PNAS

In silico learning of tumor evolution through mutational time series

Noam Auslander[a], Yuri I. Wolf[a], and Eugene V. Koonin[a,1]

[a]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894

Contributed by Eugene V. Koonin, March 15, 2019 (sent for review January 29, 2019; reviewed by Natalia L. Komarova and Andrei Zinovyev)

Cancer arises through the accumulation of somatic mutations over time. Understanding the sequence of mutation occurrence during cancer progression can assist early and accurate diagnosis and improve clinical decision-making. Here we employ long short-term memory (LSTM) networks, a class of recurrent neural network, to learn the evolution of a tumor through an ordered sequence of mutations. We demonstrate the capacity of LSTMs to learn complex dynamics of the mutational time series governing tumor progression, allowing accurate prediction of the mutational burden and the occurrence of mutations in the sequence. Using the probabilities learned by the LSTM, we simulate mutational data and show that the simulation results are statistically indistinguishable from the empirical data. We identify passenger mutations that are significantly associated with established cancer drivers in the sequence and demonstrate that the genes carrying these mutations are substantially enriched in interactions with the corresponding driver genes. Breaking the network into modules consisting of driver genes and their interactors, we show that these interactions are associated with poor patient prognosis, thus likely conferring growth advantage for tumor progression. Thus, application of LSTM provides for prediction of numerous additional conditional drivers and reveals hitherto unknown aspects of cancer evolution.

cancer progression | driver mutations | passenger mutations | machine learning | neural networks

Tumorigenesis is a multistep process characterized by accumulation of somatic mutations, which contribute to tumor growth, clinical progression, immune escape, and the development of drug resistance (1, 2). The somatic mutations found in a tumor cell are accumulated over the lifetime of the cancer patient, so that some mutations are acquired in early steps of tumorigenesis and even in premalignant cells (3). Relatively small subsets of these mutations are established tumor drivers, whereas the remainder are thought to be passengers that do not confer growth advantage or may even negatively affect tumor fitness (4–7). Although molecular and cell biology studies have revealed many mechanistic details of tumorigenesis (4, 8–10), our understanding of the tumor evolution dynamics remains limited, presumably due to the complexity of the process and the abundance of passenger events that could be randomly distributed but might exert various effects on tumor fitness and properties (11, 12).

In colorectal cancer, tumor development has been explained by a multistep model of carcinogenesis that describes the progression of a benign adenoma to a malignant carcinoma through a series of well-defined histological stages that are linked to a mutational time series, i.e., the temporal sequence of occurrence of driver mutations (13–16). Similar stepwise models have been developed for other types of adenocarcinomas (17, 18), although the temporal succession of molecular changes characterizing the progression of these tumors has not been elucidated at the level of confidence it has for colon cancer (19). Given that the somatic alterations in some tumors can be represented as a multistep sequence of events, we conjectured that time series learning algorithms could be applied to the sequence of somatic mutations, potentially revealing the complex dynamics of the tumor evolution and enabling a variety of context-specific predictions.

To this end, we employed long short-term memory (LSTM) networks (20), a type of recurrent neural network (RNN) (21) capable of learning long-term dependencies in a time series sequence. These networks have achieved major success in time series prediction tasks and for learning evolution of recurrent systems (20, 22–24). We demonstrate here the utility of applying LSTM to time-ordered mutational data in colon and lung adenocarcinomas. We define a pseudotemporal gene ranking, representing each tumor sample as a binary text, which can be used for training LSTM in a similar manner to that applied in natural language texts classification (25–27). Applied to a discrete version of the mutational data, ordered into an approximate temporal sequence, these networks can be used to predict the mutational load from a limited number of mutations and for stepwise prediction of the occurrence of following mutations in the series. Using the sequence dynamics learned by the models, we reconstruct sequences of mutations that are statistically indistinguishable from the original observations. We find that the occurrence of distinct subsets of mutations could be predicted from the timeline of the major cancer drivers. Investigation of the driver genes contributing to the prediction of each of these mutations uncovers numerous driver–interactor gene pairs that are highly and specifically enriched with different types of independently identified interactions. We further derive modules of driver genes and find that their predicted interactions are associated with poor survival rate, suggesting that the driver and associated passengers jointly promote tumor growth.

Significance

Cancer is caused by the effects of somatic mutations known as drivers. Although a number of major cancer drivers have been identified, it is suspected that many more comparatively rare and conditional drivers exist, and the interactions between different cancer-associated mutations that might be relevant for tumor progression are not well understood. We applied an advanced neural network approach to learn the sequence of mutations and the mutational burden in colon and lung cancers and to identify mutations that are associated with individual drivers. A significant ordering of driver mutations is demonstrated, and numerous, previously undetected conditional drivers are identified. These findings broaden the existing understanding of the mechanisms of tumor progression and have implications for therapeutic strategies.

Author contributions: N.A., Y.I.W., and E.V.K. designed research; N.A. performed research; N.A., Y.I.W., and E.V.K. analyzed data; and N.A. and E.V.K. wrote the paper.

Reviewers: N.L.K., University of California, Irvine; and A.Z., Institut Curie.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. Email: koonin@ncbi.nlm.nih.gov.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1901695116/-/DCSupplemental.

Published online April 23, 2019.

Results

Throughout this work, we use a discretized version of the mutational data from two tumor types, namely, colon cancer, where the stepwise evolutionary model has been established (13–16), and lung cancer, where such a model has been suggested but not widely accepted as it is in the case of colon cancer (17, 19). The snapshot mutational datasets are ordered into an approximate temporal sequence that is estimated via the training sets for each tumor type. For each classification task, we trained LSTM networks using The Cancer Genome Atlas (TCGA) (28) time-ordered mutational data for these tumor types and tested the network performance using independent datasets (Table 1).

Predicting Tumor Mutational Load from the Mutational Time Series. We aim to represent tumor mutations as a discrete timeline of mutational events such that the appearance of a mutation in the timeline would correspond to their estimated place in the tumor evolution. We thus calculate a score for every mutation as the ratio of its occurrences in the presence and in the absence of other mutations (see *Materials and Methods* for details). This form of the score is motivated by the assumption that mutations that appear late in the tumor evolution are more likely to be fixed when other mutations are present, supported by a previously established notion that cooccurring events more often take place at late stages in tumor evolution (29–32). Sorting the mutations by the scores evaluated on our training datasets, we find that this estimated order of mutation appearance is in agreement with the established succession of the key drivers in colon adenocarcinoma (*SI Appendix*, Fig. S1*A*). In particular, the APC mutation is an early driver event, followed by CTNBB1, KRAS, and SMAD4, whereas PIK3CA and TP53 mutations occur later in tumor development (13, 33). Moreover, we find that the scores significantly correlate with the ratio between the frequency of a mutation in colon adenomas and its frequency in colon carcinomas [from COSMIC database (34, 35); *SI Appendix*, Fig. S1*B*], further supporting the relevance of this score for the inference of the order of mutations.

