CANCER

Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases

Philipp Jurmeister[1,2,3]*, Michael Bockmayr[1,4,5]*, Philipp Seegerer[6], Teresa Bockmayr[1], Denise Treue[1], Grégoire Montavon[6], Claudia Vollbrecht[1,3,7], Alexander Arnold[1], Daniel Teichmann[8], Keno Bressem[9], Ulrich Schüller[4,5,10], Maximilian von Laffert[1], Klaus-Robert Müller[6,11,12], David Capper[3,8†], Frederick Klauschen[1,3†]

Head and neck squamous cell carcinoma (HNSC) patients are at risk of suffering from both pulmonary metastases or a second squamous cell carcinoma of the lung (LUSC). Differentiating pulmonary metastases from primary lung cancers is of high clinical importance, but not possible in most cases with current diagnostics. To address this, we performed DNA methylation profiling of primary tumors and trained three different machine learning methods to distinguish metastatic HNSC from primary LUSC. We developed an artificial neural network that correctly classified 96.4% of the cases in a validation cohort of 279 patients with HNSC and LUSC as well as normal lung controls, outperforming support vector machines (95.7%) and random forests (87.8%). Prediction accuracies of more than 99% were achieved for 92.1% (neural network), 90% (support vector machine), and 43% (random forest) of these cases by applying thresholds to the resulting probability scores and excluding samples with low confidence. As independent clinical validation of the approach, we analyzed a series of 51 patients with a history of HNSC and a second lung tumor, demonstrating the correct classifications based on clinicopathological properties. In summary, our approach may facilitate the reliable diagnostic differentiation of pulmonary metastases of HNSC from primary LUSC to guide therapeutic decisions.

INTRODUCTION

The prognosis of patients with head and neck squamous cell carcinoma (HNSC) is mostly limited by the presence of distant metastases that mainly occur in the lung (*1*, *2*). However, these patients are also at high risk of developing a second squamous cell carcinoma of the lung (LUSC), as both cancers share similar epidemiology and risk factors (*3*–*5*). Further compounding the problem, pulmonary metastases of HNSC and primary LUSC regularly share the same histomorphology and immunohistochemical profiles (*6*). Therefore, when a patient with a history of HNSC is diagnosed with a synchronous or metachronous squamous lung tumor, it is in many cases impossible to distinguish metastatic HNSC from a second independent tumor originating in the lung. It is thus difficult to accurately assess the proportion of HNSC metastases within these patients, which is reflected by the broad range reported in the literature from 17 to 63% of HNSC patients where an additional squamous tumor of the lung was determined to be a metastasis (*7*–*9*). However, the differentiation of these two diseases is of central importance to determining optimal treatment strategies and to correctly assessing the prognosis of the individual patient. Whereas distant metastases of HNSC are mostly incurable and patients mainly receive only palliative chemotherapy or radiotherapy, patients with locally limited LUSC normally qualify for potentially curative therapy including lung lobectomy (*10*–*12*).

With recent technical advances in molecular pathology, methods using next-generation sequencing (NGS), RNA sequencing, and proteomics have been proposed to address this diagnostic dilemma. Because mutational profiles show substantial overlap across different cancer types (*13*, *14*), it is likely that only a direct comparison of the mutational profile using NGS of both tumors may help distinguish HNSC metastases from second lung cancers in the event vast overlaps or differences between the two mutational profiles are found (*15*). However, in clinical routine, this comparative approach might fail if the primary tumor is not suitable for molecular analysis (for example, if the tumor tissue sample is used up or unavailable due to external pathology, degradation due to age, or decalcification) or if the NGS panel does not cover mutations that would allow differentiation between a common versus distinct tumor origin. More recently, RNA sequencing– or proteomics-based signatures that are specific for the tissue of origin have been suggested to avoid the need to analyze and directly compare the original HNSC and lung tumors (*7*, *16*, *17*). However, with accuracy rates below 90%, these methods are of limited utility for implementation in routine diagnostics.

The DNA methylation signatures of different tissue types are known to be quite specific, which has resulted in the recent development of promising algorithms that can characterize cancers of unknown

[1]Institute of Pathology, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, 10117 Berlin, Germany. [2]Charité Comprehensive Cancer Center, 10117 Berlin, Germany. [3]German Cancer Consortium (DKTK), Partner Site Berlin, and German Cancer Research Center (DKFZ), 69210 Heidelberg, Germany. [4]Department of Pediatric Hematology and Oncology, University Medical Center Hamburg-Eppendorf, 20251 Hamburg, Germany. [5]Research Institute Children's Cancer Center Hamburg, 20251 Hamburg, Germany. [6]Machine-Learning Group, Department of Software Engineering and Theoretical Computer Science, Technical University of Berlin, 10623 Berlin, Germany. [7]German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany [8]Department of Neuropathology, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, 10117 Berlin, Germany. [9]Department of Radiology, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, 10117 Berlin, Germany. [10]Institute of Neuropathology, University Medical Center Hamburg-Eppendorf, 20251 Hamburg, Germany. [11]Department of Brain and Cognitive Engineering, Korea University, 136-713 Seoul, South Korea. [12]Max-Planck-Institute for Informatics, 66123 Saarbrücken, Germany.
*These authors contributed equally to this work.
†Corresponding author. Email: frederick.klauschen@charite.de (F.K.); david.capper@charite.de (D.C.)

primary tumors, brain tumors, and sarcomas according to their epigenetic signatures (*18*–*20*). On the basis of these results, we aimed to develop a DNA methylation–based machine learning classifier that facilitates the diagnostic differentiation of HNSC metastases to the lung from primary LUSC.

RESULTS

Clinical cohort

In a retrospective analysis of cases from our institution, we identified 408 patients with a history of primary HNSC and a synchronous or metachronous squamous lung tumor. The results from histopathological evaluation, molecular analyses, as well as clinical and radiological information were considered and discussed by the interdisciplinary tumor board of the Charité Comprehensive Cancer Center. Figure 1A shows two example cases in which the board made a diagnosis based on conventional clinicopathological information. Although the tumors showed similar histomorphology, one case revealed concordant p16 and p53 expression in HNSC and the lung tumor and had multiple peripheral lung tumors, indicating metastatic disease. In the other case, discordant p16 expression was observed and the solitary lung tumor was located at the main bronchus, indicating a primary lung tumor. Despite thorough analysis and discussion by the board, 344 cases (84.3%) remained unsolved due to contradictory information. Consensus regarding the tumor's origin was reached in only 64 cases (15.7%) (Fig. 1B). Thirty-eight tumors (59.4%) were classified as pulmonary metastases of HNSC, and a second LUSC was assumed in 26 cases (40.6%). For 54 of these 64 specimens, a sufficient amount of tumor material was available for further analysis. We excluded three cases from the dataset as the derived DNA methylation data did not pass quality control, resulting in a clinical cohort of 51 samples, the clinical and histopathological details of which are summarized in table S1 (in data file S1). As expected, patients who were diagnosed with HNSC metastases had a significantly

