Cancer Cell International

RESEARCH

Open Access

Characterization of G2/M checkpoint classifier for personalized treatment in uterine corpus endometrial carcinoma

Yiming Liu[1], Yusi Wang[2], Shu Tan[1], Xiaochen Shi[1], Jinglin Wen[1], Dejia Chen[1], Yue Zhao[1], Wenjing Pan[3], Zhaoyang Jia[3], Chunru Lu[4*] and Ge Lou[1*]

Abstract

Background Uterine Corpus Endometrial Carcinoma (UCEC) is a highly heterogeneous tumor, and limitations in current diagnostic methods, along with treatment resistance in some patients, pose significant challenges for managing UCEC. The excessive activation of G2/M checkpoint genes is a crucial factor affecting malignancy prognosis and promoting treatment resistance.

Methods Gene expression profiles and clinical feature data mainly came from the TCGA-UCEC cohort. Unsupervised clustering was performed to construct G2/M checkpoint (G2MC) subtypes. The differences in biological and clinical features of different subtypes were compared through survival analysis, clinical characteristics, immune infiltration, tumor mutation burden, and drug sensitivity analysis. Ultimately, an artificial neural network (ANN) and machine learning were employed to develop the G2MC subtypes classifier.

Results We constructed a classifier based on the overall activity of the G2/M checkpoint signaling pathway to identify patients with different risks and treatment responses, and attempted to explore potential therapeutic targets. The results showed that two G2MC subtypes have completely different G2/M checkpoint-related gene expression profiles. Compared with the subtype C2, the subtype C1 exhibited higher G2MC scores and was associated with faster disease progression, higher clinical staging, poorer pathological types, and lower therapy responsiveness of cisplatin, radiotherapy and immunotherapy. Experiments targeting the feature gene KIF23 revealed its crucial role in reducing HEC-1A sensitivity to cisplatin and radiotherapy.

Conclusion In summary, our study developed a classifier for identifying G2MC subtypes, and this finding holds promise for advancing precision treatment strategies for UCEC.

Keywords Uterine Corpus Endometrial Carcinoma, G2/M Checkpoint, Precision Treatment, Drug Resistance, G2MCS

*Correspondence:
Chunru Lu
64579401@qq.com
Ge Lou
louge@ems.hrbmu.edu.cn
Full list of author information is available at the end of the article

Introduction

Uterine corpus endometrial carcinoma (UCEC) is a group of epithelial malignancies originating from the endometrium, commonly occurring in perimenopausal and postmenopausal women. UCEC stands as one of the most prevalent malignancies of the female reproductive system and has the third highest mortality rate among common gynecological tumors [1, 2]. Despite two-thirds of patients presenting with early stage disease at diagnosis and the overall five-year survival rate reaches 81%, patients in stages IVA and IVB have significantly lower five-year survival rates of 17% and 15%, respectively [3].

The poor prognosis of UCEC stems from the heterogeneity of tumors and limitations of the current clinical diagnostic system. The existing diagnostic model based on clinical stage, pathological type, and genomic characterization still has considerable prognostic heterogeneity among patients when dealing with some high-level UCEC clinical decisions, which present challenges in clinical diagnosis and treatment [4, 5]. In addition, the resistance of UCEC to radiotherapy and chemotherapy is another great therapeutic challenge [6], and existing therapies often fail to produce a complete and durable tumor response, ultimately leading to treatment resistance and tumor recurrence. Therefore, there is an urgent need for new molecular subtypes to more accurately distinguish patients with different clinical characteristics and provide personalized treatment. Since the publication of the Cancer Genome Atlas data (TCGA, https://portal.gdc.cancer.gov/) [7, 8], the molecular landscape of tumor has been continuously studied, providing a clearer and broader perspective on the biological heterogeneity of UCEC and its impact on prognosis and response to anti-tumor therapy. In recent years, it has been found that the heterogeneity of tumor microenvironment (TME) plays an important role in poor prognosis and treatment resistance. Exploring new molecular pathways from TME may bring new possibilities for the precision diagnosis and treatment of UCEC.

Cancer is a group of diseases characterized by uncontrolled cell division, which is tightly regulated by several conserved cell cycle regulation mechanisms to ensure accurate replication and division of genetic material [9]. Cell cycle checkpoints function as regulatory procedures for DNA replication, preventing the accumulation and spread of faulty genetic information during cell division. At least three cell cycle checkpoints, G1/S, S, and G2/M, are known to strictly control cell cycle progression. Due to the prevalence of P53 gene mutations resulting in G1 checkpoint defects, tumor cells rely primarily on S and G2/M checkpoints to repair DNA damage [10, 11]. As the last checkpoint before a cell enters mitosis, the main role of the G2 checkpoint is to monitor the state of the cell as it prepares to enter the M phase, ensuring that all DNA has been properly repaired and replicated before the cell divides. Chk1 is activated by ATR in the G2 phase results in CDC25A, -B, and -C phosphorylation, which prevents cyclin B/CDK1 activation, resulting in G2 phase arrest [12]. CDC25B/C inactivation caused by stress-induced p38MAPK/MMK activation is another important pathway leading to G2 phase arrest [13]. In addition, G2/M phase is also the most sensitive cell phase to a variety of anti-tumor drugs and ionizing radiation. When tumor cells are exposed to Cytotoxic drugs and ionizing radiation, G2/M cell checkpoints are inhibited, and the shutdown of this "relief" mechanism forces normal repair and leads to programmed cell death in the presence of DNA damage, which has become an effective strategy for sensitizing cancer anti-tumor therapy. Several drugs have been developed to target specific checkpoint components, with Wee1 tyrosine kinase playing a key role in G2/M checkpoint regulation of DNA damage repair. High levels of Wee1 have been observed in multiple cancer species such as breast cancer, leukemia, glioma, and melanoma [14–17]. Preclinical studies have shown that inhibiting Wee1 to disrupt G2/M checkpoint damage repair function can enhance the anti-tumor activity of radiotherapy and certain cytotoxic drugs. Several clinical studies have been conducted to investigate the clinical value of Wee1 inhibitors in combination with chemotherapy and radiation therapy [18–21]. Checkpoint kinase-1 (Chk1), an active transduction kinase at S and G2 checkpoints, is a potential target for anti-tumor drug development. Inactivation of Chk1 disrupts a key signaling pathway for G2-phase damage repair, reducing stress in tumor cells' resistance to DNA damage caused by radiotherapy and chemotherapy [22]. Studies have revealed that Chk1 inhibitors effectively enhance the efficacy of chemotherapy and radiotherapy without increasing cytotoxicity, demonstrating their potential to improve responsiveness in drug-resistant cell lines [23, 24]. It is important to explore the distribution of G2/M checkpoint-related pathway activity in UCEC patients and identify potential prognostic markers and therapeutic targets for UCEC sensitization.

In this study, we have identified five differentially expressed genes (DEGs) and constructed an artificial neural network (ANN) classifier to identify subtypes with high G2/M activity scores in UCEC, which are associated with poor prognosis and lower responsiveness to most anti-tumor therapies. Further clinical cohort analysis and cell experiments revealed that groups with high expression of DEGs, including KIF23, were associated with worse prognosis, and inhibiting the expression level of KIF23 could effectively increase the sensitivity of UCEC cells to radiotherapy and cisplatin treatment. These findings may

provide a new strategy for subtype diagnosis and treatment sensitization in UCEC.

Materials and methods

Data collection and processing

354 G2/M checkpoint-related genes (G2MCRGs) were downloaded from three G2/M checkpoint-related pathways of the Molecular Signatures Database (MsigDB, https://www.gsea-msigdb.org/gsea/msigdb): "HALLMARK_G2M_CHECKPOINT", "REACTOME_G2_M_CHECKPOINTS" and "BIOCARTA_G2_PATHWAY". The "TCGAbiolinks" R package was used to download data of the TCGA-UCEC cohort from the GDC database (https://portal.gdc.cancer.gov/). The study includes 580 cases of gene expression profiling data, comprising 545 cancer cases and 35 cancer-adjacent cases, out of 548 cancer patients, three patients lacked expression profiling data. 'Cancer-adjacent' refers to tissue located near the tumor that is not cancerous but may exhibit early changes relevant to tumor development. Additionally, the dataset features 548 cases of clinical characteristics and 507 cases of TCGA molecular subtype data. Gene expression profiling data were converted to log2 (TPM+1) format for subsequent analysis. The clinical characteristics distribution of the patients from TCGA is shown in Supplementary Table 1. Complete prognostic data of these patients were obtained from the study by Liu et al. [25]. The simple nucleotide variation (SNV) data of the TCGA-UCEC cohort were obtained directly from the GDC database, stored in "maf" format, and used to calculate the tumor mutation burden (TMB) for each sample, which was calculated using the following formula: TMB (mut/mb)=total mutation amount (including synonymous, non-synonymous, substitution, insertion, and deletion mutations)/size of target coding area. Copy number variation (CNV) data of the TCGA-UCEC cohort were downloaded from the UCSC Xena database (https://xena.ucsc.edu/).

Supplementary Tab S1 Baseline Data Sheet about the clinical characteristics of the TCGA-UCEC cohort.

Quantification of G2/M checkpoint pathway activity

The ssGSEA algorithm [26] based on the "GSEABase" R package and the "GSVA" R package calculated the G2/M checkpoint score (G2MCS) for each sample using 354 G2MCRGs as the input gene set. G2MCS was used to reflect G2/M checkpoint pathway activity.

Identification and evaluation of G2/M checkpoint-related NMF clustering

Based on the expression profiles of G2MCRGs, the Non-negative matrix factorization (NMF) algorithm [27] was used to divide the TCGA-UCEC cohort into different clusters (molecular subtypes). The NMF algorithm used the "Brunet" method with 10 iterations. The top point with the fastest cophenetic decline was used to determine the optimal number of clusters. t-Distributed Stochastic Neighbor Embedding (t-SNE) was used to downscale and visualize the distinguishability of the G2MCRGs expression profiles between different clusters. Difference analysis was used to compare the G2MCS between different subtypes. Kaplan–Meier (K-M) survival analysis was used to compare the differences in overall survival (OS), disease-specific survival (DSS), progression-free interval (PFI), and disease-free interval (DFI) between different G2MC subtypes. Gene set variation analysis (GSVA) was performed to compare the variation of metabolic pathways between different G2MC subtypes using 70 metabolism-related pathways screened from "c2.cp.kegg.v7.5.1.symbols.gmt" [28].

Assessment of cell infiltration abundance in TME

A marker gene set of 23 immune cells was downloaded from the TISIDB database (http://cis.hku.hk/TISIDB/data/download/CellReports.txt) [29] and assessed relative infiltration abundance of various immune cells for the TCGA-UCEC cohort by the ssGSEA algorithm. CIBERSORT is an expression profile-based deconvolution algorithm for assessing the infiltration abundance of 22 immune cells in the TME based on the "e1701" R package [30]. MCPcounter algorithm was used to assess the population abundance of the 10 immune and stromal cells infiltrating in the TME [31]. The "ESTIMATE" R package was used to calculate the StromalScore, ImmuneScore and ESTIMATEScore in the TME, where StromalScore and ImmuneScore characterize the stromal and immune components, respectively. ESTIMATEScore is the sum of them [32]. A list of 47 immune checkpoint genes [33] and 46 cytokines [34] was collected from previous study to compare the expression differences of them between different G2MC subtypes.

Subgroup analysis of clinical characteristics

TCGA typing is an important molecular typing for UCEC [35] and patients are categorized into four subtypes: copy-number high (CN_HIGH), copy-number low (CN_LOW), microsatellite instability hypermutated (MSI-H) and POLE ultramutated (POLE). CN_LOW type represents the majority of patients with histologic grades G1 and G2, POLE type has the best prognosis, and CN_HIGH type has the worst prognosis. POLE type has poorer sensitivity to conventional radiotherapy and chemotherapy. POLE type and MSI-H type are more responsive to immunotherapy. The subtype distribution graph demonstrated the differences in the distribution of the four TCGA subtypes between different G2MC

subtypes. The distributions of various important clinical characteristics in G2MC subtypes were compared, and visualized by pie charts, including "Grade", "Stage", "BMI", "Age" and "Histological type". In addition, differences of G2MCS between subgroups with different clinical characteristics were compared.

Treatment responsiveness analysis

Chemotherapy, radiotherapy, immunotherapy, targeted therapy and endocrine therapy are important non-surgical treatment modalities for patients with UCEC. The "oncoPredict" R package constructed ridge regression models based on the "GDSC2" dataset to predict the AUC values of patients for several common UCEC drugs, including cisplatin, docetaxel, paclitaxel, tamoxifen and temsirolimus [36]. The smaller the predicted AUC value, the more sensitive the patient is likely to be to the drug. Radiotherapy-related efficacy assessment based on the TCGA-UCEC cohort compared the response to radiotherapy in patients with different G2MC subtypes, and subtype-related K-M survival analysis were used to determine the impact of radiotherapy on OS. The TIDE algorithm from the Tumor Immune Dysfunction and Exclusion (TIDE) database (http://tide.dfci.harvard.edu/) was utilized to predict the response to immunotherapy [37]. The TIDE score and Exclusion score reflected the level of tumor immune escape, with higher scores representing less responsiveness to immunotherapy. The Cancer Immunome Atlas (TCIA) database (https://tcia.at/home) provided four Immunophenoscores (IPS) reflecting the responsiveness of patients in the TCGA cohort to different immunotherapies [38]. Higher IPS represents higher immunotherapy responsiveness.

Screening of subtype differential characteristic genes (SDCGs) by machine learning algorithms

Firstly, gene expression profiles difference analysis was performed to obtain subtype differentially expressed genes with $|\log FC| > 1$ and $p < 0.05$ as the screening threshold, where the p-value was corrected by False Discovery Rate. Subsequently, the "glmnet" R package was used to perform the Least absolute shrinkage and selection operator (LASSO) regression algorithm [39]. LASSO achieved the purpose of downscaling and feature screening by constructing a penalty function and compressing the zero regression coefficients. The "randomForest" R package was used to perform the Random Forest (RF) algorithm for feature screening. The default number of iterations for RF was 100. The RF model was considered robust enough when 500 decision trees were constructed. The genes were scored for importance based on "mean decrease in accuracy", and those with importance scores greater than 2 were filtered as SDCGs. The support vector machine-recursive feature elimination (SVM-RFE) algorithm based on the "e1071" R package and the "caret" R package is a posterior term selection algorithm for sequences based on the maximum interval principle of SVM, suitable for the screening of low-dimensional features [40]. Lastly, the SDCGs obtained from LASSO and RF are taken as intersection and entered into SVM-RFE model for final screening. After tenfold cross-validation, the genes corresponding to the model with the highest accuracy and the lowest error were selected as the final SDCGs.

Construction, evaluation and validation of ANN subtype classifier

Firstly, we compared the expression levels of the SDCGs (classifier genes) in a single sample with the median of all samples. For down-regulated genes, the value was 0 if the expression level is higher than the median, otherwise it was 1. The opposite was true for up-regulated genes. The gene expression profiles of the patients were transformed into [0,1] normalized "gene signatures". Then, we constructed an ANN subtype classifier based on the "gene signatures" using the "neuralnet" R package and visualized it using the "NeuralNetTools" R package [41]. The number of neurons in the hidden layer was set to two-thirds of the sum of the neurons numbers in the input layer and the output layer. The "pROC" R package was used to construct Receiver operating characteristic (ROC) curves and compute Area under curve (AUC) to evaluate the prediction accuracy of the ANN subtype classifier. The GSE120490 cohort was used to validate the clinical application significance of subtype classifier for treatment responsiveness prediction.

Genetic variation analysis

The "maftools" R package was used to analyze SNV data in "maf" format from the TCGA-UCEC cohort and to present the somatic mutation landscape by waterfall plots. RNA stemness scores (RNAss) data were downloaded from Pan-Cancer Atlas Hub (https://pancanatlas.xenahubs.net) to characterize the tumor stemness levels of samples [42]. The expression levels of four mismatch repair genes (MSH2, MSH6, MLH1, and PMS2) were used to characterize the microsatellite instability (MSI) of the samples. In addition, the CNV "gain" and "loss" frequencies of the classifier genes were calculated and displayed on the chromosomes.

Expression and prognosis exploration of classifier genes

Differences in mRNA expression of classifier genes between UCEC and normal endometrial tissues were compared based on the GEPIA2 database (http://

gepia2.cancer-pku.cn/#index). The effect of classifier gene expression on OS and recurrence-free progression (RFS) in UCEC patients was explored based on mRNA sequencing data from the Kaplan–Meier Plotter database (https://kmplot.com/analysis/). Prognostic data from the TCGA cohort were used to complement the association of classifier gene expression with DSS, PFI, and DFI.

Clinical sample collection

The study included 23 UCEC tissue samples collected from January 2019 to August 2024. All samples were obtained from the operating room of the Department of Gynecology, Harbin Medical University Cancer Hospital. The pathological diagnosis was confirmed independently by two pathologists using the latest FIGO staging criteria (2023 edition).All samples were stored in liquid nitrogen and neutral tissue fixative within 30 min after separation for subsequent studies. This study was approved by the Ethics Committee of Harbin Medical University. Informed consent was obtained from all patients. The clinical information of the patients is listed in Table 1.

Immunohistochemistry

Tumor tissues collected from 23 UCEC patients were fixed with neutral fixative at room temperature and embedded in paraffin. Paraffin-embedded tissue sections were dewaxed and rehydrated, and then blocked with 3% hydrogen peroxide at 25 °C for 25 min. Subsequently, the sections were first incubated with primary antibodies CDC7 (1:200, CUSABIO, China), ASPM (1:200, CUSABIO, China), CENPE (1:500, Sanying, China), KIF23 (1:200, CUSABIO, China), and DEPDC1B (1:200, CUSABIO, China) at 4 °C overnight, and then incubated with multimerized anti-rabbit IgG-HRP secondary antibodies at room temperature for 90 min. Finally, all samples were imaged at the same magnification under an optical microscope.

Cell culture and transfection

Human UCEC cell line (HEC-1A) was purchased from Wuhan Procell Life Technology (Wuhan, China). The cell line was cultured in McCoy's 5A medium, and the complete medium was supplemented with 10% fetal bovine serum and 1% penicillin–streptomycin. The cells were cultured in a constant temperature incubator at 37 ℃ and 5% CO_2. SiRNA was transfected to knock down the expression of KIF23 in UCEC cell line, and the targeted

Table 1 Baseline data sheet for the real-world clinical cohort

Characteristic	Levels	Numbers (%)	
		Subtype C1	Subtype C2
Age	< 64 years old	10 (77%)	8 (80%)
	≥ 64 years old	3 (23%)	2 (20%)
Grade	G1	1 (8%)	3 (30%)
	G2	4 (31%)	7 (70%)
	G3	8 (61%)	0 (0)
Clinical_Stage	I	4 (31%)	8 (80%)
	II	2 (15%)	0 (0)
	III	7 (54%)	2 (20%)
	IV	0 (0)	0 (0)
Bokhman_Classification	I	5 (39%)	9 (90%)
	II	8 (61%)	1 (10%)
Histological_Type	Endometrioid endometrial adenocarcinoma	5 (39%)	9 (90%)
	Clear cell carcinoma	2 (15%)	0 (0)
	Mixed serous and endometrioid	3 (23%)	1 (10%)
	Serous endometrial adenocarcinoma	2 (15%)	0 (0)
	Carcinosarcoma	1 (8%)	0 (0)

G1: well differentiated carcinoma with tumor solid growth area ≤ 5%; G2: moderately differentiated carcinoma with solid growth area accounting for 6%–50%; G3: Poorly differentiated carcinoma with solid growth area > 50%. Stage:—Stage I: Cancer is limited to the uterus.—Stage II: Cancer has spread from the uterus to the cervix.—Stage III: Cancer has extended beyond the uterus but is still within the pelvic area, possibly involving nearby lymph nodes.—Stage IV: Cancer has spread to the bladder or rectum, or to distant organs such as the liver or lungs. Bokhman classification: Type I: Estrogen-dependent, also known as endometrioid adenocarcinoma, which accounts for approximately 90% of endometrial cancer cases with a better prognosis. Type II: Non-estrogen-dependent, also known as non-endometrioid adenocarcinoma or special histologic subtypes of endometrial cancer, mainly including serous carcinoma, clear cell carcinoma, mucinous carcinoma, etc., accounting for about 10% of all endometrial cancers, with a poor prognosis

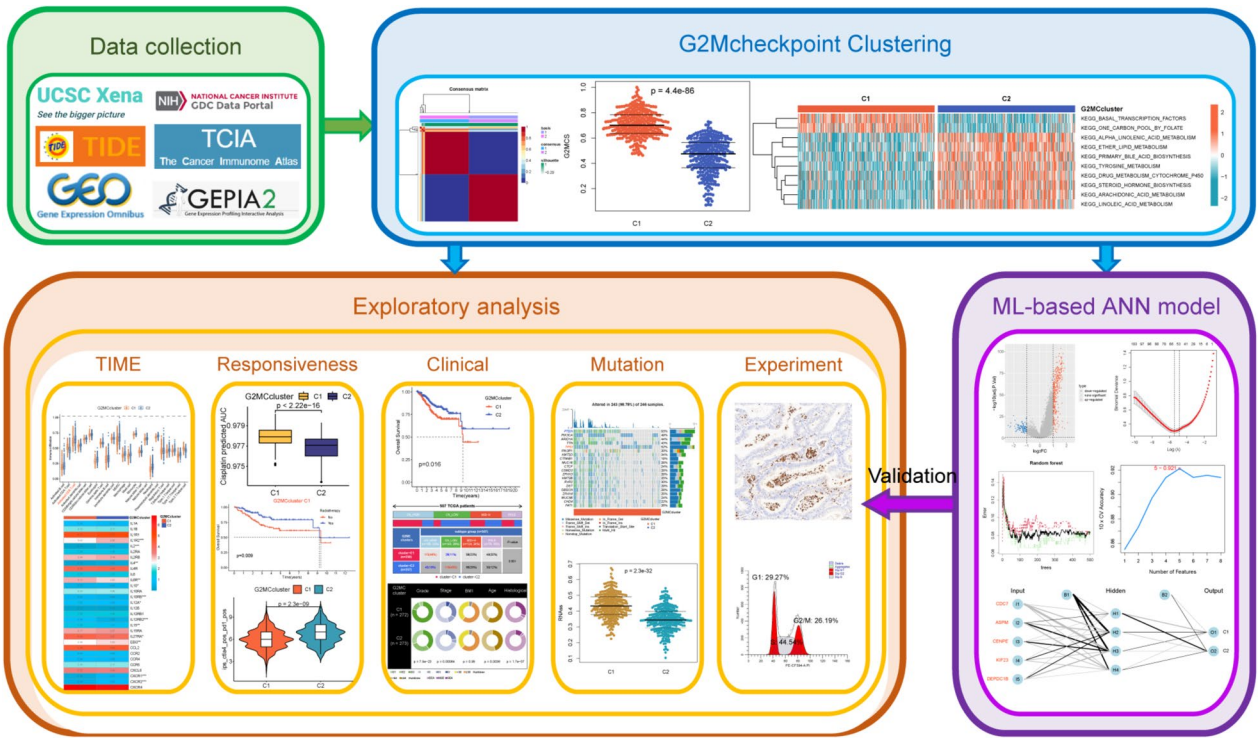


Fig. 1 The workflow diagram of this study

siRNA construct and negative control si-NC were purchased from Tsingke Biotechnology Co., Ltd. (Beijing, China). The siRNA was transfected into the cells using jetPRIME®siRNA transfection reagent (Polyplus), and the follow-up study was conducted 24 h later. The three sequences of si-KIF23 are: 5'-UGGAUUUGUACCAUUCUUCUG-3', 5'-ACUCAUUGGUUCCUUUAAGGG-3' and 5'-AAGUUUCGUUGAUAACCUGUC-3'. The sequence of si-NC is: 5'-UGGUAUUGUACCAUUCUUCUG-3'.

RNA extraction and quantitative real-time polymerase chain reaction (qRT-PCR)

Total RNA was extracted using the Accurate Steady-Pure Fast RNA Extraction Kit (AG21023, Hunan, China), and the concentration and purity of the extracted RNA were measured using NanoDrop (Thermofisher, USA). The HiScript III RT SuperMix for qPCR (#R323-01, Nanjing, China) was used for cDNA reverse transcription, and ChamQ Universal SYBR qPCR Master Mix (Vazyme#Q711) was used for real-time quantification. GAPDH was selected as the internal reference, and the $2^{-\triangle\triangle ct}$ method was applied to standardize the comparative expression levels of target genes. The primers for qRT-PCR are listed in Table S3.

Cell cycle analysis

HEC-1A cells transfected with siRNA were harvested and fixed in 75% ethanol at 4 ℃ for 24 h. After centrifugation, the cells were stained with PI for 20 min, and the fluorescence generated was measured using flow cytometry. The acquired data was further analyzed using ModFit LT software (V4.1.7) to assess cell cycle distribution.

(See figure on next page.)

Fig. 2 Identification and evaluation of G2MC Subtypes based on 353 G2MCRGs in the TCGA-UCEC cohort. **A** Differential analysis of activity of G2/M checkpoint pathway between UCEC and normal endometrial tissues based on G2MCS. **B** K-M survival analysis of G2MCS based on optimal cutoff value grouping. The ending events are OS, DSS, PFI, and DFI. **C** NMF clustering divides UCEC samples into two subtypes (k = 2) based on 353 G2MCRGs. **D** Differential analysis of activity of G2/M checkpoint pathway between the two G2MC subtypes based on G2MCS. **E** tSNE descending dimension analysis of the two G2MC subtypes. **F** K-M survival analysis reveals differences in OS, DSS, PFI, and DFI between the two G2MC subtypes. **G** GSVA between the two G2MC subtypes of the 70 metabolic pathway gene sets from KEGG. The color of the bar represents the GSVA score

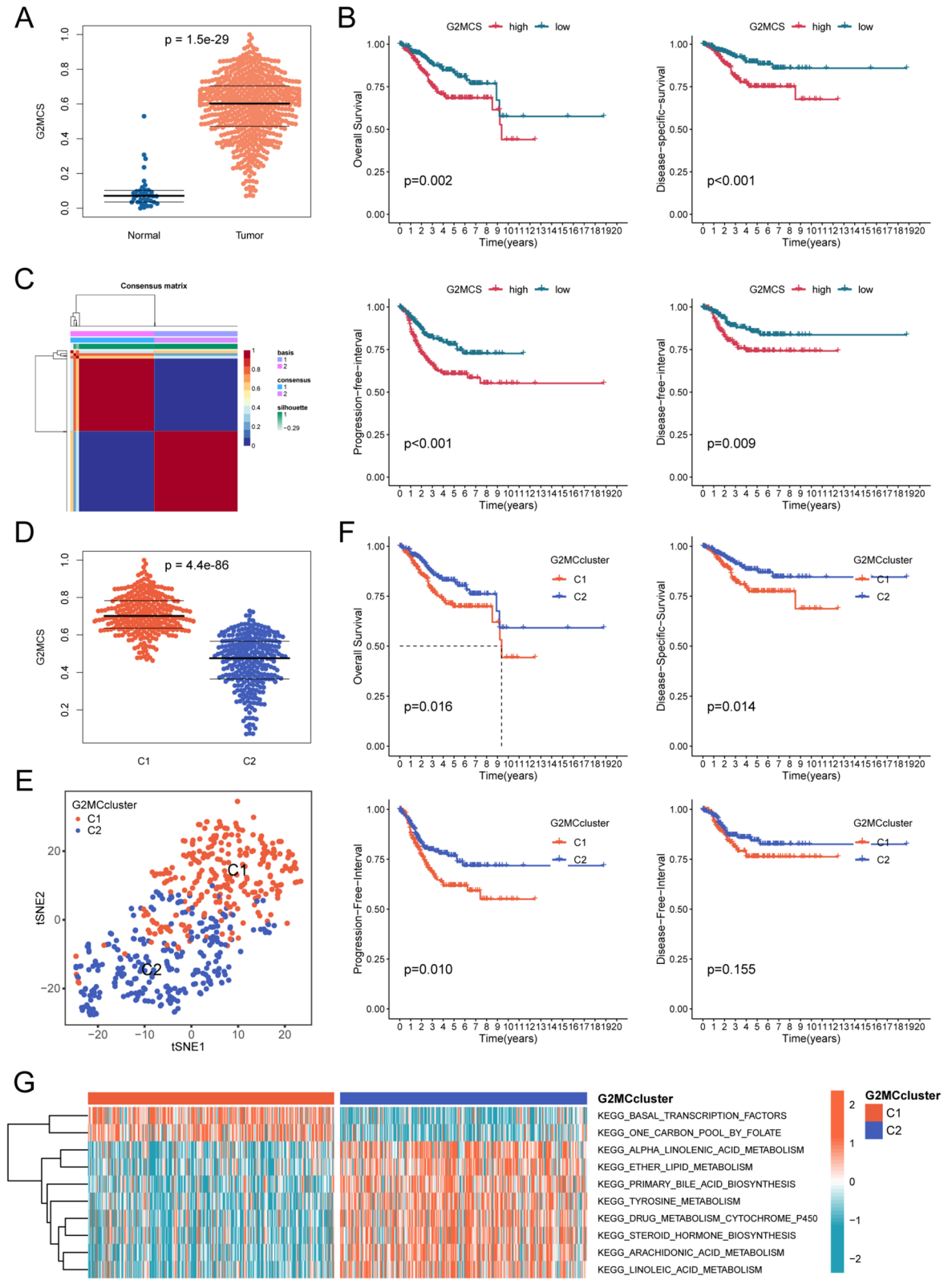


Fig. 2 (See legend on previous page.)

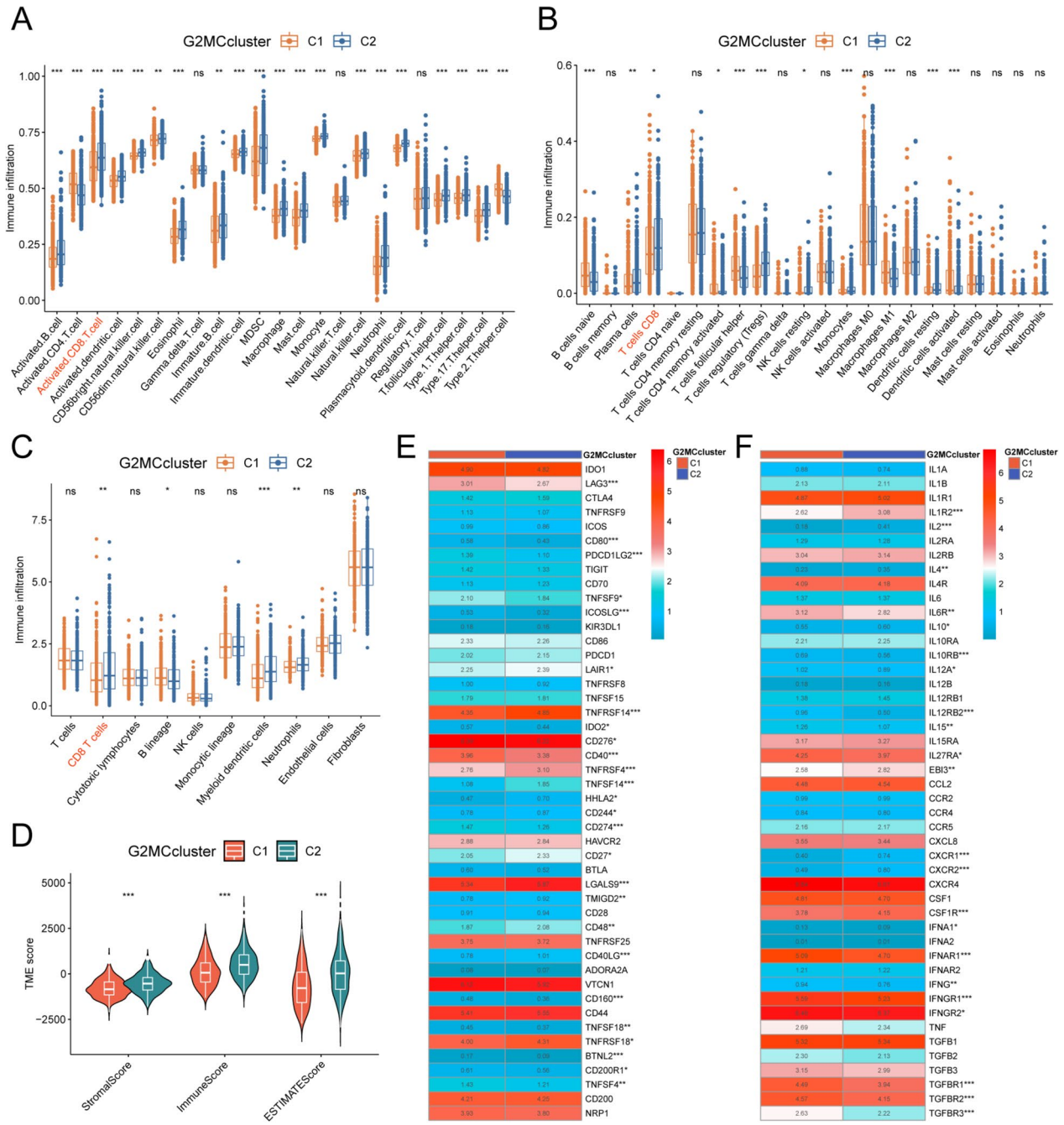


Fig. 3 TME analysis of G2MC subtypes based on the TCGA-UCEC cohort. **A**–**C** Box plots based on ssGSEA (**A**), CIBERSORT (**B**) and MCPcounter (**C**) algorithms illustrated immune cells infiltration landscapes in the TME of patients with two G2MC subtypes. Red represents CD8(+) T cells. **D** Comparison of the stromal score, immune score and ESTIMATE score between patients with subtypes C1 and C2. **E** Expression differences of 47 immune checkpoint genes between the two G2MC subtypes. **F** Expression differences of 46 cytokine genes between the two G2MC subtypes. The scale bar in the legend is the expression value of the gene. ns, no significant difference; *p < 0.05; **p < 0.01; ***p < 0.001

Anti-tumor treatment

After 24 h of cell adhesion, the treatment group was treated with cisplatin monotherapy and radiotherapy, respectively. For the chemotherapy group, cisplatin (MedChemexpress, USA) was added to the culture medium to a final concentration of 50 μmol/L for cell incubation. Cells in the radiotherapy group were irradiated using the Faxitron MultiRad 225 X-ray irradiation system (USA) at a total dose of 8 Gy, with a dose rate of 6 Gy/min. During irradiation, only the cells in the radiotherapy group were exposed, and other groups were shielded using three lead-equivalent lead plates. All cells had their culture medium replaced after six hours of treatment.

CCK-8 assay

The cells were inoculated in 96-well plates at a density of 2000 cells per well and cultured in 100 μl of complete medium. After the cells in different groups were subjected to their respective treatments for a specified period, 10 μl of CCK-8 reagent (Beyotime, Shanghai, China) was added to each well and incubated for two hours. Absorbance values at 450 nm were measured using a microplate reader (Thermofisher, USA) to determine the proliferative capacity of cells in each group.

Statistical analysis

All bioinformatics analysis in this study were performed using R (version 4.2.1), and the Perl language was used for batch processing and cleaning of data. Unless otherwise stated, difference analysis in this study was performed with the "limma" R package. The Wilcoxon test was used to compare differences between two groups, while the Kruskal–Wallis test was for three or more groups. Correlation analysis defaulted to the "pearson" method. The chi-square test was used to compare differences in rates and component ratios between groups. K-M survival analysis and log-rank tests based on the "survival" R package and "survminer" R package were used to compare survival differences between groups. A two-tailed p value of < 0.05 was considered statistically significant. Statistical analysis of all cell experiments was performed using GraphpadPrism(version9.0), and independent Student's *t*-test was used to analyze the differences between groups. All statistical tests were bilateral tests, and p values < 0.05 were considered statistically significant.

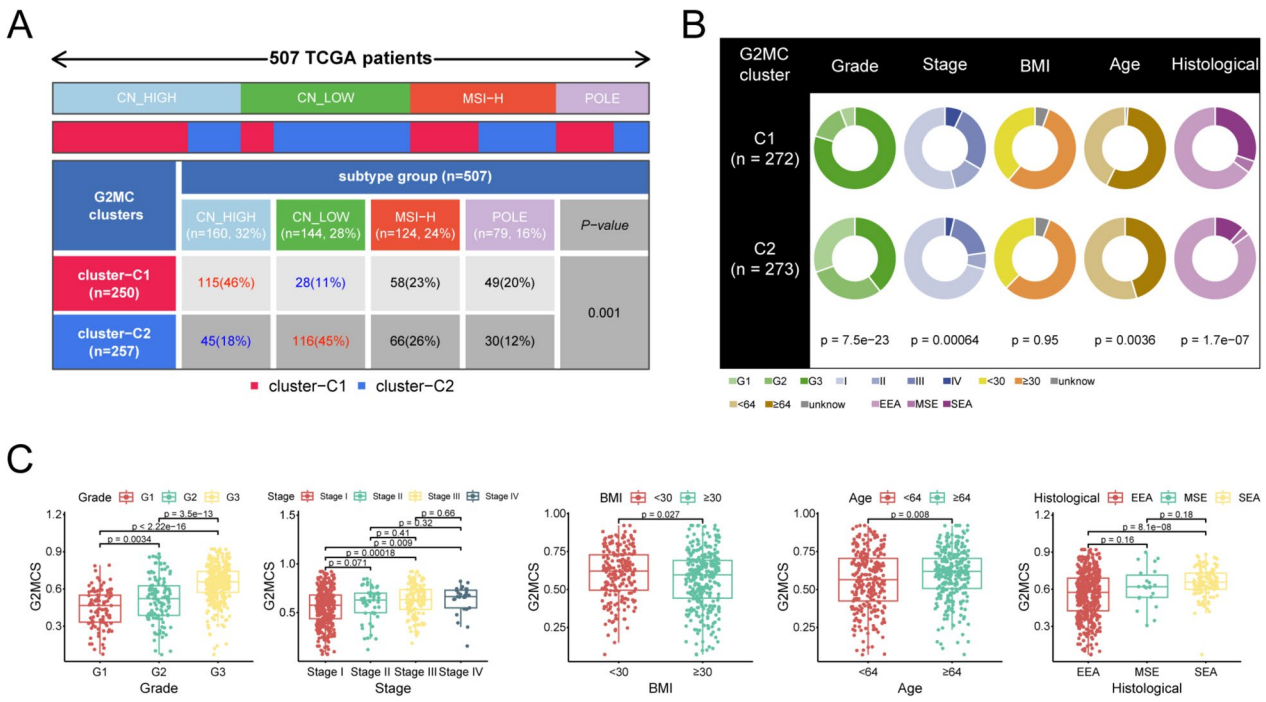


Fig. 4 Clinical characteristics analysis of G2MC subtypes based on the TCGA-UCEC cohort. **A** Distribution proportion of four TCGA molecular subtypes in two G2MC subtypes. Red represent significant high proportions, blue represents low significant proportions. **B** Pie charts combined with chi-square tests show differences in the distribution of various clinical characteristics in the two G2MC subtypes. The median age of the cohort patients is 64 (years). **C** Differences comparison of G2MCS in patients with different clinical characteristics. EEA, Endometrioid endometrial adenocarcinoma; MSE, Mixed serous and endometrioid; SEA, Serous endometrial adenocarcinoma

Results

Identification and evaluation of UCEC subtypes based on 354 G2MCRGs

The workflow of this study is shown in Fig. 1. We calculated G2MCS to reflect the activity of the G2/M checkpoint pathway for all samples of the TCGA-UCEC cohort and found that G2MCS was significantly higher in UCEC samples than in paracancer samples (Fig. 2A). Survival analysis based on optimal cutoff value grouping showed that patients in the high G2MCS group had significantly lower OS, DSS, PFI, and DFI than those in the low G2MCS group (Fig. 2B). NMF clustering based on the expression profiles of 354 G2MCRGs classified UCEC patients into two G2MC subtypes, C1 and C2 (Fig. 2C). G2MCS was significantly higher in patients with subtype C1 than in patients with subtype C2 (Fig. 2D). tSNE results showed that the G2MCRGs expression profiles of the two subtypes were distinguishable (Fig. 2E). Patients with subtype C1 had significantly lower OS, DSS, PFI, and DFI than patients with subtype C2 (Fig. 2F). GSVA results showed significant differences in the metabolic pathways between the two G2MC subtypes. The basal transcriptional activity was higher in subtype C1, and the metabolic pathways of tyrosine, α-linolenic acid, linoleic acid, and arachidonic acid were more active in subtype C2 (Fig. 2G).

TME characterization based on G2MC subtypes

The heterogeneity of TME is not only closely related to tumor immunity, but also an important factor affecting the prognosis and treatment responsiveness of patients. We evaluated TME using three algorithms, ssGSEA (Fig. 3A), CIBERSORT (Fig. 3B) and MCP-counter (Fig. 3C), which showed significant differences in immune cell infiltration abundance between the two G2MC subtypes. ssGSEA showed that the infiltration abundance of most of the immune cells in subtype C2 was significantly higher than that in subtype C1. The results of $CD8^+$ T cells, the main effector cell of tumor immunity, showed a consistent trend across the three algorithms, with higher infiltration abundance in the TME of subtype C2. We comprehensively assessed the stromal and immune components of TME by the ESTIMATE algorithm, and the StromalScore, ImmuneScore, and ESTIMATEScore were significantly higher in subtype C2 than in subtype C1 (Fig. 3D). Further, we compared the differences in the expression of 47 immune checkpoint genes and 46 cytokines in TME between the two G2MC subtypes. The results showed that the expression of immune checkpoint genes such as LAG3, CD80, PDCD1LG2, CD274, and CD40 was significantly higher in the TME of subtype C1 than that of subtype C2, whereas the opposite results were observed for TNFSF14, TNFRSF4, TNFRSF14, and TNFRSF18 (Fig. 3E). The expression of multiple inhibitory cytokines was significantly higher in the TME of subtype C1 than that of subtype C2, including IL10RB, IL12A, IL12RB2, IL27RA, TGFBR1, TGFBR2, and TGFBR3 (Fig. 3F).

Subgroup analysis of clinical characteristics associated with G2MC subtypes

TCGA subtype is an important molecular typing pattern for UCEC, guiding the precise treatment of patients. We found that the proportion of CN_HIGH subtype was the highest in patients with subtype C1 and significantly higher than subtype C2. And the proportion of CN_LOW subtype was the highest in patients with subtype C2 and significantly higher than subtype C1 (Fig. 4A). The results of the clinical characteristics combined analysis showed a predominance of G3 in subtype C1, with a significantly lower proportion of G1 and G2 than in subtype C2. Patients with subtype C1 had a lower proportion of clinical stage I and a higher proportion of the remaining clinical stages than patients with subtype C2. Patients with subtype C1 had a lower proportion of endometrioid endometrial adenocarcinoma (EEA) and a higher proportion of serous endometrial adenocarcinoma (SEA). In addition, older patients were more often clustered in subtype C1. There was no significant difference in the proportion of obesity (BMI ≥ 30) [43] between the two G2MC subtypes (Fig. 4B). Further, we

(See figure on next page.)

Fig. 5 Therapeutic responsiveness analysis based on G2MC subtypes. **A** The oncoPredict algorithm (based on GDSC2 datasource) were used to predict and compare the AUC value between two G2MC subtypes to five therapeutic drugs in the TCGA-UCEC cohort. **B** Pearson correlation analysis of G2MCS with AUC values of 5 therapeutic drugs. **C** Proportional distribution of radiotherapy response status in the two G2MC subtypes based on TCGA-UCEC-radiotherapy cohort. **D** Differential analysis of G2MCS in patients with different radiotherapy response status. **E** K-M survival analysis reveals effect of radiotherapy or not on OS in different G2MC subtypes. **F** Proportional distribution of immunotherapy response status in the two G2MC subtypes based on the TCGA-UCEC-TIDE cohort. **G** Differential analysis of G2MCS in patients with different immunotherapy response status. **H** Differential analysis of TIDE and Exclusion scores closely associated with immune escape between the two G2MC subtypes based on TIDE algorithm. **I** Differential analysis of four IPS based on the TCIA database was used to compare responsiveness to PD1 inhibitors and CTLA4 inhibitors between different G2MC subtypes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. CR, Complete response; PR, Partial response; SD, Stable disease; PD, Progressive disease

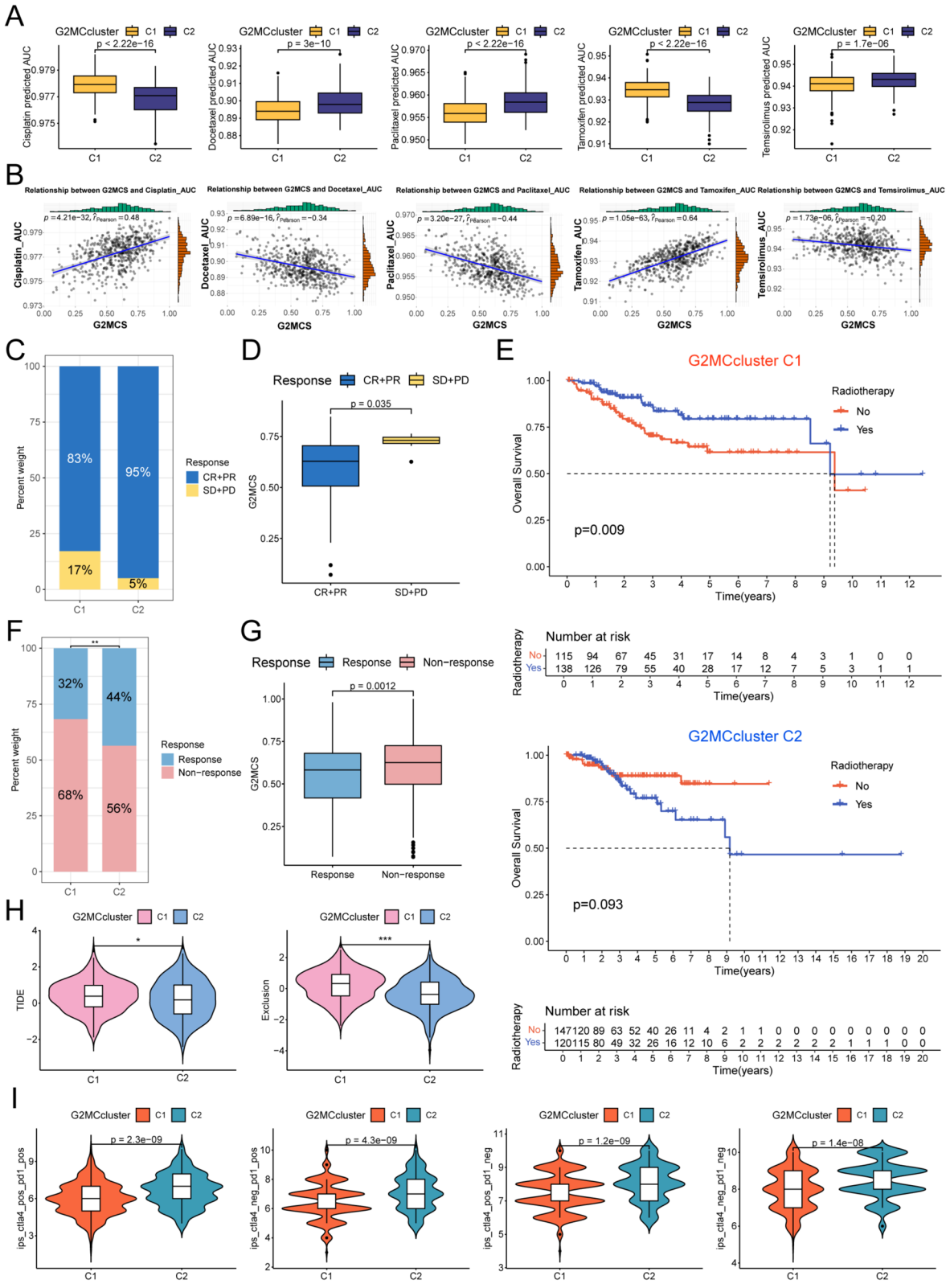


Fig. 5 (See legend on previous page.)

compared the differences in G2MCS between subgroups with different clinical characteristics and found that the higher the Grade, the higher the G2MCS. Patients with clinical stages III and IV had significantly higher G2MCS compared with stage I. Patients with SEA had significantly higher G2MCS compared with patients with EEA. In addition, older patients and non-obese patients had higher G2MCS (Fig. 4C).

Prediction of treatment responsiveness based on G2MC subtypes

Patients with UCEC are treated in a very diverse way, and systematic assessment of antitumor therapy responsiveness can help to develop precise treatment strategies based on G2MC subtypes. The results of drug sensitivity analysis showed that patients with subtype C2 had lower AUC values for cisplatin and tamoxifen and higher AUC values for docetaxel, paclitaxel, and teicoplanin compared with patients with subtype C1 (Fig. 5A). G2MCS was positively correlated with the AUC values of cisplatin and tamoxifen and negatively correlated with the AUC values of docetaxel, paclitaxel, and temsirolimus (Fig. 5B). In the TCGA-UCEC-radiotherapy cohort, patients with subtype C2 had a higher rate of objective remission (CR + PR) than subtype C1 (Fig. 5C), and patients without objective remission (SD + PD) had a significantly higher G2MCS compared with patients with objective remission (Fig. 5D). To assess the long-term benefit of radiotherapy, we used a subgroup K-M survival analysis to explore the effect of radiotherapy on OS. In patients with subtype C1, radiotherapy significantly improved patients' OS in patients with type C2, radiotherapy instead may decrease patients' OS (Fig. 5E). We assessed immune escape and predicted immunotherapy responsiveness of the TCGA-UCEC cohort based on the TIDE algorithm. Patients with subtype C1 had lower immunotherapy response rates (Fig. 5F) and higher TIDE scores and Exclusion scores compared with subtype C2 (Fig. 5G). Immunotherapy-responsive patients had significantly lower G2MCS compared with non-responders (Fig. 5H). Moreover, we validated immunotherapy responsiveness through the TCIA database. The results showed that patients with subtype C1 had significantly lower all IPS than patients with subtype C2 (Fig. 5I).

Construction, evaluation and validation of G2MC subtype classifier

In order to map the results of large-sample NMF clustering to small-sample or single-sample clinical applications, we constructed a G2MC subtype classifier based on machine learning and ANN. When the expression profiles of the classifier genes are input, the classifier can accurately classify UCEC patients into the corresponding G2MC subtypes. Firstly, we obtained 450 subtypes differentially expressed genes by difference analysis. Using subtype C2 as a control, subtype C1 had 106 down-regulated genes and 344 up-regulated genes (Fig. 6A). Subsequently, we screened and obtained 60 SDCGs by LASSO regression algorithm (Fig. 6B, C) and 42 SDCGs by RF algorithm (Fig. 6D, E), with eight overlapping genes between the two algorithms (Fig. 6F). Further screening was performed by the SVM-RFE algorithm, and the model had the highest accuracy (Fig. 6G) and the lowest error (Fig. 6H) when the number of characteristic genes were five. Five genes (CDC7, ASPM, CENPE, KIF23 and DEPDC1B) were finally screened as classifier genes. Finally, we constructed the ANN classifier with an AUC value of 0.977 for distinguishing G2MC subtypes (Fig. 6I). The specific workflow of the ANN classifier is shown in Fig. 6J.

We used the GSE120490 cohort to validate the predictive role of the G2MC subtype classifier on treatment responsiveness. Firstly, we used the subtype classifier to categorize patients of the GSE120490 cohort into subtype C1 and subtype C2, with subtype C1 having a higher G2MCS (Fig. 7A). Evaluation results based on the ssGSEA algorithm (Fig. 7B) and the MCPcounter algorithm (Fig. 7C) showed that subtype C2 had a higher infiltration level of $CD8^+$ T cells. Validation of the TIDE algorithm showed that patients with subtype C1 had a lower response rate to immunotherapy compared with

(See figure on next page.)

Fig. 6 Construction of ANN classifiers by multiple machine learning for identification of G2MC subtypes. **A** Differential analysis identified 450 DEGs of subtype C1 compared with subtype C2. **B** Ten-fold cross-validation of the coefficients of 450 DEGs in the LASSO regression model. **C** Robustness test of the LASSO regression model with varying number of DEGs revealed the highest stability when the number was 60. **D** The relationship between the number of trees and model error in the RF model. The model has the smallest error when the number of trees is 91. Green for subtype C1, red for subtype C2, black for all samples. **E** The top 20 genes were ranked by gene importance score based on mean decrease in accuracy. **F** Venn diagram illustrating the intersection of the characteristic genes of LASSO model with those of RF model. **G**, **H** Correspondence between the number of feature genes and the tenfold cross-validated accuracy (**G**) and error (**H**) in the SVM-RFE model. The model has the highest accuracy and the smallest error when the number of genes is five. **I** The ROC of ANN classifier was used to verify the predictive efficacy. **J** Schematic diagram of ANN classifier with four hidden layers. Red represents higher expression of the model gene in subtype C1 compared with subtype C2

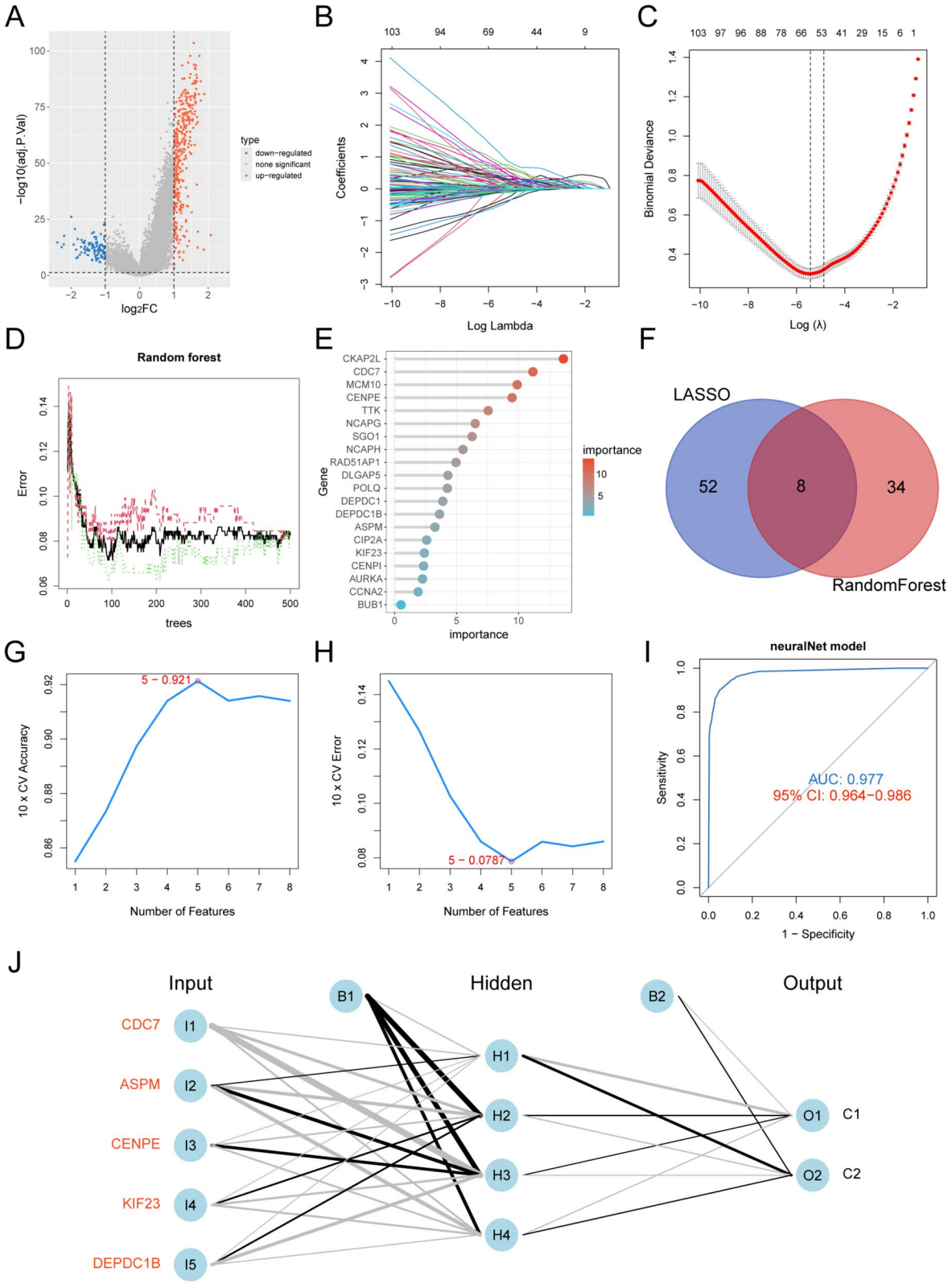


Fig. 6 (See legend on previous page.)

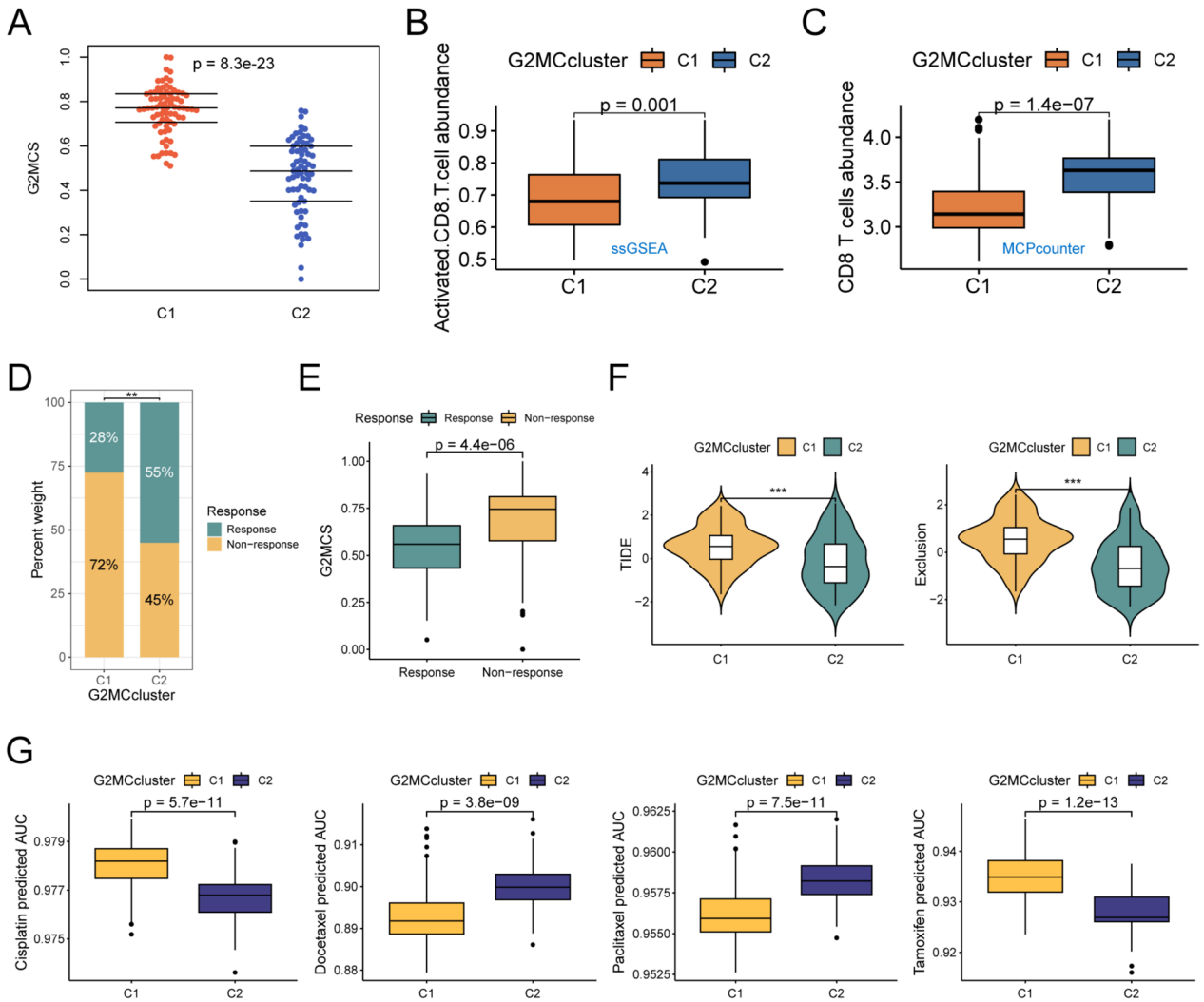


Fig. 7 Validation of ANN classifiers to identify G2MC subtypes based on GSE120490 cohort. **A** Differential analysis of G2MCS between the two G2MC subtypes. **B**, **C** Differences comparison of CD8(+) T cells' infiltration abundance between the two G2MC subtypes by ssGSEA (**B**) and MCPcounter (**C**) algorithm. **D** Proportional distribution of immunotherapy response status in the two G2MC subtypes based on the GSE120490-TIDE cohort. **E** Differential analysis of G2MCS in patients with different immunotherapy response status. **F** Differential analysis of TIDE and Exclusion scores between the two G2MC subtypes. **G** Validation of the drug sensitivity analysis showed that the AUC value of four therapeutic drugs had a trend of difference consistent with the TCGA-UCEC cohort between the two G2MC subtypes. **p < 0.01; ***p < 0.001

subtype C2 (Fig. 7D), higher TIDE scores and Exclusion scores (Fig. 7F). Immunotherapy-responsive patients had significantly lower G2MCS compared with non-responders (Fig. 7E). Validation of the drug sensitivity analysis showed that patients with subtype C2 had lower AUC values for cisplatin and tamoxifen and higher AUC values for docetaxel and paclitaxel compared with subtype C1 (Fig. 7G). This is consistent with the results of the TCGA cohort.

Genetic variation analysis associated with G2MC subtypes

The waterfall plots demonstrated the somatic mutational landscape of the two G2MC subtypes, including the 20 genes with the highest mutation frequencies in UCEC. Compared with subtype C2 (Figure S1B), patients with subtype C1 (Figure S1A) had a higher frequency of TP53 mutation (52% vs. 19%) and a lower frequency of PTEN mutation (53% vs. 74%). The results of tumor stemness

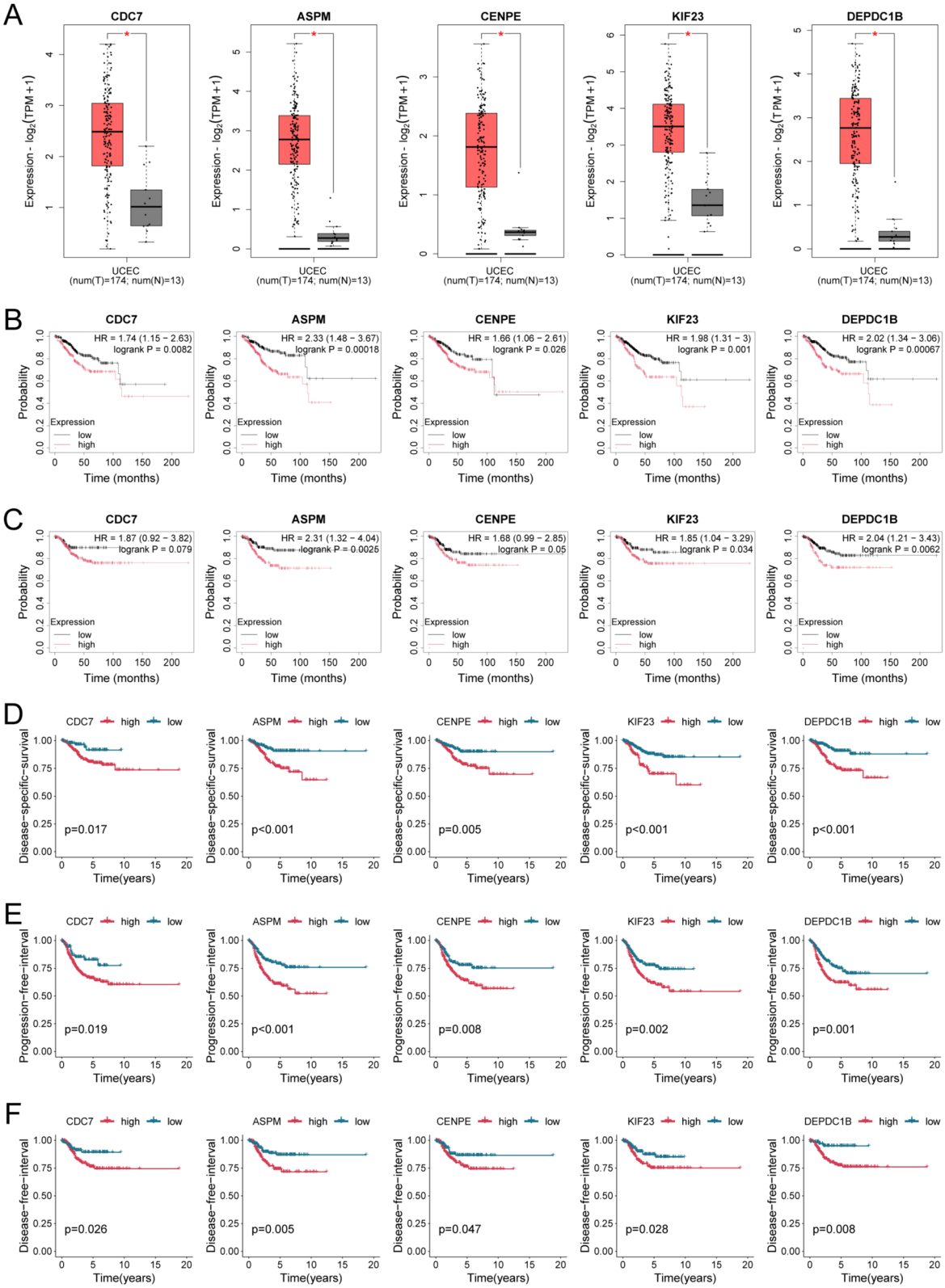


Fig. 8 Differential expression and prognostic analysis of five model genes. **A** Comparison of mRNA expression differences of five model genes between UCEC and normal endometrial tissues based on the GEPIA2 database. **B**, **C** K-M survival analysis of five model genes by optimal cutoff value grouping based on the Kaplan–Meier Plotter database. The ending events are OS (**B**) and RFS (**C**). **D**, **F** K-M survival analysis of five model genes by optimal cutoff value grouping based on the TCGA-UCEC cohort. The ending events are DSS (**D**), PFI (**E**), and DFI (**F**). *p < 0.05

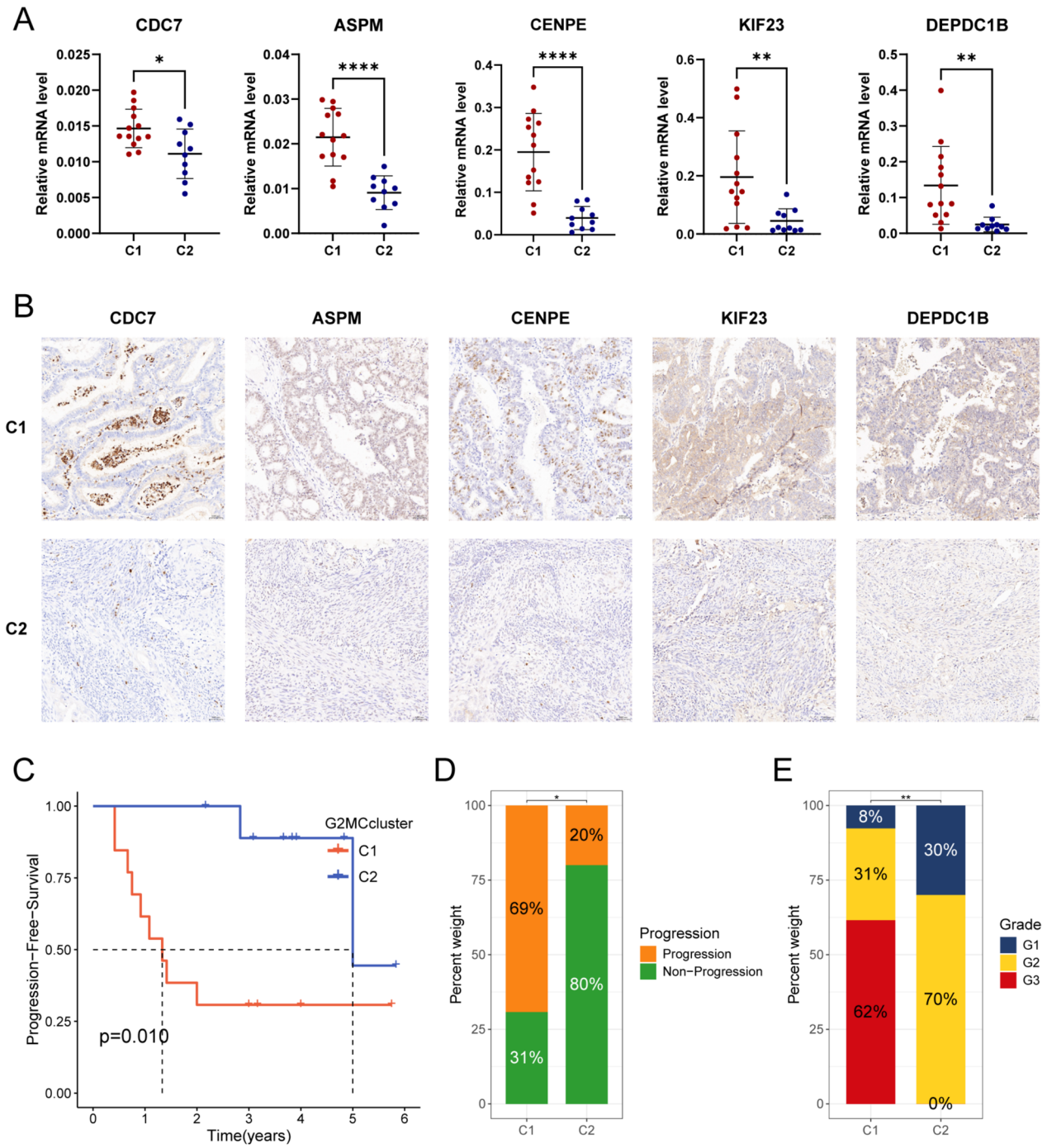


Fig. 9 Validation of real-world cohort based on the classifier. **A** The expression levels of CDC7, ASPM, CENPE, KIF23 and DEPDC1B on qRT-PCR results. **B** By inputting the mRNA levels of the 5 model genes into the classifier, the real-world UCEC cohort was identified as two G2MC subtypes. Representative immunohistochemistry images of CDC7, ASPM, CENPE, KIF23, and DEPDC1B in the two G2MC subtypes, magnification 40x. **C** K-M survival analysis of two G2MC subtypes based on optimal cutoff value grouping. The ending events are PFS. **D** The proportion of progression after treatment in the two groups of G2MC subtypes. **E** Comparing the proportion of patients with pathological grades G1 to G3 in the two G2MC subtypes in the real-world cohort. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001

analysis showed that patients with subtype C1 had significantly higher RNAss compared with subtype C2 (Figure S1C). The expression levels of all four mismatch repair genes were higher in patients with subtype C1 than subtype C2 (Figure S1D), and were positively correlated with G2MCS (Figure S1E). Five classifier genes were mutated in the UCEC samples, with ASPM having the highest mutation frequency of 15% (Figure S1F). The CNV "gain" frequency of ASPM was significantly higher than the CNV "loss" frequency, and the CNV "loss" frequency of DEPDC1B was significantly higher than the CNV "gain" frequency (Figure S1G). We then constructed the CNV landscape of all classifier genes at the chromosome (Figure S1H).

Expression and prognosis exploration of classifier genes

The results of difference analysis based on the GEPIA database showed that the mRNA expression levels of all five classifier genes were significantly higher in UCEC tissues than in normal endometrial tissues (Fig. 8A). The results of survival analysis based on the Kaplan–Meier Plotter database showed that for all five classifier genes, high expression was a prognostic risk factor for OS (Fig. 8B) and RFS (Fig. 8C, with non-significant results for CDC7 and CENPE) of patients with UCEC. We used the TCGA cohort to explore the prognostic impact of classifier gene expression from different clinical outcomes and found that for all five classifier genes, patients in the high-expression group had significantly lower DSS (Fig. 8D), PFI (Fig. 8E), and DFI (Fig. 8F) compared with the low-expression group.

Validation of the G2MC subtype classifier in the UCEC clinical cohort

Tumor tissues from a total of 23 UCEC patients were collected from the Department of Gynecology, Harbin Medical University Cancer Hospital between January 2019 and August 2024. qRT-PCR and immunohistochemistry were conducted to assess the expression levels of the characteristic genes CDC7, ASPM, CENPE, DEPDC1B, and KIF23, in order to evaluate the efficacy of the classifier. Patients were classified based on mRNA expression levels, with 13 patients classified as subtype C1 and 10 patients classified as subtype C2. In our real-world cohort, we found that the expression levels of the five characteristic molecules were significantly higher at both the nucleic acid and protein levels in the subtype C1 compared to the subtype C2 (Fig. 9A, B). Selecting disease progression as the endpoint event for survival analysis, we observed that patients in the C1 subgroup had significantly shorter PFS than those in the C2 subgroup ($p=0.01$), with a higher proportion of patients experiencing disease progression (Fig. 9C, D). Additionally, the proportion of patients with pathological grade G3 was significantly higher in the subtype C1 compared to the subtype C2 (Fig. 9E).

