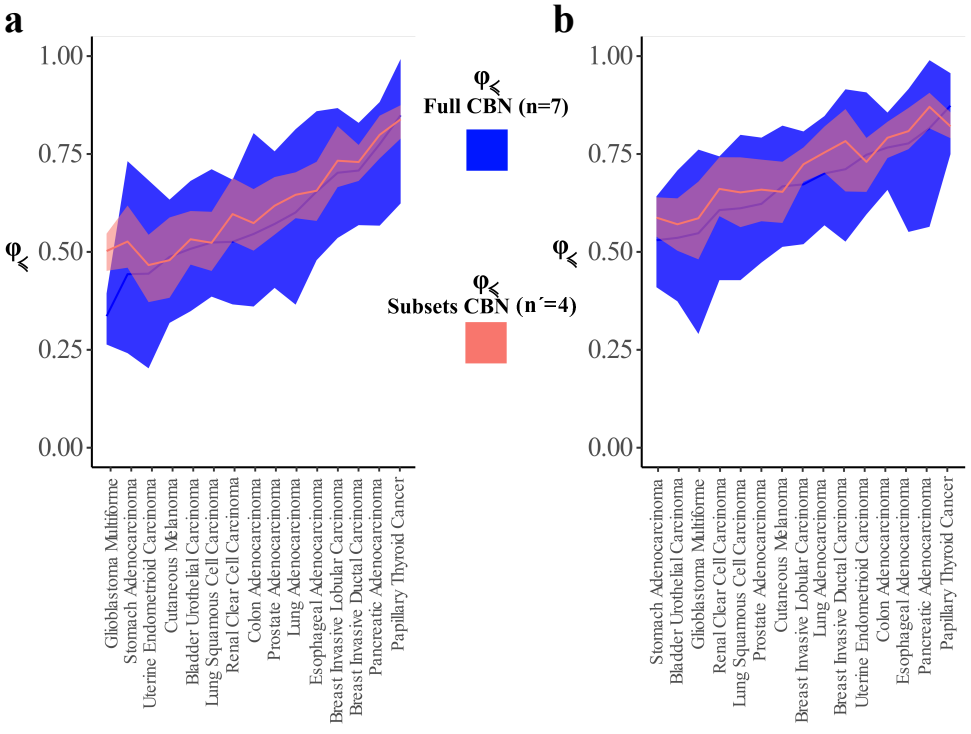

Figure S 14. Robustness of the subsets-based approximation of the CBN-based predictability ($\phi_{\preceq} subsets$). In this analysis, for each of the 15 cancer types, we define the genotypes based on the corresponding 7 most frequently mutated driver genes using TCGA dataset (panel a) or MSK dataset (panel b). For each cancer type and each dataset, we generated 100 bootstrap samples and for each sample we quantified predictability both with full CBN approach ($n = 7$) and with the subsets-based approximation of the CBN-based predictability with $n' = 4$ (Eq. 12 in the main text). In both panels, the y-axis indicates the estimated predictability and the x-axis represents the cancer type. The shaded regions correspond to the 95% confidence interval of the full CBN approach (blue region) and subsets-based CBN approach (red region).

