LOTUS: a Single- and Multitask Machine Learning Algorithm for the Prediction of Cancer Driver Genes

Olivier Collier[1,*], Véronique Stoven[2,3,4], Jean-Philippe Vert[5,2,†]

1 Modal'X, UPL, Univ Paris Nanterre, F-92000 Nanterre, France
2 MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France
3 Institut Curie, F-75248 Paris Cedex 5, France
4 INSERM U900, F-75248 Paris Cedex 5, France
5 Google Brain, F-75009 Paris, France

* olivier.collier@parisnanterre.fr, † jpvert@google.com

Abstract

Cancer driver genes, *i.e.,* oncogenes and tumor suppressor genes, are involved in the acquisition of important functions in tumors, providing a selective growth advantage, allowing uncontrolled proliferation and avoiding apoptosis. It is therefore important to identify these driver genes, both for the fundamental understanding of cancer and to help finding new therapeutic targets or biomarkers. Although the most frequently mutated driver genes have been identified, it is believed that many more remain to be discovered, particularly for driver genes specific to some cancer types.

In this paper, we propose a new computational method called LOTUS to predict new driver genes. LOTUS is a machine-learning based approach which allows to integrate various types of data in a versatile manner, including information about gene mutations and protein-protein interactions. In addition, LOTUS can predict cancer driver genes in a pan-cancer setting as well as for specific cancer types, using a multitask learning strategy to share information across cancer types.

We empirically show that LOTUS outperforms four other state-of-the-art driver gene prediction methods, both in terms of intrinsic consistency and prediction accuracy, and provide predictions of new cancer genes across many cancer types.

Author summary

Cancer development is driven by mutations and dysfunction of important, so-called cancer driver genes, that could be targeted by specific therapies. While a number of such cancer genes have already been identified, it is believed that many more remain to be discovered. To help prioritize experimental investigations of candidate genes, several computational methods have been proposed to rank promising candidates based on their mutations in large cohorts of cancer cases, or on their interactions with known driver genes in biological networks. We propose LOTUS, a new computational approach to identify genes with high oncogenic potential. LOTUS implements a machine learning

approach to learn an oncogenic potential score from known driver genes, and brings two novelties compared to existing methods. First, it allows to easily combine heterogeneous sources of information into the scoring function, which we illustrate by learning a scoring function from both known mutations in large cancer cohorts and interactions in biological networks. Second, using a multitask learning strategy, it can predict different driver genes for different cancer types, while sharing information between them to improve the prediction for every type. We provide experimental results showing that LOTUS significantly outperforms several state-of-the-art cancer gene prediction software.

Introduction

In our current understanding of cancer, tumors appear when some cells acquire functionalities that give them a selective growth advantage, allowing uncontrolled proliferation and avoiding apoptosis [1, 2]. These malignant characteristics arise from various genomic alterations including point mutations, gene copy number variants (CNVs), translocations, inversions, deletions, or aberrant gene fusions. Many studies have shown that these alterations are not uniformly distributed across the genome [3, 4], and target specific genes associated with a limited number of important cellular functions such as genome maintenance, cell survival, and cell fate [5]. Among these so-called *driver genes*, two classes have been distinguished in the literature: *tumor suppressors genes* (TSGs) and *oncogenes* (OGs) [6, Chapter 15]. TSGs, such as TP53 [7], participate in defense mechanisms against cancer and their inactivation by a genomic alteration can increase the selective growth advantage of the cell. On the contrary, alterations affecting OGs, such as KRAS [8] or ERBB2 [9], can be responsible for the acquisition of new properties that provide some selective growth advantage or the ability to spread to remote organs. Identifying driver genes is important not only from a basic biology point of view to decipher cancer mechanisms, but also to identify new therapeutic strategies and develop precision medicine approaches targeting specifically mutated driver genes. For example, Trastuzumab [10] is a drug given against breast cancer that targets the protein precisely encoded by ERBB2, which has dramatically improved the prognosis of patients whose tumors overexpress that OG.

Decades of research in cancer genomics have allowed to identify several hundreds of such cancer genes. Regularly updated databases such as the Cancer Gene Census (CGC) [11], provide catalogues of genes likely to be causally implicated in cancer, with various levels of experimental validations. Many cancer genes have been identified recently by systematic analysis of somatic mutations in cancer genomes, as provided by large-scale collaborative efforts to sequence tumors such as The Cancer Genome Atlas (TCGA) [12] or the International Cancer Genome Consortium (ICGC) [13]. Indeed, cancer genes tend to be more mutated than non-cancer genes, providing a simple guiding principle to identify them. In particular, the COSMIC database [14] is the world's largest and most comprehensive resource of somatic mutations in coding regions. It is now likely that the most frequently mutated genes have been identified [15]. However, the total number of driver genes is still a debate, and many driver genes less frequently mutated, with low penetrance, or specific to a given type of cancer are still to be discovered.

The first methods to identify driver genes from catalogues of somatic mutations simply compared genes based on somatic mutation frequencies, which was proved to be far too basic [16]. Indeed, mutations do not appear uniformly on the genome: some regions of the genome may be more affected by errors because they are more often transcribed, so that some studies actually overestimated the number of driver genes because they were expecting lower mutation rates than in reality. Mathematically, they

were formulating driver prediction as a hypothesis testing problem with an inadequate null hypothesis [17]. Several attempts have been made to adequately calibrate the null hypothesis, like [16] or [18], where it is assumed that mutations result from a mixture of several mutational processes related to different causes.

A variety of bioinformatics methods have then been developed to complete the list of pan-cancer or cancer specific driver genes. Globally, they fall into three main categories. First, a variety of "Mutation Frequency" methods such as MuSiC [19] or ActiveDriver [20] identify driver genes based on the assumption that they display mutation frequencies higher than those of a background mutation model expected for passenger mutations. However, this background rate may differ between cell types, genome positions or patients. In order to avoid such potential bias, some methods like MutSigCV [21] derive a patient-specific background mutation model, and may take into account various criteria such as cancer type, position in the genome, or clinical data. Second, "Functional impact" methods such as OncodriveFM [22] assume that driver genes have higher frequency of mutations expected to impact the protein function (usually missense mutations) than that observed in passenger genes. Third, "Pathway-based" methods consider cancer as a disease in which mutated genes occupy key roles in cancer-related biological pathways, leading to critical functional perturbations at the level of networks. For example, DriverNet [23] identifies driver genes based on their effect in the transcription networks. Although these methods tend to successfully identify the most frequently mutated genes, their overall prediction overlap is modest. Since they rely on complementary statistical strategies, one could recommend to use them in combination, as CompositeDriver allows to do [24]. The results of some of these tools are available at the Driver DB database [25].

Some methods integrate information on mutation frequency and functional impact of mutations, or other types of data such as genome position, copy number variations (CNVs) or gene expression. The underlying idea is that combining data should improve the prediction performance over tools that use a single type of information. For example, TUSON [26] or DOTS-Finder [27] combine mutation frequencies and functional impact of mutations to identify OGs and TSGs. Also in this category, the 20/20+ method [28] encodes genes with features based on their frequency and mutation types, in addition to other biological information such as gene expression level in difference cancer cell lines [29] or replication time. Then, 20/20+ predicts driver genes with a random forest algorithm, which constitutes the first attempt to use a machine learning method in this field. In [28], the authors benchmark 8 driver gene prediction methods based on several criteria including the fraction of predicted genes in CGC, the number of predicted driver genes and the consistency. Three methods proved to perform similarly on all criteria, and better than the five others: TUSON, MutSigCV, and 20/20+, validating the relevance of combining heterogeneous information to predict cancer genes.

In the present paper, we propose a new method for cancer driver gene prediction called *Learning Oncogenes and TUmor Suppressors* (LOTUS). Like 20/20+, LOTUS is a machine learning-based method, meaning that it starts from a list of known driver genes in order to "learn" the specificities of such genes and to identify new ones. In addition, LOTUS presents two unique characteristics with respect to previous work in this field. First, it combines three types of features likely to contain relevant information to predict cancer genes (mutation frequency, functional impact, and pathway-based informations). This integration of heterogeneous information is carried out in a unified mathematical and computational framework thanks to the use of kernel methods [30], and allows in principle to integrate other sources of data if available, such as transcriptomic or epigenomic information. More precisely, in our implementation, we predict cancer driver genes based not only on gene mutations features like "Mutation Frequency" and "Functional Impact" methods do, but also on known protein-protein

interaction (PPI) network like "Pathway-based" methods do. Indeed, the use of PPI information is particularly relevant since it has been reported that proteins encoded by driver genes are more likely to be involved in protein complexes and share higher "betweenness" than a typical protein [26]. Moreover, it has been successfully used by HotNet2 [31] to detect gene pathways enriched in driver genes. Second, LOTUS can predict cancer genes in a pan-cancer setting, as well as for specific cancer types, using a multitask learning strategy [32]. Although many efforts are devoted to identify cancer-specific genes based on experimental approaches, in in-silico approaches, the pan-cancer setting has been adopted by most available prediction methods, since more data are available to train models when gathering all cancer types. Prediction of drivers for specific cancer types has been less explored so far, because the number of known driver genes for a given cancer is often too small to build a reliable prediction model, and because the amount of data such as somatic mutations to train the model is smaller than in the pan-cancer setting. However, the search for cancer specific driver genes is relevant, because cancer is a very heterogeneous disease: different tumorigenic processes seem to be at work in different tissue types, and consequently, each cancer type probably has its own list of driver genes [15]. LOTUS implements a multitask algorithm that predicts new driver genes for a given cancer type based on its known driver genes, while also taking into account the driver genes known for other types of cancer according to their similarities with the considered type of cancer. Such approaches are of particular interest when the learning data are scarce in each individual tasks: they increase the amount of data available for each task and thus perform statistically better. To our knowledge, while a similar approach was used to predict disease genes across general human diseases [33], this is the first time a multitask machine learning algorithm is used for the prediction of cancer driver genes.

We compare LOTUS to four state-of-the art cancer prediction methods. We show that LOTUS outperforms the state-of-the-art in its ability to identify novel cancer genes, and clarify the benefits of heterogeneous data integration and of the multitask learning strategy to predict cancer type-specific driver genes. Finally, we provide predictions of new cancer genes according to LOTUS, as well as supporting evidence that those predictions are likely to contain new cancer genes.

Results

LOTUS, a new method for pan-cancer and cancer specific driver gene prediction

We propose LOTUS, a new method that predicts cancer driver genes. LOTUS is a machine learning-based method that estimates a scoring function to rank candidate genes by decreasing probability for them to be OGs or TSGs, given a training set of known OGs and TSGs. The score of a candidate gene is a weighted sum of similarities between the candidate gene and the known driver genes, where the weights are optimized by a one-class support vector machine (OC-SVM) algorithm. The similarities between genes are calculated based on gene features that are derived from the analysis of somatic mutation patterns in the genes (see Materials ans Methods section for a description of these features), or from the relative positions of genes in a PPI network, or from both; the mathematical framework of kernel methods allows to simply combine heterogeneous data about genes (i.e., patterns of somatic mutations and PPI information) in a single model.

