СОФИЙСКИ УНИВЕРСИТЕТ
„СВ. КЛИМЕНТ ОХРИДСКИ“

БИОЛОГИЧЕСКИ ФАКУЛТЕТ

БИОЛОГИЧЕСКИ ФАКУЛТЕТ
СОФИЙСКИ УНИВЕРСИТЕТ

SOFIA UNIVERSITY
ST. KLIMENT OHRIDSKI

FACULTY OF BIOLOGY

FACULTY OF BIOLOGY
SOFIA UNIVERSITY

Sofia University "St. Kliment Ohridski"

Faculty of Biology

Department of Genetics

GEORGI DIMITROV BLAZHEV

„A multi-omics approach in the analysis of the biological and clinical heterogeneity of some rare malignancies“

ABSTRACT

dissertation thesis for the Doctoral degree in the professional field

4.3 Biological Sciences (Genetics – Cancer Genetics)

Scientific supervisor: **assoc. prof. Velizar Shivarov, M.D.–PhD**

Sofia

2024

The PhD student, Georgi Blazhev, has developed the dissertation for the Doctoral degree as a full-time PhD student in the professional field of Biological Sciences (Genetics – Cancer Genetics) at the Faculty of Biology of Sofia University "St. Kliment Ohridski".

The dissertation was discussed at an extended meeting of the Department of Genetics, Faculty of Biology, Sofia University "St. Kliment Ohridski" held on .04.2024, based on the Rector's order No ....

The dissertation is scheduled for a public defense with a scientific jury, confirmed by Order No........................ of the Rector of Sofia University "St. Kliment Ohridski" Prof. Georgi Valchev, PhD.

Scientific Jury:

Internal scientific jury members:

1. ....

2. ....

External scientific jury members:

3. ....

4. ....

5. ....

The dissertation public defense will take place on ………….. at …………… **hours** in the building of the Faculty of Biology of Sofia University "St. Kliment Ohridski".

The materials for the defense of the dissertation are available to those interested in the library of the Faculty of Biology of Sofia University "St. Kliment Ohridski", 1164 Sofia, 8 Dragan Tsankov Blvd., as well as on the university's website.

The dissertation contains 110 pages and consists of: introduction, six chapters, contributions, appendices, bibliography, declaration of originality, and publications. In its volume, it contains 38 figures and 8 tables. The bibliography includes 173 cited literary sources.

CONTENTS

LIST OF TABLES 3

LIST OF FIGURES 4

ABBREVIATIONS USED 6

I. Introduction 7

II. Research hypothesis, aims, and tasks 9

Research hypothesis 9

Research aim 9

Research tasks 9

III. Materials and Methods 10

Datasets 10

Deriving the model 10

Gene set enrichment analysis 10

Cibersort 10

Drug sensitivity analysis 11

Integrated DNA methylation analysis 11

Common statistical procedures 11

IV. Results 13

Identification of studies for inclusion in the analysis 13

Building and initial performance of a two-gene prognostic score (2-PS) 14

Validation of the 2-PS 19

Gene Set Enrichment Analysis in Expression Profiles (GSEA) 24

Integrated DNA methylation and gene expression analysis 27

Correlation with estimated populations of infiltrating immune cells 32

The potential predictive power of the 2-PS 33

V. Discussion 35

VI. Conclusion and inferences 39

VII. Contributions 40

Original contributions 40

Confirmed contributions 40

VIII. Publications 41

Journal articles 41

Participation in scientific meetings 41

LIST OF TABLES

Table 1. Used omics datasets to build, validate, and explore the biological nature of 2-PS…............. 11

Table 2. Analysis results for differentially methylated CpG islands from the TCGA dataset. For analysis purposes, an island is defined as at least two contiguous CpG sites that were included on Illumina's Infinium 450K chip.. 28

Table 3. Results of the integrated analysis of differentially methylated CpG islands and gene expression from the TCGA dataset. For analysis purposes, an island was defined as at least two contiguous CpG sites that were included on Illumina's Infinium 450K chip…................................... 29

LIST OF FIGURES

Figure 1. Number of PubMed entries by year after searching the terms "mesothelioma AND malignant". Last search date 10-12-2023 13
Figure 2. PRISMA diagram of the process of selecting studies for inclusion in the analysis. Based on the systematic approach described in the diagram, we arrived at a selection of 3 studies containing publicly available multi-omics data from MPM patients 14
Figure 3. Diagram of the selected analytical approach. 16
Figure 4. Distribution histogram of the continuous score for the patients from the TCGA cohort. The red vertical line shows the median value used for the dichotomous (binary) stratification into groups of high- and low-scoring patients. 17
Figure 5. Univariate analysis of survival according to 2-PS in the TCGA cohort. Patients are stratified into a high score and a low score group based on the median of the continuous score. The p-value is from Cox regression analysis 17
Figure 6. ROC curve analysis for the prognostic value of the binary score in terms of overall survival in the TCGA cohort. Area under the curve (AUC) as well as sensitivity and specificity were calculated at a median value of 21,59, which was used for dichotomous separation. 18
Figure 7. Multivariate analysis of survival according to the 2-PS in the TCGA cohort. Covariates included are age, sex, histology, stage and mutational status pertaining to 4 genes: BAP1, TP53, SETD2 and NF2. 18
Figure 8. Distribution histogram of the continuous score for the patients from the Bueno cohort. The red vertical line shows the median value used for the dichotomous (binary) stratification into groups of high- and low-scoring patients. 19
Figure 9. Univariate analysis of survival according to 2-PS in the Bueno cohort. Patients are stratified into a high score and a low score group based on the median of the continuous score. The p-value is from Cox regression analysis 20
Figure 10. ROC curve analysis for the prognostic value of the binary score in terms of overall survival in the Bueno cohort. Area under the curve (AUC) as well as sensitivity and specificity were calculated at a median value of 11,78, which was used for dichotomous separation. 20
Figure 11. Multivariate analysis of survival according to the 2-PS in the Bueno cohort. Covariates included are age, sex, histology, stage and mutational status pertaining to 4 genes: BAP1, TP53, SETD2 and NF2 21
Figure 12. Distribution histogram of the continuous score for the patients from the Blum cohort. The red vertical line shows the median value used for the dichotomous (binary) stratification into groups of high- and low-scoring patients. 22
Figure 13. Univariate analysis of survival according to 2-PS in the Blum cohort. Patients are stratified into a high score and a low score group based on the median of the continuous score. The p-value is from Cox regression analysis 22
Figure 14. ROC curve analysis for the prognostic value of the binary score in terms of overall survival in the Blum cohort. Area under the curve (AUC) as well as sensitivity and specificity were calculated at a median value of 6,56, which was used for dichotomous separation. 23
Figure 15. Multivariate analysis of survival according to the 2-PS in the Blum dataset. Covariates included are age, sex, histology, stage and mutational status pertaining to 4 genes: BAP1, TP53, SETD2 and NF2. 23
Figure 16. Example result of the GSEA analysis on the TCGA dataset. (A) Enrichment plot for genes from the “Mitotic spindle” list; (B) Gene expression heatmap of the same list. Patients with a high score are shown in gray and those with a low score are shown in yellow. 24
Figure 17. Example result of the GSEA analysis on the Bueno dataset. (A) Enrichment plot for genes from the “Mitotic spindle” list; (B) Gene expression heatmap of the same list. Patients with a high score are shown in gray and those with a low score are shown in yellow. 25

Figure 18. Example result of the GSEA analysis on the Blum dataset. (A) Enrichment plot for genes from the "Mitotic spindle" list; (B) Gene expression heatmap of the same list. Patients with a high score are shown in gray and those with a low score are shown in yellow. 25
Figure 19. Bar charts of the significantly more frequently overexpressed gene lists in the patients with high scores found in at least one of the three cohorts (TCGA, Bueno, Blum) determined by GSEA analysis. Numbers indicate the count of gene lists in each of the sets... 26
Figure 20. Venn diagram of the significantly overlapping more frequently overexpressed gene lists in the patients with a high score from the three cohorts (TCGA, Bueno, Blum) determined by GSEA analysis. Numbers indicate the count of gene lists in each of the sets... 26
Figure 21. Result of the analysis of the differentially methylated CpG sites in the TCGA dataset. Heatmap of β-values for differentially expressed CpG sites (vertical) versus patients (horizontal)..... 28
Figure 22. Two-dimensional correlation plot of methylation of the CpG island chr2:113403001-113404079 versus the SLC20A1 gene expression in the TCGA dataset. Patients with high and low scores are indicated in a different color. 29
Figure 23. Two-dimensional correlation plot of methylation of the CpG island chr6:30654392-30654934 versus the KIAA1949 gene expression in the TCGA dataset. Patients with high and low scores are indicated in a different color. 30
Figure 24. Result of the analysis of the differentially methylated CpG sites in the Blum dataset. Heatmap of β-values for differentially expressed CpG sites (vertical) versus patients (horizontal)..... 31
Figure 25. Venn diagram of overlapping significantly differentially methylated CpG sites in high versus low score patients from two of the cohorts (TCGA, Blum) determined by COHCAP analysis. Numbers indicate the count of gene lists in each of the sets. 31
Figure 26. Correlation matrix of continuous score with immune infiltration score for specific immune system cell populations according to expression data from the TCGA cohort. The CIBERSORT algorithm was used to calculate the immune scores. The color of each square of the correlation matrix reflects the calculated correlation coefficient. 32
Figure 27. Correlation matrix of continuous score with immune infiltration score for specific immune system cell populations according to expression data from the Bueno cohort. The CIBERSORT algorithm was used to calculate the immune scores. The color of each square of the correlation matrix reflects the calculated correlation coefficient. 33
Figure 28. Two-dimensional correlation plots between the calculated 2-PS score and AUC for different drugs or compounds for MPM cell lines analyzed as part of the GDSC1 project. Only plots for identified significant correlations are presented... 34
Figure 29. Two-dimensional correlation plots between the calculated 2-PS score and AUC for different drugs or compounds for MPM cell lines analyzed as part of the GDSC2 project. Only plots for identified significant correlations are presented... 34

ABBREVIATIONS USED

2-PS	Two-gene prognostic score. The term "score" means result, points, etc.
DNA	Deoxyribonucleic acid
mRNA	Messenger ribonucleic acid
MPM	Malignant pleural mesothelioma
WHO	World Health Organization
AIC	Akaike information criterion
ANOVA	Analysis of variance
AUC	Area under the curve
CHiP	Chromatin immunoprecipitation
CRISPR	Clustered regularly interspaced short palindromic repeats
FDR	False discovery rate
GDC	Genomic Data Commons
GSEA	Gene set enrichment analysis
ROC	Receiver operating characteristics
RPKM	Reads per kilobase million
FRPKM	Fragments per kilobase million, fragments per kilobase of exon per million mapped fragments
PARP	Poly (ADP-ribose) polymerase
scRNA-seq	Single-cell RNA sequencing, Small conditional RNA
siRNA	Short interfering ribonucleic acid
TCGA	The Cancer Genome Atlas
TILs	Tumor infiltrating lymphocytes
TNM	TNM classification of malignant tumors

I. Introduction

Malignant diseases are currently a leading cause of death and disability and reduced quality of life. The burden of these diseases is unevenly distributed in different regions of the world. According to the latest data from the World Health Organization (WHO), the number of deaths worldwide in 2020 is about 10 million. Cancer mortality is unevenly distributed in different regions of the world but is among the highest in some Eastern European countries, including Bulgaria.

Cancer morbidity and mortality are mainly due to a few most commonly found locations with corresponding sex differences. On the other side, approximately a quarter of the newly diagnosed cancer cases in Europe are associated with various rare cancers (https://www.rarecancerseurope.org/). The objective definition of the term rare cancer, is entirely epidemiological and, in a sense, based on the incidence of the relevant type of cancer rather than the disease itself. Currently, it is accepted that as rare cancer types such nosological entities, which are characterized by an annual frequency of newly diagnosed cases in a given population below 6 per 100,000 people should be designated. According to the summaries and recommendations of the EU Joint Working Group on Rare Cancers presented in the document "Rare Cancers Agenda 2030" (https://www.esmo.org/content/download/294217/5832976/1/Rare-Cancer-Agenda-2030.pdf), the list of rare cancers is based on the so-called "Tier-1"— nosological entities with an annual incidence below <6/100,000, grouped into larger families according to location or histologic origin and childhood presentation (pediatric rare cancers are separated into a different family).

The main clinical issues related to rare cancer types arise from the basic characteristics of the rare diseases, namely: 1) clinical decisions are hindered by a lack of medical expertise and high-quality evidence from the clinical trials; 2) the healthcare system hardly serves certain territories with specialized care that these patients need; and 3) clinical trials are difficult and limited by small patient numbers, thus making it difficult to generate high-quality data.

On the other hand, for a variety of reasons, the share of rare cancers and their mortality can vary significantly even between developed countries. For example, the average 5-year survival rate for patients with rare types of cancer in the USA is 54%, in the EU - an average of 48%, in Germany - 55%, and in Bulgaria - 35%. It should be especially noted that some rare types, such as testicular carcinoma, are distinguished by extremely high survival rates with timely and accurate diagnosis and adequate treatment. On the other hand, other rare types of cancer, such as the mesotheliomas, are characterized by a 5-year survival of the order of 5-10% regardless of country and region. For this reason, the reduction of overall mortality from rare cancers worldwide is associated not only with timely and adequate care, but also with conducting thorough biomedical research, which could significantly improve diagnostic and therapeutic options, and thus clinical outcomes. In this regard, the recommendations in the document Rare Cancers Agenda 2030 encourage the implementation of clinical, epidemiological, and translational studies in the field of rare cancers, involving as many centers and patients as possible.

