Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Masterthesis

Tumor DNA methylation profiles of high-risk endometrioid endometrial carcinoma and patient recurrence

Milan Geybels

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Bioinformatics

SUPERVISOR :

Prof. dr. Ziv SHKEDY

Prof. dr. Dirk VALKENBORG

2017
2018

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Masterthesis

Tumor DNA methylation profiles of high-risk endometrioid endometrial carcinoma and patient recurrence

Milan Geybels

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Bioinformatics

SUPERVISOR :

Prof. dr. Ziv SHKEDY

Prof. dr. Dirk VALKENBORG

PREFACE

I would like to thank the teaching staff of the Master of Statistics program at Hasselt University. Many of the courses in the Master program have directly contributed to this thesis project. In particular, many thanks to my thesis advisors, Prof. dr. Ziv Shkedy and Prof. dr. Dirk Valkenborg, for their valuable input and critical questions, which improved my analyses and helped me to better understand the statistics and bioinformatics methodology.

CONTENTS

ABSTRACT

BACKGROUND: Endometrioid endometrial carcinoma (EEC) is the most common subtype of endometrial cancer. Routine clinical parameters such as tumor stage and grade provide important prognostic information for EEC patients, but combining these clinical variables with tumor biomarkers such as epigenomic alterations may significantly improve patient risk stratification for cancer recurrence. Accurately predicting the prognosis of EEC patients is important to help guide clinical decision-making for post-operative disease management strategies.

METHODS: Publically-available molecular and clinical data from The Cancer Genome Atlas (TCGA) were used. The data set included primary tumor samples from 312 surgically-treated EEC patients. Data on tumor DNA methylation levels (β-value; range: 0 – 1) at 22,443 CpG sites across the epigenome were analyzed. Based on tumor stage and grade, patients were classified as low-risk (stage I and grade 1; n = 50) or high-risk EEC (stage IV or grade 3; n = 162); and all other patients were in the intermediate-risk group (n = 100). Different (un)-supervised statistical techniques were used to identify methylation phenotypes, study differential DNA methylation at single CpGs, and generate a prognostic methylation signature.

RESULTS: Unsupervised clustering identified four distinct methylation phenotypes including a subset of tumors associated with lower methylation levels (23 percent of tumors), which was enriched for high-risk tumors (P-value < 0.0001). In total, 1,503 CpG sites were differentially methylated between high and low-risk EEC (FDR q-value < 0.01 and mean methylation β-value difference ≥ 0.05), including 15 top-ranked CpG sites with a mean methylation difference between subgroups of at least 0.2. These top-ranked CpGs were in different genes: *ARSE*, *FCRL3*, *HIST1H2BB*, *HIST1H3C*, *HLA-DOB*, *ITGB7*, *KRTAP11-1*, *REG3A*, *RNASE3*, *SLC25A35*, *TMED6*, and *TMEM101*. Using LASSO feature selection, a 56-CpG methylation signature of high versus low-risk EEC was generated (mean five-fold cross-validation misclassification error rate = 0.114). In the subset of intermediate-risk patients, this signature classified patients who developed recurrence (n = 11) versus those who remained recurrence-free (n = 84; AUC = 0.85, 95% CI: 0.71,

0.99). Further, combining the signature with routine clinical parameters for predicting EEC recurrence significantly improved the classification performance compared to the clinical model only (AUC = 0.88 vs. 0.70; likelihood-ratio test, P-value = 0.0001). In the intermediate-risk group, higher levels of the signature correlated with increased expression of genes in known pathways of cell proliferation (E2F targets, G2M checkpoint, MYC targets). Finally, in a second smaller validation set from TCGA (n = 99), the signature classified high versus low-risk EEC (misclassification error rate = 0.159), and recurrent (n = 4) versus non-recurrent disease (n = 76; AUC = 0.80; 95% CI: 0.62, 0.97).

CONCLUSION: Using data on tumor stage and grade, EEC patients were classified into low, intermediate, and high-risk groups for having a poor prognosis. A tumor DNA methylation signature for distinguishing high from low-risk EEC was generated. This study showed that, in the subset of intermediate-risk patients, the signature predicted the risk of cancer recurrence. The methylation signature, therefore, has potential as a prognostic classifier for these patients.

INTRODUCTION

Endometrial cancer is a tumor originating in the endometrium, the inner membrane of the uterus [1]. It is the most common gynecological tumor in developed countries, with an estimated 61,380 new cases and 10,920 deaths in the United States in 2017 [2]. The most common subtype of endometrial cancer is endometrioid endometrial cancer (EEC), which accounts for about 80 percent of all cases [1].

EEC tumors are classified based on tumor stage (I, II, III, IV) and grade (1, 2, 3) [1, 3]. Tumor stage provides information on the extent of the tumor. While stage I tumors are confined to the uterus, stage IV tumors have grown outside the uterus (e.g., bladder) or metastasized to distant sites in the body (e.g., lungs) [4]. Most EEC patients are diagnosed with localized tumors (stage I/II), and the standard treatment for these patients is surgery (hysterectomy) [1]. Information on tumor grade is obtained by histological examination by a pathologist. Grade 1 EEC cells are well-differentiated and are most similar to normal endometrial cells. At the other end of the extreme are grade 3 tumors, which are poorly differentiated and, therefore, aggressive.

Tumor stage and grade are important prognostic variables. However, these clinical parameters do not accurately classify all individual patients [5]. For example, EEC tumors that have the same stage and grade may behave very differently and have a different prognosis. Accurately predicting the prognosis of individual patients is important to help guide clinical decision making, and identify the best post-operative disease management strategy (e.g., adjuvant radiation, chemotherapy, no additional therapy) [1, 6].

Tumor biomarkers hold potential to improve EEC patient risk stratification [5]. Tumor biomarkers include transcriptomic changes, protein expression, genomic mutations, and epigenomic alterations. Prognostic tumor biomarkers are expected to be particularly important for EEC patients with intermediate stage/grade tumors (e.g., stage 1/2 and grade 2) as the prognosis of these patients often is unclear [5, 7].

DNA methylation is the most widely studied epigenomic mechanism [8, 9]. DNA methylation involves the addition of a methyl-group to a CG dinucleotide or CpG site. DNA methylation is related to gene transcription, as it is one mechanism to control gene expression levels. For example, higher methylation levels in a gene promoter region can suppress transcription of that gene. In cancer, including EEC, DNA methylation changes are widespread [10].

Aims and outline of the thesis

In this thesis project, publically-available clinical and molecular data from The Cancer Genome Atlas (TCGA) were used [3]. The main aim of the project was to build a tumor methylation signature for predicting EEC prognosis; specifically, for patients who have intermediate stage/grade tumors.

The outline of the thesis was as follows. First, EEC patients were classified into risk categories on the basis of tumor stage and grade. Low-risk EEC was defined as stage I and grade 1. High-risk disease was defined as stage IV or grade 3. All other patients were in the intermediate-risk category. Second, epigenome-wide DNA methylation profiles were compared between high and low-risk EEC, and a DNA methylation signature of high versus low-risk EEC was generated. Third, the methylation signature was applied in the remaining intermediate-risk patients, and used to further risk stratify these patients. Further, the DNA methylation data were also integrated with whole-genome gene expression data of the same patients' tumors to perform a gene set analysis and study correlations between CpG methylation and gene expression levels.

METHODS

Data set

The study included data from endometrioid endometrial carcinoma (EEC) patients included in The Cancer Genome Atlas (TCGA-UCEC). As described previously, TCGA primary tumor specimens were collected from newly diagnosed EEC patients who received surgical resection, and had no prior treatment for their disease [3]. All primary tumor samples were from surgical specimens and no biopsy specimens were used. Samples were restricted to those that contained at least 60% tumor nuclei by pathological review.

TCGA clinical and molecular (i.e., DNA methylation and gene expression) data were downloaded from the UCSC Xena browser (http://xena.ucsc.edu/). In total, the data set included 411 EEC patients who had a tumor sample for molecular profiling and information on pathological stage and grade. The molecular data used in this thesis project were already pre-processed and normalized. Therefore, the data also did not contain methylation or gene expression markers for which the variance across samples was equal to zero.

Tumor DNA methylation

In TCGA, two different assays were used for epigenome-wide DNA methylation profiling, i.e., the Methylation27K (No. patients = 99) and Methylation450K assay (No. patients = 312). There were 22,443 CpG sites that had methylation levels measured on both assays, and which were therefore used in the present study. This 'natural split' in the data was used to define a discovery and validation set. The largest of the two data sets (n = 312) was used as the main discovery set. The smaller set (n = 99) was used for validation.

