Genomic and epigenomic profiles distinguish pulmonary enteric adenocarcinoma from lung metastatic colorectal cancer

Ying Zuo,[a,1] Jia Zhong,[a,1] Hua Bai,[a,1] Bin Xu,[b,1] Zhijie Wang,[a] Weihua Li,[c] Yedan Chen,[d] Shi Jin,[e] Shuhang Wang,[f] Xin Wang,[a] Rui Wan,[a] Jiachen Xu,[a] Kailun Fei,[a] Jiefei Han,[g] Zhenlin Yang,[h] Hua Bao,[d] Yang Shao,[d,i] Jianming Ying,[c] Qibin Song,[b]** Jianchun Duan,[a]* and Jie Wang [a]*

[a]State Key Laboratory of Molecular Oncology, Department of Medical Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, China
[b]Cancer center, Renmin Hospital of Wuhan University, Wuhan, China
[c]Department of Pathology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, China
[d]Nanjing Geneseeq Technology Inc., Nanjing, China
[e]National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen, 518116, China
[f]GCP Center, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, China
[g]Department of Neuro-oncology, Cancer Center Beijing Tiantan Hospital, Capital Medical University, China
[h]Thoracic Surgery Department, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, China
[i]School of Public Health, Nanjing Medical University, Nanjing, China

Summary

Background As a rare subtype of lung adenocarcinoma, the diagnosis of pulmonary enteric adenocarcinoma (PEAC) remains challenging due to overlapping morphologic spectrum with lung metastatic colorectal cancer (lmCRC). However, the molecular features of PEAC as a separate lung cancer entity are poorly understood.

Methods We performed whole-exome sequencing and targeted bisulfite sequencing of 32 PEAC and 30 lmCRC to improve differential molecular characterization of the two diseases. We used machine learning methods to select key markers and developed a diagnostic classifier. In addition, we validated the classifier in the internal test cohort and an independently recruited external validation cohort with 17 PEAC and 7 lmCRC.

Findings Our results showed that *EGFR* was the key driver mutation in PEAC but at a lower prevalence compared to typical lung adenocarcinomas, whereas *ERBB2* and *KRAS* were more frequently observed in PEAC. By contrast, we observed significant enrichment of *KRAS* and *APC* mutations in lmCRC compared with PEAC. At the chromosome arm level, copy number variations in 13q, 14q, and 18p were the major chromosomal differences observed between PEAC and lmCRC. Furthermore, by comparing differentially methylated regions (DMRs), we established a neat DNA methylation-based classifier consisting of eight DMRs. This classifier correctly classified all samples in the training cohort and 95% of the samples in the internal test cohort. An external validation cohort of 24 cases recruited from multiple centers in China also reliably agreed with pathological diagnosis.

Interpretation These results provide solid evidence of PEAC-specific genomic characteristics and demonstrate the potential utility of DNA methylation markers for auxiliary diagnosis of PEAC and lmCRC.

eBioMedicine 2022;82: 104165
Published online xxx
https://doi.org/10.1016/j.ebiom.2022.104165

*Corresponding authors at: Jie Wang & Jianchun Duan, State Key Laboratory of Molecular Oncology, Department of Medical Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, 17 Pan-jia-yuan South Lane, Chaoyang District, Beijing, 100021, China.
**Corresponding author at: Qibin Song, Cancer center, Renmin Hospital of Wuhan University, 99#, Zhangzhidong Road, Wuchang district, Wuhan, Hubei Province, 430060, China.
E-mail addresses: zlhuxi@163.com (J. Wang), duanjianchun79@163.com (J. Duan), qibinsong@whu.edu.cn (Q. Song).

[1] These authors contributed equally to this work.

Funding This work was supported by National key research and development project 2019YFC1315700, CAMS Key Laboratory of Translational Research on Lung Cancer (2018PT31035), and Beijing Natural Science Foundation (7222144).

Keywords: Pulmonary enteric adenocarcinoma; Lung metastatic colorectal cancer; Machine learning model

Research in context

Evidence before this study

Pulmonary enteric adenocarcinoma (PEAC) is a rare subtype of lung cancer that presents intestinal morphology, thereby challenging the pathological diagnosis from lung metastatic colorectal cancer (lmCRC). Previous studies comparing PEAC and lmCRC mainly focused on selected drivers or hotspot mutations and were not conclusive for diagnostic differentiation. To date, no deep and comprehensive analysis of broad genomic areas has been performed to characterize PEAC. An important earlier study aiming to distinguish the two diseases used array-based DNA methylation data from public datasets of primary lung and colorectal cancers. In this study, we described PEAC-specific genomic and epigenomic features based on whole-exome sequencing and targeted bisulfite sequencing of PEAC samples to better understand this rare lung cancer entity.

Added value of this study

Our study is by far the largest cohort study to characterize PEAC. We present an in-depth description and statistical analyses of somatic mutations, copy number alterations, and DNA methylation profiles of PEAC and lmCRC and demonstrated subtype-specific features in comparison with population-matched public lung adenocarcinoma data. Furthermore, we used machine learning algorithms and carefully screened features from genomic alterations and differentially methylated regions (DMRs). Our resulting classifier is a neat eight DMR-based model that achieved above 95% accuracy in the training and internal test cohorts. The performance of this classifier was confirmed through an independently recruited multicenter validation cohort and public datasets.

Implications of all the available evidence

In 2021, the World Health Organization updated the diagnostic criteria of PEAC based on the expression of intestinal IHC markers, which greatly improved the clarity of PEAC's diagnosis. However, PEACs with positive intestinal markers but without TTF-1 or CK-7 expression can only be diagnosed by clinical exclusion of lmCRC. Therefore, an objective binary classifier is necessary to clarify ambiguous cases from pathology assessments. In addition to the rigorous characterization of PEAC-specific molecular features, our methylation-based classifier has demonstrated great potential in clinical utility to facilitate clinical decisions.

Introduction

Pulmonary enteric adenocarcinoma (PEAC) is a rare subtype of primary invasive lung adenocarcinoma that occurs in approximately 0.6% of pulmonary adenocarcinomas.[1] PEAC was first described in 1991 by Tsao and Fraser[2] and first recommended as an official adenocarcinoma classification by the International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society (IASLC/ATS/ERS) in 2011. Its diagnostic criteria were subsequently proposed by the World Health Organization (WHO) in 2015.[3,4] PEAC has been defined as a subtype of primary pulmonary adenocarcinoma with a predominant (>50%) intestinal epithelial-like component, showing either enteric differentiation immunohistochemical (IHC) markers or enteric morphology.[3] The 2021 WHO Classification has been updated to clearly define the IHC criteria for diagnosing PEAC.[5] The essential diagnostic criteria include the expression of at least one intestinal marker (CDX-2, cytokeratin 20 (CK20), HNF4α or MUC2), more than 50% of tumor histology resembling enteric morphology, and clinical exclusion of colorectal carcinoma. The desirable criteria include coexpression of thyroid transcription factor-1 (TTF-1) or CK7. However, due to the presentation of enteric features, the differential diagnosis of PEAC and lung metastatic colorectal cancer (lmCRC) is still challenging in pathology practice. An accurate pathological diagnosis is crucial for personalized primary lesion-specific therapy, prognostic evaluation, and prolonged survival. Patients with PEAC, especially those without metastasis, have an opportunity to undergo curative therapy, including radical surgery. In contrast, lmCRC is an advanced-stage disease mainly treated with palliative therapy.

Currently, the differential diagnosis of PEAC and lmCRC mainly relies on the clinical history, tumor site and pathological examination. For example, key IHC

markers of typical lung adenocarcinoma (tLUAD), i.e., TTF-1 and CK7, could be helpful for distinguishing PEAC from lmCRC.[4,6] However, expression of these markers is frequently lost in PEAC but positive in a small proportion of colorectal cancers.[7,8] Some intestinal differentiation markers, such as CDX-2, CK20, MUC2 and HNF4α, could also facilitate clinical differential diagnosis; however, they are inconsistently expressed across studies concerning PEAC,[7,9–11] and can both be positive in PEAC and lmCRC. To increase the accuracy of IHC-based diagnosis, many additional markers have been evaluated, including CDH17, SATB2, β-Catenin and Villin, but the sensitivity and specificity are still under investigation.[12,13] Therefore, there is an urgent need to identify reliable biomarkers for the accurate diagnosis of PEAC and lmCRC.

Current investigations of the genomic features of PEAC mainly focused on key driver genes, such as *EGFR* and *KRAS*, through hotspot or targeted next-generation sequencing,[12,14,15] which provided limited clues for differential diagnosis. The comprehensive genomic profiles of PEAC are still unknown. In addition to altered genomic events, aberrant epigenetic events, such as DNA methylation in promoter regions, are frequently observed in cancer cells.[16] DNA methylation is relatively stable and displays tissue-specific patterns, rendering it useful for identifying tumors of unknown origin or pathologically similar subtypes.[15,17–20]

In this study, we explored the comprehensive genomic and epigenomic profiles of PEAC by whole-exome sequencing (WES) and targeted bisulfite sequencing, and identified characteristic molecular events in PEAC. Based on DNA methylation profiles of PEAC and lmCRC, we also constructed a classifier to distinguish the two diseases, which was well confirmed in a multicenter external validation cohort.

