



HAL
open science

LOTUS: a single-and multi-task machine-learning algorithm for the prediction of cancer driver genes

Olivier Collier, Véronique Stoven, Jean-Philippe Vert

► **To cite this version:**

Olivier Collier, Véronique Stoven, Jean-Philippe Vert. LOTUS: a single-and multi-task machine-learning algorithm for the prediction of cancer driver genes. 2018. hal-01857394v1

HAL Id: hal-01857394

<https://hal.science/hal-01857394v1>

Preprint submitted on 15 Aug 2018 (v1), last revised 21 Oct 2023 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LOTUS: a single- and multi-task machine-learning algorithm for the prediction of cancer driver genes

Collier Olivier^{1,*}, Stoven Véronique^{2,3,4}, Vert Jean-Philippe^{2,3,4}

1 Modal'X, UPL, Univ Paris Nanterre, F92000 Nanterre France

2 MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, 77300 Fontainebleau, France

3 Institut Curie, 75248 Paris Cedex 5, France

4 INSERM U900, 75248 Paris Cedex 5, France

* olivier.collier@parisnanterre.fr

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. 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 informations 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 three 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 thought to be driven by some important genes that should be targeted by new treatments. Unfortunately, there is a small number of such genes, so that it is of crucial importance to design algorithms capable of finding genes with the highest oncogenic potential. Our new method analyses in particular data of mutations but also other sources of informations to establish a list of genes that should be investigated in priority. Moreover, our algorithm can differentiate between several types of cancer and share information between them to improve the prediction for every disease. We showed that in several contexts our algorithm beats its concurrents.

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 developing 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. The results of some of these tools are available at the Driver DB database (*cf.* [24]).

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 [25] or DOTS-Finder [26] combine mutation frequencies and functional impact of mutations to identify OGs and TSGs. Also in this category, the 20/20+ method [27] 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 [28] 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 [27], 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. 80

In the present paper, we propose a new method for cancer driver gene prediction 81
called *Learning Oncogenes and Tumor Suppressors* (LOTUS). Like 20/20+, LOTUS is 82
a machine learning-based method, meaning that it starts from a list of known driver 83
genes in order to "learn" the specificities of such genes in order to identify new ones. In 84
addition, LOTUS presents two unique characteristics with respect to previous work in 85
this field. First, it combines informations from all three types of informations likely to 86
contain information to predict cancer genes (mutation frequency, functional impact, and 87
pathway-based informations). This integration of heterogeneous informations is carried 88
out in a unified mathematical and computational framework thanks to the use of kernel 89
methods [29], and allows in principle to integrate other sources of data if available, such 90
as transcriptomic or epigenomic information. More precisely, in our implementation we 91
predict cancer driver genes based not only on gene mutations features like 'Mutation 92
Frequency' and 'Functional Impact' methods do, but also on known protein-protein 93
interaction (PPI) network like 'Pathway-based' methods do. Indeed, the use of PPI 94
information is particularly relevant since it has been reported that proteins encoded by 95
driver genes are more likely to be involved in protein complexes and share higher 96
'betweenness' than a typical protein [25]. Second, LOTUS can predict cancer genes in a 97
pan-cancer setting, as well as for specific cancer types, using a multitask learning 98
strategy [30]. The pan-cancer setting has been adopted by most available prediction 99
methods, since more data is available when pooling together all cancer types. The 100
cancer type-specific prediction problem has been less explored so far, because the 101
number of known driver genes for a given cancer is often too small to build a reliable 102
prediction model, and because the amount of data such as somatic mutations to train 103
the model is smaller than in the pan-cancer setting. However, the search of cancer 104
specific driver genes is relevant, because cancer is a very heterogeneous disease: different 105
tumorigenic processes seem to be at work in different tissue types, and consequently 106
every cancer type probably has its own list of driver genes (*cf.* [15]). LOTUS implements 107
a multi-task algorithm that predicts new driver genes for a given cancer type based on 108
its known driver genes, while also taking into account the driver genes known for other 109
types of cancers according to their similarities with the considered type of cancer. Such 110
approaches are of particular interest when the learning data are scarce in each 111

individual tasks: they increase the amount of data available for each task and thus
perform statistically better. To our knowledge, this is the first contribution in which a
machine learning multi-task algorithm is used for the prediction of cancer driver genes.

We compare LOTUS to the best three state-of-the art cancer prediction methods
according to [27] according to several criteria, including intrinsic consistency and
prediction accuracy on known gold standards. We show that that LOTUS outperforms
the state-of-the-art according to these criteria, 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.

Materials and methods

Pan-cancer LOTUS

LOTUS is a new machine learning-based method to predict new cancer genes, given a
list of know ones. In the simplest, pan-cancer setting, we thus assume given a list of N
known cancer genes $\{g_1, \dots, g_N\}$, 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 a also
cancer gene. Since TSGs and OGs have different characteristics, we treat them
separately and build in fact two scoring functions f_{TSG} and f_{OG} 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 [31], a classical method for novelty detection and density level set
estimation [32]. The scoring function $f(g)$ learned by a OC-SVM given a training set
 $\{g_1, \dots, g_N\}$ of known cancer genes takes the form:

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

where $\alpha_1, \dots, \alpha_N$ are weights optimized during the training of OC-SVM [31], 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 know cancer 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. This leaves a lot of freedom to create specific kernels encoding one’s prior knowledge about relevant information to predict cancer genes [29]. In addition, one can easily combine heterogeneous information in a single kernel by, e.g., summing together 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, 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, 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 [25]. We chose them because they were found to best discriminate OGs and TSGs by the TUSON algorithm [25] and were also all found among the most important features selected by the random forest algorithm used by the 20/20+ method [27]. 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 [33, 34]. More precisely, we used a diffusion kernel of the form $K = exp(-L)$, where

$L = I - D^{-1/2}AD^{-1/2}$ is the normalized Laplacian of the graph. Here A stands
 for adjacency matrix ($A_{i,j} = 1$ 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 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
 summing together the mutation and PPI kernels:

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

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 genes to those cancer 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 genes to learn a scoring
 function. Since cancer 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 cancer genes across cancer types, using the framework of
 multitask learning [30,35]. Instead of starting from a list of known cancer genes, we now
 start from a list of known (cancer gene, cancer type) pairs of the form
 $\{(g_1, d_1), \dots, (g_N, d_N)\}$, 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:

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

where K_{gene} is a kernel between genes such as the one 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 $\{(g_1, d_1), \dots, (g_N, d_N)\}$ 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 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.

