Article

Integration of Genomic and Clinical Retrospective Data to Predict Endometrioid Endometrial Cancer Recurrence

Jesus Gonzalez-Bosquet [1],*, Sofia Gabrilovich [1], Megan E. McDonald [1], Brian J. Smith [2], Kimberly K. Leslie [3], David D. Bender [1], Michael J. Goodheart [1] and Eric Devor [1]

[1] Department of Obstetrics and Gynecology, University of Iowa, 200 Hawkins Dr., Iowa City, IA 52242, USA
[2] Department of Biostatistics, University of Iowa, 145 N Riverside Dr., Iowa City, IA 52242, USA
[3] Division of Molecular Medicine, Departments of Internal Medicine and Obstetrics and Gynecology, The University of New Mexico Comprehensive Cancer Center, 915 Camino de Salud, CRF 117, Albuquerque, NM 87131, USA
* Correspondence: jesus-gonzalezbosquet@uiowa.edu; Tel.: +1-(319)-356-2160; Fax: +1-(319)-353-8363

Abstract: Endometrial cancer (EC) incidence and mortality continues to rise. Molecular profiling of EC promises improvement of risk assessment and treatment selection. However, we still lack robust and accurate models to predict those at risk of failing treatment. The objective of this pilot study is to create models with clinical and genomic data that will discriminate patients with EC at risk of disease recurrence. We performed a pilot, retrospective, case–control study evaluating patients with EC, endometrioid type: 7 with recurrence of disease (cases), and 55 without (controls). RNA was extracted from frozen specimens and sequenced (RNAseq). Genomic features from RNAseq included transcriptome expression, genomic, and structural variation. Feature selection for variable reduction was performed with univariate ANOVA with cross-validation. Selected variables, informative for EC recurrence, were introduced in multivariate lasso regression models. Validation of models was performed in machine-learning platforms (ML) and independent datasets (TCGA). The best performing prediction models (out of >170) contained the same lncRNA features (AUC of 0.9, and 95% CI: 0.75, 1.0). Models were validated with excellent performance in ML platforms and good performance in an independent dataset. Prediction models of EC recurrence containing lncRNA features have better performance than models with clinical data alone.

Keywords: endometrial cancer; recurrence; prediction; machine learning

1. Introduction

Endometrial cancer (EC) is the most common gynecologic malignancy in developed countries. It is estimated that 65,950 new uterine cancer cases will be diagnosed in the United States in 2022, accounting for 12,550 deaths [1]. Unlike other cancer types, incidence and mortality of EC have been increasing for the last 2 decades [1]. This is mainly considered due to an aging population and increased rates of obesity and metabolic syndrome [2]. Obesity contributes to an endogenous unopposed estrogen environment and is the single most important risk factor for EC [2]. The increase in EC mortality has been projected to rise another 55% by 2030 due to the obesity epidemic [3].

In addition, over the last 2 decades the evidence from important clinical trials have changed standards of treatment for low-risk and low–intermediate-risk EC (PORTEC 1, and GOG 99) [4,5], high–intermediate-risk EC (PORTEC 2 and ASTEC) [6,7], and high-risk EC (PORTEC 3, GOG 249, and GOG 258) [8–11]. Despite those advances, treatment failure occurs in approximately 10–15% of patients with early stage EC. Although non-endometrioid variants, such as serous and clear cell carcinomas, comprise <10% of all diagnoses, they account for a disproportionately high number of EC recurrences and cancer-related deaths [12]. However, the majority of treatment failures and recurrences occur in endometrioid EC type (EEC) and prognosis remains poor for these women, with

the exception of patients with isolated vaginal recurrence [12,13]. Thus, identifying patients who might benefit from additional surveillance and treatment to prevent recurrence and reduce mortality from this disease would be of great value.

The Cancer Genome Atlas (TCGA) identified molecular features that were found to categorize EC tumors into different levels of risk [14,15]. Later, the Post-Operative Radiation Therapy in Endometrial Carcinoma (PORTEC) Study Group included some of these molecular features to design its latest trial, 4a (NCT03469674). In this trial, standard adjuvant treatment with vaginal brachytherapy for women with high–intermediate-risk EC, is compared with individualized adjuvant treatment based on a molecular-integrated risk profile [16]. However, with this molecular assessment, almost 60% of patients presented a 'no specific molecular profile' (NSMP) [17]. Prior studies have also used clinical and pathological characteristics to stratify risk for recurrence and to inform adjuvant treatment [18,19]. Despite these studies, to date, there is no standard, validated, and accurate model that assesses individual risk of recurrence for patients with EC. Previous attempts reported accuracies between 60 and 73%, or area under the curve (AUC) around 80% [18,20]. Better models are needed to identify 15% of those patients with early stage EC that are going to need adjuvant treatment.

We hypothesize that integration of clinical and genomic data will improve prediction models of recurrence in EEC. The objective of this pilot study is to create models with clinical and genomic data that will discriminate patients with EEC at risk of recurrence from disease. We validated these models in independent datasets (TCGA) and machine-learning analytical platforms.

2. Results

The flow of included patients are depicted in Figure 1 and included patients characteristics in Table 1.
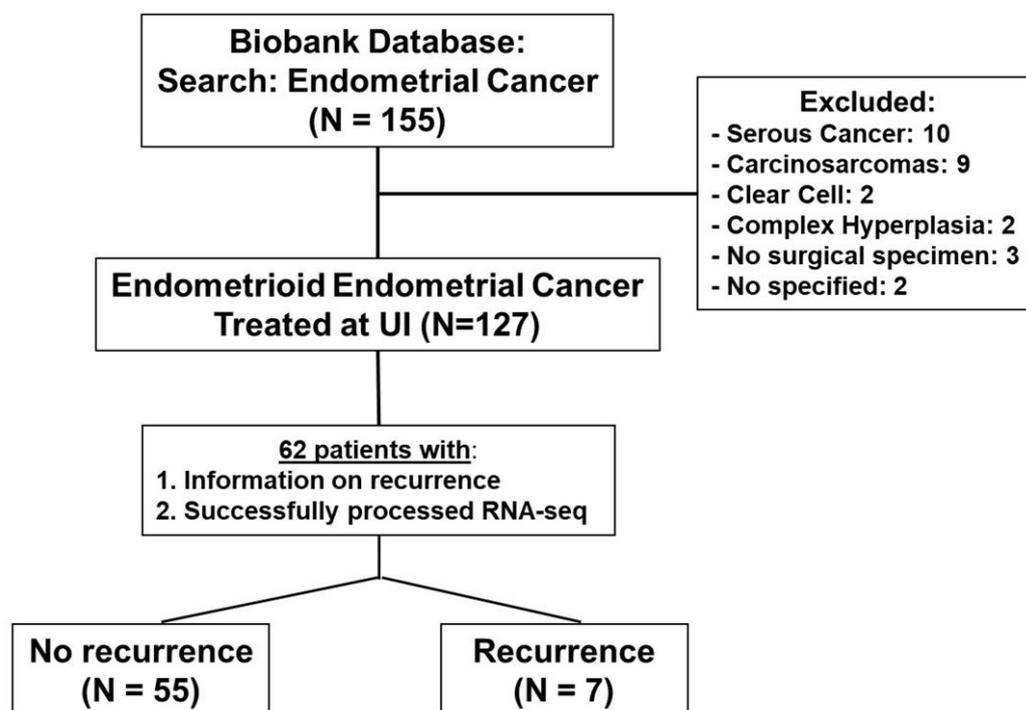

Figure 1. Flow of included patients in analysis. Of the initial 155 endometrial cancers available in the UI Biobank, 127 were confirmed to be of endometrioid histology. The rest were excluded from the study. A total of 62 patients had annotated follow-up with detailed recurrence information and with quality RNA for RNA sequencing (RNA-seq).

Table 1. Clinical patient characteristics. These are the baseline variables determined at treatment completion and included in the analysis.