Knockdown of KIF23 alters the cell cycle and treatment sensitivity of HEC-1A

As one of the feature gene members selected by our classifier, KIF23 has been confirmed as a cell cycle regulator in multiple tumors. However, research on KIF23 in the context of UCEC is still blank. In order to explore the potential influence of KIF23 on UCEC-related phenotypes, we successfully knocked down the expression level of KIF23 using siRNA (Fig. 10A). The results of CCK-8 assays showed that decreased expression of KIF23 significantly decreased the proliferation capacity of HEC-1A cells after cisplatin and radiotherapy (Fig. 10B, C), suggesting that reduced expression of KIF23 can increase the sensitivity of UCEC tumors to cisplatin and radiation. Furthermore, to explore the impact of regulating KIF23 expression on the G2M phase, we conducted cell cycle analysis using flow cytometry (Fig. 10D–G). The results revealed an increase in the proportion of the G2/M phase in the si-KIF23 group. Although there was an extension of the G2/M phase, the CCK-8 assay results suggested that the knockdown of KIF23 suppressed the activity of signaling molecules, leading to a significant reduction in the efficiency of DNA damage repair. These findings indicate that KIF23 participates in the G2/M-related signaling pathway in UCEC cell division, thereby enhancing the cells' resistance to cisplatin and radiation therapy.

(See figure on next page.)

Fig. 10 Knockdown of KIF23 by specific siRNA in the HEC-1A cell line. **A** The knockdown efficiency of three siRNAs targeting KIF23 based on PCR validation. **B** Cell proliferation curves assessing by CCK-8 assay for si-KIF23 and si-NC at 24, 48, and 72 h after receiving cisplatin and radiation, respectively. **C** Column graphs of cell proliferation levels detected by CCK-8 assay for si-KIF23 and si-NC at 24, 48, and 72 h after receiving cisplatin and radiation, respectively. **D**, **F** Proportion of cell cycle for si-NC (**D**) and si-KIF23 (**F**). **E**, **G** Cell distribution and gating situation in si-NC (**E**) and si-KIF23 (**G**) in flow cytometry. Among them, ns: no significant statistical difference, *$p<0.05$, **$p<0.01$, ***$p<0.001$, ****$p<0.0001$. Data are means ± SD, with $n=3$

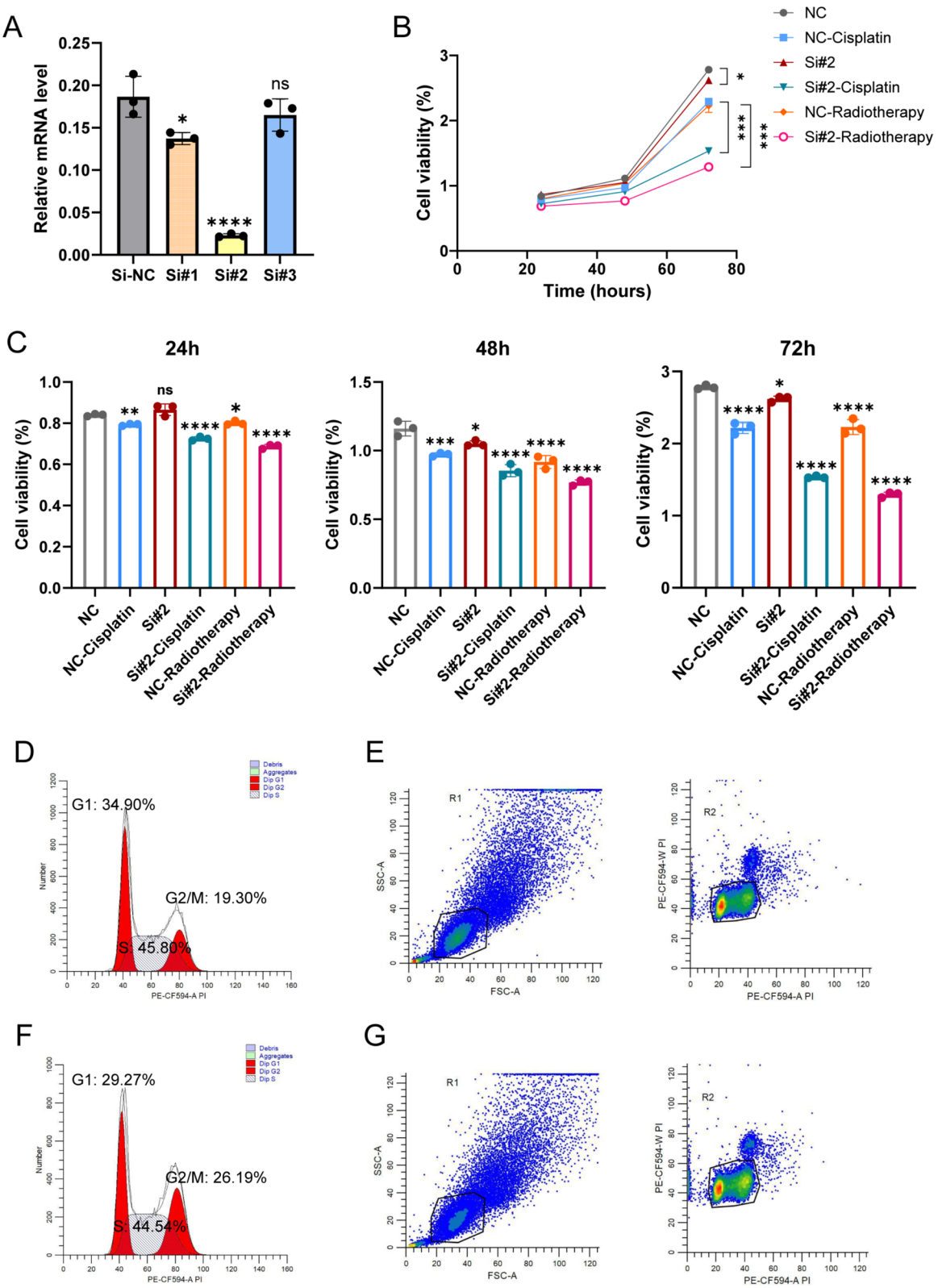


Fig. 10 (See legend on previous page.)

Discussion

UCEC is the most common gynecological malignancy in both high- and middle-income countries [5]. Despite an overall favorable prognosis, advanced-stage UCEC exhibits strong treatment resistance and a tendency for recurrence [44]. Recent exploration of the potential biological properties of UCEC has made some progress, revealing that most UCEC cases are caused by a series of somatic DNA mutations, with the most common mutations occurring in genes such as PTEN, mismatch repair genes, and TP53 [4, 5]. UCEC has been classified into four molecular subgroups based on mutation burden copy number variation, leading to the development of different treatment strategies and significantly improving the precision of treatment for this complex malignancy [45, 46]. While advancements in multi-omics technologies have greatly enhanced our understanding of UCEC heterogeneity, current molecular subtyping systems still result in different clinical outcomes for some patients with the same subtypes [47]. Therefore, the development of new molecular subtypes and the formulation of more precise anti-tumor treatment regimens are pressing clinical needs.

Tumor drug resistance results from the complex interplay of various factors, and the aberration of cell cycle regulatory mechanisms is a key driver in the development of tumor drug resistance. Overactivation of cell cycle checkpoints signaling molecules enables damaged cells to prolong their cycle arrest during replication, aiding in DNA damage repair and replication fork stability, ultimately reducing drug and radiation efficacy in killing tumor cells [9, 10]. G2/M checkpoint surveillance monitors the integrity of genetic information and inhibits damaged cells from entering mitosis, thereby maintaining efficient tumor cell division and proliferation [11]. Thus, investigating the activity of the G2/M checkpoint pathway and exploring potential signaling molecules may serve as biomarkers for predicting treatment responsiveness and prognosis in UCEC patients. At the same time, inhibiting the activity of key cell cycle molecules may prevent timely DNA damage repair in tumors, and increasing genomic stress via radiation and chemotherapy may enhance UCEC sensitivity to treatment. In this study, we utilized various machine learning methods to identify potential hub genes associated with the G2/M checkpoint pathway, with the aim of facilitating the identification of potential therapeutic targets and guiding the development of precision treatment drugs.

Based on the expression profile of 354 G2MCRGs, we scored all samples of TCGA-UCEC (G2MCS). The G2MCS of subtype C1 was significantly higher in subtype C1 than in subtype C2, indicating greater G2/M checkpoint activity. Survival and clinical feature analyses revealed that patients in subtype C1 had worse outcomes and more advanced disease than those in subtype C2. The immune infiltration of TME not only reflects the immune therapy effect of UCEC but is also closely associated with patient prognosis [33, 48, 49]. The results of tumor microenvironment analysis showed that subtype C1 had lower CD8 + T cell infiltration and higher expression of immune checkpoint genes and inhibitory cytokines, correlating with lower responsiveness to immunotherapy. In radiotherapy assessments, subtype C1 showed a lower objective response rate, but this subtype benefited from improved overall survival over time. Drug sensitivity analysis indicated that subtype C1 patients were more responsive to docetaxel, paclitaxel, and temsirolimus, while subtype C2 derived greater benefit from cisplatin and tamoxifen. These findings suggest new avenues for personalized UCEC treatment. Cisplatin, as the most classical cytotoxic drug, primarily binds to and breaks cancer cell DNA, interrupting its normal gene replication and repair, thereby inducing apoptosis of tumor cells. Radiotherapy directly or indirectly damages the genetic information of tumor cells through high-energy rays or by generating oxygen free radicals. Our results show that G2MCS is negatively correlated with sensitivity to cisplatin and radiation. Subsequent experiments confirmed that downregulating the classifier gene KIF23 increased the effectiveness of cisplatin and radiotherapy in HEC-1A cells, indicating KIF23 as a potential therapeutic target.

To enhance the applicability of G2MCS in clinical settings, we used machine learning algorithms to identify five subtype-specific feature genes and constructed a G2M classifier based on artificial neural networks (ANN), which was validated for accuracy. The predictive value of the G2M subtypes for treatment response and prognosis was confirmed in both the GSE120490 dataset and our clinical cohort, consistent with TCGA results.

Finally, our study has several limitations. Further experiments are required to investigate the regulatory mechanisms of hub genes in the G2/M checkpoint pathway. Additionally, more sample data are necessary to validate the G2/M subtype classifier, as the available clinical variables for differential feature selection are limited. Due to data constraints, other clinical risk factors, such as estrogen levels and lifestyle habits, were not included; these may interact with G2/M checkpoint activity. We conducted a differential analysis and included age as a variable, but the improved classifier's ROC curve did not show significant enhancement (Figure S2). Future efforts should focus on collecting a broader range of clinical variables to enhance the classifier's predictive performance.

Conclusion

In this study, we constructed an ANN classifier based on the activity of G2/M checkpoint pathway to identify different subtypes in UCEC patients. We found significant differences in prognosis and anti-tumor treatment responsiveness between patients with different G2MC subtypes. These research findings provide new personalized treatment strategies for UCEC patients.

Abbreviations

ANN	Artificial neural network
AUC	Area under curve
Chk1	Checkpoint kinase-1
CN_HIGH	Copy-number high
CN_LOW	Copy-number low
CNV	Copy number variation
DEGs	Differentially expressed genes
DFI	Disease-free interval
DSS	Disease-specific survival
EEA	Endometrioid endometrial adenocarcinoma
G2MC	G2/M checkpoint
G2MCRGs	G2/M checkpoint-related genes
G2MCS	G2/M checkpoint score
GSVA	Gene set variation analysis
IPS	Immunophenoscores
K-M	Kaplan–Meier
LASSO	Least absolute shrinkage and selection operator
MSI	Microsatellite instability
MsigDB	Molecular Signatures Database
MSI-H	Microsatellite instability hypermutated
NMF	Non-negative matrix factorization
OS	Overall survival
PFI	Progression-free interval
qRT-PCR	Quantitative real-time polymerase chain reaction
RF	Random Forest
RFS	Recurrence-free progression
RNAss	RNA stemness scores
ROC	Receiver operating characteristic
SEA	Serous endometrial adenocarcinoma
SNV	Simple nucleotide variation
SVM-RFE	Support vector machine-recursive feature elimination
TCGA	Cancer Genome Atlas data
TCIA	The Cancer Immunome Atlas
TIDE	Tumor Immune Dysfunction and Exclusion
TMB	Tumor mutation burden
TME	Tumor microenvironment
t-SNE	T-Distributed Stochastic Neighbor Embedding
UCEC	Uterine corpus endometrial carcinoma

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12935-025-03667-4.

Supplementary materials 1.Fig. S1. Genetic variation analysis of the G2MC subtypes and five model genes.The waterfall chart showing the landscape of gene somatic cell mutations in patients with subtype C1 and subtype C2. Red represents a significant percentage increase of important mutations, blue represents a significant percentage decrease of important mutations.Differences in RNAss that symbolizes stemness of tumor cells between patients with two G2MC subtypes.Differences in expression level of four mismatch repair genes between patients with two G2MC subtypes.Correlation matrix of G2MCS and expression levels of four mismatch repair genes.The Waterfall chart showing the types and frequencies of genetic variation in five model genes.Frequencies about gain and loss of CNV in five model genes.The Chromosome localization and the CNV landscape of five model genes. *p < 0.05; ***p < 0.001.

Additional file 2: Fig. S2. Classifier construction and validation of comprehensive clinical variables.Differences in BMI between patients with two G2MC subtypes.Differences in age between patients with two G2MC subtypes.Diagram of the improved classifier with five hidden layers, where age is included as a reference variable.The ROC of improved ANN classifier was used to verify the predictive efficacy.

Additional file 3: Tab S1. Baseline Data Sheet about the clinical characteristics of the TCGA-UCEC cohort.

Additional file 4: Tab S2. Weight parameters between nodes in ANN model to identify different G2MC subtypes.

Additional file 5: Tab S3. The primer sequences of genes.

Acknowledgements

Not applicable.

Author contributions

GL and CRL designed, supervised the study, and revision and final approval of the manuscript. YML conducted the data analysis, and wrote the manuscript. YSW,ST, XCS, JLW and DJC participated in and contributed to the experiments of this study. YZ participated in manuscript revision. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (No. 82173238), Key Program of Natural Science Foundation of Heilongjiang Province(ZD2020H007).

Availability of data and materials

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article. The complete original data and code can be downloaded from https://www.jianguoyun.com/p/DYDA1aYQyIXODBjRg9QFIAA.. No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

We confirm that the research has been carried out in accordance with the world Medical Association Declaration of Helskini. All experimental protocols were approved by the Ethics Committee of Harbin Medical University Cancer Hospital (Harbin, China) and informed consent was obtained from all patients. The Ethical permission in this study is also included in the supplementary materials.

Consent for publication

Not applicable.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author details

[1]Department of Gynecology, Harbin Medical University Cancer Hospital, Harbin, China. [2]Laboratory of Medical Genetics, Harbin Medical University, Harbin, China. [3]Second Affiliated Hospital of Harbin Medical University, Harbin, China. [4]Department of Gynecology, Suihua Maternity and Health Care Hospital, Suihua, China.

Received: 18 September 2024 Accepted: 29 January 2025
Published online: 07 February 2025

References

1. Makker V, MacKay H, Ray-Coquard I, Levine DA, Westin SN, Aoki D, et al. Endometrial cancer. Nat Rev Dis Prim. 2021;7(1):88.

2. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. Cancer J Clin. 2023;73(1):17–48.
3. Miller KD, Nogueira L, Devasia T, Mariotto AB, Yabroff KR, Jemal A, et al. Cancer treatment and survivorship statistics, 2022. Cancer J Clin. 2022;72(5):409–36.
4. Cai Y, Wang B, Xu W, Liu K, Gao Y, Guo C, et al. Endometrial cancer: genetic, metabolic characteristics, therapeutic strategies and nanomedicine. Curr Med Chem. 2021;28(42):8755–81.
5. Crosbie EJ, Kitson SJ, McAlpine JN, Mukhopadhyay A, Powell ME, Singh N. Endometrial cancer. Lancet. 2022;399(10333):1412–28.
6. Mestrallet G, Brown M, Bozkus CC, Bhardwaj N. Immune escape and resistance to immunotherapy in mismatch repair deficient tumors. Front Immunol. 2023;14:1210164.
7. Talhouk A, McConechy MK, Leung S, Li-Chang HH, Kwon JS, Melnyk N, et al. A clinically applicable molecular-based classification for endometrial cancers. Br J Cancer. 2015;113(2):299–310.
8. Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. Cancer Cell. 2018;33(4):690.
9. Matthews HK, Bertoli C, de Bruin RAM. Cell cycle control in cancer. Nat Rev Mol Cell Biol. 2022;23(1):74–88.
10. Pedroza-Garcia JA, Xiang Y, De Veylder L. Cell cycle checkpoint control in response to DNA damage by environmental stresses. Plant J. 2022;109(3):490–507.
11. Engeland K. Cell cycle arrest through indirect transcriptional repression by p53: i have a DREAM. Cell Death Differ. 2018;25(1):114–32.
12. Smith HL, Southgate H, Tweddle DA, Curtin NJ. DNA damage checkpoint kinases in cancer. Expert Rev Mol Med. 2020;22: e2.
13. Reinhardt HC, Hasskamp P, Schmedding I, Morandell S, van Vugt MA, Wang X, et al. DNA damage activates a spatially distinct late cytoplasmic cell-cycle checkpoint network controlled by MK2-mediated RNA stabilization. Mol Cell. 2010;40(1):34–49.
14. Matheson CJ, Backos DS, Reigan P. Targeting WEE1 kinase in cancer. Trends Pharmacol Sci. 2016;37(10):872–81.
15. Barnaba N, LaRocque JR. Targeting cell cycle regulation via the G2-M checkpoint for synthetic lethality in melanoma. Cell Cycle. 2021;20(11):1041–51.
16. de Nonneville A, Finetti P, Birnbaum D, Mamessier E, Bertucci F. WEE1 dependency and pejorative prognostic value in triple-negative breast cancer. Adv Sci. 2021;8(17):2101030.
17. Zhang W, Li Q, Yin R. Targeting WEE1 kinase in gynecological malignancies. Drug Des Dev Ther. 2024;18:2449–60.
18. Yap TA, Ngoi N, Dumbrava EE, Karp DD, Ahnert JR, Fu S, et al. NCI10329: Phase Ib sequential trial of agents against DNA Repair (STAR) Study to investigate the sequential combination of the Poly (ADP-Ribose) Polymerase inhibitor (PARPi) olaparib (ola) and WEE1 inhibitor (WEE1i) adavosertib (ada) in patients (pts) with DNA Damage Response (DDR)-aberrant advanced tumors, enriched for BRCA1/2 mutated and CCNE1 amplified cancers. Eur J Cancer. 2022;174:7.
19. Cai SX, Ma N, Wang X, Jiang Y, Zhang H, Guo M, et al. Discovery and development of a potent and highly selective WEE1 inhibitor IMP7068. Cancer Res. 2023;83(7):ND07.
20. Schutte T, Embaby A, Steeghs N, van der Mierden S, van Driel W, Rijlaarsdam M, et al. Clinical development of WEE1 inhibitors in gynecological cancers: a systematic review. Cancer Treatment Rev. 2023;115:102531.
21. Yap TA, Miller C, Stenehjem D, Brown EJ, Carleton M, Mirza NQ. First-in-human phase 1 study of WEE1 inhibitor APR-1051 in patients with advanced solid tumors harboring cancer-associated gene alterations. Cancer Res. 2024;84(7):CT195.
22. Merry C, Fu K, Wang J, Yeh IJ, Zhang Y. Targeting the checkpoint kinase Chk1 in cancer therapy. Cell Cycle. 2010;9(2):279–83.
23. Bucher N, Britten CD. G2 checkpoint abrogation and checkpoint kinase-1 targeting in the treatment of cancer. Br J Cancer. 2008;98(3):523–8.
24. Ahmed S, Alam W, Aschner M, Alsharif KF, Albrakati A, Saso L, et al. Natural products targeting the ATR-CHK1 signaling pathway in cancer therapy. Biomed Pharmacot. 2022;155:113797.
25. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell. 2018;173(2):400-16.e11.
26. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
27. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A. 2004;101(12):4164–9.
28. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.
29. Jia Q, Wu W, Wang Y, Alexander PB, Sun C, Gong Z, et al. Local mutational diversity drives intratumoral immune heterogeneity in non-small cell lung cancer. Nat Commun. 2018;9(1):5361.
30. Bense RD, Sotiriou C, Piccart-Gebhart MJ, Haanen J, van Vugt M, de Vries EGE, et al. Relevance of tumor-infiltrating immune cell composition and functionality for disease outcome in breast cancer. J Natl Cancer Inst. 2017;109(1):jw192.
31. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016;17(1):218.
32. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.
33. Kalbasi A, Ribas A. Tumour-intrinsic resistance to immune checkpoint blockade. Nat Rev Immunol. 2020;20(1):25–39.
34. Propper DJ, Balkwill FR. Harnessing cytokines and chemokines for cancer therapy. Nat Rev Clin Oncol. 2022;19(4):237–53.
35. Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, Shen H, et al. Integrated genomic characterization of endometrial carcinoma. Nature. 2013;497(7447):67–73.
36. Maeser D, Gruener RF, Huang RS. oncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data. Brief Bioinform. 2021;22(6):bbab260.
37. Jiang P, Gu S, Pan D, Fu J, Sahu A, Hu X, et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. Nat Med. 2018;24(10):1550–8.
38. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. Cell Rep. 2017;18(1):248–62.
39. Engebretsen S, Bohlin J. Statistical predictions with glmnet. Clin. Epigenetics. 2019;11(1):123.
40. Sanz H, Valim C, Vegas E, Oller JM, Reverter F. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. BMC Bioinformat. 2018;19(1):432.
41. Beck MW. NeuralNetTools: visualization and analysis tools for neural networks. J Stat Softw. 2018;85(11):1–20.
42. Li N, Zhan X. Integrated genomic analysis of proteasome alterations across 11,057 patients with 33 cancer types: clinically relevant outcomes in framework of 3P medicine. Epma j. 2021;12(4):605–27.
43. World Health Organization. Obesity and overweight. 2020. https://www.who.int/about/accountability/results/who-results-report-2020-2021. (Accessed 10 Dec 2024)
44. Braun MM, Overbeek-Wager EA, Grumbo RJ. Diagnosis and management of endometrial cancer. Am Fam Physician. 2016;93(6):468–74.
45. Yen TT, Wang TL, Fader AN, Shih IM, Gaillard S. Molecular classification and emerging targeted therapy in endometrial cancer. Int J Gynecol Pathol. 2020;39(1):26–35.
46. Urick ME, Bell DW. Clinical actionability of molecular targets in endometrial cancer. Nat Rev Cancer. 2019;19(9):510–21.
47. Yang Y, Wu SF, Bao W. Molecular subtypes of endometrial cancer: implications for adjuvant treatment strategies. Int J Gynaecol Obstet. 2024;164(2):436–59.

48. Zhang G, Yin Z, Fang J, Wu A, Chen G, Cao K. Construction of the novel immune risk scoring system related to CD8+ T cells in uterine corpus endometrial carcinoma. Cancer Cell Int. 2023;23(1):124.
49. Zhan L, Liu X, Zhang J, Cao Y, Wei B. Immune disorder in endometrial cancer: Immunosuppressive microenvironment, mechanisms of immune evasion and immunotherapy. Oncol Lett. 2020;20(3):2075–90.

Publisher's Note

nature medicine

Article

https://doi.org/10.1038/s41591-024-02993-w

Prediction of recurrence risk in endometrial cancer with multimodal deep learning

Received: 27 November 2023

Accepted: 11 April 2024

Published online: 24 May 2024

Check for updates

Sarah Volinsky-Fremond[1], Nanda Horeweg[2], Sonali Andani[3,4,5], Jurriaan Barkey Wolf[1], Maxime W. Lafarge[4], Cor D. de Kroon[6], Gitte Ørtoft[7], Estrid Høgdall[8], Jouke Dijkstra[9], Jan J. Jobsen[10], Ludy C. H. W. Lutgens[11], Melanie E. Powell[12], Linda R. Mileshkin[13], Helen Mackay[14], Alexandra Leary[15], Dionyssios Katsaros[16], Hans W. Nijman[17], Stephanie M. de Boer[2], Remi A. Nout[18], Marco de Bruyn[17], David Church[19,20], Vincent T. H. B. M. Smit[1], Carien L. Creutzberg[2], Viktor H. Koelzer[4,21,22] & Tjalling Bosse[1,22]

Predicting distant recurrence of endometrial cancer (EC) is crucial for personalized adjuvant treatment. The current gold standard of combined pathological and molecular profiling is costly, hampering implementation. Here we developed HECTOR (histopathology-based endometrial cancer tailored outcome risk), a multimodal deep learning prognostic model using hematoxylin and eosin-stained, whole-slide images and tumor stage as input, on 2,072 patients from eight EC cohorts including the PORTEC-1/-2/-3 randomized trials. HECTOR demonstrated C-indices in internal (n = 353) and two external (n = 160 and n = 151) test sets of 0.789, 0.828 and 0.815, respectively, outperforming the current gold standard, and identified patients with markedly different outcomes (10-year distant recurrence-free probabilities of 97.0%, 77.7% and 58.1% for HECTOR low-, intermediate- and high-risk groups, respectively, by Kaplan–Meier analysis). HECTOR also predicted adjuvant chemotherapy benefit better than current methods. Morphological and genomic feature extraction identified correlates of HECTOR risk groups, some with therapeutic potential. HECTOR improves on the current gold standard and may help delivery of personalized treatment in EC.

EC is the most common gynecological malignancy in high-income countries and is increasing in incidence[1]. Although most women with localized disease are cured by surgery, 10–20% develop distant recurrence[2], which is typically incurable. Adjuvant chemotherapy can reduce this risk, at the expense of toxicity[3,4]. Thus, current guidelines recommend such adjuvant treatment based on a combination of clinicopathological risk factors (for example, histological subtype, grade, lymphovascular space invasion (LVSI), FIGO (International Federation of Gynaecology and Obstetrics) tumor stage) and, if available, the molecular classification of EC. The last identifies patients with favorable and unfavorable outcomes defined by *POLE* mutation (*POLE*mut) or p53 abnormality (p53abn), respectively, and intermediate outcomes characterized by mismatch repair deficiency (MMRd) or no specific molecular profile (NSMP)[5–8]. Recent efforts have been made to combine clinicopathological and molecular factors[9]; however, in practice, challenges remain as a result of the complexity of combining an increasing number of factors, high-interobserver variability in the assessment of histopathological factors, and costs and turnaround-times of molecular testing. In addition, histological slides contain lots of visual information, some with prognostic potential[10], that is only partly captured in the grading and tumor histotyping by pathologists.

Deep learning (DL) models, including those using digitized hematoxylin and eosin (H&E)-stained tumor slides, have shown great promise

A full list of affiliations appears at the end of the paper. e-mail: t.bosse@lumc.nl

in the prediction of molecular alterations[11–13], cell composition[14] and prognosis[15–21], outperforming standard pathologist-based assessment. This is particularly true of the latest generation of self-supervised learning and whole-slide image (WSI) prediction DL models, which use attention-based networks[22], graphs[15,19] or (vision) transformers[23,24] to provide more granular and interpretable image representation. In addition, multimodal DL models for prognosis prediction are promising to outperform unimodal approaches that solely rely on the morphological information provided by H&E WSIs[16,21]. We previously developed a DL model, image-based (im) four molecular classes in EC (im4MEC), to accurately predict the molecular EC classification from tumor H&E WSIs, and showed that image-based molecular classes predicted prognosis[11]. Others have classified EC binary recurrence[25] or used uni-/multimodal DL models to predict EC overall survival[15,16,19,21] (concordance indices (C-indices) of 0.629–0.687), but these have relied on more detailed tumor profiling, such as multiplex immunofluorescence staining[25] or the combination of H&E WSIs with genomic and/or transcriptomic data[16], neither of which is deliverable in clinical practice at present. Thus, there remains a pressing unmet need for a method that can predict EC distant recurrence from input data generated as part of routine clinical diagnostics.

In the present study, we report the development and evaluation of HECTOR (Fig. 1)—a multimodal DL model to predict distant recurrence from H&E WSI and anatomical stage for postsurgical women with EC—across eight EC cohorts including three large randomized trials[3,26–31].

Results

EC cohorts

HECTOR is a two-step DL model wherein the first step consists of self-supervised tumor image representational learning and the second of the distant recurrence prediction task (Fig. 1).

To train and validate the distant recurrence prediction task of HECTOR, we collected and curated tumor-containing, H&E-stained WSIs of the hysterectomy specimen and comprehensive clinicopathological datasets, molecular and clinical distant recurrence data for 2,072 patients with tumor stages (FIGO 2009) I–III EC across eight cohorts, including the PORTEC-1, -2 and -3 randomized trials[3,26–30] (Extended Data Fig. 1; study CONSORT diagram shown as Supplementary Figs. 1 and 2 and Supplementary Tables 1 and 2). Of these, two population-based cohorts were held out as two external test sets: patients treated at the University Medical Center Groningen[31] (UMCG; $n = 160$ patients) and the Leiden University Medical Center (LUMC; $n = 151$ patients) where the LUMC external test set also simulates a diagnostic scenario with up to three tumor blocks per patient. The remaining patients were divided randomly into a 20% held-out internal test set ($n = 353$) and 80% training set ($n = 1{,}408$) where fivefold crossvalidation was performed. The median duration of follow-up in the training set, internal test set, UMCG external test set and LUMC external test was 7.8, 8.4, 5.3 and 2.9 years, respectively, during which 246 (17.5%), 62 (17.6%), 14 (8.8%) and 24 (15.9%) patients had distant recurrence. Importantly, patients who underwent chemotherapy, predominantly the experimental treatment arm of the PORTEC-3 randomized trial ($n = 225$), were excluded from training because this treatment influences distant recurrence risk[3,4] (Extended Data Fig. 1). These PORTEC-3 patients were, however, used for downstream analysis of adjuvant chemotherapy benefit by HECTOR.

To train HECTOR's self-supervised learning step (which requires a large imaging dataset without outcome data), we enriched the training set with one additional cohort of the TCGA-UCEC[32] (The Cancer Genome Atlas Uterine Corpus Endometrial Carcinoma) as well as the WSIs that were excluded for the distant recurrence task owing to cancer metastasized at diagnosis (FIGO 2009, stage IV) or missing outcome ($n = 1{,}862$; Methods).

Altogether, including the two training steps and the downstream analyses, the present study comprised tumor data from 2,751 patients.

HECTOR design and performance

To design HECTOR and obtain the most performant DL model for prediction of distant recurrence based on the highest C-index[33], we conducted ablation studies on the fivefold crossvalidation (Supplementary Table 3). HECTOR's first step comprises a vision transformer for patch-level, self-supervised representational learning (Fig. 1a). HECTOR's second step is a multimodal, three-arm architecture to predict distant recurrence-free probabilities (Fig. 1b). The three-arm architecture fuses prognostic information from the H&E-stained WSI of the tumor-containing uterine section, the image-based molecular class as predicted by im4MEC directly from the H&E WSI[11] and the surgically assessed anatomical stage (as three-tiered based on the FIGO 2009 system, wherein stage I indicates a tumor confined in the uterus, stage II a cervical extent and stage III beyond, including vaginal, adnexal, pelvic and lymph nodes)[34]. To do this, we combined attention-based multiple instance learning with Embedding layers to map the discrete risk factors (the image-based molecular class and anatomical stage) to a higher-dimensional continuous vector space, with the importance of each factor controlled by gating-based attention[16,35]. Ablation studies (Supplementary Table 3) also included multitask learning[36], with a second training objective predicting the image-based molecular class instead of the frozen im4MEC, or replacing attention-based multiple instance learning with DL models that integrate spatial information of the patches, such as transformer[23] and attention-based graph neural network[15]. These two architectures did not outperform attention-based multiple instance learning for this task. Further details are provided in Methods and a summary of the HECTOR configuration is provided in Supplementary Tables 4 and 5.

HECTOR demonstrated a mean C-index of 0.795 (95% confidence interval (CI): 0.768–0.822) on fivefold crossvalidation. Notably, the addition of the image-based molecular class arm as predicted by im4MEC to the H&E WSI (referred to as two-arm or one-arm model, respectively) boosted performance from 0.775 (95% CI: 0.748–0.802) to 0.782 (95% CI: 0.759–0.805) with no need for extra input data. Adding the anatomical stage (as three-tiered FIGO 2009, stage I, II or III) further improved the C-index to 0.795 (95% CI: 0.768–0.822), yielding the final architecture of HECTOR (Fig. 2a). The cumulative area under the receiver operating curve (AUC)[37] and integrated Brier score[38] are reported in Supplementary Table 6. We also observed that HECTOR concentrated high attention to fewer regions while ignoring large parts of the H&E WSI compared with a model relying on the H&E WSI (Extended Data Fig. 2).

On the unseen internal test set, HECTOR obtained a C-index of 0.789 and, on the UMCG external test set, a C-index of 0.828. The performance in the LUMC external test set is depicted in 'Performance with multiple WSIs'.

To aid clinical interpretation, we first defined categorical HECTOR risk groups as quartiles of the continuous risk scores in the training set. The groups from the first two quartiles were then combined for simplification because these had very similar clinical outcomes in the training set (distant recurrence-free probabilities of 98.1% and 95.8% by Kaplan–Meier analysis, respectively; Supplementary Fig. 3) and applied on to the internal and external test sets. Second, we computed the hazard ratio (HR) of HECTOR using a Cox's proportional hazard (CPH) model with both continuous and categorical HECTOR risk scores as the independent variable and time to distant recurrence as the dependent variable.

HECTOR showed strong prognostic value as a continuous variable in the training test set (HR = 5.06; 95% CI: 4.35–5.89; $P = 9.00 \times 10^{-99}$), the internal test set (HR = 2.69; 95% CI: 2.07–3.49; $P = 1.31 \times 10^{-13}$) and the UMCG external test set (HR = 5.84; 95% CI: 3.06–11.14; $P = 8.37 \times 10^{-8}$). On the internal test set, 10-year distant recurrence-free probabilities for HECTOR low- ($n = 175$), intermediate- ($n = 82$) and high- ($n = 96$) risk groups were 97.0% (95% CI: 0.930–0.988), 77.7% (95% CI: 0.670–0.854) and 58.1% (95% CI: 0.469–0.677), respectively (log rank $P = 1.78 \times 10^{-10}$;

https://doi.org/10.1038/s41591-024-02993-w

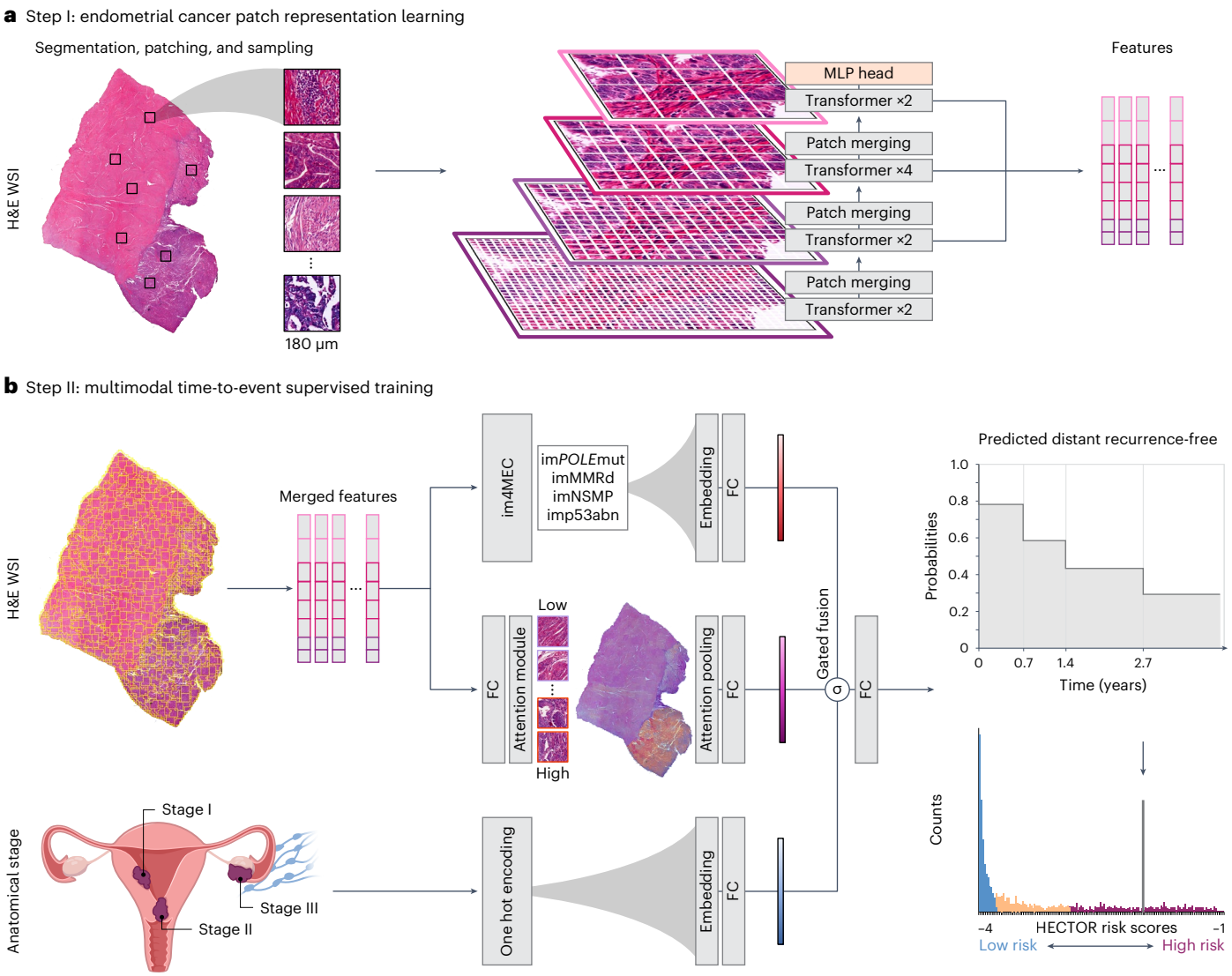


Fig. 1 | Overview of HECTOR. a, Tissue segmented from the H&E WSI of EC, subsequently patched at 180 µm. A multistage vision transformer[60] was trained using self-supervised learning by randomly sampling patches from WSIs of 1,862 patients, excluding any patients of the internal and external test sets. Patch-level features are extracted from the last eight transformer blocks. **b**, HECTOR taking the H&E WSI and the (FIGO 2009) anatomical stage I–III category as inputs. Extracted patch-level features are spatially and semantically averaged. The patch features are passed into both an attention-based multiple instance learning model and the im4MEC DL model (with all layers frozen), which predicts the molecular class from the H&E WSI as im*POLE*mut, imMMRd, imNSMP or imp53abn[11]. Both the anatomical stage category and image-based molecular class are fed through the Embedding layers. Gating-based attention is applied on the resulting three embeddings[16,35], followed by a Kronecker product for fusion. The −log(likelihood loss) was used to predict the distant recurrence-free probability function over discrete time[61]. Risk scores were defined as the integrated predicted probabilities. MLP, multilayer perceptron; FC, Fully Connected layer.

Fig. 2d). The corresponding HR for HECTOR high- and intermediate-risk groups in the internal set, using the HECTOR low-risk group as the reference, were 15.63 (95% CI: 6.58–37.13; $P = 4.81 \times 10^{-10}$) and 7.67 (95% CI: 3.06–19.22; $P = 1.37 \times 10^{-5}$), respectively. In the UMCG external test set, a similar stratification was observed with 5-year distant recurrence-free probabilities for HECTOR low- ($n = 102$), intermediate- ($n = 44$), and high- ($n = 14$) risk groups of 93.9% (95% CI: 0.859–0.974), 91.4% (95% CI: 0.756–0.972) and 19.0% (95% CI: 0.0097–0.553), respectively (log rank $P = 5.56 \times 10^{-10}$; Supplementary Fig. 4). The corresponding HR for the HECTOR intermediate group in the UMCG external test set was 2.26 (95% CI: 0.61–8.42; $P = 0.225$) and in the high-risk group was 20.42 (95% CI: 5.92–70.50; $P = 2.00 \times 10^{-6}$), respectively.

Comparison with current prognostic gold standard

We compared DL-based risk scores (that is, the one-, two-arm and HECTOR models) with the current standards for EC prognostication comprising clinicopathological risk factors and the molecular EC classification on the fivefold crossvalidation (Fig. 2a). For this, we first compared C-indices by type of input required: (1) a 'base' CPH model including variables defined by pathologists using H&E images alone (histological subtype, grade and LVSI); (2) the base model plus anatomical stage; and (3) the base model plus anatomical stage and molecular EC class. In the fivefold crossvalidation, given the H&E-based input data, the one- and two-arm model discrimination was superior to the base CPH model (C-index = 0.681; 95% CI: 0.624–0.738). HECTOR model discrimination was superior to the base CPH model plus anatomical stage which used the same inputs (C-index = 0.716; 95% CI: 0.672–0.761) and better or as good as the base CPH model plus anatomical stage and molecular EC class (C-index = 0.762; 95% CI: 0.732–0.791), which requires sequencing, immunohistochemistry (IHC) and expert pathology.

We further compared HECTOR prognostic values against current clinicopathological and molecular risk factors in multivariable

https://doi.org/10.1038/s41591-024-02993-w

a

Required input	Model	Fivefold cross-validation mean C-index [95% CI]
H&E tumor slide	Unimodal DL	0.775 [0.748–0.802]
	Two-arm DL	0.782 [0.759–0.805]
	CPH	0.681 [0.624–0.738]
H&E tumor slide + anatomic stage (2009 FIGO I-II-III)	HECTOR	0.795 [0.768–0.822]
	CPH	0.716 [0.672–0.761]
H&E tumor slide + anatomic stage (2009 FIGO I-II-III) + molecular class	CPH	0.762 [0.732–0.791]

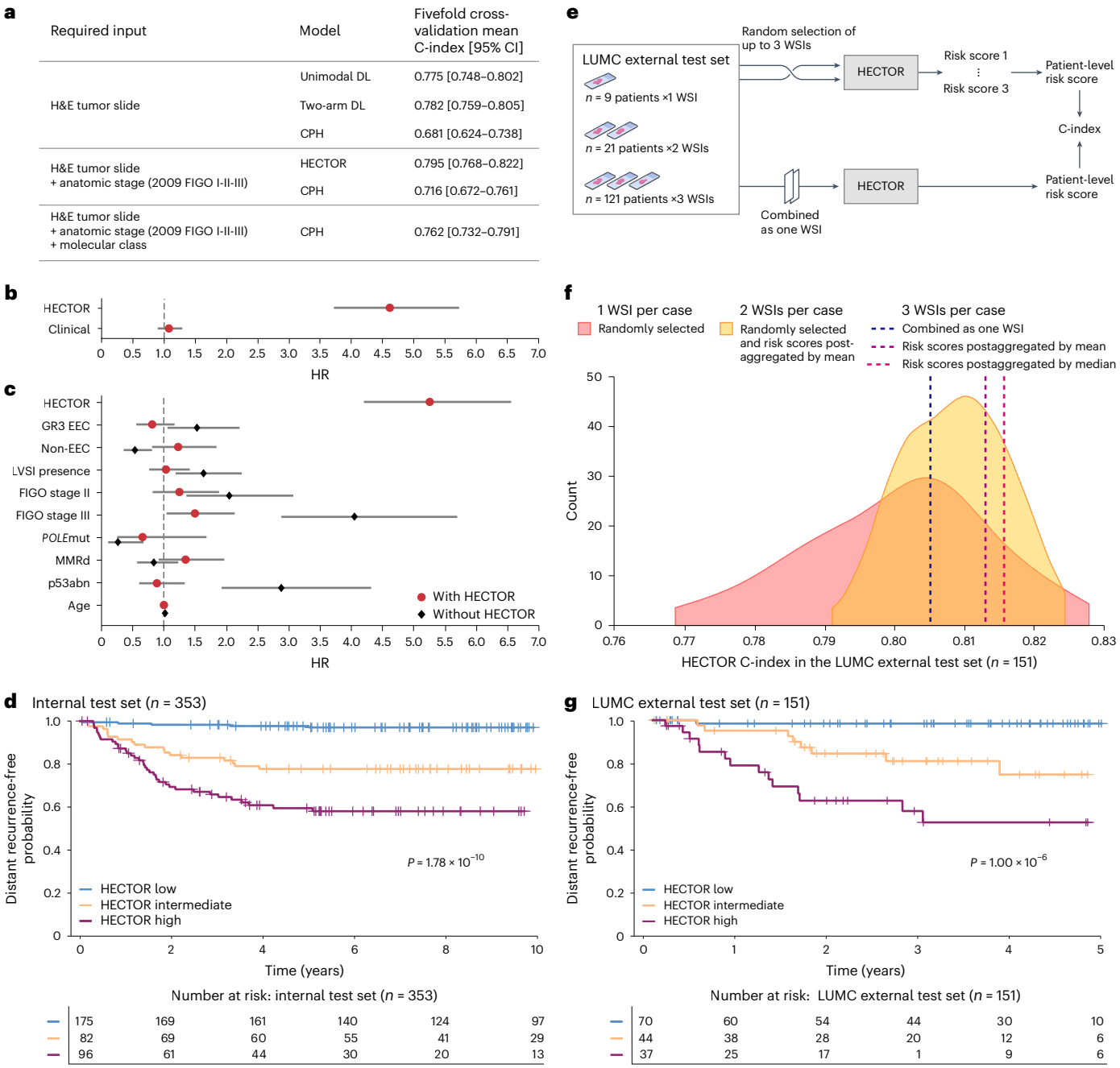


Fig. 2 | Performance of HECTOR. a, Comparison of HECTOR performance using the C-index with alternative unimodal and two-arm DL models and CPH models fitted on clinicopathological and molecular risk factors. **b**, Comparison of prognostic values between HECTOR and clinicopathological and molecular risk factors combined into one risk score in a multivariable analysis. Data are presented as the HRs and 95% CIs (n = 1,254 patients). **c**, Residual prognostic value of all established clinicopathological and molecular risk factors when using HECTOR-predicted risk scores in a multivariable analysis. Data are presented as the HRs and 95% CIs (n = 1,254 patients). **d**, The 10-year distant recurrence-free probability analysis using the Kaplan–Meier method by HECTOR risk groups in the internal test set and log rank test P value. **e**, Experiments conducted in the LUMC external test set (n = 151 patients) with the input of multiple WSIs. **f**, C-index of HECTOR in the LUMC external test set randomly using one to three WSIs for all patients and repeating the experiment 100×. **g**, The 5-year distant recurrence-free probability analysis using the Kaplan–Meier method by HECTOR risk groups when using up to three WSIs (postaggregated by median) in the LUMC external test set and log rank test P value. GR3, grade 3; EEC, endometrioid.

analysis using HECTOR continuous risk scores as the independent variable. HECTOR retained prognostic values in multivariable models in which known risk factors (histological subtype, grade, LVSI, FIGO 2009 stage I–III, age, molecular class) combined as one risk score (referred to as the CLINICAL risk score) were not prognostic (HECTOR HR = 4.62 (95% CI: 3.72–5.73; $P = 5.02 \times 10^{-44}$) versus CLINICAL HR = 1.08 (95% CI: 0.90–1.30; $P = 0.402$)) (Fig. 2b). Similar multivariable analysis, including risk factors as individual variables, showed independent prognostic value of HECTOR (HR = 5.26; 95% CI: 4.21–6.56; $P = 2.30 \times 10^{-48}$), with only FIGO 2009 stage III disease retaining statistical significance (HR = 1.50; 95% CI: 1.05–2.14; $P = 0.026$) (Fig. 2c). Other known risk factors were no longer prognostic after inclusion of the HECTOR risk score, suggesting that these factors were captured by HECTOR. For instance, the *POLE*mut and p53abn molecular classes derived from ground-truth sequencing and IHC, respectively—HR = 0.66 (95% CI: 0.26–1.69; $P = 0.384$) and HR = 0.90 (95% CI: 0.61–1.34; $P = 0.616$)—and

histological factors such as LVSI (HR: 1.05; 95% CI: 0.77–1.42, $P = 0.776$) would not be of additive prognostic value for the prediction of distant recurrence.

Given the current prognostic gold standards that would classify p53abn EC as high-risk tumors and MMRd and NSMP as intermediate-risk tumors with heterogeneous outcomes, we validated the capacity of HECTOR to refine prognosis within the MMRd, NSMP and p53abn molecular classes in the training and internal test sets. In particular, the HECTOR low-risk group also identified about 5.3% (16 out of 300) of p53abn EC cases with excellent prognosis in the entire dataset (Supplementary Fig. 5). Along these lines, we estimated the number of patients with markable different risk classification between HECTOR and the ESGO-ESTRO-ESP 2021 guidelines[5] which combine clinicopathological and molecular factors (Supplementary Fig. 6). Among all patients with intermediate- to high-risk tumors based on the guidelines (and no report of distant recurrence), 48.2% (552 cases out of 1,146) of patients were predicted to be HECTOR low risk and 16.9% (62 cases out of 366) were predicted to be HECTOR low risk among high-risk tumors only. Among all guideline-based low-to-high intermediate-risk tumors, 11.2% (131 out of 1,170) of patients were predicted to be HECTOR high risk and 4.9% (14 out of 287) when restricting to only low-risk tumors.

Performance with multiple WSIs

To evaluate the prognostic value and robustness of HECTOR in a second real-world external test set, we leveraged the fact that most cases in the LUMC cohort had multiple tumor-containing H&E WSIs derived from different tissue blocks per patient (121 of 151 cases had 3 WSIs, 21 had 2 and 9 had 1; Fig. 2e). This enabled us to validate the external performance of HECTOR in a diagnostic setting and subsequently test robustness to selection of the H&E WSI. The initial evaluation, using a HECTOR score derived from random selection of a single WSI per patient repeated 100×, demonstrated a mean C-index of 0.802 (95% CI: 0.799–0.804) for prediction of distant recurrence on the LUMC external test set (Fig. 2f).

HECTOR performance and risk stratification were slightly improved by the addition of further WSIs (taking per-patient HECTOR risk scores as either the mean or the median scores across WSIs) with C-indices of 0.810 (95% CI: 0.808–0.811) with up to 2 WSIs per patient, and 0.813 or 0.815 with up to 3 WSIs (Fig. 2f). A different method was tested wherein the WSIs were combined as one single input bag of images, yielding a C-index of 0.805. The 5-year distant recurrence-free probabilities using the median of HECTOR risk scores per patient were 98.4% (95% CI: 0.891–0.998) in HECTOR low risk ($n = 70$), 74.8% (95% CI: 0.534–0.874) in HECTOR intermediate risk ($n = 44$) and 52.6% (95% CI: 0.323–0.694) in HECTOR high risk ($n = 37$; log rank $P = 1.00 \times 10^{-6}$) (Fig. 2g and Supplementary Fig. 7). The corresponding HR (for the continuous HECTOR risk score) was 3.73 (95% CI: 2.34–5.96; $P = 3.17 \times 10^{-8}$) and (for the categorical high risk versus intermediate risk) 34.51 (95% CI: 4.52–263.39; $P = 6.37 \times 10^{-4}$) versus 15.08 (95% CI: 1.91–119.16; $P = 0.010$). Furthermore, HECTOR performance in patient stratification of the LUMC external test set extended to overall survival (5-year probabilities of 88.4% (95% CI: 0.769–0.944), 69.9% (95% CI: 0.468–0.845) and 47.0% (95% CI: 0.289–0.633) for low, intermediate and high risk, respectively; Supplementary Fig. 8).

Potential confounding by intratumoral heterogeneity also appeared to be minimal because 85 cases out of the 142 cases with more than 1 WSI had consistent HECTOR risk group predictions across the WSIs and only 3 cases with 3 WSIs had a different predicted HECTOR risk group for each WSI (Supplementary Figs. 9–12 and Supplementary Notes p16).

Association with prognostic factors and input contribution

DL prognostic models may provide information on the correlates or features that determine clinical outcome. Initial analysis of the internal test set by multiple linear regression (Fig. 3a,b) revealed that lower HECTOR risk scores were associated with established favorable risk factors of endometrioid (EEC) histological subtype, grade 1 and *POLE*mut EC, and higher HECTOR risk scores with unfavorable factors, including non-EEC histological subtypes, grade 3, FIGO stage III, LVSI, p53abn EC, estrogen receptor negativity and L1 cell adhesion molecule (L1CAM) positivity (Supplementary Tables 7–9 and Supplementary Fig. 13). MMRd EC, grade 2 and FIGO 2009 stage II were spread throughout the risk score axis and were not statistically significant.

For deeper explainability, we evaluated the impact of the H&E WSI, im4MEC and anatomical stage on the prediction, that is, whether each modality decreased (negative contribution) or increased (positive contribution) the HECTOR risk scores of developing distant recurrence. We used the normalized Integrated Gradient (IG) values for the H&E WSIs, and differences in predicted risk scores with fixed value of im4MEC or FIGO anatomical stage for the same case in the internal test set. The H&E WSIs mainly had a positive contribution with values linearly increasing alongside HECTOR risk scores (Fig. 3c and Supplementary Fig. 14). We also noted higher magnitude of contributions toward grade 3 EEC or non-EEC histological subtypes and LVSI (Fig. 3d). Both observations may indicate that unfavorable morphological features captured in H&E WSIs are a strong driver of risk score predictions. The use of image-based molecular class and FIGO 2009 stage I–III was consistent with domain expertise in EC with im*POLE*mut and imMMRd mainly decreasing and imp53abn strongly increasing the HECTOR risk scores given accurate predictions (Fig. 3e, Supplementary Table 8 and Supplementary Fig. 15) and higher anatomical stage increasing the HECTOR risk scores (Fig. 3f and Supplementary Fig. 16).

These analyses enabled us to dissect data of the six patients with distant recurrence predicted as HECTOR low risk in the internal test set (Supplementary Table 10 and Supplementary Fig. 17). Experimental tests, in which the image-based molecular class was replaced by the true molecular class, showed no effect of misclassification by im4MEC in these instances on to the HECTOR risk group. Review of the single WSI input by an expert gynecopathologist revealed that, at least in two cases, WSIs were missing unfavorable visual features that were reported in the pathology report (substantial LVSI or high-grade tumoral areas). We also noted three cases predicted as HECTOR high risk with a *POLE* mutation. Although the same experiment confirmed that the image-based molecular class had little or no effect in the HECTOR predictions of these instances, these three cases all had notably FIGO 2009, stage II or III disease (Supplementary Table 11).

Morphological correlates of outcome risk

To identify the prognostic morphological features that may have been used by HECTOR, the top 5% regions of the H&E WSIs with the highest impact on the risk scores (decreasing and increasing) were extracted and reviewed by an expert gynecopathologist in the internal test set (Fig. 4a and Supplementary Figs. 18–22). Within the HECTOR low-risk group, the morphological features decreasing the risk score were identified as smooth luminal borders, inflamed stroma and intraepithelial lymphocytes, intraepithelial neutrophils and abundant compact normal myometrium without tumor. Morphological features increasing the risk score in the HECTOR high-risk group were a ragged luminal tumor surface (also referred to as hobnailing), LVSI, solid tumor growth with marked nuclear atypia, desmoplastic stromal reaction and the presence of mitotic figures (Fig. 4a). Within the HECTOR low-risk group, we observed morphological features with positive contribution, although relatively less common, as surface changes mimicking hobnailing, retraction artifacts mimicking LVSI, loose myometrium with edema mimicking desmoplasia and solid tumor growth with scattered high-grade nuclear atypia (Extended Data Fig. 3a).

Mitotic activity, inflammatory cell density and the size of the tumor nuclei were quantified using DL-based image analysis tools (Fig. 4b and Methods). More inflammatory cells were present in the

https://doi.org/10.1038/s41591-024-02993-w

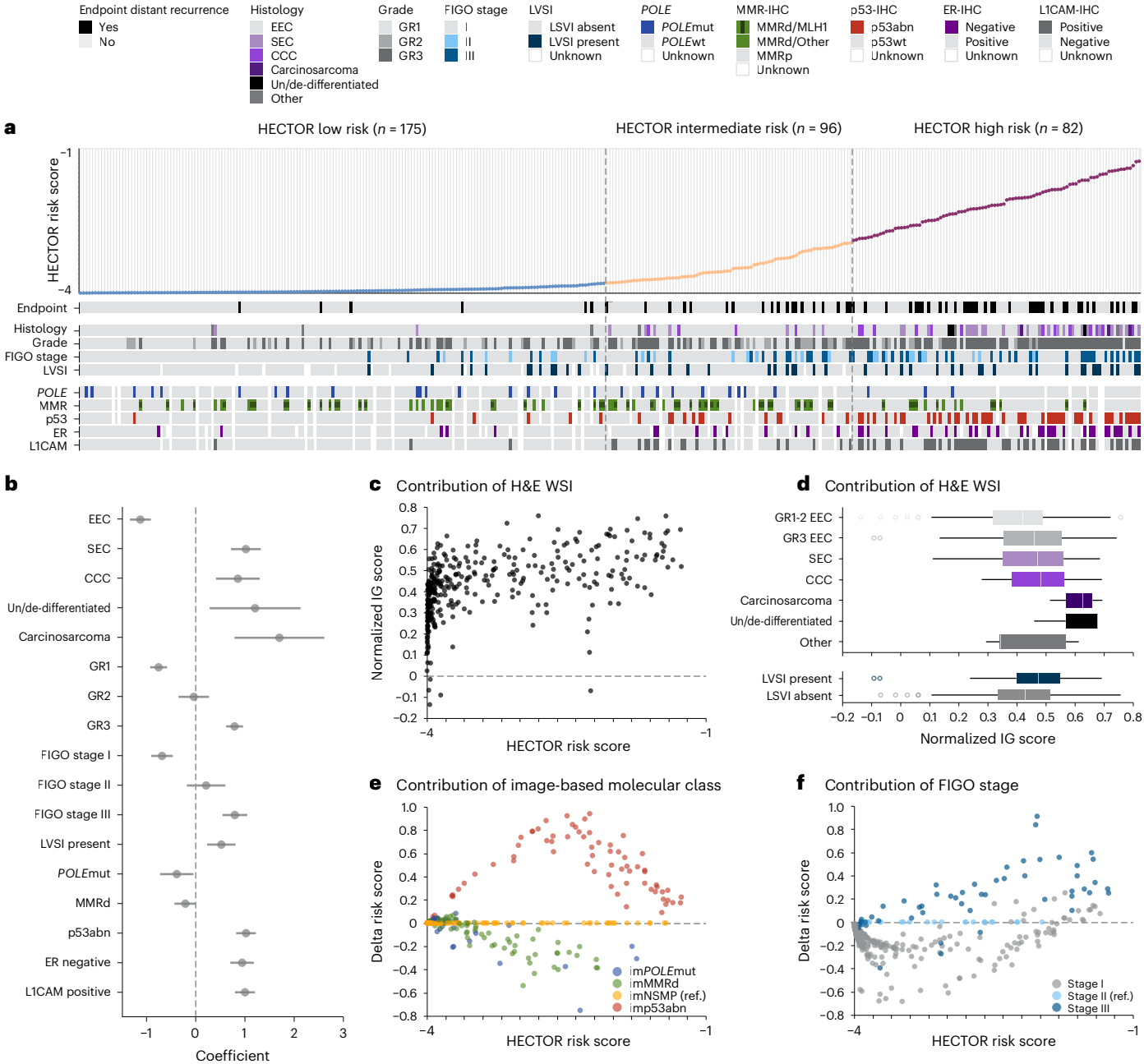


Fig. 3 | HECTOR explainability by analysis of HECTOR risk score with prognostic factors and analysis of input contribution. a, Heatmap of established prognostic factors for patients included in the internal test set ($n = 353$ patients) ordered by predicted HECTOR risk scores. Cases with multiple alterations in *POLE*, MMR and/or p53 are shown. Cases lacking any of these three specific molecular alterations are considered as NSMP according to the World Health Organization 2020 classification of female genital tumors[62]. **b**, Association of the prognostic factors and continuous HECTOR risk scores using multiple single linear regression with the HECTOR continuous risk scores as the dependent variable. Data are presented as the coefficients of the linear regression and 95% CIs ($n = 353$ patients). **c**, Analysis of the contribution to the HECTOR risk scores of the WSI modality in the internal test set ($n = 353$ patients), using the IG method[63]. The IG values of the patches were normalized and averaged by WSI. **d**, IG-normalized values of the WSIs stratified by histological subtypes (top) and presence of LVSI (bottom) in the internal test set ($n = 353$ patients). The box plots are defined by the center tick as the median value, the lower and upper parts of the box as the first (Q1) and third (Q3) quartiles, respectively, and the bounds of whiskers are (Q1 − 1.5 × IQR, Q3 + 1.5 × IQR) where IQR is the interquartile range (Q3 − Q1). Any outlier points beyond the whiskers are displayed with point marks. **e**, The contribution of the image-based molecular classes to the continuous HECTOR risk score in the internal test set, using the imNSMP as the reference (ref.) group. The difference in predicted risk score is computed between the risk score given by the image-based molecular class and the one produced by using imNSMP. **f**, The contribution of FIGO 2009 stage to the continuous HECTOR risk score in the internal test set, using FIGO 2009 stage II as the reference group. CCC, clear cell; GR1–3, grades 1–3; SEC, serous; wt, wild-type.

top 5% regions decreasing the risk scores and this effect was more pronounced in the HECTOR low-risk group ($P = 0.011$). A higher mitotic density and larger tumor nuclei were found in the top 5% regions in the HECTOR high-risk group (both $P < 0.001$). These results remained consistent across image-based molecular classes and FIGO 2009 stages I–III (Supplementary Figs. 23–25) and when filtering in regions containing tumor cells (Supplementary Fig. 26). In a quantitative spatial analysis, we computed the overlap of the top 5% regions with the tumor and invasive border areas (Extended Data Fig. 3b). The latter showed that the regions increasing the risk scores were picked out more from the

https://doi.org/10.1038/s41591-024-02993-w

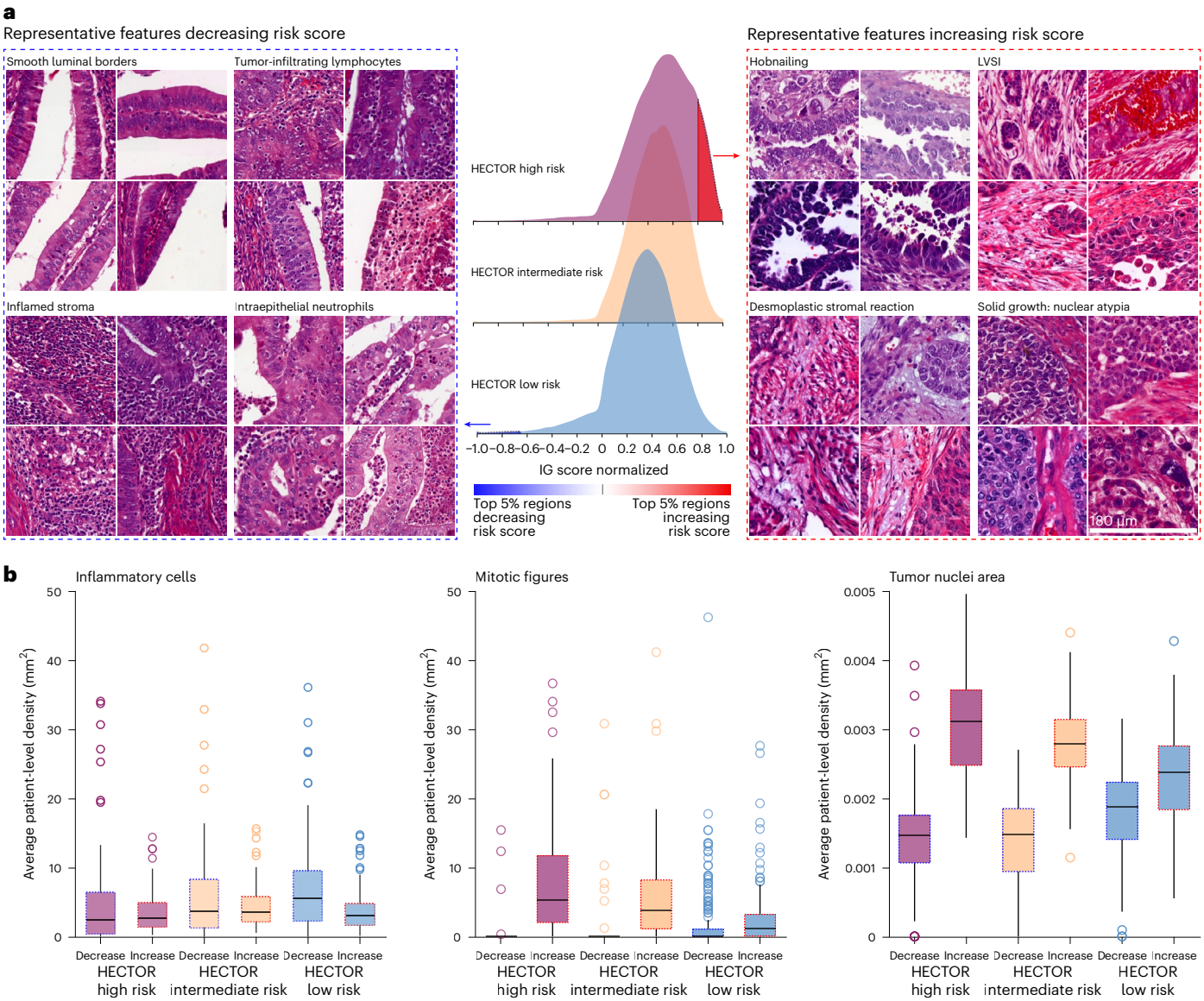


Fig. 4 | Morphological features contributing to HECTOR risk scores. a, The top 5% of the regions increasing and decreasing the risk score, from the IG method[63], extracted for qualitative review and quantitative analysis. A representative selection of four patches for each morphological subtype (each selected from a different patient) showed the increasing risk score in the HECTOR high-risk group (right). A representative selection of four patches for each morphological subtype (each selected from a different patient) showed the decreasing risk score in the HECTOR low-risk group (left). Each patch is 180 × 180 µm². **b**, Among the top 5% regions, decreasing and increasing the risk score, inflammatory cells, mitotic figures and the tumor nuclei area detected and computed with DL-based image analysis tools[14,64]. The average by patient is reported in the internal test set (n = 353). The box plots are defined by the center tick as the median value, the lower and upper parts of the box Q1 and Q3 quartiles, respectively, and the bounds of whiskers are (Q1 − 1.5 × IQR, Q3 + 1.5 × IQR). Any outlier points beyond the whiskers are displayed with point marks.

tumor than from the invasive border area. Tumor and invasive border areas contributed almost the same in regions decreasing the risk scores, notably in the HECTOR low-risk group.

Genomic alterations, immune and transcriptional signatures

For comprehensive analysis of the molecular correlates of HECTOR risk scores, we analyzed the TCGA-UCEC (n = 381 FIGO, stage I–III ECs) dataset (Fig. 5 and Supplementary Fig. 27). Coding driver mutations in *ARID1A*, *CTCF*, *CTNNB1*, *FGFR2*, *KRAS* and *PTEN* were enriched in the HECTOR low-risk group (all $P < 0.005$), whereas *PPP2R1A* and *TP53* mutations were more frequent in the HECTOR high-risk group ($P = 2.19 \times 10^{-3}$ and $P = 2.81 \times 10^{-7}$, respectively) (Fig. 5a and Supplementary Table 12). Using transcriptional data, we performed an analysis of CIBERSORT-defined lymphocyte populations using multiple linear regression (Fig. 5b). This revealed that increasing HECTOR scores were positively correlated with memory B cells ($P = 0.008$), activated dendritic cells ($P < 0.001$) and resting mast cells ($P = 0.029$), and inversely correlated with $CD8^+$ T cells ($P < 0.001$), follicular helper T cells ($P < 0.001$), regulatory T cells ($P < 0.001$) and natural killer (NK) cell activation ($P = 0.049$). Notably, these associations were independent of EC molecular class and tumor mutational burden (TMB) (Supplementary Table 13). Further transcriptomic analysis (Fig. 5c, Supplementary Fig. 27c and Supplementary Table 15) confirmed that variation in lymphocyte populations was reflected in the differential expression of canonical immune cell markers, including *CD1C*, *BTLA* and *CD40LG* (enriched in the HECTOR low-risk cases). HECTOR high-risk tumors also demonstrated upregulation of genes predictive of worse outcomes in EC, including *L1CAM* and *CLDN6*, whereas HECTOR low-risk cases showed upregulation of genes associated with hormone signaling (*C1orf64* and *OVGP1*).

https://doi.org/10.1038/s41591-024-02993-w

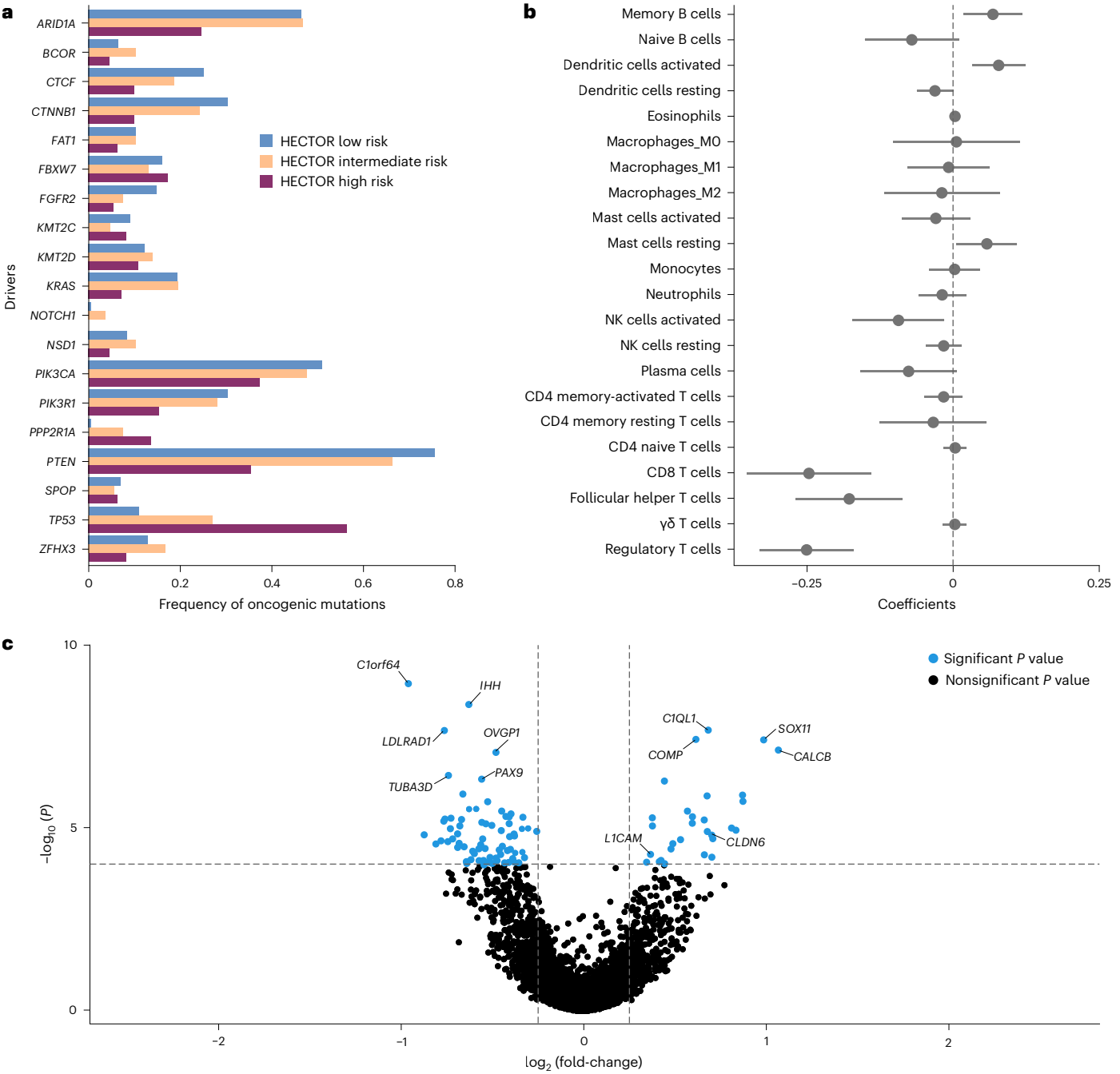


Fig. 5 | Genomic and transcriptomic correlations of HECTOR risk groups using TCGA-UCEC (*n* = 381). a, Analysis of the mutational frequency of the top 19 genes recognized as key oncogenic alterations in EC for each HECTOR risk group. **b**, Association of HECTOR risk score with the immune activation gene using multiple single linear regressions (Methods). Data are presented as the coefficients of the linear regression and 95% CIs ($n = 381$). **c**, Differential gene expression of HECTOR high-risk versus HECTOR low-risk TCGA-UCEC cases. *P* values of the likelihood ratio test were adjusted using the Benjamini–Hochberg FDR and statistical significance accepted <0.050.

