Journal of Pathology
J Pathol 2022; **256:** 378–387
Published online 20 January 2022 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/path.5845

ORIGINAL ARTICLE

Machine learning models predict the primary sites of head and neck squamous cell carcinoma metastases based on DNA methylation

Maximilian Leitheiser[1], David Capper[2,3], Philipp Seegerer[4,5], Annika Lehmann[1], Ulrich Schüller[6,7,8], Klaus-Robert Müller[4,9,10,11], Frederick Klauschen[1,5,11,12], Philipp Jurmeister[1,3,12†*] and Michael Bockmayr[1,6,8,13†*]

[1] *Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Institute of Pathology, Berlin, Germany*
[2] *Department of Neuropathology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin, Germany*
[3] *German Cancer Consortium (DKTK), Partner Site Berlin, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[4] *Machine-Learning Group, Department of Software Engineering and Theoretical Computer Science, Technical University of Berlin, Berlin, Germany*
[5] *Aignostics GmbH, Berlin, Germany*
[6] *Department of Pediatric Hematology and Oncology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany*
[7] *Institute of Neuropathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany*
[8] *Research Institute Children's Cancer Center Hamburg, Hamburg, Germany*
[9] *Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea*
[10] *Max-Planck-Institute for Informatics, Saarbrücken, Germany*
[11] *BIFOLD – Berlin Institute for the Foundations of Learning and Data, Berlin, Germany*
[12] *Faculty of Medicine, LMU München, Institute of Pathology, Munich, Germany*
[13] *Mildred Scheel Cancer Career Center HaTriCS4, University Medical Center Hamburg-Eppendorf, Hamburg, Germany*

**Correspondence to: P Jurmeister, Institute of Pathology, Ludwig-Maximilians-Universität München, Thalkirchner Street 36, 80337 Munich, Germany or M Bockmayr, Institute of Pathology, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany. E-mail: philipp.jurmeister@med.uni-muenchen.de or michael.bockmayr@charite.de*

†*These authors contributed equally to this work.*

Abstract

In head and neck squamous cell cancers (HNSCs) that present as metastases with an unknown primary (HNSC-CUPs), the identification of a primary tumor improves therapy options and increases patient survival. However, the currently available diagnostic methods are laborious and do not offer a sufficient detection rate. Predictive machine learning models based on DNA methylation profiles have recently emerged as a promising technique for tumor classification. We applied this technique to HNSC to develop a tool that can improve the diagnostic work-up for HNSC-CUPs. On a reference cohort of 405 primary HNSC samples, we developed four classifiers based on different machine learning models [random forest (RF), neural network (NN), elastic net penalized logistic regression (LOGREG), and support vector machine (SVM)] that predict the primary site of HNSC tumors from their DNA methylation profile. The classifiers achieved high classification accuracies (RF = 83%, NN = 88%, LOGREG = SVM = 89%) on an independent cohort of 64 HNSC metastases. Further, the NN, LOGREG, and SVM models significantly outperformed p16 status as a marker for an origin in the oropharynx. In conclusion, the DNA methylation profiles of HNSC metastases are characteristic for their primary sites, and the classifiers developed in this study, which are made available to the scientific community, can provide valuable information to guide the diagnostic work-up of HNSC-CUP.

Keywords: head and neck squamous cell carcinoma; DNA methylation; machine learning; cancer of unknown primary

Received 26 July 2021; Revised 24 October 2021; Accepted 6 December 2021

Conflict of interest statement: FK and KRM are co-founders of Aignostics. PS is employed at Aignostics. DC is listed as an 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. No other potential conflicts of interests were declared.

Introduction

Comprising about 4.6% of all cancers, head and neck cancers are the eighth most common malignancy worldwide [1]. Up to 9% of head and neck cancers initially present as cervical lymph node metastases with unknown primary, i.e. their primary tumor could not be identified in the routine diagnostic work-up. Squamous cell carcinomas account for up to 75% of tumors presenting in such a constellation [2].

While these head and neck squamous cell cancers (HNSCs) of unknown primary (HNSC-CUPs) have a more favorable prognosis than the general class of cancers of unknown primary (CUPs) [3], their therapy remains challenging. The identification of a primary tumor significantly improves the overall survival of these patients by enabling a more specific therapeutic approach [4,5]. Most importantly, precautionary irradiation of putative primary sites, which is otherwise the recommended procedure in many cases [6], can be avoided. Further, the identified primaries often have a low local tumor stage (see, for example, Figure 1A–D) and can potentially be resected with clear surgical margins [6].

The current American Society of Clinical Oncology (ASCO) guidelines [6] recommend an extensive diagnostic work-up for HNSC-CUP. After the diagnosis has been established from a biopsy or resection of the suspicious neck mass, a thorough physical examination, computed tomography (CT), and positron emission tomography-CT (PET-CT) imaging should be performed. If necessary, complete operative evaluation of the upper aerodigestive tract, including biopsies of any suspicious mucosal sites and in some cases diagnostic tonsillectomy, is recommended. With traditional panendoscopic methods, primary detection rates of 50–60% are reported for the combination of the described procedures. In more recent publications, the use of transoral laser microsurgery or transoral robotic surgery in the evaluation of the upper aerodigestive tract showed improved rates, ranging from 63% to 90% [7]. From a pathologist's perspective, squamous cell carcinomas of different primary sites do not show any specific histomorphological features. Therefore, the only tissue-based method that can assist in determining the primary site of an HNSC-CUP is the detection of Epstein–Barr virus (EBV) or human papillomavirus (HPV), both of which are associated with tumors of the nasopharynx and oropharynx, respectively [4]. In its entirety, the recommended diagnostic work-up is costly, invasive, and time-intensive, and is still not sufficient to identify the primary tumor in all cases. Thus, further techniques to assist the identification of the primary tumor in HNSC-CUP cases are needed.

In recent years, machine learning (ML) techniques have contributed to diagnosis in pathology [8–13]. Models based on DNA methylation profiles have been used to classify CNS tumors [14] and to predict the origin of neuroendocrine tumors [15] as well as CUPs [16]. In the latter, the set of tumors is heterogeneous with respect to histology and primary site, and the prediction classes are therefore broad. In particular, there was no further differentiation within the class of squamous cell tumors. In a recent study, we found first indications that squamous cell carcinomas from different subregions of the head and neck area exhibit characteristic DNA methylation profiles [17].