		Recurrent (N = 7)	Non-Recurrent (N = 55)	*p*-Value
Age	(average)	61	61	0.983
BMI	(average)	35.2	37.1	0.627
**Charlson Index ** **				0.720
	Low (1–3)	0	9	
	Medium (4–6)	7	37	
	High (>6)	0	5	
Personal History	DM	1	13	0.582
	Heart atack	0	1	0.995
	CHF	0	1	0.995
	Stroke	0	2	0.996
	Pulmonary disease	0	8	0.994
	Other cancers	0	8	0.994
Grade				0.731
	1	2	23	
	2	4	20	
	3	1	10	
Lymphovascular involvement				0.208
	No	4	42	
	Yes	3	11	
MI	(average)	69.3	36.6	0.029 *
Cytology				0.450
	No	5	50	
	Yes	1	4	
Stage				0.009 *
	I	1	43	
	II	1	3	
	III	5	6	
	IV	0	3	
Adjuvant Radiation (any type)				0.150
	No	3	39	
	Yes	4	16	

Tobacco use, pre-operative Hgb, creatinine, WBC, albumin, type of surgery -open or minimally invasive-, surgical complications (including blood loss), and length of stay were not significantly different. BMI: body mass index; MI: myometrial invasion. * Statistically significant with *p*-value < 0.05. ** Charlson Comorbidity Index is a measure of the prognostic burden of all associated morbidities to predict mortality and is the most validated measure of the prognostic impact of multiple chronic illnesses (www.charlsoncomorbidity.com (accessed on 30 July 2022)).

2.1. Creation of Prediction Models of EEC Recurrence

After RNA extraction, sequencing and analysis, we determined a series of genomics features that were used for the prediction analysis: (a) from the extracted transcriptome: gene, long non-coding RNA (lncRNA) and single exon expression; (b) we determined genomic variation, including single nucleotide variation (SNV), copy number variation by gene (CNV), and copy number variation by chromosomal region. Additionally, we identified structural variation (SV), including fusion transcripts (FT), retained introns (RI), novel exon/junction (NEJ), and unknown or previously not reported SV (UNK). After the univariate analysis with cross-validation of all genomic features, we found those characteristics that were more informative of EC recurrence (Figure 2). These significant features were later introduced in prediction models of recurrence.
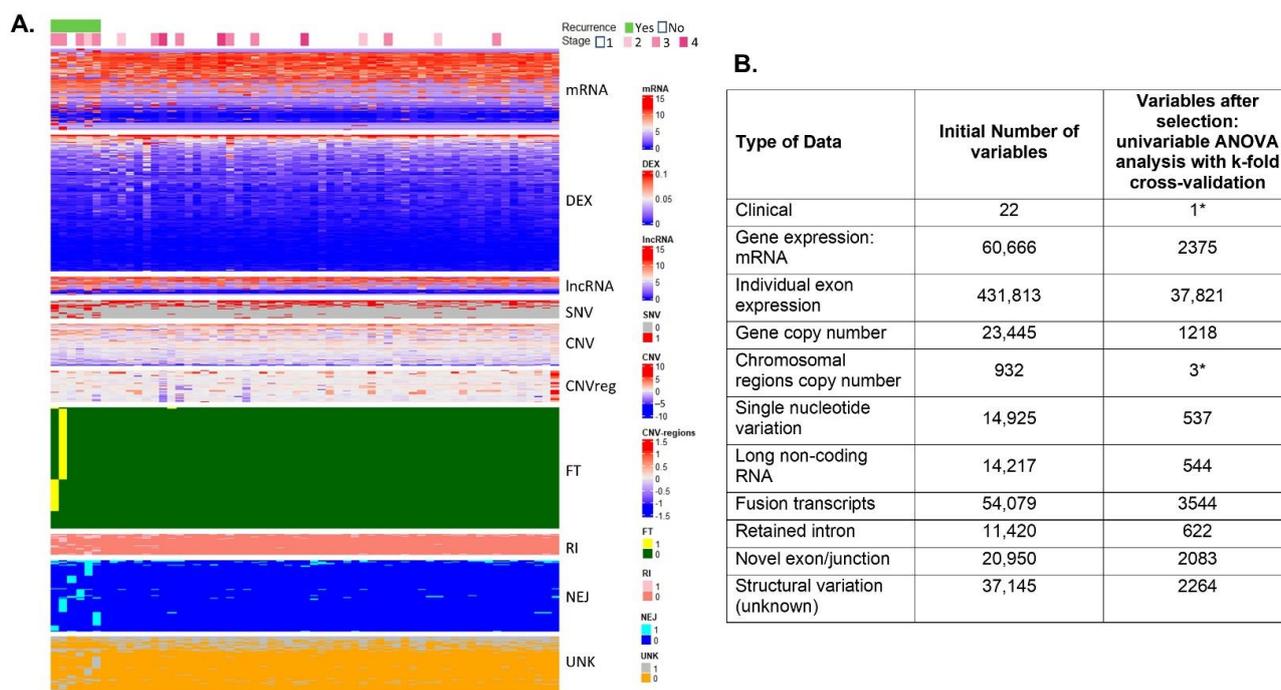
Figure 2. Included patients. (**A**) Heatmap of selected variables after univariate ANOVA analysis. Representation of the significant variables after univariate analysis ($p < 0.05$) for different types of genomic data. Recurrent cases are at the left side of the panel (under the light green bar); non-recurrent cases are at the right (white bar). Of the 22 clinical features introduced in the lasso analysis, only stage was informative for recurrence (upper part of the panel, color coded from 1 to 4). Transcriptome: DEX: exon expression; lncRNA: long non-coding RNA; Exp: gene expression. Genomic variation: SNV: single nucleotide variation; CNV: gene copy number by gene; CNVreg: copy number by chromosomal region; Structural variation: FT: Fusion transcripts; RI: Retained intron; NEJ: Novel exon/junction; U-SV: Unknown SV. At the right side of the panel are the labels and the color-coded range of values for all genomic variables. (**B**) Variable selection and variables after univariate analysis. To reduce the number of variables, we used univariate analysis of all data with ANOVA to select the variables that were more informative for prediction of response, with a *p*-value < 0.05 (3rd column). * Lasso regression was performed directly with no pre-reduction with ANOVA because the smaller number of variables in two types of data: clinical data and copy number by chromosomal region. Graphics were generated with R package *ComplexHeatmap* [21].

Next, we built prediction models for recurrence. Initially we constructed them with only one feature. Then, we made models with two and three different sets of variables. Adding four or more variables did not improve prediction models and added complexity to the system (Figure 3). In total, we built over 170 models to predict EEC recurrence (Supplementary Figures S1 and S2). In Figure 3, we represented the 30 models with one or two types of variables with the best performance measured by AUC (Figure 3A), and the 30 best performing models with three types of variables (Figure 3B). If we consider that clinical data is the best way to assess risk of recurrence up to now, potentially superior models are those with a performance, measured by AUC, over 0.75, which is the basic clinical model performance (Figure 3A in lighter blue).