We then used the time-ordered mutational data to predict the overall mutational load in the respective tumors and evaluate the number of mutations required for an accurate prediction. To this end, we trained LSTM networks aiming to predict high vs. low mutational load (*Materials and Methods*) from a time series of mutations. Given a discrete time series of mutation occurrences, the LSTM network is trained using the series up to a time t and is applied to the left-out test to produce scores that reflect the probability of each sample in the test to have a high mutational load. Starting from the final time point (the last mutation in the series, likely occurring late in the tumor evolution), we find that prediction of the mutational load saturates with high performance (AUC > 0.95) with fewer than 100 mutations for both colon and lung adenocarcinomas (Fig. 1*A*). The sets of genes that contribute to the mutational load prediction significantly overlap between colon and lung cancers (P value ≈ 0 for the last 100 genes). Notably, this overlap includes genes that encode some of the longest human proteins that perform various organizing roles in either intracellular or intercellular interactions, such as titin (TTN), mucin-16 (MUC16), and nesprin-1 (SYNE1). TTN is one of the most commonly mutated early drivers in colon cancer (36). MUC16 is implicated in the progression of several cancers, apparently, via interactions with the immune system, and is emerging as an important target for cancer therapy (37). SYNE1, a cytoskeleton organizer, although less thoroughly characterized, appears to contribute to DNA damage response and, thus, to genome instability and tumorigenesis (38). Thus, the genes that contribute to mutation load prediction in both types of cancer seem to reveal common biological themes. Moreover, the scores assigned by the LSTM using only the 20 latest mutations (i.e., the 20 mutations with the highest order scores) as a sequence are highly correlated with the observed mutational load in all test sets (Fig. 1 *B–D*). Using the time series from the earliest time point, however, results in almost random prediction performance with similar number of mutations (*SI Appendix*, Fig. S2), suggesting that the ultimate mutational burden of a tumor depends primarily on mutations that occur late in tumor evolution. Training linear classifiers to predict the mutational load from the same sequences of mutations or randomly selected mutations (*Materials and Methods*) resulted in significantly inferior performance compared with the LSTM (Fig. 1 *E–G*), suggesting that the LSTM networks learn complex dynamics within the mutational data that could not be captured by conventional classification approaches.

The scores assigned to predict the mutational load are also associated with clinical phenotypes. For the colon cancer test sets, the scores assigned by LSTM trained with the 20 latest mutations in the sequence are significantly higher in samples of primary tumors assigned with poor grade vs. moderate grade, in high vs. low microsatellite instability, and in high vs. low CpG island methylation phenotypes (Fig. 1 *H–K*). In the lung cancer test set, higher scores are associated with lower progression-free survival (PFS; Fig. 1*L*).

Predicting Occurrence of Mutations in the Sequence and Simulated Data Analysis. We then explored the possibility of predicting the occurrence of mutations in the course of tumor evolution from other mutations in the time series. To this end, the LSTM networks were trained to predict the occurrence of each mutation M_t in the time series, using the time series of mutations from the latest mutation up to the M_{t+1} mutation (thus predicting earlier mutations from later ones). The occurrence of most mutations in the sequence could be predicted with good accuracy, with the median AUC = 0.88, 0.73, and 0.69 for the first and second colon test sets and for the lung test set, respectively (Fig. 2*A*). Mutations for which occurrence could not be predicted (AUC < 0.5; 2 and 17% of the mutations in colon and lung cancers, respectively) were significantly less common than those that were readily predictable (rank-sum P value = 0.001 for colon and 2.4e-106 for lung; Dataset S1), implying that the occurrence of these low-frequency mutations is not linked to tumor progression.

More unexpected than the poor prediction for low-frequency mutations, the prediction accuracy for mutations in known cancer driver genes was significantly lower than that for mutations in all other genes, in both tumor types (Fig. 2 *B* and *C*). This observation is maintained when including only genes that are frequently mutated (*SI Appendix*, Fig. S3). These findings indicate that the driver mutations that determine the course of tumorigenesis could not be readily predicted from the rest of the mutational landscape of a given tumor type. In contrast, many passenger mutations tend to be linked to specific drivers (39) and

Table 1. Datasets used for training and testing, for colon and lung adenocarcinomas

Dataset	Colon	Lung
Training	COAD (n = 253)	LUAD (n = 506)
Test	Colorectal adenocarcinoma [DFCI (69), n = 612] Colorectal adenocarcinoma [Genentech (70), n = 72]	Lung adenocarcinoma [Broad (71), n = 183]

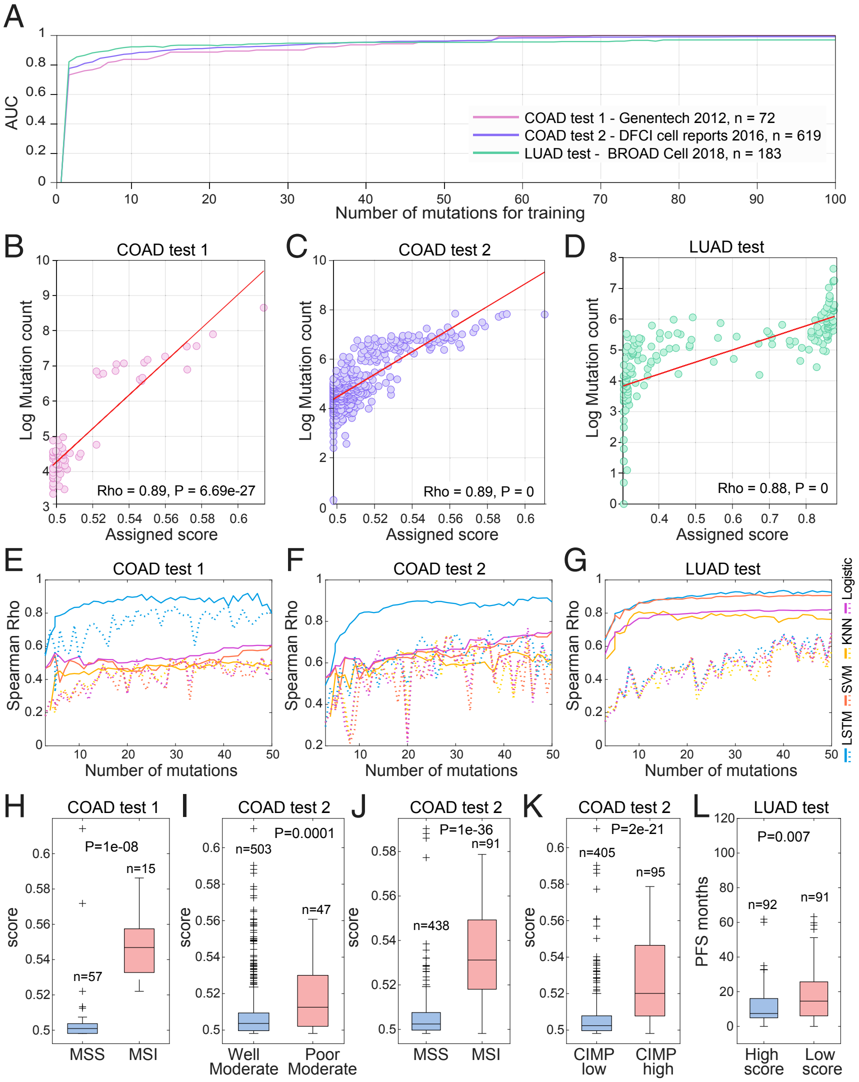

Fig. 1. Prediction of tumor mutational load from the mutational time series. (*A*) The test AUCs (*y* axis) obtained for training LSTMs on different lengths of mutation sequences (*x* axis) when starting from the latest ordered mutation, for colon and lung test sets. (*B–D*) Correlation between the score assigned by LSTMs with the 20 latest mutations in the time series (*x* axis), and the true mutational load (*y* axis, log transformed). (*E–G*) Spearman correlation between the scores assigned by different learning models and the observed mutational load (*y* axis) when using different number of mutations from these that are ordered latest in the sequence (up to 50 mutations, *x* axis). The dashed lines show the results for classifiers trained on randomly selected mutations (rather than the ordered sequence of mutations that is shown by solid lines). (*H–K*) Scores assigned by LSTM using the last 20 mutations for colon cancer patients with different clinical characteristics. (*L*) PFS of lung cancer patients in the test set, of samples assigned with high vs. low (using the median) scores with the last 20 mutations. COAD, colon adenocarcinoma; LUAD, lung adenocarcinoma.

thus could be predicted with confidence. Notably, however, the subset of driver mutations for which good prediction accuracy was achieved included DNA repair genes, such as MLH1, MSH2, MSH6, PMS2, and MUTYH for colon cancer and DDX5 and CHD1L for lung cancer, conceivably due to their effect on other mutations in the sequence (Dataset S1).