To date, most applications of ML methods to DNA methylation analysis have been based on random forest (RF) models [14–16,18]. While these were successful, support vector machines (SVMs), elastic net penalized logistic regression (LOGREG), and neural network (NN) models showed better results in two recent studies [17,19].

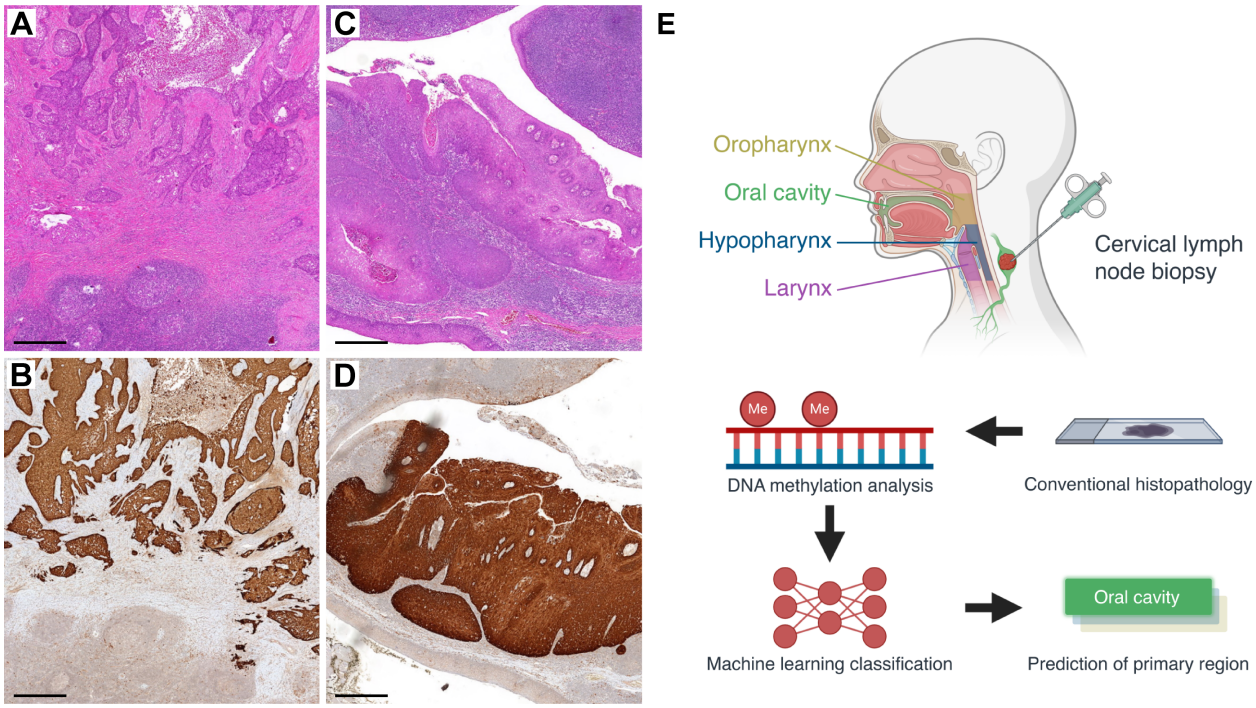

Figure 1. Histopathology and anatomy of HNSC. (A) H&E and (B) p16 immunohistochemistry images of an HNSC lymph node metastasis from the validation set. (C) H&E and (D) p16 immunohistochemistry images of the corresponding primary T1 tumor located in the tonsil. Scale bars, 50 µm. (E) Illustration of the workflow for the developed classifiers, including the relevant anatomical subregions of the head and neck.

The goal of the present study was to develop a DNA methylation-based classifier able to further differentiate the regions of origin for HNSC metastases. Such a classifier has the potential to be used in the diagnostic work-up of HNSC-CUPs and could facilitate faster and more frequent identification of a primary tumor and thus better patient outcome.

Materials and methods

Study design

We developed four different machine learning classifiers that predict the primary site of HNSC tumors ('oral cavity', 'oropharynx', 'hypopharynx or larynx') from their DNA methylation profile. They were trained on a reference cohort of HNSC primary tumors ($n = 405$) and applied to an independent validation cohort of HNSC metastases ($n = 64$) to compare their performance.

The data used in this study included the previously published datasets GSE87053 [20], GSE95036 [21], and GSE124052 [17] from the Gene Expression Omnibus (GEO) [22] and the HNSC dataset (TCGA HNSC) from The Cancer Genome Atlas (TCGA) [23,24]. This was complemented by data from a DNA methylation analysis of patient samples from our archives (hereafter 'study dataset'). An overview of these datasets and their characteristics is given in Table 1.

For inclusion in the reference cohort, all HNSC primary tumor samples from the included datasets were considered. After exclusion of 63 samples from the TCGA HNSC dataset due to an unclear primary site, this yielded a reference cohort of 405 samples (see Table 2). The validation cohort was composed of all HNSC metastasis samples in the included datasets. For four samples, no clear primary site could be assigned, and one sample did not pass the quality control for its DNA methylation data. After exclusion of these samples, the validation cohort contained 64 samples (see Table 2).

Patient selection, samples, and clinical data

For inclusion in the study dataset, we identified 49 suitable patients with cervical lymph node metastases of HNSC using the electronic patient files and the electronic database of the Charité – University Hospital Berlin. We specifically selected cases that were initially diagnosed as CUPs but had their primary tumor site identified in the further course of the disease through additional diagnostic procedures. Additional clinical data for these cases are given in supplementary material, Table S1.

Formalin-fixed and paraffin-embedded (FFPE) tissue was retrieved from the archives of the Institute of Pathology at the Charité – University Hospital Berlin and clinical data were extracted from the electronic patient files. Ethics approval was granted by the local ethics committee (EA1/122/18).

Annotation of primary sites

The primary sites of HNSC tumors were categorized into three classes: oral cavity, oropharynx, and hypopharynx or larynx (Figure 1E). These individual regions and their anatomical subsites were defined according to the current guidelines of the WHO [25]. Hypopharynx and larynx were combined into one class due to the small number of samples with a hypopharyngeal primary site ($n = 29$) and their anatomical adjacency.

Additionally, all HNSC tumors were annotated with regard to their organ of origin. The following organs were considered for annotation: oral mucosa, tongue (composed of oral tongue in the oral cavity and base of tongue in the oropharynx), pharyngeal wall (ranging over oropharynx and hypopharynx), tonsil, and larynx.