Notably, all best models included lncRNA data. Moreover, the model with only lncRNA had an AUC of 0.9 (95% CI 0.75–1.0) and adding more clinical or genomic data to the model did not improve the performance (Figure 3). No matter how many types of variables were added to lncRNA data, the multivariate regression lasso model ended up with the same five lncRNAs: ENSG00000274840, ENSG00000240137, ENSG00000250137, ENSG00000253622, and ENSG00000258240. So, comparing all models, the simplest model with only lncRNA turned out to have one of the best performances, with an AUC of 0.9,

and the same final five variables at the end of the lasso analysis than more complicated models with more types of data (Figure 4). This simplest, best-performing model would be the one with potential to improve the only clinical model.
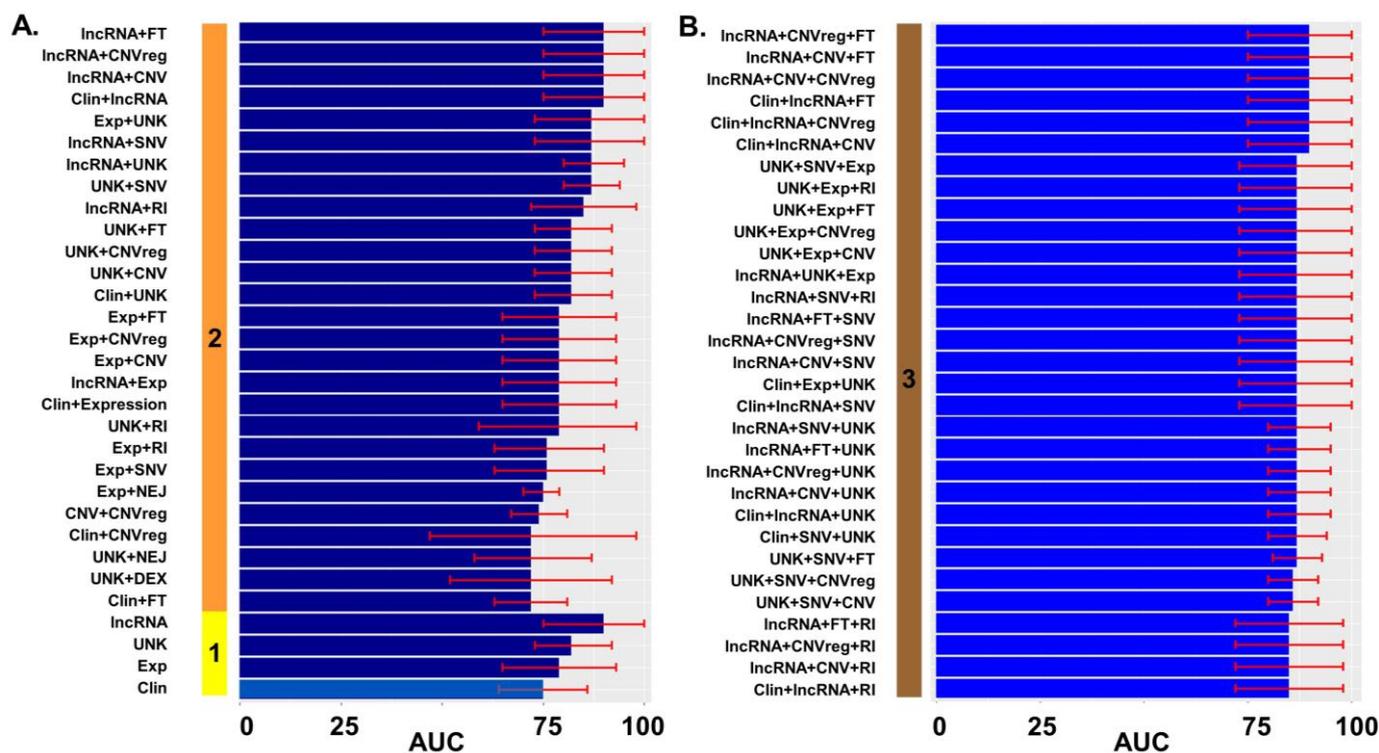

Figure 3. Performance of prediction models of EEC recurrence. (**A**) The solid vertical bar represents the number of types of data: 1 (yellow): only one variable was included in the model; 2 (orange): combination of two types of variables; (**B**) The solid vertical maroon bar represents the combination of three types of variables. Different performances on both panels are displayed in ascending order. The *x* axis is AUC as a percentage (0–100%). The red error mark displays the 95% confidence interval (CI). Overall, 72 models with different combinations of data were tested. We only displayed the best (**A**) 30 models for combinations of one and two variables and (**B**) 30 best models for combinations of three types of variables. Exp: gene expression; DEX: exon expression; lncRNA: long non-coding RNA; SNV: single nucleotide variation; CNV: gene copy number by gene; CNVreg: copy number by chromosomal region; FT: Fusion transcripts; RI: Retained intron; NEJ: Novel exon/junction; UNK: Unknown SV. Graphics were generated with R package *ggplot*.

2.2. *Validation of Prediction Models of EEC Recurrence*

Validation of the best performing model was performed with different analytical platforms and with an independent dataset (TCGA).

2.2.1. Validation of Models with Machine Learning (ML)

We validated the best model with two different ML analytical platforms. The first one used *TensorFlow,* and we tested the model with (Figure 5A) and without (Figure 5B) FIGO Stage. The performance of both were excellent, with AUC of 1.00, and accuracies over 85%. For the second ML platform we used the suite MATLAB and its ML App. The App has over 30 ML methods than can be used in parallel to assess the accuracy of a model. Again, both AUC and accuracies were excellent (100%) for the model with lncRNA data and FIGO Stage (Figure 5C) and for the one with only five lncRNAs (Figure 5D).
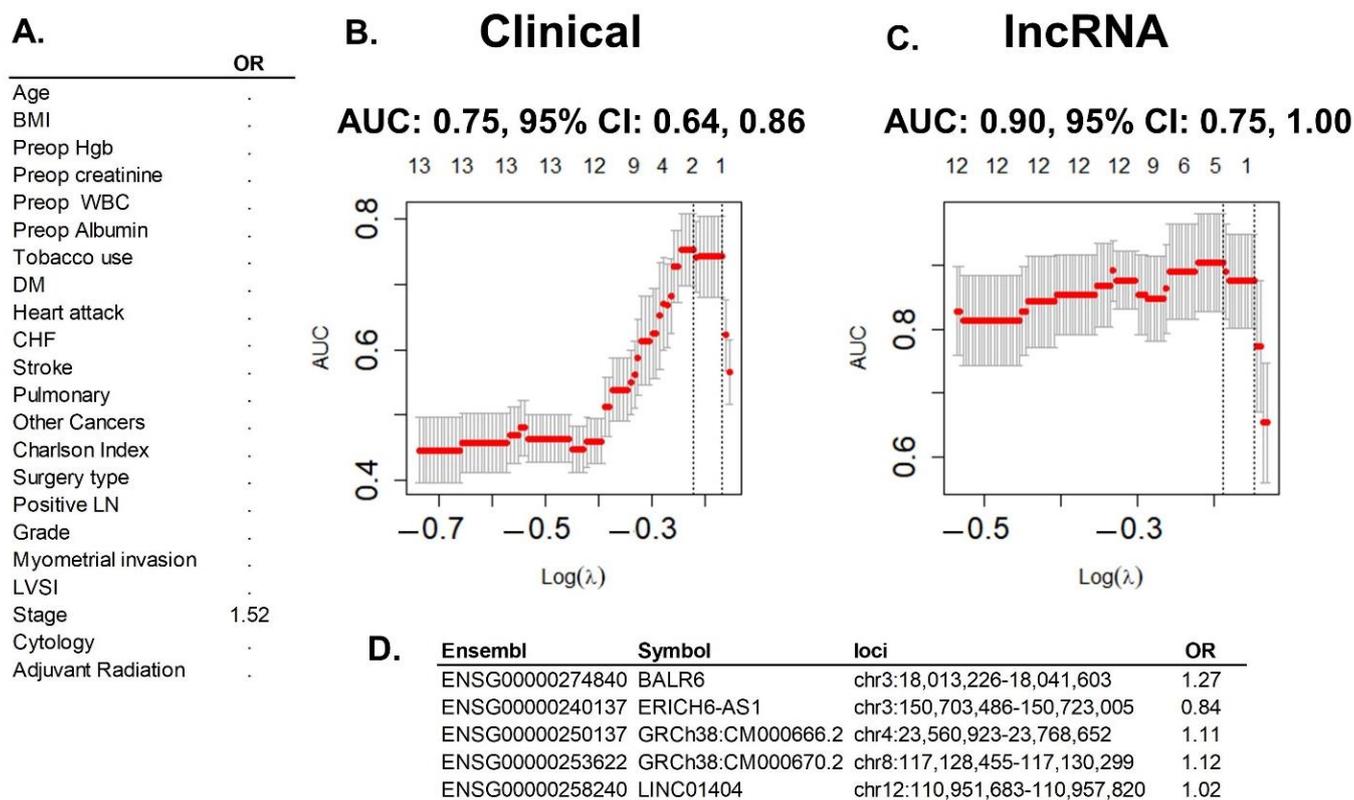
A.