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:

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

which makes a compromise between the two extreme uniform and Dirac kernels [30]. 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 similarities with other cancers. Therefore, known cancer 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 our prior hypothesis about how similar cancer genes are likely to be between different cancers. To create K_{cancer} , we first represent each cancer type as a 50-dimensional binary vector as follows. The first 15 bits correspond to a list of cancer type characteristics used in COSMIC to describe tumors: adenocarcinoma, benign, blastoma, carcinoma, gastro-intestinal stromal tumour, germ cell tumour, glioma, leukemia, lymphoma, melanoma, meningioma, myeloma, neuro-endocrine, sarcoma, stromal. The last 35 components correspond to localization characteristics also used in COSMIC to describe tumors: bile ducts, bladder, blood vessels, bone, bone marrow, breast, central nervous system, cervix, colorectal, endocrine glands, endometrium, eye, gall bladder, germ cell, head and neck, heart, intestine, kidney, liver, lung, lymphocytes, mouth, muscle, nerve, oesophagus, ovary, pancreas, pituitary glands, prostate, salivary glands, skin, soft tissue, stomach, tendon, thyroid. A disease might be assigned one or several types and be associated to one or several locations. For example, we associated neurofibroma with one localization "nerve" and two types "benign" and "sarcoma", so that neurofibroma is described by a vector with three 1's and forty-seven 0's. For each disease, we constructed the list of binary features by documenting every disease in the literature. The corresponding vectors encoding the considered disease are given in Supplementary Materials. Finally, if $\Psi(d) \in \mathbb{R}^{50}$ 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.:

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

Data

198

In all experiments, we restrict ourselves to the total set of 17948 genes considered in the TUSON, 20/20 and MutsigCV papers, as candidate driver genes. Somatic mutations were collected from COSMIC [14], TCGA (<http://cancergenome.nih.gov/>) and [18]. This dataset contains a total of 1195223 mutations across 8207 patients. We obtained

199

200

201

202

the PPI network from the HPRD database release 9 from April 13, 2010 [36]. It contains 39239 interactions among 7931 proteins. As for known pan-cancer driver genes, we consider three lists in our experiments: (i) the TUSON train set, 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 [27] to train the 20/20+ method, contains 53 OGs and 60 TSGs; finally, (iii) the CGCv84 train set consists of two broader lists that we extracted from (CGC release v84 of the COSMIC database): the list of all 136 dominant driver genes in the CGC database that were not reported through translocations (i.e. OGs), and the list of all 138 recessive driver genes in the CGC database that were not reported through translocations (i.e. TSGs). For cancer type-specific lists of driver genes, we only consider the CGCv84 train set. We distinguished 175 diseases based on the available annotations describing patients in COSMIC, using as few interpretations as possible: for example, we merged together diseases corresponding to obvious synonyms like singular and plural forms of the same cancer name. 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, 1 to 20 TSGs/OGs were known in CGCv84. We considered only diseases with at least 4 known TSGs or OGs available, in order to have enough learning data points to perform a cross-validation scheme, which led us to consider 27 diseases for TSG prediction and 22 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 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 $\{2^{-5/2}, 2^{-4/2}, \dots, 2^{5/2}\}$ that minimizes the mean CE over the folds.

Other driver prediction methods

We compare the performance of LOTUS to three other state-of-the-art methods: MutsigCV [21], which is a frequency-based method, and TUSON [25] and 20/20+ [27] that combine frequency and functional information.

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 [27]. 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 MutSig 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.

Code and data availability

We implemented all methods in R using in particular the kernlab package for OC-SVM [37]. The code and data to reproduce all experiments are available at <http://members.cbio.mines-paristech.fr/~ocollier/lotus.html>.

Results

279

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

280

281

We propose LOTUS, a new method to predict cancer driver genes. LOTUS is a machine learning-based method that estimates a scoring function predicting the probability that a given candidate gene is an OG or a TSG, 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 cancer genes, where the weights are optimized by a OC-SVM algorithm. The similarities themselves are derived from the analysis of somatic mutation patterns in the genes, 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.

282

283

284

285

286

287

288

289

290

291

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 in the same way across all cancer types, and a new strategy that shares more information across similar cancer types. More details about the mathematical formulation and algorithms implemented in LOTUS are provided in the Methods section.

292

293

294

295

296

297

298

299

In the following, we assess the performance of LOTUS in the pan-cancer regime, where we compare it to three state-of-the-art methods (TUSON, MutsigCV and 20/20+), and in the cancer type specific regime, where we illustrate the importance of the multitask learning strategies.