Recently, it has been shown that LSTM RNNs can be used to generate complex sequences by simply predicting one data point at a time (24). We hence reasoned that similar technique could be utilized to generate the mutational time sequence and thus could be employed for mutational data reconstruction. We used the LSTM scores to predict the occurrence of each mutation in the time series (from the latest to the earliest), to determine the occurrence of one mutation at a step in a simulated time series (see *Materials and Methods* for details). In this manner, we reconstructed 100 mutational samples for colon cancer and 100 samples for lung cancer (Datasets S2–S3). The K-means clustering analysis did not separate the simulated data from the real data (*SI Appendix*, Fig. S4 *A* and *B*), and principal component analysis (PCA) showed notable similarity between the simulated datasets and the real ones (Fig. 2 *D* and *E*), which is maintained when considering only frequently mutated genes (*SI Appendix*, Fig. S5). This similarity was observed also when using the tSNE dimensionality reduction method (40) (*SI Appendix*, Fig. S4 *C* and *D*). Considering the patterns of occurrence of the frequently mutated cancer drivers (those with the frequency of mutation in the top 10%) in the simulated data, we identified clear similarities to the observed patterns in the original data, such as the mutual exclusion of APC, KRAS, and TP53 in colon cancer and of KRAS and MGA in lung cancer (Fig. 2 *F* and *G* and *SI Appendix*, Fig. S6).

To evaluate the effect of the actual order of the mutations in the sequence function on the prediction of the preceding mutations, we repeated this analysis for the 300 last mutations, in

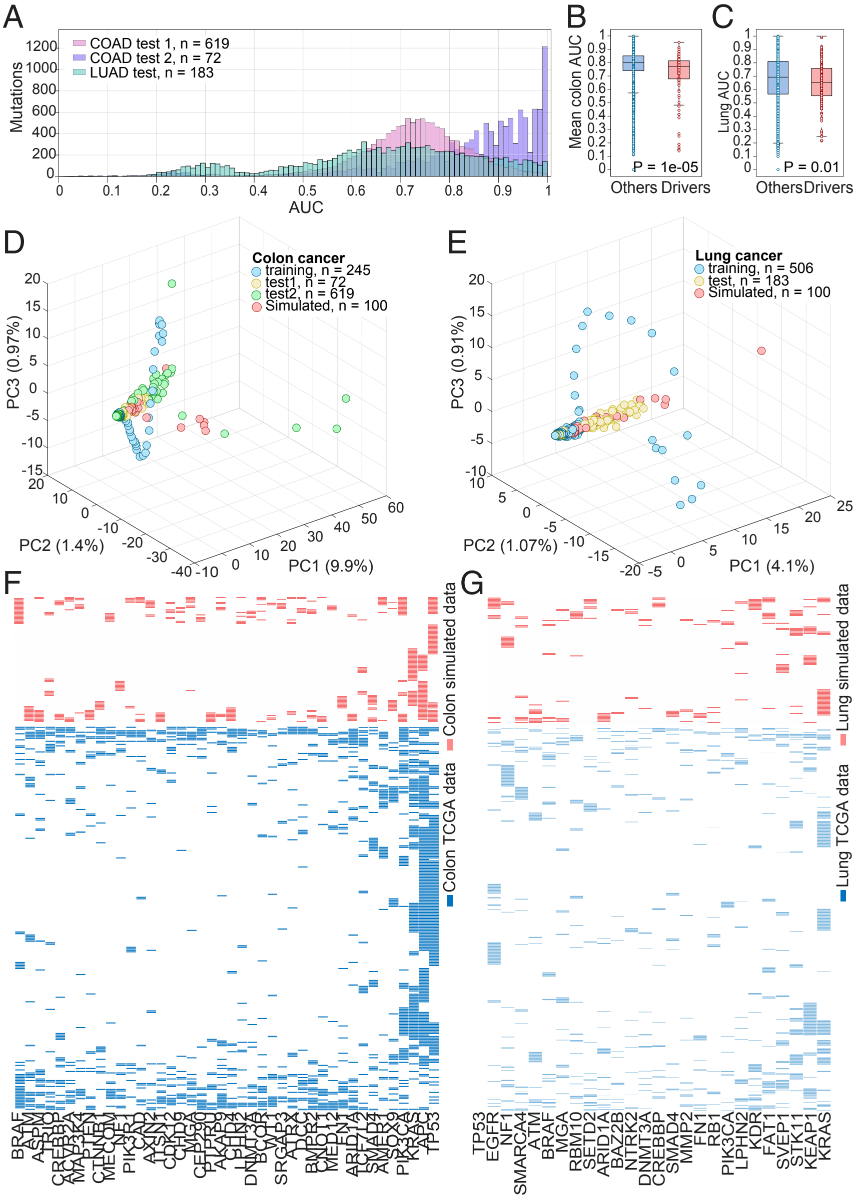

Fig. 2. Prediction and generation of the sequence of mutations. (*A*) Histogram of mutations count (*y* axis) for each performance level (AUC; *x* axis) for mutation prediction in a sequence for colon cancer test sets (pink and purple bars for test sets 1 and 2, respectively) and for lung cancer test set (green bars). (*B*) Mean AUC of mutation prediction in the sequence for the two colon test sets: comparison of drivers with all other genes. (*C*) AUC of mutation prediction in the sequence for the lung test set: comparison of drivers with all other genes. (*D* and *E*) Scatter plots of PC1–PC3 obtained by PCA applied to the combined mutational data from all datasets used and the simulated samples, for colon and lung cancers, respectively. The percentage of variance explained by each PC is indicated in parentheses. (*F* and *G*) Presence–absence patterns for the high-frequency cancer drivers in the reconstructed mutational samples (red) and the TCGA mutational data (blue) for colon and lung cancers, respectively. The samples are ordered by the hierarchical clustering results, with the Euclidean distance metric and average linkage.

both colon and lung cancers, after randomly permuting the order of mutations in the sequence up to each predicted mutation (10 random permutations for each prediction task). The comparison of the results with those obtained with the original, ordered sequence of mutations shows a dramatic drop in the prediction accuracy (paired rank-sum P value < 0.06, 0.02, and 9e-6 for colon test set1, colon test set 2, and lung test set, respectively; *SI Appendix*, Fig. S7). This is a strong indication that the actual order of mutations is important for predicting the mutational sequence.

Predicting Associations Between Major Cancer Drivers and Other Genes. Given that most mutations are predicted with high accuracy from the time sequence, we investigated the possibility that some mutations (currently classified as passenger) might also be predicted from the ordered time sequence of mutations in cancer driver genes which play key roles in tumor development. Mutations that might be predicted through their association with major cancer drivers are of obvious interest because they could potentially contribute to different aspects of tumorigenesis. To explore such potential associations, we selected well-characterized, major cancer drivers in each tumor type studied ($n = 42$ for colon and $n = 26$ for lung; *Materials and Methods*) and utilized the discrete time series of their occurrences to predict the occurrence of other mutations. This analysis yielded 354 genes for colon cancer and 273 genes for lung cancer in which mutations could be predicted robustly with high

AUC from multiple points in the time series of the major drivers (*Materials and Methods* and Dataset S4).

Given that these mutations are accurately predicted using a short mutational sequence that includes only the major drivers, we hypothesized that the respective genes interact with these drivers. To further characterize the potential functional connections between the major drivers and the identified associated genes, we first determined which driver contributed to the prediction of each of the identified driver-associated mutations (*Materials and Methods*) to generate a list of driver–interactor pairs for colon and lung cancers (Fig. 3*A* and Dataset S5). A STRING interactions enrichment analysis (41, 42) (*Materials and Methods*) showed that for 68 and 39 predicted driver–interactor pairs in colon and lung cancers, respectively, the interaction is validated under the multiple criteria implemented in STRING (hypergeometric *P* value ≈ 0 and 5.2675e-04 for colon and lung, respectively). When the major cancer drivers were analyzed individually, we found significant (*P* value < 0.05) STRING enrichment between about 22% of both colon and lung cancer major drivers and their predicted interactors, a result that is unlikely to be obtained by chance (Fig. 3*A*; permutation *P* value < 0.001 for both colon and lung).