	OR
Age	.
BMI	.
Preop Hgb	.
Preop creatinine	.
Preop WBC	.
Preop Albumin	.
Tobacco use	.
DM	.
Heart attack	.
CHF	.
Stroke	.
Pulmonary	.
Other Cancers	.
Charlson Index	.
Surgery type	.
Positive LN	.
Grade	.
Myometrial invasion	.
LVSI	.
Stage	1.52
Cytology	.
Adjuvant Radiation	.

B. **Clinical**

AUC: 0.75, 95% CI: 0.64, 0.86

C. **lncRNA**

AUC: 0.90, 95% CI: 0.75, 1.00

D.

Ensembl	Symbol	loci	OR
ENSG00000274840	BALR6	chr3:18,013,226-18,041,603	1.27
ENSG00000240137	ERICH6-AS1	chr3:150,703,486-150,723,005	0.84
ENSG00000250137	GRCh38:CM000666.2	chr4:23,560,923-23,768,652	1.11
ENSG00000253622	GRCh38:CM000670.2	chr8:117,128,455-117,130,299	1.12
ENSG00000258240	LINC01404	chr12:110,951,683-110,957,820	1.02

Figure 4. Best prediction model and comparison with clinical model. (**A**) The lasso multivariate regression model with clinical variables: out of the two clinical variables included in the model, only Stage remained as informative for prediction of EEC recurrence after the analysis, with an increased risk for recurrence as stage increases. (**B**) Graphic representation of the clinical lasso analysis: superior margin reflects number of variables; left margin reflects performance of the model measured in AUC (area under the curve); lower margin reflects lambda tunning parameter chose by cross-validation to optimize the model. The optimized AUC was 0.75 (95% CI: 0.68, 0.86), between the dotted lines. (**C**) Graphic representation of the lncRNA data lasso analysis (same margins and design as before): optimized AUC of 0.9 (95% CI: 0.75, 1.00). (**D**) The lasso multivariate regression model with lncRNA data: out of the 544 clinical variables included in the model, five single lncRNAs remained as informative for prediction of EEC recurrence. Four of them increased risk (OR > 1) and one protected from recurrence (OR < 1). Graphics were generated with R package glmnet.

In summary, in this validation analysis, the best predictive model for EEC recurrence seems to be robust enough throughout different analytical methods and platforms. Additionally, adding clinical data (FIGO stage) to lncRNA data did not improve the performance of the validation model.

2.2.2. Validation of Models with TCGA Dataset

Finally, we validated our best model with an independent, publicly available dataset, TCGA, with 406 EEC patients, 346 non-recurrent, and 60 recurrent. Patients' characteristics of this dataset were similar to our study population and can be reviewed in the Supplementary Figure S3. After extracting lncRNA data from the original BAM files, we selected the five lncRNAs that were driving the prediction model and tested them in the *TensorFlow* and MATLAB ML platforms. It had an accuracy of 78% and 86%, respectively, with also good AUC of 0.68 and 0.78 (Supplementary Figure S4). Therefore, our best model also performed well in an independent dataset (TCGA).
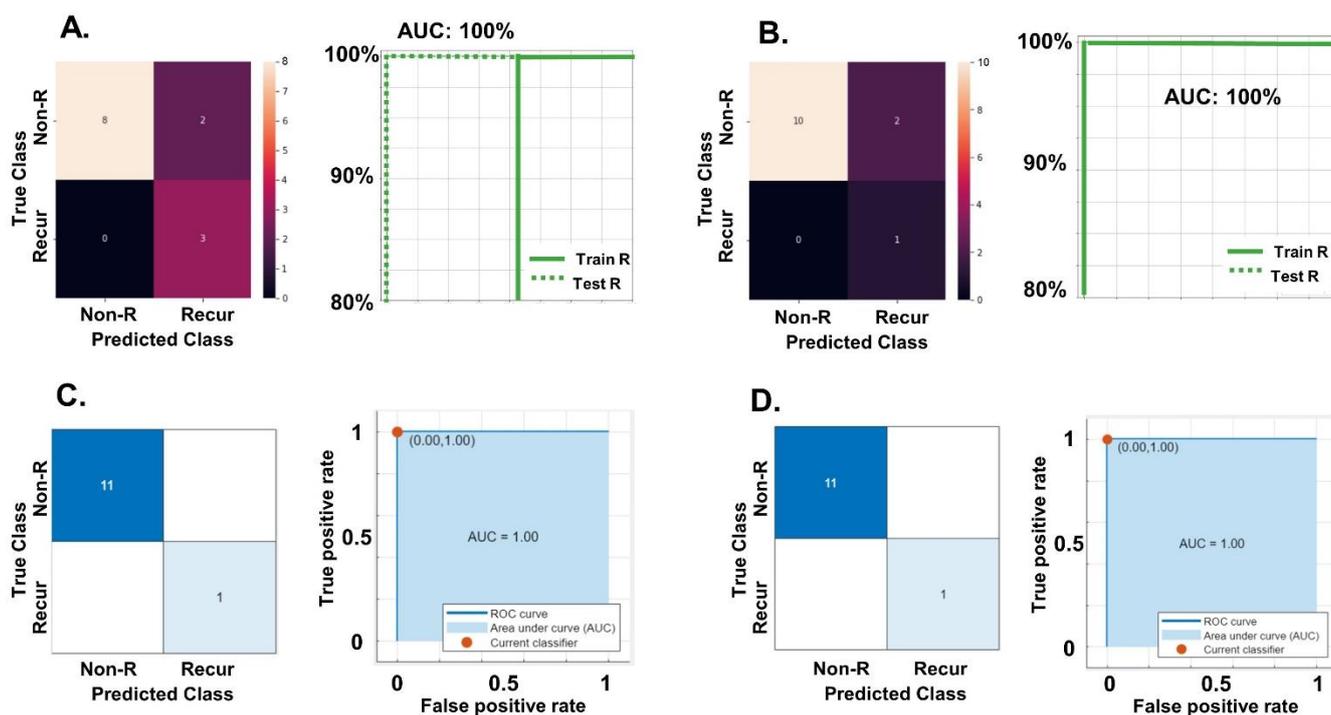
Figure 5. (**A**) Validation of the best model with lncRNA data and FIGO Stage clinical variable (the most informative in the clinical model). These are the results after training and validating the model with TensorFlow in 75% of the samples, and then performing testing on the remaining 25% of the data. We are showing the testing results. The left panel shows the confusion matrix representing the true (True Class) versus the predicted values (Predicted Class). The right panel is an ROC graphic: true positives in the *x* axis, false positives in the *y* axis, and AUC results. Train R: results of unbalanced (or re-sampling) model training; Test R: results of re-sampling model testing. AUC of 1.00 and accuracy of 0.92. Recur: recurrent; Non-R: non recurrent. (**B**) Validation of the best model with only lncRNA data and no clinical variables. As before, we are showing testing results in 25% of the data, after training and validating have been performed with TensorFlow. Left and right panels are as before. AUC of 1.00 and accuracy of 0.85. (**C**) Validation of the best model with lncRNA data and FIGO Stage clinical variable performed with MATLAB platform. We are showing testing results in 20% of the data, after training and validating have been performed. MATLAB offers over 30 methods for its machine learning (ML) App. In four of them the accuracy of testing was 100%, as shown in the graphic. Specifically, this is the coarse Gaussian SVM (support vector machines) method. Left and right panels are as before. AUC of 1.00 and accuracy of also of 1. (**D**) Validation of the best model with only lncRNA data and no clinical variables with MATLAB platform. Parameters are as before. This time, most methods had a testing accuracy of 100%. Showing the linear SVM method. Left and right panels are as before. AUC of 1.00 and accuracy of also of 1.