Driven by the obvious greatest unmet medical need to improve overall survival in mesotheliomas, with this work, we have focused our efforts precisely on the main nosological entity of this group of diseases – malignant pleural mesothelioma (MPM).

II. Research hypothesis, aims, and tasks

Research hypothesis

Based on the above, we hypothesized that by using the already available multi-omics data, it is possible to derive a new gene expression-based score that has prognostic and predictive value in MPM patients.

Research aim

Derive and validate a novel gene expression-based prognostic score in MPM patients.

Research tasks

1. Identify published studies in MPM patients with publicly available transcriptomic data and at least one other type of omics data (e.g. genomic, epigenomic).
2. Derive a gene expression-based score using data from the most extensive study (with the largest amount of omics data) as training data.
3. Determine the prognostic value of the score relating to other clinical data from the patients in the training dataset.
4. Validate the derived score based on the transcriptomic data from the remaining identified studies.
5. Determine the prognostic value of the score in relation to other clinical data from the patients in the validation dataset.
6. Test whether the derived score defines specific subgroups of patients based on the gene expression profile in each of the datasets that were used.
7. Test whether the derived score defines specific subgroups of patients based on the DNA methylation profile in each of the available epigenomic datasets that were used.
8. Test whether the derived score correlates with the infiltration profile by specific immune system cell types using deconvolution techniques of transcriptomic data.
9. Test whether the derived score correlates with sensitivity to certain drugs based on publicly available data from in vitro studies with MPM cell lines.

III. Materials and Methods

Datasets

- Data from TCGA from GDC – Genomics Data Commons Portal (https://portal.gdc.cancer.gov/)
- The European Genome-phenome Archive (EGA) (https://ega-archive.org/) (EGAD00001001915)
- Trim Galore v. 0.6.3
- FastQC v. 0.72
- HISAT2 (v. 2.1)
- featureCounts (v. 1.6.4)
- *limma* package for R
    - RMA
- Array Express (E-MTAB-6877)

Deriving the model

- DepMap (https://depmap.org/portal/)
- *rbsurv* package for the R statistical environment
    - Akaike Information Criterion (AIC)
- *survival* package for R
    - Cox regression coefficients
- Cutoff Finder package for R (https://molpathoheidelberg.shinyapps.io/CutoffFinder_v1/)
    - Receiver operating chrarecteristics (ROC) curves
- *survminer* package for R

Gene set enrichment analysis

- Gene set enrichment analysis (GSEA) developed by Broad Institute (http://www.broadinstitute.org/gsea/index.jsp) (standalone version GSEA 4.0.3)
- Oncogenic gene ontology signatures from the Molecular Signatures Database (MSigDB) (https://www.gsea-msigdb.org/gsea/msigdb/human/collections.jsp#H)

Cibersort

- The Immune Landscape of Cancer (https://gdc.cancer.gov/about-data/publications/panimmune)
- CIBERSORTx (https://cibersortx.stanford.edu/index.php)
- the *cor* function from the *stats* package for R
- *corplot* package for R

Drug sensitivity analysis

- RNA-Seq data for MPM in the Genomics of Drug Sensitivity in Cancer (GDSC) project
- ArrayExpress (E-MTAB-3983, https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-3983)
- Area under the curve (AUC)
- ANOVA models
- two-sided p-value

Integrated DNA methylation analysis

- Data from TCGA from GDC
- ArrayExpress (E-MTAB-6884)
- β-values
- COHCAP algorithm implemented through the corresponding R package (*COHCAP* v. 1.48.0)
- p-values
- False discovery rate (FDR)
- Integrative Genomics Viewer (https://www.igv.org/)

Common statistical procedures

- Chi-sqared
- Two-samples t-test
- Wilcoxon-Mann-Whitney test

Table 1. Used omics datasets to build, validate, and explore the biological nature of 2-PS.

Dataset	Data type	Technology	Accession number	Patients (n)	Purpose
TCGA	RNA-Seq	Illumina HiSeq 2000	phs000178	87	Training
TCGA	DNA methylation	Infinium HumanMethylation 450 BeadChip	phs000178	87	Exploratory
Bueno	RNA-Seq	Illumina HiSeq 2000	EGAD00001001915 EGAS000010015631	211	Validation
Blum	Expression microarray	HG-U133 Plus 2.0 chip	E-MTAB-6877	67	Validation
Blum	DNA methylation	Infinium HumanMethylation 450 BeadChip	E-MTAB-6884	67	Exploratory

IV. Results

Identification of studies for inclusion in the analysis

The most easily accessible database of scientific publications in the field of medicine and biology is the US National Library of Medicine which is freely accessible online. Its literature division is referred to as PubMed, for short. As shown in Fig. 1. the number of publications deposited in PubMed on mesothelioma has increased rapidly over the past 30 years, and is currently around 600 per year. For this reason, we chose PubMed as the only source to search literature sources that might contain a description about omics data from MPM patients. Due to the rapid technological development in the field of omics technologies, we decided to limit our search temporally to publications from the period between 01-01-2015 and 31-12-2020. In identifying and screening records from PubMed, we followed the international consensus to perform systematic analyzes of PRISMA meta-analyses, as detailed in Fig. 2.

Figure 1. Number of PubMed entries by year after searching the terms "mesothelioma AND malignant". Last search date 10-12-2023

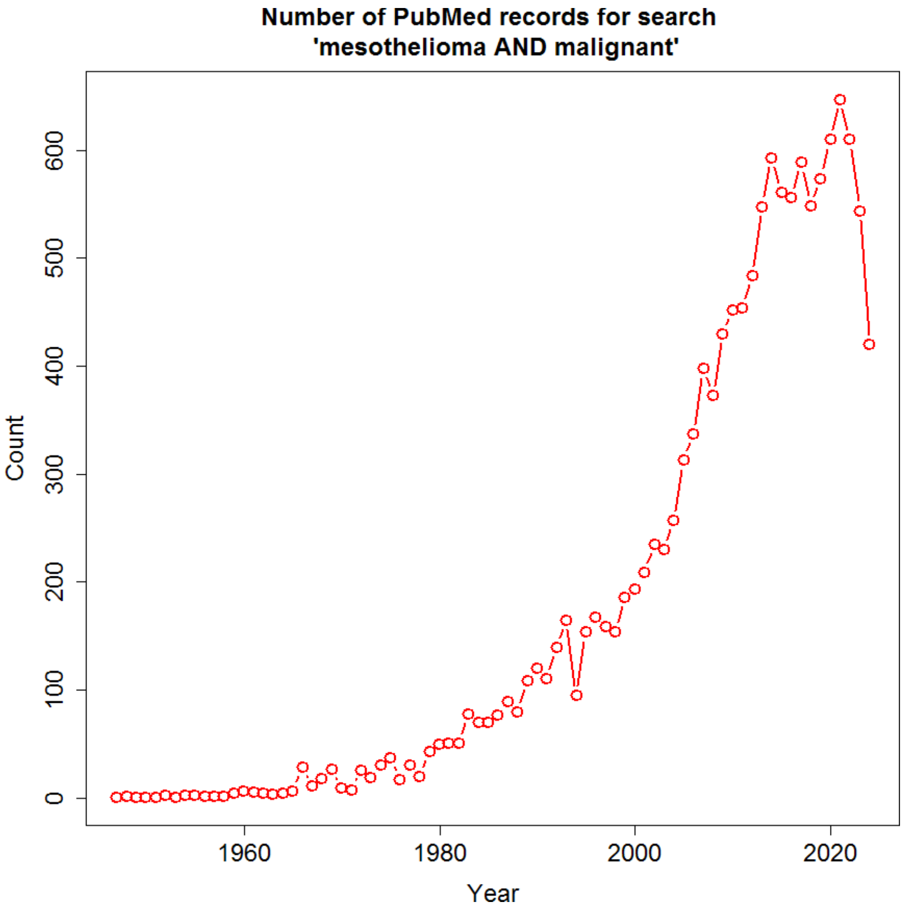
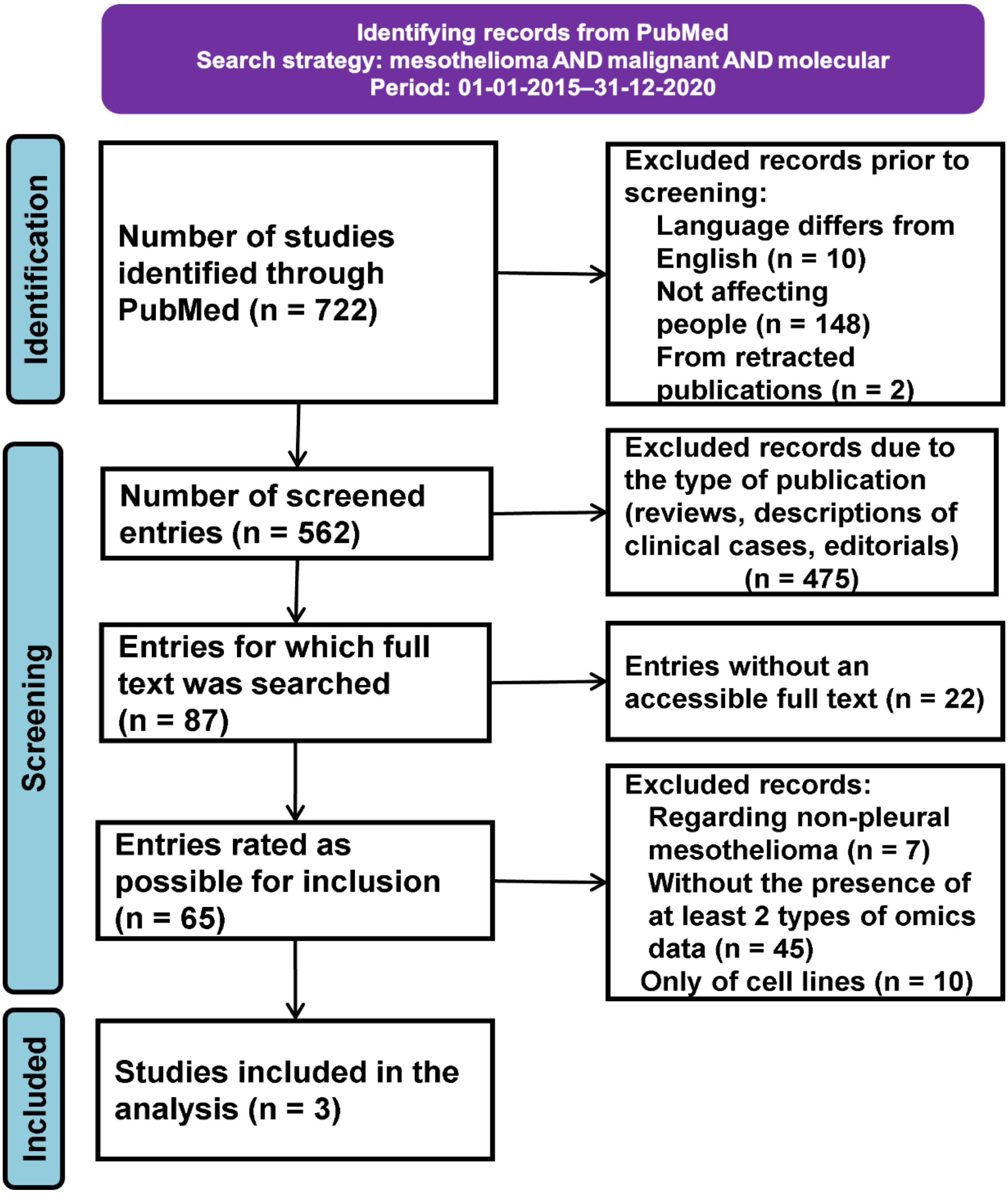


Figure 2. PRISMA diagram of the process of selecting studies for inclusion in the analysis. Based on the systematic approach described in the diagram, we arrived at a selection of 3 studies containing publicly available multi-omics data from MPM patients


Identifying records from PubMed
Search strategy: mesothelioma AND malignant AND molecular
Period: 01-01-2015–31-12-2020
Identification
Number of studies identified through PubMed (n = 722)
Excluded records prior to screening:
Language differs from English (n = 10)
Not affecting people (n = 148)
From retracted publications (n = 2)
Screening
Number of screened entries (n = 562)
Excluded records due to the type of publication (reviews, descriptions of clinical cases, editorials) (n = 475)
Entries for which full text was searched (n = 87)
Entries without an accessible full text (n = 22)
Entries rated as possible for inclusion (n = 65)
Excluded records:
Regarding non-pleural mesothelioma (n = 7)
Without the presence of at least 2 types of omics data (n = 45)
Only of cell lines (n = 10)
Included
Studies included in the analysis (n = 3)

Building and initial performance of a two-gene prognostic score (2-PS)

We applied the Robust Likelihood-Based Survival Modeling with Microarray Data to the training dataset (TCGA dataset) with genes that MPM cell lines were shown to be dependent on (Fig. 3). The algorithm selected the best-performing prognostic model based on the lowest value of the Akaike Information Criterion (AIC). The model chosen consists of two genes – *GOLT1B* and *MAD2L1*. The estimated Cox regression coefficients (ln(HR)) for *GOLT1B* and *MAD2L1* were 1,403 and 0,945, respectively. The continuous score for each sample in every dataset was calculated as the sum of expression values for each of the genes in the model multiplied by the regression coefficient. In univariate analysis, the continuous score was prognostic for the overall survival (Fig. 5). We further defined a binary score using as a cut-off the median of the continuous score for all samples (Fig. 4). In univariate analysis, the binary score also showed a significant prognostic value with Area Under the Curve value of the Receiver Operator Characteristics (ROC) analysis of 0,67 (Fig. 6). Additionally, we evaluated the performance of the binary score in a multivariate model with age, sex, stage, histology and mutational status as covariates while retaining independent prognostic value (Fig. 7).