DNA methylation levels are represented as β-values, which range from 0 (completely unmethylated) to 1 (completely methylated). Note that although a single CpG is either methylated or not (0 or 1), a tissue sample always contains a mixture of cells [11]. This may include tumor, normal (e.g., epithelial, stromal), and infiltrating immune cells. A β-value

therefore always has a value between 0 and 1, which represents the proportion of methylated CpGs in the sample.

Tumor gene expression

The study also used tumor gene expression data from TCGA (exon expression RNAseq – IlluminaGA). Gene expression data were available for 20,359 genes. Of the 312 patients in the discovery set who had tumor DNA methylation data, 208 patients also had tumor gene expression data. The expression levels are represented as reads per kilobase of exon model per million mapped reads.

Defining risk groups

Based on (AJCC) tumor stage (I, II, III, IV) and (FIGO) grade (1, 2, 3), the EEC patients were classified into low, intermediate, and high-risk groups for having adverse cancer outcomes (e.g., recurrence/relapse or cancer death). The definition of high-risk disease was pathological stage IV or grade 3. The definition of low-risk EEC was pathological stage I and grade 1. All other patients were classified as having an intermediate-risk for having a poor prognosis.

Unsupervised hierarchical clustering

Using a Euclidean distance matrix of the DNA methylation data (*dist* in R), unsupervised hierarchical clustering of the samples was performed (*hclust* in R). The 5% most variable CpG sites were used as input for this analysis (i.e., the markers with the largest standard deviation [SD]; n = 1,122). Ward's minimum variance method was used for clustering, which aims to minimize the total within-cluster variance.

A dendrogram of the clustering results was generated, and methylation clusters were defined by cutting the dendrogram at a specific height. The optimal value for the height of the dendrogram was obtained by visual inspection. The methylation data were then visualized using a heatmap (*pheatmap* in R).

Identifying differentially methylated CpG sites in high versus low-risk EEC

Linear models were used to test for differential DNA methylation at single CpG sites in high compared to low-risk EEC (*limma* in R/Bioconductor) [12]. The *limma* procedure uses an empirical Bayes method to moderate the standard errors of the estimated log-fold changes. This has the effect of borrowing information from the full the set of markers to help with inference about each individual marker. All 22,443 CpG sites were used as input for the analysis, and filtering was not performed [13]. The false discovery rate (FDR) procedure was used to adjust for multiple testing, and FDR q-values were computed [14]. Genomic annotation data (e.g., gene name, chromosome, location in gene, epigenomic location) for all CpG sites were downloaded from: https://support.illumina.com/array/array_kits/infinium_humanmethylation450_beadchip_kit/downloads.html.

Generating a methylation signature of high versus low-risk EEC

A DNA methylation signature of high versus low-risk EEC was generated using feature selection and regularization with LASSO (least absolute shrinkage and selection operator) regression (*glmnet* in R) [15]. The 30% most variable CpG sites were used as input for the analysis (n = 6,384). A logistic LASSO regression procedure was utilized where high (coded as 1) versus low-risk EEC (coded as 0) was the response variable. LASSO logistic regression is implemented by minimizing

$$L + \lambda \sum_{j=1}^{p} |\beta_j| \, .$$

Here, L is the negative log likelihood from logistic regression, $\lambda \geq 0$ is a tuning parameter to be determined separately, p represents the CpG markers, and β_j is the model coefficient for the j^{th} CpG marker. The second term in the formula, called a shrinkage penalty (ℓ_1), has the effect of shrinking some model coefficients to be exactly zero. As such, the procedure produces a model based on a subset of the CpG markers (with non-zero coefficients) for distinguishing high from low-risk EEC.

Five-fold cross-validation (CV) was used to identify the optimal value for the tuning parameter (log λ) for classification and feature selection (grid search using default *glmnet* settings). Specifically, the value for log λ that corresponded with the lowest five-fold CV misclassification error rate was chosen. The misclassification error rate was calculated as follows: the predicted probabilities were categorized using 0.5 as a cut-point; where a value larger than 0.5 represents predicted high-risk EEC, and a value lower than 0.5 represents predicted low-risk EEC. The misclassification (error) rate is the sum of all false positive and false negative findings divided by the total number of samples.

This process was repeated 500 times, each time using a different CV split, which resulted in 500 DNA methylation models. The average CV misclassification error rate was calculated. Average LASSO coefficients of the CpG markers across all 500 repetitions were calculated. The CpG markers that had a model coefficient different from zero in at least half of the 500 repetitions were included in the final methylation signature. The methylation signature was then calculated as follows: for each patient, the methylation level of each selected CpG was multiplied by its corresponding LASSO coefficient; and the sum of these products was computed.

In an additional analysis, gene expression levels (20,359 genes) of the same patients' tumors were used. Correlations between methylation levels of the CpGs included in the signature and gene expression levels of the same genes the signature CpGs map to were investigated. FDR q-values were calculated.

Applying the methylation signature in intermediate-risk patients

Using the intermediate-risk patients, which were not used to generate the methylation signature, associations between the signature and tumor stage and grade were evaluated. Boxplots of the methylation signature in subgroups based on stage and grade were generated.

Predicting cancer recurrence

Associations between the signature and cancer recurrence were studied. For this, a receiver operating characteristic (ROC) analysis was performed (*pROC* and *ROCR* in R). A 95% confidence

interval (CI) for the area under the curve (AUC) was computed using 2,000 stratified bootstrap replicates. The performance of the signature was compared to that of a clinical model based on tumor stage and grade. Likelihood-ratio (LR) tests were used to compare the clinical model versus a model based on the clinical variables and the methylation signature combined (*lmtest* in R).

Gene set analysis

A gene set analysis was performed to identify sets of genes that are significantly up- or downregulated with higher levels of the methylation signature [16]. For this analysis, gene expression levels (20,359 genes) of the same patients' tumors were used. The *camera* method was used, which is implemented as part of *limma* in R/Bioconductor. First, gene expression levels of all individual genes were modelled in relation to the methylation signature. Second, these analytical results were compared to the predefined HALLMARK gene sets (n = 50) [17], to determine whether specific sets of genes (e.g., genes involved in DNA repair) are more up- or down-regulated with higher levels of the methylation signature. Barcodeplots were generated for the top-enriched pathways, and FDR q-values were computed.

Independent validation of the methylation signature

The validation set was used to further evaluate the signature. First, boxplots of the signature by risk category (low, intermediate, high) were generated. Second, using the subset of high and low-risk EEC patients, a logistic regression model was fit of the signature as a predictor in relation to high-risk EEC (coded as 1; low-risk EEC was coded as 0). The predicted probabilities were categorized using 0.5 as a cut-point; where a value larger than 0.5 represents predicted high-risk EEC, and a value lower than 0.5 represents predicted low-risk EEC. Using these predicted classes, a confusion matrix was generated and the misclassification error rate was computed as described previously.

After that, the signature was studied in relation to cancer recurrence. For this, an ROC analysis was performed.

RESULTS

Patient characteristics

Baseline characteristics

The discovery data set included 50 low-risk (stage I and grade 1) and 162 high-risk patients (stage IV or grade 3). There were 100 intermediate-risk patients. Note that the large number of high-risk tumors is the result of oversampling more advanced and aggressive tumors in TCGA [3]. **Table 1** shows selected baseline characteristics of the study participants by risk group. The mean age of the 312 EEC patients was 62.7 years (SD = 11.7), and this was not substantially different for patients in the low, intermediate, and high-risk category (ANOVA, P-value = 0.4).

Follow-up information on cancer recurrence

The median follow-up time for disease recurrence was 2.0 years (interquartile range [IQR]: 1.2, 3.8). The number of patients who experienced disease recurrence in the low, intermediate, and high-risk group was 3, 11, and 30, respectively (**Table 1**).

Hierarchical clustering identified four methylation phenotypes

Figure 1A shows a dendrogram of hierarchical clustering of the tumor samples based on epigenome-scale DNA methylation data. Four distinct clusters or methylation phenotypes were identified: C1 (n = 87), C2 (n = 53), C3 (n = 100), and C4 (n = 72). **Figure 1B** shows a heatmap of the methylation levels with the samples grouped by methylation cluster. Average methylation levels were significantly different between the methylation clusters (ANOVA, P-value < 0.0001): C1 = 0.39, C2 = 0.53, C3 = 0.50, C4 = 0.35. Age was not associated with methylation cluster (ANOVA, P-value = 0.08). **Figure 1C** shows the proportion of samples by risk category in each methylation cluster. Cluster 4 included the highest proportion of high-risk tumors and lowest proportion of low-risk tumors (chi-square test, P-value < 0.0001).