The node degrees of the major cancer drivers in the network of STRING-validated interactions vary substantially within the networks inferred for each tumor type and differ between the colon and lung networks for shared drivers, with the mean degree of 3.5 for colon and 3 for lung (Fig. 3 *C* and *D*). In both networks, SMAD4, a gene encoding a protein involved in the TGF-beta signaling pathway (43), is highly connected. Most of the genes connected with SMAD4 in these networks are also involved in TGF-beta signaling, but the specific subsets of these genes differ between the colon and lung networks. Several of these interactions have been reported previously. In particular, MAP2K4 has been identified as a conditional tumor suppressor in lung adenocarcinomas but not in colon cancer (44), whereas mutations in HIF1A are associated with poor prognosis in colon cancer (45); furthermore, HIF1A protein physically interacts with SMAD4 under hypoxic conditions in colon cancer cell lines (46). The degree of TP53 is considerably higher in the lung cancer network than in the colon cancer network, possibly due to different TP53 mutants that are observed in these tumors (47) that have different interacting partners (48). Notably, in both the colon and the lung networks, TP53, SMAD4, ATM, and NF1 belong to the same strongly connected network module, whereas major tumor-specific cancer drivers such as APC (colon) and EGFR (lung) are disconnected.

Next, we systematically assessed whether the predicted interactors of the major cancer drivers are involved in the same or similar processes with the corresponding drivers. Using GO (Gene Ontology) enrichment (49, 50), we find that for both colon and lung cancers, all major drivers share a significant (with hypergeometric *P* value < 0.05) overlap of the sets of GO processes with their predicted interactors (Fig. 4 *A* and *B*). This level of enrichment in shared processes is unlikely to be reached by chance (permutation *P* < 0.001 for both colon and lung).

For the 13 major drivers that are shared between colon and lung cancers, we identified 1,119 GO significantly enriched processes (enrichment *P* value < 0.01; Dataset S6). For the predicted interactors of these major drivers, 240 GO processes are highly enriched in the case of colon cancer, and 229 processes are highly enriched for lung cancer. Of these, 133 GO processes are shared between the interactors from colon and lung cancers (hypergeometric *P* value ≈ 0), and among them, 34 GO processes are shared with the set of processes enriched among drivers (hypergeometric *P* value = 0.02; Fig. 4*C* and Dataset S6). Among the GO processes that are shared between these major cancer drivers and their interactors are cell motility pathways, regulation of cell development, differentiation and growth factor stimulation, and mesenchyme development processes. It appears plausible that the predicted interactions of diverse genes with the major drivers promote tumor growth and aggressiveness through the modulation of those processes.

Driver–Interactor Modules Are Associated with Patients' Survival. We next investigated the bipartite network of major drivers and their predicted interactors. In an attempt to infer the contributions of these interactions to tumor fitness through patients' survival, we first sought to identify modules of major drivers that share

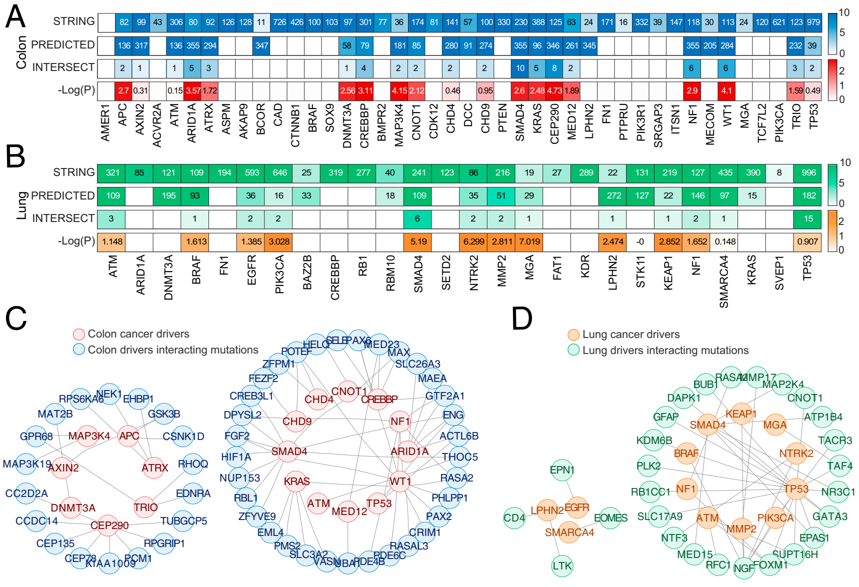

Fig. 3. STRING-validated interactions of major drivers in colon and lung cancers. (*A* and *B*) Heatmaps showing, for each major colon and lung cancer driver, respectively, the number of STRING interactions within the mutational data (first row), the number of predicted interactions (second row), the number of interactions in the intersection (third row), and the log-transformed hypergeometric *P* value (fourth row). (*C* and *D*) The networks of STRING-validated interactions for colon and lung cancers, respectively.

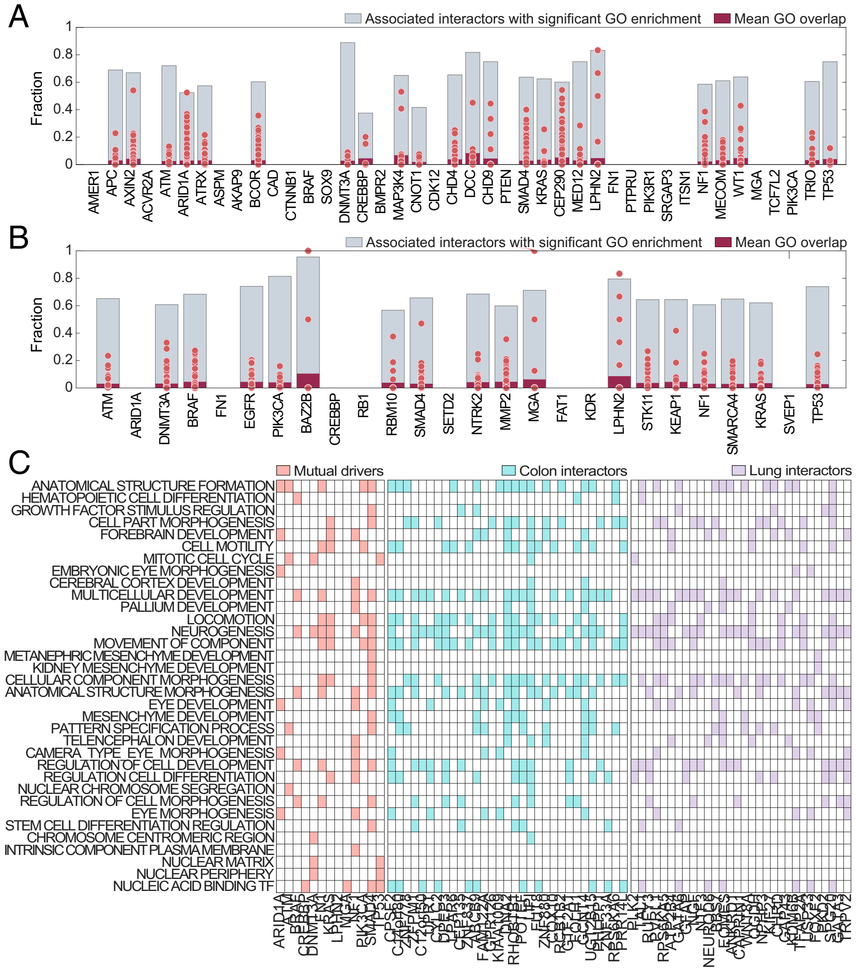

Fig. 4. GO enrichment of the predicted interactions of major cancer drivers. (*A* and *B*) The gray bars show the fraction of GO processes associated with each major cancer driver that are significantly shared with its predicted interactors, for colon and lung cancers, respectively. The dot plots show the percentage of overlap of GO processes between each major driver and its predicted interactors (the red bar shows the mean of this distribution). (*C*) Heatmaps for GO processes enriched with the shared colon and lung major drivers and their interactors (presented are the interactors that are most strongly associated with these GO processes; for full information, see Dataset S6).