Methods

Clinical cohorts

For molecular profiling and model establishment, 32 patients with PEAC and 30 patients with lmCRC from the Cancer Hospital, Chinese Academy of Medical Sciences (Beijing, China) were included and separated into training and test cohorts based on their time of diagnosis. Specifically, formalin-fixed paraffin-embedded (FFPE) tumor samples of 22 PEAC and 20 lmCRC patients diagnosed between March 2011 and October 2017 were collected as the training cohort, and 10 PEAC and 10 lmCRC patients diagnosed between November 2017 and February 2019 were collected as the test cohort. Matched normal tissue samples were also obtained. For the external multicenter validation, we collected 17 PEAC and 7 lmCRC samples, including samples from 8 PEAC and 6 lmCRC patients prospectively admitted in our center and our affiliated Shenzhen center (Cancer Hospital Chinese Academy of Medical Sciences, Shenzhen Center) between March 2019 and May 2020, and archived FFPE samples of 9 PEAC and 1 lmCRC from the Renmin Hospital of Wuhan University (Table 1). The samples in all three clinical cohorts were diagnosed based on surgical specimens. The IHC staining of the PEAC and lmCRC with sufficient specimens was conducted according to manufacturer's instructions, using IHC antibodies including TTF-1 (RRID: AB_2888646), CK7 (RRID: AB_1658454), CK20 (RRID: AB_1658454), CDX-2 (RRID: AB_2819184), MUC2 (RRID: AB_10578542), and HNF4α (RRID: AB_10818555). Each sample was histologically reassessed by two experienced pathologists according to the classification criteria of the 2021 WHO guidelines. The researchers were blinded to the clinical diagnosis when analyzing the external validation cohort. All PEAC patients underwent either endoscopic evaluation or PET-CT, or endoscopic evaluation plus abdominal CT or MRI for the clinical exclusion of colorectal cancer. The clinical data of these patients, including sex, age at diagnosis, smoking status, and tumor staging (American Joint Committee on Cancer (AJCC) 7th edition) were obtained from the medical records. The study was approved by the Ethics Committees of the Cancer Hospital Chinese Academy of Medical Sciences, the Renmin Hospital of Wuhan University and Cancer Hospital Chinese Academy of Medical Sciences, Shenzhen Center (ethical approval number NCC-008634), and written consent forms were obtained from all patients. For comparison with tLUAD, we obtained whole-exome sequencing data of an East Asian tLUAD cohort from the cBioPortal website (https://www.cbioportal.org/datasets).[21]

Library preparation and sequencing

Genomic DNA was extracted from FFPE samples using the *QIAamp* DNA FFPE Tissue Kit (Qiagen) and fragmented by an M220 Focused-ultrasonicator (Covaris) into approximately 250 bp fragments. A whole-genome library was prepared using the KAPA Hyper Prep Kit (KAPA Biosystems). Whole exome capture was performed using the SureSelect Human All Exon V6 (Agilent Technologies) according to manufacturer's protocol. Captured libraries were amplified with Illumina p5 (50 AAT GAT ACG GCG ACC ACC GA 30) and p7 (50 CAA GCA GAA GAC GGC ATA CGA GAT 30) primers in KAPA HiFi HotStart ReadyMix (KAPA Biosystems), and purified using Agencourt AMPure XP beads. The enriched libraries were quantified by qPCR using the KAPA Library Quantification Kit (KAPA Biosystems) and sequenced using the Illumina HiSeq 4000 platform as paired 125 bp reads. The mean coverage of tumor and normal samples was 104.6X and 57.3X, respectively.

Mutation calling

Trimmomatic was used to trim adaptors with a sliding window quality control to remove low -quality reads

Feature	Training cohort		Test cohort		Validation cohort		Total	
	PEAC	lmCRC	PEAC	lmCRC	PEAC	lmCRC	PEAC (*N*=49)	lmCRC (*N*=37)
Sex — *n/N* (%)								
Male	15/22 (68.2)	12/20 (60.0)	5/10 (50.0)	7/10 (70.0)	12/17 (70.6)	5/7 (71.4)	32/49 (65.3)	24/37 (64.9)
Female	7/22 (31.8)	8/20 (40.0)	5/10 (50.0)	3/10 (30.0)	5/17 (29.4)	2/7 (28.6)	17/49 (34.7)	13/37 (35.1)
Age at diagnosis — yr								
Median	58.5	58	59	66	65	63	60	61
Range	25-82	39-77	30-70	42-76	48-83	55-72	25-83	39-77
Smoker — *n/N* (%)[a]								
Yes	14/22 (63.6)	7/20 (35.0)	4/10 (40.0)	2/10 (20.0)	10/15 (66.7)	3/7 (42.9)	28/47 (59.6)	12/37 (32.4)
No	8/22 (36.4)	13/20 (65.0)	6/10 (60.0)	8/10 (80.0)	5/15 (33.3)	4/7 (57.1)	19/47 (40.4)	25/37 (67.6)
Tumor staging[a,b]								
I	9/22 (40.9)	0/20 (0)	5/10 (50.0)	0/10 (0)	4/15 (26.7)	0/7 (0)	18/47 (38.3)	0/37 (0)
II	3/22 (13.6)	0/20 (0)	4/10 (40.0)	0/10 (0)	4/15 (26.7)	0/7 (0)	11/47 (23.4)	0/37 (0)
III	10/22 (45.5)	0/20 (0)	0/10 (0)	0/10 (0)	6/15 (40.0)	0/7 (0)	16/47 (34.0)	0/37 (0)
IV	0/22 (0)	20/20 (100)	1/10 (10.0)	10/10 (100)	1/15 (6.7)	7/7 (100)	2/47 (4.3)	37/37 (100)
TTF-1 — *n/N* (%)[a]								
Positive	14/22 (63.6)	0/18 (0)	7/10 (70.0)	2/9 (22.2)	7/16 (43.8)	0/6 (0)	28/48 (58.3)	2/33 (6.1)
Negative	8/22 (36.4)	18/18 (100)	3/10 (30.0)	7/9 (77.8)	9/16 (56.3)	6/6 (100)		
CK7 — *n/N* (%)[a]								
Positive	20/21 (95.2)	0/15 (0)	9/9 (100)	0/9 (0)	16/16 (100)	0/7 (0)	45/46 (97.8)	0/31 (0)
Negative	1/21 (4.8)	15/15 (100)	0/9 (0)	9/9 (100)	0/16 (0)	7/7 (100)		
CK20 — *n/N* (%)[a]								
Positive	16/22 (72.7)	18/18 (100)	7/10 (70.0)	9/9 (100)	9/15 (60.0)	7/7 (100)	32/47 (68.1)	34/34 (100)
Negative	6/22 (27.3)	0/18 (0)	3/10 (30.0)	0/9 (0)	6/15 (40.0)	0/7 (0)		
CDX-2 — *n/N* (%)[a]								
Positive	18/22 (81.8)	18/18 (100)	7/10 (70.0)	9/9 (100)	11/16 (68.8)	7/7 (100)	36/48 (75.0)	34/34 (100)
Negative	4/22 (18.2)	0/18 (0)	3/10 (30.0)	0/9 (0)	5/16 (31.3)	0/7 (0)		
MUC2 — *n/N* (%)[a]								
Positive	7/12 (58.3)	14/16 (87.5)	6/6 (100)	8/8 (100)	3/11 (27.3)	7/7 (100)	16/29 (55.2)	29/31 (93.5)
Negative	5/12 (41.7)	2/16 (12.5)	0/6 (0)	0/8 (0)	8/11 (72.7)	0/7 (0)		
HNF4α — *n/N* (%)[a]								
Positive	20/20 (100)	16/16 (100)	7/7 (100)	10/10 (100)	6/6 (100)	6/6 (100)	33/33(100)	32/32 (100)
Negative	0/20 (0)	0/16 (0)	0/7 (0)	0/10 (0)	0/6 (0)	0/6 (0)		

***Table 1*: Baseline characteristics of PEAC and lmCRC patients in the training, test and multicenter validation cohorts.**

[a] Some patients' information was unavailable due to unattained medical records or insufficient quantity of samples.

[b] American Joint Committee on Cancer (AJCC) 7th edition.

(quality reading below 20) and N bases from the FASTQ files. Clean paired-end reads were then aligned to the reference human genome (build hs37d5) using the Burrows-Wheeler Aligner (BWA), and PCR deduplication was performed using Picard. GATK3 was used to perform indel local realignment and base quality-score recalibration. The matching of tumor and normal sample pairs was confirmed for the same single nucleotide polymorphism (SNP) fingerprint using VCF2LR (GeneTalk). Subsequently, samples with a mean depth <30X after removing deduplicate reads were removed. Cross-sample contamination was estimated using ContEst (Broad Institute) by evaluating the likelihood of detecting alternate alleles of SNPs reported in the 1000G database. Somatic SNVs and insertions/deletions (INDELs) were called using VarDict (Ver 1.5.4). The SNVs and INDELs of the protein-coding genes were further filtered using the following criteria: i) minimum ≥4 variant supporting reads and ≥2% variant allele frequency (VAF) supporting the variant; ii) removed if present in >1% population frequency in the 1000G or ExAC database; and iii) filtered through an internally collected list of recurrent sequencing errors (≥3 variant reads and ≤20% VAF in at least 30 of ~2000 normal samples) on the same sequencing platform. The final mutations were annotated using vcf2maf. The tumor mutational burden (TMB) was estimated from the total number of missense mutations in the 32Mb human genome coding region. For the mutational signature analysis, both synonymous and nonsynonymous single base substitutions (SBSs) were extracted and mapped to the 72 mutational signatures from the Catalog of Somatic Mutation in Cancer (COSMIC) database with the R package sigminer (v1.2.1).[22,23] The signatures

