OPEN ACCESS

Original research

DNA-methylation signature accurately differentiates pancreatic cancer from chronic pancreatitis in tissue and plasma

Yenan Wu,[1,2] Isabelle Seufert ,[1,2] Fawaz N Al-Shaheri,[1,3] Roman Kurilov,[4] Andrea S Bauer,[1] Mehdi Manoochehri,[1] Evgeny A Moskalev,[5] Benedikt Brors,[4] Christin Tjaden,[6] Nathalia A Giese,[6] Thilo Hackert,[6] Markus W Büchler,[6] Jörg D Hoheisel [1]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2023-330155).

For numbered affiliations see end of article.

Correspondence to
Dr Jörg D Hoheisel, Functional Genome Analysis, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany; j.hoheisel@dkfz.de

TH, MWB and JDH contributed equally.

YW, IS and FNA-S are joint first authors.

Received 24 April 2023
Accepted 31 August 2023
Published Online First 14 September 2023

To cite: Wu Y, Seufert I, Al-Shaheri FN, *et al*. *Gut* 2023;**72**:2344–2353.

ABSTRACT

Objective Pancreatic ductal adenocarcinoma (PDAC) is a lethal malignancy. Differentiation from chronic pancreatitis (CP) is currently inaccurate in about one-third of cases. Misdiagnoses in both directions, however, have severe consequences for patients. We set out to identify molecular markers for a clear distinction between PDAC and CP.

Design Genome-wide variations of DNA-methylation, messenger RNA and microRNA level as well as combinations thereof were analysed in 345 tissue samples for marker identification. To improve diagnostic performance, we established a random-forest machine-learning approach. Results were validated on another 48 samples and further corroborated in 16 liquid biopsy samples.

Results Machine-learning succeeded in defining markers to differentiate between patients with PDAC and CP, while low-dimensional embedding and cluster analysis failed to do so. DNA-methylation yielded the best diagnostic accuracy by far, dwarfing the importance of transcript levels. Identified changes were confirmed with data taken from public repositories and validated in independent sample sets. A signature of six DNA-methylation sites in a CpG-island of the protein kinase C beta type gene achieved a validated diagnostic accuracy of 100% in tissue and in circulating free DNA isolated from patient plasma.

Conclusion The success of machine-learning to identify an effective marker signature documents the power of this approach. The high diagnostic accuracy of discriminating PDAC from CP could have tremendous consequences for treatment success, once the result from still a limited number of liquid biopsy samples would be confirmed in a larger cohort of patients with suspected pancreatic cancer.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Distinction between patients with pancreatic ductal adenocarcinoma (PDAC) and chronic pancreatitis (CP) is currently incorrect for about one-third of patients.
⇒ Every misdiagnosis is likely to have severe consequences for a patient.

WHAT THIS STUDY ADDS

⇒ Using machine-learning, we were able to extract from genome-wide information on DNA-methylation and messenger RNA/microRNA expression in tissue a validated signature of six DNA-methylation features that allowed fully accurate diagnosis.
⇒ The methylation variations exhibit a diagnostic robustness that is likely to be critical for real-life application.
⇒ Discrimination worked with identical accuracy in plasma samples, suggesting potential for non-invasive diagnosis by liquid biopsy.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Once replicated on a larger number of patients with PDAC and CP and confirmed in a clinical trial, translation could substantially affect patient management and prognosis.

INTRODUCTION

The 5-year survival rate of patients with pancreatic ductal adenocarcinoma (PDAC) is approximately 9%, dropping to 3% in metastatic PDAC.[1] Reasons are late diagnosis, misdiagnosis and inherent therapy resistance. An important risk factor for PDAC is chronic pancreatitis (CP),[2] a persisting fibro-inflammatory disorder of the exocrine pancreas.[3]

About 5%–6% of patients with CP develop PDAC. Usually, initial diagnosis of PDAC and CP is done by imaging. Sensitivity and specificity in the diagnosis of pancreatic lesions have been reported as 89% and 90% for CT, 89% and 89% for MRI, 91% and 72% for positron emission tomography/CT; for endoscopic ultrasonography-guided fine-needle aspiration they are 89% and 81%.[4] Imaging-based diagnosis yields partially unclear differentiation, however, since focal pancreatic masses exist in both PDAC and CP.[5] Sensitivity and specificity of distinguishing PDAC and CP are commonly around 65% only.[6] In consequence, PDAC may be wrongly diagnosed as CP and an urgently needed treatment of patients, who face a very short survival period after diagnosis, may get delayed. Conversely, CP might be

misdiagnosed as PDAC, leading to unnecessary pancreas resection. Nearly 10% of patients who underwent surgical resection under the suspicion of pancreatic cancer have been reported to have CP instead.[7] Processes are needed to differentiate between them more reliably; improved diagnosis will facilitate the selection of appropriate therapeutic options. Biomarkers in serum or plasma are an option. Carbohydrate antigen 19-9 (CA19-9) is the most widely used serum marker for the diagnosis of PDAC. However, elevated CA19-9 levels were also detected in patients with CP, liver cirrhosis and cholangitis.[8] Many other blood biomarkers have been described, but they are not yet reliable enough for routine diagnostics.[9]

In recent years, machine-learning techniques together with the advent of large-scale, whole-genome data have promoted the identification of accurate biomarkers for cancer diagnosis.[10 11] Their application has been useful in the establishment of biomarker models for improved cancer diagnosis and therapy surveillance.[12] Concerning PDAC, previous studies focused on the identification of cancer-specific biomarkers in comparison with healthy, inflammation-free samples based on a single marker type.[13–15] Here, we introduce a random-forest machine-learning approach that takes into account datasets of different 'omic' types: messenger RNA (mRNA), microRNA (miRNA) and DNA-methylation profiles. They were compared and combined in order to select optimal biomarkers for a differential diagnosis of patients with PDAC and CP. The most promising biomarkers were substantiated and validated in independent sample cohorts, yielding a diagnostic panel of 100% accuracy in both tissue and plasma samples.

MATERIALS AND METHODS

Patient samples

Pancreatic tissue samples were collected during surgery. Initially, 345 samples (238 PDAC; 64 CP; 43 healthy individuals) were studied; for validation, another 48 samples (24 PDAC; 24 CP) were analysed. All samples were collected at the Department of Surgery of Heidelberg University Hospital during the period February 2002 to October 2009. Details about individual patient groups and related clinical information are provided in online supplemental table S1. All samples had been evaluated a second time by experienced pathologists to check the initial diagnosis. No statistical methods were used to predetermine sample size. Different from other studies,[13 16] all healthy control samples were pancreata obtained from organ donors, whose pancreas was eventually not used for transplantation. Intentionally, no tumour-adjacent non-tumour tissue was used as it has a molecular profile that at the RNA level is similar to the tumour tissue to quite an extent, although looking normal pathologically.[17] All samples were snap-frozen in liquid nitrogen directly after resection and stored at −80°C.

For validating diagnostic performance by liquid biopsy, 3 mL plasma were collected from eight patients with CP and eight with PDAC at the Department of Surgery of Heidelberg University Hospital during the period February 2019 to September 2020; samples were snap-frozen in liquid nitrogen and stored at −80°C. There was no overlap with patients, whose tissues were studied. The respective disease condition was confirmed pathologically. The patients' clinical details are listed in online supplemental table S1.

mRNA and miRNA expression profiling in tissue samples

mRNA isolation and expression profiling has been described in detail.[17] Samples were analysed on the Sentrix Human-6v3 Whole Genome Expression BeadChip (Sentrix Human WG-6; Illumina). The raw data were quantile normalised and $\log_2$-transformed. We performed differential expression analysis using the R package LIMMA (V.3.40.5) to detect differences of the PDAC sample group with the CP and healthy (N) sample groups, respectively.[18] The data are accessible at the public database ArrayExpress (ID: E-MTAB-1791; password: rpqqrysi).

miRNA expression analysis has also been published previously.[19] The Geniom Realtime Analyzer (febit biomed) with the Geniom Biochip miRNA homo sapiens was used. The data were processed as reported and deposited in the public database Gene Expression Omnibus (GEO; GSE24279).