interactors. To this end, we performed a heuristic search for driver modules, aiming to cover the maximum number of drivers in each tumor type. Specifically, we searched for the maximum partition of each tumor graph into disjoint subgraphs, such that each subgraph is complete (more precisely, the relation between the modules of drivers and interactors is complete; see *Materials and Methods* for details). In colon cancer, we identified 3 mutually exclusive modules of drivers (with mutually exclusive pairwise interactions), which together cover 22 of the 23 major colon cancer drivers with predicted interactions (all but the DCC gene), and 47 interactors (Fig. 5 *A*–*C*). For each module, we then identified the TCGA colon samples in which the given module is highly mutated (see *Materials and Methods* for details). We performed Kaplan–Meier survival analysis comparing the survival curves between samples with high vs. low number of mutations of the predicted interactors within each module, conditional on the drivers in the respective module being highly mutated. Strikingly, we found that the high mutation rate of interactors of each of the three modules is associated with poor survival when the driver component of the module is mutated (Fig. 5 *D*–*F*). For some of these shared interactors, we also find that individual mutations are significantly associated with lower survival rate in the context where their driver module is highly mutated (Fig. 5 *G*–*I*). Among these, SIX4 expression has been shown to correlate with lymph node metastasis and late stage and unfavorable prognosis of colorectal cancer (51), and PBX3 mutations have been identified in colorectal tumor cells undergoing epithelial–mesenchymal transition and have been shown to be associated with poor prognosis (52).

For the lung adenocarcinoma mutational data, we detected two mutually exclusive modules of drivers (with mutually exclusive pairwise interactions) that together cover 10 of the 18 major lung cancer drivers with predicted interactions, with 47 interactors of these drivers (Fig. 6 *A* and *B*). Similarly to the observations on colon cancer, a high number of mutations in the interactors of each driver module is associated with poor survival in samples where the corresponding drivers are mutated (Fig. 6 *C* and *E*) and with lower PFS (Fig. 6 *D* and *F*). For four of the predicted interactors of module 1 (but none for module 2), we also find that some of the individual interacting mutations are significantly associated with lower survival rate when drivers from this module are mutated (Fig. 6*G*). One of such interactors is EPN1, an Epsin family member shown to regulate tumor progression (53).

Discussion

Most epithelial cancers are preceded by premalignant lesions, which frequently display mutations in cancer driver genes (54–57).

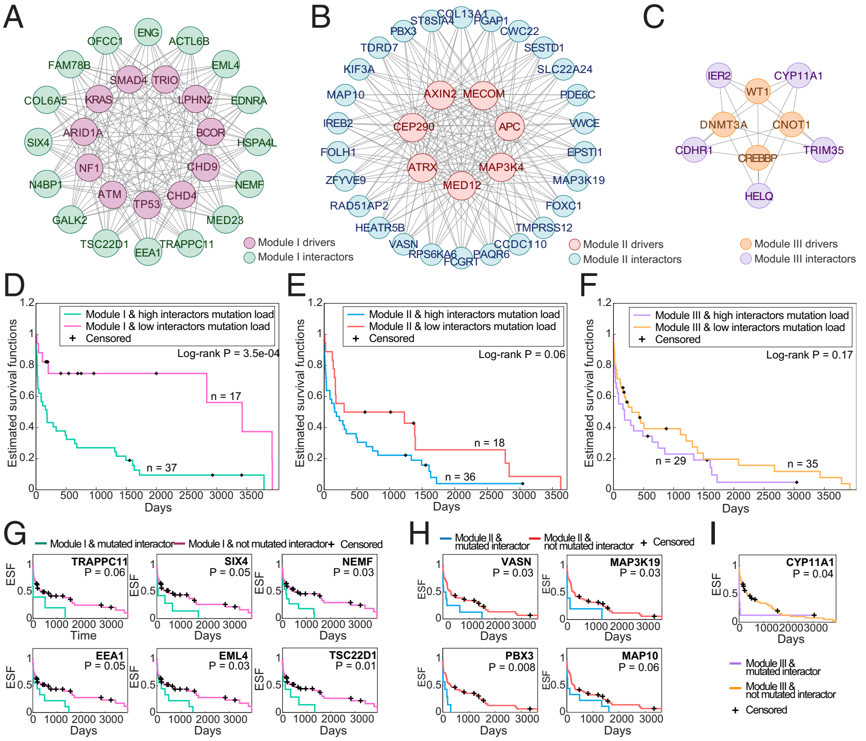

Fig. 5. Modules of drivers and interactors in colon cancer. (*A*–*C*) The complete networks of interactions between modules of major colon cancer drivers (modules I–III, respectively) and their predicted shared interactors. (*D*–*F*) Kaplan–Meier survival curves of TCGA colon cancer samples with high mutation rate of drivers modules I–III, respectively, with high vs. low number of mutation in the interactors of these modules (defined by the median). (*G*–*I*) Kaplan–Meier survival curves of TCGA colon cancer samples with vs. without mutations in individual interactors of these modules.

Nevertheless, early and late events are not universally characterized in most epithelial tumors, with the exception of colorectal cancer, where tumor progression has been thoroughly dissected in terms of stepwise accumulation of somatic mutations. This phenomenon starts from transformation of normal epithelium to an adenoma, proceeding to in situ carcinoma and ultimately to invasive and metastatic tumor, where specific mutations mark each step in this tumorigenic transformation (17, 19). It is hence reasonable to surmise that time series prediction approaches could be valuable when applied to the sequence of these genetic events in tumors.

RNNs, and particularly gated RNN architectures such as LSTMs, have recently shown promising results in learning long-term dependencies of sequences for multiple tasks of classification (58, 59) and for data labeling and synthesis (24, 60). Here we show that the mutational time series could be utilized via LSTM networks to achieve good performance in otherwise difficult prediction tasks. Using the estimated order of mutations appearance in tumor evolution, we demonstrate that end point conditions, such as the mutational burden and clinical phenotypes, could be predicted from a limited number of mutations. The nonlinear relations learned by the networks, together with the discrete representation of the data, enable performance that is significantly superior to the previous models built for this task (61, 62). The model can learn intricate dynamics of the mutational sequence in tumor evolution that can subsequently be used for the reconstruction of mutation sequences. When more data become available and the relevant neural network models are further refined, similar approaches could be applied to reconstruct data on actual DNA sequences and could thus extensively contribute to our understanding of tumor evolution. It is worth pointing out that applications of LSTMs generally take advantage of much larger data for training because in these, the input alphabet as well as possible labels are from a much larger range, thus substantially increasing the size of the dataset required for training. In our analysis, discretizing the mutational data and maintaining discrete labels (i.e., both input size and labels are always of size 2) generates a simple enough problem that can be approached even with limited amounts of data (Table 1). We also found that LSTMs perform much better than linear classifiers, such as support vector machine (SVM), for the prediction of the mutational load. This is likely to be the case because LSTMs learn nonlinear relationships between mutations in the given sequence, rather than defining a linear separating hyperplane, with the underlining assumption that the mutational load can be predicted using a linear function of mutations.