A large number of CpG sites were differentially methylated in high versus low-risk EEC

For the following analysis, the intermediate-risk patients were excluded. Differentially methylated CpG sites were identified by comparing high (n = 162) versus low-risk tumors (n = 50). In total, 1,503 CpGs were differentially methylated between the groups (FDR q-value < 0.01 and mean methylation β-value difference ≥ 0.05; **Figure 2A**). Of these, 15 CpGs had a mean methylation difference of at least 0.2 (**Table 2**).

The majority of the significant CpGs (n = 1,503) had a higher mean methylation level in high-risk tumors (67%). Compared to the non-significant CpGs, the significant CpGs were enriched in Open Sea regions, and were less commonly found in CpG Island regions (**Figure 2B**; chi-square test, P-value < 0.0001).

A 56-CpG methylation signature of high versus low-risk EEC

Using a repeated LASSO procedure (No. repetitions = 500), a DNA methylation signature of high versus low-risk EEC was generated (**Figure 3A**). The average five-fold CV misclassification error rate from repeated LASSO logistic regression was 0.114 (median = 0.113; IQR: 0.104, 0.123; **Figure 3B**). The final methylation signature included 56 CpG markers (**Figure 3C-D)**. Three of the CpG markers were also in the list of 15 top-ranked differentially methylated CpG sites (**Table 2**): *ITGB7* cg08374799, *ARSE* cg11964613, and *HIST1H3C* cg25438963. **Figure 3E** shows a heatmap of the methylation levels of the CpGs in the signature. Methylation signature levels were then calculated as described in the Methods and using the LASSO coefficients in **Table 3**. **Figure 3F** shows that average levels of the methylation signature increase with each higher risk group.

Applying the methylation signature in the remaining intermediate-risk patients

The methylation signature was then applied in patients with intermediate-risk tumors (n = 100). In this subset of patients, higher signature levels were associated with higher grade (G2 vs. G1; t-test, P-value = 0.02; **Figure 4B**), but not stage (ANOVA, P-value = 0.89; **Figure 4A**). The majority of patients in the intermediate-risk group had stage I and grade 2 tumors (n = 64; **Figure 4C**).

The methylation signature predicted cancer recurrence

Mean signature levels where higher in patients who experienced recurrence (n = 11) compared to those who had no evidence of recurrence (n = 84; **Figure 4D**). The signature revealed an AUC for recurrence of 0.85 (95% CI: 0.71, 0.99; **Figure 4E**). The clinical model based on tumor stage and grade had an AUC of 0.70, and adding the signature to this model significantly improved the model fit (LR test, P-value = 0.0001), which corresponded with a 18% improvement in the AUC.

The signature was then applied in patients with stage I and grade 2 tumors (n = 64). Importantly, in this subset of patients who have the same grade and stage, the classification performance of the signature for cancer recurrence (No. events = 7) remained statistically significant (AUC = 0.82; 95% CI: 0.60, 1.00; **Figure 4F**).

Higher levels of the methylation signature were associated with increased cell proliferation

Whole-genome gene expression data (20,359 genes) of the same patients' tumors were used to perform a gene set analysis of the methylation signature. **Figure 4G-I** show barcodeplots for the top 3 enriched HALLMARK gene sets (total No. gene sets = 50), which are cell proliferation gene sets [17]. As such, higher signature levels correlate with higher expression of genes involved in cell proliferation. The top 10 HALLMARK gene sets are shown in **Table 4**. Further, a different pathway analysis tool, iPathwayGuide, was used, which confirmed that higher signature levels are associated with increased DNA replication and cell cycle progression.

Methylation levels of about half of the signature CpGs were associated with gene expression levels

Gene expression data of the same patients' tumors were then used to study correlations between methylation levels of the 56 CpGs included in the signature and expression levels of the genes these CpGs map to. The 56 CpGs map to 59 genes (thus, some CpG are located in more than one gene). 26 of the CpGs/genes had a significant correlation (FDR q-value < 0.05; **Table 5**). Twenty-four of these 26 correlations were inverse.

The methylation signature predicted high-risk EEC and cancer recurrence in an independent data set

The signature was then applied in an additional EEC cohort from TCGA (No. patients = 99; **Table 6**). Average signature levels increased with each higher risk category (ANOVA, P-value < 0.0001; **Figure 5A**). The signature distinguished high from low-risk EEC with a misclassification error rate of 0.159 (**Figure 5B**). Four low-risk tumors and six high-risk tumors were misclassified.

There were only four patients who experienced disease recurrence in this data set (median follow-up time = 4.9 years; IQR: 2.4, 6.3). Despite the small number of events, the signature statistically significantly classified the patients who experienced recurrence versus those who remained recurrence-free during follow-up (AUC = 0.80; 95% CI: 0.62, 0.97; **Figure 5C-D**). Because of the small number of recurrence events, further testing of the signature in combination with clinical prognostic parameters was not performed.

DISCUSSION

Cancer patients with intermediate clinical-pathological features, in particular tumor stage and grade, typically have an unclear prognosis [18-20]. Finding biomarkers to better risk stratify these patients is therefore important. A few previous studies have generated prognostic biomarkers specifically for intermediate-risk cancer patients, including patients with thyroid cancer, leukemia, and breast cancer [21-23].

In this thesis project, information on tumor stage and grade were used to classify EEC patients into low, intermediate, and high-risk groups. The study demonstrated, for the first time, that a DNA methylation signature of high versus low-risk EEC can be used as a prognostic classifier in the remaining intermediate-risk patients.

The DNA methylation signature, which was developed using supervised learning with the LASSO method, included CpGs in different genes that are involved in various biological processes such as signaling, transcription regulation, DNA repair, immunity, and developmental processes. Some of the CpGs are in known cancer-related genes (e.g., *CDX2*, *RAD54*, *VAV1* [24-26]). Further, a few of the signature CpGs are in genes involved in cell proliferation (e.g., *BUB3*, *CETN1* [27, 28]). Increased cell proliferation is a key driver of tumor growth and progression [29]. The study also showed that in the intermediate-risk patients, higher levels of the signature correlated with higher expression of cell proliferation genes (E2F targets, G2M checkpoint, and MYX targets [17]). Therefore, the signature included CpGs in major cellular pathways and captures important biological variation associated with cell cycle proliferation and, therefore, tumor progression.

In addition to supervised learning, the present study also applied unsupervised statistical clustering, and, as such, identified four tumor clusters or phenotypes in the overall data set. Although these methylation phenotypes were statistically significantly associated with EEC risk category, this clustering did not accurately distinguish high from low-risk tumors, and it is therefore not a strong independent prognostic classifier. This is not surprising as unsupervised techniques are not typically used to generate predictive signatures for clinical decision making [30].

Strengths and limitations

An important strength of the project is the large overall number of EEC patients, and the availability of a validation set to independently test the classification performance of the methylation signature generated in the discovery data set. Further, TCGA includes multiple omics data types from the same tumors. In this project, tumor DNA methylation data were integrated with tumor RNAseq data to obtain further biological insights into the role of the methylation signature in EEC progression.

A limitation of the study is the short follow-up for cancer recurrence. It is therefore not unlikely that some of the patients who were classified as recurrence-free, actually experienced recurrence after the end of follow-up; suggesting endpoint misclassification. Note, however, that this did not influence the methylation signature, because the definitions of high (stage IV or grade 3) and low-risk EEC (stage I and grade 1) were based on tumor stage and grade only. In addition to the short follow-up, the TCGA cohort has a retrospective design, and prospective cohorts are preferred for prognostic biomarker evaluation.

Future research

The present study applied the LASSO method to generate a methylation signature, and did not consider other supervised statistical learning procedures (e.g., random forests [31]), which will generate different signatures from the methylation data with a potentially different classification performance. Further studies are needed to determine if these alternative methods produce a methylation signature with an even better classification performance than the LASSO-based methylation signature in the present rapport.

Because of the short follow-up time in TCGA, it is important to further examine the association of the methylation signature with recurrence in additional independent patient cohorts with long-term and complete follow-up on cancer recurrence (at least 5 years). Further, the ultimate goal of any biomarker is to classify patients into distinct risk categories. In this study, the signature was used as a continuous parameter to predict recurrence. Further research on the

prognostic signature should therefore also include finding the best cut-point for classification; i.e., the cut-point that provides the highest biomarker sensitivity and specificity (e.g., Youden Index [32]).

General conclusion

The tumor methylation signature has potential as a prognostic biomarker for EEC patients with intermediate stage/grade tumors, which represent a large subset of all patients. Accurately predicting the prognosis of EEC patients is important to help guide clinical decision making and identify the best post-operative disease management strategy for each patient.

