Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features

Edwin Cuppen (ecuppen@umcutrecht.nl)
University Medical Center Utrecht https://orcid.org/0000-0002-0400-9542
Luan Nguyen
University Medical Center Utrecht https://orcid.org/0000-0003-2788-6625
Arne van Hoeck
University Medical Center Utrecht https://orcid.org/0000-0002-6570-1452

Article

Keywords: tissue of origin, cancers of unknown primary, whole genome sequencing, CUPLR

Posted Date: October 21st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-956886/v1

Version of Record: A version of this preprint was published at Nature Communications on July 11th, 2022. See the published version at https://doi.org/10.1038/s41467-022-31666-w.

Abstract

Tumor tissue of origin (TOO) is an important factor for guiding treatment decisions. However, TOO cannot be determined for ~3% of metastatic cancer patients and are categorized as cancers of unknown primary (CUP). As whole genome sequencing (WGS) of tumors is now transitioning from the research domain to diagnostic practice in order to address the increasing demand for biomarker detection, its use for detection of TOO in routine diagnostics also starts becoming within reach. While proof of concept for the use of genome-wide features has been demonstrated before, more complex WGS mutation features, including structural variant (SV) driver and passenger events, have never been integrated into TOO-classifiers even though they bear highly characteristic links with tumor TOO. Using a uniformly processed dataset containing 6820 whole-genome sequenced primary and metastatic tumors, we have developed Cancer of Unknown Primary Location Resolver (CUPLR), a random forest based TOO classifier that employs 502 features based on simple and complex somatic driver and passenger mutations. Our model is able to distinguish 33 cancer (sub)types with an overall accuracy of 91% and 89% based on cross-validation (n=6139) and hold out set (n=681) predictions respectively. We found that SV derived features increase the accuracy and utility of TOO classification for specific cancer types. To ensure that predictions are human-interpretable and suited for use in routine diagnostics, CUPLR reports the top contributing features and their values compared to cohort averages. The comprehensive output of CUPLR is complementary to existing histopathological procedures and may thus improve diagnostics for patients with CUP.

Introduction

Cancers of unknown primary (CUPs) is an umbrella term for advanced stage metastatic tumors for which the tumor tissue of origin (TOO) cannot be conclusively determined based on routine diagnostics (typically via histopathology [1]), and there is also a significant fraction of patients with indeterminate or differential diagnoses, especially with poorly differentiated tumors [2]. Patients with uncertain TOO diagnosis suffer from a lack of therapeutic options as primary cancer type classification is a dominant factor in guiding treatment decisions [3].

Thus far, the most performant TOO classifiers have been developed on data from whole genome sequencing (WGS) [4, 5], whole transcriptome sequencing (WTS) [6, 7] or methylation profiling [8]. While gene expression and methylation profiling for cancer diagnostics is not yet widespread [9, 10], WGS is being increasingly adopted, driven by its utility for comprehensive identification of actionable mutations [11]. WGS-based TOO classifiers are thus at this moment more viable for application in the clinic. Recently developed WGS-based classifiers [4, 5] were shown to outperform targeted or whole exome sequencing based approaches [12, 13] due to being able to utilize mutations from all genomic regions. The main features employed by these classifiers included mutational signatures which are patterns of somatic mutations resulting from exogenous or endogenous mutational processes (e.g. C>T mutations due to ultraviolet radiation exposure in melanoma) [14], as well as regional mutational density (RMD) which represents the genomic distribution of somatic mutations that are associated with tissue type

specific chromatin states, whereby late-replicating closed chromatin regions show increased mutation rates [15]. However, not all WGS based features are yet fully explored for TOO classification including complex mutagenic features such as viral DNA integrations, driver gene fusions, and other complex structural events (e.g. chromothripsis), as well as non-mutagenic features such as gender, all of which have been shown to be correlated with specific tumor type(s). Indeed, human papillomavirus (HPV) sequence insertions are specifically and frequently found in cervical and head and neck cancer [14], *KIAA1549-BRAF* fusions in pilocytic astrocytomas [13], and liposarcomas frequently harbor *FUS-DDIT3* fusions [15] and chromothripsis events [16].

Here we describe the development of CUPLR (Cancer of Unknown Primary Location Resolver), a TOO classifier that integrates current state-of-the-art WGS based mutation features as well as thus far unexplored complex structural variant (SV) features. CUPLR comprises an ensemble of binary random forest classifiers that each discriminate one of 33 cancer types with an overall accuracy of 0.91. We find that while RMD and mutational signatures were highly predictive of cancer type (in line with existing classifiers [4, 5]), the incorporation of SV features improves prediction performance for cancer types that currently lack highly informative features. Furthermore, we have ensured that the output of CUPLR, namely the prediction probabilities and the features supporting each prediction, are human interpretable to facilitate diagnostic use and clinical decision making with CUPLR.

Results

Extraction of genomic features

To develop CUPLR, we constructed a harmonized dataset from two large pan-cancer WGS datasets from the Hartwig Medical Foundation (Hartwig) and Pan-Cancer Analysis of Whole Genomes consortium (PCAWG). The raw sequencing reads were analyzed with the same mutation calling pipeline to construct a catalogue of uniformly called simple and complex mutations. The harmonized dataset consisted of tumors from 6820 patients across 33 different cancer types (Figure 1a). In contrast to many previously published papers [4, 5, 12, 13], this dataset includes a large proportion of samples taken from metastatic lesions, which is relevant for TOO classification as CUP samples are by definition from patients with metastatic cancer.

A wide range of features (n=4122) were extracted for classifying cancer type based on driver/passenger and simple/complex mutations (Figure 1c). First, we determined the presence of gain of function (amplifications and activating mutations) and loss of function (deep deletions and biallelic loss) events in 205 cancer associated genes. These genes were selected based on having enrichment of gain and/or loss of function events in at least one cancer type (see methods). Second, we calculated the mutational load of single base substitutions (SBS), double base substitutions (DBS), and indels for each sample. Third, we determined the number of contributing mutations to the SBS, DBS and indel signatures from the COSMIC catalog [14]. Fourth, the number of SBSs in each 1Mb bin across the genome (n=3071) was calculated to determine the RMD [17]. Mutational signatures and RMD were normalized by the mutational

load of the respective mutation type to account for differences in mutational load across samples. Fifth, for each sample we determined the copy number change of each chromosome arm relative to the genome ploidy [18]. Sixth, gender was determined based on sex chromosome ploidy [19]. Lastly, we parsed the called simple and complex structural variants to determine: (i) the SV load per sample; (ii) the number of large deletions and duplications stratified by size, as well as the number of long interspersed nuclear element (LINE) insertions, normalized by the total number of SVs; (iii) the presence of complex SV events; and (iv) the presence of gene fusions and viral sequence insertions [19, 20].

Classifier training

The extracted genomic features were then used to develop CUPLR, a classifier consisting of two components (Figure 1d). Firstly, an ensemble of binary random forest classifiers each discriminates one cancer type versus other cancer types (i.e. one-vs-rest). We chose to use an ensemble of binary classifiers as opposed to one multiclass classifier so that feature selection could be performed per cancer type, since different features are important for each cancer type. Additionally, we chose to use random forests over other algorithms (e.g. neural networks) as they can natively handle different feature types (continuous, Boolean, categorical, etc.) without requiring feature values to be scaled, which also improves model interpretability. The second component of CUPLR is an ensemble of isotonic regressions to calibrate the probabilities produced by each random forest. Random forests tend to be overconfident at probabilities towards 0 and underconfident at probabilities towards 1, and this bias varies between random forests [21]. The calibration we have performed here ensures that probabilities are comparable between random forests. Furthermore, calibration allows for the probabilities to have the following intuitive interpretation: a probability of e.g. 0.8 means that there is an 80% chance of a prediction being correct (this relationship does not hold for raw random forest probabilities).

We used 6139 samples for training and held out 681 samples as an independent test set, with both having the same cancer type proportions. Training of the main random forest ensemble involved several steps. Briefly, due to the sheer number of RMD bins (3071) as well as their sparsity, non-negative matrix factorization (NMF) was for each cancer type performed on the RMD bins to reduce these to 44 cancer type-specific RMD profiles. Then, for each cancer type, univariate feature selection was performed (to remove irrelevant features) with 502 features ultimately being selected (219 numeric and 283 boolean; **Supplementary data 3**). This was followed by class resampling (to alleviate imbalances in the number of samples for each cancer type), and subsequently training of the binary random forest itself (**Supplementary figure 1**). The above training procedure was applied to all samples of the training set to produce the final random forest ensemble. The random forest ensemble training procedure was then subjected to stratified 15-fold cross-validation to obtain cancer type probabilities for the training samples. These probabilities were then used to train the ensemble of isotonic regressions for calibrating the random forest probabilities (Figure 1b).

Performance of CUPLR

To assess the performance of CUPLR, we used the cancer type predictions based on the isotonic regression calibrated cross-validation probabilities, as well as the predictions upon applying CUPLR to the held out test set (Figure 1b). Both training (n=6139) and test sets (n=681) had the same cancer type proportions. Based on cross-validation predictions (Figure 2), CUPLR could classify all samples with an accuracy (i.e. fraction correctly classified) of 91%. Similar performance was observed for the test set across cancer types, with an accuracy of 89% across all samples (**Supplementary figure 5**). Certain cancer types show differences in test and CV accuracy which was due to low sample sizes (**Supplementary figure 6**).