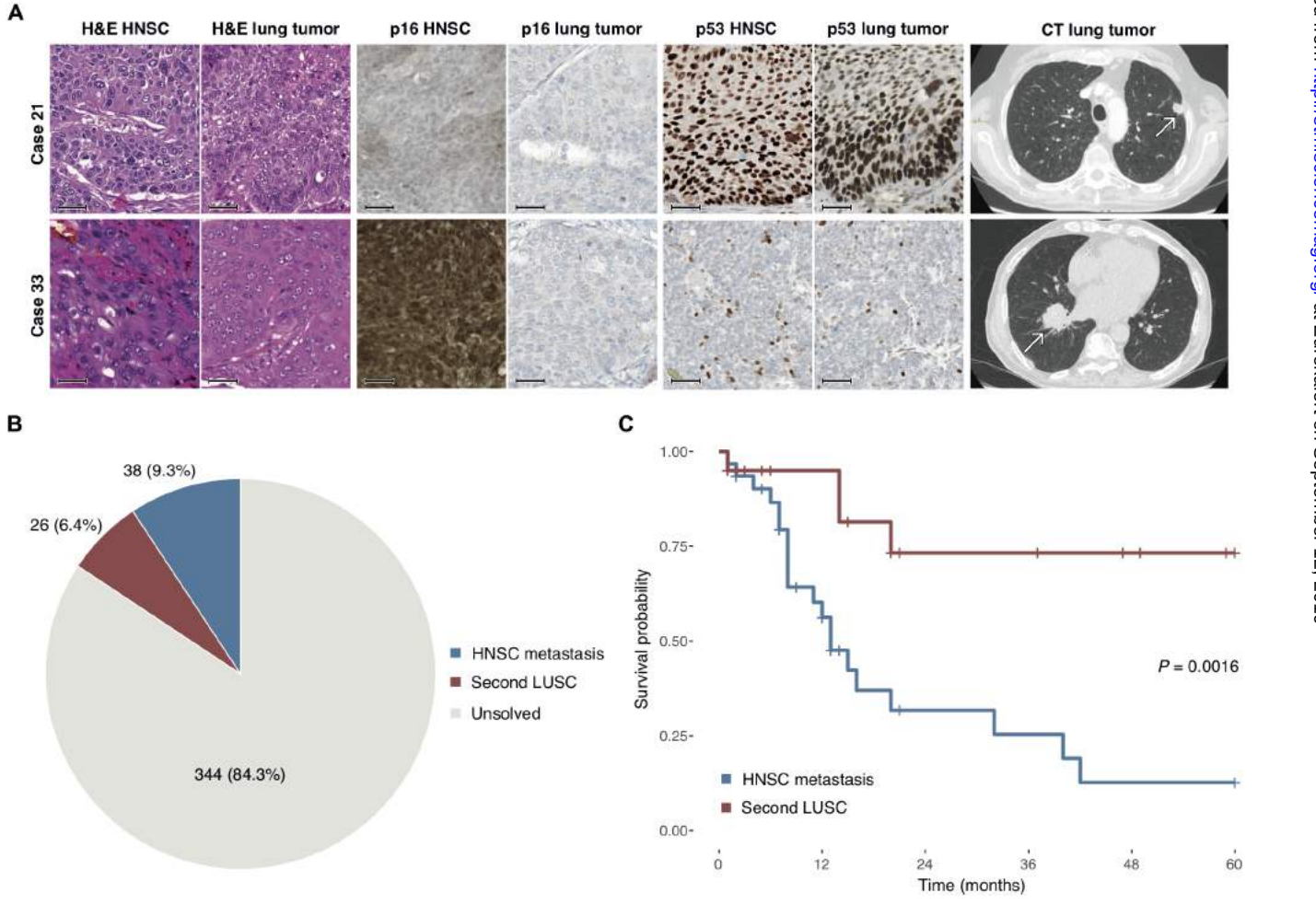

Fig. 1. Retrospective evaluation of patients with HNSC with a synchronous or metachronous squamous lung tumor. (**A**) Two cases of HNSC patients with an additional squamous lung tumor. In case 21, a pulmonary metastasis of HNSC was diagnosed because of similar morphology on hematoxylin and eosin (H&E)–stained slides, concordant immunohistochemistry of p16 (negative) and p53 (mutated pattern), as well as the peripheral localization of the tumor (arrow). Case 33 was diagnosed as a second LUSC because of different p16 expression (HNSC positive, lung tumor negative) and the central localization of the tumor (arrow) with connection to a main bronchus. Scale bars, 50 μm. CT, computed tomography. (**B**) In our retrospective analysis, we identified 408 patients with a history of HNSC and a synchronous or metachronous lung tumor. After reviewing all available clinicopathological data and discussion within the tumor board, 26 (6.4%) were considered as second LUSC and 38 (9.3%) as HNSC metastases. (**C**) Kaplan-Meier plot of 51 cases that were solved on the basis of clinicopathological data and had enough tumor tissue available for further analysis.

shorter disease-specific survival than those with a second LUSC ($P = 0.0016$; Fig. 1C).

Development and comparison of different machine learning classifiers on primary tumor samples

As a first step, we used primary HNSC and LUSC samples to identify characteristic epigenetic signatures of the respective tumor type. This reference cohort ($n = 1071$) consisted of primary HNSC, primary LUSC, and normal lung tissue samples from The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO) databases. By also including normal lung tissue, we provided a control mechanism to recognize samples that were contaminated by adjacent lung stroma. Otherwise, a pulmonary metastasis of HNSC with a low tumor cell content might have falsely been considered as a LUSC.

In a *t*-distributed stochastic neighbor embedding (*t*-SNE) analysis of this reference cohort, HNSC and LUSC samples formed two roughly different groups with considerable overlap and no clear separation (Fig. 2A). We observed two distinct groups for the normal lung tissue specimens, representing mixed lung tissue from the TCGA dataset and alveolar epithelial cells from a GEO dataset. Moreover, two LUSC samples fell into the normal lung tissue group. We found a more pronounced separation of LUSC versus HNSC in cases with higher tumor purity as assessed by the tumor purity estimation method "ESTIMATE" (*21*) based on gene expression data (Fig. 2B), as well as the DNA methylation–based "InfiniumPurify" (*22*) method (Fig. 2C). Additional methods to estimate tumor purity are shown in fig. S1. LUSC cases with low tumor cellularity resembled normal lung tissue, as expected. With regards to human papillomavirus (HPV) status, HPV-positive cases accumulated in a distinct subgroup (Fig. 2D), suggesting that these tumors were epigenetically different from HPV-negative specimens. The samples in the HPV subgroup also had a relatively low incidence of *TP53* mutations (Fig. 2E). Among HNSC and LUSC, we found no subgroups that associated with shorter overall survival (Fig. 2F). Annotation of the HNSC site of origin revealed three relatively distinct subgroups representing tumor originating from the oropharynx, the oral cavity, and the larynx (Fig. 2G). Similar to *t*-SNE, hierarchical clustering revealed that normal lung samples were distinct from the tumor samples, but there was no clear separation of HNSC and LUSC (fig. S2).

For further downstream analysis, we selected the 2000 most variable CpG sites. In a gene enrichment analysis, 128 of 22,132 tested Gene Ontology (GO) categories were significantly enriched in the 2000 top variable CpG sites after multiple testing correction with the Benjamini-Hochberg (BH) method (table S2 in data file S1). The enriched categories included GO terms related to tissue differentiation such as "system development" (p-BH = 5.98×10^{-12}), "embryonic morphogenesis" (p-BH = 9.43×10^{-8}), and "cell differentiation" (p-BH = 0.001). GO terms representing transcriptional factor activity,

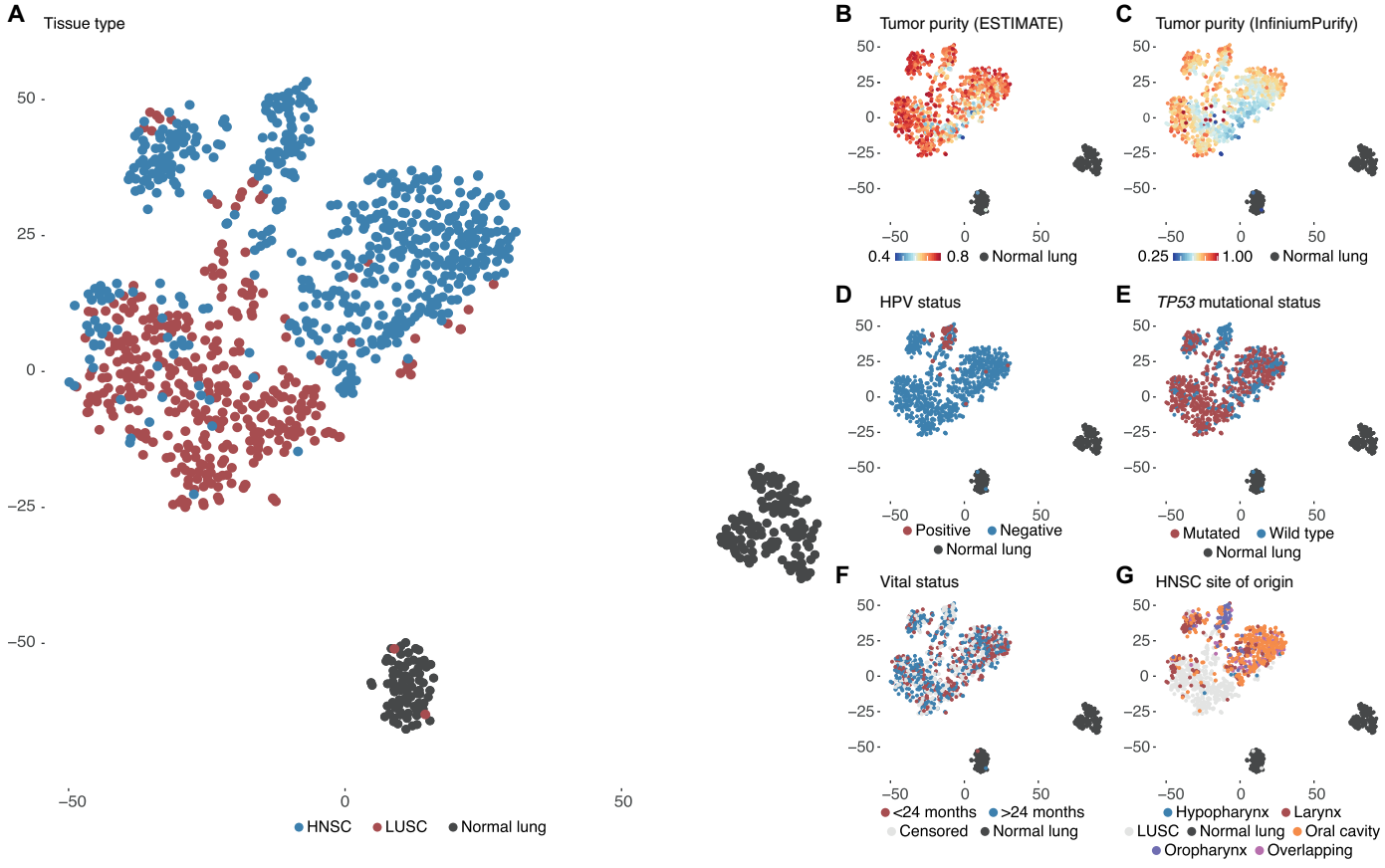

Fig. 2. *t*-SNE plots of the training cohort, consisting of primary HNSC, LUSC, and normal lung tissue. (**A**) General *t*-SNE plot showing the tissue type. (**B** and **C**) *t*-SNE plots showing the association with estimated tumor cell purity using the RNA-based "ESTIMATE" method (B) and the DNA methylation–based "InfiniumPurify" method (C). (**D**) *t*-SNE plot showing the human papillomavirus (HPV) status of the training cohort. (**E**) *t*-SNE plot visualizing the *TP53* mutational status. (**F**) Annotation of survival time reveals no prognostic subgroups. For the survival analysis, patients with an event-free follow-up of less than 24 months were censored. (**G**) *t*-SNE plot showing the distribution of the different HNSC tumor origins.