Another salient feature of LOTUS is its ability to work in a pan-cancer setting, as well as to predict driver genes specific to individual cancer types. In the later case, we use a multitask learning strategy to jointly learn scoring functions for all cancer types

by sharing information about known driver genes in different cancer types. We test both a default multitask learning strategy, that shares information uniformly across all cancer types, and a new strategy that shares information across cancer types according to their degree of similarity. More details about the mathematical formulation and algorithms implemented in LOTUS are provided in the Material and Methods section.

In the following, we assess the performance of LOTUS first in the pan-cancer regime, i.e. in the single task setting, and compare it to four state-of-the-art methods (TUSON, MutSigCV, 20/20+ and DiffMut), and second in the cancer type specific regime, where we illustrate the importance of the multitask learning strategies.

Cross-validation performance for pan-cancer driver gene prediction

We first study the pan-cancer regime, where cancer is considered as a single disease, and where we search for driver genes involved in at least one type of cancer. Several computational methods have been proposed to solve this problem in the past, and we compare LOTUS with four well-known state-of-the-art methods [28]: MutSigCV [21], which is a frequency-based method, TUSON [26], 20/20+ [28], which combines frequency and functional information, and DiffMut [34], which takes the mutation patterns of genes into account.

Among these five methods, we can distinguish on the one side, the unsupervised methods MutSigCV and DiffMut that score candidate genes independently of any training set of known drivers, and the supervised methods LOTUS, TUSON and 20/20+ that make predictions based on a training set of known driver genes.

In addition, all methods use gene descriptors that are calculated based on a mutation databases, and therefore, changing the mutation database will change the prediction performance.

In order to make fair comparison between LOTUS and the other four methods, we performed several experiments in which LOTUS is trained with the training set of the TUSON (respectively 20/20+) paper when compared to TUSON (respectively 20/20+). In addition, in all experiments, the gene features calculated for LOTUS were based on the same mutation databases as those used by the other methods in their respective papers.

Therefore, for a fair comparison between LOTUS and TUSON, we use the mutation database available on the website of the authors along with their training sets of OGs and TSGs provided in [26]. We evaluate the performance of LOTUS on this dataset by 5-fold cross-validation (CV) repeated twice (see Methods). For TUSON, we use the prediction results available in [26] and evaluate the consistency errors (CE) as the mean number of non-driver genes that are ranked before known driver genes of the TUSON train sets.

For a fair comparison between LOTUS and 20/20+, we use the mutation database of 20/20+ and the training sets of OGs and TSGs provided by the authors on their website [28]. We evaluate the performance of LOTUS as above. However we note that the 20/20+ score itself is obtained by a bootstrap procedure similar to our cross-validation approach [28].

For a fair comparison between LOTUS and MutSigCV, we use the example mutation database available only for lung squamous cell tumours. Since MutSigCV does not use a train set of driver genes, we trained LOTUS with known OGs and TSGs available in CGCv86 for lung squamous cell tumours. MutSigCV provides a ranked list of genes that does not distinguish TSG and OG. Therefore, the consistency error (CE) is obtained by averaging the numbers of non-driver genes ranked before each driver genes in the train sets used for LOTUS.

Finally, for a fair comparison between LOTUS and DiffMut, we use the 20/20+ mutation dataset for both methods. LOTUS is trained with the 20/20+ training sets of OGs and TSGs. We run DiffMut using the latest version of the algorithm, and we compute the CE related to the 20/20+ training sets of OGs and TSGs.

The CE for OGs and TSGs are presented in Table 1 for TUSON, in Table 2 for 20/20+ and DiffMut, and in Table 3 for MutSigCV. When analyzing these results, one should keep in mind that the total number of cancer driver genes is still a subject of debate, but it is expected to be much lower than the size of the test set (which depends on the method but is of the order of $18,000$), and it should rather be in the range of a few hundreds. Therefore, consistency errors above a few thousand can be considered as poor performance results.

These results show that LOTUS strongly outperforms all other algorithms in term of CE, for both TSG and OG predictions. More precisely, for OG predictions, TUSON is about three times better than MutSigCV, three times better than TUSON, twice better than 20/20+ and five times better than DiffMut, in terms of CE. For TSG predictions, the reduction in CE with LOTUS is two-fold, five-fold, two-fold and five-fold compared to MutSigCV, TUSON, 20/20+ and DiffMut, respectively. Note that the performance is overall much better in the two first experiments, which are also easier because they provide larger mutational data.

It is interesting to note that, for all methods except in the MutSigCV experiment, the performances obtained for OG do not reach those obtained for TSG, suggesting that OG prediction is a more difficult problem than TSG prediction. This reflects the fundamental difference between TSG mutations and OG mutations: the first lead to loss-of-function and can pile up, while the second are gain-of-function mutations and have a much more subtle nature. In addition, gain-of-function can also be due to overexpression of the OG, which can arise from other mechanisms than gene mutation. One way to improve the OG prediction performance may be to include descriptors better suited to them, such as copy number. Moreover, as mutations affecting OGs are not all likely to provide them with new functionalities, many mutations on OGs present in the database and used here might not bear information on OGs. Therefore, relevant information on OGs is scarce, which makes OG prediction more difficult. In addition, the data themselves might also contribute to difference in performance between TSG and OG prediction. For example, in the case of the TUSON train set, although the TSG and OG train sets both contain 50 genes, the mutation matrix that we used to build the gene features contains $13,525$ mutations affecting TSGs and $7,717$ mutations affecting OGs. Therefore, the data are richer for TSG, which might contribute to the difference in prediction performance.

The benefits of combining mutations and PPI informations

LOTUS, 20/20+, MutSigCV, DiffMut and TUSON differ not only by the algorithm they implement, but also by the type of data they use to make predictions: in particular, TUSON and 20/20+ use only mutational data while LOTUS uses PPI information in addition to mutational data. To highlight the contributions of the algorithm and of the PPI information to the performance of LOTUS, we ran LOTUS with $K_{genes} = K_{mutation}$, or $K_{genes} = K_{PPI}$, *i.e.,* with only mutation information, or only PPI information.

The good results of the PPI kernel could be due to the fact that driver genes are important nodes in the PPI network because they are already well studied in the literature. In order to rule out this possibility, we also run LOTUS with the kernel K_{degree} defined by:

$$\left(K_{degree}\right)_{i,j} = d_i d_j,$$

where d_i is the degree of i in the PPI network, *i.e.,* its number of neighbors in the network.

The results are presented in Table 4 and Table 5 respectively for OG and TSG. The last column of these Tables recalls the performance obtained when mutation and PPI information are both used (values reported from Table 1, Table 2 and Table 3).

These results show that, both for OG and TSG, using both mutation and PPI information dramatically improves the prediction performance over using only one of them. This underlines the fact that mutation and PPI bear complementary information that are both useful for the prediction tasks. The performances obtained with only PPI information are similar for OG and TSG, which seems to indicate that this information contributes similarly to both prediction tasks. On the contrary, the performances obtained using only mutation information are much better for TSG than for OG. This is consistent with the above comment that mutation information is more abundant in the database and more relevant in nature for TSG than for OG. It is also consistent with the fact that using $K_{mutation}$ alone outperforms using K_{PPI} alone for TSGs, while the opposite is observed for OGs. Finally, we see that the degree kernel has in almost all cases worse performance than the PPI kernel, which confirms that the number of neighbors alone contains less relevant information in relation with the driver prediction problem than the K_{PPI} kernel does.

Furthermore, we examined the first predictions (excluding already known driver genes) of LOTUS with the 20/20 datasets, when both the mutation and the PPI kernels are used. Among the first 50 TSGs (described by the number of frameshift, LOF and splice mutations), 26 have less than 3 mutations of each kind, 4 predicted TSGs even having no mutation at all. This demonstrates that LOTUS predictions strongly benefit from the PPI information, and that some of these genes would have never been detected using mutation data only.

Performance on CGCv86 prediction in the pan-cancer regime

We now evaluate the generalization properties of the different methods on new unseen data as external test set. This could mitigate the potential bias in the evaluation of the performance of TUSON, DiffMut and 20/20+ based on cross-validation experiments, as in the previous paragraph. We now evaluate the performance of the different methods when predicting supposedly "difficult" new cancer genes (an independent test set), which have only been added recently in CGCv86. We train on the one hand LOTUS and TUSON with the TUSON mutation database and driver gene train sets, and on the other hand LOTUS, DiffMut and 20/20+ with the 20/20 mutation database and driver gene train sets. Then, we make predictions on the remaining genes in COSMIC, and count how many driver genes in CGCv86 appear among the 20, 50 and 100 first predictions. Note that the driver genes from the train sets were excluded from the predictions. The results are shown in Tables 6-9 and are illustrated by corresponding ROC curves, see Figures 1 and 2.

First, we observe that TUSON outperforms LOTUS in almost all these experiments. Second, LOTUS outperforms DiffMut in all experiments. Third, LOTUS is better than 20/20+ for TSG detection, and the contrary holds for OGs. Generally speaking, the first predictions of TUSON and 20/20+ are more reliable than LOTUS's, but, as shown in Fig 1 and 2, but LOTUS outperforms all the methods when all genes are considered, and not only the first 20 to 100 genes.

The good performance of TUSON and 20/20+ for the top ranked genes compared to those of LOTUS could be explained by the fact that, all genes in CGCv86 so far have been reported through analysis of mutation data (cf. CGC web page: 'The Cancer Gene Census (CGC) is an ongoing effort to catalogue those genes which contain mutations that have been causally implicated in cancer'). This interpretation would also explain

why LOTUS hardly agrees with the other methods when comparing the top ranked genes. Indeed, for the 20 top predictions of LOTUS (excluding training sets), the intersection with TUSON consists only in two TSGs and one OG, the intersection with DiffMut consist of one TSG and two OGs, and the intersection with 20/20+ consists in four TSGs and one OG. Since LOTUS is the only method, among those considered here, that uses PPI information in addition to mutation data, this could explain less overlap between its predictions and those of the other methods, and a lower overlap with CGC for the top ranked genes, since no PPI information was used either to establish the CGC database.

Analysis of new driver genes predicted by LOTUS

We tested the ability of LOTUS to make new driver gene predictions. We trained LOTUS with the CGCv86 train set, made predictions over the complete COSMIC database (19,320 genes including the training sets). The complete results are given in Supplementary Table 3. Complete analysis of the predicted OGs and TSGs rankings is out of the scope of this paper. However, we considered the 22/21 best ranked TSGs and OGs, and made bibliographic search in order to look for independent information that could validate these predictions.

For most of the 22 best ranked predicted OGs, abundant literature reports implication in various cancers. Some genes were known to promote cancer or be therapeutic targets or biomarkers for some cancer types. It is not possible to make a full review of literature for each of these genes, but we cite below some of the most relevant papers.

Twelve out of the 22 best ranked genes are involved in transcription regulation at various levels, a mechanism that is invariably perturbed in cancer.