Adjuvant chemotherapy response prediction by HECTOR

The investigation of whether HECTOR could predict the benefit of chemotherapy for distant recurrence risk was conducted using the PORTEC-3 randomized trial[3]. In this trial, patients with high-risk stage I–III EC were randomized to concurrent and adjuvant external beam radiotherapy with or without platinum- and paclitaxel-based chemotherapy. HECTOR risk scores were predicted on all PORTEC-3 cases for whom WSI was available ($n = 442$), which included the patients who underwent chemotherapy ($n = 225$). Importantly, these 225 cases had not been used in either training or test sets (Extended Data Fig. 4, Supplementary Table 14 and Supplementary Fig. 28). Analysis of distant recurrence-free probabilities by treatment arm and HECTOR demonstrated a statistically significant interaction between chemotherapy and HECTOR risk score as either a continuous or a categorical variable ($P_{INTERACTION} = 0.014$ and $P_{INTERACTION} = 0.064$, respectively).

We examined this in detail across HECTOR risk groups (Fig. 6a). Within HECTOR low- ($n = 92$) and HECTOR intermediate-risk ($n = 177$) groups, outcomes were similarly favorable in both treatment arms, as evidenced by similar probability of EC distant recurrence (log rank $P = 0.244$ and 0.807, respectively). In contrast, among women classified as HECTOR high risk ($n = 173$), those who received adjuvant

https://doi.org/10.1038/s41591-024-02993-w

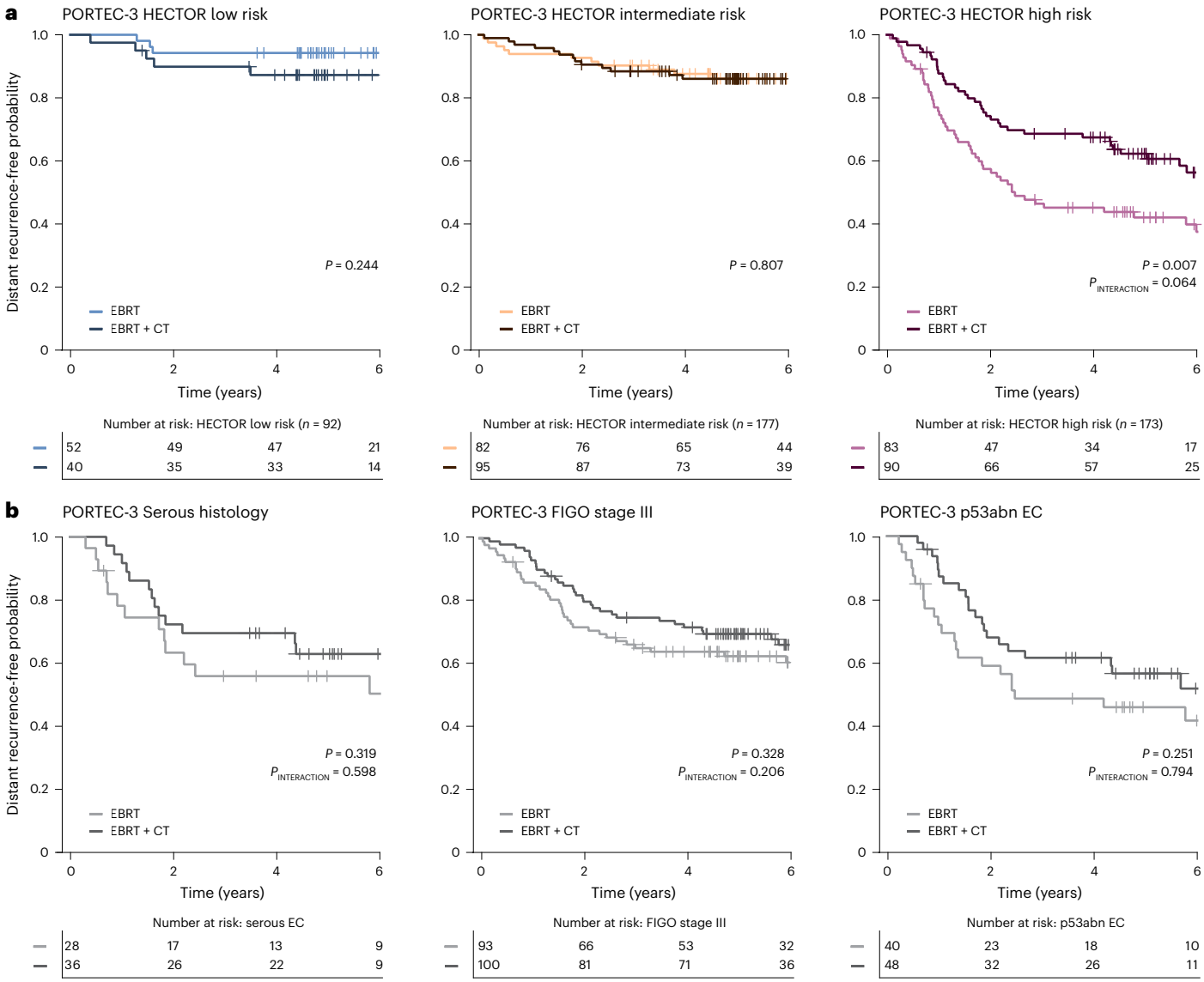


Fig. 6 | Impact of the addition of adjuvant chemotherapy to external beam radiotherapy on distant recurrence in the PORTEC-3 randomized trial by HECTOR risk group. a, The 6-year distant recurrence-free probability by Kaplan–Meier analysis and log rank test *P* value shown for each HECTOR risk group stratified by randomly allocated treatment. The *P* value of the interaction term using categorical HECTOR risk group is shown. There was also a significant interaction between the HECTOR continuous risk scores and the treatment ($P_{INTERACTION}$ = 0.014). **b**, For comparison with HECTOR selection, distant recurrence-free probability by Kaplan–Meier analysis from the PORTEC-3 trial for different gold standard prognostic factors in EC relying on serous histology, the FIGO 2009 stage III and the p53abn molecular class is shown. The log rank test and interaction term *P* values are displayed. EBRT, external beam radiotherapy; CT, chemotherapy.

chemotherapy had significantly improved distant recurrence-free probabilities compared with those treated with external beam radiotherapy alone (5-year distant recurrence-free probability of 62.2% (95% CI: 0.511–0.715) versus 42.0% (95% CI: 0.311–0.526); log rank $P = 0.007$; HR = 0.561 (95% CI: 0.366–0.862; $P = 0.008$)). Exploratory analysis suggested that the predictive accuracy was greater than that provided by prognostic factors currently used to identify patients with high-risk tumors who were likely to benefit from adjuvant chemotherapy, including serous histological subtype, FIGO 2009 stage III and the p53abn molecular class (Fig. 6b). Further exploratory analyses suggested that HECTOR also identified patients who benefited from adjuvant chemotherapy within the NSMP and MMRd molecular classes (Supplementary Figs. 29 and 30). These results remained consistent when sub-stratifying by the image-based molecular class arm of HECTOR (Supplementary Fig. 31). Thus, HECTOR demonstrated significant predictive utility that may exceed that offered by current methods.

Discussion

HECTOR, a DL model trained and validated in 2,072 patients with stage I–III EC[3,26–31], with long-term follow-up, predicts postoperative distant recurrence risk using only H&E-stained tumor slide(s) of the hysterectomy specimen and anatomical stage. HECTOR obtained C-indices of 0.789, 0.828 and 0.815 in three unseen test sets for distant recurrence outcome. Its performance is on a par with clinically implemented prognostic DL tools in other cancer types (C-indices of 0.714 and 0.744 for colorectal cancer recurrence[39], AUC of 0.78 for 10-year prostate cancer distant recurrence[40]) and also favorably compares with molecular prognostic assays such as OncotypeDX (C-index of 0.641 for 10-year breast cancer distant recurrence[41]). Notably, HECTOR outperformed the current diagnostic gold standard of combined pathological and molecular analysis for distant recurrence risk prediction, and was also found to be predictive of adjuvant chemotherapy benefit in the PORTEC-3 randomized trial[3]. Pending prospective validation, our results suggest

that HECTOR may have the potential to be a highly effective tool for individualized prognostication of women with EC, while delivering shorter turnaround times and reducing testing costs. HECTOR may also enable biomarker discoveries for improving targeted treatment decision-making.

HECTOR performance is the result of a new multimodal, integrative, three-arm architecture which leveraged prognostic information from the H&E WSI, the image-based molecular class from im4MEC[11] and anatomical stage[34]. This multimodal architecture outperformed alternative DL models using only H&E-based information, corroborating other studies[16,42]. It is interesting that nesting of the im4MEC model within HECTOR boosted the performance, in contrast to other studies where integration of copy number variation or transcriptomics did not improve prediction of overall survival in EC[16]. We demonstrated that the prognostic value of categorical clinical risk factors, such as the anatomical stage, can be learned end to end by the DL model to increase predictive accuracy. HECTOR takes a step toward integrating patient-level imaging, image-based molecular and clinical insights, which may benefit similar studies in other cancer types where unimodal DL models have been developed on images only[17,20,39].

Our preliminary investigations of model explainability and risk score correlates offer good prospects to improve our understanding of the biology of EC and other cancer types. For example, the association of HECTOR low-risk scores with immune cell infiltrate is consistent with data showing better prognosis of immune-infiltrated EC[10], although at present it is unclear whether HECTOR directly quantified lymphocyte subtypes such as T cells from H&E WSIs. The upregulation of *CLDN6* in HECTOR high-risk ECs is consistent with this being a predictor of distant recurrence[43]. Cases with combined HECTOR high risk and *CLDN6* upregulated could be actionable as a chimeric antigen receptor T cell target[44]. Although desmoplastic stromal reaction is known to predict bad prognosis in colorectal cancer, the association that we describe in the present study has not previously been reported in EC[45]. Whether this represents a morphological readout of *L1CAM* overexpression[46] is presently unclear. We also confirmed well-established, unfavorable histopathological risk factors in EC aligning with higher HECTOR risk scores[5]. Thus, we expect the outperformance of standard histopathology by HECTOR probably being driven by the nonlinear combination of each factor and, more importantly, the noncategorical processing of the visual information from the WSIs.

HECTOR's design holds considerable promise for scaling to clinical implementation because it is built on two broadly available and cost-effective inputs routinely obtained in diagnostics: one H&E-stained tumor slide from which we used the image-based rather than the true molecular classes and high-level clinical information of the tumor extension at diagnosis (to the cervix or beyond the uterus excluding distant) which is independent of an evolving FIGO staging system[9]. After appropriate validation in a prospective clinical trial setting, HECTOR may have great potential to individualize triage of women with EC in the adjuvant setting from low to high risk of distant recurrence. Subsequent treatment decision-making by clinicians could be guided accordingly because HECTOR low-risk prediction could provide a means to de-escalate adjuvant treatment or to encourage adjuvant systemic therapy recommendation for patients predicted to be HECTOR high risk (such as chemotherapy[3,4] or targeted therapies in clinical trials[47–49]). The therapeutic guidance within HECTOR high risk can be supported by selective targeted molecular testing such as MMRd or even DL-based molecular predictions given a good accuracy[11]. Although our data support that HECTOR could reduce under- and over-treatment for women with EC, it would also spare challenges and expenses of resource-limited environments where molecular testing and expert pathologist review are difficult or not feasible. We speculate that future technical improvements of HECTOR could be an extension of its inputs to consecutive digitized H&E-stained hysterectomy sections followed by three-dimensional reconstruction[50], routinely performed IHC-stained WSIs[51], preoperative radiology images[52] or a clinical report encoding patient-level clinical information[53]. Moreover, DL-based assessment of the anatomical stage by leveraging histology images of the cervical, ovarian and (or radiology images of) lymph node sections would make HECTOR independent of pathology review.

Our study has several strengths. Our total cohort of 2,751 patients, including 3 randomized trials, makes this one of the largest DL-based prognostic studies in EC performed to date. Our state-of-the-art multimodal DL methodology allowed us to leverage prognostic information from multiple factors, including those beyond the H&E image alone. Expert pathology review and molecular profiling enabled us to benchmark our methodology against the current gold standard in risk stratification of EC. Limitations of our study are that our current model based on multiple instance learning is unaware of the spatial relationship between regions and was not designed to leverage information between multiple WSIs, both of which may improve performance[54,55]; although context-aware architectures have not been found to improve performance in this task. In addition, complex interactions of the morphology, molecular and anatomical stage may be further optimized by experimenting with other early-to-late fusion techniques[42], or learning more generalizable morpho-molecular representations using pretext tasks. Some patients in the study did not undergo surgical staging lymphadenectomy[26,27], a consideration that may have introduced some noise in the anatomical stage input and may explain the residual prognostic value of advanced disease stage III in multivariable analysis. Given that *POLE*mut EC mutations rarely metastasize[56], we acknowledge the possibility that the risk may be overestimated in these rare instances by HECTOR. Furthermore, not all morphological correlates observed in the H&E regions (for example, structural changes) were quantified in the present study owing to the lack of available labeled datasets that could have been used for training DL-based, EC-specific image analysis tools. Importantly, HECTOR performance needs further validation both in unselected cohorts more diverse than the ones of largely European ancestry that we examined and in prospective trials. As such, prospective validation will be conducted first in the PORTEC-4a trial[57]. Moreover, as the therapeutic landscape of EC is rapidly evolving, the most suitable adjuvant systemic therapy for HECTOR high-risk patients needs to be continuously validated[4,58] or (prospectively) explored in other randomized trials[47–49,59].

In summary, validation and extension of HECTOR could help delivery of precision medicine to advance prognostication of women with stage I–III EC who underwent primary surgery, with improvement worldwide on both systemic therapy recommendation and treatment de-escalation.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-02993-w.

References

1. Crosbie, E. J. et al. Endometrial cancer. *Lancet* **399**, 1412–1428 (2022).
2. Ørtoft, G., Lausten-Thomsen, L., Høgdall, C., Hansen, E. S. & Dueholm, M. Lymph-vascular space invasion (LVSI) as a strong and independent predictor for non-locoregional recurrences in endometrial cancer: a Danish Gynecological Cancer Group Study. *J. Gynecol. Oncol.* **30**, e84 (2019).
3. de Boer, S. M. et al. Adjuvant chemoradiotherapy versus radiotherapy alone in women with high-risk endometrial cancer (PORTEC-3): patterns of recurrence and post-hoc survival analysis of a randomised phase 3 trial. *Lancet Oncol.* **20**, 1273–1285 (2019).

4. Hogberg, T. et al. Sequential adjuvant chemotherapy and radiotherapy in endometrial cancer—results from two randomised studies. *Eur. J. Cancer* **46**, 2422–2431 (2010).
5. Concin, N. et al. ESGO/ESTRO/ESP guidelines for the management of patients with endometrial carcinoma. *Int. J. Gynecol. Cancer* **31**, 12–39 (2021).
6. Abu-Rustum, N. et al. Uterine neoplasms, version 1.2023, NCCN Clinical Practice Guidelines in Oncology. *J. Natl Compr. Cancer Netw.* **21**, 181–209 (2023).
7. Oaknin, A. et al. Endometrial cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann. Oncol.* **33**, 860–877 (2022).
8. Harkenrider, M. M. et al. Radiation therapy for endometrial cancer: an American Society for Radiation Oncology clinical practice guideline. *Pract. Radiat. Oncol.* **13**, 41–65 (2023).
9. Berek, J. S. et al. FIGO staging of endometrial cancer: 2023. *Int. J. Gynecol. Obstet.* **162**, 383–394 (2023).
10. Horeweg, N. et al. Prognostic integrated image-based immune and molecular profiling in early-stage endometrial cancer. *Cancer Immunol. Res.* **8**, 1508–1519 (2020).
11. Fremond, S. et al. Interpretable deep learning model to predict the molecular classification of endometrial cancer from haematoxylin and eosin-stained whole-slide images: a combined analysis of the PORTEC randomised trials and clinical cohorts. *Lancet Digit. Health* **5**, e71–e82 (2023).
12. Lafarge, M. W. & Koelzer, V. H. Towards computationally efficient prediction of molecular signatures from routine histology images. *Lancet Digit. Health* **3**, e752–e753 (2021).
13. Sirinukunwattana, K. et al. Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning. *Gut* **70**, 544–554 (2021).
14. Graham, S. et al. Hover-Net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* **58**, 101563 (2019).
15. Lee, Y. et al. Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. *Nat. Biomed. Eng.* https://doi.org/10.1038/s41551-022-00923-0 (2022).
16. Chen, R. J. et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* **40**, 865–878.e6 (2022).
17. Wulczyn, E. et al. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ Digit. Med.* **4**, 71 (2021).
18. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N. & Huang, J. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Med. Image Anal.* **65**, 101789 (2020).
19. Chen, R. J. et al. Whole slide images are 2D point clouds: context-aware survival prediction using patch-based graph convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* Vol. 12908 (eds de Bruijne, M. et al.) 339–349 (Springer Cham, 2021).
20. Courtiol, P. et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
21. Chen, R. J. et al. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 3995–4005 (IEEE, 2021); https://ieeexplore.ieee.org/document/9710773
22. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *Proc. of the 35th International Conference on Machine Learning* Vol. 80 (eds Dy, J. & Krause, A.) 2127–2136 (PMLR, 2018).
23. Wagner, S. J. et al. Transformer-based biomarker prediction from colorectal cancer histology: a large-scale multicentric study. *Cancer Cell* **41**, 1650–1661.e4 (2023).
24. Using AI to improve the molecular classification of brain tumors. *Nat. Med.* **29**, 793–794 (2023).
25. Jiménez-Sánchez, D. et al. Weakly supervised deep learning to predict recurrence in low-grade endometrial cancer from multiplexed immunofluorescence images. *NPJ Digit. Med.* **6**, 48 (2023).
26. Creutzberg, C. L. et al. Surgery and postoperative radiotherapy versus surgery alone for patients with stage-1 endometrial carcinoma: multicentre randomised trial. PORTEC study group. post operative radiation therapy in endometrial carcinoma. *Lancet* **355**, 1404–1411 (2000).
27. Nout, R. A. et al. Vaginal brachytherapy versus pelvic external beam radiotherapy for patients with endometrial cancer of high-intermediate risk (PORTEC-2): an open-label, non-inferiority, randomised trial. *Lancet* **375**, 816–823 (2010).
28. Stelloo, E. et al. Refining prognosis and identifying targetable pathways for high-risk endometrial cancer; a TransPORTEC initiative. *Mod. Pathol.* **28**, 836–844 (2015).
29. Jobsen, J. J. et al. Outcome of endometrial cancer stage IIIA with adnexa or serosal involvement only. *Obstet. Gynecol. Int.* **2011**, 962518 (2011).
30. Ørtoft, G. et al. Location of recurrences in high-risk stage I endometrial cancer patients not given postoperative radiotherapy: a Danish gynecological cancer group study. *Int. J. Gynecol. Cancer* **29**, 497–504 (2019).
31. Workel, H. H. et al. CD103 defines intraepithelial $CD8^+$ $PD1^+$ tumour-infiltrating lymphocytes of prognostic significance in endometrial adenocarcinoma. *Eur. J. Cancer* **60**, 1–11 (2016).
32. Kandoth, C. et al. Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73 (2013).
33. Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. & Wei, L. J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* **30**, 1105–1117 (2011).
34. Pecorelli, S. Revised FIGO staging for carcinoma of the vulva, cervix, and endometrium. *Int. J. Gynaecol. Obstet.* **105**, 103–104 (2009).
35. Zadeh, A., Chen, M., Poria, S., Cambria, E. & Morency, L.-P. Tensor fusion network for multimodal sentiment analysis. In *Proc. 2017 Conference on Empirical Methods in Natural Language Processing* 1103–1114 (Association for Computational Linguistics, 2017).
36. Mormont, R., Geurts, P. & Maree, R. Multi-task pre-training of deep neural networks for digital pathology. *IEEE J. Biomed. Health Inform.* **25**, 412–421 (2021).
37. Lambert, J. & Chevret, S. Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves. *Stat. Methods Med. Res.* **25**, 2088–2102 (2016).
38. Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.* **18**, 2529–2545 (1999).
39. Pai, R. K. et al. Quantitative pathologic analysis of digitized images of colorectal carcinoma improves prediction of recurrence-free survival. *Gastroenterology* **163**, 1531–1546.e8 (2022).
40. Esteva, A. et al. Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials. *NPJ Digit. Med.* **5**, 71 (2022).
41. Pece, S. et al. Comparison of StemPrintER with Oncotype DX recurrence score for predicting risk of breast cancer distant recurrence after endocrine therapy. *Eur. J. Cancer* **164**, 52–61 (2022).
42. Jaume, G. et al. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).

43. Kojima, M. et al. Aberrant claudin-6-adhesion signaling promotes endometrial cancer progression via estrogen receptor α. *Mol. Cancer Res.* **19**, 1208–1220 (2021).
44. Mackensen, A. et al. CLDN6-specific CAR-T cells plus amplifying RNA vaccine in relapsed or refractory solid tumors: the phase 1 BNT211-01 trial. *Nat. Med.* https://doi.org/10.1038/s41591-023-02612-0 (2023).
45. Ueno, H. et al. Prognostic value of desmoplastic reaction characterisation in stage II colon cancer: prospective validation in a phase 3 study (SACURA trial). *Br. J. Cancer* **124**, 1088–1097 (2021).
46. Corrado, G. et al. Endometrial cancer prognosis correlates with the expression of L1CAM and miR34a biomarkers. *J. Exp. Clin. Cancer Res.* **37**, 139 (2018).
47. Mirza, M. R. et al. Dostarlimab for primary advanced or recurrent endometrial cancer. *N. Engl. J. Med.* **388**, 2145–2158 (2023).
48. Makker, V. et al. Lenvatinib plus pembrolizumab for advanced endometrial cancer. *N. Engl. J. Med.* **386**, 437–448 (2022).
49. Eskander, R. N. et al. Pembrolizumab plus chemotherapy in advanced endometrial cancer. *N. Engl. J. Med.* **388**, 2159–2170 (2023).
50. Kiemen, A. L. et al. Tissue clearing and 3D reconstruction of digitized, serially sectioned slides provide novel insights into pancreatic cancer. *Med* **4**, 75–91 (2023).
51. Foersch, S. et al. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nat. Med.* **29**, 430–439 (2023).
52. Braman, N. et al. Deep orthogonal fusion: multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021* (eds de Bruijne, M. et al.) 667–677 (Springer, 2021).
53. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
54. Jaume, G., Song, A. H. & Mahmood, F. Integrating context for superior cancer prognosis. *Nat. Biomed. Eng.* **6**, 1323–1325 (2022).
55. Song, A. H. et al. Analysis of 3D pathology samples using weakly supervised AI. *Cell* **187**, 2502–2520.e17 (2024).
56. León-Castillo, A. et al. Molecular classification of the PORTEC-3 trial for high-risk endometrial cancer: impact on prognosis and benefit from adjuvant therapy. *J. Clin. Oncol.* **38**, 3388–3397 (2020).
57. van den Heerik, A. S. V. M. et al. PORTEC-4a: international randomized trial of molecular profile-based adjuvant treatment for women with high–intermediate risk endometrial cancer. *Int. J. Gynecol. Cancer* **30**, 2002–2007 (2020).
58. Kuoppala, T. et al. Surgically staged high-risk endometrial cancer: randomized study of adjuvant radiotherapy alone vs. sequential chemo-radiotherapy. *Gynecol. Oncol.* **110**, 190–195 (2008).
59. RAINBO Research Consortium. Refining adjuvant treatment in endometrial cancer based on molecular features: the RAINBO clinical trial program. *Int. J. Gynecol. Cancer* **33**, 109–117 (2022).
60. Li, C. et al. Efficient self-supervised vision transformers for representation learning. In *International Conference on Learning Representations* (ICLR, 2022); https://openreview.net/forum?id=fVu3o-YUGQK
61. Zadeh, S. G. & Schmid, M. Bias in cross-entropy-based training of deep survival networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 3126–3137 (2021).
62. Höhn, A. K. et al. 2020 WHO classification of female genital tumors. *Geburtshilfe Frauenheilkd.* **81**, 1145–1153 (2021).
63. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *Proc. of the 34th International Conference on Machine Learning* Vol. 70 (eds Precup, D. & Teh, Y. W.) 3319–3328 (PMLR, 2017).
64. Lafarge, M. W. & Koelzer, V. H. in *Mitosis Domain Generalization and Diabetic Retinopathy Analysis* (eds.Sheng, B. & Aubreville, M.) 226–233 (Springer Nature Switzerland, 2023).

[1]Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands. [2]Department of Radiation Oncology, Leiden University Medical Center, Leiden, The Netherlands. [3]Department of Computer Science, ETH Zurich, Zurich, Switzerland. [4]Department of Pathology and Molecular Pathology, University Hospital, University of Zurich, Zurich, Switzerland. [5]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [6]Department of Gynecology and Obstetrics, Leiden University Medical Center, Leiden, The Netherlands. [7]Department of Gynecology, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark. [8]Department of Pathology, Herlev University Hospital, Herlev, Denmark. [9]Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands. [10]Department of Radiation Oncology, Medisch Spectrum Twente, Enschede, The Netherlands. [11]Maastricht Radiation Oncology, MAASTRO, Maastricht, The Netherlands. [12]Department of Clinical Oncology, Barts Health NHS Trust, London, UK. [13]Department of Medical Oncology, Peter MacCallum Cancer Center, Melbourne, Victoria, Australia. [14]Department of Medical Oncology and Hematology, Odette Cancer Center Sunnybrook Health Sciences Center, Toronto, Ontario, Canada. [15]Department Medical Oncology, Gustave Roussy Institute, Villejuif, France. [16]Department of Surgical Sciences, Gynecologic Oncology, Città della Salute and S Anna Hospital, University of Turin, Turin, Italy. [17]Department of Obstetrics and Gynecology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands. [18]Department of Radiotherapy, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, The Netherlands. [19]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [20]Oxford NIHR Comprehensive Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [21]Institute of Medical Genetics and Pathology, University Hospital Basel, Basel, Switzerland. [22]These authors contributed equally: Viktor H. Koelzer, Tjalling Bosse. ✉e-mail: t.bosse@lumc.nl

Methods

Ethics statement

The PORTEC-1, PORTEC-2 (NCT00376844) and PORTEC-3 (NCT00411138) study protocols were approved by the Medical Ethical Committee Leiden, Den Haag, Delft and the medical ethics committees at participating centers. Studies were conducted in accordance with the principles of the Declaration of Helsinki. Ethical permissions for the retrospective use of the clinical trials and retrospective cohorts (TransPORTEC study, Medisch Spectrum Twente (MST)) were obtained by the Medical Ethical Committee Leiden (nos. B21.065 and B21.011), as well as the LUMC cohort (nWMO-D4-2023-002) and the Danish Cohort by the Center for Regional Udvikling, De Videnskabsetiske Komiteer (H-16025909). All study participants of the clinical trials provided informed consent. The ethical boards have provided a waiver for informed consent for the other studies. For the UMCG cohort, the medical ethical committee granted permission for the use of the data and provided a waiver for informed consent owing to the observational nature of the study.

Cohorts

We used formalin-fixed paraffin-embedded (FFPE) tumor material and clinicopathological data of patients with EC from three randomized trials and six clinical cohorts. We included study participants of the female sex, independent of gender identity.

The PORTEC-1 trial recruited 714 women with early stage intermediate-risk EC from 1990 to 1997, and after primary surgery, randomly assigned to pelvic external beam radiotherapy or no adjuvant treatment[26]. The PORTEC-2 trial randomized 427 women with early stage, high- to intermediate-risk EC between 2000 and 2006 to external beam radiotherapy or vaginal brachytherapy[27]. The PORTEC-3 randomized trial included 660 women with stage I–III high-risk EC from 2006 and 2013, and randomly allocated them to pelvic external beam radiotherapy alone or external beam radiotherapy combined with concurrent and adjuvant chemotherapy[3]. The retrospective TransPORTEC study included 116 high-risk EC tumors from international patients using the same inclusion criteria as the PORTEC-3 from 5 institutions (LUMC and UMCG, the Netherlands; University College London and St Mary's Hospital, Manchester, UK; and Institute Gustave Roussy, Villejuif, France)[28]. The prospective cohort of MST included 257 patients with stage I–III high-risk EC, with the same inclusion criteria as PORTEC-3, who were treated between 1987 and 2015 at MST, Enschede in the Netherlands[29]. The Danish cohort consisted of 451 patients with high-grade EC who were prospectively registered in the Danish gynecological cancer database[30]. The UMCG cohort is a population-based cohort consisting of patients treated at the UMCG between 1984 and 2004, that is, 278 patients with follow-up data collected until 2010 (ref. 31). The LUMC cohort is a retrospectively collected, population-based cohort of 222 patients diagnosed and treated at the LUMC between 2012 and 2021. Finally, the publicly available TCGA-UCEC cohort[32] of 529 patients was downloaded from the cBioPortal[65,66].

Datasets

One representative H&E-stained slide of the hysterectomy specimen was included for each patient depending on the availability of the tumor material (Supplementary Figs. 1 and 2, and Supplementary Tables 1, 2 and 14). For the LUMC cohort, we collected three diagnostic H&E-stained tumor slides per patient case with EC, each from a different FFPE tumor tissue block. H&E slides were scanned at ×40 magnification using two scanners 3Dhistech P250 (resolution 0.19 µm per pixel) and 3Dhistech P1000 (resolution 0.24 µm per pixel). Any image provided in the manuscript is an unprocessed scan. Qualitative review was conducted on all WSIs by our expert pathologist, after which cases with no tumor, poor tissue quality and out-of-focus scanning issues were excluded, yielding 2,560 cases with at least one WSI per case (CONSORT chart in Supplementary Figs. 1 and 2).

In the present study, some cases were excluded from the supervised training of HECTOR based on the following criteria: (1) missing time to distant recurrence follow-up data, (2) FIGO 2009 stage IV[34] because they already have distant recurrence at time of diagnosis and (3) treatment with adjuvant chemotherapy because it may have lowered the risk of distant recurrence[3,4]. The categorical anatomical stages I, II and III are defined following the FIGO 2009 classification[34]. Hence, it represents a tumor confined in the uterus (stage I), a tumor spread to the cervical stroma (stage II) or to the vagina, adnexa, pelvis and lymph nodes (stage III) at diagnosis. Distant recurrence in the adjuvant setting was defined as any recurrence outside the pelvis. Hence, distant recurrence included abdominal metastasis and para-aortic lymph node metastasis. Time to distant recurrence was defined to start at randomization (for PORTEC-1, -2 and -3) or date of primary surgery (MST, TransPORTEC study, Danish, UMCG and LUMC cohort) and to end at the date of the diagnosis of metastasis, or the date of last follow-up or death in patients without metastasis. We also stress that adjuvant chemotherapy was not the standard of care at the time the clinical cohorts were collected and that the vast majority of patients treated with adjuvant chemotherapy originated from the PORTEC-3 randomized trial ($n = 225$).

Following the aforementioned criteria, 2,072 cases were included for the supervised train–test split: 584 from PORTEC-1 (ref. 26), 395 from PORTEC-2 (ref. 27), 217 from PORTEC-3 (ref. 3), 67 from the TransPORTEC study[28], 226 from the MST cohort[29], 272 from the Danish cohort[30], 160 from the UMCG cohort[31] and 151 from the LUMC cohort. Then we held out one internal test set and two external test sets, all representing an unselected population. The internal test set was obtained by randomly sampling 20% of the supervised training set, stratified by discrete time intervals and censorship status to ensure the presence of enough events across time ($n = 353$, of which 116 were from PORTEC-1, 100 from PORTEC-2, 43 from PORTEC-3, 13 from the TransPORTEC study, 35 from the MST cohort and 46 from the Danish cohort; median follow-up of 8.45 years with 62 events). The first external test set is the UMCG cohort ($n = 160$ patients; 5.32-year median follow-up time with 14 events). The second external test set is the LUMC cohort ($n = 151$ patients: 121 with 3 WSIs, 21 with 2 WSIs and 9 with 1 WSI; 2.90-year median follow-up time with 24 events). Finally, the remaining 1,408 WSIs were used for supervised training of HECTOR (468 from PORTEC-1, 295 from PORTEC-2, 174 from PORTEC-3, 54 from the TransPORTEC study, 191 from the MST cohort and 226 from the Danish cohort; median follow-up of 7.77 years with 246 events).

In addition, the HECTOR risk scores were predicted on the previously excluded, chemotherapy-treated cases from the PORTEC-3 randomized trial[3] ($n = 225$), as well as the patients with stages I–III from TCGA-UCEC ($n = 381$).

For the self-supervised learning, we used only the 1,408 WSIs already reserved for supervised training, and thus strictly limited to only those that were not part of the internal and external test sets. In addition, the self-supervised learning training was enriched by cases with any stage of disease, whose treatment or distant recurrence outcome data were unknown ($n = 454$ of which 31 from the TransPORTEC study, 5 from the MST cohort, 16 from the Danish cohort and 402 from TCGA-UCEC), resulting in 1,862 cases for self-supervised learning.

Performance evaluation

Hyperparameter optimization and model comparisons (including architecture choices for patch representational learning with self-supervised learning) were evaluated on the supervised downstream task guided by the C-index metric[33] (using a tau = 10 years and scikit-survival Python package (v.0.17.2)). To this end, a fivefold crossvalidation routine was performed on the 1,408 WSIs reserved for supervised training. The most performant architecture and hyperparameters were selected based on the highest mean C-index over the five folds. The final model, referred to as HECTOR, is then retrained on the full training set and evaluated on to the internal and the two external test sets (UMCG and LUMC). The cumulative AUC[37] and Brier scores[38] were additionally computed.

Given the fact that the LUMC external test set contains up to three WSIs per case, as opposed to one in the internal test set and the UMCG external test set, we performed multiple experiments to derive patient-level risk scores using random sampling. First, we randomly selected one WSI per case and repeated this experiment 100×, yielding a mean C-index and CI. Second, we randomly selected up to two WSIs for each case when available, then averaged with the mean the two risk scores per patient and repeated it 100×. Third, we selected all available WSIs of the external test set with up to three WSIs per case when available and computed the mean and median of the two or three risk scores. In an additional experiment, we combined each patient's WSIs by merging the patch features from all available WSIs into a single feature bag.

WSI preprocessing

WSI segmentation was performed using Otsu thresholding. Nonoverlapping patching was performed at 180 µm and patches were resized to 256 × 256 pixels^2. On average, this procedure generated a bag of 10,185 patches per WSI.

Vision transformer-based patch representational learning

We followed advancements in self-supervised learning by adopting vision transformer-based DL models that are capable of learning fine-grained, patch-level representation at multiple resolutions. For this, we trained EsVIT[60] and compared it with CtransPath[67], an alternative model trained on the histopathology domain (Supplementary Table 3). We modified the initial proposed four-stage Swin[68], transformer-based architecture of EsVIT to capture cell- and region-level tissue information and to fit our computational resources. The patch size of stage 1 was doubled to 8 pixels to reduce the sequence length and increase field of view to capture cell views. In stages 2–4, we kept the two-factor feature map merging rate and resized the input images to 256 × 256 pixels^2 instead of 224 × 224 pixels^2 to avoid indivisible patch size at stage 4. Finally, the number of stacked transformers in stage 3 was reduced from six to four and the rest were kept to two. The first embedding dimension remained unchanged at 96 and the number of attention heads by stage was also kept unchanged, that is, 3, 6, 12 and 24 (Supplementary Table 4).

A dataset of 3,702,447 patches was curated by randomly extracting up to 2,000 patches per WSI at 180 µm resized to 256 × 256 pixels^2 from the 1,862 WSIs appointed for self-supervised learning. Thereafter, the modified EsVIT was trained on 3 Nvidia RTX 8000 GPUs (graphic processing units) with a batch size of 128 for 100 epochs with a window of 14 to encourage learning of long-term dependencies between patches. For performance improvement, we also used the view- and region-level prediction DINO (self-distillation with no labels) heads with no weight normalization and frozen layers at first epoch and the default output dimension of 65,536 (ref. 60). We followed the EsVIT authors' recommendations with a smaller batch size by increasing the momentum teacher to 0.9996 and starting with the initial teacher temperature of 0.04. The teacher temperature was adjusted halfway through training from 0.04 to 0.02 for further loss decrease. We optimized with AdamW and default parameters, default optimization routines of the learning rate (linear warm-up for ten epochs followed by cosine scheduler to 1×10^{-6}) and weight decay (cosine scheduler from 0.04 to 0.4). The data augmentation was used exactly as done in the original publication[60].

After the training was completed, the patch-level features were extracted from the attention heads of the stacked transformers at each stage. For our downstream task, we observed an improvement by extracting the last 8 blocks compared with the default last 4 mentioned in the publication[60], yielding feature vectors of size 3,456 (Supplementary Table 3).

Multimodal DL prognostic model

To build the multimodal model for distant recurrence prediction task, ablation studies were first performed using the H&E WSI modality only (referred to as H&E-based, one-arm model) followed by integrating the image-based molecular classes derived from the H&E-based predictions of im4MEC[11] (referred to as two-arm model) and the categorical stage (hence referred to as HECTOR). This section describes HECTOR with Supplementary Table 5 summarizing the architecture and training parameters, whereas 'Ablation studies' provides further details about some training experiments and the choice of the architecture.

The H&E-based, one-arm model takes as input the bag of 180-µm patch-level features of size 3,456 extracted from EsVIT[60], where the number of patches per bag varies. To train toward time-to-event data and given a batch size of one of the attention-based multiple instance learning (AttentionMIL) model, the time scale was discretized into four intervals based on the quartiles of the distribution of uncensored patients and the −log(likelihood loss) was used[61].

Within the AttentionMIL model, we reported a slight performance increase by adding another WSI preprocessing step. Specifically, WSI morphological information was spatially and semantically compressed by averaging highly correlated, nearby patch-level features using a L2 norm threshold of three patches and a cosine similarity of 0.8. This step reduced the bag of features from 10,185 patches on average to 1,723 at 180 µm (Supplementary Table 3). Each mean patch-level feature is compressed by 3 Fully Connected layers gradually down to 512. The attention module computes attention scores on latent features reduced to 256 before pooling, resulting in a slide-level embedding of size 512.

To leverage the well-established prognostic value of the molecular class (here image-based derived from the H&E-based predictions of im4MEC[11]) and the categorical (FIGO 2009) stage I, II and III variable, and given the AttentionMIL model computes an H&E slide-level embedding from the patches, we experimented with intermediate-to-late fusion to integrate slide-level, image-based molecular class and patient-level anatomical stage information at the H&E slide-level embedding. We proposed an approach of first encoding each categorical risk factor to higher-dimensional vector space with a learnable Embedding layer of size 16 followed by Elu activation function and one Fully Connected layer of size 8. Next, a gating-based attention mechanism with bilinear product was applied on the embeddings from different modalities to weight the importance of each modality based on ref. 16. To capture all interactions and retain unimodal embeddings, one was appended to the attention-weighted embeddings and then fused using the Kronecker product[35]. It is important to note that, for using the image-based molecular class as an input modality for HECTOR, we retrained the im4MEC model on the training set specifically designed for the present study. This was done to avoid any information leakage because some cases used for training the original im4MEC model were used as testing on validation in the present study.

The final multimodal embedding was further reduced by using two Fully Connected layers of size 256 and 128 before the survival categorical head of a Fully Connected layer with output size as the number of discrete time intervals. Each Fully Connected layer in the architecture was followed by a dropout of 0.25 and a ReLU activation function.

HECTOR was trained for 24 epochs with an initial learning rate of 3×10^{-5} decayed by a factor of 10 at epochs 2, 5 and 15. The Adam optimizer was used with default parameters and a weight decay of 1×10^{-5}. HECTOR was also developed by adapting sections of open access repositories[11,16,21].

Ablation studies

To find first the optimal architecture to predict distant recurrence from the H&E modality (one-arm model), three state-of-art WSI classification architectures were adapted to our distant recurrence prediction task: AttentionMIL[22], a Graph Attention Network following ref. 15, with a radius up to 32 connected patch nodes and a transformer architecture following ref. 23. Both of these architectures were adapted from their open access repository. They were both trained on the same feature bags extracted using EsVIT with a batch size of one and the same

discrete survival loss (−log(likelihood loss)). We found that the AttentionMIL architecture yielded a higher C-index than the Graph Attention Network and the transformer in this prognostic task while featuring far lower computational complexity (Supplementary Table 3), which corroborates the findings of ref. 15 for TCGA-UCEC.

To incorporate the image-based molecular class predicted by im4MEC from the H&E WSIS, experiments included: (1) transfer learning in which the AttentionMIL backbone was pretrained toward the molecular class and subsequently fine-tuned on the prognostic task; (2) multitask learning in which a second training objective was added to predict the image-based molecular class in addition to the prognosis; and (3) fusion of the image-based molecular class derived from the frozen im4MEC model (as extracted from either an intermediate layer or the final predicted categorical class, followed by an Embedding layer and attention gate). In experiment 2, a second classification head was implemented which was trained using the weighted sum of the survival loss (−log(likelihood loss)) and the cross-entropy classification loss. The weight factor was considered as a hyperparameter and was optimized using the fivefold crossvalidation. Experiment 3 which consisted of the inclusion of the predicted categorical class using an Embedding layer and attention gate resulted in the highest mean C-index (Supplementary Table 3).

Experiments around fusing the stage category included notably training with the extended FIGO 2009 taxonomy or a reduced three-class taxonomy (I, II and II) followed by an Embedding layer and attention gate, the latter achieving the highest C-index (Supplementary Table 3).

Association with clinicopathological data analysis

We performed multiple single linear regression analyses using the HECTOR continuous risk scores as the dependent variable and the clinicopathological data as the regressor. Statistical tests were two sided with statistical significance accepted with P values <0.050. Regression coefficients and exact P values have been reported in Supplementary Table 7.

Input contribution

The IG method[63] was used to measure the contribution of the WSI and to identify the patches within a WSI relevant to the prediction of the hazard function. Given the discrete time intervals, IG scores were averaged over the four neuron targets. The IG baseline for feature missingness was represented as patch-level features derived from white patches. All IG scores were patient-wise normalized between −1 and +1 while maintaining the sign and the IG score of zero, and further averaged to get a WSI-level IG score. Positive IG value toward 1 means that it contributed positively to increase the risk score, whereas negative means it contributed to decrease the risk score. Selection of representative patches was performed once by an expert pathologist within the top 5% patches, increasing and decreasing the risk scores for each case.

The contribution of the predicted image-based molecular class by im4MEC and the FIGO stage was calculated by fixing the stage- and image-based molecular class values with the value of our choice (referred to as the 'reference group') followed by computing the difference in predicted risk scores. Similar to the IG method, a positive or negative difference means a positive or negative contribution to the risk score, respectively.

Cell-level composition

As part of the explainability section of HECTOR to quantify visual features of extracted patches with high contribution, we first used the cell segmentation and classification Hover-Net[14] DL model to obtain inflammatory cell counts, retrained on EC-specific WSIs[11]. Then, mitotic figures were detected with a pan-cancer DL-based detector[64] that was fine-tuned on EC tissue for the purpose of the present study. Fine-tuning was performed by extending the original training set[69] with additional data points that we internally annotated in 10 WSIs from the PORTEC datasets selected to cover the variability of EC histological types. Region-level inflammatory and mitotic activity density were defined as absolute count normalized by the area in square millimeters and further averaged over the number of regions to obtain a patient-level density value. The size of tumor nuclei was reported in mm^2 and averaged by patient. The statistical association between the HECTOR risk scores and the patient-level quantity of visual features was tested with linear regressions within the regions of interest, that is, the regions with either a negative or a positive contribution. Statistical tests were two sided with statistical significance accepted for P values <0.050. The coefficients of linear regressions and exact P values were the following: coefficient −0.0109 (95% CI: −0.019 to −0.002), $P = 0.011$, for the patient-level inflammatory density within the negative regions; and coefficient 0.0447 (95% CI: 0.033–0.057), $P = 1.96 \times 10^{-12}$ for the patient-level mitotic density within the positive regions; coefficient 377.916 (95% CI: 297.677–458.155), $P = 3.10 \times 10^{-19}$, for the patient-level tumor nuclei area within the positive regions.

Outcome analysis

Analysis of distant recurrence-free probabilities was conducted according to the Kaplan–Meier method and the two-sided log rank test with statistical significance accepted for $P < 0.050$. Cutoffs for the HECTOR risk groups were defined by taking the quantiles (25%, 50% and 75%) of the distribution of HECTOR risk scores in the training set only. In the training set, the first two groups (<25% and between 25% and 50%) did not show any major difference in prognosis and were therefore merged into one group named the HECTOR low-risk group. As a result, we defined the HECTOR low-risk group as cases with a risk score below the median risk score value of the training set, the HECTOR intermediate-risk group as those with a risk score between median and third quartile values of the training set and the HECTOR high-risk group as those with a risk score greater than the third quartile value of the training set. These same cutoff values were applied to the unseen internal, UMCG and LUMC external test sets, and the TCGA-UCEC and PORTEC-3.

To compare the DL model performance with well-established clinicopathological risk factors, we fitted CPH models on these clinicopathological risk factors in EC and calculated the corresponding C-index. First, we used risk factors that can be visually assigned on histological slides: the histological subtype, the grade and LVSI. Then we added the FIGO 2009 stage I–III variable. Finally, we included the molecular class of EC (*POLE*mut, MMRd, NSMP and p53abn). To maintain consistency within validation sets in the fivefold crossvalidation and the internal test sets, missing molecular class (115 out of 1,408 in crossvalidation and 38 out of 353 in the internal test set) was imputed using mean substitution.

To estimate HECTOR's prognostic value as compared to the clinicopathological risk factors, we computed HRs using CPH with HECTOR continuous risk scores. For these analyses, we included all cases with a complete set of clinicopathological and molecular risk factors ($n = 1,254$). First, we corrected the HECTOR risk scores for all clinicopathological risk factors combined into one risk score in a multivariable analysis. To this end, a CPH model was first fitted on to these clinicopathological risk factors. Then, the derived risk scores, referred to as 'clinical', were calculated by taking the linear combination of the CPH coefficients and the variables. In the second analysis, we corrected HECTOR's continuous risk scores for the histological subtype, the grade, LVSI, stage, the molecular class and, in addition, L1CAM and age as continuous data in a multivariable analysis.

The histological subtype categorical variable was processed as grade 3 EEC versus the reference group low-grade EEC and non-EEC versus the reference EEC. The reference group for molecular class was NSMP and stage I for the FIGO 2009 stage variable.

All statistical tests were two sided with statistical significance accepted for P values <0.050.

Genomic and transcriptomic correlation analysis

To analyze the frequency of driver mutations by HECTOR risk groups, the genomic features were extracted from ref. 70 using MC3 MAF (mutation annotation format) data. The mutational status of the top 19 oncogenic drivers in EC was downloaded from the cBioPortal portal[65,66] and annotated by OncoKB[71]. The statistical comparison of proportions with oncogenic mutations between HECTOR risk groups was performed using the two-sided χ^2 tests for each individual gene with $P < 0.050$ accepted as significant. Exact P values and sample size are reported in Supplementary Table 12.

The association between the HECTOR continuous risk scores and each immune cell subset was performed using the $\log_2$(transformed proportion of the immune cell subset) as a fraction of the whole tumor, using the leukocyte fraction values. Linear regressions were performed with the HECTOR continuous risk scores as the independent variable. In addition, we tested the associations by correcting for the molecular class and TMB as additional independent variables. Two-sided P values <0.050 are accepted as significant. Regression coefficients and exact P values have been reported in Supplementary Table 13.

Messenger RNA sequencing (mRNA-seq) and clinical data from TCGA-UCEC were downloaded from firebrowse.org. Differentially expressed genes were assessed between HECTOR high-risk and HECTOR low-risk cases by DESeq2 (ref. 72) (v.1.40.1). Genes with a likelihood ratio test P value adjusted using a Benjamini–Hochberg false discovery rate (FDR) were accepted if <0.050 (Supplementary Table 15).

Analysis of adjuvant chemotherapy effect

We predicted the HECTOR risk scores for the patients included in the PORTEC-3 (ref. 3) treatment arm who did receive concurrent and adjuvant chemotherapy ($n = 225$) and, thus, who had been previously left out from training and any test sets. The effect of the combination of adjuvant chemotherapy and external beam radiotherapy over external beam radiotherapy alone was analyzed by: (1) analyzing distant recurrence-free probabilities by treatment arm stratified by HECTOR risk group and measuring group-wise treatment effect with the Kaplan–Meier method and the two-sided log rank test and/or HR of treatment variable with the univariable Cox's model; (2) calculating the statistical significance of the interaction term between the HECTOR continuous risk scores and the treatment binary variable; and (3) calculating the statistical significance of the interaction term between the HECTOR high-risk group and the treatment binary variable (corrected for HECTOR intermediate-risk group and using HECTOR low-risk group as a reference group). To measure the statistical significance of the interaction term defined as the HECTOR risk score (continuous or categorical) multiplied by the treatment binary variable, a multivariable Cox's regression analysis was performed. Similar analyses were performed to test the interaction between serous histological subtype and the chemotherapy treatment binary variable (corrected for EEC and clear cell histological subtype), and the FIGO 2009 stage III (corrected for stages I–II) and p53abn (corrected for MMRd, NSMP as a reference group and *POLE*mut tumors removed to reach convergence).

All statistical tests were two sided with statistical significance accepted with P values <0.050.

Software and packages

EsVIT and HECTOR were implemented with Pytorch (v.1.8.1 and v.1.10.0, respectively). IG was implemented with Captum Python package (v.0.6.0), metrics such as the C-index with scikit-survival Python package (v.0.17.2), CPH models and the Kaplan–Meier method with Lifelines Python package (v.0.27.1), χ^2 tests with Scipy Python package (v.1.5.2), boxplot visualizations with altair Python package (v.4.2.0) and linear regression with statsmodels Python package (v.0.13.5). Differentially expressed genes were performed using DESeq2 (v.1.40.1)[72] and R v.4.3.0 (2023-04-21 ucrt). Additional packages for image processing included Openslide Python package (v.1.1.2), OpenCV (v.4.3.0.36) and Pillow (v.7.2.0). Annotations were done with QuPath (v.0.4.1).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The tumor material and datasets generated during or analyzed in the present study are not publicly available owing to restrictions by privacy laws. Data and tumor material from PORTEC-1, PORTEC-2, PORTEC-3, MST and the TransPORTEC study are held by the PORTEC study group and the international TransPORTEC consortium. Data and tumor material from the Danish cohort are held by the coauthor of this article, G.Ø. Data and tumor material from the UMCG cohort are held by the coauthors of this article, H.W.N. and M.d.B., and from the LUMC by the coauthors N.H. and T.B. Requests for sharing of all data and material should be addressed to the corresponding author within 15 years of the date of publication of this article and include a scientific proposal. Depending on the specific research proposal, the TransPORTEC consortium (PORTEC-3 and TransPORTEC study) or the PORTEC study group (PORTEC-1, PORTEC-2 and MST) or coauthors G.Ø., H.W.N. and M.d.B., or N.H. and T.B., will determine when, for how long, for which specific purposes and under which conditions the requested data can be made available, subject to ethical consent. Requests for data access will be processed within a 3-month timeframe. TCGA-UCEC images, mutational status and clinical data are publicly available via the cBioPortal[65,66] for Cancer Genomics at https://www.cbioportal.org/study/clinicalData?id=ucec_tcga_pan_can_atlas_2018. The mRNA-seq data of the TCGA-UCEC were downloaded from http://firebrowse.org/?cohort=UCEC.

Code availability

The code base is available at https://github.com/AIRMEC/HECTOR.

References

65. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
66. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).
67. Wang, X. et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**, 102559 (2022).
68. Liu, Z. et al. Swin transformer: hierarchical vision transformer using shifted windows. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)* 9992–10002 (IEEE, 2021); https://ieeexplore.ieee.org/document/9710580
69. Aubreville, M. et al. MItosis DOmain Generalization Challenge 2022. *Zenodo* https://doi.org/10.5281/zenodo.6362337 (2022).
70. Thorsson, V. et al. The immune landscape of cancer. *Immunity* **48**, 812–830.e14 (2018).
71. Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.* https://doi.org/10.1200/PO.17.00011 (2017).
72. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

Acknowledgements

This work was supported by a translational research project grant from the Hanarth Foundation and the Swiss Federal Institutes of Technology (strategic focus area of personalized health and related technologies; grant no. 2021-367) and a grant from the Promedica Foundation (no. F-87701–41–01) during the conduct of the study. The PORTEC-1, PORTEC-2 and PORTEC-3 trials were funded by grants from the

Dutch Cancer Society (DCS; grant nos. CKTO 90–01, CKTO 2001–04 and CKTO 2006–04, respectively). We first and foremost thank the participants in these studies and those who donated a tumor sample for translational research. We are grateful to the international and local investigators and the data management teams who recruited and followed the women who participated in these studies, and to the many pathologists who collected samples for the PORTEC-1, PORTEC-2 and PORTEC-3 randomized trial biobanks, as well as the TransPORTEC Research Consortium for the establishment of the TransPORTEC study. We thank the investigators of the prospective MST cohort and G.Ø. and E.H. and investigators of the Danish cohort. We are indebted to T. Rutten and N. ter Haar, LUMC, for excellent technical support, slide collection and scanning. We thank L. Vermij, A. Leon-Castillo and E. Stelloo for the contribution to molecularly classifying the samples. We thank V. S. Hadnagy, University Hospital Zurich, for contributing to the annotation of the EC image dataset used to develop the mitosis detector for the present study. We also thank the Light Microscopy team of the Cell and Chemical Biology Department, LUMC, for the technical support and use of the 3DHISTECH P250 scanner, and the Netherlands Cancer Institute for use of their 3DHISTECH P1000 scanner. We acknowledge and thank the SHARK team, the computational cluster of the LUMC, for their technical support and the installation of the Nvidia RTX 8000 GPUs. We also thank K. Yost for her work and support with the figures. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Author contributions

S.V.-F., N.H., V.H.K. and T.B. conceived the study design. S.V.-F. designed the model and trained in its use. S.V.-F., S.A. and J.B.W. provided the coding, and implementation and technical support. S.V.-F. and N.H. acquired the data. S.V.-F., N.H., S.A., J.B.W., M.W.L., J.D., M.d.B., D.C., C.L.C., V.H.K. and T.B. analyzed and interpreted the data. S.V.-F. drafted the paper and the figures. S.V.-F., N.H., S.A., J.B.W., M.W.L., M.d.B., D.C., V.H.K. and T.B. substantially reviewed the paper. All authors critically reviewed the paper and the results and approved the final version.

Competing interests

S.V.-F., N.H., V.H.K. and T.B. are co-inventors on the patent application no. 23315438.4 related to the present study. N.H. declares having received research grants from the DCS and Varian (paid to the institution) unrelated to the present study. C.D.d.K. declares KWF and ZonMW grants unrelated to the project. A.L. received funded research unrelated to the present study from AZ, Clovis, GSK, MSD, Ability, Zentalis, Agenus, Lovance, Sanofi, Roche, OSEimmuno and BMS, is an advisory board member or consultant for AZ, Clovis, GSK, MSD, Merck Serono, Ability, Zentalis, Agenus and Blueprint, and received honoraria and compensation for expenses from AZ, Clovis and GSK. R.A.N. declared research grants unrelated to the present study to the institution from Elekta, Varian, Accuray and Sensius, and is an advisory board member of MSD. M.d.B. received grants from the DCS, the European Research Council, Health Holland, Mendus, BioNovion, Aduro Biotech, Vicinivax, Genmab and IMMIOS (all paid to the institute) unrelated to the present study, received nonfinancial support from BioNTech, Surflay Nanotec and Merck Sharp & Dohme, and is a stock option holder in Sairopa. D.C. is on an advisory board of MSD, received research funding unrelated to the project of HalioDx and Veracyte (to TransSCOT consortium), is a spouse of an Amgen employee, is affiliated to the Wellcome Centre for Human Genetics and National Institute for Health and Care Research (NIHR) Oxford Biomedical Research Centre (BRC), and received funding from Oxford NIHR Comprehensive BRC and a Cancer Research UK (CRUK) Advanced Clinician Scientist Fellowship (C26642/A27963). C.L.C. received grants from the DCS for the PORTEC-1,-2,-3,-4a, RAINBO trials and research grant for translational work on PORTEC unrelated to the present study, and has leadership roles in and is chair of GCIG Endometrial Cancer Committee. V.H.K. declared being an invited speaker for Sharing Progress in Cancer Care and Indica Labs, is on the advisory board of Takeda and sponsored research agreements with Roche and IAG, all unrelated to the present study. T.B. received grants unrelated to this work by the DCS. The other authors declare no competing interests.

Additional information

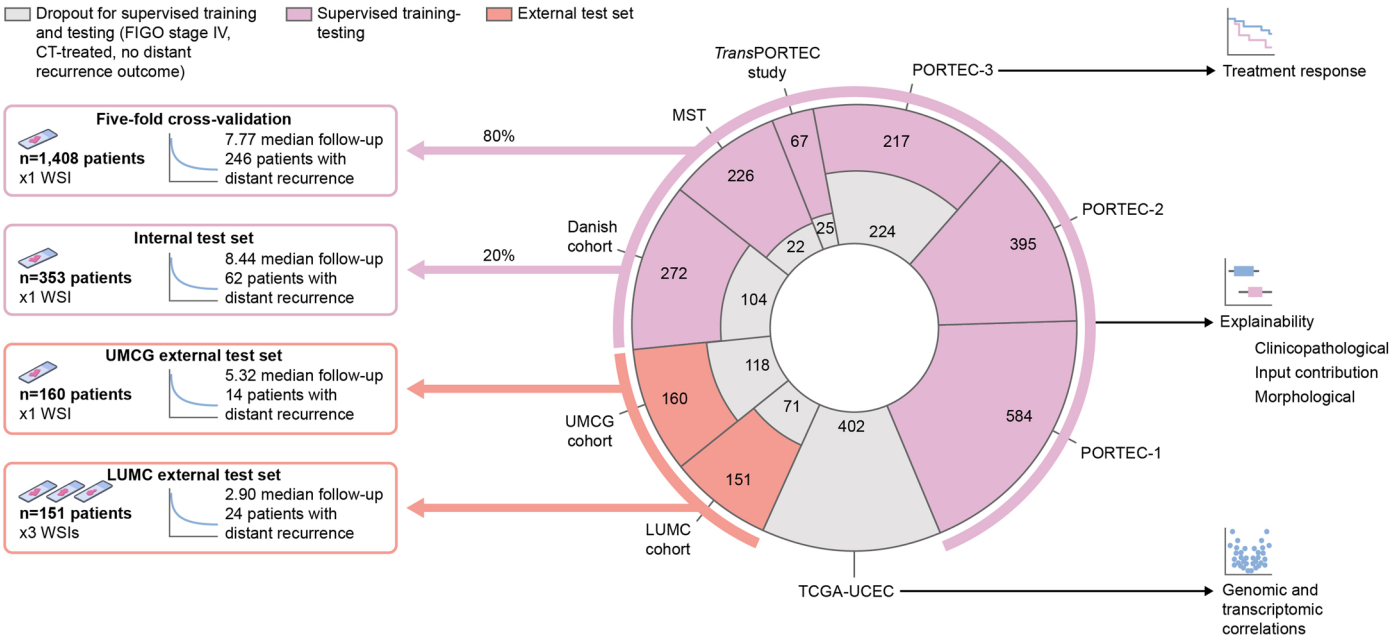
Extended data is available for this paper at https://doi.org/10.1038/s41591-024-02993-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-02993-w.

Correspondence and requests for materials should be addressed to Tjalling Bosse.

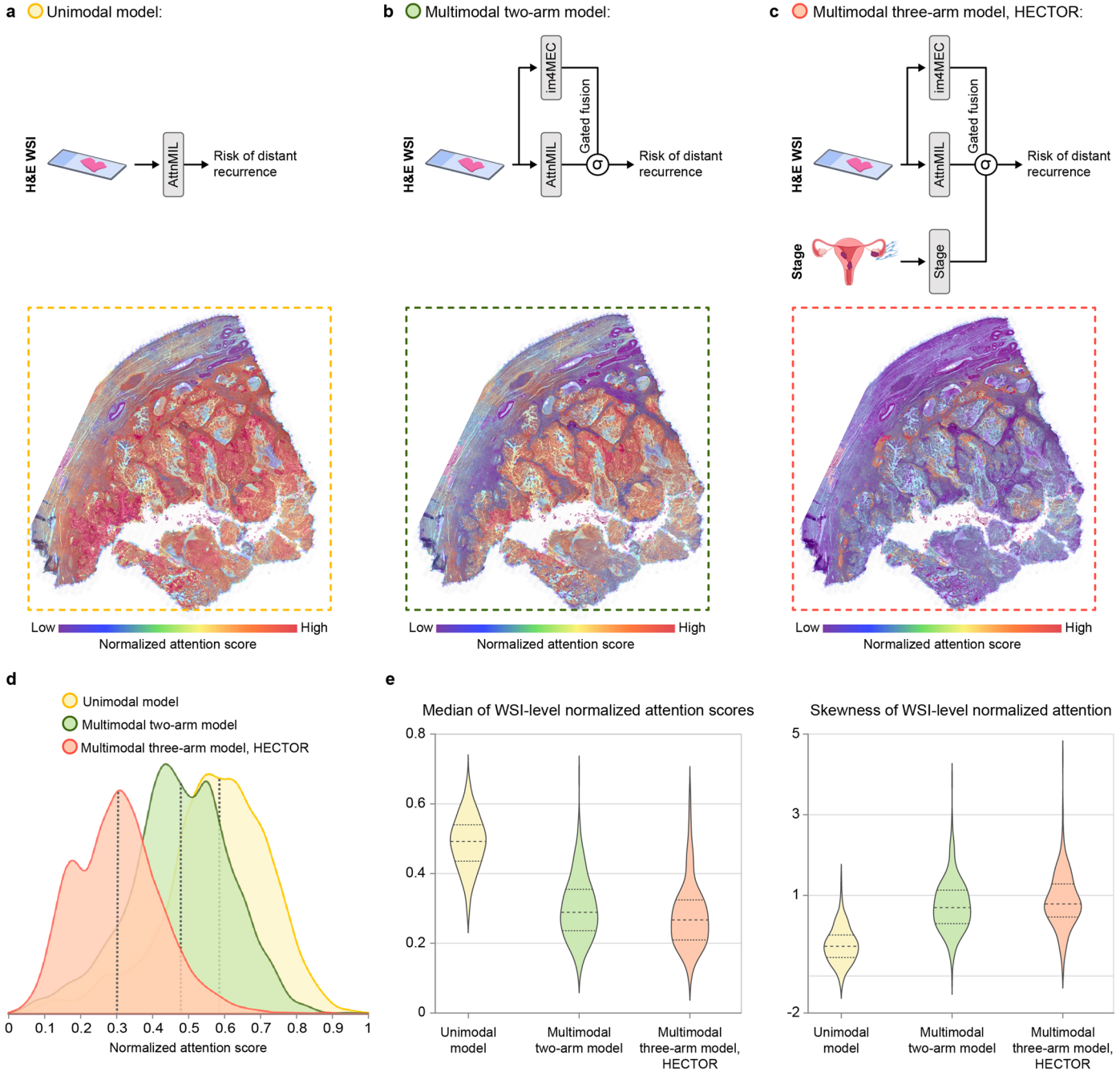
Peer review information *Nature Medicine* thanks Ming Lu, Amit Oza, Antonio Raffone and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Lorenzo Righetto and Ulrike Harjes, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-024-02993-w



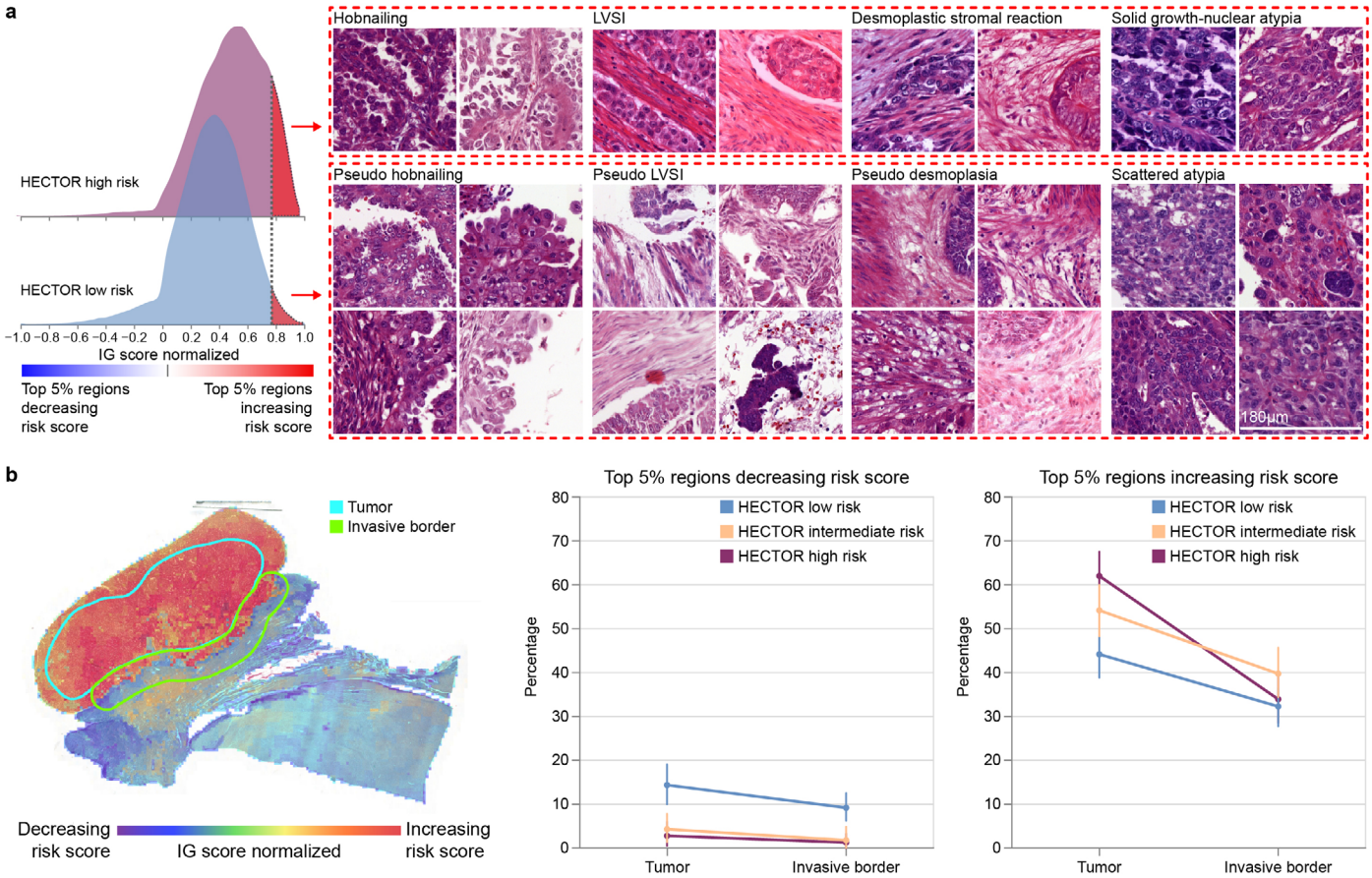
Extended Data Fig. 1 | Overview of the data split and downstream analyses performed in this study. One representative WSI per patient from an Formalin-Fixed Paraffin-Embedded (FFPE) block was included. 20% of cases meeting inclusion criteria were randomly held out for an internal test set ($n = 353$). The remaining 80% was used for five-cross validation ($n = 1,408$ patients). This training dataset was enriched with dropped WSIs of FIGO 2009 stage IV cases or those with missing outcome such as the TCGA-UCEC cohort[21] for training with self-supervised learning ($n = 1,862$). Two cohorts were held out as external test sets, the UMCG external test set ($n = 160$) and the LUMC external test set ($n = 151$). The LUMC external test set contains up to three FFPE blocks per case. More details for training and data split are provided in Methods. Altogether, including the two training steps and all downstream analyses, this comprehensive analysis comprised data of 2,751 tumors of women. CT, chemotherapy.

https://doi.org/10.1038/s41591-024-02993-w



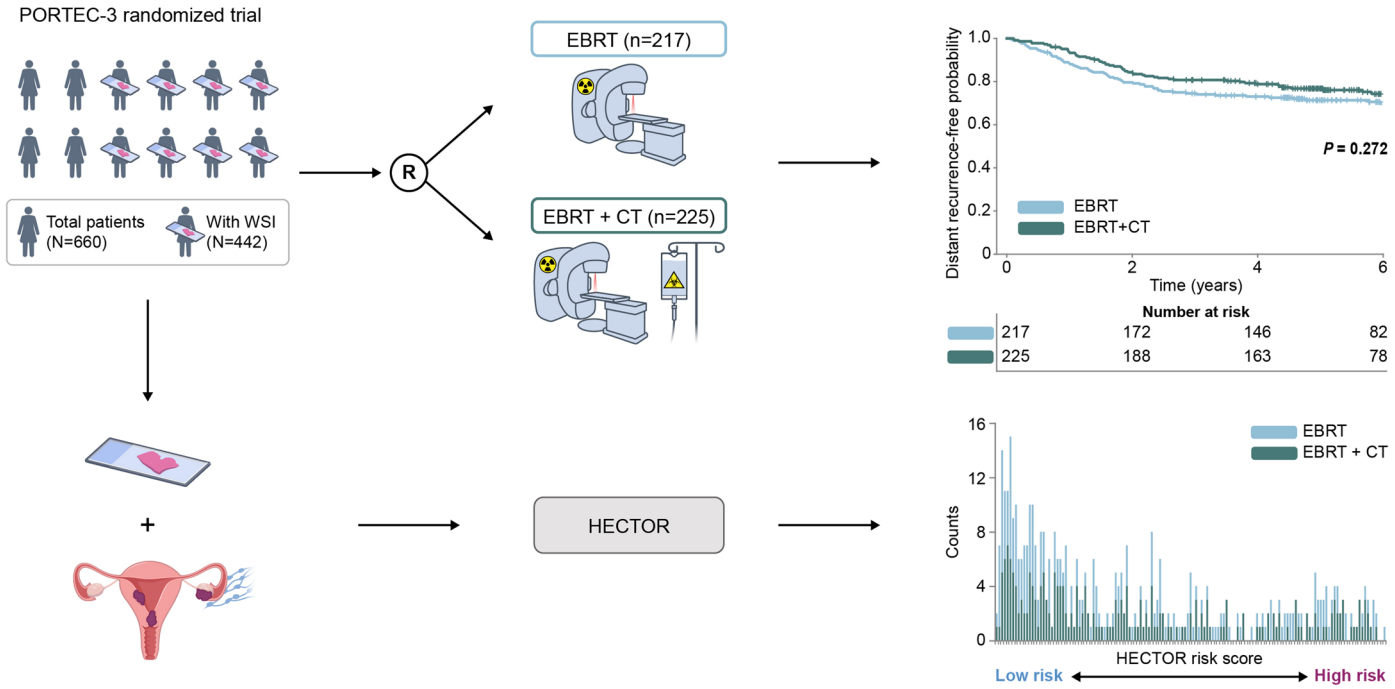
Extended Data Fig. 2 | Shifts of attention scores from unimodal to multimodal model. a, Model using only H&E WSI (unimodal) and a corresponding example of the normalized attention scores shown as overlaid on the H&E WSI as a heatmap where red is high attention score and blue low attention score. **b**, The two-arm model with H&E WSI and image-based molecular class predicted by im4MEC, and a corresponding example of the normalized attention scores shown as overlaid on the H&E WSI. **c**, The multimodal three-arm HECTOR model with H&E WSI, image-based molecular class, and stage, and a corresponding example of the normalized attention scores shown as overlaid on the H&E WSI. **d**, Density plot of the normalized attention scores of the heatmap shown in a,b,c for each model. **e**, Quantitative analysis of the distribution shift between the three models in the internal test set ($n = 353$ patients) using the WSI-level skewness and median of the normalized attention scores.

https://doi.org/10.1038/s41591-024-02993-w



Extended Data Fig. 3 | Morphological features increasing risk score in HECTOR high versus low risk group and quantitative spatial analysis. **a**, A representative selection of four patches for each morphological subtype (each selected from a different patient) increasing the risk score in the HECTOR low risk group as compared to the features increasing the risk score in the HECTOR high risk. Each patch is 180 × 180 µm. **b**, Spatial analysis of top 5% regions decreasing and increasing the risk score in all WSIs of the LUMC test set based on the manually annotated areas: tumor and invasive border. (left) An example showing the annotation of the tumor area and invasive border of one WSI and heatmap showing the contribution of the regions using the IG methods. (right) The relative contribution of these two annotated areas averaged by WSI shown for each HECTOR risk group. Data are presented as the mean values and standard deviation (n = 414 WSIs).

https://doi.org/10.1038/s41591-024-02993-w



Extended Data Fig. 4 | Overview of the PORTEC-3 randomized trial and analysis of treatment response prediction by HECTOR. In PORTEC-3, 660 evaluable patients were randomized (1:1) between adjuvant external beam radiotherapy (EBRT) alone and external beam radiotherapy in combination with concurrent and adjuvant chemotherapy (CT). For 442 patients whose WSI was available, HECTOR risk scores were inferred. HECTOR risk groups cutoffs were kept the same as the training set (Methods).

nature portfolio

Corresponding author(s): Dr. Tjalling Bosse

Last updated by author(s): Apr 4, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Scanning of images was conducted with the 3D Histech P250 and P1000 scanner at 40x magnification. Images were read and pre-processed with Openslide Python package (version 1.1.2), OpenCV (version 4.3.0.36), and Pillow (version 7.2.0). Annotations were done with QuPath (version 0.4.1).
Data analysis	The custom deep learning model (HECTOR) was developed and trained using Pytorch (version 1.8.1 for the self-supervised learning and version 1.10.0 otherwise). Integrated Gradient was implemented with Captum Python package (version 0.6.0); metrics such as the concordance-index with scikit-survival Python package (version 0.17.2); Cox Proportional Hazard models and Kaplan Meier's method with Lifelines Python package (version 0.27.1); Chi square tests with Scipy Python package (version 1.5.2); Boxplots visualizations with altair Python package (version 4.2.0); Linear regression with statsmodels Python package (version 0.13.5). Differentially expressed genes was performed with DESeq272 (version 1.40.1) and R version 4.3.0 (2023-04-21 ucrt). We made publicly available the code at https://github.com/AIRMEC/HECTOR.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The tumor material and datasets generated during or analyzed in this study are not publicly available due to restrictions by privacy laws. Data and tumor material from PORTEC-1, PORTEC-2, PORTEC-3, MST, the TransPORTEC study, are held by the PORTEC study group and the international TransPORTEC consortium. Data and tumor material from the Danish cohort are held by the coauthor of this article G.O. Data and tumor material from the UMCG cohort are held by the coauthors of this article H.N and M.B; LUMC by the co-authors N.H and T.B. Requests for sharing of all data and material should be addressed to the corresponding author within 15 years of the date of publication of this Article and include a scientific proposal. Depending on the specific research proposal, the TransPORTEC consortium (PORTEC-3 and TransPORTEC study) or the PORTEC study group (PORTEC-1, PORTEC-2, MST), or co-author G.O., H.N and M.B, or N.H and T.B, will determine when, for how long, for which specific purposes, and under which conditions the requested data can be made available, subject to ethical consent. Requests for data access will be processed within a 3-month timeframe. TCGA-UCEC images, mutational status and clinical data are publicly available via the cBioPortal65,66 for Cancer Genomics at https://www.cbioportal.org/study/clinicalData?id=ucec_tcga_pan_can_atlas_2018. mRNA-seq data of the TCGA-UCEC were downloaded from http://firebrowse.org/?cohort=UCEC.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We do not report on sex and gender. The findings of this study relate to endometrial cancer and apply to biologically female individuals. We reported in the methods "We included study participants of the female sex, independent of gender identity."
Reporting on race, ethnicity, or other socially relevant groupings	We do not report on race, ethnicity or socially relevant groupings, nor have data related to this.
Population characteristics	All population characteristics of any cohort used are described in the Supplemental Figure 2 and Supplemental Tables 1,2,14, in which we report the following characteristics : age, type of tumor, tumor stage, molecular characteristic of the tumor (POLE mutation, Mismatch repair deficient, p53 abnormality), adjuvant treatment received, and median follow-up time.
Recruitment	Cohorts used are the three PORTEC (1/2/3) randomized trials in which recruitment followed the design protocol of the clinical trials as reported in their original publication as well as in the Methods section. The PORTEC-1 trial recruited 714 women with early-stage intermediate risk EC from 1990 to 1997, and after primary surgery, randomly assigned to pelvic external beam radiotherapy or no adjuvant treatment. The PORTEC-2 trial randomized 427 women with early-stage high-intermediate risk EC between 2000 to 2006 to external beam radiotherapy or vaginal brachytherapy. The PORTEC-3 randomized trial included 660 women with stage I-III high risk EC from 2006 and 2013, and randomly allocated them to pelvic external beam radiotherapy alone or external beam radiotherapy combined with concurrent and adjuvant chemotherapy. The retrospective TransPORTEC study included 116 high-risk EC tumors from international patients using the same inclusion criteria as the PORTEC-3 from five institutions (Leiden University Medical Center, The Netherlands; University Medical Center Groningen, The Netherlands; University College London, United Kingdom; St Mary's Hospital, Manchester, United Kingdom; and Institute Gustave Roussy, Villejuif, France). The prospective cohort of Medisch Spectrum Twente (MST) included 257 patients with stage I-III high risk EC, with the same inclusion criteria as the PORTEC-3, who were treated between 1987 and 2015 at MST, Enschede in the Netherlands. The Danish cohort consisted of 451 high-grade EC of patients who were prospectively registered in the Danish gynecological cancer database. The Leiden cohort is a retrospectively collected population-based cohort of 222 patients diagnosed and treated at the Leiden University Medical Center between 2012 and 2021. This study excluded patients if tumor data or material was missing such as an image of the tumor, or missing follow-up data.
Ethics oversight	The PORTEC-1, PORTEC-2 (NCT00376844) PORTEC-3 (NCT00411138) study protocols were approved by the Medical Ethical Committee Leiden – Den Haag – Delft and the medical ethics committees at participating centers. Studies were conducted in accordance with the principles of the Declaration of Helsinki. Ethical permissions for the retrospective use of the clinical trials and retrospective cohorts (TransPORTEC study, MST), were obtained by the Medical Ethical Committee Leiden – Den Haag – Delft (numbers B21.065 and B21.011) as well as the LUMC Cohort (nWMO-D4-2023-002), and the Danish Cohort by the Center for Regional Udvikling – De Videnskabsetiske Komiteer (H-16025909). All study participants of the clinical trials provided informed consent. The ethical boards have provided a waiver for informed consent for the other studies. For the UMCG cohort, the medical ethical committee granted permission of the use of the data and provided a waiver for informed consent due to the observational nature of the study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Deep learning-based methods benefit from the largest possible datasets for a better training and relatively large test sets. No minimum sample size was calculated in our case as we possessed a sufficiently large cohort of 2,072 patients for training and testing the model. As for the split between training, validation and testing, we followed standard split by sampling 20% for internal test set, and held out two external test sets. As a result, the training dataset had 1,408 patients (with 246 clinical events), the internal test 353 patients (with 62 clinical events) and the first external test set 151 patients (with 24 clinical events) and the second test set contained 160 patients (with 14 events), which is sufficiently large for having enough events in each set and evaluating the accuracy of the model or using Kaplan-Meir's methods.
Data exclusions	Exclusion criteria were pre-established before development and testing of the models based on 1) requirements for training the model and 2) clinical knowledge of the disease based on previous publication. First exclusion criteria are the absence of tumor material or/and tumor data; artifacts in the digitized tumor slide such as out of focus areas. In our specific study, where the supervised Deep learning-based developed model is trained to predict distant recurrence-free probability, patients that already had distant recurrence at diagnosis (that is FIGO stage IV) and then the ones who received adjuvant chemotherapy were excluded from training-testing. This is because adjuvant chemotherapy likely reduces this risk as shown in previous clinical publications. Moreover in our dataset in which treatment is known for any patient, the far majority of patients treated with chemotherapy comes from the PORTEC-3 randomized trial and chemotherapy is not standard of care in the Netherlands, and rarely given. Therefore, any bias which would exclude a specific type of tumor is very unlikely, as a matter of fact, all the patients included in this study cover all tumor types, all stages I to III, and all molecular types. These specificities were all reported in the manuscript in the Methods as well as in the supplemental data with, for instance, a flow chart indicating exact number of patients being excluded.
Replication	We used 5 fold-cross validation. Furthermore we showed similar performance in 5 fold-cross validation, the internal and external test sets.
Randomization	The external test sets were blindly and randomly held-out. The internal test set was randomly sampled from the entire training set. Similarly, the 5 fold-cross validation split was performed randomly.
Blinding	Our manuscript describes the development and performance of a deep learning model. The developed model was tested one time after development, in one internal and one external test set and performance was reported. The tumor slide images of internal and external test sets were therefore completely unseen by the model, and no optimization on the internal nor external test set was performed. Furthermore, the internal and external test sets were blindly and randomly held-out. Specifically, tumor characteristics in each test set were not analyzed before testing the model performance in these test sets.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	The PORTEC-1 (there is no registration as clinical trial registration did not exist in the 90s. The PORTEC-1 study was supported by the grant CKVO 90–01), PORTEC-2 (NCT00376844) PORTEC-3 (NCT00411138)

Study protocol	For the clinical trials that were included in this study, that is the PORTEC-1, PORTEC-2, PORTEC-3, we can provide the protocols as they are not available online. The PORTEC-1 protocol is in Dutch and the PORTEC-2 and PORTEC-3 in english.
Data collection	The PORTEC-1 trial recruited 714 women with early-stage intermediate risk EC from 1990 to 1997, and after primary surgery, randomly assigned to pelvic external beam radiotherapy or no adjuvant treatment. 19 departments of radiation oncology in the Netherlands took part. The patients were evaluated and treated by their local radiation oncologist. Central blocked randomisation by telephone was done at the Daniel den Hoed Cancer Centre trial office. The PORTEC-2 trial randomized 427 women with early-stage high-intermediate risk EC between 2000 to 2006 to external beam radiotherapy or vaginal brachytherapy. 19 Dutch radiation oncology departments participated. Patient details and answers about eligibility questions were entered by the data managers of the participating centres. Eligibility check and randomisation were done on the basis of the original pathology diagnosis. Central review of the pathology was done to assess histological type, stage, and grade. The PORTEC-3 randomized trial included 660 women with stage I-III high risk EC from 2006 and 2013, and randomly allocated them to pelvic external beam radiotherapy alone or external beam radiotherapy combined with concurrent and adjuvant chemotherapy. 103 centres (oncology centres, university hospitals, regional hospitals, or radiation oncology centres with referrals from regional hospitals) from six clinical trial groups which collaborated in the Gynaecological Cancer Intergroup. Participating groups were the National Cancer Research Institute (NCRI; UK), Australia and New Zealand Gynaecologic Oncology Group (ANZGOG; Australia and New Zealand), Mario Negri Gynaecologic Oncology Group (MaNGO; Italy), Canadian Cancer Trials Group (CCTG; Canada), and Fedegyn (France). Central pathology review was done by reference gynaecopathologists (as appointed by each participating group before the start of the trial) to determine final eligibility. The slides and blocks were sent to each participating group's central review pathologists at one gynaecological pathology review site (in France and Italy), two sites (in the UK and the Netherlands), or five to six sites (in Australia and New Zealand, and Canada), with the result of the review confirming the patient's eligibility for the trial being sent to the local investigators within 1 week.
Outcomes	This is a deep learning-based study. We predefined the primary outcomes as the performance of the model measured by the concordance-index between the predicted risk score of distant recurrence and the true time to distant recurrence. Secondary outcomes were the survival area under the curve (AUC) and the Brier scores. Additionally, we analyzed the distant recurrence-free probabilities and stratification with the Kaplan Meier's method, the log-rank test and Cox regression analyses.