DNA-methylation profiling in tissue samples

DNA was isolated as reported.[17] Genome-wide DNA-methylation was analysed using the Illumina Infinium 450k DNA-methylation platform (Illumina). The raw data were preprocessed using the standard workflow of RnBeads.[20] Normalisation was done using the subset-quantile within-array normalisation method (SWAN).[21] Differences with a Benjamini-Hochberg adjusted p value <0.01 and $|\log2FC|>0.5$ were considered significant when comparing sample groups. The data are accessible at ArrayExpress (DNA-methylation profiling ID: E-MTAB-3855; password: pyzqdbii).

Low-dimensional embedding and visualisation

PDAC, N and CP data were visualised using the low-dimensional embedding by Uniform Manifold Approximation and Projection (UMAP); all mRNA transcripts, miRNAs and CpG probes were considered independent features. UMAP embeddings were calculated with the R (V.3.6.1) package UMAP (V.0.2.2.0) using default settings.[22]

Feature selection

Feature selection for random-forest modelling was performed once on the complete multi-omic dataset using the following strategy: for RNA data, all probe sets on the microarrays were annotated to RefSeq-IDs. For probe sets with identical RefSeq-ID, the arithmetic mean was used. For replicate experiments, the arithmetic mean was calculated. For all datasets, missing feature values were substituted by the overall feature median. The top 1000 significantly differential features were determined by decreasing order of intergroup difference of significant features (mRNA and miRNA expression data: logarithmic fold change; DNA-methylation data: mean group difference). For multi-'omic' datasets, the top 1000 features from each 'omic' dataset were combined to generate a list of 2000 features. In all lists, correlated features were removed by applying a threshold on the calculated Pearson's correlation coefficients. For DNA-methylation, a threshold of 0.9 was used; for mRNA and miRNA expression, a threshold of 0.7 was applied.

Cross-validated training of predictive models

Separate random-forest models were trained for differentiation of PDAC versus N and PDAC versus CP samples using the R (V.3.6.1) package caret (V.6.0-81) as described below[23]: (i) we partitioned the data into 10 equal folds and performed 10 independent training/test runs. In each run, seven folds were used for training; the remaining three were used for testing; (ii) multiple random-forest models were trained with the R package ranger (V.0.11.0) on the 10 training datasets separately.[24] For each consecutive model, another uncorrelated feature was added from the filtered and ranked feature list; (iii) accuracy, sensitivity,

specificity and the area under the curve (AUC) were calculated. AUC calculation was conducted using the R package pROC (V.1.14.0).[25] For final model scores, the metrics were averaged across all folds; (iv) the best model of each single-'omic' and multi-'omic' dataset was determined by comparing AUC values. Undersampling, oversampling and the Synthetic Minority Oversampling Technique[26] were conducted and their performance was compared with the original unbalanced datasets to evaluate the influence of class imbalance on the accuracy of classification.

Validation of PDAC and N differentiation using publicly available datasets

The predictive models of PDAC versus N were validated using datasets available at the public repositories GEO and The Cancer Genome Atlas (TCGA). Not all markers were present in the public datasets because of differences in the analysis platforms. For expression, dataset GSE62452 was downloaded from GEO, which contains data of 69 PDAC and 61 adjacent non-tumour tissue samples. It was processed using the oligo package[27] and quantile normalised. For DNA-methylation, we used GSE49149 with results from 155 PDAC and 19 adjacent non-tumour tissue samples. The raw data were preprocessed using the minfi package[28] and SWAN.[21] The best performing models were validated in the public datasets using the R package caret. Accuracy, sensitivity, specificity and AUC values were calculated, and receiver operating characteristic (ROC) curves were generated using the R package pROC. The predictive markers of the models were validated by performing two-sided t-tests.

Validation of DNA-methylation markers for PDAC and CP differentiation with independent sample set

For validating DNA-methylation markers that differentiate PDAC from CP samples, MethyLight qPCR was performed on samples from an independent patient cohort (online supplemental table S1). DNA from 24 PDAC and 24 CP tissue samples was bisulfite converted (EpiTect bisulfite kit; Qiagen). Pilot experiments were performed using calibration DNA of 100%, 75%, 50%, 25%, 12.5% and 0% methylation (EpiTect PCR Control DNA Set; Qiagen). MethyLight qPCR was done in 10 μL containing 5 μL 2× EpiTect MethyLight Master Mix (without ROX) (EpiTect MethyLight PCR kit; Qiagen), 400 nM forward and reverse primers, 200 nM probe and 3 μL bisulfite converted DNA template. Simultaneously, the C-LESS-C1 primer/probe set was used in each reaction as internal control for normalisation.[29] MethyLight qPCR was performed in triplicates on a LightCycler 480 (Roche) with a pre-amplification incubation of 95°C for 5 min, followed by 45 cycles of 95°C for 15 s and the primer/probe set-specific annealing temperature (online supplemental table S2) for 30 s. Raw data were analysed using the LightCycler software (Roche). To evaluate the DNA-methylation level of the region of interest, the methylation index was calculated. DNA-methylation index$=\log_2[2^{(-\Delta\Delta Ct)}]$, $\Delta\Delta Ct = \Delta Ct_{\text{sample (ROI-MIP)}} - \Delta Ct_{\text{100\% methylation control sample (ROI-MIP)}}$; ROI stands for 'region of interest primer/probe', and MIP for 'methylation independent primer/probe', which is the C-LESS-C1 primer/probe set. DNA-methylation indices were scaled to a range from 0 to 1.

The statistical analyses of MethyLight qPCR results were performed using the GraphPad Prism V.6 software (GraphPad Software). Mean and SD are reported. Comparison of the normally distributed variable's values in the two groups was performed using t-test. All p values were two-sided, and $p<0.05$ was considered statistically significant. A predictive model was trained on the initial microarray data as outlined previously. To apply the model to the MethyLight qPCR data, we normalised them to the range of 0–1 to match the distribution of methylation microarray CpG level values. The predictive power was determined using the R package caret.

Co-methylation analysis of CpG islands

For visualising the overall DNA-methylation levels of genomic ROI and for estimating DNA-methylation correlation between different CpG sites within genomic regions, the R package coMET was applied.[30] The DNA-methylation 450k array data from 26 PDAC samples and 12 CP samples were used. Gene tracks from ENSEMBL and CpG island tracks from UCSC based on GRCh37/hg19 were selected as genomic annotations. Spearman's correlation coefficients of DNA-methylation levels and disease state were calculated for all CpG sites within the selected gene regions. We evaluated the diagnostic power of selected co-methylated CpG sites by a cross-validated prediction of patient disease states in the original DNA-methylation 450k array data.

Analysing DNA-methylation levels in healthy tissues

Whole-genome bisulfite sequencing data from breast, oesophagus, heart, lung, muscle, pituitary, skin and thyroid tissues were available at the Genotype-Tissue Expression (GTEx) repository. Beta values of the CpG sites in the protein kinase C beta type gene (*PRKCB*) region chr16:23,836,004-23,836,682 (GRCh38r/hg38) were retrieved and visualised in the GTEx portal. Additionally, DNA-methylation data available at TCGA was explored. DNMIVD[31] and MethyHC[32] were used for analysing 23 or 33 non-tumorous tissue types, respectively.