The region and organ of the primary site were determined based on all clinical data available for each sample (see Table 1 and section entitled 'Patient selection, samples, and clinical data'). For 63 samples in the reference cohort and four samples in the validation cohort, no clear region of the primary site was identified, and they were excluded from further analysis. The resulting distribution of annotated regions is given in Table 2 and a full list of the annotations is provided in supplementary material, Table S2.

Tissue preparation and DNA methylation analysis

For samples from the study dataset, DNA methylation analysis was performed on FFPE tissue with a tumor cell content of at least 60%. Semi-automated DNA extraction was performed using the Maxwell RSC FFPE Plus DNA Purification Kit (Promega, Fitchburg, WI, USA) on a Maxwell RSC 16 instrument (Promega) according to protocols supplied by the manufacturer. DNA quantities were measured using the Qubit HS DNA assay (Thermo Fisher Scientific, Waltham, MA, USA). The EpiTect Fast DNA Bisulfite Kit (Qiagen, Venlo, The Netherlands) was used to perform DNA bisulfite

Table 1. Reference and validation cohort: included datasets and their characteristics.

Cohort	Dataset	Chip design	Tumor type	Tissue type	IDAT source	Annotation source
Reference	GSE87053	450k	Primary	Frozen	GEO	GEO metadata
	GSE95036	450k	Primary	FFPE	GEO	GEO metadata
	TCGA HNSC	450k	Primary	Frozen	TCGA	TCGA pathology reports
	GSE124052 (prim)	EPIC	Primary	FFPE	GEO	GEO metadata + medical records
Validation	GSE124052 (met)	EPIC	LU met	FFPE	GEO	GEO metadata + medical records
	Study dataset	EPIC	LN met	FFPE	Sample analysis	Medical records

prim, primary tumors; met, metastases; LU, lung; LN, lymph node.

Table 2. Reference and validation cohort: case numbers and distribution of primary sites.

Cohort	Dataset	No. considered	ROO NA	Failed QC	No. used	OC	ORO	H&L
Reference		468	63	0	405	234 (58%)	57 (14%)	114 (28%)
	GSE87053	11	0	0	11	11 (100%)	0 (0%)	0 (0%)
	GSE95036	11	0	0	11	3 (27%)	5 (45%)	3 (27%)
	TCGA HNSC	442	63	0	379	217 (57%)	51 (13%)	111 (29%)
	GSE124052 (prim)	4	0	0	4	3 (75%)	1 (25%)	0 (0%)
Validation		69	4	1	64	25 (39%)	25 (39%)	14 (22%)
	GSE124052 (met)	20	4	0	16	3 (19%)	8 (50%)	5 (31%)
	Study dataset	49	0	1	48	22 (46%)	17 (35%)	9 (19%)

prim, primary tumors; met, metastases; ROO, region of origin; NA, not available; QC, quality control; OC, oral cavity; ORO, oropharynx; H&L, hypopharynx or larynx.

conversion. Following DNA restoration for FFPE samples (Infinium HD FFPE DNA Restore Kit; Illumina, San Diego, CA, USA), DNA methylation analysis was carried out using the Infinium MethylationEPIC BeadChip (Illumina), according to the manufacturer's instructions. Stained bead array chips were analyzed using the iScan platform (Illumina).

Data preprocessing

All data processing was performed in R (version 3.6.1; R Foundation for Statistical Computing, Vienna, Austria). Using the *minfi* library [26], raw IDAT files were imported; Noob normalization was performed [27]; and beta values were computed. Beta values whose corresponding intensities on probe-level failed a z-test with a P value threshold of 0.01 against background signal were masked. To allow for consistent analysis of data from EPIC and 450k chips, each sample was reduced to the CpG sites present in both designs ($n = 452\,453$). Further, CpG sites on sex chromosomes were excluded ($n = 10\,583$) as well as a group of underperforming CpG sites ($n = 54\,898$) following the recommendation in ref 28. Samples with more than 10% missing beta values were excluded (five samples, see Table 2). Finally, missing or masked beta values were imputed by adopting the values of the CpG site with the closest genomic position.

Classifier development

A dimensionality reduction was performed on the training set, in which each sample was reduced to the 2000 most variable CpG sites (results for alternative numbers of CpG sites are shown in supplementary material, Figure S1) across the training set. The chosen CpG sites are provided in supplementary material, Table S3. Then, LOGREG, RF, SVM, and NN models were trained and evaluated using five-fold cross-validation on the training set. Class-balanced partitions for the cross-validation were created with the package *caret* [29]. All models were configured to return probability scores. From those, predictions were obtained by selecting the class with the maximum probability score. The hyperparameters and the packages used are listed in supplementary material, Table S4. The optimal set of hyperparameters was selected by minimal categorical cross-entropy loss during cross-validation. The final model of each type was then trained on the full training set with its optimal hyperparameters. While chip design and tissue type are known to be a possible cause of batch effects in DNA methylation data, batch effect correction for these factors did not significantly affect classification performance in the cross-validation and was therefore not used subsequently (supplementary material, Table S5).

Classifier validation

The four classifiers were applied to the validation set after a dimensionality reduction to the 2000 CpG sites previously selected on the training set. Differences in their results were verified in Cochran's Q test using the package *RVAideMemoire*. Fisher's exact test was used to assess the independence of prediction accuracy and metastasis site (lung, lymph node) for each classifier. Further, the performance of each classifier was compared with p16-based prediction in the binary discrimination between oropharynx and non-oropharynx.

t-SNE plots

To visualize the high-dimensional methylation profiles, t-distributed stochastic neighbor embeddings (t-SNE) [30] were computed using the package *Rtsne*. Embeddings were computed on the reference cohort only, and on the reference and validation cohort combined with perplexity 20, with 1000 iterations and a 50-dimensional principal component analysis.

Tumor purity estimation

The proportion of tumor cells was estimated from DNA methylation profiles using the ESTIMATE-based method provided by the package *RFpurify* [31]. A comparison of these results with estimates obtained by using the package *InfiniumPurify* [32] is presented in the supplementary material, Figure S2.

p16 status and HPV genotyping

Immunohistochemical evaluation was performed on the Ventana BenchMark XT automated slide stainer (Roche Tissue Diagnostics, Tucson, AZ, USA) according to the manufacturer's instructions. The D7C1M p16 antibody (Cell Signaling Technology, Danvers, MA, USA) was used diluted 1:1000. Samples with strong nuclear or cytoplasmic staining in ≥70% of tumor cells were

considered positive. HPV genotyping was performed as described previously using the HPV Type 3.5 C LCD array (Chipron, Berlin, Germany) [17].