3. Discussion

Our pilot study found that all best prediction models of EEC recurrence contained lncRNA features, and specifically five lncRNAs. The simplest, best performing model contained only five lncRNAs and was as accurate as more complex models. Furthermore, the lncRNA model was superior to the model with only clinical data (AUC of 0.9 versus 0.75, respectively) and with a 95% CI that reached 1.0. If these results were to be validated in future studies this would result in an accurate and robust model that could discriminate which EEC patients would be at risk of initial treatment failure at the time of surgery. This would leave healthcare professional with plenty of time to design alternative adjuvant treatments to prevent these outcomes.

As we hypothesized previously, integration of clinical and genomic data improves prediction models of EEC recurrence, with AUC performances of 0.9, and CI reaching 1.00.

Integration of complex data is a difficult task, but the results could be very valuable [22,23]. It requires a constant dialogue between basic scientists, clinicians, statisticians and bioinformaticians to build models that are clinically and scientifically meaningful [24]; models that predict clinically significant outcomes that can be actioned upon. Our goal was to identify EEC patients at risk of recurrence who might benefit from additional surveillance and treatment to prevent relapse and reduce mortality from this disease. Our prediction models achieved that goal.

As noted initially, EC trials incorporated pathological prognosticators to determine postoperative radiation, as initial attempts to individualized therapy [4–11,25]. However, treatment failures still seemed to burden low-risk patients determined by pathology, though the percentage was superior in high-risk women [12,13]. Molecular profiling of EC promised improvement in risk assessment and treatment selection, especially after the TCGA initiative [14,15]. TCGA described four groups that had different molecular features: POLE ultra-mutated, microsatellite-instability-hypermutated, copy-number-low, and copy-number-high EC. Later, these groups were modified to make the molecular determination more feasible and affordable—the Proactive Molecular Risk Classifier or *ProMisE* [26]. The resulting groups seemed to correlate well with disease prognosis [17,26], and have been used to design new EC trials [16]. However, several questions remain to be addressed. The first one is the high number of unclassifiable EC, as much as 59% by one of these studies [17], and whether pathology prognostic factors should be applied to these cases. Additionally, when reviewing the prediction performance of these molecular features, prediction of recurrence for all models varied from 60–75% [27], which are not superior to clinical models [18,20]. Recent ESGO/ESTRO/ESP guidelines for EC management [28] stated that molecular classification could impact clinical management, especially in cases with high-grade/high-risk disease. However, they recognized that the molecular classifier is not perfect, and there is room for improvement for those patients with low-risk and/or unclassifiable molecular features. This is where our prediction model could help, in EEC patients with seemingly low-risk disease and unclassifiable molecular features, which are the majority.

LncRNAs have regulatory functions [29,30]. They participate in epigenetic regulation, maintain chromatin structure, and modulate transcription [30,31]. There is increasing reporting of the function that lncRNA have in the development and progression of EC [32]. The development of EC is a complicated biological process and lncRNAs may act as oncogenes or tumor suppressors. Their expression may contribute to EC transformation and the subsequent progression. Gene expression experiments have previously demonstrated that a large number of lncRNA expression is altered in EC [33]. Therefore, it is not surprising that some lncRNAs were selected in the univariate analysis because of their difference in expression between recurrent and non-recurrent samples. Furthermore, some of the lncRNAs present in our best model have been described previously in several cancers, ENSG00000274840, ENSG00000240137, and ENSG00000253622 [34–36], and ENSG00000250137 have been associated with increased BMI [37]. Our best model may be reflecting the underlying biological characteristics of the recurrent EEC phenotype.

The strengths of our study include our use of a comprehensive genomic characterization of EEC samples, including genomic variation and structural variation to determine the best prediction model of EEC recurrence. Additionally, we used proven statistical methods that employed internal validation with cross-validation for feature selection to avoid model over-fitting. Finally, we performed external validation with diverse analytical platforms, including the use of ML, and different and independent datasets of EEC (TCGA) analyzed with identical methods and software, also described previously [38,39]. It should be mentioned that genomic variation, specifically SNV and CNV, are better determined with DNA sequencing. Extracting genomic variation from RNAseq is an estimation of the real somatic variation (~75% of the variants) [40,41], but it served the purpose of this study and prevented cost scalation. In the end, no SNV or CNV were in the best model of EEC recurrence prediction.

One of the potential limitations of the study is the relativity low number of recurrent EEC that have complete clinical and genomic data in our dataset. Recurrence of EEC range between 10–15% [12], and even larger genomic databases, such as TCGA, have a low and unbalanced number of recurrent cases that may affect any prediction model. To adjust for this low and unbalanced number of recurrent cases, we validated all models with ML analytics that specifically account for this issue by resampling data during model training, validation, and testing [42]. Our best model is the result of a well-studied EEC population [43–46] that has been well annotated and followed over the years. We are a state-sponsored University, which receives and serves the vast majority of patients with gynecological cancer in the State of Iowa. However, as the population of Iowa is predominantly white, 92% of patients included in the study were white. The lack of diversity of our selected subjects is a potential limitation of our analysis that may influence the generalizability of the study to other states where there is more diversity. In a previous study comparing our population with TCGA population, we identified differences in the admixture of both cohorts [46]. These differences may have an effect on the performance and validation of the model outside Iowa. Other limitation of the study may arise from inherited biases of retrospective studies, mainly recall biases. Due to losses in follow-up, some of the recurrences could be under-reported. However, we are more concerned with the surveillance of TCGA EEC patients that may influence the performance of validation analysis. Finally, we have to be aware of potential overfitting of our recurrence prediction model, either to our own population and/or to our own data. To avoid this issue, the performance of the model has to be confirmed in a new prospective set of diverse EEC patients where the phenotype is known confidently. Until then, the model should not be used clinically.

4. Materials and Methods

4.1. Study Design

We performed a retrospective, single institution, case–control study in which we included 62 patients with EEC from 1991 to 2010 available in our biobank with pre-operative and intra-operative clinical data. RNA was extracted from tumor specimens and processed as detailed below to obtain the necessary genomic data. Clinical and genomic data were then combined to create predictive models using statistical learning to identify criteria which accurately predicted recurrence for EC patients.

4.2. Ethics and Tissue Procurement

Tissue samples and clinical outcome data were obtained from the Department of Obstetrics and Gynecology Biobank (IRB, ID#200209010), which is part of the Women's Health Tissue Repository (WHTR, IRB, ID#201804817). All tissues archived in the Gynecologic Oncology Biobank (herein termed Biobank) were originally obtained from adult patients under informed consent in accordance with University of Iowa IRB guidelines. Tumor samples were collected, reviewed by a board-certified pathologist, flash-frozen, and then the diagnosis was confirmed in paraffin. All experimental protocols were approved by the University of Iowa Biomedical IRB-01.

4.3. Clinical Data Procurement

Clinical data was extracted from the electronic medical record. Table 1 summarizes the baseline clinical and pathologic characteristics. Only data that were available by the end of initial treatment were used in the development of predictive models. Pre-operative characteristics included age at diagnosis, body mass index (BMI), pre-operative hemoglobin, serum creatinine, albumin, comorbidities (coronary artery disease, diabetes mellitus, congestive heart disease, history of cardiovascular accident, tobacco use), and Charlson morbidity index. Intraoperative characteristics included type of surgery (laparoscopic, robotic, laparotomy, vaginal), operative time, and estimated blood loss. Post-operative characteristics extracted included final pathology diagnosis, disease stage, estrogen and progesterone

receptor status, surgical complications, adjuvant therapy (including types of radiation therapy), and recurrence of disease.

For the purposes of this study, we defined disease recurrence as EEC diagnosed in any location after completing treatment and a subsequent period with no evidence of disease. Of the 62 EEC patients with clinical and genomic information included in the study, 7 had a recurrence of disease. Two patients had an advanced stage with persistence of disease after initial treatment and, by the study definition, they were considered as non-recurrent. All patients recurred within 5 years; 86% of them experienced a recurrence within 2 years of initial treatment. Differences between clinical variables among both study groups were assessed by logistic regression (significance at *p*-value < 0.05).