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-04368-5 OPEN

Interpretable dimensionality reduction of single cell transcriptome data with deep generative models

Jiarui Ding [1,2,3,4], Anne Condon [1] & Sohrab P. Shah [1,2,3,5]

Single-cell RNA-sequencing has great potential to discover cell types, identify cell states, trace development lineages, and reconstruct the spatial organization of cells. However, dimension reduction to interpret structure in single-cell sequencing data remains a challenge. Existing algorithms are either not able to uncover the clustering structures in the data or lose global information such as groups of clusters that are close to each other. We present a robust statistical model, scvis, to capture and visualize the low-dimensional structures in single-cell gene expression data. Simulation results demonstrate that low-dimensional representations learned by scvis preserve both the local and global neighbor structures in the data. In addition, scvis is robust to the number of data points and learns a probabilistic parametric mapping function to add new data points to an existing embedding. We then use scvis to analyze four single-cell RNA-sequencing datasets, exemplifying interpretable two-dimensional representations of the high-dimensional single-cell RNA-sequencing data.

[1] Department of Computer Science, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. [2] Department of Molecular Oncology, BC Cancer Agency, Vancouver, BC V5Z 1L3, Canada. [3] Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC V6T 2B5, Canada. [4] Present address: Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [5] Present address: Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, USA. Correspondence and requests for materials should be addressed to J.D. (email: jding@broadinstitute.org) or to S.P.S. (email: sshah@bccrc.ca)

Categorizing cell types comprising a specific organ or disease tissue is critical for comprehensive study of tissue development and function[1]. For example, in cancer, identifying constituent cell types in the tumor microenvironment together with malignant cell populations will improve understanding of cancer initialization, progression, and treatment response[2, 3]. Technical developments have made it possible to measure the DNA and/or RNA molecules in single cells by single-cell sequencing[4–15] or protein content by flow or mass cytometry[16, 17]. The data generated by these technologies enable us to quantify cell types, identify cell states, trace development lineages, and reconstruct the spatial organization of cells[18, 19]. An unsolved challenge is to develop robust computational methods to analyze large-scale single-cell data measuring the expression of dozens of protein markers to all the mRNA expression in tens of thousands to millions of cells in order to distill single-cell biology[20–23].

Single-cell datasets are typically high dimensional in large numbers of measured cells. For example, single-cell RNA-sequencing (scRNA-seq)[19, 24–26] can theoretically measure the expression of all the genes in tens of thousands of cells in a single experiment[9, 10, 14, 15]. For analysis, dimensionality reduction projecting high-dimensional data into low-dimensional space (typically two or three dimensions) to visualize the cluster structures[27–29] and development trajectories[30–33] is commonly used. Linear projection methods such as principal component analysis (PCA) typically cannot represent the complex structures of single-cell data in low dimensional spaces. Nonlinear dimension reduction, such as the *t*-distributed stochastic neighbor embedding algorithm (t-SNE)[34–39], has shown reasonable results for many applications and has been widely used in single-cell data processing[1, 40, 41]. However, t-SNE has several limitations[42]. First, unlike PCA, it is a non-parametric method that does not learn a parametric mapping. Therefore, it is not natural to add new data to an existing t-SNE embedding. Instead, we typically need to combine all the data together and rerun t-SNE. Second, as a non-parametric method, the algorithm is sensitive to hyperparameter settings. Third, t-SNE is not scalable to large datasets because it has a time complexity of $O(N^2D)$ and space complexity of $O(N^2)$, where N is the number of cells and D is the number of expressed genes in the case of scRNA-seq data. Fourth, t-SNE only outputs the low-dimensional coordinates but without any uncertainties of the embedding. Finally, t-SNE typically preserves the local clustering structures very well given proper hyperparameters, but more global structures such as a group of subclusters that form a big cluster are missed in the low-dimensional embedding.

In this paper, we introduce a robust latent variable model, scvis, to capture underlying low-dimensional structures in scRNA-seq data. As a probabilistic generative model, our method learns a parametric mapping from the high-dimensional space to a low-dimensional embedding. Therefore, new data points can be directly added to an existing embedding by the mapping function. Moreover, scvis estimates the uncertainty of mapping a high-dimensional point to a low-dimensional space that adds rich capacity to interpret results. We show that scvis has superior distance preserving properties in its low-dimensional projections leading to robust identification of cell types in the presence of noise or ambiguous measurements. We extensively tested our method on simulated data and several scRNA-seq datasets in both normal and malignant tissues to demonstrate the robustness of our method.

Results

Modeling and visualizing scRNA-seq data. Although scRNA-seq datasets have high dimensionality, their intrinsic dimensionalities are typically much lower. For example, factors such as cell type and patient origin explain much of the variation in a study of metastatic melanoma[3]. We therefore assume that for a high-dimensional scRNA-seq dataset $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N$ with N cells, where $\mathbf{x}_n$ is the expression vector of cell n, the $\mathbf{x}_n$ distribution is governed by a latent low-dimensional random vector $\mathbf{z}_n$ (Fig. 1a). For visualization purposes, the dimensionality d of $\mathbf{z}_n$ is typically two or three. We assume that $\mathbf{z}_n$ is distributed according to a prior, with the joint distribution of the whole model as $p(\mathbf{z}_n \mid \boldsymbol{\theta})p(\mathbf{x}_n \mid \mathbf{z}_n, \boldsymbol{\theta})$. For simplicity, we can choose a factorized standard normal distribution for the prior $p(\mathbf{z}_n \mid \boldsymbol{\theta}) = \prod_{i=1}^d \mathcal{N}\left(z_{n,i}|0, \mathbf{I}\right)$. The distribution $p(\mathbf{x}_n|\boldsymbol{\theta}) = \int p(\mathbf{z}_n|\boldsymbol{\theta})p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\theta})\mathrm{d}\mathbf{z}_n$ can be a complex multimodal high-dimensional distribution. To represent complex high-dimensional distributions, we assume that $p(\mathbf{x}_n \mid \mathbf{z}_n, \boldsymbol{\theta})$ is a location-scale family distribution with location parameter $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_n)$ and scale parameter $\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{z}_n)$; both are functions of $\mathbf{z}_n$ parameterized by a neural network with parameter $\boldsymbol{\theta}$. The inference problem is to compute the posterior distribution $p(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\theta})$, which is however intractable to compute. We therefore use a variational distribution $q(\mathbf{z}_n \mid \mathbf{x}_n, \phi)$ to approximate the posterior (Fig. 1b). Here $q(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\phi})$ is a multivariate normal distribution with mean $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}_n)$ and standard deviation $\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x}_n)$. Both parameters are (continuous) functions of $\mathbf{x}_n$ parameterized by a neural network with parameter $\boldsymbol{\phi}$. To model the data distribution well (with a high likelihood of $\int p(\mathbf{z}_n|\boldsymbol{\theta})p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\theta})\mathrm{d}\mathbf{z}_n$), the model tends to assign similar posterior distributions $p(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\theta})$ to cells with similar expression profiles. To explicitly encourage cells with similar expression profiles to be proximal (and those with dissimilar profiles to be distal) in the latent space, we add the t-SNE objective function on the latent $\mathbf{z}$ distribution as a constraint. More details about the model and the inference algorithms are presented in the Methods section. The scvis model is implemented in Python using Tensorflow[43] with a command-line interface and is freely available from https://bitbucket.org/jerry00/scvis-dev.

Single-cell datasets. We analyzed four scRNA-seq datasets in this study[1, 3, 9, 44]. Data were mostly downloaded from the single-cell portal[45]. Two of these datasets were originally used to study intratumor heterogeneity and the tumor microenvironment in metastatic melanoma[3] and oligodendroglioma[44], respectively. One dataset was used to categorize the mouse bipolar cell populations of the retina[1], and one dataset was used to categorize all cell types in the mouse retina[9]. For all the scRNA-seq datasets, we used PCA (as a noise-reduction preprocessing step[1, 19]) to project the cells into a 100-dimensional space and used the projected coordinates in the 100-dimensional spaces as inputs to scvis. We also used two mass cytometry (CyTOF) datasets consisting of bone marrow mononuclear cells from two healthy adult donors H1 and H2[17]. For CyTOF data, since their dimensionality (32) is relatively low, we directly used these data as inputs to scvis.

Experimental setting and implementation. The variational approximation neural network has three hidden layers (l_1, l_2, and l_3) with 128, 64, and 32 hidden units each, and the model neural network has five hidden layers $\left(l_1', l_2', l_3', l_4', \text{and } l_5'\right)$ with 32, 32, 32, 64, and 128 units each. We use the exponential linear unit activation function as it has been shown to speed up the convergence of optimization[46] and the Adam stochastic optimization algorithm with a learning rate of 0.01[47]. Details about the influence of these hyperparameters on results are presented in the Methods section. The time complexity to compute the t-SNE loss is quadratic in terms of the number of data points. Consequently, we use mini-batch optimization and set the mini-batch size to 512 (cells). We expect that a large batch of data could be better in

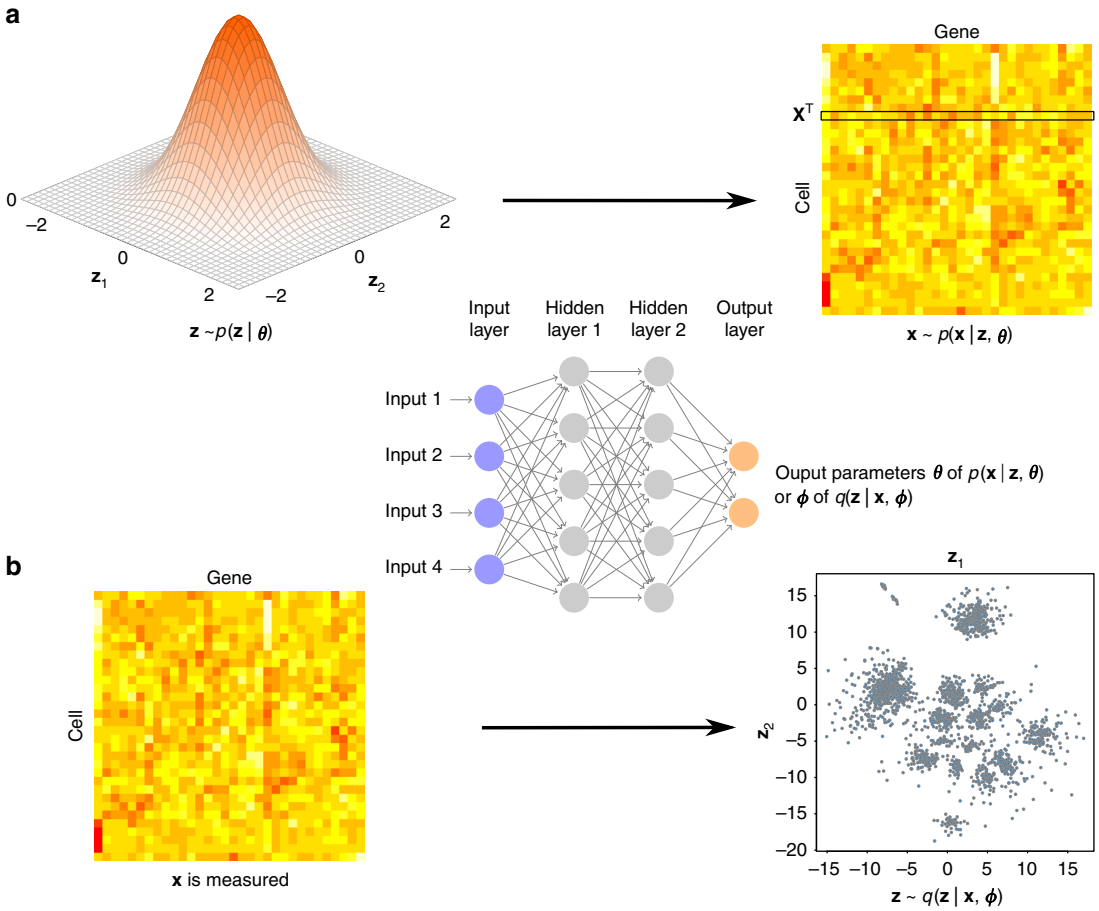


Fig. 1 Overview of the scvis method. **a** scvis model assumptions: given a low-dimensional point drawn from a simple distribution, e.g., a two-dimensional standard normal distribution, a high-dimensional gene expression vector of a cell can be generated by drawing a sample from the distribution $p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\theta})$. The heatmap represents a cell–gene expression matrix, where each row is a cell and each column is a gene. Color encodes the expression levels of genes in cells. The data-point-specific parameters $\boldsymbol{\theta}$ are determined by a model neural network. The model neural network (a feedforward neural network) consists of an input layer, several hidden layers, and an output layer. The output layer outputs the parameters $\boldsymbol{\theta}$ of $p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\theta})$. **b** scvis inference: given a high-dimensional gene expression vector of a cell (a row of the heatmap), scvis obtains its low-dimensional representation by sampling from the conditional distribution $q(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\phi})$. The data-point-specific parameters $\boldsymbol{\phi}$ are determined by a variational inference neural network. The inference neural network is also a feedforward neural network and its output layer outputs the parameters $\boldsymbol{\phi}$ of $q(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\phi})$. Again, the heatmap represents a cell–gene expression matrix. The scatter plot shows samples drawn from the variational posterior distributions $q(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\phi})$

estimating the high-dimensional data manifold, however we found that 512 cells work accurately and efficiently in practice. We run the Adam stochastic gradient descent algorithm for 500 epochs for each dataset with at least 3000 iterations by default. For large datasets, running 500 epochs is computationally expensive, we therefore run the Adam algorithm for a maximum of 30,000 iteration or two epochs (which ever is larger). We use an L2 regularizer of 0.001 on the weights of the neural networks to prevent overfitting.

Benchmarking scvis against t-SNE on simulated data. To demonstrate that scvis can robustly learn a low-dimensional representation of the input data, we first simulated data in a two-dimensional space (for easy visualization) as in Fig. 2a. The big cluster on the left consisted of 1000 points and the five small clusters on the right each had 200 points. The five small clusters were very close to each other and could roughly be considered as a single big cluster. There were 200 uniformly distributed outliers around these six clusters. For each two-dimensional data point with coordinates (x, y), we then mapped it into a nine-dimensional space by the transformation $(x+y, x-y, xy, x^2, y^2, x^2y, xy^2, x^3, y^3)$. Each of the nine features was then divided by its corresponding maximum absolute value.

Although t-SNE (with default parameter setting, we used the efficient Barnes-Hut t-SNE[34] R wrapper package[48]) uncovered the six clusters in this dataset, it was still challenging to infer the overall layout of the six clusters (Fig. 2b). t-SNE by design preserves local structure of the high-dimensional data, but the "global" structure is not reliable. Moreover, for the uniformly distributed outliers, t-SNE put them into several compact clusters, which were adjacent to other genuine clusters.

The scvis results, on the other hand, better preserved the overall structure of the original data (Fig. 2c): (1) The five small clusters were on one side, and the big cluster was on the other side. The relative positions of the clusters were also preserved. (2) Outliers were scattered around the genuine clusters as in the original data. In addition, as a probabilistic generative model, scvis not only learned a low-dimensional representation of the input data but also provided a way to quantify the uncertainty of the low-dimensional mapping of each input data point by its log-likelihood. We colored the low-dimensional embedding of each data point by its log-likelihood (Fig. 2d). We can see that generally scvis put most of its modeling power to model the five

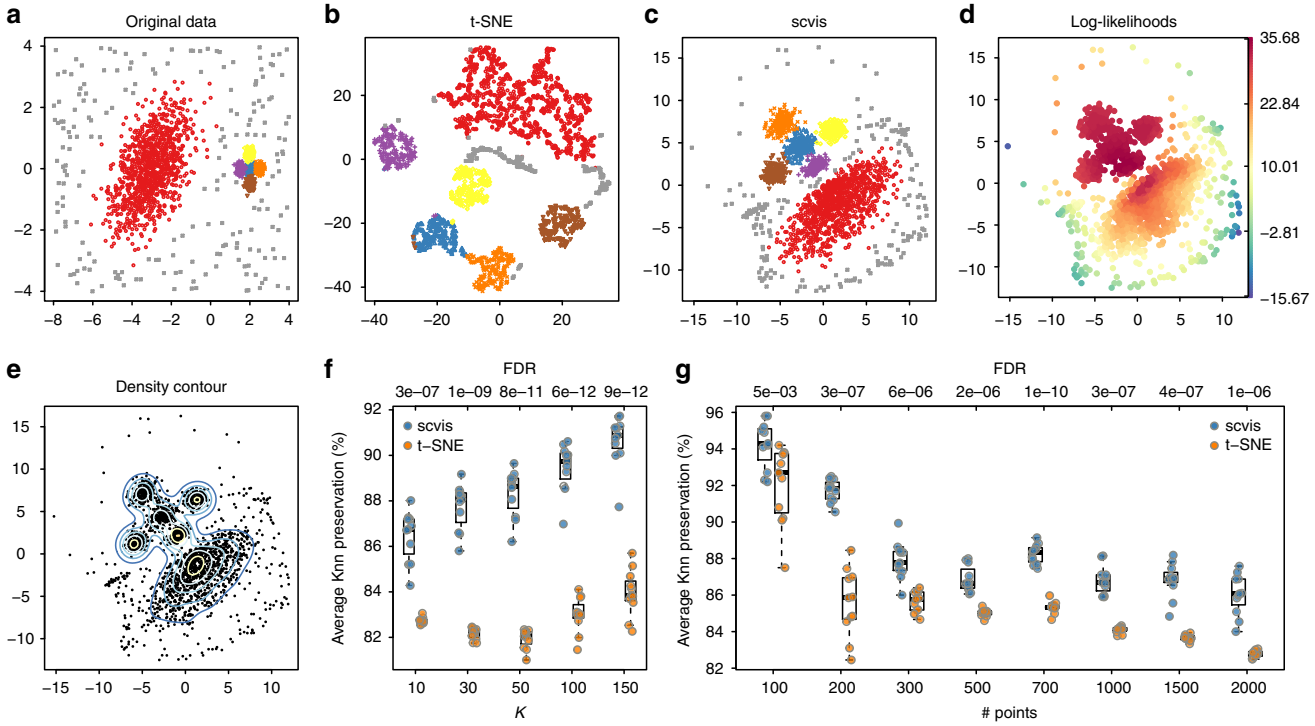


Fig. 2 Benchmarking scvis against t-SNE on synthetic data. **a** The original 2200 two-dimensional synthetic data points (points are colored by cluster labels. The randomly distributed outliers are also colored in a distinct color, same for **b**, **c**), **b** t-SNE results on the transformed nine-dimensional dataset with default perplexity parameter of 30, **c** scvis results, **d** coloring points based on their log-likelihoods from scvis, **e** the kernel density estimates of the scvis results, where the density contours are represented by lines, **f** average *K*-nearest neighbor preservations across ten runs for different *K*s, and **g** the average *K*-nearest neighbor preservations ($K = 10$) for different numbers of subsampled data points from ten repeated runs. The numbers at the top are the adjusted *p*-values (FDR, one-sided Welch's *t*-test) comparing the average *K*nn preservations from scvis with those from t-SNE. Boxplots in **f**, **g** denote the medians and the interquartile ranges (IQRs). The whiskers of a boxplot are the lowest datum still within 1.5 IQR of the lower quartile and the highest datum still within 1.5 IQR of the upper quartile

compact clusters, while the outliers far from the five compact clusters tended to have lower log-likelihoods. Thus, by combining the log-likelihoods and the low-dimensional density information (Fig. 2e), we can better interpret the structure in the original data and uncertainty over the projection.

The low-dimensional representation may change for different runs because the scvis objective function can have different local maxima. To test the stability of the low-dimensional representations, we ran scvis ten times. Generally, the two-dimensional representations from the ten runs (Supplementary Fig. 1a–j) showed similar patterns as in Fig. 2c. As a comparison, we also ran t-SNE ten times, and the results (Supplementary Fig. 1k–t) showed that the layouts of the clusters were less preserved, e.g., the relative positions of the clusters changed from run to run. To quantitatively compare scvis and t-SNE results, we computed the average *K*-nearest neighbor (*K*nn) preservations across runs for $K \in \{10, 30, 50, 100, 150\}$. Specifically, for the low-dimensional representation from each run, we constructed *K*nn graphs for different *K*s. We then computed the *K*nn graph from the high-dimensional data for a specific *K*. Finally, we compared the average overlap of the *K*nn graphs from the low-dimensional representations with the *K*nn graph from the high-dimensional data for a specific *K*. For scvis, the median *K*nn preservations monotonically increased from 86.7% for $K = 10$, to 90.9% for $K = 150$ (Fig. 2f). For t-SNE, the median *K*nn preservations first decreased from 82.7% for $K = 10$ to 82.1% for $K = 50$ (consistent with t-SNE preserving local structures) and then increased to 84.0% for $K = 150$. Thus scvis preserved *K*nn more effectively than t-SNE.

To test how scvis performs on smaller datasets, we subsampled the nine-dimensional synthetic dataset. Specifically, we subsampled 100, 200, 300, 500, 700, 1000, 1500, and 2000 points from the original dataset and ran scvis 11 times on each subsampled dataset. We then computed the *K*nn preservations ($K = 10$) and found that the *K*nn preservations from the scvis results were significantly higher than those from t-SNE results (false discovery rate (FDR) <0.01 for all the subsampled datasets, one-sided Welch's *t*-test, Fig. 2g). scvis performs very well on all the subsampled datasets (Supplementary Fig. 2a–h). Even with just 100 data points, the two-dimensional representation (Supplementary Fig. 2a) preserved much of the structure in the data. The log-likelihoods estimated from the subsampled data also recapitulated the log-likelihoods from the original 2200 data points (Supplementary Fig. 3a–h). The t-SNE results on the subsampled datasets (Supplementary Fig. 2i–p) generally revealed the clustering structures. However, the relative positions of the five clusters and the big cluster were largely inaccurate.

To test the performance of scvis when adding new data to an existing embedding, we increased by tenfold the number of points in each cluster and the number of outliers (for a total of 22,000 points) using a different random seed. The embedding (Fig. 3a, b) was very similar to that of the 2200 training data points in Fig. 2c, d. We trained *K*nn classifiers on the embedding of the 2200 training data for $K \in \{5, 9, 17, 33, 65\}$ and used the trained classifiers to classify the embedding of the 22,000 points, repeating 11 times. Median accuracy (the proportion of points correctly assigned to their corresponding clusters) was 98.1% for $K = 5$ and 94.8% for $K = 65$. The performance decreased mainly

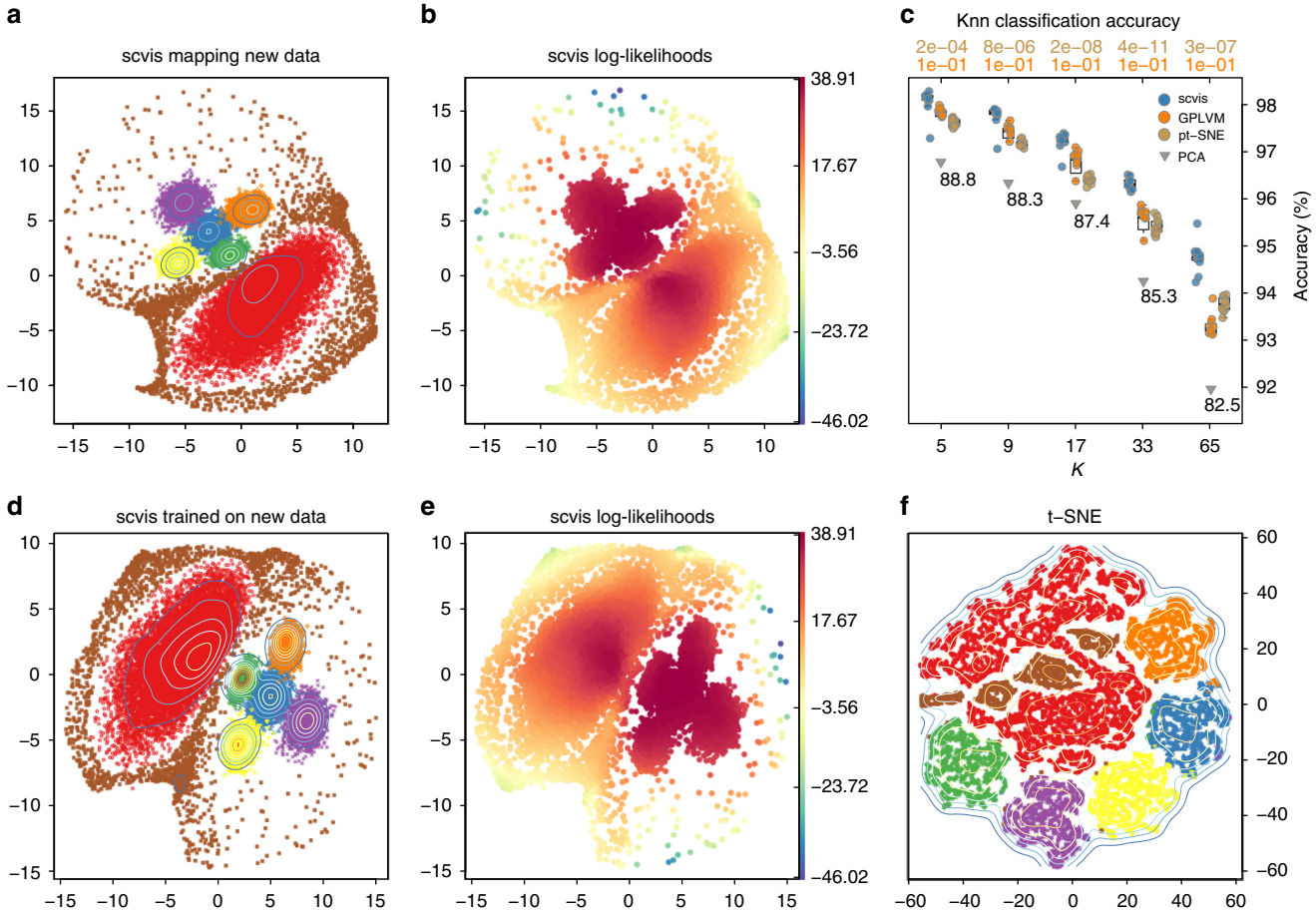


Fig. 3 Benchmarking scvis against GPLVM, parametric t-SNE, and PCA to embed 22,000 synthetic out-of-sample data. **a** scvis mapping 22,000 new data points based on the learned probabilistic mapping function from the 2200 training data points, **b** the estimated log-likelihoods, and **c** the average *K*-nearest neighbor classification accuracies for different *K*s across 11 runs, the classifiers were trained on the 11 embeddings from the 2200 points. The numbers at the top are the FDR (one-sided Mann–Whitney *U*-test) comparing the *K*-nearest neighbor classification accuracy from scvis with those from GPLVM (orange, bottom) and those from parametric t-SNE (golden, top). Notice that, for GPLVM, two runs produced bad results and were not plotted in the figure. Boxplots denote the medians and the interquartile ranges (IQR). The whiskers of a boxplot are the lowest datum still within 1.5 IQR of the lower quartile and the highest datum still within 1.5 IQR of the upper quartile. **d** scvis results on the larger dataset with the same perplexity parameter as used in Fig. 2; **e** scvis log-likelihoods on the larger dataset; and **f** t-SNE results on the larger dataset

because, for a larger *K*, the outliers were wrongly assigned to the six genuine clusters.

We then benchmarked scvis against Gaussian process latent variate model[49] (GPLVM, implemented in the GPy[50] package), parametric t-SNE[51] (pt-SNE), and PCA on embedding the 22,000 out-of-sample data points. We used the 11 scvis models trained on the small nine-dimensional synthetic dataset with 2200 data points to embed the larger nine-dimensional synthetic data with 22,000 data points. Similarly, we trained 11 GPLVM models and pt-SNE models on the small nine-dimensional synthetic dataset and applied these models to the bigger synthetic dataset. To compare the abilities of the trained models to embed unseen data, we trained *K*nn classifiers on the two-dimensional representations (of the small 2200 data points) outputted from different algorithms. These *K*nn classifiers were used to classify the two-dimensional coordinates of the 22,000 data points outputted from different algorithms. scvis was significantly better than GPLVM and pt-SNE for different Ks (Fig. 3c, two runs of GPLVM produced bad results and were not plotted in the figure, FDR < 0.05, one-sided Mann–Whitney *U*-test). For PCA, because the model is unique for a given dataset, we generated unique two-dimensional coordinates for the 22,000 out-of-sample data points. The *K*nn classifiers trained on the PCA coordinates were worse than those from scvis, GPLVM, and pt-SNE in terms of the mean classification accuracies for different Ks.

As a non-parametric dimension reduction method, t-SNE was sensitive to hyperparameter setting, especially the perplexity parameter (the effective number of neighbors, see the Methods section for details). The optimal perplexity parameter increased as the total number of data points increased. In contrast, as we adopted mini-batch for training scvis by subsampling, e.g., 512 cells each time, scvis was less sensitive to the perplexity parameter as we increase the total number of training data points because the number of cell is fixed at 512 at each training step. Therefore, scvis performed well on approximately an order of magnitude larger dataset (Fig. 3d, e), without changing the perplexity parameter for scvis. For this larger dataset, the t-SNE results (Fig. 3f) were difficult to interpret without the ground-truth cluster information, because it was already difficult to see how many clusters in this dataset, not to mention to uncover the overall structure of the data. Although by increasing the perplexity parameter, the performance of t-SNE became better (Supplementary Fig. 4), the outliers still formed distinct clusters, and it remains difficult to set this parameter in practice.

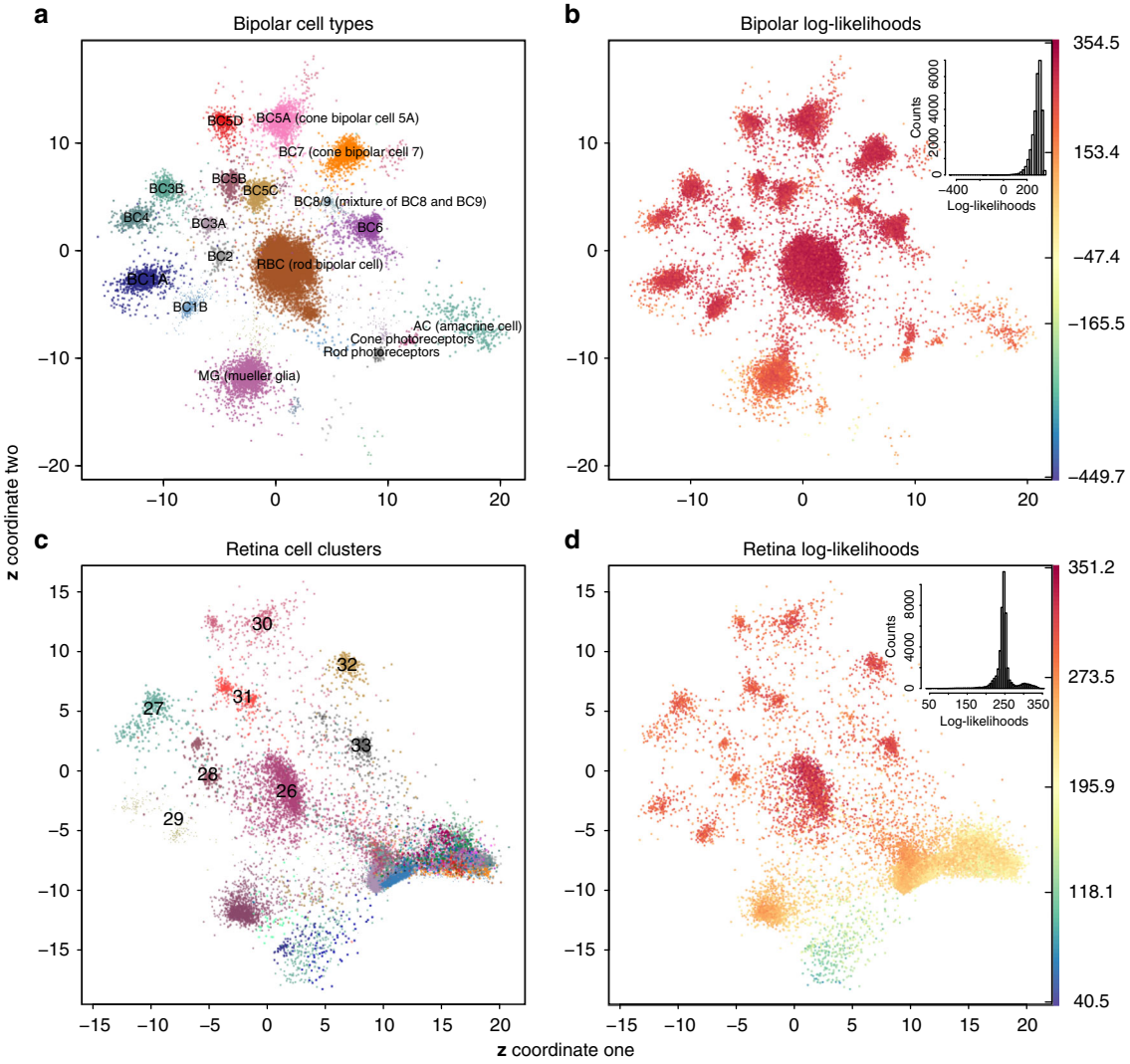


Fig. 4 Learning a probabilistic mapping function from the bipolar data and applying the function to the independently generated mouse retina dataset. **a** scvis learned two-dimensional representations of the bipolar dataset, **b** coloring each point by the estimated log-likelihood, **c** the whole mouse retina dataset was directly projected to a two-dimensional space by the probabilistic mapping function learned from the bipolar data, and **d** coloring each point from the retina dataset by the estimated log-likelihood

Learning a parametric mapping for a single-cell dataset. We next analyzed the scvis learned probabilistic mapping from a training single-cell dataset and tested how it performed on unseen data. We first trained a model on the mouse bipolar cell of the retina dataset[1] and then used the learned model to map the independently generated mouse retina dataset[9]. The two-dimensional coordinates from the bipolar dataset captured much information in this dataset (Fig. 4a). For example, non-bipolar cells such as amacrine cells, Mueller glia, and photoreceptors were at the bottom, the rod bipolar cells were in the middle, and the cone bipolar cells were on the top left around the rod bipolar cells. Moreover, the "OFF" cone bipolar cells (BC1A, BC1B, BC2, BC3A, BC3B, BC4) were on the left and close to each other, and the "ON" cone bipolar cells (BC5A-D, BC6, BC7, BC8/9) were at the top. Cell doublets and contaminants (accounting for 2.43% of the cells comprised eight clusters[1], with distinct color and symbol combinations in Fig. 4a but not labeled) were rare in the bipolar datasets, and they were mapped to low-density regions in the low-dimensional plots (Fig. 4a).

Consistent with the synthetic data (Fig. 2), t-SNE put the "outlier" cell doublets and contaminants into very distinct compact clusters (Supplementary Fig. 5a, t-SNE coordinates from Shekhar et al.[1]). In addition, although t-SNE mapped cells from different cell populations into distinct regions, more global organizations of clusters of cells were missed in the t-SNE embedding. The "ON" cone bipolar cell clusters, the "OFF" cone bipolar cell clusters, and other non-bipolar cell clusters were mixed together in the t-SNE results.

The bipolar cells tended to have higher log-likelihoods than non-bipolar cells such as amacrine cells, Mueller glia, and photoreceptors (Fig. 4b), suggesting that the model used most of its power to model the bipolar cells, while other cell types were not modeled as well. The embedded figure at the top right corner shows the histogram of the log-likelihoods. The majority of the points exhibited high log-likelihoods (with a median of 292.4). The bipolar cells had significantly higher log-likelihoods (median log-likelihood of 298.4) relative to non-bipolar cells (including amacrine cells, Mueller glia, rod and cone photoreceptors) (median log-likelihood of 223.6; one-sided Mann–Whitney *U*-test FDR < 0.001; Supplementary Fig. 5b). The amacrine cells had the lowest median log-likelihood (median log-likelihood for amacrine cells, Mueller glia, rod and cone photoreceptors were 226.4, 187.3, 222.7, and 205.4, respectively; Supplementary Fig. 5b).

We benchmarked scvis against GPLVM, pt-SNE, and PCA on embedding out-of-sample scRNA-seq data, performing a five-fold cross-validation analysis on the bipolar dataset. Specifically, we partitioned the bipolar dataset into five roughly equal size subsamples and held out one subsample as out-of-sample evaluation data, using the remaining four subsamples as training data to learn different models. We then trained *K*nn classifiers on the two-dimensional representations of the training data and then used the *K*nn classifiers to classify the two-dimensional representations of the out-of-sample evaluation data. The process was repeated five times with each of the five subsamples used exactly once as the out-of-sample validation data. scvis was significantly better than pt-SNE, GPLVM, and PCA on embedding the out-of-samples (Supplementary Fig. 6a, b, FDR < 0.05, one-sided Welch's *t*-test).

We used the learned probabilistic mapping from the bipolar cells to map the independent whole-retina dataset[9]. We first projected the retina dataset to the subspace spanned by the first 100 principal direction vectors of the bipolar dataset and then mapped each 100-dimensional vector to a two-dimensional space based on the learned scvis model from the bipolar dataset. The bipolar cell clusters in the retina dataset identified in the original study[9] (clusters 26–33) tended to be mapped to the corresponding bipolar cell subtype regions discovered in the study[1] (Fig. 4c). Although Macosko et al.[9] only identified eight subtypes of bipolar cells, all the recently identified 14 subtypes of bipolar cells[1] were possibly present in the retina dataset as can be seen from Fig. 4c, i.e., cluster 27 (BC3B and BC4), cluster 28 (BC2 and BC3A), cluster 29 (BC1A and BC1B), cluster 30 (BC5A and BC5D), cluster 31 (BC5B and BC5C), and cluster 33 (BC6 and BC8/9).

Interestingly, there was a cluster just above the rod photoreceptors (Fig. 4c) consisting of different subtypes of bipolar cells. In the bipolar dataset, cell doublets or contaminants were mapped to this region (Fig. 4a). We used densitycut[52] to cluster the two-dimensional mapping of all the bipolar cells from the retina dataset to detect this mixture of bipolar cell cluster (Supplementary Fig. 5c, where the 1535 high-density points in this cluster were labeled with red circles). To test whether this mixture cell population was an artifact of the projection, we randomly drew the same number of data points from each bipolar subtype as in the mixture cluster and computed the *K*nns of each data point (here K was set to $\log_2(1535) = 11$). We found that the 11 nearest neighbors of the points from the mixture clusters were also mostly from the mixture cluster (median of 11 and mean of 10.8), while for the randomly selected points from the bipolar cells, a relatively small number of points of their 11 nearest neighbors (median of 0 and mean of 0.2) were from the mixture cluster. The results suggest that the bipolar cells in the mixture cluster were substantially different from other bipolar cells. Finally, this mixture of bipolar cells had significantly lower log-likelihoods compared with other bipolar cells (one-sided Mann–Whitney *U*-test *p*-value <0.001, Supplementary Fig. 5d).

Non-bipolar cells, especially Mueller glia cells, were mapped to the corresponding regions as in the bipolar dataset (Fig. 4c). Photoreceptors (rod and cone photoreceptors accounting for 65.6 and 4.2% of all the cells from the retina[9]) were also mapped to their corresponding regions as in the bipolar dataset (Supplementary Fig. 5e). The amacrine cells (consisting of 21 clusters) together with horizontal cells and retinal ganglion cells were mapped to the bottom right region (Supplementary Fig. 5f); all the amacrine cells were assigned the same label and the same color.

As in the training bipolar data, the bipolar cells in the retina dataset also tended to have high log-likelihoods, and other cells tended to have relatively lower log-likelihoods (Fig. 4d). The embedded plot on the top right corner shows a bimodal distribution of the log-likelihoods. The "Other" cells types (horizontal cells, retina ganglion cells, microglia cells, etc) that were only in the retina dataset had the lowest log-likelihoods (median log-likelihoods of 181.7, Supplementary Fig. 5d).

It is straightforward to project scRNA-seq to a higher than two-dimensional space. To evaluate how scvis performs on higher-dimensional maps, we projected the bipolar data to a three-dimensional space. We obtained better average log-likelihood per data point, i.e., 255.1 versus 253.3 (from the last 100 iterations) by projecting the data to a three-dimensional space compared to projecting the data to a two-dimensional space (Supplementary Fig. 7). In addition, the average $\mathbb{KL}$ divergence was smaller (2.7 versus 4.1 from the last 100 iterations) by projecting the data to a three-dimensional space.

Finally, to demonstrate that scvis can be used for other types of single-cell data, we learned a parametric mapping from the CyTOF data H2 and then directly used the mapping to project the CyTOF data H1 to a two-dimensional space. As can be seen from Supplementary Fig. 8a, all the 14 cell types were separated (although CD16+ and CD16− NK cells have some overlaps), and CD4 T cells and CD8 T cells clusters are adjacent to each other. Moreover, the high quality of the mapping carried over to the CyTOF data H1 (72,463 cells, Supplementary Fig. 8a, b).

Tumor microenvironments and intratumor heterogeneity. We next used scvis to analyze tumor microenvironments and intratumor heterogeneity. The oligodendroglioma dataset consists of mostly malignant cells (Supplementary Fig. 9a). We used densitycut[52] to cluster the two-dimensional coordinates to produce 15 clusters (Supplementary Fig. 9b). The non-malignant cells (microglia/macrophage and oligodendrocytes) formed two small clusters on the left and each consisted of cells from different patients. We therefore computed the entropy of each cluster based on the cells of origin (enclosed bar plot). As expected, the non-malignant clusters (cluster one and cluster five) had high entropies. Cluster 12 (cells mostly from MGH53 and MGH54) and cluster 14 (cells from MGH93 and MGH94) also had high entropies (Fig. 5a). The cells in these two clusters consisted of mostly astrocytes (Fig. 5b; the oligodendroglioma cells could roughly be classified as oligodendrocyte, astrocyte, or stem-like cells.) Interestingly, cluster 15 had the highest entropy, and these cells had significant higher stem-like scores (one-sided Welch's *t*-test *p*-value $<10^{-12}$). We also colored cells by the cell-cycle scores (G1/S scores, Supplementary Fig. 9c; G2/M scores, Supplementary Fig. 9d) and found that these cells also had significantly higher G1/S scores (one-sided Welch's *t*-test *p*-value $<10^{-12}$) and G2/M scores (one-sided Welch's *t*-test *p*-value $<10^{-9}$). Therefore, cluster 15 cells consisted of mostly stem-like cells, and these cells were cycling.

Malignant cells formed distinct clusters even if they were from the same patient (Fig. 5a). We next colored each malignant cell by its lineage score[44] (Fig. 5b). The cells in some clusters highly expressed the astrocyte gene markers or the oligodendrocyte gene markers. The stem-like cells tended to be rare and they could link "outliers" connecting oligodendrocyte and astrocyte cells in the two-dimensional scatter plots (Fig. 5b). In addition, some clusters of cells consisted of mixtures of cells (e.g., both oligodendrocyte and stem-like cells), suggesting that other factors such as genetic mutations and epigenetic measurements would be required to fully interpret the clustering structures in the dataset.

For the melanoma dataset, the authors profiled both malignant cells and non-malignant cells[3]. The malignant cells originated from different patients were mapped to the bottom left region (Fig. 5c). These malignant cells were further subdivided by the patients of origin (Fig. 5d). Similar to the oligodendroglioma

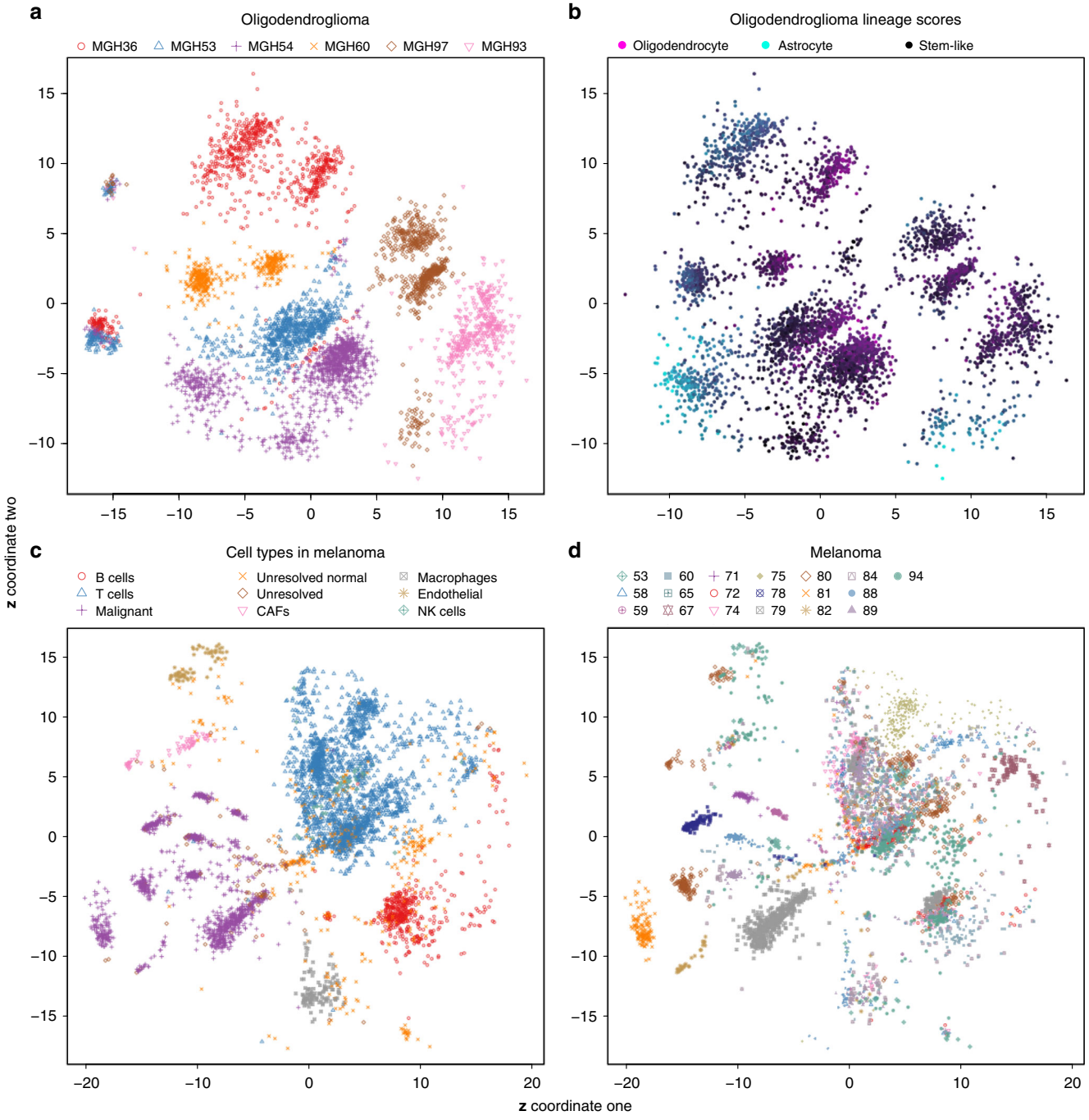


Fig. 5 scvis learned low-dimensional representations. **a** The oligodendroglioma dataset, each cell is colored by its patient of origin, **b** the oligodendroglioma dataset, each cell is colored by its linage score from Tirosh et al.[44], **c** the melanoma dataset, each cell is colored by its cell type, and **d** the melanoma dataset, each cell is colored by its patient of origin

dataset, non-malignant immune cells such as T cells, B cells, and macrophages, even from different patients, tended to be grouped together by cell types instead of patients of origin of the cells (Fig. 5c, d), although for some patients (e.g., 75, 58, and 67, Fig. 5d), their immune cells showed patient-specific bias. We did a differential expression analysis of patient 75 T cells and other patient T cells using limma[53]. Most of the top 100 differently expressed genes were ribosome genes (Supplementary Fig. 10a), suggesting that batch effects could be detectable between patient 75 T cells and other patient T cells.

Interestingly, as non-malignant cells, cancer-associated fibroblasts (CAFs) were mapped to the region adjacent to the malignant cells. The endothelial cells were just above the CAFs (Fig. 5d). To test whether these cells were truly more similar with the malignant cells than with immune cells, we first computed the average principal component values in each type of cells and did a hierarchical clustering analysis (Supplementary Fig. 10b). Generally, there were two clusters: one cluster consisted of the immune cells and the "Unsolved normal" cells, while the other cluster consisted of CAFs, endothelial cells, malignant cells, and the "Unsolved" cells, indicating CAFs and endothelial cells were

more similar to malignant cells (they had high PC1 values) than to the immune cells.

Discussion

We have developed a novel method, scvis, for modeling and reducing dimensionality of single-cell gene expression data. We demonstrated that scvis can robustly preserve the structures in high-dimensional datasets, including in datasets with small numbers of data points.

Our contribution has several important implications for the field. As a probabilistic generative model, scvis provides not only the low-dimensional coordinate for a given data point but also the log-likelihood as a measure of the quality of the embedding. The log-likelihoods could potentially be used for outlier detection, e.g., for the bipolar cells in Fig. 4b, the log-likelihood histogram shows a long tail of data points with relatively low log-likelihoods, suggesting some outliers in this dataset (the non-bipolar cells). The log-likelihoods could also be useful in mapping new data. For example, although horizontal cells and retinal ganglion cells were mapped to the region adjacent to/overlap the region occupied by amacrine cells, these cells exhibited low log-likelihoods, suggesting that further analyses were required to elucidate these cell types/subtypes.

scvis preserves the "global" structure in a dataset, greatly enhancing interpretation of projected structures in scRNA-seq data. For example, in the bipolar dataset, the "ON" bipolar cells were close to each other in the two-dimensional representation in Fig. 4a, and similarly, the "OFF" bipolar cells were close to each other. For the oligodendroglioma dataset, the cells can be first divided into normal cells and malignant cells. The normal cells formed two clusters, with each cluster of cells consisting of cells from multiple patients. The malignant cells, although from the same patient, formed multiple clusters with cell clusters from the same patient adjacent to each other. Adjacent malignant cell clusters from different patients tended to selectively express the oligodendrocyte marker genes or the astrocyte marker genes. For the metastatic melanoma dataset, malignant cells from different patients, although mapped to the same region, formed clusters based on the patient origin of the cells, while immune cells from different patients tended to be clustered together by cell types. From the low-dimensional representations, we can hypothesize that the CAFs were more "similar" to the malignant cells than to the immune cells.

Other methods, e.g., the SIMLR algorithm, improve the t-SNE algorithm[54] by learning a similarity matrix between cells, and the similarity matrix is used as the input of t-SNE for dimension reduction. However, SIMLR is computationally expensive because its objective function involves large matrix multiplications (an $N \times N$ kernel matrix multiplying an $N \times N$ similarity matrix, where N is the number of cells). In addition, although the learned similarity matrix could help clustering analyses, it may distort the manifold structure as demonstrated in the t-SNE plots on the learned similarity matrix[54] because the SIMLR objective function encourages forming clusters. The DeepCyTOF[55] framework has a component that uses a denoising autoencoder (trained on the cells with few or without zeros events) to filter CyTOF data to minimize the influence of dropout noises in single-cell data. The purpose of DeepCyTOF is quite different from that of scvis to model and visualize the low-dimensional structures in high-dimensional single-cell data. The most similar approach for scvis may be the parametric t-SNE algorithm[51], which uses a neural network to learn a parametric mapping from the high-dimensional space to a low dimension. However, parametric t-SNE is not a probabilistic model, the learned low-dimensional embedding is difficult to interpret, and there are no likelihoods to quantify the uncertainty of each mapping.

In conclusion, the scvis algorithm provides a computational framework to compute low-dimensional embeddings of scRNA-seq data while preserving global structure of the high-dimensional measurements. We expect scvis to model and visualize structures in scRNA-seq data while providing new means to biologically interpretable results. As technical advances to profile the transcriptomes of large numbers of single cells further mature, we envisage that scvis will be of great value for routine analysis of large-scale, high-resolution mapping of cell populations.

Methods

A latent variable model of single-cell data. We assume that the gene expression vector $\mathbf{x}_n$ of cell n is a random vector and is governed by a low-dimensional latent vector $\mathbf{z}_n$. The graphical model representation of this latent variable model (with N cells) is shown in Fig. 6a. The $\mathbf{x}_n$ distribution could be a complex high-dimensional distribution. We assume that it follows a Student's t-distribution given $\mathbf{z}_n$:

$$p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\theta}) = \mathcal{T}(\mathbf{x}_n|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_n), \boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{z}_n), \boldsymbol{\nu}) \quad (1)$$

where both $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\cdot)$ and $\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\cdot)$ are functions of $\mathbf{z}$ given by a neural network with parameter $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$ is the degree of freedom parameter and learned from data. The marginal distribution $p(\mathbf{x}_n|\boldsymbol{\theta}) = \int p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\theta}) p(\mathbf{z}_n|\boldsymbol{\theta}) \mathrm{d}\mathbf{z}_n$ can model a complex high-dimensional distribution.

We are interested in the posterior distribution of the low-dimensional latent variable given data: $p(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\theta})$, which is intractable to compute. To approximate the posterior, we use the variational distribution $q(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\phi}) = \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}_n), \mathrm{diag}(\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x}_n))\right)$ (Fig. 6b). Both $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\cdot)$ and $\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\cdot)$ are functions of $\mathbf{x}$ through a neural network with parameter $\boldsymbol{\phi}$. Although the number of latent variables grows with the number of cells, these latent variables are governed by a neural network with a fixed set of parameters $\boldsymbol{\phi}$. Therefore, even for datasets with large number of cells, we still can efficiently infer the posterior distributions of latent variables. The model coupled with the variational inference is called the variational autoencoder[56, 57].

Now the problem is to find the variational parameter $\boldsymbol{\phi}$ such that the approximation $q(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\phi})$ is as close as possible to the true posterior distribution $p(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\theta})$. The quality of the approximation is measured by the Kullback–Leibler (

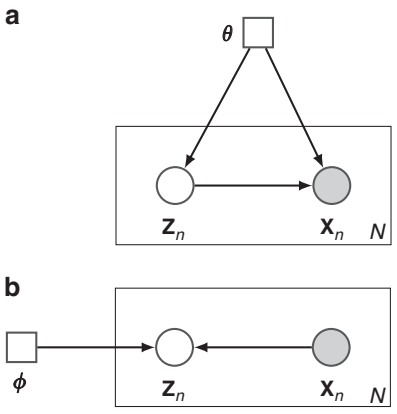


Fig. 6 The scvis directed probabilistic graphical model and the variational approximation of its posterior. Circles represent random variables. Squares represent deterministic parameters. Shaded nodes are observed, and unshaded nodes are hidden. Here we use the plate notation, i.e., nodes inside each box will get repeated when the node is unpacked (the number of repeats is on the bottom right corner of each box). Each node and its parents constitute a family. Given the parents, a random variable is independent of the ancestors. Therefore, the joint distribution of all the random variables is the product of the family conditional distributions. **a** The generative model to generate data $\mathbf{x}_n$, and **b** the variational approximation $q(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\phi})$ to the posterior $p(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\theta})$

$\mathbb{KL}$) divergence[58]

$$\begin{aligned}&\mathbb{KL}(q(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\phi})||p(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\theta}))\\&=\int q(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\phi})\log\frac{q(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\phi})}{p(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\theta})}\mathrm{d}\mathbf{z}_n\\&=\int q(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\phi})\log\frac{q(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\phi})p(\mathbf{x}_n|\boldsymbol{\theta})}{p(\mathbf{z}_n,\mathbf{x}_n|\boldsymbol{\theta})}\mathrm{d}\mathbf{z}_n\\&=\mathbb{E}_{q(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\phi})}[\log q(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\phi})]\\&-\mathbb{E}_{q(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\phi})}[\log p(\mathbf{z}_n,\mathbf{x}_n|\boldsymbol{\theta})]+logp(\mathbf{x}_n|\boldsymbol{\theta})\end{aligned}\tag{2}$$

$$\begin{aligned}&=\mathbb{KL}[q(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\phi})||p(\mathbf{z}_n|\boldsymbol{\theta})]\\&-\mathbb{E}_{q(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\phi})}[\log p(\mathbf{x}_n|\mathbf{z}_n,\boldsymbol{\theta})]+\log p(\mathbf{x}_n|\boldsymbol{\theta})\end{aligned}\tag{3}$$

The term $\mathbb{E}_{q(\mathbf{z}_n|\mathbf{x}_n,\varphi)}[\log p(\mathbf{z}_n,\mathbf{x}_n|\boldsymbol{\theta})]-\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_n,\varphi)}[\log q(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\phi})]$ in Eq. (2) is the evidence lower bound (ELBO) because it is a lower bound of $logp(\mathbf{x}_n\mid\boldsymbol{\theta})$ as the $\mathbb{KL}$ divergence on the left hand side is non-negative. We therefore can do maximum-likelihood estimation of both $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ by maximizing the ELBO. Notice that in the Bayesian setting, the ELBO is a lower bound of the evidence log $p(\mathbf{x}_n)$ as the parameters $\boldsymbol{\theta}$ are also latent random variables.

Both the prior $p(\mathbf{z}_n\mid\boldsymbol{\theta})$ and the variational distribution $q(\mathbf{z}_n\mid\mathbf{x}_n,\boldsymbol{\phi})$ in the ELBO of the form in Eq. (3) are distributions of $\mathbf{z}_n$. In our case, we can compute the $\mathbb{KL}$ term analytically because the prior is a multivariate normal distribution, and the variational distribution is also a multivariate normal distribution given $\mathbf{x}_n$. However, typically there is no closed-form expression for the integration $\mathbb{E}_{q(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\phi})}[\log p(\mathbf{x}_n\mid\mathbf{z}_n,\boldsymbol{\theta})]$ because we should integrate out $\mathbf{z}_n$ and the parameters of the model $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_n)$ and $\mathrm{diag}(\boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{z}_n))$ are functions of $\mathbf{z}_n$. Instead, we can use Monte Carlo integration and obtain the estimated evidence lower bound for the nth cell:

$$\mathrm{ELBO}_n=-\mathbb{KL}(q(\mathbf{z}_n|\mathbf{x}_n,\boldsymbol{\phi})||p(\mathbf{z}_n|\boldsymbol{\theta}))+\frac{1}{L}\sum_{l=1}^{L}\log p(\mathbf{x}_n|\mathbf{z}_{n,l},\boldsymbol{\theta})\tag{4}$$

where $\mathbf{z}_{n,l}$ is sampled from $q(\mathbf{z}_n\mid\mathbf{x}_n,\boldsymbol{\phi})$ and L is the number of samples. We want to take the partial derivatives of the ELBO w.r.t. the variational parameter $\boldsymbol{\phi}$ and the generative model parameter $\boldsymbol{\theta}$ to find a local maximum of the ELBO. However, if we directly sample points from $q(\mathbf{z}_n\mid\mathbf{x}_n,\boldsymbol{\phi})$, it is impossible to use the chain rule to take the partial derivative of the second term of Eq. (4) w.r.t $\boldsymbol{\phi}$ because $\mathbf{z}_{n,l}$ is a number. To use gradient-based methods for optimization, we indirectly sample data from $q(\mathbf{z}_n\mid\mathbf{x}_n,\boldsymbol{\phi})$ using the "reparameterization trick"[56, 57]. Specifically, we first sample $\boldsymbol{\varepsilon}_l$ from a easy to sample distribution $\boldsymbol{\epsilon}_l\sim p(\boldsymbol{\epsilon}|\boldsymbol{\alpha})$, e.g., a standard multivariate Gaussian distribution for our case. Next we pass $\boldsymbol{\epsilon}_l$ through a continuous function $g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon},\mathbf{x}_n)$ to get a sample from $q(\mathbf{z}_n\mid\mathbf{x}_n,\boldsymbol{\phi})$. For our case, if $q(\mathbf{z}_n\mid\mathbf{x}_n,\boldsymbol{\phi})=\mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\varphi}}(\mathbf{x}_n),\ \mathrm{diag}\ (\boldsymbol{\sigma}_{\boldsymbol{\phi}}\ (\mathbf{x}_n))\right)$, then $g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon},\mathbf{x}_n)=\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}_n)+\mathrm{diag}\left(\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x}_n)\right)\times\boldsymbol{\epsilon}$.

Adding regularizers on the latent variables. Given *i.i.d* data $\mathcal{D}=\{\mathbf{x}_n\}_{n=1}^N$, by maximizing the $\sum_n\mathrm{ELBO}_{n=1}^N$, we can do maximum-likelihood estimation of the model parameters $\boldsymbol{\theta}$ and the variational distribution parameters $\boldsymbol{\phi}$. Although $p(\mathbf{z}_n\mid\boldsymbol{\theta})p(\mathbf{x}_n\mid\mathbf{z}_n,\boldsymbol{\theta})$ may model the data distribution very well, the variational distribution $q(\mathbf{z}_n\mid\mathbf{x}_n,\boldsymbol{\phi})$ is not necessarily good for visualization purposes. Specifically, it is possible that there are no very clear gaps among the points from different clusters. In fact, to model the data distribution well, the low-dimensional $\mathbf{z}$ space tends to be filled such that all the $\mathbf{z}$ space is used in modeling the data distribution. To better visualize the manifold structure of a dataset, we need to add regularizers to the objective function in Eq. (4) to encourage forming gaps between clusters and at the same time keeping nearby points in the high-dimensional space nearby in the low-dimensional space. Here we use the non-symmetrized t-SNE[34–39] objective function.

The t-SNE algorithm preserves the local structure in the high-dimensional space after dimension reduction. To measure the "localness" of a pairwise distance, for a data point i in the high-dimensional space, the pairwise distance between i and another data point j is transformed to a conditional distribution by centering an isotropic univariate Gaussian distribution at i

$$p_{j|i}=\frac{\exp\left(-\mathbf{x}_i-\mathbf{x}_j^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\mathbf{x}_i-\mathbf{x}_k^2/2\sigma_i^2\right)}\tag{5}$$

The point-specific standard deviation σ_i is a parameter that is computed automatically in such a way that the perplexity ($2^{-\sum_j p_{j|i}\log_2 p_{j|i}}$) of the conditional distribution $p_{j|i}$ equals a user defined hyperparameter (e.g., typically 30[48]). We set $p_{i|i}=0$ because only pairwise similarities are of interest.

In the low-dimensional space, the conditional distribution $q_{j|i}$ is defined similarly and $q_{i|i}$ is set to 0. The only difference is that an unscaled univariate Student's t-distribution is used instead of an isotropic univariate Gaussian distribution as in the high-dimensional space. Because in the high-dimensional space more points can be close to each other than in the low-dimensional space (e.g., only two points can be mutually equidistant in a line, three points in a two-dimensional plane, and four points in a three-dimensional space), it is impossible to faithfully preserve the high-dimensional pairwise distance information in the low-dimensional space if the intrinsic dimensionality of the data is bigger than that of the low-dimensional space. A heavy tailed Student's t-distribution allows moderate distances in the high-dimensional space to be modeled by much larger distances in the low-dimensional space to prevent crushing different clusters together in the low-dimensional space[34].