300

301

302

303

Cross-validation performance for pan-cancer driver gene prediction

304

305

We first study the pan-cancer regime where cancer is considered as a single disease, and where we search for driver genes involved in some types of cancer. Several

306

307

computational methods have been proposed to solve this problem in the past, and we 308
compare LOTUS with the top 3 methods in terms of performance according to a recent 309
benchmark [27]: MutsigCV [21], which is a frequency-based method, and TUSON [25] 310
and 20/20+ [27], which combine frequency and functional information. 311

While MutsigCV and TUSON are unsupervised methods that scores candidate genes 312
independently of any training set of known drivers, 20/20+ depends on a training set, 313
just like LOTUS. To perform a comparison as fair as possible between different methods, 314
we collected the training sets of TUSON and 20/20+, and evaluated the performance of 315
LOTUS on each of these datasets by 5-fold CV repeated twice (see Methods). For 316
TUSON and 20/20+, we used the prediction results available in the corresponding 317
papers, in order to evaluate the CE error as the mean number of non-driver genes in the 318
test set that are ranked before driver genes of the TUSON and 20/20 train sets. These 319
ranks were obtained by training these two algorithms on their respective train set, i.e. 320
the TUSON and 20/20 train sets. We note that this gives an advantage to TUSON and 321
20/20+ compared to LOTUS, since for the former two methods the training set is used 322
both to define the score and to assess the performance, while for TUSON the k -fold CV 323
procedure ensures that different genes are used to train the model and to test its 324
performance. However we note that the 20/20+ score itself is obtained by a bootstrap 325
procedure similar to our cross-validation approach [27]. This allows us to make fair 326
comparisons between TUSON, MutsigCV and LOTUS (trained on TUSON train set), 327
on the one hand, and between 20/20+, MutsigCV and LOTUS (trained on 20/20 train 328
set), on the other hand. We further note that MutsigCV also provides a ranked list of 329
genes, but does not make the difference between TSG and OG. Therefore, it is not 330
dependent from a train set, and the CE in this case is obtained by averaging the 331
numbers of non-driver genes ranked before each driver genes in the considered train set. 332

The consistency errors CE (see Methods) for the different methods and the different 333
training sets are presented in Table 1 for OGs and in Table 2 for TSGs. When analyzing 334
these results, one should keep in mind that the total number of cancer driver genes is 335
still a subject of debate, but it is expected to be much lower than the size of the test set 336
of 17849 genes, and it should rather be in the range of a few hundreds. Therefore, 337
consistency errors above a few thousand can be considered as poor performance results. 338

These results show that LOTUS significantly outperforms all other algorithms in 339

Train set \ Method	MutsigCV	TUSON	20/20+	LOTUS
TUSON train set	4 489	3 286	×	931
20/20 train set	5 823	×	1 831	819

Table 1. Consistency error for OG prediction in the pan-cancer setting, for different methods (columns) and different gold standard sets of known OG (rows).

Train set \ Method	MutsigCV	TUSON	20/20+	LOTUS
TUSON train set	1 443	626	×	130
20/20 train set	2 447	×	845	514

Table 2. Consistency error for TSG prediction in the pan-cancer setting, for different methods (columns) and different gold standard sets of known TSG (rows).

term of CE , both for TSG and OG predictions. The performances of TUSON and 20/20+ are in the same range, although we should keep the above remark in mind. The results also show that MutsigCV does not perform as well as the three other methods, at least on the datasets used here.

It is interesting to note that, for all methods, 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 would 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 13525 mutations affecting TSGs and 7717 mutations affecting OGs. Therefore, the data were richer for TSG, which might have contributed to the difference in prediction performance.

The benefits of combining mutations and PPI informations

LOTUS, 20/20+, MutsigCV 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 results are presented in Table 3 and Table 4 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 and Table 2).

Train set \ Kernel	$K_{mutation}$	K_{PPI}	$K_{mutation} + K_{PPI}$
TUSON train set	2 333	1 565	931
20/20 train set	2 072	2 013	819

Table 3. 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 (rows).

Train set \ Kernel	$K_{mutation}$	K_{PPI}	$K_{mutation} + K_{PPI}$
TUSON train set	388	1 645	130
20/20 train set	901	1 858	514

Table 4. 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 (rows).

These results show that, both for OG and TSG, using both mutation and PPI information dramatically improves the prediction performance over using only one type of them. This underlines the fact that mutation and PPI are complementary informations 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.

Performance on CGCv84 prediction in the pan-cancer regime

We now evaluate the generalization properties of the different methods on new unseen data as external test set. We trained LOTUS with the full 20/20 or TUSON train sets, made predictions on the COSMIC database, and evaluated the CE on the CGCv84 database, under the assumption that this database is enriched in driver genes (a criterion that was also used in [27]). We compared these CE to those of TUSON (for the TUSON train set) and 20/20+ (for the 20/20 train set). For LOTUS, TUSON and 20/20+, genes belonging to their corresponding train set were removed from the CGCv84 database before calculating the CE. For MutsigCV, the CE was calculated based on the ranked list of genes provided in the corresponding paper [21], removing genes of the TUSON train set from CGCv84 database when MutsigCV is compared to TUSON and LOTUS (Table 5), and removing genes from the 20/20 train set from CGCv84 when MutsigCV is compared to 20/20+ and LOTUS (Table 6). These results are illustrated by the corresponding ROC curves, see Fig 1 and 2.

Driver type \ Method	MutsigCV	TUSON	LOTUS
TSG	6282	6820	3769
OG	7359	7179	2333

Table 5. CE obtained on the CGCv84 data set with the TUSON train set.

Driver type \ Method	MutsigCV	20/20+	LOTUS
TSG	7000	4846	3949
OG	7000	3949	2872

Table 6. CE obtained on the CGCv84 data set with the 20/20 train set.

As observed in Table 5, Table 6 and Figure 4, LOTUS presents better prediction *CE*, and therefore better generalization properties on CGCv84, than the three other methods. Interestingly, the two machine learning-based algorithms, LOTUS and 20/20+ (CE between 4846 and 7359), perform better than the two other algorithms (CE between 2333 and 3949).