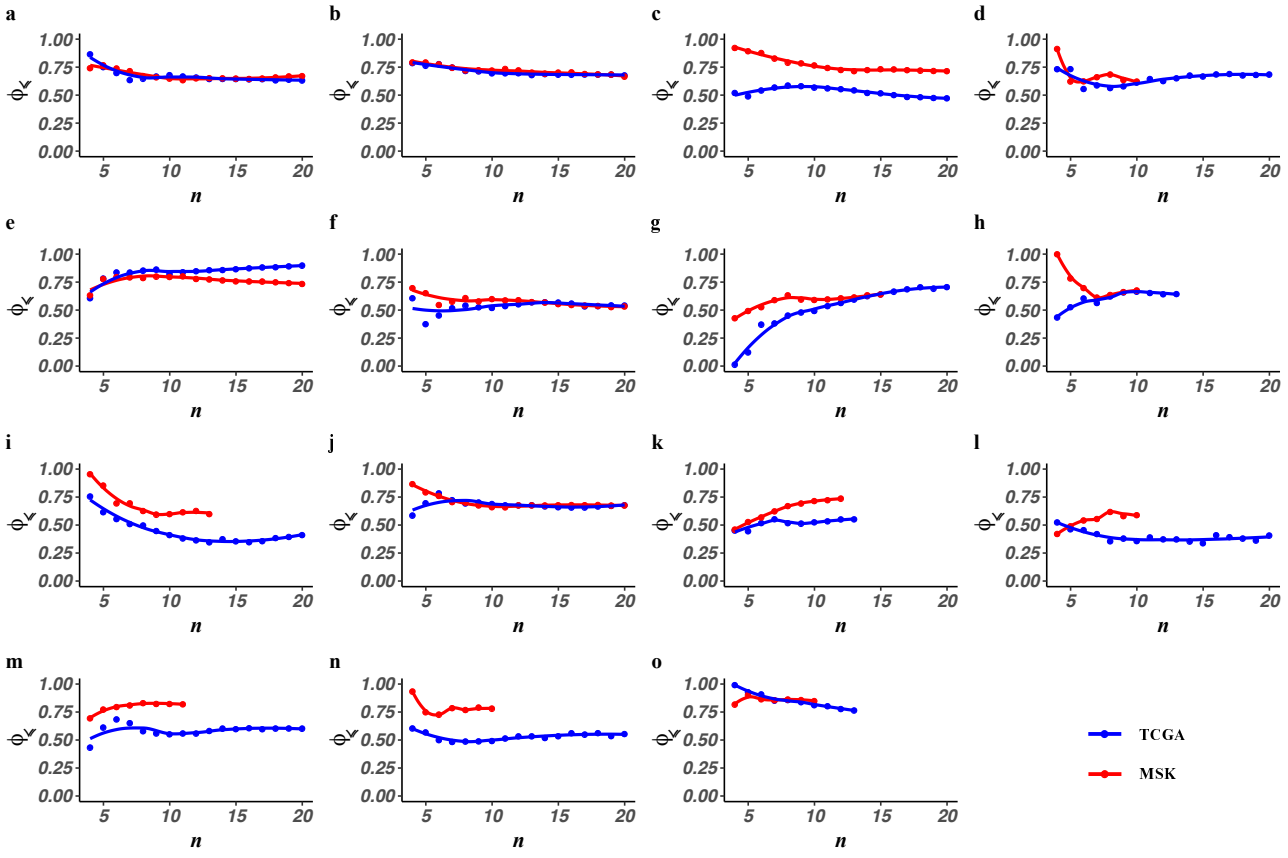

Figure S 15. Predictability as a function of the number of driver genes (n). Each panel corresponds to a given cancer type: a) Lung Adenocarcinoma, b) Breast Invasive Ductal Carcinoma, c) Colon Adenocarcinoma, d) Prostate Adenocarcinoma, e) Pancreatic Adenocarcinoma, f) Bladder Urothelial Carcinoma, g) Glioblastoma Multiforme, h) Renal Clear Cell Carcinoma, i) Cutaneous Melanoma, j) Breast Invasive Lobular Carciboma, k) Lung Squamous Cell Carcinoma, l) Stomach Adenocarcinoma, m) Esophageal Adenocarcinoma, n) Uterine Endometrioid Carcinoma, and o) Papillary Thyroid Cancer. Vertical axis indicates the estimated CBN-based predictability using the subsets approach (with $n' = 4$) and the x-axis shows the number of driver genes considered for defining the genotypes (n), which varies from $n = 4$ to $n = 20$. Blue and red curves correspond respectively to the TCGA and MSK datasets. As n increases, the preditability converges to a fixed value in all cancer types. Note that for some cancer types, especially in MSK datasets, n does not reach 20, because the number of mutated driver genes is less than 20 for the given cancer type and dataset. Furthermore, it is noticeable that all the curvesat least reach up to $n = 10$, and from $n = 10$ upwards, the estimated predictability remains almost constant in all cancer types and datasets.