The low-dimensional embedding coordinates $\{\mathbf{z}_i\}_{i=1}^N$ are obtained by minimizing the $\mathbb{KL}$ divergence between the sum of conditional distributions:

$$\begin{aligned}&\sum_i\mathbb{KL}\left(p_{\cdot|i}||q_{\cdot|i}\right)=\sum_{i=1}^N\sum_{j=1,j\neq i}^N p_{j|i}\log\frac{p_{j|i}}{q_{j|i}}\\&=\sum_{i=1}^N\sum_{j=1,j\neq i}^N p_{j|i}\log p_{j|i}-\sum_{i=1}^N\sum_{j=1,j\neq i}^N p_{j|i}\log q_{j|i}\\&\propto-\sum_{i=1}^N\sum_{j=1,j\neq i}^N p_{j|i}\log\frac{\left(1+\left\|\mathbf{z}_i-\mathbf{z}_j\right\|^2/\nu\right)^{-\frac{\nu+1}{2}}}{\sum_{k,k\neq i}\left(1+\left\|\mathbf{z}_i-\mathbf{z}_k\right\|^2/\nu\right)^{-\frac{\nu+1}{2}}}\\&\propto-\sum_{i=1}^N\sum_{j=1,j\neq i}^N p_{j|i}\log\left(1+\left\|\mathbf{z}_i-\mathbf{z}_j\right\|^2/\nu\right)^{-\frac{\nu+1}{2}}\\&+\sum_{i=1}^N\sum_{j=1,j\neq i}^N p_{j|i}\log\sum_{k,k\neq i}\left(1+\left\|\mathbf{z}_i-\mathbf{z}_k\right\|^2/\nu\right)^{-\frac{\nu+1}{2}}\\&\propto-\sum_{i=1}^N\sum_{j=1,j\neq i}^N p_{j|i}\log\left(1+\left\|\mathbf{z}_i-\mathbf{z}_j\right\|^2/\nu\right)^{-\frac{\nu+1}{2}}\end{aligned}\tag{6}$$

$$+\sum_{i=1}^N\log\sum_{k,k\neq i}\left(1+\mathbf{z}_i-\mathbf{z}_k^2/\nu\right)^{-\frac{\nu+1}{2}}\tag{7}$$

Here v is the degree of freedom of the Student's t-distribution, which is typically set to one (the standard Cauchy distribution) or learned from data. Equation (6) is a data-dependent term (depending on the high-dimensional data) that keeps nearby data points in the high-dimensional data nearby in the low-dimensional space[37]. Equation (7) is a data-independent term that pushes data points in the low-dimensional space apart from each other. Notice that the t-SNE objection function[34] minimizes the $\mathbb{KL}$ divergence of the joint distribution defined as the symmetrized condition distributions $p_{i,j}=(p_{i|j}+p_{j|i})/(2\times N)$ and $q_{i,j}=(q_{i|j}+q_{j|i})/(2\times N)$. t-SNE has shown excellent results on many visualization tasks such as visualizing scRNA-seq data and CyTOF data[40].

The final objective function is a weighted combination of the ELBO of the latent variable model and the above asymmetric t-SNE objective function:

$$\arg\min_{\boldsymbol{\theta},\boldsymbol{\phi}}\left(-\sum_{n=1}^N\mathrm{ELBO}_n+\alpha\sum_{n=1}^N\mathbb{KL}\left(p_{\cdot|n}||q_{\cdot|n}\right)\right)\tag{8}$$

The parameter α is set to the dimensionality of the input high-dimensional data because the magnitude of the log-likelihood term in the ELBO scales with the dimensionality of the input data. The perplexity parameter is set to ten for scvis.

Sensitivity of scvis on cell numbers. To test the performance of scvis on scRNA-seq datasets with small numbers of cells, we ran scvis on subsampled data from the bipolar dataset. The bipolar dataset consists of six batches of datasets. We only used the cells from batch six (6221 cells in total after removing cell doublets and contaminants) to remove batch effects. Specifically, we subsampled 1, 2, 3, 5, 10, 20, 30, and 50% of the bipolar dataset from batch six (62, 124, 187, 311, 622, 1244, 1866, and 3110 cells, respectively). Then we computed the principal components from the subsampled data, and ran scvis using the top 100 PCs (for the cases with $M<100$ cells, we used the top M PCs). For each subsampled dataset, we ran scvis ten times with different seeds. We used exactly the same parameter setting for all the datasets. Therefore, except for the models trained on the cases with <100 cells, all other models have the same number of parameters.

When the number of training data is small (e.g., 62 cells, 124 cells, or 187 cells), only the large clusters such as cluster one and cluster two are distinct from the rest (Supplementary Figs. 11 and 12). As we increased the number of subsampled data points, some small clusters of cells can be recovered. At 622 cells, many cell clusters can be recovered. The *K*nn classification accuracies in Supplementary Fig. 13 (trained on the two-dimensional representations of the subsampled data and tested on the two-dimensional representations of the remaining cells from batch six) shows relatively high mean accuracies of 84.4, 84.5, 84.1, and 81.2% for K equals to 5, 9, 17, and 33. When we subsampled 622 cells as in Supplementary Fig. 11e, the cells in cluster 22 were not present in the 622 cells. However, when we used the model trained on these 622 cells to embed the remaining 5599 cells (6221–622), cluster 22 cells were mapped to the "correct" region that was adjacent to cluster 20 cells and bridged cluster 20 and other clusters as in Supplementary Fig. 11d, f–h. Interestingly, with smaller numbers of cells (cell numbers ≤622), *K*nn classifiers trained on the two-dimensional scvis coordinates were better than those trained using the 100 principal components (one-sample t-test FDR < 0.05; Supplementary Fig. 12, the red color triangles represent the *K*nn accuracy by using the original 100 PCs).

It seems that there is no noticeable overfitting even with small numbers of cells as can be seen from the two-dimensional plots from Supplementary Figs. 11 and 12 and the *K*nn classification accuracies in Supplementary Fig. 13. To decrease the possibility of overfitting, we used Student's *t*-distributions instead of Gaussian distributions for the model neural networks. In addition, we used relatively small neural networks (a three-layer inference network with 128, 64, and 32 units and a five-layer model network with 32, 32, 32, 64, and 128 units), which may decrease the chances of overfitting.

Sensitivity of scvis on hyperparameters. scvis has several hyperparameters such as the number of layers in the variational inference and the model neural networks, the layer sizes (the number of units in each layer), and the α parameters in Eq. (8). We established that scvis is typically robust to these hyperparameters. We first tested the influence of the number of layers using batch six of the bipolar dataset. We learned scvis models with different layers from the 3110 subsampled data points from the batch six bipolar dataset. Specifically, we tested these variational influence neural networks and the model neural network layer combinations (the first number is the number of layers in the variational influence neural network, and the second number is the number of layers in the model neural network): (10, 1), (10, 3), (10, 5), (10, 7), (10, 10), (7, 10), (5, 10), (3, 10), and (1, 10). These models performed reasonably well such that the cells from the same cluster are close to each other in the two-dimensional spaces (Supplementary Fig. 14). When the number of layers in the variational inference neural network was fixed to ten, for some types of cells, their two-dimensional embeddings were close to each other and formed curve-like structures as can be seen from Supplementary Fig. 14e. The reason for this phenomenon could be that the variational influence network underestimated the variances of the latent **z** posterior distributions or the optimization was not converged. On the contrary, when the influence networks have smaller number of layers (<10), we did not see these curve structures (Supplementary Fig. 14f–i). The out-of-sample mapping results in Supplementary Fig. 15 show similar results.

We computed the *K*nn classification accuracies of the out-of-samples. As before, the *K*nn classifiers were trained on the two-dimensional coordinates of the training subsampled data, and the classifiers were used to classify the two-dimensional coordinates of the out-of-sample data. The parameter setting did influence scvis performance, i.e., for each $K \in \{5,9,17,33,65,129,257\}$, the trained scvis models did significantly different (Supplementary Fig. 16, FDR < 0.05, one-way analysis of variance (ANOVA) test). To find out which parameter combinations led to inferior or superior performance, we then compared the classification accuracies of each model with the most complex model with both ten layers of variational influence neural networks and ten layers of model networks. The FDR (two-sided Welch's *t*-test) at the top of each subfigure of Supplementary Fig. 16 shows that, except for $K = 257$, all the models with one layer of variational influence neural networks did significantly worse than those from the most complex model (FDR < 0.05, two-sided Welch's *t*-test). Similarly, the models with three layers of variational influence neural networks did significantly worse than those from the most complex model when $K \in \{5, 9, 17, 33, 65\}$. While for other models, their performances were not statistically different from those of the most complex models.

We next examined the influence of the layer sizes of the neural networks. The number of layers was fixed at ten for both the variational influence neural networks and the model neural networks; the number of units in each layer was set to 8, 16, 32, 64, and 128. All layers of the inference and the model neural networks had the same size. All models successfully embedded both the training data and the out-of-sample test data (Supplementary Fig. 17). However, the layer size parameter did influence scvis performance, i.e., the *K*nn classifiers on the out-of-sample data did significantly different (Supplementary Fig. 18, FDR < 0.05, one-way ANOVA test). The FDR (two-sided Mann–Whitney *U*-test) at the top of each subfigure of Supplementary Fig. 18 shows that all models with layer size of eight did significantly worse than those from the most complex model using 128 units (FDR < 0.05). Similarly, the models with layer size of 16 did significantly worse than those from the most complex model when $K \in \{5, 9, 17, 33, 65, 129\}$ (FDR < 0.05). While for other models, their performances were not statistically different from those from the most complex models. Notice that, at layer size of 64, the mapping functions from one run were worse than others in embedding the out-of-sample data. However, there was no significant difference in the log-likelihoods from the repeated ten runs (Supplementary Fig. 19, one-way ANOVA *p*-value = 0.741).

For the α weight parameter in Eq. (8), we set α relative to the dimensionality of the input data. We set $\alpha = 0$, 0.5, 1.0, 1.5, 2.0, 10.0, inf times of the dimensionality of the input data. When $\alpha =$ inf, the trained models did significantly worse than the models trained with the default α equals to the dimensionality of the input data for $K \in \{5, 9, 17, 33, 65\}$ (FDR ≤ 0.05, two-sided Welch's *t*-test, Supplementary Figs. 20, 21 and 22). Also, when $\alpha = 0$, the trained models were significantly worse than the models trained with the default α equaling to the dimensionality of the input data for all *K*s (FDR ≤ 0.05, two-sided Welch's *t*-test, Supplementary Fig. 22). For $\alpha = 0$, we performed an extra comparison by using the synthetic nine-dimensional data, showing that when *K* was large (≥65), setting $\alpha = 9$ (the dimensionality of the input data) did significantly better than letting $\alpha = 0$ (Supplementary Fig. 23, FDR ≤ 0.05, one-sided Welch's *t*-test).

Computational complexity analysis. The scvis objective function involves the asymmetrical t-SNE objective function. The most time-consuming part is to compute the pairwise distances between two cells in a mini-batch that takes $O(TN^2D + TN^2d)$ time, where N is the mini-batch size (we use $N = 512$ in this study), D is the dimensionality of the input data, d is the dimensionality of the low-dimensional latent variables (e.g., $d = 2$ for most cases), and T is the number of iterations. For our case, we first use PCA to project the scRNA-seq data to a 100-dimensional space, so $D = 100$. For a feedforward neural network with L hidden layers, and the number of neurons in layer l is n_l, the time complexity to train the neural network is $O\left(NT\sum_{i=0}^{L} n_{l+1} * n_l\right)$, where $n_0 = D$ and n_{L+1} is the size of the output layer. For the model neural network, we use a five hidden layer (with layer size 32, 32, 32, 64, and 128) feedforward neural network. The input layer size is d and the output layer size is D. For the variational inference network, we use a three hidden layer (with layer size 128, 64, and 32) feedforward neural network. The size of the input layer is D and the size of the output layer is d. For space complexity, we need to save the weights and bias of each neuron $O\left(\sum_{i=0}^{L} n_{l+1} * n_l\right)$. We also need to save the $O(N^2)$ pairwise distances and the data of size $O(ND)$ in a mini-batch.

The original t-SNE algorithm is not scalable to large datasets (with tens of thousands of cells to millions of cells) because it needs to compute the pairwise distances between any two cells (taking $O(M^2D + M^2T)$ time and $O(M^2)$ space, where M is the total number of cells and T is the number of iterations). Approximate t-SNE algorithms are typically more scalable in both time and space. For example, BH t-SNE only computes the distance between a cell and its *K*nns. Therefore, BH t-SNE takes $O(M \log(M))$ time and $O(M \log(M))$ space, where we assume K is in the order of $O(\log(M))$.

We next experimentally compare the scalability of scvis and BH t-SNE (the widely used Rtsne package[48]) by using the 1.3 million cells from 10X genomics[59]. However, BH t-SNE did not finish in 24 h and we terminated it. On the contrary, scvis produced visually reasonable results in <33 min (after 3000 mini-batch training, Supplementary Fig. 24a). Therefore, scvis can be much more scalable than BH t-SNE for very large datasets. As we increased the number of training batches, we can see slightly better separations in clusters as in Supplementary Fig. 24b–f. The time used to train scvis increased linearly in the number of training mini-batches (Supplementary Fig. 24h). However, for small datasets, BH t-SNE can be more efficient than scvis. For example, for the melanoma dataset with only 4645 cells, scvis still took 24 min to run 3000 mini-batches, while BH t-SNE finished in only 28.9 s. All the experiments were conducted using a Mac computer with 32 GB of RAM, 4.2 GHz four-core Intel i7 processor with 8 MB cache.

Finally, when the mapping function is trained, mapping new cells takes only $O\left(M\sum_{l=0}^{L} n_{l+1} * n_l\right)$ time, where M is the number of input cells. Also, because each data point can be mapped independently, the space complexity could be only $O\left(\sum_{l=0}^{L+1} n_{l+1} * n_l\right)$. As an example, it took only 1.5 s for a trained scvis model to map the entire 1.3 million cells from 10X genomics.

Datasets. The oligodendroglioma dataset measures the expression of 23,686 genes in 4347 cells from six *IDH1* or *IDH2* mutant human oligodendrolioma patients[44]. The expression of each gene is quantified as $\log_2$ (TPM/10+1), where "TPM" standards for "transcripts per million"[60]. Through copy number estimations from these scRNA-seq measurements, 303 cells without detectable copy number alterations were classified as normal cells. These normal cells can be further grouped into microglia and oligodendrocyte based on a set of marker genes they expressed. Two patients show subclonal copy number alterations.

The melanoma dataset is from sequencing 4645 cells isolated from 19 metastatic melanoma patients[3]. The cDNAs from each cell were sequenced by an Illumina NextSeq 500 instrument to 30 bp pair-end reads with a median of ~150,000 reads per cell. The expression of each gene (23,686 genes in total) is quantified by $\log_2$ (TPM/10+1). In addition to malignant cells, the authors also profiled immune cells, stromal cells, and endothelial cells to study the whole-tumor multi-cellular ecosystem.

The bipolar dataset consists of low-coverage (median depth of 8200 mapped reads per cell) Drop-seq sequencing[9] of 27,499 mouse retinal bipolar neural cells from a transgenic mouse[1]. In total, 26 putative cells types were identified by clustering the first 37 principal components of all the 27,499 cells. Fourteen clusters can be assigned to bipolar cells, and another major cluster is composed of Mueller glia cells. These 15 clusters account for about 96% of all the 27,499 cells. The remaining 11 clusters (comprising of only 1060 cells) include rod photoreceptors, cone photoreceptors, amacrine cells, and cell doublets and contaminants[1].

The retina dataset consists of low-coverage Drop-seq sequencing[9] of 44,808 cells from the retinas of 14-day-old mice. By clustering the two-dimensional t-SNE embedding using DBSCAN[61]—a density-based clustering algorithm, the authors identified 39 clusters after merging the clusters without enough differentially expressed genes between any two clusters.

The 10X Genomics neural cell dataset consist of 1,306,127 cells from cortex, hippocampus, and subventricular zones of two E18 C57BL/6 mice. The cells were sequenced on 11 Illumina Hiseq 4000 machines to produce 98 bp reads[59].

For the mass cytometry dataset H1[17], manual gating assigned 72,463 cells to 14 cell types based on 32 measured surface protein markers. Manual gating assigned 31,721 cells to the same 14 cell populations from H2 based on the same 32 surface protein markers.

Statistical analysis. All statistical analyses were performed using the R statistical software package, version 3.4.3. Boxplots denote the medians and the interquartile ranges (IQRs). The whiskers of a boxplot are the lowest datum still within 1.5 IQR of the lower quartile and the highest datum still within 1.5 IQR of the upper quartile. The full datasets were superimposed to boxplots. For datasets with non-normal distribution (e.g., outliers), non-parametric tests were used. To account for unequal variances, Welch's *t*-test was used for pairwise data comparison. Adjusted *p*-values <0.05 (FDR, the Benjamini–Hochberg procedure[62]) were considered to be significant.

Code availability. The scvis v0.1.0 Python package is available freely from bitbucket: https://bitbucket.org/jerry00/scvis-dev.

Data availability. The scRNA-seq data that support the findings of this study are available in Gene Expression Omnibus with the identifiers (bipolar: GSE81905, retina: GSE63473, oligodendroglioma: GSE70630, metastatic melanoma: GSE72056, E18 mouse neural cells: GSE93421). The E18 mouse neural cells are freely available from 10X Genomics[59]. The other scRNA-seq data are publicly available from the single-cell portal[45]. The mass cytometric data can be downloaded from cytoback[63]. The synthetic data used in this study are available from bitbucket repo: https://bitbucket.org/jerry00/scvis-dev.

Received: 16 July 2017 Accepted: 25 April 2018
Published online: 21 May 2018

References

1. Shekhar, K. et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* **166**, 1308–1323 (2016).
2. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
3. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
4. Navin, N. et al. Tumor evolution inferred by single cell sequencing. *Nature* **472**, 90–94 (2011).
5. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
6. Jaitin, D. A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
7. Islam, S. et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
8. Macaulay, I. C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
9. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
10. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
11. Hashimshony, T. et al. Cel-seq2: sensitive highly-multiplexed single-cell RNA-seq. *Genome Biol.* **17**, 77 (2016).
12. Gierahn, T. M. et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).
13. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
14. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
15. Rosenberg, A. B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
16. Bendall, S. C. et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).
17. Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
18. Regev, A. et al. The human cell atlas. *Elife* https://doi.org/10.7554/eLife.27041 (2017).
19. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
20. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).
21. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* **14**, 565–571 (2017).
22. Bacher, R. et al. SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* **14**, 584–586 (2017).
23. Qiu, X. et al. Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14**, 309–315 (2017).
24. Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
25. Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643 (2017).
26. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133–145 (2015).
27. Angerer, P. et al. destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32**, 1241–1243 (2015).
28. Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16**, 241 (2015).
29. DeTomaso, D. & Yosef, N. FastProject: a tool for low-dimensional analysis of single-cell RNA-seq data. *BMC Bioinforma.* **17**, 315 (2016).
30. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
31. Setty, M. et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **34**, 637–645 (2016).
32. Campbell, K. R. & Yau, C. Probabilistic modeling of bifurcations in single-cell gene expression data using a bayesian mixture of factor analyzers. *Wellcome Open Res.* **2**, 19 (2017).
33. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *bioRxiv* https://doi.org/10.1101/128843 (2017).
34. Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
35. Hinton, G. E. & Roweis, S. T. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems 15* (eds Becker, S., Thrun, S. & Obermayer, K.) 857–864 (MIT Press, Cambridge, 2003).
36. Cook, J., Sutskever, I., Mnih, A. & Hinton, G. E. Visualizing similarity data with a mixture of maps. In *Proc. Eleventh International Conference on Artificial Intelligence and Statistics*, vol. 2 of *Proceedings of Machine Learning Research* (eds Meila, M. & Shen, X.) 67–74 (PMLR, San Juan, Puerto Rico, 2007).
37. Carreira-Perpinán, M. A. The elastic embedding algorithm for dimensionality reduction. In *Proc. 27th International Conference on Machine Learning* 167–174 (Haifa, Israel, 2010).
38. Yang, Z., Peltonen, J. & Kaski, S. Scalable optimization of neighbor embedding for visualization. In *Proc. 30th International Conference on Machine Learning* (eds Dasgupta, S. & McAllester, D.) 127–135 (PMLR, Atlanta, Georgia, 2013).
39. Maaten, L. v. d. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
40. Amir, E.-a.D. et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**, 545–552 (2013).
41. Zurauskiene, J. & Yau, C. pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* **17**, 140 (2016).
42. Wattenberg, M., ViÉgas, F. & Johnson, I. How to use t-SNE effectively. *Distill* http://distill.pub/2016/misread-tsne (2016).
43. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous systems. Preprint at https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45166.pdf (2015).
44. Tirosh, I. et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).
45. Tickle, T. et al. Single cell portal. https://portals.broadinstitute.org/single_cell (2017).
46. Clevert, D.-A., Unterthiner, T. & Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). In *4th International Conference for Learning Representations* (San Juan, Puerto Rico, 2016).
47. Kingma, D. P. & Ba, J. L. Adam: a method for stochastic optimization. In *3rd International Conference for Learning Representations* (San Diego, CA, 2015).
48. Krijthe, J. H. Rtsne: t-distributed stochastic neighbor embedding using Barnes-Hut implementation. https://github.com/jkrijthe/Rtsne, R package version 0.13 (2015).
49. Lawrence, N. D. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems 16* (eds Thrun, S., Saul, L. K. & Schölkopf, B.) 329–336 (Cambridge, MIT Press, 2004).
50. GPy. GPy: A gaussian process framework in python. http://github.com/SheffieldML/GPy (2012).
51. Maaten, L. Learning a parametric embedding by preserving local structure. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, vol. 5 of *Proceedings of Machine Learning Research* (eds van Dyk, D. & Welling, M.) 384–391 (PMLR, Clearwater Beach, Florida, 2009).
52. Ding, J., Shah, S. & Condon, A. densityCut: an efficient and versatile topological approach for automatic clustering of biological data. *Bioinformatics* **32**, 2567–2576 (2016).

53. Smyth, G. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor* (eds Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A. & Dudoit, S.) 397–420 (Springer, New York, 2005).
54. Wang, B., Zhu, J., Pierson, E. & Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* **14**, 414–416 (2017).
55. Li, H. et al. Gating mass cytometry data by deep learning. *Bioinformatics* **33**, 3423–3430 (2017).
56. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. In *Proc. 2nd International Conference on Learning Representations* (Banff, Alberta, 2014).
57. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *Proc. 31st International Conference on Machine Learning* (eds Xing, E. P. & Jebara, T.) 1278–1286 (PMLR, Beijing, 2014).
58. Kullback, S. & Leibler, R. A. On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).
59. 10X Genomics. 1.3 million brain cells from E18 mice. https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons (2017).
60. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).
61. Ester, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD'96 Proc. Second International Conference on Knowledge Discovery and Data Mining* (eds Simoudis, E., Han, J. & Fayyad, U.) 226–231 (AAAI Press, Portland, Oregon, 1996).
62. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **57**, 289–300 (1995).
63. Levine, J. H. et al. Phenograph. https://www.cytobank.org/nolanlab/reports/Levine2015.html (2015).

Acknowledgements

This work was supported by a Discovery Frontiers project grant, "The Cancer Genome Collaboratory", jointly sponsored by the Natural Sciences and Engineering Research Council (NSERC), Genome Canada (GC), the Canadian Institutes of Health Research (CIHR), and the Canada Foundation for Innovation (CFI) to S.P.S. In addition, we acknowledge generous long-term funding support from the BC Cancer Foundation. The S.P.S. group receives operating funds from the Canadian Breast Cancer Foundation, the Canadian Cancer Society Research Institute (impact grant 701584 to S.P.S.), the Terry Fox Research Institute (grant 1021, The Terry Fox New Frontiers Program Project Grant in the Genomics of Forme Fruste Tumours: New Vistas on Cancer Biology and Treatment, and grant 1061, The Terry Fox New Frontiers Program Project Grant in Overcoming Treatment Failure in Lymphoid Cancers), CIHR (grant MOP-115170 to S.P.S.), and CIHR Foundation (grant FDN-143246 to S.P.S.). S.P.S. is supported by Canada Research Chairs. S.P.S. is a Michael Smith Foundation for Health Research scholar. S.P.S is a Susan G. Komen Scholar.

Author contributions

J.D.: project conception, software implementation, and data analysis; J.D., A.C., and S.P.S.: algorithm development and manuscript writing; S.P.S.: project conception, oversight, and senior responsible author.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-04368-5.

Competing interests: S.P.S is a shareholder of Contextual Genomics Inc. The remaining authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

genes

MDPI

Article

Diagnosis and Prediction of Endometrial Carcinoma Using Machine Learning and Artificial Neural Networks Based on Public Databases

Dongli Zhao [1,†], Zhe Zhang [2,†], Zhonghuang Wang [3,4], Zhenglin Du [3], Meng Wu [5], Tingting Zhang [1], Jialu Zhou [1], Wenming Zhao [3,4,*,‡] and Yuanguang Meng [1,2,6,*,‡]

1 Department of Obstetrics & Gynecology, Chinese People's Liberation Army (PLA) Medical School, No. 28, Fuxing Road, Haidian District, Beijing 100853, China; zz18813066251@163.com (D.Z.); 17865190928@163.com (T.Z.); kalozzhou@163.com (J.Z.)
2 Department of Obstetrics and Gynecology, Seventh Medical Center of Chinese PLA General Hospital, No. 5, Nanmencang, Dongsishitiao, Dongcheng District, Beijing 100700, China; zhangzhe301@126.com
3 National Genomics Data Center & CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Building 104, Courtyard 1, Beichen West Road, Chaoyang District, Beijing 100101, China; wangzhonghuang17m@big.ac.cn (Z.W.); duzhl@big.ac.cn (Z.D.)
4 University of Chinese Academy of Sciences, 19 Yuquan Road (a), Shijingshan District, Beijing 100049, China
5 Medical College, Graduate School of Nankai University, No. 94, Weijin Road, Nankai District, Tianjin 300110, China; awumeng_1992@163.com
6 Department of Gynecology and Obstetrics, Chinese PLA General Hospital, No. 28, Fuxing Road, Haidian District, Beijing 100853, China
* Correspondence: zhaowm@big.ac.cn (W.Z.); meng6512@vip.sina.com (Y.M.)
† These authors contributed equally to this work.
‡ These authors jointly supervised this work.

Abstract: Endometrial carcinoma (EC), a common female reproductive system malignant tumor, affects thousands of people with high morbidity and mortality worldwide. This study was aimed at developing a prediction model for the diagnosis of EC in the general population. First, we obtained datasets GSE63678, GSE106191, and GSE115810 from the Gene Expression Omnibus (GEO) database, dataset GSE17025 from the GEO database, and the RNA sequence of EC from The Cancer Genome Atlas (TCGA) database to constitute the training, test, and validation groups, respectively. Subsequently, the 96 most significantly differentially expressed genes (DEGs) were identified and analyzed for function and pathway enrichment in the training group. Next, we acquired the disease-specific genes by random forest and established an artificial neural network for the diagnosis. Receiver operating characteristic (ROC) curves were utilized to identify the signature across the three groups. Finally, immune infiltration was analyzed to reveal tumor-immune microenvironment (TIME) alterations in EC. The top 96 DEGs (77 down-regulated and 19 up-regulated genes) were primarily enriched in the interleukin-17 signaling pathway, protein digestion and absorption, and transcriptional misregulation in cancer. Subsequently, 14 characterizing genes of EC were identified by random forest. In the training, test, and validation groups, the artificial neural network was constructed with high diagnostic accuracies of 0.882, 0.864, and 0.839, respectively, and areas under the ROC curve (AUCs) of 0.928, 0.921, and 0.782, respectively. Finally, resting and activated mast cells were found to have increased in TIME. We constructed an artificial diagnostic model with excellent reliability for EC and uncovered variations in the immunological ecosystem of EC through integrated bioinformatics approaches, which might be potential diagnostic targets for EC.

Keywords: endometrial carcinoma; GEO; TCGA; random forest; receiver operating characteristic curve

Citation: Zhao, D.; Zhang, Z.; Wang, Z.; Du, Z.; Wu, M.; Zhang, T.; Zhou, J.; Zhao, W.; Meng, Y. Diagnosis and Prediction of Endometrial Carcinoma Using Machine Learning and Artificial Neural Networks Based on Public Databases. *Genes* **2022**, *13*, 935. https://doi.org/10.3390/genes13060935

Academic Editors: Piero Fariselli and Peixin Dong

Received: 7 April 2022
Accepted: 20 May 2022
Published: 24 May 2022

1. Introduction

Endometrial carcinoma (EC), a malignancy of the inner epithelial lining of the uterus, is a common neoplasm in women worldwide, with increasing rates of incidence and disease-associated mortality in recent years [1,2], seriously threatening women's physical and mental health. Most cases of early EC are cured by surgery alone or with adjuvant therapy. However, many cases of EC are diagnosed in the advanced stage at the first consultation and are associated with a poor prognosis. Although the survival rate of patients has increased, owing to molecular targeted therapy, no targeted gene mutations have been explored in advanced EC [3–5].

Currently, EC is diagnosed mainly based on clinical symptoms; physical findings; results of laboratory investigations, transvaginal ultrasound, pelvic ultrasonography, endometrial biopsy with hysteroscopy, and imaging (computed tomography, positron emission tomography/computed tomography, and magnetic resonance imaging); and some biomarkers (e.g., CA125 and HE4) [6–9]. The purpose of these investigations is to examine the endometrial cells, determine the disease extent, and detect the presence/absence of metastasis. Although these methods have good sensitivity for the diagnosis of EC, they have disadvantages, such as poor specificity (particularly transvaginal ultrasound), invasiveness, pain, and high cost. Therefore, improved examination techniques are urgently required, and target genes seem to be appropriate candidates.

Owing to advancements in computer technology and the introduction of sequencing technology, studies have promoted our understanding of cellular and genetic changes during oncogenesis and yielded more targeted and individualized treatment choices [10–12]. Machine learning, a component of artificial intelligence, using computer technology to simulate human intellect, can make predictions using mathematical algorithms after being trained with data. Deep learning, a branch of machine learning, focuses on making forecasts using a multilayer neural network algorithm and can expand model predictions exponentially with increased data volume and dimension, making it suitable for large-scale data analyses. Thus, deep learning can generate meaningful insights and discern relevant traits from genomic data. Genomic analyses have revealed novel biological targets for EC. The genetic bases of cancer progression and therapeutic response have been extensively studied, and the developments of next-generation sequencing and machine learning have yielded opportunities to systematically assess differentially expressed genes (DEGs) [11–13]. Moreover, large public databases, such as the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA), have provided abundant cancer genome sequencing data, which have improved our understanding of molecular changes in oncogenesis. However, due to the lack of multi-omics data, studies on the genomic analysis of EC focusing on gene expression or immune response are few. Since RNA sequencing of tumor tissues is usually performed to characterize gene expression and tumor immune microenvironment (TIME) cells, many datasets have estimated the abundance of DEGs and TIME cells in neoplastic tissue [13–16].

This study was aimed at identifying the signature genes in EC using machine learning, constructing a diagnostic model using an artificial neural network, and verifying the model in three EC cohorts. Finally, the changes in TIME during EC were confirmed.

2. Materials and Methods

2.1. Data Collection and Pre-Processing

Table 1 demonstrates the datasets utilized in this study. The gene expression datasets GSE63678, GSE106191, and GSE115810 were obtained from GEO (https://www.ncbi.nlm.nih.gov/geo/) (accessed on 3 March 2022), merged, and corrected for the batch effect to constitute the training group. The dataset GSE17025 from GEO and the gene expression of EC from TCGA (https://www.cancer.gov/) (accessed on 3 March 2022) were accessed to constitute the test and validation groups, respectively. Our study complied with the publication guidelines laid down by GEO and TCGA. No ethics committee approval was required.

Table 1. Composition of the datasets and component of patients enrolled in this study.

	Train Group			Test Group	Validation Group
	GSE106191	**GSE115810**	**GSE63678**	**GSE17025**	**TCGA**
Sample Count	97	27	35	103	583
Normal	64	3	5	12	35
Cancer	33	24	7	91	548
Enrollment	97	27	12	103	583

2.2. Exploration of DEGs and Functional Enrichment

The 96 DEGs across the EC and para-cancer samples in the training group were calculated using the R package "limma", which employed the empirical Bayesian method and the moderated Wilcox test to assess differences in gene expression. Subsequently, heatmaps and volcanic maps were drawn using the R package with an absolute log2 fold change $\geq$0.8 and an adjusted p-value < 0.05. For the next functional analysis of the 96 DEGs in EC, the R package "clusterProfiler" was used to perform the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGGs) enrichment analyses. The GO analysis mainly comprised the biological process, cellular component, and molecular function. For the functional enrichment analysis, statistical significance was set at $p < 0.05$, and the R packages "enrichplot" and "ggplot2" were used.

2.3. Construction of Metascape and the Protein-Protein Interaction (PPI) Network

In addition, we also analyzed gene sets using the online toolkit WebGestalt (http://www.webgestalt.org/) (accessed on 3 March 2022); performed enrichment analyses using Metascape (http://metascape.org/) (accessed on 3 March 2022), Reactome, and WikipathwayCancer; and investigated a protein–protein interaction (PPI) network using the STRING (https://cn.string-db.org/) (accessed on 3 March 2022) database.

2.4. Selection of the Signature Genes and Construction of the Diagnostic Prediction Model

Random forest analyses were performed, and characteristic DEGs were selected based on the point at which the error of cross validation was the least. The setting seed was 123,456, and the ntree was 500. Subsequently, the characteristic genes were assigned a gene importance score, and those with a score >0.9 were selected and visualized by the R packages "limma" and "pheatmap". Next, we clustered the samples according to the expression of DEGs in the training group and found that the samples were divided into two clusters, similar to carcinoma and paraneoplastic samples.

Subsequently, we assigned scores to the specific DEGs to eliminate batch effects in samples. Up-regulated genes greater than the median value were scored 1, whereas the rest were scored 0; similarly, down-regulated genes lesser than the median value were scored 1, whereas the rest were scored 0. The artificial neural network model for the EC diagnosis was constructed from three types of layers: the input layer, with the scores of 14 genes; the hidden layers, with the scores and weights of genes; and the output layer, with the results for control and experimental samples. The R package "NeuralNetTools" was applied for the procedure with a seed of 12,345,678. Similarly, the selected DEGs and the constructed artificial neural network were applied to the test and validation groups. Unlike the other two groups, the control samples enrolled in the test group comprised tissues of other uterine pathologic types, whereas the samples in the experimental group comprised tissues of early EC. In addition, we constructed a receiver operating characteristic (ROC) curve using the R package "pROC" and assessed the area under the ROC curve (AUC) for the diagnostic model across the three cohorts.

2.5. Identification of TIME

In the analysis of immune cell infiltration, a total of 22 immune cells were identified by the CIBERSORT algorithm and screened using the R packages "e1071", "preprocessCore", and "CIBERSORT.R" at $p < 0.05$. The correlation between the immune cells was calculated using the R package "corrplot." Moreover, the different distribution of immune cells between EC and normal tissues was measured and presented as a violin plot.

3. Results

3.1. DEGs and Functional Enrichment Analysis Results in EC

Based on the filer criteria, a total of 96 DEGs were found between EC and normal samples in the training group and analyzed. There were 19 up-regulated (e.g., MMP12 and CCL20) and 77 down-regulated (e.g., SFP4, OGN, OSR2, FOXL2, and IGFBP4) genes (Figure 1A,B). The top 10 GO terms revealed that the DEGs were mainly involved in collagen-containing extracellular matrix organization and signaling receptor activator activity (Figure 1C). KEGG terms demonstrated that the 96 DEGs were mainly involved in the interleukin-17 (IL-17) signaling pathway, protein digestion and absorption, and transcriptional misregulation in cancer (Figure 1D); thereby playing important roles in inflammatory and immune processes and the occurrence and development of tumors. All enrichment analysis results were closely related to TIME.

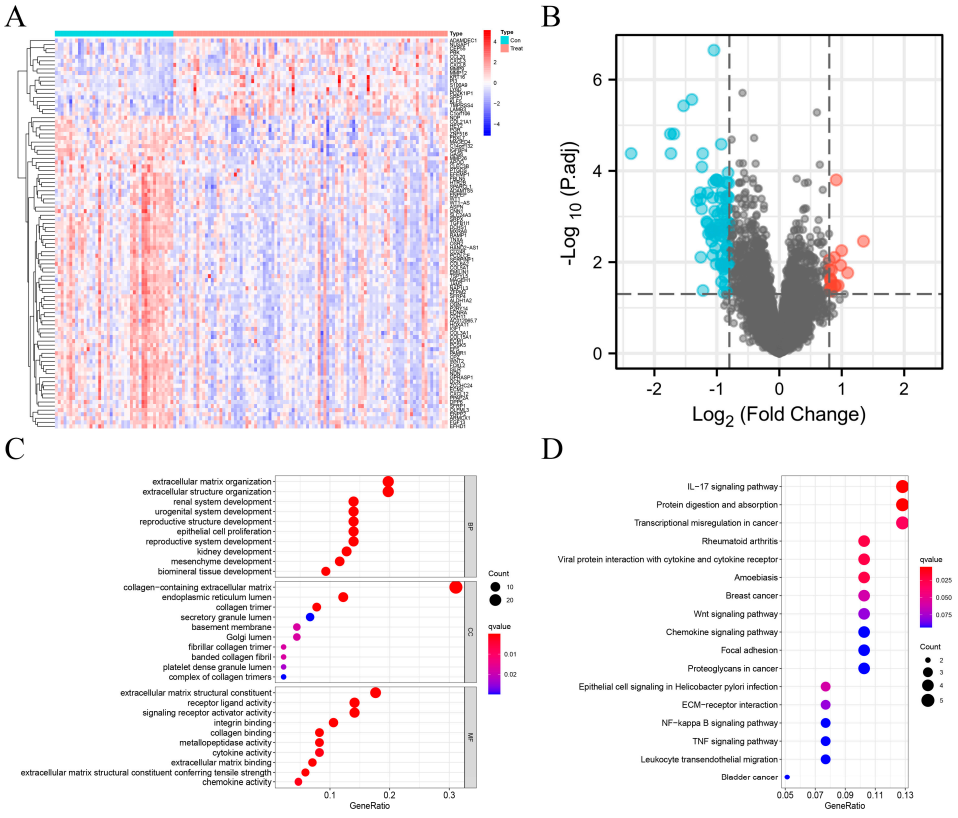


Figure 1. Identification of 96 DEGs in EC in the training group. (**A**) Heatmap of DEGs. The columns represent samples, and rows represent genes. The red color represents up-regulation, and the blue color represents down-regulation. $|\log2FC| > 0.8$, p-value < 0.05. (**B**) Volcanic map of DEGs. The red, blue, and black colors represent up-regulated, down-regulated, and undifferentiated genes, respectively. $|\log2FC| > 0.8$, p-value < 0.05. (**C**) Top 10 biological processes, cellular components, and molecular functions with the most significant p-value. (**D**) All KEGG enrichment results of DEGs.

3.2. Metascape and PPI Network Analysis Results

A network diagram was created based on Metascape analysis. Spots represented functions or pathways. Larger and connected points represented the presence of more similar genes between the functions or pathways. The NABA_CORE_MATRISOME gene set contained many genes encoding extracellular matrix organization and extracellular matrix-associated proteins activated in EC, while the NABA_MATRISOME_ASSPCIATED gene set contained many genes encoding vascular development, tissue morphogenesis, and growth regulation (Figure 2A). Figure 2B shows the top 50 function enrichments. Subsequently, the enrichment analyses of DisGeNET and PaGenBase revealed that the DEGs were primarily specialized in endometrial neoplasms and the uterus (Figure 2C,D), consistent with this study. During the pathogenesis of EC, epigenetic changes in pathogenic genes were mainly regulated by transcription factors EP300, RELA, JUN, SP1, NFKB1, ERG, HDAC1, CEBPA, FOS, and HIF1A (Figure 2E), which play important roles in inflammation, cell proliferation, transformation, differentiation, apoptosis, and immune response. In addition, the PPI network showed a relationship between different genes and proteins in the three sub-modules (Figure 2F). The NABA_CORE_MATRISOME sub-module included COL21A1, COL5A1, COL6A2, COL3A1, and COL15A1, which can identify the structural components of the extracellular matrix to provide tensile strength; the extracellular matrix organization sub-module included SPP1, IGFBP4, GAS6, MXRA8, and SPARCL1, which could enable proteins and/or the extracellular matrix; the NABA_MATRISOME_ASSOCIATED sub-module included P2RY14, CXCL8, CCL20, CXCL3, and CXCL12, which could enable protein binding and chemokine activity.

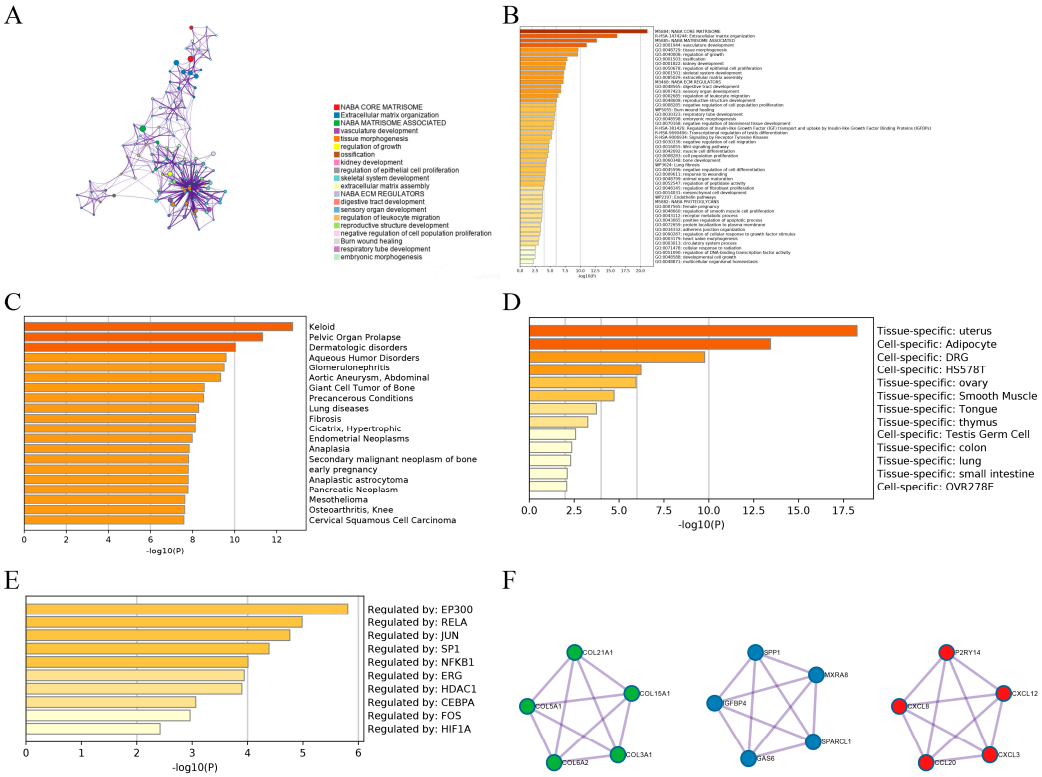


Figure 2. PPI network based on Metascape. (**A**) Network diagrams of the enrichment pathway and process of EC. (**B**) Bar plot of the enrichment pathway and process of EC. (**C**) Bar plot of enrichment on DisGeNET. (**D**) Bar plot of enrichment on PaGenBase. (**E**) Bar chart of enrichment on TRRUST. (**F**) Three sub-modules of PPI.

3.3. Exploration of Characteristic DEGs and Diagnostic Prediction Model of EC

We conducted a random forest analysis to identify the characteristic DEGs. The black line and horizontal and vertical axes represented the error value of the samples, number of trees, and cross-validation error, respectively (Figure 3A). Figure 3B shows the importance of genes. After re-validating DEGs, all 14 EC-signature DEGs with a score >0.9 were enrolled, including three up-regulated (MMP12, MMP9, and ADAMDEC1) and 11 down-regulated (OGN, FOXL2, IGFBP4, DCHS1, ENPP2, ALDH1A2, ADAMTS5, MXRA8, EFEMP1, EFS, and ENPEP genes (Figures 3C and 4). In the diagnostic prediction model, the control and experimental samples were aggregated, which signified that the expression of the pathogenic genes was distinguished between the normal and EC samples (Figure 3D). In addition, for the training, test, and validation groups, the AUCs were 0.928, 0921, and 0.782, respectively, and the accuracies were 0.882, 0.864 and 0.839, respectively (Figure 5A–C, Table 2); implying that the EC diagnostic prediction model could be used as an independent diagnostic predictor of EC.

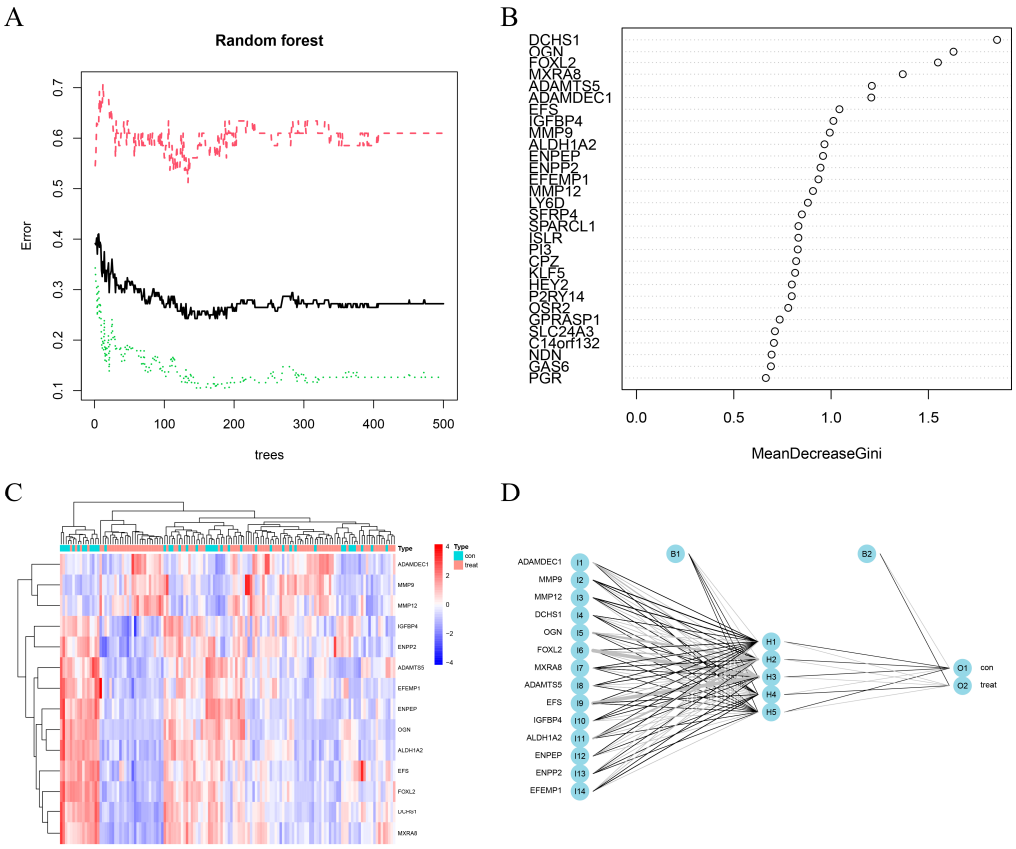


Figure 3. Selection of signature genes by machine learning and construction of a diagnostic prediction model by artificial neural network. (**A**) Construction of random forest. (**B**) Exploring signature genes of EC based on gene importance scores. (**C**) Heatmap of 14 characteristic DEGs. (**D**) Process of constructing artificial neural network.

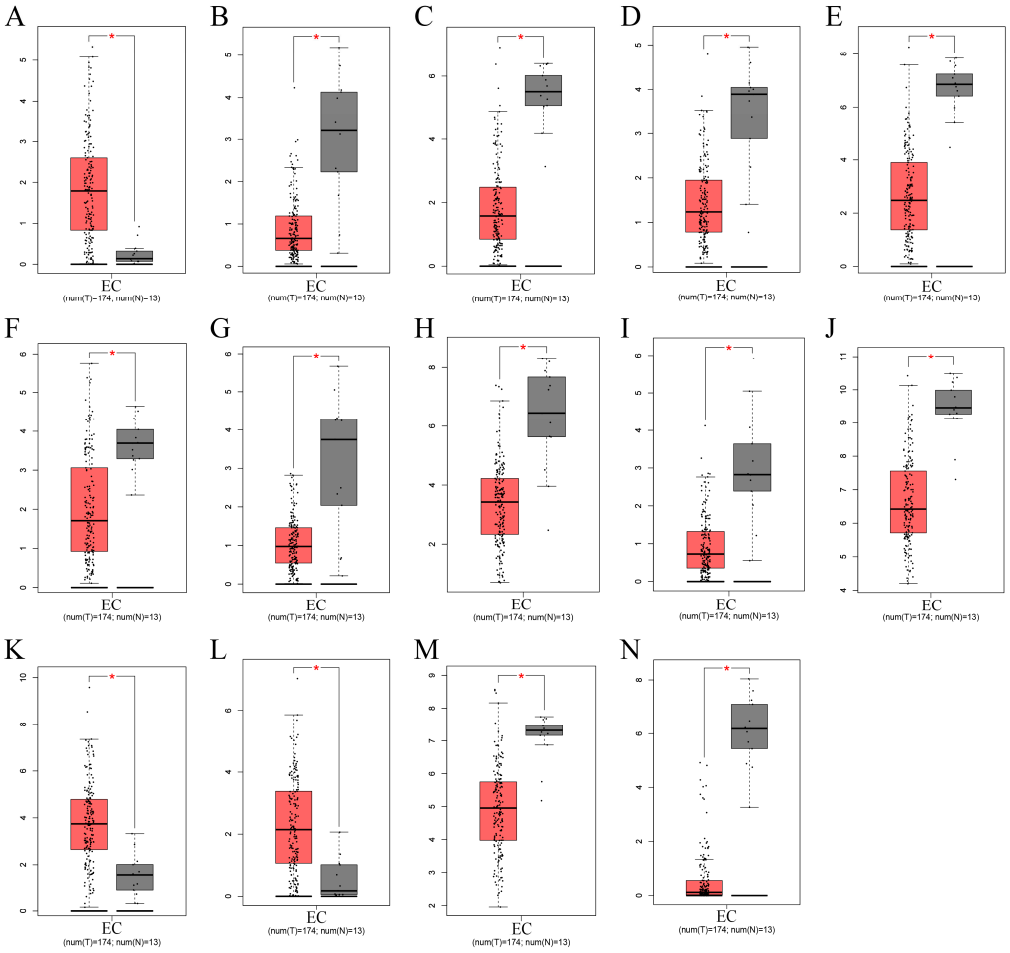


Figure 4. Box diagram of 14 characteristic genes in EC and healthy controls with *p*-value < 0.01. (**A**–**N**) ADAMDEC1, ADAMTS5, ALDH1A2, DCHS1, EFEMP1, EFS, ENPEP, ENPP2, FOXL2, IGFBP4, MMP9, MMP12, MXRA8, and OGN. The red color represents EC, and the black color represents healthy controls. * means *p*-value < 0.01.

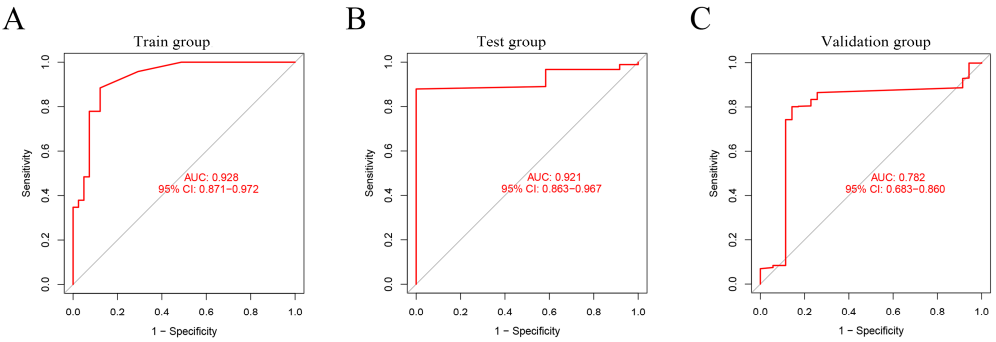


Figure 5. ROC curves of the three groups. (**A**) Training group. (**B**) Test group. (**C**) Validation group.

Table 2. Neural Diagnostic for the training, test and validation cohorts.

		Training Group		Test Group		Validation Group	
		Normal	Cancer	Normal	Cancer	Normal	Cancer
Prediction results	Normal	32	7	12	14	26	85
	Cancer	9	88	0	77	9	463
	Normal Accuracy	0.780		1.000		0.743	
	Cancer Accuracy	0.926		0.846		0.845	
	Accuracy	0.882		0.864		0.839	

3.4. TIME of EC

Figure 6A shows the 22 categories of immunocytes in each sample. Resting and activated mast cells, neutrophils, macrophage M1s, activated NK cells, and eosinophils were relatively abundant in EC. Figure 6B shows the correlation in infiltration of immune cells. The greater the absolute value of the number, the stronger the correlation coefficients, with red and blue colors representing positive and negative correlations, respectively. Activated and resting mast cells showed a strong negative correlation, with a correlation coefficient of −0.54. Activated mast cells and NK cells showed a negative correlation, with a correlation coefficient of −0.43. The activated T cells CD4 and CD8 showed a strong positive correlation, with a correlation coefficient of 0.36 (Figure 6B). In summary, resting and activated mast cells, neutrophils, macrophage M1s, activated NK cells, and eosinophils in EC and normal samples were significantly different (Figure 6A–C); high expressions of activated mast cells, macrophage M1, and neutrophils and low expressions of resting mast cells, activated NK cells, and eosinophils were found in EC.

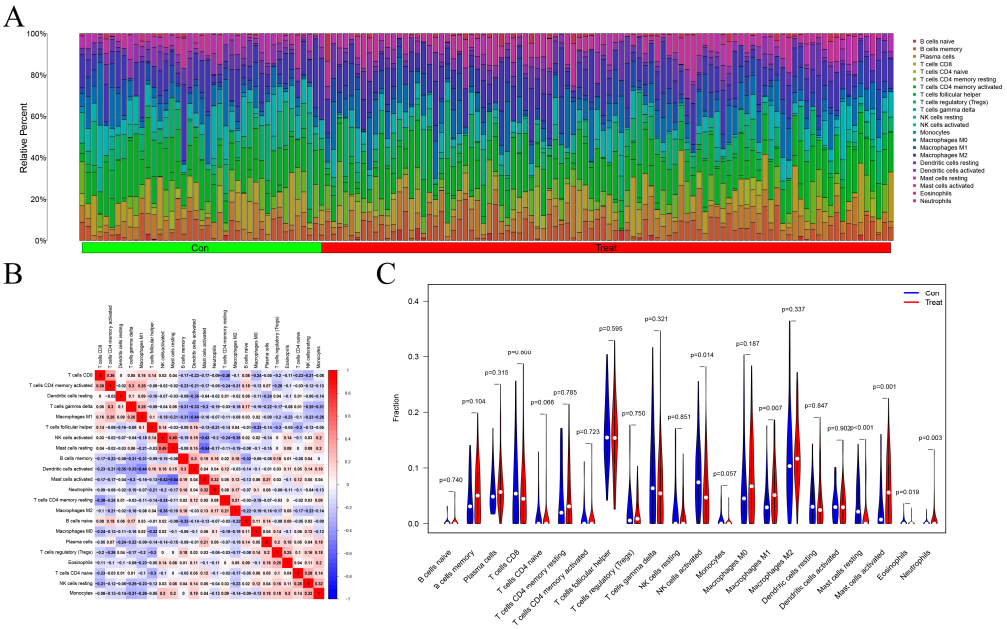


Figure 6. Tumor-immune microenvironment of EC. (**A**) Histogram of 22 types of immune cells in EC and healthy controls. (**B**) Correlation of immune cells in EC. (**C**) Violin image of immune cells.

4. Discussion

At present, EC is diagnosed mainly based on clinical symptoms, physical findings, results of laboratory investigations, and imaging examination. Endometrial biopsy under hysteroscopy seems to be the best method for the diagnosis of benign EC [17,18]. Fertility retention technology can effectively improve the quality of life of gynecological cancer patients, and has become the goal and hope for cancer survivors to live a better life [19]. Studies based on systems biology proteomics have highlighted the exact potential molecular mechanisms associated with SLN and EC grades [20,21]. The aim of

these investigations is to examine the endometrial cells, determine the disease extent, and detect the presence/absence of metastasis. Although the accuracy of the diagnosis and treatment of EC has made great progress in recent years, the molecular mechanism remains unknown. Abnormal gene expression and immune response in TIME play active roles in tumor occurrence, development, invasion, metastasis, and recurrence and are key considerations influencing tumor prognosis [16,22–24]. Endometrial biopsy under hysteroscopy seems to be the best method for the diagnosis of benign EC. In this study, we focused on transcriptional data from GEO and TCGA to identify the complex correlations of the signature genes for EC with a diagnosis and to build a diagnostic prediction model of EC, involving 14 signature genes by random forest and artificial neural network analyses, which distinguished patients with EC from the general population to guide diagnosis and treatment.

We obtained 96 DEGs, including 19 up-regulated and 77 down-regulated genes, and investigated sophisticated biological functions using GO and KEGG analyses in the training group. The outcome indicated that the DEGs were mainly enriched in extracellular matrix and structure organizations, and involved in the IL-17 signaling pathway, protein digestion and absorption, and transcriptional mis-regulation in TIME. These results indicated that changes in gene expression could be conducive to tumor remodeling and promote chronic inflammation, tumor progression, metastasis, and immune escape. We also obtained ten transcription factors, including EP300, RELA, JUN, SP1, NFKB1, ERG, HDAC1, CEBPA, FOS, and HIF1A, which regulated gene expression and played important roles in inflammation, cell proliferation, transformation, differentiation, apoptosis, and immune response [25–30]. In addition, the PPI network mainly showed a relationship between different genes and proteins among the three sub-modules. The NABA_CORE_MATRISOME sub-module comprised COL21A1, COL5A1, COL6A2, COL3A1, and COL15A1, which could identify structural components of the extracellular matrix to provide tensile strength. The extracellular matrix organization sub-module comprised SPP1, IGFBP4, GAS6, MXRA8, and SPARCL1, which could enable proteins and extracellular matrix. The NABA_MATRISOME_ASSOCIATED sub-module comprised P2RY14, CXCL8, CCL20, CXCL3, and CXCL12, which could enable protein binding and chemokine activity.

To obtain a good neural network model, we found 14 characteristic genes for EC by the machine learning method random forest. A diagnostic prediction model for EC was constructed using the artificial neural network, which may be widely applied to the formulation of diagnosis and treatment models for EC. In the model, expressions of MMP12, MMP9, and ADAMDEC1 were increased in EC, and those of OGN, FOXL2, IGFBP4, DCHS1, ENPP2, ALDH1A2, ADAMTS5, MXRA8, EFEMP1, EFS, and ENPEP were decreased in EC. MMP12 and MMP9 were related to cancer development, progression, and survival through various pathological processes and play essential roles in tumor invasion and metastasis [31–34]. Therefore, MMP12 knockdown inhibited proliferation and invasion of nasopharyngeal and lung cancers. Overexpression of ADAMDEC1 is correlated with tumor progression, inflammation, immunotherapeutic response, and a poor prognosis in many cancers [35–37]. Under-expressed OGN and EFS, compared to the normal samples, improved survival, reduced tumor recurrence, and reversed the epithelial to mesenchymal transition by inhibiting EGFR/AKT/Zeb-1 in tumors [38,39]. In a previous study, FOXL2 was considered for molecular diagnostic testing in ovarian adult granulosa cell and microcystic stromal cancers [40]. IGFBP-4 plays an important role in tumor growth regulation by inhibiting IGF actions [41]. Although these feature genes are widely expressed in tumors, according to previous reports, further research is required to clarify the gene function in the pathology of carcinoma, particularly EC. According to the traditional model, EC is divided into types 1 and 2, with certain classic mutations between the two types. Type 1 has mutations in PTEN, ARID1A, PIK3CA, and KRAS, while type 2 has mutations in TP53. Currently, EC is mainly diagnosed based on uterine curettage or biopsy findings. Some data suggest that the susceptibility of endometrial biopsy for EC is 52–94% [42–46].

The accuracy of differentiation of EC in other studies was slightly lower than our model (Table 2) [21,47,48]. Particularly, the test group comprised non-cancerous uterine pathologic types and early EC. The diagnostic rate of 100% in the non-cancerous group confirmed the efficacy of our diagnostic model for early EC in the test group. Thus, the model in the training and test groups showed a good effect, while that in the validation group showed an average effect. The 14 feature genes were key potential biomarkers of EC, but further studies are required to verify the results.

In addition, we also focused on TIME of EC and found that high expressions of activated mast cells, macrophage M1s, and neutrophils, and low expressions of resting mast cells, NK cells, and activated eosinophils played vital roles in EC. Multiple studies have documented that mast cells, neutrophils, macrophage M1s, NK cells, and eosinophils play a protective role during cancer progression, such as inflammatory responses, development of blood vessels, apoptosis, proliferation, invasion, and immune evasion [49–56].

However, this study has some limitations. First, the RNA sequencing data were only obtained from public databases. Second, although we validated the predictive performance of the EC diagnosis, further investigation is required for accurate validation. Further basic and clinical studies should be performed to validate the outcome and find a simpler, faster, and more economic approach.

5. Conclusions

In our study, we identified 14 genes involved in EC, verified them, based on GEO and TCGA, and established a robust diagnostic prediction model for EC through an artificial neural network, which was promising for the exploration of new diagnostic tools. The diagnostic model possessed excellent sensitivity and specificity, demonstrating the capability of diagnosing early EC. We also discovered that activated and resting mast cells were important and inversely correlated in EC. These results could serve as a basis for extensive cohorts in the future.

Author Contributions: All authors made important contributions to the study design, data acquisition, and data analysis; formal analysis, D.Z. and Z.W.; investigation, D.Z. and Z.W.; writing—original draft preparation, D.Z. and Z.Z.; writing—review and editing, D.Z. and Z.D.; visualization, M.W., T.Z. and J.Z.; supervision W.Z. and Y.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

Endometrial Carcinoma: EC; Gene Expression Omnibus: GEO; The Cancer Genome Atlas: TCGA; Differentially Expressed Genes: DEGs; Tumor Immune Microenvironment: TIME; Support Vector Machine: SVM; Protein–Protein Interaction: PPI; Gene Ontology: GO; Kyoto Encyclopedia of Genes and Genomes: KEGG Receiver Operating Characteristic: ROC; Area Under Curve: AUC.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
2. Koh, W.J.; Abu-Rustum, N.R.; Bean, S.; Bradley, K.; Campos, S.M.; Cho, K.R.; Chon, H.S.; Chu, C.; Cohn, D.; Crispens, M.A.; et al. Uterine Neoplasms, Version 1.2018, NCCN Clinical Practice Guidelines in Oncology. *J. Natl. Compr. Cancer Netw.* **2018**, *16*, 170–199. [CrossRef] [PubMed]

3. Brooks, R.A.; Fleming, G.F.; Lastra, R.R.; Lee, N.K.; Moroney, J.W.; Son, C.H.; Tatebe, K.; Veneris, J.L. Current recommendations and recent progress in endometrial cancer. *CA Cancer J. Clin.* **2019**, *69*, 258–279. [CrossRef] [PubMed]
4. Bolivar, A.M.; Luthra, R.; Mehrotra, M.; Chen, W.; Barkoh, B.A.; Hu, P.; Zhang, W.; Broaddus, R.R. Targeted next-generation sequencing of endometrial cancer and matched circulating tumor DNA: Identification of plasma-based, tumor-associated mutations in early stage patients. *Mod. Pathol.* **2019**, *32*, 405–414. [CrossRef]
5. Bell, D.W.; Ellenson, L.H. Molecular Genetics of Endometrial Carcinoma. *Annu. Rev. Pathol.* **2019**, *14*, 339–367. [CrossRef]
6. McKenney, J.K.; Longacre, T.A. Low-grade endometrial adenocarcinoma: A diagnostic algorithm for distinguishing atypical endometrial hyperplasia and other benign (and malignant) mimics. *Adv. Anat. Pathol.* **2009**, *16*, 1–22. [CrossRef] [PubMed]
7. Gimpelson, R.J.; Rappold, H.O. A comparative study between panoramic hysteroscopy with directed biopsies and dilatation and curettage. A review of 276 cases. *Am. J. Obstet. Gynecol.* **1988**, *158*, 489–492. [CrossRef]
8. Antonsen, S.L.; Jensen, L.N.; Loft, A.; Berthelsen, A.K.; Costa, J.; Tabor, A.; Qvist, I.; Hansen, M.R.; Fisker, R.; Andersen, E.S.; et al. MRI, PET/CT and ultrasound in the preoperative staging of endometrial cancer—A multicenter prospective comparative study. *Gynecol. Oncol.* **2013**, *128*, 300–308. [CrossRef]
9. Duk, J.M.; Aalders, J.G.; Fleuren, G.J.; de Bruijn, H.W. CA 125: A useful marker in endometrial carcinoma. *Am. J. Obstet. Gynecol.* **1986**, *155*, 1097–1102. [CrossRef]
10. Sone, K.; Toyohara, Y.; Taguchi, A.; Miyamoto, Y.; Tanikawa, M.; Uchino-Mori, M.; Iriyama, T.; Tsuruga, T.; Osuga, Y. Application of artificial intelligence in gynecologic malignancies: A review. *J. Obstet. Gynaecol. Res.* **2021**, *47*, 2577–2585. [CrossRef]
11. Hamamoto, R. Application of Artificial Intelligence for Medical Research. *Biomolecules* **2021**, *11*, 90. [CrossRef] [PubMed]
12. Hamamoto, R.; Komatsu, M.; Takasawa, K.; Asada, K.; Kaneko, S. Epigenetics Analysis and Integrated Analysis of Multiomics Data, Including Epigenetic Data, Using Artificial Intelligence in the Era of Precision Medicine. *Biomolecules* **2019**, *10*, 62. [CrossRef] [PubMed]
13. Welford, S.M.; Gregg, J.; Chen, E.; Garrison, D.; Sorensen, P.H.; Denny, C.T.; Nelson, S.F. Detection of differentially expressed genes in primary tumor tissues using representational differences analysis coupled to microarray hybridization. *Nucleic Acids Res.* **1998**, *26*, 3059–3065. [CrossRef] [PubMed]
14. Albaradei, S.; Thafar, M.; Alsaedi, A.; Van Neste, C.; Gojobori, T.; Essack, M.; Gao, X. Machine learning and deep learning methods that use omics data for metastasis prediction. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 5008–5018. [CrossRef]
15. Jiménez-Sánchez, D.; Ariz, M.; Chang, H.; Matias-Guiu, X.; de Andrea, C.E.; Ortiz-de-Solórzano, C. NaroNet: Discovery of tumor microenvironment elements from highly multiplexed images. *Med. Image Anal.* **2022**, *78*, 102384. [CrossRef]
16. Ruan, T.; Wan, J.; Song, Q.; Chen, P.; Li, X. Identification of a Novel Epithelial-Mesenchymal Transition-Related Gene Signature for Endometrial Carcinoma Prognosis. *Genes* **2022**, *13*, 216. [CrossRef]
17. Vitale, S.G.; Riemma, G.; Carugno, J.; Chiofalo, B.; Vilos, G.A.; Cianci, S.; Budak, M.S.; Lasmar, B.P.; Raffone, A.; Kahramanoglu, I. Hysteroscopy in the management of endometrial hyperplasia and cancer in reproductive aged women: New developments and current perspectives. *Transl. Cancer Res.* **2020**, *9*, 7767–7777. [CrossRef]
18. Prip, C.M.; Stentebjerg, M.; Bennetsen, M.H.; Petersen, L.K.; Bor, P. Risk of atypical hyperplasia and endometrial carcinoma after initial diagnosis of non-atypical endometrial hyperplasia: A long-term follow-up study. *PLoS ONE* **2022**, *17*, e0266339. [CrossRef]
19. La Rosa, V.L.; Garzon, S.; Gullo, G.; Fichera, M.; Sisti, G.; Gallo, P.; Riemma, G.; Schiattarella, A. Fertility preservation in women affected by gynaecological cancer: The importance of an integrated gynaecological and psychological approach. *Ecancermedicalscience* **2020**, *14*, 1035. [CrossRef]
20. Aboulouard, S.; Wisztorski, M.; Duhamel, M.; Saudemont, P.; Cardon, T.; Narducci, F.; Lemaire, A.S.; Kobeissy, F.; Leblanc, E.; Fournier, I.; et al. In-depth proteomics analysis of sentinel lymph nodes from individuals with endometrial cancer. *Cell Rep. Med.* **2021**, *2*, 100318. [CrossRef]
21. Della Corte, L.; Giampaolino, P.; Mercorio, A.; Riemma, G.; Schiattarella, A.; De Franciscis, P.; Bifulco, G. Sentinel lymph node biopsy in endometrial cancer: State of the art. *Transl. Cancer Res.* **2020**, *9*, 7725–7733. [CrossRef] [PubMed]
22. Rousset-Rouviere, S.; Rochigneux, P.; Chrétien, A.S.; Fattori, S.; Gorvel, L.; Provansal, M.; Lambaudie, E.; Olive, D.; Sabatier, R. Endometrial Carcinoma: Immune Microenvironment and Emerging Treatments in Immuno-Oncology. *Biomedicines* **2021**, *9*, 632. [CrossRef] [PubMed]
23. Zheng, M.; Hu, Y.; Gou, R.; Li, S.; Nie, X.; Li, X.; Lin, B. Development of a seven-gene tumor immune microenvironment prognostic signature for high-risk grade III endometrial cancer. *Mol. Ther. Oncolytics* **2021**, *22*, 294–306. [CrossRef]
24. Chen, Y.; Lee, K.; Liang, Y.; Qin, S.; Zhu, Y.; Liu, J.; Yao, S. A Cholesterol Homeostasis-Related Gene Signature Predicts Prognosis of Endometrial Cancer and Correlates With Immune Infiltration. *Front. Genet.* **2021**, *12*, 763537. [CrossRef] [PubMed]
25. Ahn, S.H.; Edwards, A.K.; Singh, S.S.; Young, S.L.; Lessey, B.A.; Tayade, C. IL-17A Contributes to the Pathogenesis of Endometriosis by Triggering Proinflammatory Cytokines and Angiogenic Growth Factors. *J. Immunol.* **2015**, *195*, 2591–2600. [CrossRef]
26. Miossec, P.; Korn, T.; Kuchroo, V.K. Interleukin-17 and type 17 helper T cells. *N. Engl. J. Med.* **2009**, *361*, 888–898. [CrossRef]
27. Cornelius, D.C.; Lamarca, B. TH17- and IL-17- mediated autoantibodies and placental oxidative stress play a role in the pathophysiology of pre-eclampsia. *Minerva Ginecol.* **2014**, *66*, 243–249.
28. Liu, L.; Chen, F.; Xiu, A.; Du, B.; Ai, H.; Xie, W. Identification of Key Candidate Genes and Pathways in Endometrial Cancer by Integrated Bioinformatical Analysis. *Asian Pac. J. Cancer Prev.* **2018**, *19*, 969–975.
29. Gorczynski, R.M. IL-17 Signaling in the Tumor Microenvironment. *Adv. Exp. Med. Biol.* **2020**, *1240*, 47–58.

