Accurate Discrimination of 23 Major Cancer Types via Whole Genome Somatic Mutation Patterns

Wei Jiao[1]*, Paz Polak[2]*, Rosa Karlic[3], Gad Getz[2], Lincoln Stein[1,4]‡, for the PCAWG Pathology and Clinical Correlates Working Group and the ICGC/TCGA Pan-cancer Analysis of Whole Genomes Network

[1] Ontario Institute for Cancer Research, Toronto, ON, Canada M5G0A3
[2] The Broad Institute of MIT and Harvard, Cambridge, MA, USA 02142
[3] University of Zagreb, Horvatovac 102a, Zagbreb Croatia
[4] Department of Molecular Genetics, University of Toronto, Toronto, ON Canada

* These authors contributed equally to the work
‡ Corresponding author

Abstract

The two strongest factors predicting a human cancer's clinical behaviour are the primary tumour's anatomic organ of origin and its histopathology. However, roughly 3% of the time a cancer presents with metastatic disease and no primary can be determined even after a thorough radiological survey. A related dilemma arises when a radiologically defined mass is sampled by cytology yielding cancerous cells, but the cytologist cannot distinguish between a primary tumour and a metastasis from elsewhere.

Here we use whole genome sequencing (WGS) data from the ICGC/TCGA PanCancer Analysis of Whole Genomes (PCAWG) project to develop a machine learning classifier able to accurately distinguish among 23 major cancer types using information derived from somatic mutations alone. This demonstrates the feasibility of automated cancer type discrimination based on next-generation sequencing of clinical samples. In addition, this work opens the possibility of determining the origin of tumours detected by the emerging technology of deep sequencing of circulating cell-free DNA in blood plasma.

Introduction

Human cancers are distinguished by their anatomic organ of origin and their histopathology. For example, lung squamous cell carcinoma originates in the lung and has a histology similar to the normal squamous epithelium that lines bronchi and bronchioles. Together these two criteria, which jointly reflect the tumour's cell of origin, are the single major predictor of the natural history of the disease, including the age at which the tumour typically manifests, its risks factors, its growth rate, pattern of invasion and metastasis, response to therapy, and overall prognosis. A tumour's type is generally determined by a histopathologist who examines microscopic stained sections of the tumour. An increasing number of tumour types are

subclassified with molecular markers that have been demonstrated to distinguish among subtypes with clinically distinct features.

We now know, based on recent large-scale exome and genome sequencing studies, that major tumour types have dramatically different patterns of somatic mutation.[1-4] For example, ovarian cancers are distinguished by a high rate of genomic rearrangements,[5] chronic myelogenous leukemias carry a nearly pathognomonic structural variation involving a t(9;22) translocation leading to a BCR-ABL fusion transcript,[6] melanomas have high rates of C>T and G>A transition mutations due to UV damage,[7] and pancreatic ductal adenocarcinomas have near-universal activating mutations in the KRAS gene.[8]

This paper asks whether we can use machine learning techniques to accurately determine tumour organ of origin and histology using the patterns of somatic mutation identified by whole genome DNA sequencing. The primary motivation of this effort was to demonstrate the feasibility of a next-generation sequencing (NGS) based diagnostic tool for tumour type identification. Studies have shown that site-directed therapy based on the tumour's cell of origin is more effective than broad-spectrum chemotherapy;[9] however it is not always straightforward to determine where a metastatic tumour comes from. In the most extreme case, a pathologist may be presented with the challenge of determining the source of a poorly differentiated metastatic cancer when multiple imaging studies have failed to identify the primary ("cancer of unknown primary," CUPS).[10] More frequently, the pathologist must distinguish between two or more biologically distinct but histologically similar tumour types, such as the class of "small round cell" tumours.[11]

In current practice, pathologists use histological criteria and a series of immunohistochemical stains to determine such tumours' histological type and site of origin,[12] but this process can be complex and time-consuming, and some tumours are so poorly differentiated that they no longer express the cell-type specific proteins needed for unambiguous immunohistochemical classification. A simple NGS-based sequencing and analysis protocol for tumour type determination that could be applied to a variety of fresh and fixed clinical specimens would be a useful adjunct to existing histopathological techniques. It might also be helpful for characterising cytological specimens, such as those obtained via needle aspiration.

In designing these experiments, we speculated that DNA-sequence based tumour type identification would be most accurate when triangulated from three major categories of mutational feature: (1) the topological distribution of somatic passenger mutations, which reflect the epigenetic state of the tissue of origin ; (2) the distribution of somatic mutation types, which reflect environmental and genetic exposures of the cell of origin; and (3) the driver genes and pathways that are altered in the tumour. Our results indicate that each of these categories provides sufficient information to discriminate many, but not all, tumour types, and that combinations of all three feature categories usually outperform the individual ones.

Results

Our overall strategy was to develop a series of features representative of the three general categories described in the introduction. Using the Pan-cancer Analysis of Whole Genomes

(PCAWG) data set, which consists of >2,800 primary tumours across 38 forms of cancer subjected to paired whole genome sequencing,[4] we evaluated each individual feature type for its ability to accurately predict the cancer type. We then evaluated the accuracy of machine learning classifiers built on combinations of feature types. Finally, the best performing classifier was validated against an independent set of tumour genomes to determine overall predictive accuracy. For comparison, we determined the performance of DNA-based tumour type classification based on whole exome sequencing (WXS). Lastly, we examined patterns of misclassification errors to identify cases in which tumour types share biology or contain latent type heterogeneity.

Tumour Type Groupings

The full PCAWG data set consists of tumours from 2834 donors comprising 34 main histopathological tumour types. However, the tumour type groups are unevenly represented, and several have inadequate numbers of specimens to adequately train and test a classifier. We chose a minimum cutoff of 35 donors per tumour type. In some cases, the same donor had contributed both primary and metastatic tumour specimens to the PCAWG data set. In such cases, we used the primary tumour for training and evaluation. The resulting initial set consisted of 2606 tumours spanning 25 major types.

During preliminary development of the classifier, we noted that the tumour types chronic lymphocytic leukemia (Lymph-CLL) and B-cell non-Hodgkins lymphoma (Lymph-BNHL) were misclassified as each other more than 50% of the time. Both tumour types arise from malignant B-cells, and are interrelated by a phenomenon known as Richter's transformation in which CLL transforms into diffuse large B-cell lymphoma in 5-10% of patients.[13] We also suspect that these two disease entities share a common trajectory of genomic alterations. Consequently, we pooled these two tumour types into a group labeled ***Lymphoid.*** Similarly, we found considerable misclassification among samples representing the myeloproliferative neoplasm (Myeloid-MPN) and acute myeloid leukemia (Myeloid-AML), again possibly reflecting the common progression of the former to the latter[14] and/or overlapping biology. We pooled these two PCAWG types into a single grouping labeled ***Myeloid.*** These merging steps reduced the number of major tumour types to 23, still comprising 2606 tumours (Table 1).

Classification using Single Mutation Feature Types

We first evaluated a series of tumour type classifiers based on a single class of feature derived from the tumour mutation profile. For each feature class we developed a random forest classifier using the forward feature selection method (Online Methods). Each classifier's input is the mutational profile for an individual tumour specimen, and its output is the probability estimate that the specimen belongs to the type under consideration. We trained each classifier using a randomly selected set of 75% of samples drawn from the corresponding tumour type. To determine the most likely type for a particular tumour samples, we applied its mutational profile to each of the 23 type-specific classifiers, and selected the type whose classifier emitted the highest probability. To evaluate the overall accuracy of the system, we stratified sampling the data sets in four folds, trained on three quarters of the data set and tested against each of the

other quarter specimens. We report accuracy (recall, precision and F1) using the average of all four test data sets (Online methods).

We selected a total of seven mutational feature types spanning three major categories (Table 2):

Mutation Distribution. The somatic mutation rate in cancers varies tremendously from one region of the genome to the next.[2] In whole genome sequencing, a major covariate of this regional variation in whole genome sequences is the epigenetic state of the tumour's likely cell of origin, with 74-86% of the variance in the mutation density being explained by histone marks and other chromatin features related to open versus closed chromatin.[15] This suggests that tumours with the same cell of origin will have a similar topological distribution of mutations across the genome. To capture this, we divided the genome into ~3000 1 Mbp bins and created features corresponding to the number of somatic mutations per bin normalized to the total number of somatic mutations. Mutation rate profiles were created independently for somatic substitutions (SNV), indels, somatic copy number alterations (CNA), and other structural variations (SV). Note that the preponderance of variants used for this analysis are non-functional passenger mutations.

Mutation Type. The type of the mutation and its nucleotide neighbors, for example G{C>T}C, is a strong indicator of the exposure history of the cell of origin to extrinsic and endogenous factors that promote mutational processes.[16] This in turn can provide information on the tumour's organ of origin. For example, skin cancers have mutation types strongly correlated with UV light-induced DNA damage. Reasoning that similar tumour types will have similar mutational exposure profiles, we generated a series of features that represented the normalized frequencies of each potential nucleotide change in the context of its 5' and 3' neighbors. Like the mutation distribution, the variants that contribute to this feature category are mostly passengers.