A

Study	HNSC	LUSC	Normal lung	Material type	Array	Data type	Preprocessing method
GSE56044	0	23	0	Frozen	450k	Raw IDAT files	Noob
GSE79556	83	0	0	FFPE	450k	Raw IDAT files	Noob
GSE95036	11	0	0	FFPE	450k	Raw IDAT files	Noob
GSE66836	0	0	19	Frozen	450k	Raw IDAT files	Noob
GSE39279	0	122	0	Frozen	450k	Preprocessed beta values	Illumina
GSE87053	11	0	0	Frozen	450k	Raw IDAT files	Noob
Charité	5	5	0	FFPE	EPIC	Raw IDAT files	Noob

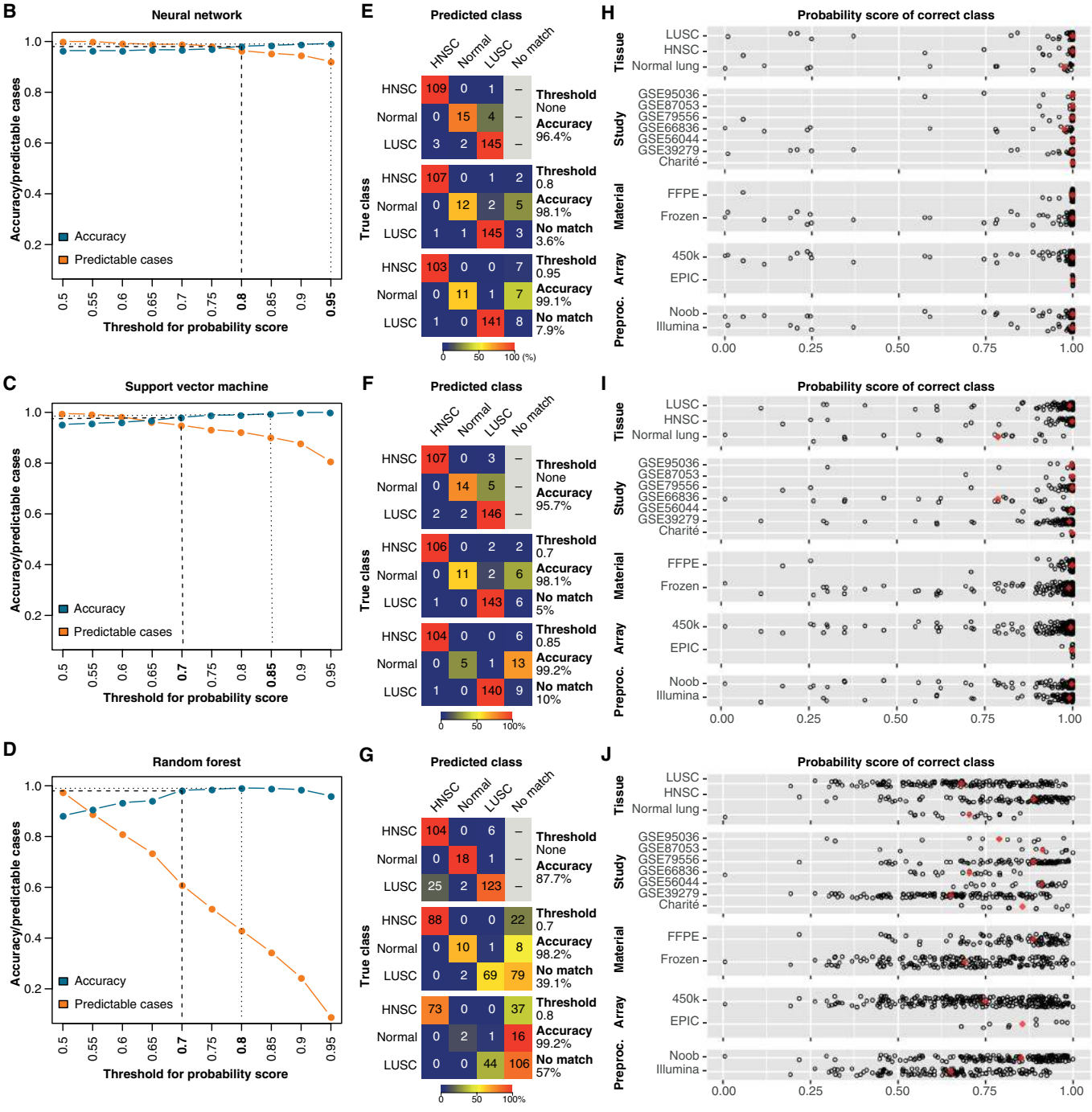

Fig. 3. Classification results of the machine learning algorithms on an independent validation cohort. (**A**) Overview of the samples in the validation cohort ($n = 279$). (**B** to **D**) Threshold analysis for the prediction score of the neural network (B), the support vector machine (C), and the random forest (D). (**E** to **G**) Confusion matrices for the neural network (E), the support vector machine (F), and the random forest (G) comparing the true class with the predicted class for the HNSC ($n = 110$), LUSC ($n = 150$), and normal lung tissue ($n = 19$) samples. (**H** to **J**) Distribution of probability scores (range, 0 to 1) for the correct class for the neural network (H), the support vector machine (I), and the random forest (J) over different tissue types, datasets, material types [formalin-fixed paraffin-embedded (FFPE) versus frozen], DNA methylation chip arrays (450k and EPIC), and preprocessing (Preproc.) methods (Noob versus Illumina).

such as "sequence-specific DNA binding," were also significantly enriched (p-BH = 6.47×10^{-12}).

Using these 2000 CpG sites, we developed three different machine learning classifiers based on artificial neural networks, support vector machines, and random forests. We trained these algorithms to classify tumor samples with respect to their organ of origin based on their DNA methylation profile and provide probability scores of the classification results. All three classifiers were tuned using fivefold cross-validation on the reference cohort. We then applied the resulting models to an independent validation cohort consisting of primary HNSC and LUSC tumor samples from different GEO datasets and our own studies, amounting to 279 samples in total (Fig. 3A). This publicly available validation dataset comprised DNA methylation data from diverse sources, including different material types, DNA methylation array chips, and preprocessing methods.

Classification accuracy

Assigning each sample to the category with the highest probability score, the artificial neural network and the support vector machine correctly classified 96.4 and 95.7% of all cases in the validation cohort, respectively. The random forest fell short of these results, achieving an accuracy rate of only 87.8%. The corresponding three-class areas under the curve (AUCs) were 0.9934, 0.9915, and 0.9708 for the artificial neural network, the support vector machine, and the random forest classifier, respectively. Further, the positive predictive values were 96.4% (HNSC) and 96.7% (LUSC) for the neural network, 98.1% (HNSC) and 94.8% (LUSC) for the support vector machine, and 80.6% (HNSC) and 94.6% (LUSC) for the random forest classifier. To explore whether higher prediction accuracies could be achieved for subsets of cases, we included the confidence of the predictions (probability scores) in our analysis and excluded samples with low scores, which were considered to be unclassifiable ("no match"). By increasing the threshold for the minimally required score, the accuracies could be increased to more than 99% for subsets of samples for all three algorithms. As the distribution of the probability scores was different for each of the three methods, we defined thresholds yielding an accuracy of 98 and 99% if the corresponding samples had a probability score above this threshold calculated by the neural network (Fig. 3B), the support vector machine (Fig. 3C), or the random forest (Fig. 3D). Using probability score cutoffs of 0.8 and 0.95, the accuracy rates of the artificial neural network increased to 98.1 and 99.2%, respectively, while 96.4 and 92.1% of the samples still had probability scores above the threshold (Fig. 3E). The remaining samples were considered not classifiable (no match). For the support vector machine, using thresholds of 0.7 and 0.85, accuracies of 98.1 and 99.2% were achieved for 95.0 and 90.0% of the samples, respectively (Fig. 3F). For the random forest algorithm, with thresholds 0.7 and 0.8, accuracies of 98.2 and 99.2% were reached for 60.9 and 43.0% of all samples, respectively (Fig. 3G). This shows that although high subset prediction accuracies could be achieved for all three methods, the artificial neural network approach resulted in high accuracy predictions for the largest number of cases.

