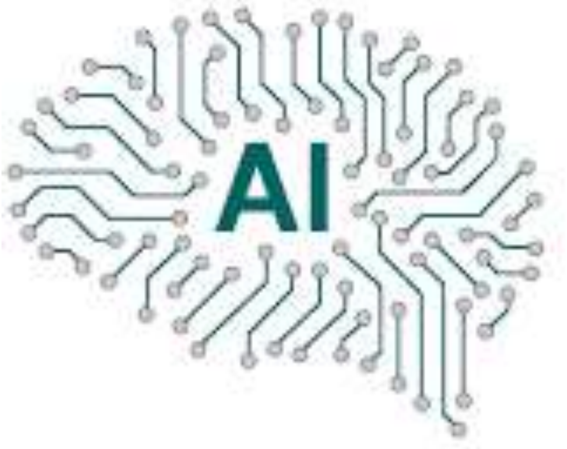
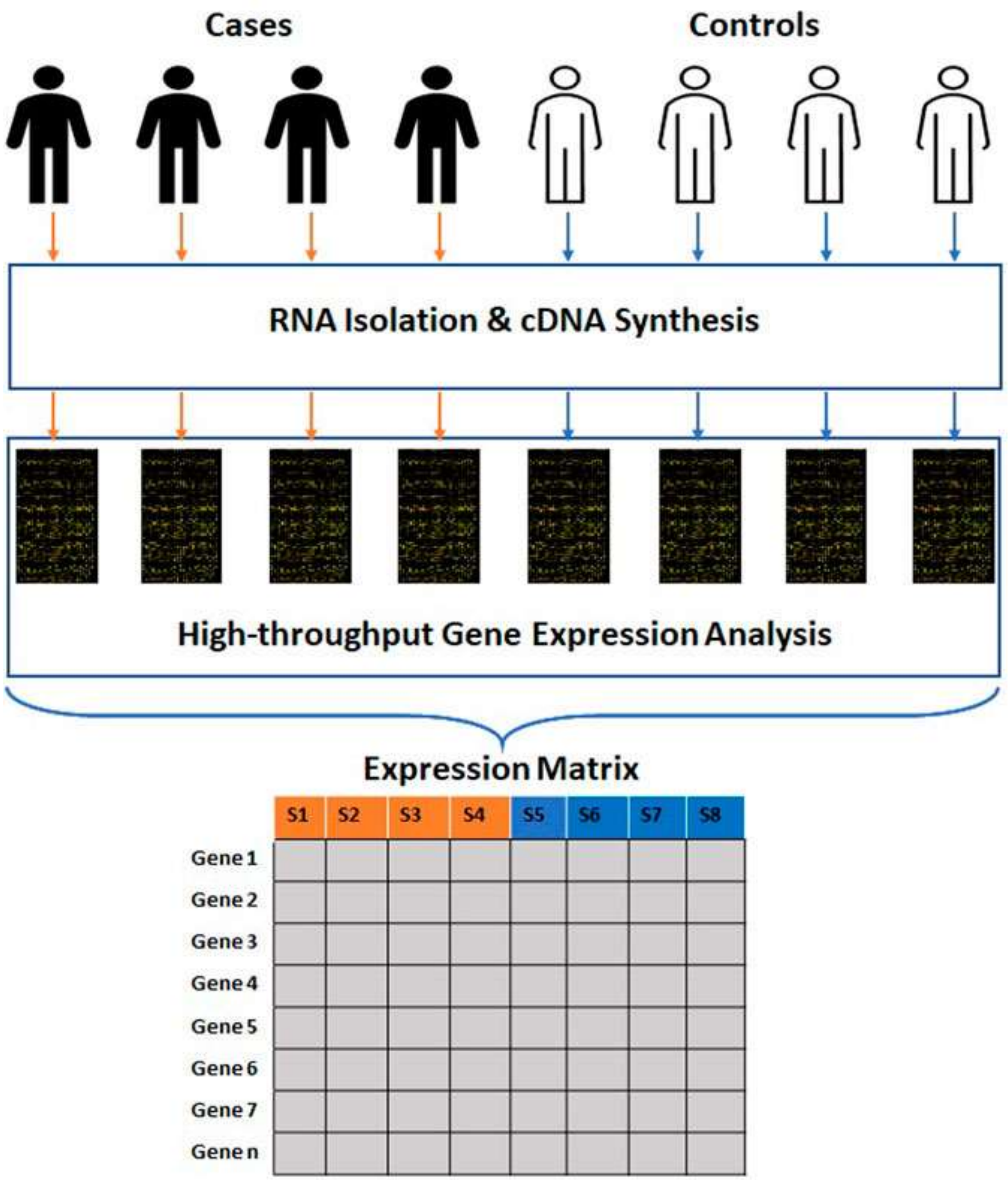


Role of AI in Unlocking Gene Expression Patterns

Presented by:
Neetu Tyagi,
PhD Research Scholar,
ICGEB, New Delhi

Understanding gene expression data


Cases
Controls
RNA Isolation & cDNA Synthesis
High-throughput Gene Expression Analysis
Expression Matrix

	S1	S2	S3	S4	S5	S6	S7	S8
Gene 1								
Gene 2								
Gene 3								
Gene 4								
Gene 5								
Gene 6								
Gene 7								
Gene n								

Gene expression data is a biological representation of various transcriptions and other chemicals found inside a cell at a given time.

Questions it will answer ?

- ✓ Which genes are active and to what extent in a particular biological sample?
- ✓ How do genes contribute to various traits, diseases, and physiological conditions?

Bhandari, N et al., (2022)

Gene expression data types

- ✓ Microarray data
- ✓ RNA-seq data
- ✓ Single cell RNA-seq data

- ➢ Each gene expression data type has its advantages and limitations.

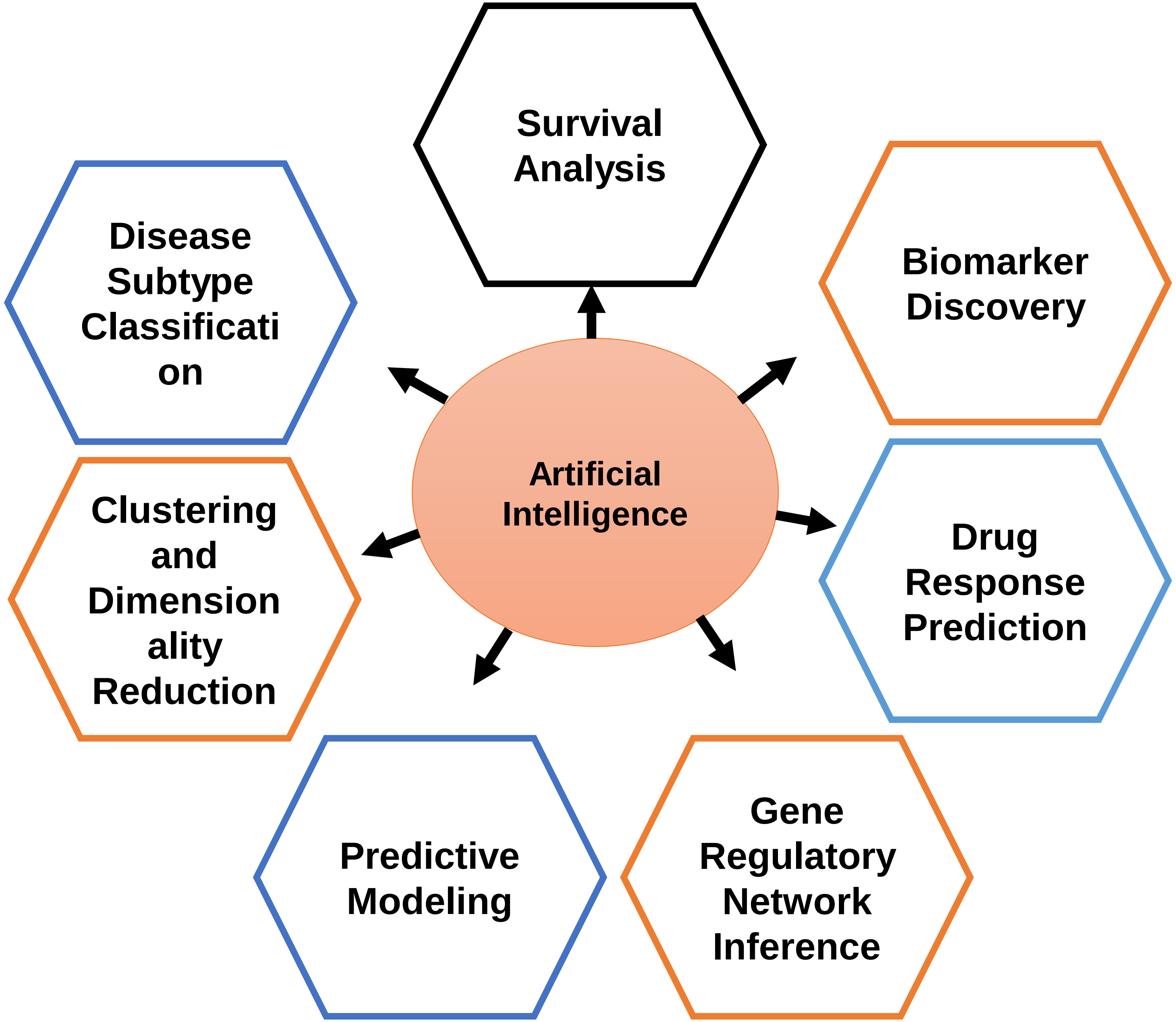
- ➢ The choice of method depends on the research question, the number of genes to be analyzed, the available sample material, and the desired level of accuracy and throughput.