4.4. Genomic Analysis

4.4.1. Included Subjects

A cohort of 127 patients diagnosed with EEC at UI was assembled under approval by the Institutional Review Board (IRB# 201607815). Only patients with a confirmed EEC diagnosis, with clinical follow-up and biological specimens with quality RNA for sequencing, were included in the study. The flow diagram in Figure 1 summarizes patients included in this study.

4.4.2. RNA Isolation and Sequencing

RNA was then isolated from these tumor specimens. RNA extraction, processing and sequencing have been described previously [39,43]. In brief, total cellular RNA was extracted from primary tumor tissue using the mirVana (Thermo Fisher, Waltham, MA, USA) RNA purification kit. The RNA yield and quality were assessed with Trinean Dropsense 16 spectrophotometer and Agilent Model 2100 bioanalyzer. RNA quality was determined to be adequate if the sample had an RNA integrity number (RIN) of 7.0 or greater. Samples that were of adequate quality were then sequenced. 500ng of RNA was quantified by Qubit measurement (Thermo Fisher). RNA was then converted to cDNA and ligated to sequencing adaptors with Illumina TriSeq stranded total RNA library preparation (Illumina, San Diego, CA, USA). cDNA samples were then sequenced with the Illumina HiSeq 4000 genome sequencing platform using 150 bp paired-end SBS chemistry. All sequencing was performed at the Genome Facility at the University of Iowa Institute of Human Genetics (IIHG).

4.4.3. Data Preprocessing

STAR was used to align the RNAseq reads to the human reference genome (version hg38) [47]. We then created BAM files after alignment. *FeatureCount* was used to measure gene expression [48]. The *DESeq2* package was used to import, normalize, and prepare the gene counts for analysis [49]. *ENSEMBL* was used to annotate single exons within the gene expression alignment analysis. Exon expression was then evaluated using the *DEXSeq* package [50]. BAM files for each sample were used to estimate SNV discovery and base-calling against the human genome reference utilizing *SAMtools* and *BCFtools* for sorting and indexing. Results were filtered for duplicates, known non-synonymous single-nucleotide variants, and synonymous variants and then annotated with *ANNOVAR*. Gene CNV were estimated using *SAMtools* and *superFreq* [40]. BAM files were then used to identify lncRNA, as described previously [51,52]. Lastly, fusion transcripts were determined using the *STAR-Fusion* suite from fastq files [53]. Supplementary Figure S5 depicts each program used for RNA processing and the identification of various genomic components.

4.5. Statistical Analysis

Most genomic data were used as continuous variables, except SNV and structural variation features, which were used as dichotomous variables. To select those variables most informative for prediction of EEC recurrence, we used univariate analysis with ANOVA (*p* < 0.05) and cross-validation with 10 replicates for each fold, as implemented by

the *caret* R package and detailed previously [39]. Significant predictive variables were then used in a multivariate *lasso* regression prediction model (statistical learning). Thus, poorly annotated variables were removed from model construction.

4.5.1. Creation of Prediction Models of EEC Recurrence with Statistical Learning

Significant variables from the univariate analysis were then incorporated into multivariate *lasso* (least absolute shrinkage and selection operator) regression prediction models of recurrence. Initial models included only significant variables from one category of clinical or genomic data (i.e., lncRNA expression, gene expression, CNV, etc.). Variables were then progressively combined to create more complex prediction models. Multivariate prediction models were fit with *lasso* as implemented in the *glmnet* R package [54], and detailed previously [38,39]. Performances of prediction models were measured with area under the receiver operating characteristics curve (AUC) and their respective 95% confidence intervals (CI) and estimated with 1000 replicates of ten-fold cross-validation to avoid over-fitting. Bias-corrected and accelerated bootstrap CIs were computed for each model. AUC of 0.5 indicates no predictive ability of a model and 1.0 represents perfect predictive performance.

4.5.2. Validation of Predictive Models with Machine-Learning Methods

For validation of the best prediction models of recurrence in a machine-learning platform we used *TensorFlow* [55] in a *Jupyter* notebook with a *Keras* application programming interface (API) [42]. *TensorFlow* code was modified from a tutorial (found here: https://www.tensorflow.org/tutorials (accessed on 25 October 2022)). Training, validating, and testing were performed to account for weights of the outcomes as well as for unbalanced data (mainly for complete vs. optimal patients). Additionally, we validated the best prediction models in *MATLAB* machine learning app, where there are over 20 classifier methods. Model from UI were validated in this new analytical platform and later was validated in EEC TCGA dataset.

4.5.3. Validation of Predictive Models with Independent Data, TCGA

Data from TCGA dataset for endometrial EC were downloaded from the National Cancer Institute (NCI) database in accordance with TCGA Human Subject Protection and Data Access Policies, adopted by the NCI and the National Human Genome Research Institute (NHGRI). Data were downloaded with the NCI database of genotypes and phenotypes approval (dbGaP#16003) as previously described [43]. Patients with non-endometrioid histology were excluded. Clinical and molecular data were obtained from 406 patients diagnosed with EEC, of which 60 experienced recurrence of disease as defined above (Supplementary Table S4. Original downloaded BAM files were then used to identify lncRNAs, as described previously [51,52].The best-performing parameters were used to fit a final score of that model to the entire TCGA cohort [39]. Performances measured by AUC between 0.8–0.9 were considered 'very good'; performances between 0.9–1 were considered 'excellent'.

5. Conclusions

Prediction models containing lncRNA features have better performance, measured by AUC, than models with clinical data alone. These models must be validated in prospective manner and different populations before their use in clinical settings.

Author Contributions: Conceptualization, J.G.-B. and E.D.; methodology, J.G.-B., B.J.S. and E.D.; validation, E.D. and J.G.-B.; formal analysis, J.G.-B.; investigation, E.D. and S.G.; resources, J.G.-B., E.D., K.K.L. and M.J.G.; data curation, E.D., M.E.M. and J.G.-B.; writing—original draft preparation, J.G.-B.; writing—review and editing, J.G.-B., E.D., M.E.M., K.K.L., M.J.G., S.G. and D.D.B.; visual-

References

1. Siegel, R.L.; Miller, K.D.; Fuchs, H.E.; Jemal, A. Cancer statistics, 2022. *CA Cancer J. Clin.* **2022**, *72*, 7–33. [CrossRef]
2. Kalampokas, E.; Giannis, G.; Kalampokas, T.; Papathanasiou, A.A.; Mitsopoulou, D.; Tsironi, E.; Triantafyllidou, O.; Gurumurthy, M.; Parkin, D.E.; Cairns, M.; et al. Current Approaches to the Management of Patients with Endometrial Cancer. *Cancers* **2022**, *14*, 4500. [CrossRef] [PubMed]
3. Sheikh, M.A.; Althouse, A.D.; Freese, K.E.; Soisson, S.; Edwards, R.P.; Welburn, S.; Sukumvanich, P.; Comerci, J.; Kelley, J.; LaPorte, R.E.; et al. USA endometrial cancer projections to 2030: Should we be concerned? *Future Oncol.* **2014**, *10*, 2561–2568. [CrossRef] [PubMed]
4. Creutzberg, C.L.; van Putten, W.L.; Koper, P.C.; Lybeert, M.L.; Jobsen, J.J.; Warlam-Rodenhuis, C.C.; De Winter, K.A.; Lutgens, L.C.; van den Bergh, A.C.; van de Steen-Banasik, E.; 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* **2000**, *355*, 1404–1411. [CrossRef] [PubMed]
5. Keys, H.M.; Roberts, J.A.; Brunetto, V.L.; Zaino, R.J.; Spirtos, N.M.; Bloss, J.D.; Pearlman, A.; Maiman, M.A.; Bell, J.G.; Gynecologic Oncology, G. A phase III trial of surgery with or without adjunctive external pelvic radiation therapy in intermediate risk endometrial adenocarcinoma: A Gynecologic Oncology Group study. *Gynecol. Oncol.* **2004**, *92*, 744–751. [CrossRef] [PubMed]
6. Nout, R.A.; Smit, V.T.; Putter, H.; Jurgenliemk-Schulz, I.M.; Jobsen, J.J.; Lutgens, L.C.; van der Steen-Banasik, E.M.; Mens, J.W.; Slot, A.; Kroese, M.C.; 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* **2010**, *375*, 816–823. [CrossRef] [PubMed]
7. Barton, D.P.; Naik, R.; Herod, J. Efficacy of systematic pelvic lymphadenectomy in endometrial cancer (MRC ASTEC Trial): A randomized study. *Int. J. Gynecol. Cancer: Off. J. Int. Gynecol. Cancer Soc.* **2009**, *19*, 1465. [CrossRef] [PubMed]
8. de Boer, S.M.; Powell, M.E.; Mileshkin, L.; Katsaros, D.; Bessette, P.; Haie-Meder, C.; Ottevanger, P.B.; Ledermann, J.A.; Khaw, P.; D'Amico, R.; 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.* **2019**, *20*, 1273–1285. [CrossRef]