Next, we analyzed the distribution of the probability scores across different tissue types, studies, material types [formalin-fixed paraffin-embedded (FFPE) versus frozen], DNA methylation chip arrays (EPIC versus 450k), and preprocessing methods for the neural network (Fig. 3H), the support vector machine (Fig. 3I), and the random forest (Fig. 3J). The normal lung tissue samples reached slightly lower scores than the tumor samples for all three algorithms. Besides that, the study, material type, DNA methylation chip array, and preprocessing method had no major influence on prediction accuracies for the artificial neural network and the support vector machine, demonstrating that these methods are robust to possible confounding factors. Although the classifiers had only been trained on frozen samples, the FFPE specimens were assigned to the correct class with even higher accuracy than the frozen samples. Unlike the artificial neural

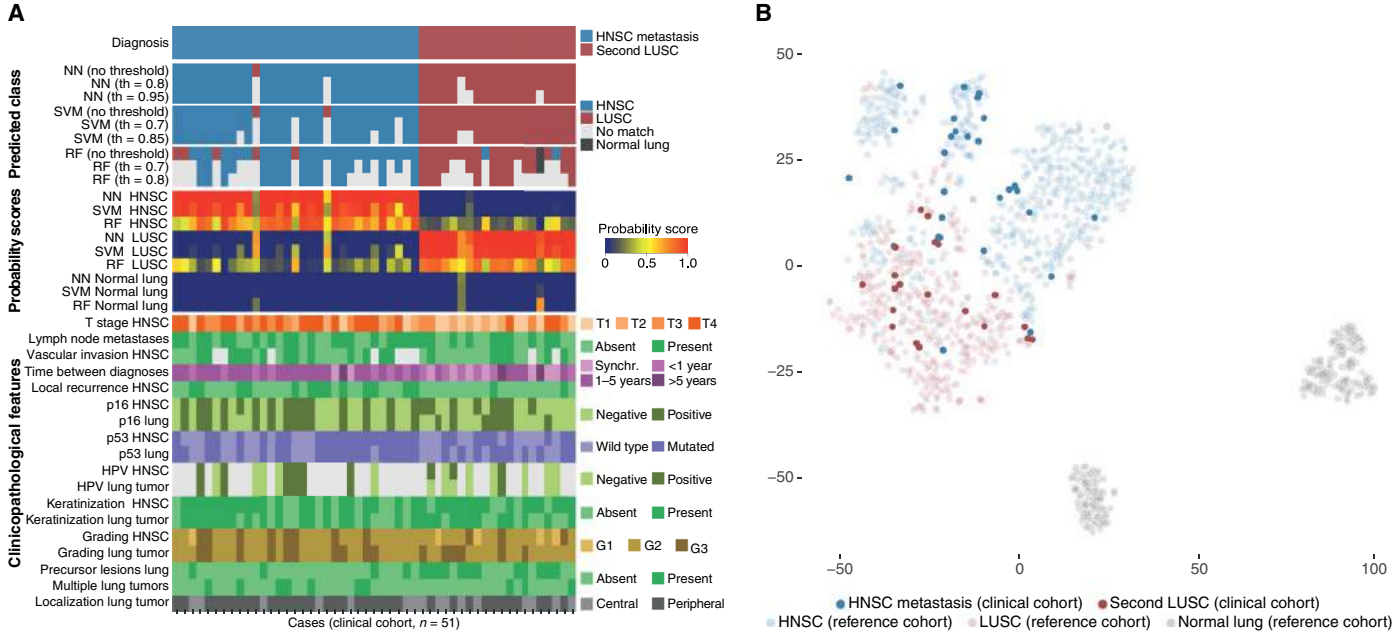

Fig. 4. Classification results in the clinical cohort. (**A**) Heat map of classification results for the neural network (NN), support vector machine (SVM), and random forest (RF) alongside with clinicopathological features that were considered to reach consent by the tumor board. (**B**) *t*-SNE analysis of the samples from the clinical cohort ($n = 51$) and the reference cohort ($n = 1071$) featuring LUSC ($n = 354$), HNSC ($n = 528$), and normal lung samples ($n = 189$).

network and the support vector machine, random forest scores were influenced by the abovementioned factors, which might indicate that this method generalizes less well to data types that differ from those that have been used as a training set.

Validation in an independent clinical cohort

As the three different classifiers were developed and validated in this study on primary tumor samples, we aimed to further verify our results on an additional cohort resembling a clinical setting. As there is no gold standard to differentiate HNSC and LUSC, we selected an independent clinical cohort consisting of 51 suitable patients from the archives of the Charité Institute of Pathology (Fig. 4A). In a *t*-SNE plot including all cases from the reference cohort and the clinical cohort, the clinical validation cases tended to accumulate in the reference cohort groups that were expected based on their clinicopathological annotation (Fig. 4B).

Applying our classifiers to this clinical cohort yielded raw accuracies of 98.0, 96.1, and 84.3% and two-class AUCs (LUSC versus HNSC) of 1.0, 1.0, and 0.976 for the artificial neural network, support vector machine, and random forest, respectively (Fig. 4A). The positive predictive values were 100% (HNSC) and 95.5% (LUSC) for the neural network, 100% (HNSC) and 91.3% (LUSC) for the support vector machine, and 92.6% (HNSC) and 78.2% (LUSC) for the random forest classifier. All misclassified samples had comparably low probability scores (table S1 in data file S1). Considering those samples with prediction scores above the previously defined threshold, the accuracy, positive predictive value, and the AUC were increased to 100% (fig. S3). Again, the artificial neural network achieved the highest prediction accuracy for the largest number of samples. As expected, the classification results were also associated with disease-specific survival, except for the random forest classifier (fig. S4).

Furthermore, we applied our classifier to the four TCGA patients with LUSC that also had a history of HNSC. These samples were not part of the reference or the validation cohort, as we excluded patients with previous malignancies that could metastasize to the lung and represent or mimic squamous differentiation. Two tumors were classified as HNSC metastases with high prediction scores in all three machine learning methods (table S3 in data file S1), suggesting that these tumors might have been misdiagnosed. Our results are supported by the clinical follow-up data of these cases, as the two patients with a predicted HNSC metastasis had a relatively short overall survival despite having a relatively low clinical stage if the tumor was considered a primary LUSC (table S3 in data file S1). In addition, HPV16 was detected in one of these samples, representing the only HPV-positive tumor in the entire TCGA LUSC cohort (*23*), further indicating that this specimen is most likely a pulmonary metastasis of the previously diagnosed HNSC.

DISCUSSION

With this study, we successfully demonstrate that DNA methylation profiling in conjunction with machine learning solves the diagnostic problem of differentiating lung metastases of squamous cell carcinomas of the head and neck from primary lung cancers arising in patients with previous or simultaneous HNSC. In the clinical database of the Charité University Hospital Berlin, we identified a substantial number of patients with HNSC who had a synchronous or metachronous squamous lung tumor. In line with common clinical knowledge, we demonstrated that currently established diagnostic methods such as histomorphology, immunohistochemistry, and thorough clinical and radiological investigation failed to resolve this problem in the majority of cases.

Previous studies that tried to address this issue using mRNA expression or proteomics achieved accuracy rates between 82.9 and 86.8% (*7*, *16*, *17*). In our study, we were able to demonstrate the superiority of DNA methylation profiling in differentiating primary squamous cell carcinomas from the lung region and the head and neck region. We also compared different machine learning techniques to identify the most suitable method to interpret our DNA methylation data. Even our worst-performing model equaled the best accuracy rates of previous studies that were based on proteomics (*7*). Using an artificial neural network, we were able to correctly classify 96.4% of primary tumor samples from an independent and diverse validation cohort. The classification accuracy can be further improved for subsets of patients by using probability scores indicating the confidence of the results provided by the classifier.

This leads to the second major aspect of our study, in which we compared the performance of random forest, support vector machine, and neural network–based classifiers to interpret DNA methylation data. Different methods have previously been used to develop DNA methylation–based diagnostic classifiers, including random forests and prediction analysis of microarrays (PAM) classifiers (*19*, *20*, *24*). The most successful and commonly used classifiers have been based on random forests (*19*). In these previous studies, most *t*-SNE analyses revealed clearly separated groups for each entity or subgroup. However, in our study, unsupervised techniques such as *t*-SNE or hierarchical clustering failed to clearly separate HNSC and LUSC samples. The application of a random forests classifier on our dataset resulted in only mediocre overall accuracy. Furthermore, the use of tail probabilities led to the exclusion of more than 50% of the samples with random forests. Therefore, we also trained artificial neural network and support vector machine classifiers. Overall, the artificial neural network was slightly superior to the support vector machine and both methods largely excel the random forest classifier. In particular, the artificial neural network yielded better results across datasets from different sources, which might be due to batch effects. This advantage is of tremendous importance for application in a diagnostic setting, which requires robust results independent of potentially confounding factors such as interlaboratory variability or other batch effects. Moreover, neural networks can be pretrained on unlabeled data, which might, even more, increase the prediction accuracies in future studies.