Validating diagnostic performance of cell-free DNA-methylation changes in liquid biopsy

Cell-free DNA (cfDNA) was isolated from 1 to 2 mL of patient plasma using the QIAamp MinElute ccfDNA Mini Kit (Qiagen) according to the manufacturer's instructions and recovered in 60 μL ultra-clean water. DNA concentration was determined using the Qubit dsDNA HS Assay Kit (Invitrogen). For bisulfite conversion, cfDNA was treated with the EpTect Bisulfite Kit (Qiagen). Afterwards, 50–100 ng converted cfDNA were amplified with the EpiTect Whole Bisulfitome Kit (Qiagen).

For the preparation of next-generation sequencing libraries, 1 µg amplified DNA was used for an end-repair reaction (NEBNext Ultra End Repair/dA-Tailing Module; New England Biolabs). Adapters were added (NEBNext Ultra II DNA Library Prep Kit for Illumina), and the DNA fragments then amplified with index primers (NEBNext Multiplex Oligos for Illumina). For library enrichment, the NEBNext Ultra II DNA Library Prep Kit for Illumina was used. Each library was prepared and diluted to an equal molar concentration of 10 nM. Libraries were sequenced on an Illumina Novaseq 6000 S4 instrument using paired-end 150 bp reads. The data were processed with the nf-core/methylseq pipeline (V.2.3.0)[33] using Nextflow (V.22.10.6).[34] In the pipeline, we applied the BWAmeth aligner,[35] the GRCh38 reference sequence and methyldackel quantification. DNA-methylation levels were quantified as M values.[36 37]

RESULTS

Clustering does not exhibit sufficient diagnostic power for discriminating PDAC and CP

Initially, we used tissue samples from 345 patients: 238 with PDAC; 64 with CP and 43 healthy individuals (N) (for more details see online supplemental table S1). All samples had been

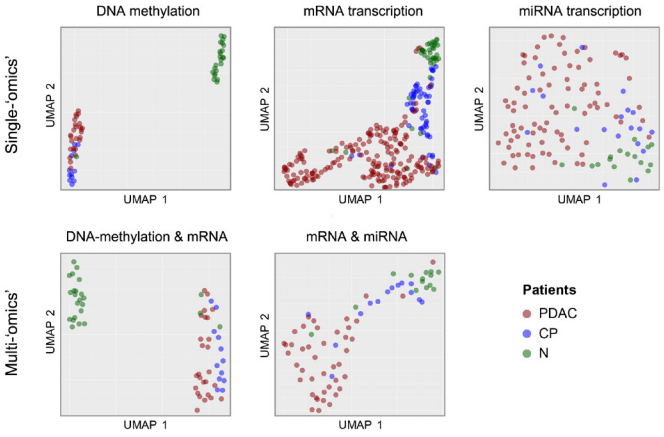

Figure 1 Low-dimensional embedding of cohort samples using UMAP dimensionality reduction of different 'omic' datasets. Results are shown that were obtained by applying DNA-methylation, mRNA expression or miRNA expression data, respectively, or by using combined mRNA expression and DNA-methylation or combined mRNA and miRNA expression data. Individual tissue samples are colour-coded as indicated: PDAC, red; CP, blue; N, green. CP, chronic pancreatitis; mRNA, messenger RNA; miRNA, microRNA; N, healthy individuals; PDAC, pancreatic ductal adenocarcinoma; UMAP, Uniform Manifold Approximation and Projection.

evaluated a second time by experienced pathologists to check the initial diagnosis. Quality and utility of the material have been reported before.[17 19 38] Genome-wide information about DNA-methylation, mRNA and miRNA expression was produced. Besides exploring each dataset individually, we combined mRNA and DNA-methylation data as well as mRNA and miRNA data to two multi-'omic' datasets. All CpG methylation sites, mRNA and miRNA transcripts were considered independent features. Patient data of the three single-'omic' and two multi-'omic' analyses were embedded in a low-dimensional space for visualisation (figure 1) using UMAP dimensionality reduction.[22] To detect confounding factors within our patient cohort, clustering of patients with equal co-variables, such as age, sex and tumour localisation, were evaluated. No bias by these factors could be detected (online supplemental figure S1). Since especially DNA-methylation has been reported to be associated with smoking, diabetic status and alcohol consumption, we additionally looked at these three factors, too. Again, there was no apparent bias (online supplemental figure S2 and S3).

Clustering according to medical condition (figure 1) indicated global differences over all 'omic' levels. Especially, N samples are distinct from PDAC in all five datasets. In contrast, CP tissues are embedded more closely to PDAC. For mRNA, miRNA and combined mRNA and miRNA data, CP samples fall in-between PDAC and N tissues; some CP samples are even embedded in the PDAC cluster. Similarly, several CP samples are located in the PDAC cluster in the DNA-methylation and combined mRNA and DNA-methylation data. However, PDAC and CP are clearly separated from the N cluster. The UMAP embeddings confirmed that CP and PDAC share greater similarities on a molecular level than PDAC and N do. This molecular resemblance represents a challenge for establishing reliable cancer-specific markers.[39]

To overcome this, we used a random-forest-based machine-learning approach for identifying molecular features that may unambiguously differentiate PDAC from CP. An overview of the entire process is presented in figure 2. First, a machine-learning procedure was set up and evaluated. Once established, it was used on tissues to evaluate the performance of different molecular classifier types. The best classifier was pursued further by validation on an independent samples set and finally on material isolated from blood samples in order to demonstrate its wide applicability and accuracy.

Establishment of machine-learning workflow

While the study's objective was differentiating PDAC from CP, we first used PDAC and N samples to establish the machine-learning workflow (figure 2). Random-forest models were trained separately on the single-'omic' and multi-'omic' datasets as described in the 'Materials and methods' section. No significant effect of class size imbalances was observed (online supplemental figure S4). An ROC curve was calculated for each model, from which the AUC values were determined: they were used to define the best model of each dataset. The trained models for the differentiation of PDAC and N exhibited cross-validated AUCs between 0.85 and 0.98, when features of the respective dataset were added consecutively (figure 3A). A predictive model based on DNA-methylation showed the highest AUC value—0.980—with the lowest number of features, namely cg02964172 (gene *GCNT2*, gene body) and cg17184704 (intergenic, chr. 10: 11727286). The next best models were trained on combined DNA-methylation and mRNA data (AUC of 0.977, 5 features), the miRNA and mRNA data (AUC of 0.955, 4 features), the miRNA data (AUC of 0.980, 12 features) and the mRNA data (AUC of 0.946, 14 features) (online supplemental table S3).

We focused on the DNA-methylation marker panel, since it showed the best performance and required the lowest number of features. Unsupervised hierarchical clustering based on the model's two features showed a clear separation of PDAC and N samples (figure 3B); clustering for the other models is shown in online supplemental figure S5A-D. To test the reliability of the DNA-methylation markers, the random-forest model was applied to the dataset GSE49149 that is available at the public GEO repository. It represents data generated from PDAC and normal controls and yielded a sensitivity of 0.981 and specificity of 0.579 in diagnosing the medical status of patients, with an AUC value of 0.810 (figure 3C). The low specificity was not surprising since the controls (N samples) in the GEO dataset were adjacent non-tumour tissues of patients with PDAC, while the markers had been defined in our dataset with controls that all were pancreata from healthy donors. It is known that they differ substantially.[17] Individually, the DNA-methylation level of cg02964172 in PDAC was significantly higher than in N samples, while the opposite was true for cg17184704 (figure 3D). The results of validating other 'omic' models with public data can be found in online supplemental table S4 and online supplemental figure S6. The results confirmed the potential of random-forest modelling for biomarker identification.