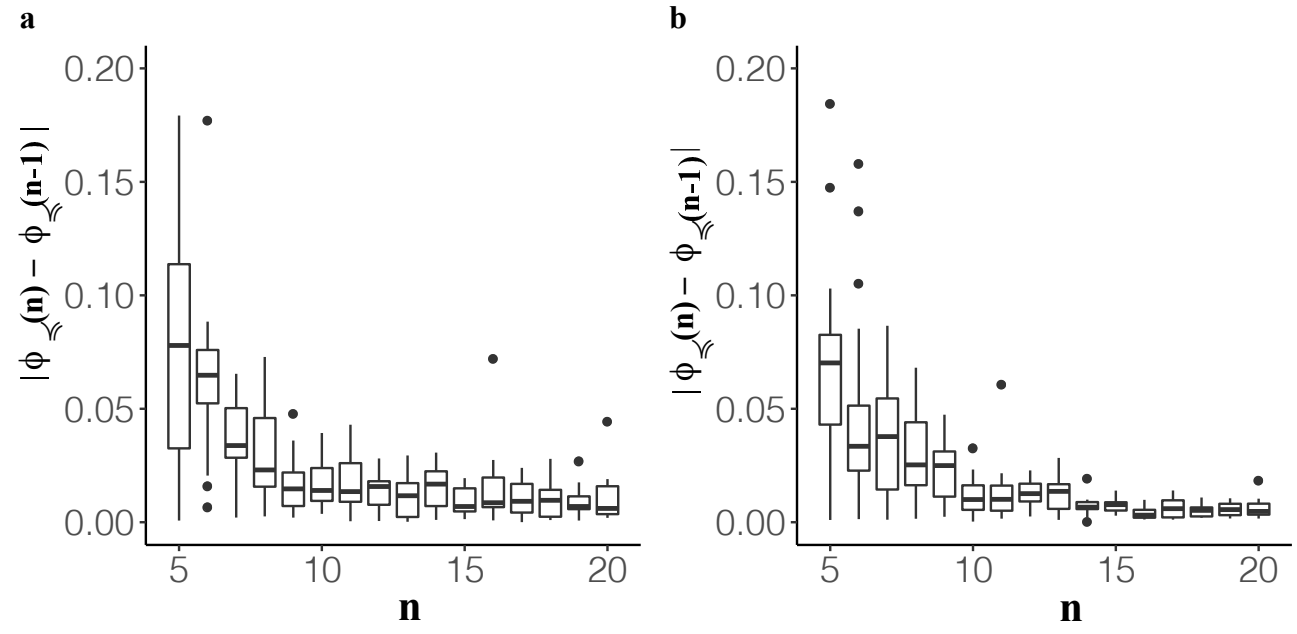

Figure S 16. Convergence of the predictability as a function of n. In both panels, the vertical axis shows the difference between the predictability of consecutive n ($\mid \phi_{\preceq}(n) - \phi_{\preceq}(n-1) \mid$) as a function of the number of considered driver genes (n). Each box represents the distribution of ($\mid \phi_{\preceq}(n) - \phi_{\preceq}(n-1) \mid$) among the 15 cancer types based on A) TCGA and B) MSK datasets. Note that in $n = 10$, ($\mid \phi_{\preceq}(n) - \phi_{\preceq}(n-1) \mid$) converges to the minimum level in both datasets.