A limitation of our work is the fact that DNA methylation analysis is not yet as widely deployed as other molecular methods (for example, NGS), restricting the immediate broad diagnostic utility of our classifier. However, several publications recently demonstrated the emerging importance of this method to correctly characterize different cancers such as brain tumors or sarcomas (*19*, *20*). Therefore, a broader establishment of DNA methylation analysis is probably imminent. Another disadvantage of this technique is the relatively large amount of optimal DNA input (500 ng) recommended by the manufacturer. In theory, this could potentially limit the applicability of this approach with respect to small biopsies. However, it has been shown that DNA methylation analysis of small samples still delivers reliable results (*25*). In line with these reports, we successfully analyzed and classified 15 biopsy specimens with a DNA input as low as 110 ng. Another limitation of DNA methylation analysis is the potential interference with adjacent benign tissue. As evident from our *t*-SNE analysis, low tumor cellularity could potentially complicate the classification

of the tumor origin by causing low probability scores. Tumor purity has already been described as a limiting factor for the classification in brain tumors, leading to a dropout rate of approximately 4% (*19*). However, tumor purity is routinely checked histologically and low probability scores or classification as normal tissue points the pathologist to problematic cases. Last, the system has so far not been trained for the classification of squamous cell carcinomas from other anatomic sites. To what extent our approach can be applied to identify other tumor origins will have to be analyzed in future studies.

Despite these limitations, the method described in this study delivers a so far unmatched accuracy in distinguishing pulmonary metastases of HNSC from primary LUSC. In contrast to DNA methylation–based classifiers for cancers of unknown primary, our approach is specifically tailored to differentiate HNSC from LUSC. By this means, we were able to achieve accuracy rates that could actually qualify this approach for use in routine diagnostics. Furthermore, by using a reference approach, we can avoid the disadvantages of currently established analyses, which rely on a comparative evaluation (for example, immunohistochemistry, HPV genotyping, and NGS) of both tumors. In contrast to our proposed reference-based single analysis, comparative analyses of the HNSC and the lung tumor are more costly and often tissue samples of the original HNSC tumor are not readily available.

We were also able to further validate our results in an independent clinical cohort of patients with a history of HNSC and a synchronous or metachronous squamous lung tumor who underwent diagnosis and treatment at our institution. We successfully tested our classification against the fact that metastatic disease is prognostically unfavorable compared to a second (primary) lung tumor.

A gene set enrichment analysis revealed that the 2000 most variant CpG sites that were chosen for the classifier training were significantly enriched for GO terms related to tissue differentiation. This suggests that specific epigenetic marks acquired during tissue differentiation might be particularly important for the DNA methylation–based classification.

In summary, our study presents a highly accurate and robust machine learning–based DNA methylomics approach capable of differentiating pulmonary metastases of HNSC from primary LUSC. In addition, our study demonstrated that, compared to random forests, support vector machines and particularly neural networks were superior in solving this diagnostic dilemma.

MATERIALS AND METHODS

Study design

Our main research objective was to build a diagnostic classifier able to distinguish pulmonary metastases of HNSC from LUSC based on their DNA methylation profiles. DNA methylation array data were obtained from publicly availably sources and FFPE tissue from our archives. The experimental design was retrospective. We compared three different machine learning algorithms (artificial neural networks, support vector machines, and random forests) to identify the optimal method. All algorithms were trained on a reference cohort of primary tumors and normal lung tissue ($n = 1071$), with 528 HNSC, 354 LUSC, and 74 normal lung tissue samples obtained from TCGA as well as 115 normal tissue samples from the GEO. We then evaluated the classifiers on a validation cohort of 110 primary HNSC, 150 primary LUSC, and 19 normal lung tissue samples obtained from six GEO datasets as well as from the archives of the Institute of Pathology at the Charité University Hospital Berlin. We further validated our method in an independent clinical cohort of 51 patients from our archives with synchronous or metachronous squamous cell lung tumor whose diagnosis could reliably be determined based on clinicopathological and molecular characteristics. No information about the samples in the validation cohort or the clinical cohort was used at any point for the development of the classifiers.

Patients and samples

Cases with a history of HNSC and a synchronous or metachronous squamous lung tumor ($n = 408$) were identified using the electronic patient files and the electronic database of the Charité University Hospital Berlin. All cases were discussed by the interdisciplinary lung or head and neck tumor boards of the Charité Comprehensive Cancer Center, which consists of medical oncologists, radiation oncologists, surgeons, pathologists, and radiologists. For all cases, a variety of clinical, radiological, histopathological, and molecular data were discussed (Fig. 4A and table S1 in data file S1). Consent regarding the lung tumor's origin was reached in 64 of 408 samples.

Matching FFPE tissue available for 54 cases was retrieved from the archives of the Institute of Pathology at the Charité University Hospital Berlin. All samples were reevaluated on the basis of hematoxylin and eosin (H&E) stains, and additional immunohistochemical investigations were performed if needed. Clinical data were extracted from the electronic patient files and the Berlin and Brandenburg Clinical Cancer Registry.

Immunohistochemistry

Immunohistochemical staining was performed on a VENTANA BenchMark XT automated slide stainer according to the manufacturer's instructions. Antibodies, their manufacturers, as well as concentrations and scoring systems are listed in table S4.

DNA extraction

Representative tumor areas were identified using light microscopy. If necessary, macrodissection was performed to reach a tumor cell content of at least 70%. Semi-automated DNA extraction was performed according to the manufacturer's instructions (Maxwell RSC FFPE Plus DNA Purification Kit, Custom, Promega). DNA quantities were measured using Qubit HS DNA assay (Thermo Fisher Scientific).

HPV genotyping

HPV genotyping was performed using the HPV Type 3.5 C LCD array (Chipron) according to the manufacturer's instructions. Data on HPV genotyping in the TCGA cohort were derived from (*23*).

Next-generation sequencing

Ion AmpliSeq Library Kit 2.0 (Thermo Fisher Scientific) was used to perform library preparation with 10 ng of genomic DNA using the Ion AmpliSeq Cancer Hotspot Panel v2 (Thermo Fisher Scientific). The final library was quantified with the Ion Library Quantitation Kit (Thermo Fisher Scientific). Samples were multiplexed and amplified on Ion Spheres Particles with Ion 530 Kit-Chef and were sequenced using Ion 530 Chip (Thermo Fisher Scientific) with an adapted standard protocol using 330 flows (*26*).

Tumor purity estimation

Estimations for the tumor purity were based on gene expression profiles ("ESTIMATE"), somatic copy number data ("ABSOLUTE"), DNA methylation ("LUMP" and "InfiniumPurity"), (*21*, *22*), visual

quantification of H&E slides, as well as a consensus measurement considering all of the mentioned methods ("consensus measurement of purity estimations") (*21*).

DNA methylation analysis

The Infinium HD FFPE DNA Restore Kit was used for DNA restoration of FFPE samples. DNA methylation analysis was performed using the Illumina Infinium MethylationEPIC BeadChip, according to protocols supplied by the manufacturer.

Reference cohort

Raw DNA methylation data (IDAT files) from 528 primary HNSC, 370 primary LUSC, and 74 normal lung tissue samples were obtained from TCGA (*27*–*29*). LUSC cases with a documented previous or synchronous tumor that could potentially result in a pulmonary metastasis of squamous-like differentiation (for example, HNSC, squamous cell skin cancer, and melanoma) were excluded from the dataset (n = 16). Patients with previous malignancies that did not mimic squamous cell carcinomas (such as lymphoma) or that could not metastasize to the lung (for example, basal cell carcinoma) were not excluded. We completed the reference cohort with a dataset of alveolar epithelial cells from 115 patients (GSE85566) (*30*). Only samples from the reference cohort were used for tuning of the classification algorithms and definition of the most variable CpG sites.

Validation cohort

We tested the classification algorithms on DNA methylation data from 105 primary HNSC, 145 primary LUSC, and 19 normal lung tissue samples from additional six different GEO datasets (GSE56044, GSE79556, GSE95036, GSE66836, GSE39279, and GSE87053) (*31*–*36*). For GSE39279, raw IDAT files were not available; therefore, we used the beta values preprocessed with the Illumina method. Furthermore, we also analyzed five primary HNSC and five primary LUSC samples from our own archives.