Figure 3. Diagram of the selected analytical approach.

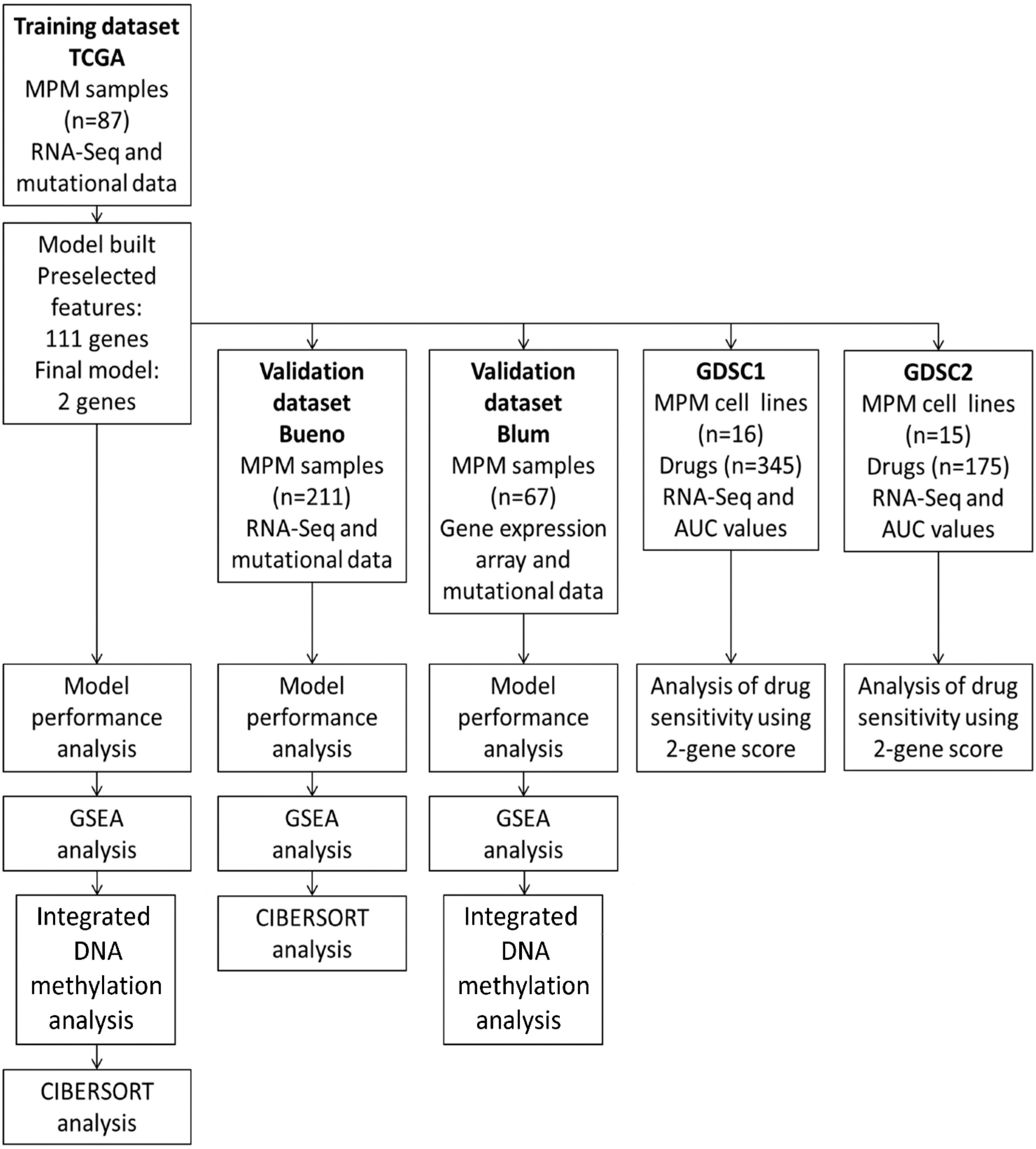

Training dataset TCGA
MPM samples (n=87)
RNA-Seq and mutational data
Model built
Preselected features: 111 genes
Final model: 2 genes
Validation dataset Bueno
MPM samples (n=211)
RNA-Seq and mutational data
Validation dataset Blum
MPM samples (n=67)
Gene expression array and mutational data
GDSC1
MPM cell lines (n=16)
Drugs (n=345)
RNA-Seq and AUC values
GDSC2
MPM cell lines (n=15)
Drugs (n=175)
RNA-Seq and AUC values
Model performance analysis
Model performance analysis
Model performance analysis
Analysis of drug sensitivity using 2-gene score
Analysis of drug sensitivity using 2-gene score
GSEA analysis
GSEA analysis
GSEA analysis
Integrated DNA methylation analysis
CIBERSORT analysis
Integrated DNA methylation analysis
CIBERSORT analysis

Figure 4. Distribution histogram of the continuous score for the patients from the TCGA cohort. The red vertical line shows the median value used for the dichotomous (binary) stratification into groups of high- and low-scoring patients.

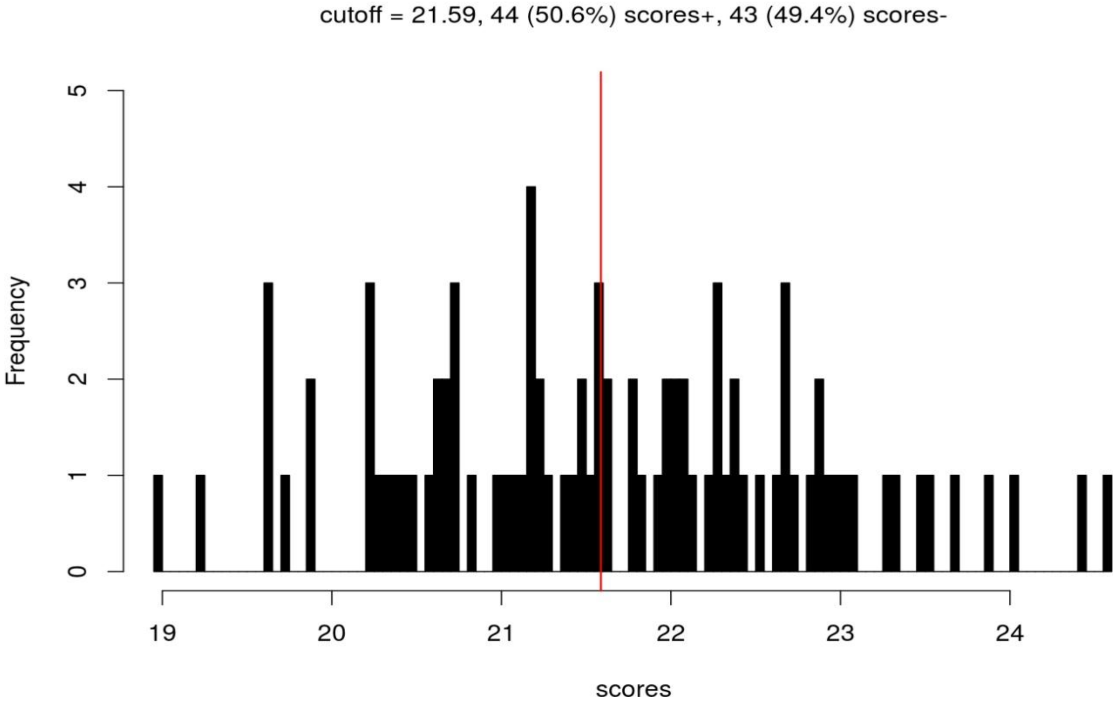


Figure 5. Univariate analysis of survival according to 2-PS in the TCGA cohort. Patients are stratified into a high score and a low score group based on the median of the continuous score. The p-value is from Cox regression analysis

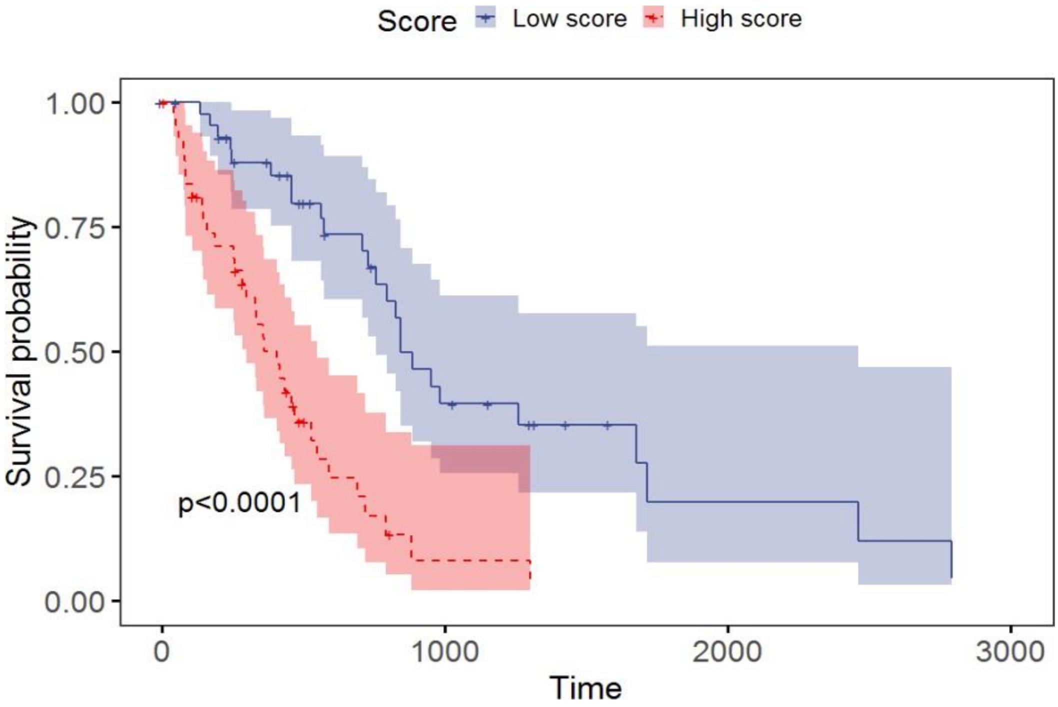


Figure 6. ROC curve analysis for the prognostic value of the binary score in terms of overall survival in the TCGA cohort. Area under the curve (AUC) as well as sensitivity and specificity were calculated at a median value of 21,59, which was used for dichotomous separation.

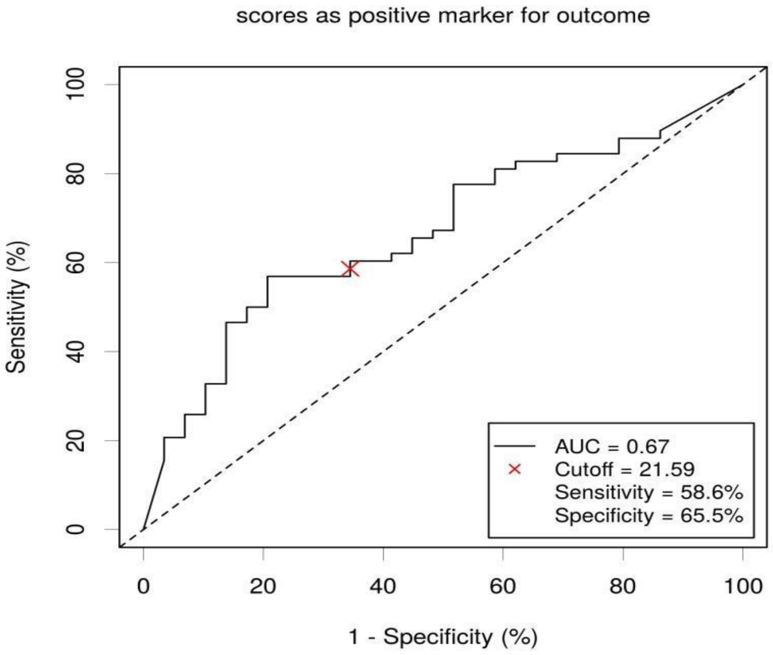
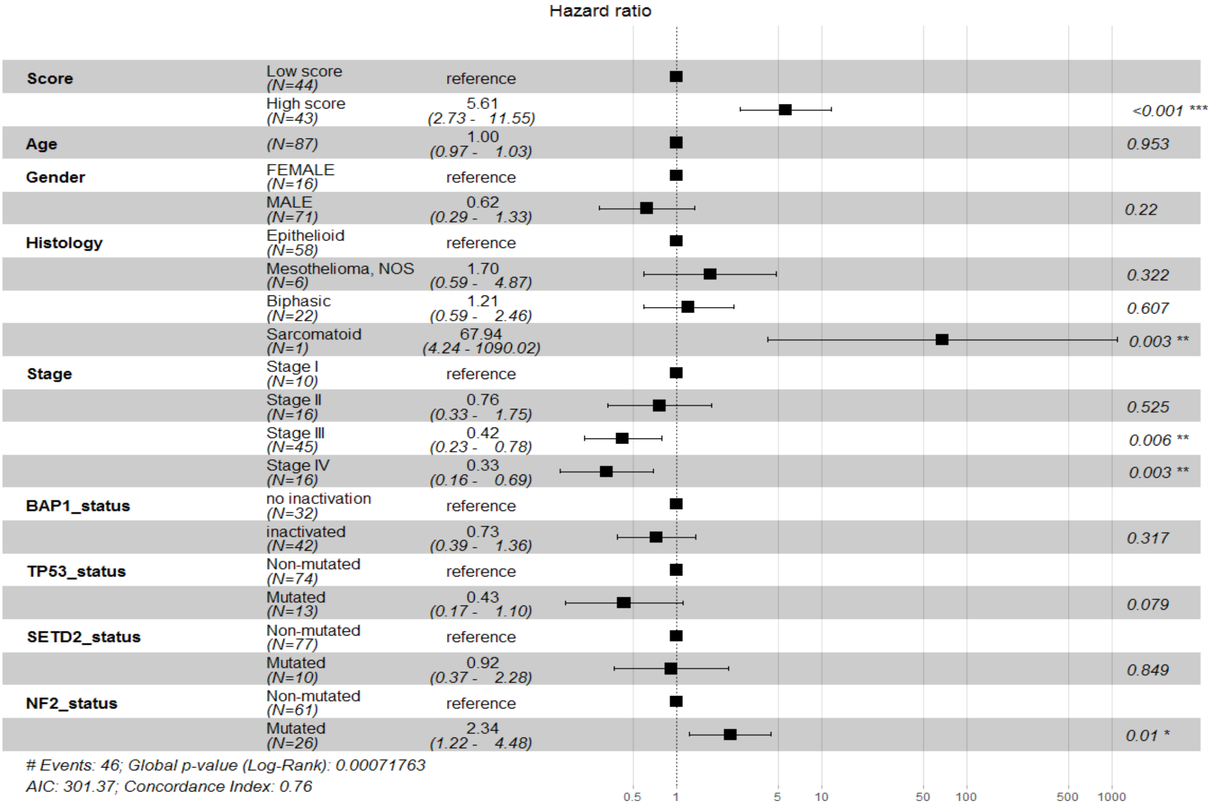


Figure 7. Multivariate analysis of survival according to the 2-PS in the TCGA cohort. Covariates included are age, sex, histology, stage and mutational status pertaining to 4 genes: BAP1, TP53, SETD2 and NF2.