30. Lee, T.I.; Young, R.A. Transcriptional regulation and its misregulation in disease. *Cell* **2013**, *152*, 1237–1251. [CrossRef]
31. Gialeli, C.; Theocharis, A.D.; Karamanos, N.K. Roles of matrix metalloproteinases in cancer progression and their pharmacological targeting. *FEBS J.* **2011**, *278*, 16–27. [CrossRef] [PubMed]
32. Zheng, J.; Chu, D.; Wang, D.; Zhu, Y.; Zhang, X.; Ji, G.; Zhao, H.; Wu, G.; Du, J.; Zhao, Q. Matrix metalloproteinase-12 is associated with overall survival in Chinese patients with gastric cancer. *J. Surg. Oncol.* **2013**, *107*, 746–751. [CrossRef]
33. Brun, J.L.; Cortez, A.; Lesieur, B.; Uzan, S.; Rouzier, R.; Daraï, E. Expression of MMP-2, -7, -9, MT1-MMP and TIMP-1 and -2 has no prognostic relevance in patients with advanced epithelial ovarian cancer. *Oncol. Rep.* **2012**, *27*, 1049–1057. [CrossRef] [PubMed]
34. Wang, X.; Chen, T. CUL4A regulates endometrial cancer cell proliferation, invasion and migration by interacting with CSN6. *Mol. Med. Rep.* **2021**, *23*, 23. [CrossRef] [PubMed]
35. Liu, X.; Huang, H.; Li, X.; Zheng, X.; Zhou, C.; Xue, B.; He, J.; Zhang, Y.; Liu, L. Knockdown of ADAMDEC1 inhibits the progression of glioma in vitro. *Histol. Histopathol.* **2020**, *35*, 997–1005. [PubMed]
36. Zhu, W.; Shi, L.; Gong, Y.; Zhuo, L.; Wang, S.; Chen, S.; Zhang, B.; Ke, B. Upregulation of ADAMDEC1 correlates with tumor progression and predicts poor prognosis in non-small cell lung cancer (NSCLC) via the PI3K/AKT pathway. *Thorac. Cancer* **2022**, *13*, 1027–1039. [CrossRef]
37. Ahn, S.B.; Sharma, S.; Mohamedali, A.; Mahboob, S.; Redmond, W.J.; Pascovici, D.; Wu, J.X.; Zaw, T.; Adhikari, S.; Vaibhav, V.; et al. Potential early clinical stage colorectal cancer diagnosis using a proteomics blood test panel. *Clin. Proteom.* **2019**, *16*, 34. [CrossRef]
38. Lomnytska, M.I.; Becker, S.; Hellman, K.; Hellström, A.C.; Souchelnytskyi, S.; Mints, M.; Hellman, U.; Andersson, S.; Auer, G. Diagnostic protein marker patterns in squamous cervical cancer. *Proteom. Clin. Appl.* **2010**, *4*, 17–31. [CrossRef]
39. Hu, X.; Li, Y.Q.; Li, Q.G.; Ma, Y.L.; Peng, J.J.; Cai, S.J. Osteoglycin (OGN) reverses epithelial to mesenchymal transition and invasiveness in colorectal cancer via EGFR/Akt pathway. *J. Exp. Clin. Cancer Res. CR* **2018**, *37*, 41. [CrossRef]
40. Rabban, J.T.; Karnezis, A.N.; Devine, W.P. Practical roles for molecular diagnostic testing in ovarian adult granulosa cell tumour, Sertoli-Leydig cell tumour, microcystic stromal tumour and their mimics. *Histopathology* **2020**, *76*, 11–24. [CrossRef]
41. Baxter, R.C. IGF binding proteins in cancer: Mechanistic and clinical insights. *Nat. Rev. Cancer* **2014**, *14*, 329–341. [CrossRef] [PubMed]
42. Long, S. Endometrial Biopsy: Indications and Technique. *Primary care* **2021**, *48*, 555–567. [CrossRef] [PubMed]
43. Reijnen, C.; Visser, N.C.M.; Bulten, J.; Massuger, L.; van der Putten, L.J.M.; Pijnenborg, J.M.A. Diagnostic accuracy of endometrial biopsy in relation to the amount of tissue. *J. Clin. Pathol.* **2017**, *70*, 941–946. [CrossRef] [PubMed]
44. Kunaviktikul, K.; Suprasert, P.; Khunamornpong, S.; Settakorn, J.; Natpratan, A. Accuracy of the Wallach Endocell endometrial cell sampler in diagnosing endometrial carcinoma and hyperplasia. *J. Obstet. Gynaecol. Res.* **2011**, *37*, 483–488. [CrossRef] [PubMed]
45. Guido, R.S.; Kanbour-Shakir, A.; Rulin, M.C.; Christopherson, W.A. Pipelle endometrial sampling. Sensitivity in the detection of endometrial cancer. *J. Reprod. Med.* **1995**, *40*, 553–555.
46. Laban, M.; Nassar, S.; Elsayed, J.; Hassanin, A.S. Correlation between pre-operative diagnosis and final pathological diagnosis of endometrial malignancies; impact on primary surgical treatment. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **2021**, *263*, 100–105. [CrossRef]
47. Della Corte, L.; Vitale, S.G.; Foreste, V.; Riemma, G.; Ferrari, F.; Noventa, M.; Liberto, A.; De Franciscis, P.; Tesarik, J. Novel diagnostic approaches to intrauterine neoplasm in fertile age: Sonography and hysteroscopy. *Off. J. Soc. Minim. Invasive Ther.* **2021**, *30*, 288–295. [CrossRef]
48. Heremans, R.; Van den Bosch, T.; Valentin, L.; Wynants, L.; Pascual, M.A.; Fruscio, R.; Testa, A.C.; Buonomo, F.; Guerriero, S.; Epstein, E.; et al. Ultrasound features of endometrial pathology in women without abnormal uterine bleeding: Results from the International Endometrial Tumor Analysis Study (IETA3). *Ultrasound Obstet. Gynecol.* **2022**. [CrossRef]
49. Johansson, A.; Rudolfsson, S.; Hammarsten, P.; Halin, S.; Pietras, K.; Jones, J.; Stattin, P.; Egevad, L.; Granfors, T.; Wikström, P.; et al. Mast cells are novel independent prognostic markers in prostate cancer and represent a target for therapy. *Am. J. Pathol.* **2010**, *177*, 1031–1041. [CrossRef]
50. Sinnamon, M.J.; Carter, K.J.; Sims, L.P.; Lafleur, B.; Fingleton, B.; Matrisian, L.M. A protective role of mast cells in intestinal tumorigenesis. *Carcinogenesis* **2008**, *29*, 880–886. [CrossRef]
51. Fleischmann, A.; Schlomm, T.; Köllermann, J.; Sekulic, N.; Huland, H.; Mirlacher, M.; Sauter, G.; Simon, R.; Erbersdobler, A. Immunological microenvironment in prostate cancer: High mast cell densities are associated with favorable tumor characteristics and good prognosis. *Prostate* **2009**, *69*, 976–981. [CrossRef] [PubMed]
52. Coffelt, S.B.; Wellenstein, M.D.; de Visser, K.E. Neutrophils in cancer: Neutral no more. *Nat. Rev. Cancer* **2016**, *16*, 431–446. [CrossRef] [PubMed]
53. Shojaei, F.; Singh, M.; Thompson, J.D.; Ferrara, N. Role of Bv8 in neutrophil-dependent angiogenesis in a transgenic model of cancer progression. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 2640–2645. [CrossRef] [PubMed]
54. Spiegel, A.; Brooks, M.W.; Houshyar, S.; Reinhardt, F.; Ardolino, M.; Fessler, E.; Chen, M.B.; Krall, J.A.; DeCock, J.; Zervantonakis, I.K.; et al. Neutrophils Suppress Intraluminal NK Cell-Mediated Tumor Cell Clearance and Enhance Extravasation of Disseminated Carcinoma Cells. *Cancer Discov.* **2016**, *6*, 630–649. [CrossRef]

55. Boutilier, A.J.; Elsawa, S.F. Macrophage Polarization States in the Tumor Microenvironment. *Int. J. Mol. Sci.* **2021**, *22*, 6995. [CrossRef]
56. Jhunjhunwala, S.; Hammer, C.; Delamarre, L. Antigen presentation in cancer: Insights into tumour immunogenicity and immune evasion. *Nat. Rev. Cancer* **2021**, *21*, 298–312. [CrossRef]

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 27 June 2024
DOI 10.3389/fimmu.2024.1418508

OPEN ACCESS

EDITED BY
Pengpeng Zhang,
Nanjing Medical University, China

REVIEWED BY
Chunyan Niu,
Southeast University, China
Yuquan Chen,
Monash University, Australia

*CORRESPONDENCE
Jiangtao Fan
jt_fan2018@163.com
Jingxin Mao
2230040@cqmpc.edu.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 16 April 2024
ACCEPTED 12 June 2024
PUBLISHED 27 June 2024

CITATION
Wei C, Lin S, Huang Y, Wei Y, Mao J and Fan J (2024) Integrated machine learning identifies a cellular senescence-related prognostic model to improve outcomes in uterine corpus endometrial carcinoma. *Front. Immunol.* 15:1418508. doi: 10.3389/fimmu.2024.1418508

Integrated machine learning identifies a cellular senescence-related prognostic model to improve outcomes in uterine corpus endometrial carcinoma

Changqiang Wei[1†], Shanshan Lin[1†], Yanrong Huang[1†], Yiyun Wei[1], Jingxin Mao[2*] and Jiangtao Fan[1*]

[1]Department of Obstetrics and Gynecology, The First Affiliated Hospital of Guangxi Medical University, Guangxi, China, [2]Department of Science and Technology Industry, Chongqing Medical and Pharmaceutical College, Chongqing, China

Background: Uterine Corpus Endometrial Carcinoma (UCEC) stands as one of the prevalent malignancies impacting women globally. Given its heterogeneous nature, personalized therapeutic approaches are increasingly significant for optimizing patient outcomes. This study investigated the prognostic potential of cellular senescence genes(CSGs) in UCEC, utilizing machine learning techniques integrated with large-scale genomic data.

Methods: A comprehensive analysis was conducted using transcriptomic and clinical data from 579 endometrial cancer patients sourced from the Cancer Genome Atlas (TCGA). A subset of 503 CSGs was assessed through weighted gene co-expression network analysis (WGCNA) alongside machine learning algorithms, including Gaussian Mixture Model (GMM), support vector machine - recursive feature elimination (SVM-RFE), Random Forest, and eXtreme Gradient Boosting (XGBoost), to identify key differentially expressed cellular senescence genes. These genes underwent further analysis to construct a prognostic model.

Results: Our analysis revealed two distinct molecular clusters of UCEC with significant differences in tumor microenvironment and survival outcomes. Utilizing cellular senescence genes, a prognostic model effectively stratified patients into high-risk and low-risk categories. Patients in the high-risk group exhibited compromised overall survival and presented distinct molecular and immune profiles indicative of tumor progression. Crucially, the prognostic model demonstrated robust predictive performance and underwent validation in an independent patient cohort.

Conclusion: The study emphasized the significance of cellular senescence genes in UCEC progression and underscored the efficacy of machine learning in developing reliable prognostic models. Our findings suggested that targeting cellular senescence holds promise as a strategy in personalized UCEC treatment, thus warranting further clinical investigation.

KEYWORDS

UCEC, cellular senescence, machine learning, MYBL2, CPEB1

1 Introduction

Uterine Corpus Endometrial Carcinoma (UCEC) stands as one of the most prevalent malignancies in gynecology. In China, its incidence ranks second only to cervical cancer (1). In 2023, an estimated 66,200 new cases and 13,030 deaths are projected in the United States (2).

The pathogenesis and classification of UCEC have garnered considerable attention in medical research. It is primarily categorized into two types based on biological characteristics and clinical behavior: Type I (estrogen-dependent) and Type II (non-estrogen dependent) endometrial carcinoma. Recent studies have further delineated it into four molecular subtypes: POLE ultramutated, microsatellite instability, copy-number stability, and p53 abnormal types (3). This molecular classification enriches our comprehension of UCEC heterogeneity and forms the basis for devising personalized treatment strategies (4).

Early-stage endometrial cancer commonly involves total hysterectomy and bilateral salpingo-oophorectomy (5), whereas in cases of advanced or recurrent endometrial cancer, surgery remains crucial but must be supplemented with systemic treatments such as chemotherapy, immunotherapy, targeted therapy, and endocrine therapy (6). Recent studies have concentrated on molecular markers like mutations in the PTEN, PIK3CA, ARID1A, and KRAS genes, prevalent in Type I endometrial cancers, which foster tumor growth and survival (7). Type II cancers often manifest mutations in the p53 gene and amplification of the HER2 gene (8). These findings aid in delineating distinct biological features and therapeutic targets for various tumor types.

Targeted therapies, including PI3K and mTOR inhibitors, have become essential in UCEC treatment, significantly improving outcomes for certain patients (9). For individuals exhibiting microsatellite instability or mismatch repair deficiencies, immune checkpoint inhibitors like PD-1/PD-L1 present novel therapeutic possibilities (10). The efficacy of these strategies highlights the significance of personalized medicine in UCEC treatment. However, challenges persist in precisely identifying eligible patients and devising novel medications.

Cellular senescence constitutes a multifaceted biological process involving alterations in gene expression, DNA damage accumulation, protein function loss, and cell cycle arrest (11). Serving as a critical tumor-suppressing mechanism, it inhibits cancer by constraining the proliferation of damaged or mutated cells (12). Nonetheless, the accumulation of senescent cells can foster tumor progression via the secretion of pro-inflammatory and pro-tumorigenic factors (13). Studies have demonstrated the pivotal roles of senescence-associated genes, such as p53, RB, and PTEN, in cancer development (11). Targeting SASP factors presents a novel perspective for certain cancer treatments (14, 15).

Studies utilizing public databases such as TCGA and Gene Expression Omnibus have pinpointed specific genes linked to the prognosis and treatment responses of UCEC (16–18). These genes can potentially serve as novel biomarkers for refining prognostic models. Currently, research on cellular senescence related to UCEC remains limited. Employing advanced bioinformatics to investigate the relationship between cellular senescence genes and UCEC is imperative for patient stratification and the identification of new therapeutic targets and immune treatment strategies.

2 Materials and methods

2.1 Data and patient collection

Figure 1 illustrates the methodology employed in this research. Transcriptomic and clinical data for 579 endometrial cancer patients, comprising 544 UCEC cases and 35 control subjects, were obtained from the TCGA database (https://portal.gdc.cancer.gov/). A total of 503 cellular senescence genes were sourced from the CSGene database (https://csgene.bioinfominzhao.org/index.html, Supplementary Table 1).

Furthermore, 20 endometrial cancer tissues and 20 non-cancerous endometrial tissues were collected from the First Affiliated Hospital of Guangxi Medical University. All UCEC diagnoses were confirmed by experienced pathologists, with pertinent clinical details provided in Supplementary Table 2. Following surgery, tissues were promptly transferred to a petri dish using forceps and rinsed thoroughly with physiological saline to eliminate surrounding blood clots. Subsequently, approximately 5g samples were dissected using a surgical blade for subsequent RT-

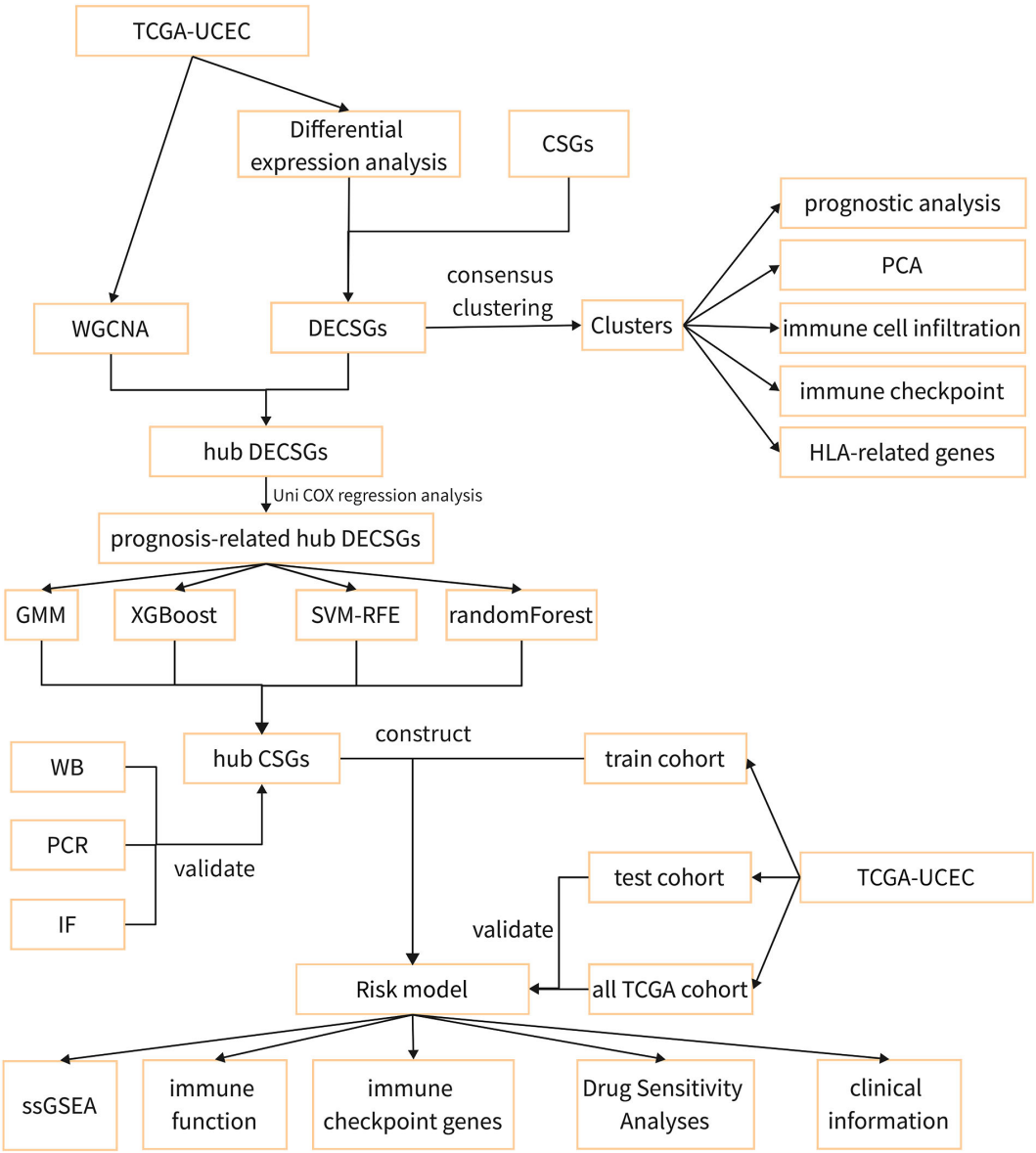


FIGURE 1
The flow diagram of the study.

qPCR and Western blot experiments. Additionally, roughly 10g of tissue was placed in 4% paraformaldehyde fixative, fixed for 24 hours, and then subjected to dehydration and paraffin embedding for sectioning. The study received ethical approval (No. 2023-S033–01), and all participants provided informed consent before undergoing surgery.

2.2 Differential expression analysis

In this study, we utilized the "limma" package (19) in R software to perform differential expression analysis on the UCEC dataset. The filtering criteria were set as: $|\log_2 \text{FoldChange}| \geq 1.5$, and $P<0.05$. Subsequently, the differentially expressed genes and cellular senescence genes were intersected to yield a series of Differentially Expressed Cellular Senescence Genes (DECSGs).

2.3 Consensus clustering and subtype analysis

To identify UCEC subtypes associated with DECSGs, we utilized the "ConsensusClusterPlus" R package for consensus clustering analysis (20). This approach evaluated consistency across multiple clustering runs to determine a more stable final clustering structure, commonly employed in data analysis and bioinformatics. The clustering criteria were as follows: enhanced correlation within subtypes post-clustering, and weakened

correlation between subtypes. We ensured the reliability of our results through 1,000 iterations and utilized the Probably Approximately Correct (PAC) method to determine the optimal number of clusters. Specifically, the PAC method initially generated a set of random datasets and conducted cluster analysis on these datasets to obtain a range of random cluster numbers. The PAC value quantified the dissimilarity between observed clustering results and random clustering results. A higher PAC value indicated greater dissimilarity between the observed clustering structure and random results, indicating a more robust and reliable clustering structure.

Subsequently, we employed principal component analysis (PCA) to discern variations in gene expression patterns among the clusters. Additionally, we conducted differential expression analysis across the clusters and utilized the "ClusterProfiler" (21) and "org.Hs.eg.db" packages to explore potential biological mechanisms through Gene Ontology (GO), Kyoto Encyclopedia of Genomes (KEGG), and Gene Set Enrichment Analysis (GSEA). Furthermore, the "survival" and "survminer" packages (22) were utilized to analyze the overall survival (OS) and progression-free survival (PFS) rates across the different clusters. The tumor microenvironment (TME) of endometrial cancer was assessed using the "estimate" package to understand its characteristics deeply. Based on the "CIBERSORT" package (23), we analyzed the infiltration levels of 22 immune cell types to identify differences in immune cell infiltration across clusters. Lastly, we investigated the expression differences in key immune checkpoint genes and human leukocyte antigen (HLA)-related genes between clusters. This exploration aimed to elucidate mechanisms by which tumors evade immune surveillance, providing valuable insights for the development of novel immunotherapeutic strategies.

2.4 Co-expression network construction

WGCNA was conducted using the "WGCNA" package (24) to construct a scale-free network associated with clinical phenotypes. The process commenced with hierarchical clustering to filter the cases, followed by the selection of an appropriate soft threshold to construct a weighted adjacency matrix. This matrix was then transformed into a topological overlap matrix (TOM), represented with colors and module eigengenes. Additionally, the Pearson correlation coefficient between the module eigengenes and clinical features was calculated to unveil potential links between gene expression patterns and clinical manifestations.

2.5 Cox regression analysis and machine learning algorithms

In this study, we intersected genes from key modules identified by WGCNA with DECSGs to pinpoint key DECSGs. Patients from the TCGA database with complete clinical information and survival times exceeding 30 days were selected for univariate Cox regression analysis to identify prognostically relevant DECSGs.

To accurately identify hub genes associated with UCEC, we employed four machine learning algorithms: GMM, SVM-RFE, Random Forest, and XGBoost. Firstly, GMM analysis was conducted utilizing the "SimDesign" package (25). This method examined the probability distribution of gene expression data and fit it to multiple Gaussian distributions, revealing complex underlying biological information. Subsequently, the SVM-RFE method (26) was implemented using the "e1071," "kernlab," and "caret" packages. This technique constructed a model based on SVM and optimized the feature set by recursively removing the least impactful features. Next, we employed the Random Forest algorithm via the "randomForest" package and the XGBoost algorithm using the "xgboost" package (27, 28). Random Forest is a robust ensemble learning algorithm that builds multiple decision trees and combines their predictions to enhance model accuracy and robustness, widely utilized in classification and regression tasks. XGBoost is an efficient ensemble learning algorithm that incrementally constructs decision trees and corrects errors to optimize model performance, identifying core features. The common genes identified by these algorithms were determined to be the core DECSGs. Finally, the relationship between these core DECSGs and the prognosis of endometrial cancer was analyzed using the external survival prognosis database Kaplan-Meier Plotter (https://kmplot.com/analysis/index.php?p=background).

2.6 Construction and validation of the cellular senescence-relate risk score model

UCEC samples were randomly divided into a training set and a testing set at a ratio of 7:3. Based on the expression of key DECSGs, a prognostic model was constructed within the training set using the LASSO Cox regression method.

This methodology entails an initial fitting of gene expression data and survival time via LASSO regression, followed by cross-validation utilizing the "cv.glmnet" function. Subsequently, the "coef" function is utilized to extract and compute the weights of the selected genes within the model. The model predicts patient survival prognosis through the calculation of a risk score, formulated as: Risk score = Σ (Xi*Yi), where X represents the coefficient of each gene in the model, and Y denotes the expression level of the corresponding gene. Within the training set, UCEC samples were stratified into high-risk and low-risk clusters based on the risk score. Kaplan-Meier survival analysis was employed to compare the OS between these groups, thereby validating the performance of the risk score model. ROC curve analysis, facilitated by the "timeROC" package (29), was conducted to assess the model's accuracy in predicting patient survival rates. Finally, the model's accuracy was further validated utilizing the independent testing set from TCGA, as well as the entire TCGA dataset.

2.7 Differences in immune characteristics and molecular biology between the high-risk and low-risk groups

Using the "GSEABase" and "GSVA" packages, we analyzed the infiltration fractions and immune-related functions of tumor-infiltrating immune cells in UCEC cases. Differences in immune cell infiltration between low-risk and high-risk groups were compared employing the Wilcoxon test. Moreover, the correlation between the risk score and the expression levels of immune checkpoint genes was investigated using Pearson correlation coefficients. Furthermore, comparisons of risk scores across different stages, grades, and subgroups were conducted to assess the prognostic value of the risk score.

2.8 Drug sensitivity analyses

To investigate the association between chemotherapeutic responsiveness and the risk score model, we employed the "oncoPredict" package (30), leveraging data from the Genomics of Drug Sensitivity in Cancer (GDSC) database (www.cancerRxgene.org). This enabled an analysis of drug sensitivity. Subsequently, we conducted comparative analyses of IC50 values across two distinct groups to assess differential therapeutic outcomes, with the aim of identifying potentially efficacious drugs for the treatment of UCEC.

2.9 Reverse transcription quantitative polymerase chain reaction

Total RNA was extracted using TRIzol reagent (Takara, Japan) and reverse-transcribed into cDNA. PCR was performed using the SYBR Green Master Mix kit (Qiagen, Germany), with the expression level of glyceraldehyde 3-phosphate dehydrogenase (GAPDH) serving as the internal reference. The primer sequences were provided in Table 1. The experiment was conducted with at least three technical replicates. We employed the $2^{-\Delta\Delta CT}$ method to calculate the relative mRNA expression levels of hub genes. A CT value difference within 0.5 between replicate wells of the same sample was considered acceptable for analysis.

TABLE 1 The primers of hub DEERGs and GAPDH.

Gene name	Primer orientation	Sequences
MYBL2	Forward	CTTGAGCGAGTCCAAAGACTG
	Reverse	AGTTGGTCAGAAGACTTCCCT
CPEB1	Forward	GTCCTCCCAAAGGTAATATGCC
	Reverse	TGCAGAGCACCGACAAACA
GAPDH	Forward	CAGGAGGCATTGCTGATGAT
	Reverse	GAAGGCTGGGGCTCATTT

2.10 Western blotting

Cells and clinical samples were lysed with RIPA lysis buffer (Solarbio, China), and the protein concentrations were quantified with a BCA protein quantification kit (NCM Biotech, China). The protein samples were then loaded onto a 10% SDS-PAGE gel for electrophoretic separation, followed by transfer to PVDF membranes (Millipore, USA). After blocking with 5% BSA (Solarbio, China) for 1 hour, the membranes were washed three times with Tris-buffered saline containing 0.1% Tween-20 (TBST), with each wash lasting 5 minutes. Next, the PVDF membrane was incubated overnight at 4°C with specific primary antibodies (anti-β-actin, Sigma, USA, 1/10000; MYBL2, Abcam, UK,1/1000; CPEB1, abways, China, 1/1000). The following day, the membrane was incubated for 1 hour at room temperature with HRP-conjugated goat anti-rabbit IgG. Finally, the target protein band was visualized by laser scanning (Thermo Fisher, USA).

2.11 Immunofluorescence assay

Clinical samples were prepared into slides and deparaffinized in xylene, followed by rehydrated in 100% ethanol and sequentially dehydrated in 95%, 85%, and 75% ethanol concentrations. Antigen retrieval was carried out using sodium citrate in a microwave. To block endogenous peroxidases, the samples were treated with 3% hydrogen peroxide (H2O2), followed by incubation in a 3% Bovine Serum Albumin (BSA) solution (Solarbio, China) for blocking purposes. Subsequently, the tissues were incubated with primary antibodies (MYBL2, Abcam, UK, 1/200; CPEB1, abways, China, 1/200)) overnight at 4°C. After the primary antibody incubation, the tissues underwent incubation with secondary antibodies (Goat Anti-Rabbit IgG H&L/AF555 and Goat Anti-Mouse IgG H&L/AF488) for 1 hour at room temperature. DAPI (Solarbio, China) was added, and the samples were briefly incubated before being washed with phosphate-buffered saline. Finally, images were acquired at 400-fold magnification using a confocal microscope (Nikon AIR, Japan).

2.12 Statistical analysis

Data processing, analysis, and visualization were conducted using R (version 4.3.0) and GraphPad Prism (Version 9.4). Differential analysis in R was primarily conducted utilizing the "limma" package. Visualization of data was predominantly achieved through the "ggplot2", "ggpubr", and "enrichplot" packages. Time-dependent ROC curves were calculated and plotted using the "timeROC" package, facilitating comparisons between different models. Statistical comparisons of experimental results between different groups were executed using the Wilcoxon test, with statistical significance set as a p-value of less than 0.05.

3 Results

3.1 Identification of different expression cellular senescence genes

Differential expression analysis was performed on the TCGA-UCEC dataset. The findings revealed 1,132 upregulated genes and 3,839 downregulated genes in the endometrial carcinoma tissues compared to the control group (Figures 2A, B). Intersection analysis of differentially expressed genes with those associated with cellular senescence identified a total of 104 DECSGs (Figure 2C).

3.2 Construction and analysis of cellular senescence gene-related molecular clusters for UCEC

Consensus clustering was conducted based on the expression of DECSGs. As shown in the Figures 3A–C and Supplementary Figure 1, the PAC algorithm determined the optimal number of clusters to be k=2, yielding clusters denoted as C1 (n=229) and C2 (n=315). PCA affirmed the robust intergroup segregation between cluster C2 and cluster C1 (Figure 3D). Subsequent differential analysis of these subtypes identified 1,375 genes exhibiting differential expression. GO enrichment analysis underscored the significant involvement of these DEGs in pathways vital for nuclear division, precise chromosome segregation, and cytoskeleton functions (Figure 3E; Supplementary Table 3). Moreover, KEGG pathway analysis delineated their predominant roles in cell cycle regulation, motor proteins, cellular senescence, and protein digestion and absorption processes (Figure 3F). GSEA further elucidated that cluster C2 is significantly associated with pivotal biological processes encompassing the cell cycle, focal adhesion, pathways pertinent to cancer, spliceosome activity, and ubiquitin-mediated proteolysis (Figure 3G).

Survival analysis between the clusters revealed that patients in cluster C2 exhibit a shorter OS and PFS compared to those in cluster C1 (Figures 3H, I). Analysis of the tumor microenvironment indicated that cluster C2 demonstrates lower immune scores, stromal scores, and ESTIMATE scores, alongside higher tumor purity (Figures 3J–M). Further exploration of the immune landscapes among UCEC patients in the two clusters involved calculating the relative proportions of immune cells using the CIBERSORT algorithm. In comparison to cluster C1, cluster C2 exhibited significantly elevated levels of infiltration by follicular helper T cells, M1 macrophages, M2 macrophages, and activated dendritic cells, while levels of CD8 T cells and regulatory T cells (Tregs) were diminished (Figure 4A).

The majority of immune checkpoint genes (CD274, SIGLEC15, HAVCR2, TIGIT, LAG3, and PDCD1LG2) were highly expressed in cluster C2, while CTLA4 and PDCD1 showed no significant statistical difference between the two risk groups (Figure 4B). Furthermore, the expression levels of most HLA-related genes were significantly elevated in cluster C2, with the exception of HLA-L, which demonstrated decreased expression (Figure 4C).

3.3 Screening of hub prognostic DEGs

In the WGCNA, a β value of 7 (R^2 = 0.75) was chosen to construct a scale-free network (Figures 5A–C), resulting in the identification of 15 modules (Figure 5D). Among these, the darkgreen, royal blue, and salmon modules exhibited the highest correlation with endometrial carcinoma and were selected as hub modules (Figure 5E). By intersecting the WGCNA results with DECSGs, 40 critical genes were identified. Through univariate Cox regression model analysis, 20 DECSGs that displayed prognostic significance were singled out (Figure 6A). Further refinement was conducted using machine learning algorithms to identify hub prognostic DECSGs from these 20 genes, ensuring a more focused selection of genes with significant prognostic value. The XGBoost algorithm ultimately identified 7 central genes with a Gain > 0.01 (Figure 6B). In the GMM regression analysis, after 2^{20} iterations for 20 genes, the model with the highest accuracy (AUC=0.99) was determined, comprising 8 key genes (Figure 6C). In the SVM-RFE process, the classifier error was minimized when the number of signatures was reduced to 6; thus, these 6 genes were identified as central signatures (Figures 6E, F). The Random Forest algorithm, by integrating multiple decision trees, ultimately identified 12 genes with importance scores >1.0 as central features

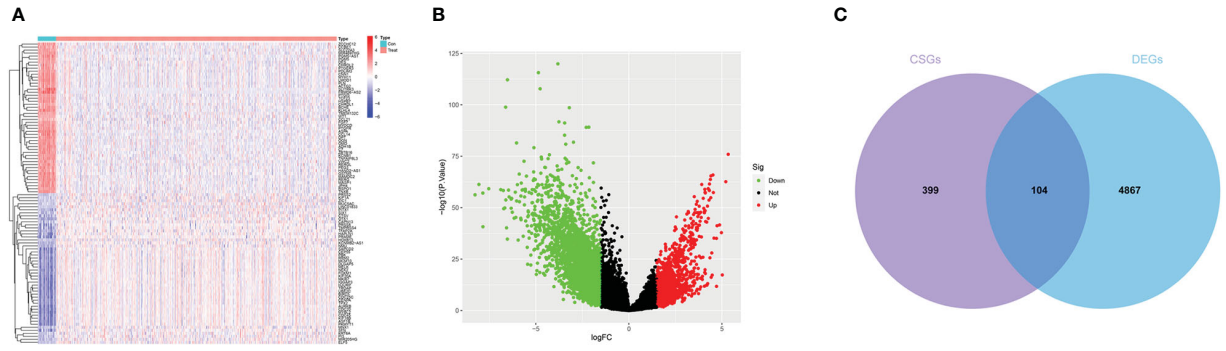


FIGURE 2
(A, B) The heatmap and volcano plot of differential analysis. **(C)** Intersection map of cell senescence genes and differentially expressed genes.

10.3389/fimmu.2024.1418508

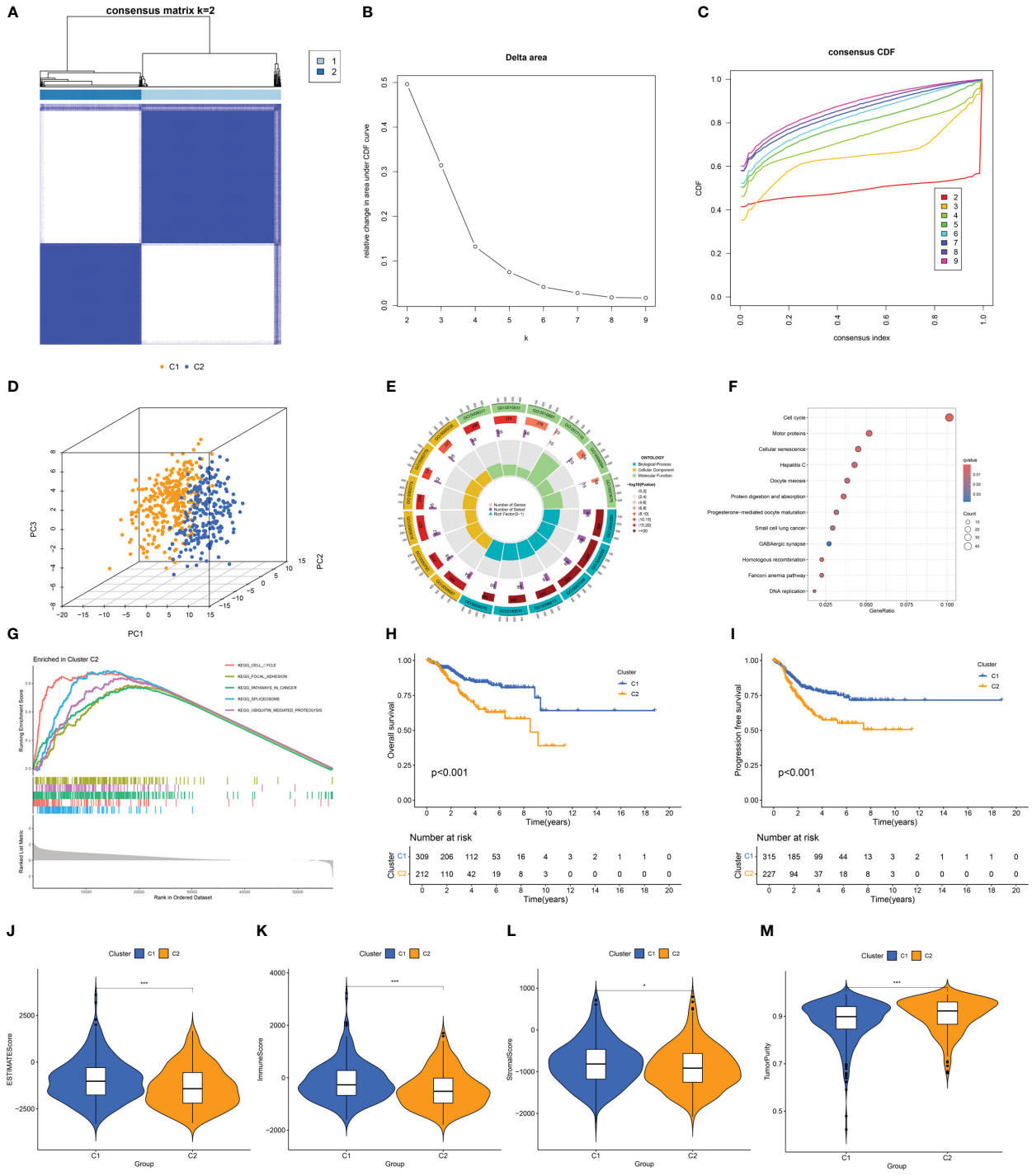


FIGURE 3
(A) Consensus clustering matrix when k = 2. **(B)** Relative alterations in CDF delta area curves. **(C)** Consensus CDF curves when k=2 to 9. **(D)** Three-dimensional Principal Component Analysis delineating the segregation between Cluster C1 and Cluster C2. **(E–G)** GO term enrichment, KEGG pathway analysis, and GSEA results in two clusters. **(H, I)** The difference in OS and PFS between the two clusters. **(J–M)** Differences in ESTIMATEScore, immune scores, stromal scores, and tumor purity between the two clusters (*p<0.05; ***p<0.001).

(Figures 6G, H). The intersection of these selected feature genes identified CPEB1 and MYBL2 as hub prognostic DECSGs (Figure 6D). Survival analyses from the Kaplan-Meier Plotter database revealed a significant decrease in OS of patients with endometrial carcinoma as the expression levels of CPEB1 and MYBL2 increased (Figures 6I, J). Compared to the control group, the expression of MYBL2 was upregulated in endometrial carcinoma, whereas CPEB1 expression was downregulated (Figures 6K, L). ROC curve analysis showed the areas under the curve (AUC) values for CPEB1 and MYBL2 are 0.979 and 0.974, respectively, indicating excellent diagnostic value for UCEC (Supplementary Figure 2).

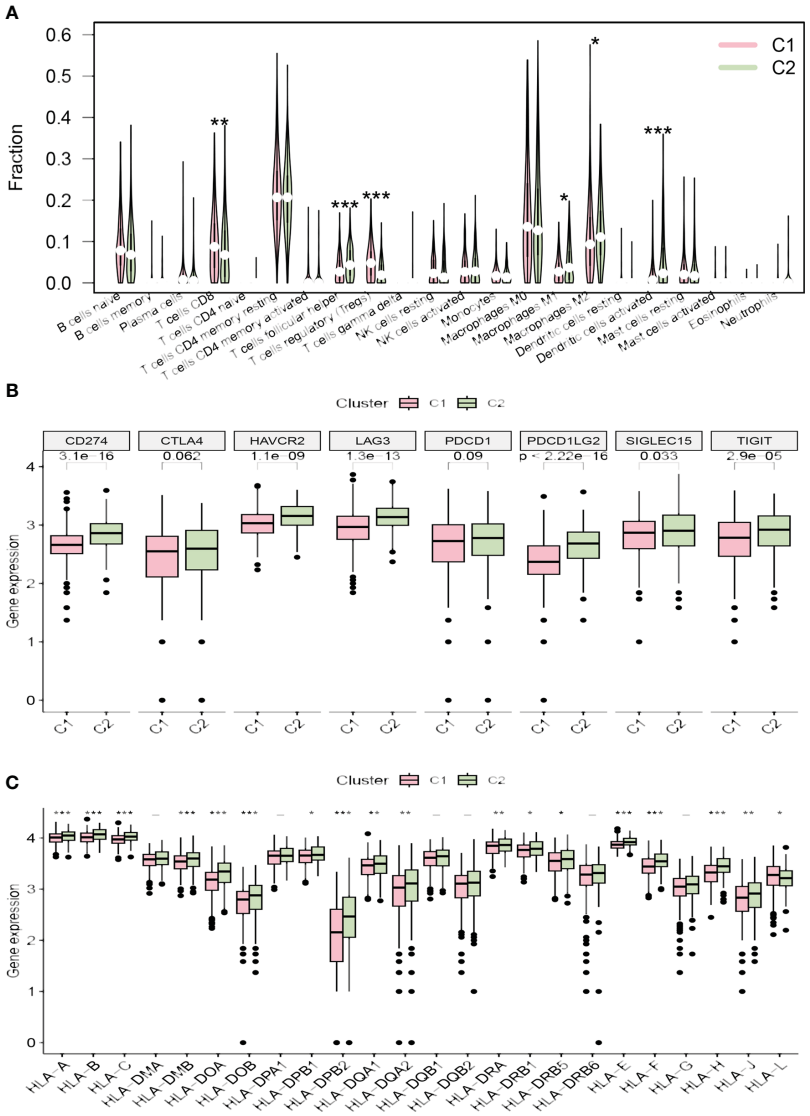


FIGURE 4
(A) The diagram of the difference in immune cell infiltration levels between the two clusters. **(B, C)** The different expression levels of immune checkpoint genes and HLA-related genes in two clusters, respectively (*p < 0.05; **p < 0.01; ***p < 0.001).

3.4 Development and validation of a novel cellular senescence-related prognostic model

After random division, the TCGA training set included 355 patients, while the testing set comprised 156 patients. Utilizing CPEB1 and MYBL2, a risk model incorporating two hub gene risk features was developed through LASSO Cox regression analysis in the TCGA training set (Figures 7A, B). The risk score was calculated as follows: Risk score = (0.1279 × expression of MYBL2) + (0.0879 × expression of CPEB1). UCEC patients were then categorized into high-risk and low-risk groups based on the median risk score. Figures 7C, F, I showed the distribution of risk scores and survival times across the training cohort, testing cohort, and the entire TCGA cohort. Survival analysis results demonstrated a positive correlation between higher risk scores and increased mortality in the training cohort, test cohort, and the entire TCGA cohort. According to Kaplan-Meier analysis, the overall survival of the high-risk group was significantly shorter than that of the low-risk group, indicating a worse prognosis for the high-risk group (Figures 7D, G, J). ROC curves demonstrated that the AUC for the 3-year time-dependent ROC for the three cohorts were 0.624, 0.768, and 0.661, respectively, indicating that the prognostic model exhibits good predictive performance (Figures 7E, H, K).

3.5 Evaluation of TME and drug sensitivity between the two risk score groups

The results obtained from the ssGSEA algorithm revealed distinctive immune infiltration patterns between the high-risk and low-risk groups. Specifically, compared to the low-risk group, the

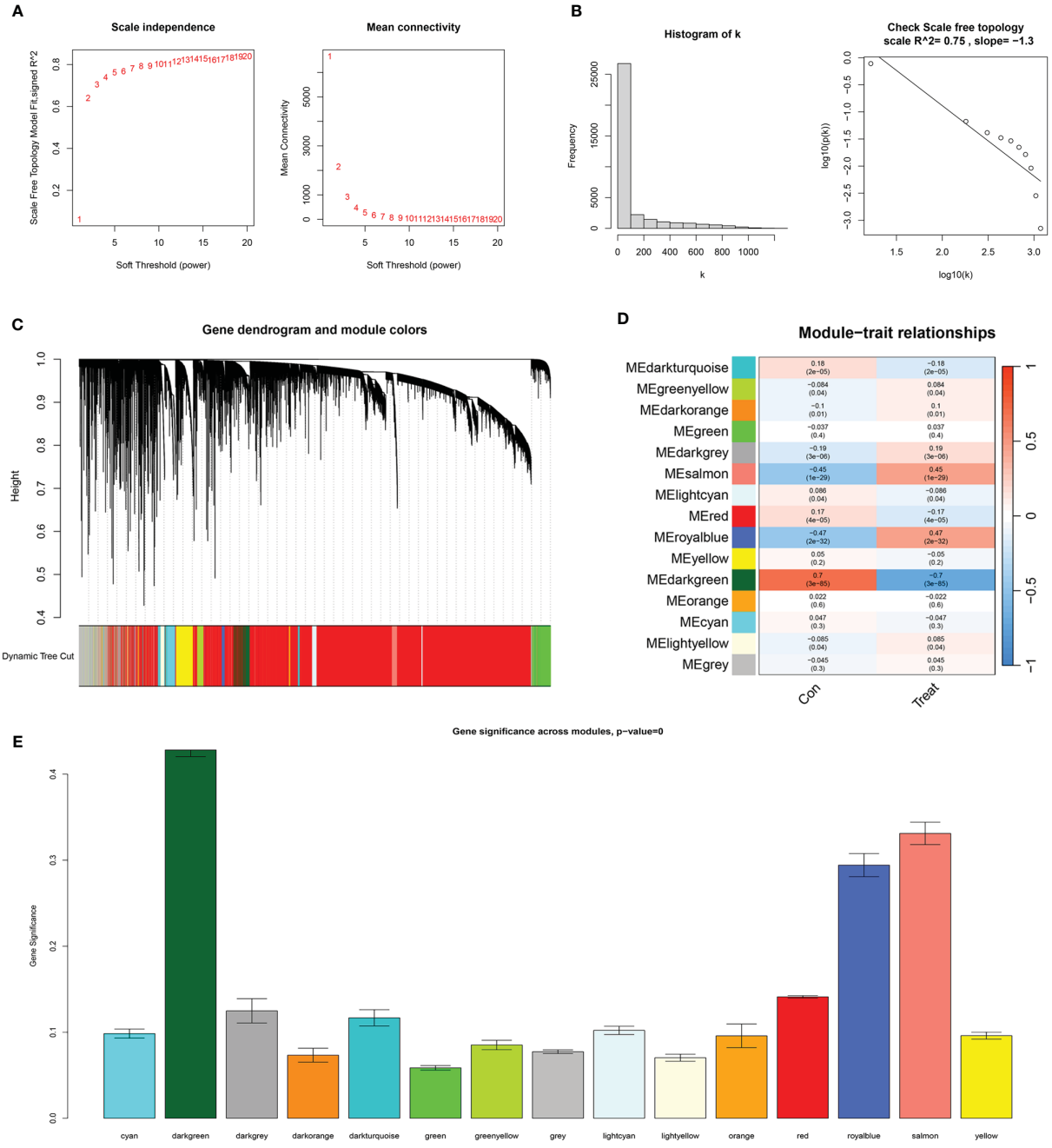


FIGURE 5
WGCNA results. **(A)** The scale-free fit index for various soft-thresholding powers (β) and the mean connectivity for various soft-thresholding powers. **(B)** Histogram of connectivity distribution and the scale-free topology when β=7. **(C)** Dendrogram of genes clustered via the dissimilarity measure. **(D)** Heatmap of the correlation between module and clinical traits. **(E)** Bar plot of gene significance across WGCNA modules.

high-risk group exhibited a unique immune infiltration pattern characterized by significantly lower abundance of most tumor-infiltrating immune cells, except for natural killer cells (Figure 8A). Regarding immune function activity, apart from macrophages and parainflammation, most immune functions were significantly higher in the low-risk group compared to the high-risk group (Figure 8B).

Additionally, our differential analysis of IC50 values between the groups revealed notable differences. Specifically, the IC50 values for Trametinib, PD0325901, Dactolisib, Docetaxel, and Camptothecin were substantially higher in the high-risk group compared to the low-risk group (Figures 8C–G). This suggests that patients with lower risk scores may derive enhanced benefits from these drugs. Conversely, IC50 values for Vincristine, BI-2536, BMS-754807, Bortezomib, and Daporinad were found to be lower in the high-risk group (Figures 8H–L), indicating that these drugs might be particularly effective for patients classified as high risk. These insights highlight the importance of risk stratification in tailoring chemotherapeutic strategies to individual patient profiles, potentially optimizing treatment outcomes.

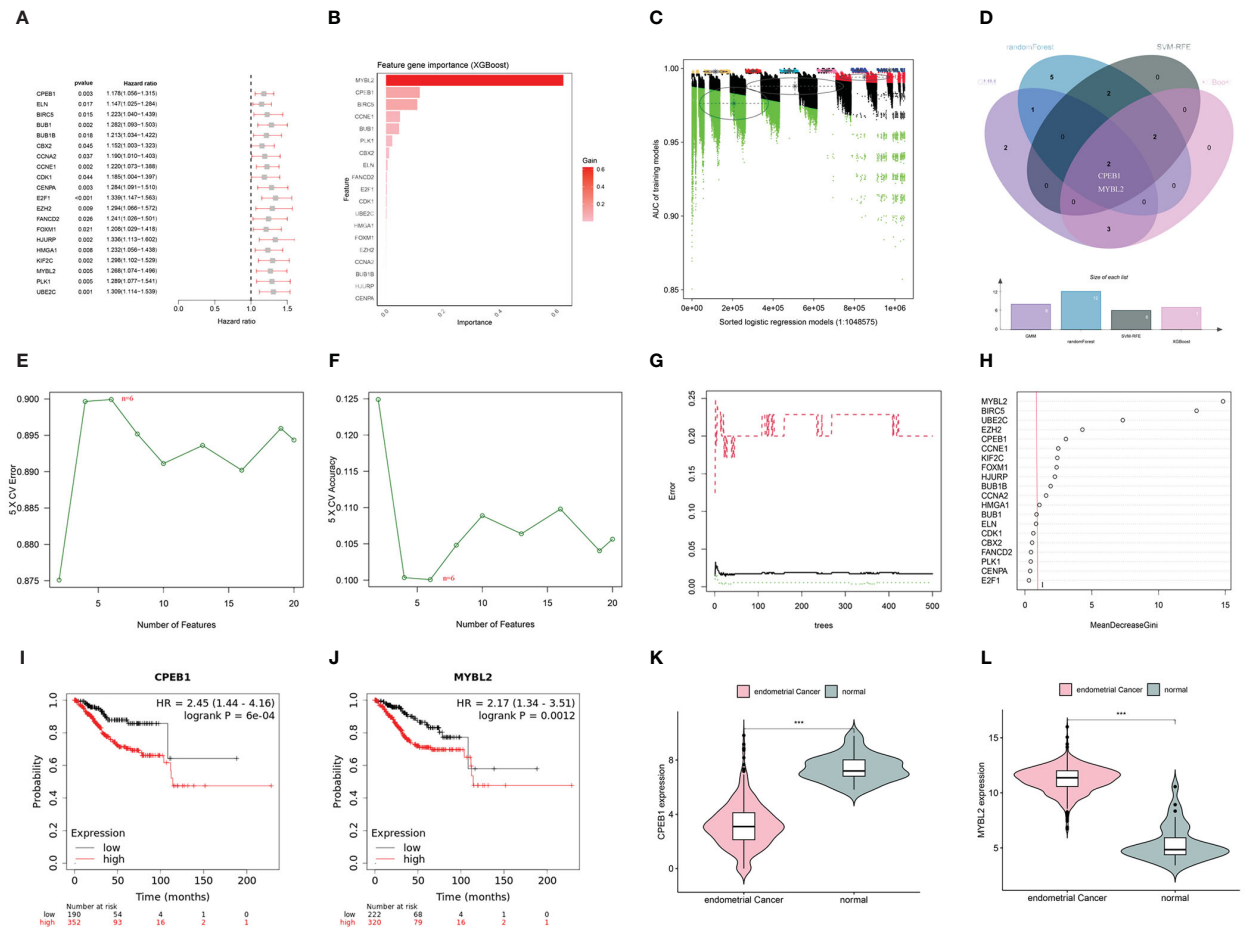


FIGURE 6
(A) Univariate COX analysis shows 20 genes associated with overall survival. **(B)** Screening of diagnostic biomarkers based on XGBoost algorithm (n=7). **(C)** Variable selection in GMM model (n=8); **(D)** Venn diagram of four machine learning results. **(E, F)** Through SVF- RFE algorithm selects the best biomarkers (n=6). **(G, H)** Important features selected by random forest algorithm (n=12). **(I, J)** K-M curves of CPEB1 and MYBL2 in UCEC. **(K, L)** Violin plot show the expression levels of CPEB1 and MYBL2 in TCGA-UCEC cohort(***p < 0.001).

3.6 Correlation of risk scores with clinical information, cellular senescence-related subtypes and immune checkpoints

We conducted a comparison of risk score levels across clinical stages and grades in patients. In the TCGA-UCEC dataset, we observed that higher grades were associated with higher risk scores (Figure 9A). Regarding clinical stages, risk scores for patients in stages II, III, and IV were significantly higher than those in stage I. However, there were no statistical differences in risk scores between stages II, III, and IV (Figure 9B). Subsequently, we explored the correlation between the expression levels of immune checkpoint genes and prognostic risk scores. Notably, there was a significant difference in risk scores between the two subtypes established through cellular senescence genes (Figure 9C). As illustrated in Figure 9D, the expression of most immune checkpoint genes, except for CTLA4, was positively correlated with risk scores. An alluvial diagram illustrated the variations in cellular senescence-related clusters, risk scores, and life states (Figure 9E).

3.7 Verification of the expression of CPEB1 and MYBL2

We conducted further analysis to assess the relative mRNA and protein expression levels of the hub genes CPEB1 and MYBL2 in clinical samples. PCR results indicated that at the transcriptomic level, the relative mRNA expression of MYBL2 was significantly higher in UCEC compared to normal tissue (Figure 10A), while the relative expression of CPEB1 was significantly down-regulated in UCEC (Figure 10B). Results from WB analyses (Figures 10C, D) and immunofluorescence staining (Figures 10E, F) corroborated these findings, demonstrating that the protein expression levels of the two hub genes were consistent with the RT-qPCR results (Figures 9A–D).

4 Discussion

Uterine corpus endometrial carcinoma has been demonstrated to exhibit high levels of heterogeneity (31). The tumor

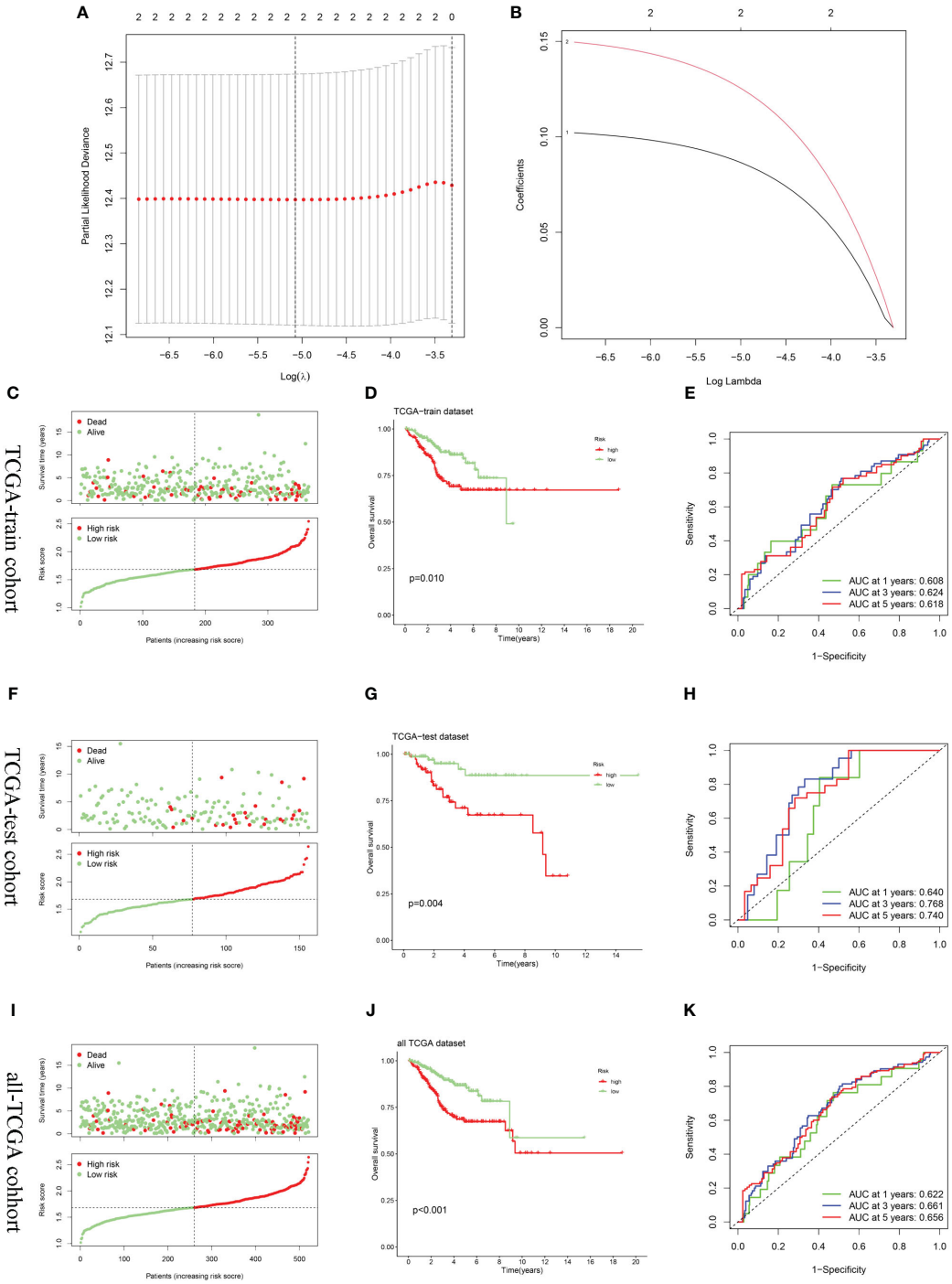


FIGURE 7
Construction and validation of the risk score model. **(A, B)** Constructed a prognostic model in the TCGA-train cohort through LASSO COX regression analysis. **(C, F, I)** Risk scores distribution and survival status of each patient in the TCGA-train cohort, TCGA-train cohort, and all-TCGA cohort, respectively. **(D, G, J)** Kaplan–Meier curves for the OS of the two subtypes in the TCGA-train cohort, TCGA-train cohort, and all-TCGA cohort, respectively. **(E, H, K)** ROC curves illustrated the predictive efficacy of the risk score for 1-, 3-, and 5-year survival in the TCGA-train cohort, TCGA-train cohort, and all-TCGA cohort, respectively.

microenvironment, comprising malignant, immune, endothelial, and stromal components (32), plays a pivotal role in the progression of the cancer and its sensitivity to therapeutic agents (33). The molecular attributes of endometrial cancer cells, along with the composition and dynamics of the tumor microenvironment, significantly influence these processes.

The widespread utilization of genomic sequencing has generated a plethora of biological data, offering enhanced diagnostic and prognostic capabilities across various malignancies. In recent years, researchers have developed diverse prognostic models utilizing gene expression profiles sourced from databases, employing a range of bioinformatics analysis

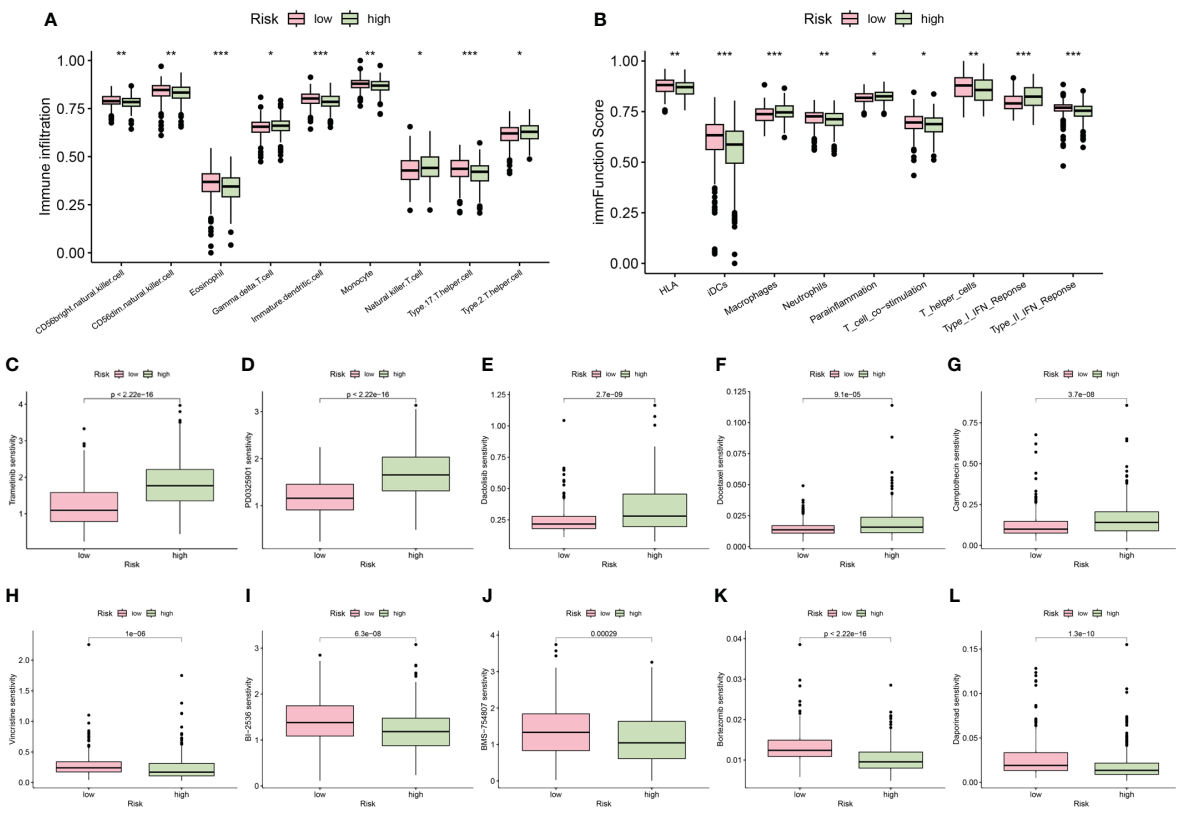


FIGURE 8
The differences of immune infiltrating cells **(A)** and immune function **(B)** between high- and low- risk groups. **(C–L)** Chemotherapy and immunotherapy sensitivity prediction between the low-risk and the high-risk groups (*p < 0.05; **p < 0.01; ***p < 0.001).

methodologies. These models have provided valuable insights into guiding personalized treatment strategies for UCEC (34, 35).

Cellular senescence plays a crucial role in maintaining tissue stability, internal equilibrium, and serves as a natural mechanism to prevent cancer. However, under certain conditions, it can also promote tumor development (36). It has been closely associated with the onset and progression of various diseases and serves as an effective means of stratifying cancer patients (37).

Previous studies have investigated the association between cellular senescence and endometrial cancer. Gao et al. (38) conducted a bioinformatics study focusing on the role of cell senescence-related genes in UCEC and made significant progress. However, their study has certain limitations. Primarily, although they utilized various datasets from TCGA-UCEC and GEO to expand the sample size for analysis, it's worth noting that GSE119041 dataset includes cases of undifferentiated uterine sarcoma. UCEC encompasses pure endometrioid cancer as well as carcinomas with high-risk endometrial histology, including sarcoma. Sarcomas represent uncommon subtypes with a generally poorer prognosis, and the TCGA-UCEC dataset comprises only a limited number of sarcoma cases. Incorporating data from GSE119041 into the analysis may lead to unreliable conclusions.

In our study, all samples were sourced from TCGA-UCEC, avoiding heterogeneity between diseases and samples, as well as batch effects stemming from different datasets. Unlike previous approaches that solely relied on LASSO regression to select feature genes, we employed a stepwise selection process for UCEC feature genes using methods such as WGCNA, Cox regression, and machine learning. Our findings hold promise as diagnostic and prognostic markers for UCEC. WGCNA facilitated the identification of co-expression gene modules in cancer samples, offering a refined and systematic perspective on understanding the molecular mechanisms of cancer by establishing network relationships between genes. Furthermore, the utilization of machine learning, especially in managing and analyzing large biomedical datasets, significantly enhanced the accuracy of analysis and the performance of predictive models. Leveraging these advanced algorithms allowed for the more precise identification of genes closely associated with UCEC. Lastly, we conducted multidimensional experimental validations including PCR, WB, and IF, thereby further confirming the abnormal expression of hub genes. Our study results yielded divergent findings from Gao et al., expanding the realm of research on cell senescence genes and their implications in endometrial cancer.

In this study, we conducted an in-depth exploration of the relationship between UCEC and cellular senescence genes. Utilizing 104 differentially expressed cellular senescence genes, we performed a consensus clustering analysis, ultimately categorizing UCEC into two clusters. We observed significant differences between clusters C1 and C2 in terms of biological functions, prognostic outcomes, tumor microenvironment, immune cell infiltration, immune

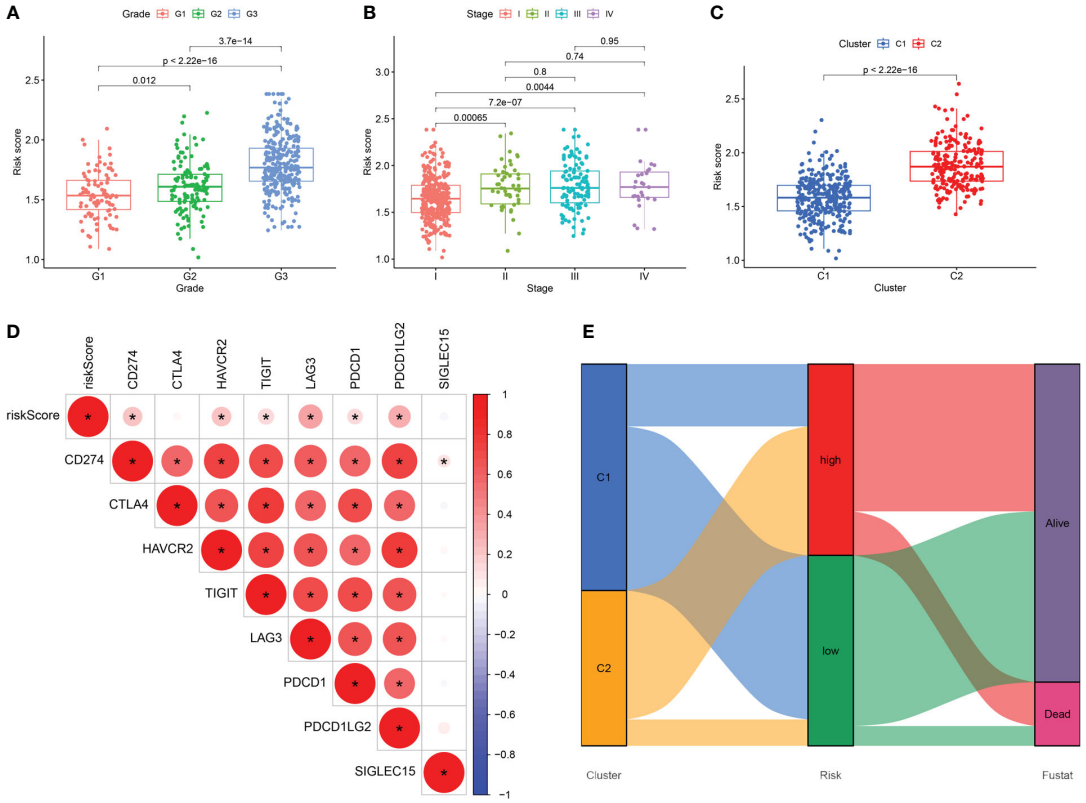


FIGURE 9
(A–C) The difference in risk scores between pathologic grades, clinical stages, and the two subtypes. **(D)** Correlation between the expression levels of immune checkpoint genes and risk score. **(E)** Alluvial diagram of subtype distributions and prognosis of UCEC patients. *, means p-values less than 0.05.

checkpoints, and HLA gene expression. This underscores the presence of substantial tumor heterogeneity within UCEC. The KEGG results indicated that the differentially expressed genes in clusters C1 and C2 were primarily implicated in the cellular senescence pathway, highlighting the pivotal role of cellular senescence genes in UCEC. Furthermore, both KEGG and GSEA analyses indicated the activation of the cell cycle pathway.

In cluster C2, we speculated that aberrant expression of cellular senescence genes may enable damaged or potentially malignant cells to evade senescence defenses and enter a state of uncontrolled proliferation. This not only disrupted crucial cell cycle checkpoints but may also impact the expression and activity of cyclin-dependent kinases (CDKs) and cyclins, as well as their inhibitors, thereby enhancing tumor cells' ability to override growth inhibitory signals. This propensity for unbridled proliferation facilitated the rapid expansion of cluster C2 tumor cells, exacerbating genomic instability and promoting the survival and division of DNA-damaged cells. Consequently, this promoted the malignant transformation of the C2 cluster, ultimately resulting in poor prognosis.

In the tumor microenvironment of cluster C2, we noted a higher tumor purity alongside a lower immune score. Furthermore, most of the HLA class I and class II molecules in cluster C2 were found to be upregulated. HLA class I molecules typically present endogenous antigens to CD8+ T cells, while HLA class II molecules present exogenous antigens to CD4+ T cells (39). Generally, increased expression of HLA molecules should facilitate more effective T-cell-mediated immune responses, thereby enhancing the recognition and elimination of tumor cells, ultimately improving patients' prognosis (40). However, the results from CIBERSORT analysis revealed a decrease in the infiltration levels of CD8 T cells and regulatory T cells in cluster C2, with no significant difference observed in CD4 T cells. Conversely, the proportion of follicular helper T cells, M1 macrophages, and activated dendritic cells was found to increase.

Follicular helper T cells, primarily found in secondary lymphoid tissues, play a pivotal role in facilitating B cells interactions, thereby promoting antibody production and the formation of memory B cells (41). M1 macrophages represent an activated state of macrophages that bolster immune responses by eliminating tumor cells and pathogens (42). Activated dendritic cells capture and present antigens, thereby initiating immune responses in T cells and B cells (43). In cluster C2, combined with the upregulation of most immune checkpoint genes, these immune checkpoint molecules, typically expressed on the surface of immune cells, possessed the capacity to inhibit the activation and proliferation of T cells, fostering a tumor-promoting environment conducive to immune evasion (44). We speculated that despite adequate antigen presentation in cluster C2, the predominant influence of immune checkpoint molecules in UCEC progression renders related T cell

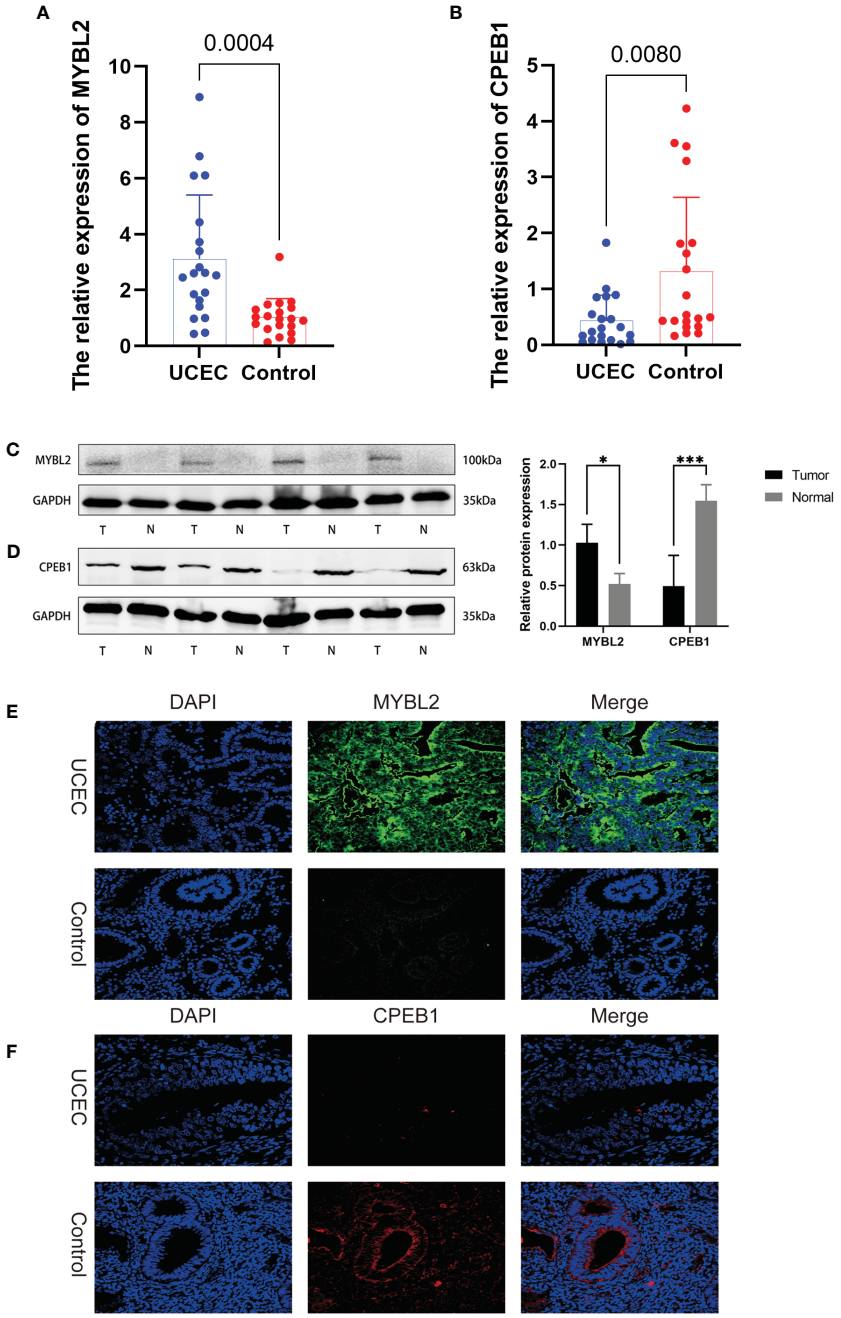


FIGURE 10
The expression levels of 2 hub genes in UCEC tissues and normal tissues were validated by RT-qPCR, WB, and immunofluorescence. **(A, B)** RT-qPCR. **(C, D)** WB assay. **(E, F)** immunofluorescence. *, and ***, means p-values less than 0.05, and 0.001, respectively.

activation ineffective. Moreover, under the influence of abnormally high expression of immune checkpoint molecules, although follicular helper T cells and M1 macrophages showed an increased proportion, their functionality may be compromised by the immunosuppressive environment, thus limiting their anti-tumor activity. Consequently, the anti-tumor immune response in cluster C2 appeared weakened, thereby facilitating tumor growth and dissemination. This underscored the potential utility of immune checkpoint inhibitors in patients within Cluster C2, as these therapeutic agents may help restore the anti-tumor immune response and impede tumor progression.