Location (GRCh37/hg19)	DMR length	CpG Island/Shore/Shelf	Methylation Status	Promoter	Genes body (Genetic element)
chr6:10555801-10556300	500	-	Hyper	GCNT2	GCNT2(UTR)
					GCNT2(exon)
chr17:46707701-46707900	200	Shelf	Hyper	-	HOXB7(transcript)
					HOXB-AS4 (transcript)
chr17:63554501-63554600	100	Shore	Hyper	-	AC004805.1(UTR)
					AC004805.1(exon)
					AXIN2(CDS)
					AXIN2(exon)
chr17:46697501-46697700	200	Shore	Hyper	-	HOXB7(transcript)
chr7:27178801-27179600	800	Shore	Hyper	-	AC004080.6 (transcript)
					HOXA3(transcript)
					HOXA-AS3 (transcript)
chr21:40195001-40195200	200	Shore	Hyper	-	ETS2(UTR)
					ETS2(exon)
chr2:10445001-10445100	100	Shore	Hyper	-	HPCAL1(transcript)
chr19:30162701-30162800	100	Shore	Hypo	-	PLEKHF1 (transcript)

***Table 2*: Chromosomal locations, reference genes and genetic elements corresponding to the eight DMRs, and the methylation status of PEAC relative to lmCRC of each DMR.**
Hypo, hypomethylated; Hyper, hypermethylated; DMR, differentially methylated region.

were grouped into 10 categories according to their proposed aetiologies.

Identification of copy number alterations

Copy number analysis of the WES data was performed using FACETS (Ver 0.5.13). Copy number alteration (CNA) events for amplification or deletion were assigned based on the sample-ploidy adjusted copy number calculated by the FACETS algorithm as previously described.[24] In brief, chromosome arm-level CNA gain was identified if the segments of gain accounted for more than 60% of the total segments of the corresponding chromosome arm. Arm-level CNA loss was identified if the segments of loss accounted for more than 60% of the total segments of the given chromosome. For focal CNA, segments contributing to deep amplification and deep deletion events were considered for analysis. Fisher's exact test was used to compare arm-level CNA differences with FDR adjustment of the p values. Significantly amplified or deleted focal CNA regions in each cancer were identified using the Genome Identification of Significant Targets in Cancer (GISTIC, v.2.0) algorithm with slightly relaxed parameters considering our sample size as follows: q-value<0.25, log2 ratio=0.2, broad = 1, brlen = 0.6, and genegistic = 1.

Bisulfite treatment and targeted bisulfite sequencing

The extracted genomic DNA was fragmented as previously described. After end repair and A-tailing, adapters with methylated cytosines were ligated onto each DNA fragment. Bisulfite treatment of unmethylated DNA was then performed using the EZ DNA Methylation-Lightning Kit (Zymo Research). A bisulfite-converted DNA library was constructed using the Accel-NGS Methyl-Seq DNA Library Kit for Illumina platforms (Swift Biosciences) and hybridized with the SeqCap Epi CpGiant probe pool from the SeqCap Epi CpGiant Enrichment Kit (ROCHE). Noncomplementary library fragments were washed away, leaving a recovered library of interest targeting over 5.5 million CpGs.[25] The post-capture library was amplified using LM-PCR oligos, and the enriched libraries were sequenced using the Illumina HiSeq 4000 platform with a mean effective coverage of approximately 38.9X.

DNA methylation marker prescreening

Differentially methylated regions (DMRs) were analyzed using the methylKit package in R (v1.2.0),[26] the CpG clusters were initially divided into 100 bp windows with a sliding step of 100 bp based on in-house assay validation. The methylation level of each DMR was determined by dividing the total methylated cytosines by the total number of CpGs within each window. A relatively stringent cutoff of a minimum 0.2 methylation difference was used for DMR calling with q-value-based false discovery rate (FDR) at a 0.05 significance level controlled by the Sliding Linear Model (SLIM). DMRs with low sequencing depth and low-quality reads were filtered. The promoter region was defined as 1.5 kb upstream and 500 kb downstream of the transcription

start site (TSS) and the genetic elements were annotated according to the GENCODE gene annotation file (gencode.v29lift37.annotation.gtf). The DMRs were prescreened by comparing PEAC with lmCRC and with normal lung tissue in the training cohort. Contiguous DMRs from the prescreened DMRs were subsequently joined as a single DMR, resulting in 204 DMRs of 100 bp to 800 bp long as candidate markers for feature selection.

Establishment and validation of the diagnostic classifier

Starting with 711 genes mutated in either cancer type, 77 arm-level CNA events or 204 characteristic DMRs in the training cohort for each feature category, we first applied random forest (RF) algorithm-based recursive feature elimination with ten-fold cross validation to select the best sizes of variable subsets. The overall ranking of the variables was calculated by the summed importance rank from 200 repeats. This method suggested 11 DMRs, 37 CNAs or one genetic mutation (*APC* mutation), respectively as best subsets. The second method we employed was the Least Absolute Shrinkage and Selection Operator (LASSO). One standard error of minimum lambda from a 10-fold cross-validation was used to minimize the loss of the mean standard error and provide the most stringent feature selection (for DMR and mutation variables). Alternatively, minimum lambda was chosen for tuning the CNA features, as using the more parsimonious lambda would leave us with zero variable. Eight DMRs, 10 CNAs or *APC* mutation were kept for each category of features in this method.

Overlapping features obtained from the two methods, resulting in the *APC* mutation alone, 10 CNAs or 8 DMRs from each feature category, were then used for the development of mutation-based, CNA-based, or DMR-based binary classification models, respectively, by random forest. To further verify the optimal DMR subset, we also trained a support vector machine (SVM) model with linear kernel setting in the R package e1071,[27] and an XGBoost model with binary logistic objective in the R package XGBoost[28] as methylation classifiers. For XGBoost, a prediction probability between zero (PEAC) and one (lmCRC) was calculated for each sample, and the final classification was determined by its closeness to 0 or 1, i.e., using 0.5 as the cutoff. The performance of the classifiers was evaluated by their sensitivity, specificity and overall accuracy.

For cross-platform evaluation of the selected DMR features, 503 tLUAD samples and 453 CRC samples profiled using the Illumina Infinium HumanMethylation450 BeadChip array were accessed through The Cancer Genome Atlas (TCGA) database. Additionally, 14 PEAC and 4 lmCRC samples profiled using the Illumina Infinium MethylationEPIC BeadChip (850K) array were obtained from the Gene Expression Omnibus (GEO) repository (GSE116699).[15] The methylation levels of the 204 prescreened DMRs in these public datasets were estimated by averaging methylation of CpG that fell within each DMR.

Statistical analysis

All statistical analyses and data visualization were performed using R (version 3.6.3), including the randomForest,[29] e1071,[27] XGBoost,[28] caret,[30] maftools,[31] gplots (heatmap.2)[32] and ggplot2[33] packages. Sample clustering on heatmaps was measured using Euclidean distances. Fisher's exact test was used to evaluate differences in mutations or CNAs between the PEAC and lmCRC cohorts. The Wilcoxon rank-sum test was used to assess the statistical significance of each DMR. Two-tailed P values <0.05 were considered indicative of statistical significance.

Role of funding sources

The funders only provided funding, and had no role in the study design, data collection, data analysis, interpretation, and writing of the report.

Results

Clinicopathological features of the patients

In this study, we retrospectively collected archived FFPE samples of 32 PEAC and 30 lmCRC patients from the Cancer Hospital, Chinese Academy of Medical Sciences (Beijing, China) for molecular biomarker discovery. The diagnosis of all cases was confirmed based on the 2021 WHO criteria to ensure accurate classifier establishment. We also collected a multicenter external cohort of 17 PEAC and 7 lmCRC patients for classifier validation. Figure 1 illustrates the design of this study.

Our three cohorts had similar clinical characteristics. Between the two cancer types, there were significantly more smokers among the PEAC patients than among lmCRC (p=0.016 by Fisher's exact test). Other clinical factors, including age and sex, were equally distributed (Table 1). Although TTF-1 is a primary IHC marker of lung cancers, two lmCRC patients (2/33, 6.1%) displayed weak TTF-1 staining. However, these two patients had confirmed colorectal tumors and all lmCRC patients were negative for CK7. Of the intestinal IHC markers, median to strong staining of CK20 (34/34, 100%), CDX-2 (34/34, 100%), and HNF4α (32/32, 100%) was observed in all patients, and weak to median staining of MUC2 was observed in 29/31 patients (93.5%). Conversely, the expression of these IHC markers varied among PEAC samples, indicating the clinical difficulty of its diagnosis. For PEAC, TTF-1 was positive in 28/48 (58.3%) patients and CK7 was positive in 45/46 (97.8%) patients, whereas positive staining of