REFERENCES

1. Morice P, Leary A, Creutzberg C, Abu-Rustum N, Darai E: **Endometrial cancer**. *Lancet* 2016, **387**(10023):1094-1108.
2. Siegel RL, Miller KD, Jemal A: **Cancer Statistics, 2017**. *CA Cancer J Clin* 2017, **67**(1):7-30.
3. Cancer Genome Atlas Research N, Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, Shen H, Robertson AG, Pashtan I, Shen R *et al*: **Integrated genomic characterization of endometrial carcinoma**. *Nature* 2013, **497**(7447):67-73.
4. Kurra V, Krajewski KM, Jagannathan J, Giardino A, Berlin S, Ramaiya N: **Typical and atypical metastatic sites of recurrent endometrial carcinoma**. *Cancer Imaging* 2013, **13**:113-122.
5. Piulats JM, Guerra E, Gil-Martin M, Roman-Canal B, Gatius S, Sanz-Pamplona R, Velasco A, Vidal A, Matias-Guiu X: **Molecular approaches for classifying endometrial carcinoma**. *Gynecol Oncol* 2017, **145**(1):200-207.
6. Gupta V, McGunigal M, Prasad-Hayes M, Kalir T, Liu J: **Adjuvant radiation therapy is associated with improved overall survival in high-intermediate risk stage I endometrial cancer: A national cancer data base analysis**. *Gynecol Oncol* 2017, **144**(1):119-124.
7. Winham WM, Lin D, Stone PJ, Nucci MR, Quick CM: **Architectural versus nuclear atypia-defined FIGO grade 2 endometrial endometrioid adenocarcinoma (EEC): a clinicopathologic comparison of 154 cases with clinical follow-up**. *Int J Gynecol Pathol* 2014, **33**(2):120-126.
8. Farkas SA, Sorbe BG, Nilsson TK: **Epigenetic changes as prognostic predictors in endometrial carcinomas**. *Epigenetics* 2017, **12**(1):19-26.
9. Schubeler D: **Function and information content of DNA methylation**. *Nature* 2015, **517**(7534):321-326.
10. Jones PA, Issa JP, Baylin S: **Targeting the cancer epigenome for therapy**. *Nat Rev Genet* 2016, **17**(10):630-641.
11. Aran D, Sirota M, Butte AJ: **Systematic pan-cancer analysis of tumour purity**. *Nat Commun* 2015, **6**:8971.
12. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies**. *Nucleic Acids Res* 2015, **43**(7):e47.
13. Bourgon R, Gentleman R, Huber W: **Independent filtering increases detection power for high-throughput experiments**. *Proc Natl Acad Sci U S A* 2010, **107**(21):9546-9551.
14. Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *Proc Natl Acad Sci U S A* 2003, **100**(16):9440-9445.
15. Tibshirani R: **Regression shrinkage and selection via the Lasso**. *J Roy Stat Soc B Met* 1996, **58**(1):267-288.
16. Wu D, Smyth GK: **Camera: a competitive gene set test accounting for inter-gene correlation**. *Nucleic Acids Res* 2012, **40**(17):e133.
17. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P: **The Molecular Signatures Database (MSigDB) hallmark gene set collection**. *Cell Syst* 2015, **1**(6):417-425.

18. Di Costanzo GG, Tortora R: **Intermediate hepatocellular carcinoma: How to choose the best treatment modality?** *World J Hepatol* 2015, **7**(9):1184-1191.
19. Kamat AM, Witjes JA, Brausi M, Soloway M, Lamm D, Persad R, Buckley R, Bohle A, Colombel M, Palou J: **Defining and treating the spectrum of intermediate risk nonmuscle invasive bladder cancer**. *J Urol* 2014, **192**(2):305-315.
20. Kane CJ, Eggener SE, Shindel AW, Andriole GL: **Variability in Outcomes for Patients with Intermediate-risk Prostate Cancer (Gleason Score 7, International Society of Urological Pathology Gleason Group 2-3) and Implications for Risk Stratification: A Systematic Review**. *Eur Urol Focus* 2017.
21. Brennan K, Holsinger C, Dosiou C, Sunwoo JB, Akatsu H, Haile R, Gevaert O: **Development of prognostic signatures for intermediate-risk papillary thyroid cancer**. *BMC Cancer* 2016, **16**(1):736.
22. Chretien AS, Fauriat C, Orlanducci F, Rey J, Borg GB, Gautherot E, Granjeaud S, Demerle C, Hamel JF, Cerwenka A *et al*: **NKp30 expression is a prognostic immune biomarker for stratification of patients with intermediate-risk acute myeloid leukemia**. *Oncotarget* 2017, **8**(30):49548-49563.
23. Ignatov T, Eggemann H, Burger E, Fettke F, Costa SD, Ignatov A: **Moderate level of HER2 expression and its prognostic significance in breast cancer with intermediate grade**. *Breast Cancer Res Treat* 2015, **151**(2):357-364.
24. Ghamrasni SE, Cardoso R, Li L, Guturi KK, Bjerregaard VA, Liu Y, Venkatesan S, Hande MP, Henderson JT, Sanchez O *et al*: **Rad54 and Mus81 cooperation promotes DNA damage repair and restrains chromosome missegregation**. *Oncogene* 2016, **35**(37):4836-4845.
25. Katzav S: **Vav1: A Dr. Jekyll and Mr. Hyde protein--good for the hematopoietic system, bad for cancer**. *Oncotarget* 2015, **6**(30):28731-28742.
26. Olsen J, Espersen ML, Jess P, Kirkeby LT, Troelsen JT: **The clinical perspectives of CDX2 expression in colorectal cancer: a qualitative systematic review**. *Surg Oncol* 2014, **23**(3):167-176.
27. Grabsch H, Takeno S, Parsons WJ, Pomjanski N, Boecking A, Gabbert HE, Mueller W: **Overexpression of the mitotic checkpoint genes BUB1, BUBR1, and BUB3 in gastric cancer--association with tumour cell proliferation**. *J Pathol* 2003, **200**(1):16-22.
28. Hart PE, Glantz JN, Orth JD, Poynter GM, Salisbury JL: **Testis-specific murine centrin, Cetn1: genomic characterization and evidence for retroposition of a gene encoding a centrosome protein**. *Genomics* 1999, **60**(2):111-120.
29. Evan GI, Vousden KH: **Proliferation, cell cycle and apoptosis in cancer**. *Nature* 2001, **411**(6835):342-348.
30. Nuyten DS, Hastie T, Chi JT, Chang HY, van de Vijver MJ: **Combining biological gene expression signatures in predicting outcome in breast cancer: An alternative to supervised classification**. *Eur J Cancer* 2008, **44**(15):2319-2329.
31. Ho TK: **Random Decision Forests**. *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC* 1995:278-282.
32. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF: **Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection**. *Biom J* 2008, **50**(3):419-430.

ABBREVIATIONS

AUC, area under the curve
CI, confidence interval
CV, cross-validation
EEC, endometrioid endometrial cancer
FDR, false discovery rate
FIGO, International Federation of Gynaecology and Obstetrics
IQR, interquartile range
LASSO, least absolute shrinkage and selection operator
LR, likelihood-ratio
ROC, receiver operating characteristic
SD, standard deviation
TCGA, The Cancer Genome Atlas
TSS, transcription start site
TSS1500, the region from 200 to 1,500 nucleotides upstream of the TSS
TSS200, the region from 200 nucleotides upstream to the TSS itself
UCEC, uterine corpus endometrial carcinoma
UCSC, University of California, Santa Cruz

TABLES

Table 1

Selected baseline characteristics of study participants (n = 312) by risk category

	Low-risk (stage I and grade 1) n = 50	**Intermediate-risk**[a] n = 100	**High-risk (stage IV or grade 3)** n = 162
Mean age, years (SD)	62.64 (11.40)	61.53 (11.00)	63.43 (12.19)
Stage I tumors	50	64	101
Stage II tumors		12	17
Stage III tumors		24	33
Stage IV tumors			11
Grade 1 tumors	50	12	
Grade 2 tumors		88	1
Grade 3 tumors			161
Recurrence events[b]			
No	46	84	111
Yes	3	11	30
Missing	1	5	21

[a] This included all patients not in the low or high-risk group.

[b] Median follow-up time for recurrence was 2.0 years (IQR: 1.2, 3.8). Note that data on cancer recurrence was not used to define risk groups.