In summary, the observed upregulation of HLA genes in cluster C2, combined with the decrease in CD8+ T cells and Treg levels, alongside the heightened expression of immune checkpoint genes, revealed a complex immune regulatory network. While theoretically, this network should enhance anti-tumor immune responses, it may inadvertently lead to immune suppression due to tumor cells' strategies for immune evasion. This phenomenon

underscored the importance of emphasizing the value of immune checkpoint inhibitors in exploring immune-based therapeutic strategies for UCEC, aiming to circumvent these inhibitory mechanisms within the tumor microenvironment.

Through the application of WGCNA and Cox regression analysis, in conjunction with a series of advanced machine learning algorithms, we successfully identified CPEB1 and MYBL2 and developed a prognostic risk model. Internal validation results indicated that patients with high-risk scores exhibited significantly worse OS across the training cohort, testing cohort, and the entire TCGA cohort. Furthermore, we observed significant variations in risk scores across two clusters, clinical stages, and grades. These findings suggested that the prognostic risk model holds substantial clinical value in identifying high-risk patients.

MYBL2, a member of the MYB transcription factor family, plays a crucial role in regulating the cell cycle, particularly during DNA replication and mitosis. As a central regulator in tumorigenesis, MYBL2 is involved in the proliferation, apoptosis, and differentiation of cancer cells. Elevated expression of MYBL2 in various tumors is often associated with poor prognosis (45, 46), rendering it a potential therapeutic target in cancer treatment. As a prognostic indicator of unfavorable outcomes in osteosarcoma and a universal marker for immune infiltration across various cancers, MYBL2 exerts regulatory control over proliferation, tumor advancement, and immune cell infiltration within osteosarcoma and broader cancer contexts (47). In clear cell renal carcinoma, MYBL2 promotes malignant characteristics and impedes apoptosis through activation of the hedgehog signaling pathway (48). Within gastric cancer, MYBL2 modulates DNA damage via UBEC2 activation, thereby promoting tumor progression and resistance to cisplatin therapy (49). In ovarian cancer, the MYBL2-CCL2 axis promotes tumor progression and confers resistance to PD-1 therapy by inducing immunosuppressive macrophages (50). In colorectal cancer, MYBL2 expedites cancer progression through an interactive feed-forward activation with E2F2 (51). In our investigation, we observed upregulated expression of MYBL2 in UCEC tissues, thus suggesting its potential utility as a prognostic marker for this malignancy.

CPEB1, also known as Cytoplasmic Polyadenylation Element Binding Protein 1, exerts influence over the stability and translation of its target mRNA molecules, significantly impacting fundamental cellular processes such as growth, differentiation, and apoptosis (52). The expression and function of CPEB1 have garnered considerable attention due to its diverse expression patterns and roles across various types of cancer (53). Research into colorectal cancer metastasis has revealed a novel tumor-suppressive role for CPEB1. High methylation of the CPEB1 promoter, restricting chromatin accessibility and transcription factor binding, diminishes its expression, thereby influencing colorectal cancer progression (54). Additionally, studies have demonstrated that CPEB1 can directly target SIRT1, suppressing its translation and mediating cancer stemness *in vitro* and *in vivo*, suggesting its potential as a therapeutic target in hepatocellular carcinoma (HCC) (55). Overall, recent research has increasingly recognized the multifaceted role of CPEB1 in cellular processes and its impact on various cancers. Currently, there is a lack of research on CPEB1 in the context of endometrial cancer in the existing literature. Our analysis revealed downregulation of CPEB1 expression in endometrial cancer, a finding supported by PCR, WB, and IF assays. While we are the first to report its association with endometrial cancer, further experimental investigations are warranted to fully elucidate the underlying mechanisms.

Carboplatin, in combination with paclitaxel, has emerged as the frontline chemotherapy regimen for endometrial cancer (56). Nonetheless, substantial variability exists among patients in their responses to chemotherapy. Through drug sensitivity analysis, we have identified several drugs that hold promise for UCEC treatment. Significant differences in IC50 values of these drugs observed between distinct risk groups indicate the substantial predictive capacity of our model in predicting drug responses among patients with endometrial cancer.

Immunotherapy, particularly checkpoint inhibitors, has demonstrated high efficacy and generally favorable safety and tolerability profiles. In several clinical trials, checkpoint inhibitors have shown substantial therapeutic effects in patients with recurrent endometrial cancer, especially in those unresponsive to chemotherapy (57). Moreover, studies indicate that the use of checkpoint inhibitors can significantly enhance long-term survival rates in endometrial cancer patients characterized by specific molecular markers (58). Currently, immunotherapy drugs are increasingly being incorporated into the clinical management of endometrial cancer. PD-1 inhibitors, such as pembrolizumab and dostarlimab, have shown efficacy in treating unresectable or metastatic solid tumors with MSI-H or dMMR status. Concurrently, PD-L1 inhibitors, including atezolizumab and avelumab, are under evaluation in clinical trials for their potential in endometrial cancer therapy. Combination therapy, such as pembrolizumab combined with multikinase inhibitors like lenvatinib, is being utilized for endometrial cancer patients experiencing disease progression after prior systemic therapy. Moreover, CTLA-4 inhibitors like ipilimumab are being investigated in combination with PD-1 inhibitors to assess their efficacy in endometrial cancer treatment (59). The advent of immune checkpoint inhibitors (ICIs) has significantly transformed the therapeutic landscape for endometrial cancer, highlighting the substantial immune heterogeneity within UCEC (60). Additionally, a recent review revealed that the addition ICIs to chemotherapy can improve PFS in the overall population compared to chemotherapy alone (61). New treatment guidelines are also being formulated to explore the use of immune checkpoint inhibitors across the four molecular categories of endometrial cancer and their potential prognostic effects (62). However, not all endometrial cancer patients respond favorably to checkpoint inhibitors, particularly those with microsatellite stable (MSS) tumors or low tumor mutational burden (63). Additionally, the high costs and potential toxicities associated with these therapies limit their accessibility to all UCEC patients. Our analysis unveiled that cluster C2 exhibits elevated levels of immune checkpoint genes and a positive correlation between risk scores and immune

checkpoint expression, suggesting that patients in the high-risk group may derive greater benefits from treatment with immune checkpoint inhibitors.

Our advanced bioinformatics analyses, based on a prognostic model centered on cellular senescence genes, provide novel perspectives on UCEC and present opportunities for personalized immune therapies to advance treatment strategies. Nevertheless, our study is not without limitations. Firstly, its retrospective nature and reliance on bioinformatics methodologies underscore the need for further investigations with larger patient cohorts to enhance the generalizability of the results. Additionally, while we validated the dysregulated expression of hub genes at the transcriptomic and proteomic levels, understanding their biological functions and interactions within the tumor microenvironment, particularly with regard to immune checkpoints, necessitates additional experimental exploration.

In summary, our diverse bioinformatics analyses based on senescence-associated genes have unveiled two distinct molecular subtypes of UCEC exhibiting significantly different tumor microenvironments and prognoses. Moreover, the prognostic risk model we established has demonstrated remarkable efficacy in predicting the prognosis and responsiveness to chemotherapy among UCEC patients, indicating its potential clinical applicability.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.gdc.cancer.gov/projects/TCGA-UCEC.

Ethics statement

The studies involving humans were approved by the Medical Ethics Committee of the First Affiliated Hospital of Guangxi Medical University. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the collection of clinical specimens during the ethical approval process.

Author contributions

CW: Conceptualization, Data curation, Methodology, Software, Visualization, Writing – original draft. SL: Validation, Conceptualization, Data curation, Formal analysis, Visualization, Writing – original draft. YH: Conceptualization, Data curation, Validation, Visualization, Writing – original draft. YW: Data curation, Validation, Writing – original draft, Investigation, Methodology, Software. JM: Software, Supervision, Visualization, Conceptualization, Writing – review & editing, Formal analysis, Methodology. JF: Funding acquisition, Supervision, Validation, Resources, Project administration, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by grants from the National Natural Science Foundation of China (Nos. 81960464) and the construction of clinical intervention protocols Guangxi Key R & D program (Guike AB22080045).

Acknowledgments

Thanks to the R software development team and the TCGA databases for providing many biological information data, and express our gratitude to the support from the National Natural Science Foundation of China.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2024.1418508/full#supplementary-material

References

1. Zheng RS, Zhang SW, Sun KX, Chen R, Wang SM, Li L, et al. [Cancer statistics in China, 2016]. *Zhonghua Zhong Liu Za Zhi*. (2023) 45:212–20. doi: 10.3760/cma.j.cn112152–20220922–00647

2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. (2021) 71:209–49. doi: 10.3322/caac.21660

3. Mills AM, Liou S, Ford JM, Berek JS, Pai RK, Longacre TA. Lynch syndrome screening should be considered for all patients with newly diagnosed endometrial cancer. *Am J Surg Pathol.* (2014) 38(11):1501–9. doi: 10.1097/pas.0000000000000321

4. Betella I, Fumagalli C, Raviele PR, Schivardi G, Vitis LAD, Achilarre MT, et al. *A Novel Algorithm to Implement the Molecular Classification According to the New Esgo/Estro/Esp 2020 Guidelines for Endometrial Cancer*, Vol. 32. (2022). pp. 993–1000. Hudson Place, New Jersey, USA: International Journal of Gynecologic Cancer. doi: 10.1136/ijgc-2022–003480.

5. Morice P, Leary A, Creutzberg C, Abu-Rustum N, Darai E. Endometrial cancer. *Lancet*. (2016) 387:1094–108. doi: 10.1016/s0140–6736(15)00130–0

6. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. *CA Cancer J Clin*. (2021) 71:7–33. doi: 10.3322/caac.21654

7. Xue Y, Dong Y, Lou Y, Lv Q, Shan W, Wang C, et al. Pten mutation predicts unfavorable fertility preserving treatment outcome in the young patients with endometrioid endometrial cancer and atypical hyperplasia. *J Gynecol Oncol*. (2023) 34:e53. doi: 10.3802/jgo.2023.34.e53

8. Vermij L, Léon-Castillo A, Singh N, Powell ME, Edmondson RJ, Genestie C, et al. P53 immunohistochemistry in endometrial cancer: clinical and molecular correlates in the portec-3 trial. *Mod Pathol*. (2022) 35:1475–83. doi: 10.1038/s41379-022-01102-x

9. Janku F, Wheler JJ, Westin SN, Moulder SL, Naing A, Tsimberidou AM, et al. Pi3k/akt/mtor inhibitors in patients with breast and gynecologic Malignancies harboring pik3ca mutations. *J Clin Oncol*. (2012) 30:777–82. doi: 10.1200/jco.2011.36.1196

10. Peters I, Marchetti C, Scambia G, Fagotti A. New windows of surgical opportunity for gynecological cancers in the era of targeted therapies. *Int J Gynecol Cancer*. (2024) 34:352–62. doi: 10.1136/ijgc-2023–004580

11. Kumari R, Jat P. Mechanisms of cellular senescence: cell cycle arrest and senescence associated secretory phenotype. *Front Cell Dev Biol*. (2021) 9:645593. doi: 10.3389/fcell.2021.645593

12. Chambers CR, Ritchie S, Pereira BA, Timpson P. Overcoming the senescence-associated secretory phenotype (Sasp): A complex mechanism of resistance in the treatment of cancer. *Mol Oncol*. (2021) 15:3242–55. doi: 10.1002/1878–0261.13042

13. Schmitt CA, Wang B, Demaria M. Senescence and cancer - role and therapeutic opportunities. *Nat Rev Clin Oncol*. (2022) 19:619–36. doi: 10.1038/s41571–022-00668–4

14. Colucci M, Zumerle S, Bressan S, Gianfanti F, Troiani M, Valdata A, et al. Retinoic acid receptor activation reprograms senescence response and enhances anti-tumor activity of natural killer cells. *Cancer Cell*. (2024) 42:646–61.e9. doi: 10.1016/j.ccell.2024.02.004

15. Guo Y, Wang S, Dong Y, Liu Y. Attenuation of pro-tumorigenic senescent secretory phenotype by stn, a novel derivative of stevioside, potentiates its inhibitory activity on hepatocellular carcinoma. *Food Chem Toxicol*. (2024) 184:114371. doi: 10.1016/j.fct.2023.114371

16. Talhouk A, McConechy MK, Leung S, Li-Chang HH, Kwon JS, Melnyk N, et al. A clinically applicable molecular-based classification for endometrial cancers. *Br J Cancer*. (2015) 113:299–310. doi: 10.1038/bjc.2015.190

17. Liu W, Ma J, Zhang J, Cao J, Hu X, Huang Y, et al. Identification and validation of serum metabolite biomarkers for endometrial cancer diagnosis. *EMBO Mol Med*. (2024) 16(4):988–1003. doi: 10.1038/s44321–024-00033–1

18. Hong S, Fu N, Sang S, Ma X, Sun F, Zhang X. Identification and validation of irf6 related to ovarian cancer and biological function and prognostic value. *J Ovarian Res*. (2024) 17:64. doi: 10.1186/s13048–024-01386–4

19. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res*. (2015) 43:e47. doi: 10.1093/nar/gkv007

20. Wilkerson MD, Hayes DN. Consensusclusterplus: A class discovery tool with confidence assessments and item tracking. *Bioinformatics*. (2010) 26:1572–3. doi: 10.1093/bioinformatics/btq170

21. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. Clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)*. (2021) 2:100141. doi: 10.1016/j.xinn.2021.100141

22. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*, Vol. 65. (2013). pp. 843–4. Tiergartenstrasse 17, Heidelberg, Germany: Springer Science & Business Media. doi: 10.2307/25053321.

23. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. (2015) 12:453–7. doi: 10.1038/nmeth.3337

24. Langfelder P, Horvath S. Wgcna: an R package for weighted correlation network analysis. *BMC Bioinf*. (2008) 9:559. doi: 10.1186/1471–2105-9–559

25. Chalmers R, Adkins M. *Writing Effective and Reliable Monte Carlo Simulations with the Simdesign Package*. (2020). Dr. Penfield Avenue, Montreal, Canada: Tutorials in Quantitative Methods for Psychology. doi: 10.20982/tqmp.16.4.p248.

26. Das P, Roychowdhury A, Das S, Roychoudhury S, Tripathy S. Sigfeature: novel significant feature selection method for classification of gene expression data using support vector machine and T statistic. *Front Genet*. (2020) 11:247. doi: 10.3389/fgene.2020.00247

27. Breiman L, Breiman L. Cutler RAJJoCM. *Random Forests Mach Learning*. (2001) 2:199–228.

28. Chen T, Guestrin C. *Xgboost: A Scalable Tree Boosting System*. San Francisco, California, USA: ACM (2016). doi: 10.1145/2939672.2939785

29. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med*. (2013) 32:5381–97. doi: 10.1002/sim.5958

30. Maeser D, Gruener RF, Huang RS. Oncopredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data. *Brief Bioinform*. (2021) 22. doi: 10.1093/bib/bbab260

31. Tang X, Hu Y. The role of tcga molecular classification in clear cell endometrial carcinoma. *Front Oncol*. (2023) 13:1147394. doi: 10.3389/fonc.2023.1147394

32. Pradhan R, Kundu A, Kundu CN. The cytokines in tumor microenvironment: from cancer initiation-elongation-progression to metastatic outgrowth. *Crit Rev Oncol Hematol*. (2024) 196:104311. doi: 10.1016/j.critrevonc.2024.104311

33. Huang Y, Chen Z, Shen G, Fang S, Zheng J, Chi Z, et al. Immune regulation and the tumor microenvironment in anti-pd-1/pdl-1 and anti-ctla-4 therapies for cancer immune evasion: A bibliometric analysis. *Hum Vaccin Immunother*. (2024) 20:2318815. doi: 10.1080/21645515.2024.2318815

34. Liu J, Wu Z, Sun R, Nie S, Meng H, Zhong Y, et al. Using mrnasi to identify prognostic-related genes in endometrial carcinoma based on wgcna. *Life Sci*. (2020) 258:118231. doi: 10.1016/j.lfs.2020.118231

35. Wu E, Fan X, Tang T, Li J, Wang J, Liu X, et al. Biomarkers discovery for endometrial cancer: A graph convolutional sample network method. *Comput Biol Med*. (2022) 150:106200. doi: 10.1016/j.compbiomed.2022.106200

36. Regulski MJ. Cellular senescence: what, why, and how. *Wounds*. (2017) 29:168–74.

37. Zeng Q, Gong Y, Zhu N, Shi Y, Zhang C, Qin L. Lipids and lipid metabolism in cellular senescence: emerging targets for age-related diseases. *Ageing Res Rev*. (2024) 97:102294. doi: 10.1016/j.arr.2024.102294

38. Gao L, Wang X, Wang X, Wang F, Tang J, Ji J. A prognostic model and immune regulation analysis of uterine corpus endometrial carcinoma based on cellular senescence. *Front Oncol*. (2022) 12:1054564. doi: 10.3389/fonc.2022.1054564

39. Juanes-Velasco P, Landeira-Viñuela A, Acebes-Fernandez V, Hernández ÁP, Garcia-Vaquero ML, Arias-Hidalgo C, et al. Deciphering human leukocyte antigen susceptibility maps from immunopeptidomics characterization in oncology and infections. *Front Cell Infect Microbiol*. (2021) 11:642583. doi: 10.3389/fcimb.2021.642583

40. Aptsiauri N, Garrido F. The challenges of hla class I loss in cancer immunotherapy: facts and hopes. *Clin Cancer Res*. (2022) 28:5021–9. doi: 10.1158/1078–0432.Ccr-21–3501

41. Qin L, Waseem TC, Sahoo A, Bieerkehazhi S, Zhou H, Galkina EV, et al. Insights into the molecular mechanisms of T follicular helper-mediated immunity and pathology. *Front Immunol*. (2018) 9:1884. doi: 10.3389/fimmu.2018.01884

42. Liu J, Geng X, Hou J, Wu G. New insights into M1/M2 macrophages: key modulators in cancer progression. *Cancer Cell Int*. (2021) 21:389. doi: 10.1186/s12935–021-02089–2

43. Mellman I. Dendritic cells: master regulators of the immune response. *Cancer Immunol Res*. (2013) 1:145–9. doi: 10.1158/2326–6066.Cir-13–0102

44. Guo Z, Zhang R, Yang AG, Zheng G. Diversity of immune checkpoints in cancer immunotherapy. *Front Immunol*. (2023) 14:1121285. doi: 10.3389/fimmu.2023.1121285

45. Liu W, Shen D, Ju L, Zhang R, Du W, Jin W, et al. Mybl2 promotes proliferation and metastasis of bladder cancer through transactivation of cdca3. *Oncogene*. (2022) 41:4606–17. doi: 10.1038/s41388-022-02456-x

46. Li Q, Wang M, Hu Y, Zhao E, Li J, Ren L, et al. Mybl2 disrupts the hippo-yap pathway and confers castration resistance and metastatic potential in prostate cancer. *Theranostics*. (2021) 11:5794–812. doi: 10.7150/thno.56604

47. Qiu X, He H, Zeng H, Tong X, Zhang C, Liu Y, et al. Integrative transcriptome analysis identifies mybl2 as a poor prognosis marker for osteosarcoma and a pan-cancer marker of immune infiltration. *Genes Dis*. (2024) 11:101004. doi: 10.1016/j.gendis.2023.04.035

48. Yang W, Chen H, Ma L, Wei M, Xue X, Li Y, et al. The oncogene mybl2 promotes the Malignant phenotype and suppresses apoptosis through hedgehog signaling pathway in clear cell renal cell carcinoma. *Heliyon*. (2024) 10:e27772. doi: 10.1016/j.heliyon.2024.e27772

49. Long J, Zhu B, Tian T, Ren L, Tao Y, Zhu H, et al. Activation of ubec2 by transcription factor mybl2 affects DNA damage and promotes gastric cancer progression and cisplatin resistance. *Open Med (Wars)*. (2023) 18:20230757. doi: 10.1515/med-2023–0757

50. Pan B, Wan T, Zhou Y, Huang S, Yuan L, Jiang Y, et al. The mybl2-ccl2 axis promotes tumor progression and resistance to anti-pd-1 therapy in ovarian cancer by inducing immunosuppressive macrophages. *Cancer Cell Int*. (2023) 23:248. doi: 10.1186/s12935–023-03079–2

51. Fan X, Wang Y, Jiang T, Liu T, Jin Y, Du K, et al. B-myb accelerates colorectal cancer progression through reciprocal feed-forward transactivation of E2f2. *Oncogene*. (2021) 40:5613–25. doi: 10.1038/s41388–021-01961–9

52. Drisaldi B, Colnaghi L, Levine A, Huang Y, Snyder AM, Metzger DJ, et al. Cytoplasmic polyadenylation element binding proteins cpeb1 and cpeb3 regulate the translation of fosb and are required for maintaining addiction-like behaviors induced by cocaine. *Front Cell Neurosci*. (2020) 14:207. doi: 10.3389/fncel.2020.00207

53. Kochanek DM, Wells DG. Cpeb1 regulates the expression of mtdh/aeg-1 and glioblastoma cell migration. *Mol Cancer Res*. (2013) 11:149–60. doi: 10.1158/1541–7786.Mcr-12–0498

54. Shao K, Pu W, Zhang J, Guo S, Qian F, Glurich I, et al. DNA hypermethylation contributes to colorectal cancer metastasis by regulating the binding of cebpb and tfcp2 to the cpeb1 promoter. *Clin Epigenet*. (2021) 13:89. doi: 10.1186/s13148-021-01071-z

55. Xu M, Fang S, Song J, Chen M, Zhang Q, Weng Q, et al. Cpeb1 mediates hepatocellular carcinoma cancer stemness and chemoresistance. *Cell Death Dis*. (2018) 9:957. doi: 10.1038/s41419–018-0974–2

56. Abu-Rustum N, Yashar C, Arend R, Barber E, Bradley K, Brooks R, et al. Uterine neoplasms, version 1.2023, nccn clinical practice guidelines in oncology. *J Natl Compr Canc Netw*. (2023) 21:181–209. doi: 10.6004/jnccn.2023.0006

57. Mutlu L, Harold J, Tymon-Rosario J, Santin AD. Immune checkpoint inhibitors for recurrent endometrial cancer. *Expert Rev Anticancer Ther*. (2022) 22:249–58. doi: 10.1080/14737140.2022.2044311

58. Mahdi H, Chelariu-Raicu A, Slomovitz BM. Immunotherapy in endometrial cancer. *Int J Gynecol Cancer*. (2023) 33:351–7. doi: 10.1136/ijgc-2022–003675

59. Peng H, He X, Wang Q. Immune checkpoint blockades in gynecological cancers: A review of clinical trials. *Acta Obstet Gynecol Scand*. (2022) 101:941–51. doi: 10.1111/aogs.14412

60. Bhangoo MS, Boasberg P, Mehta P, Elvin JA, Ali SM, Wu W, et al. Tumor mutational burden guides therapy in a treatment refractory pole-mutant uterine carcinosarcoma. *Oncologist*. (2018) 23:518–23. doi: 10.1634/theoncologist.2017–0342

61. Bartoletti M, Montico M, Lorusso D, Mazzeo R, Oaknin A, Musacchio L, et al. Incorporation of anti-pd1 or anti pd-L1 agents to platinum-based chemotherapy for the primary treatment of advanced or recurrent endometrial cancer. *A Meta-Analysis Cancer Treat Rev*. (2024) 125:102701. doi: 10.1016/j.ctrv.2024.102701

62. Bruchim I, Capasso I, Polonsky A, Meisel S, Salutari V, Werner H, et al. New therapeutic targets for endometrial cancer: A glimpse into the preclinical sphere. *Expert Opin Ther Targets*. (2024) 28:29–43. doi: 10.1080/14728222.2024.2316739

63. van der Woude H, Hally KE, Currie MJ, Gasser O, Henry CE. Importance of the endometrial immune environment in endometrial cancer and associated therapies. *Front Oncol*. (2022) 12:975201. doi: 10.3389/fonc.2022.975201

ARTICLE

OPEN

doi:10.1038/nature12113

Integrated genomic characterization of endometrial carcinoma

The Cancer Genome Atlas Research Network*

We performed an integrated genomic, transcriptomic and proteomic characterization of 373 endometrial carcinomas using array- and sequencing-based technologies. Uterine serous tumours and ~25% of high-grade endometrioid tumours had extensive copy number alterations, few DNA methylation changes, low oestrogen receptor/progesterone receptor levels, and frequent *TP53* mutations. Most endometrioid tumours had few copy number alterations or *TP53* mutations, but frequent mutations in *PTEN*, *CTNNB1*, *PIK3CA*, *ARID1A* and *KRAS* and novel mutations in the SWI/SNF chromatin remodelling complex gene *ARID5B*. A subset of endometrioid tumours that we identified had a markedly increased transversion mutation frequency and newly identified hotspot mutations in *POLE*. Our results classified endometrial cancers into four categories: *POLE* ultramutated, microsatellite instability hypermutated, copy-number low, and copy-number high. Uterine serous carcinomas share genomic features with ovarian serous and basal-like breast carcinomas. We demonstrated that the genomic features of endometrial carcinomas permit a reclassification that may affect post-surgical adjuvant treatment for women with aggressive tumours.

Endometrial cancer arises from the lining of the uterus. It is the fourth most common malignancy among women in the United States, with an estimated 49,500 new cases and 8,200 deaths in 2013 (ref. 1). Most patients present with low-grade, early-stage disease. The majority of patients with more aggressive, high-grade tumours who have disease spread beyond the uterus will progress within 1 year (refs 2, 3). Endometrial cancers have been broadly classified into two groups[4]. Type I endometrioid tumours are linked to oestrogen excess, obesity, hormone-receptor positivity, and favourable prognosis compared with type II, primarily serous, tumours that are more common in older, non-obese women and have a worse outcome. Early-stage endometrioid cancers are often treated with adjuvant radiotherapy, whereas serous tumours are treated with chemotherapy, similar to advanced-stage cancers of either histological subtype. Therefore, proper subtype classification is crucial for selecting appropriate adjuvant therapy.

Several previous reports suggest that *PTEN* mutations occur early in the neoplastic process of type I tumours and co-exist frequently with other mutations in the phosphatidylinositol-3-OH kinase (PI(3)K)/AKT pathway[5,6]. Other commonly mutated genes in type I tumours include *FGFR2*, *ARID1A*, *CTNNB1*, *PIK3CA*, *PIK3R1* and *KRAS*[7–9]. Microsatellite instability (MSI) is found in approximately one-third of type I tumours, but is infrequent in type II tumours[10]. *TP53*, *PIK3CA* and *PPP2R1A* mutations are frequent in type II tumours[11,12]. Most of these studies have been limited to DNA sequencing only with samples of heterogeneous histological subtypes and tumour grades. We present a comprehensive, multiplatform analysis of 373 endometrial carcinomas including low-grade endometrioid, high-grade endometrioid, and serous carcinomas. This integrated analysis provides key molecular insights into tumour classification, which may have a direct effect on treatment recommendations for patients, and provides opportunities for genome-guided clinical trials and drug development.

Results

Tumour samples and corresponding germline DNA were collected from 373 patients, including 307 endometrioid and 66 serous (53) or mixed histology (13) cases. Local Institutional Review Boards approved all tissue acquisition. The clinical and pathological characteristics of the samples generally reflect a cross-section of individuals with recurrent endometrial cancer[2,3] (Supplementary Table 1.1). The median follow-up of the cohort was 32 months (range, 1–195 months); 21% of the patients have recurred, and 11% have died. Comprehensive molecular analyses were performed at independent centres using six genomic or proteomic platforms (Supplementary Table 1.2). MSI testing performed on all samples using seven repeat loci (Supplementary Table 1.3) found MSI in 40% of endometrioid tumours and 2% of serous tumours.

Somatic copy number alterations

Somatic copy number alterations (SCNAs) were assessed in 363 endometrial carcinomas. Unsupervised hierarchical clustering grouped the tumours into four clusters (Fig. 1a). The first three copy-number clusters were composed almost exclusively (97%) of endometrioid tumours without significant differences in tumour grades. Cluster 1 tumours were nearly devoid of broad SCNAs, averaging less than 0.5% genome alteration, with no significant recurrent events. Cluster 1 tumours also had significantly increased non-synonymous mutation rates compared to all others (median 7.2×10^{-6} versus 1.7×10^{-6} mutations per megabase (Mb), $P < 0.001$). Copy-number clusters 2 and 3 consisted mainly of endometrioid tumours, distinguished by more frequent 1q amplification in cluster 3 than cluster 2 (100% of cluster 3 tumours versus 33% of cluster 2 tumours) and worse progression-free survival ($P = 0.003$, log-rank versus clusters 1 and 2; Fig. 1b).

Most of the serous (50 out of 53; 94%) and mixed histology (8 out of 13; 62%) tumours clustered with 36 (12%) of the 289 endometrioid tumours, including 24% of grade 3 and 5% of grade 1 or 2, into copy-number cluster 4; a single group characterized by a very high degree of SCNAs (Supplementary Fig. 2.1; focal SCNAs with false discovery rate (FDR) < 0.15, and Supplementary Data 2.1). Cluster 4 tumours were characterized by significantly recurrent previously reported focal amplifications of the oncogenes *MYC* (8q24.12), *ERBB2* (17q12) and *CCNE1* (19q12)[13], and by SCNAs previously unreported in endometrial cancers including those containing *FGFR3* (4p16.3) and *SOX17* (8q11.23). Cluster 4 tumours also had frequent *TP53* mutations (90%),

*Lists of participants and their affiliations appear at the end of the paper.

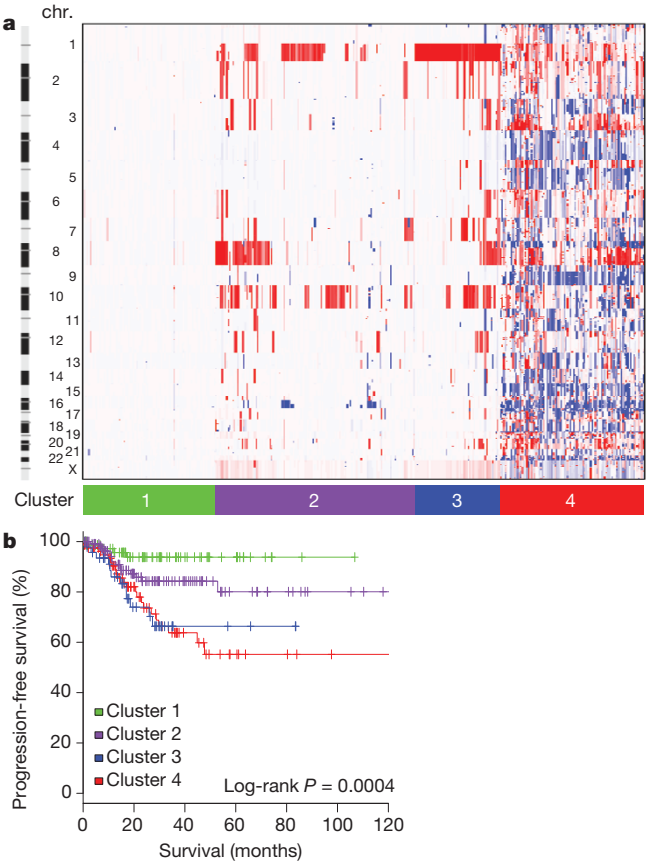


Figure 1 | SCNAs in endometrial carcinomas. **a**, Tumours were hierarchically clustered into four groups based on SCNAs. The heat map shows SCNAs in each tumour (horizontal axis) plotted by chromosomal location (vertical axis). Chr., chromosome. **b**, Kaplan–Meier curves of progression-free survival for each copy-number cluster.

little MSI (6%), and fewer *PTEN* mutations (11%) than other endometrioid tumours (84%). Overall, these findings suggest that a subset of endometrial tumours contain distinct patterns of SCNAs and mutations that do not correlate with traditional tumour histology or grade.

As expected, tumours in the 'serous-like' cluster (cluster 4) had significantly worse progression-free survival than tumours in the endometrioid cluster groups ($P = 0.003$, log-rank, Fig. 1b). Potential therapeutically relevant SCNAs included the cluster 2 15q26.2 focal amplification, which contained *IGF1R*; and cluster 4 amplifications of *ERBB2*, *FGFR1* and *FGFR3*, and *LRP1B* deletion, which was recently associated with resistance to liposomal doxorubicin in serous ovarian cancer[14].

Exome sequence analysis

We sequenced the exomes of 248 tumour/normal pairs. On the basis of a combination of somatic nucleotide substitutions, MSI and SCNAs, the endometrial tumours were classified into four groups (Fig. 2a, b): (1) an ultramutated group with unusually high mutation rates (232×10^{-6} mutations per Mb) and a unique nucleotide change spectrum; (2) a hypermutated group (18×10^{-6} mutations per Mb) of MSI tumours, most with *MLH1* promoter methylation; (3) a group with lower mutation frequency (2.9×10^{-6} mutations per Mb) and most of the microsatellite stable (MSS) endometrioid cancers; and (4) a group that consists primarily of serous-like cancers with extensive SCNA (copy-number cluster 4) and a low mutation rate (2.3×10^{-6} mutations per Mb). The ultramutated group consisted of 17 (7%) tumours exemplified by an increased C→A transversion frequency, all with mutations in the exonuclease domain of *POLE*, and an improved progression-free survival (Fig. 2a, c). *POLE* is a catalytic subunit of DNA polymerase epsilon involved in nuclear DNA replication and repair. We identified hotspot mutations in *POLE* at Pro286Arg and Val411Leu present in 13 (76%) of the 17 ultramutated samples. Significantly mutated genes (SMGs) identified at low FDRs (Q) in this subset included *PTEN* (94%, $Q = 0$), *PIK3R1* (65%, $Q = 8.3 \times 10^{-7}$), *PIK3CA* (71%, $Q = 9.1 \times 10^{-5}$), *FBXW7* (82%, $Q = 1.4 \times 10^{-4}$), *KRAS* (53%, $Q = 9.2 \times 10^{-4}$) and *POLE* (100%, $Q = 4.2 \times 10^{-3}$). Mutation rates in *POLE* mutant endometrial and previously reported ultramutated colorectal tumours exceeded those found in any other lineage including lung cancer and melanoma[15–17]. Germline susceptibility variants have been reported in *POLE* (Leu424Val) and *POLD1* (Ser478Asn), but were not found in our endometrial normal exome-seq reads[18].

The MSI endometrioid tumours had a mutation frequency approximately tenfold greater than MSS endometrioid tumours, few SCNAs, frameshift deletions in *RPL22*, frequent non-synonymous *KRAS* mutations, and few mutations in *FBXW7*, *CTNNB1*, *PPP2R1A* and *TP53*. The MSS, copy-number low, endometrioid tumours had an unusually high frequency of *CTNNB1* mutations (52%); the only gene with a higher mutation frequency than the MSI samples. The copy-number high group contained all of the remaining serous cases and one-quarter of the grade 3 endometrioid cases. Most of these tumours had *TP53* mutations and a high frequency of *FBXW7* (22%, $Q = 0$) and *PPP2R1A* (22%, $Q = 1.7 \times 10^{-16}$) mutations, previously reported as common in uterine serous but not endometrioid carcinomas. Thus, a subset of high-grade endometrioid tumours had similar SCNAs and mutation spectra as uterine serous carcinomas, suggesting that these patients might benefit from treatment approaches that parallel those for serous tumours.

There were 48 genes with differential mutation frequencies across the four groups (Fig. 2d and Supplementary Data 3.1). *ARID5B*, a member of the same AT-rich interaction domain (ARID) family as *ARID1A*, was more frequently mutated in MSI (23.1%) than in either MSS endometrioid (5.6%) or high SCNA serous tumours (0%), a novel finding for endometrial cancer. Frameshifting *RPL22* indels near a homopolymer at Lys 15 were almost exclusively found in the MSI group (36.9%). The *TP53* mutation frequency (>90%) in serous tumours differentiated them from the endometrioid subtypes (11.4%). However, many (10 out of 20; 50%) endometrioid tumours with a non-silent *TP53* mutation also had non-silent mutations in *PTEN*, compared to only 1 out of 39 (2.6%) serous tumours with non-silent *TP53* mutations. Although *TP53* mutations are not restricted to serous tumours, the co-existing *PTEN* mutations in the endometrioid cases suggest a distinct tumorigenic mechanism.

Comparisons of 66 SMGs between traditional histological subtypes are provided (Supplementary Methods 3), and SMGs across other subcohorts can be found in Supplementary Data 3.2. The spectrum of *PIK3CA* and *PTEN* mutations in endometrial cancer also differed from other solid tumours (Supplementary Methods 3). Integrated analysis may be useful for identifying histologically misclassified cases. For example, a single serous case was identified without a *TP53* mutation or extensive SCNAs and with a *KRAS* mutation and high mutation rate. After re-review of the histological section, the case was deemed consistent with a grade 3 endometrioid tumour, demonstrating how molecular analysis could reclassify tumour histology and potentially affect treatment decisions.

Multiplatform subtype classifications

All of the endometrial tumours were examined for messenger RNA expression ($n = 333$), protein expression ($n = 293$), microRNA expression ($n = 367$), and DNA methylation ($n = 373$) (Supplementary Methods 4–7). Unsupervised k-means clustering of mRNA expression from RNA sequencing identified three robust clusters termed 'mitotic', 'hormonal' and 'immunoreactive' (Supplementary Fig. 4.1) that were significantly correlated with the four integrated clusters; *POLE*, MSI, copy-number low and copy-number high ($P < 0.0001$). Supervised analysis identified signature genes of the *POLE* cluster ($n = 17$) mostly involved in cellular metabolism (Fig. 3a). Among the few signature genes

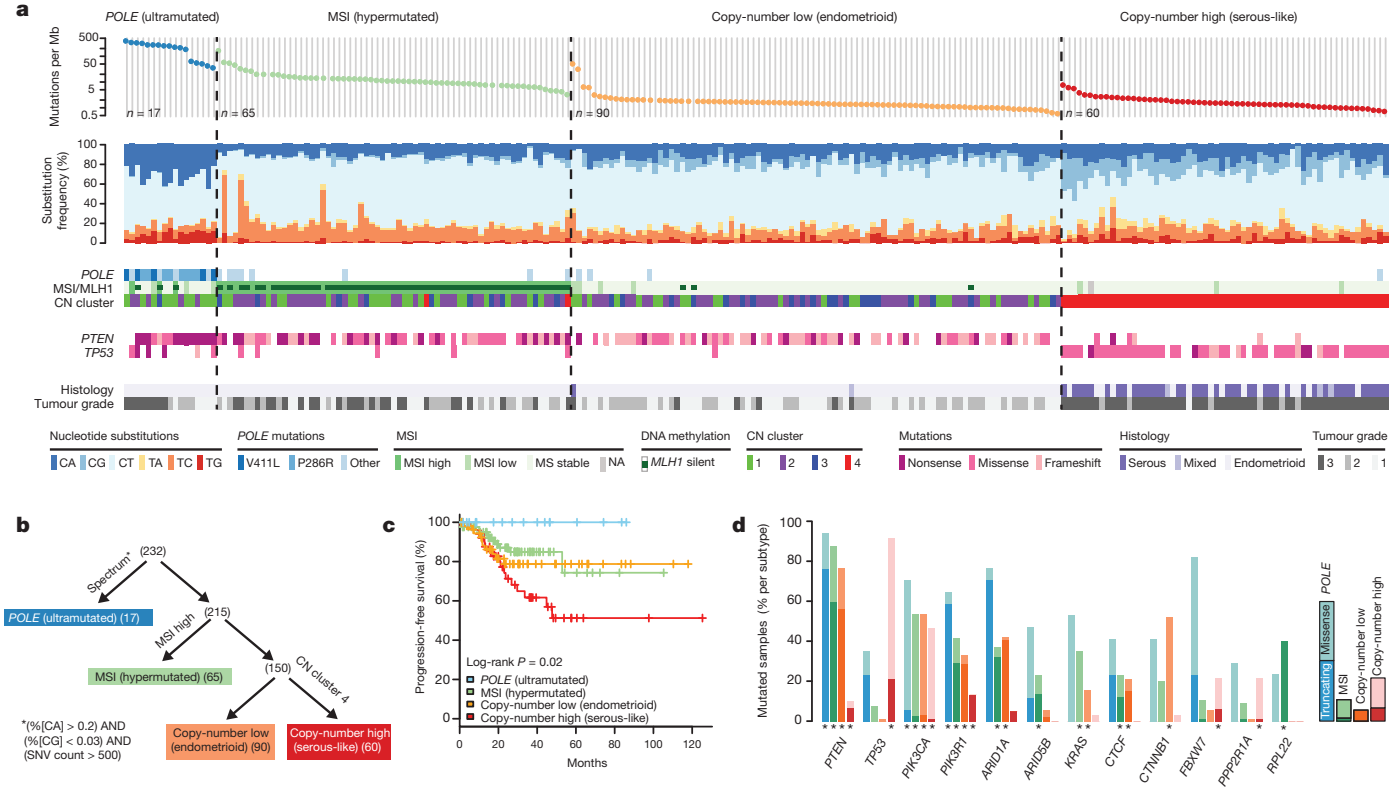


Figure 2 | Mutation spectra across endometrial carcinomas. **a**, Mutation frequencies (vertical axis, top panel) plotted for each tumour (horizontal axis). Nucleotide substitutions are shown in the middle panel, with a high frequency of C-to-A transversions in the samples with *POLE* exonuclease mutations. CN, copy number. **b**, Tumours were stratified into the four groups by (1) nucleotide substitution frequencies and patterns, (2) MSI status, and (3) copy-number cluster. SNV, single nucleotide variant. **c**, *POLE*-mutant tumours have significantly better progression-free survival, whereas copy-number high tumours have the poorest outcome. **d**, Recurrently mutated genes are different between the four subgroups. Shown are the mutation frequencies of all genes that were significantly mutated in at least one of the four subgroups (MUSiC, asterisk denotes FDR < 0.05).

in the MSI cluster was decreased *MLH1* mRNA expression, probably due to its promoter methylation. Increased progesterone receptor (*PGR*) expression was noted in the copy-number low cluster, suggesting responsiveness to hormonal therapy. The copy-number high cluster, which included most of the serous and serous-like endometrioid tumours, exhibited the greatest transcriptional activity exemplified by increased cell cycle deregulation (for example, *CCNE1*, *PIK3CA*, *MYC* and *CDKN2A*) and *TP53* mutation (Supplementary Figs 4.2 and 4.3). This is consistent with reports that increased CDKN2A can distinguish serous from endometrioid carcinomas[19]. Approximately 85% of cases in the copy-number high cluster shared membership with the 'mitotic' mRNA subtype.

Supervised clustering of the reverse phase protein array (RPPA) expression data was consistent with loss of function for many of the mutated genes (Fig. 3b). *TP53* was frequently mutated in the copy-number high group ($P = 2.5 \times 10^{-27}$) and its protein expression was also increased, suggesting that these mutations are associated with increased expression. By contrast, *PTEN* ($P = 2.8 \times 10^{-19}$) and *ARID1A* ($P = 1.2 \times 10^{-6}$) had high mutation rates in the remaining groups, but their expression was decreased, suggesting inactivating mutations in both genes. The copy-number high group also had decreased levels of phospho-AKT, consistent with downregulation of the AKT pathway. The copy-number low group had raised RAD50 expression, which is associated with DNA repair, explaining some of the differences between the copy-number high and low groups. The *POLE* group had high expression of ASNS and CCNB1, whereas the MSI tumours had both high phospho-AKT and low PTEN expression.

Unsupervised clustering of DNA methylation data generated from Illumina Infinium DNA methylation arrays revealed four unique subtypes (MC1–4) that support the four integrative clusters. A heavily methylated subtype (MC1) reminiscent of the CpG island methylator phenotype (CIMP) described in colon cancers and glioblastomas[20–22] was associated with the MSI subtype and attributable to promoter hypermethylation of *MLH1*. A serous-like cluster (MC3) with minimal DNA methylation changes was composed primarily of serous tumours and some endometrioid tumours (Supplementary Fig. 7.1) and contained most of the copy-number high tumours.

Integrative clustering using the iCluster framework returned two major clusters split primarily on serous and endometrioid histology highlighting *TP53* mutations, lack of *PTEN* mutation and encompassing almost exclusively copy-number high tumours[23] (Supplementary Fig. 8.1). We developed a new clustering algorithm, called SuperCluster, to derive overall subtypes based on sample cluster memberships across all data types (Supplementary Fig. 9.1). SuperCluster identified four clusters that generally confirmed the contributions of individual platforms to the overall integrated clusters. No major batch effects were identified for any platform (Supplementary Methods 10).

Structural aberrations

To identify somatic chromosomal aberrations, we performed low-pass, paired-end, whole-genome sequencing on 106 tumours with matched normals. We found recurrent translocations involving genes in several pathways including WNT, EGFR–RAS–MAPK, PI(3)K, protein kinase A, retinoblastoma and apoptosis. The most frequent translocations (5 out of 106) involved a member of the BCL family (*BCL2*, *BCL7A*, *BCL9* and *BCL2L11*). Four of these were confirmed by identification of the translocation junction point and two were also confirmed by high-throughput RNA sequencing (RNA-Seq). In all cases the translocations result in in-frame fusions and are predicted to result in activation or increased expression of the BCL family members (Supplementary Fig. 3.2). Translocations involving members of the BCL family leading to reduced apoptosis have been

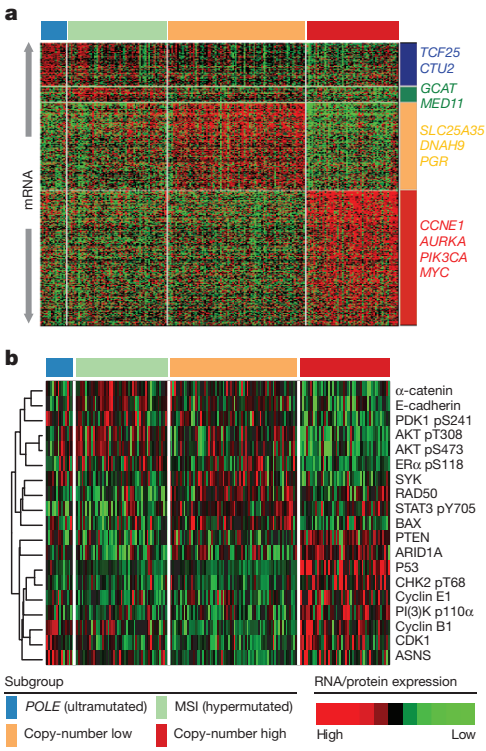


Figure 3 | Gene expression across integrated subtypes in endometrial carcinomas. **a**, Supervised analysis of ~1,500 genes significantly associated with integrated subtypes. **b**, Heat map of protein expression clusters, supervised by integrated subtypes. Samples are in columns; genes or proteins are in rows.

described in other tumour types[24] and our results suggest that similar mechanisms may be operative here.

Pathway alterations

Multiple platform data were integrated to identify recurrently altered pathways in the four endometrial cancer integrated subgroups. Because of the high background mutation rate and small sample size, we excluded the *POLE* subgroup from this analysis. Considering all recurrently mutated, homozygously deleted, and amplified genes, we used MEMo[25] to identify gene networks with mutually exclusive alteration patterns in each subgroup. The most significant module was found in the copy-number low group and contained *CTNNB1*, *KRAS* and *SOX17* (Fig. 4a). The very strong mutual exclusivity between mutations in these three genes suggests that alternative mechanisms activate WNT signalling in endometrioid endometrial cancer. Activating *KRAS* mutations have been shown to increase the stability of β-catenin via glycogen synthase kinase 3β (GSK-3β), leading to an alternative mechanism of β-catenin activation other than adenomatous polyposis coli degradation[26]. *SOX17*, which mediates proteasomal degradation of β-catenin[27,28], is mutated exclusively in the copy-number low group (8%) at recurrent positions (Ala96Gly and Ser403Ile) not previously described. Other genes with mutually exclusive alteration patterns in this module were *FBXW7*, *FGFR2* and *ERBB2* (ref. 29). *ERBB2* was focally amplified with protein overexpression in 25% of the serous or serous-like tumours, suggesting a potential role for human epidermal growth factor receptor 2 (HER2)-targeted inhibitors. A small clinical trial of trastuzumab found no activity in endometrial carcinoma, but accrued few HER2 fluorescence *in situ* hybridization (FISH)-amplified serous carcinomas[30].

PIK3CA and *PIK3R1* mutations were frequent and showed a strong tendency for mutual exclusivity in all subgroups, but unlike other tumour types, they co-occurred with *PTEN* mutations in the MSI and copy-number low subgroups as previously reported[5,9] (Fig. 4b). The copy-number high subgroup showed mutual exclusivity between alterations of all three genes. Overall, 93% of endometrioid tumours had mutations that suggested potential for targeted therapy with PI(3)K/AKT pathway inhibitors.

Consensus clustering of copy number, mRNA expression and pathway interaction data for 324 samples yielded five PARADIGM clusters with distinct pathway activation patterns[31] (Fig. 4c and Supplementary Methods 11). PARADIGM cluster 1 had the lowest level of MYC pathway activation and highest level of WNT pathway activation, consistent with its composition of copy-number low cases having frequent *CTNNB1* mutations. PARADIGM cluster 3 was composed predominantly of the copy-number high cases, with relatively high MYC/MAX signalling but low oestrogen receptor/FOXA1 signalling and p53 activity. Only *TP53* truncation and not missense mutations were implicated as loss-of-function mutations, suggesting different classes of p53 mutations may have distinct signalling consequences. PARADIGM cluster 5 was enriched for hormone receptor expression.

Comparison to ovarian and breast cancers

The clinical and pathologic features of uterine serous carcinoma and high-grade serous ovarian carcinoma (HGSOC) are quite similar. HGSOC shares many similar molecular features with basal-like breast carcinoma[32]. Focal SCNA patterns were similar between these three tumour subtypes and unsupervised clustering identified relatedness (Fig. 5a and Supplementary Fig. 12.1). Supervised analysis of transcriptome data sets showed high correlation between tumour subtypes (Supplementary Fig. 12.2). The MC3 DNA methylation subtype with minimal DNA methylation changes was also similar to basal-like breast and HGSOCs (Supplementary Fig. 12.3). A high frequency of *TP53* mutations is shared across these tumour subtypes (uterine serous, 91%; HGSOC, 96%; basal-like breast, 84%)[33,34], as is the very low frequency of *PTEN* mutations (uterine serous, 2%; HGSOC, 1%; basal-like breast, 1%). Differences included a higher frequency of *FBXW7*, *PPP2R1A* and *PIK3CA* mutations in uterine serous compared to basal-like breast and HGSOCs (Fig. 5b). We showed that uterine serous carcinomas share many molecular features with both HGSOCs and basal-like breast carcinomas, despite more frequent mutations, suggesting new opportunities for overlapping treatment paradigms.

Discussion

This integrated genomic and proteomic analysis of 373 endometrial cancers provides insights into disease biology and diagnostic classification that could have immediate therapeutic application. Our analysis identified four new groups of tumours based on integrated genomic data, including a novel *POLE* subtype in ~10% of endometrioid tumours. Ultrahigh somatic mutation frequency, MSS, and common, newly identified hotspot mutations in the exonuclease domain of *POLE* characterize this subtype. SCNAs add a layer of resolution, revealing that most endometrioid tumours have few SCNAs, most serous and serous-like tumours exhibit extensive SCNAs, and the extent of SCNA roughly correlates with progression-free survival.

Endometrial cancer has more frequent mutations in the PI(3)K/AKT pathway than any other tumour type studied by The Cancer Genome Atlas (TCGA) so far. Endometrioid endometrial carcinomas share many characteristics with colorectal carcinoma including a high frequency of MSI (40% and 11%, respectively), *POLE* mutations (7% and 3%, respectively) leading to ultrahigh mutation rates, and frequent activation of WNT/CTNNB1 signalling; yet endometrial carcinomas have novel exclusivity of *KRAS* and *CTNNB1* mutations and a distinct mechanism of pathway activation. Uterine serous carcinomas share many similar characteristics with basal-like breast and HGSOCs; three tumour types with high-frequency non-silent *TP53* mutations and extensive SCNA. However, the high frequency of *PIK3CA*, *FBXW7*, *PPP2R1A* and *ARID1A* mutations in uterine serous carcinomas are not found in basal-like breast and HGSOCs. The frequency of mutations in *PIK3CA*, *FBXW7* and *PPP2R1A* was ~30% higher than in a recently

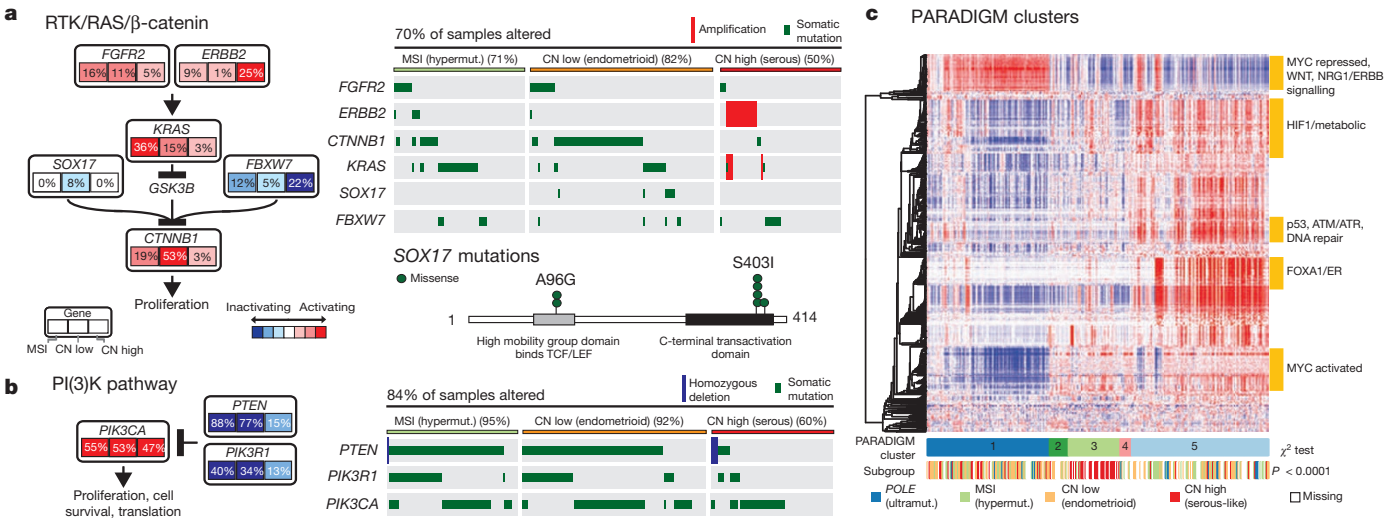


Figure 4 | Pathway alterations in endometrial carcinomas. **a**, The RTK/RAS/β-catenin pathway is altered through several mechanisms that exhibit mutually exclusive patterns. Alteration frequencies are expressed as a percentage of all cases. The right panel shows patterns of occurrence. **b**, The PI(3)K pathway has mutually exclusive *PIK3CA* and *PIK3R1* alterations that frequently co-occur with *PTEN* alterations in the MSI and copy-number low subgroups. **c**, Heat map display of top 1,000 varying pathway features within PARADIGM consensus clusters. Samples were arranged in order of their consensus cluster membership. The genomic subtype for each sample is displayed below the consensus clusters.

reported study of 76 uterine serous carcinomas[11], but similar to another study[12]. Uterine serous carcinomas have *ERBB2* amplification in 27% of tumours and *PIK3CA* mutations in 42%, which provide translational opportunities for targeted therapeutics.

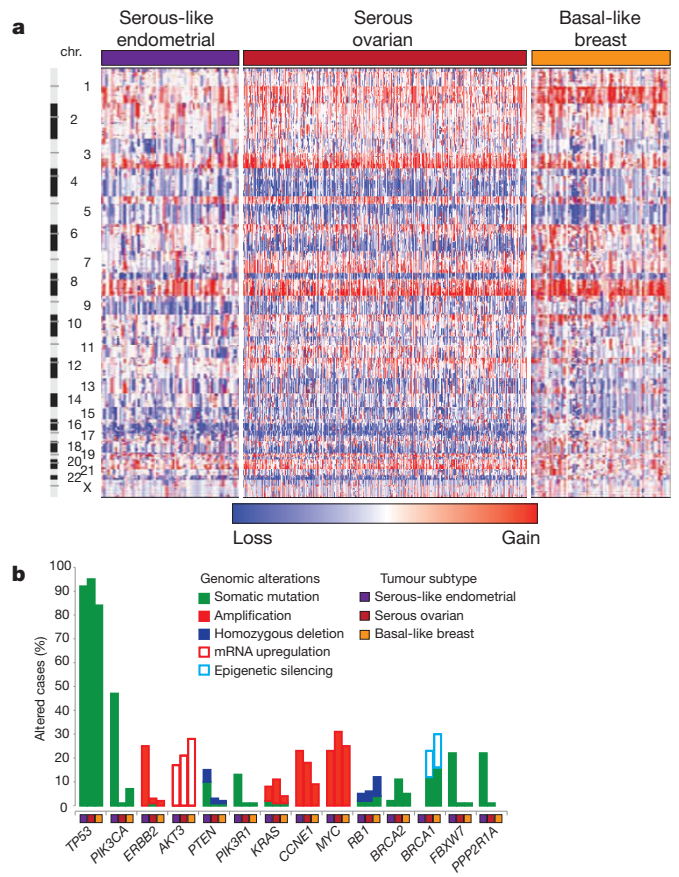


Figure 5 | Genomic relationships between endometrial serous-like, ovarian serous, and basal-like breast carcinomas. **a**, SCNAs for each tumour type. **b**, Frequency of genomic alterations present in at least 10% of one tumour type.

Early stage type I endometrioid tumours are often treated with adjuvant radiotherapy, whereas similarly staged type II serous tumours are treated with chemotherapy. High-grade serous and endometrioid endometrial carcinomas are difficult to subtype correctly, and intra-observer concordance among speciality pathologists is low[7,34–36]. Our molecular characterization data demonstrate that ~25% of tumours classified as high-grade endometrioid by pathologists have a molecular phenotype similar to uterine serous carcinomas, including frequent *TP53* mutations and extensive SCNA. The compelling similarities between this subset of endometrioid tumours and uterine serous carcinomas suggest that genomic-based classification may lead to improved management of these patients. Clinicians should carefully consider treating copy-number-altered endometrioid patients with chemotherapy rather than adjuvant radiotherapy and formally test such hypotheses in prospective clinical trials. Furthermore, the marked molecular differences between endometrioid and serous-like tumours suggest that these tumours warrant separate clinical trials to develop the independent treatment paradigms that have improved outcomes in other tumour types, such as breast cancer.

METHODS SUMMARY

Biospecimens were obtained from 373 patients after Institutional Review Board-approved consents. DNA and RNA were co-isolated using a modified AllPrep kit (Qiagen). We used Affymetrix SNP 6.0 microarrays to detect SCNAs in 363 samples and GISTIC analysis to identify recurrent events[37]. The exomes of 248 tumours were sequenced to a read-depth of at least ×20. We performed low-pass whole-genome sequencing on 107 tumours to a mean depth of ×6. Consensus clustering was used to analyse mRNA, miRNA, RPPA and methylation data with methods previously described[38–40]. Integrated cross-platform analyses were performed using MEMo, iCluster and PARADIGM[25,31].

Received 10 December 2012; accepted 21 March 2013.

1. Siegel, R., Naishadham, D. & Jemal, A. Cancer statistics, 2013. *CA Cancer J. Clin.* **63,** 11–30 (2013).
2. Fleming, G. F. *et al.* Phase III trial of doxorubicin plus cisplatin with or without paclitaxel plus filgrastim in advanced endometrial carcinoma: a Gynecologic Oncology Group Study. *J. Clin. Oncol.* **22,** 2159–2166 (2004).
3. Sutton, G. *et al.* Whole abdominal radiotherapy in the adjuvant treatment of patients with stage III and IV endometrial cancer: a gynecologic oncology group study. *Gynecol. Oncol.* **97,** 755–763 (2005).
4. Lax, S. F. & Kurman, R. J. A dualistic model for endometrial carcinogenesis based on immunohistochemical and molecular genetic analyses. *Verh. Dtsch. Ges. Pathol.* **81,** 228–232 (1997).

5. Cheung, L. W. *et al.* High frequency of *PIK3R1* and *PIK3R2* mutations in endometrial cancer elucidates a novel mechanism for regulation of PTEN protein stability. *Cancer Discov.* **1,** 170–185 (2011).
6. Levine, R. L. *et al. PTEN* mutations and microsatellite instability in complex atypical hyperplasia, a precursor lesion to uterine endometrioid carcinoma. *Cancer Res.* **58,** 3254–3258 (1998).
7. McConechy, M. K. *et al.* Use of mutation profiles to refine the classification of endometrial carcinomas. *J. Pathol.* **228,** 20–30 (2012).
8. Byron, S. A. *et al. FGFR2* point mutations in 466 endometrioid endometrial tumors: relationship with MSI, *KRAS*, *PIK3CA*, *CTNNB1* mutations and clinicopathological features. *PLoS ONE* **7,** e30801 (2012).
9. Urick, M. E. *et al. PIK3R1* (p85α) is somatically mutated at high frequency in primary endometrial cancer. *Cancer Res.* **71,** 4061–4067 (2011).
10. Zighelboim, I. *et al.* Microsatellite instability and epigenetic inactivation of *MLH1* and outcome of patients with endometrial carcinomas of the endometrioid type. *J. Clin. Oncol.* **25,** 2042–2048 (2007).
11. Kuhn, E. *et al.* Identification of molecular pathway aberrations in uterine serous carcinoma by genome-wide analyses. *J. Natl. Cancer Inst.* **104,** 1503–1513 (2012).
12. Le Gallo, M. *et al.* Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. *Nature Genet.* **44,** 1310–1315 (2012).
13. Salvesen, H. B. *et al.* Integrated genomic profiling of endometrial carcinoma associates aggressive tumors with indicators of PI3 kinase activation. *Proc. Natl Acad. Sci. USA* **106,** 4834–4839 (2009).
14. Cowin, P. A. *et al.* LRP1B deletion in high-grade serous ovarian cancers is associated with acquired chemotherapy resistance to liposomal doxorubicin. *Cancer Res.* **72,** 4060–4073 (2012).
15. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
16. Govindan, R. *et al.* Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150,** 1121–1134 (2012).
17. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–196 (2010).
18. Palles, C. *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genet.* **45,** 136–144 (2013).
19. Bartosch, C. *et al.* Endometrial carcinomas: a review emphasizing overlapping and distinctive morphological and immunohistochemical features. *Adv. Anat. Pathol.* **18,** 415–437 (2011).
20. Toyota, M. *et al.* CpG island methylator phenotype in colorectal cancer. *Proc. Natl Acad. Sci. USA* **96,** 8681–8686 (1999).
21. Hinoue, T. *et al.* Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res.* **22,** 271–282 (2012).
22. Noushmehr, H. *et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17,** 510–522 (2010).
23. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25,** 2906–2912 (2009).
24. Hockenbery, D., Nunez, G., Milliman, C., Schreiber, R. D. & Korsmeyer, S. J. Bcl-2 is an inner mitochondrial membrane protein that blocks programmed cell death. *Nature* **348,** 334–336 (1990).
25. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22,** 398–406 (2012).
26. Li, J., Mizukami, Y., Zhang, X., Jo, W. S. & Chung, D. C. Oncogenic K-*ras* stimulates Wnt signaling in colon cancer through inhibition of GSK-3β. *Gastroenterology* **128,** 1907–1918 (2005).
27. Zorn, A. M. *et al.* Regulation of Wnt signaling by Sox proteins: XSox17 α/β and XSox3 physically interact with β-catenin. *Mol. Cell* **4,** 487–498 (1999).
28. Sinner, D. *et al.* Sox17 and Sox4 differentially regulate β-catenin/T-cell factor activity and proliferation of colon carcinoma cells. *Mol. Cell. Biol.* **27,** 7802–7815 (2007).
29. Pollock, P. M. *et al.* Frequent activating FGFR2 mutations in endometrial carcinomas parallel germline mutations associated with craniosynostosis and skeletal dysplasia syndromes. *Oncogene* **26,** 7158–7162 (2007).
30. Fleming, G. F. *et al.* Phase II trial of trastuzumab in women with advanced or recurrent, HER2-positive endometrial carcinoma: a Gynecologic Oncology Group study. *Gynecol. Oncol.* **116,** 15–20 (2010).
31. Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26,** i237–i245 (2010).
32. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
33. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–615 (2011).
34. Clarke, B. A. & Gilks, C. B. Endometrial carcinoma: controversies in histopathological assessment of grade and tumour cell type. *J. Clin. Pathol.* **63,** 410–415 (2010).
35. Yemelyanova, A. *et al.* Utility of p16 expression for distinction of uterine serous carcinomas from endometrial endometrioid and endocervical adenocarcinomas: immunohistochemical analysis of 201 cases. *Am. J. Surg. Pathol.* **33,** 1504–1514 (2009).
36. Gilks, C. B., Oliva, E. & Soslow, R. A. Poor inter-observer reproducibility in the diagnosis of high-grade endometrial carcinoma. *Am. J. Surg. Pathol.* **91,** 248A (2012).
37. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12,** R41 (2011).
38. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11,** 367 (2010).
39. Houseman, E. A. *et al.* Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics* **9,** 365 (2008).
40. Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA* **101,** 4164–4169 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements We wish to thank all patients and families who contributed to this study. We thank M. Sheth and L. Lund for administrative coordination of TCGA activities, G. Monemvasitis for editing the manuscript, and C. Gunter for critical reading of the manuscript. This work was supported by the following grants from the US National Institutes of Health: 5U24CA143799-04, 5U24CA143835-04, 5U24CA143840-04, 5U24CA143843-04, 5U24CA143845-04, 5U24CA143848-04, 5U24CA143858-04, 5U24CA143866-04, 5U24CA143867-04, 5U24CA143882-04, 5U24CA143883-04, 5U24CA144025-04, U54HG003067-11, U54HG003079-10 and U54HG003273-10.

Author Contributions The TCGA Research Network contributed collectively to this study. Biospecimens were provided by the tissue source sites and processed by the biospecimen core resource. Data generation and analyses were performed by the genome sequencing centres, cancer genome characterization centres and genome data analysis centres. All data were released through the data coordinating centre. The National Cancer Institute and National Human Genome Research Institute project teams coordinated project activities. We also acknowledge the following TCGA investigators who made substantial contributions to the project: N.S. (manuscript coordinator); J. Gao (data coordinator); C.K. and L. Ding (DNA sequence analysis); W.Z. and Y.L. (mRNA sequence analysis); H.S. and P.W.L. (DNA methylation analysis); A.D.C. and I.P. (copy number analysis); S.L. and A. Hadjipanayis (translocations); N.S., N.W. G.C., C.C.B. and C.Y. (pathway analysis); Andy C. and A.G.R. (miRNA sequence analysis); R. Broaddus, P.J.G., G.B.M. and R.A.S. (pathology and clinical expertise); G.B.M., H.L. and R.A. (reverse phase protein arrays); P.J.G. and R.B. (disease experts); G.B.M. and R.K. (manuscript editing); D.A.L. and E.R.M. (project chairs).

Author Information The primary and processed data used to generate the analyses presented here are deposited at the Data Coordinating Center (https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp); all of the primary sequence files are deposited in CGHub (https://cghub.ucsc.edu/). Sample lists, data matrices and supporting data can be found at: (https://tcga-data.nci.nih.gov/docs/publications/ucec_2013/). The data can be explored via the cBio Cancer Genomics Portal (http://cbioportal.org). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.A.L. (levine2@mskcc.org).

The Cancer Genome Atlas Research Network (Participants are arranged by area of contribution and then by institution.)