Driver Gene/Pathway. Some tumour types are distinguished by high frequencies of alterations in particular driver genes and pathways. For example, melanomas have a high frequency of BRAF gene mutations,[17] while pancreatic cancers are distinguished by KRAS mutations.[8] We captured this in two ways: (1) whether a known or suspected driver event, which includes mutations in coding sequences, long non-coding RNA, and micro-RNAs, are contained in the tumour, and (2) whether there was an impactful coding mutation in any gene belonging to a known or suspected driver pathway. We did not attempt to account for alterations in cis-regulatory regions. In all we created ~2000 driver pathway-related features describing potential gene and pathway alterations for each tumour. By definition, this feature includes only somatic mutational events that act as drivers.

Figure 1, columns 1-7, describe the accuracy for a set of 161 classifiers built using single feature categories. The accuracy of individual classifiers ranges widely across tumour and feature categories, with an F1 (harmonic mean of recall and precision) of 0.48 and a range from 0.00 to 0.99. Thirteen tumour types had at least one well-performing classifier that achieved an F1 of 0.80: Lymphoid, CNS-GBM, CNS-PiloAstro, Prost-AdenoCA, ColoRect-AdenoCA, Skin-Melanoma, Panc-AdenoCA, Kidney-RCC, Liver-HCC, Lung-SCC, Kidney-ChRCC, CNS-Medullo and Myeloid. Three classifiers performed poorly, with no classifier achieving an accuracy of 0.6:

Bone-Osteosarc, Stomach-AdenoCA and Uterus-AdenoCA. The remaining eight tumour types had classifiers achieving accuracies between 0.60 and 0.80.

In general, the highest accuracies were observed for features related to mutation type and distribution. Only Panc-AdenoCA achieved an accuracy of at least 0.80 for classifiers based on features related to driver genes/pathways, suggesting that few of the other tumour types examined here have sufficiently distinctive alterations in driver pathways.

Classification using Combinations of Mutation Feature Types

To test whether we could improve classifier accuracy, we built tumour-specific classifiers that combined feature categories. Using the three best-performing feature types for each tumour, we built, trained and tested random forest classifiers for each combination of two and three types (Figure 1, column 8; Supplementary Figure 1). In all cases, classifiers built by combining feature types showed improved accuracy relative to those built on single feature types, as assessed by cross-validation. However, the extent of improvements vary from one cancer type to the other. For example, Bone-Osteosarc, which failed to achieve an accuracy greater than 0.56 with any individual feature, achieved an F1 of 0.70 for a combination of three feature types. However, for Lymphoma, Skin-Melanoma and Liver-HCC, which already achieved high levels of accuracy with single feature types, combining types barely improved the classification accuracy.

For each tumour type, we selected the best classifier from among those based on single features types or feature combinations, resulting in a final set of 23 tumour-specific classifiers. This merged classifier was used for all subsequent testing and validation.

Figure 2 shows the performance of the merged classifier when tested against held out tumours (mean of 4 test runs). Overall, the accuracy for the complete set of 23 tumour types was 0.86 (classification accuracy), but there was considerable variation for individual tumours types. Sensitivity/recall ranged from 0.35 (Stomach-AdenoCA) to 1.0 (Lymphoid). The precision, which reflects the proportion of true positives among the calls made by a classifier, and is sensitive to the number of positives in the data set, was somewhat lower, with rates ranging from 0.44 (Stomach-AdenoCA) to 0.97 (Liver-HCC and Lymphoid).

Fourteen tumour types achieved accuracies of at least 0.80, including all 13 of the types that met this threshold for single-feature types, plus Head-SCC. Three tumour type classifiers, Ovary-AdenoCA, Uterus-AdenoCA, and Stomach-AdenoCA performed poorly and failed to achieved an accuracy of at least 0.60. The remaining six types had middling classifier accuracies ranging from 0.69 to 0.78.

By design, the combined tumour type classifier consists of one classifier for each tumour type, which when given the features of an unknown tumour emits a probability between 0 and 1 that the unknown tumour belongs to the type. The test tumour is then assigned to the type whose classifier gives the highest relative probability. This means that for each unknown tumour there is a natural ranking of the top pick, the second-best pick, and so forth. We asked how often the correct type was within the top N picks. As shown in Figure 3 the correct choice was ranked at the top 86% of the time, was within the top two choices 93% of the time, and among the top three choices 95% of the time.

Tumour-Specific Features

We asked which features the merged classifiers selected from each tumour type to identify biologically-distinguishing characteristics (Supplementary Tables 1 and 2). The majority of classifiers selected mutation type and/or mutation distribution as the most influential features, emphasizing the importance of exposures and epigenetically-related cell-of-origin marks in distinguishing tumour types. For some tumour types, mutation distribution features dominated. Features from mutation distribution accounted for more than 90% of the features selected by the classifiers for liver, breast, esophageal adenocarcinoma, melanoma, ovarian serous adenocarcinoma, the lymphoid tumour category, and both types of lung cancer. The stomach adenocarcinoma and myeloid classifiers are both dominated by mutation type features, which account for nearly two-thirds of their selected features. The classifiers for other tumour types, including both types of pancreatic cancer, prostate adenocarcinoma, and thyroid adenocarcinoma, used a more balanced mixture of mutation type and mutation distribution.

In some cases the classifiers identified individual gene-related features and focal variants that distinguish one tumour type from others. In pilocytic astrocytoma, the strongest feature was a SV hotspot on chromosome 7 that reflects a BRAF fusion transcript present in two-thirds of cases of this tumour type.[18] The classifiers for pancreatic ductal adenocarcinoma and high-grade serous ovarian carcinoma picked out oncogenic mutations in KRAS and TP53 as discriminative features. Activating mutations of KRAS are present in ~95% of pancreatic cancers,[8] and a roughly similar proportion of inactivating TP53 mutations are present in ovarian cancer.[19] The lung squamous cell carcinoma classifier picked up recurrent point mutations in the NFE2L2 (NRF2) transcription factor mutations. This gene has previously been identified as recurrently mutated in Lung-SCC and is a marker of poor prognosis.[20] In the poorly-performing uterine adenocarcinoma classifier, mutations in the PIK3R1, PTEN and PPP2R1A genes were selected. Although mutations in each of these genes is associated with uterine cancer,[21] they are also common in ovarian and breast cancer, which may help explain the uterine adenocarcinoma misclassification pattern described in the following section.

The thyroid adenocarcinoma classifier picked up multiple features relating to copy number amplifications on chromosome 8 spanning the c-Myc gene. We found this curious since such amplifications are not characteristic of this tumour type, but are common in many other tumour types. On further inspection we found this to be a negatively weighted feature; the absence of this amplification has positive predictive value for the Thy-AdenoCa classifier.

Patterns of Misclassification

Misclassifications produced by the set of classifiers are not haphazard, but instead seem to reflect overlapping biological characteristics. For example, esophageal adenocarcinoma (Eso-AdenoCA), is misclassified as gastric adenocarcinoma (Stomach-AdenoCA) 35% of the time, and Stomach-AdenoCA cases are misclassified as Eso-AdenoCA 20% of the time. Examining the features selected for the esophageal and stomach cancer classifiers, we find that mutation substitution are the major predictive features for both tumour types, and both classifiers picked up a subset of C->A, C->G substitutions mutations as features. Those substitutions are commonly observed in both stomach and esophageal cancer and comprise Signature 17 in the COSMIC catalogue of mutational signatures.[14] In addition to stomach and

esophageal tumours, a small number of B-cell lymphomas exhibit Signature 17, which may explain why a fraction of Lymphoid tumours are misclassified as Eso-AdenoCA by our classifier. It is also possible that a subset of esophageal tumours are arising at the gastroesophageal junction (GEJ), which some consider to be a distinct subset of esophageal tumours with molecular characteristics similar to gastric adenocarcinoma.[22.]

Another frequent pattern of misclassifications is confusion between squamous cell carcinoma of the lung (Lung-SCC) and squamous cell carcinoma of the head and neck (Head-SCC). While Head-SCC samples were classified accurately 80% of the time and had no systematic misclassification patterns, only 69% of Lung-SCC samples were correctly classified, and were instead called Head-SCC 23% of the time. This pattern of misclassification appears to be related to sharing of the tobacco smoking signature (Signature 4) in both tumour types. The asymmetry of the misclassification errors appears to be the result of the Head-SCC classifier's selection of T{C>TG}T and T{C>TG}A mutations as a discriminating feature in Head-SCC. In our sample set, this pattern of mutation is enriched in squamous cell carcinomas of the head and neck, but not in lung. These mutations have been associated with APOBEC-mediated mutation processes.[16]

The classifier for invasive ductal adenocarcinoma of the breast (Breast-AdenoCA) has high sensitivity (97% of breast cancers are classified correctly), but relatively low precision (70%). This classifier incorrectly identifies as breast cancer 33%, 17%, 11% and 8% of Ovary-AdenoCA, Uterus-AdenoCA, Lung-AdenoCA and Thy-AdenoCA cases respectively. We speculated that this lack of specificity is the result of breast cancer's highly heterogeneous molecular subtypes,[23] among which is a basal subtype that shares molecular characteristics with high grade serous ovarian adenocarcinoma.[24] Of the 280 Breast-AdenoCA samples in PCAWG, 182 had associated RNA-seq profiling data and had been subtyped by the PCAWG Pathology and clinical correlates working group using the PAM50 molecular classification system.[25] From this list, we retrained a breast cancer classifier using Breast-AdenoCA samples having no basal subtype. When applied to non-breast cancers, the misclassification rate was reduced to 1, 3, 0 and 0% of ovarian, uterine, lung and thyroid tumours respectively, supporting the hypothesis that the inclusion of the basal subtype of breast cancer contributed to its classifier's reduced precision.