Differentiating PDAC and CP tissue samples

For the classification of PDAC and CP, we employed the established machine-learning workflow. The five different 'omic' datasets were used to construct predictive models. A model consisting of four DNA-methylation features—cg15506157 (killer cell lectin like receptor G2 (*KLRG2*), TSS200 (0–200 bases upstream of transcription start site)), cg03306374 (*PRKCB*, first exon/5'-untranslated region (UTR)), cg21294301 (intergenic, chr. 1: 8120055) and cg27341866 (C19orf35, gene body)—was the most accurate for differentiating PDAC and CP with an AUC value of 1.00 (figure 4A). The best model of the combined mRNA and DNA-methylation dataset also yielded an AUC of

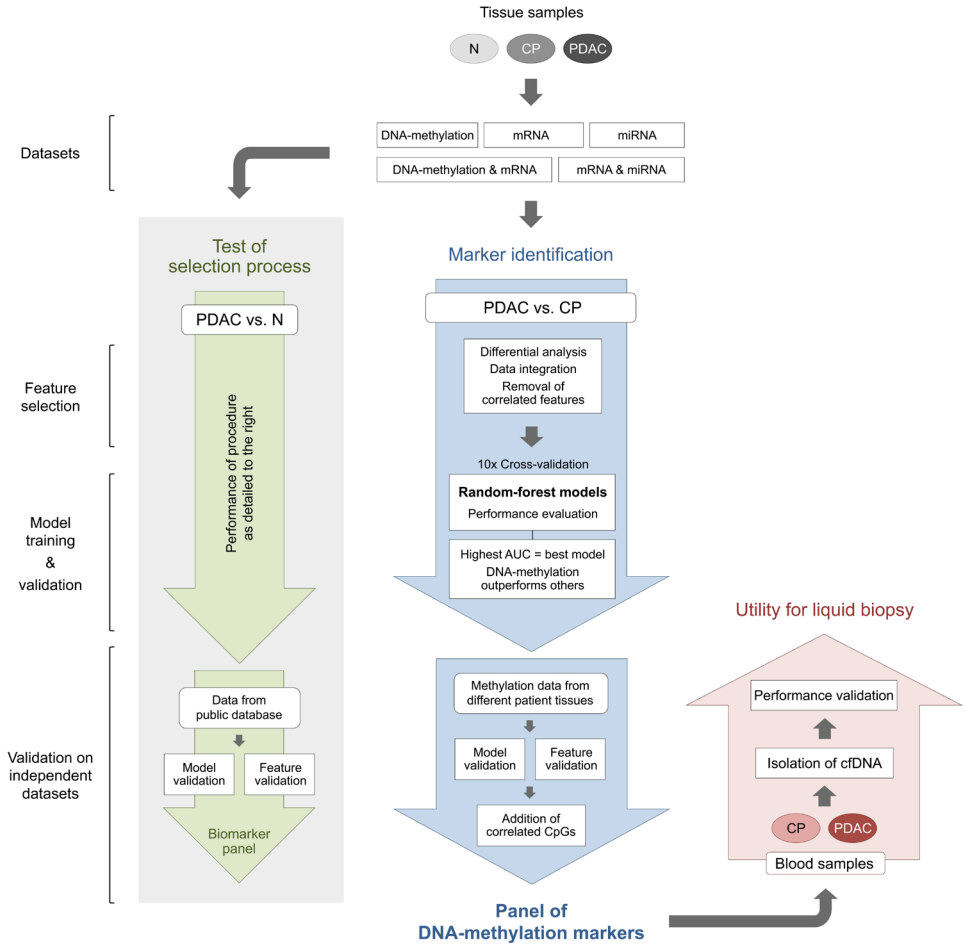

Figure 2 Schematic workflow of biomarker selection process. Single-'omic' and multi-'omic' data from PDAC, CP and N tissue samples were collected and analysed. The workflow consists of four main steps: feature selection, model training, internal validation and validation on independent datasets. AUC, area under the curve; cfDNA, cell-free DNA; CP, chronic pancreatitis; mRNA, messenger RNA; miRNA, microRNA; N, healthy individuals; PDAC, pancreatic ductal adenocarcinoma.

1.00. However, 11 features were required: the DNA-methylation features cg15506157 (*KLRG2*, TSS200), cg21294301 (intergenic, chr. 1: 8120055), cg27341866 (C19orf35, gene body), cg11141652 (*GSTTP1*, TSS1500), cg11792281 (*NLK*, gene body), cg15138289 (*HLA-DPB1*, gene body), cg05137263 (*NR0B1*, TSS200), cg13686615 (intergenic, chr. 2: 71503742) and cg05795005 (*LIN7C/BDNFOS*, TSS200/gene body) as well as the mRNA features NM_005980.2 (*S100P*), and NM_001008218.1 (*AMY1B*). The best models from the mRNA dataset (AUC of 0.962 with 54 features), the combined mRNA and miRNA dataset (AUC of 0.953 with 13 features) and the miRNA dataset (AUC of 0.783 with 23 features) performed worse than the DNA-methylation and combined DNA-methylation and mRNA models. The features of the best model for each dataset can be found in online supplemental table S5.

The highest AUC value with the smallest feature number was achieved by the model trained on DNA-methylation. Also the second-best model, based on mRNA and DNA-methylation, was dominated by methylation markers. It has been shown that DNA-methylation profiling is highly robust and stable even in DNA isolated from poor-quality material.[40] Therefore, we selected five DNA-methylation biomarkers with high performance from these two models for independent validation: cg03306374 (in gene *PRKCB*), cg05795005 (*LIN7C*), cg11792281 (*NLK*), cg15506157 (*KLRG2*) and cg27341866 (C19orf35) (online supplemental figure S7A). They exhibited the lowest p values and SD in differential PDAC versus CP analysis. Unsupervised hierarchical clustering revealed that the five markers separated PDAC from CP accurately (figure 4B). Clustering results based on the other 'omic' models are shown in online supplemental figure S7B-E.

Validation of DNA-methylation features for discriminating PDAC from CP

Since no public DNA-methylation profiling datasets were available for PDAC versus CP differentiation, we validated the five DNA-methylation markers by MethyLight qPCR on DNA from an independent cohort of 24 patients with PDAC and 24 patients with CP (online supplemental table S1). Besides employing another technology in order to check for methodological bias, using an entirely independent set of samples should compensate for any potential overestimation of accuracy during the initial analysis. DNA-methylation levels of cg03306374 (*PRKCB*) and cg15506157 (*KLRG2*) were significantly higher in PDAC compared with CP samples ($p<0.0001$), while the other markers did not show a significant difference (figure 4C). For the two significantly different CpG sites, we observed several outliers in the PDAC and CP sample groups. We revisited the histology evaluation of these patients. For cg15506157, for example, the six outliers in the PDAC sample group did not exhibit clear-cut cancer in the second pathological evaluation and the three CP

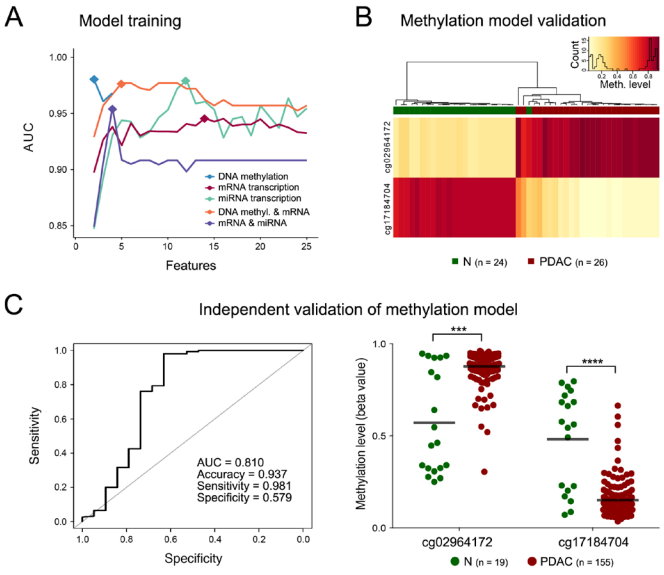

Figure 3 Evaluating the machine-learning performance on PDAC and N samples. (A) Comparison of 10-fold cross-validated model performances. For each 'omic' dataset, the relationship between the model performance measured by AUC value and the feature number is indicated. For each analysis, the highest AUC value achieved with the smallest possible number of features is marked by a rectangle. (B) Unsupervised hierarchical clustering of the two CpG marker sites cg02964172 and cg17184704 as defined by the best performing DNA-methylation model. (C) Left, the ROC curve of the diagnostic prediction model is shown as calculated with the two CpG sites in the public validation dataset GSE49149. On the right, the methylation levels of the two CpGs are shown as determined in PDAC and N samples from the public validation dataset. Mean methylation values are indicated by horizontal lines. Differential analysis was performed by t-test (***p<0.001, ****p<0.0001). AUC, area under the curve; mRNA, messenger RNA; miRNA, microRNA; N, healthy individuals; PDAC, pancreatic ductal adenocarcinoma; ROC, receiver operating characteristic.