Table 2

Fifteen top-ranked differentially methylated CpG sites in high versus low-risk EEC with a methylation β-value difference of at least 20 percent[a]

CpG ID	Chr.	Gene	Location in gene	Epigenomic location	FDR q-value
cg08374799	12	*ITGB7*	TSS200	Open Sea	1.03E-07
cg11964613	X	*ARSE*	5'UTR;1stExon	Open Sea	1.03E-07
cg07911673	17	*SLC25A35*	1stExon	N_Shore	2.60E-07
cg12259256	17	*TMEM101*	TSS200	Island	8.97E-07
cg16148454	16	*TMED6*	1stExon	Open Sea	1.69E-06
cg07014174	21	*KRTAP11-1*[b]	1stExon	Open Sea	1.59E-05
cg24240626	2	*REG3A*[b]	5'UTR;1stExon	Open Sea	1.78E-05
cg22376897	X	*ARSE*	5'UTR	Open Sea	3.13E-05
cg18368125	16	*TMED6*	TSS200	Open Sea	4.35E-05
cg07525077	14	*RNASE3*[b]	Body	Open Sea	6.56E-05
cg04576021	6	*HLA-DOB*	Body	Open Sea	8.42E-05
cg25259754	1	*FCRL3*[b]	Body	Open Sea	1.15E-04
cg25438963	6	*HIST1H3C*	1stExon	Island	1.99E-04
cg07636178	6	*HIST1H3C*	TSS200	N_Shore	2.52E-04
cg21250296	6	*HIST1H2BB;HIST1H3C*	TSS1500	Island	4.24E-04

[a] Table is sorted by FDR q-value.

[b] These four CpGs had a lower mean methylation level in high compared to low-risk tumors (i.e., hypomethylation). All other CpGs had a higher mean methylation level in high compared to low-risk tumors (i.e., hypermethylation).

Table 3

Fifty-six CpG sites included in the methylation signature of high-risk EEC[a]

CpG ID	Chr.	Gene	Location in gene	Epigenomic location	Mean β low-risk	Mean β high-risk	LASSO coefficient
cg02055963	13	*CDX2*	TSS1500	Island	0.16	0.32	3.31
cg26581729	9	*NPDC1*	Body	Island	0.31	0.51	2.86
cg27270684	7	*FKBP9L*	TSS200;TSS1500;Body	Open Sea	0.23	0.38	2.60
cg06415153	12	*PITPNM2*	5'UTR	Open Sea	0.39	0.52	2.35
cg08374799	12	*ITGB7*	TSS200	Open Sea	0.16	0.37	1.48
cg14244577	16	*DDX19B*	TSS200	Open Sea	0.29	0.39	1.41
cg26323655	8	*RAD54B*	Body	Open Sea	0.72	0.81	1.24
cg11964613	X	*ARSE*	5'UTR;1stExon	Open Sea	0.36	0.60	0.76
cg08090640	17	*IFI35*	Body	Open Sea	0.21	0.40	0.73
cg26571739	19	*VAV1*	TSS200	Open Sea	0.14	0.23	0.71
cg00930194	5	*PROP1*	TSS1500	Open Sea	0.56	0.66	0.64
cg11981631	11	*ABCC8*	Body	Island	0.08	0.13	0.57
cg25021247	3	*AMT;NICN1*	TSS200;3'UTR	Open Sea	0.55	0.71	0.54
cg14597908	20	*GNASAS;GNAS*	Body;1stExon;5'UTR	N_Shore	0.43	0.50	0.47
cg16175725	12	*HNF1A*	1stExon	Island	0.48	0.66	0.42
cg02620769	12	*CCDC65*	1stExon;5'UTR	Open Sea	0.12	0.23	0.40
cg07251857	15	*ALPK3*	1stExon	Island	0.67	0.75	0.32
cg05155595	2	*ANXA4*	5'UTR	Open Sea	0.47	0.64	0.28
cg13878010	3	*ADCY5*	1stExon	Island	0.18	0.31	0.28
cg15572745	14	*NRXN3*	5'UTR	Open Sea	0.31	0.35	0.27
cg24471894	9	*KIAA0020*	5'UTR	Open Sea	0.19	0.37	0.27
cg15021292	5	*PIK3R1*	TSS1500	Open Sea	0.67	0.74	0.27
cg25438963	6	*HIST1H3C*	1stExon	Island	0.27	0.52	0.26
cg16516400	1	*FAM89A*	TSS1500	S_Shore	0.25	0.40	0.25
cg13323752	12	*SLC2A14*	TSS200	Island	0.29	0.34	0.19
cg21870884	1	*GPR25*	1stExon	Island	0.18	0.28	0.14
cg19246110	19	*ZNF671*	1stExon;5'UTR	Island	0.12	0.28	0.13
cg23152772	9	*FIBCD1*	Body	Island	0.34	0.44	0.11
cg20483374	11	*C1QTNF5;MFRP*	5'UTR;3'UTR	Island	0.18	0.36	0.10
cg11319389	20	*TOX2*	TSS200;Body	Island	0.30	0.44	0.07
cg20770175	2	*COL3A1*	Body	Open Sea	0.63	0.76	0.06
cg08203715	11	*ST3GAL4*	1stExon;5'UTR	S_Shore	0.49	0.40	-0.09
cg12529228	1	*NHLH1*	1stExon;5'UTR	N_Shelf	0.85	0.77	-0.13
cg14467840	1	*S100A13;S100A1*	TSS1500;5'UTR	Open Sea	0.79	0.68	-0.13
cg26530341	8	*TNFRSF10A*	TSS1500	S_Shore	0.72	0.55	-0.21

cg22035229	1	*MSH4*	1stExon;5'UTR	Island	0.90	0.82	-0.24
cg23765993	20	*SPINLW1*	TSS200;TSS1500	Open Sea	0.85	0.71	-0.25
cg26065841	15	*CHAC1*	TSS1500	N_Shore	0.73	0.62	-0.26
cg09324116	X	*GEMIN8*	TSS1500	Island	0.31	0.21	-0.28
cg16612699	11	*OR8B8*	TSS1500	Open Sea	0.55	0.46	-0.30
cg04058169	10	*BUB3*	TSS1500	N_Shore	0.82	0.72	-0.37
cg03264209	16			S_Shore	0.61	0.52	-0.37
cg23101680	13	*SPERT*	Body	Open Sea	0.68	0.53	-0.42
cg25119415	1	*MNDA*	TSS1500	Open Sea	0.54	0.38	-0.45
cg08085267	17	*C17orf57*	5'UTR	S_Shore	0.19	0.15	-0.46
cg26757722	22	*CACNG2*	1stExon;5'UTR	N_Shore	0.68	0.63	-0.47
cg21229055	X	*GPM6B*	TSS1500	S_Shore	0.36	0.28	-0.52
cg23264413	19	*PSG4*	TSS1500	Open Sea	0.76	0.69	-0.65
cg14704941	11	*CSRP3*	TSS1500	Open Sea	0.89	0.83	-0.87
cg03022541	3	*DNAJB8*	TSS1500	Open Sea	0.72	0.60	-0.92
cg12288726	7	*ARF5*	TSS1500	Island	0.78	0.68	-0.97
cg18129786	3	*ZNF445*	TSS1500	S_Shore	0.88	0.83	-1.03
cg03020597	X	*SLITRK2*	TSS1500	N_Shore	0.34	0.23	-1.26
cg05111110	11	*PC;LRFN4*	Body;1stExon	Island	0.85	0.76	-1.33
cg26738010	18	*CETN1*	TSS200	Island	0.67	0.58	-1.49
cg10305797	19	*KRTDAP*	TSS1500	Open Sea	0.61	0.48	-1.90

[a] Table is sorted by LASSO coefficient.

Table 4

Gene set analysis of the methylation signature in intermediate-risk patients: top-10 enriched HALLMARK gene sets[a,b]

Gene set	No. genes in set	Direction	P-value	FDR q-value
HALLMARK E2F TARGETS	188	Up	5.19E-39	2.60E-37
HALLMARK G2M CHECKPOINT	189	Up	1.71E-29	4.27E-28
HALLMARK MYC TARGETS V1	192	Up	5.01E-14	8.34E-13
HALLMARK MTORC1 SIGNALING	196	Up	3.70E-11	4.62E-10
HALLMARK MYC TARGETS V2	58	Up	1.08E-08	1.08E-07
HALLMARK MITOTIC SPINDLE	198	Up	2.02E-07	1.68E-06
HALLMARK ALLOGRAFT REJECTION	200	Down	4.77E-06	3.41E-05
HALLMARK OXIDATIVE PHOSPHORYLATION	197	Up	2.42E-05	1.51E-04
HALLMARK EPITHELIAL MESENCHYMAL TRANSITION	196	Down	2.09E-04	1.16E-03
HALLMARK DNA REPAIR	141	Up	4.45E-04	2.23E-03

[a] Barcodeplots for the top-3 enriched gene sets are shown in Figure 4 (G-I).