Among them, four genes act through chromatic remodeling (PYGO1, PYGO2, EP300, DOT1L). PYGO1 and PYGO2 are co-activators of the Wnt signaling pathway and they increase target genes transcription [35]. In particular, PYGO1 is involved in colorectal cancer [36], while PYGO2 was shown to be a tumor promoter in mice [37]. EP300 is an histone acetyltransferase, and defects in this gene's function by mutations or alterations in expression contributes to cancer phenotype [38]. DOT1L is an histone methyltransferase, and it is a known therapeutic target [39].

Six genes are transcription factors or repressors (MSEI1, MSEI2, MSEI3, TFEC, NKX2-2, ZIK1). MSEI1, MSEI2, and MSEI3 are involved in the etiology, progression and metastatic evolution of some cancer types such as prostate cancer [40], or leukemia [41, 42]. TFEC belongs to the microphthalmia family (MiT/TFE) of leucine zipper transcription factors, and the latest research on proteins of this family decipher their mechanisms in cancer development [43]. NKX2-2 is homeobox containing transcription factor. While its close homologue NKX2-1 is a known oncogene absent from the CGCv86 database [44], NKX2-2 was identified to be a critical target gene in Ewing's sarcoma development [45]. Aberrant methylation of the promoter of gene ZIK1 is observed in colon cancer [46] and in intestinal metaplasia [47]. This gene belongs to the ZNF family that has not been studies into sufficient detail because of its complexity, but recent studies establish them as new oncologic biomarkers or therapeutic targets [48].

DROSHA and ELF1 participate to transcription regulation via microRNA regulation and as elongation factor, respectively. Alteration in microRNAs expression is a frequent finding in human cancers. DROSHA is involved in the miRNA depletion observed in lung cancer, and alterations in this gene was shown to have a remarkable impact in lung cancer [36]. ELF1 directly plays a role in the mechanism of eIF6 release that is corrupted in inherited and sporadic leukemias [49].

Besides these twelve transcription-related genes, eight other best ranked OGs are known to be related to cancer by various mechanisms. Among them, FGF6 and FGF5 are members of the fibroblast growth factors (FGF) that are well known players contributing to tumor progression [50]. MOB3B (or MOB1) is a pivotal kinase player in the Hippo tumor suppressor pathway, and mutations in this gene is associated to prostate cancer susceptibility and agressive tumors [51]. GALTLN11 is a member of the GALNT family of enzymes that catalyse O-linked glycosylation, a family proteins that strongly promotes liver tumor growth after a shift from the endoplasmic reticulum to the Golgi [52]. FBXW7 is a member of the F-box protein family involved in phosphorylation-dependent ubiquitination, and mutations in this gene are detected in ovarian and breast cancer cell lines [53].

Although not exhaustive, these findings indicate that the best ranked oncogenes predicted by LOTUS are realistic OG for some cancer types.

Among the 21 best ranks tumor suppressors, APOM is actually a known TSG for hepatocellular carcinoma [54].

Five of the 21 best ranked TSGs are genes coding for proteins involved in DNA repair, a role closely related to genome maintenance and cancer [55, 56]. These genes are EXO1 [57], ERCC1 [58], GTF2H1 [59], and MDC1 [60], and DGCR8 [61]. Besides their DNA repair functions, many studies related to these genes are available in the literature (in addition to those cited here), underlying their protective role in various types of cancers, which provides additional clues for them to be confident TSG candidates. In addition to these five genes, the transcription factor ZNF521 was shown to regulate the expression of the BRCA1, a well known TSG involved in DNA repair.

Three genes are involved in immune system response to cancer cells by similar mechanisms, PDCD1, KLRG1, and MUC16. PDCD1 (or PD-1, for Programmed Cell Death 1) is expressed in T-cell lymphocytes that contribute to kill cancer cells. Cancer cells can escape from T lymphocytes attack by over expression of the PD-L1 protein, which binds to PD-1 and consequently induce tolerance from T lymphocytes [61]. Therefore, PD-1 can be viewed as a TSG, since mutations in this gene might prevent binding of cancer cells PD-L1 proteins. Similarly, KLRG1 gene encode a natural killer (NK) cell lectin-like receptor that drives lysis of tumor cells by NK cells, and epigenetic repression of KLRG1 expression favors breast cancer tumorigenesis and cell survival [62]. MUC16 encode for a protein from the mucin family. Mucins are O-glycosylated proteins forming a protective mucous barrier. They can bind to various receptors of immune cells including NK cells. Aberrant over-expression and glycosylation of mucins (including MUC16) in various malignancies facilitate oncogenic events to escape from the immune response [63].

For six other genes of various functions, we found recent publications indicating that they could potentially act as TSGs (SPTA1, GALNT5, PIWIL1, PIWIL4, SNX5, ADAM6). SPTA1 encodes a protein that links the plasma membrane to the actin cytoskeleton and functions in the determination of cell shape, arrangement of transmembrane proteins, and organization of organelles. Mutations in this gene was found to play a role in glioblastoma [64]. A non-coding RNA directed against GALNT5 is overexpressed in gastric cancer, inhibiting the translation of its target gene, and the level of expression of this non-coding RNA is correlated with cancer progression and metastasis [65]. These results are consistent with a TSG role of GALNT5 in gastric cancer. PIWIL1 and PWIL4 genes encode for proteins that play important roles in stem cell self-renewal, RNA silencing and translational regulation, and recent papers illustrate that they could be epigenetic TSG genes. Aberrant methylation of the promoter region of PIWIL1 plays a role in the development of lung adenocarcinoma [66], while decreased expression levels of PIWIL1 and PIWIL4 is associated with worse survival in renal cell carcinoma patients [67]. SNX5 is involved in intracellular

trafficking, and loss or decreased expression of this gene promotes thyroid cancer progression [68]. A long coding RNA that silences ADAM6 was found to be overexpressed in lung adenocarcinoma, which is consistent with a potential TSG role for ADAM6 in this tumor type [69].

Interestingly, for three best ranked predicted TSGs, bibliographic search provided clues that they indeed play a role in cancer, but that they would rather behave as OG. These genes are CENPU [70], FXYD2 [71], ANXA9 [72]. In fact, the literature provides other examples of genes able to switch from oncogenes to tumor suppressor genes, depending on the context [73], which could be the case for these genes. In most cases, the cited papers (and others) observe that over-expression of these genes are observed in various types of cancers. One assumption could be that variations of their levels of expression might lead to switch between TSG and OG roles.

Taken together, these results show that, among the top TSG and OG ranked by LOTUS, many genes are indeed involved in cancer, and that LOTUS predictions correspond to relevant genes that are reliable candidates as cancer driver genes, at least for some tumor types.

Identification of cancer-specific driver genes with multitask LOTUS

In this section, we do not consider cancer as a single disease, but as a variety of diseases with different histological types that can affect various organs. Indeed, an important question in cancer research is to identify driver genes for each type of cancer. One way to solve this problem is to use a prediction method that is trained only with driver genes known for the considered cancer. Such single-task methods may however display poor performance because the number of known drivers per cancer is often too small to derive a reliable model. Indeed, scarce training data lead to a potential loss of statistical power as compared to the problem of identification of pan-cancer driver genes where data available for all cancers are used.

In this context, we investigate two multitask versions of LOTUS, where we predict driver genes for a given cancer based on the drivers known for this cancer but also on all driver genes known for other cancer types. For a given cancer type, this may improve driver genes prediction by limiting the loss of statistical power compared to the aforementioned single-task approach.

For that purpose, we derive a list of 30 cancer diseases from the 20/20 mutation dataset as explained in Methods. This complete list is available in Supplementary Table 1. As expected, many cancer types have only few known cancer genes (Figure 3).

Since we want to evaluate the performance of LOTUS in a cross-validation scheme, we only consider diseases with more than 4 known driver genes in order to be able to run a 2-fold CV scheme. This leads us to keep 27 cancer types for TSG prediction and 27 for OG prediction. Note however that, for each cancer type, prediction are made while sharing all the driver genes known for the 30 diseases, according to their similarities with these cancer types.

The 2-fold CV consistency error of LOTUS for the 27 considered cancer types is presented in Tables 10 (for TSG) and 11 (for OG). We compare four variants of LOTUS, as explained in Methods: (1) single-task LOTUS treats each disease in turn independently from the others using only the mutation data related to the considered disease to calculate gene features, and only the driver genes known for this disease are used to train the algorithm; (2) Aggregation LOTUS is also a single-task version of LOTUS, but gene features are calculated using the complete mutation database of gene mutations in all cancers. In addition, for each disease, the train set consists of known drivers for all the other cancers and have of the drivers known for the considered disease.

Then the prediction performance are calculated for the other half of known drivers for this disease, which constitute the test set in the 2-fold cross validation scheme. Therefore, Aggregation LOTUS is a single-task algorithm that uses much richer information than the basic Single-task LOTUS; (3 and 4) Two multitask versions of LOTUS use either a standard multitask strategy that does not take into account the relative similarities between diseases (Multitask LOTUS), or a more refined multitask strategy where driver gene information is shared between cancer types according to their similarity based on biological information (Multitask LOTUS2). Finally, we compare these performances with those of DiffMut, as a single-task method using only the mutation data related to the considered disease, as for single-task LOTUS.

For most diseases (25/27 for TSG, 27/27 for OG), single-task LOTUS and DiffMut lead to the worst CE, confirming the difficulty to treat each cancer type individually, due to the small number of known driver genes and to the smaller mutation database available for each cancer type type. Interestingly, Aggregation LOTUS often leads to a strong improvement in performance. This shows that different cancer types often share some common mechanisms and driver genes, and therefore, simply using all the available information as in a pan-cancer paradigm improves the performance of driver gene prediction for each disease. However, in many cases, the multitask LOTUS and LOTUS2 algorithms lead to an additional improvement over Aggregation LOTUS, LOTUS2 leading in general to the best results (in 17 types out of 27 for TSG prediction, and in 17 types out of 27 for OG prediction) . On average, the decrease in CE between Aggregate LOTUS and LOTUS2 is of 20% for OG and 18% for TSG. The improvement in performance observed between Aggregate LOTUS and LOTUS2 shows that, besides some driver mechanisms common to many cancers, each cancer presents some specific driver mechanisms that can only be captured by prediction methods able to integrate some biological knowledge about the different diseases. The above results show that multitask algorithms allowing to share information between cancers according to their biological similarities such as LOTUS2, rather than on more naive rules, better capture these specific driver genes. They also show that the kernel $K_{diseases} = K_{descriptors}$ built on disease descriptors contains some relevant biological information to compare diseases.

To measure how different the predictions of LOTUS2 are for each cancer type, we compared the first 50 predictions for each type. Aggregating all predictions for TSGs (respectively OGs) results in 210 genes (respectively 224 genes), which shows that various cancer types share some drivers, but that the prediction lists are different. Indeed, some drivers with high penetrance (such as TP53) are expected to be found in most cancer types, whereas other drivers are more specific to given organs or cell types, in particular since all genes are not expressed in all cell types.

In addition, multitask algorithms based on task descriptors (here, disease descriptors) appear to be promising in order to include prior knowledge about diseases and share information according to biological features characterizing the diseases.