Classifier Accuracy Across an Independent Collection of Cancer Whole Genomes

A distinguishing characteristic of the PCAWG data set is its use of a uniform computational pipeline for sequence alignment, quality filtering, and variant calling. In real world settings, however, the data set used to train the classifier may be called using a different set of algorithms than the test data. To assess the accuracy of DNA-based tumour identification when applied to a more heterogeneous data set, we applied the classifier trained on PCAWG samples to an independent validation set of 1,600 cancer whole genomes assembled from a series of published non-PCAWG projects. The validation set spans 14 distinct tumour types assembled from 21 publications or databases (Table 3). We were unable to collect sufficient numbers of independent tumour genomes representing nine of the 23 types in the merged classifier, including colorectal cancer, thyroid adenocarcinoma and lung squamous cell carcinoma. In addition, only SNV calls (somatic point mutations) and small indels were available for these genomes; hence features relating to copy number and structural variations were marked as

unavailable. SNV coordinates were lifted from GRCh38 to GRCh37 when necessary. With the exception of a set of brain tumour samples in the validation set, which is explained below, a comparison of the mutation burden among each tumour type cohort revealed no significant differences between the PCAWG and validation data sets (Supplementary Figure 2) that would suggest strong batch effects due to sequencing coverage or analytic methods.

As shown in Figure 4, the classifier accuracy for the tumour types included in the validation data set ranged from 51-87% (overall F1 score 0.80). Following the trend observed for the PCAWG data set, the Breast-AdenoCA, Skin-Melanoma, CNS-Medullo, and Prost-Adeno tumour types were classified with at least 80% accuracy. A set of CNS gliomas was accurately classified 51% of the time, and the remaining tumour types were classified correctly in 60-79% of cases. The majority of errors mirrored the pattern of misclassifications previously observed within the PCAWG samples.

We were initially puzzled that the set of 34 CNS glioma samples from the validation data set overwhelmingly matched to the pediatric piloastrocytoma model rather than to the CNS-GBM model. However, on further investigation, we discovered that the CNS glioma samples represent a mixture of low- and high-grade pediatric gliomas, including piloastrocytomas.[26-28] The SNV mutation burden of these pediatric gliomas is also similar to CNS-PiloAstro and significantly lower than adult CNS-GBM (Supplementary Figure 2).

Application to Whole Exome Sequencing

Lastly, we asked how well the DNA based classifier would perform on reduced representation sequencing, such as whole exome sequencing (WXS). To evaluate this, we generated a series of synthetic whole exomes from the PCAWG testing set by retaining only those somatic mutations falling within annotated UTR and CDS regions. From the synthetic whole exomes we were able to derive classification features related to altered genes, pathways and mutation types, but lacked features relating to copy number alterations, structural variations, and passenger mutation distribution. These features were then applied to the classifiers built from WGS data. As expected, the classifier performance was degraded for all tumour types. Lymphoid, Liver-HCC and Breast-AdenoCA, each of which had accuracies of >0.95 in the WGS data set, had accuracies of 0.55, 0.87, and 0.48 in the WXS data set respectively. Overall, accuracy of the ensemble of classifiers was 61% (Supplementary Figure 3).

Code Availability

The R code developed for training and testing the classifier, along with documentation and trained models for the 23 tumour types are available from GitHub at [URL TO COME]. The code is distributed under the Apache Version 2.0 Open Source license (https://www.apache.org/licenses/LICENSE-2.0).

Discussion

In this paper, we used the largest collection of uniformly processed cancer whole genomes assembled to date to develop a supervised machine learning system capable of accurately distinguishing 23 major tumour types based solely on features that can be derived from DNA

sequencing. The accuracy of the system overall was 86%, with 14 of the 23 tumour types achieving recalls of 80% or higher. When the tumour type predictions were ranked according to likelihood, the correct prediction was found among the top three rankings 95% of the time.

While the accuracy of the classifier was primarily assessed using an internal held-out data set strategy, the application of an entirely independent WGS tumour validation set representing 14 of the tumour types achieved a satisfactory overall accuracy of 0.80 (F1-metric). The limitation of the validation experiment set is that the mutation calling was performed using a non-uniform set of alignment and variant-calling protocols, and only SNV-based features were available for our use. We expect to see better predictive power from an independent set of whole genomes on which full PCAWG-level variant calling has been performed.

The topographic distribution of somatic passenger mutations across the genome was by far the single most predictive class of features for the ensemble of classifiers, followed by mutation type and driver gene/pathway. On this basis, we expect to see degraded performance on exome sequencing, in which only 3-5% of the genome is sampled and the presence of purifying selection across many protein-coding sites makes unbiased determination of the topographic distribution of passenger mutations challenging to determine. Indeed, when we attempted to build a classifier based only on features that can be derived from WXS, our overall accuracy was reduced from 0.86 to 0.61 (F1 metric).

Previous work in the area of DNA-based tumour type identification has used targeted gene panel and whole exome sequencing strategies.[29,30] The targeted gene-based approach described in Tothill ***et al***[29] can discriminate a handful of tumour types that have distinctive driver gene profiles, and can identify known therapeutic response biomarkers, but does not have broader applicability to the problem of tumour typing. In contrast, the whole exome sequencing approach reported in Chen ***et al***[30] used a machine learning approach similar to ours to discriminate among 17 tumour types based on somatically altered gene, mutation type, chromosome and altered pathway. This approach achieved a F1 of 0.62 overall, and an F1 of 0.70 for five primary sites (colon, liver, skin, pancreas and lung), which is similar to what we observe in the simulated WXS data set. We demonstrate here that the addition of whole genome sequencing data and a richer set of somatic mutation types, such as structural alterations, substantially improves discriminative ability across a wider spectrum of tumour types. Indeed, an advantage of this study is that the comprehensive data set allows us to systematically evaluate the accuracy of classifiers built on top of features derived from different sequencing strategies.

There was considerable variability in the classification accuracy among tumour types. In some cases the poorly performing type-specific classifiers appear to have been confused by biological similarities across types. In at least the case of breast cancer tumour subtype heterogeneity within the training set also appears to have degraded the precision of the classifier. The issue of subtype heterogeneity is likely to be a more general problem. One of several examples is thyroid adenocarcinoma, in which the papillary and follicular forms are known to have distinct patterns of molecular alteration,[33] but due to the limited size of the data set available we were forced to pool these histological subtypes. We expect to see improvements in the classifier after training it with subtype-specific training sets.

In addition to training using larger subtype-specific tumour cohorts, there are other potential ways to improve classifier performance. Most notably, in the current implementation we did not make any use of germline information. However, many germline cancer risk alleles increase the risk of developing specific tumour types, for example BRCA1 mutations for breast and ovarian cancer, and APC mutations for colon cancer. Adding information on cancer-associated genetic loci would likely improve classifier accuracy among the subset of patients carrying such germline risk alleles.

Cancer of unknown primary site (CUPS) is a heterogeneous set of cancers diagnosed when a patient presents with metastatic disease, but despite extensive imaging, pathological and molecular studies the primary cannot be determined.[10] CUPS accounts for 3-5% of cancers, making it the seventh to eighth most frequent type of cancer and the fourth most common cause of cancer death.[31] Even at autopsy, the primary cannot be identified roughly 70% of the time,[31] suggesting regression of the primary in many CUPS cases. CUPS is a clinical dilemma, because therapeutic options are largely driven by tissue of origin, and site-directed therapy is more effective than broad-spectrum chemotherapy.[33] Recent studies of molecular typing of biopsies of CUPS patients using mRNA or miRNA expression profiling, have demonstrated the ability to identify a putative primary greater than 90% of the time,[33-35] while a prospective clinical trial of CUPS patients treated according to the tissue of origin specified by a RT-PCR based RNA expression profiling technique showed a modest improvement in overall outcomes relative to standard empiric treatment.[36]

A more common occurrence is the diagnostic challenge faced by surgical pathologists facing a differential diagnosis. For example, a lung nodule in a female patient biopsied by fine needle aspiration may reveal a poorly differentiated adenocarcinoma that could be a primary lung tumour or alternatively metastatic breast cancer. A "small round cell tumour" in a pediatric patient might be lymphoma, Ewing's sarcoma, Wilm's tumour, a neuroendocrine tumour, or melanoma. While conventional immunohistochemistry using a series of monoclonocal antibodies is generally successful at distinguishing these histologically-similar tumour types, the process can be time-consuming and labour-intensive, and the decision tree varies according to the differential diagnosis.[37] Further, immunohistochemistry can be confounded by the loss of antigens in poorly differentiated tumours.[38]

Given the increasing likelihood that in the near future most cancers will be subject to routine genomic profiling to identify actionable mutations, it is attractive to consider the possibility of simultaneously deriving the cancer type using an automated computational protocol. This would serve as an adjunct to histopathological diagnosis, and could also be used as a quality control check to flag the occasional misdiagnosis or to find genetically unusual tumours. More forward-looking is the prospect of accurately determining the site of origin of circulating cell-free tumour DNA detected in the plasma using so-called "liquid biopsies."[39] As genome sequencing technologies continue to increase in sensitivity and decrease in cost, there are realistic prospects for blood tests to detect early cancers in high risk individuals.[40] The ability to suggest the site and histological type of tumours detected in this way would be invaluable for informing the subsequent diagnostic workup.