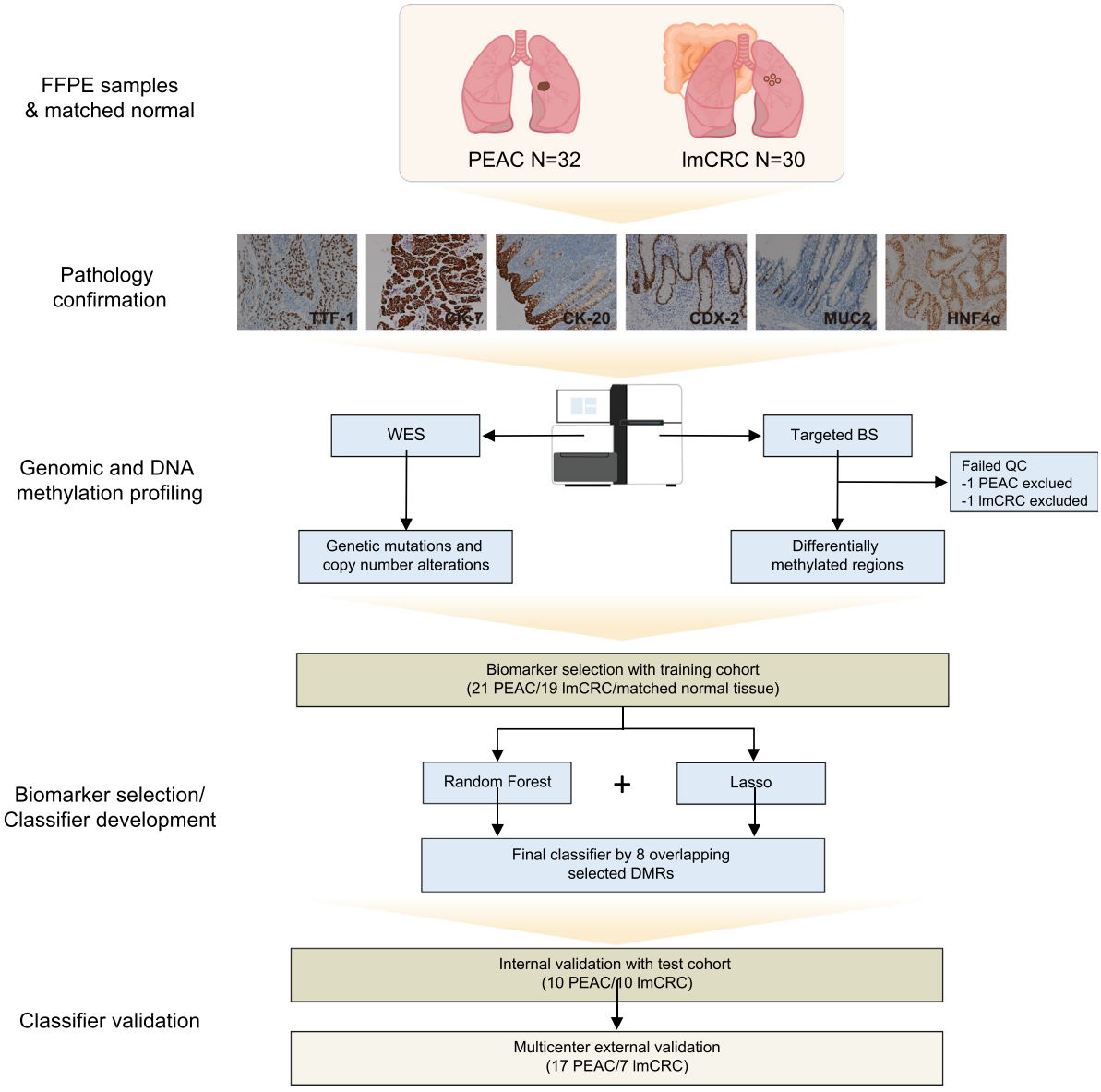


Figure 1. Graphical summary of the study design and data analyses. FFPE tumor samples from 22 patients with PEAC and 20 patients with lmCRC were obtained for pathological confirmation. Whole-exome sequencing and targeted bisulfite sequencing data were performed to compare and screen genetic mutation, CNA, and DNA methylation markers. One PEAC sample and one lmCRC sample in the training cohort failed the quality control for targeted bisulfite sequencing possibly due to insufficient sample input and were removed from subsequent analyses. Random forest (RF) and Least Absolute Shrinkage and Selection Operator (LASSO) were applied to select important markers and construct the diagnostic classifiers. These classifiers were tested in an independent test cohort of 10 PEAC patients and 10 lmCRC patients, and validated in an independent multicenter external cohort for diagnostic prediction. Pathology confirmation is illustrated with pictures of positive controls of each IHC marker. PEAC, pulmonary enteric adenocarcinoma; lmCRC, lung metastatic colorectal cancer; WES, whole exome sequencing; BS, bisulfite sequencing; LASSO, least absolute shrinkage and selection operator. QC, quality control.

CK20, CDX-2, MUC2, and HNF4α was observed in 32/47 (68.1%), 36/48 (75%), 16/29 (55.2%), and 33/33 (100%) patients, respectively, albeit the overall staining of these intestinal markers in PEAC was weaker than that in lmCRC (Figure 2, Table 1, and Table S1).

Somatic DNA alterations

To explore molecular biomarkers to differentiate the two diseases, we first performed WES using DNA extracted from the tumor and matched normal tissue samples of each patient. In PEAC, the frequently mutated cancer

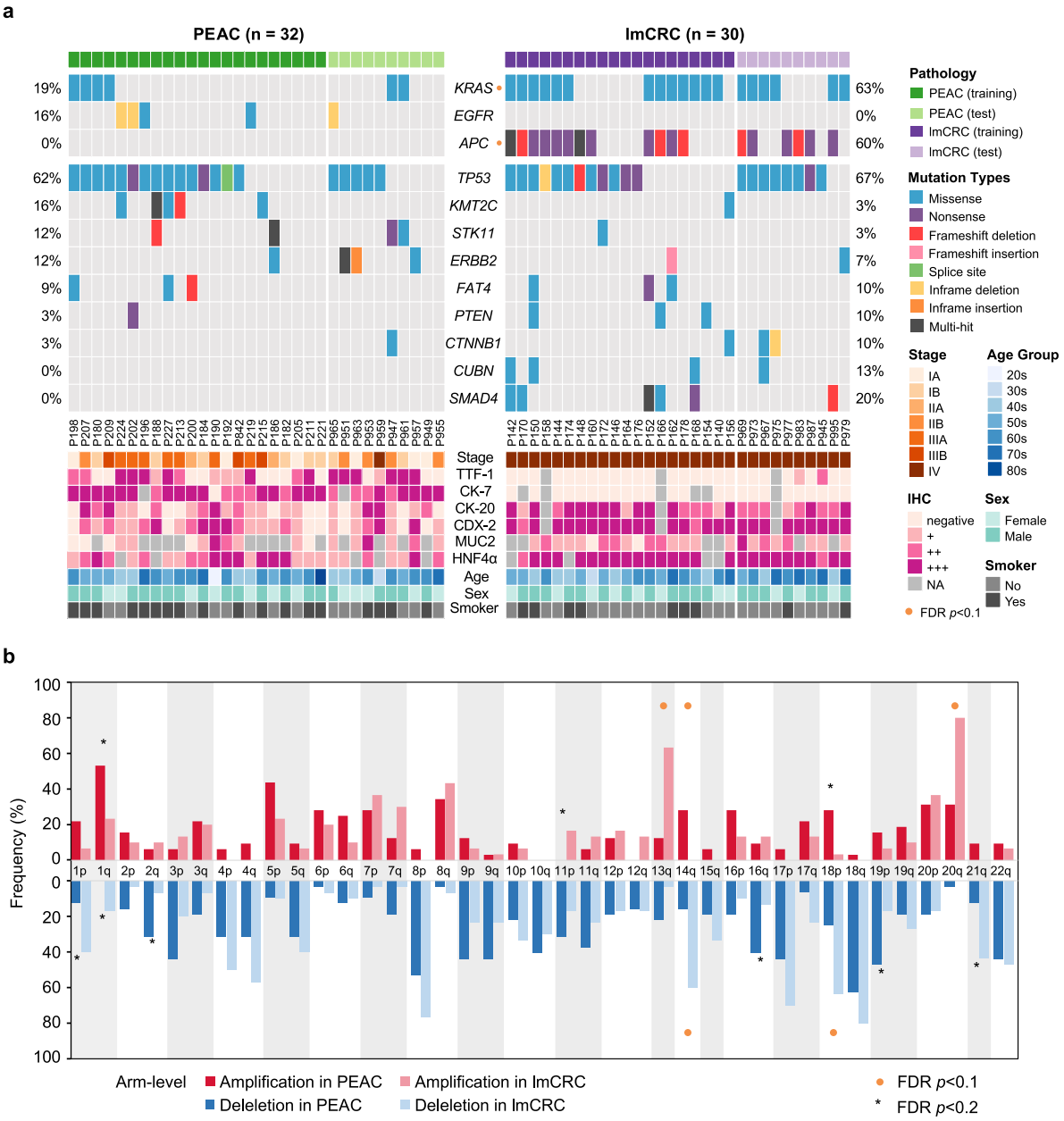


Figure 2. Genomic and clinical features of PEAC and lmCRC patients. (a) Each patient is presented in a single column organized by their clinical diagnosis and the corresponding cohort in our study. The occurrence of the selected genes is presented with the percentage of the variation frequencies in either cancer type. For mutations, cancer-related genes with more than 10% mutation frequency in either the PEAC or lmCRC cohort are shown. Immunostaining of key clinical diagnostic biomarkers, including TTF-1, CK7, CK20, CDX-2, MUC2 and HNF4α is presented based on the staining intensity (negative to +++). Other clinical categories include tumor stage, age, sex, and smoking history. (b) Frequency of chromosome arm-level copy number alterations. Darker colors represent events in PEAC (red, amplification; blue, deletion) and lighter colors are used to represent lmCRC. Arm-level alterations with adjusted p values <0.1 by Fisher's exact test are marked with orange dots. PEAC, pulmonary enteric adenocarcinoma; lmCRC, lung metastatic colorectal cancer.