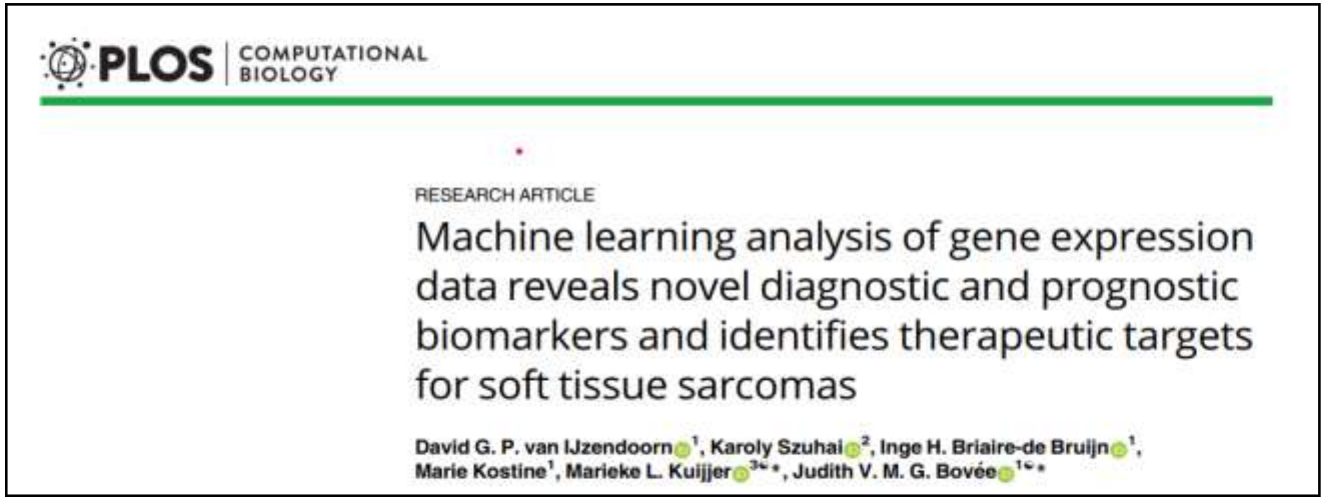
Comparison of Microarray and RNA-Seq technologies

S.No.	Basis		RNA-Seq
1.	**Technology**		High-throughput sequencing
2.	**Dynamic Range**		Broader
3.	**Probe Design**		Does not rely on predefined probes
4.	**Data Format**		Raw sequencing reads
5.	**Quantification Accuracy**		Higher accuracy
6.	**Detection of Low-Abundance Transcripts**		Detect low-abundance transcripts more effectively
7.	**Cost and Throughput**		Higher cost & provides a comprehensive view of the transcriptome
8.	**Data Analysis Complexity**		Complex analysis
9.	**Experimental Flexibility**		Data can be reanalyzed for various purposes

- ✓ scRNA-seq offers a high-resolution view of gene expression at the single-cell level, enabling the study of cellular heterogeneity and rare cell populations.
- ✓ **Steps include:** Data preprocessing, Dimensionality reduction, Cell clustering, Marker gene identification, Cell trajectory analysis, Functional enrichment analysis, visualization, downstream analysis.

Key applications of AI/ML in gene expression data analysis


Survival Analysis
Disease Subtype Classificati on
Biomarker Discovery
Artificial Intelligence
Clustering and Dimension ality Reduction
Drug Response Prediction
Predictive Modeling
Gene Regulatory Network Inference

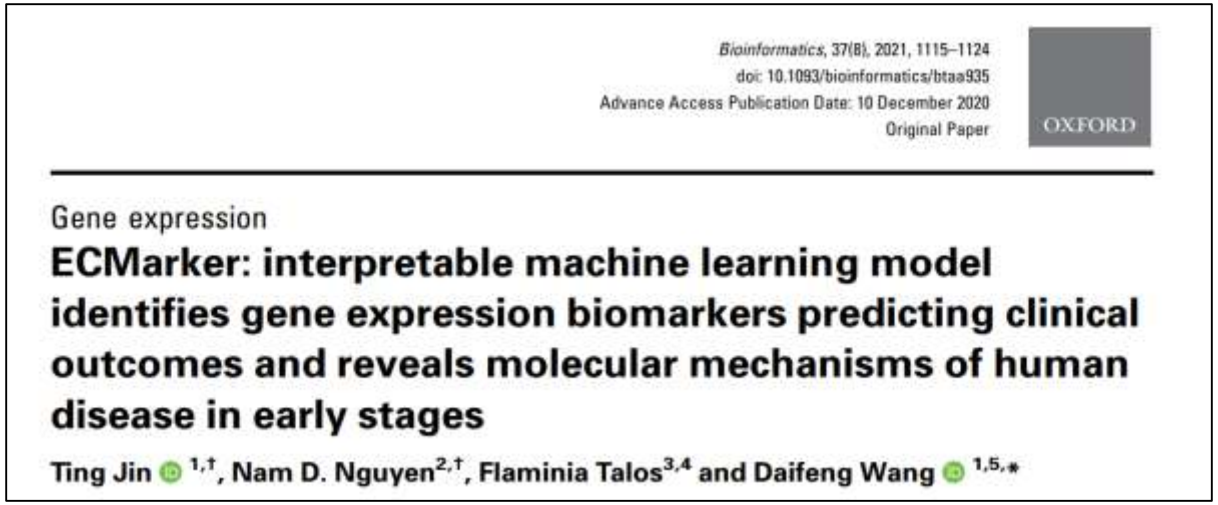


PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas

David G. P. van IJzendoorn[1], Karoly Szuhai[2], Inge H. Briaire-de Bruijn[1], Marie Kostine[1], Marieke L. Kuijjer[3☯*], Judith V. M. G. Bovée[1☯*]



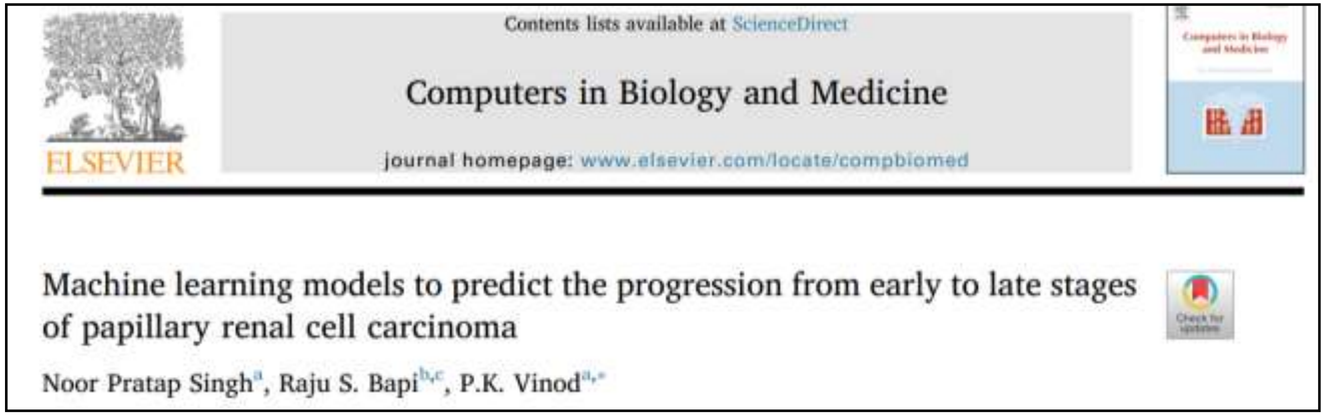
Bioinformatics, 37(8), 2021, 1115–1124
doi: 10.1093/bioinformatics/btaa935
Advance Access Publication Date: 10 December 2020
Original Paper

OXFORD

Gene expression

ECMarker: interpretable machine learning model identifies gene expression biomarkers predicting clinical outcomes and reveals molecular mechanisms of human disease in early stages