In summary, this is the first study to demonstrate the potential of whole genome sequencing to distinguish major cancer types on the basis of somatic mutation patterns alone. Future studies will focus on improving the classifier performance by training with larger numbers of samples, subdividing tumour types into major molecular subtypes, adding new feature types, and adapting the technique to work with clinical specimens such as those from formalin-fixed, paraffin-embedded biopsies and cytologies.

Acknowledgements

We would like to thank Irina Kalatskaya, Quang Trinh, Quaid Morris and Jared Simpson for their helpful comments during preparation of this manuscript. We also gratefully acknowledge the assistance of Ludmil B. Alexandrov Mi Ni Huang, Arnoud Boot, Steven G. Rozen and Michael R. Stratton in providing the set of independent WGS SNV calls used for validation. WJ and LS are supported by funding from the Province of Ontario, Canada.

Literature Cited

1. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MD, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, Ding L. Mutational landscape and significance across 12 major cancer types. Nature 502:333-9 (2013).
2. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortés ML, Auclair D, Saksena G, Voet D, Noble M, DiCara D, Lin P, Lichtenstein L, Heiman DI, Fennell T, Imielinski M, Hernandez B, Hodis E, Baca S, Dulak AM, Lohr J, Landau DA, Wu CJ, Melendez-Zajgla J, Hidalgo-Miranda A, Koren A, McCarroll SA, Mora J, Lee RS, Crompton B, Onofrio R, Parkin M, Winckler W, Ardlie K, Gabriel SB, Roberts CW, Biegel JA, Stegmaier K, Bass AJ, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499:214-8 (2013).
3. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nat Genet. 45:1127-33 (2013).
4. Campbell P, Getz G, Korbel J, Stuart J, Stein L and PCAWG Network. Pan-cancer analysis of whole genomes. https://doi.org/10.1101/162784 (2017).
5. Patch AM, Christie EL, Etemadmoghadam D, Garsed DW, George J, Fereday S, Nones K, Cowin P, Alsop K, Bailey PJ, Kassahn KS, Newell F, Quinn MC, Kazakoff S, Quek K, Wilhelm-Benartzi C, Curry E, Leong HS; Australian Ovarian Cancer Study Group., Hamilton A, Mileshkin L, Au-Yeung G, Kennedy C, Hung J, Chiew YE, Harnett P, Friedlander M, Quinn M, Pyman J, Cordner S, O'Brien P, Leditschke J, Young G, Strachan K, Waring P, Azar W, Mitchell C, Traficante N, Hendley J, Thorne H, Shackleton M, Miller DK, Arnau GM, Tothill RW, Holloway TP, Semple T, Harliwong I, Nourse C, Nourbakhsh E, Manning S, Idrisoglu S, Bruxner TJ, Christ AN, Poudel B, Holmes O, Anderson M, Leonard C, Lonie A, Hall N, Wood S, Taylor DF, Xu Q, Fink JL, Waddell N, Drapkin R, Stronach E, Gabra H, Brown R, Jewell A, Nagaraj SH, Markham E, Wilson PJ, Ellul J, McNally O, Doyle

MA, Vedururu R, Stewart C, Lengyel E, Pearson JV, Waddell N, deFazio A, Grimmond SM, Bowtell DD. Whole-genome characterization of chemoresistant ovarian cancer. Nature 521:489-94 (2015).

6. Kurzrock R, Kantarjian H M, Druker BJ, Talpaz M. Philadelphia chromosome-positive leukemias: From basic mechanisms to molecular therapeutics. Ann. Int. Med. 138: 819–30. (2003).
7. Hayward NK, Wilmott JS, Waddell N, Johansson PA, Field MA, Nones K, Patch AM, Kakavand H, Alexandrov LB, Burke H, Jakrot V, Kazakoff S, Holmes O, Leonard C, Sabarinathan R, Mularoni L, Wood S, Xu Q, Waddell N, Tembe V, Pupo GM, De Paoli-Iseppi R, Vilain RE, Shang P, Lau LMS, Dagg RA, Schramm SJ, Pritchard A, Dutton-Regester K, Newell F, Fitzgerald A, Shang CA, Grimmond SM, Pickett HA, Yang JY, Stretch JR, Behren A, Kefford RF, Hersey P, Long GV, Cebon J, Shackleton M, Spillane AJ, Saw RPM, López-Bigas N, Pearson JV, Thompson JF, Scolyer RA, Mann GJ. Whole-genome landscapes of major melanoma subtypes. Nature 545:175-80 (2017).
8. Biankin AV, Waddell N, Kassahn KS, Gingras MC, Muthuswamy LB, Johns AL, Miller DK, Wilson PJ, Patch AM, Wu J, Chang DK, Cowley MJ, Gardiner BB, Song S, Harliwong I, Idrisoglu S, Nourse C, Nourbakhsh E, Manning S, Wani S, Gongora M, Pajic M, Scarlett CJ, Gill AJ, Pinho AV, Rooman I, Anderson M, Holmes O, Leonard C, Taylor D, Wood S, Xu Q, Nones K, Fink JL, Christ A, Bruxner T, Cloonan N, Kolle G, Newell F, Pinese M, Mead RS, Humphris JL, Kaplan W, Jones MD, Colvin EK, Nagrial AM, Humphrey ES, Chou A, Chin VT, Chantrill LA, Mawson A, Samra JS, Kench JG, Lovell JA, Daly RJ, Merrett ND, Toon C, Epari K, Nguyen NQ, Barbour A, Zeps N; Australian Pancreatic Cancer Genome Initiative., Kakkar N, Zhao F, Wu YQ, Wang M, Muzny DM, Fisher WE, Brunicardi FC, Hodges SE, Reid JG, Drummond J, Chang K, Han Y, Lewis LR, Dinh H, Buhay CJ, Beck T, Timms L, Sam M, Begley K, Brown A, Pai D, Panchal A, Buchner N, De Borja R, Denroche RE, Yung CK, Serra S, Onetto N, Mukhopadhyay D, Tsao MS, Shaw PA, Petersen GM, Gallinger S, Hruban RH, Maitra A, Iacobuzio-Donahue CA, Schulick RD, Wolfgang CL, Morgan RA, Lawlor RT, Capelli P, Corbo V, Scardoni M, Tortora G, Tempero MA, Mann KM, Jenkins NA, Perez-Mancera PA, Adams DJ, Largaespada DA, Wessels LF, Rust AG, Stein LD, Tuveson DA, Copeland NG, Musgrove EA, Scarpa A, Eshleman JR, Hudson TJ, Sutherland RL, Wheeler DA, Pearson JV, McPherson JD, Gibbs RA, Grimmond SM. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature 491:399-405 (2012).
9. Greco FA. Molecular diagnosis of the tissue of origin in cancer of unknown primary site: useful in patient management. Curr Treat Options Oncol.14:634-42 (2013).
10. Pavlidis N, Khaled H, Gaafar R. A mini review on cancer of unknown primary site: A clinical puzzle for the oncologists. J Adv Res. 6:375-82 (2015)
11. Tsokos M. Peripheral primitive neuroectodermal tumours. Diagnosis, classification, and prognosis. Perspect. Pediatr. Pathol. 16:27–98 (1992).
12. D'cruze L, Dutta R, Rao S, R A, Varadarajan S, Kuruvilla S. The role of immunohistochemistry in the analysis of the spectrum of small round cell tumours at a tertiary care centre. J Clin Diagn Res. 7:1377-82 (2013).
13. Jain P, and O'Brien S. Richter's transformation in chronic lymphocytic leukemia. Oncology 26:1146-52 (2012).