High misclassification rates for certain cancer types could likely be explained by shared cancer type characteristics (Figure 2a). This could be due to a common developmental origin, such as with Uterus being misclassified as Ovary (12%) due to both being gynecological cancers [22], and Biliary being misclassified as Pancreas (31%) and Liver (8%) due to being cancers of the foregut [23, 24]. Cancer subtypes were also often misclassified as other subtypes, which was the case for Lung_SC (small cell lung cancer) towards Lung_NSC (non-small cell lung cancer) (37%); Kidney_Papillary towards Kidney_ClearCell (48%); and Sarcoma_Leiomyo (46%) and Sarcoma_Lipo (10%) towards Sarcoma_Other (sarcomas other than leiomyosarcoma, liposarcoma or gastrointestinal stromal tumors). Lastly, Gastrointestinal_NET (gastrointestinal neuroendocrine tumors) was occasionally misclassified as Pancreas_NET (pancreatic neuroendocrine tumors) (12%) which may, at least partially, reflect cancer type mis-annotation amongst these samples due to neuroendocrine tumors having similar morphological features and shared anatomical location [25].

Thus far, we have assessed performance based on whether the highest probability cancer type is the correct cancer type (i.e. top-1 accuracy; Figure 2a,b). However, if we consider whether the correct cancer type is amongst the top-2 highest probabilities (i.e. top-2 accuracy; Figure 2c), overall accuracy increases from 91–95%, with the greatest increases being for the cancer subtypes including Lung_SC (57–88%), Kidney_Papillary (48–83%), and Sarcoma_Leiomyo (46–89%). A large gain in accuracy was also observed for Biliary (41–80%) which was often misclassified as Pancreas. Similar increases in accuracy were seen based on predictions on the held out test set (**Supplementary figure 7**). The top-2 (and even top-3) probabilities of CUPLR are particularly useful for differential diagnosis purposes to narrow down potential TOOs when routine diagnostics are not fully conclusive.

Added predictive value of SV related features

When examining the most important feature types from each random forest within CUPLR (Figure 3a), RMD profiles ('rmd') were consistently the most predictive of cancer type (in line with the findings from Jiao *et al.* 2020 [4]), as well mutational signatures ('sigs') including those with known cancer type associations such as SBS4 (associated with smoking [14]) in lung cancer (Figure 3b). As these mutation features are derived from genome-wide SBSs and indels, we assessed whether the presence of certain confounding factors that affect the SBS and indel genomic landscape (including DNA repair deficiencies [26, 27], chemotherapy treatment [28, 29], smoking history [30]) may lead to more incorrect predictions.

However, these confounding factors showed minimal impact on classification performance (**Supplementary notes**, **Supplementary table 1, Supplementary figure 12**).

In addition to RMD profiles and mutational signatures, gender (as expected) was highly important for predicting cancers of the reproductive organs including breast, cervical, ovarian, prostate, and uterine cancer (Figure 3b, **Supplementary figure 8**). Notably, SV related features were important for classifying certain cancer types (Figure 3b). This included known cancer type specific gene fusions such as *TMPRSS2-ERG* in Prostate [31], *KIAA1549-BRAF* in CNS_PiloAstro (pilocytic astrocytomas) [32], *CTNNB1-PLAG1* and *MYBL1-NFIB* in HeadAndNeck_SG (salivary gland cancer) [33, 34], and *FUS-DDIT3* in Sarcoma_Lipo (liposarcoma) [35]. Of note, while *EML4-ALK* fusions are known to be specific to non-small cell lung cancer [36], this fusion does not appear amongst the top 15 most important features in Figure 3b since it has low feature importance (**Supplementary data 4**) due to only occurring in a minor proportion of non-small cell lung cancer samples (**Supplementary figure 11**). We also find known viral DNA integrations as important features, including human papillomavirus (viral_ins.HPV) in Cervix [37] and Merkel cell polyomavirus (viral_ins.MCPyV) in Skin_Other (non-melanoma skin cancer) [38]. Lastly, the number of SVs in the largest complex SV cluster (sv.complex.largest_cluster) which we use as a proxy for the presence of chromothripsis was also predictive for liposarcomas, which are known to frequently harbor chromothripsis events. However, in contrast to the features mentioned above, the presence of chromothripsis alone is not sufficient for classifying a tumor as liposarcoma as chromothripsis is also highly prevalent in other tumor types [16].

To further assess the added value of SV related features, we excluded entire feature types from the training and examined the decrease in classifier performance (**Supplementary figure 9**). Indeed, removing gene fusion features ('fusion') resulted in a large drop in accuracy for HeadAndNeck_SG (salivary gland cancer; 43–23%) and Sarcoma_Lipo (liposarcoma; 80–63%), with removal of the viral integration features ('viral_ins') leading to decreased accuracy for Cervix (88–65%). Exclusion of simple and complex SV features ('sv') led to lower accuracy for Sarcoma_Lipo (80–73%) as well as Sarcoma_Leiomyo (leiomyosarcoma; 50–45%), likely due to the presence of chromothripsis (sv.complex.largest_cluster; Figure 3b) being important for the separation of these two cancer types, which are often misclassified as each other (Figure 2a). Similarly, removal of the simple and complex SV features ('sv') resulted in a large drop in accuracy for Lung_SC (small cell lung cancer; 53–41%) (**Supplementary figure 9**), likely because 1-10kb deletions and 10-100kb duplications distinguish Lung_NSC from Lung_SC (Figure 3, **Supplementary figure 8**).

When compared to existing WGS-based CUP classifiers (**Supplementary figure 10**), CUPLR also showed improved performance for cancer types where SV related features were important, including for CNS_PiloAstro, Lung_NSC, and Prostate. While Sarcoma_Lipo had no comparable cancer type in existing classifiers, prediction of all sarcoma subtypes by CUPLR (except for Sarcoma_Leiomyo) outperformed osteosarcoma prediction by PCAWG consortium classifier [4] (PCAWG-NN; Bone-Osteosarc) and sarcoma prediction from the Salvadores *et al.* [5] classifier (Salvadores-SVM; SARC). Overall, CUPLR performs similarly for the remaining cancer types (except for head and neck, myeloid and pancreatic

neuroendocrine and thyroid cancers) when compared to existing classifiers, with CUPLR also being able to classify more cancer (sub)types.

In summary, the incorporation of all feature types resulted in the best performance, with SV related features being highly important for specific cancer types. To our knowledge, we are the first to show the added value of SV related features for cancer type classification.

Graphical prediction report

Aside from cancer type probabilities, CUPLR also outputs explanations as to which features support these probabilities. These allow one to verify the predictions based on existing knowledge, and could be included in diagnostic reports to support decision making, e.g. in molecular tumor boards. The feature explanations are based on feature contribution calculations which enable feature importance determination at the sample level (rather than at the cohort level as in Figure 3). Specifically, feature contributions represent the total gain (or loss) in probability upon passing a feature through a random forest [39]. To ease the interpretation of CUPLR's output, we have implemented a graphical report (Figure 4) which can be generated per patient that shows the cancer type probabilities (left panel), feature contributions for the top features for the top cancer types (middle panel). Also shown are the corresponding feature values in the patient in relation to the average feature value amongst patients in the training set (right panel).

We will use patient DO35949 as an example demonstration of the graphical report (Figure 4a). Since the CNS_PiloAstro probability (1.00) was much higher than the probability of other cancer types, only one cancer type (i.e. panel row) is shown, with up to 3 panel rows being shown when more than one cancer type probability is high (such as for patient HMF004199A; Figure 4b). The top feature for DO35949 was the presence of a KIAA1549-BRAF fusion (middle panel) which is on average found in 74% of CNS_PiloAstro patients (blue label), but 0% in all other patients (grey label). Since this feature is of Boolean type, a feature value of 1 (red label) indicates the presence of the *KIAA1549-BRAF* fusion in DO35949 (whereas a feature value of 0 would indicate absence). On the other hand, the mutational load of SNVs, DBSs and indels (mut_load.snv, mut_load.dbs, mut_load.indel) is on average lower in DO35949 (red label) and CNS_PiloAstro patients (blue label) compared to all other patients (grey label). Conversely, the proportion of 1-10Mb duplications (sv.DUP_[1e+06,1e+07]), is on average higher in DO35949 (red label) and CNS_PiloAstro patients (blue label) compared to all other patients (grey label). For all features, a red arrow (going right or left)highlights whether the feature value in DO35949 is higher or lower (respectively) than that in all other patients. Additionally, this relationship is also indicated in text in the feature contribution panel (middle panel) for non-Boolean features.

Patient HMF004199A (Figure 4b) is an example of a situation where more than one cancer type probability is high, with the probability of Lung_NSC (non-small cell lung cancer) being 0.84 and Lung_SC (small cell lung cancer) being 0.75 (Figure 4b). Here, two feature panels are shown for both of these cancer types to aid with resolving this uncertainty. Since *RB1* loss is specific to small cell lung cancer and

rarely occurs in non-small cell lung cancer [40, 41], it is more likely that the correct predicted cancer type is Lung_SC and not Lung_NSC. Indeed, this patient was diagnosed with small cell lung cancer.

This graphical report presents the detailed machine learning-based classification output of CUPLR in a human readable format. We acknowledge that the output is highly detailed, which is inevitable due to the large amounts of data used by the algorithm. However, as these details may not be necessary in all circumstances, we have implemented the option in the software to hide the feature contribution and/or feature value panels in the final graphical output.