Results

DNA methylation analysis of the reference cohort

To identify characteristic epigenetic signatures for HNSC tumors with respect to their primary site, we first analyzed the DNA methylomes of a reference cohort of primary tumors ($n = 405$). The general characteristics of the dataset and the distribution of classes are listed in Tables 1 and 2, respectively.

A t-SNE of the reference cohort revealed three distinct groups related to the region of the primary sites (Figure 2A). The largest, central group was mainly composed of samples from the oral cavity but also contained a considerable number of samples from the hypopharynx and larynx concentrated at its top. Additionally, several samples from the oropharynx were scattered across this group. The upper right group almost exclusively consisted of samples from the oropharynx, whereas the lower left one contained mostly samples from the hypopharynx or larynx with a considerable number of exceptions from the oral cavity. For the organ of origin, the relation to the grouping in the t-SNE plot was less pronounced (Figure 2B). In particular, samples from the tongue, which is part of both the oral cavity (oral tongue) and the oropharynx (base of tongue), did not form a distinct subset but were distributed across the respective regional groups. This indicated that the region of origin explains the described grouping better than the organ of origin. We observed no relation between the grouping of the samples and their original dataset, thus ruling out substantial batch effects concerning tissue preparation or chip design (Figure 2C and Table 1). HPV-positive samples were mostly located within the aggregation of oropharynx samples, where they did not form a distinct subgroup (Figure 2D). Samples with a higher tumor cell content were more clearly separated than samples with lower purity. The latter were mostly found in the center of the large group, thus resembling samples from the oral cavity, which was the class with the highest number of samples (Figure 2E).

Classifier development on the reference cohort

Using the reference cohort, we trained four different classifiers, based on an NN, an SVM, a LOGREG, and an RF model, to predict primary sites from DNA methylation data (Figure 1E). The optimal hyperparameters for each model were determined in a five-fold cross-validation (supplementary material, Table S4). All models achieved high training accuracies on the reference cohort (NN = SVM = 93%, LOGREG = 92%, RF = 91%).

DNA methylation analysis of the validation cohort

We compiled an independent validation cohort of 64 HNSC metastases, comprising 48 cervical lymph node metastases from the study dataset (e.g. Figure 1A–D) and 16 pulmonary metastases from our previous work (Table 2).

In a combined t-SNE plot of the reference and validation cohort (Figure 3A), the samples from the validation cohort integrated well into the existing groups according to their region of origin, with only a few exceptions. Of

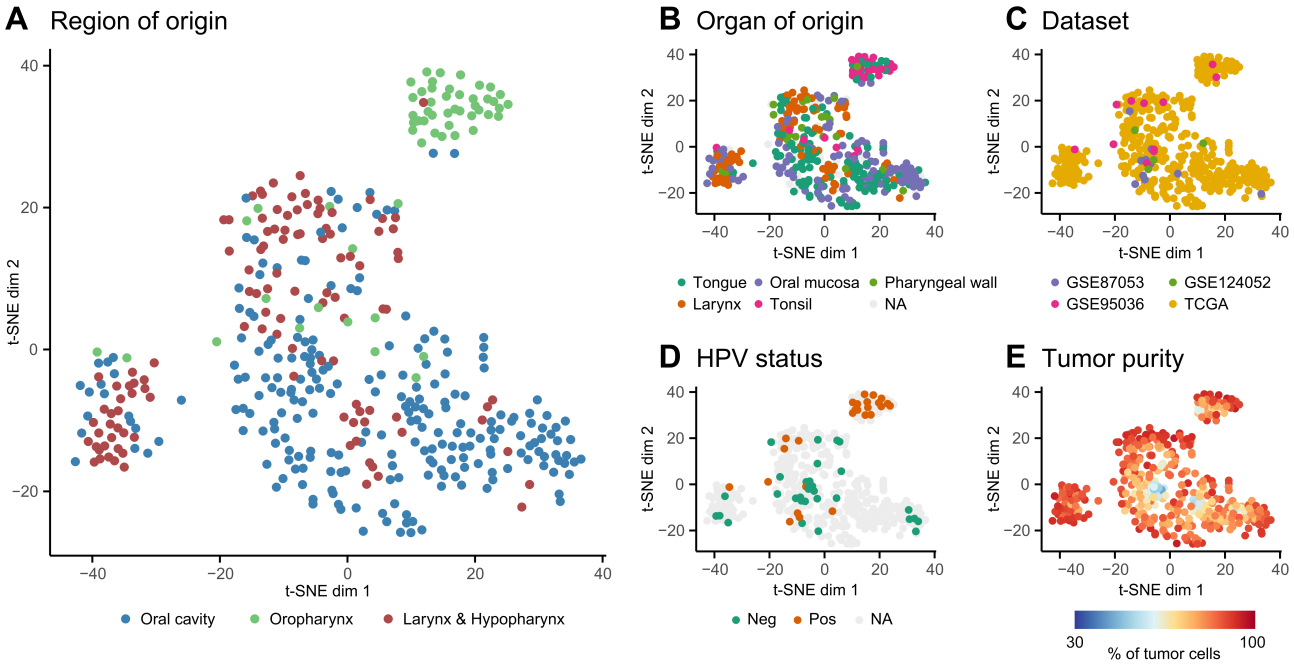

Figure 2. Reference cohort of primary HNSC tumors. Two-dimensional representation of the reference cohort samples ($n = 405$) based on a t-SNE computed from their DNA methylation profiles. Individual samples are color-coded according to (A) the region of origin as used in the classification, (B) the organ of origin, (C) the dataset, (D) HPV status, and (E) a tumor purity estimation.

note, most metastasis samples from the hypopharynx or larynx were located in the concentration of primary samples from the same region within the central group. In contrast, the left group, which was mainly formed by primary samples from the hypopharynx or larynx, contained no corresponding metastasis samples. We did not observe distinct groups for lymph node and lung metastasis specimens.

Classification results on the validation cohort

The four classifiers developed on the reference cohort predicted the primary site in the validation cohort with overall accuracies of 88% for the NN, 89% for both the SVM and LOGREG, and 83% for the RF model (Figure 3C). Although the accuracy of the RF classifier is clearly lower than the others, a Cochran Q test showed no significant differences between all classifiers ($p = 0.45$). Notably, 28% ($n = 7$) of oropharynx samples were falsely classified as oral cavity by the RF model, leading to a sensitivity of only 64% for the oropharynx class. The other classifiers did not show a similar tendency for false classifications in any class (Figure 3C).