14. Tefferi A, Vainchenker W. Myeloproliferative neoplasms: molecular pathophysiology, essential clinical understanding, and treatment strategies. J Clin Oncol. 29:573-82 (2011).
15. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, Reynolds A, Rynes E, Vlahoviček K, Stamatoyannopoulos JA, Sunyaev SR. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature 518:360-4. (2015).
16. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjörd JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Jäger N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, López-Otín C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdés-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR; Australian Pancreatic Cancer Genome Initiative.; ICGC Breast Cancer Consortium.; ICGC MMML-Seq Consortium.; ICGC PedBrain., Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR. Signatures of mutational processes in human cancer. Nature. 500:415-21 (2013).
17. Pollock PM, Meltzer PS. A genome-based strategy uncovers frequent BRAF mutations in melanoma. Cancer Cell. 2:5-7 (2002).
18. Jones DT, Kocialkowski S, Liu L, Pearson DM, Backlund LM, Ichimura K, Collins VP. Tandem duplication producing a novel oncogenic BRAF fusion gene defines the majority of pilocytic astrocytomas. Cancer Res. 2008 Nov 1;68(21):8673-7.
19. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011 Jun 29;474(7353):609-15.
20. Sasaki H, Hikosaka Y, Okuda K, Kawano O, Moriyama S, Yano M, Fujii Y. NFE2L2 gene mutation in male Japanese squamous cell carcinoma of the lung. J Thorac Oncol. 2010 Jun;5(6):786-9.
21. Cherniack AD, Shen H, Walter V, Stewart C, Murray BA, Bowlby R, Hu X, Ling S, Soslow RA, Broaddus RR, Zuna RE, Robertson G, Laird PW, Kucherlapati R, Mills GB, Weinstein JN, Zhang J, Akbani R, Levine DA. Integrated Molecular Characterization of Uterine Carcinosarcoma. Cancer Cell 2017 Mar 13;31(3):411-423.
22. Rüschoff J. Adenocarcinoma of the GEJ: gastric or oesophageal cancer? Recent Results Cancer Res. 196:107-13 (2012).
23. Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J, Shi B. Breast cancer intrinsic subtype classification, clinical use and future trends. Am J Cancer Res. 2015 Sep 15;5(10):2929-43.
24. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 490:61-70 (2012)
25. Nielsen TO, Parker JS, Leung S, Voduc D, Ebbert M, Vickery T, Davies SR, Snider J, Stijleman IJ, Reed J, Cheang MC, Mardis ER, Perou CM, Bernard PS, Ellis MJ. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. Clin Cancer Res. 2010 Nov 1;16(21):5222-32.

26. Wu G, Diaz AK, Paugh BS, Rankin SL, Ju B, Li Y, Zhu X, Qu C, Chen X, Zhang J, Easton J, Edmonson M, Ma X, Lu C, Nagahawatte P, Hedlund E, Rusch M, Pounds S, Lin T, Onar-Thomas A, Huether R, Kriwacki R, Parker M, Gupta P, Becksfort J, Wei L, Mulder HL, Boggs K, Vadodaria B, Yergeau D, Russell JC, Ochoa K, Fulton RS, Fulton LL, Jones C, Boop FA, Broniscer A, Wetmore C, Gajjar A, Ding L, Mardis ER, Wilson RK, Taylor MR, Downing JR, Ellison DW, Zhang J, Baker SJ. The genomic landscape of diffuse intrinsic pontine glioma and pediatric non-brainstem high-grade glioma. Nat Genet. 2014 May;46(5):444-450.
27. Zhang J, Wu G, Miller CP, Tatevossian RG, Dalton JD, Tang B, Orisme W, Punchihewa C, Parker M, Qaddoumi I, Boop FA, Lu C, Kandoth C, Ding L, Lee R, Huether R, Chen X, Hedlund E, Nagahawatte P, Rusch M, Boggs K, Cheng J, Becksfort J, Ma J, Song G, Li Y, Wei L, Wang J, Shurtleff S, Easton J, Zhao D, Fulton RS, Fulton LL, Dooling DJ, Vadodaria B, Mulder HL, Tang C, Ochoa K, Mullighan CG, Gajjar A, Kriwacki R, Sheer D, Gilbertson RJ, Mardis ER, Wilson RK, Downing JR, Baker SJ, Ellison DW. Whole-genome sequencing identifies genetic alterations in pediatric low-grade gliomas. Nat Genet. 2013 Jun;45(6):602-12.
28. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, Morozova O, Newton Y, Radenbaugh A, Pagnotta SM, Anjum S, Wang J, Manyam G, Zoppoli P, Ling S, Rao AA, Grifford M, Cherniack AD, Zhang H, Poisson L, Carlotti CG Jr, Tirapelli DP, Rao A, Mikkelsen T, Lau CC, Yung WK, Rabadan R, Huse J, Brat DJ, Lehman NL, Barnholtz-Sloan JS, Zheng S, Hess K, Rao G, Meyerson M, Beroukhim R, Cooper L, Akbani R, Wrensch M, Haussler D, Aldape KD, Laird PW, Gutmann DH; TCGA Research Network., Noushmehr H, Iavarone A, Verhaak RG. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell. 164:550-63 (2016).
29. Tothill RW, Li J, Mileshkin L, Doig K, Siganakis T, Cowin P, Fellowes A, Semple T, Fox S, Byron K, Kowalczyk A, Thomas D, Schofield P, Bowtell DD. Massively-parallel sequencing assists the diagnosis and guided treatment of cancers of unknown primary. J Pathol.231:413-23 (2013).
30. Chen Y, Sun J, Huang LC, Xu H, Zhao Z. Classification of Cancer Primary Sites Using Machine Learning and Somatic Mutations. Biomed Res Int. 2015:491502 (2015).
31. Pavlidis N, Briasoulis E, Hainsworth J, Greco FA. Diagnostic and therapeutic management of cancer of an unknown primary. Eur J Cancer. 39:1990-2005 (2003).
32. [26] Ferracin M, Pedriali M, Veronese A, Zagatti B, Gafà R, Magri E, Lunardi M, Munerato G, Querzoli G, Maestri I, Ulazzi L, Nenci I, Croce CM, Lanza G, Querzoli P, Negrini M. MicroRNA profiling for the identification of cancers with unknown primary tissue-of-origin. J Pathol. 225:43-53 (2011).
33. Greco FA. Molecular diagnosis of the tissue of origin in cancer of unknown primary site: useful in patient management. Curr Treat Options Oncol.14:634-42 (2013).
34. Ferracin M, Pedriali M, Veronese A, Zagatti B, Gafà R, Magri E, Lunardi M, Munerato G, Querzoli G, Maestri I, Ulazzi L, Nenci I, Croce CM, Lanza G, Querzoli P, Negrini M. MicroRNA profiling for the identification of cancers with unknown primary tissue-of-origin. J Pathol. 225:43-53 (2011).
35. Bridgewater J, van Laar R, Floore A, Van'T Veer L. Gene expression profiling may improve diagnosis in patients with carcinoma of unknown primary. Br J Cancer. 98:1425-30 (2008).

36. Hainsworth JD, Rubin MS, Spigel DR, Boccia RV, Raby S, Quinn R, Greco FA. Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: a prospective trial of the Sarah Cannon research institute. J Clin Oncol. 31:217-23 (2013).
37. D'cruze L, Dutta R, Rao S, R A, Varadarajan S, Kuruvilla S. The role of immunohistochemistry in the analysis of the spectrum of small round cell tumours at a tertiary care centre. J Clin Diagn Res. 7:1377-82 (2013).
38. Bahrami A, Truong LD, Ro JY. Undifferentiated tumor: true identity by immunohistochemistry. Arch Pathol Lab Med. 2008 Mar;132(3):326-48.
39. Chu D, Park BH. Liquid biopsy: unlocking the potentials of cell-free DNA. Virchows Arch. May 2 (2017).
40. Han X, Wang J, Sun Y. Circulating tumour DNA as Biomarkers for Cancer Detection. Genomics Proteomics Bioinformatics. 15:59-72 (2017).
41. David Croft, Gavin O'Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, Steven Jupe, Irina Kalatskaya, Shahana Mahajan, Bruce May, Nelson Ndegwa, Esther Schmidt, Veronica Shamovsky, Christina Yung, Ewan Birney, Henning Hermjakob, Peter D'Eustachio, and Lincoln Stein. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Research. 39:D691-D697. (2011)

Figures

Figure 1. Comparison of tumour type classifiers using single and multiple feature types. Each cell represents the accuracy (F1 score) of an individual classifier for the task for distinguishing a particular tumor type from all other tumor types in the testing set. Columns 1-7 represent the accuracy for single feature types. The rightmost column indicates the accuracy of the best model constructed using multiple feature types. Rows are sorted using hierarchical clustering to group together tumour types that have a similar distribution of single-feature classifier accuracies.

Figure 2. Confusion matrix displaying the accuracy of the merged classifier using a held out portion of the PCAWG data set for evaluation. Each row corresponds to the true tumor type; Columns correspond to the predictions emitted by each of the classifiers. Cells are labeled with the proportion of tumors of a particular type that were called by each type-specific classifier (accuracy). The sensitivity and precision of each classifier is shown in the color bars at the top and left sides of the matrix. All values represent the mean of 10 runs with randomly selected testing data sets.

Figure 3. Frequency with which the correct tumour type was contained within the top X predictions. Using a held-out portion of the PCAWG data set, we calculated how frequently the corrected tumour type was present among the top ranked X predictions.