Feature contributions aid in clarifying the primary tumor location of CUPs

To showcase how CUPLR could be used in a real life clinical setting, we applied CUPLR to 149 tumors for which primary tumor location was unknown from the Hartwig dataset, and examined the top predicted cancer type together with the top contributing features for each sample (**Supplementary data 2**). Of the 75 samples with a probability ≥0.8 for the top cancer type, the cancer type for 29 of these samples could be confidently determined purely based on these samples exhibiting features with well-known cancer type associations. This included 22 samples with signatures of smoking (SBS4 and ID13 [14]) and predicted as Lung_NSC, 1 sample with signatures of ultraviolet radiation (SBS7, DBS1 [14]) exposure and predicted as Skin_Melanoma, 5 samples with *APC* mutations and predicted as Colorectum [42], and 1 sample with a *TMPRSS2-ERG* fusion [31] and predicted as Prostate.

An additional 15 of the 75 samples could be confidently annotated based on having known cancer type specific features in conjunction with having a high contribution of the RMD profile of the predicted cancer type. For example, *KRAS* mutations most commonly occur in pancreatic cancer but are also frequent in other cancer types such as colorectal or lung cancer (**Supplementary figure 11**) [43]. For 3 samples with a *KRAS* mutation and predicted as Pancreas, the top contributing feature was rmd.Pancreas.1 (RMD profile of pancreatic cancer) thereby giving us confidence that the Pancreas classification was correct. Likewise, in conjunction with the respective RMD profile, 10 samples could be clarified as Gastric based on the presence of the signature of ROS damage (SBS17) [14], LINE insertions [44], and/or fragile site deletions [45]; 1 sample as HeadAndNeck_Other based on having HPV integration [46]; and 1 sample as Uterus based on having a *PIK3R1* mutation [47] (**Supplementary data 2**). For the remaining 31 of the 75 samples, no known cancer type specific features could be identified, although the RMD profiles were the primary contributors of the top cancer type probability for these samples. Most of these samples were predicted as cancer types with highly specific RMD profiles respective to each cancer type, including Breast, Colorectum, Liver, Kidney_ClearCell, Ovary, and Urothelial (**Supplementary data 2**, **Supplementary figure 11**), and classification of cancer types achieved ≥0.9 accuracy (Figure 2a). The majority of these samples are thus likely correctly annotated, although additional evidence would be required for validation and final diagnosis. This could for example be based on histopathological examinations, but these were unfortunately not available for the retrospectively analyzed samples that were included here. However, such information would be readily available in routine diagnostics.

Discussion

Here, we have developed a tissue of origin classifier (CUPLR) for the analysis and diagnostics of CUPs using a large uniformly processed WGS dataset of both primary and metastatic tumors. Our classifier is to our knowledge the first to incorporate genomic features derived from SVs, as well as provide human interpretable explanations alongside each prediction which allows for manual resolving of CUPs, especially in cases of lower scoring samples or for samples for which multiple tumor types are suggested by CUPLR.

Current state-of-the-art WGS-based classifiers, including those by Jiao *et al.* [4] and Salvadores *et al.* [5], achieve high accuracy (≥80%) for over three quarters of the cancer types they predict by primarily employing RMD and mutational signatures as features, which are derived from simple mutations. CUPLR builds upon these classifiers, with the inclusion of SV and driver gene related features improving performance for certain cancer types, such as for pilocytic astrocytoma and prostate cancer. Of note, genomic-based TOO classifiers published thus far have only used data from primary tumors [4, 5, 12, 13]. However, since CUPs are by definition metastatic tumors, the inclusion of 4401 metastatic tumors with known tissue of origin for training CUPLR may make it a more suitable tool for the purpose of clarifying CUPs. We do acknowledge however that the metastatic tumor data used here may harbor treatment effects which are absent in treatment-naive CUPs, such as driver mutations in *AR* [48] or *ESR1* [49] conferring resistance to therapy or presence of characteristic mutations induced by for example 5-fluorouracil [28] or radiotherapy [29, 50]. Identification and removal of such treatment associated features could potentially improve TOO classification. Additionally, CUPLR is able to distinguish the largest number of cancer (sub)types (33 cancer types) compared to existing WGS-based classifiers, with the Jiao [4] and Salvadores [5] classifiers discriminating 24 and 18 cancer types respectively, and also more than a recently published whole histology slide image based classifier which discriminates 18 cancer types [51]. We do acknowledge that discriminating even more cancer types is warranted, but this is currently limited by the amount of available training samples that were sequenced and uniformly bioinformatically analyzed. For example, thymic cancer or lung neuroendocrine tumors had too few (<20) samples to be included as separate classes for training (**Supplementary data 1**). Furthermore, certain cancer types could be divided into their subtypes, such as Myeloid (currently only with 31 samples; Figure 2a) into acute myeloid leukemia and multiple myeloma [52], and sarcomas in a broader range of subtypes [53]. The availability of more WGS data for less frequent, but also (ultra-)rare cancers would allow for training of an updated CUPLR model that classifies additional cancer types and subtypes.

While CUPLR achieved overall excellent classification accuracy (91%), which is similar to or better than other WGS-based classifiers (**Supplementary figure 10**), it should be noted that this is driven by several large sub-cohorts of common cancer types (e.g. breast and colorectal cancer) and that performance for certain cancer types is still suboptimal. For cancer subtypes that CUPLR has difficulty discriminating, such as small cell versus non-small cell lung cancer and papillary versus clear cell renal carcinoma, additional information from histopathological staining and immunohistochemistry could be used to clarify these cases. Here, the application of artificial intelligence-based histology image analysis [51]

could further improve the accuracy and reliability of resolving CUPs. Clinical metadata, such as biopsy location and metastasis grade [54], when used together with CUPLR can also aid with clarifying primary tumor location. For instance, there may be uncertainty as to whether a tumor with human papillomavirus DNA integration was a head and neck cancer or cervical cancer based on CUPLR predictions (e.g. DO15430, **Supplementary data 2**), since human papillomavirus DNA integration is characteristic of these two cancer types [37]. However, in clinical practice, it will always be known if the tumor biopsy was taken from the upper body and if the lesions were local, to be able to conclude that this tumor can only be of head and neck origin.

Given that RMD [4, 5], mutational signatures [55], and SVs [20, 56] are still active areas of research, we expect improvements to these features could also lead to better TOO classification. Here, we have demonstrated that extraction of cancer type specific RMD profiles is possible from raw mutation counts, similarly as was done for mutational signatures [14]. However, CUPLR does not heavily rely on RMD profiles for classification of certain cancer types such as liposarcoma and non-melanoma skin cancer (Figure 3b), potentially because more training samples are required to extract more stable and informative RMD profiles which could improve classification for these cancer types. Improvements to TOO classification may also come from extraction of more comprehensive mutational signatures, for example by incorporating information of mutation timing or genome localization [57, 58]. Development of more sophisticated signature extraction methods may also allow for quantification of low signal tissue type specific signatures, such as SBS88 (associated with colibactin induced DNA damage in the colon) which has only been extracted from colon healthy tissue [59, 60] but not cancer tissue likely because other mutational processes in cancer tissue mask the presence of this signature (**Supplementary figure 8, Supplementary figure 11**). Lastly, while CUPLR uses a wide range of features derived from SVs (including gene fusions, viral DNA integrations, LINE insertions, structural deletion and duplication size, and chromothripsis), there is still an opportunity to explore other SV related features to improve TOO classification, such as SV signatures [56].

Aside from WGS data, gene expression [6, 7, 61] and methylation [8] data have also been used to develop TOO classifiers, and inclusion of such data could further improve cancer type classification, although there may be strong redundancy with RMD information derived from WGS. As application of gene expression and methylation profiling for cancer diagnostics is still in its infancy [9, 10] and the inclusion of additional procedures would likely increase costs and time, its potential added value should thus be well balanced against these extra burdens. Furthermore, reliable gene expression profiling depends on the integrity and quality of RNA [62], which remains a challenge for routine clinical samples. Thus, to facilitate clinical implementation of CUPLR in diagnostic centers that have started implementing WGS in routine cancer diagnostics, the model was trained solely on WGS data despite RNA-seq data being available for subsets of samples of both the Hartwig [63] and PCAWG datasets [43].

Given the gradually increasing adoption of WGS in routine diagnostics [11], CUPLR serves as a viable complementary tool to standard procedures for determining TOO (e.g. histopathological staining and immunohistochemistry). CUPLR can be run from the output of open source tools and is freely available

as an R package at https://github.com/UMCUGenetics/CUPLR. The trained model as well as the code for generating the input features is provided to enable prediction on new samples and for facilitating diagnostic use.

Methods

Datasets

We have used whole genome sequencing data from cancer patients from two cohorts: the Hartwig Medical Foundation (Hartwig) cohort and the Pan-Cancer Analysis of Whole Genomes (PCAWG) cohort.

The Hartwig cohort contained patient data for which re-use for cancer research was consented by the patients as part of clinical studies (NCT01855477, NCT02925234) that were approved by the medical ethical committees (METC) of the University Medical Center Utrecht and the Netherlands Cancer Institute, and was provided under data transfer agreement (DR-104) by Hartwig Medical Foundation. The Hartwig cohort included 4776 metastatic tumor samples from 4635 patients. For patients with multiple biopsies that were taken at different timepoints, patient IDs were suffixed by ‘A’ for the first biopsy, ‘B’ for the second biopsy, etc (e.g. HMF001423A, HMF001423B).