Ting Jin[1,†], Nam D. Nguyen[2,†], Flaminia Talos[3,4] and Daifeng Wang[1,5,*]



Contents lists available at ScienceDirect

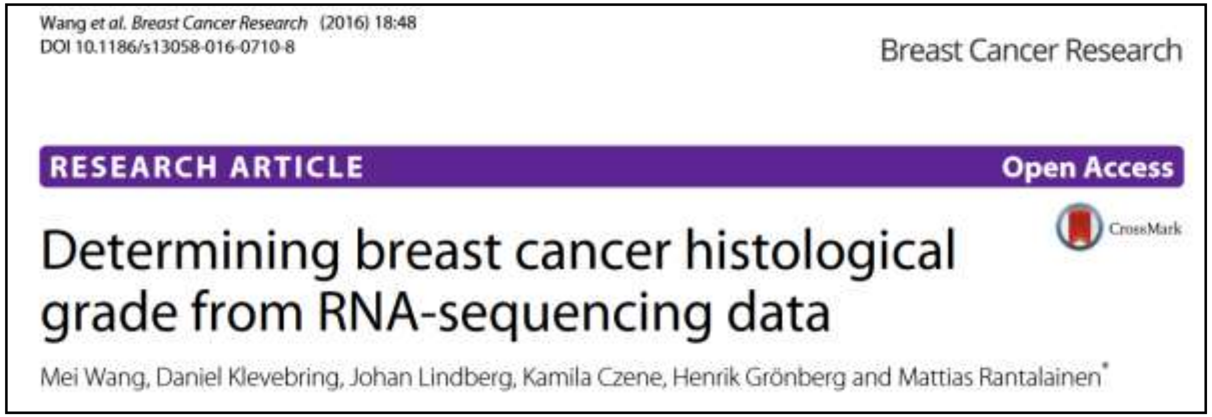
Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

ELSEVIER

Machine learning models to predict the progression from early to late stages of papillary renal cell carcinoma

Noor Pratap Singh[a], Raju S. Bapi[b,c], P.K. Vinod[a,*]



Wang *et al. Breast Cancer Research* (2016) 18:48
DOI 10.1186/s13058-016-0710-8

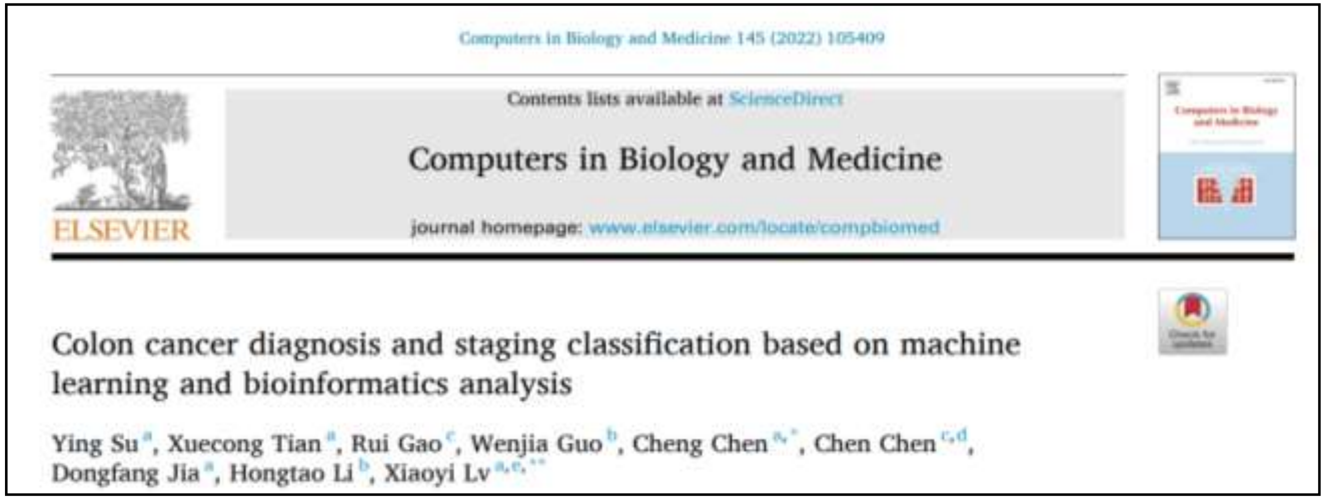
Breast Cancer Research

RESEARCH ARTICLE Open Access

CrossMark

Determining breast cancer histological grade from RNA-sequencing data

Mei Wang, Daniel Klevebring, Johan Lindberg, Kamila Czene, Henrik Grönberg and Mattias Rantalainen[*]



Computers in Biology and Medicine 145 (2022) 105409

Contents lists available at ScienceDirect

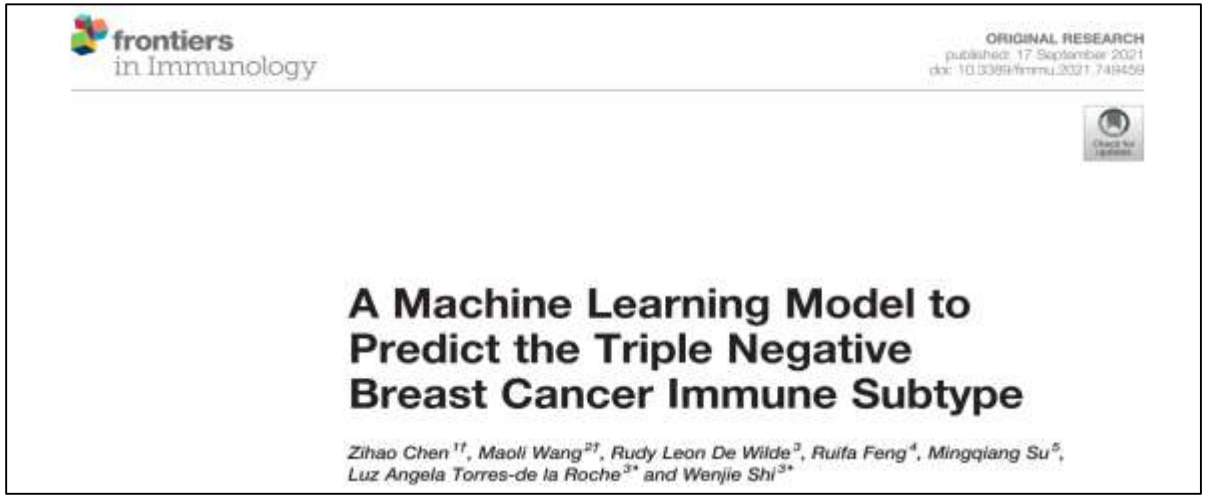
Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

ELSEVIER

Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis

Ying Su[a], Xuecong Tian[a], Rui Gao[c], Wenjia Guo[b], Cheng Chen[a,*], Chen Chen[c,d], Dongfang Jia[a], Hongtao Li[b], Xiaoyi Lv[a,e,**]



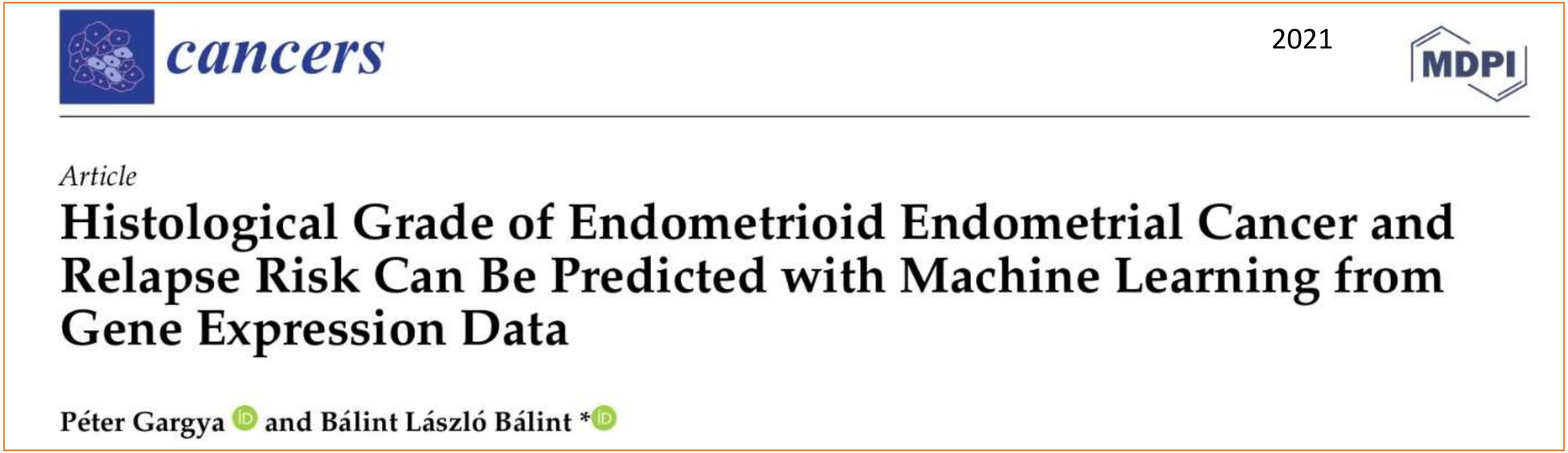
frontiers in Immunology

ORIGINAL RESEARCH
published: 17 September 2021
doi: 10.3389/fimmu.2021.749459

A Machine Learning Model to Predict the Triple Negative Breast Cancer Immune Subtype