[b] The total number of HALLMARK gene sets is 50 [17].

Table 5

Correlations methylation and expression levels of the CpG/genes included in the methylation signature[a]

CpG ID	Gene name	Pearson correlation	P-value	FDR q-value
cg07251857	*ALPK3*	-0.85	1.97E-59	1.16E-57
cg16175725	*HNF1A*	-0.79	1.60E-46	4.71E-45
cg19246110	*ZNF671*	-0.78	5.69E-44	1.12E-42
cg11964613	*ARSE*	-0.75	1.32E-38	1.95E-37
cg25021247	*AMT*	-0.74	5.62E-37	6.63E-36
cg14467840	*S100A1*	-0.61	5.49E-23	5.40E-22
cg25438963	*HIST1H3C*	-0.60	6.37E-22	5.37E-21
cg02620769	*CCDC65*	-0.52	9.97E-16	7.35E-15
cg08090640	*IFI35*	-0.49	3.12E-14	2.05E-13
cg26581729	*NPDC1*	-0.49	5.94E-14	3.51E-13
cg05155595	*ANXA4*	-0.47	6.46E-13	3.18E-12
cg08203715	*ST3GAL4*	-0.46	2.37E-12	1.08E-11
cg05111110	*LRFN4*	-0.43	7.13E-11	3.00E-10
cg14467840	*S100A13*	-0.38	2.05E-08	8.07E-08
cg20483374	*MFRP*	-0.36	7.98E-08	2.94E-07
cg03020597	*SLITRK2*	-0.35	2.27E-07	7.88E-07
cg02055963	*CDX2*	-0.33	1.37E-06	4.48E-06
cg25021247	*NICN1*	-0.31	5.86E-06	1.82E-05
cg26323655	*RAD54B*	-0.25	2.11E-04	5.93E-04
cg20770175	*COL3A1*	-0.24	5.87E-04	1.57E-03
cg15572745	*NRXN3*	-0.21	2.07E-03	5.32E-03
cg08374799	*ITGB7*	-0.20	3.47E-03	8.54E-03
cg08085267	*C17orf57*	-0.18	8.61E-03	2.03E-02
cg18129786	*ZNF445*	-0.17	1.55E-02	3.51E-02
cg26530341	*TNFRSF10A*	-0.13	5.40E-02	1.14E-01
cg26065841	*CHAC1*	-0.13	6.47E-02	1.32E-01
cg11319389	*TOX2*	-0.12	7.70E-02	1.51E-01
cg14597908	*GNAS*	-0.12	8.23E-02	1.57E-01
cg05111110	*PC*	-0.12	8.83E-02	1.63E-01
cg12288726	*ARF5*	-0.09	1.93E-01	3.17E-01
cg23264413	*PSG4*	-0.09	2.13E-01	3.39E-01
cg15021292	*PIK3R1*	-0.08	2.24E-01	3.48E-01
cg09324116	*GEMIN8*	-0.08	2.47E-01	3.74E-01
cg04058169	*BUB3*	-0.08	2.74E-01	4.04E-01

cg00930194	*PROP1*	-0.06	3.71E-01	5.08E-01
cg16516400	*FAM89A*	-0.06	3.80E-01	5.09E-01
cg22035229	*MSH4*	-0.05	4.34E-01	5.69E-01
cg27270684	*FKBP9L*	-0.05	4.47E-01	5.74E-01
cg13323752	*SLC2A14*	-0.04	6.02E-01	7.27E-01
cg26757722	*CACNG2*	-0.04	6.04E-01	7.27E-01
cg11981631	*ABCC8*	-0.03	6.33E-01	7.47E-01
cg14597908	*GNASAS*	-0.02	7.22E-01	8.19E-01
cg03022541	*DNAJB8*	-0.02	7.83E-01	8.58E-01
cg16612699	*OR8B8*	-0.02	8.10E-01	8.69E-01
cg26571739	*VAV1*	0.00	9.85E-01	9.85E-01
cg21229055	*GPM6B*	0.01	8.93E-01	9.08E-01
cg26738010	*CETN1*	0.01	8.60E-01	8.90E-01
cg23101680	*SPERT*	0.02	8.27E-01	8.72E-01
cg14244577	*DDX19B*	0.02	7.85E-01	8.58E-01
cg23152772	*FIBCD1*	0.03	7.00E-01	8.09E-01
cg06415153	*PITPNM2*	0.05	4.83E-01	6.06E-01
cg13878010	*ADCY5*	0.07	3.04E-01	4.27E-01
cg21870884	*GPR25*	0.07	2.82E-01	4.06E-01
cg12529228	*NHLH1*	0.10	1.72E-01	2.89E-01
cg14704941	*CSRP3*	0.10	1.38E-01	2.40E-01
cg25119415	*MNDA*	0.11	1.03E-01	1.83E-01
cg24471894	*KIAA0020*	0.14	4.45E-02	9.73E-02
cg10305797	*KRTDAP*	0.31	7.47E-06	2.20E-05
cg23765993	*SPINLW1*	0.49	8.78E-14	4.71E-13

[a] Table is sorted by Pearson correlation of CpG methylation and gene expression levels (lowest to highest).

Table 6

Selected baseline characteristics of patients in the validation data set (n = 99) by risk category

	Low-risk (stage I and grade 1)	**Intermediate-risk[a]**	**High-risk (stage IV or grade 3)**
	n = 31	n = 36	n = 32
Mean age, years (SD)	61.06 (9.58)	62 (12.57)	62.97 (10.98)
Stage I tumors	31	24	20
Stage II tumors		5	2
Stage III tumors		7	7
Stage IV tumors			3
Grade 1 tumors	31	5	
Grade 2 tumors		31	
Grade 3 tumors			52
Recurrence events[b]			
No	24	33	19
Yes	1	1	2
Missing	6	2	11

[a] This included all patients not in the low or high-risk group.

[b] Median follow-up time for recurrence was 4.9 years (IQR: 2.4, 6.3). Note that data on cancer recurrence was not used to define risk groups.

FIGURES

Figure 1

Unsupervised hierarchical clustering of the tumor samples based on DNA methylation levels. **(A)** Dendrogram of hierarchical clustering of the tumor samples. The orange boxes highlight four main methylation clusters or phenotypes. **(B)** Heatmap of methylation levels. The rows are the 5% most variable CpGs (largest SD), which were used as input for hierarchical clustering. The columns are the samples grouped by methylation cluster. Information on selected patient variables (i.e., age at diagnosis, risk group) was added at the top of the heatmap by means of colored bars. **(C)** Proportion of samples by risk category in each methylation cluster or phenotype.