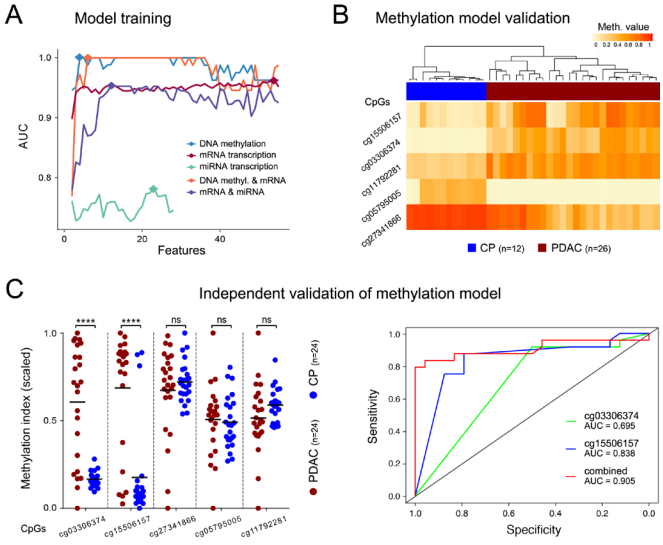

Figure 4 Identification and validation of biomarkers for the differentiation of PDAC and CP. (A) Comparison of 10-fold cross-validated model performances. For each 'omic' dataset, the relationship between the model performance measured by AUC value and the feature number is indicated. The highest AUC value with the smallest possible number of features is marked by a rectangle. (B) Unsupervised hierarchical clustering of five methylation markers selected from the best predictive models in the DNA-methylation and the combined DNA-methylation and mRNA expression datasets. (C) Left, the normalised methylation index is shown of five selected methylation markers in 24 PDAC and 24 CP samples. Experimental validation was by MethyLight qPCR. The mean values of normalised methylation indexes are indicated by horizontal lines. Differential analysis was performed by t-test (****p<0.0001; ns, not significant). In the right panel, ROC curves are shown of single and combined (red) predictive markers cg03306374 (green) and cg15506157 (blue) as calculated from the independent validation dataset. AUC, area under the curve; CP, chronic pancreatitis; mRNA, messenger RNA; miRNA, microRNA; PDAC, pancreatic ductal adenocarcinoma.

outliers did not show a distinct but only a moderate CP pathology. Irrespective of this, however, the power of cg03306374 and cg15506157 was evaluated with models trained on the original DNA-methylation profiling data. The sites could differentiate PDAC from CP samples individually with an AUC of 0.695 and 0.838, respectively (figure 4D). The combined model of both methylation markers discriminated PDAC from CP with an AUC of 0.905.

Regional co-methylation around cg03306374 enables fully accurate diagnostics

Site cg03306374 represents a CpG in the first exon/5'-UTR of the *PRKCB*; cg15506157 is a site in the 200 bp region upstream of the transcription start of gene *KLRG2*. Analysing clusters of co-methylated CpGs could be more comprehensive and robust than analysing individual CpG sites. We used the 450k DNA-methylation data to study the correlation between methylation levels of multiple CpG sites within the genomic ROI and the association between their methylation levels and the disease state. The correlation matrix in figure 5A shows co-methylation patterns between 38 CpGs within the genomic region of *PRKCB*. Sites located in the CpG island at the 5'-end were positively correlated with PDAC-hypermethylated cg03306374 relative to CP samples. More distant CpG sites outside of the CpG island showed predominantly negative correlation. Five positively correlated CpG sites were significantly associated with the PDAC phenotype. Co-methylation patterns were also observed at 17 CpG sites in the *KLRG2* gene region (figure 5B). The sites located in the CpG island at the 5'-end of *KLRG2* were positively correlated with the PDAC-hypermethylated cg15506157 relative to CP samples. Three were significantly associated with the PDAC disease status.

Methylation of a particular CpG may vary in individual patients, even if it exhibits a significant diagnostic performance overall. We therefore investigated whether combining cg03306374 and the five co-methylated CpGs in *PRKCB* (cg03156893, cg03217795, cg05436658, cg09507526, cg21370856) as well as cg15506157 and the three highly correlated CpGs in *KLRG2* (cg00699934, cg00919016, cg05224190) into a panel of 10 methylation sites may improve the robustness of classifying individual patients. Using 24 PDAC and 24 CP samples, the classifiers of these DNA-methylation markers could differentiate PDAC from CP with absolute accuracy (figure 6). We observed an AUC value of 1.00 with the 10 CpG sites instead of 0.905 with only cg03306374 and cg15506157. As a matter of fact, already the six sites in *PRKCB* produced an AUC of 1.00 if used on their own, while the four CpGs in *KLRG2* yielded an AUC of 0.934 on their own. This documents that the six *PRKCB* methylation sites are

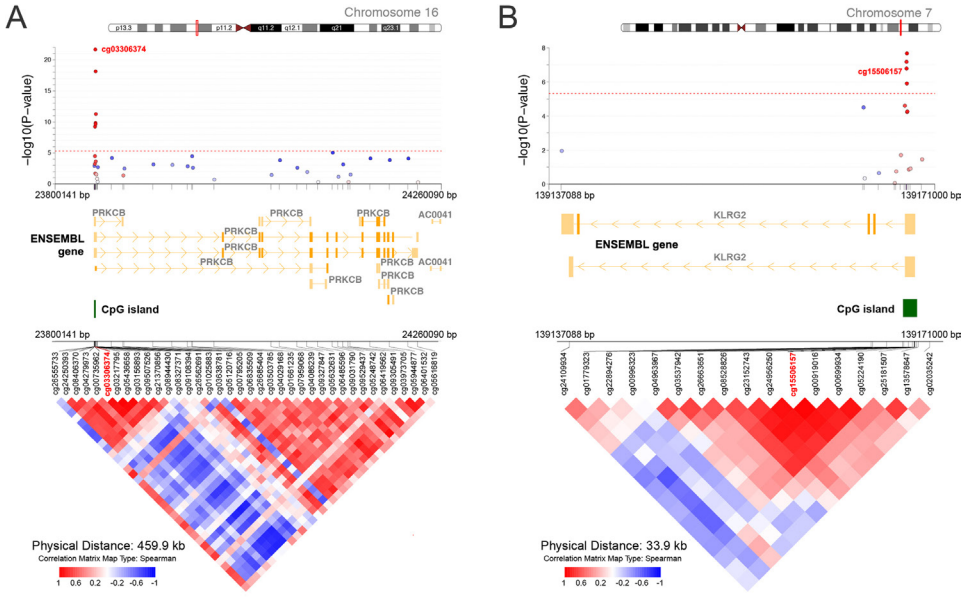

Figure 5 Regional plot of an epigenome-wide association analysis of cg03306374 and cg15506157 in PDAC and CP tissues. Co-methylation patterns were found in the PRKCB (A) and KLRG2 gene regions (B), respectively. At the very top, the gene position within the respective chromosome is shown. Below, each dot stands for a particular CpG site. Its genomic position is indicated along the X-axis. The negative log-transformed p value reflects the association of a CpG site with the disease status. The panels in the middle show the ENSEMBL annotation tracks including genes/transcripts and the direction of transcription. Also, the positions of the CpG sites (vertical lines) are indicated. The lower panels present Spearman's correlation coefficients of DNA-methylation levels between selected CpG sites in the two genomic regions. The colour scheme of the heatmap is reflected also in the association panel at the top with respect to correlation to the reference CpG sites cg03306374 and cg15506157. Blue colouring of a dot stands for low correlation values; red colouring indicates high correlation values. CP, chronic pancreatitis; PDAC, pancreatic ductal adenocarcinoma.