The PCAWG cohort consisted of 2650 patient tumors, and access for raw sequencing data for these tumors was granted via the Data Access Compliance Office (DACO) Application Number DACO-1050905 on 6 October 2017 and via https://console.cancercollaboratory.org download portal on 4 December 2017.

Sample selection for developing CUPLR were based on several criteria. First, for patients with multiple biopsies, the biopsy with the highest tumor purity was selected. Second, samples with <0.2 tumor purity were removed as somatic variant calling was not reliable for these samples. Third, PCAWG samples that have been grey- or blacklisted by the PCAWG consortium were removed (https://dcc.icgc.org/releases/PCAWG/donors_and_biospecimens). Fourth, samples with low mutational load were removed (≤35 single base substitutions), except for medulloblastoma and pilocytic astrocytoma samples. Lastly, cancer type cohorts with <20 samples were removed.

Variant calling

Variant calling for both the Hartwig and PCAWG cohorts was performed using the Hartwig Medical Foundation Platinum pipeline (v5) (https://github.com/hartwigmedical/pipeline5). Briefly, reads were mapped to GRCh37 using BWA (v0.7.17). GATK (v3.8.0) Haplotype Caller was used for calling germline variants in the reference sample. SAGE (v2.2) was used to call somatic single and multi nucleotide variants as well as indels. GRIDSS (v2.9.3) was used to call simple and complex structural variants. PURPLE (v2.53) combines B-allele frequency (BAF) from AMBER (v3.3), read depth ratios from COBALT (v1.7), and structural variants from GRIDSS to estimate copy number profiles, variant allele frequency (VAF) and variant clonality. PURPLE also determines sample gender based on sex chromosome ploidy.

LINX (v1.14) interprets structural variants (to identify simple and complex structural events) from PURPLE, and also detects gene fusions, viral DNA integrations, and homozygously disrupted genes.

Extraction of features

Regional mutational density

RMD was defined as the number of somatic SBSs in each 1Mb bin across the genome (n=3071), normalized by the total number of SBSs in the sample. Extraction of RMD profiles from these RMD bins was performed within the CUPLR training procedure (**Supplementary figure 1a**) using non-negative matrix factorization (NMF). For details, see **Methods: CUPLR training procedure**.

Mutational signatures

The number of somatic mutations falling into the 96 SBS, 78 DBS and, 83 indel contexts (as described in COSMIC: https://cancer.sanger.ac.uk/signatures/) was determined using the R package mutSigExtractor (https://github.com/UMCUGenetics/mutSigExtractor, v1.23). To obtain the mutational signature contributions for each sample, the mutation context counts were fitted to the COSMIC catalog of mutational signatures using the *nnlm()* function from the NNLM R package.

The contributions of the child signatures SBS7a, SBS7b, SBS7c and SBS7d were summed to yield the parent signature SBS7. Similarly SBS10a-d and SBS17a-b were merged to yield SBS10 and SBS17. Lastly, the SBS, DBS and indel signature contributions were normalized by the total number of SBSs, DBSs and indels respectively.

Chromosome arm and overall genome ploidy

Chromosome arm and genome ploidies were determined in a similar method as described by Taylor *et al.* 2018 [18].

Somatic copy number (CN) segments (called by PURPLE) were split by their respective chromosome arms. Only chr1-22 and chrX were included. All chromosomes have *p* and *q* arms, except for chr13, chr14, chr15, chr21, and chr22 which are considered to only have the *q* arm. For each chromosome arm, the CN values of each segment were converted to integer values. The arm coverage of each CN integer value was then determined (e.g. 70% of the arm has a CN of 2, 20% a CN of 1 and 10% a CN of 3). The CN with the highest coverage was assigned as the preliminary arm CN. The most common CN across all arms was assigned as the genome CN.

Two filtering steps were then performed to obtain the final arm CN values. For each arm, if the CN with the highest coverage has <50% coverage, and if any of the CN values equal the genome CN, then assign that CN as the final CN of the arm. Else, assign the genome CN as the final CN of the arm.

To determine the CN gains and losses of each arm, the difference between the arm CN and genome CN was calculated. Arms with CN differences >1 were considered to have a gain, and <1 a loss.

Features extracted from LINX output

LINX combines simple mutations (point mutations and indels), structural variants, and copy number variants to resolve simple and complex structural rearrangements, and subsequently identify gene driver events, gene fusions and detect viral DNA integrations.

The presence of 4 types of gene driver events were determined from the output of LINX: (i) amplifications, (ii) deep deletions, (ii) biallelic loss, and (iv) monoallelic hits (pathogenic mutation in one allele). Amplified genes were marked as *likelihoodMethod=='AMP'* by LINX. Genes with deep deletions were marked as *likelihoodMethod=='DEL'*. Genes with biallelic loss were marked as *biallelic==TRUE*. Genes with a monoallelic pathogenic mutation were marked as *biallelic==FALSE* and *driver=="MUTATION"*, and we selected only those with *driverLikelihood>=0.9* (referring to the likelihood of the mutation being impactful as determined by the dndscv R package [64]). LINX determines the presence of driver events for 462 genes. Thus, there are 462 genes x 4 driver types = 1848 gene driver features in total. We then performed preliminary feature selection to reduce the computational resources required for training CUPLR. Here, one-sided Fisher's exact tests were performed and Cramer's V values were calculated. Only genes where at least one driver type had a p-value <0.01 and Cramer's V ≥0.1 were kept (205 genes x 4 driver type = 820 gene driver features).

Gene fusions belonged to 3 categories: (i) well known fusion pairs (e.g. *TMPRSS2-ERG*), (ii) immunoglobulin heavy chain (IGH) locus fusions, and (iii) fusions with promiscuous gene partners (e.g. *BCR*). *IGH* fusions were grouped into a single feature as these are characteristic events in lymphoid cancers [65]. Fusions with one promiscuous gene partner were grouped by gene (e.g. *RUNX1_ETS2* and *RUNX1_RCAN1* would both fall under the *RUNX1_** feature). Fusions with two promiscuous gene partners were split into two separate features (e.g. *SLC45A3_MYC* would become the features *SLC45A3_** and **_MYC*). Only fusions that were marked as *reported==TRUE* by LINX (i.e. reported in literature) were selected. We then performed preliminary feature selection due to the large number of possible fusions present in our dataset (n=512). Here, one-sided Fisher's exact tests were performed and Cramer's V values were calculated. Only fusions with p-value <0.01 and Cramer's V >=0.1 were kept (45 fusions).

For the viral DNA integrations present in our dataset, we merged virus strains into 9 virus categories: adeno-associated virus (AAV), Epstein-Barr virus (EBV), hepatitis B virus (HBV), hepatitis C virus (HCV), human immunodeficiency virus (HIV), human papillomavirus (HPV), herpes simplex virus (HSV), human T-lymphotropic virus (HTLV), and Merkel cell polyomavirus (MCPyV). For example, *human papillomavirus type 16* and *human papillomavirus type 18* would be both grouped as *human papillomavirus*.

LINX chains individual SVs into SV clusters and classifies these clusters into various event types. Clusters can have one SV (for simple events such as deletions and duplications), or multiple SVs. We defined SV load as the total number of SV clusters. We quantified deletions and duplications (*ResolvedType is 'DEL' or 'DUP'*) stratified by length (1–10 kb, 10–100 kb, 100kb–1Mb, 1–10Mb, >10 Mb), as well as long interspersed nuclear element (LINE) insertions (*ResolvedType=='LINE'*), and normalized these counts by

SV load. We defined the number of double minutes as the number of clusters where *InDoubleMinute==TRUE*. We defined the number of foldback events as the number of clusters where *InfoStart=='FOLDBACK' or InfoEnd=='FOLDBACK' or ResolvedType=='RESOLVED_FOLDBACK'*. For clusters where *ResolvedType=='COMPLEX'*, we defined the complex SV load as the total number of these clusters, and the largest complex SV cluster as the cluster with most SVs.

CUPLR training procedure

Extraction of regional mutational density profiles

To extract the cancer specific RMD profiles from the 3071 RMD bins, a multistep procedure involving non-negative matrix factorization (NMF) (**Supplementary figure 2**) was performed prior to classifier training (**Supplementary figure 1ai**).

All NMF runs described below are performed with the *nnmf()* function from the NNLM R package (v0.4.4) with the parameters *loss='mkl'* and *max.iter=2000*.

For each cancer type cohort, an NMF rank search was done to determine the optimum rank (i.e. number of RMD profiles) (**Supplementary figure 2a**). For ranks 1 to 10, NMF was performed 50 times on a random subset of 100 samples from the cohort (or if the cohort contained less than 100 samples, all samples from that cohort were used) with 10% of the values randomly removed. The missing values were then imputed and the mean squared error (MSE) of these imputed values was calculated. This method of calculating MSE is described by the authors of the NNLM R package [66]. The median of the MSE across the 50 NMF iterations was then calculated. The rank search thus results in 10 MSE values across the 10 ranks searched. The optimum rank was the rank before the increase in log10(MSE) was >0.2%, and NMF was then performed using the optimum rank and without removing random values to produce the RMD profiles for the cancer type cohort (**Supplementary figure 2b**).

The above procedure thus yields a different set of RMD profiles for each cancer type cohort. However, some RMD profiles across related cancer types (e.g. pancreas and biliary cancer) may actually be equivalent RMD profiles. Hierarchical clustering (using Pearson correlation as a distance measure) was thus performed to group similar RMD profiles across all cancer type cohorts. The resulting dendrogram was cut at a height of 0.1 (using the R function *cutree()*), whereby RMD profiles under this height were grouped and considered the same profile. From each of the groups, one profile was greedily selected to yield the final set of RMD profiles (**Supplementary figure 2c**).