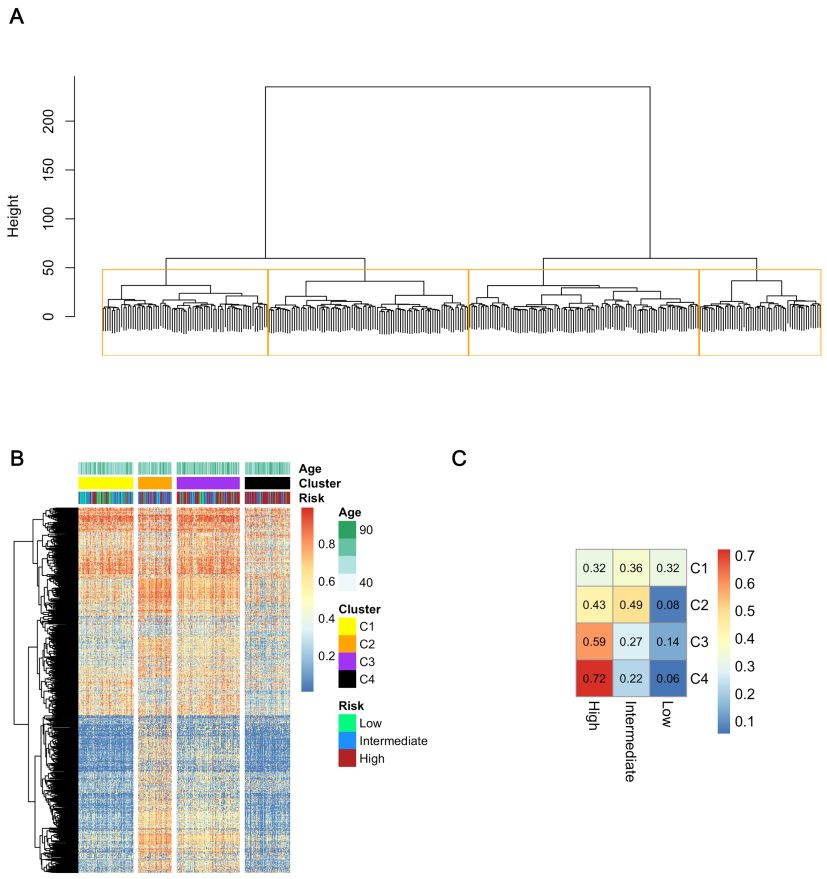


Figure 2

Testing for differential DNA methylation at single CpG sites in high versus low-risk EEC. (A) Volcano plot showing the results from statistical testing. Each dot represents a CpG. The CpGs shown in blue/orange have an FDR q-value less than 0.01 (n = 5,132). Further, the CpGs shown in dark blue and orange have a significant q-value and a mean methylation β-value difference of at least 0.05 (n = 1,503). The CpGs shown in orange have a significant q-value and a mean methylation difference of at least 0.2 (n = 15). **(B)** Proportion of samples by epigenomic location in the group of non-significant (n = 20,940) and significant CpGs (FDR q-level < 0.01 and mean β-value difference ≥ 0.05; n = 1,503).

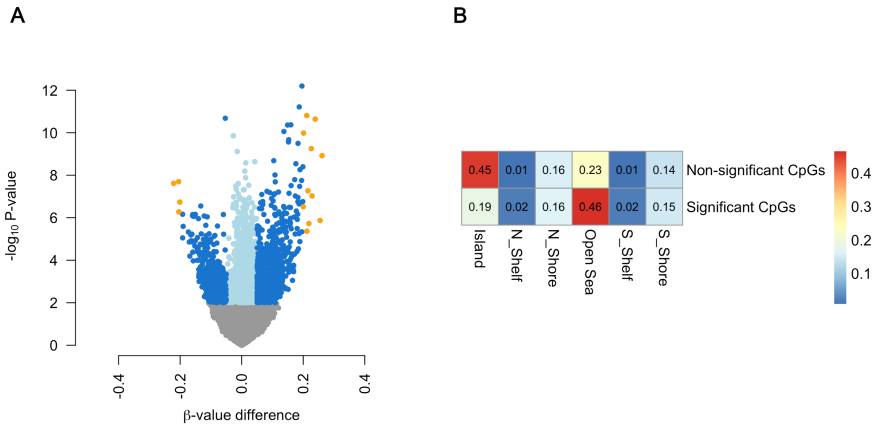


Figure 3

Building a DNA methylation signature of high versus low-risk EEC. (A) Coefficient path from LASSO regularization for classifying high versus low-risk EEC. The colored lines represent individual CpGs, and the y-axis shows the value of the coefficient associated with each CpG, which is a function of log λ (x-axis). The optimal value for log λ for classification and feature selection was identified using five-fold CV. This procedure was repeated 500 times, each time randomly selecting a different CV split, which resulted in 500 DNA methylation models. **(B)** Distribution of the CV misclassification error rate from repeated LASSO. The vertical dashed line represents the average misclassification error. **(C)** The final methylation signature was built using the CpG markers that were selected (i.e., coefficient different from zero) in at least half of the 500 repetitions. This resulted in the selection of 56 markers (shown in red). The total number of unique CpG markers in any of the 500 LASSO models was 110. The median number of CpGs with non-zero model coefficients across all 500 models was 58 (range: 24, 88). **(D)** Shrinkage of the LASSO model coefficients of the 56 signature CpGs. The x-axis shows the LASSO coefficients. The y-axis shows the coefficients from logistic regression models including single CpG markers, and high versus low-risk EEC as the response. **(E)** Heatmap of DNA methylation levels of the 56 CpG sites included in the signature (rows). The samples are grouped by risk category (columns). The CpGs were clustered based on Euclidean distance and the complete linkage method. **(F)** Boxplots of the methylation signature by risk group. The signature was calculated for each patient using the methylation β-values of the 56 CpGs and their LASSO coefficients as explained in the Methods.

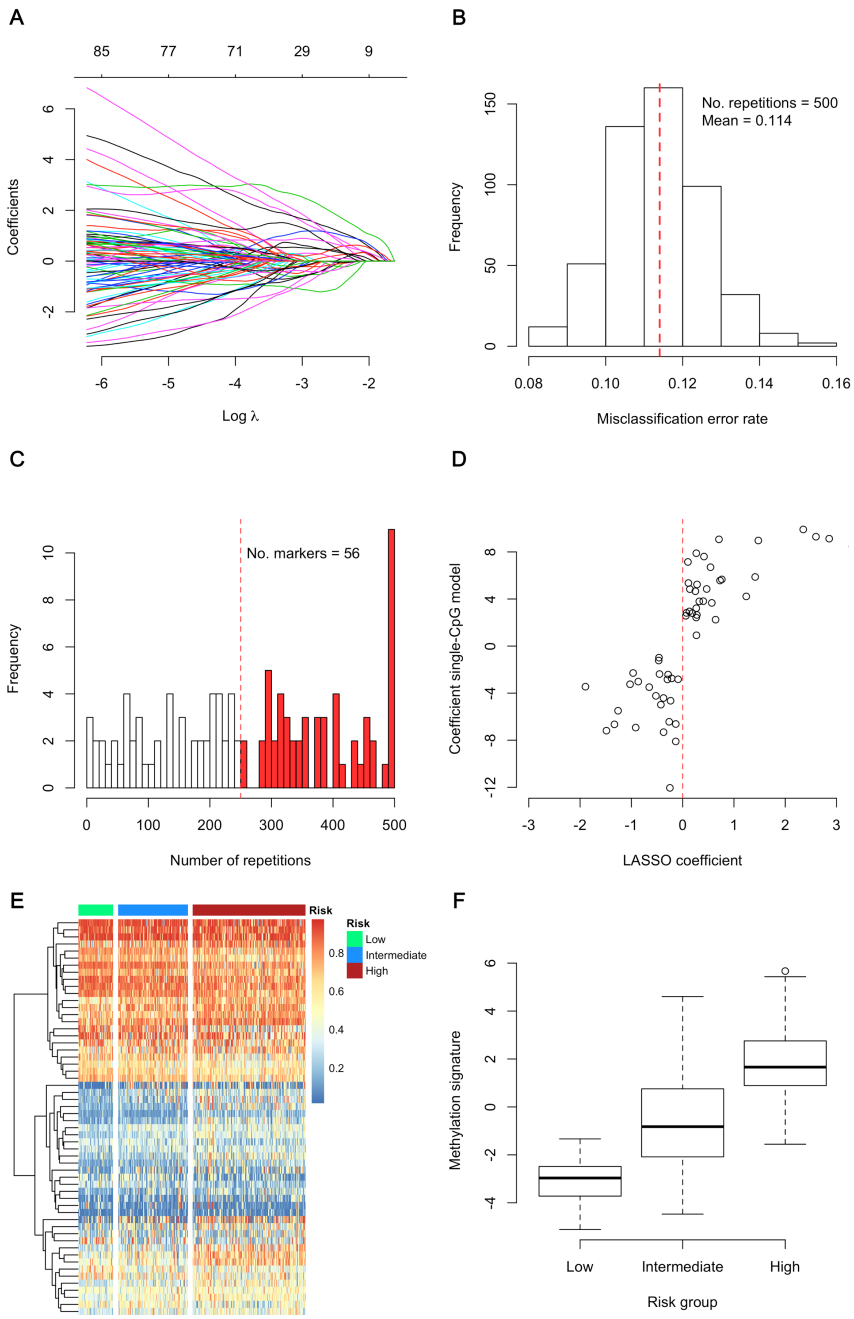

A
85
77
71
29
9
Coefficients
Log λ
B
No. repetitions = 500
Mean = 0.114
Frequency
Misclassification error rate
C
No. markers = 56
Frequency
Number of repetitions
D
Coefficient single-CpG model
LASSO coefficient
E
Risk
Low
Intermediate
High
F
Methylation signature
Low
Intermediate
High
Risk group

Figure 4

Applying the methylation signature of high versus low-risk EEC in the remaining intermediate-risk patients. (A-B) Boxplots of the methylation signature by stage and grade. **(C)** Number of intermediate-risk patients by stage and grade. **(D)** Boxplots of the signature by disease recurrence status. **(E)** ROC analysis of different predictive models for recurrence (yes vs. no). The black curve represents the clinical model based on tumor stage and grade. The blue curve is the model based on the methylation signature only. The red curve represents the combined model that includes the clinical variables and the signature. The AUC for each model and associated 95% CI are shown in the figure. The P-value from the LR-test comparing the clinical model versus the model based on both the clinical variables and the signature is shown as well. **(F)** ROC analysis of the signature in relation to recurrence in patients with stage I and grade 2 tumors only (n = 64). **(G-I)** Higher levels of the methylation signature correlated with increased expression of cell proliferation genes (E2F targets, G2M checkpoint, and MYC targets). Barcodeplots for the top-3 HALLMARK gene sets are shown.