Finally, note that we did not try to run TUSON, MutSigCV or 20/20+ to search for cancer specific driver genes in the single-task setting (they cannot be run in the multi-task setting). Indeed, according to the results of pan-cancer studies in the single-task setting, they do not perform as well as single-task LOTUS. Considering that single-task LOTUS and DiffMut were far from reaching the performance of multi-task LOTUS for prediction of cancer specific driver genes, TUSON, MutSigCV or 20/20+ are not expected to reach these performance either.

Discussion

Our work demonstrates that LOTUS outperforms several state-of-the-art methods on all tested situations for driver gene prediction. This improvement results from various

aspects of the LOTUS algorithm. First, LOTUS allows to include the PPI network information as independent prior biological knowledge. In the single-task setting, we proved that this information has significance for the prediction of cancer driver genes. Because LOTUS is based on kernel methods, it is well suited to integrate other data from multiple sources such as protein expression data, information from chip-seq, HiC or methylation data, or new features for mutation timing as designed in [74]. Further development could involve the definition of other gene kernels based on such type of data, and combine them with our current gene kernel, in order to evaluate their relevance in driver gene prediction.

We also showed how LOTUS can serve as a multitask method. It relies on a disease kernel that controls how driver gene information is shared between diseases. Interestingly, we showed that building a kernel based on independent biological prior knowledge about disease similarity leads on average to the best prediction performance with respect to single-task algorithms, and also with respect to a more generic and naive multitask learning strategy that does not incorporate knowledge about the cancer types. Again, the kernel approach leaves space for integration of other types and possibly more complex biological sources of information about diseases. Our multitask approach thus allows to make prediction for cancer types with very few known driver genes, which would be less reliable with the single-task methods. We considered here only diseases with at least 4 known driver genes, in order to perform cross-validation studies, which was necessary to evaluate the methods. However, it is important to note that in real-case studies, at the extreme, both versions of multitask LOTUS could make driver gene prediction for the 30 cancer types, including those for which no driver gene is known.

LOTUS is a machine learning algorithm based on one-class SVM. In fact, the most classical problem in machine learning is binary classification, where the task is to classify observations into two classes (positives and negatives), based on training sets $\mathcal{P}$ of known positives and $\mathcal{N}$ of known negatives. Driver gene detection can be seen as binary classification of TSGs vs. neutral genes, and of OGs vs. neutral genes. However, although the $\mathcal{P}$ set is composed of known driver genes, it is not straightforward to build the $\mathcal{N}$ set because we cannot claim that some genes cannot be drivers. Thus, driver gene detection should rather be seen as binary classification problem with only one training set $\mathcal{P}$ of known positives. This problem is classically called PU learning (for Positive-Unknown), as opposed to PN learning (for Positive-Negative).

The classical way to solve PU learning problems is to choose a set $\mathcal{N}$ of negatives among the unlabeled data and apply a PN learning method. For example, one can consider all unknown items as negatives (some of which may be reclassified afterwards as positives), or randomly choose bootstrapped sets of negatives among the unknown, like in [33]. Both methods assume that a minority of the unlabeled items are in fact positives, which is expected for driver genes.

The one-class SVM algorithm [75] can also be used as a PU learning method, in which a virtual item is chosen as the training set of negatives. We preferred this approach because in preliminary studies, we found that it had slightly better performances than PU learning methods and was also faster.

For LOTUS, as for all machine learning algorithm, the set of known driver genes is critical: if this set is poorly chosen (*i.e.,* if some genes were wrongly reported as driver genes, or more likely, if the reported genes are not the best driver genes), the best algorithm might not minimize the consistency error CE. To circumvent this problem, we propose two new approaches for future developments: one could build a multi-step algorithm that iteratively removes some genes from the positive set and labels them as unknown, and relabel as positives some of the best ranked unknown genes. We believe that such an algorithm would make the set of positives converge to a more relevant list. Alternatively, one could assign (finite) scores to the known driver genes before

performing classification and increment these scores at each step.

Materials and methods

Pan-cancer LOTUS

LOTUS is a new machine learning-based method to predict new cancer driver genes, given a list of know ones. In the simplest, pan-cancer setting, we consider a list of N known driver genes $\{g_1, \ldots, g_N\}$, and the goal of LOTUS is to learn from them a scoring function $f(g)$, for any other gene g, that predicts how likely it is that g is also a cancer gene. Since TSGs and OGs have different characteristics, we treat them separately and build two scoring functions f_{TSG} and f_{OG} that are trained from lists of know TSGs and OGs, respectively.

LOTUS learns the scoring function $f(g)$ with a one-class support vector machine (OC-SVM) algorithm [75], a classical method for novelty detection and density level set estimation [76]. The scoring function $f(g)$ learned by a OC-SVM given a training set $\{g_1, \ldots, g_N\}$ of known cancer genes takes the form:

$$f(g) = \sum_{i=1}^{N} \alpha_i K(g_i, g)\,, \tag{1}$$

where $\alpha_1, \ldots, \alpha_N$ are weights optimized during the training of OC-SVM [75], and $K(g, g')$ is a so-called *kernel* function that quantifies the similarity between any two genes g and g'. In other words, the score of a new gene g is a weighted combination of its similarities with the known driver genes.

The kernel K encodes the similarity among genes. Mathematically, the only constraint that K must fulfill is that it should be a symmetric positive definite function [30]. This leaves a lot of freedom to create specific kernels encoding prior knowledge about relevant information to predict driver genes. In addition, one can easily combine heterogeneous information in a single kernel by, e.g., summing two kernels based on different sources of data. In this work, we restrict ourselves to the following basic kernels, and leave for future work a more exhaustive search of optimization of kernels for cancer gene prediction.

- *Mutation kernel.* Given a large data set of somatic mutations in cohorts of cancer patients, we characterize each gene g by a vector $\Phi_{mutation}(g) \in \mathbb{R}^3$ encoding 3 features. For OG prediction the three features are the number of damaging missense mutations (defined as in [26] as mutations with a Polyphen2 score larger than 0.447), the total number of missense mutations, and the entropy of the spatial distribution of the missense mutations on each gene. For TSG prediction, the features are the number of frameshift mutations, the number of LOF mutations (defined as the nonsense and frameshift mutations), and the number of splice site mutations. These features were calculated as proposed by [26]. We chose them because they were found to best discriminate OGs and TSGs by the TUSON algorithm [26] and were also all found among the most important features selected by the random forest algorithm used by the 20/20+ method [28]. Given two genes g and g' represented by their 3-dimensional vectors $\Phi(g)$ and $\Phi(g')$, we then define the mutation kernel as the inner product between these vectors:

  $$K_{mutation}(g, g') = \Phi_{mutation}(g)^\top \Phi_{mutation}(g')\,.$$

  Notice that using $K_{mutation}$ as a kernel in OC-SVM produces a scoring function (1) which is simply a linear combination of the three features used to define the vector $\Phi_{mutation}$.

- *PPI kernel.* Given an undirected graph with genes as vertices, such as a PPI network, we define a PPI kernel K_{PPI} as a graph kernel over the network [77, 78]. More precisely, we used a diffusion kernel of the form $K_{PPI} = \exp_M(-L)$, where $L = I - D^{-1/2}AD^{-1/2}$ is the normalized Laplacian of the graph and $\exp_M$ is the matrix exponential function. Here I is the identity matrix, A stands for the adjacency matrix of the graph ($A_{i,j} = 1$ if vertices i and j are connected, 0 otherwise) and D for the diagonal matrix of degrees ($D_{ii} = \sum_{j=1}^{n} A_{ij}$). Intuitively, two genes are similar according to K_{PPI} when they are close and well connected through several routes to each other on the PPI network, hence learning a OC-SVM with K_{PPI} allows to diffuse the information about cancer genes over the network.

- *Integrated kernel.* In order to train a model that incorporates informations about both mutational features and PPI, we create an integrated gene kernel by simply averaging the mutation and PPI kernels:

$$K_{gene}(g, g') = \left(K_{mutation}(g, g') + K_{PPI}(g, g')\right)/2\,.$$

  While more complex kernel combination strategies such as multiple kernel learning could be considered, we restrict ourselves to this simple kernel addition scheme to illustrate the potential of our approach for heterogeneous data integration.

Multitask LOTUS for cancer type-specific predictions

The pan-cancer LOTUS approach can also be used for cancer-specific predictions, by restricting the training set of known cancer driver genes to those genes known to be driver in a particular cancer type. However, for many cancer types, only few driver genes have been validated, creating a challenging situation for machine learning-based methods like LOTUS that rely on a training set of known driver genes to learn a scoring function. Since cancer driver genes of different cancer types are likely to have similar features, we propose instead to learn jointly cancer type-specific scoring functions by sharing information about known driver genes across cancer types, using the framework of multitask learning [32, 33]. Instead of starting from a list of known driver genes, we now start from a list of known (cancer gene, cancer type) pairs of the form $\{(g_1, d_1), \ldots, (g_N, d_N)\}$, where a sample (g_i, d_i) means that gene g_i is a known cancer gene in disease d_i. Note that a given gene (and a given cancer type) may of course appear in several such pairs.

The extension of OC-SVM to the multitask setting is straightforwardly obtained by creating a kernel for (gene, disease) pairs of the form:

$$K_{pair}\left((g, d), (g', d')\right) = K_{gene}(g, g') \times K_{disease}(d, d')\,,$$

where K_{gene} is a kernel between genes such as that used in pan-cancer LOTUS and $K_{disease}$ is a kernel between cancer types described below. We then simply run the OC-SVM algorithm using K_{pair} as kernel and $\{(g_1, d_1), \ldots, (g_N, d_N)\}$ as training set, in order to learn a cancer type-specific scoring function of the form $f(g, d)$ that estimates the probability that g is a cancer gene for cancer type d.

The choice of the disease kernel $K_{disease}$ influences how information is shared across cancer types. One extreme situation is to take the uniform kernel $K_{uniform}(d, d') = 1$ for any d, d'. In that case, no distinction is made between diseases, and all known cancer driver genes are pooled together, recovering the pan-cancer setting (with the slight difference that genes may be counted several times in the training set if they appear in several diseases). Another extreme situation is to take the Dirac kernel $K_{Dirac}(d, d') = 1$ if $d = d'$, 0 otherwise. In that case, no information is shared across

cancer types, and the joint model over (gene, disease) pairs is equivalent to learning independently a model for each disease, as in the single-task approach.

In order to leverage the benefits of multitask learning and learn disease-specific models by sharing information across diseases, we consider instead the following two disease kernels:

- First, we consider the standard multitask learning kernel:

$$K_{multitask}(d, d') = \left(K_{uniform}(d, d') + K_{Dirac}(d, d')\right)/2\,,$$

  which makes a compromise between the two extreme uniform and Dirac kernels [32]. Intuitively, for a given cancer type, prediction of driver genes is made by assigning twice more weight to the data available for this cancer than to the data available for all other cancer types.