A potentially important contribution of this approach is the identification of interactors of the major cancer drivers. Here we predict many such interactors and show that they are significantly enriched with STRING interactions and show nonrandom overlap of GO processes with the corresponding major drivers. Anecdotally, at least some of the better characterized interactors were found to be involved in the same pathway with the corresponding driver. In contrast, we find that the predicted interactors of each drivers are not located on similar chromosomal arms (*SI Appendix*, Fig. S8), suggesting that these are mostly functional, rather than physical interactions. Furthermore, and most strikingly, we found that mutations in these predicted interactors, in the presence of the corresponding driver mutations, are associated with poor survival and, thus, are likely to confer growth advantage at the respective steps of tumor progression. We identified unique modules of major drivers and their interactors for both colon and lung cancer. The strong correlation with patients' survival suggests that these interactions are clinically relevant and, if further tested, could potentially be used for patients' stratification and clinical decision-making. The identified interactors can be regarded as secondary drivers whose oncogenic activity is conditional to and associated with the occurrence of

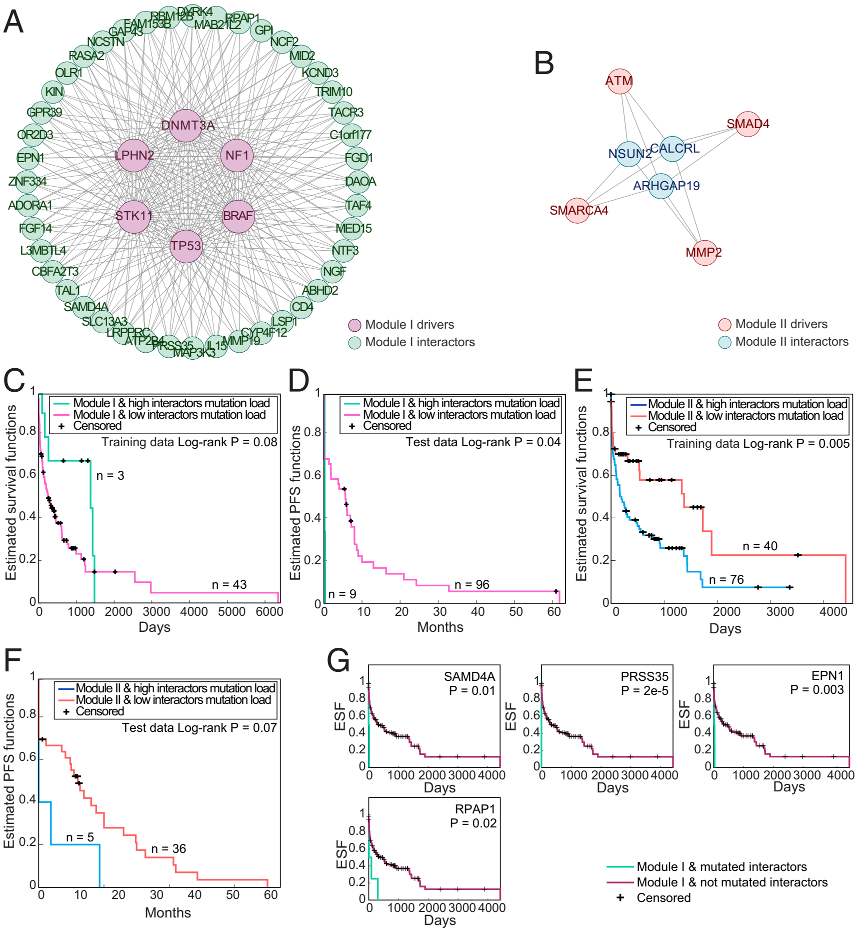

Fig. 6. Modules of drivers and interactors in lung cancer. (*A* and *B*) The complete networks of interactions between modules of major lung cancer drivers (modules 1 and 2, respectively) and their predicted shared interactors. (*C*–*F*) Kaplan–Meier survival curves of TCGA colon cancer samples (overall survival; *C* and *E*) and the lung cancer test set samples (PFS; *D* and *F*) of samples with high mutation rate of driver modules I and II, respectively, with high vs. low number of mutation in the interactors of these modules (defined by the median). (*G*) Kaplan–Meier survival curves of TCGA colon cancer samples with high mutation rate of driver module I (defined by the median), with vs. without mutations of individual interactors of module I.

mutations in major drivers for the respective cancer types. Hence, these conditional drivers could be readily predicted notwithstanding the low accuracy of prediction for most of the major drivers themselves.

To summarize, in this work, we present the application of LSTM for learning the stepwise sequence of mutations in tumors. This approach is shown to efficiently tackle several tasks that are not amenable to standard techniques, such as prediction of the occurrence of mutations and reconstruction of mutational data. Our findings reveal a unidirectional relation between driver and passenger mutations: drivers determine the course of tumorigenesis, and their occurrence is difficult to predict from the rest of the mutational landscape, whereas passengers are often linked to specific drivers and so can be predicted with confidence. Thus, drivers are indeed in the driver's seat and bring with them a host of associated passengers, some of which could be secondary, conditional drivers. The present results support the notion that long-term dependencies between genes involved in tumorigenesis and cancer progression are widespread in tumor evolution and can be learned from the mutational sequence using LSTM networks and similar approaches. We show that this notion holds for colon cancer, where the stepwise process of mutation acquisition is established, but also for lung cancer for which such a stepwise model has been suggested but remains controversial. Similar strategies could be readily employed for other tumor types and different types of biological data to advance our understanding of tumor initiation and progression through the dissection of the sequence of evolutionary events.

Materials and Methods

Mutation Data. Mutational data of colon and lung adenocarcinoma from TCGA were used for training throughout this work and were obtained from Xena browser (63). Datasets used for testing were obtained from cBioPortal (64, 65). The datasets are summarized in Table 1, spanning 1,626 samples overall, each derived from a distinct tumor.

Driver Mutation Lists for Lung and Colon Cancers. The union of driver genes obtained from (66) and (67) was used for lung and colon cancers (Dataset S7).

Preprocessing and Sorting Mutational Data. To generate the binary sequence of mutations, we first discretize the mutational data assigning 1 for each nonsynonymous mutation. For each cancer type, we then consider all genes that are mutated at least once in all datasets used (totals of 12,322 and 12,327 mutated genes for colon and lung cancer, respectively).

To sort the mutations of colon and lung adenocarcinoma by their estimated temporal order, we evaluate the following function for each TCGA training dataset:

$$Order\_Score(gi) = \sum_{all\ genes\ gj} \frac{\sum_{all\ samples\ s} s_{gi} = 1 \,\&\, s_{gj} = 1}{\sum_{all\ samples\ s} s_{gi} = 1 \,\&\, s_{gj} = 0}, \qquad [1]$$

where $sg_i = 1$ if sample *s* has a mutation in gene *gi*. We calculate the order score for each considered gene using the TCGA datasets that are used for training and use that score to sort the test datasets. Order scores calculated using the test datasets were found to significantly correlate with that derived from the training set for both colon and lung cancers (*SI Appendix*, Fig. S9).

LSTM Machines. Each LSTM network unit defines a time *t* in a sequence (the subscript *t* denotes the mutation that is ordered *t* in the tumor evolution via the order score defined above), and is composed of the following components:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f), \qquad [2]$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i), \qquad [3]$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o), \qquad [4]$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \sigma_c(W_f x_t + U_f h_{t-1} + b_f), \qquad [5]$$

$$h_t = f_t \odot \sigma_c(c_t), \qquad [6]$$

where the initial values are $c_0 = 0$ and $h_0 = 0$. $\odot$ denotes the Hadamard product.

x_t are the input vectors to the LSTM unit (an ordered sequence of mutations). f_t, i_t, and o_t are the activation vectors for the forget gate, input gate and output gate, respectively. h_t is the output vector of the LSTM unit, and c_t is cell state vector. *W* and *U* are the weight matrices, and *b* represents the bias matrices that are learned during training. σ represents the nonlinear functions, where σ_g is the gate activation, sigmoid function $\sigma_g(x) = (1 + e^{-x})^{-1}$, and σ_c is the state activation, *tanh* (hyperbolic tangent) function,$\tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1}$.

All LSTM networks used in this work are sequence-to-label LSTMs with five hidden layers and were trained using Adam optimizer (68), where the maximum number of epochs for training was set to 100 for the mutational load prediction and to 10 for all other prediction tasks. The minibatch size used for each training iteration was set to 27, with a standard gradient-clipping threshold set to 1.