By summing up the false classifications across the four classifiers for each sample, four problematic samples with more than two false classifications were identified (Figure 3B); they made up 50% (4/8), 57% (4/7), 43% (3/7), and 36% (4/11) of the total misclassified samples for the NN, SVM, LOGREG, and RF models, respectively. Of these four problematic cases, three were p16-negative oropharynx samples. Further analysis showed that the sensitivity of all classifiers was indeed remarkably lower on the subset of p16-negative oropharynx samples (NN = SVM = LOGREG = 75%, RF = 50%) than on p16-positive oropharynx samples (NN = SVM = LOGREG = 100%, RF = 77%).

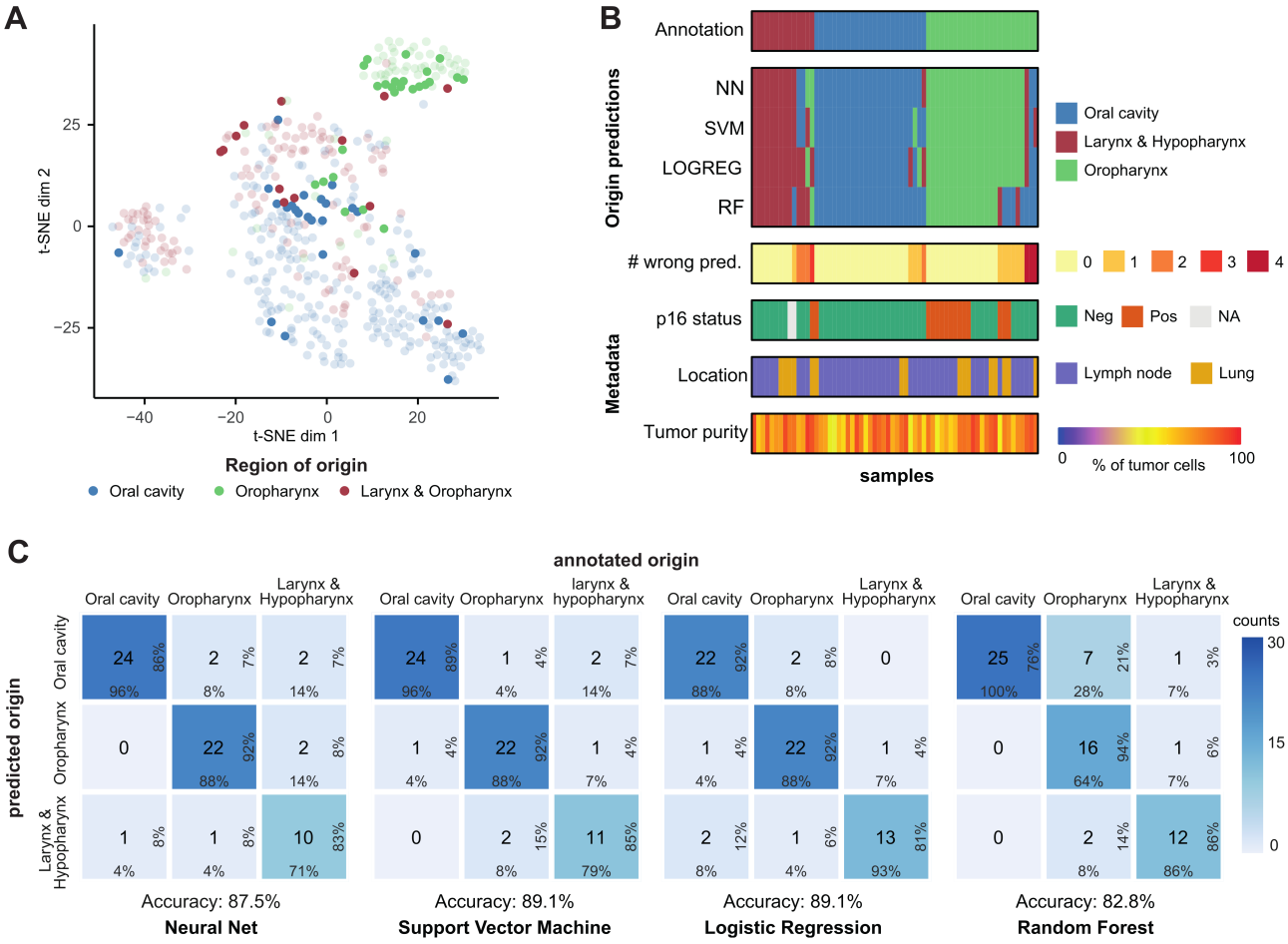

Figure 3. Classification results on the validation cohort of HNSC lung and lymph node metastases. (A) Two-dimensional representation of the full dataset ($n = 469$) based on a t-SNE computed from the DNA methylation profiles of its samples. The full dataset contains the reference cohort ($n = 405$) of primary HNSC tumors and the validation cohort ($n = 64$) of HNSC metastases in lung and lymph nodes. Samples from the reference cohort are displayed in transparent colors, and samples from the validation cohort in opaque colors. Individual samples are color-coded according to the region of origin as used in the classification. (B) Heatmap of the validation cohort ($n = 64$) showing the annotated region of origin, the predictions of the four classifiers, the total number of wrong predictions across the classifiers, the p16 status, the location of the metastases, and an estimation of tumor purity for every sample. (C) Confusion matrices for the result of the four classifiers on the validation cohort ($n = 64$) showing the relationship between the annotated and the predicted regions of origin. The main numbers show absolute counts of the respective cases; row and column percentages are displayed on the right and at the bottom of each square. The row and column percentages of the diagonal entries are the sensitivity and positive predictive value of the corresponding class, respectively. The overall accuracy of each classifier on the validation cohort is given below the confusion matrix.

Table 3. Binary classification results for oropharyngeal origin of metastases and comparison of accuracies by metastasis site.

	Binary classification (ORO versus non-ORO)			Multiclass classification (OC, ORO, H&L): comparison by metastasis site	
	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy lung (%)	Accuracy lymph node (%)
NN	92	88	95	88	88
SVM	92	88	95	88	90
LOGREG	92	88	95	94	88
RF	84	97	64	69	88
p16 status	77	52	95	-	-

OC, oral cavity; ORO, oropharynx; H&L, hypopharynx or larynx.