Clinical cohort

We identified 54 samples from patients with a history of HNSC and a synchronous or metachronous squamous cell lung tumor whose diagnosis could reliably be determined based on histopathological investigation, molecular analyses, and clinical and radiological data. Thirty-four cases were regarded as pulmonary metastases of HNSC and 20 as a second LUSC. These samples were analyzed using the Illumina Infinium MethylationEPIC BeadChip. Three samples were excluded after data generation because of low global quality scores (median detection P value > 0.01). Case 43 was initially suspected to be a pulmonary metastasis because of the detection of HPV16 in the primary HNSC as well as the lung tumor. However, there was a considerably long time interval between the diagnosis of both cancers (121 months) and a discrepancy regarding p53 immunohistochemistry, and the patient was still alive after more than 5 years of follow-up. To resolve this case, we performed comparative NGS. We discovered mutations in the *FGFR3*, *RB1*, and *NOTCH1* gene only in the HNSC sample as well as a *TP53* mutation that was exclusive to the lung tumor (table S5). There was no overlap of pathogenic mutations, thus further supporting the pulmonary origin of this tumor despite the presence of HPV16.

Classifier development

We trained artificial neural networks, a support vector machine, and a random forest classifier on the reference cohort and tuned the parameters using fivefold cross-validation. The validation cohort and the clinical cohort were not used at any point for variable selection or for model selection. To determine the minimal number of required CpG sites (N_{sites}), all models were trained on the top 100, 200, 500, 1000, 2000, 5000, 10,000, and 20,000 most variable CpG sites. Neural networks were built with the R package keras (*37*). The optimal number of hidden layers (search range: 1, 2, 3), neurons per hidden layer (64, 128, 256), the dropout ratio (0.1, 0.3, 0.5), and the nonlinearity (tanh, ReLU) were determined by fivefold cross-validation. The Adam optimizer (*38*) was used with default parameters and a cross-entropy loss. A support vector machine with Gaussian radial basis function (RBF) kernel was used as implemented in the R package kernlab (*39*). Optimal parameters for the variance of the RBF kernel sigma ($1/N_{sites} \times 2^{-5,\ldots,5}$) and the amount of regularization C ($2^{0,\ldots,5}$) were selected using fivefold cross-validation and cross-entropy loss. A random forest classifier with 2000 trees was used as implemented in the R package randomForest (*40*). The inclusion of more trees did not reduce cross-validation loss. The number of features at each split mtry was optimized with fivefold cross-validation (mtry = $\lceil \sqrt{(N_{sites})} \rceil \times 2^{-5,\ldots,5}$), with cross-entropy loss and the sample size sampsize set to (189, 189, 189). All other parameters were kept at the default values.

Selection of final classifier models

With increasing numbers of sites, the cross-validation loss reached a plateau at 2000 CpG sites for all three models (fig. S5). Hence, with 2000 CpG sites, the models had low cross-validation error and comparatively low complexity and were thus less prone to overfitting than more complex models including more CpG sites, in line with machine learning theory (*41*). Therefore, we selected the top 2000 variable CpG sites to train all three classification models.

An artificial neural network with three hidden layers of 64 units and a dropout ratio of 0.3 achieved the lowest cross-validation loss and was subsequently used. The final classifier was an ensemble obtained by averaging the output probabilities of the neural networks trained on the five different cross-validation folds. In our experiments, the ensemble had, on average, a higher accuracy than the individual models on both the validation and clinical cohort. Ensembles are known to reduce the classifier variance and thus the risk for overfitting; this makes them attractive to use specifically in domains where only moderate amounts of data are available (*38*, *41*, *42*). For support vector machines, the optimal parameters were sigma = 6.25×10^{-5} and C = 32. For the random forest classifier, the cross-validation resulted in mtry = 1408. Probability scores were computed for the prediction task. All fitted models are available in the Supplementary Materials.

Rejection class analysis

To increase classification accuracy, we defined a rejection class based on thresholds of minimally required probability scores. The probability scores computed by the classifier provide a measure of the confidence of the classifier results. As the output scores of the classifiers are not calibrated, the distribution of these probabilities in the validation cohort differed between the different algorithms; hence, it was not possible to define a unique probability threshold for all three methods. Therefore, for each threshold in 0.5, 0.55, …, 0.95, we computed the accuracy if we excluded samples with probability scores below the threshold (Fig. 3). This allowed defining method-specific thresholds, which resulted in a prediction accuracy of more than 98 and 99%, respectively, for all samples with scores above these

thresholds. Specimens below the threshold were considered not classifiable (no match).

t-SNE, heat maps, receiver operating characteristics, and survival analysis

All *t*-SNE plots were performed using the R package Rtsne (*43*) with 2000 iterations and a perplexity of 30. Heat maps were created using the ComplexHeatmap (*42*) package using average linkage and Euclidian distance as similarity measure. Receiver operating characteristic curves (ROCs) and AUCs were computed with the R package pROC (*44*). Multi-class AUCs were computed as defined by Hand and Till (*45*).

Gene set enrichment analysis

We performed gene set enrichment analysis of the top 2000 variable CpG sites using the function gometh from the R-package missMethyl (*46*). Multiple testing correction was applied using the BH method (false discovery rate < 5%).

Statistical analysis

All data analyses were performed using the statistical programming language R (*47*), including the packages caret (*48*), kernlab (*39*), randomForest (*40*), keras (*37*), and Rtsne (*43*). DNA methylation data were processed using the minfi (*49*) package and single-sample normal-exponential out-of-band (Noob) normalization (*50*). As described previously (*19*), we excluded CpG sites associated with single-nucleotide polymorphisms or the sex chromosomes as well as sites that have previously been reported as cross-reactive (*19*, *51*). Survival curves were generated using the Kaplan-Meier method and tested for significance using the log-rank test. *P* values of <0.05 were considered statistically significant.

SUPPLEMENTARY MATERIALS

stm.sciencemag.org/cgi/content/full/11/509/eaaw8513/DC1

Fig. S1. *t*-SNE analysis of additional methods to estimate tumor purity in the reference cohort.
Fig. S2. Heat map analysis of the top 2000 variable CpG sites from the reference cohort.
Fig. S3. Overview of the probability scores of the three machine learning algorithms on the clinical cohort.
Fig. S4. Prognostic relevance of the diagnosis predicted by the three different classifiers in the clinical cohort.
Fig. S5. Dependency of the cross-validation loss in the reference cohort on the number of CpG sites included in the classification model for the three different classification methods.
Data file S1 contains tables S1 to S3:
Table S1. Detailed metadata and probability scores for the samples included in the clinical cohort.
Table S2. Gene enrichment analysis of the 2000 top variable CpG sites of the reference cohort.
Table S3. Metadata and detailed classification results for LUSC cases from TCGA with previous HNSC.
Table S4. Antibodies used for immunohistochemistry.
Table S5. Results from comparative NGS of cases 31 and 43.