Figure 4. Confusion matrix displaying the accuracy of the merged classifier on an independent validation set. Each row corresponds to the true tumor type; Columns correspond to the predictions emitted by each of the classifiers. Cells are labeled with the proportion of tumors of a particular type that were called by each type-specific classifier. The sensitivity and precision of each classifier is shown in the color bars at the top and left side of the matrix. Note that the validation data set did not contain any representatives of the nine tumour types to the right of CNS-PiloAstro.

Tables

Table 1: Distribution of tumour types in the PCAWG training and test data sets

Abbreviation	Organ system	Tumor Type	Tumors	Merged Abbreviation	Samples
Liver-HCC	LIVER	Liver hepatocellular carcinoma	326	Liver-HCC	326
Prost-AdenoCA	PROSTATE GLAND	Prostate adenocarcinoma	286	Prost-AdenoCA	286
Panc-AdenoCA	PANCREAS	Pancreatic adenocarcinoma	241	Panc-AdenoCA	241
Breast-AdenoCA	BREAST	Breast adenocarcinoma	198	Breast-AdenoCA	198
CNS-Medullo	BRAIN, & CRANIAL NERVES, & SPINAL CORD,	Medulloblastoma	146	CNS-Medullo	146
Kidney-RCC	KIDNEY	Renal cell carcinoma (proximal tubules)	144	Kidney-RCC	144
Ovary-AdenoCA	OVARY	Ovarian adenocarcinoma	113	Ovary-AdenoCA	113
Skin-Melanoma	SKIN	Skin melanoma	107	Skin-Melanoma	107
Lymph-BNHL	LYMPH NODES	Mature B-cell lymphoma	105	Lymphoid	200
Lymph-CLL	BLOOD, BONE MARROW, & HEMATOPOIETIC SYS	Chronic lymphocytic leukemia	95		
Eso-AdenoCA	ESOPHAGUS	Esophageal adenocarcinoma	98	Eso-AdenoCA	98
CNS-PiloAstro	BRAIN, & CRANIAL NERVES, & SPINAL CORD,	Pilocytic astrocytoma	89	CNS-PiloAstro	89
Panc-Endocrine	PANCREAS	Pancreatic neuroendocrine tumor	85	Panc-Endocrine	85
Stomach-AdenoCA	STOMACH	Gastric adenocarcinoma	75	Stomach-AdenoCA	75
ColoRect-AdenoCA	LARGE INTESTINE, (EXCL. APPENDIX)	Colorectal adenocarcinoma	60	ColoRect-AdenoCA	60
Head-SCC	GUM, FLOOR OF MOUTH, & OTHER MOUTH	Head/neck squamous cell carcinoma	57	Head-SCC	57
Uterus-AdenoCA	UTERUS, NOS	Uterine adenocarcinoma	51	Uterus-AdenoCA	51
Lung-SCC	LUNG & BRONCHUS	Lung squamous cell carcinoma	48	Lung-SCC	48
Thy-AdenoCA	THYROID GLAND	Thyroid	48	Thy-AdenoCA	48

		adenocarcinoma			
Kidney-ChRCC	KIDNEY	Renal cell carcinoma (distal tubules)	45	Kidney-ChRCC	45
Bone-Osteosarc	BONES & JOINTS	Sarcoma, bone	44	Bone-Osteosarc	44
CNS-GBM	BRAIN, & CRANIAL NERVES, & SPINAL CORD,	Diffuse glioma	41	CNS-GBM	41
Lung-AdenoCA	LUNG & BRONCHUS	Lung adenocarcinoma	38	Lung-AdenoCA	38
Myeloid-AML	BLOOD, BONE MARROW, & HEMATOPOIETIC SYS	Acute myeloid leukemia	11	Myeloid	
Myeloid-MPN	BLOOD, BONE MARROW, & HEMATOPOIETIC SYS	Myeloproliferative neoplasm	55		66
			2606		**2606**

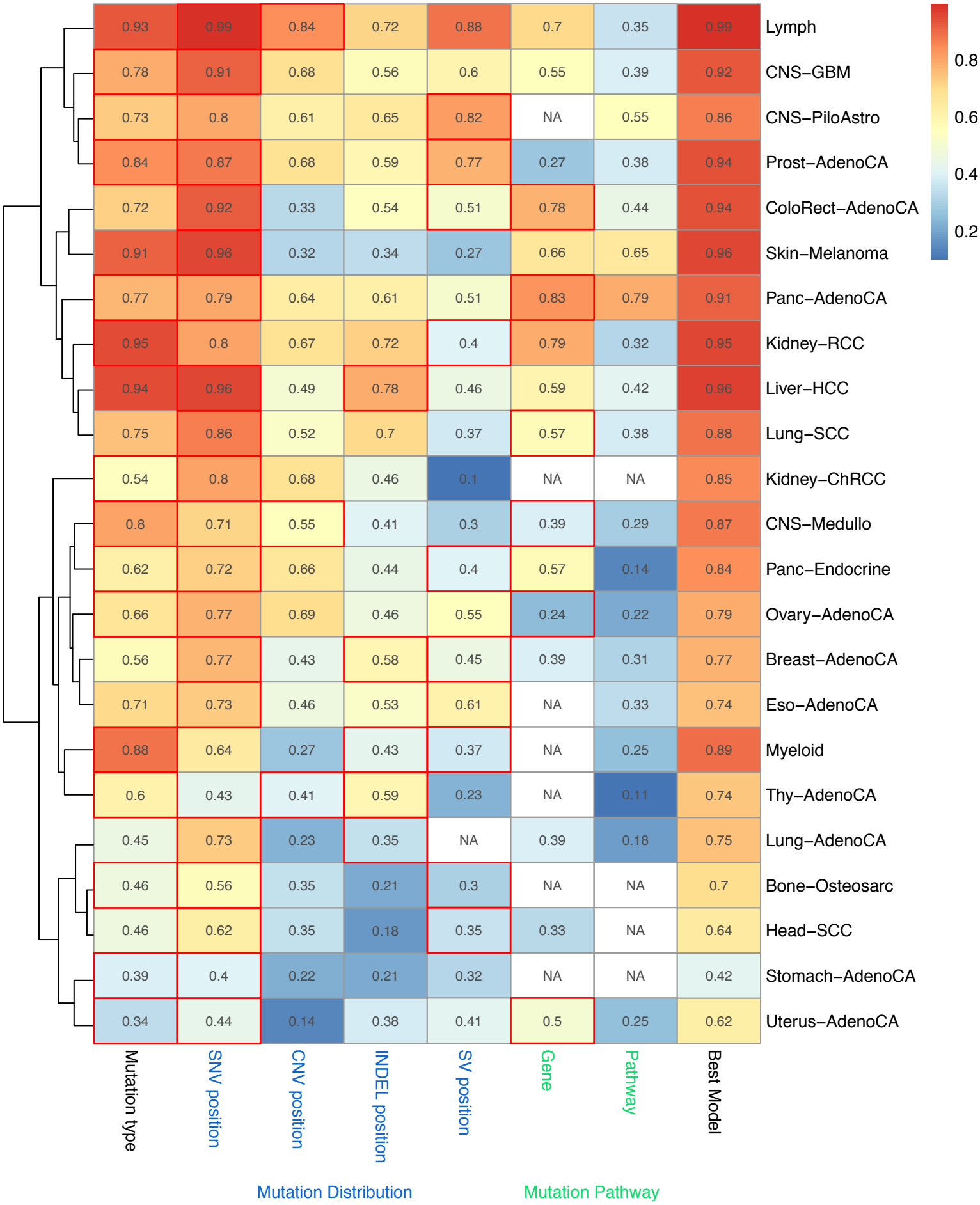
Table 2: WGS feature types used in classifiers