genes included *TP53* (62%), *KRAS* (19%), *EGFR* (16%), *ERBB2* (12%) and *STK11* (12%), whereas lmCRC showed a high prevalence of *TP53* (67%), *KRAS* (63%), *APC* (60%) and *SMAD4* (20%) (Figure 2). Of these, we found significant enrichment of *APC* ($p<0.001$ by Fisher's exact test), *KRAS* ($p<0.001$ by Fisher's exact test) and *SMAD4* ($p=0.01$ by Fisher's exact test) mutations in lmCRC (Figure S2a). As expected,

EGFR-sensitive mutations (exon 19 deletion and exon 21 L858R) were exclusively found in PEAC (16%), but at a frequency significantly lower than that in tLUAD ($p<0.001$ by Fisher's exact test) (Figure S2b). Notably, one of the *EGFR*-mutant patients carried concurrent L858R/T790M at the time of diagnosis. We also found a slightly higher frequency of *ERBB2* mutations in PEAC (12%) than in tLUAD (4%) ($p=0.11$ by Fisher's exact test). These included two patients with *ERBB2* exon 20 insertions, and two with *ERBB2* missense mutations at V659E and D769Y, and were mutually exclusive from *EGFR* or *KRAS* mutations (Figure 2). The prevalence of *KRAS* was also slightly higher than that in tLUAD, but the difference was not statistically significant. The mutants of *KRAS* in PEAC included two G12C, two G12D, one G12S, and one concurrent Q61H/R68W. Importantly, *APC* was also observed in approximately 6% of tLUADs, but none in the PEAC cohort. Overall, we observed low TMB in this study, with no difference between PEAC and lmCRC (Figure S1b). Interestingly, we found an enrichment of age-related (clocklike) signatures (SBS1 and SBS5) in lmCRC, whereas APOBEC (SBS2 and SBS13) contributed to a larger proportion of mutations in PEAC (Figure S1c).

At the chromosome arm level, significantly amplified 13q and 20q, and significantly deleted 14q and 18p were detected in lmCRC, whereas significantly amplified 14q was observed in PEACs (all with FDR<0.1) (Figure 2b). For gene-level CNAs, we found significant enrichment of *BCL2L1* gain ($p<0.001$ by Fisher's exact test) and *E2F1* gain ($p=0.017$ by Fisher's exact test) exclusively in lmCRCs (both with FDR<0.1) (Figure S2c). Moreover, many focal CNA changes were detected by the GISTIC 2.0 algorithm. Significant focal amplification of 11q13.5 and focal deletion peaks of 6p22.2, 11p15.4 and 19q13.43 were enriched in PEAC, whereas focal amplification of 12p13.32 and focal deletion of 6p22.2 was observed in lmCRC (Figure S3). Despite these CNA differences, the overall fraction of genome altered was similar between PEAC and lmCRC, as indicated by the levels of chromosome instability in Figure S2d.

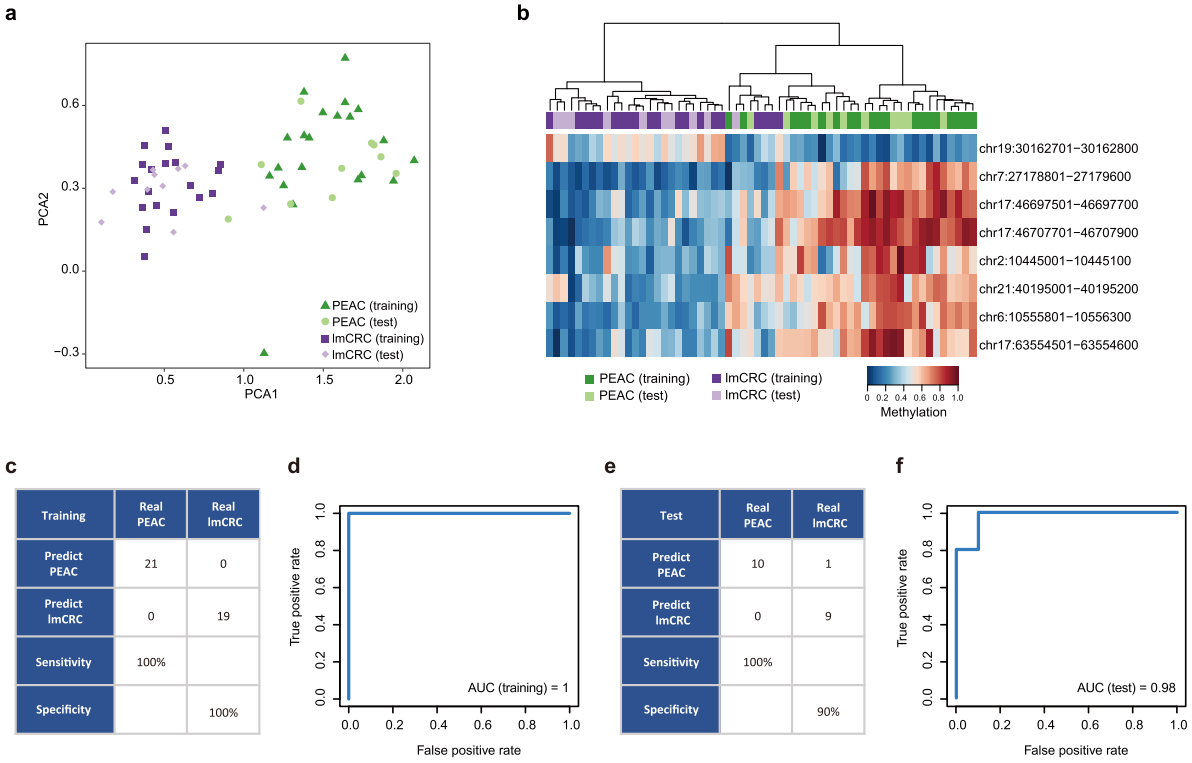
Identification of differentially methylated regions

Next, we performed DNA methylation sequencing to explore disease-specific DNA methylation changes in PEAC and lmCRC. Overall, one PEAC sample and one lmCRC sample in the training cohort failed the quality control for targeted bisulfite sequencing possibly due to insufficient sample input and were removed from subsequent analyses. DMRs were extracted by comparing tumor and matched normal tissues of these two diseases. In the training cohort (21 PEAC and 19 lmCRC), with a focus on PEAC, we identified a total of 14235 DMRs by comparing the methylation levels of PEAC tumors with those in corresponding normal lung tissues, and 19791 DMRs by comparing PEAC tumors with lmCRC tumors. In total, 1094 overlapping DMRs were considered PEAC-specific and subsequently joined as a single DMR if they were adjacent to one another, and those with an overall sequencing coverage <50X were filtered. This step helped increase the feature diversity and avoid repeated hits from adjacent regions. The resulting 204 DMRs were used as candidate markers for classifier construction (Table S2).

Relative hypermethylation in PEAC was detected in 141/204 DMRs and hypomethylation was detected in the remaining 63 DMRs. Some top-ranked regions were hypermethylated within the promoter of genes, including glucosaminyl (N-Acetyl) transferase 2 (I Blood Group) (*GCNT2*), homeobox A4 (*HOXA4*), homeobox A9 (*HOXA9*), activin A receptor-like type 1 (*ACVRL1*), microRNA 196a-1 (*MIR196A1*) and Rho guanine nucleotide exchange factor 15 (*ARHGEF15*) (Table S2). For cross-platform comparison, we evaluated whether the candidate DMRs contained any CpG sites covered by the 450K methylation chip array in the TCGA cohort. We found that 81/204 DMRs contained zero CpG sites and the remaining DMRs were represented by limited number of CpGs (ranging from 1 to11 cg sites). The estimated mean methylation level of each DMR in TCGA LUAD is also shown (Table S2).

Development and validation of the diagnostic classifiers

Although the genetic and epigenetic biomarkers described above could well characterize PEAC, an integrated diagnostic classifier is still essential for evaluating the molecular heterogeneity to assist decision making. Therefore, we constructed independent classifiers with gene mutation, CNA or DMR markers alone, or with combined molecular events. The candidate features included all 711 genes mutated in either cancer type, 77 arm-level CNA events, and 204 characteristic DMRs. Two feature selection methods were performed to reduce dimensionality and remove redundant genetic or epigenetic features, resulting in 8 DMRs, 10 CNAs, and the *APC* mutation in each category (Figure S4, Table S3). First, we established separate classifiers with only DMR, CNV, or mutational features. With the DMR features, the principal component analysis (PCA) of selected DMRs demonstrated two distinct groups that represented the PEAC and lmCRC samples, which were better separated than with all 204 DMRs (Figure 3a and Figure S5a). Similarly, unsupervised hierarchical clustering based on eight DMR markers also presented a clearer separation of the PEAC and lmCRC tumors than the clustering of all 204 DMRs (Figure 3b and Figure S5b). Of these eight DMRs, only one showed higher methylation level in lmCRC, whereas the methylation levels of the others were higher in PEAC (Figure S6).