Score	Low score *(N=44)*	reference	
	High score *(N=43)*	5.61 *(2.73 - 11.55)*	*<0.001 ****
Age	*(N=87)*	1.00 *(0.97 - 1.03)*	*0.953*
Gender	FEMALE *(N=16)*	reference	
	MALE *(N=71)*	0.62 *(0.29 - 1.33)*	*0.22*
Histology	Epithelioid *(N=58)*	reference	
	Mesothelioma, NOS *(N=6)*	1.70 *(0.59 - 4.87)*	*0.322*
	Biphasic *(N=22)*	1.21 *(0.59 - 2.46)*	*0.607*
	Sarcomatoid *(N=1)*	67.94 *(4.24 - 1090.02)*	*0.003 ***
Stage	Stage I *(N=10)*	reference	
	Stage II *(N=16)*	0.76 *(0.33 - 1.75)*	*0.525*
	Stage III *(N=45)*	0.42 *(0.23 - 0.78)*	*0.006 ***
	Stage IV *(N=16)*	0.33 *(0.16 - 0.69)*	*0.003 ***
BAP1_status	no inactivation *(N=32)*	reference	
	inactivated *(N=42)*	0.73 *(0.39 - 1.36)*	*0.317*
TP53_status	Non-mutated *(N=74)*	reference	
	Mutated *(N=13)*	0.43 *(0.17 - 1.10)*	*0.079*
SETD2_status	Non-mutated *(N=77)*	reference	
	Mutated *(N=10)*	0.92 *(0.37 - 2.28)*	*0.849*
NF2_status	Non-mutated *(N=61)*	reference	
	Mutated *(N=26)*	2.34 *(1.22 - 4.48)*	*0.01 **

# Events: 46; Global p-value (Log-Rank): 0.00071763
AIC: 301.37; Concordance Index: 0.76

Validation of the 2-PS

To validate our 2-gene prognostic score we used two recent publicly available datasets with RNA-Seq data (n=211) (Bueno) and planar expression array (n=67) (Blum). The estimated continuous score in the Bueno dataset showed a clear prognostic value and was further converted to a binary one using the median of the continuous score as a cut-off. Analogous to the training dataset, the binary score in this validation dataset also had prognostic power (Fig. 8) with AUC of the ROC analysis of 0,75. (Fig. 10). Similar to the extensive multivariate model for the Bueno dataset, the binary score (Fig. 9) was still of independent prognostic value (Fig. 11).

.

Figure 8. Distribution histogram of the continuous score for the patients from the Bueno cohort. The red vertical line shows the median value used for the dichotomous (binary) stratification into groups of high- and low-scoring patients.

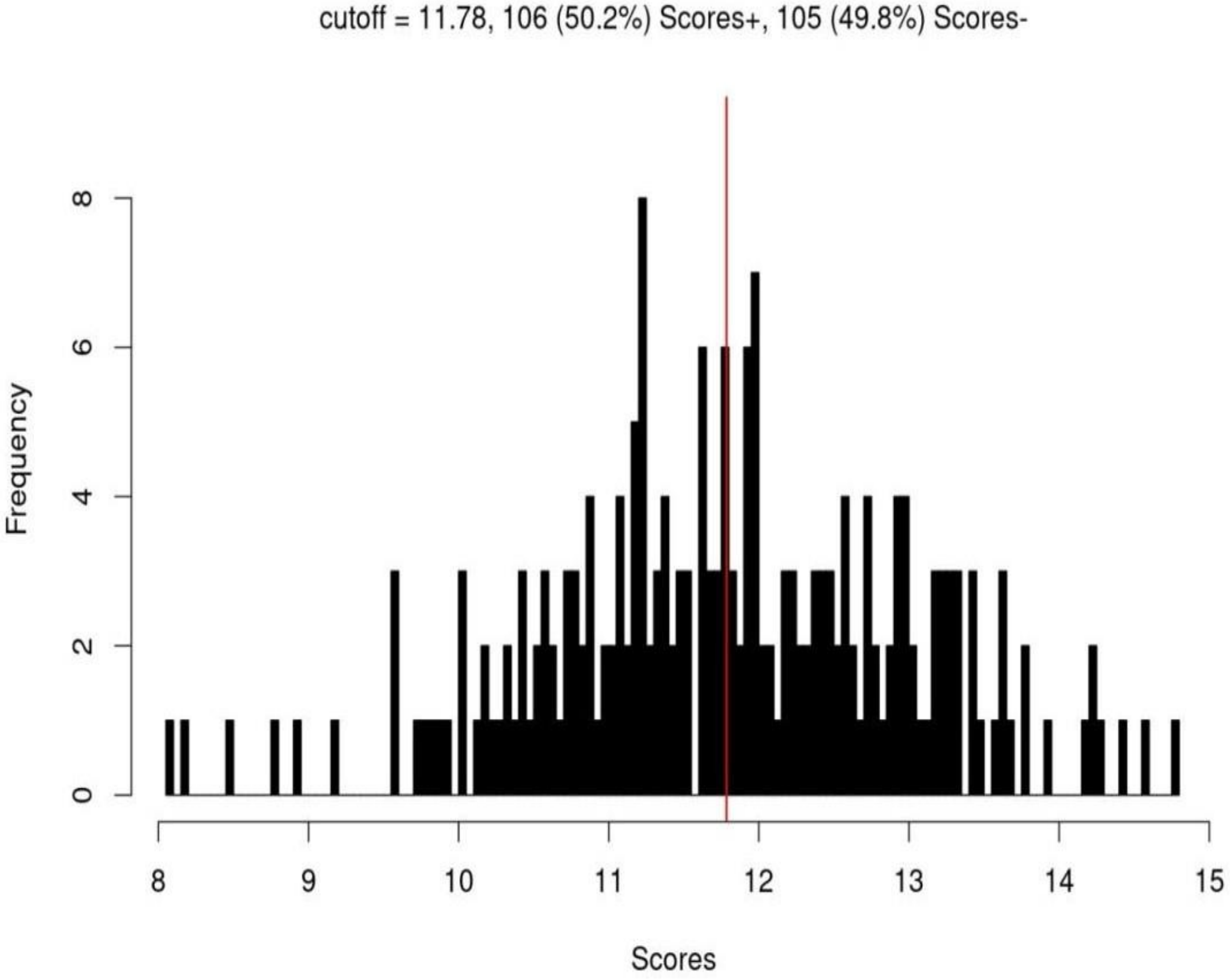


Figure 9. Univariate analysis of survival according to 2-PS in the Bueno cohort. Patients are stratified into a high score and a low score group based on the median of the continuous score. The p-value is from Cox regression analysis

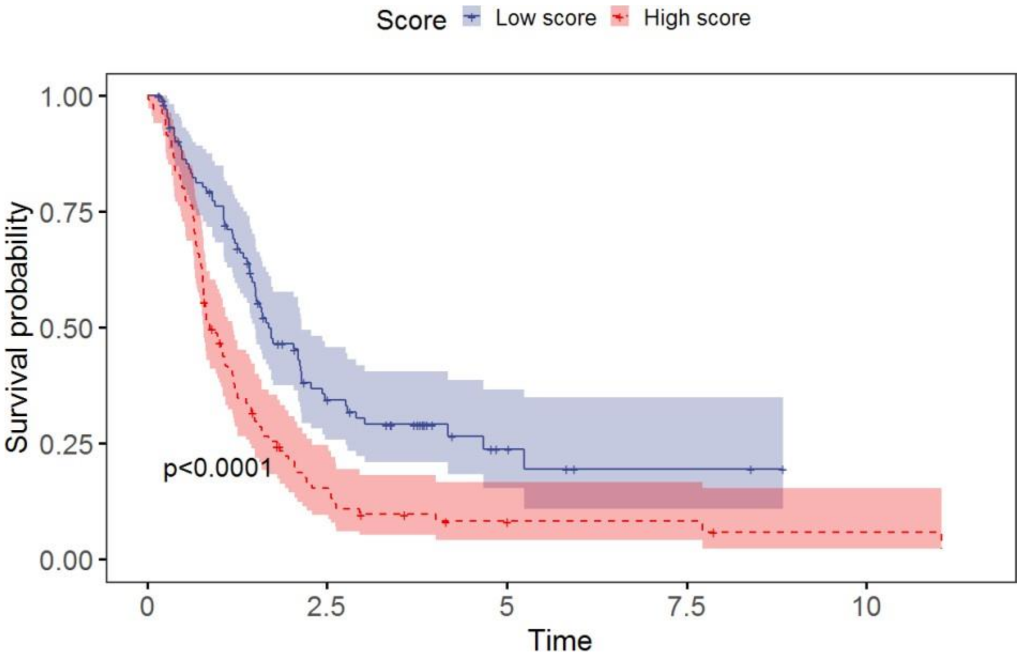
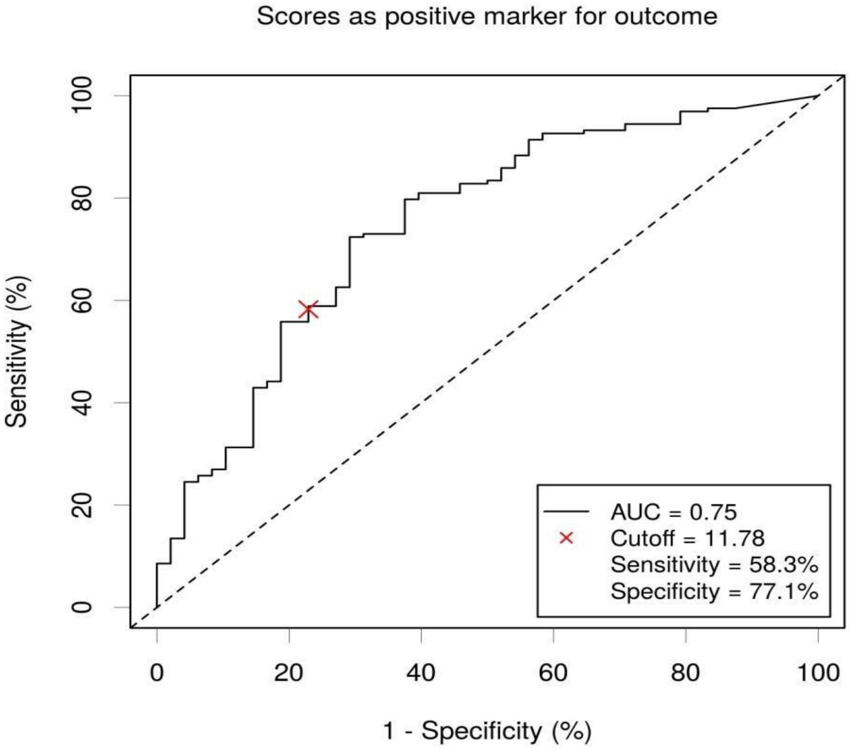
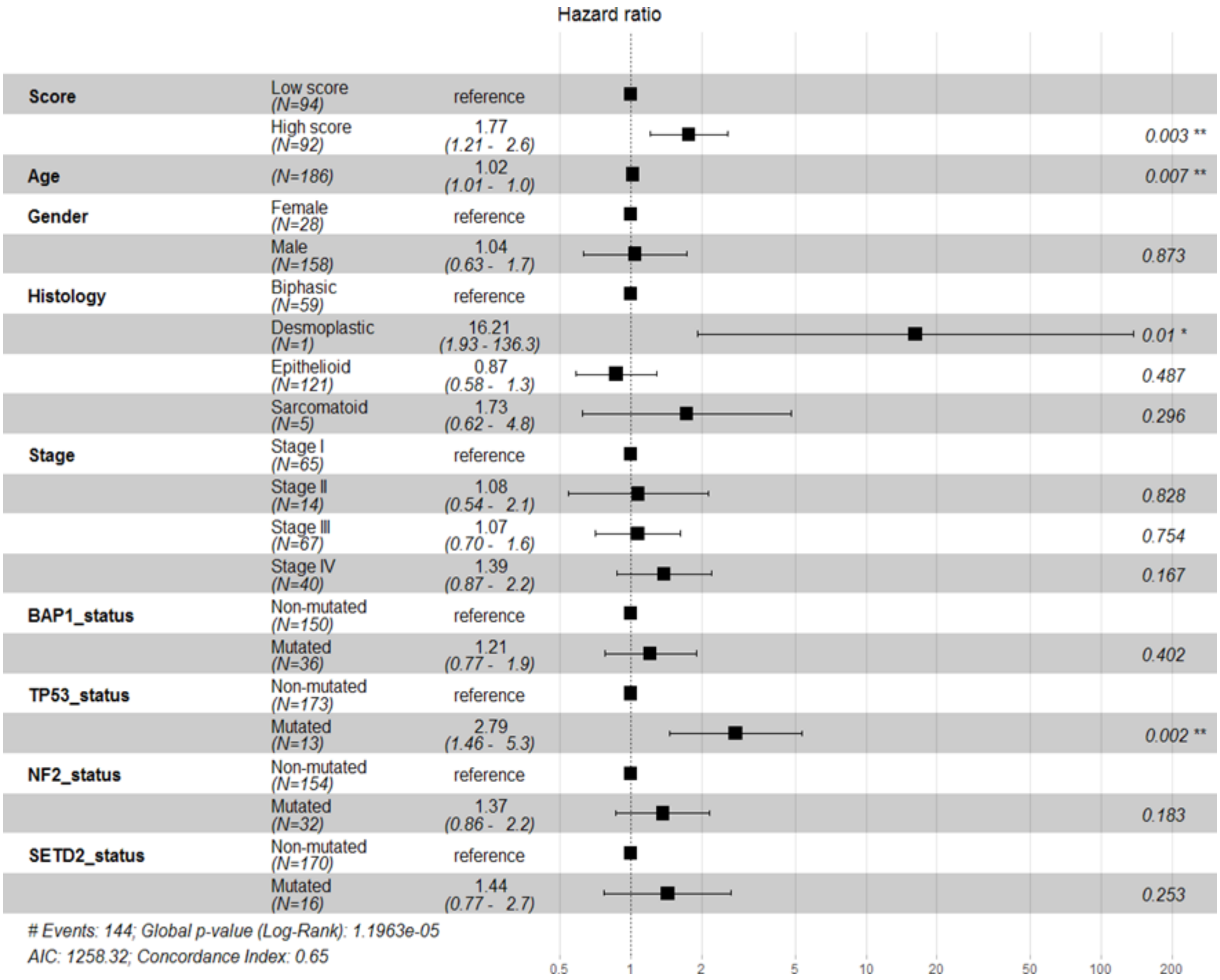