REFERENCES AND NOTES

1. S. Wiegand, A. Zimmermann, T. Wilhelm, J. A. Werner, Survival after distant metastasis in head and neck cancer. *Anticancer Res.* **35**, 5499–5502 (2015).
2. R. P. Takes, A. Rinaldo, C. E. Silver, M. Haigentz Jr., J. A. Woolgar, A. Triantafyllou, V. Mondin, D. Paccagnella, R. de Bree, A. R. Shaha, D. M. Hartl, A. Ferlito, Distant metastases from head and neck squamous cell carcinoma. Part I. Basic aspects. *Oral Oncol.* **48**, 775–779 (2012).
3. X. Gao, S. G. Fisher, N. Mohideen, B. Emami, Second primary cancers in patients with laryngeal cancer: A population-based study. *Int. J. Radiat. Oncol. Biol. Phys.* **56**, 427–435 (2003).
4. R. P. Dikshit, P. Boffetta, C. Bouchardy, F. Merletti, P. Crosignani, T. Cuchi, E. Ardanaz, P. Brennan, Risk factors for the development of second primary tumors among men after laryngeal and hypopharyngeal carcinoma. *Cancer* **103**, 2326–2333 (2005).
5. Y.-B. Hsu, S.-Y. Chang, M.-C. Lan, J.-L. Huang, S.-K. Tai, P.-Y. Chu, Second primary malignancies in squamous cell carcinomas of the tongue and larynx: An analysis of incidence, pattern, and outcome. *J. Chin. Med. Assoc.* **71**, 86–91 (2008).
6. T. C. Pereira, S. M. Share, A. V. Magalhães, J. F. Silverman, Can we tell the site of origin of metastatic squamous cell carcinoma? An immunohistochemical tissue microarray study of 194 cases. *Appl. Immunohistochem. Mol. Morphol.* **19**, 10–14 (2011).
7. H. Bohnenberger, L. Kaderali, P. Ströbel, D. Yepes, U. Plessmann, N. V. Dharia, S. Yao, C. Heydt, S. Merkelbach-Bruse, A. Emmert, J. Hoffmann, J. Bodemeyer, K. Reuter-Jessen, A.-M. Lois, L. Dröge, P. Baumeister, C. Walz, L. Biggemann, R. Walter, B. Häupl, F. Comoglio, K.-T. Pan, S. Scheich, C. Lenz, S. Küffer, F. Bremmer, J. Kitz, M. Sitte, T. Beißbarth, M. Hinterthaner, M. Sebastian, J. Lotz, H.-U. Schildhaus, H. Wolff, B. C. Danner, C. Brandts, R. Büttner, M. Canis, K. Stegmaier, H. Serve, H. Urlaub, T. Oellerich, Comparative proteomics reveals a diagnostic signature for pulmonary head-and-neck cancer metastasis. *EMBO Mol. Med.* **10**, e8428 (2018).
8. T. W. Geurts, P. M. Nederlof, M. W. M. van den Brekel, L. J. van't Veer, D. de Jong, A. A. M. Hart, N. van Zandwijk, H. Klomp, A. J. M. Balm, M.-L. F. van Velthuysen, Pulmonary squamous cell carcinoma following head and neck squamous cell carcinoma: Metastasis or second primary? *Clin. Cancer Res.* **11**, 6608–6614 (2005).
9. M. G. C. T. van Oijen, F. G. J. L. vd Straat, M. G. J. Tilanus, P. J. Slootweg, The origins of multiple squamous cell carcinomas in the aerodigestive tract. *Cancer* **88**, 884–893 (2000).
10. J. Vermorken, P. Specenier, Optimal treatment for recurrent/metastatic head and neck cancer. *Ann. Oncol.* **21**, vii252–vii261 (2010).
11. F. Al-Shahrabani, D. Vallböhmer, S. Angenendt, W. T. Knoefel, Surgical strategies in the therapy of non-small cell lung cancer. *World J. Clin. Oncol.* **5**, 595–603 (2014).
12. B. J. Flehinger, M. Kimmel, M. R. Melamed, The effect of surgical treatment on survival from early lung cancer: Implications for screening. *Chest* **101**, 1013–1018 (1992).
13. D. Heim, G. Montavon, P. Hufnagl, K.-R. Müller, F. Klauschen, Computational analysis reveals histotype-dependent molecular profile and actionable mutation effects across cancers. *Genome Med.* **10**, 83 (2018).
14. D. Heim, J. Budczies, A. Stenzinger, D. Treue, P. Hufnagl, C. Denkert, M. Dietel, F. Klauschen, Cancer beyond organ and tissue specificity: Next-generation-sequencing gene mutation data reveal complex genetic similarities across major cancers. *Int. J. Cancer* **135**, 2362–2369 (2014).
15. V. Pareek, J. N. Sharma, Y. Eng, S. M. Keller, R. V. Smith, X. Guo, C. D. Shah, L. M. Gay, J. A. Elvin, J. Suh, J.-A. Vergilio, P. Stephens, J. S. Ross, V. A. Miller, B. Halmos, M. Haigentz, Distinguishing head and neck cancer metastasis from second primary squamous lung cancer in the genomic era. *J. Clin. Oncol.* **34**, e17506 (2016).
16. A. Lal, R. Panos, M. Marjanovic, M. Walker, E. Fuentes, G. J. Kubicek, D. W. Henner, L. J. Buturovic, M. Halks-Miller, A gene expression profile test to resolve head & neck squamous versus lung squamous cancers. *Diagn. Pathol.* **8**, 44 (2013).
17. S. G. Talbot, C. Estilo, E. Maghami, I. S. Sarkaria, D. Pham, P. O-charoenrat, N. D. Socci, I. Ngai, D. Carlson, R. Ghossein, A. Viale, B. J. Park, V. W. Rusch, B. Singh, Gene expression profiling allows distinction between primary and metastatic squamous cell carcinomas in the lung. *Cancer Res.* **65**, 3063–3071 (2005).
18. S. Moran, A. Martínez-Cardús, S. Sayols, E. Musulén, C. Balañá, A. Estival-Gonzalez, C. Moutinho, H. Heyn, A. Diaz-Lagares, M. de Moura, G. M. Stella, P. M. Comoglio, M. Ruiz-Miró, X. Matias-Guiu, R. Pazo-Cid, A. Antón, R. Lopez-Lopez, G. Soler, F. Longo, I. Guerra, S. Fernandez, Y. Assenov, C. Plass, R. Morales, J. Carles, D. Bowtell, L. Mileshkin, D. Sia, R. Tothill, J. Tabernero, J. M. Llovet, M. Esteller, Epigenetic profiling to classify cancer of unknown primary: A multicentre, retrospective analysis. *Lancet Oncol.* **17**, 1386–1395 (2016).
19. D. Capper, D. T. W. Jones, M. Sill, V. Hovestadt, D. Schrimpf, D. Sturm, C. Koelsche, F. Sahm, L. Chavez, D. E. Reuss, A. Kratz, A. K. Wefers, K. Huang, K. W. Pajtler, L. Schweizer, D. Stichel, A. Olar, N. W. Engel, K. Lindenberg, P. N. Harter, A. K. Braczynski, K. H. Plate, H. Dohmen, B. K. Garvalov, R. Coras, A. Hölsken, E. Hewer, M. Bewerunge-Hudler, M. Schick, R. Fischer, R. Beschorner, J. Schittenhelm, O. Staszewski, K. Wani, P. Varlet, M. Pages, P. Temming, D. Lohmann, F. Selt, H. Witt, T. Milde, O. Witt, E. Aronica, F. Giangaspero, E. Rushing, W. Scheurlen, C. Geisenberger, F. J. Rodriguez, A. Becker, M. Preusser, C. Haberler, R. Bjerkvig, J. Cryan, M. Farrell, M. Deckert, J. Hench, S. Frank, J. Serrano, K. Kannan, A. Tsirigos, W. Brück, S. Hofer, S. Brehmer, M. Seiz-Rosenhagen, D. Hänggi, V. Hans, S. Rozsnoki, J. R. Hansford, P. Kohlhof, B. W. Kristensen, M. Lechner, B. Lopes, C. Mawrin, R. Ketter, A. Kulozik, Z. Khatib, F. Heppner, A. Koch, A. Jouvet, C. Keohane, H. Mühleisen, W. Mueller, U. Pohl, M. Prinz, A. Benner, M. Zapatka, N. G. Gottardo, P. Driever, C. M. Kramm, H. L. Müller, S. Rutkowski, K. von Hoff, M. C. Frühwald, A. Gnekow, G. Fleischhack, S. Tippelt, G. Calaminus, C.-M. Monoranu, A. Perry, C. Jones, T. S. Jacques, B. Radlwimmer, M. Gessi, T. Pietsch, J. Schramm, G. Schackert, M. Westphal, G. Reifenberger, P. Wesseling, M. Weller, V. Collins, I. Blümcke, M. Bendszus, J. Debus, A. Huang, N. Jabado, P. A. Northcott, W. Paulus, A. Gajjar, G. W. Robinson, M. D. Taylor, Z. Jaunmuktane, M. Ryzhova, M. Platten, A. Unterberg, W. Wick, M. A. Karajannis, M. Mittelbronn, T. Acker, C. Hartmann, K. Aldape, U. Schüller, R. Buslei, P. Lichter, M. Kool, C. Herold-Mende, D. W. Ellison, M. Hasselblatt, M. Snuderl,