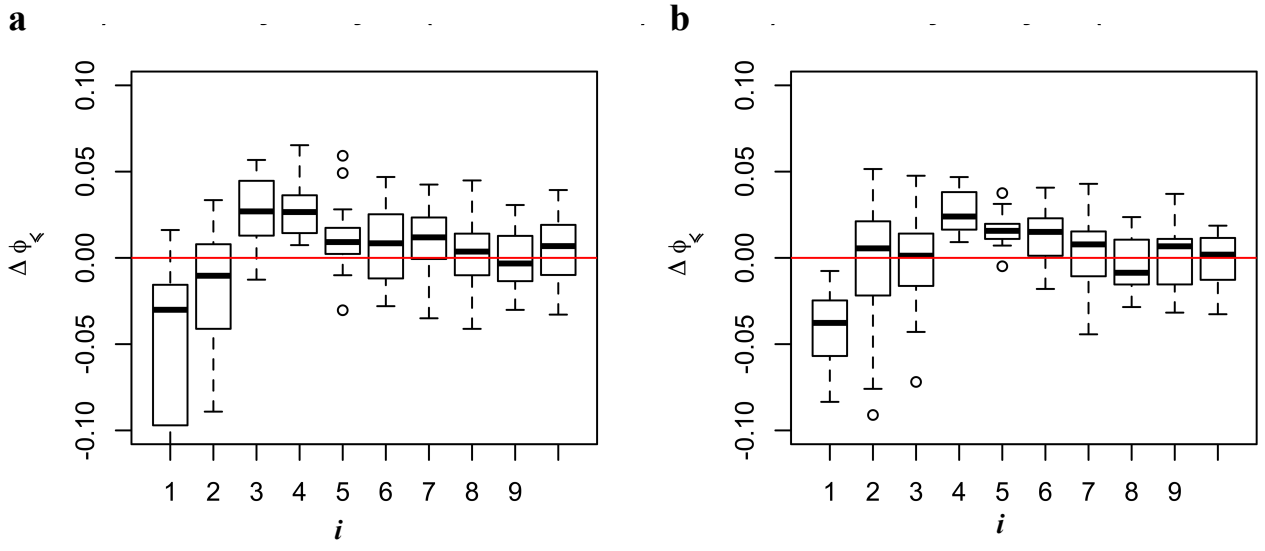

Figure S 17. Sensitivity of the CBN based estimation of predictability ($\phi_{\preceq}$) to gene removal. The vertical axes represent the difference ($\Delta\phi_{\preceq}$) between the estimated predictability after and before the removal of the i^{th} frequently mutated driver gene (x-axis). The boxes indicate the distribution of ($\Delta\phi_{\preceq}$) among the 15 cancer types using a) TCGA dataset and b) MSK dataset. Boxes span 25-th to 75-th percentile, and whiskers indicate maxima and minima.

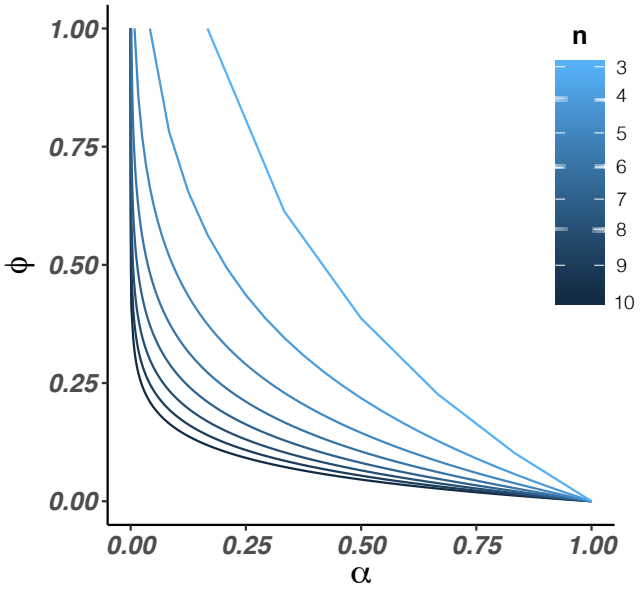

Figure S 18. Predictability as a function of the fraction of feasible mutational pathway (α). Assuming equal pathway probability for all feasible mutational pathways (see text S9), each curve shows the predictability (ϕ, y-axis) as a function of the fraction of feasible pathways (α, x-axis, see equation (11) in text S9) for a given n color coded according to the color bar.

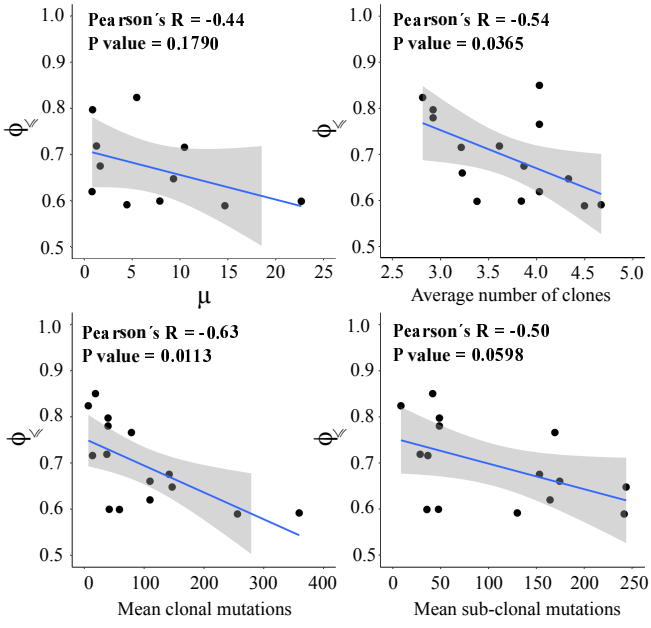

Figure S 19. Predictability, mutation frequency and intra-tumor heterogeneity. In all panels, each point corresponds to a given cancer type and the vertical axis indicates the estimated predictability ($\phi_{\preceq}$). The horizontal axis shows a) the average mutation frequency per mega base-pairs (determined by analyzing more than 3000 samples of the Broad Institute ((Lawrence et al., 2013)), b) the average number of clones, c) mean clonal mutations, and d) mean sub-clonal mutations according to (Raynaud et al., 2018). The blue lines are the linear regression models surrounded by a shaded confidence interval region. The predictability is approximated for CBNs according to (Eq. 12), with $n = 10$ and $n' = 4$ on MSK dataset. Note that in panel a only 11 cancer types are included in this analysis, because the Broad Institute dataset (Lawrence et al., 2013) covered only 11 of our 15 cancer types.