Figure 10. ROC curve analysis for the prognostic value of the binary score in terms of overall survival in the Bueno cohort. Area under the curve (AUC) as well as sensitivity and specificity were calculated at a median value of 11,78, which was used for dichotomous separation.



.

Figure 11. Multivariate analysis of survival according to the 2-PS in the Bueno cohort. Covariates included are age, sex, histology, stage and mutational status pertaining to 4 genes: BAP1, TP53, SETD2 and NF2



Following the same procedure, we analyzed the performance of the estimated continuous score in the Blum data set. In univariate analysis, the continuous and binary score (Fig. 12) defined by the cut-off of the median for the continuous score, was proven a significant independent prognostic factor with AUC of the ROC curve analysis of 0,85 (Fig. 14). The multivariate model for the Blum dataset was built using the binary score (Fig. 13), sex, age, stage, and histological subtype of the disease. The mutational data was excluded, as those were not publicly available. The model demonstrated the independent prognostic value of the binary score (Fig. 15).

Figure 12. Distribution histogram of the continuous score for the patients from the Blum cohort. The red vertical line shows the median value used for the dichotomous (binary) stratification into groups of high- and low-scoring patients.

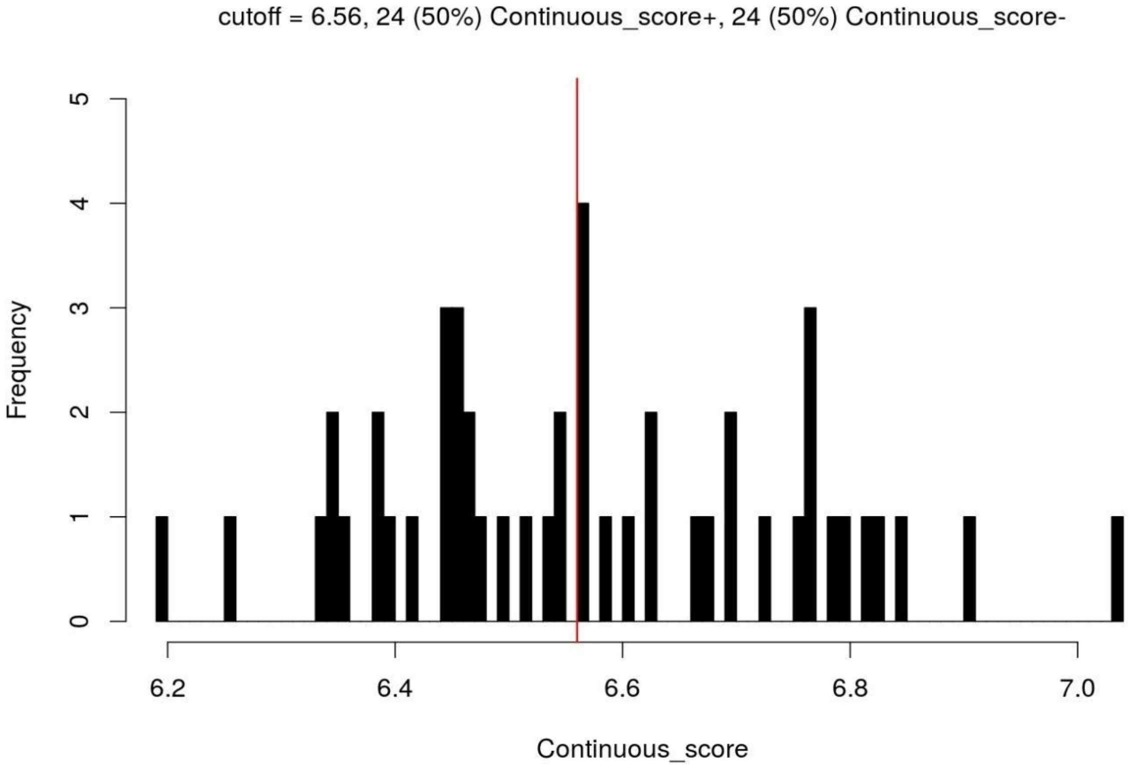


Figure 13. Univariate analysis of survival according to 2-PS in the Blum cohort. Patients are stratified into a high score and a low score group based on the median of the continuous score. The p-value is from Cox regression analysis

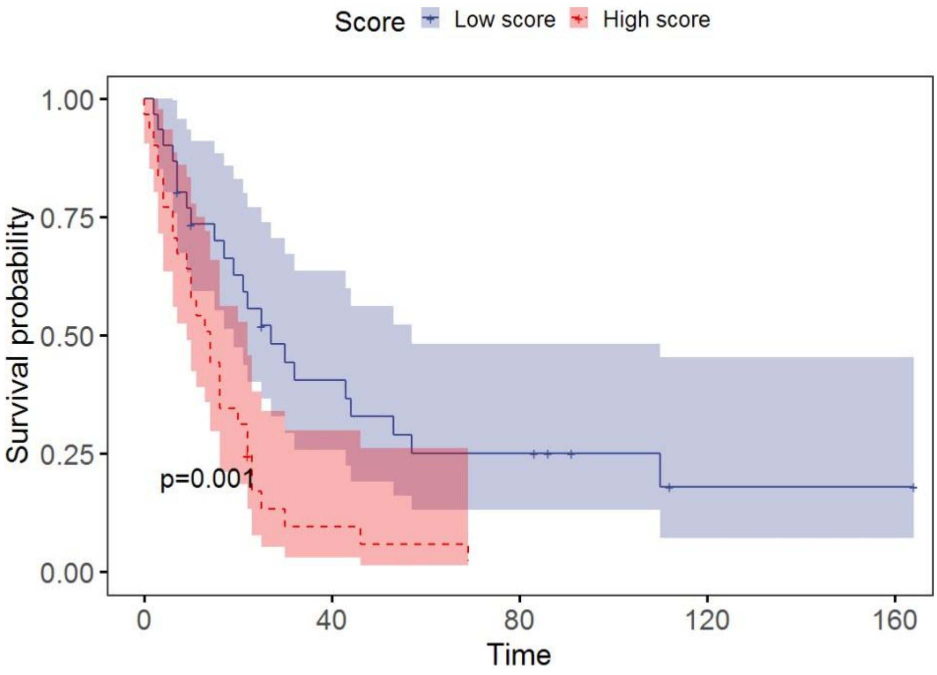


Figure 14. ROC curve analysis for the prognostic value of the binary score in terms of overall survival in the Blum cohort. Area under the curve (AUC) as well as sensitivity and specificity were calculated at a median value of 6,56, which was used for dichotomous separation.

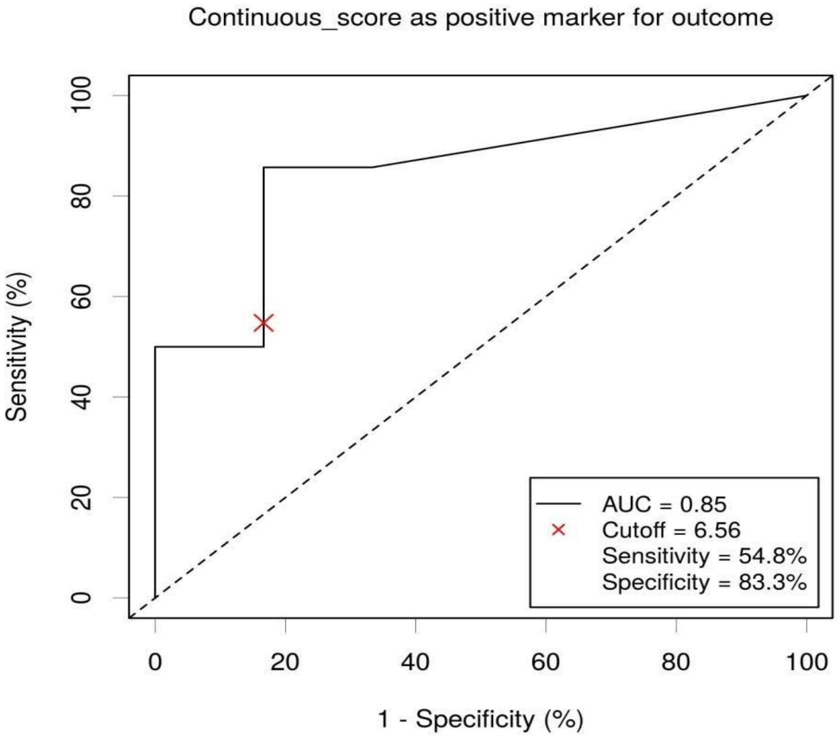
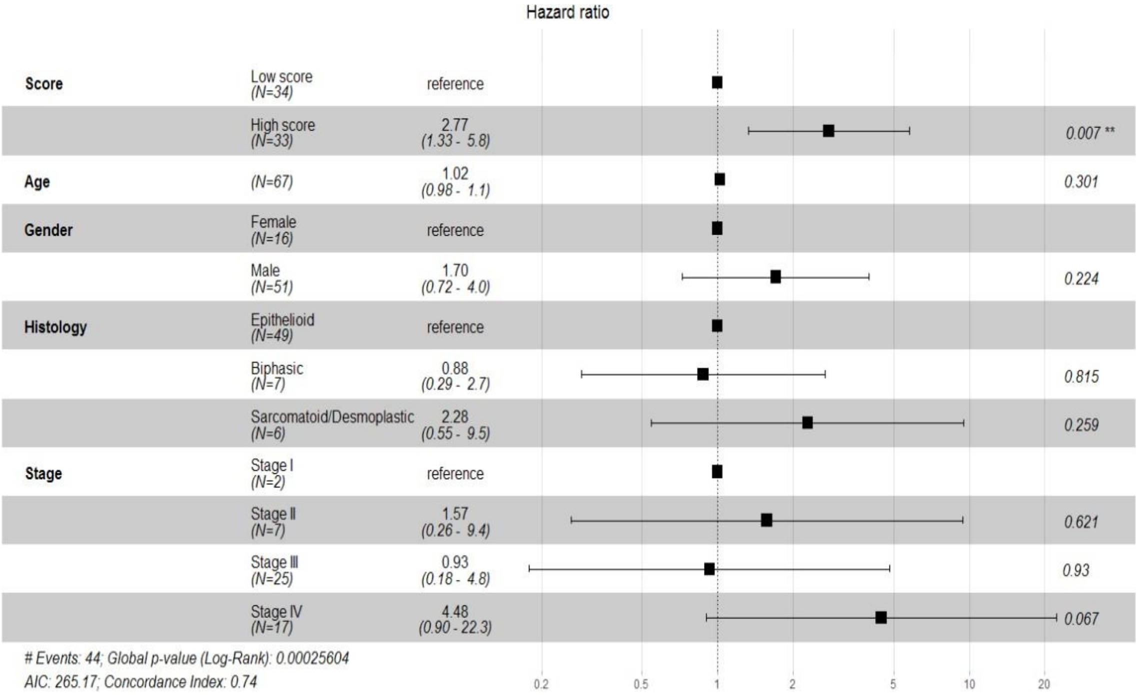


Figure 15. Multivariate analysis of survival according to the 2-PS in the Blum dataset. Covariates included are age, sex, histology, stage and mutational status pertaining to 4 genes: BAP1, TP53, SETD2 and NF2.