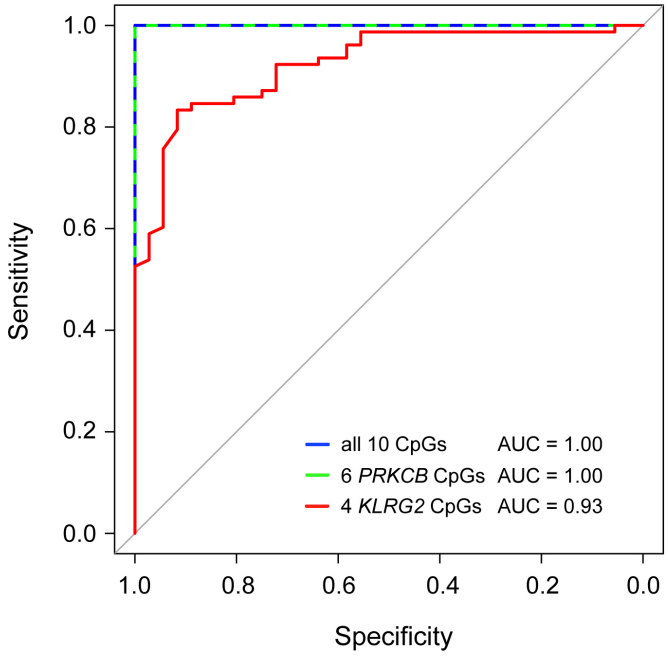

Figure 6 Differentiation between PDAC and CP based on DNA-methylation. ROC curves are shown that were calculated on the basis of differential DNA-methylation at cg15506157 and the three other CpGs in the KLRG2 gene region (red line), cg03306374 and the five surrounding DNA-methylation sites in the PRKCB gene region (green line) or a combination of all 10 sites (blue line). AUC, area under the curve; CP, chronic pancreatitis; PDAC, pancreatic ductal adenocarcinoma.

sufficiently accurate. Furthermore, they represent a marker panel that is robust enough in its diagnostic performance to compensate for other, less accurate diagnostic information, such as the one resulting from the *KLRG2* CpGs, and may therefore well be suited for clinical diagnostics.

Diagnosis in liquid biopsy samples

To evaluate the performance of the methylation signature in liquid biopsy, we performed whole-genome bisulfite sequencing of cfDNA that was isolated from plasma samples collected from patients with PDAC and CP (online supplemental table S1). On purpose, the samples were from patients, who had not been studied as part of the tissue sample analyses so as to assure independent confirmation. We did initial analyses by pyrosequencing, droplet-PCR and targeted next-generation sequencing. They all required PCR amplification of particular regions on bisulfite-converted cfDNA templates. This proved to be technically challenging because of the limited amount of initial material as well as its fragmentation, which was further exacerbated on bisulfite treatment. We therefore performed whole-genome bisulfite sequencing eventually. The rationale was to avoid bias introduced by PCR-based enrichment of particular target regions. In accordance with the results of the tissue analysis, the comparison of PDAC and CP samples (online supplemental table S6) produced clear differences in methylation level for all six CpG sites of the *PRKCB* gene individually (figure 7); only cg21370856 was not quite significant after false discovery rate adjustment. Also for the CpG sites in *KLRG2*, the degree of DNA-methylation was higher in PDAC than CP samples. For all four of them, however, the increase was not significant. These results are consistent with

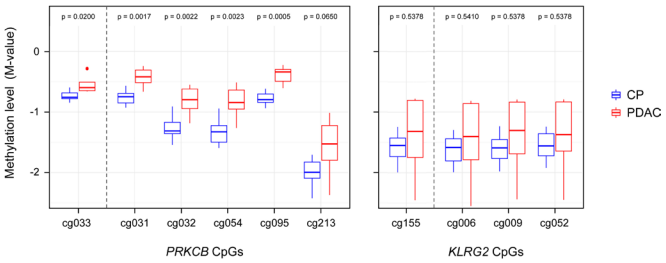

Figure 7 Diagnostic accuracy of DNA-methylation in liquid biopsy samples. Variation in DNA-methylation in cell-free DNA isolated from patient plasma is presented as box plot for the six CPG sites in gene PRKCB and the four CpG sites in KLRG2. The M-value was chosen as a measure of methylation.[54] At the top, the respective false discovery rate adjusted p values are given. Only the first three numbers of the CpG identifiers are indicated: cg033: cg03306374; cg031: cg03156893; cg032: cg03217795; cg054: cg05436658; cg095: cg09507526; cg213: cg21370856; cg155: cg15506157; cg006: cg00699934; cg009: cg00919016; cg052: cg05224190. CP, chronic pancreatitis; PDAC, pancreatic ductal adenocarcinoma.

the diagnostic performance in tissue, in which the *PRKCB* CpGs clearly outperformed the *KLRG2* sites.

Already most individual CpGs would have made discrimination possible between PDAC and CP, although not with absolute accuracy. For improving accuracy and robustness further so that it could potentially be sufficient for clinical application, we used the methylation values of the six *PRKCB* CpG sites to train a support vector machine (SVM) classifier.[41] The data were randomly split into training and validation sets (each made of 4× PDAC and 4× CP). During the training process, a combination of three parameters (cost, gamma, epsilon) was optimised to avoid overfitting and underfitting. The classifier with fixed parameter settings was subsequently assessed on the validation data. This process was performed repeatedly. With SVM parameters of cost=0.1, gamma=0.03 and epsilon=0.9, an AUC value of 100% for discriminating PDAC from CP was obtained from an ROC curve analysis.

Whole-genome bisulfite sequencing data were analysed that had been generated from healthy tissues (breast, oesophagus, heart, lung, muscle, pituitary, skin and thyroid) and deposited at the public repository GTEx. Beta values of the *PRKCB* CpG sites ranged from 0.00 to 0.06, indicating that there is no significant DNA-methylation in normal tissues. Additionally, DNA-methylation data from 20 healthy tissue types available at TCGA was explored; there was no record for cg05436658. No beta value >0.1 was found for the *PRKCB* CpGs, again demonstrating low methylation of the CpGs. In comparison, mean beta values >0.6 were found in PDAC tissues. These results suggest that the DNA-methylation variation found in cfDNA of PDAC patients is tumour-specific.

DISCUSSION

CP is a long-term inflammation of the pancreas that alters the organ's normal structure and functionality and has long been recognised as a risk factor for PDAC.[2] Some pseudo-tumours are not due to CP but areas of focal lobulocentric atrophy with nearby PanIN. However, because of its predominance, we focused on CP. Because of molecular similarities, it complicates the diagnosis of PDAC. This has implications on the survival of patients with PDAC, who face a very short mean survival period of few months after initial diagnosis. Sensitivity and specificity in the differentiation of pancreatic cancer and CP are commonly about 60%–65%.[6] Even ultrasonography-guided fine-needle aspiration does not perform better than imaging.[4] Visualisation of the mRNA, miRNA and DNA-methylation profiles by UMAP embedding corroborates the difficulty of distinguishing between PDAC and CP. While it showed a clear difference between PDAC and N samples, a joint cluster of PDAC and CP samples was produced with all 'omic' datasets. This is probably due to many dysregulated pathways that are common to CP and PDAC.[42]

With the help of machine-learning methods, one can analyse molecular characteristics in a comprehensive manner and reveal underlying phenotype-genotype relationships,[12 43] thus facilitating marker identification. We evaluated the diagnostic power of different 'omic' datasets. Advanced integration methods are available, such as correlation-based,[44] association-based,[45] prior knowledge-based[46] and model-based[47] integration. Since we aimed at identifying multi-'omic' features independent of their potential connection to features of other datasets, we applied a simple concatenation of the multi-'omic' feature spaces. This enabled us to treat them as independent features during the subsequent machine learning. As opposed to low-dimensional embedding, the machine-learning identified potential biomarker panels for the differentiation of PDAC and CP.