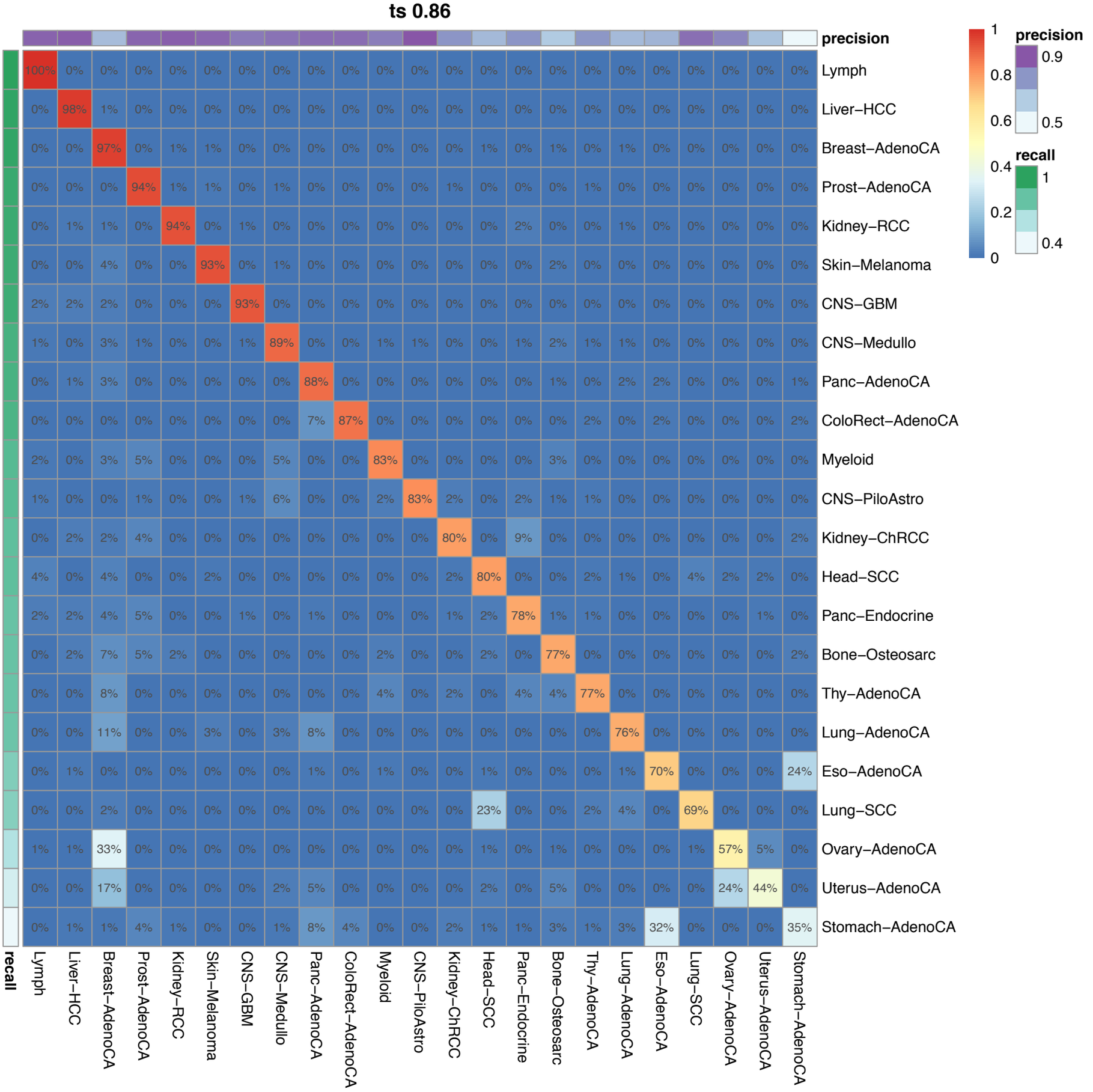
Feature Category	Feature Name	Feature Count	Description
Mutation Distribution	SNV-BIN	2939	Number of SNVs per 1 Mbp bin, and per chromosome,, normalized against total number of SNVs per sample
	CNA-BIN	2826	Number of CNAs per 1 Mbp bin
	SV-BIN	2929	Number of SVs per 1 Mbp bin, and per chromosome,, normalized against total number of SV per sample
	INDEL-BIN	2939	Number of SNVs per 1 Mbp bin, and per chromosome,, normalized against total number of INDEL per sample
Mutation Type	MUT-WGS	150	Type of single nucleotide substitution,, double, and triple nucleotide substitution (plus its adjacent nucleotide neighbors)
Driver Gene/Pathway	GEN	554	Presence of a impactful mutation in a suspected driver gene
	MOD	1865	Presence of an impactul mutation in a gene belonging to a suspected driver pathway

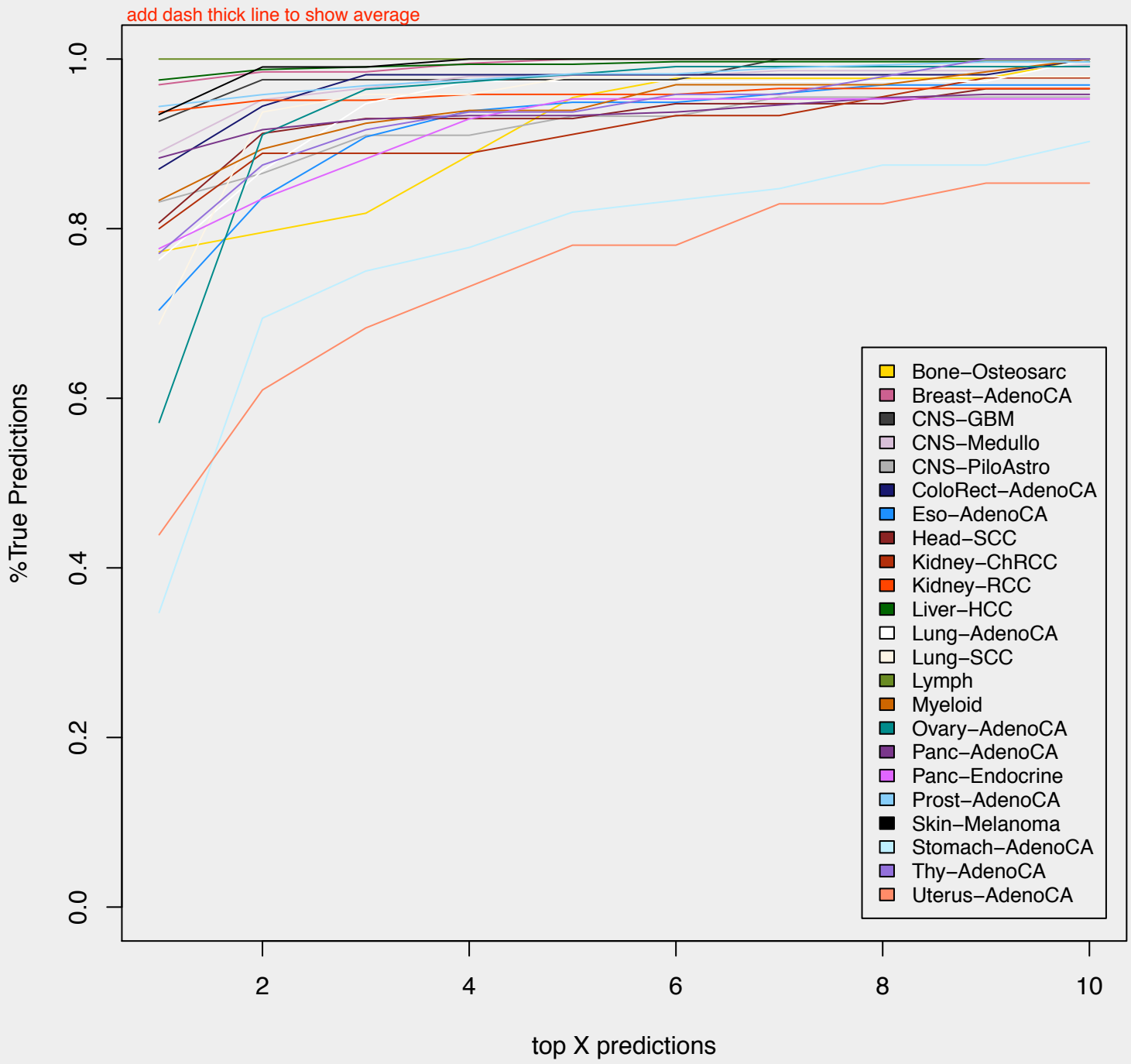
Table 3. Distribution and source of tumour types contained within the validation data set