Zihao Chen[1†], Maoli Wang[2†], Rudy Leon De Wilde[3], Ruifa Feng[4], Mingqiang Su[5], Luz Angela Torres-de la Roche[3] and Wenjie Shi[3*]*

Case study: AI application in tumor grade classification



cancers

2021

MDPI

Article

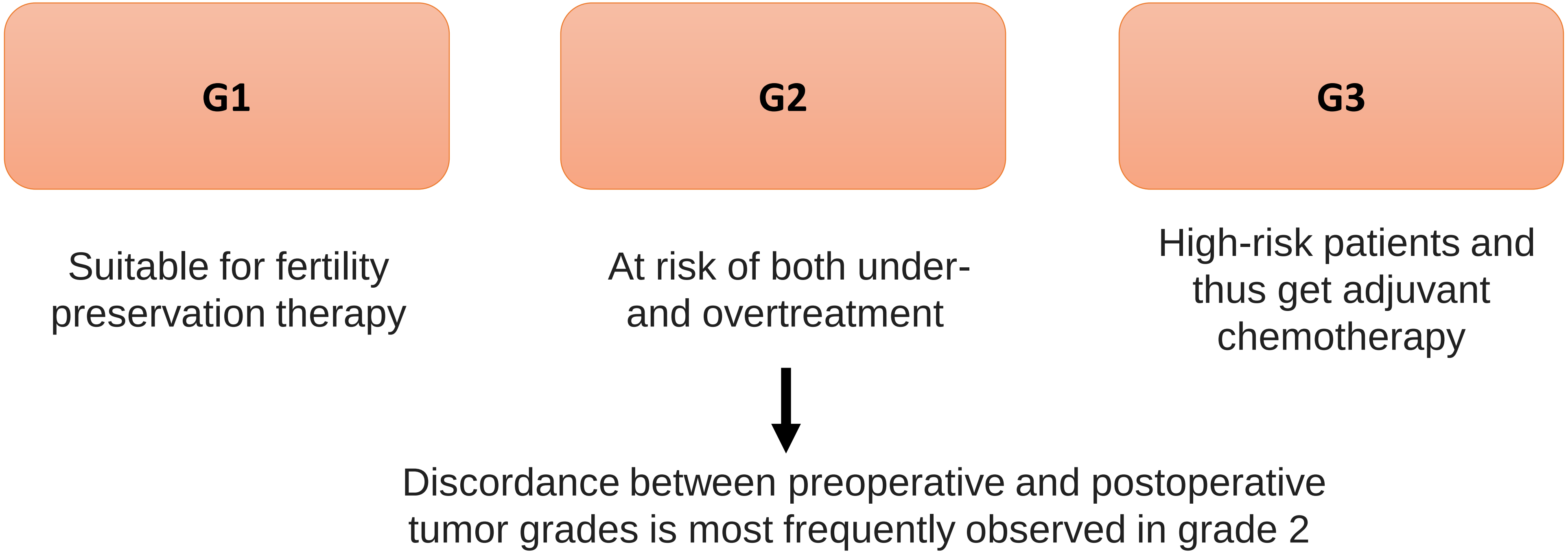
Histological Grade of Endometrioid Endometrial Cancer and Relapse Risk Can Be Predicted with Machine Learning from Gene Expression Data

Péter Gargya and Bálint László Bálint *

Article download link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8430924/
Code link: https://github.com/gargyapeter/ucec_ml_grade2021.git

Introduction

- Endometrial carcinoma represents the fourth most frequent type of malignancy among women in developed countries.
- The tumor grade serves as an independent marker of survival and may have an impact on therapy in line.
- Three groups are there:



G1

Suitable for fertility preservation therapy

G2

At risk of both under- and overtreatment

G3

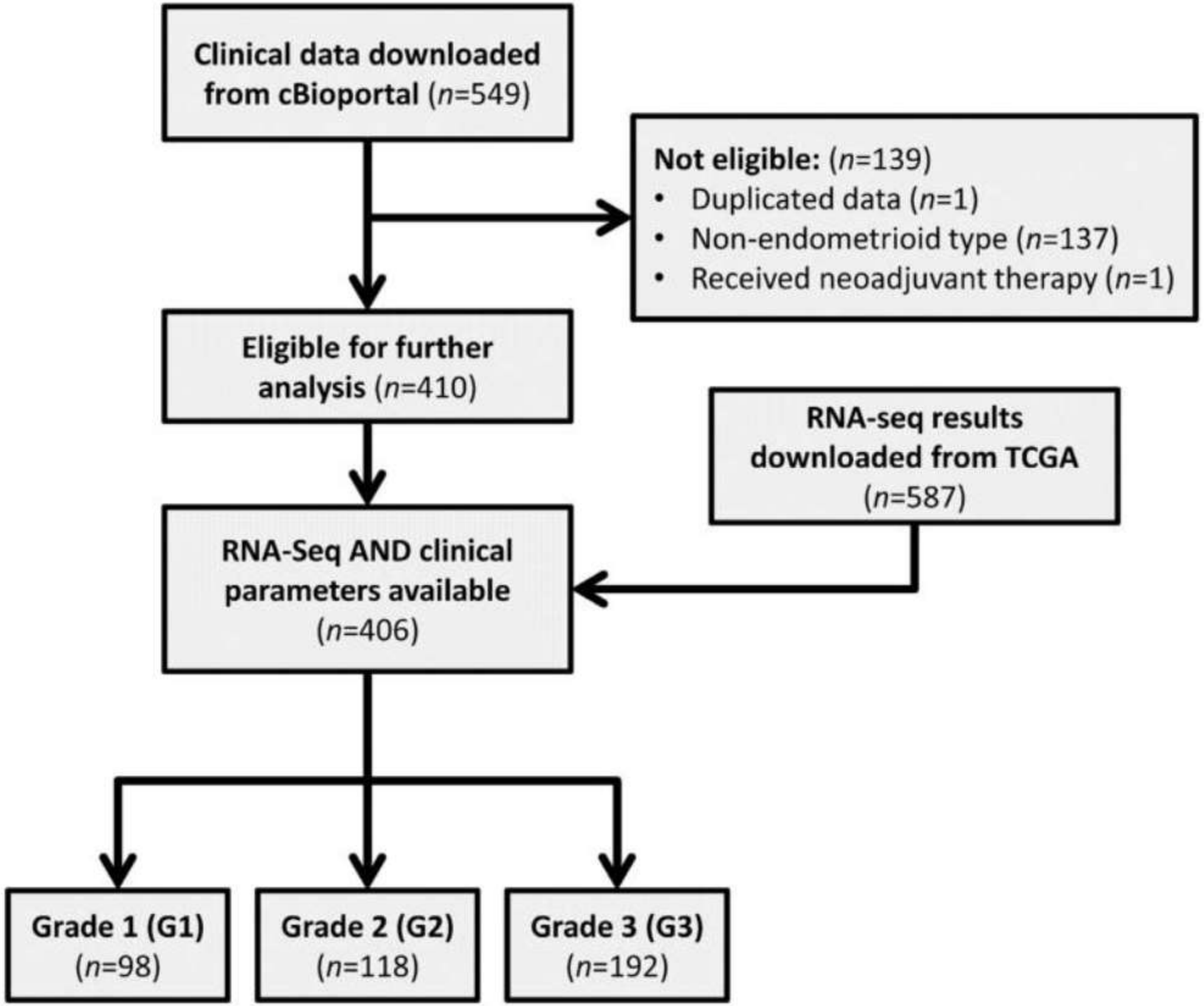
High-risk patients and thus get adjuvant chemotherapy

Discordance between preoperative and postoperative tumor grades is most frequently observed in grade 2

Rationale of the study

To develop a method to separate the G1 and G3 patients and further divide G2 patients into high-risk and low-risk subgroups based on their global gene expression profiles

Sample filtering



Clinical data downloaded from cBioportal (*n*=549)

Not eligible: (*n*=139)

- Duplicated data (*n*=1)
- Non-endometrioid type (*n*=137)
- Received neoadjuvant therapy (*n*=1)

Eligible for further analysis (*n*=410)

RNA-seq results downloaded from TCGA (*n*=587)

RNA-Seq AND clinical parameters available (*n*=406)

Grade 1 (G1) (*n*=98)

Grade 2 (G2) (*n*=118)

Grade 3 (G3) (*n*=192)

Fig. 1. **(a)** Workflow of the sample filtering and pre-processing.