To obtain the RMD profile contributions for each sample, the RMD bins were fitted to the RMD profiles using the *nnlm()* function from the NNLM R package.

Random forest ensemble training

The main component of CUPLR comprises an ensemble of binary random forests that each discriminate one cancer type (**Supplementary figure 1aii**). The below text describes the training procedure for each cancer type random forest.

First, univariate feature selection was performed to remove irrelevant features (**Supplementary figure 1aiii**). Pairwise testing was done to compare feature values from samples of the target cancer type (case group) versus the remaining samples (control group). For numeric features, p-values were determined using Wilcoxon rank sum tests, and effect sizes were calculated using Cliff's delta. For boolean features, p-values were determined using Fisher's exact tests, and effect sizes were calculated using Cramer's V. Depending on the feature, alternative hypotheses for the Wilcoxon rank sum tests and Fisher's exact tests were one or two sided. See **Supplementary data 3** for details on which features are numeric or boolean, as well as which alternative hypothesis was used. Features were kept which had $p<0.01$ and effect size ≥ 0.1. The number of features kept was capped to 100 features.

Second, random oversampling was performed for the case group which always contained fewer samples than the control group, which were randomly undersampled (**Supplementary figure 1aiv**). A grid search was performed to determine the optimal pair of 4 oversampling and 4 undersampling ratios. These ratios were automatically determined as follows: i) calculate the geometric mean between the case and control group sample sizes; ii) the 4 resampling ratio values are logarithmically spaced between the geometric mean and the case group sample size or the control group sample size. For each over-/undersampling ratio pair, stratified 10-fold cross-validation (CV) was performed, after which the area under the precision recall curve (AUPRC) was calculated. The pair with the highest AUPRC was chosen and the resampling was applied.

Lastly, a random forest was trained (**Supplementary figure 1av**) using the *randomForest* R package (v4.6-14) with default settings. A filter is applied to the probabilities produced by the random forest based on sample gender, where breast, ovary and cervix probabilities are set to 0 for male samples, and prostate probabilities are set to 0 for female samples. Local increments were calculated for the random forest using the *rfFC* R package (v1.0) to enable downstream calculation of feature contributions [39].

Isotonic regression training

The entire random forest ensemble training procedure was then subjected to stratified 15-fold cross-validation which allows every sample to be excluded from the training set in order to obtain cancer type probabilities for the training samples (**Supplementary figure 1b**). These cross-validation probabilities were then used to train an ensemble of isotonic regressions using the *isoreg* R function (one per cancer type random forest) to calibrate the probabilities produced by the random forest ensemble (**Supplementary figure 1c, Supplementary figure 3**).

Random forests tend to be overconfident at probabilities towards 0 and underconfident at probabilities towards 1 [21], and this bias varies between random forests (**Supplementary figure 4**). In other words, a probability of e.g. 0.8 from one random forest does not correspond to a probability of 0.8 from another random forest. Probability calibration greatly reduced this bias ensuring that predictions across the random forests are comparable (**Supplementary figure 4**).

Performance evaluation

To assess the performance of CUPLR, we used the cancer type predictions based on the isotonic regression calibrated cross-validation probabilities, as well as by applying the final CUPLR model to a validation set whereby 10% of samples were held out from the full training set. Accuracy was calculated for each cancer type (as well as across all cancer types), which was the fraction of samples where the top predicted cancer type was the actual cancer type for the respective sample. Top-N accuracy was also calculated, which was the fraction of samples where the actual cancer type was amongst the top N predicted cancer types for the respective sample. 'Accuracy' is equivalent to 'top-1 accuracy'.

Data availability

Metastatic WGS data and corresponding metadata have been obtained from the Hartwig Medical Foundation and provided under data request numbers DR-104. Both WGS data and metadata is freely available for academic use from the Hartwig Medical Foundation through standardized procedures and request forms can be found at https://www.hartwigmedicalfoundation.nl. For access to identifying data (e.g. germline or raw read data) for the Pan-Cancer Analysis of Whole Genomes (PCAWG) cohort, researchers will need to request access via the ICGC Data Access Compliance Office (DACO; https://daco.icgc.org/). The extracted features for each sample and used to develop CUPLR is available at https://doi.org/10.5281/zenodo.5547228. All other data are available within the article, Supplementary Information or available from the authors upon request.

Code availability

The Hartwig Medical Foundation Platinum (https://github.com/hartwigmedical/pipeline5) is used for germline and somatic variant calling, as well as post-processing procedures such as identification of simple and complex structural rearrangements, annotation of driver gene mutation events, and detection of gene fusions and viral DNA integrations. CUPLR can be run from the output of this pipeline, and is available as an R package at https://github.com/UMCUGenetics/CUPLR. The code used for data processing and generating the figures is also available in this repository.

Declarations

Acknowledgements

This research was supported by an unrestricted grant of Stichting Hanarth Fonds, The Netherlands. This publication and the underlying study have been made possible partly on the basis of the data that Hartwig Medical Foundation and the Center of Personalised Cancer Treatment (CPCT) have made available to the study.

Author contributions

L.N. performed analyses, wrote/edited the paper. A.V.H. conceived the study, performed analyses, wrote/edited the paper. E.C. edited the paper and provided discussion. E.C. and A.V.H. supervised the study. All authors proofread, made comments, and approved the paper.

Competing interests

The authors declare no competing interests.

Supplementary data

Supplementary data 1: Metadata for tumors in the Hartwig and PCAWG datasets, including cancer type label, HRD and MSI status, and inclusion/exclusion in the training and holdout sets.

Supplementary data 2: Output from CUPLR based on cross-validation predictions on the training set, and predictions on the holdout set and on cancer of unknown primary samples.

Supplementary data 3: Description of each feature used by CUPLR.

Supplementary data 4: Feature importance for each random forest within CUPLR.

References

1. Anderson GG, Weiss LM. Determining tissue of origin for metastatic cancers: meta-analysis and literature review of immunohistochemistry performance. Appl Immunohistochem Mol Morphol. 2010;18: 3–8.
2. Pavlidis N, Pentheroudakis G. Cancer of unknown primary site. Lancet. 2012;379: 1428–1435.
3. Greco FA. Molecular diagnosis of the tissue of origin in cancer of unknown primary site: useful in patient management. Curr Treat Options Oncol. 2013;14: 634–642.

4. Jiao W, Atwal G, Polak P, Karlic R, Cuppen E, PCAWG Tumor Subtypes and Clinical Translation Working Group, et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nat Commun. 2020;11: 728.
5. Salvadores M, Mas-Ponte D, Supek F. Passenger mutations accurately classify human tumors. PLoS Comput Biol. 2019;15: e1006953.
6. Zhao Y, Pan Z, Namburi S, Pattison A, Posner A, Balachander S, et al. CUP-AI-Dx: A tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. EBioMedicine. 2020;61: 103030.
7. Grewal JK, Tessier-Cloutier B, Jones M, Gakkhar S, Ma Y, Moore R, et al. Application of a Neural Network Whole Transcriptome-Based Pan-Cancer Method for Diagnosis of Primary and Metastatic Cancers. JAMA Netw Open. 2019;2: e192597.
8. Moran S, Martínez-Cardús A, Sayols S, Musulén E, Balañá C, Estival-Gonzalez A, et al. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. Lancet Oncol. 2016;17: 1386–1395.
9. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet. 2016;17: 257–271.
10. Kim H, Wang X, Jin P. Developing DNA methylation-based diagnostic biomarkers. J Genet Genomics. 2018;45: 87–97.
11. Nangalia J, Campbell PJ. Genome Sequencing during a Patient's Journey through Cancer. N Engl J Med. 2019;381: 2145–2156.
12. Dietlein F, Eschner W. Inferring primary tumor sites from mutation spectra: a meta-analysis of histology-specific aberrations in cancer-derived cell lines. Hum Mol Genet. 2014;23: 1527–1537.
13. Marquard AM, Birkbak NJ, Thomas CE, Favero F, Krzystanek M, Lefebvre C, et al. TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen. BMC Med Genomics. 2015;8: 58.
14. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578: 94–101.
15. Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature. 2012;488: 504–507.
16. Cortés-Ciriano I, Lee JJ-K, Xi R, Jain D, Jung YL, Yang L, et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nat Genet. 2020;52: 331–341.
17. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence M, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015;518: 360–364.
18. Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, et al. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell. 2018;33: 676–689.e3.