Training	Real PEAC	Real lmCRC
Predict PEAC	21	0
Predict lmCRC	0	19
Sensitivity	100%	
Specificity		100%

Test	Real PEAC	Real lmCRC
Predict PEAC	10	1
Predict lmCRC	0	9
Sensitivity	100%	
Specificity		90%

Figure 3. Classification of PEAC and lmCRC based on methylation analysis. (a) Principal component analysis (PCA) based on eight selected DMRs showing the coordinates of two principal components of individual patients. The separation of lmCRC and PEAC can be observed as two groups. (b) Unsupervised hierarchical clustering based on eight DMRs in both the training and test cohorts. The colored bar indicates the methylation level. (c, e) Confusion matrices summarizing the sensitivity and specificity of the DMR diagnostic classifier in the training and test cohorts. (d, f) Receiver operating characteristic curves of the diagnostic classifier in the training and test cohorts. PEAC, pulmonary enteric adenocarcinoma; lmCRC, lung metastatic colorectal cancer; AUC, area under the curve.

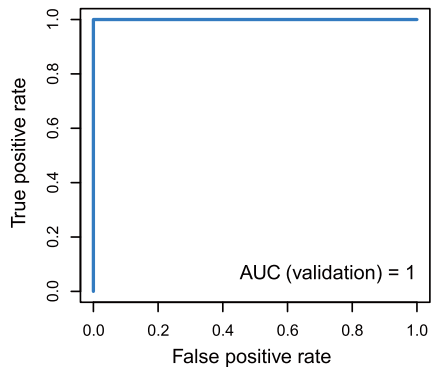
Then, we constructed a predictive model by the random forest algorithm using the eight DMR markers to distinguish between PEAC and lmCRC. The classifier successfully classified all lmCRC and PEAC cases in the training cohort, yielding a total accuracy of 100% (40/40) (AUC=1) (Figure 3c, d, and Figure S7a). Applying this model to the internal test cohort, 90% (9/10) of the lmCRC cases and 100% (10/10) of the PEAC cases were correctly predicted with an accuracy of 95% (19/20) (AUC=0.98) (Figure 3e, f). Specifically, one lmCRC patient, P979, was predicted to have PEAC (Figure S7b). Furthermore, we ran these samples through the SVM and XGBoost machine learning algorithms to assess the robustness of the selected markers. Both SVM and XGBoost produced an accuracy of 100% in the training cohort. In the internal test cohort, SVM generated the same predictions as RF, whereas XGBoost exhibited a slightly lower sensitivity (90%) and overall accuracy (90%). Notably, P979, who was incorrectly classified by the RF model, was repeatedly misclassified using SVM and XGBoost (Figure S8).

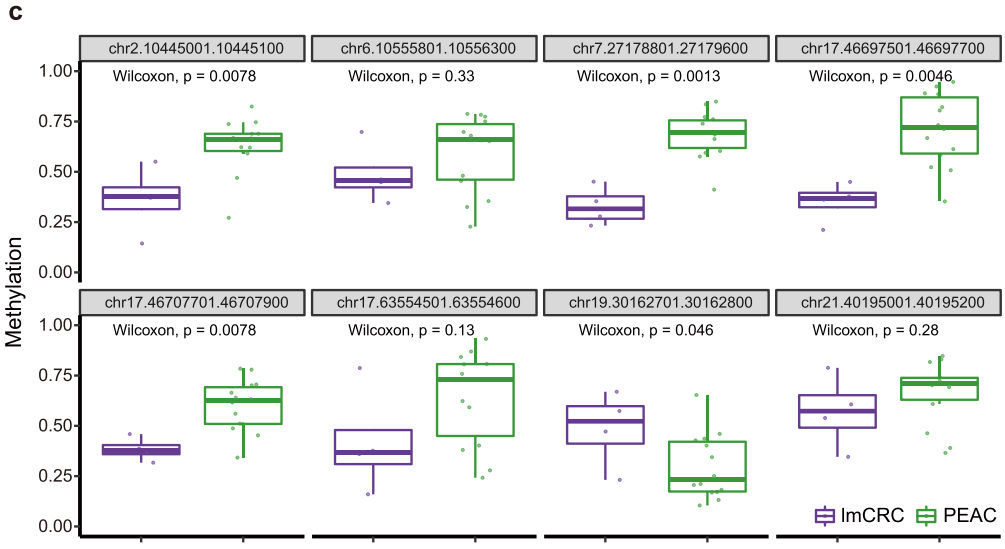
Despite the distinct CNA patterns observed between PEAC and lmCRC, the classifier based solely on CNA markers did not separate the two diseases well (training cohort, sensitivity, 81.0%; specificity, 78.9%; overall accuracy, 80%; internal test cohort, sensitivity, 60%; specificity, 90%; overall accuracy, 75%) (Figure S9b and Table S4). Additionally, all *APC*-wildtype lmCRC were falsely categorized, indicating that a single genetic mutation was not informative enough for disease classification (Figure S9a and Table S4). Furthermore, combination of different molecular events, including the *APC* mutation, 10 CNAs and 8 DMRs generated the same prediction results as the pure DMR-based classifier (training cohort, sensitivity, 100%; specificity, 100%; overall accuracy, 100%; internal test cohort, sensitivity, 100%; specificity, 90%; overall accuracy, 95%) (Figure S9c). Therefore, only DMR features were considered for the final classifier construction and evaluation.

Finally, to validate the performance of our classifier, we first performed external validation with samples collected from multiple centers in China, including 17 PEAC and 7 lmCRC patients. Remarkably, all cases agreed with their clinical diagnosis (Figure 4a and b). Given the lack of public PEAC cases and DNA

a

Validation	Real PEAC	Real lmCRC
Predict PEAC	17	0
Predict lmCRC	0	7
Sensitivity	100%	
Specificity		100%





d

GEO cohort	Real PEAC	Real lmCRC
Predict PEAC	14	1
Predict lmCRC	0	3
Sensitivity	100%	
Specificity		75%

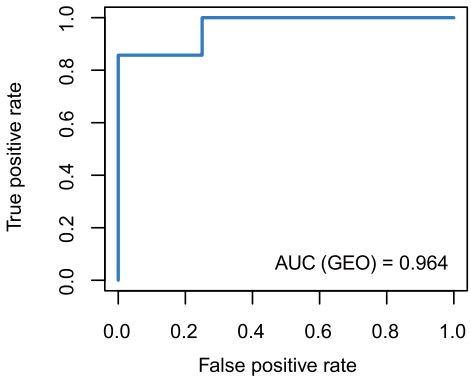


Figure 4. External validation of the classifier. (a, b) Confusion matrix and receiver operating characteristic (ROC) curve summarizing the results of the diagnostic prediction of an independent multicenter external validation cohort of 17 PEAC and 7 lmCRC samples. (c) Box plots demonstrating the estimated methylation levels in each featured region in a GEO cohort of 14 PEAC and 4 lmCRC samples. P values were extracted from the Wilcoxon rank sum test. (d, e) confusion matrix and ROC curve summarizing the results of the diagnostic prediction of 14 PEAC and 4 lmCRC samples from the GEO cohort. PEAC, pulmonary enteric adenocarcinoma; lmCRC, lung metastatic colorectal cancer; AUC, area under the curve.

methylation data using NGS platforms, we further evaluated the potential generalizability of the classifier by estimating the approximate methylation levels of the eight DMRs using the mean methylation level of CpG sites in methylation array-based GEO and TCGA data. Specifically, a total of 15 CpG sites from the 450K array overlapped with the eight DMRs, whereas no CpG from the selected sites used to classify the GEO samples in the previous study overlapped with the eight DMRs.[15] In the GEO cohort, similar differences in the methylation levels of the eight DMRs were observed, although three did not demonstrate statistical significance probably due to the small sample size (Figure 4c). Using the estimated methylation levels for prediction, only sample GSM3258624 (lmCRC) was misclassified as PEAC, yielding an accuracy of 94.4% (sensitivity=100%, specificity=75%, AUC=0.964) (Figure 4d and 4e). Similar results were observed in TCGA samples, with markedly different methylation levels of the eight DMRs between LUAD and CRC (Figure S10a). Classifier resulted in the misclassification of 80/453 CRCs and 30/503 LUAD, giving sensitivity 94.0%, specificity of 82.3% and AUC of 0.947 (Figure S10b and S10c). We further assessed the documented clinicopathological status of these TCGA tumors, including age at diagnosis, sex, smoking history, and tumor staging. Interestingly, a significantly higher proportion of stage I-III CRCs (19.6%) were misclassified than stage IV tumors (6.3%) (p=0.011 by Fisher's exact test), probably because the classifier was specifically trained for advanced CRCs. Other patient characteristics were balanced between the correctly and incorrectly classified samples.

Discussion

PEAC is a rare subtype of lung adenocarcinoma that is easily misdiagnosed as lmCRC in the clinic due to their shared pathological presentation. The mainstay criteria for differential diagnosis rely on enteric and pulmonary IHC markers and clinical history. However, the diagnosis of tumors with clinical features resembling each other is still challenging, especially when clinical history might not always be informative. The 2021 WHO classification guideline for lung cancers has greatly improved the IHC criteria for PEAC diagnosis. However, current pathology assessment still relies heavily on pathologists' experience, which might result in difficulty in disease classification. In this study, by analyzing whole exome sequencing and targeted bisulfite sequencing data, we identified PEAC-specific genetic and epigenetic markers to distinguish PEAC from lmCRC. We also successfully established a binary diagnostic classifier with selected DMRs and showed its potential generalizability through internal and external validations.