Comparison of accuracies for pulmonary and lymph node metastases

The overall accuracy did not show a significant difference between samples from the lung and the lymph nodes for any of the classifiers in a pairwise Fisher's exact test (P values: NN = 1.00, SVM = 1.00, LOGREG = 0.67, RF = 0.12). The NN, SVM, and LOGREG models achieved accuracies of around 90% on both subsets, whereas the accuracy of the RF model was markedly worse on pulmonary metastases, with 69%, than on lymph node metastases, with 88% (Table 3).

Binary classification (oropharynx versus non-oropharynx) and p16 status

In the binary classification task of discriminating between oropharyngeal and non-oropharyngeal origin, the NN, SVM, and LOGREG models all achieved an accuracy of 92%, a sensitivity of 88%, and a specificity of 95% (Table 3). The predictions of the RF model had a slightly lower accuracy of 84%, with a remarkably high sensitivity of 97% and low specificity of 64%. Prediction based purely on p16 status showed contrary results, with low sensitivity (52%) and high specificity (95%). Further, all classifiers had a higher overall accuracy than p16-based prediction, with only 77%.

Discussion

Currently recommended practices for the diagnosis and treatment of HNSC-CUPs [6] include laborious and invasive measures to identify the occult primary tumor. Identification of the primary site is highly desirable in HNSC-CUPs as it allows for better treatment options and increases patient survival [4,5]. In particular, the primary tumor can then be targeted directly for locoregional control, and harmful procedures such as the precautionary irradiation of potential primary areas can be avoided. The classification algorithms based on the DNA methylation signature of these tumors have the potential to accelerate this process and increase the detection rate of the primary tumor. Our approach can easily be implemented in the diagnostic work-up of HNSC-CUPs. Biopsy or resection of the involved lymph node is usually performed at the very beginning of the diagnostic process. After the conventional histopathological work-up has established the diagnosis of squamous cell carcinoma, the same tissue can be used for DNA methylation analysis. The most prominent use-case for the proposed techniques is HNSC-CUPs in the upper cervical lymph nodes (levels 1–3), which usually originate from the head and neck region and are most commonly squamous cell cancers [2]. Other promising applications of the proposed methods are pulmonary HNSC metastases. However, for their diagnosis as HNSC cancer, they first need to be distinguished from primary squamous cell cancer of the lung. For this purpose, another DNA methylation-based classifier was developed recently [17].

To our knowledge, tests for EBV and HPV are the only tissue-based methods that have been proposed for the prediction of primary sites in HNSC-CUPs. EBV status is reported to have a sensitivity and specificity of around 90% [2] in predicting a primary site in the nasopharynx. HPV status, and p16 status as a surrogate marker, achieve high accuracies and specificities of about 90% and 98%, respectively, as predictors for oropharyngeal primary sites in cervical squamous cell lymph node metastases [33,34]. However, they are not able to identify HPV-negative oropharyngeal HNSCs, which constitute a distinct tumor entity [35], leading to a sensitivity of only about 70% [33,34]. Both HPV and EBV status are limited in their use because they are only indicative of a single region of origin. The DNA methylation-based classifiers developed in this study can predict primary sites in all regions present in the data. Further, the NN, SVM, and LOGREG models clearly outperformed purely p16-based prediction on our data in terms of overall accuracy (92% versus 77%) and sensitivity (88% versus 52%) and were also superior to the results for HPV- and p16-status in the literature described above.

Gene expression [36–38] and protein profiling [39–41] have been successfully applied to the classification of other tumor entities by tissue of origin. A recent study [17] showed the superiority of a DNA methylation-based classifier over both of these approaches in distinguishing HNSC metastases from primary lung carcinoma, suggesting that this method is more promising for the goal of the present study. Gene expression and proteomic methods are preferably performed on fresh-frozen samples as they are negatively affected by deterioration of RNA and proteins in FFPE tissue. DNA methylation-based methods can be applied equally well to fresh-frozen and FFPE tissue, which is an advantage in practicability, as FFPE tissue samples are more readily available in routine diagnostic pathology.

We developed the classifiers on a dataset consisting exclusively of primary tumor samples and applied them to metastases with good results. This supports the assumption that the epigenetic profiles of primary HNSC tumors are preserved in their metastases. Similar results from other publications indicate that this might be true for a large class of tumor entities [15–17]. Further, we found no significant differences in the overall accuracies of all four classifiers on the subset of cervical lymph node metastases compared with the pulmonary metastases in the validation set. A strength of this study is that the cervical lymph node metastases in the validation set of this study were in fact initially diagnosed as CUPs. CUPs are hypothesized to be distinct tumor entities that differ from their non-CUP counterparts in biological characteristics relating, for example, to early dissemination and slow growth of the primary [42]. This may be reflected in their DNA methylation profiles and make their classification more challenging since the machine learning models were trained on primary tumors. The performance of the classifiers on this subset indicates that they are robust to this potential effect.

The second aspect of this study is the comparison of NN, SVM, LOGREG, and RF models for the classification of HNSC tumors based on their DNA methylation profiles. In most publications on similar classification tasks, a single RF model with additional calibration methods has been used successfully [14–16,18]. However, in a comparison of NN, SVM, and RF models for the task of distinguishing pulmonary HNSC metastases from primary SCC of the lung, the RF model performed considerably worse than the other two classifiers [17]. A comparison of a multitude of methods for the classification of central nervous system tumor entities showed, among other results and put in simplified terms, that a LOGREG model performed best of all methods and also substantially better than an RF model [19]. In the present study, the RF model also had the lowest overall accuracy (83%) on the validation set by a considerable margin, whereas the other methods achieved similar results (NN = 88%, SVM = LOGREG = 89%). However, the difference in classifier performance on the validation set of our study was not statistically significant. In the binary classification task for oropharyngeal primary sites, all models but RF achieved significantly better accuracies than a purely p16-based prediction. Thus, in line with the result of the previously reported comparisons, NN, SVM, and LOGREG models show advantages over the RF model for the classification task in this study.