S. Brandner, A. Korshunov, A. von Deimling, S. M. Pfister, DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).
20. C. Koelsche, W. Hartmann, D. Schrimpf, D. Stichel, S. Jabar, A. Ranft, D. E. Reuss, F. Sahm, D. T. W. Jones, M. Bewerunge-Hudler, M. Trautmann, T. Klingebiel, C. Vokuhl, M. Gessler, E. Wardelmann, I. Petersen, D. Baumhoer, U. Flucke, C. Antonescu, M. Esteller, S. Fröhling, M. Kool, S. M. Pfister, G. Mechtersheimer, U. Dirksen, A. von Deimling, Array-based DNA-methylation profiling in sarcomas with small blue round cell histology provides valuable diagnostic information. *Mod. Pathol.* **31**, 1246–1256 (2018).
21. D. Aran, M. Sirota, A. J. Butte, Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015).
22. X. Zheng, N. Zhang, H.-J. Wu, H. Wu, Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biol.* **18**, 17 (2017).
23. K.-W. Tang, B. Alaei-Mahabadi, T. Samuelsson, M. Lindh, E. Larsson, The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat. Commun.* **4**, 2513 (2013).
24. K. Brennan, J. L. Koenig, A. J. Gentles, J. B. Sunwoo, O. Gevaert, Identification of an atypical etiological head and neck squamous carcinoma subtype featuring the CpG island methylator phenotype. *EBioMedicine* **17**, 223–236 (2017).
25. V. Hovestadt, M. Remke, M. Kool, T. Pietsch, P. A. Northcott, R. Fischer, F. M. G. Cavalli, V. Ramaswamy, M. Zapatka, G. Reifenberger, S. Rutkowski, M. Schick, M. Bewerunge-Hudler, A. Korshunov, P. Lichter, M. D. Taylor, S. M. Pfister, D. T. W. Jones, Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour material using high-density DNA methylation arrays. *Acta Neuropathol.* **125**, 913–916 (2013).
26. J. Budczies, M. Bockmayr, D. Treue, F. Klauschen, C. Denkert, Semiconductor sequencing: How many flows do you need? *Bioinformatics* **31**, 1199–1203 (2015).
27. The Cancer Genome Atlas Network, Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
28. The Cancer Genome Atlas Network, Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015).
29. The Cancer Genome Atlas Network, Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
30. J. Nicodemus-Johnson, R. A. Myers, N. J. Sakabe, D. R. Sobreira, D. K. Hogarth, E. T. Naureckas, A. I. Sperling, J. Solway, S. R. White, M. A. Nobrega, D. L. Nicolae, Y. Gilad, C. Ober, DNA methylation in lung cells is associated with asthma endotypes and genetic risk. *JCI Insight* **1**, e90151 (2016).
31. A. Karlsson, M. Jönsson, M. Lauss, H. Brunnström, P. Jönsson, Å. Borg, G. Jönsson, M. Ringnér, M. Planck, J. Staaf, Genome-wide DNA methylation analysis of lung carcinoma reveals one neuroendocrine and four adenocarcinoma epitypes associated with patient outcome. *Clin. Cancer Res.* **20**, 6127–6140 (2014).
32. A. M. Lim, N. C. Wong, R. Pidsley, E. Zotenko, J. Corry, A. Dobrovic, S. J. Clark, D. Rischin, B. Solomon, Genome-scale methylation assessment did not identify prognostic biomarkers in oral tongue carcinomas. *Clin. Epigenetics* **8**, 74 (2016).
33. D. D. Esposti, A. Sklias, S. C. Lima, S. Beghelli-de la Forest Divonne, V. Cahais, N. Fernandez-Jimenez, M.-P. Cros, S. Ecsedi, C. Cuenin, L. Bouaoun, G. Byrnes, R. Accardi, A. Sudaka, V. Giordanengo, H. Hernandez-Vargas, L. F. R. Pinto, E. Van Obberghen-Schilling, Z. Herceg, Unique DNA methylation signature in HPV-positive head and neck squamous cell carcinomas. *Genome Med.* **9**, 33 (2017).
34. M. Bjaanæs, T. Fleischer, A. R. Halvorsen, A. Daunay, F. Busato, S. Solberg, L. Jørgensen, E. Kure, H. Edvardsen, A.-L. Børresen-Dale, O. T. Brustugun, J. Tost, V. Kristensen, Å. Helland, Genome-wide DNA methylation analyses in lung adenocarcinomas: Association with EGFR, KRAS and TP53 mutation status, gene expression and prognosis. *Mol. Oncol.* **10**, 330–343 (2016).
35. J. Sandoval, J. Mendez-Gonzalez, E. Nadal, G. Chen, J. F. Carmona, S. Sayols, S. Moran, H. Heyn, M. Vizoso, A. Gomez, M. Sanchez-Cespedes, Y. Assenov, F. Müller, C. Bock, M. Taron, J. Mora, L. A. Muscarella, T. Liloglou, M. Davies, M. Pollan, M. J. Pajares, W. Torre, L. M. Montuenga, E. Brambilla, J. K. Field, L. Roz, M. L. Iacono, G. V. Scagliotti, R. Rosell, D. G. Beer, M. Esteller, A prognostic DNA methylation signature for stage i non–small-cell lung cancer. *J. Clin. Oncol.* **31**, 4140–4147 (2013).
36. B. Basu, J. Chakraborty, A. Chandra, A. Katarkar, J. R. K. Baldevbhai, D. Chowdhury, J. G. Ray, K. Chaudhuri, R. Chatterjee, Genome-wide DNA methylation profile identified a unique set of differentially methylated immune genes in oral squamous cell carcinoma patients in India. *Clin. Epigenetics* **9**, 13 (2017).
37. J. Allaire, F. Chollet, R interface to keras; https://keras.io.
38. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv:1412.6980 [cs.LG] (22 December 2014).
39. A. Karatzoglou, A. Smola, K. Hornik, A. Zeileis, kernlab—An S4 package for kernel methods in R. *J. Stat. Softw.* **11**, 1–20 (2004).
40. A. Liaw, M. Wiener, Classification and regression by RandomForest. *R News* **2**, 18–22 (2002).
41. V. Vapnik, *The Nature of Statistical Learning Theory* (Springer, 1995).
42. Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
43. J. H. Krijthe, Rtsne: T-distributed stochastic neighbor embedding using a Barnes-Hut implementation; https://github.com/jkrijthe/Rtsne.
44. X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, M. Müller, pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
45. D. J. Hand, R. J. Till, A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach. Learn.* **45**, 171–186 (2001).
46. B. Phipson, J. Maksimovic, A. Oshlack, missMethyl: An R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* **32**, 286–288 (2016).
47. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2017); www.R-project.org/.
48. M. Kuhn, caret: Classification and regression training. R package version 6.0-71; https://CRAN.R-project.org/package=caret.
49. M. J. Aryee, A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen, R. A. Irizarry, Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
50. T. J. Triche Jr., D. J. Weisenberger, D. Van Den Berg, P. W. Laird, K. D. Siegmund, Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* **41**, e90 (2013).
51. Y. A. Chen, M. Lemire, S. Choufani, D. T. Butcher, D. Grafodatskaya, B. W. Zanke, S. Gallinger, T. J. Hudson, R. Weksberg, Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).

Acknowledgments: We thank J. Staaf for providing raw IDAT files from primary LUSC samples. We thank H. Bläker for providing excellent consultations during the histomorphological reevaluation. We gratefully acknowledge the excellent technical assistance of A. Förster, V. Arnemann, B. Meyer-Bartell, P. Wolkenstein, and P. Jank. The results shown here are, in part, based on data generated by the TCGA Research Network (http://cancergenome.nih.gov/). **Funding:** This work was supported by the European Fund for Regional Development (EFRE) and the Federal State of Berlin (1303/2013) (to P.J.), the German Federal Ministry of Education and Research (MALT3, BBDC, BZML) (to M.B., F.K., K.-R.M., and P.S.), the Fördergemeinschaft Kinderkrebszentrum Hamburg (to M.B. and U.S.), the Berlin Institute of Health (BIH) Charité Clinician Scientist Program funded by the Charité–Universitätsmedizin Berlin and the Berlin Institute of Health (to M.v.L.), and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Government of South Korea (2017-0-00451; to K.-R.M.). **Author contributions:** M.B., D.C., F.K., and P.J. designed the research goals and aims. The methodology was developed and designed by M.B., D.C., G.M., P.J., and P.S. The formal analysis was performed by M.B., P.J., P.S., and K.-R.M. M.B., P.J., D. Teichmann, D. Treue, and C.V. performed the investigations. Resources were provided by D.C., F.K., and K.-R.M. Data curation was performed by A.A., K.B., M.B., T.B., P.J., D. Teichmann, and C.V. M.B. and P.J. wrote the original draft of the paper. All authors reviewed and edited the manuscript. Visualizations were created by M.B. and P.J. D.C., F.K., and K.-R.M. supervised the project. **Competing interests:** D.C. is listed as inventor on the patent application "DNA-methylation based method for classifying tumor species" (PCT/EP2016/055337) filed by Deutsches Krebsforschungszentrum Stiftung des öffentlichen Rechts and Ruprecht-Karls-Universität Heidelberg. **Data and materials availability:** All data associated with this paper can be found in the main text or the Supplementary Materials. IDAT files of the samples analyzed in this paper have been deposited in GEO (GSE124052). The code and the data necessary to reproduce the main analyses from this paper are available at http://doi.org/10.6084/m9.figshare.8182376.

Submitted 4 February 2019
Accepted 22 August 2019
Published 11 September 2019
10.1126/scitranslmed.aaw8513

Citation: P. Jurmeister, M. Bockmayr, P. Seegerer, T. Bockmayr, D. Treue, G. Montavon, C. Vollbrecht, A. Arnold, D. Teichmann, K. Bressem, U. Schüller, M. von Laffert, K.-R. Müller, D. Capper, F. Klauschen, Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Sci. Transl. Med.* **11**, eaaw8513 (2019).

Science Translational Medicine

Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases

Philipp Jurmeister, Michael Bockmayr, Philipp Seegerer, Teresa Bockmayr, Denise Treue, Grégoire Montavon, Claudia Vollbrecht, Alexander Arnold, Daniel Teichmann, Keno Bressem, Ulrich Schüller, Maximilian von Laffert, Klaus-Robert Müller, David Capper and Frederick Klauschen

Sci Transl Med **11**, eaaw8513.
DOI: 10.1126/scitranslmed.aaw8513

Discriminating lung primary tumors and metastases

Pulmonary metastases of head and neck squamous cell carcinoma (HNSC) are currently difficult to distinguish from primary lung squamous cell carcinomas (LUSCs). Differentiating these tumor types has important clinical implications, as whether the lung tumor is primary or has spread can affect the treatment options offered to a patient. Here, Jurmeister *et al.* developed a machine learning algorithm that exploits the differential DNA methylation observed in primary LUSC and metastasized HNSC tumors in the lung. Their method was able to discriminate between these two tumor types with high accuracy across multiple cohorts, suggesting its potential as a clinical diagnostic tool.