9. Randall, M.E.; Filiaci, V.; McMeekin, D.S.; von Gruenigen, V.; Huang, H.; Yashar, C.M.; Mannel, R.S.; Kim, J.W.; Salani, R.; DiSilvestro, P.A.; et al. Phase III Trial: Adjuvant Pelvic Radiation Therapy Versus Vaginal Brachytherapy Plus Pacli-taxel/Carboplatin in High-Intermediate and High-Risk Early Stage Endometrial Cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **2019**, *37*, 1810–1818. [CrossRef]

10. Matei, D.; Filiaci, V.; Randall, M.E.; Mutch, D.; Steinhoff, M.M.; DiSilvestro, P.A.; Moxley, K.M.; Kim, Y.M.; Powell, M.A.; O'Malley, D.M.; et al. Adjuvant Chemotherapy plus Radiation for Locally Advanced Endometrial Cancer. *N. Engl. J. Med.* **2019**, *380*, 2317–2326. [CrossRef]

11. Simon, R.; Lam, A.; Li, M.C.; Ngan, M.; Menenzes, S.; Zhao, Y. Analysis of gene expression data using BRB-ArrayTools. *Cancer Inf.* **2007**, *3*, 11–17.

12. Del Carmen, M.G.; Boruta, D.M., 2nd; Schorge, J.O. Recurrent endometrial cancer. *Clin. Obs. Gynecol.* **2011**, *54*, 266–277. [CrossRef] [PubMed]

13. Restaino, S.; Dinoi, G.; La Fera, E.; Gui, B.; Cappuccio, S.; Campitelli, M.; Vizzielli, G.; Scambia, G.; Fanfani, F. Recurrent Endometrial Cancer: Which Is the Best Treatment? Systematic Review of the Literature. *Cancers* **2022**, *14*, 4176. [CrossRef] [PubMed]

14. Cancer Genome Atlas Research, N.; Kandoth, C.; Schultz, N.; Cherniack, A.D.; Akbani, R.; Liu, Y.; Shen, H.; Robertson, A.G.; Pashtan, I.; Shen, R.; et al. Integrated genomic characterization of endometrial carcinoma. *Nature* **2013**, *497*, 67–73. [CrossRef]

15. Alexa, M.; Hasenburg, A.; Battista, M.J. The TCGA Molecular Classification of Endometrial Cancer and Its Possible Impact on Adjuvant Treatment Decisions. *Cancers* **2021**, *13*, 1478. [CrossRef]

16. van den Heerik, A.; Horeweg, N.; Nout, R.A.; Lutgens, L.; van der Steen-Banasik, E.M.; Westerveld, G.H.; van den Berg, H.A.; Slot, A.; Koppe, F.L.A.; Kommoss, S.; 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 Off. J. Int. Gynecol. Cancer Soc.* **2020**, *30*, 2002–2007. [CrossRef]

17. Stelloo, E.; Nout, R.A.; Osse, E.M.; Jurgenliemk-Schulz, I.J.; Jobsen, J.J.; Lutgens, L.C.; van der Steen-Banasik, E.M.; Nijman, H.W.; Putter, H.; Bosse, T.; et al. Improved Risk Assessment by Integrating Molecular and Clinicopathological Factors in Early-stage Endometrial Cancer-Combined Analysis of the PORTEC Cohorts. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **2016**, *22*, 4215–4224. [CrossRef]

18. Versluis, M.A.; de Jong, R.A.; Plat, A.; Bosse, T.; Smit, V.T.; Mackay, H.; Powell, M.; Leary, A.; Mileshkin, L.; Kitchener, H.C.; et al. Prediction model for regional or distant recurrence in endometrial cancer based on classical pathological and immunological parameters. *Br. J. Cancer* **2015**, *113*, 786–793. [CrossRef]

19. Devor, E.J.; Miecznikowski, J.; Schickling, B.M.; Gonzalez-Bosquet, J.; Lankes, H.A.; Thaker, P.; Argenta, P.A.; Pearl, M.L.; Zweizig, S.L.; Mannel, R.S.; et al. Dysregulation of miR-181c expression influences recurrence of endometrial endometrioid adenocarcinoma by modulating NOTCH2 expression: An NRG Oncology/Gynecologic Oncology Group study. *Gynecol. Oncol.* **2017**, *147*, 648–653. [CrossRef]

20. Creutzberg, C.L.; van Stiphout, R.G.; Nout, R.A.; Lutgens, L.C.; Jurgenliemk-Schulz, I.M.; Jobsen, J.J.; Smit, V.T.; Lambin, P. Nomograms for prediction of outcome with or without adjuvant radiation therapy for patients with endometrial cancer: A pooled analysis of PORTEC-1 and PORTEC-2 trials. *Int. J. Radiat. Oncol. Biol. Phys.* **2015**, *91*, 530–539. [CrossRef]

21. Gu, Z.; Eils, R.; Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **2016**, *32*, 2847–2849. [CrossRef] [PubMed]

22. Falzone, L.; Scandurra, G.; Lombardo, V.; Gattuso, G.; Lavoro, A.; Distefano, A.B.; Scibilia, G.; Scollo, P. A multidisciplinary approach remains the best strategy to improve and strengthen the management of ovarian cancer (Review). *Int. J. Oncol.* **2021**, *59*, 1–14. [CrossRef] [PubMed]

23. Boron, D.; Zmarzly, N.; Wierzbik-Stronska, M.; Rosinczuk, J.; Mieszczanski, P.; Grabarek, B.O. Recent Multiomics Approaches in Endometrial Cancer. *Int. J. Mol. Sci.* **2022**, *23*, 1237. [CrossRef] [PubMed]

24. Emons, G.; Steiner, E.; Vordermark, D.; Uleer, C.; Bock, N.; Paradies, K.; Ortmann, O.; Aretz, S.; Mallmann, P.; Kurzeder, C.; et al. Interdisciplinary Diagnosis, Therapy and Follow-up of Patients with Endometrial Cancer. Guideline (S3-Level, AWMF Registry Number 032/034-OL, April 2018)—Part 2 with Recommendations on the Therapy and Follow-up of Endometrial Cancer, Palliative Care, Psycho-oncological/Psychosocial Care/Rehabilitation/Patient Information and Healthcare Facilities. *Geburtshilfe Frauenheilkd* **2018**, *78*, 1089–1109.