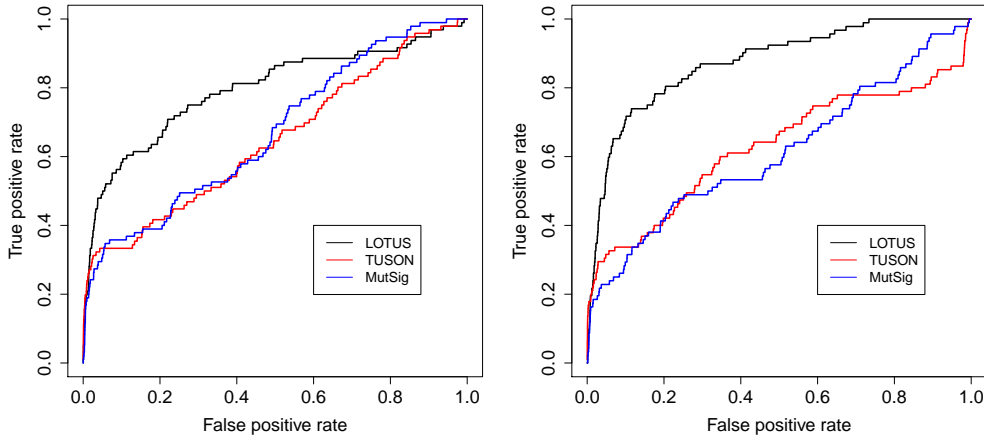


Fig 1. ROC curves for TSGs (left) and OGs (right) and the TUSON train set.

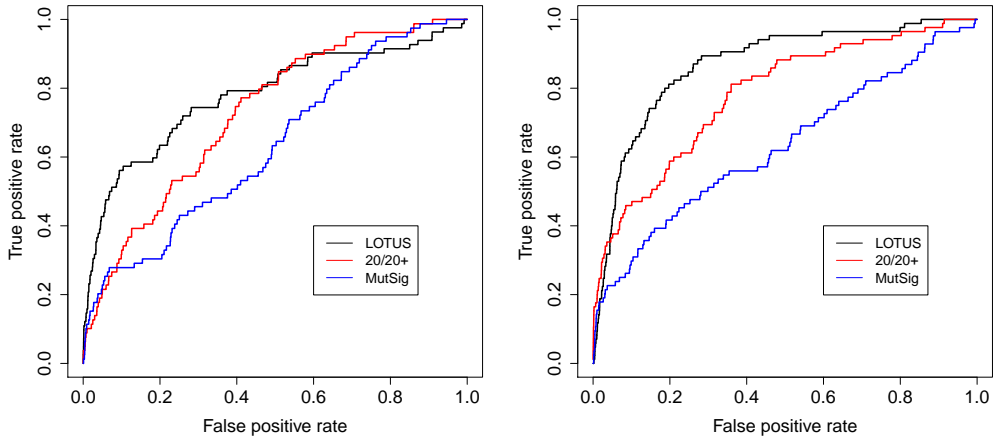


Fig 2. ROC curves for TSGs (left) and OGs (right) and the 20/20 train set.

Analysis of new driver genes predicted by LOTUS

We now test the ability of LOTUS to make new driver gene predictions. We trained LOTUS with the CGCv84 train set, made predictions over the complete COSMIC database (17948 genes). The complete results are given in Supplementary Table 3.

We tried to validate some of these predictions based on independent sources of information. Complete analyses of the predicted OG and TSG rankings is out of the scope of this paper. However, we considered the 20 best ranked TSGs and OGs.

Among the 20 best ranked TSGs, 4 genes are actually known TSGs that were not included yet in CGCv84: PTEN [38], FAT1 [39], STAG1 [40], TRAP1 [41].

Interestingly, 8 genes out of these 20 best ranked TSGs are genes coding for proteins 411
involved in DNA repair, a role closely related to genome maintenance and cancer [42,43]. 412
These genes are EXO1 [44], ERCC1 [45], GTF2H1 and GTF2H4 (both involved in the 413
TFIIH complex [46]), NTHL1 [47], ATR [48], RAD52 [49] and RPA4 [50]. In addition to 414
these clues referring to the DNA repair functions, many additional studies related to 415
these genes are available in the literature, underlining their role in various types of 416
cancers, which provides another clue for them to be confident TSG candidates. In 417
particular, mutations in NTHL1 are known to predispose to colorectal cancer, which is 418
an additional argument in favor of NTHL1 being a strong candidate TSG [51,52]. 419

For 2 additional genes, we found recent publications indicating that they could 420
potentially act as TSG, at least in given tumor types. A non-coding RNA directed 421
against GALNT5 is overexpressed in gastric cancer, inhibiting the translation of its 422
target gene, and the level of expression of this non-coding RNA is correlated with 423
cancer progression and metastasis [53]. These results are consistent with a TSG role of 424
GALNT5 in gastric cancer. In the case of PIWIL1, a recent paper concluded that it was 425
an epidriver gene for lung adenocarcinoma, which means that aberrant methylation of 426
its promoter region plays a role in the development of this cancer [54]. 427

Among the 20 best ranked putative OGs, 3 genes are actually known OGs at least 428
for some types of cancers, and not yet included in CGCv84: MAP3K1 [55], PLCE1 [56], 429
FGF5 [57]. 430

One gene, GATA3, is known to behave either as an OG or as a TSG, depending on 431
the genetic context of the disease [58]. In fact, the literature provides other examples of 432
genes able to switch from oncogenes to tumor suppressor genes, depending on the 433
context [59]. In line with this remark, 3 genes among the 20 best ranked OGs are 434
known TSGs. They could in fact have a potential property to be OG or TSG, 435
depending on the context: PIK3R1 [60], APC [61], TP53 [62]. 436

Mutations in the 6th ranked HTPO gene seems to be causal in some cancer types, 437
where it could therefore be considered as an oncogene [63]. 438

Finally 4 genes are known to be associated to cancer development and progression in 439
some cancer types, are studied as biomarkers or as therapeutic targets, which indicates 440
that they could indeed be credible oncogene candidates: PPARP10 [64], HTR2B [65], 441
STAP2 [66], FXVD2 [67]. 442

Taken together, these results show that LOTUS was able to retrieve among the top ranked genes, known driver genes that were absent in the training set. They also show that LOTUS suggested high confidence driver genes for which many references about their implication in cancer are available.