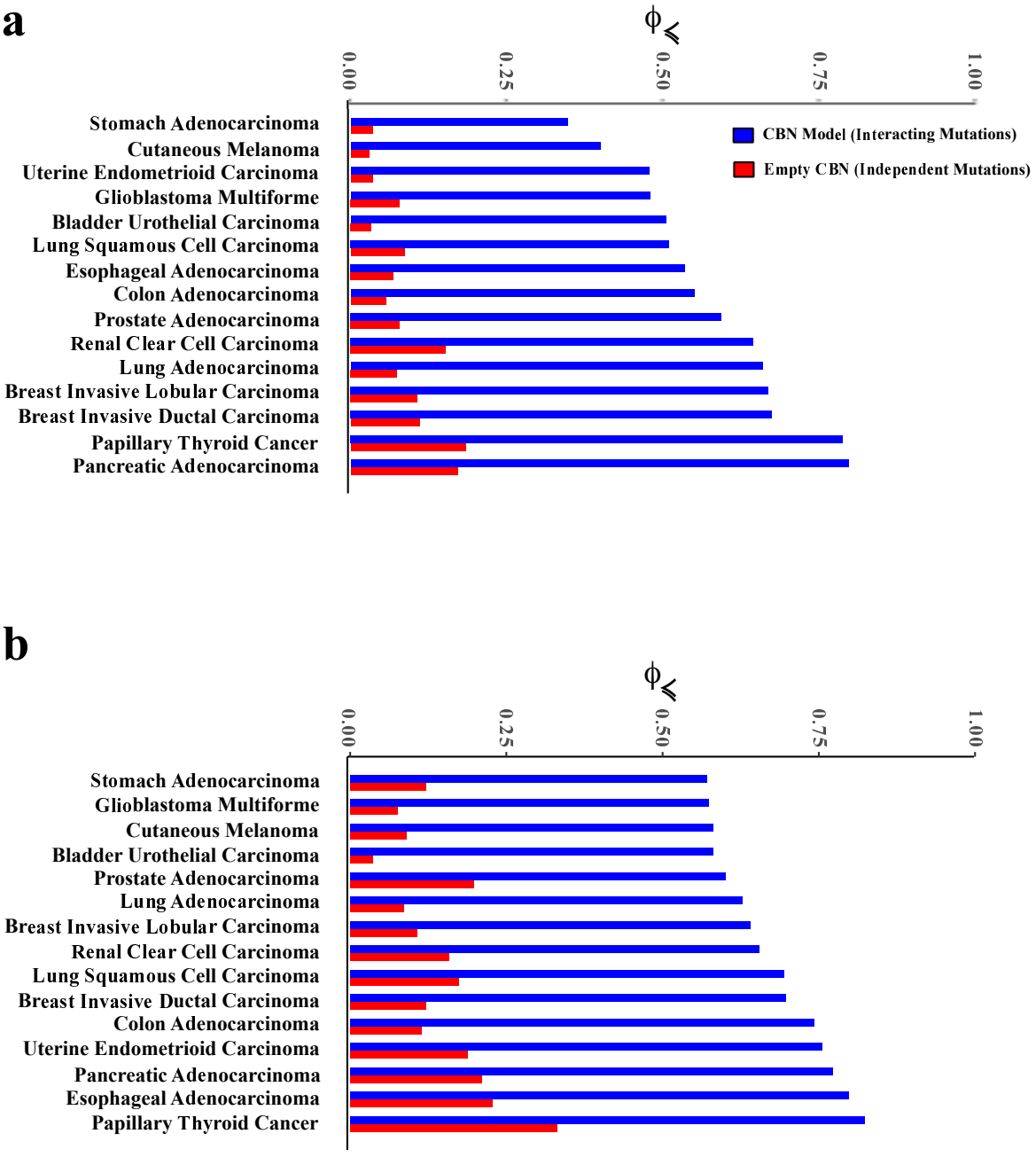

Figure S 20. Empty CBN-based quantification of the predictability: interacting versus independent mutations. The horizontal axes in the bar plots show the predictability quantified i) based on CBN (blue bars, Eq. 12 in the main text, with $n = 10$ and $n' = 4$), and ii) based on empty CBN (red bars) for each cancer type specified in the vertical axes for a) TCGA dataset and b) MSK dataset. Note that in an empty CBN, the mutations are assumed to be independent and hence no restrictions on their ordering exist; For more details see text S10.