Genome sequencing centres: Broad Institute Gad Getz[1], Stacey B. Gabriel[1], Kristian Cibulskis[1], Eric Lander[1], Andrey Sivachenko[1], Carrie Sougnez[1], Mike Lawrence[1]; **Washington University in St Louis** Cyriac Kandoth[2], David Dooling[2], Robert Fulton[2], Lucinda Fulton[2], Joelle Kalicki-Veizer[2], Michael D. McLellan[2], Michelle O'Laughlin[2], Heather Schmidt[2], Richard K. Wilson[2], Kai Ye[2], Li Ding[2], Elaine R. Mardis[2]

Genome characterization centres: British Columbia Cancer Agency Adrian Ally[3], Miruna Balasundaram[3], Inanc Birol[3], Yaron S. N. Butterfield[3], Rebecca Carlsen[3], Candace Carter[3], Andy Chu[3], Eric Chuah[3], Hye-Jung E. Chun[3], Noreen Dhalla[3], Ranabir Guin[3], Carrie Hirst[3], Robert A. Holt[3], Steven J. M. Jones[3], Darlene Lee[3], Haiyan I. Li[3], Marco A. Marra[3], Michael Mayo[3], Richard A. Moore[3], Andrew J. Mungall[3], Patrick Plettner[3], Jacqueline E. Schein[3], Payal Sipahimalani[3], Angela Tam[3], Richard J. Varhol[3], A. Gordon Robertson[3]; **Broad Institute** Andrew D. Cherniack[1], Itai Pashtan[1,4,5], Gordon Saksena[1], Robert C. Onofrio[1], Steven E. Schumacher[1], Barbara Tabak[1], Scott L. Carter[1], Bryan Hernandez[1], Jeff Gentry[1], Helga B. Salvesen[1,6,7], Kristin Ardlie[1], Gad Getz[1], Wendy Winckler[1], Rameen Beroukhim[1,8], Stacey B. Gabriel[1], Matthew Meyerson[1,8]; **Harvard Medical School/Brigham & Women's Hospital/MD Anderson Cancer Center** Angela Hadjipanayis[9], Semin Lee[10], Harshad S. Mahadeshwar[11], Peter Park[10,12], Alexei Protopopov[11], Xiaojia Ren[9], Sahil Seth[11], Xingzhi Song[11], Jiabin Tang[11], Ruibin Xi[10], Lixing Yang[10], Dong Zeng[11], Raju Kucherlapati[9], Lynda Chin[1,11], Jianhua Zhang[11]; **University of North Carolina** J. Todd Auman[13,14], Saianand Balu[15], Tom Bodenheimer[15], Elizabeth Buda[15], D. Neil Hayes[15,16], Alan P. Hoyle[15], Stuart R. Jefferys[15], Corbin D. Jones[17], Shaowu Meng[15], Piotr A. Mieczkowski[18], Lisle E. Mose[15], Joel S. Parker[15], Charles M. Perou[15,18,19], Jeff Roach[20], Yan Shi[15], Janae V. Simons[15], Mathew G. Soloway[15], Donghui Tan[15], Michael D. Topal[15,19], Scot Waring[15], Junyuan Wu[15], Katherine A. Hoadley[15,18]; **University of Southern California & Johns Hopkins** Stephen B. Baylin[21], Moiz S. Bootwalla[22], Phillip H. Lai[22], Timothy J. Triche Jr[22], David J. Van Den Berg[22], Daniel J. Weisenberger[22], Peter W. Laird[22], Hui Shen[22]

Genome data analysis centres: Broad Institute Lynda Chin[1,11], Jianhua Zhang[11], Gad Getz[1], Juok Cho[1], Daniel DiCara[1], Scott Frazer[1], David Heiman[1], Rui Jing[1], Pei Lin[1], Will Mallard[1], Petar Stojanov[1], Doug Voet[1], Hailei Zhang[1], Lihua Zou[1], Michael Noble[1], Mike

Lawrence[1]; **Institute for Systems Biology** Sheila M. Reynolds[23], Ilya Shmulevich[23]; **Memorial Sloan-Kettering Cancer Center** B. Arman Aksoy[24], Yevgeniy Antipin[24], Giovanni Ciriello[24], Gideon Dresdner[24], Jianjiong Gao[24], Benjamin Gross[24], Anders Jacobsen[24], Marc Ladanyi[25], Boris Reva[24], Chris Sander[24], Rileen Sinha[24], S. Onur Sumer[24], Barry S. Taylor[26], Ethan Cerami[24], Nils Weinhold[24], Nikolaus Schultz[24], Ronglai Shen[27]; **University of California, Santa Cruz/Buck Institute** Stephen Benz[28], Ted Goldstein[28], David Haussler[28], Sam Ng[28], Christopher Szeto[28], Joshua Stuart[28], Christopher C. Benz[29], Christina Yau[29]; **The University of Texas MD Anderson Cancer Center** Wei Zhang[30,31], Matti Annala[30,31,32], Bradley M. Broom[33], Tod D. Casasent[33], Zhenlin Ju[33], Han Liang[33], Guoyan Liu[30,31], Yiling Lu[34], Anna K. Unruh[33], Chris Wakefield[33], John N. Weinstein[33], Nianxiang Zhang[33], Yuexin Liu[30,31], Russell Broaddus[31], Rehan Akbani[33], Gordon B. Mills[34]

Biospecimen core resource: Nationwide Children's Hospital Christopher Adams[35], Thomas Barr[35], Aaron D. Black[35], Jay Bowen[35], John Deardurff[35], Jessica Frick[35], Julie M. Gastier-Foster[35,36], Thomas Grossman[35], Hollie A. Harper[35], Melissa Hart-Kothari[35], Carmen Helsel[35], Aaron Hobensack[35], Harkness Kuck[35], Kelley Kneile[35], Kristen M. Leraas[35], Tara M. Lichtenberg[35], Cynthia McAllister[35], Robert E. Pyatt[35], Nilsa C. Ramirez[35,36], Teresa R. Tabler[35], Nathan Vanhoose[35], Peter White[35], Lisa Wise[35], Erik Zmuda[35]

Tissue source sites: Asterand Nandita Barnabas[37], Charlenia Berry-Green[37], Victoria Blanc[37], Lori Boice[38], Michael Button[37], Adam Farkas[37], Alex Green[37], Jean MacKenzie[37], Dana Nicholson[37]; **British Columbia Cancer Agency** Steve E. Kalloger[39,40], C. Blake Gilks[39,40]; **Cedars-Sinai Medical Center** Beth Y. Karlan[41], Jenny Lester[41], Sandra Orsulic[41]; **Christiana Care** Mark Borowsky[42], Mark Cadungog[42], Christine Czerwinski[42], Lori Huelsenbeck-Dill[42], Mary Iacocca[42], Nicholas Petrelli[42], Brenda Rabeno[42], Gary Witkin[42]; **Cureline** Elena Nemirovich-Danchenko[43], Olga Potapova[43], Daniil Rotin[43]; **Duke University** Andrew Berchuck[44]; **Gynecologic Oncology Group** Michael Birrer[45], Phillip DiSaia[46], Laura Monovich[47]; **International Genomics Consortium** Erin Curley[48], Johanna Gardner[48], David Mallery[48], Robert Penny[48]; **Mayo Clinic** Sean C. Dowdy[49], Boris Winterhoff[49], Linda Dao[50], Bobbie Gostout[49], Alexandra Meuter[49], Attila Teoman[49]; **Memorial Sloan-Kettering Cancer Center** Fanny Dao[51], Narciso Olvera[51], Faina Bogomolniy[51], Karuna Garg[52], Robert A. Soslow[52], Douglas A. Levine[51]; **N. N. Blokhin Russian Cancer Research Center** Mikhail Abramov[53]; **Ontario Tumour Bank** John M. S. Bartlett[54], Sugy Kodeeswaran[54], Jeremy Parfitt[55]; **St Petersburg Academic University** Fedor Moiseenko[56]; **University Health Network** Blaise A. Clarke[57]; **University of Hawaii** Marc T. Goodman[58,59], Michael E. Carney[58], Rayna K. Matsuno[58]; **University of North Carolina** Jennifer Fisher[38], Mei Huang[38], W. Kimryn Rathmell[15], Leigh Thorne[38], Linda Van Le[38]; **University of Pittsburgh** Rajiv Dhir[60], Robert Edwards[60], Esther Elishaev[60], Kristin Zorn[60]; **The University of Texas MD Anderson Cancer Center** Russell Broaddus[31]; **Washington University School of Medicine** Paul J. Goodfellow[36,61], David Mutch[61]

Disease analysis working group: Nikolaus Schultz[24], Yuexin Liu[30,31], Rehan Akbani[33], Andrew D. Cherniack[1], Ethan Cerami[24], Nils Weinhold[24], Hui Shen[22], Katherine A. Hoadley[15,18], Ari B. Kahn[62], Daphne W. Bell[63], Pamela M. Pollock[64], Chen Wang[65], David A.Wheeler[66], Eve Shinbrot[66], Beth Y. Karlan[41], Andrew Berchuck[44], Sean C. Dowdy[49], Boris Winterhoff[49], Marc T. Goodman[58,59], A. Gordon Robertson[3], Rameen Beroukhim[1,8], Itai Pashtan[1,4,5], Helga B. Salvesen[1,6,7], Peter W. Laird[22], Michael Noble[1], Joshua Stuart[28], Li Ding[2], Cyriac Kandoth[2], C. Blake Gilks[39,40], Robert A. Soslow[52], Paul J. Goodfellow[36,61], David Mutch[61], Russell Broaddus[31], Wei Zhang[30,31], Gordon B. Mills[34], Raju Kucherlapati[9], Elaine R. Mardis[2], Douglas A. Levine[51]

Data coordination centre: Brenda Ayala[62], Anna L. Chu[62], Mark A. Jensen[62], Prachi Kothiyal[62], Todd D. Pihl[62], Joan Pontius[62], David A. Pot[62], Eric E. Snyder[62], Deepak Srinivasan[62], Ari B. Kahn[62]

Project team: National Cancer Institute Kenna R. Mills Shaw[67], Margi Sheth[67], Tanja Davidsen[67], Greg Eley[68]; Martin L. Ferguson[69], John A. Demchok[67], Liming Yang[67]; **National Human Genome Research Institute** Mark S. Guyer[70], Bradley A. Ozenberger[70], Heidi J. Sofia[70]

Writing committee: Cyriac Kandoth[2], Nikolaus Schultz[24], Andrew D. Cherniack[1], Rehan Akbani[33], Yuexin Liu[30,31], Hui Shen[22], A. Gordon Robertson[3], Itai Pashtan[1,4,5], Ronglai Shen[27], Christopher C. Benz[29], Christina Yau[29], Peter W. Laird[22], Li Ding[2], Wei Zhang[30,31], Gordon B. Mills[34], Raju Kucherlapati[9], Elaine R. Mardis[2] & Douglas A. Levine[51]

[1]The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University Cambridge, Massachusetts 02142, USA. [2]The Genome Institute, Washington University, St Louis, Missouri 63108, USA. [3]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia V5Z, Canada. [4]Department of Radiation Oncology, Dana-Farber Cancer Institute and Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [5]Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. [6]Department of Obstetrics and Gynecology, Haukeland University Hospital, 5021 Bergen, Norway. [7]Department of Clinical Medicine, University of Bergen, 5020 Bergen, Norway. [8]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. [9]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [10]Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115, USA. [11]Institute for Applied Cancer Science, Department of Genomic Medicine, University of Texas MD Anderson Cancer Center, Houston, Texas 77054, USA. [12]Informatics Program, Boston Children's Hospital, Boston, Massachusetts 02115, USA. [13]Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [14]Institute for Pharmacogenetics and Individualized Therapy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [15]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [16]Department of Internal Medicine, Division of Medical Oncology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [17]Department of Biology, University of North Carolina at Chapel Hill, North Carolina 27599, USA. [18]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [19]Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [20]Research Computing Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [21]Cancer Biology Division, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University, Baltimore, Maryland 21231, USA. [22]University of Southern California Epigenome Center, University of Southern California, Los Angeles, California 90089, USA. [23]Institute for Systems Biology, Seattle, Washington 98109, USA. [24]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [25]Human Oncology and Pathogenesis Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [26]Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, California 94158, USA. [27]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [28]Department of Biomolecular Engineering and Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA. [29]Buck Institute for Age Research, Novato, California 94945, USA. [30]Cancer Genomics Core Laboratory, University of Texas MD Anderson Cancer Center, Houston, Texas 77054, USA. [31]Department of Pathology, University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [32]Tampere University of Technology Korkeakoulunkatu 10, FI-33720 Tampere, Finland. [33]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [34]Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [35]The Research Institute at Nationwide Children's Hospital, Columbus, Ohio 43205, USA. [36]The Ohio State University, Columbus, Ohio 43210, USA. [37]Asterand, Detroit, Michigan 48202, USA. [38]University of North Carolina, Chapel Hill, North Carolina 27599, USA. [39]OvCaRe British Columbia, British Columbia Cancer Agency, Vancouver, British Columbia V5Z 4E6, Canada. [40]Department of Pathology & Laboratory Medicine, The University of British Columbia, Vancouver, British Columbia V6T 2B5, Canada. [41]Women's Cancer Program at the Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, USA. [42]Helen F Graham Cancer Center at Christiana Care, Newark, Delaware 19713, USA. [43]Cureline, Inc., South San Francisco, California 94080, USA. [44]Duke University Medical Center, Duke Cancer Institute, Durham, North Carolina 27710, USA. [45]Harvard Medical School, Massachusetts General Hospital Cancer Center, Boston, Massachusetts 02114, USA. [46]University of California Medical Center, Irvine, Orange California 92868, USA. [47]GOG Tissue Bank, The Research Institute at Nationwide Children's Hospital, Columbus, Ohio 43205, USA. [48]International Genomics Consortium, Phoenix, Arizona 85004, USA. [49]Department of OB Gyn, Division of Gynecologic Oncology, Mayo Clinic, Rochester, Minnesota 55905, USA. [50]Department of Pathology, Mayo Clinic, Rochester, Minnesota 55905, USA. [51]Gynecology Service, Department of Surgery, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [52]Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [53]N. N. Blokhin Russian Cancer Research Center RAMS, Moscow 115478, Russia. [54]Ontario Tumour Bank, Ontario Institute for Cancer Research, Toronto, Ontario M5G 0A3, Canada. [55]Ontario Tumour Bank, London Health Sciences Centre, London, Ontario N6A 5A5, Canada. [56]St Petersburg Academic University, St Petersburg 199034, Russia. [57]Department of Pathology, University Health Network, Toronto, Ontario M5G 2C4, Canada. [58]University of Hawaii, Honolulu, Hawaii 96813, USA. [59]Cedars-Sinai Medical Center, Los Angeles, California 90024, USA. [60]University of Pittsburgh, Pittsburgh, Pennsylvania 15213, USA. [61]Washington University School of Medicine, St Louis, Missouri 63110, USA. [62]SRA International, Fairfax, Virgina 22033, USA. [63]Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [64]Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane 4059, Australia. [65]Department of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota 55905, USA. [66]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA. [67]The Cancer Genome Atlas Program Office, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [68]Scimentis, LLC, Atlanta, Georgia 30666, USA. [69]MLF Consulting, Arlington, Maryland 02474, USA. [70]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.

ERRATUM

doi:10.1038/nature12325

Erratum: Integrated genomic characterization of endometrial carcinoma

The Cancer Genome Atlas Research Network

Nature **497,** 67–73 (2013); doi:10.1038/nature12113

In the 'Results' section of this Article, the range in the sentence "The median follow-up of the cohort was 32 months (range, 1–19 months); 21% of the patients have recurred, and 11% have died." should have been 1–195 months. This error has been corrected in the HTML and PDF versions of the paper.

A variational autoencoder trained with priors from canonical pathways increases the interpretability of transcriptome data

Bin Liu[1], Bodo Rosenhahn[2], Thomas Illig[1,3], David S. DeLuca[1*]

1 Hannover Medical School, Biomedical Research in Endstage and Obstructive Lung Disease Hannover (BREATH), German Center for Lung Research, Hannover, Lower Saxony, Germany
2 Institut für Informationsverarbeitung (TNT), Leibniz University Hannover, Hannover, Lower Saxony, Germany,
3 Hannover Unified Biobank, Hannover Medical School, Hannover, Lower Saxony, Germany

* DeLuca.David@mh-hannover.de

Abstract

Interpreting transcriptome data is an important yet challenging aspect of bioinformatic analysis. While gene set enrichment analysis is a standard tool for interpreting regulatory changes, we utilize deep learning techniques, specifically autoencoder architectures, to learn latent variables that drive transcriptome signals. We investigate whether simple, variational autoencoder (VAE), and beta-weighted VAE are capable of learning reduced representations of transcriptomes that retain critical biological information. We propose a novel VAE which utilizes priors from biological data to direct the network to learn a representation of the transcriptome that is based on understandable biological concepts.

After training five different autoencoder architectures on 22310 transcriptomes, we benchmarked their performance on organ and disease classification tasks on separate selection of 5577 test samples. Every tested architecture succeeded in reducing the transcriptomes to 50 latent dimensions, which captured enough variation for accurate reconstruction. The simple, fully connected autoencoder, performs best across the benchmarks, but lacks the characteristic of having directly interpretable latent dimensions. The beta-weighted, prior-informed VAE implementation is able to solve the benchmarking tasks, and provide semantically accurate latent features equating to biological pathways.

This study opens a new direction for differential pathway analysis in transcriptomics with increased transparency and interpretability.

Author summary

The ability to measure the human transcriptome has been a critical tool to studying health and disease. However, transcriptomes data sets are too large and complex for direct human interpretation. Deep learning techniques such as autoencoders are capable of distilling high-level features from complex data. However, even if deep learning models find patterns, these patterns are not necessarily represented in a way that humans can easily understand. By bringing in the prior knowledge of biological pathways, we have trained the model to "speak the language" of the biologist, and

represent complex transcrtomes, in simpler concepts that are already familiar to biologists. We can then apply the tool to compare for example samples from lung cancer cells to healthy cells, and show which biological processes are perturbed.

Introduction

Transcriptomics is a powerful tool in characterizing cellular activity under various conditions, which allows researchers to discover the underlying associations between transcripts or genes and pathological or environmental factors. Therefore, transcriptomics data are widely applicable in multiple areas of biomedical research, varying from understanding disease mechanisms [1], detecting biomarkers [2,3], to tissue-specific regulatory gene identification [4]. The broad application of the technology leads to generating a considerable amount of transcriptomic sequencing data and constructing a few specific public platforms hosting the relevant biological data sets, such as ArrayExpress [5] and NCBI GEO [6]. In all these biomedical tasks, human-understandable interpretation of the experimental transcriptomics data serves as the pivotal component to understanding the underlying biology. However, this interpretation remains a challenge in the face of large and complex data sets. Here we explore the potential for machine learning models to learn a simplified representation of the transcriptome in terms of commonly understood biological processes, and thus increase the interpretability.

While gene set enrichment analysis (GSEA) [7] is a standard tool for interpreting regulatory changes to the transcriptome, the method highly relies on a list of well-detected differentially expressed genes (DEGs) between the conditions of interest. Furthermore, most of the state-of-the-art models for differential expression analysis (DEA) are based on the linear assumption across samples, the representatives of which include the models from limma [8,9], DESeq2 [10], and Seurat [11]. However, variation in measured expression levels, whether of biological or technical origin, may not always behave linearly. Given the potential for synergistic effects between genes, this assumption of linearity might lead to a loss of power. Complex sources of variation are also present in combined public data sets, consisting of a large number of samples from multiple sources, and influenced by non-biological factors such as batches or processing centers. These limitations daunt the exploration of large-scale transcriptomics data for answering fundamental biological questions.

The goal of this study is to utilize deep learning techniques to bring transparency into which biological process patterns are represented in a transcriptome data set, and thus to facilitate the interpretation of experimental results. Deep learning with artificial neural networks has witnessed rapid development in recent years and outperformed the traditional approaches in handling massive and complex data from multiple areas because of its high-level feature extraction at a non-linear space and objective data processing [12–16]. This development also enables more possibilities in understanding the transcriptomics data and has successfully contributed, for example to drug repurposing and development [17,18], phenotype classification [19] and genomics functional characterization [20] with a more in-depth understanding of transcriptomics data.

We are specifically interested in autoencoders as a class of methods which can reduce the dimensionality and learn the major features in complex data [21]. Autoencoders achieve this by passing the data through a bottleneck layer (a layer with fewer nodes than in input layer) and optimizing the model with the objective of generating an output that is as similar as possible to the input. These methods have been explored in the context of transcriptomes in a few representative publications, including [20,22–26], demonstrating that interesting biological features are captured in the latent space of the

model. However, these studies take different approaches to the challenge of associating latent representations with human-understandable biological concepts. [22] implemented an autoencoder and interpreted that latent space by correlating latent features with phenotypes post hoc. [20] on the other hand sought to constrain the latent space to represent known pathways by restricting the network connectivity, i.e., each latent node represents a pathway, and only genes known to be involved in that pathway are connected to the node upstream. In the study of [22], the network is free to learn any representation from that data, but the burden of interpretation is left to post hoc analysis. In the case of [20], the network is restricted directly in its architecture based on gene set definitions.

Here, we see an opportunity to implement a solution that finds a middle path in which the network is encouraged to learn a latent representation based on known biological concepts, but still has the freedom to learn relationships among genes from the data. Specifically, we propose an autoencoder variant using a novel technique of introducing pathway-informed priors. The basis for this approach is the Bayesian framework implemented in variational autoencoders (VAE) [27]. The VAE framework inherently provides the opportunity to involve priors in the training process to learn latent representations. With the introduction of biologically meaningful priors, our approach here aims to integrate prior knowledge from the domain (here, we use Hallmark pathways defined by MSigDB as an example) and still retain the flexibility of the data-driven deep learning approach. This approach is most comparable to those presented by Zhao [25] and Lotfollahi [26], with the commonality that the goal is to produce latent features that correspond to known biological concepts. However, compared to our approach of incorporating canonical pathway knowledge as priors, both Zhao [25] and Lotfollahi [26] provide pathway definitions to constrain the decoder architecture. The prior-based approach provides an additional opportunity to calibrate the strength of the effect of prior following the established beta-VAE approach ([28]).

Specifically, We make use of a hyperparameter, beta, in a similar way described by [28], which can add weight to the influence of the priors on the training solution. Thus, using the beta, we can control the extent to which the model conforms to pathway concepts previously defined by MSigDB versus being free to define latent variables in any way that best encodes the transcriptome in a reduced representation. The fine-tuning of this hyperparameter enables control over the tension between direct biological interpretability and the ability to deviate from canon or find new patterns.

In this study, we implement and compare several standard autoencoder implementations: (i) fully connected autoencoder (simpleAE) [29], (ii) variational autoencoder (simpleVAE), (iii) beta-VAE (beta-simpleVAE), as well as (iv) novel derivatives using prior biological data: priorVAE, and (v) beta-priorVAE. In order to benchmark the performance of the series of models proposed here, we perform tissue and disease classification (e.g., adenocarcinoma, small cell lung cancer (SCLC)) based on the latent variables discovered in the models. This study explores the feasibility of the prior-based VAE approach in increasing the transparency and interpretability of transcriptomes, as well as how the hyperparameter beta controls the balance of using prior information versus learning novel patterns directly from the data.

Materials and methods

Data sets and preprocessing

The data set employed for the training of the model and subsequent analyses was downloaded from ArrayExpress [5], with Accession ID E-MTAB-3732 [30]. This data set comprises 27,887 Affymetrix HG-U133Plus2 arrays, sourced publicly. All samples

underwent quality control filtering and were annotated for disease status and cell line information. The data was normalized using fRMA [31–34] within the R Bioconductor platform by the data set's author. The data set contains samples from healthy individuals, those with diseases (including cancer), and cell lines. The original data set has been divided into a training and a test set at an 8:2 ratio, with stratification based on the source organs of the samples. The curated gene signature data set, used in the previous generation, is the Hallmark gene set, sourced from Human MSigDB Collections [35].

The transcriptome data were prepared as input into the autoencoders in three variations. In the first input variation, the data were fed into the model on the transcriptome level without further processing. For the second input variation, the transcripts were collapsed into the gene level using the platform annotation offered by Gene Expression Omnibus (GEO) [6]. The normalized expression levels of the transcripts were transformed into the original level by applying a power function. The original expressions were averaged, and a log2 transformation was processed on the mean values.

For the third input variation, the goal was to decrease the number of trainable parameters in the models further. A community detection algorithm was applied to the gene expression level. An unsupervised nearest neighbors learning was applied to the absolute level of correlation between each pair of genes using the NearestNeighbors function from the sklearn Python package [36]. A graph of k-Neighbors was computed to make a graph of gene relationships based on the expression levels. We then use the Leiden algorithm [37] to detect the communities in this graph. A total of 2032 communities were defined using a resolution value of 0.02. A single gene was chosen as a representative of each community for use as the final input. Community representatives were based on the criteria of having the highest sum of correlation with all the other genes in the same community.

Model architectures

The experiment systematically trained and compared five architectures of autoencoders, including one fully-connected autoencoder without prior information (simpleAE) and four variational autoencoders (VAE) architectures. Besides the variational autoencoder with unit Gaussian prior (simpleVAE) and the beta-constrained variational autoencoder with unit Gaussian prior (beta-simpleVAE), we also presented a novel technique of introducing pathway-informed priors (priorVAE) and tested the influence of the hyperparameter beta over this biological relevant prior VAE (beta-priorVAE).

We construct three fully-connected linear layers in the encoder of all the autoencoder architectures (number of trainable parameters equal to the dimension of features, 1000 and 100, respectively). The leaky rectified linear unit (Leaky ReLU) serves as the activation function between encoder layers.

The bottleneck layer of the simpleAE is a dense layer with 50 dimensions. For the VAEs, the bottleneck is implemented as two fully-connected, 50-dimension layers: one for learning σs and one for learning μs, which are then sampled using the reparametrization trick before going on to the decoder. The decoders in all models mirror the three dense layers of the encoders. The Softplus activation function is added to the last layer of the decoder for reconstruction to output values on the same scale as gene expression input values.

The loss function for the simpleAE is the mean squared error (MSE) between the decoder output and the original input values. For VAE, the loss function has two terms: (i) the reconstruction loss, also MSE, and (ii) the KL divergence between the latent distributions and unit Gaussian prior distributions. For the prior- and beta-priorVAE implementations, the loss functions are described in detail below.

The preparation of pathway-informed priors

Pathway-informed prior distributions were generated for each sample in the form of (σ^2) and (μ) parameters using a bootstrapping procedure. We defined μ as the average gene expression level of genes within each pathway definition, and σ as the variance across bootstrapping iterations. Pathway definitions were taken from MSigDB Hallmarks [35].

Loss function for biological priors

Generally, variational autoencoders take the form found in Fig 1 and have been previously described in detail [27]. In short, it has been established that neural network training can perform variational inference when the loss function takes the form:

$$= \mathbb{E}_{q_\phi(z|x)}[logp_\theta(x|z)] - KL(q_\phi(z|x)||p(z)) = L_{recon} + L_{KL} = L_{VAEarchitectures}$$

Here, $x \in \mathbb{R}$ is a vector of expression values for a sample in the set of all samples, X. The generative model $p_\theta(x|z)$ is learned, where z ($z \in \mathbb{R}$) are latent variables with a prior $p(z)$ such that z can generate the observed data x.

Fig 1. A conceptual illustration of variations autoencoder architectures. A: An overview of the system architecture with different input levels. A pathway-derived prior is generated on the transcript or gene level, as described in the following sections. The training and post hoc are repeated with input on the transcript, gene, and community levels. B: As a latent variable framework, the model assumes that latent variables (Z) are determinants of the measured data (X). To learn Z, $p(Z|X)$ is approximated by $q(Z|X)$, and modeled as the encoder portion of the network. The latent values represent probability distributions, which are implemented in the bottle-neck layer as values for mu (μ) and sigma (σ). Finally, the decoder is conceptually equivalent to $p(X|Z)$.

The reconstruction loss term L_{recon} can be measured by the mean squared error (MSE) between the input and the reconstructed output. In most VAE implementations, a Gaussian distribution is used for $p(z)$ to make a tractable KL calculation. The assumption of unit Gaussian, $\mathcal{N}(0, I)$, as priors leads to the simplified expression:

$$KL(q_\phi(z|x), p(z)) = \frac{1}{2}\left[-\left(\log\sigma^2 + 1\right) + \sigma^2 + \mu^2\right]$$

In order to implement pathway-informed priors with any parameter values, the KL divergence had to be implemented more generically for any two Gaussian distributions to measure the distance between $q_\phi(z|x)$ and $p(z)$, where $q_\phi(z|x) \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $p(z)$ $\sim \mathcal{N}(\mu_2, \sigma_2^2)$:

$$\begin{aligned} & KL(q_\phi(z|x), p(z)) \\ = & -\int p(z)\log q_\phi(z|x)dx + \int p(z)\log p(x)dx \\ = & \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \end{aligned}$$

The models beta-simpleVAE and beta-priorVAE involve the addition of hyperparameter β ($\beta > 1$) to put more weight on the L_{KL} term, as described in [28]:

$$\begin{aligned} & L_{beta-VAEarchitectures} \\ = & \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta KL(q_\phi(z|x)||p(z)) \\ = & L_{recon} + \beta L_{KL} \end{aligned}$$

Results

Input layer

The Affymetrix HG-U133Plus2 chip captures expression at the transcript level, providing an input space of roughly 50,000 features. Concerned that this would result in a parameter space that was too large for the available data, we experimented with collapsing the inputs to the gene level and even further to 'representative genes' using a network-based clustering technique. A comparison of the choice of inputs can be found in S1 Fig. We concluded that gene-level input was a sufficient reduction in the parameter space based on the results. The results reported below are based on this gene-level input.

Learning latent representations with several autoencoder architectures

Five autoencoder variations were trained on the full set of transcriptomes provided by the ArrayExpression dataset, E-MTAB-3732, containing 27887 samples. As shown in Table 1, common to each architecture is a 50-dimensional latent space. For the models, simpleAE, simpleVAE, and beta-simpleVAE the 50 latent dimensions were learned strictly from the data, without attributing any prior biological concepts, and are simply enumerated 1 through 50. For the priorVAE and beta-priorVAE, the 50 latent nodes are associated with the 50 pathways found in the MSigDB Hallmarks gene sets, and accordingly, each latent node can be labeled with the gene set name. To evaluate the training results, we use three types of benchmarking strategies, as seen in Table 2. (i) an evaluation of reconstruction performance, (ii) the Kullback–Leibler (KL) divergence between the latent distributions and the prior distributions, and (iii) the combined total loss.

Table 1. An overview of the structures of the five architectures.

Model	Latent Nodes	Beta	Prior	Loss Function
SimpleAE	50 values, unlabeled	none	none	MSE
SimpleVAE	50 Gaussians, unlabeled	none	unit Gaussian	$MSE + \frac{1}{2}\left[-\left(\log\sigma^2+1\right)+\sigma^2+\mu^2\right]$
Beta-SimpleVAE	50 Gaussians, unlabeled	beta = 250	unit Gaussian	$MSE + \beta * (\frac{1}{2}\left[-\left(\log\sigma^2+1\right)+\sigma^2+\mu^2\right])$
PriorVAE	50 Gaussians, pathway-derived	none	pathway names	$MSE + (\log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2})$
Beta-PriorVAE	50 Gaussians, pathway-derived	beta = 250	pathway names	$MSE + \beta * (\log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2})$

Table 2. Benchmark results for the five architectures on the test set.

Model	Recon. Loss	KL(w.β)	KL(n.β)	Total Loss (w.β)
SimpleAE	2674.6	NA	NA	NA
SimpleVAE	2780.8	184.5	184.5	2965.3
Beta-SimpleVAE	5899.7	7.4	1846.8	7746.5
PriorVAE	2758.5	160.3	160.3	2918.8
Beta-PriorVAE	5740.2	8.1	2012.9	7753.1

Loss function values broken down into reconstruction loss, KL divergences with β (w.β), without β (n.β)

Taking these metrics together, it is clear that the introduction of hyperparameter β results in a lower KL divergence at the cost of a slightly worse reconstruction loss. The introduction of priors resulted in little change to the loss terms compared to their non-prior counterparts.

To further explore reconstruction performance, we computed correlation coefficients between input transcriptomes vs. the output transcriptomes. A complete pairwise correlation analysis across samples shows how similar input and output transcriptomes are to each other in the context of the natural variability across samples (Fig 2). Each model was able to reproduce reasonable output transcriptomes, which correlated with R between ***0.97 and 0.8***. In most cases, the closest pairwise correlations were between the input and output models, although this was not the case for many samples in the beta-simpleVAE, and a single sample in the beta-priorVAE. Together with the performance shown in Fig 2, it is clear that increasing the beta hyperparameter and emphasizing the influence of the priors comes at a cost to reconstruction performance.

Fig 2. Reconstruction performances using correlation coefficients between input and output transcriptomes. A-E: The clustered pair-wise correlation heatmaps of the selected input and their reconstructed output for A: simpleAE, B: simpleVAE, C: priorVAE, D: beta-simpleVAE, E: and beta-priorVAE. Selected input samples and their corresponding reconstruction output are enumerated as 1-20. T represents the input train sample and R represents the reconstructed output. F: The average correlation between the input and its corresponding reconstruction output.

Ideally, the latent dimension should contain enough information to capture the essential biological features. To systematically evaluate how well the models can perform in this regard, we have used the latent features as input into classification models for three validation tasks: (i) distinguishing tissue types (ii) distinguishing healthy vs. adenocarcinoma, and (iii) distinguishing adenocarcinoma from small cell lung cancer (SCLC). The classification performance after five-fold cross-validation is reported in Fig 3, 4, and 5.

Fig 3. The results and performance of the AE models on the tissue classification. A-E: The t-distributed stochastic neighbor embedding (t-sne) representations of the latent values (μ) representation of samples from different tissue origins for A: simpleAE, B: simpleVAE, C: priorVAE, D: beta-simpleVAE, and E: beta-priorVAE. F: The average precision score of the multivariate logistic regression models sourcing from the latent representation of the models across the eight selected tissue types.

Fig 4. The results and performance of the AE models on the classification of healthy vs. adenocarcinoma samples. A-E: The biplots for the top two principle components (PCs) distinguishing the adenocarcinoma samples and the healthy lung samples for A: simpleAE, B: simpleVAE, C: priorVAE, D: beta-simpleVAE, and E: beta-priorVAE. F: The average precision score of the multivariate logistic regression models sourcing from the latent representation of the models on the task of healthy vs. adenocarcinoma classification.

For the benchmarking task of classifying tissue types, eight target organs were selected based on available samples. In this classification task, the simpleAE, simpleVAE and priorVAE perform with an average precision above 80%. Beta-weighting leads to worse classification performance in both beta-simpleVAE and beta-priorVAE. However, beta-priorVAE performs much better in this validation task than the former (roughly 0.76 vs. less than 0.6).

For the task of distinguishing adenocarcinoma from healthy controls, all five architectures perform comparably well, with precision between 98.6 and 100 percent. Four architectures, (all but beta-simpleVAE) can still distinguish the adenocarcinoma

Fig 5. The results of the prior variants (priorVAE and beta-priorVAE) on the classification of lung cancer subtypes (adenocarcinoma vs. small cell lung cancer). A and B: Biplots and principle components analysis (PCA) results of priorVAE and beta-priorVAE. C and D: Heatmaps of the latent results based on selected disease samples for priorVAE and beta-priorVAE. Data is scaled across the fifty latent dimensions. E and F: Two-sided t-test results of the pathway dimensions with the two disease statuses. The significance levels (the p-values) are transformed in the -log10 form and shown on the y-axis. The size of the points represents the correlation between the prior and the latent μ, with a larger size indicating a higher level of correlation. The dimensions with a negative correlation are colored in red. G: The average precision score of the multivariate logistic regression models sourcing from the latent representation of the models on the task of adenocarcinoma vs. SCLC classification, showing that each model type is well suited for this classification task.

cell line, a subpopulation of the non-small cell lung cancer (NSCLC), from the small cell lung cancer (SCLC) cell lines with 100 percent of average precision. The performance of beta-simpleVAE latent results as a classifier slightly decreased from 98.6% to 97.5%, and the standard deviation (sd) across the five-fold validation increased from 0.016 to 0.049.

Influence of priors on learned latent representation

The goal of bringing pathway-derived priors into the model is to produce latent representations that both accurately represent the full transcriptome and also directly correspond to recognizable biological concepts. However, these goals are at odds, as we reported in (Fig 6 and S2 Fig). In the case of the priorVAE, the latent variable *allograft rejection* does retain a high correlation with the prior scores (R = 0.82). However, all the other 49 pathways are with a correlation weaker than 0.5, and only six pathways are to some extent correlated with the prior, even with a far less stringent threshold of 0.4. 21 out of 50 pathways correlate negatively with the prior scores. In contrast, the latent variables of the beta-priorVAE retain their connection to their pathways, with 48 of 50 pathways correlating higher than 0.6. Fig 6C indicates that the beta-priorVAE provides latent values with a high level of semantic meaning, providing a direct means for interpreting complex transcriptome data sets.

Fig 6. The semantic meaningfulness of the latent variables in the prior-based models, shown as the correlation between the biological priors and the latent μ of prior-based models on the test set. The correlation of each dimension is shown in (A) for priorVAE and (B) for beta-priorVAE. Subplot (C) summarizes these correlations to directly compare the semantic interpretability of the two models.

The beta-priorVAE provides a simplified representation of the transcriptome in terms of 50 features corresponding to 50 pathways. A comparative analysis across samples and conditions is now possible directly on the bases of these features. For the priorVAE model, the biplots of adenocarcinoma and healthy samples indicate that *allograft rejection* and *xenobiotic metabolism* are major features associated with disease (Fig 4). However, for beta-priorVAE *interferon alpha response*, and *MYC targets v2* are the main distinguishing features. In a direct comparison of adenocarcinoma and SCLC, the biplots show that *angiogenesis* and *spermatogenesis* are the two pathways with the highest load to the first principle component (PC) in priorVAE, and *MYC targets v2* and *spermatogenesis* in the case of beta-priorVAE.

A more direct comparison can be made by performing a two-tailed t-test on the values from the two lung cancer phenotype conditions. The results indicate that the top

three most differentially expressed dimensions in priorVAE are: *glycolysis*, *angiogenesis*, and *apoptosis*; for beta-priorVAE, the top three are *coagulation*, *spermatogenesis*, and *angiogenesis*.

These differences between priorVAE and beta-priorVAE are explained to a large extent by the fact that the beta-priorVAE produces latent variables which track more closely with the pathway-based priors (Fig 6). A direct interpretation of the involvement of a given pathway is only possible when that latent feature is highly correlated with the prior distributions. In the case of adeno vs. small cell, the most distinguishing features, *glycolysis* and *angiogenesis* have only weak correlations with their priors: R = 0.043 and -0.075, respectively. In contrast, for the beta-priorVAE, the most statistically significant feature distinguishing adeno from the small cells is *coagulation*, which is a feature that is also highly correlated with its prior (R=0.96). Therefore, it is only for the beta-priorVAE model that the labels on the latent features retain the meaning of the original pathways. Even for this model, however, it should be noted that the second most statistically significant differentiator is reported as *spermatogenesis*, which is in fact the latent value that has the lowest correlation out of all latent variables for this model (R=0.095). Therefore, it must be concluded that the model detected a feature that is a critical source of variation distinguishing adeno from small cell samples but that this feature is not represented in the set of 50 MSigDB Hallmark pathways.

The comparison between the priorVAE and beta-priorVAE show clearly a trade-off between capturing the biological variability in the models' latent space, but meanwhile, adhering to prior biological concepts found in the set of Hallmark pathways. To further investigate the effect of hyperparameter beta on performance, we ran the benchmarks across a range of values for beta (S3 Fig) The classification performance seems to decrease consistently with an increasing beta, although for beta values up to 100, the trend is close to flat. This implies that we can find a beta value that balances the need to capture biological features and, at the same time, adhere to the pathway labels provided via the priors.

DISCUSSION

The results of these experiments demonstrate that autoencoders are capable of generating a simplified representation of a transcriptome that still retains the key biological information necessary to differentiate different cells under different conditions. Furthermore, it is also possible to constrain the training process in a way that forces the network to find a latent representation corresponding to human-understandable biological concepts. Here we have achieved this by taking advantage of the VAE framework, which allows for integrating prior knowledge. There is a trade-off between efficiently representing the complexity of a transcriptome, and adhering to a panel of chosen biological concepts, in our case, defined by 50 Hallmark pathways.

The primary goal of utilizing pathway-based priors in the priorVAE and beta-priorVAE models was to generate a latent space that would be immediately interpretable to a biologist because the model will describe a transcriptome in terms of features that are familiar to a biologist. However, we have observed that latent features do not always retain the identity of their associated priors. In the case of the priorVAE (i.e. beta = 1), the model has substantial freedom to deviate from the pathway priors, and in fact, does so for many features. However, by boosting the requirement to adhere to the pathway concepts with the beta hyperparameter, the immediate biological interpretation of the latent space is achievable (Fig 6C). A notable exception is the case of *spermatogenesis*, which was the feature that had the lowest correlation with the prior. This indicates that the model does require some freedom to discover major sources of variation beyond those pre-defined by the 50 MSigDB Hallmarks. This could indicate a

limitation inherent to the 50 Hallmarks, or it could be a limitation of the idea of relying only on known pathways as a source of variation across transcriptomes. There could also be technical sources of variation that need to be accounted for that would not be represented in a pathway database. The presence of unanticipated sources of variation indicates that an interesting future direction for this research would be to include a few "wild-card" nodes in the latent space, with unit-Gaussian priors that are not driven by pathway data. This modification would allow the modeling to account for "unexpected" sources of variation while at the same time utilizing prior pathway information when possible. The burden of interpreting wild-card nodes would be placed on additional *post hoc* analysis.

The beta-priorVAE with the current setting shows an overall satisfying correlation between the latent variables and prior scores, which enables an interpretation of the biological pathways involved in the chosen vignette. Based on the t-test results (Fig 5), *coagulation* is the most significant differentiator between the two lung cancer phenotypes (adenocarcinoma as the representative NSCLC and the SCLC). While coagulation function is associated with the prognosis in NSCLC patients as described in [39, 40], both [41] and [42] reported an absence of similar correlation between coagulation and the SCLC prognosis. The literature is, therefore, consistent with the latent representation in that the concept of coagulation is a differentiating feature between the two diseases. Thus, the evaluation supports the feasibility of such architecture in making transcriptomes more intuitively transparent and interpretable.

Although our primary motivation for including priors in our VAE was to make the latent space directly interpretable, the traditional motivation for including priors was to increase model accuracy. For most of our benchmarks, model accuracy generally decreased when we increased the emphasis on priors via the beta parameter. This implies that the reconstruction portion of the loss function is the main driver of performance in these benchmarks in comparison to the KL divergence term. However, an interesting point of comparison in our experiments is between the beta-simpleVAE and beta-priorVAE, which have the same emphasis on priors (i.e., the same betas), but in the latter model, prior biological knowledge is incorporated. Table 2 and S3 Fig show an increase in performance when using a biological prior vs. a unit Gaussian prior, indicating that in this local comparison, prior biological information can be beneficial.

We have provided evidence that the key sources of biological variation are captured in the latent space. At the primary level, the successful reconstruction as demonstrated by the reconstruction loss, as well as high correlation coefficients between inputs and outputs, indicate that the latent representations are reliable. This is further supported by the fact that cancer types and tissue types can be distinguished using only the latent features. However, the performance for classifying tissues is surprisingly mediocre. The question is whether this is a limitation of the information found in the latent representation or an inadequate classification procedure. The latter scenario is supported by the high reconstruction accuracy and the fact that the multivariate logistic regression maybe be inadequate without proper feature selection.

Conclusion

The training and post hoc experiments demonstrate that (i) autoencoder models can find simplified representations of transcriptomes that still retain biological information, (ii) using pathway-derived priors, we can encourage the models to find latent representations that still adhere to concepts that are familiar to biologists, and (iii) latent features can provide a direct means of comparison among samples and conditions that can provide an immediate biological interpretation. This area of research should be explored further, with attention to alternate pathway definitions to define the priors

and thus the latent space, additional model architectures, and integration into bioinformatic workflows.

Supporting information

S1 Fig. The performance of reconstruction correlation and the biplots for healthy vs. adenocarcinoma classification on the level of community-, gene- and transcript-level input.

S2 Fig. The scatter plot of the priors (x-axis) and the latent μ (y-axis) of all test samples for A: priorVAE and B: beta-priorVAE.

S3 Fig. The average precision score of the beta-simpleVAE and the beta-priorVAE models on the tissue classification with different values for hyperparameter beta.

Data availability

The source code for this paper is available on GitHub with the following link: https://github.com/BinLiu9205/deepRNA_autoencoder.git and on figshare with the following link: https://figshare.com/articles/software/deepRNA_autoencoder/22227217

References

1. Kotula-Balak, M., Duliban, M., Gurgul, A., Krakowska, I., Grzmil, P., Bilinska, B., and Wolski, J. K. (2021) Transcriptome analysis of human Leydig cell tumours reveals potential mechanisms underlying its development. *Andrologia,* **53**(11), e14222.

2. Kim, S. H., Kim, J. H., Lee, S. J., Jung, M. S., Jeong, D. H., and Lee, K. H. (2022) Minimally invasive skin sampling and transcriptome analysis using microneedles for skin type biomarker research. *Skin Research and Technology,* **28**(2), 322–335.

3. Dubois, J., Rueger, J., Haubold, B., Far, R. K.-K., and Sczakiel, G. (2021) Transcriptome analyses of urine RNA reveal tumor markers for human bladder cancer: Validated amplicons for RT-qPCR-based detection. *Oncotarget,* **12**(10), 1011.

4. Consortium, G. (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science,* **369**(6509), 1318–1330.

5. Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., et al. (2007) ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic acids research,* **35**(suppl_1), D747–D750.

6. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (11, 2012) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research,* **41**(D1), D991–D995.

7. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences,* **102**(43), 15545–15550.

8. Smyth, G. K. (2005) Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor* pp. 397–420 Springer.

9. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research,* **43**(7), e47–e47.

10. Love, M. I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology,* **15**(12), 1–21.

11. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015) Spatial reconstruction of single-cell gene expression data. *Nature biotechnology,* **33**(5), 495–502.

12. Zeleznik, R., Foldyna, B., Eslami, P., Weiss, J., Alexander, I., Taron, J., Parmar, C., Alvi, R. M., Banerji, D., Uno, M., et al. (2021) Deep convolutional neural networks to predict cardiovascular risk from computed tomography. *Nature communications,* **12**(1), 1–9.

13. Yao, D., Zhi-li, Z., Xiao-feng, Z., Wei, C., Fang, H., Yao-ming, C., and Cai, W.-W. (2022) Deep hybrid: multi-graph neural network collaboration for hyperspectral image classification. *Defence Technology,*.

14. Gaur, L., Bhatia, U., Jhanjhi, N., Muhammad, G., and Masud, M. (2021) Medical image-based detection of COVID-19 using deep convolution neural networks. *Multimedia systems,* pp. 1–10.

15. Miles, C., Bohrdt, A., Wu, R., Chiu, C., Xu, M., Ji, G., Greiner, M., Weinberger, K. Q., Demler, E., and Kim, E.-A. (2021) Correlator convolutional neural networks as an interpretable architecture for image-like quantum matter data. *Nature Communications,* **12**(1), 1–7.

16. Sharma, P. K., Bisht, I., and Sur, A. (2021) Wavelength-based attributed deep neural network for underwater image restoration. *ACM Journal of the ACM (JACM),*.

17. Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., and Zhavoronkov, A. (2016) Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular pharmaceutics,* **13**(7), 2524–2530.

18. Luo, Q., Mo, S., Xue, Y., Zhang, X., Gu, Y., Wu, L., Zhang, J., Sun, L., Liu, M., and Hu, Y. (2021) Novel deep learning-based transcriptome data analysis for drug-drug interaction prediction with an application in diabetes. *BMC bioinformatics,* **22**(1), 1–15.

19. Hong, J., Hachem, L. D., and Fehlings, M. G. (2022) A deep learning model to classify neoplastic state and tissue origin from transcriptomic data. *Scientific reports,* **12**(1), 1–7.

20. Chen, H.-I. H., Chiu, Y.-C., Zhang, T., Zhang, S., Huang, Y., and Chen, Y. (2018) GSAE: an autoencoder with embedded gene-set nodes for genomics functional characterization. *BMC systems biology,* **12**(8), 45–57.

21. Liou, C.-Y., Cheng, W.-C., Liou, J.-W., and Liou, D.-R. (2014) Autoencoder for words. *Neurocomputing,* **139**, 84–96.

22. Way, G. P. and Greene, C. S. (2018) Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium* World Scientific pp. 80–91.

23. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (December, 2018) Deep generative modeling for single-cell transcriptomics. *Nature Methods,* **15**(12), 1053–1058 Number: 12 Publisher: Nature Publishing Group.

24. Ding, J., Condon, A., and Shah, S. P. (May, 2018) Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature Communications,* **9**(1), 2002 Number: 1 Publisher: Nature Publishing Group.

25. Zhao, Y., Cai, H., Zhang, Z., Tang, J., and Li, Y. (September, 2021) Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nature Communications,* **12**(1), 5261 Number: 1 Publisher: Nature Publishing Group.

26. Lotfollahi, M., Rybakov, S., Hrovatin, K., Hediyeh-zadeh, S., Talavera-López, C., Misharin, A. V., and Theis, F. J. (February, 2023) Biologically informed deep learning to query gene programs in single-cell atlases. *Nature Cell Biology,* **25**(2), 337–350 Number: 2 Publisher: Nature Publishing Group.

27. Doersch, C. (2016) Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908,*.

28. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2016) beta-vae: Learning basic visual concepts with a constrained variational framework.

29. Rumelhart, D. E., Hinton, G. E., and Williams, R. J., Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science (1985).

30. Torrente, A. A comprehensive human expression map.

31. McCall, M. N., Bolstad, B. M., and Irizarry, R. A. (2010) Frozen robust multiarray analysis (fRMA). *Biostatistics,* **11**(2), 242–253.

32. McCall, M. N., Uppal, K., Jaffee, H. A., Zilliox, M. J., and Irizarry, R. A. (2011) The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic acids research,* **39**(suppl_1), D1011–D1015.

33. Margus, L., Wolfgang, H., et al. (2011) Assessing affymetrix GeneChip microarray quality. BMC.

34. McCall, M. N., Jaffee, H. A., and Irizarry, R. A. (2012) fRMA ST: frozen robust multiarray analysis for Affymetrix Exon and Gene ST arrays. *Bioinformatics,* **28**(23), 3153–3154.

35. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015) The molecular signatures database hallmark gene set collection. *Cell systems,* **1**(6), 417–425.

36. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research,* **12**, 2825–2830.

37. Traag, V. A., Waltman, L., and Van Eck, N. J. (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports,* **9**(1), 5233.

38. Asperti, A. and Trentin, M. (2020) Balancing reconstruction error and Kullback-Leibler divergence in Variational Autoencoders. *IEEE Access,* **8**, 199440–199448.

39. Qi, Y. and Fu, J. (2017) Research on the coagulation function changes in non small cell lung cancer patients and analysis of their correlation with metastasis and survival. *J buon,* **22**(2), 462–467.

40. Sotiropoulos, G. P., Dalamaga, M., Antonakos, G., Marinou, I., Vogiatzakis, E., Kotopouli, M., Karampela, I., Christodoulatos, G. S., Lekka, A., and Papavassiliou, A. G. (2018) Chemerin as a biomarker at the intersection of inflammation, chemotaxis, coagulation, fibrinolysis and metabolism in resectable non-small cell lung cancer. *Lung Cancer,* **125**, 291–299.

41. Gabazza, E. C., Taguchi, O., Yamakami, T., Machishi, M., Ibata, H., and Suzuki, S. (1993) Correlation between increased granulocyte elastase release and activation of blood coagulation in patients with lung cancer. *Cancer,* **72**(7), 2134–2140.

42. Gezelius, E., Flou Kristensen, A., Bendahl, P., Hisada, Y., Risom Kristensen, S., Ek, L., Bergman, B., Wallberg, M., Falkmer, U., Mackman, N., et al. (2018) Coagulation biomarkers and prediction of venous thromboembolism and survival in small cell lung cancer: A sub-study of RASTEN-A randomized trial with low molecular weight heparin. *PLoS One,* **13**(11), e0207387.

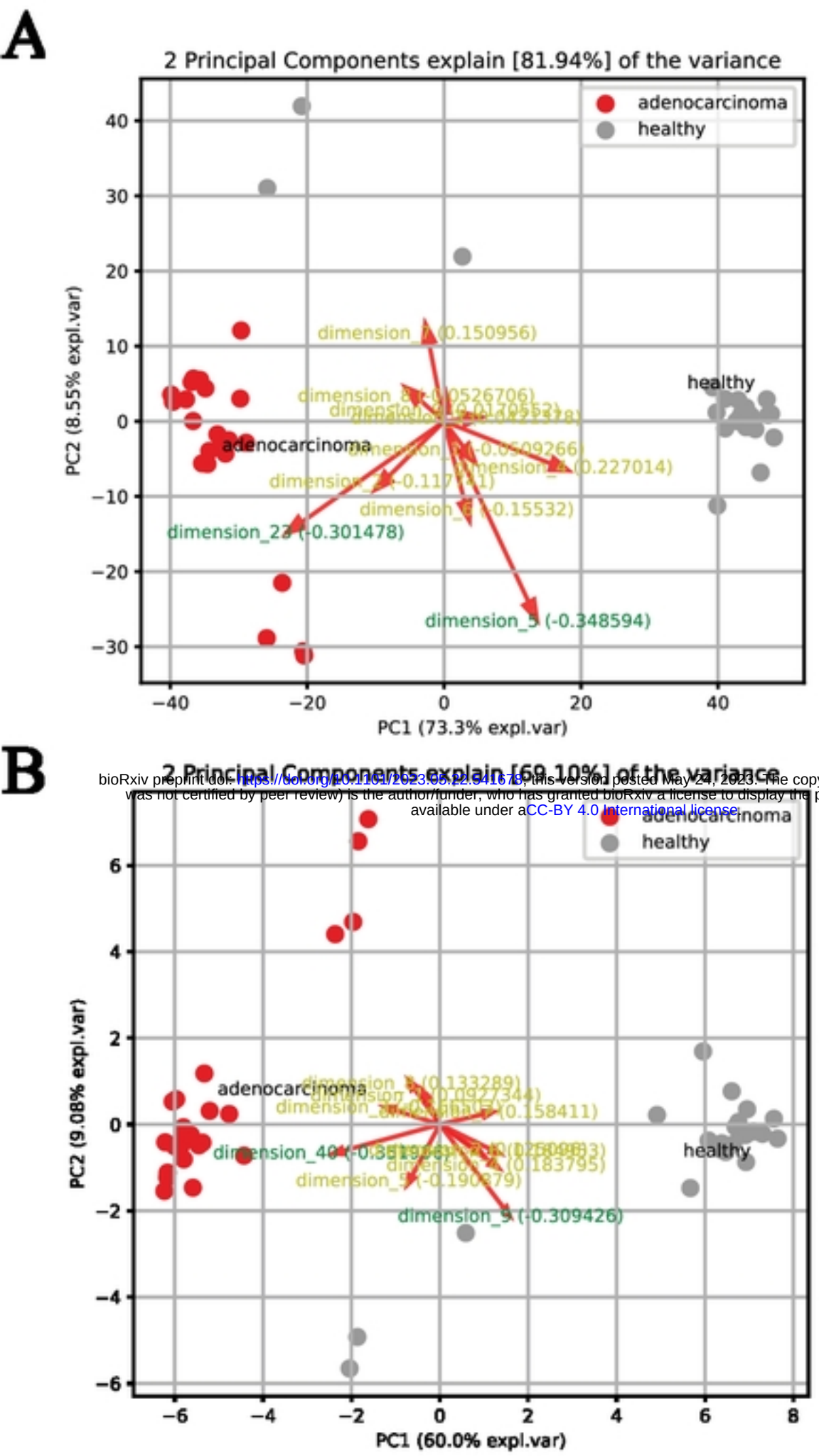


Figure 4

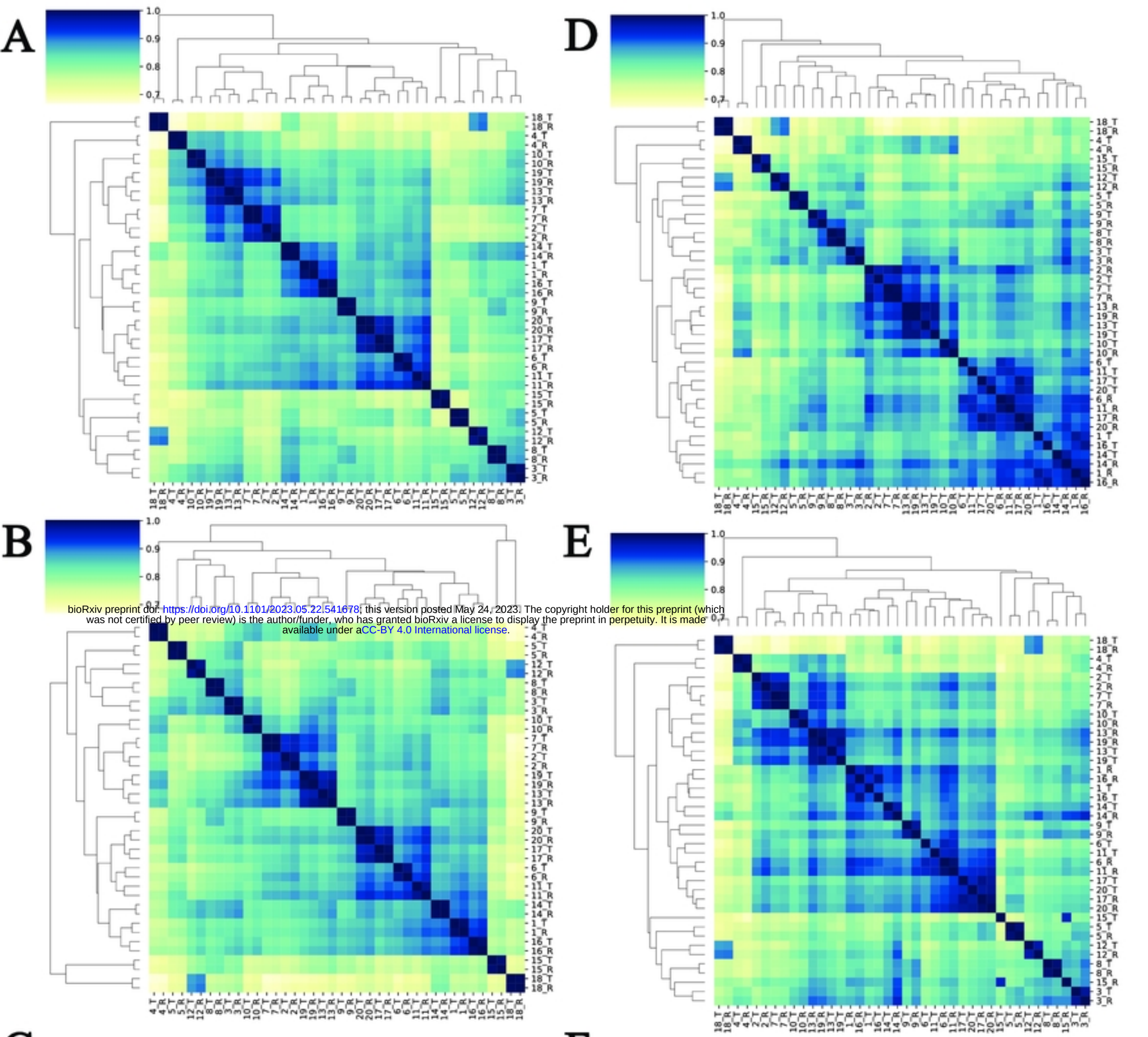
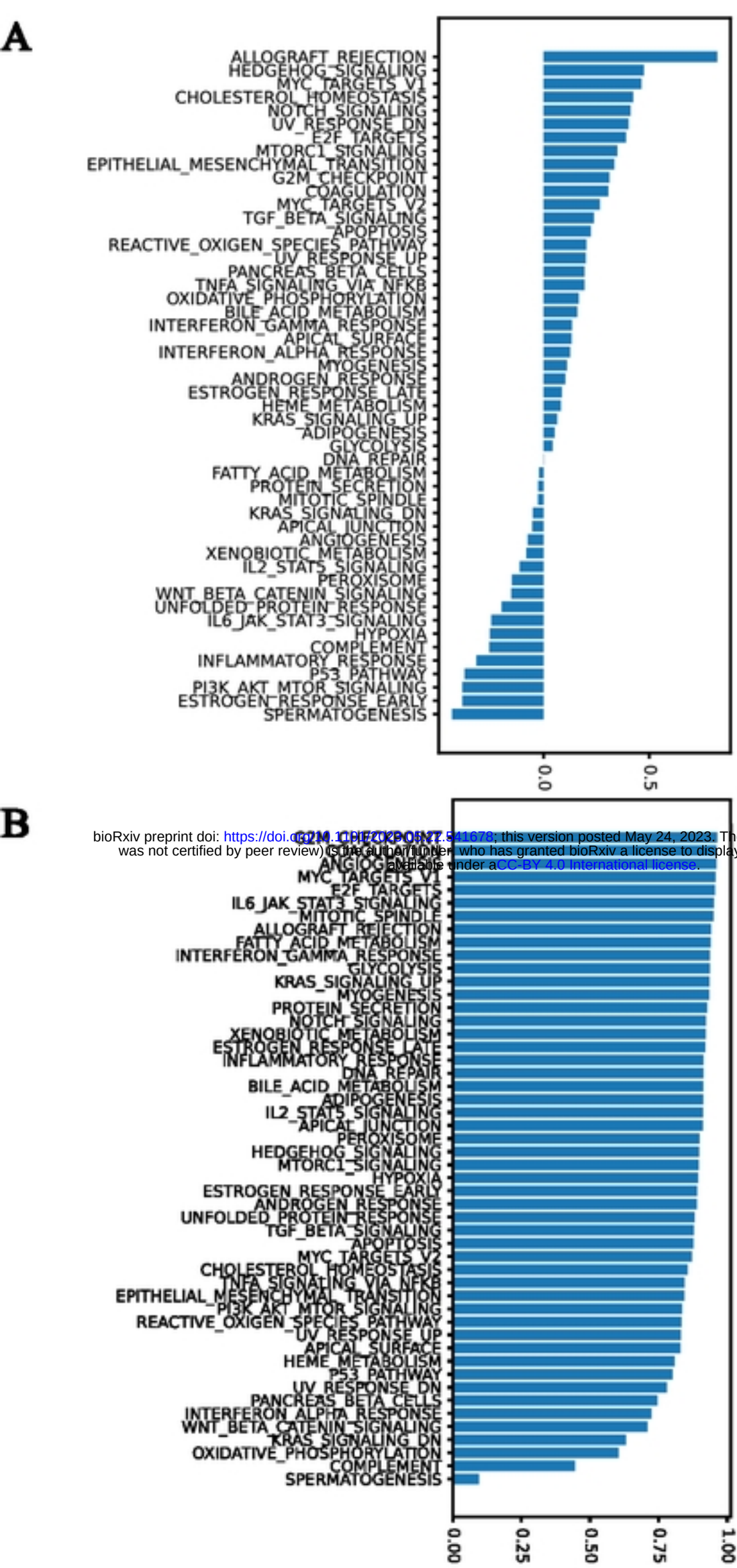


Figure 2



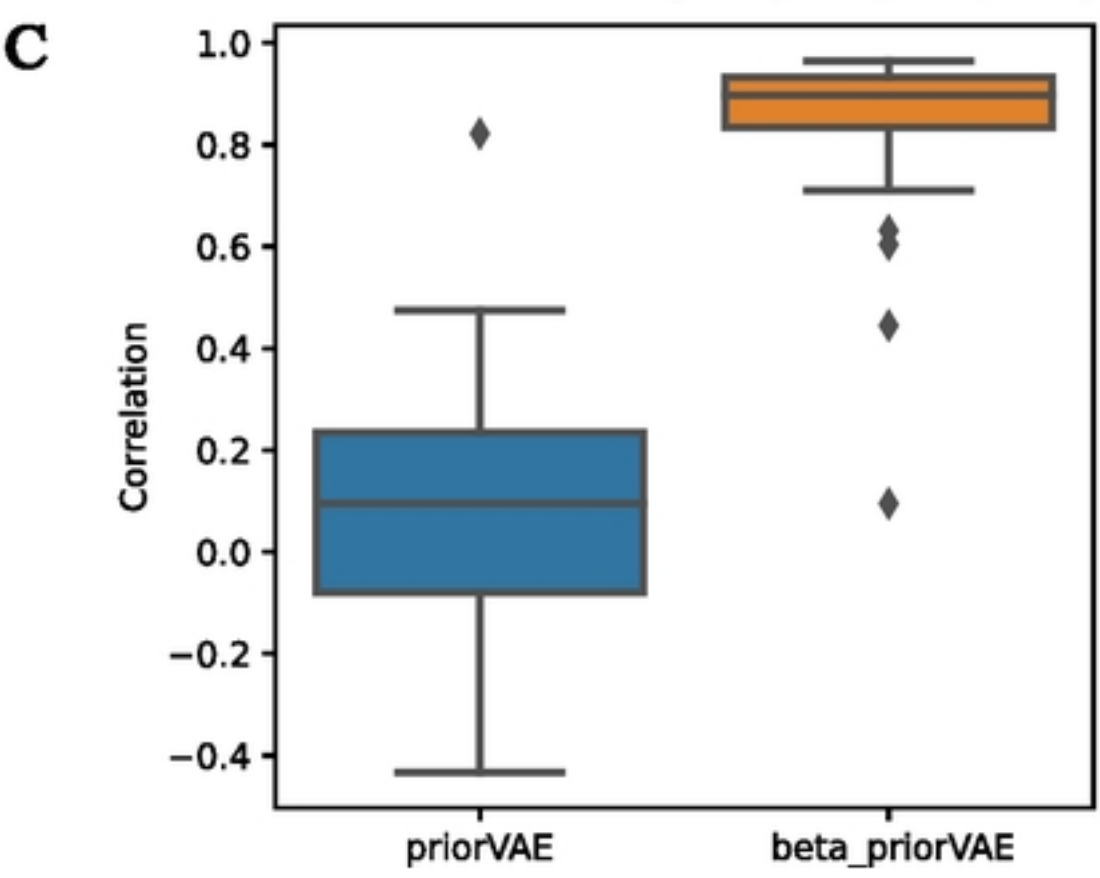


Figure 6

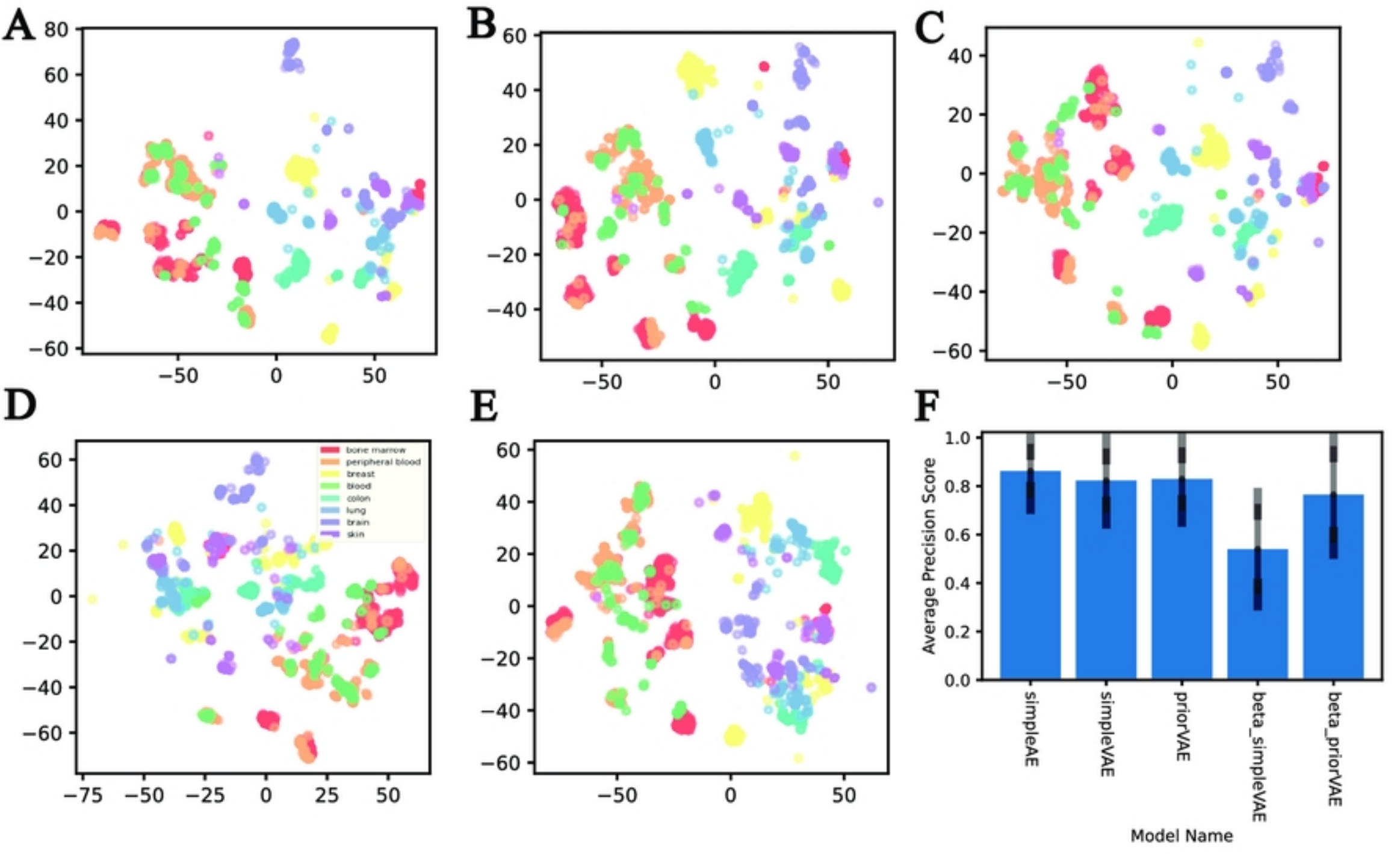

A
B
C
D
E
F
bone marrow
peripheral blood
breast
blood
colon
lung
brain
skin
Average Precision Score
simpleAE
simpleVAE
priorVAE
beta_simpleVAE
beta_priorVAE
Model Name

Figure 3

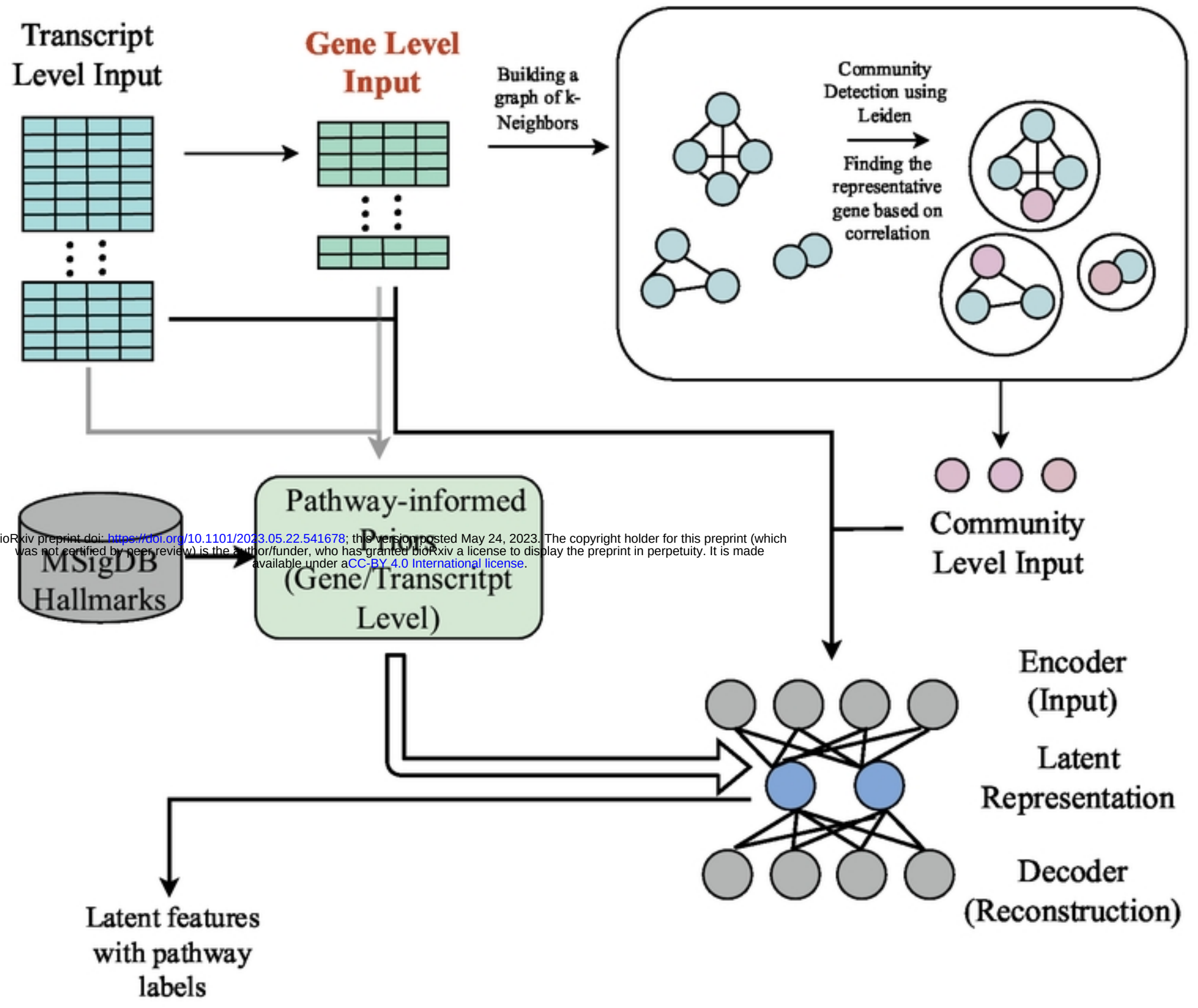
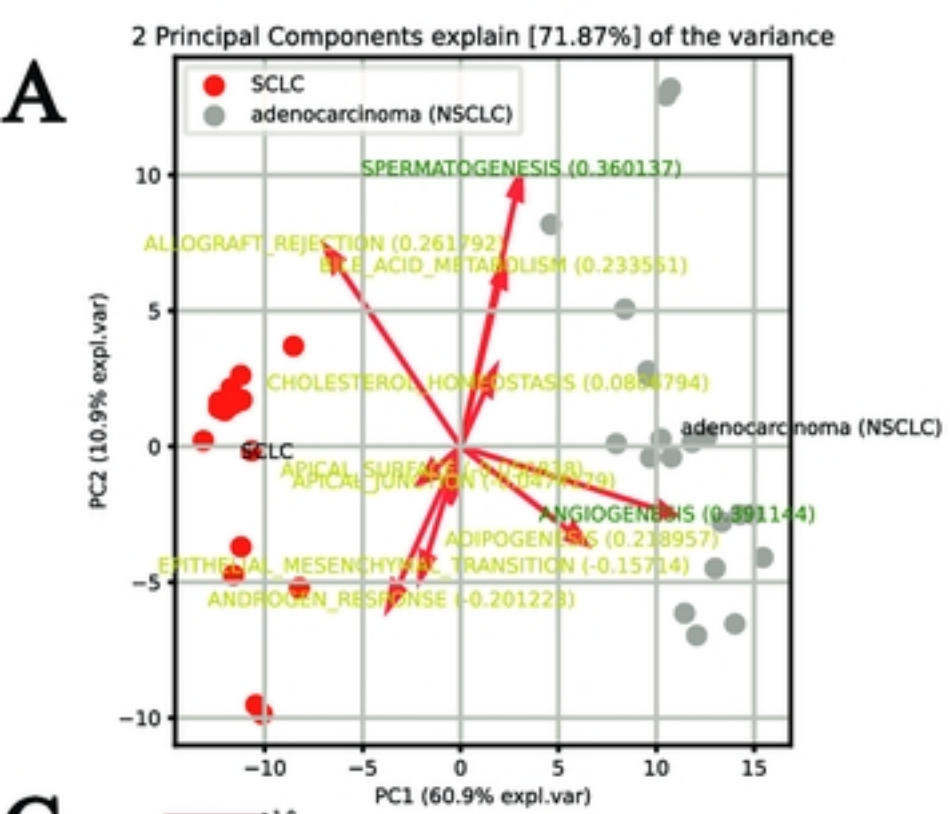


Figure 1



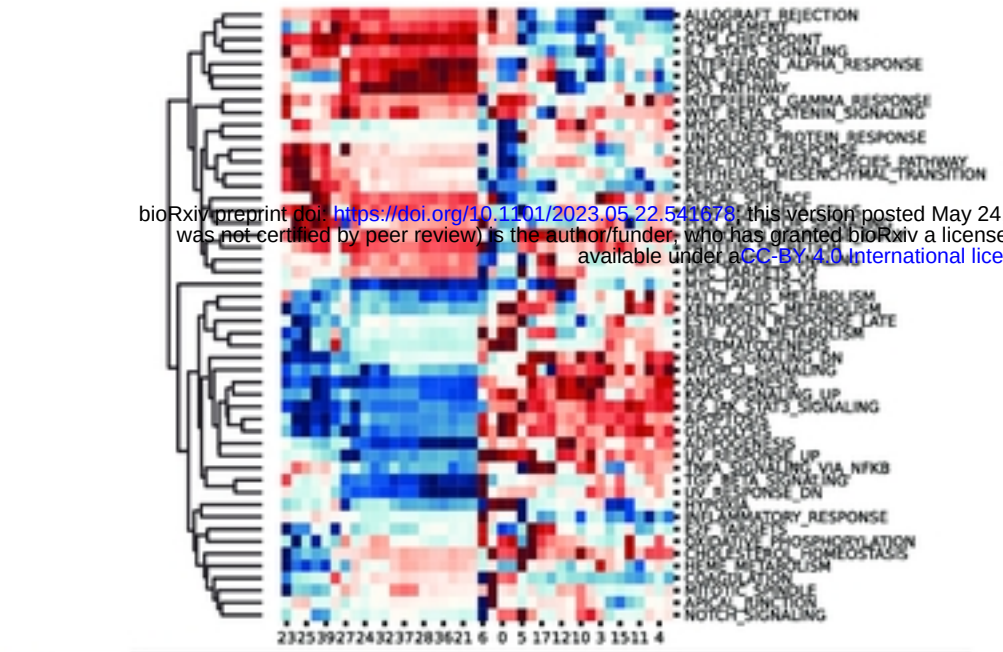


Figure 5