Identification of cancer-specific driver genes with multi-task LOTUS

In this section, we do not consider cancer as a single disease, but as a variety of diseases with different histological types and that can affect various organs. It is then important to identify driver genes for each type of cancer. The classical 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 were data available for all cancers are used.

In this context, we investigate the 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 derived a list of 175 cancer diseases from COSMICv84 as explained in Methods. This complete list is available in Supplementary Table 1. As expected, many cancer types have only few known cancer genes (Figures 3 and 4). Also note that not every disease has known OGs and TSGs at the same time.

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 22 for OG prediction. Note however that prediction are made for these 27 and 22 cancer types while sharing all the driver genes known for the 175 diseases (according to their similarities with these 27 and 22 cancer types).

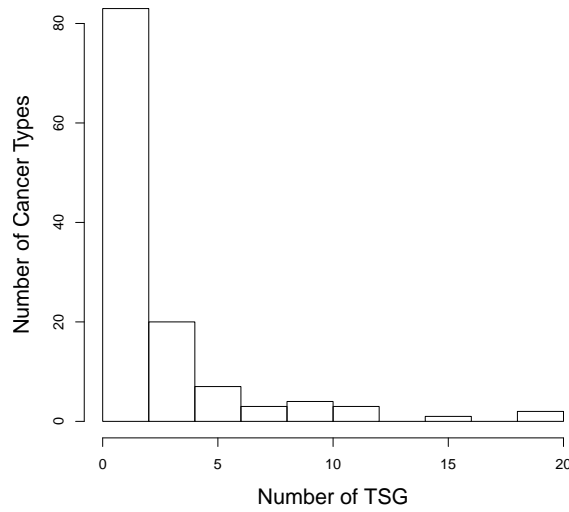


Fig 3. Histogram of the number of TSG per cancer type

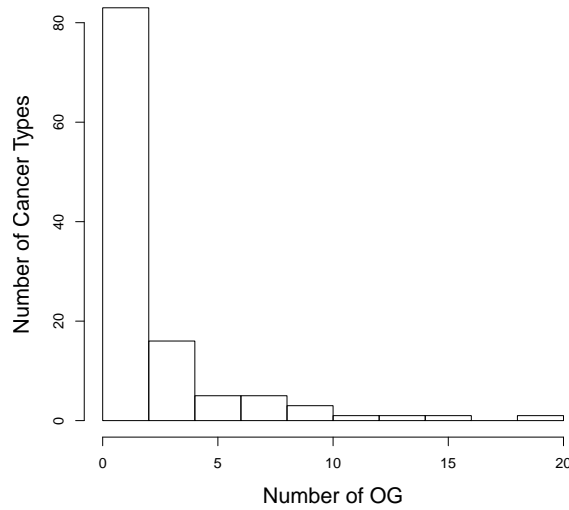


Fig 4. Histogram of the number of OG per cancer type

The 2-fold CV consistency error of LOTUS for each of those cancer types is 473
 presented in Tables 7 (for TSG) and 8 (for OG). Here we compare four variants of 474
 LOTUS, as explained in Methods: single-task LOTUS treats each disease in turn 475
 independently from the others; aggregation LOTUS applies a pan-cancer prediction by 476
 pooling together the known genes of all cancer types; and the two multitask versions of 477
 LOTUS use either a standard multitask strategy that do not take into account the 478

relative similarities between diseases (multitask TUSON), or a more refined multitask strategy where similar cancer types share more information than non-similar ones (multitask TUSON2).

Disease	Number of TSGs	Single-Task LOTUS	Aggregation LOTUS	Multi-Task LOTUS	Multi-Task LOTUS2
AML	15	1552	655	678	525
breast	20	1308	1149	1151	1131
colon carcinoma	7	943	71	67	51
colorectal	19	811	75	47	43
DLBCL	5	633	568	546	602
endometrial	9	77	77	54	33
gastric	4	2414	27	73	55
glioblastoma	4	87	87	89	93
glioma	8	1693	64	47	42
hepatocellular carcinoma	6	158	102	86	57
leukemia	11	1172	59	81	31
lymphoma	4	2069	88	62	42
MDS	4	5095	222	178	154
medulloblastoma	9	1427	333	333	320
melanoma	12	874	36	64	26
NSCLC	4	300	68	53	35
osteosarcoma	4	2539	67	99	61
ovary	11	171	48	49	40
pancreatic	8	174	85	39	54
paraganglioma	5	14699	1993	2446	2404
pheochromocytoma	6	12135	78	114	87
renal	5	2845	76	87	107
renal cell carcinoma	6	2932	48	33	26
skin basal cell	9	725	48	71	24
skin squamous cell	9	687	56	65	19
T-ALL	5	767	831	833	855
Wilms tumour	4	1154	224	231	227

Table 7. Consistency errors (CE) for prediction of disease specific TSGs in the multi-task setting.