Two DNA-methylation markers—cg03306374 (*PRKCB*) and cg15506157 (*KLRG2*)—yielded an AUC of 0.905 on validation. These markers have already been reported in the list of top 20 differentially methylated CpG sites when comparing PDAC with N tissue data from TCGA.[48] Here, we could show that a predictive model containing these two DNA-methylation markers is able to classify correctly PDAC and CP tissue samples, which is a more demanding challenge. Taking into account the two CpGs and the differentially methylated region (DMR) around them, we improved diagnostic performance to 100% accuracy, a level that was also obtained with only the six methylations sites in gene *PRKCB*. The not entirely accurate distinction of the four CpGs in gene *KLRG2* did not affect accuracy on combination with the six *PRKCB* sites. This documents a degree of information redundancy of the *PRKCB* CpGs that is likely to warrant assay robustness in real-life applications. Furthermore, the six *PRKCB* CpGs were found to be equally informative in liquid biopsy samples. There was no apparent difference between treatment-naïve patients and others. However, the sample number is too small for identification of more subtle effects. The result suggests a diagnostic approach that is potentially superior to imaging and based on an essentially non-invasive process. However, the sample number in our study is clearly not sufficient for a proof of clinical utility. Towards this end, a clinical trial in a multicentric set-up has to be performed with a substantially larger number of samples.

The DMR around cg03306374 is a CpG island in the first exon/5'-UTR region of *PRKCB*. The protein kinase C (PKC) family has come to the focus of cancer research, since the receptor plays a role for tumour-promoting phorbol esters.[49] A recent mouse model study revealed that not the inactivation of PKC but its activation suppresses tumour growth.[50] *PRKCB* regulates the expression of PKCB, PRKCB1 and PRKCB2.[51] Interestingly, the expression of *PRKCB1* was upregulated in PDAC and CP compared with N tissue samples, while it did not significantly differ between PDAC and CP. The biological effect of the DMR in the *PRKCB* promoter remains elusive. A recent pilot study on pancreatic juice also identified hypermethylation in the *PRKCB* gene region to be associated with differentiation of PDAC and CP, although with an AUC of only 0.77.[52]

In conclusion, the identified DNA-methylation signature may significantly improve the quality of tumour diagnostics in

patients with suspected pancreatic cancer. However, two major points still need to be met towards clinical application. Other analysis platforms but sequencing exist that yield good results.[53] However, whole-genome bisulfite sequencing was required in our analysis, since processes that need PCR-amplification of the region of interest did not perform well. Most likely, this was due to a combination of limiting factors: the number of cfDNA copies that can be isolated from blood samples of patients with PDAC is very small; the bisulfite treatment further breaks the already fragmented cfDNA; and finally, representing a CpG island, the DNA of interest is A:T-rich after bisulfite conversion and does not yield good results with respect to PCR-based amplification. Alternative amplification processes could overcome this problem or bisulfite sequencing could be automated further and thus become cheap enough for routine diagnostics. Second, translation to clinical routine could be achieved only after a successful multicentre evaluation of the results on a substantially larger patient cohort and a clinical trial.

The results add to the growing body of evidence that DNA-methylation by its epigenetic nature may be better suited to act as disease-specific marker than the more stable genetic factors on the one hand or the more complexly regulated transcriptional variations on the other hand. The machine-learning algorithm used for selecting the biomarkers could be applied to many other diseases, for which relevant data exist.

Author affiliations
[1]Division of Functional Genome Analysis, German Cancer Research Center (DKFZ), Heidelberg, Germany
[2]Faculty of Biosciences, Heidelberg University, Heidelberg, Germany
[3]Medical Faculty Heidelberg, University of Heidelberg, Heidelberg, Germany
[4]Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany
[5]Institute of Pathology, Universitätsklinikum Erlangen, Friedrich Alexander Universität Erlangen-Nürnberg, Erlangen, Germany
[6]Department of Surgery, Heidelberg University Hospital, Heidelberg, Germany

Acknowledgements We thank the EPZ-Pancobank (Biobank of the European Pancreas Centre at the Department of General Surgery, University Hospital Heidelberg) for provision of the human specimens and clinical data. The DKFZ Sequencing Core Facility was instrumental for the reported sequence analyses.

Contributors YW was involved in the study's conceptualisation, improved critical methods, performed and analysed experiments, interpreted results and was involved in writing the initial manuscript and its final editing. IS was involved in the study's conceptualisation, developed machine learning software tools, performed data analysis and interpretation and was involved in writing the initial manuscript and its final editing. FNA-S improved methods for liquid biopsy analyses, performed the cell-free DNA (cfDNA) experiments, analysed and interpreted data and was involved in writing the initial manuscript and its final editing. RK supported the development and implementation of software tools. ASB performed data curation, analysis and interpretation. MM was involved in methylation data analysis and interpretation. EAM performed experiments for cfDNA isolation and developed processes for their analysis. BB supported and led and supported developments, implementation and use of software tools. CT obtained patient material and related clinical information and was involved in data curation. NAG provided resources and related clinical information, performed validation experiments, interpreted results and was involved in the final manuscript editing. TH was involved in the study's conceptualisation, performed project administration, contributed samples and information, co-supervised the study and was involved in the final manuscript editing. MWB was involved in the study's conceptualisation, performed project administration, contributed resources, co-supervised the study and was involved in the final manuscript editing. JDH was involved in the study's conceptualisation, analysed and interpreted data, performed project administration and coordination, co-supervised the study, was involved in writing the initial manuscript and its final editing and is responsible for the overall content as guarantor.

Funding The EPZ-Pancobank was supported by the Heidelberger Stiftung Chirurgie and the German Federal Ministry of Education and Research (BMBF) grants 01ZX1305C, 01ZX1605C, 01KT1506 and 01EY1701. Financial support of Yenan Wu by a CSC-fellowship (201406350109) and Fawaz Al-Shaheri by a DAAD-fellowship (91559475) is gratefully acknowledged.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study was approved by ethics committee at the University of Heidelberg (references 301/2001 and S-708/2019). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request.