19. Cameron DL, Baber J, Shale C, Papenfuss AT, Valle-Inclan JE, Besselink N, et al. GRIDSS, PURPLE, LINX: Unscrambling the tumor genome via integrated analysis of structural variation and copy number. bioRxiv. 2019. p. 781013. doi:10.1101/781013
20. Cameron DL, Baber J, Shale C, Valle-Inclan JE, Besselink N, van Hoeck A, et al. GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. Genome Biol. 2021;22: 202.
21. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. Proceedings of the 22nd international conference on Machine learning. New York, NY, USA: Association for Computing Machinery; 2005. pp. 625–632.
22. Lim YK, Padma R, Foo L, Chia YN, Yam P, Chia J, et al. Survival outcome of women with synchronous cancers of endometrium and ovary: a 10 year retrospective cohort study. J Gynecol Oncol. 2011;22: 239–243.
23. Henson DE, Schwartz AM, Nsouli H, Albores-Saavedra J. Carcinomas of the pancreas, gallbladder, extrahepatic bile ducts, and ampulla of vater share a field for carcinogenesis: a population-based study. Arch Pathol Lab Med. 2009;133: 67–71.
24. Sell S, Dunsford HA. Evidence for the stem cell origin of hepatocellular carcinoma and cholangiocarcinoma. Am J Pathol. 1989;134: 1347–1363.
25. Oronsky B, Ma PC, Morgensztern D, Carter CA. Nothing But NET: A Review of Neuroendocrine Tumors and Carcinomas. Neoplasia. 2017;19: 991–1002.
26. Cortes-Ciriano I, Lee S, Park W-Y, Kim T-M, Park PJ. A molecular portrait of microsatellite instability across multiple cancers. Nat Commun. 2017;8: 15180.
27. Nguyen L, W M Martens J, Van Hoeck A, Cuppen E. Pan-cancer landscape of homologous recombination deficiency. Nat Commun. 2020;11: 5584.
28. Christensen S, Van der Roest B, Besselink N, Janssen R, Boymans S, Martens JWM, et al. 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. Nat Commun. 2019;10: 4571.
29. Pich O, Muiños F, Lolkema MP, Steeghs N, Gonzalez-Perez A, Lopez-Bigas N. The mutational footprints of cancer therapies. Nat Genet. 2019;51: 1732–1740.
30. Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, et al. Mutational signatures associated with tobacco smoking in human cancer. Science. 2016;354: 618–622.
31. Tomlins SA, Laxman B, Varambally S, Cao X, Yu J, Helgeson BE, et al. Role of the TMPRSS2-ERG gene fusion in prostate cancer. Neoplasia. 2008;10: 177–188.
32. Faulkner C, Ellis HP, Shaw A, Penman C, Palmer A, Wragg C, et al. BRAF Fusion Analysis in Pilocytic Astrocytomas: KIAA1549-BRAF 15-9 Fusions Are More Frequent in the Midline Than Within the Cerebellum. J Neuropathol Exp Neurol. 2015;74: 867–872.
33. Aström AK, Voz ML, Kas K, Röijer E, Wedell B, Mandahl N, et al. Conserved mechanism of PLAG1 activation in salivary gland tumors with and without chromosome 8q12 abnormalities: identification of SII as a new fusion partner gene. Cancer Res. 1999;59: 918–923.

34. Mitani Y, Liu B, Rao PH, Borra VJ, Zafereo M, Weber RS, et al. Novel MYBL1 Gene Rearrangements with Recurrent MYBL1-NFIB Fusions in Salivary Adenoid Cystic Carcinomas Lacking t(6;9) Translocations. Clin Cancer Res. 2016;22: 725–733.
35. Göransson M, Andersson MK, Forni C, Ståhlberg A, Andersson C, Olofsson A, et al. The myxoid liposarcoma FUS-DDIT3 fusion oncoprotein deregulates NF-kappaB target genes by interaction with NFKBIZ. Oncogene. 2009;28: 270–278.
36. Sasaki T, Rodig SJ, Chirieac LR, Jänne PA. The biology and treatment of EML4-ALK non-small cell lung cancer. Eur J Cancer. 2010;46: 1773–1780.
37. Psyrri A, DiMaio D. Human papillomavirus in cervical and head-and-neck cancer. Nat Clin Pract Oncol. 2008;5: 24–31.
38. Dworkin AM, Tseng SY, Allain DC, Iwenofu OH, Peters SB, Toland AE. Merkel cell polyomavirus in cutaneous squamous cell carcinoma of immunocompetent individuals. J Invest Dermatol. 2009;129: 2868–2874.
39. Palczewska A, Palczewski J, Robinson RM, Neagu D. Interpreting random forest classification models using a feature contribution method. arXiv [cs.LG]. 2013. Available: http://arxiv.org/abs/1312.1121
40. van Meerbeeck JP, Fennell DA, De Ruysscher DKM. Small-cell lung cancer. Lancet. 2011;378: 1741–1755.
41. Niederst MJ, Sequist LV, Poirier JT, Mermel CH, Lockerman EL, Garcia AR, et al. RB loss in resistant EGFR mutant lung adenocarcinomas that transform to small-cell lung cancer. Nat Commun. 2015;6: 6377.
42. Kwong LN, Dove WF. APC and its modifiers in colon cancer. Adv Exp Med Biol. 2009;656: 85–106.
43. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020;578: 82–93.
44. Rodriguez-Martin B, Alvarez EG, Baez-Ortega A, Zamora J, Supek F, Demeulemeester J, et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nat Genet. 2020;52: 306–319.
45. Secrier M, Li X, de Silva N, Eldridge MD, Contino G, Bornschein J, et al. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. Nat Genet. 2016;48: 1131–1141.
46. Sabatini ME, Chiocca S. Human papillomavirus as a driver of head and neck cancers. Br J Cancer. 2020;122: 306–314.
47. Cheung LWT, Hennessy BT, Li J, Yu S, Myers AP, Djordjevic B, et al. High frequency of PIK3R1 and PIK3R2 mutations in endometrial cancer elucidates a novel mechanism for regulation of PTEN protein stability. Cancer Discov. 2011;1: 170–185.
48. Watson PA, Arora VK, Sawyers CL. Emerging mechanisms of resistance to androgen receptor inhibitors in prostate cancer. Nat Rev Cancer. 2015;15: 701–711.

49. Zundelevich A, Dadiani M, Kahana-Edwin S, Itay A, Sella T, Gadot M, et al. ESR1 mutations are frequent in newly diagnosed metastatic and loco-regional recurrence of endocrine-treated breast cancer and carry worse prognosis. Breast Cancer Res. 2020;22: 16.
50. Kocakavuk E, Anderson KJ, Varn FS, Johnson KC, Amin SB, Sulman EP, et al. Radiotherapy is associated with a deletion signature that contributes to poor outcomes in patients with cancer. Nat Genet. 2021;53: 1088–1096.
51. Lu MY, Chen TY, Williamson DFK, Zhao M, Shady M, Lipkova J, et al. AI-based pathology predicts origins for cancers of unknown primary. Nature. 2021;594: 106–110.
52. Ghobrial IM, Detappe A, Anderson KC, Steensma DP. The bone-marrow niche in MDS and MGUS: implications for AML and MM. Nat Rev Clin Oncol. 2018;15: 219–233.
53. Katz D, Palmerini E, Pollack SM. More Than 50 Subtypes of Soft Tissue Sarcoma: Paving the Path for Histology-Driven Treatments. Am Soc Clin Oncol Educ Book. 2018;38: 925–938.
54. Brierley J, O'Sullivan B, Asamura H, Byrd D, Huang SH, Lee A, et al. Global Consultation on Cancer Staging: promoting consistent understanding and use. Nat Rev Clin Oncol. 2019;16: 763–771.
55. Koh G, Degasperi A, Zou X, Momen S, Nik-Zainal S. Mutational signatures: emerging concepts, caveats and clinical applications. Nat Rev Cancer. 2021. doi:10.1038/s41568-021-00377-7
56. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020;578: 112–121.
57. Vöhringer H, Van Hoeck A, Cuppen E, Gerstung M. Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. Nat Commun. 2021;12: 3628.
58. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, et al. The evolutionary history of 2,658 cancers. Nature. 2020;578: 122–128.
59. Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature. 2019;574: 532–537.
60. Pleguezuelos-Manzano C, Puschhof J, Rosendahl Huber A, van Hoeck A, Wood HM, Nomburg J, et al. Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. Nature. 2020;580: 269–273.
61. Monzon FA, Medeiros F, Lyons-Weiler M, Henner WD. Identification of tissue of origin in carcinoma of unknown primary with a microarray-based gene expression test. Diagn Pathol. 2010;5: 3.
62. Houseley J, Tollervey D. The many pathways of RNA degradation. Cell. 2009;136: 763–776.
63. Hartwig Medical Foundation data access request guide. [cited 20 Aug 2021]. Available: https://hartwigmedical.github.io/documentation/data-access-request-guide.html
64. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017;171: 1029–1041.e21.
65. Schmitz R, Renné C, Rosenquist R, Tinguely M, Distler V, Menestrina F, et al. Insights into the multistep transformation process of lymphomas: IgH-associated translocations and tumor

suppressor gene mutations in clonally related composite Hodgkin's and non-Hodgkin's lymphomas. Leukemia. 2005;19: 1452–1458.