Pre-processing details

✓ TCGAbiolink 2.12.6 package, downloaded the level three RNA-sequencing data altogether totaling 588 samples.

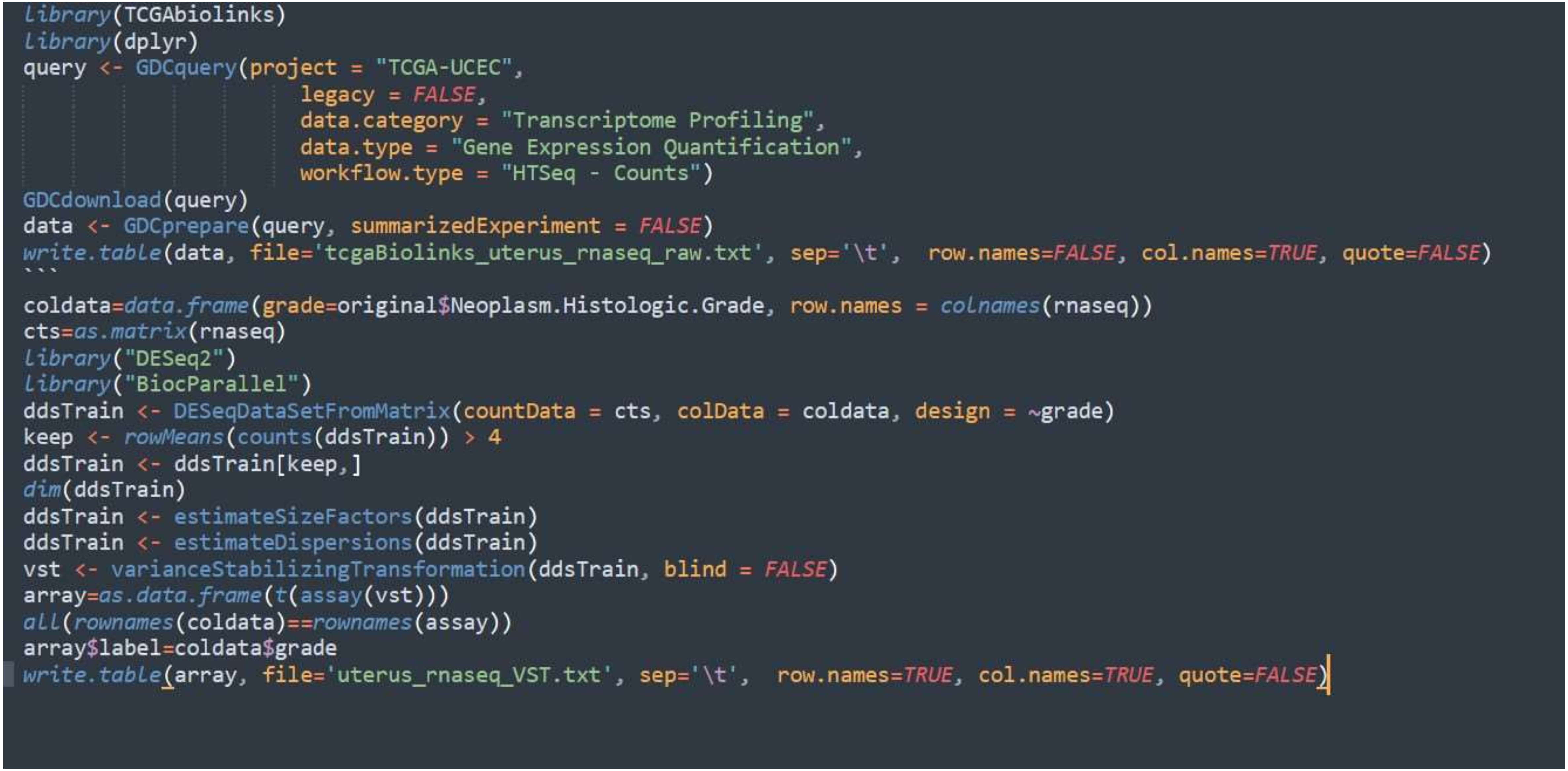
✓ The clinical data of the 546 patients of the samples were downloaded from www.cbioportal.org (accessed on 28 October 2020).

✓ Of the remaining 406 people, the G2 patients were separated and the G1 and G3 patients were managed together.

✓ From the total transcript (60,488), the low read counts were removed, and then with the left transcripts (24,349) normalized the gene expression matrix of the merged G1 and G3 groups with the help of the varianceStabilizingTransformation() function.

✓ Normalized the G2 group the same way based on the parameters received for the previous group.

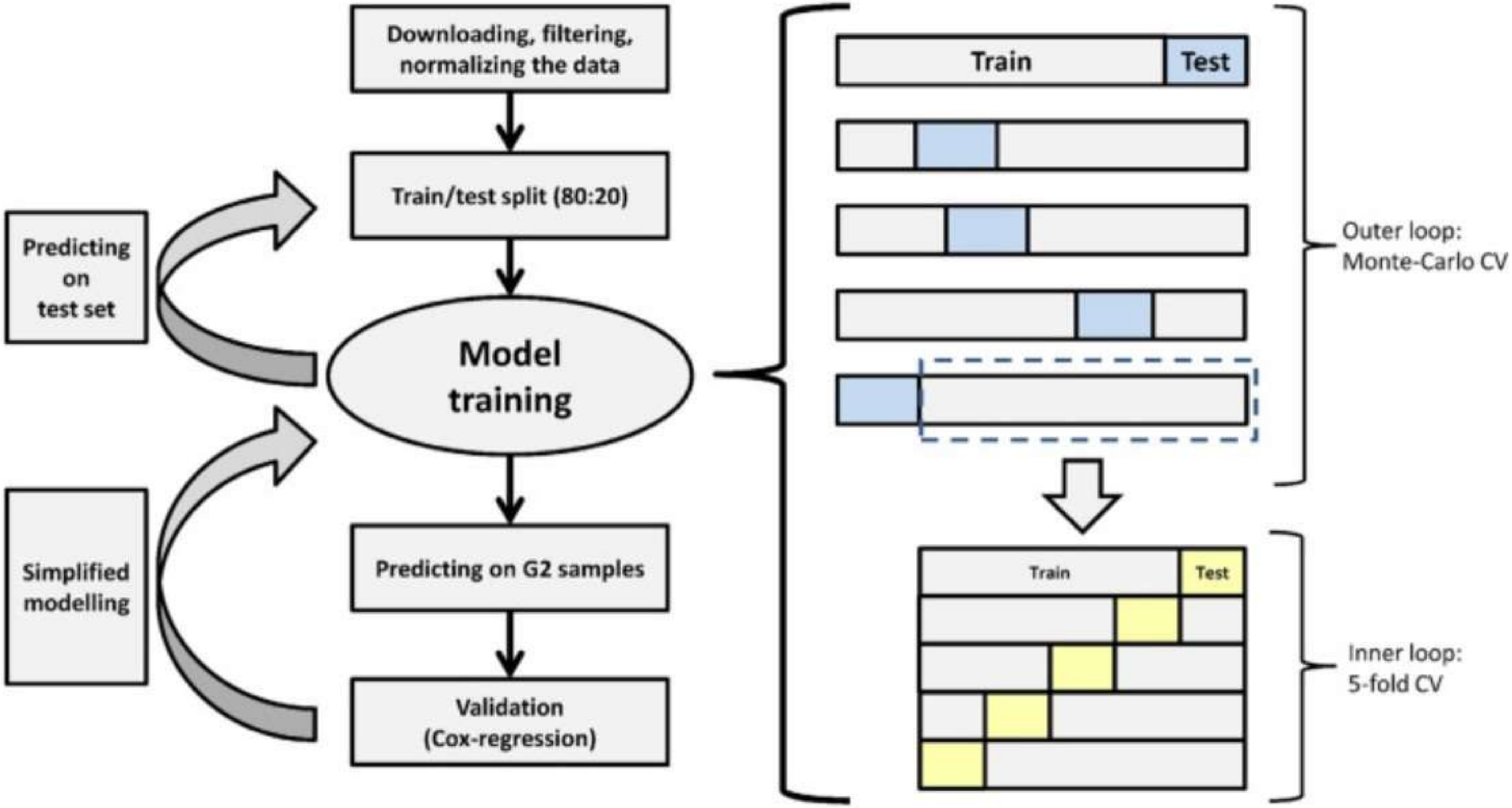
R code for pre-processing



~~~
library(TCGAbiolinks)
library(dplyr)
query <- GDCquery(project = "TCGA-UCEC",
                      legacy = FALSE,
                      data.category = "Transcriptome Profiling",
                      data.type = "Gene Expression Quantification",
                      workflow.type = "HTSeq - Counts")
GDCdownload(query)
data <- GDCprepare(query, summarizedExperiment = FALSE)
write.table(data, file='tcgaBiolinks_uterus_rnaseq_raw.txt', sep='\t',  row.names=FALSE, col.names=TRUE, quote=FALSE)
```