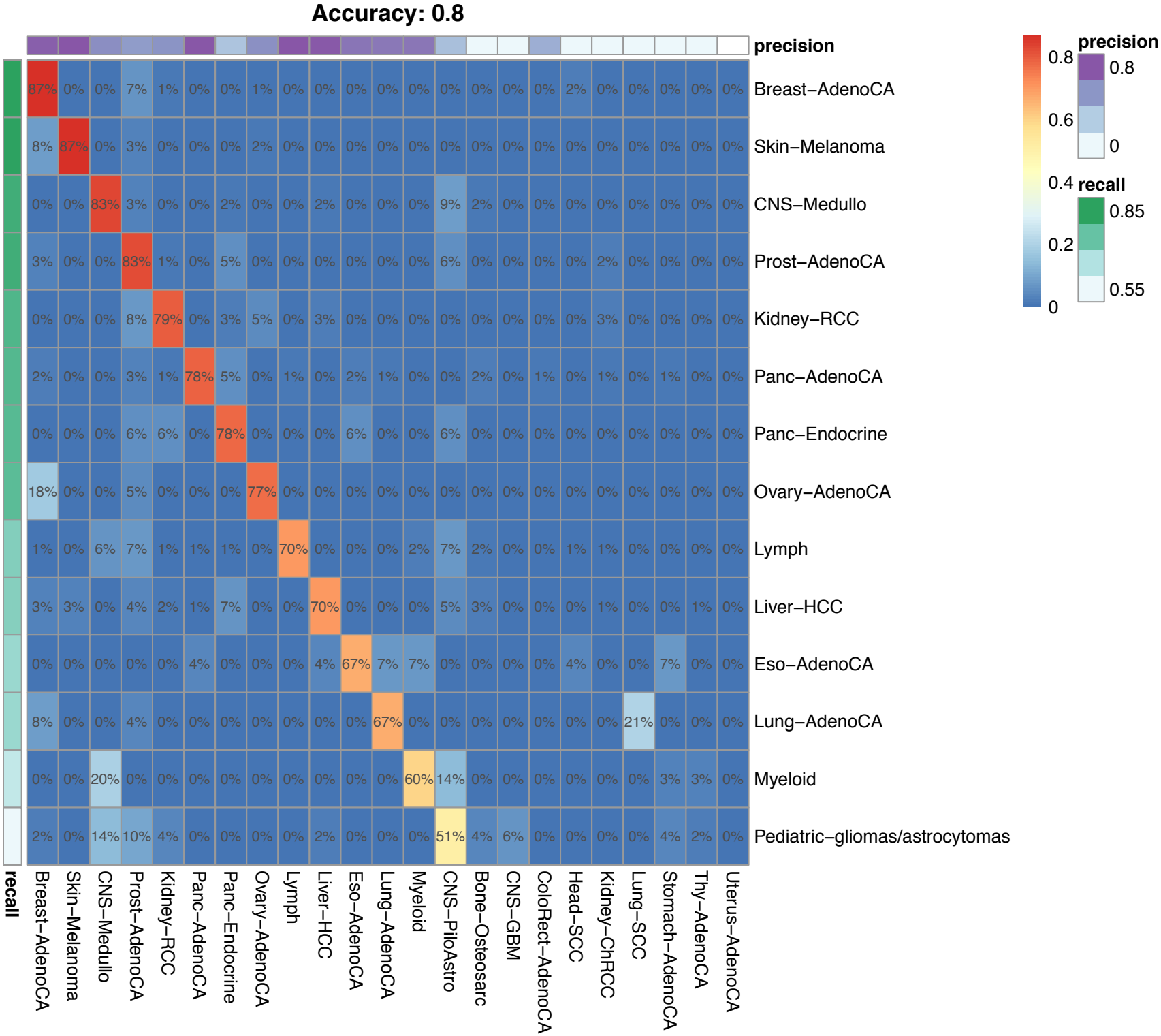
Source	Year	Cancer Type	Genome Version	#Samples
doi:10.1038/nature08629	2009	Lung-SCC	GRCh38	1
doi:10.1038/nature08658	2009	Skin-Melanoma	GRCh38	1

doi:10.1038/nature09744	2011	Prost-AdenoCA	GRCh38	8
doi:10.1016/j.cell.2012.06.023	2012	Myeloid	GRCh38	13
doi:10.1016/j.cell.2012.08.029	2012	Lung-AdenoCA	GRCh37	24
doi:10.1038/nature10738	2012	Myeloid	GRCh38	11
doi:10.1038/nature11213	2012	CNS-Medullo	GRCh38	11
doi:10.1038/ng.2468	2012	Lymphoid	GRCh37	1
doi:10.1056/NEJMoa1106968	2012	Myeloid	GRCh38	9
doi:10.1038/nature12477	2013	Myeloid, Breast-AdenoCA, CNS-Medullo	GRCh37	130
doi:10.1038/ng.2611	2013	CNS-GBM	GRCh38	34
doi:10.1038/ng.2699	2013	Kidney-RCC	GRCh38	14
doi:10.1073/pnas.1314608110	2013	Lymphoid	GRCh38	4
doi:10.1101/gr.154492.113	2013	Liver-HCC	GRCh37	78
doi:10.1038/ng.2938	2014	CNS-GBM	GRCh38	33
doi:10.1056/NEJMoa1403088	2014	Lymphoid	GRCh38	32
doi:10.1186/1476-4598-13-141	2014	ColoRect-AdenoCA	GRCh38	1
doi:10.1038/nature17676	2016	Breast-AdenoCA	GRCh37	455
doi:10.1038/ng.3547	2016	Liver-HCC	GRCh37	1
ICGC (https://dcc.icgc.org/)	Release 25, 8Jun2017	Bone-Osteosarc, Lymphoid, ColoRect-AdenoCA, Eso-AdenoCA, Liver-HCC, Skin-Melanoma, Ovary-AdenoCA, Panc-AdenoCA, Panc-Endocrine, Prost-AdenoCA, Kidney-RCC	GRCh37	658
COSMIC	v82, 3Aug2017	CNS-Medullo, Myeloid, Lymphoid, Kidney-RCC Liver-HCC, Eso-AdenoCA, Panc-AdenoCA, Prost-AdenoCA	GRCh38	81


Lymph 0.93 0.99 0.84 0.72 0.88 0.7 0.35 0.99
CNS−GBM 0.78 0.91 0.68 0.56 0.6 0.55 0.39 0.92
CNS−PiloAstro 0.73 0.8 0.61 0.65 0.82 NA 0.55 0.86
Prost−AdenoCA 0.84 0.87 0.68 0.59 0.77 0.27 0.38 0.94
ColoRect−AdenoCA 0.72 0.92 0.33 0.54 0.51 0.78 0.44 0.94
Skin−Melanoma 0.91 0.96 0.32 0.34 0.27 0.66 0.65 0.96
Panc−AdenoCA 0.77 0.79 0.64 0.61 0.51 0.83 0.79 0.91
Kidney−RCC 0.95 0.8 0.67 0.72 0.4 0.79 0.32 0.95
Liver−HCC 0.94 0.96 0.49 0.78 0.46 0.59 0.42 0.96
Lung−SCC 0.75 0.86 0.52 0.7 0.37 0.57 0.38 0.88
Kidney−ChRCC 0.54 0.8 0.68 0.46 0.1 NA NA 0.85
CNS−Medullo 0.8 0.71 0.55 0.41 0.3 0.39 0.29 0.87
Panc−Endocrine 0.62 0.72 0.66 0.44 0.4 0.57 0.14 0.84
Ovary−AdenoCA 0.66 0.77 0.69 0.46 0.55 0.24 0.22 0.79
Breast−AdenoCA 0.56 0.77 0.43 0.58 0.45 0.39 0.31 0.77
Eso−AdenoCA 0.71 0.73 0.46 0.53 0.61 NA 0.33 0.74
Myeloid 0.88 0.64 0.27 0.43 0.37 NA 0.25 0.89
Thy−AdenoCA 0.6 0.43 0.41 0.59 0.23 NA 0.11 0.74
Lung−AdenoCA 0.45 0.73 0.23 0.35 NA 0.39 0.18 0.75
Bone−Osteosarc 0.46 0.56 0.35 0.21 0.3 NA NA 0.7
Head−SCC 0.46 0.62 0.35 0.18 0.35 0.33 NA 0.64
Stomach−AdenoCA 0.39 0.4 0.22 0.21 0.32 NA NA 0.42
Uterus−AdenoCA 0.34 0.44 0.14 0.38 0.41 0.5 0.25 0.62
0.8
0.6
0.4
0.2
Mutation type
SNV position
CNV position
INDEL position
SV position
Gene
Pathway
Best Model
Mutation Distribution
Mutation Pathway


ts 0.86
precision
recall
Lymph
Liver–HCC
Breast–AdenoCA
Prost–AdenoCA
Kidney–RCC
Skin–Melanoma
CNS–GBM
CNS–Medullo
Panc–AdenoCA
ColoRect–AdenoCA
Myeloid
CNS–PiloAstro
Kidney–ChRCC
Head–SCC
Panc–Endocrine
Bone–Osteosarc
Thy–AdenoCA
Lung–AdenoCA
Eso–AdenoCA
Lung–SCC
Ovary–AdenoCA
Uterus–AdenoCA
Stomach–AdenoCA
1
0.8
0.6
0.4
0.2
0
0.9
0.5
1
0.4


add dash thick line to show average
%True Predictions
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
top X predictions
Bone−Osteosarc
Breast−AdenoCA
CNS−GBM
CNS−Medullo
CNS−PiloAstro
ColoRect−AdenoCA
Eso−AdenoCA
Head−SCC
Kidney−ChRCC
Kidney−RCC
Liver−HCC
Lung−AdenoCA
Lung−SCC
Lymph
Myeloid
Ovary−AdenoCA
Panc−AdenoCA
Panc−Endocrine
Prost−AdenoCA
Skin−Melanoma
Stomach−AdenoCA
Thy−AdenoCA
Uterus−AdenoCA


Accuracy: 0.8
precision
recall
Breast−AdenoCA
Skin−Melanoma
CNS−Medullo
Prost−AdenoCA
Kidney−RCC
Panc−AdenoCA
Panc−Endocrine
Ovary−AdenoCA
Lymph
Liver−HCC
Eso−AdenoCA
Lung−AdenoCA
Myeloid
Pediatric−gliomas/astrocytomas
CNS−PiloAstro
Bone−Osteosarc
CNS−GBM
ColoRect−AdenoCA
Head−SCC
Kidney−ChRCC
Lung−SCC
Stomach−AdenoCA
Thy−AdenoCA
Uterus−AdenoCA
0.8
0.6
0.4
0.2
0
0.85
0.55