In the above table, AML stands for acute myeloid leukemia, DLBCL for diffuse large B-cell lymphoma, MDS for myelodysplastic syndromes, NSCLC for non-small cell lung cancer and T-ALL for T-cell acute lymphoblastic cancer.

For most diseases, single-task LOTUS leads to the worst CE. 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 in a pan-cancer paradigm improves the performance of driver gene prediction for each cancer type. However, in many cases, the multi-task LOTUS and LOTUS2 algorithms lead to an additional improvement over

Disease	Number of OGs	Single-Task LOTUS	Aggregation LOTUS	Multi-Task LOTUS	Multi-Task LOTUS2
ALL	9	1637	873	856	796
AML	20	1447	606	600	578
bladder	5	636	83	32	54
breast	8	2250	121	134	91
CLL	8	2598	824	814	825
colorectal	12	2018	68	32	27
DLBCL	5	107	355	353	327
endometrial	6	616	40	28	26
gastric	9	112	40	25	15
glioblastoma	8	3452	74	60	54
glioma	6	613	761	773	769
head and neck	6	320	71	51	39
lymphoma	4	5651	79	61	77
MDS	9	5071	86	109	82
melanoma	14	1420	281	276	295
MM	4	3122	77	37	60
NSCLC	15	2281	280	126	149
ovary	8	3194	57	37	32
prostate	8	845	162	126	154
Spitzoid tumour	4	183	68	38	48
T-ALL	4	8436	2041	2047	2046
WM	4	203	162	160	78

Table 8. Consistency errors (CE) for prediction of disease specific OGs in the multi-task setting

In the above table, ALL stands for acute lymphocytic leukemia, AML for acute myeloid leukemia, CLL for chronic lymphocytic leukemia, DLBCL for diffuse large B-cell lymphoma, MDS for myelodysplastic syndromes, MM for multiple myeloma, NSCLC for non-small cell lung cancer, T-ALL for T-cell acute lymphoblastic cancer and WM for Waldenstrom macroglobulinemia.

Aggregation LOTUS, LOTUS2 leading in general to the best results. On average, the decrease in CE between Aggregate LOTUS and LOTUS2 is of 23% for OG and 17% for TSG (the smaller the CE , the better). 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 diseases. The above results show that multi-task 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 information to compare diseases.

Taken together, these results show that multi-task machine learning algorithms like LOTUS are interesting approaches to predict cancer specific driver genes. In addition, multi-task 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. Indeed, according to the results of pan-cancer studies in the single-task setting, they do not perform as well as single-task LOTUS. Moreover, they are not adapted, as such, to the multitask setting.

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 had 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 [68]. 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 multi-task method. It relies on a disease kernel that controls how driver gene information is shared between diseases. With LOTUS2 (where $K_{diseases} = (K_{descriptors} + I + D)/3$), 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 (single-task LOTUS or Aggregation LOTUS), and with respect to the more naive multi-task LOTUS algorithm where $K_{diseases} = (I + D)/2$. Again, the kernel approach leaves space for integration of other types and possibly more complex biological sources

of information about diseases. Our multi-task approach thus allows to make prediction 528
for cancer types with very few known driver genes, which would be less reliable with the 529
single-task methods. We considered here only diseases with at least 4 known driver 530
genes, in order to perform cross-validation studies, which was necessary to evaluate the 531
methods. However, it is important to note that in real-case studies, at the extreme, 532
both versions of multi-task LOTUS could make driver gene prediction for cancer types 533
for which no driver gene is known. 534

Among the 175 diseases derived from the COSMIC database (see Table 1), we kept 535
only 27 cancer types for TSG prediction and 22 for OG prediction, for which at least 536
four driver genes were available. However, inspection of the 175 disease names indicates 537
that there might be diseases that could be grouped (for example "colorectal" and 538
"colorectal adenocarcinoma", or "skin" with "skin basal cell" or "skin squamous cell"), 539
which would have allowed to enlarge the training sets and possibly improve the 540
predictions. Future directions could be to have experts analyze and potentially modify 541
this disease list, in order to optimize the training sets, or help to derive finer disease 542
descriptors. 543

LOTUS is a machine learning algorithm based on one-class SVM. In fact, the most 544
classical problem in machine learning is binary classification, where the task is to 545
classify observations into two classes (positives and negatives), based on training sets \mathcal{P} 546
of known positives and \mathcal{N} of known negatives. Driver gene detection can be seen as 547
binary classification of TSGs vs. neutral genes, and of OGs vs. neutral genes. However, 548
although the \mathcal{P} set is composed of known driver genes, it is not straightforward to build 549
the \mathcal{N} set because we cannot claim that some genes cannot be drivers. Thus, driver 550
gene detection should rather be seen as binary classification problem with only one 551
training set \mathcal{P} of known positives. This problem is classically called PU learning 552
(for Positive-Unknown), as opposed to PN learning (for Positive-Negative). 553

The classical way to solve PU learning problems is to choose a set \mathcal{N} of negatives 554
among the unlabeled data and apply a PN learning method. For example, one can 555
consider all unknown items as negatives (some of which may be reclassified afterwards 556
as positives), or randomly choose bootstrapped sets of negatives among the unknown, 557
like in [35]. Both methods assume that a minority of the unlabeled items are in fact 558
positives, which is expected for driver genes. 559

The one-class SVM algorithm [31] 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 add 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.

Supporting information

S1 Table List of cancer types (CGC v84). Cancer types derived from COSMIC annotations along with their numbers of associated OG and TSG. The resulting names are sometimes very general and sometimes very specific, and some redundancies may be present, because we chose to add as little interpretation as possible.

S2 Table Description of cancer types (CGC v84) . Descriptors of all cancer types according to their localizations and types that are used to compute the disease kernel used by LOTUS2. The general scheme is: "cancer type" ("localizations"/"types").