coldata=data.frame(grade=original$Neoplasm.Histologic.Grade, row.names = colnames(rnaseq))
cts=as.matrix(rnaseq)
library("DESeq2")
library("BiocParallel")
ddsTrain <- DESeqDataSetFromMatrix(countData = cts, colData = coldata, design = ~grade)
keep <- rowMeans(counts(ddsTrain)) > 4
ddsTrain <- ddsTrain[keep,]
dim(ddsTrain)
ddsTrain <- estimateSizeFactors(ddsTrain)
ddsTrain <- estimateDispersions(ddsTrain)
vst <- varianceStabilizingTransformation(ddsTrain, blind = FALSE)
array=as.data.frame(t(assay(vst)))
all(rownames(coldata)==rownames(assay))
array$label=coldata$grade
write.table(array, file='uterus_rnaseq_VST.txt', sep='\t',  row.names=TRUE, col.names=TRUE, quote=FALSE)
~~~

Model development



Downloading, filtering, normalizing the data

Train/test split (80:20)

Model training

Predicting on G2 samples

Validation (Cox-regression)

Predicting on test set

Simplified modelling

Fig. 1. (b) Flowchart of data modelling.

Results

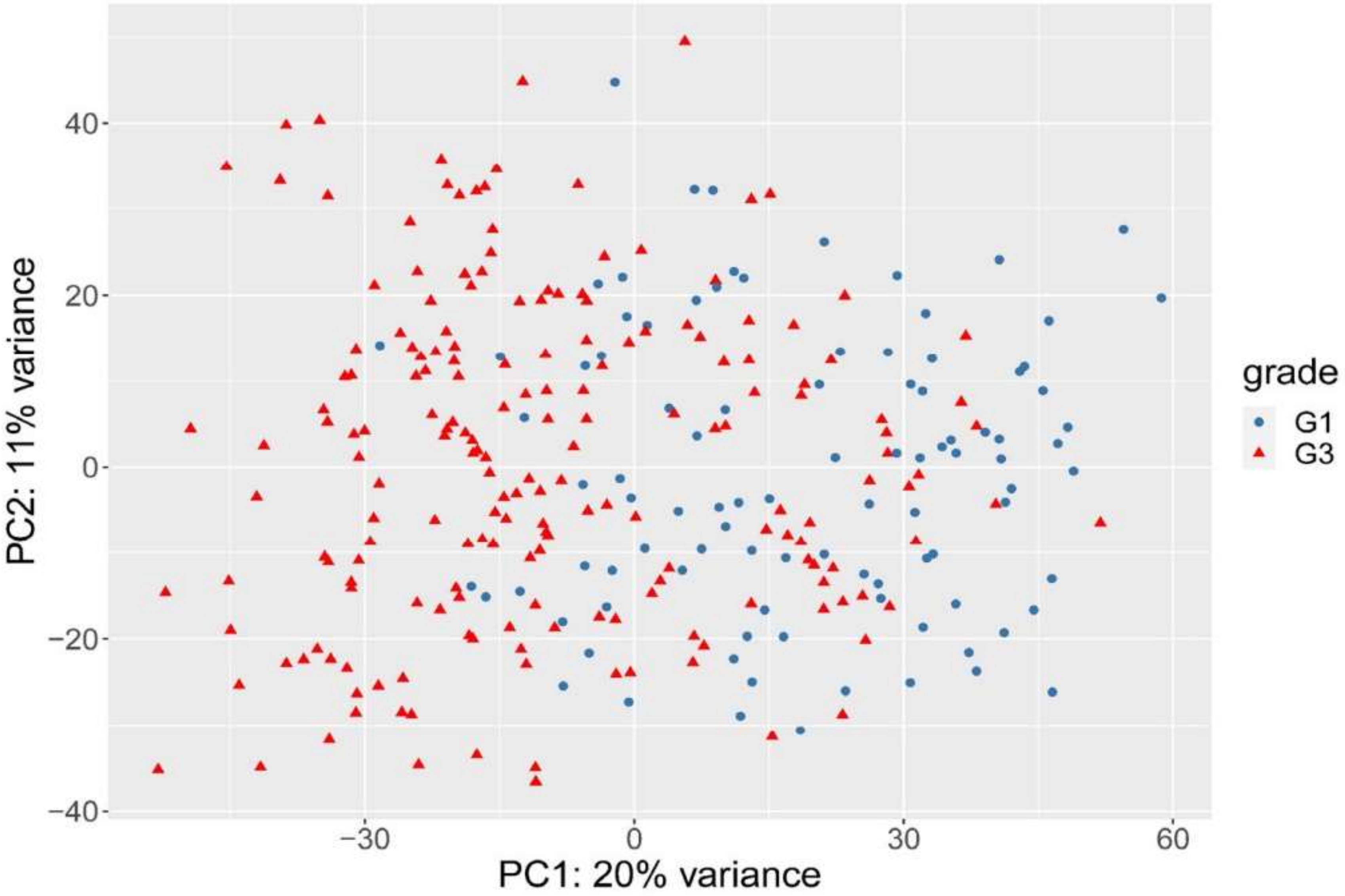
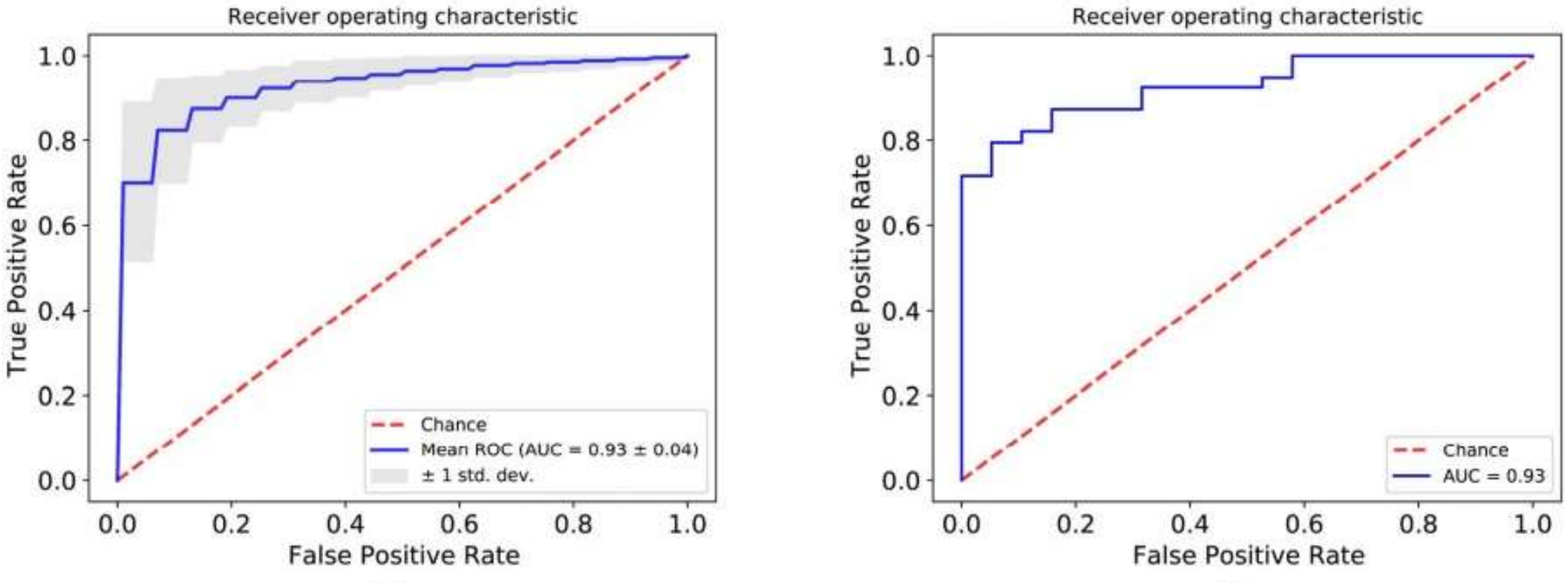


Fig. 2. Principal component analysis of G1 (blue dots) and G3 (red triangles) groups



(a)

(b)

True Grade			
	G1	14	5
	G3	5	34
		G1	G3
		Predicted Grade	