As a special subtype of lung adenocarcinoma, PEAC demonstrates a characteristic genetic profile and occurrence of lung cancer drivers, which might inform therapeutic options. According to previous literatures, *KRAS* is among the most frequently altered genes in PEAC, with a greatly varied frequency ranging from 7.7% to 60.0%, probably due to limited sample sizes.[12,14,15] In the current study, despite being one of the top PEAC driver mutations, *KRAS* mutation was significantly less mutated in PEACthan in lmCRC. Nevertheless, future clinical efficacy in *KRAS* G12C-mutant PEAC might be achieved by sotorasib, as shown by the latest data of CodeBreaK100 study.[34] Moreover, *EGFR* sensitive mutations were exclusively found in five PEAC patients, which is similar to a previous Chinese report on PEAC.[35] Less frequent *EGFR* mutations have been described in other European studies possibly due to different mutation profiles across ethnic groups.[12,14,15,36] Importantly, as most PEACs are diagnosed at early stages, the *EGFR* mutation status could guide the choice of adjuvant or neoadjuvant therapies.[37,38] Although no *ALK, MET* or *ROS1* alterations were identified in our cohorts and others,[15] we found a higher prevalence of *ERBB2* mutations in PEAC, suggesting potential benefit from ado-trastuzumab emtansine or trastuzumab deruxtecan at advanced stages.[39,40] Additionally, as a known driver of CRC, the *APC* mutation was found to be a major biomarker to distinguish lmCRC from PEAC, at a frequency similar to those in previous reports.[41,42] However, *APC* is found in approximately 6% of all lung cancers,[21] suggesting that more data might be necessary to prove its association with PEAC specifically.

To date, limited research has been conducted to investigate CNAs in PEAC. Overall, the CNA landscape of PEAC was similar to that of tLUAD.[43] However, we found distinct CNA profiles between PEAC and lmCRC. In concordance with previous studies, our results demonstrated the frequently observed arm-level amplifications of 13q and 20q, and deletions of 14q and 18p in CRCs, which are less frequent in lung adenocarcinomas.[41,43] Focal amplification of 12p13.32 spanning across *CCND2* was identified in both the TCGA CRC data and ours.[41] However, in both cancer types, broader spectra of focal alterations were reported by TCGA, probably due to larger sample sizes, which affected the statistical significance in the GISTIC analysis.

Our analysis of DNA methylation revealed that PEAC-specific epigenetic features enable robust diagnostic classifier development. For high-dimensional bioinformatics or genomic data, such as DNA methylation, machine learning methods can quickly reduce dimensionality and extract key molecular features for biomarker identification and classifier construction. Recently, many studies[15,17] have taken advantage of public databases (e.g., TCGA and GEO) of known primary lesions to identify epigenetic biomarkers and train diagnostic models to analyze the origin of tumors on the consensus of DNA methylation's tissue-specificity.[18,44]

However, models established based on features from primary tumors might affect the correct classification of metastatic lesions. Stromal cells in the tumor microenvironment could bring mixed methylation patterns and confuse the classifier. For example, an lmCRC biopsy with low tumor fraction might over-resemble the lung tissue and be misclassified as PEAC because the normal para-tumor tissue of lmCRC harbors methylation features from the lung rather than the colon.[45] Importantly, the final eight DMRs did not overlap with the top 10000 CpGs selected using TCGA data, indicating that the most important features used to distinguish PEAC from lmCRC might differ from those that distinguished primary tLUAD and CRC.[15] Hence, the external validation with these public datasets resulted in slightly lower overall accuracies. Importantly, to reduce interference from adjacent tissue in this study, normal lung tissues were also sequenced and used to screen PEAC-specific DMR features. Another limitation of using public microarray-based DNA methylation data is the experimental and inter-individual variability induced by discrete CpG sites and a relatively narrow scope of epigenome analysis.[46] Here, we employed targeted bisulfite sequencing technique to overcome both the low genome coverage of methylation arrays and the high cost and inefficiency of whole-genome bisulfite sequencing while preserving the single-base resolution, general reproducibility, and comparable results of these two platforms.[47,48]

Notably, approximately 40% of the DMRs we identified covered CpG sites novel to the existing databases. Of the top PEAC-specific DNA methylation features, the highest ranked DMR located on chromosome 6 corresponded to the promoter hypermethylation of *GCNT2*, indicating possible gene suppression. Remarkably, a previous study reported that *GCNT2* hypomethylation, which promoted the activation of this gene, was closely related to the lymph node metastasis of primary CRC.[49–51] We also found aberrant methylation in a few members of the HOX cluster of genes, including *HOXA9* and *HOXA4*. HOX genes were highly conserved during evolution and are involved in numerous aspects of biological processes. Recent studies have shown their roles in cancer predisposition and oncogenesis.[52,53] *MIR196A* is a microRNA that closely interacts with several members within the HOX family, where methylation-induced silencing of these miRNAs is frequently observed in colon cancers.[53,54]

A major concern of our study is whether the sample size was large enough for solid binary classifier building. Given the low incidence of PEAC and limited public datasets, we carefully selected the suitable algorithm and further validated our model through a multicenter external cohort and public datasets. We employed an RF-based classifier as it is known as the most compatible method for training high dimensional data and overcoming overfitting problems from small sample sizes.[55,56] Several studies have demonstrated the feasibility to distinguish confusing tumor types or early tumor screening using RF-based models.[57,58] Through rigorous biomarker screening, we developed a neat classifier of eight DNA methylation-based markers to distinguish PEAC from lmCRC, and were able to achieve near-perfect accuracies in both training, test and external validation cohorts. Although different testing platforms were used to assess DNA methylation, the methylation classifier composed of the eight DMRs successfully distinguished PEAC from lmCRC in the fourteen publicly available PEAC samples,[15] further supporting the robustness of our classifier. We believed that a near perfect performance of this classifier is essential for a rare disease, such as PEAC. For future clinical application, we could design probes targeting the selected DMR regions (total 2200bp) instead of broad genome coverage, improving the cost-effectiveness of this model to facilitate clinical diagnosis of PEAC.

To the best of our knowledge, this study is the largest multi-omics study to directly compare PEAC with its "unrelated twin", lmCRC. We demonstrated PEAC-specific genomic and epigenetic features with a proof-of-concept DMR-based classifier to facilitate the diagnosis of the two diseases. With additional clinical validation in the future, this diagnostic classifier could become a powerful tool for auxiliary diagnosis of PEAC.

Contributors

Conception and design: Jie Wang, Qibin Song, Ying Zuo, Hua Bai, Jia Zhong, Bin Xu. Development of methodology: Ying Zuo, Zhijie Wang, Jia Zhong, Yedan Chen, Hua Bao. Acquisition of data: Jianchun Duan, Shi Jin, Shuhang Wang, Weihua Li, Bin Xu, Rui Wan, Jiachen Xu, Zhenlin Yang. Analysis and interpretation of data: Ying Zuo, Hua Bai, Yedan Chen, Shi Jin, Xin Wang, Kailun Fei, Jiefei Han. Writing, review and/or revision of the manuscript: Jie Wang, Qibin Song, Ying Zuo, Hua Bai, Zhijie Wang, Jianchun Duan, Weihua Li, Yedan Chen. Administrative, technical, or material support: Hua Bai, Zhijie Wang, Jianming Ying, Yang Shao. Study supervision: Jie Wang, Qibin Song. All authors read and approved the final version of the manuscript. Ying Zuo, Jia Zhong, Weihua Li, and Yedan Chen have accessed and verified the data. Jie Wang and Qibin Song were responsible for the decision to submit the manuscript.

Data sharing statement

Public WES data of Asian lung adenocarcinoma can be acquired from the cbioportal website (https://www.cbioportal.org/study?id=luad_oncosg_2020). Array-based DNA methylation data of primary CRC and primary LUAD can be acquired from the TCGA Research

Network (https://portal.gdc.cancer.gov/). Array-based DNA methylation data of PEAC and lmCRC can be acquired from the GEO repository (GSE116699). Sequencing data and code for this study can be obtained by contacting the corresponding author at zlhuxi@163.com.

Declaration of interests

YC, HB, and YS are employees of Nanjing Geneseeq Technology Inc. Other authors declare no potential conflicts of interest.

Acknowledgements

We thank Bo Zheng, Long Wang, Lintong Zheng, Runze Xu of Department of Pathology, National Cancer Center for preparing the samples. We thank Shuang Chang from Nanjing Geneseeq Technology Inc. for the valuable help in processing bisulfite sequencing data and reviewing the diagnostic classifier. And we thank Zhengyi Zhao for proofreading the revised manuscript. This work was supported by National key research and development project 2019YFC1315700, CAMS Key Laboratory of Translational Research on Lung Cancer (2018PT31035), and Beijing Natural Science Foundation (7222144).

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2022.104165.