S3 Table TSG and OG rankings for LOTUS with the 20/20, the TUSON and the CGCv84 datasets. Note that the training sets were removed every time.

Acknowledgments

585

Olivier Collier's work has been conducted as part of the project Labex MME-DII
(ANR11-LBX-0023-01).

586

587

References

588

1. D. HANAHAN AND R. WEINBERG *The hallmarks of cancer*. Cell, 100(1), 57-70,
2000. 589
590
2. D. HANAHAN AND R. WEINBERG *The hallmarks of cancer: the next generation*.
Cell, 144, 646-674, 2011. 591
592
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. 593
594
595
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. 596
597
598
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. 599
600
601
6. G.M. COOPER *The cell: a molecular approach, 2nd edition*. Sunderland (MA):
Sinauer Associates, 2000. 602
603
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. 604
605
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. 606
607
608
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. 609
610
611

10. C.A. HUDIS *Trastuzumab—mechanism of action and use in clinical practice*. N Engl J Med, 357(1), 39-51, 2007. 612
613
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. 614
615
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. 616
617
618
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. 619
620
621
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. 622
623
624
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. 625
626
627
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. 628
629
630
17. *Comprehensive genomic characterization of squamous cell lung cancers*. Nature,
489.7417: 519-52, 2012. 631
632
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. 633
634
635
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. 636
637
638

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. 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.
25. 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.
26. 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.
27. C.J. TOKHEIM, N. PAPADOPOULOS, K.W. KINZLER, B. VOGELSTEIN AND R. KARCHIN *Evaluating the evaluation of cancer driver genes*. PNAS, 113(50):14330–14335, 2016.
28. 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.
29. B. SCHÖLKOPF ET AL. *Kernel methods in computational biology*. MIT Press, 2004.

30. T. EVGENIOU, C. MICCHELLI AND M. PONTIL *Learning multiple tasks with kernel methods*. JMLR, 6:615–637, 2005. 667
668
31. B. SCHÖLKOPF, R. WILLIAMSON, A. SMOLA, J. SHAWE-TAYLOR, J. PLATT *Support vector method for novelty detection*. NIPS 1999, 582-588, 1999. 669
670
32. 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. 671
672
33. R.I. KONDOR AND J. LAFFERTY *Diffusion kernels on graphs and other discrete input spaces*. ICML,3:315-322, 2002. 673
674
34. L. COWEN, T. IDEKER, B.J. RAPHAEL AND R. SHARAN *Network propagation: a universal amplifier of genetic associations*. Nature Rev Genet, 2017. 675
676
35. 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. 677
678
679
36. T.S.K. PRASSAD, R. GOEL, K. KANDASAMY, S. KEERTHIMUKAR, S. KUMAR, S. MATHIVANAN ET AL. *Human Protein Reference Database - 2009 update*. Nucleic Acids R, 37, D767-72, 2009. 680
681
682
37. A. KARATZOGLOU, A. SMOLA, K. HORNIK AND A. ZEILEIS *kernelab – An S4 Package for Kernel Methods in R*. JSS, 11-9, 1-20, 2004. 683
684
38. M.S. SONG, L. SALMENA AND P.P. PANDOLFI *The functions and regulation of the PTEN tumour suppressor*. Nature Rev, Molecular Cell Biology, 13(5), 283–96, 2012. 685
686
687
39. L.G. MORRIS, A.M. KAUFMAN, Y. GONG, D. RAMASWAMI, L.A. WALSH, Ş. TURCAN ET AL. *Recurrent somatic mutation of FAT1 in multiple human cancers leads to aberrant Wnt activation*. Nature Genet, 45(3), 253–61, 2013. 688
689
690
40. L. BENEDETTI, M. CEREDA, L. MONTEVERDE, N. DESAI AND F.D. CICCARELLI *Synthetic lethal interaction between the tumour suppressor STAG2 and its paralog STAG1*. Oncotarget, 8(23), 37619–32, 2017. 691
692
693

41. D. MATASSA, I. AGLIARULO, R. AVOLIO, M. LANDRISCINA AND F. ESPOSITO 694
TRAP1 Regulation of Cancer Metabolism: Dual Role as Oncogene or Tumor 695
Suppressor. Genes, 9(4), 195, 2018. 696
42. Y.K. CHAE, J.F. ANKER, B.A. CARNEIRO, S. CHANDRA, J. KAPLAN, A. 697
KALYAN ET AL. *Genomic landscape of DNA repair genes in cancer*. Oncotarget, 698
7(17), 23312–21, 2016. 699
43. A. TORGOVNICK AND B. SCHUMACHER *DNA repair mechanisms in cancer* 700
development and therapy. Front Genet, 6–157, 2015. 701
44. J. GENSCHER, L.R. BAZEMORE AND P.J. MODRICH *Human exonuclease I is* 702
required for 5' and 3' mismatch repair. J Biol Chem, 277:13302–11, 2002. 703
45. M. MANANDHAR, K.S. BOULWARE AND R.D. WOOD *The ERCC1 and ERCC4* 704
(XPF) genes and gene products. Gene, 569(2):153–161, 2015. 705
46. M. OKUDA, NAKAZAWA, C. GUO, T. OGI AND Y. NISHIMURA *Common TFIID* 706
recruitment mechanism in global genome and transcription-coupled repair 707
subpathways. Nucleic Acids Res, 45(22):13043–55, 2017. 708
47. R. ASPINWALL, D.G. ROTHWELL, T. ROLDAN-ARJONA, C. ANSELMINO, C.J. 709
WARD, J.P. CHEADLE ET AL. *Cloning and characterization of a functional* 710
human homolog of Escherichia coli endonuclease III. PNAS, 94:109–114, 1997. 711
48. A. KUMAGAI, J. LEE, H.Y. YOO AND W.G. DUNPHY *TopBP1 activates the* 712
ATR-ATRIP complex. Cell, 124(5):943–55, 2006. 713
49. M.S. PARK, D.L. LUDWIG, E. STIGGER AND S.H. LEE *Physical interaction* 714
between human RAD52 and RPA is required for homologous recombination in 715
mammalian cells. J Biol Chem, 271:18996–19000, 1996. 716
50. A.C. MASSON, R. ROY, D.T. SIMMONS AND M.S. WOLD *Functions of* 717
alternative replication protein A in initiation and elongation. Biochem, 718
49:5919–28, 2010. 719
51. R.P. KUIPER AND N. HOOGERBRUGGE *NTHL1 defines novel cancer syndrome*. 720
Oncotarget, 6(33):34069–70, 2015. 721