66. Lin X, Boutros PC. Optimization and expansion of non-negative matrix factorization. BMC Bioinformatics. 2020;21: 7.

Figures

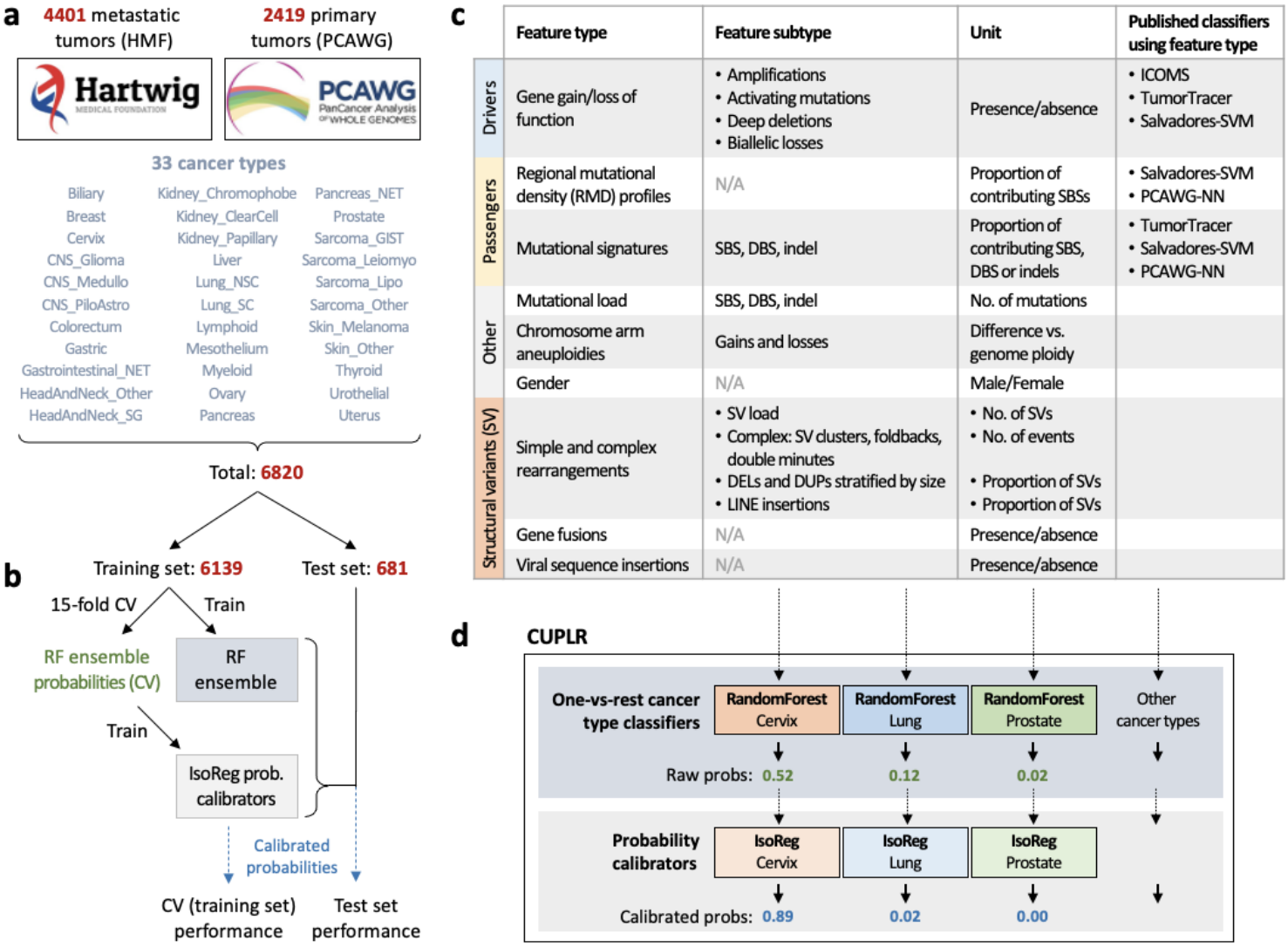

Figure 1

Cancer of Unknown Primary Location Resolver (CUPLR) classifies 33 different cancer types using features derived from all mutation types. (a) CUPLR was developed using whole-genome sequencing data 4401 metastatic tumors from the Hartwig Medical Foundation (Hartwig) and 2419 primary tumors from the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium, totaling 6820 samples across 33 different cancer types. (b) 6139 samples were used to train CUPLR. The whole training set was used to train the final random forest ensemble. 15-fold cross-validation was performed to obtain the random forest cancer type probabilities on the training set, which were then used to train the ensemble of isotonic

regressions (for probability calibration). CUPLR is composed of the random forest and isotonic regression ensembles as shown in (d). The performance of CUPLR was assessed using the calibrated cross-validation probabilities as well as probabilities obtained by applying CUPLR to the test set. (c) A summary of the genomic features extracted from the whole-genome sequencing data and used by CUPLR. The names of the published classifiers refer to the following studies; ICOMS: Inferring Cancer Origins from Mutation Spectra, Dietlen et al. 2014 [12]; TumorTracer: Marquard et al. 2015 [13]; Salvadores-SVM: support vector machine by Salvadores et al. 2019 [5]; PCAWG-NN: PCAWG neural network by Jiao et al. 2020 [4]. Abbreviations; RF: random forest, IsoReg: isotonic regression, CV: cross-validation, SBS: single base substitutions, DBS: double base substitutions, SV: structural variants, DEL: structural deletions, DUP: structural duplications, LINE: long interspersed nuclear element.

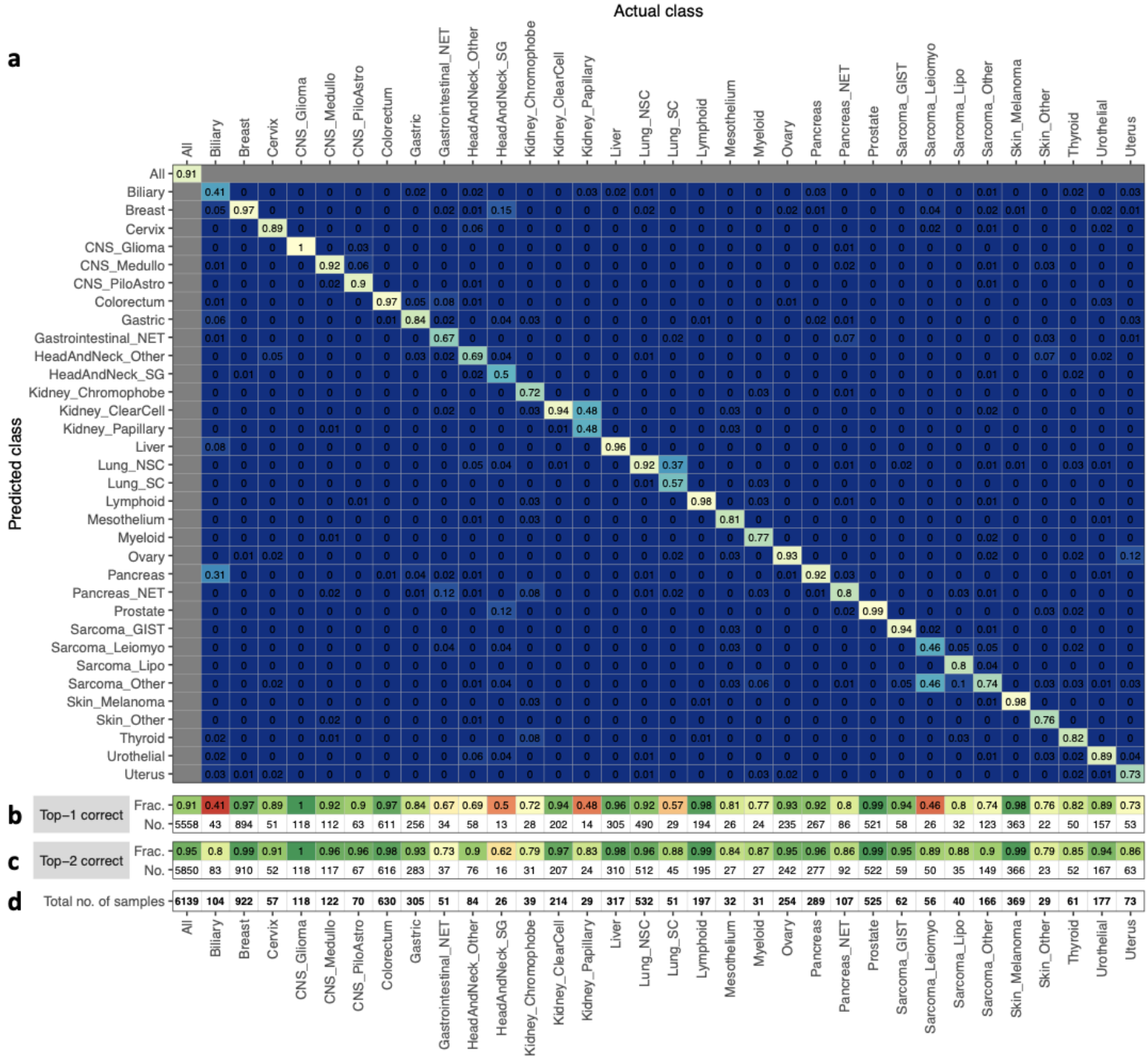

Figure 2

Accuracy of CUPLR based on cross-validation predictions. (a) A heatmap showing the accuracy of CUPLR where columns represent the fraction of samples in a cancer type cohort predicted as a particular cancer type. The diagonal represents the fraction of samples correctly predicted as a particular cancer type (i.e. accuracy). (b) The fraction and number of correctly classified samples, equivalent to the diagonal in (a). (c) The fraction and number of correctly classified samples when considering the 2 highest probability cancer types. (d) The total number of samples for each cancer type. Abbreviations; CNS: central nervous system; CNS_Medullo: medulloblastoma; CNS_PiloAstro: pilocytic astrocytoma; Gastrointestinal_NET: gastrointestinal neuroendocrine tumor; HeadAndNeck_SG: salivary gland cancer; HeadAndNeck_Other: head and neck cancers other than salivary gland cancer; Lung_NSC: non-small cell lung cancer; Lung_SC: small cell lung cancer; Pancreas_NET: pancreatic neuroendocrine tumor; Sarcoma_GIST: gastrointestinal stromal tumor; Sarcoma_Leiomyo: leiomyosarcoma; Sarcoma_Lipo: liposarcoma; Sarcoma_Other: sarcomas other than leiomyosarcoma, liposarcoma or gastrointestinal stromal tumors; Skin_Other: non-melanoma skin cancer.