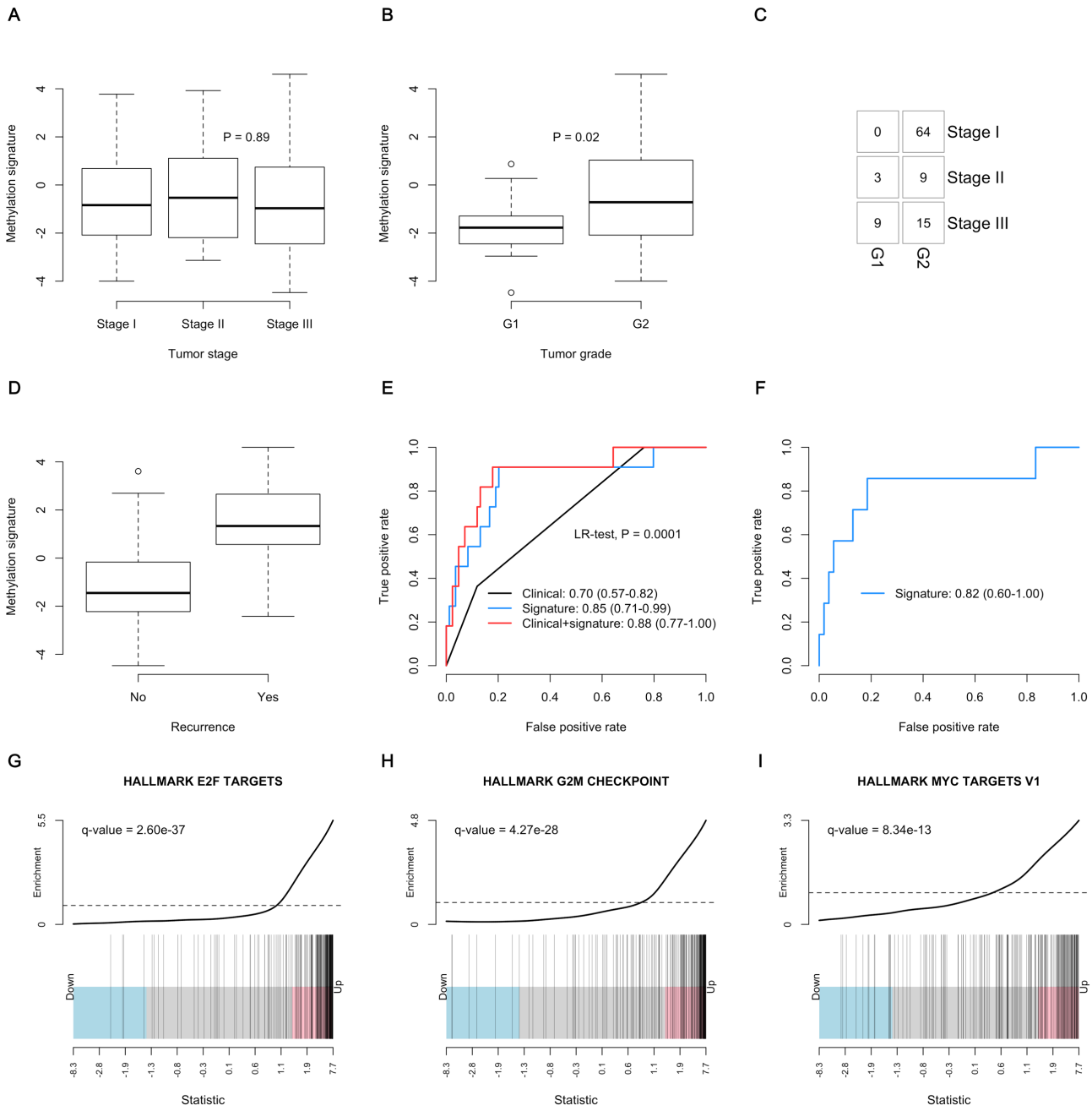
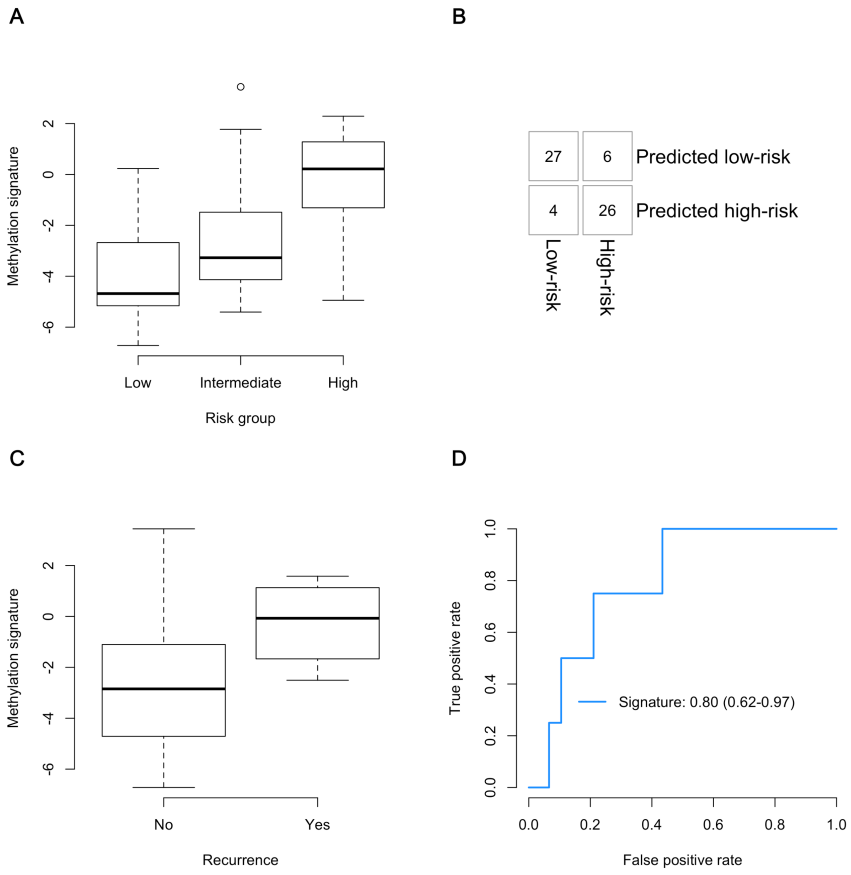

A
Methylation signature
P = 0.89
Stage I
Stage II
Stage III
Tumor stage
B
Methylation signature
P = 0.02
G1
G2
Tumor grade
C
0
64
Stage I
3
9
Stage II
9
15
Stage III
G1
G2
D
Methylation signature
No
Yes
Recurrence
E
True positive rate
LR-test, P = 0.0001
Clinical: 0.70 (0.57-0.82)
Signature: 0.85 (0.71-0.99)
Clinical+signature: 0.88 (0.77-1.00)
False positive rate
F
True positive rate
Signature: 0.82 (0.60-1.00)
False positive rate
G
HALLMARK E2F TARGETS
q-value = 2.60e-37
Enrichment
Down
Up
Statistic
H
HALLMARK G2M CHECKPOINT
q-value = 4.27e-28
Enrichment
Down
Up
Statistic
I
HALLMARK MYC TARGETS V1
q-value = 8.34e-13
Enrichment
Down
Up
Statistic

Figure 5

Validation of the methylation signature in an independent data set. (A) Boxplots of the signature by risk group. **(B)** Confusion matrix based on classifying high and low-risk EEC tumors using the methylation signature. The misclassification error rate was 0.159. **(C)** Boxplots of the signature by recurrence status in all patients (n_{yes} = 4; n_{no} = 76). **(D)** ROC analysis of the methylation signature for predicting recurrence (yes vs. no). The AUC and its 95% CI are shown.



Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Tumor DNA methylation profiles of high-risk endometrioid endometrial carcinoma and patient recurrence

Richting: **Master of Statistics-Bioinformatics**
Jaar: **2018**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Geybels, Milan

Datum: **22/01/2018**