- Second, we test a more elaborate multitask version where we implement the idea that a given cancer might share various degrees of biological similarities with other cancers. Therefore, known driver genes for other cancers should be shared with those of the considered cancer based on this similarity. Hence, we create a specific disease kernel $K_{cancer}(d, d')$ to capture how similar two cancer types are. To create K_{cancer}, we first represent each cancer type as a 43-dimensional binary vector as follows. The first 12 bits correspond to a list of cancer type characteristics used in COSMIC to describe tumors: adenocarcinoma, adenoma, blastoma, carcinoma, glioma, leukemia, lymphoma, medulloblastoma, melanoma, myeloma, rhabdomyosarcoma, sarcoma. The last 31 components correspond to localization characteristics also used in COSMIC to describe tumors: adrenal glands, astrocytes, B-cell, bladder, bone, breast, cervix, central nervous system, colon, ducts, endometrium, eye, head and neck, heart, kidney, liver, lung, lymphocytes, mucosa, muscle, nerve, oesophagus, ovary, pancreas, prostate, salivary glands, skin, soft tissue, squamous cell, stomach, T-cell, thyroid. A disease might be assigned one or several types and be associated to one or several locations. For example, Melanoma is associated with a single type ("melanoma") and four localizations ("skin", "mucosa", "eye" and "head and neck"), so that Melanoma is described by a vector with five 1's and thirty-eight 0's. For each disease, we construct the list of binary features by documenting every disease in the literature. The corresponding vectors encoding the considered disease are given in Supplementary Table S2. Finally, if $\Psi(d) \in \mathbb{R}^{43}$ denotes the binary vector representation of disease d, we create the disease kernel as a simple inner product between these vectors, combined with the standard multitask kernel, i.e.:

$$K_{cancer}(d, d') = \left(\Psi(d)^\top \Psi(d') + K_{uniform}(d, d') + K_{Dirac}(d, d')\right)/3\,.$$

Data

When comparing LOTUS to TUSON, we use a dataset of somatic mutations collected from COSMIC [14], TCGA (http://cancergenome.nih.gov/) and [18], that was used in [26]. This dataset contains a total of $1,195,223$ mutations across $8,207$ patients affecting $18,843$ genes.

When comparing LOTUS to DiffMut and 20/20+, we use a dataset of somatic mutations borrowed from [28]. This dataset contains a total of $729,205$ mutations across $7,916$ patients affecting $19,320$ genes.

When comparing LOTUS to MutSigCV, we use an example dataset available on GenePattern. This dataset contains a total of $137,343$ mutations across 177 patients of lung squamous cell carcinoma affecting $16,885$ genes.

We obtained the PPI network from the HPRD database release 9 from April 13, 2010 [79]. It contains 39,239 interactions among 7,931 proteins. As for known pan-cancer driver genes, we consider three lists in our experiments: (i) the TUSON train set, proposed in [26], consists of two high confidence lists of 50 OGs and 50 TSGs extracted from CGC (release v71) based on several criteria, in particular excluding driver genes reported through translocations; (ii) the 20/20 train set, proposed in [28] to train the 20/20+ method, contains 53 OGs and 60 TSGs; finally, (iii) the CGCv86 train set consists of two broader lists that we extracted from CGC release v86 of the COSMIC database: we consider as OGs the genes annotated as "oncogene", "oncogene, TSG", "oncogene, fusion", "oncogene, TSG, fusion", and as TSGs the genes annotated as "TSG", "oncogene, TSG", "TSG, fusion", "oncogene, TSG, fusion". For cancer type-specific lists of driver genes, we only consider the CGCv86 train sets. We distinguished 30 diseases based on the available annotations describing patients in the mutation matrix, only merging "Kidney Chromophobe", "Kidney Papillary Cell Carcinoma" and "Kidney Clear Cell Carcinoma" into "Kidney Cancer", "DLBCL" and "Lymphoma B-Cell" into "Lymphoma B-Cell" and neglecting the unspecific "CARC". The names of these diseases and their numbers of associated TSGs and OGs can be found in Supplementary Table 1. For each of the resulting diseases, 0 to 56 TSGs/OGs were known in CGCv86. We considered only diseases with at least 4 known TSGs or OGs available, in order to have enough learning data points to perform a two-fold cross-validation scheme, which led us to consider 27 diseases for TSG prediction and 27 for OG prediction.

Experimental protocol

To assess the performance of a driver gene prediction method on a given gold standard of known driver genes, we score all genes in the COSMIC database and measure how well the known driver genes are ranked. For that purpose, we plot the receiver operating characteristic (ROC) curve, considering all known drivers as positive examples and all other genes in COSMIC as negative ones, and define the consistency error (CE) as

$$CE = \#\mathcal{N} \times (1 - AUC),$$

where $\#\mathcal{N}$ is the number of negative genes, and AUC denotes the area under the ROC curve. In other words, CE measures the mean number of "non-driver" genes that the prediction method ranks higher than known driver genes. Hence, a perfect prediction method should have $CE = 0$, while a random predictor should have a CE near $\#\mathcal{N}/2$.

To estimate the performance of a machine learning-based prediction method that estimates a scoring function from a training set of known driver genes, we use k-fold cross-validation (CV) for each given gold standard set of known driver genes. In k-fold CV, the gold standard set is randomly split into k subsets of roughly equal sizes. Each subset is removed from the gold standard in turn, the prediction method is trained on the remaining $k-1$ subsets, and its CE is estimated considering the subset left apart as positive examples, and all other genes of COSMIC not in the gold standard set as negative examples. A mean ROC curve and mean CE is then computed from the k resulting ROC curves. This computation is repeated several times to consider several possibly different partitions of the gold standard set.

Tuning of parameters

Each version of LOTUS depends on a unique parameter, the regularization parameter C of the OC-SVM algorithm. Each time a LOTUS model is trained, its C parameter is optimized by 5-fold CV on the training set, by picking the value in a grid of candidate values $\{2^{-5/2}, 2^{-4/2}, \ldots, 2^{5/2}\}$ that minimizes the mean CE over the folds.

Other driver prediction methods

We compare the performance of LOTUS to four other state-of-the-art methods: MutSigCV [21], which is a frequency-based method, TUSON [26] and 20/20+ [28] that combine frequency and functional information, and DiffMut that analyses mutation profiles on genes.

MutSigCV searches driver genes among significantly mutated genes which adjusts for known covariates of mutation rates. The method estimates a background mutation rate for each gene and patient, based on the observed silent mutations in the gene and noncoding mutations in the surrounding regions. Incorporating mutational heterogeneity, MutSigCV eliminates implausible driver genes that are often predicted by simpler frequency-based models. For each gene, the mutational signal from the observed non-silent counts are compared to the mutational background. The output of the method is an ordered list of all considered genes as a function of a p-value that estimates how likely this gene is to be a driver gene.

TUSON uses gene features that encode frequency mutations and functional impact mutations. The underlying idea is that the proportion of mutation types observed in a given gene can be used to predict the likelihood of this gene to be a cancer driver. After having identified the most predicting parameters for OGs and TSGs based on a train set (called the TUSON train set in the present paper), TUSON uses a statistical model in which a p-value is derived for each gene that characterizes its potential as being an OG or a TSG, then scores all genes in the COSMIC database, to obtain two ranked lists of genes in increasing orders of p-values for OGs and TSGs.

The 20/20+ method encodes genes based on frequency and mutation types, and other biological information. It uses a train set of OGs and TSGs (called the 20/20 train set in the present paper) to train a random forest algorithm. Then, the random forest is used on the COSMIC database and the output of the method is again a list of genes ranked according to their predicted score to be a driver gene [28]. We did not implement this method, so we decided to evaluate its performance only on its original training set: the 20/20 dataset. Moreover, we applied the same method to compute the CE as for MutSigCV and TUSON, which should actually give an advantage to 20/20+, since it is harder to make predictions in a cross-validation loop using a smaller set of known driver genes.

DiffMut uses a dataset of somatic mutations and a dataset of healthy genomes, but no training sets of known driver genes. It compares the mutation profiles on a gene in the mutation dataset with the nucleotide variation profile in the healthy genomes, and computes for every gene a score that allows to rank all genes according to their potential as OG or TSG.

Code and data availability

We implemented LOTUS and performed all experiments in R using in particular the kernlab package for OC-SVM [80]. The code and data to reproduce all experiments are available at `http://members.cbio.mines-paristech.fr/~ocollier/lotus.html`.

Acknowledgments

This work was supported the European Research Council grant ERC-SMAC- 280032 (OC, JPV) and the Labex MME-DII ANR11-LBX-0023-01 (OC).

References

1. D. Hanahan and R. Weinberg *The hallmarks of cancer.* Cell, 100(1), 57-70, 2000.

2. D. Hanahan and R. Weinberg *The hallmarks of cancer: the next generation.* Cell, 144, 646-674, 2011.

3. L. Ding, G. Getz, D.A. Wheeler, E.R. Mardis, M.D. McLellan, K. Cibulkis et al. *Somatic mutations affect key pathways in lung adenocarcinoma.* Nature, 455(7216), 1069-1075, 2008.

4. R.D. Morin, M. Mendez-Lago, A.J. Mungall, R. Goya, K.L. Mungall, R.D. Corbett et al. *Frequent mutation of histone modifying genes in non-Hodgkin lymphoma.* Nature, 476(7360), 298-303, 2012.

5. J.G. Paez, P.A. Jänne, J.C. Lee, S. Tracy, H. Greulich, S. Gabriel et al. *EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy.* Science, 304(5676), 1497-1500, 2004.

6. G.M. Cooper *The cell: a molecular approach, 2nd edition.* Sunderland (MA): Sinauer Associates, 2000.

7. P.L. Chen, Y.M. Chen, R. Bookstein and W.H. Lee *Genetic mechanisms of tumor suppression by the human p53 gene.* Science, 250(4987), 1576-1580, 1990.

8. M.L. Gemignani, A.C. Schlaerth, F. Bogomolniy, R.R. Barakat, O. Lin, R. Soslow et al. (2003) *Role of KRAS and BRAF gene mutations in mucinous ovarian carcinoma.* Gynecol Oncol, 90(2003), 378-381, 2003.

9. A.L. Schechter, D.F. Stern, L. Vaidyanathan, S.J. Decker, J.A. Drebin, M.I. Greene et al. *The neu oncogene: an erb-B-related gene encoding an 185,000-M tumor antigen.* Nature, 312:513–516, 1984.

10. C.A. Hudis *Trastuzumab–mechanism of action and use in clinical practice.* N Engl J Med, 357(1), 39-51, 2007.

11. P. Futreal, L. Coin, M. Marshall, T. Down, T., Hubbard, R. Wooster et al. *A census of human cancer genes.* Nat Rev Cancer, 4, 177-183, 2004.

12. J.N. Weinstein, E.A. Collisson, G.B. Mills, K.M. Shaw, B.A. Ozenberger, K. Ellrott et al. *The Cancer Genome Atlas Pan-Cancer Analysis Project* Nature Genet, 45(10):1113–1120, 2013.

13. J. Zhang, J. Baran, A. Cros, J.M. Guberman, S. Haider, J. Hsu et al. *International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data.* Database (Oxford), 2011.

14. S.A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal et al. *COSMIC: somatic cancer genetics at high-resolution* Nucleic Acids Res, 45, D777-D783, 2017.