52. I. TOMLINSON *The Mendelian colorectal cancer syndromes*. Ann Clin Biochem, 722
52(6):690–692, 2015. 723
53. 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*. 724
Oncogene, 2018. 725
726
54. K. XIE, K. ZHANG, J. KONG, C. WANG, Y. GU, C. LIANG ET AL. 727
*Cancer-testis gene PIWIL1 promotes cell proliferation, migration, and invasion in
lung adenocarcinoma*. Cancer Med, 7(1):157–166, 2018. 728
729
55. P.J. STEPHENS, P.S. TARPEY, H. DAVIES, P. VAN LOO, C. GREENMAN, D.C. 730
WEDGE ET AL. *The landscape of cancer genes and mutational processes in breast
cancer*. Nature, 486(7403):400–4, 2012. 731
732
56. S. ZHAI, C. LIU, L. ZHANG, J. ZHU, J. GUO, J. ZHANG ET AL. *PLCE1
Promotes Esophageal Cancer Cell Progression by Maintaining the Transcriptional
Activity of Snail*. Neoplasia, 19(3):154–164, 2017. 733
734
735
57. S. ALLERSTORFER, G. SONVILLA, H. FISCHER, S. SPIEGL-KREINECKER, C. 736
GAUGLHOFER, U. SETINEK ET AL. *FGF5 as an oncogenic factor in human
glioblastoma multiforme: autocrine and paracrine activities*. Oncogene, 737
27(30):4180–90, 2008. 738
739
58. H. COHEN, R. BEN-HAMO, M. GIDONI, I. YITZHAKI, R. KOZOL, A. 740
ZILBERGER ET AL. *Shift in GATA3 functions, and GATA3 mutations, control
progression and clinical presentation in breast cancer*. Breast Cancer Res, 741
16(6):464, 2014. 742
743
59. 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, 744
2014. 745
746
60. L.X. YAN, Y.H. LIU, J.W. XIANG, Q.N. WU, L.B. XU, X.L. LUO ET AL. 747
*PIK3R1 targeting by miR-21 suppresses tumor cell migration and invasion by
reducing PI3K/AKT signaling and reversing EMT, and predicts clinical outcome
of breast cancer*. Int J Oncol, 2016. 748
749
750

61. A.C. LESKO, K.H. GOSS, F.F. YANG, A. SCHWERTNER, I. HULUR, K. ONEL 751
ET AL. *The APC tumor suppressor is required for epithelial cell polarization and* 752
three-dimensional morphogenesis. Biochim Biophys Acta, 1853(3):711–23, 2015. 753
62. E. KOTLER, O. SHANI, G. GOLDFELD, M. LOTAN-POMPAN, O. TARCIC, A. 754
GERSHONI ET AL. *A Systematic p53 Mutation Library Links Differential* 755
Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. 756
Mol Cell, 71(1):178–190, 2018. 757
63. M.E. HOUWING, E.A. KOOPMAN-COENEN, R. KERSSEBOO, S. GOOSKENS, I.M. 758
APPEL, S.T. ARENTSEN-PETERS ET AL. *Somatic thrombopoietin (THPO) gene* 759
mutations in childhood myeloid leukemias. Int J Hematol, 102(1):140–3, 2015. 760
64. T. EKBLAD, A.E. LINDGREN, C.D. ANDERSSON, R. CARABALLO, A.G. 761
THORSELL, T. KARLBERG ET AL.(2015) *Towards small molecule inhibitors of* 762
mono-ADP-ribosyltransferases. Eur J Med Chem, 95:546–51, 2015. 763
65. S. TEN HOORN, A. TRINH, J. DE JONG, L. KOENS AND L. VERMEULEN 764
Classification of Colorectal Cancer in Molecular Subtypes by 765
Immunohistochemistry. Methods Mol Biol, 1765:179–191, 2018. 766
66. Y. KITAI, M. IWAKAMI, K. SAITOH, S. TOGI, S. ISAYAMA, Y. SEKINE ET AL. 767
STAP-2 protein promotes prostate cancer growth by enhancing epidermal growth 768
factor receptor stabilization. J Mol Biol, 292(47):19392–99, 2017. 769
67. K.L. NG, C. MORAIS, A. BERNARD, N. SAUNDERS, H. SAMARATUNGA, G. 770
GOBE ET AL. *A systematic review and meta-analysis of immunohistochemical* 771
biomarkers that differentiate chromophobe renal cell carcinoma from renal 772
oncocytoma. J Clin Pathol, 69(8):661–71, 2016. 773
68. T. SAKOPARNIG, P. FRIED ET N. BEERENWINKEL *Identification of constrained* 774
cancer driver genes based on mutation timing. PLoS Comput Biol, 775
11(1):e1004027, 2015. 776