Gene Set Enrichment Analysis in Expression Profiles (GSEA)

Based on the observation that our novel 2-PS exhibited a similar prognostic value in both training and the validation datasets, the score may correlate with specific gene expression profile. We performed GSEA using predefined cancer hallmarks signatures corresponding to the main characteristics of cancer from the MSig database. For each of the datasets we obtained several enriched signatures in the high score patient subgroups as follows: TCGA (n=37), Bueno (n=34), and Blum (n=34). Example plots showing enrichment of the "mitotic spindle" profile in both patients with high scores and all three cohorts are presented on Fig. 16, 17 and 18. It was evident that there was a more significant overlap between the overexpressed signatures in the TCGA and Bueno and slightly less so between any of those two and the Blum dataset. However, there were a total of 25 signatures that were commonly overexpressed in high-score patients from the three cohorts (Fig. 19 and Fig. 20). Most of them were related to DNA repair and DNA damage response and control of mitotic cell division.

Figure 16. Example result of the GSEA analysis on the TCGA dataset. (A) Enrichment plot for genes from the "Mitotic spindle" list; (B) Gene expression heatmap of the same list. Patients with a high score are shown in gray and those with a low score are shown in yellow.

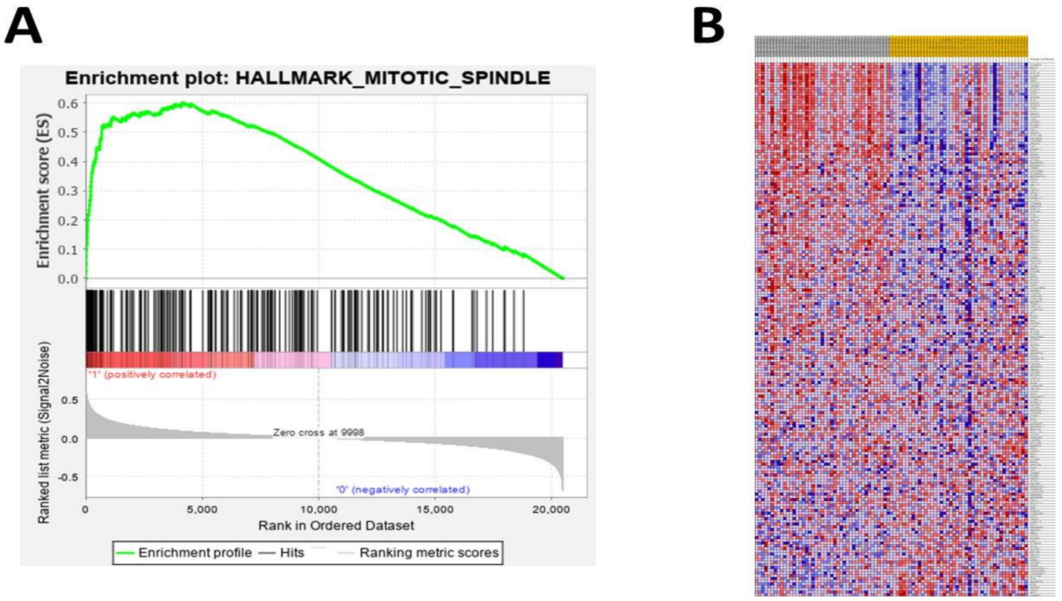


Figure 17. Example result of the GSEA analysis on the Bueno dataset. (A) Enrichment plot for genes from the "Mitotic spindle" list; (B) Gene expression heatmap of the same list. Patients with a high score are shown in gray and those with a low score are shown in yellow.

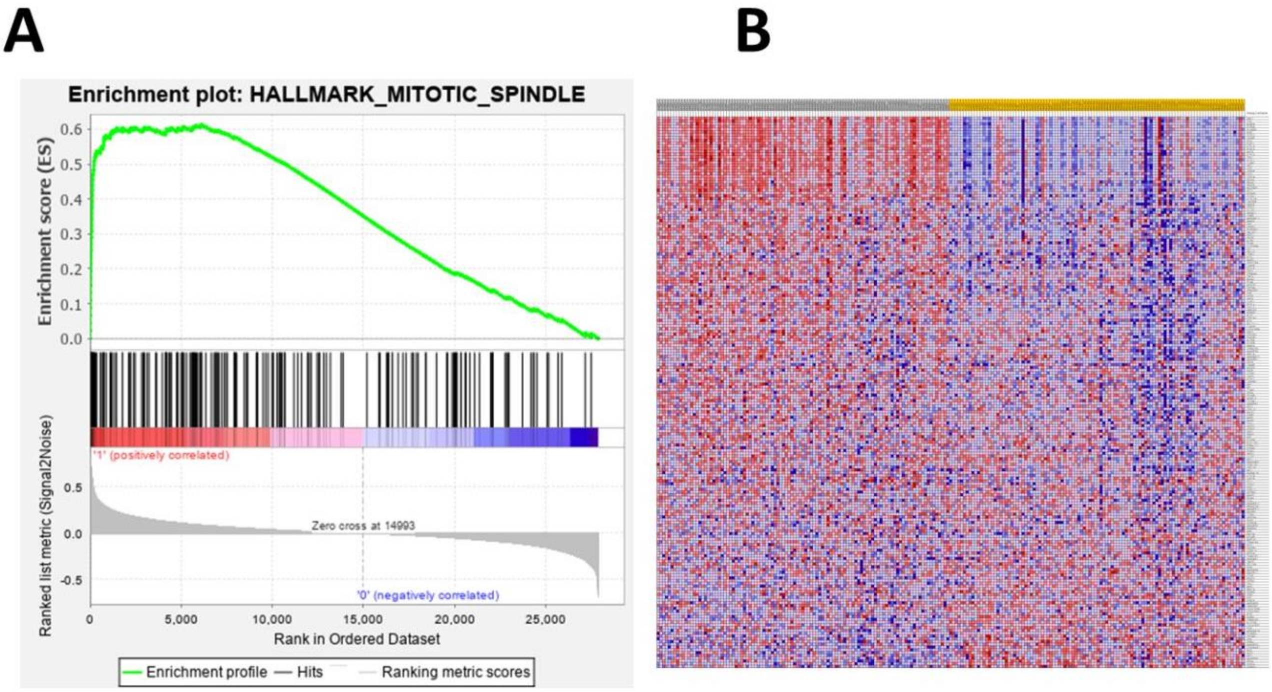


Figure 18. Example result of the GSEA analysis on the Blum dataset. (A) Enrichment plot for genes from the "Mitotic spindle" list; (B) Gene expression heatmap of the same list. Patients with a high score are shown in gray and those with a low score are shown in yellow.

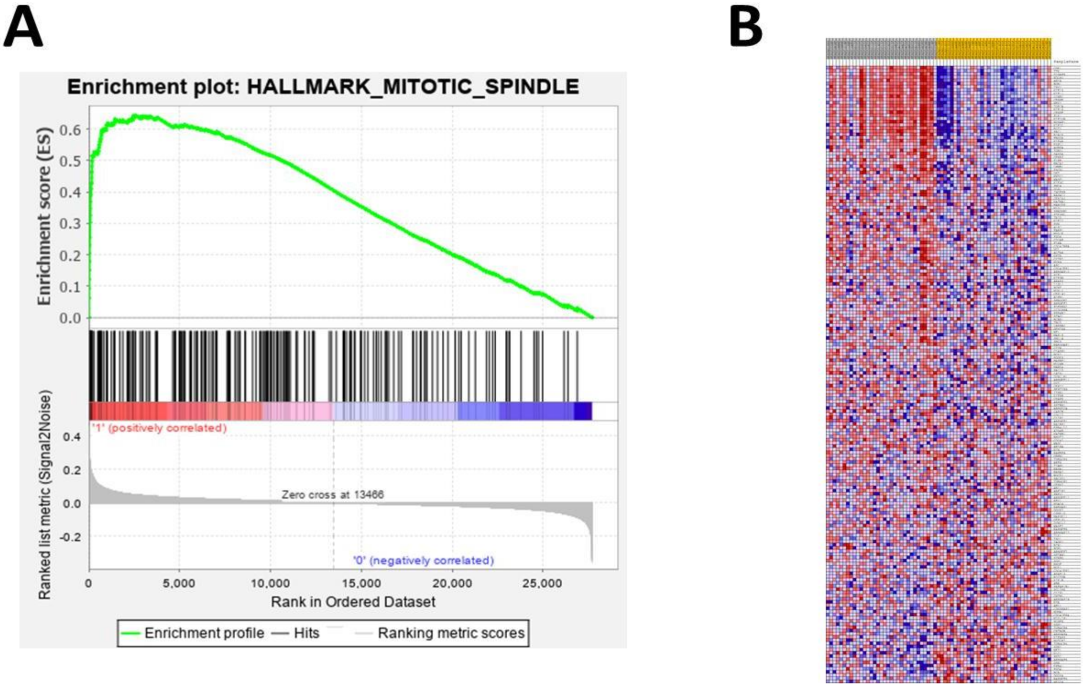


Figure 19. Bar charts of the significantly more frequently overexpressed gene lists in the patients with high scores found in at least one of the three cohorts (TCGA, Bueno, Blum) determined by GSEA analysis. Numbers indicate the count of gene lists in each of the sets.

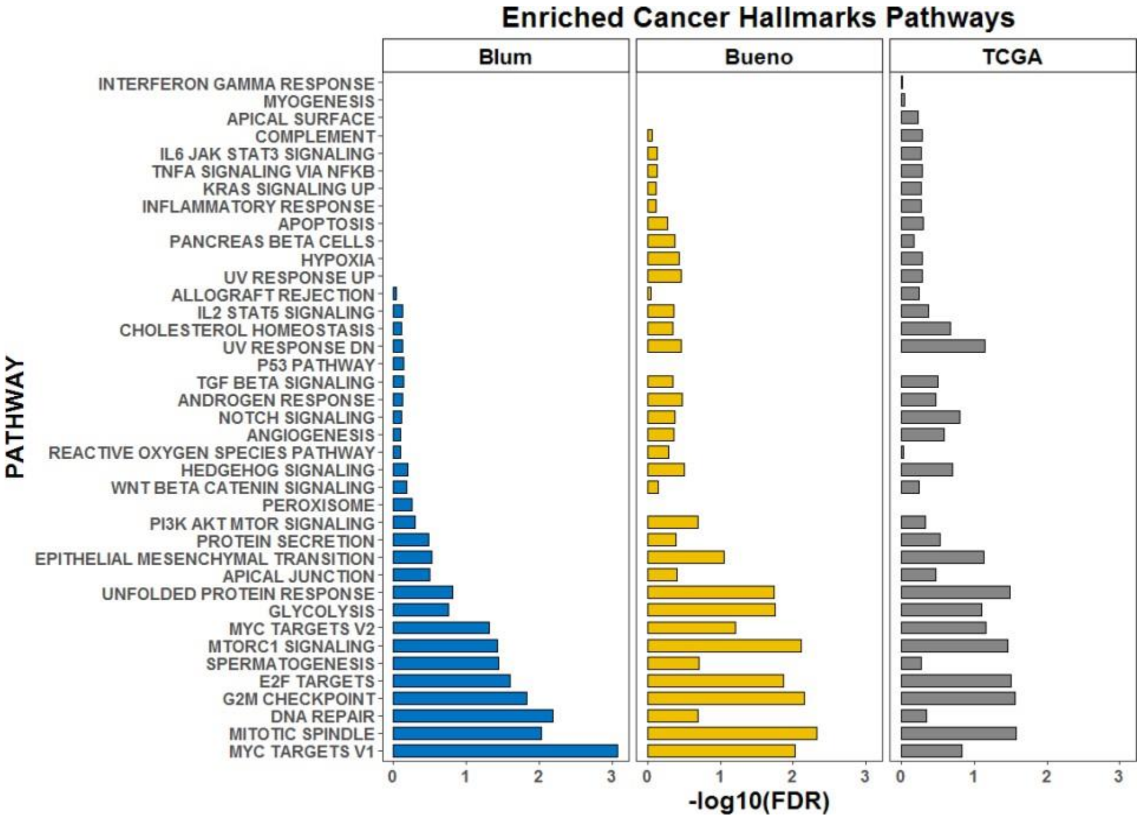
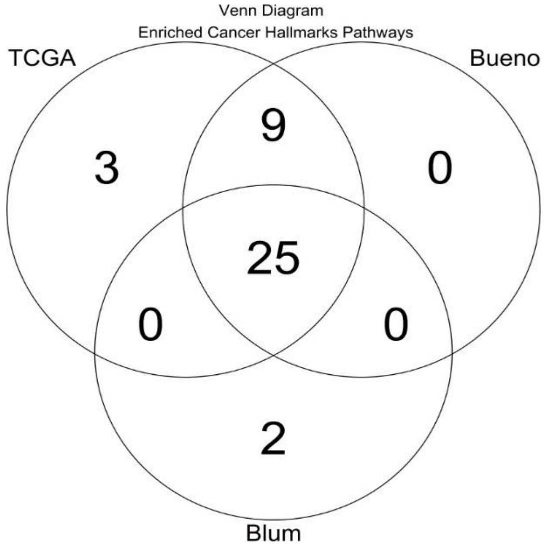


Figure 20. Venn diagram of the significantly overlapping more frequently overexpressed gene lists in the patients with a high score from the three cohorts (TCGA, Bueno, Blum) determined by GSEA analysis. Numbers indicate the count of gene lists in each of the sets.