25. Creasman, W.T.; Morrow, C.P.; Bundy, B.N.; Homesley, H.D.; Graham, J.E.; Heller, P.B. Surgical pathologic spread patterns of endometrial cancer. A Gynecologic Oncology Group Study. *Cancer* **1987**, *60*, 2035–2041. [CrossRef]

26. Talhouk, A.; McConechy, M.K.; Leung, S.; Yang, W.; Lum, A.; Senz, J.; Boyd, N.; Pike, J.; Anglesio, M.; Kwon, J.S.; et al. Confirmation of ProMisE: A simple, genomics-based clinical classifier for endometrial cancer. *Cancer* **2017**, *123*, 802–813. [CrossRef]

27. Talhouk, A.; McConechy, M.K.; Leung, S.; Li-Chang, H.H.; Kwon, J.S.; Melnyk, N.; Yang, W.; Senz, J.; Boyd, N.; Karnezis, A.N.; et al. A clinically applicable molecular-based classification for endometrial cancers. *Br. J. Cancer* **2015**, *113*, 299–310. [CrossRef]

28. Concin, N.; Matias-Guiu, X.; Vergote, I.; Cibula, D.; Mirza, M.R.; Marnitz, S.; Ledermann, J.; Bosse, T.; Chargari, C.; Fagotti, A.; et al. ESGO/ESTRO/ESP guidelines for the management of patients with endometrial carcinoma. *Int. J. Gynecol. Cancer Off. J. Int. Gynecol. Cancer Soc.* **2021**, *31*, 12–39. [CrossRef]

29. Li, N.; Yu, K.; Lin, Z.; Zeng, D. Identifying immune subtypes of uterine corpus endometrial carcinoma and a four-paired-lncRNA signature with immune-related lncRNAs. *Exp. Biol. Med.* **2022**, *247*, 221–236. [CrossRef]

30. Bhan, A.; Soleimani, M.; Mandal, S.S. Long Noncoding RNA and Cancer: A New Paradigm. *Cancer Res.* **2017**, *77*, 3965–3981. [CrossRef]

31. Jiang, M.C.; Ni, J.J.; Cui, W.Y.; Wang, B.Y.; Zhuo, W. Emerging roles of lncRNA in cancer and therapeutic opportunities. *Am. J. Cancer Res.* **2019**, *9*, 1354–1366. [PubMed]

32. Li, B.L.; Wan, X.P. The role of lncRNAs in the development of endometrial carcinoma. *Oncol. Lett.* **2018**, *16*, 3424–3429. [CrossRef] [PubMed]

33. Yang, L.; Zhang, J.; Jiang, A.; Liu, Q.; Li, C.; Yang, C.; Xiu, J. Expression profile of long non-coding RNAs is altered in endometrial cancer. *Int. J. Clin. Exp. Med.* **2015**, *8*, 5010–5021. [PubMed]

34. Du, Z.; Yu, T.; Sun, M.; Chu, Y.; Liu, G. The long non-coding RNA TSLC8 inhibits colorectal cancer by stabilizing puma. *Cell Cycle* **2020**, *19*, 3317–3328. [CrossRef]

35. Lv, X.; Liu, L.; Li, P.; Yuan, Y.; Peng, M.; Jin, H.; Qin, D. Constructing a Novel Signature Based on Immune-Related lncRNA to Improve Prognosis Prediction of Cervical Squamous Cell Carcinoma Patients. *Reprod. Sci.* **2022**, *29*, 800–815. [CrossRef]

36. Rodriguez-Malave, N.I.; Fernando, T.R.; Patel, P.C.; Contreras, J.R.; Palanichamy, J.K.; Tran, T.M.; Anguiano, J.; Davoren, M.J.; Alberti, M.O.; Pioli, K.T.; et al. BALR-6 regulates cell growth and cell survival in B-lymphoblastic leukemia. *Mol. Cancer* **2015**, *14*, 214. [CrossRef]

37. Chiang, K.M.; Chang, H.C.; Yang, H.C.; Chen, C.H.; Chen, H.H.; Lee, W.J.; Pan, W.H. Genome-wide association study of morbid obesity in Han Chinese. *BMC Genet.* **2019**, *20*, 97. [CrossRef]

38. Cardillo, N.; Devor, E.J.; Pedra Nobre, S.; Newtson, A.; Leslie, K.; Bender, D.P.; Smith, B.J.; Goodheart, M.J.; Gonzalez-Bosquet, J. Integrated Clinical and Genomic Models to Predict Optimal Cytoreduction in High-Grade Serous Ovarian Cancer. *Cancers* **2022**, *14*, 3554. [CrossRef]

39. Gonzalez Bosquet, J.; Devor, E.J.; Newtson, A.M.; Smith, B.J.; Bender, D.P.; Goodheart, M.J.; McDonald, M.E.; Braun, T.A.; Thiel, K.W.; Leslie, K.K. Creation and validation of models to predict response to primary treatment in serous ovarian cancer. *Sci. Rep.* **2021**, *11*, 5957. [CrossRef]

40. Flensburg, C.; Sargeant, T.; Oshlack, A.; Majewski, I.J. SuperFreq: Integrated mutation detection and clonal tracking in cancer. *PLoS Comput. Biol.* **2020**, *16*, e1007603. [CrossRef]

41. Flensburg, C.; Oshlack, A.; Majewski, I.J. Detecting copy number alterations in RNA-Seq using SuperFreq. *Bioinformatics* **2021**, *37*, 4023–4032. [CrossRef] [PubMed]

42. Mohammad, N.; Muad, A.M.; Ahmad, R.; Yusof, M. Accuracy of advanced deep learning with tensorflow and keras for classifying teeth developmental stages in digital panoramic imaging. *BMC Med. Imaging* **2022**, *22*, 66. [CrossRef] [PubMed]

43. Miller, M.D.; Salinas, E.A.; Newtson, A.M.; Sharma, D.; Keeney, M.E.; Warrier, A.; Smith, B.J.; Bender, D.P.; Goodheart, M.J.; Thiel, K.W.; et al. An integrated prediction model of recurrence in endometrial endometrioid cancers. *Cancer Manag. Res.* **2019**, *11*, 5301–5315. [CrossRef] [PubMed]

44. Dai, D.; Thiel, K.W.; Salinas, E.A.; Goodheart, M.J.; Leslie, K.K.; Gonzalez Bosquet, J. Stratification of endometrioid endometrial cancer patients into risk levels using somatic mutations. *Gynecol. Oncol.* **2016**, *142*, 150–157. [CrossRef] [PubMed]

45. Salinas, E.A.; Miller, M.D.; Newtson, A.M.; Sharma, D.; McDonald, M.E.; Keeney, M.E.; Smith, B.J.; Bender, D.P.; Goodheart, M.J.; Thiel, K.W.; et al. A Prediction Model for Preoperative Risk Assessment in Endometrial Cancer Utilizing Clinical and Molecular Variables. *Int. J. Mol. Sci.* **2019**, *20*, 1205. [CrossRef]

46. Miller, M.D.; Devor, E.J.; Salinas, E.A.; Newtson, A.M.; Goodheart, M.J.; Leslie, K.K.; Gonzalez-Bosquet, J. Population Substructure Has Implications in Validating Next-Generation Cancer Genomics Studies with TCGA. *Int. J. Mol. Sci.* **2019**, *20*, 1192. [CrossRef]

47. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [CrossRef]

48. Liao, Y.; Smyth, G.K.; Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **2014**, *30*, 923–930. [CrossRef]

49. Anders, S.; Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **2010**, *11*, R106. [CrossRef]

50. Anders, S.; Reyes, A.; Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **2012**, *22*, 2008–2017. [CrossRef]

51. Sun, Z.; Nair, A.; Chen, X.; Prodduturi, N.; Wang, J.; Kocher, J.P. Author Correction: UClncR: Ultrafast and comprehensive long non-coding RNA detection from RNA-seq. *Sci. Rep.* **2018**, *8*, 5124. [CrossRef] [PubMed]

52. Cardillo, N.; Russo, D.; Newtson, A.; Reyes, H.; Lyons, Y.; Devor, E.; Bender, D.; Goodheart, M.J.; Gonzalez-Bosquet, J. Identification of Novel lncRNAs in Ovarian Cancer and Their Impact on Overall Survival. *Int. J. Mol. Sci.* **2021**, *22*, 1079. [CrossRef] [PubMed]

53. Haas, B.J.; Dobin, A.; Stransky, N.; Li, B.; Yang, X.; Tickle, T.; Bankapur, A.; Ganote, C.; Doak, T.G.; Pochet, N.; et al. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv* **2017**, 120295. [CrossRef]

54. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef]

55. Developers, T. TensorFlow. Available online: https://zenodo.org/record/5949169 (accessed on 5 December 2022). [CrossRef]