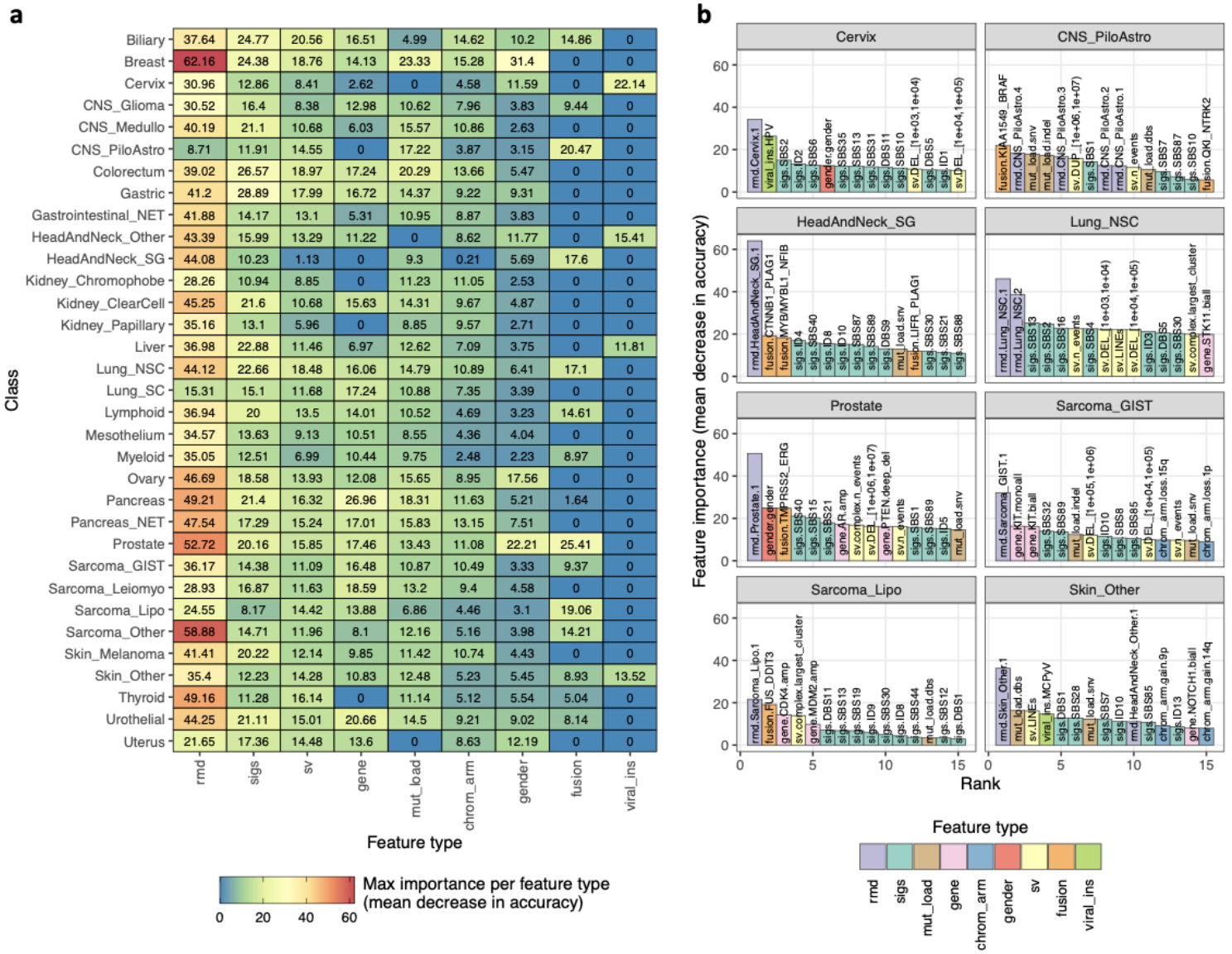

Figure 3

Features most predictive of cancer type. (a) Maximum feature importance per feature type for each cancer type random forest from CUPLR. Feature type definitions; rmd: regional mutation density profiles; sigs: mutational signatures; mut_load: total number of single base substitutions, double base substitutions or indels, gene: presence of gene gain or loss of function events; chrom_arm: chromosome arm copy number difference compared to overall genome ploidy; gender: sample gender as determined by sex chromosome copy number; sv: structural variants; fusion: presence of gene fusions; viral_ins: presence of viral sequence insertions. For a full description of each feature, see Supplementary data 3. (b) Importance of the top 15 features for selected cancer type random forests. Feature names are in the form {feature type}.{feature name}.

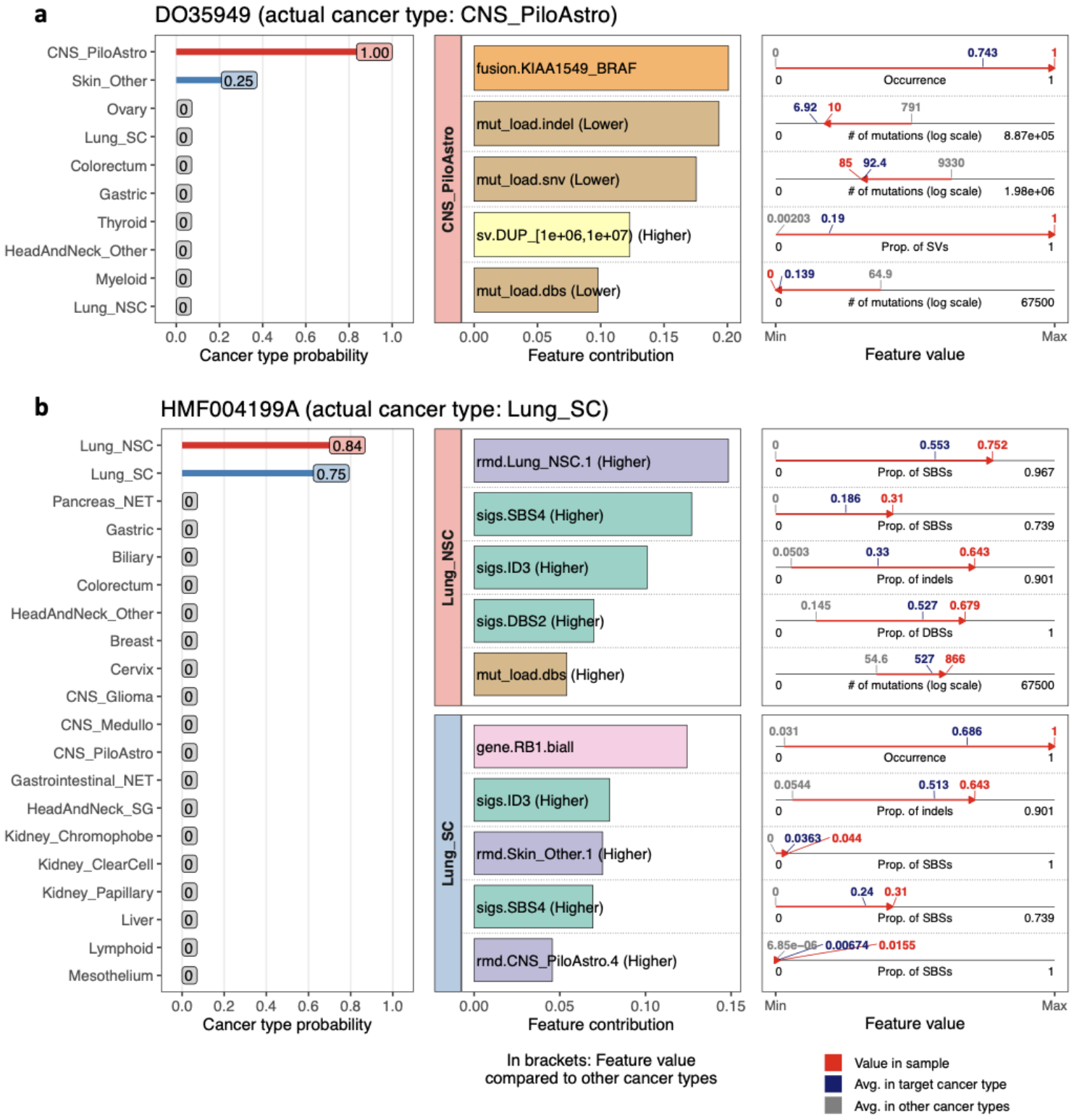

Figure 4

Graphical report for CUPLR predictions. Example reports are shown for two patients from the holdout set: (a) DO35949, annotated as having pilocytic astrocytoma (CNS_PiloAstro), and (b) HMF004199A, annotated as having small cell lung cancer (Lung_SC). The leftmost panels show the predicted cancer type probabilities. In the middle panels, contributions of the top features are shown for the top predicted cancer types. When there is uncertainty in the cancer type probabilities such as in (b), more feature

contribution panel rows are shown. In the rightmost panel, the feature values in the patient are plotted in relation to the average feature value amongst patients in the training set with and without the target cancer type. For non-Boolean features (i.e. not those where the values are present or absent), the red arrows indicate whether the feature value in the patient is higher or lower than in patients without the target cancer type. For a full description of each feature, see Supplementary data 3.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S1samplemetadata.xlsx
- S2predsumm.xlsx
- S3featuredescriptions.xlsx
- S4featimp.xlsx
- SupplemenentaryInformation.docx