Training LSTMs to Predict the Mutational Load. We train LSTMs for the sequence up to each time point when starting once from the mutation ordered last, and second from the mutation ordered first in the sequence. The two-categorical labels $L_s \in \{0,1\}$ are defining low (0, <median) vs. high mutational load (1, >median). For each time point *t*, an $LSTM_t$ is trained on the training set using the sequence up to *t* and is then applied to the test set to predict the mutational load using the time sequence of mutations up to *t*. For each time point *t*, the resulting scores are used to evaluate the performance via two measures: (*i*) the AUC_t resulting from an ROC curve, predicting the low vs. high categories in the test, and (*ii*) the Spearman rank correlation coefficient ρ_t between the $LSTM_t$ scores that are assigned to each test sample and the actual mutational load of the sample.

Training KNN, SVM, and Logistic Regression Classifiers to Predict the Mutational Load. KNN (with K = 5), SVM (using linear kernel), and logistic regression classifiers were trained on the training sets using (*i*) sequences of the latest ordered mutations that were used as input to the LSTMs and (*ii*) sequences of 50 randomly selected mutations. These were trained to predict low vs. high mutational load as described for the LSTMs and applied to the test sets, and the resulting classification scores were correlated with the true mutational load via Spearman rank correlation.

Training LSTMs to Predict the Occurrence of Succeeding Mutations in the Time Sequence. To predict the occurrence of a mutation ordered *t* in the sequence, we train LSTM for the sequence starting from the mutation ordered last up to time point $t+1$ using the training sets and test their performance for predicting these occurrences in the test sets, where the two-categorial labels $L_s \in \{0,1\}$ are defining the occurrence of the mutation.

Training LSTM Networks to Simulate Mutational Data. To synthesize mutational data, we use the full mutational data for each cancer type. We reconstruct 100 simulated mutational samples for each tumor type, one mutation at a step, from the last-ordered mutation to the first.

For the last-ordered mutation we randomly assign 1 to reconstructed samples corresponding to the frequency of the mutation in the genuine datasets. To assign the occurrence of every other mutation *t* to the reconstructed samples, we train $LSTM_t$ to predict the occurrence of *t* from the sequence starting from the mutation ordered last up to time point $t+1$ and apply $LSTM_t$ to the simulated sequence (synthesized up to time point $t+1$), to obtain a vector of scores predicting the occurrence of mutation *t* in each reconstructed sample. We then use $freq_t$, the frequency of the *t* ordered mutation in the genuine datasets, and assign 1 to the $freq_t$ mutations that were assigned with highest scores by $LSTM_t$ when applied to the reconstructed sequences.

PCA was applied to the integration of all datasets (Fig. 2 *D* and *E* and *SI Appendix*, Fig. S5 *A–D*) or, when inferring the PCA coefficient, without the training sets (*SI Appendix*, Fig. S5 *E* and *F*).

Training LSTMs to Identify Mutations That Interact with the Major Cancer Drivers and Assigning Major Drivers to Interacting Mutations. We select the driver genes in which mutations are observed frequently in our training sets (top 0.1 percentile). These genes are defined as major drivers and are used as an ordered sequence of mutations (*n* = 42 for colon and *n* = 26 for lung; Dataset S4). The sequence of occurrences of these major drivers is used to predict the occurrence of other mutations, excluding those with very low frequency (genes that are mutated in three samples or less) as the prediction of those could be obtained easily by chance. We hence trained 42 LSTMs for colon and 26 for lung (using the sequence of major drivers up to each time point). The genes that could be predicted with AUC > 0.85 for the test set repeatedly, from multiple locations in the sequence of drivers, are selected as predicted interactors of the major drivers.

We then use the scores produced by the LSTMs where the driver-interacting gene is well predicted and correlate them with the occurrence of each of the major drivers. The major drivers whose occurrence is significantly correlated (Spearman *P* value < 0.05) with the LSTM scores predicting a given driver-interacting gene are combined with it into driver–interactor pairs. A detailed graphical schema describing the steps of this analysis can be found in *SI Appendix*, Fig. S10

Enrichment with STRING Interactions and GO Analysis. To investigate whether the pairs of major drivers and their predicted interactors are enriched with established interactions, we performed the following analyses: (*i*) STRING enrichment, in which hypergeometric enrichment analysis was performed for each major driver gene, to find if its LSTM-predicted interactors are enriched with its interactors from the STRING database, and (*ii*) GO enrichment, where for each major driver, we calculated the percentage of its associated interactors that share a significant number of GO processes with it (hypergeometric *P* value < 0.05) and the mean percentage of overlapping GO pathways with its interactors. We then calculate an empirical *P* value from 1,000 repetitions, with drivers randomly assigned to the identified interacting mutations (maintaining the same degree).

Modules of Major Drivers and Interacting Genes. To cover as many drivers as possible, we created modules via a heuristic search using 10,000 repetitions of the following genetic algorithm. Starting with a randomly selected major driver, in each round, we randomly selected a major driver that had not yet been added to the module and added it to the module if its addition did not decrease the number of mutual module-interactors by more than 20%. The round ended when 100 random selections were not added to the current module or when the number of interactors of a module was less than 3. Finally, we investigated the 10,000 modules and selected those that together cover maximal number of major drivers such that the modules were mutually exclusive with respect to both drivers and interactors.

Survival Analyses. All Kaplan–Meier analyses are performed by comparing the survival of patients with high scores (module mutational count larger than median) to those with low scores, using a one-sided log-rank test.

Code Availability. All code was implemented in MATLAB_R2018a using Deep Learning Toolbox and is publicly available through GitHub: https://github.com/noamaus/LSTM-Mutational-series.

ACKNOWLEDGMENTS. We thank E.V.K. group members for helpful discussions. The authors' research is supported by the Intramural Research Program of the National Institutes of Health (National Library of Medicine).

1. Vogelstein B, Kinzler KW (1993) The multistep nature of cancer. *Trends Genet* 9:138–141.
2. Farber E (1984) The multistep nature of cancer development. *Cancer Res* 44: 4217–4223.
3. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. *Nature* 458: 719–724.
4. Vogelstein B, et al. (2013) Cancer genome landscapes. *Science* 339:1546–1558.