A limitation of this study lies in the classes of primary sites that the models were trained on. First, nasopharyngeal HNSC samples were not present in the dataset and thus cannot be predicted by the classifiers. This is due to the fact that no DNA methylation datasets of nasopharyngeal HNSC cases are currently publicly available, and patients do not present at our institutions in sufficient numbers since they are rare outside of Asia (<1% of new HNSC cases in 2012) [43]. Second, we used a merged hypopharynx and larynx class because the number of available hypopharyngeal HNSC samples was too small. However, due to the anatomical adjacency of these two regions, a correct prediction of the combined class still provides helpful information in a clinical setting, e.g. by narrowing the irradiation target, guiding imaging techniques, and enabling targeted biopsies of likely tumor sites. When more data for naso- or hypo-pharyngeal HNSCs become available, the classifiers can be extended to predict their primary site without a change of methodology. Further analysis of the classification results revealed that, for all classifiers, HPV/p16-negative oropharynx samples are classified with lower accuracy than p16-positive ones and make up a large proportion of the total misclassifications. HPV status was shown to affect the DNA methylation profiles of oropharyngeal HNSCs [21,44], which makes the classification of this entity more challenging. Finally, the different DNA methylation patterns of HNSCs could be linked not only to tumor localization but also to aspects of tumor biology that have prognostic or therapeutic value. Response and follow-up data were not available here, but prospective analyses of prognosis and therapy response based on DNA methylation within clinical studies are promising future research.

In conclusion, we demonstrate that a classification of the primary region of HNSC metastases by their DNA methylation profiles is possible and provide a set of classifiers that achieve good results on a validation cohort representative of possible applications. Further, we conclude that the NN, SVM, and LOGREG models are more suitable for this classification task than the RF model. The classifiers, which have been made publicly available, can be a useful tool in the search for the primary tumors of HNSC-CUPs.

Acknowledgements

We gratefully acknowledge the expert technical assistance of Peggy Wolkenstein, Ines Koch, Daniel Teichmann, Anne Reichstein, and Carola Geiler. The results shown here are, in part, based on data generated by the TCGA Research Network. Parts of Figure 1 were created with BioRender.com (RRID: SCR_018361). Michael Bockmayr is a fellow of the Mildred Scheel Cancer Career Center Hamburg/Deutsche Krebshilfe.

MB was supported in part by University Medical Center Hamburg-Eppendorf. PJ is a participant in the Berlin Institute of Health (BIH) Charité Digital Clinician Scientist Program funded by the Charité – Universitätsmedizin Berlin, the Berlin Institute of Health, and the German Research Foundation (DFG). KRM was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korean Government (No. 2017-0-00451, Development of BCI Based Brain and Cognitive Computing Technology for Recognizing User's Intentions using Deep Learning, and No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University); by the German Ministry for Education and Research (BMBF) under

Grants 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18025A and 01IS18037A; and by the German Research Foundation (DFG) under Grant Math+, EXC 2046/1, Project ID 390685689. PS was supported by the German Ministry for Education and Research (BMBF) under Grant Patho234 031L0207D. US was supported by the Fördergemeinschaft Kinderkrebszentrum Hamburg. Additional funding was provided by the German Cancer Consortium (DKTK), partner site Berlin.

Author contributions statement

ML, PJ, MB and FK were responsible for the study concept and design. ML, PJ, MB, KRM and FK developed the methodology. ML performed formal analysis of the data. All the authors contributed to the investigation. ML, DC, US, FK, PJ and MB provided resources. ML, PJ, and MB performed data curation. ML wrote the original manuscript draft. All the authors assisted with review and revision of the paper. ML and PJ produced the figures. FK, PJ and MB provided supervision. KRM, FK and MB acquired funding for the project. All the authors read and approved the final version of the manuscript.

Data availability statement

IDAT files of the samples which have been newly analyzed in the course of this research project (study dataset) have been deposited at the GEO repository GSE171994. All other IDAT files considered in this paper have been previously published and can be found at the GEO repositories GSE87053, GSE95036, and GSE124052, and in the TCGA HNSC dataset. The R code to reproduce the main analyses from this paper is available on figshare (https://figshare.com/s/fb62d93b87d3da08681e).

References

1. Sung H, Ferlay J, Siegel RL, *et al*. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021; **71:** 209–249.
2. Müller von der Grün J, Tahtali A, Ghanaati S, *et al*. Diagnostic and treatment modalities for patients with cervical lymph node metastases of unknown primary site – current status and challenges. *Radiat Oncol* 2017; **12:** 82.
3. Pavlidis N, Pentheroudakis G, Plataniotis G. Cervical lymph node metastases of squamous cell carcinoma from an unknown primary site: a favourable prognosis subset of patients with CUP. *Clin Transl Oncol* 2009; **11:** 340–348.
4. Davis KS, Byrd JK, Mehta V, *et al*. Occult primary head and neck squamous cell carcinoma: utility of discovering primary lesions. *Otolaryngol Head Neck Surg* 2014; **151:** 272–278.
5. Haas I, Hoffmann TK, Engers R, *et al*. Diagnostic strategies in cervical carcinoma of an unknown primary (CUP). *Eur Arch Otorhinolaryngol* 2002; **259:** 325–333.
6. Maghami E, Ismaila N, Alvarez A, *et al*. Diagnosis and management of squamous cell carcinoma of unknown primary in the head and neck: ASCO guideline. *J Clin Oncol* 2020; **38:** 2570–2596.
7. Geltzeiler M, Doerfler S, Turner M, *et al*. Transoral robotic surgery for management of cervical unknown primary squamous cell carcinoma: updates on efficacy, surgical technique and margin status. *Oral Oncol* 2017; **66:** 9–13.
8. Binder A, Bockmayr M, Hägele M, *et al*. Morphological and molecular breast cancer profiling through explainable machine learning. *Nat Mach Intell* 2021; **3:** 355–366.
9. Campanella G, Hanna MG, Geneslaw L, *et al*. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019; **25:** 1301–1309.
10. Klauschen F, Müller KR, Binder A, *et al*. Scoring of tumor-infiltrating lymphocytes: from visual estimation to machine learning. *Semin Cancer Biol* 2018; **52:** 151–157.
11. Seegerer P, Binder A, Saitenmacher R, *et al*. Interpretable deep neural network to predict estrogen receptor status from haematoxylin–eosin images. In *Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-Art and Future Challenges. Lecture Notes in Computer Science*, Holzinger A, Goebel R, Mengel M, *et al*. (eds). Springer International Publishing: Cham, 2020; 16–37.
12. Stenzinger A, Alber M, Allgäuer M, *et al*. Artificial intelligence and pathology: from principles to practice and future applications in histomorphology and molecular profiling. *Semin Cancer Biol* 2021. https://doi.org/10.1016/j.semcancer.2021.02.011.
13. Pfeifer B, Saranti A, Holzinger A. Network module detection from multi-modal node features with a greedy decision forest for actionable explainable AI. arXiv.org 2021; 2108.11674 [Not peer reviewed].
14. Capper D, Jones DTW, Sill M, *et al*. DNA methylation-based classification of central nervous system tumours. *Nature* 2018; **555:** 469–474.
15. Hackeng WM, Dreijerink KMA, de Leng WWJ, *et al*. Genome methylation accurately predicts neuroendocrine tumor origin: an online tool. *Clin Cancer Res* 2021; **27:** 1341–1350.
16. Moran S, Martínez-Cardús A, Sayols S, *et al*. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *Lancet Oncol* 2016; **17:** 1386–1395.
17. Jurmeister P, Bockmayr M, Seegerer P, *et al*. Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Sci Transl Med* 2019; **11:** eaaw8513.
18. Jurmeister P, Schöler A, Arnold A, *et al*. DNA methylation profiling reliably distinguishes pulmonary enteric adenocarcinoma from metastatic colorectal cancer. *Mod Pathol* 2019; **32:** 855–865.
19. Maros ME, Capper D, Jones DTW, *et al*. Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data. *Nat Protoc* 2020; **15:** 479–512.
20. Basu B, Chakraborty J, Chandra A, *et al*. Genome-wide DNA methylation profile identified a unique set of differentially methylated immune genes in oral squamous cell carcinoma patients in India. *Clin Epigenetics* 2017; **9:** 13.
21. Degli Esposti D, Sklias A, Lima SC, *et al*. Unique DNA methylation signature in HPV-positive head and neck squamous cell carcinomas. *Genome Med* 2017; **9:** 33.
22. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002; **30:** 207–210.
23. Clark K, Vendt B, Smith K, *et al*. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging* 2013; **26:** 1045–1057.
24. Zuley ML, Jarosz R, Kirk S, *et al*. Radiology Data from The Cancer Genome Atlas Head–Neck Squamous Cell Carcinoma [TCGA-HNSC] collection. The Cancer Imaging Archive. 2016. http://doi.org/10.7937/K9/TCIA.2016.LXKQ47MS.
25. World Health Organization. In *International Classification of Diseases for Oncology (ICD-O)* (3rd edn, 1st Revision edn), Fritz A,