15. B. Vogelstein, N. Papadopoulos, V.E. Velculescu, S. Zhou, L.A. Diaz and K.W. Kinzler *Cancer Genome Landscapes.* Science, 339(6127):1546–1558, 2013.

16. M. Lawrence, P. Stojanov, P. Polak, G.V. Kryukov, K. Cibulkis, A. Sivachenko et al. *Mutational heterogeneity in cancer and the search for new cancer associated genes.* Nature, 499, 214-218, 2013.

17. The Cancer Genome Atlas Research Network *Comprehensive genomic characterization of squamous cell lung cancers.* Nature, 489.7417: 519-52, 2012.

18. L. Alexandrov, S. Nik-Zainal, D. Wedge, S. Aparicio, S. Behjati, A. Biankin et al. *Signatures of mutational processes in human cancer.* Nature, 500, 415-421, 2013.

19. N.D. Dees, Q. Zhang, C. Kandoth, M.C. Wendl, W. Schierding, D.C. Koboldt et al. *Identifying mutational significance in cancer genomes.* Genome Res, 22(8): 1589-1598, 2012.

20. J. Reimand and G.D. Bader *Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers.* Mol Syst Biol, 9:637, 2013.

21. M.S. Lawrence, P. Stojanov, C.H. Mermel, J.T. Robinson, L.A. Garraway, T.R. Golub et al. *Discovery and saturation analysis of cancer genes across 21 tumor types.* Nature, 505(7484): 495–501, 2014.

22. A. Gonzalez-Perez and N. Lopez-Bigas *Functional impact bias reveals cancer drivers.* Nucleic Acids Res, 40(21), 2012.

23. A. Bashashati, G. Haffari, J. Ding, G. Ha, K. Lui, J. Rosner et al. *DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer.* Genome Biol, 13(12):R124, 2012.

24. M.H. Bailey, C. Tolkheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe et al. *Comprehensive characterization of cancer driver genes and mutations.* Cell, 173:371–385, 2018.

25. I.F. Chung, C.Y. Chen, S.C. Su, C.Y. Li, K.J. Wu, H.W. Wang et al. *DriverDBv2: a database for human cancer driver gene research.* Nucleic Acids Res, 44(D1):D975-9, 2016.

26. T. Davoli, A. Xu, K. Mengwasser, L. Sack, J. Yoon, P. Park et al. *Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome.* Cell, 155(4), 948-962, 2013.

27. G.E.M Melloni, A.G.E. Ogier, S. de Pretis, L. Mazzarella, M. Pelizzola, P.G. Pelicci et al. *DOTS-Finder: a comprehensive tool for assessing driver genes in cancer genomes.* Genome Med, 6(6):44, 2014.

28. C.J. Tokheim, N. Papadopoulos, K.W. Kinzler, B. Vogelstein and R. Karchin *Evaluating the evaluation of cancer driver genes.* Proc Natl Acad Sci U S A, 113(50):14330–14335, 2016.

29. J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A.A. Margolin, S. Kim et al. *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 483(7391):603-7, 2012.

30. B. Schölkopf et al. *Kernel methods in computational biology.* MIT Press, 2004.

31. M.D.M. Leiserson, F. Vandin, H.-T. Wu, J.R. Robson, J.V. Eldridge, J.L. Thomas et al. *Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes.* Nature Genetics, 47:106–114, 2015.

32. T. Evgeniou, C. Micchelli and M. Pontil *Learning multiple tasks with kernel methods.* J Mach Learn Res, 6:615–637, 2005.

33. F. Mordelet and J.-P. Vert *ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples.* BMC Bioinformatics, 12(1), 389, 2011.

34. P.F. Przytycki and M. Singh *Differential analysis between somatic mutation and germline variation profiles reveals cancer-related genes.* Genome Medicine, 9–79, 2017.

35. B. Frank, M. Hoffmeister, N. Klopp, T. Illig, J. Chang-Claude and H. Brenner *Single nucleotide polymorphisms in Wnt signaling and cell death pathway genes and susceptibility to colorectal cancer.* Carcinogenesis, 31(8):1381–6, 2010.

36. T. Frixa, A. Sacconi, M. Cioce, G. Roscilli, F.F. Ferrara, L. Aurisicchio et al. *MicroRNA-128-3p-mediated depletion of Drosha promotes lung cancer cell migration.* Carcinogenesis, 39(2):293–304, 2018.

37. S.B. Talla and F.H. Brembeck *The role of Pygo2 for Wnt-catenin signaling activity during intestinal tumor initiation and progression.* Oncotarget, 7(49):80612–32, 2016.

38. N. Attar, S.K. Kurdistani *Exploitation of EP300 and CREBBP Lysine Acetyltransferases by Cancer.* Cold Spring Harbor Perspect in Medicine, 1;7(3), 2017.

39. L. Morera, M. Lübbert and M. Jung *Targeting histone methyltransferases and demethylases in clinical trials for cancer therapy.* Clinical Epigenetics, 8:57, 2016.

40. R.R. Bhanvadia, C. VanOpstall, H. Brechka, N.S. Barashi, M. Gillard, E.M. McAuley et al. *MEIS1 and MEIS2 Expression and Prostate Cancer Progression: A Role For HOXB13 Binding Partners in Metastatic Disease.* Clinical Cancer Research, 1;24(15):3668–80, 2018.

41. T. Hara, M. Schwieger, R. Kazama, S. Okamoto, K. Minehata, M. Ziegler et al. *Acceleration of chronic myeloproliferation by enforced expression of Meis1 or Meis3 in Icsbp-deficient bone marrow cells.* Oncogene, 19;27(27):3865–9, 2008.

42. S. Mohr, C. Doebele, F. Comoglio, T. Berg, J. Beck, H. Bohnenberger et al. *Hoxa9 and Meis1 Cooperatively Induce Addiction to Syk Signaling by Suppressing miR-146a in Acute Myeloid Leukemia.* Cancer Cell, 31(4):549–562, 2017.

43. M. Yang, E. Liu, L. Tang, Y. Lei, X. Sun, J. Hu et al. *Emerging roles and regulation of MiT/TFE transcriptional factors.* Cell Communication and Signaling, 16(1):31, 2018.

44. T. Yamaguchi, Y. Hosono, K. Yanagisawa and T. Takahashi *NKX2-1/TTF-1: an enigmatic oncogene that functions as a double-edged sword for cancer cell survival and progression.* Cancer Cell, 23(6):718–23, 2013.

45. R. Smith, L.A. Owen, D.J. Trem, J.S. Wong, J.S. Whangbo, T.R. Golub et al. *Expression profiling of EWS/FLI identifies NKX2.2 as a critical target gene in Ewing's sarcoma.* Cancer Cell, 9(5):405–16, 2006.

46. S.C. Borinstein, M. Conerly, S. Dzieciatkowski, S. Biswas, M.K. Washington, P. Trobridge et al. *Aberrant DNA methylation occurs in colon neoplasms arising in the azoxymethane colon cancer model.* Molecular Carcinogenesis, 49(1):94–103, 2009.

47. M. Mihara, Y. Yoshida, T. Tsukamoto, K. Inada, Y. Nakanishi, Y. Yagi et al. *Methylation of multiple genes in gastric glands with intestinal metaplasia: A disorder with polyclonal origins.* American Journal of Pathology, 169(5):1643–5, 2006.

48. M. Gladych, R. Cylwa, K. Kielczewski, P. Biecek, T. Liloglou, A. Mackiewicz et al. *The expression signature of cancer-associated KRAB-ZNF factors identified in TCGA pan-cancer transcriptomic data.* Molecular Oncology, 2018.

49. F. Weis, E. Giodice, M. Churcher, L. Jin, C. Hilcenko, C.C. Wong et al. *Mechanism of eIF6 release from the nascent 60S ribosomal subunit.* Natural Structural and Molecular Biology, 22(11):914–9, 2015.

50. M. Presta, P. Chiodelli, A. Giacomini, M. Rusnati and R. Ronca *Fibroblast growth factors (FGFs) in cancer: FGF traps as a new therapeutic approach.* Pharmacology and Therapeutics, 179:171–187, 2017.

51. E.A. Kim, Y.H. Kim, H.W. Kang, H.Y. Yoon, W.T. Kim, S.J. Yun et al. *Lower Levels of Human MOB3B Are Associated with Prostate Cancer Susceptibility and Aggressive Clinicopathological Characteristics.* Journal of Korean Medical Science, 30(7):937–42, 2015.

52. A.T. Nguyen, J. Chia, M. Ros, K.M. Hui, F. Saltel and F. Bard *Organelle Specific O-Glycosylation Drives MMP14 Activation, Tumor Growth, and Metastasis.* Cancer Cell, 32(5):639–53, 2017.

53. J. Zhao, Y. Wang, C. Mu, Y. Xu and J. Sang *MAGEA1 interacts with FBXW7 and regulates ubiquitin ligase-mediated turnover of NICD1 in breast and ovarian cancer cells.* Oncogene, 36(35):5023–5034, 2017.

54. Y.W. Hu, Z.P. Chen, X.M. Hu, J.Y. Zhao, J.L. Huang, X. Ma et al. *The miR-573/apoM/Bcl2A1-dependent signal transduction pathway is essential for hepatocyte apoptosis and hepatocarcinogenesis.* Apoptosis, 20(10):1321–37, 2015.

55. Y.K. Chae, J.F. Anker, B.A. Carneiro, S. Chandra, J. Kaplan, A. Kalyan et al. *Genomic landscape of DNA repair genes in cancer.* Oncotarget, 7(17), 23312–21, 2016.

56. A. Torgovnick and B. Schumacher *DNA repair mechanisms in cancer development and therapy.* Front Genet, 6, 157, 2015.

57. J. Genschel, L.R. Bazemore and P.J. Modrich *Human exonuclease I is required for 5' and 3' mismatch repair.* J Biol Chem, 277:13302–11, 2002.

58. M. Manandhar, K.S. Boulware and R.D. Wood *The ERCC1 and ERCC4 (XPF) genes and gene products.* Gene, 569(2):153–161, 2015.

59. M. Okuda, Nakazawa, C. Guo, T. Ogi and Y. Nishimura *Common TFIIH recruitment mechanism in global genome and transcription-coupled repair subpathways.* Nucleic Acids Res, 45(22):13043–55, 2017.

60. H. Zhou, Y. Qin, S. Ji, J. Ling, J. Fu, Z. Zhuang et al. *SOX9 activity is induced by oncogenic Kras to affect MDC1 and MCMs expression in pancreatic cancer.* Oncogene, 37(7):912–23, 2018.

61. A. Salmaninejad, V. Khoramshahi, A. Azani, E. Soltaninejad, S. Aslani, M.R. Zamani et al. *PD-1 and cancer: molecular mechanisms and polymorphisms.* Immunogenetics, 70(2):73–86, 2018.

62. F. Casciello, F. Al-Ejeh, G. Kelly, D.J. Brennan, S.F. Ngiow, A. Young et al. *G9a drives hypoxia-mediated gene repression for breast cancer cell survival and tumorigenesis.* Proceedings of National Academy of Sciences of United States of America, 114(27):7077–82, 2017.