(c)

Fig. 3. ROC curves of the cross-validation rounds **(a)** and the test data **(b).** The blue line represents the mean AUC value and the grey area represents the standard deviation. **(c)** Confusion matrix of the test data.

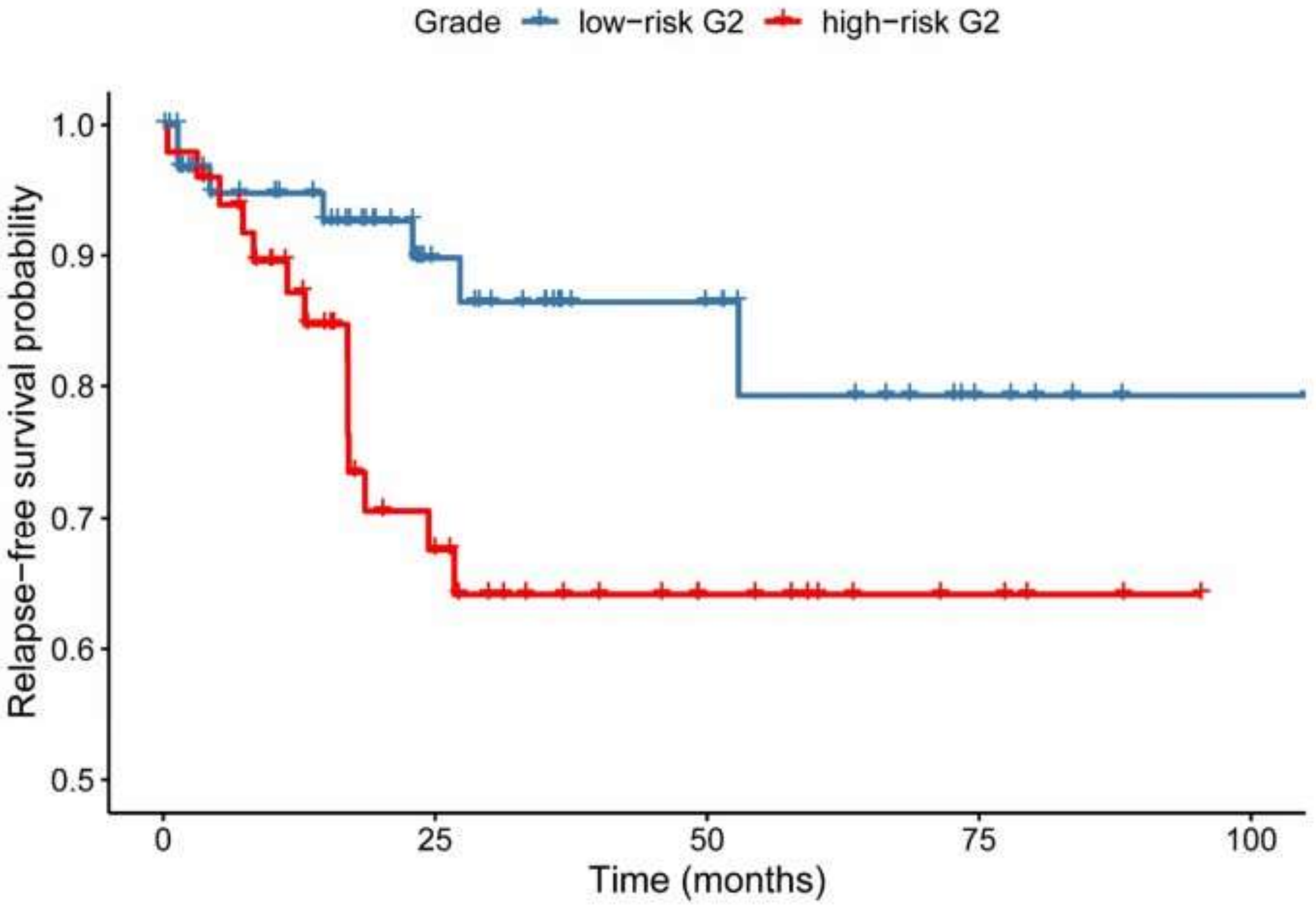
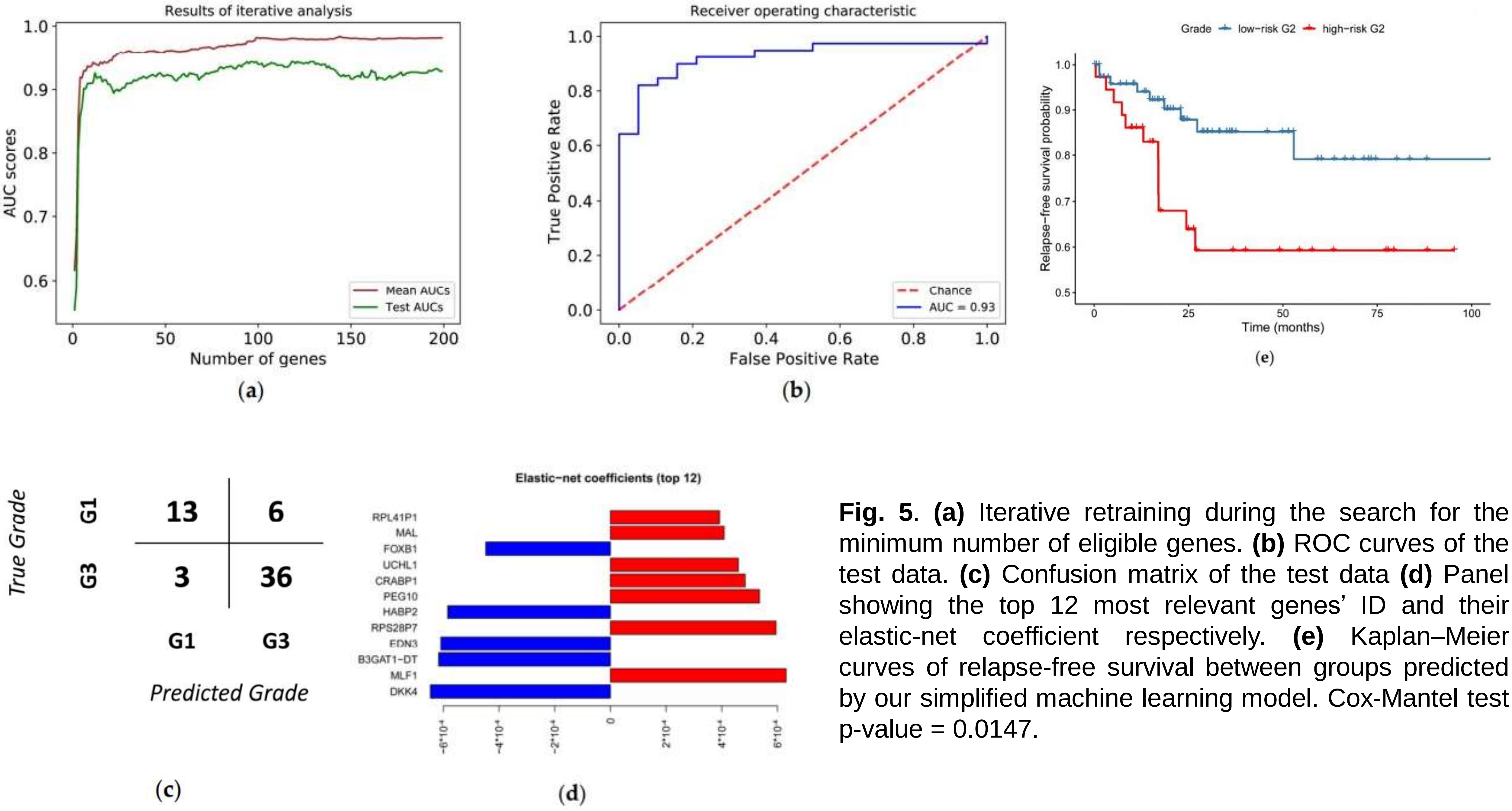
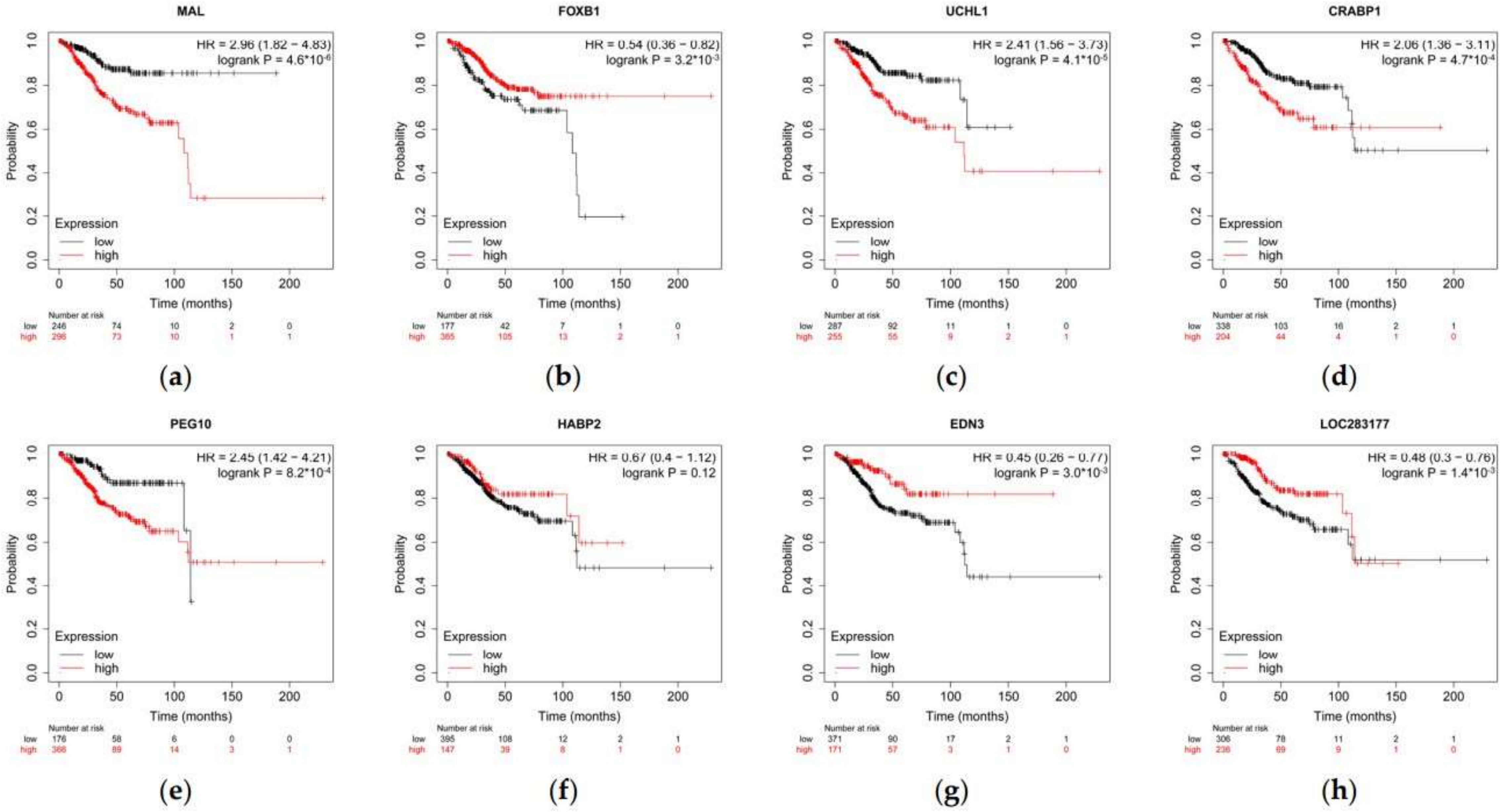


Fig. 4. Kaplan–Meier curves of relapse-free survival between groups predicted by machine learning model. Blue: low-risk G2, red: high-risk G2. Cox-Mantel test p-value 0.037.



G1	13	6
G3	3	36
	G1	G3

Fig. 5. **(a)** Iterative retraining during the search for the minimum number of eligible genes. **(b)** ROC curves of the test data. **(c)** Confusion matrix of the test data **(d)** Panel showing the top 12 most relevant genes' ID and their elastic-net coefficient respectively. **(e)** Kaplan–Meier curves of relapse-free survival between groups predicted by our simplified machine learning model. Cox-Mantel test p-value = 0.0147.

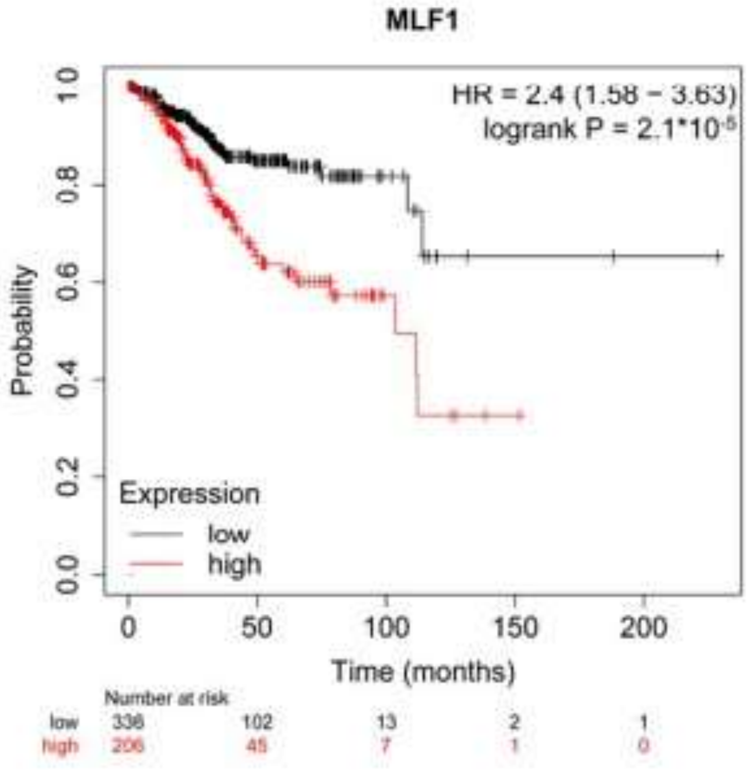


(a) (b) (c) (d)

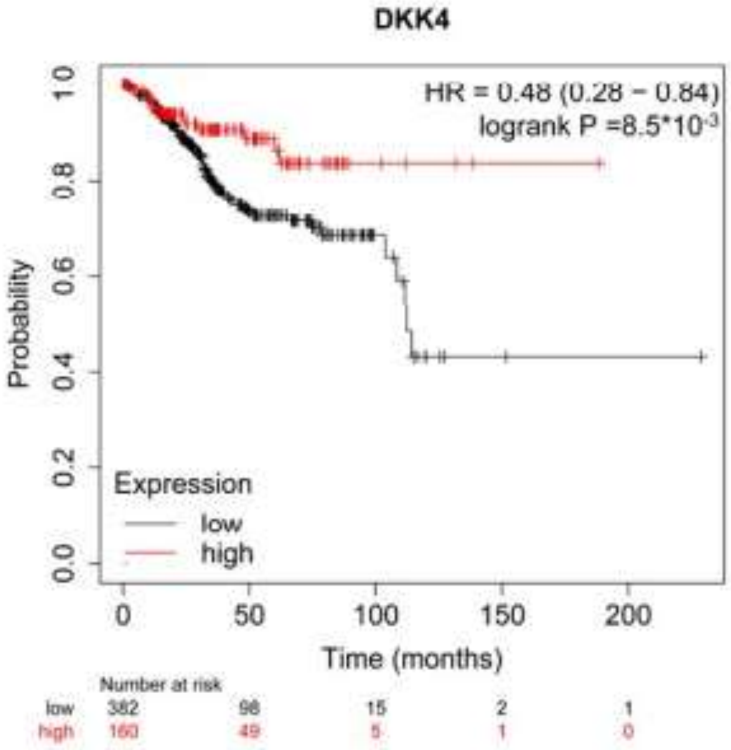
(e) (f) (g) (h)

Fig. 6. Kaplan–Meier curves of overall survival between groups with high gene expression and low gene expression of selected genes.

Kaplan–Meier curves for selected genes



(i)



(j)

- ✓ High expression of FOXB1, HABP2, EDN3, B3GAT1- DT, and DKK4 was associated with low-risk phenotype.

- ✓ High expression of RPL41P1, MAL, UCHL1, CRABP1, PEG10, RPS28P7, and MLF1 was responsible for more clinically aggressive behavior

Key findings

✓ RNA-sequencing data from the TCGA project and machine learning to develop a model which can correctly classify grade 1 and grade 3 samples. They used the trained model on grade 2 patients to subdivide them into low-risk and high-risk groups.

✓ With iterative retraining, they selected the most relevant 12 transcripts to build a simplified model without losing accuracy. Both models had a high AUC of 0.93.

Significance

✓ The approach overcomes the subjective components of the histological evaluation.

✓ The developed method can be automated to perform a prescreening of the samples before a final decision is made by pathologists.

✓ A translational approach based on machine learning methods could allow for better therapeutic planning for grade 2 endometrial cancer patients.

Thank you…. ☺