ORCID iDs
Isabelle Seufert http://orcid.org/0000-0001-9811-3836
Jörg D Hoheisel http://orcid.org/0000-0002-1583-5049

REFERENCES

1 Siegel RL, Miller KD, Fuchs HE, *et al*. Cancer statistics, 2021. *CA Cancer J Clin* 2021;71:7–33.
2 Park W, Chawla A, O'Reilly EM. Pancreatic cancer: a review. *JAMA* 2021;326:851–62.
3 Majumder S, Chari ST. Chronic pancreatitis. *Lancet* 2016;387:1957–66.
4 Treadwell JR, Zafar HM, Mitchell MD, *et al*. Imaging tests for the diagnosis and staging of pancreatic adenocarcinoma: a meta-analysis. *Pancreas* 2016;45:789–95.
5 Zhang L, Sanagapalli S, Stoita A. Challenges in diagnosis of pancreatic cancer. *World J Gastroenterol* 2018;24:2047–60.
6 de Icaza E, López-Cervantes M, Arredondo A, *et al*. Likelihood ratios of clinical, laboratory and image data of pancreatic cancer: Bayesian approach. *J Eval Clin Pract* 2009;15:62–8.
7 De Castro SMM, De Nes LCF, Nio CY, *et al*. Incidence and characteristics of chronic and lymphoplasmacytic sclerosing pancreatitis in patients scheduled to undergo a pancreatoduodenectomy. *HPB* 2010;12:15–21.
8 Duffy MJ, Sturgeon C, Lamerz R, *et al*. Tumor markers in pancreatic cancer: a European group on tumor markers (EGTM) status report. *Ann Oncol* 2010;21:441–7.
9 Al-Shaheri FN, Alhamdani MSS, Bauer AS, *et al*. Blood biomarkers for differential diagnosis and early detection of pancreatic cancer. *Cancer Treat Rev* 2021;96:102193.
10 Xu R-H, Wei W, Krawczyk M, *et al*. Circulating tumour DNA-methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nat Mater* 2017;16:1155–61.
11 Capper D, Jones DTW, Sill M, *et al*. DNA-methylation-based classification of central nervous system tumours. *Nature* 2018;555:469–74.
12 Kourou K, Exarchos TP, Exarchos KP, *et al*. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 2015;13:8–17.
13 Nones K, Waddell N, Song S, *et al*. Genome-wide DNA-methylation patterns in pancreatic ductal adenocarcinoma reveal epigenetic deregulation of SLIT-ROBO, ITGA2 and MET signaling. *Int J Cancer* 2014;135:1110–8.
14 Klett H, Fuellgraf H, Levit-Zerdoun E, *et al*. Identification and validation of a diagnostic and prognostic multi-gene biomarker panel for pancreatic ductal adenocarcinoma. *Front Genet* 2018;9:108.
15 Lomberk G, Blum Y, Nicolle R, *et al*. Distinct epigenetic landscapes underlie the pathobiology of pancreatic cancer subtypes. *Nat Commun* 2018;9:1978.
16 Waddell N, Pajic M, Patch A-M, *et al*. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* 2015;518:495–501.
17 Bauer AS, Nazarov PV, Giese NA, *et al*. Transcriptional variations in the wider peritumoral tissue environment of pancreatic cancer. *Int J Cancer* 2018;142:1010–21.
18 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.
19 Bauer AS, Keller A, Costello E, *et al*. Diagnosis of pancreatic ductal adenocarcinoma and chronic pancreatitis by measurement of microRNA abundance in blood and tissue. *PLoS One* 2012;7:e34151.
20 Assenov Y, Müller F, Lutsik P, *et al*. Comprehensive analysis of DNA-methylation data with RnBeads. *Nat Methods* 2014;11:1138–40.
21 Maksimovic J, Gordon L, Oshlack A. SWAN: subset-quantile within array normalization for Illumina Infinium humanmethylation450 Beadchips. *Genome Biol* 2012;13:R44.

22 Konopka T. R-package: Umap. uniform manifold approximation and projection. 2020. Available: https://cran.r-project.org/web/packages/umap/umap.pdf [Accessed 25 Aug 2021].
23 Kuhn M. Caret: classification and regression training package. R package version: 6.0-77. 2017. Available: https://cran.microsoft.com/snapshot/2017-09-17/web/packages/caret/index.html [Accessed 25 Aug 2021].
24 Wright MN, Ziegler A. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *arXiv:150804409* 2015.
25 Robin X, Turck N, Hainard A, *et al*. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;12:77.
26 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: synthetic minority over-sampling technique. *Jair* 2002;16:321–57.
27 Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* 2010;26:2363–7.
28 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;30:1363–9.
29 Yu M, Carter KT, Makar KW, *et al*. Methylight droplet digital PCR for detection and absolute quantification of infrequently methylated Alleles. *Epigenetics* 2015;10:803–9.
30 Martin TC, Yet I, Tsai PC, *et al*. coMET: visualization of regional epigenome-wide association scan results and DNA co-methylation patterns. *BMC Bioinformatics* 2015;16:131.
31 Ding W, Chen J, Feng G, *et al*. DNMIVD: DNA methylation interactive visualization database. *Nucleic Acids Res* 2020;48:D856–62.
32 Huang H-Y, Li J, Tang Y, *et al*. Methhc 2.0: information repository of DNA methylation and gene expression in human cancer. *Nucleic Acids Res* 2021;49:D1268–75.
33 Ewels PA, Peltzer A, Fillinger S, *et al*. The NF-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol* 2020;38:276–8.
34 Di Tommaso P, Chatzou M, Floden EW, *et al*. Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;35:316–9.
35 Pedersen BS, Eyring K, De S, *et al*. Fast and accurate alignment of long Bisulfite-Seq reads. *arXiv:14011129* 2014.
36 Xie C, Leung Y-K, Chen A, *et al*. Differential methylation values in differential methylation analysis. *Bioinformatics* 2019;35:1094–7.
37 Kruppa J, Sieg M, Richter G, *et al*. Estimands in epigenome-wide association studies. *Clin Epigenet* 2021;13:98.
38 Rizzato C, Campa D, Giese N, *et al*. Pancreatic cancer susceptibility loci and their role in survival. *PLoS One* 2011;6:e27921.
39 Thomas H. Regenerative medicine: bioengineering the common bile duct. *Nat Rev Gastroenterol Hepatol* 2017;14:504–5.
40 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;125:913–6.
41 Zhang C, Al-Shaheri FN, Alhamdani MSS, *et al*. Blood-based diagnosis and risk stratification of patients with pancreatic intraductal papillary mucinous neoplasm (IPMN). *Clin Cancer Res* 2023;29:1535–45.
42 Maisonneuve P, Lowenfels AB. Chronic pancreatitis and pancreatic cancer. *Dig Dis* 2002;20:32–7.
43 Malta TM, Sokolov A, Gentles AJ, *et al*. Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell* 2018;173:338–54.
44 Kwon M-S, Kim Y, Lee S, *et al*. Integrative analysis of multi-omics data for identifying multi-markers for diagnosing pancreatic cancer. *BMC Genomics* 2015;16 Suppl 9:S4.
45 Zhang Y-W, Zheng Y, Wang J-Z, *et al*. Integrated analysis of DNA-methylation and mRNA expression profiling reveals candidate genes associated with cisplatin resistance in non-small cell lung cancer. *Epigenetics* 2014;9:896–909.
46 Koh HWL, Fermin D, Vogel C, *et al*. iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. *NPJ Syst Biol Appl* 2019;5:22.
47 Argelaguet R, Velten B, Arnol D, *et al*. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data SETS. *Mol Syst Biol* 2018;14:e8124.
48 Mishra NK, Guda C. Genome-wide DNA-methylation analysis reveals molecular subtypes of pancreatic cancer. *Oncotarget* 2017;8:28990–9012.
49 Castagna M, Takai Y, Kaibuchi K, *et al*. Direct activation of calcium-activated, phospholipid-dependent protein kinase by tumor-promoting Phorbol esters. *J Biol Chem* 1982;257:7847–51.
50 Antal CE, Hudson AM, Kang E, *et al*. Cancer-associated protein kinase C mutations reveal kinase's role as tumor suppressor. *Cell* 2015;160:489–502.
51 Newton AC. Protein kinase C: structure, function, and regulation. *J Biol Chem* 1995;270:28495–8.
52 Kisiel JB, Raimondo M, Taylor WR, *et al*. New DNA-methylation markers for pancreatic cancer: discovery, tissue validation, and pilot testing in pancreatic juice. *Clin Cancer Res* 2015;21:4473–81.
53 BLUEPRINT consortium. Quantitative comparison of DNA-methylation assays for biomarker development and clinical applications. *Nat Biotechnol* 2016;34:726–37.
54 Du P, Zhang X, Huang C-C, *et al*. Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 2010;11:587.