63. R. Bhatia, S.K. Gautam, A. Cannon, C. Thompson, B.R. Hall, A. Aithal et al. *Cancer-associated mucins: role in immune modulation and metastasis.* Cancer and Metastasis Reviews, 2019.

64. V. Patil, J. Pal and K. Somasundaram *Elucidating the cancer-specific genetic alteration spectrum of glioblastoma derived cell lines from whole exome and RNA sequencing.* Oncotarget, 6(41):43452–71, 2015.

65. H. Guo, L. Zhao, B. Shi, J. Bao, D. Zheng, B. Zhou et al. *GALNT5 uaRNA promotes gastric cancer progression through its interaction with HSP90.* Oncogene, 1, 2018.

66. K. Xie, K. Zhang, J. Kong, C. Wang, Y. Gu, C. Liang et al. *Cancer-testis gene PIWIL1 promotes cell proliferation, migration, and invasion in lung adenocarcinoma.* Cancer Med, 7(1):157–166, 2018.

67. R. Iliev, M. Stanik, M. Fedorko, A. Poprach, P. Vychytilova-Faltejskova, K. Slava et al. *Decreased expression levels of PIWIL1, PIWIL2, and PIWIL4 are associated with worse survival in renal cell carcinoma patients.* OncoTargets and Therapy, 9:217–22, 2016.

68. S. Jitsukawa, R. Kamekura, K. Kawata, F. Ito, A. Sato, H. Matsumiya et al. *Loss of sorting nexin 5 stabilizes internalized growth factor receptors to promote thyroid cancer progression.* Journal of Pathology, 243(3):342–353, 2017.

69. L. Li, M. Peng, W, Xue, Z. Fan, T. Wang, J. Lian et al. *Integrated analysis of dysregulated long non-coding RNAs/microRNAs/mRNAs in metastasis of lung adenocarcinoma.* Journal of Translational Medicine, 27;16(1):372, 2018.

70. S.Y. Lin, Y.B. Lv, G.X. Mao, X.J. Chen and F. Peng *The effect of centromere protein U silencing by lentiviral mediated RNA interference on the proliferation and apoptosis of breast cancer.* Oncology Letters, 16(5):6721–8, 2018.

71. I.L. Hsu, C.Y. Chou, Y.Y. Wu, J.E. Wu, C.H. Liang, Y.T. Tsai et al. *Targeting FXYD2 by cardiac glycosides potently blocks tumor growth in ovarian clear cell carcinoma.* Oncotarget, 7(39):62925–38, 2016.

72. S. Yu, H. Bian, X. Gao and L. Gui *Annexin A9 promotes invasion and metastasis of colorectal cancer and predicts poor prognosis.* Journal of Molecular Medicine, 41(4):2185–92, 2018.

73. C. Lobry, P. Oh, M.R. Mansour, A.T. Look and I. Aifantis *Notch signaling: switching an oncogene to a tumor suppressor.* Blood, 123(16):2451–9, 2014.

74. T. Sakoparnig, P. Fried et N. Beerenwinkel *Identification of constrained cancer driver genes based on mutation timing.* PLoS Comput Biol, 11(1):e1004027, 2015.

75. B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, J. Platt *Support vector method for novelty detection.* Mach Learn Interpret Neuroimaging (1999), 582-588, 1999.

76. R. Vert and J.-P. Vert *Consistency and convergence rates of one-class SVMs and related algorithms.* J. Mach. Learn. Res., 7:817-54, 2006.

77. R.I. Kondor and J. Lafferty *Diffusion kernels on graphs and other discrete input spaces.* Proc Int Conf Mach Learn,3:315-322, 2002.

78. L. Cowen, T. Ideker, B.J. Raphael and R. Sharan *Network propagation: a universal amplifier of genetic associations.* Nature Rev Genet, 2017.

79. T.S.K. Prassad, R. Goel, K. Kandasamy, S. Keerthimukar, S. Kumar, S. Mathivanan et al. *Human Protein Reference Database - 2009 update.* Nucleic Acids Res, 37, D767-72, 2009.

80. A. Karatzoglou, A. Smola, K. Hornik and A. Zeileis *kernlab – An S4 Package for Kernel Methods in R.* J Stat Softw, 11-9, 1-20, 2004.

Figure legends

Fig 1. ROC curves for TSGs (left) and OGs (right) and the TUSON train set.

Fig 2. ROC curves for TSGs (left) and OGs (right) and the 20/20 train set.

Fig 3. Distribution of the number of TSGs (left) and OGs (right) per cancer type.

Supporting information legends

S1 Table **List of cancer types.** Cancer types derived from annotations in the 20/20 mutation dataset along with their numbers of associated OG and TSG.

S2 Table **Description of cancer types.** Descriptors of all cancer types according to their localizations and types that are used to compute the disease kernel used by LOTUS2.

S3 Table **TSG and OG rankings for LOTUS with the 20/20, the TUSON and the CGC v86 datasets.** Note that the training sets were removed every time.

Tables

Driver type \ Method	TUSON	LOTUS
OG	3,286	**990**
TSG	626	**127**

Table 1. Comparison of Consistency Errors for OG and TSG prediction between TUSON and LOTUS.

Driver type \ Method	20/20+	DiffMut	LOTUS
OG	1,831	4,254	**782**
TSG	845	2,537	**468**

Table 2. Comparison of Consistency Errors for OG and TSG prediction between 20/20+, DiffMut and LOTUS.

Driver type \ Method	MutSigCV	LOTUS
OG	6,294	**1,929**
TSG	7,232	**2,990**

Table 3. Comparison of Consistency errors for OG and TSG prediction between MutSigCV and LOTUS.

Train set \ Kernel	$K_{mutation}$	K_{PPI}	K_{degree}	$K_{mutation} + K_{PPI}$
TUSON datasets	2,904	1,574	1,659	**990**
20/20 datasets	2,453	1,642	1,774	**782**
MutSig datasets	2,292	1,450	**1,306**	1,929

Table 4. Consistency error of LOTUS for OG prediction in the pan-cancer setting, with different gene kernels (columns) and different gold standard sets of known OGs and mutations (rows).

Train set \ Kernel	$K_{mutation}$	K_{PPI}	K_{degree}	$K_{mutation} + K_{PPI}$
TUSON datasets	393	1,413	1,965	**127**
20/20 datasets	971	2,460	2,994	**468**
MutSig datasets	4,335	4,017	4,253	**2,990**

Table 5. Consistency error of LOTUS for TSG prediction in the pan-cancer setting, with different gene kernels (columns) and different gold standard sets of known TSGs and mutations (rows).

Method \ Number of considered predictions	20	50	100
LOTUS	5	10	19
TUSON	**13**	**25**	**35**

Table 6. Number of predictions belonging to the TSGs in CGCv86, when the methods are run with the TUSON mutation database and train set.

Method \ Number of considered predictions	20	50	100
LOTUS	3	7	**16**
TUSON	**7**	**11**	12

Table 7. Number of predictions belonging to the OGs in CGCv86, when the methods are run with the TUSON mutation database and train set.

Method \ Number of considered predictions	20	50	100
LOTUS	4	**10**	**16**
DiffMut	1	4	10
20/20+	**7**	9	**16**

Table 8. Number of predictions belonging to the TSGs in CGCv86, when the methods are run with the 20/20 mutation database and train set.

Method \ Number of considered predictions	20	50	100
LOTUS	3	5	9
DiffMut	2	3	4
20/20+	**10**	**15**	**20**

Table 9. Number of predictions belonging to the OGs in CGCv86, when the methods are trained with the 20/20 mutation database and train set.

Disease	Number of TSGs	DiffMut	Single-Task LOTUS	Aggregation LOTUS	Multitask LOTUS	Multitask LOTUS2
ALL	38	7,122	1,431	783	709	**631**
Astrocytoma	17	7,605	2,612	49	38	**0**
BladUroCarc	9	4,852	2052	173	138	**115**
BreastAdeno	34	3,250	1,837	801	**770**	778
CLL	21	4,253	1,336	895	**894**	921
Colorectal	53	7,600	3,640	911	870	**825**
EndomCarc	18	5,831	1,222	89	82	**59**
GliobMulti	24	4,776	2,771	166	191	**153**
HNSC	24	5,819	3,051	681	**571**	595
Kidney Cancer	19	6,133	2,766	2,512	**2,474**	**2,474**
LAML	56	5,947	1,936	1,483	1,451	**1,328**
LiverHepCarc	10	3,768	602	221	**156**	172
Low-Grade Glioma	17	6,047	2,712	49	38	**0**
LungAdeno	30	6,773	4,712	339	341	**323**
LungSquaCarc	16	6,829	3,868	53	57	**25**
LungSmallCarc	28	8,883	5,746	58	68	**17**
Lymphoma B-Cell	37	6,383	**1,754**	2,238	2,284	2,252
Medulloblastoma	14	6,692	1,123	265	247	**230**
Melanoma	31	6,719	2,467	459	365	**233**
Multiple Myeloma	7	5,165	**2,871**	3,754	3,871	3,683
Ovarian	22	6,481	2,632	724	627	**593**
PancAdeno	13	3,140	1,777	140	**118**	123
ProstAdeno	7	6,565	2,345	457	**371**	514
Rhabd	6	4,871	1,957	181	111	**26**
Soft-Tissue Sarcoma	23	8,572	4,447	2,008	1,992	**1,970**
StomAdeno	17	6,530	2,878	331	**319**	322
ThyrCarc	8	10,352	2,834	**1,222**	1,325	1,538

Table 10. *CE* for prediction of disease specific TSGs in the multitask setting. ALL stands for Acute Lymphocytic Leukemia, BladUroCarc for Bladder Urothelial Carcinoma, BreastAdeno for Breast Adenocarcinoma, CLL for Chronic Lymphocytic Leukemia, EndomCarc for Endometrial Carcinoma, GliobMulti for Glioblastoma Multiform, HNSC for Head and Neck Squamous Cell Carcinoma, LAML for Acute Myeloid Leukemia, LiverHepCarc for Liver Hepatocellular Carcinoma, LungAdeno for Lung Adenocarcinoma, LungSquaCarc for Lung Squamous Cell Carcinoma, LungSmallCarc for Lung Small Cell Carcinoma, PancAdeno for Pancreatic Adenocarcinoma, ProstAdeno for Prostate Adenocarcinoma, Rhabd for Rhabdomyosarcoma, StomAdeno for Stomach Adenocarcinoma and ThyrCarc for Thyroid Carcinoma.