Integrated DNA methylation and gene expression analysis

We analyzed whether the score defined epigenetically distinct disease subtypes using available DNA methylation data obtained by Infinium 450K bead chip (Illumina) for patients from the TCGA and Blum cohorts. For the analysis, we used the COHCAP algorithm, which allows the integration of DNA methylation data with gene expression data to identify CpG sites and islands that are differentially methylated, but the level of methylation correlates inversely with the level of gene expression of adjacent genes, which helps identify potentially functionally relevant sites for epigenetic regulation of gene expression. DNA methylation profiles of 87 patients from the TCGA cohort were analyzed, identifying 428 differentially methylated CpG sites between high and low score patients (Fig. 21), located in three differentially methylated CpG islands (Table 2), 2 of which inversely correlated in the gene expression of the adjacent genes – *SLC20A1* and *KIAA1949* (Fig. 22 and 23, Table 3).

Figure 21. Result of the analysis of the differentially methylated CpG sites in the TCGA dataset. Heatmap of β-values for differentially expressed CpG sites (vertical) versus patients (horizontal)

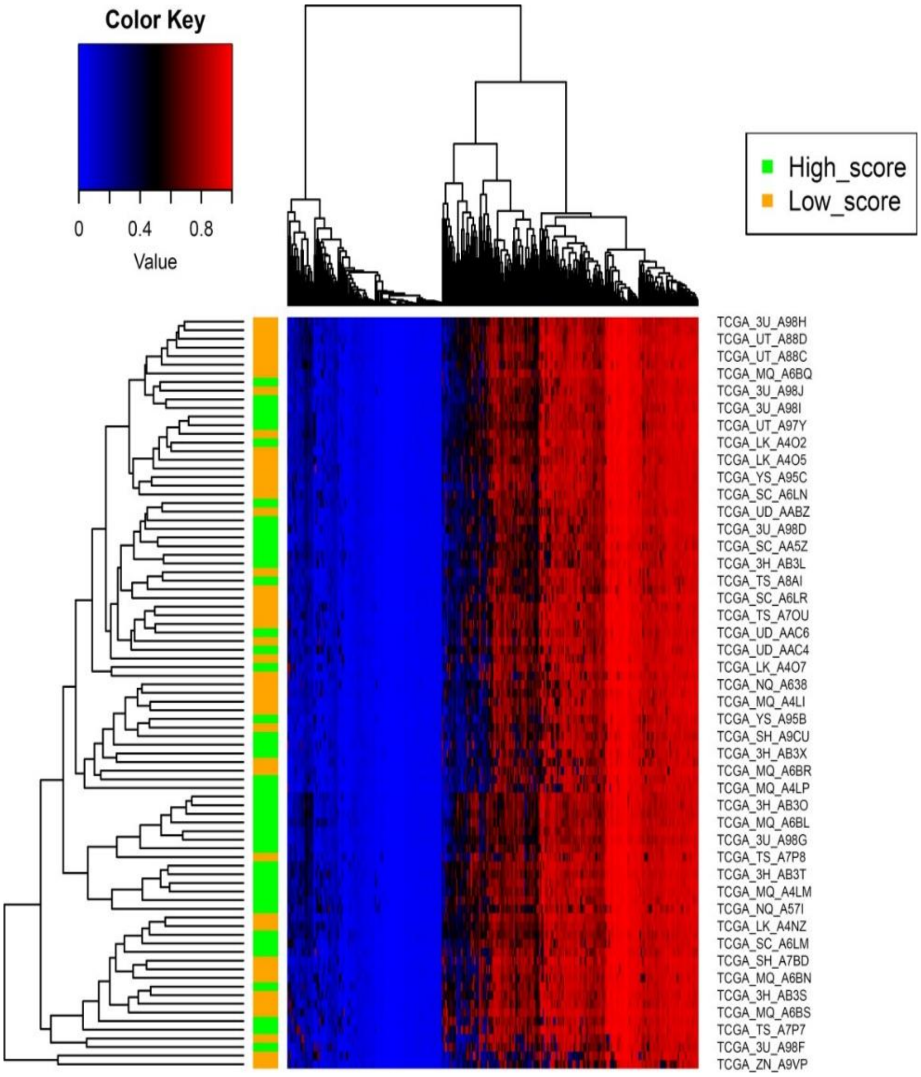


Table 2. Analysis results for differentially methylated CpG islands from the TCGA dataset. For analysis purposes, an island is defined as at least two contiguous CpG sites that were included on Illumina's Infinium 450K chip

CpG island coordinates	Genes	Mean β value for patients with a high score	Mean β value for patients with a low score	Difference in mean β values High vs. Low score	Nominal p-value for the island	FDR value for the island	Number of CpG sites
chr11:2923301-2923817	*SLC22A18; SLC22A18AS*	0,166689202	0,342562598	-0,175873397	1,15E-06	3,46E-06	2
chr2:113403001-113404079	*SLC20A1*	0,217591513	0,364831843	-0,14724033	9,22E-05	0,00013824	2
chr6:30654392-30654934	*KIAA1949*	0,216549012	0,381515783	-0,164966771	0,000248645	0,00024864	13

Table 3. Results of the integrated analysis of differentially methylated CpG islands and gene expression from the TCGA dataset. For analysis purposes, an island was defined as at least two contiguous CpG sites that were included on Illumina's Infinium 450K chip.

CpG island coordinates	Direction of methylation	Genes	Correlation coefficient	Nominal p-value	FDR value
chr2:113403001-113404079	Decreased methylation	**SLC20A1**	**-0,364768107**	**0,0005135**	**0,000513**
chr6:30654392-30654934	Decreased methylation	**KIAA1949**	**-0,778921734**	**6,55E-19**	**1,31E-18**

Figure 22. Two-dimensional correlation plot of methylation of the CpG island chr2:113403001-113404079 versus the SLC20A1 gene expression in the TCGA dataset. Patients with high and low scores are indicated in a different color.

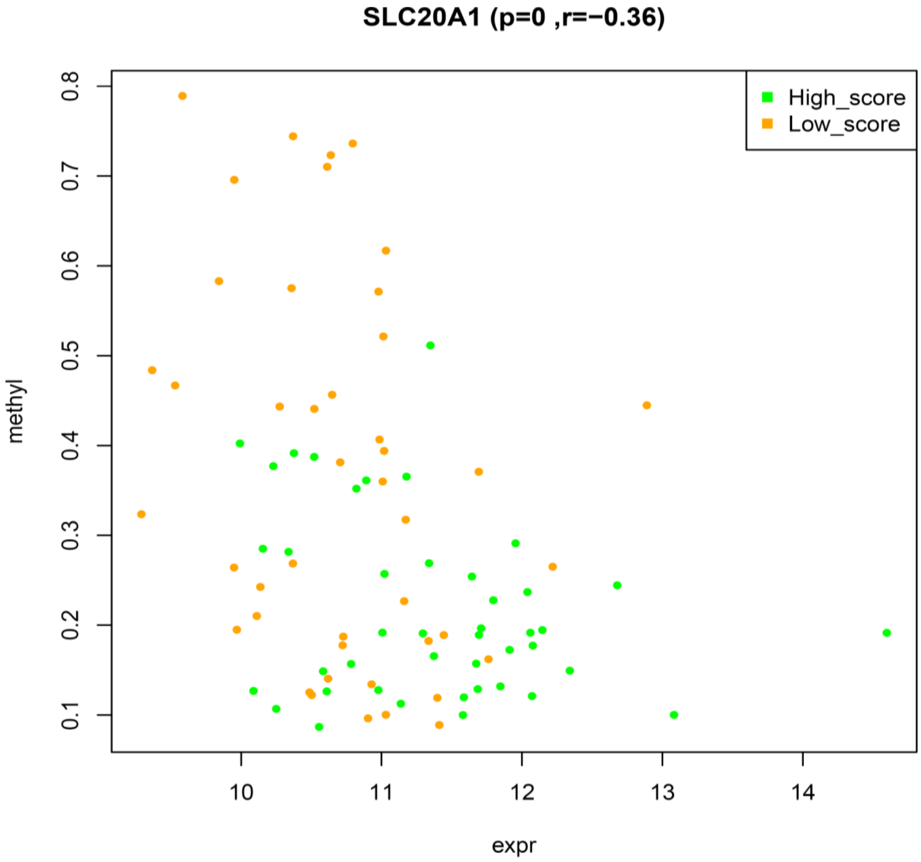
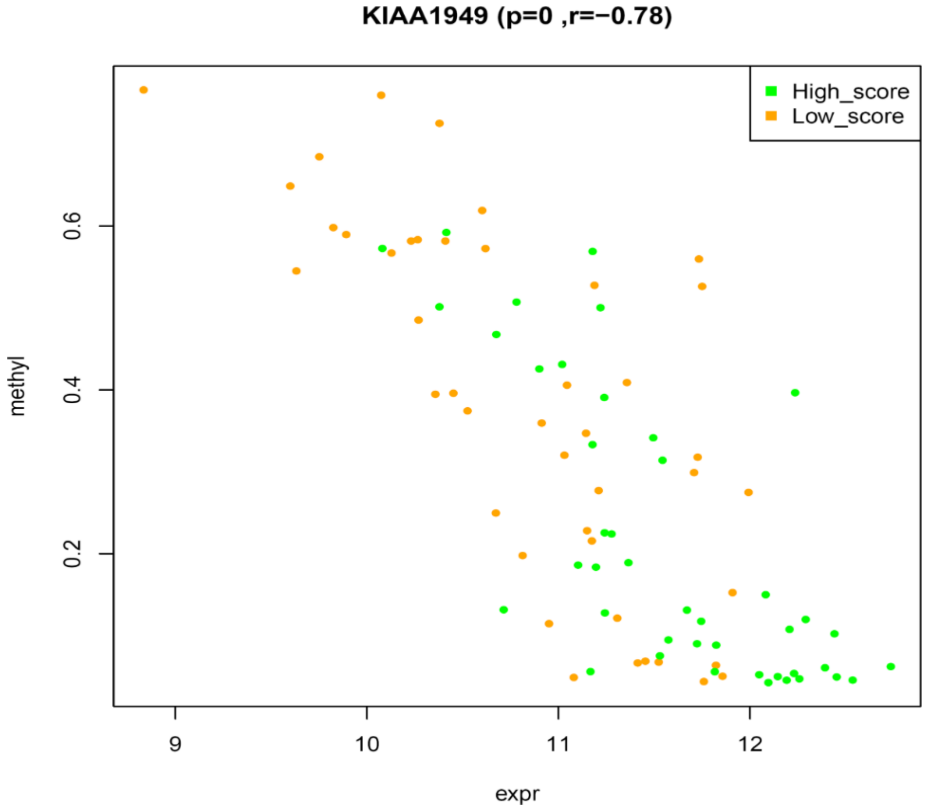


Figure 23. Two-dimensional correlation plot of methylation of the CpG island chr6:30654392-30654934 versus the KIAA1949 gene expression in the TCGA dataset. Patients with high and low scores are indicated in a different color.



Using the same method, we analyzed 62 samples from the Blum dataset. Thus, 526 differentially methylated CpG sites were identified (Fig. 24), of which only a small fraction (n=38) was identical to those from the TCGA dataset (Fig. 25). 39 CpG islands were identified that showed differential methylation and an inverse correlation with gene expression levels of adjacent genes.