Percy C, Jack A, *et al*. (eds). World Health Organization, 2013. https://apps.who.int/iris/handle/10665/96612.
26. Aryee MJ, Jaffe AE, Corrada-Bravo H, *et al*. Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 2014; **30:** 1363–1369.
27. Triche TJ Jr, Weisenberger DJ, Van Den Berg D, *et al*. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res* 2013; **41:** e90.
28. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res* 2017; **45:** e22.
29. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw* 2008; **28:** 1–26.
30. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008; **9:** 2579–2605.
31. Johann PD, Jäger N, Pfister SM, *et al*. RF_Purify: a novel tool for comprehensive analysis of tumor-purity in methylation array data based on random forest regression. *BMC Bioinformatics* 2019; **20:** 428.
32. Qin Y, Feng H, Chen M, *et al*. InfiniumPurify: an R package for estimating and accounting for tumor purity in cancer methylation research. *Genes Dis* 2018; **5:** 43–45.
33. Begum S, Gillison ML, Ansari-Lari MA, *et al*. Detection of human papillomavirus in cervical lymph nodes: a highly effective strategy for localizing site of tumor origin. *Clin Cancer Res* 2003; **9:** 6469–6475.
34. El-Mofty SK, Zhang MQ, Davila RM. Histologic identification of human papillomavirus (HPV)-related squamous cell carcinoma in cervical lymph nodes: a reliable predictor of the site of an occult head and neck primary carcinoma. *Head Neck Pathol* 2008; **2:** 163–168.
35. Gillison ML. Human papillomavirus-associated head and neck cancer is a distinct epidemiologic, clinical, and molecular entity. *Semin Oncol* 2004; **31:** 744–754.
36. Pillai R, Deeter R, Rigl CT, *et al*. Validation and reproducibility of a microarray-based gene expression test for tumor identification in formalin-fixed, paraffin-embedded specimens. *J Mol Diagn* 2011; **13:** 48–56.
37. Erlander MG, Ma XJ, Kesty NC, *et al*. Performance and clinical evaluation of the 92-gene real-time PCR assay for tumor classification. *J Mol Diagn* 2011; **13:** 493–503.
38. Tothill RW, Kowalczyk A, Rischin D, *et al*. An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. *Cancer Res* 2005; **65:** 4031–4040.
39. Bockmayr T, Erdmann G, Treue D, *et al*. Multiclass cancer classification in fresh frozen and formalin-fixed paraffin-embedded tissue by DigiWest multiplex protein analysis. *Lab Invest* 2020; **100:** 1288–1299.
40. Zhang PW, Chen L, Huang T, *et al*. Classifying ten types of major cancers based on reverse phase protein array profiles. *PLoS One* 2015; **10:** e0123147.
41. Bloom GC, Eschrich S, Zhou JX, *et al*. Elucidation of a protein signature discriminating six common types of adenocarcinoma. *Int J Cancer* 2007; **120:** 769–775.
42. Rassy E, Assi T, Pavlidis N. Exploring the biological hallmarks of cancer of unknown primary: where do we stand today? *Br J Cancer* 2020; **122:** 1124–1132.
43. Pezzuto F, Buonaguro L, Caponigro F, *et al*. Update on head and neck cancer: current knowledge on epidemiology, risk factors, molecular features and novel therapies. *Oncology* 2015; **89:** 125–136.
44. van Kempen PM, Noorlag R, Braunius WW, *et al*. Differences in methylation profiles between HPV-positive and HPV-negative oropharynx squamous cell carcinoma. *Epigenetics* 2014; **9:** 194–203.
45. Ritchie ME, Phipson B, Wu D, *et al*. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015; **43:** e47.

Reference 45 is cited only in supplementary material.

SUPPLEMENTARY MATERIAL ONLINE

Figure S1. Effect of CpG site number on cross-validation error

Figure S2. Comparison of tumor purity estimation methods on the reference cohort

Table S1. Clinical data for cases in the study dataset

Table S2. Primary region and primary organ annotations for all samples

Table S3. List of 2000 CpG sites with highest variance in the reference cohort

Table S4. Parameters for classifier development per model type

Table S5. Effects of batch correction on cross-validation metrics on the reference cohort