Disease	Number of OGs	DiffMut	Single-Task LOTUS	Aggregation LOTUS	Multitask LOTUS	Multitask LOTUS2
ALL	52	8,479	2,649	1,232	1,269	**1,145**
Astrocytoma	13	7,847	2,894	75	63	**13**
BladUroCarc	10	5,324	1,578	210	**139**	140
BreastAdeno	19	2,672	1,371	852	806	**792**
CLL	19	4,582	3,821	1,537	1,501	**1,462**
Colorectal	23	4,043	3,376	818	784	**758**
EndomCarc	8	5,112	1,671	122	128	**105**
GliobMulti	22	4,915	2,539	143	128	**106**
HNSC	23	4,539	2,917	**1,305**	1,500	1,504
Kidney Cancer	11	5,774	1,903	**543**	600	763
LAML	56	4,990	2,623	1,418	1,408	**1,307**
Low-Grade Glioma	10	3,753	1,541	46	33	**3**
LungAdeno	26	4,510	2,038	84	79	**40**
LungSmallCarc	6	3,243	2,129	1,061	**666**	864
LungSquaCarc	24	5,737	1,641	67	54	**17**
Lymphoma B-Cell	34	4,765	2,424	1,712	1,714	**1,669**
Medulloblastoma	5	7,165	58	93	34	**25**
Melanoma	35	3,377	1,925	1,576	1,550	**1,525**
Multiple Myeloma	9	3,466	2,870	**1,823**	1,877	2,026
Neuroblastoma	5	5,298	3,830	2,078	**2,077**	2,101
Ovarian	12	6,371	3,606	1,256	**869**	870
PancAdeno	6	1,464	1,142	498	426	**302**
ProstAdeno	13	6,523	2,451	**955**	1,599	1,475
Rhabd	7	8,265	1,978	172	104	**30**
Soft-Tissue Sarcoma	38	8,886	2,480	**2,424**	2,466	2,444
StomAdeno	10	2,235	750	**85**	127	97
ThyrCarc	8	8,407	2,656	547	612	**494**

Table 11. *CE* for prediction of disease specific OGs in the multitask setting. ALL stands for Acute Lymphocytic Leukemia, BladUroCarc for Bladder Urothelial Carcinoma, BreastAdeno for Breast Adenocarcinoma, CLL for Chronic Lymphocytic Leukemia, EndomCarc for Endometrial Carcinoma, GliobMulti for Glioblastoma Multiform, HNSC for Head and Neck Squamous Cell Carcinoma, LAML for Acute Myeloid Leukemia, LungAdeno for Lung Adenocarcinoma, LungSquaCarc for Lung Squamous Cell Carcinoma, LungSmallCarc for Lung Small Cell Carcinoma, PancAdeno for Pancreatic Adenocarcinoma, ProstAdeno for Prostate Adenocarcinoma, Rhabd for Rhabdomyosarcoma, StomAdeno for Stomach Adenocarcinoma and ThyrCarc for Thyroid Carcinoma.

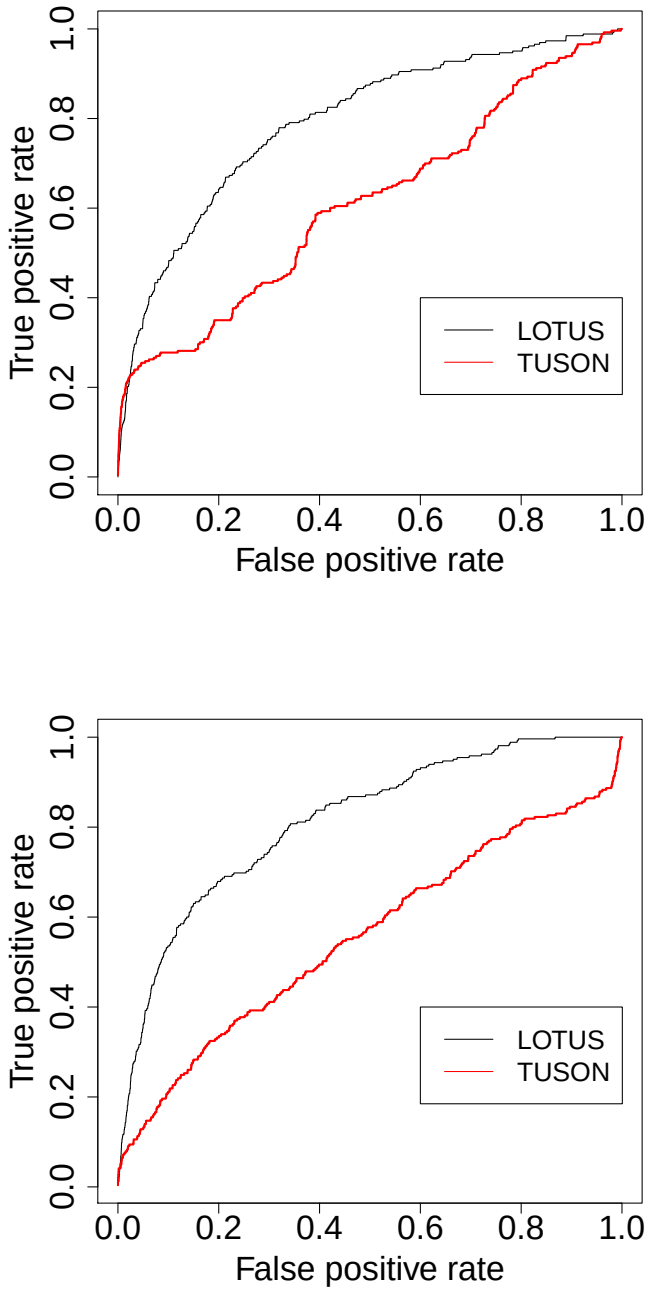

Fig 1. ROC curves for TSGs (top) and OGs (bottom) for LOTUS and TUSON, run with the TUSON mutation database train sets.

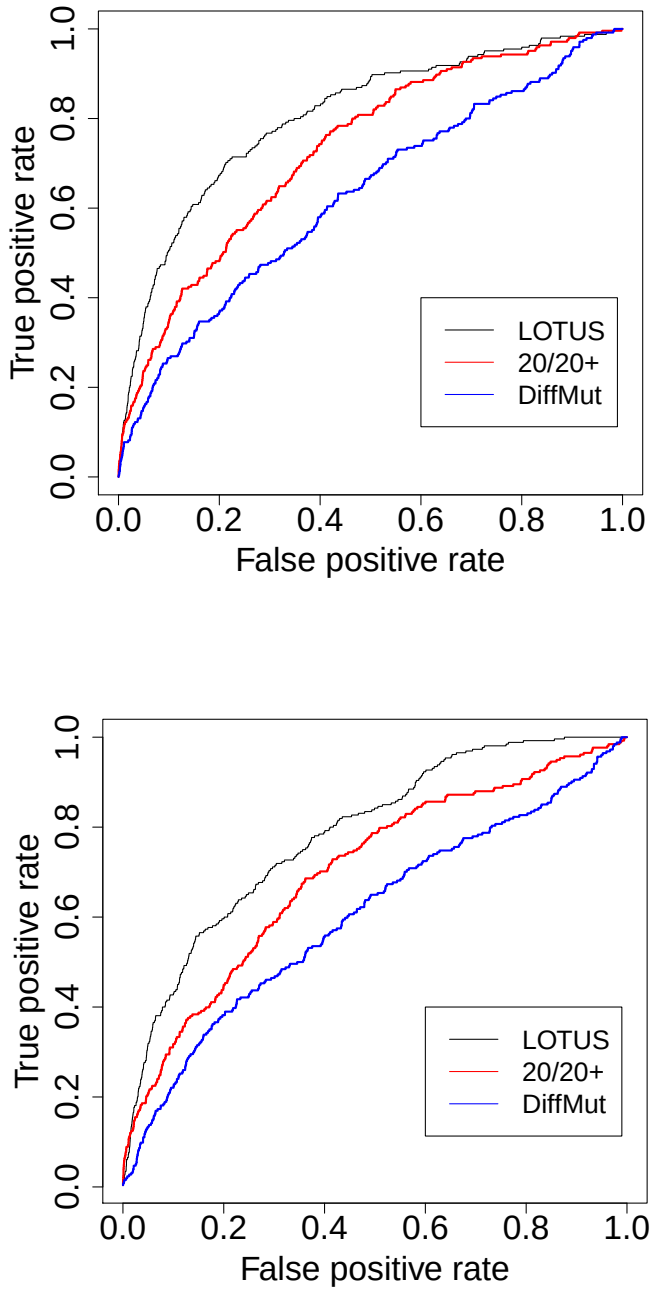

Fig 2. ROC curves for TSGs (top) and OGs (bottom) for LOTUS, 20/20+ and DiffMut, run with the 20/20 mutation database and train sets.

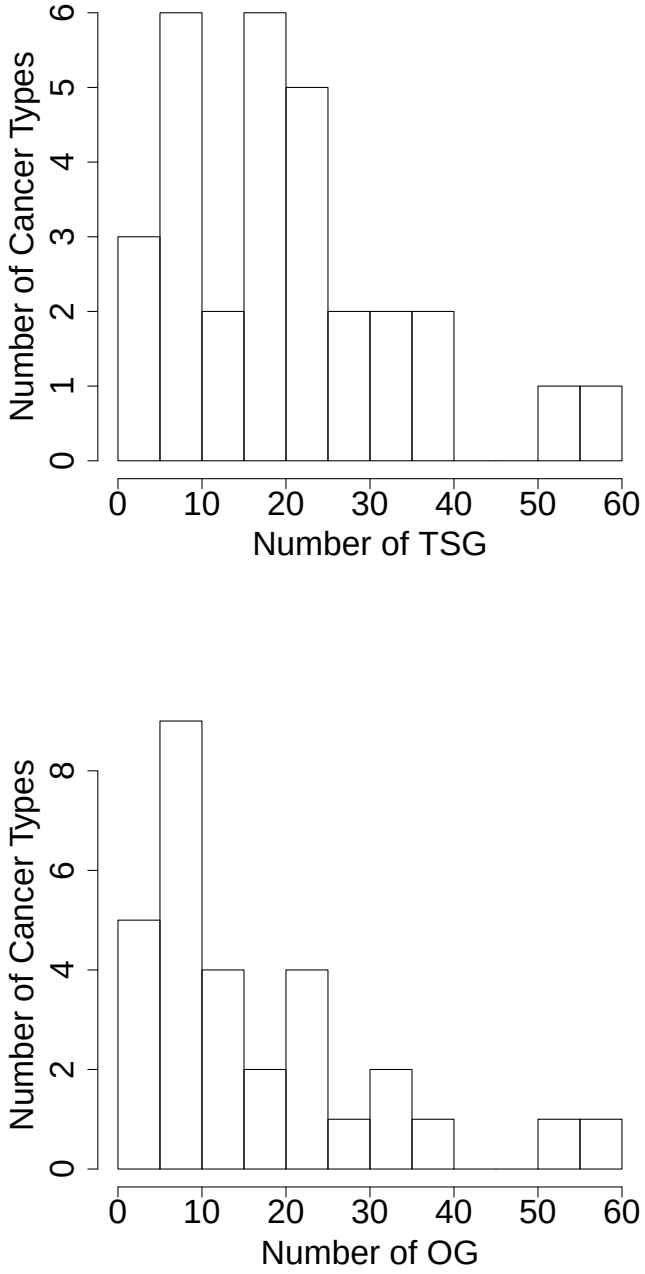

Fig 3. Number of TSGs (top) and OGs (bottom) per cancer type: for example, 3 cancer types have between 0 and 5 TSGs.