References

1 Inamura K, Satoh Y, Okumura S, et al. Pulmonary adenocarcinomas with enteric differentiation. *Am J Surg Pathol*. 2005;29(5):660–665.
2 Tsao M-S, Fraser RS. Primary pulmonary adenocarcinoma with enteric differentiation. *Cancer*. 1991;68:1754–1757.
3 Travis WD, Brambilla E, Nicholson AG, et al. The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. *J Thorac Oncol*. 2015;10(9):1243–1260.
4 Travis WD, Brambilla E, Noguchi M, et al. International association for the study of lung cancer/American thoracic society/European respiratory society international multidisciplinary classification of lung adenocarcinoma. *J Thorac Oncol*. 2011;6(2):244–285.
5 Nicholson AG, Tsao MS, Beasley MB, et al. The 2021 WHO classification of lung tumors: impact of advances since 2015. *J Thorac Oncol*. 2022;17(3):362–387.
6 Remo A, Zanella C, Pancione M, et al. Lung metastasis from TTF-1 positive sigmoid adenocarcinoma. pitfalls and management. *Pathologica*. 2013;105(2):69–72.
7 Li H, Cao W. Pulmonary enteric adenocarcinoma: a literature review. *J Thorac Dis*. 2020;12(6):3217–3226.
8 Miyaoka M, Hatanaka K, Iwazaki M, Nakamura N. CK7/CK20 double-negative pulmonary enteric adenocarcinoma with histopathological evaluation of transformation zone between enteric adenocarcinoma and conventional pulmonary adenocarcinoma. *Int J Surg Pathol*. 2018;26(5):464–468.
9 Hatanaka K, Tsuta K, Watanabe K, Sugino K, Uekusa T. Primary pulmonary adenocarcinoma with enteric differentiation resembling metastatic colorectal carcinoma: a report of the second case negative for cytokeratin 7. *Pathol Res Pract*. 2011;207(3):188–191.
10 Bayrak R, Yenidunya S, Haltas H. Cytokeratin 7 and cytokeratin 20 expression in colorectal adenocarcinomas. *Pathol Res Pract*. 2011;207(3):156–160.
11 Li HC, Schmidt L, Greenson JK, Chang AC, Myers JL. Primary pulmonary adenocarcinoma with intestinal differentiation mimicking metastatic colorectal carcinoma: case report and review of literature. *Am J Clin Pathol*. 2009;131(1):129–133.
12 Nottegar A, Tabbò F, Luchini C, et al. Pulmonary adenocarcinoma with enteric differentiation: immunohistochemistry and molecular morphology. *Appl Immunohistochem Mol Morphol*. 2018;26(6):383–387.
13 Matsushima J, Yazawa T, Suzuki M, et al. Clinicopathological, immunohistochemical, and mutational analyses of pulmonary enteric adenocarcinoma: usefulness of SATB2 and beta-catenin immunostaining for differentiation from metastatic colorectal carcinoma. *Hum Pathol*. 2017;64:179–185.
14 Nottegar A, Tabbo F, Luchini C, et al. Pulmonary adenocarcinoma with enteric differentiation: dissecting oncogenic genes alterations with DNA sequencing and FISH analysis. *Exp Mol Pathol*. 2017;102(2):276–279.
15 Jurmeister P, Scholer A, Arnold A, et al. DNA methylation profiling reliably distinguishes pulmonary enteric adenocarcinoma from metastatic colorectal cancer. *Mod Pathol*. 2019;32(6):855–865.
16 Saghafinia S, Mina M, Riggi N, Hanahan D, Ciriello G. Pan-cancer landscape of aberrant DNA methylation across human tumors. *Cell Rep*. 2018;25(4):1066–1080.e1068.
17 Capper D, Jones DTW, Sill M, et al. DNA methylation-based classification of central nervous system tumours. *Nature*. 2018;555(7697):469–474.
18 Lokk K, Modhukur V, Rajashekar B, et al. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol*. 2014;15(4):r54.
19 Xia D, Leon AJ, Cabanero M, et al. Minimalist approaches to cancer tissue-of-origin classification by DNA methylation. *Mod Pathol*. 2020;33(10):1874–1888.
20 Kang S, Li Q, Chen Q, et al. CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biol*. 2017;18(1):53.
21 Chen J, Yang H, Teo ASM, et al. Genomic landscape of lung adenocarcinoma in East Asians. *Nat Genet*. 2020;52(2):177–186.
22 Wang S, Li H, Song M, et al. Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes. *PLoS Genet*. 2021;17(5):e1009557.
23 Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. *Nature*. 2020;578(7793):94–101.
24 Riaz N, Havel JJ, Makarov V, et al. Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell*. 2017;171(4):934–949.e916.
25 Wendt J, Rosenbaum H, Richmond TA, Jeddeloh JA, Burgess DL. Targeted bisulfite sequencing using the SeqCap epi enrichment system. *Methods Mol Biol*. 2018;1708:383–405.
26 Akalin A, Kormaksson M, Li S, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*. 2012;13(10):R87.
27 Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: misc functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-0.1. Available at: https://CRAN.R-project.org/package=caret. 2019.
28 Chen T, He T, Benesty M, et al. xgboost: Extreme Gradient Boosting. R package version 1.4.1.1. Available at: https://CRAN.R-project.org/package=xgboost. 2021.
29 Liaw A, Wiener M. Classification and regression by randomForest. *R News*. 2002;2(3):18–22.
30 Kuhn M. Building Predictive Models in R Using the caret Package. *J Stat Software*. 2008;28(5):1–26.
31 Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res*. 2018;28(11):1747–1756.
32 Warnes GR, Bolker B, Bonebakker L, et al. gplots: various R Programming Tools for Plotting Data. R package version 3.0.1.1. Available at: https://CRAN.R-project.org/package=gplots. 2019.
33 Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York; 2016.
34 Dy G, Govindan R, Velcheti V, et al. Long-term outcomes with sotorasib in pretreated KRAS p.G12C-mutated NSCLC: 2-year analysis of CodeBreaK 100. *AACR Annual Meeting. 2022; Abstract CT008: Presented April 10*. 2022.

35 Zhao L, Huang S, Liu J, Zhao J, Li Q, Wang HQ. Clinicopathological, radiographic, and oncogenic features of primary pulmonary enteric adenocarcinoma in comparison with invasive adenocarcinoma in resection specimens. *Medicine*. 2017;96(39):e8153.
36 Gou LY, Wu YL. Prevalence of driver mutations in non-small-cell lung cancers in the People's Republic of China. *Lung Cancer*. 2014;5:1–9.
37 Zhong WZ, Wang Q, Mao WM, et al. Gefitinib versus vinorelbine plus cisplatin as adjuvant treatment for stage II-IIIA (N1-N2) EGFR-mutant NSCLC (ADJUVANT/CTONG1104): a randomised, open-label, phase 3 study. *Lancet Oncol*. 2018;19(1):139–148.
38 Wu YL, Tsuboi M, He J, et al. Osimertinib in resected EGFR-mutated non-small-cell lung cancer. *N Engl J Med*. 2020;383(18):1711–1723.
39 Li BT, Shen R, Buonocore D, et al. Ado-trastuzumab emtansine for patients with HER2-mutant lung cancers: results from a phase II basket trial. *J Clin Oncol*. 2018;36(24):2532–2537.
40 Li BT, Smit EF, Goto Y, et al. Trastuzumab deruxtecan in HER2-mutant non-small-cell lung cancer. *N Engl J Med*. 2022;386(3):241–251.
41 Network TCGA. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330–337.
42 Seshagiri S, Stawiski EW, Durinck S, et al. Recurrent R-spondin fusions in colon cancer. *Nature*. 2012;488(7413):660–664.
43 Network TCGAR. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511(7511):543–550.
44 Bormann F, Rodríguez-Paredes M, Lasitschka F, et al. Cell-of-origin DNA methylation signatures are maintained during colorectal carcinogenesis. *Cell Rep*. 2018;23(11):3407–3418.
45 Hoadley KA, Yau C, Hinoue T, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*. 2018;173(2):291–304.e296.
46 Fan S, Tang J, Li N, et al. Integrative analysis with expanded DNA methylation data reveals common key regulators and pathways in cancers. *NPJ Genom Med*. 2019;4:2.
47 Teh AL, Pan H, Lin X, et al. Comparison of methyl-capture sequencing vs. infinium 450K methylation array for methylome analysis in clinical samples. *Epigenetics*. 2016;11(1):36–48.
48 Ziller MJ, Gu H, Muller F, et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature*. 2013;500(7463):477–481.
49 Nakamura K, Yamashita K, Sawaki H, et al. Aberrant methylation of GCNT2 is tightly related to primary CRC. *Anticancer Res*. 2015;35:1411–1422.
50 Dimitroff CJ. I-branched carbohydrates as emerging effectors of malignant progression. *Proc Natl Acad Sci USA*. 2019;116(28):13729–13737.
51 Chao CC, Wu PH, Huang HC, et al. Downregulation of miR-199a/b-5p is associated with GCNT2 induction upon epithelial-mesenchymal transition in colon cancer. *FEBS Lett*. 2017;591(13):1902–1917.
52 Bhatlekar S, Fields JZ, Boman BM. HOX genes and their role in the development of human cancers. *J Mol Med*. 2014;92(8):811–823.
53 Shah N, Sukumar S. The Hox genes and their roles in oncogenesis. *Nat Rev Cancer*. 2010;10(5):361–371.
54 Milevskiy MJG, Gujral U, Del Lama Marques C, et al. MicroRNA-196a is regulated by ER and is a prognostic biomarker in ER+ breast cancer. *Br J Cancer*. 2019;120(6):621–632.
55 Qi Y. Random forest for bioinformatics. In: Zhang C, Ma Y, eds. *Ensemble Machine Learning: Methods and Applications*. Springer Science+Business Media, LLC; 2012:307–323.
56 Chen X, Ishwaran H. Random forests for genomic data analysis. *Genomics*. 2012;99(6):323–329.
57 Hovestadt V, Remke M, Kool M, et al. Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour material using high-density DNA methylation arrays. *Acta Neuropathol*. 2013;125(6):913–916.
58 Klassen M, Cumming M, Saldaña-González G. Investigation of random forest performance with cancer microarray data. *Computers and their Applications*. 2008;64–69.