5. Pleasance ED, et al. (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 463:191–196.
6. McFarland CD, Korolev KS, Kryukov GV, Sunyaev SR, Mirny LA (2013) Impact of deleterious passenger mutations on cancer progression. *Proc Natl Acad Sci USA* 110:2910–2915.
7. Persi E, Wolf YI, Leiserson MDM, Koonin EV, Ruppin E (2018) Criticality in tumor evolution and clinical outcome. *Proc Natl Acad Sci USA* 115:E11101–E11110.
8. Tanaka T (2009) Colorectal carcinogenesis: Review of human and experimental animal studies. *J Carcinog* 8:5.
9. Loeb LA, Harris CC (2008) Advances in chemical carcinogenesis: A historical review and prospective. *Cancer Res* 68:6863–6872.
10. Knudson AG (2001) Two genetic hits (more or less) to cancer. *Nat Rev Cancer* 1:157–162.
11. Teixeira MR, Heim S (2005) Multiple numerical chromosome aberrations in cancer: What are their causes and what are their consequences? *Semin Cancer Biol* 15:3–12.
12. Gillies RJ, Verduzco D, Gatenby RA (2012) Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. *Nat Rev Cancer* 12:487–493.
13. Vogelstein B, et al. (1988) Genetic alterations during colorectal-tumor development. *N Engl J Med* 319:525–532.
14. Armaghany T, Wilson JD, Chu Q, Mills G (2012) Genetic alterations in colorectal cancer. *Gastrointest Cancer Res* 5:19–27.
15. Fearon ER (1992) Genetic alterations underlying colorectal tumorigenesis. *Cancer Surv* 12:119–136.
16. Lurje G, Zhang W, Lenz H-J (2007) Molecular prognostic markers in locally advanced colon cancer. *Clin Colorectal Cancer* 6:683–690.
17. Noguchi M (2010) Stepwise progression of pulmonary adenocarcinoma–Clinical and molecular implications. *Cancer Metastasis Rev* 29:15–21.
18. Sanada Y, et al. (2006) Histopathologic evaluation of stepwise progression of pancreatic carcinoma with immunohistochemical analysis of gastric epithelial transcription factor SOX2: Comparison of expression patterns between invasive components and cancerous or nonneoplastic intraductal components. *Pancreas* 32:164–170.
19. Yatabe Y, Borczuk AC, Powell CA (2011) Do all lung adenocarcinomas follow a stepwise progression? *Lung Cancer* 74:7–11.
20. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9:1735–1780.
21. Williams RJ, Zipser D (1989) A learning algorithm for continually running fully recurrent neural networks. *Neural Comput* 1:270–280.
22. Gers FA, Schmidhuber J, Cummins F (2000) Learning to forget: Continual prediction with LSTM. *Neural Comput* 12:2451–2471.
23. Schmidhuber J, Wierstra D, Gomez F (2005) Evolino: Hybrid neuroevolution/optimal linear search for sequence learning. *IJCAI International Joint Conference on Artificial Intelligence,* (IJCAI, Edinburgh), pp 853–858.
24. Graves A (2013) Generating sequences with recurrent neural networks. arXiv:13080850.
25. Sundermeyer M, Schlueter R, Ney H (2012) LSTM neural networks for language modeling. *INTERSPEECH 2012*.
26. Lai S, Xu L, Liu K, Zhao J (2015) Recurrent convolutional neural networks for text classification. *AAAI '15 Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (AAAI Press, Menlo Park, CA), pp 2267–2273.
27. Sutskever I, Martens J, Hinton G (2011) Generating text with recurrent neural networks. *Proceedings of the 28th International Conference on Machine Learning (ICML'11),* (Omnipress, Bellevue, WA), pp 1017–1024.
28. Weinstein JN, et al.; Cancer Genome Atlas Research Network (2013) The cancer genome Atlas pan-cancer analysis project. *Nat Genet* 45:1113–1120.
29. Attolini CS, et al. (2010) A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proc Natl Acad Sci USA* 107:17604–17609.
30. Cheng YK, et al. (2012) A mathematical methodology for determining the temporal order of pathway alterations arising during gliomagenesis. *PLOS Comput Biol* 8:e1002337.
31. Desper R, et al. (2000) Distance-based reconstruction of tree models for oncogenesis. *J Comput Biol* 7:789–803.
32. Höglund M, et al. (2001) Multivariate analyses of genomic imbalances in solid tumors reveal distinct and converging pathways of karyotypic evolution. *Genes Chromosomes Cancer* 31:156–171.
33. Lin SH, et al. (2018) The somatic mutation landscape of premalignant colorectal adenoma. *Gut* 67:1299–1305.
34. Bamford S, et al. (2004) The COSMIC (catalogue of somatic mutations in cancer) database and website. *Br J Cancer* 91:355–358.
35. Tate JG, et al. (2018) COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids Res* 47:D941–D947.
36. Wolff RK, et al. (2018) Mutation analysis of adenomas and carcinomas of the colon: Early and late drivers. *Genes Chromosomes Cancer* 57:366–376.
37. Aithal A, et al. (2018) MUC16 as a novel target for cancer therapy. *Expert Opin Ther Targets* 22:675–686.
38. Sur I, Neumann S, Noegel AA (2014) Nesprin-1 role in DNA damage response. *Nucleus* 5:173–191.
39. Bozic I, et al. (2010) Accumulation of driver and passenger mutations during tumor progression. *Proc Natl Acad Sci USA* 107:18545–18550.
40. van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9:2579–2605.
41. Szklarczyk D, et al. (2015) STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 43:D447–D452.
42. Szklarczyk D, et al. (2017) The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* 45:D362–D368.
43. Zawel L, et al. (1998) Human Smad3 and Smad4 are sequence-specific transcription activators. *Mol Cell* 1:611–617.
44. Ahn Y-H, et al. (2011) Map2k4 functions as a tumor suppressor in lung adenocarcinoma and inhibits tumor cell invasion by decreasing peroxisome proliferator-activated receptor γ2 expression. *Mol Cell Biol* 31:4270–4285.
45. Baba Y, et al. (2010) HIF1A overexpression is associated with poor prognosis in a cohort of 731 colorectal cancers. *Am J Pathol* 176:2292–2301.
46. Papageorgis P, et al. (2011) Smad4 inactivation promotes malignancy and drug resistance of colon cancer. *Cancer Res* 71:998–1008.
47. Harris CC (1996) p53 tumor suppressor gene: At the crossroads of molecular carcinogenesis, molecular epidemiology, and cancer risk assessment. *Environ Health Perspect* 104:435–439.
48. Freed-Pastor WA, Prives C (2012) Mutant p53: One name, many proteins. *Genes Dev* 26:1268–1286.
49. The Gene Ontology Consortium (2017) Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res* 45:D331–D338.
50. Ashburner M, et al.; The Gene Ontology Consortium (2000) Gene ontology: Tool for the unification of biology. *Nat Genet* 25:25–29.
51. Li G, Hu F, Luo X, Hu J, Feng Y (2017) SIX4 promotes metastasis via activation of the PI3K-AKT pathway in colorectal cancer. *PeerJ* 5:e3394.
52. Lamprecht S, et al. (2018) PBX3 is part of an EMT regulatory network and indicates poor outcome in colorectal cancer. *Clin Cancer Res* 24:1974–1986.
53. Tessneer KL, et al. (2013) Epsin family of endocytic adaptor proteins as oncogenic regulators of cancer progression. *J Cancer Res Updates* 2:144–150.
54. Ryan BM, Faupel-Badger JM (2016) The hallmarks of premalignant conditions: A molecular basis for cancer prevention. *Semin Oncol* 43:22–35.
55. Shin DM, et al. (1994) Activation of p53 gene expression in premalignant lesions during head and neck tumorigenesis. *Cancer Res* 54:321–326.
56. Viola MV, Fromowitz F, Oravez S, Deb S, Schlom J (1985) ras oncogene p21 expression is increased in premalignant lesions and high grade bladder carcinoma. *J Exp Med* 161:1213–1218.
57. Enomoto T, et al. (1991) K-ras activation in premalignant and malignant epithelial lesions of the human uterus. *Cancer Res* 51:5308–5314.
58. Lipton ZC, Berkowitz J, Elkan C (2015) A critical review of recurrent neural networks for sequence learning. arXiv:1506.00019.
59. Shi X, et al. (2015) Convolutional LSTM network: A machine learning approach for precipitation nowcasting. arXiv:1506.04214.
60. Graves A, Fernandez S, Gomez F, Schmidhuber J (2006) Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning (ICML),* (ACM, New York), pp 369–376.
61. Lyu GY, Yeh YH, Yeh YC, Wang YC (2018) Mutation load estimation model as a predictor of the response to cancer immunotherapy. *NPJ Genomic Med* 3:12.
62. Roszik J, et al. (2016) Novel algorithmic approach predicts tumor mutation load and correlates with immunotherapy clinical outcomes using a defined gene mutation set. *BMC Med* 14:168.
63. Goldman M, et al. (2018) The UCSC Xena platform for cancer genomics data visualization and interpretation. bioRxiv:10.1101/326470.
64. Gao J, et al. (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 6:pl1.
65. Cerami E, et al. (2012) The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2:401–404.
66. Bailey MH, et al.; MC3 Working Group; Cancer Genome Atlas Research Network (2018) Comprehensive characterization of cancer driver genes and mutations. *Cell* 173:371–385.e18.
67. Rubio-Perez C, et al. (2015) In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell* 27:382–396.
68. Kingma DP, Ba JL (2015) Adam: A method for stochastic optimization. arXiv:1412.6980.
69. Giannakis M, et al. (2016) Genomic correlates of immune-cell infiltrates in colorectal carcinoma. *Cell Rep* 15:857–865.
70. Seshagiri S, et al. (2012) Recurrent R-spondin fusions in colon cancer. *Nature* 488:660–664.
71. Imielinski M, et al. (2012) Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* 150:1107–1120.