Figure 24. Result of the analysis of the differentially methylated CpG sites in the Blum dataset. Heatmap of β-values for differentially expressed CpG sites (vertical) versus patients (horizontal)

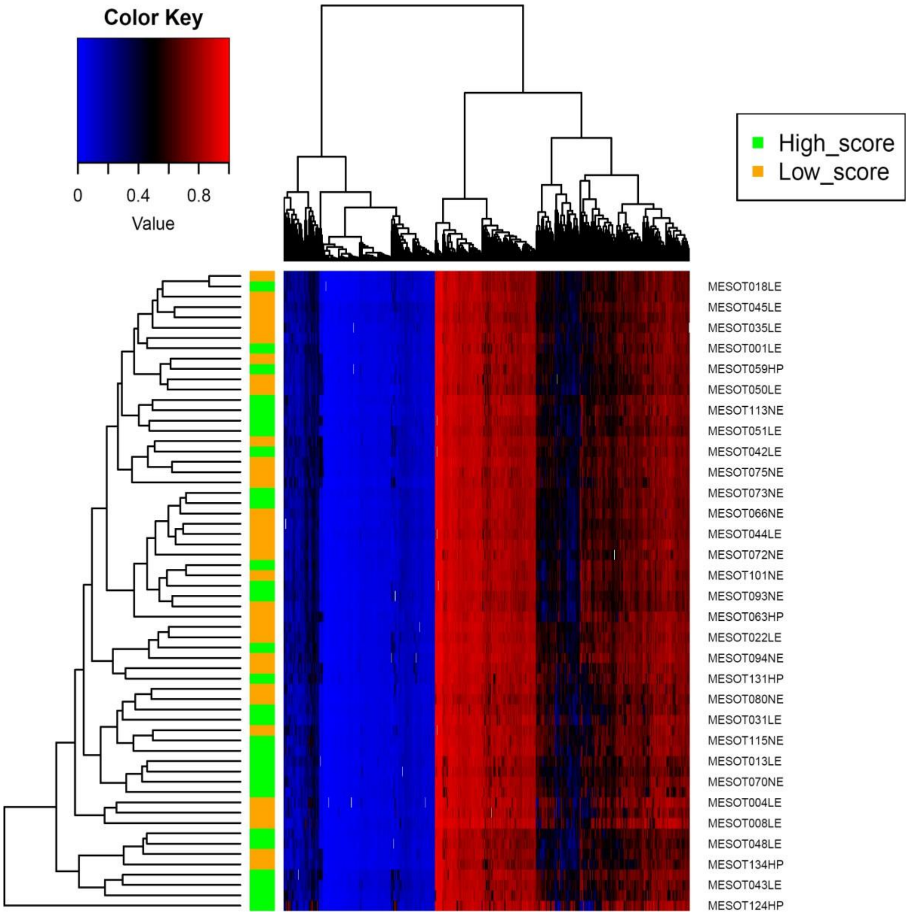
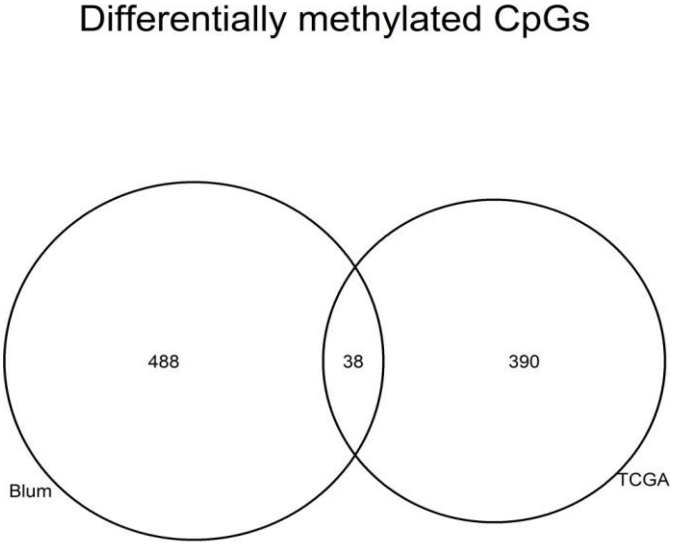


Figure 25. Venn diagram of overlapping significantly differentially methylated CpG sites in high versus low score patients from two of the cohorts (TCGA, Blum) determined by COHCAP analysis. Numbers indicate the count of gene lists in each of the sets.



Correlation with estimated populations of infiltrating immune cells

Infiltrating immune cells are a major player in the immune response against cancer and can be used as a prognostic and predictive marker. We analyzed whether the 2-PS correlated with specific immune cell subtype infiltration in MPM. We used the inferred infiltrating immune cell fractions using the CIBERSORT algorithm using TCGA (Fig. 26) and Bueno (Fig. 27) datasets. Notably for both datasets, the continuous prognostic score showed a positive correlation with CD8+ T cell fraction as well as with M1 and M2 macrophage fractions.

Figure 26. Correlation matrix of continuous score with immune infiltration score for specific immune system cell populations according to expression data from the TCGA cohort. The CIBERSORT algorithm was used to calculate the immune scores. The color of each square of the correlation matrix reflects the calculated correlation coefficient.

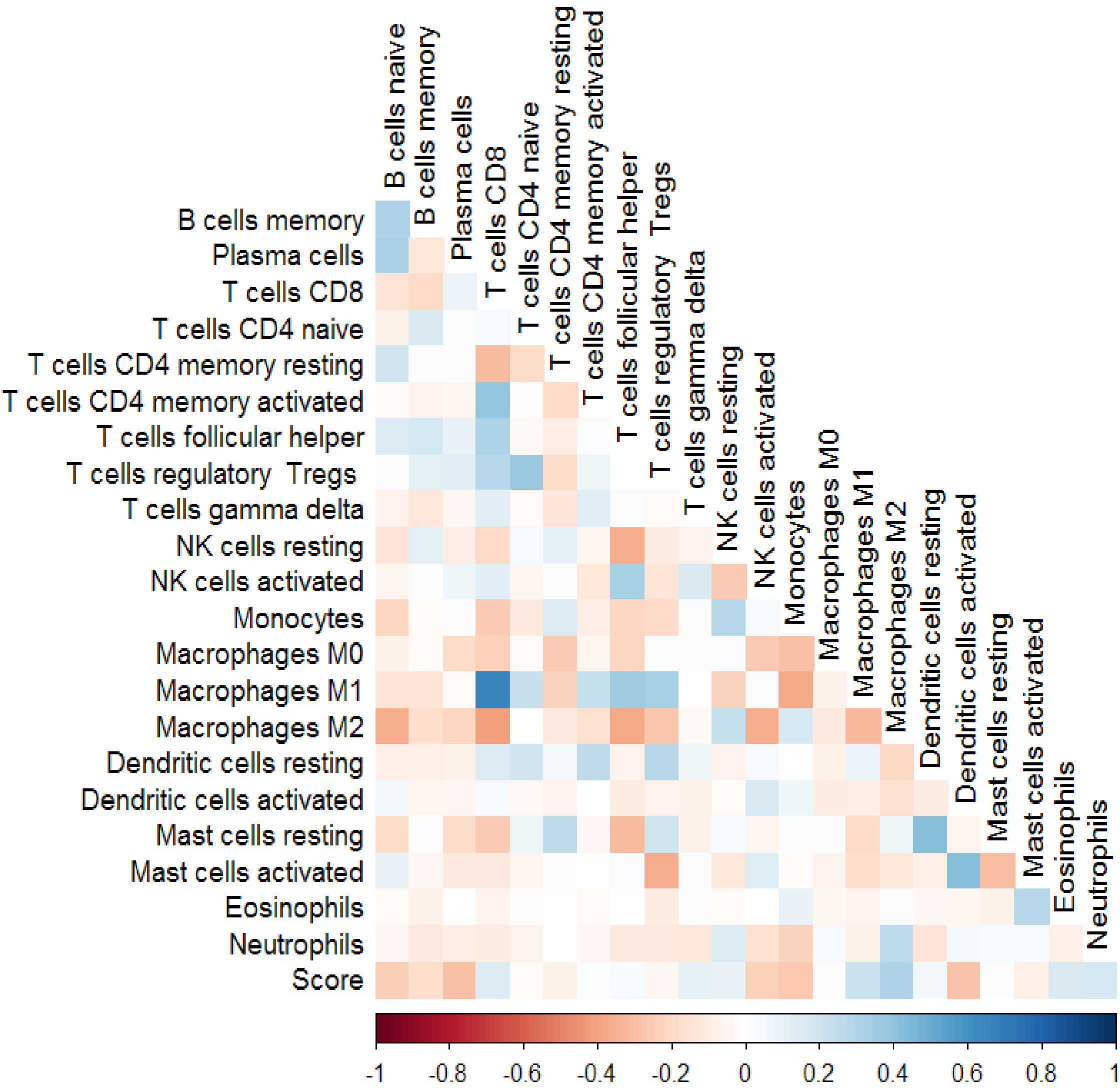
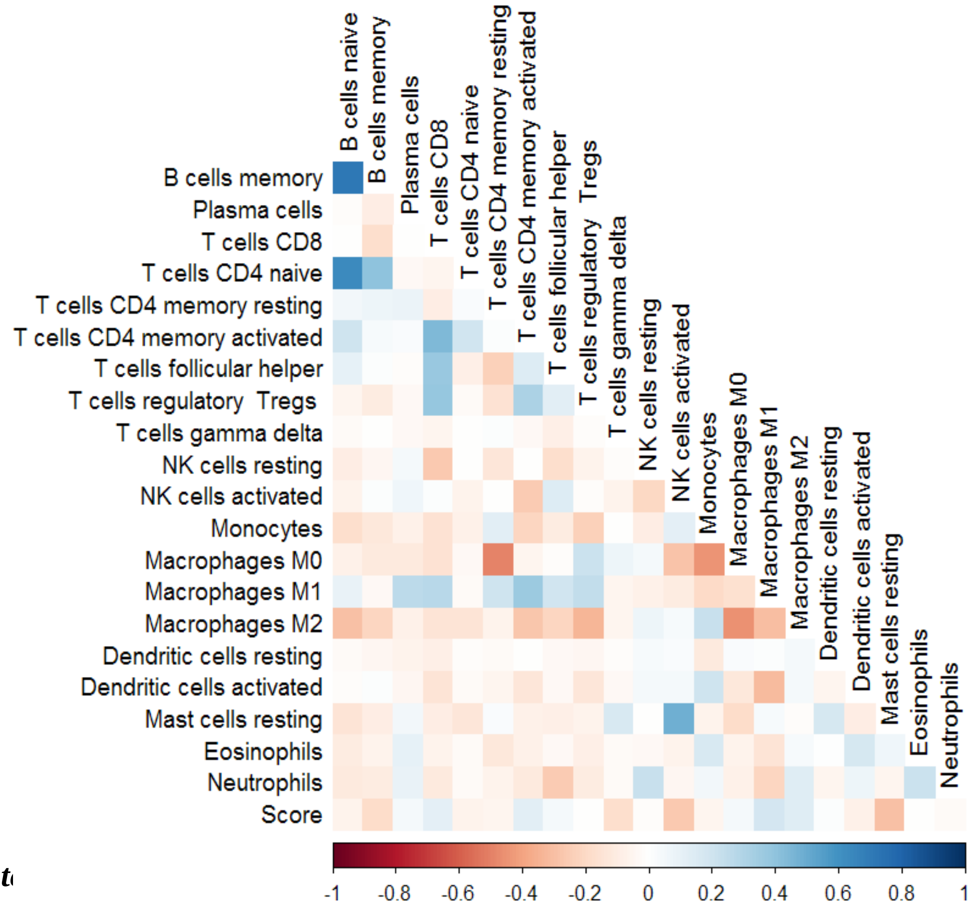


Figure 27. Correlation matrix of continuous score with immune infiltration score for specific immune system cell populations according to expression data from the Bueno cohort. The CIBERSORT algorithm was used to calculate the immune scores. The color of each square of the correlation matrix reflects the calculated correlation coefficient.



The pot

We calculated the 2-PS for each of the MPM cell lines included in the Genomics of Drug Sensitivity in Cancer (GDSC) project - GDSC1 (n=16) and GDSC2 (n=15). We subsequently tested for correlation between the 2-PS and the sensitivity to each of the drugs tested in both projects using AUC values with which we defined the response to 11 drugs from the GDSC1 dataset showing significant correlation with 2-PS of the tested mesothelioma cell lines (Fig. 28), whereas for the GDSC2 dataset the number of such significant correlations was 18 (Fig. 29). This analysis revealed a correlation of the 2-PS with the response to commonly used drugs in mesothelioma management such as cisplatin (R= - 0.51, p=0.046), gemcitabine (R=0.69, p=0.019) and vinblastine (R=0.63, p=0.037).

Figure 28. Two-dimensional correlation plots between the calculated 2-PS score and AUC for different drugs or compounds for MPM cell lines analyzed as part of the GDSC1 project. Only plots for identified significant correlations are presented.

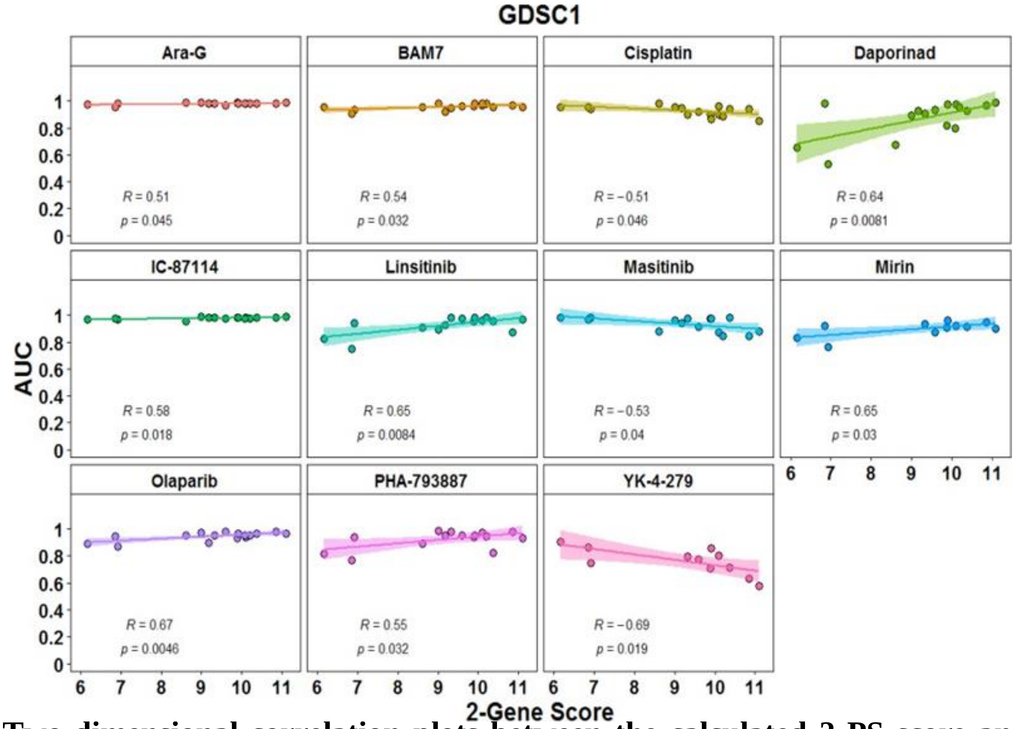
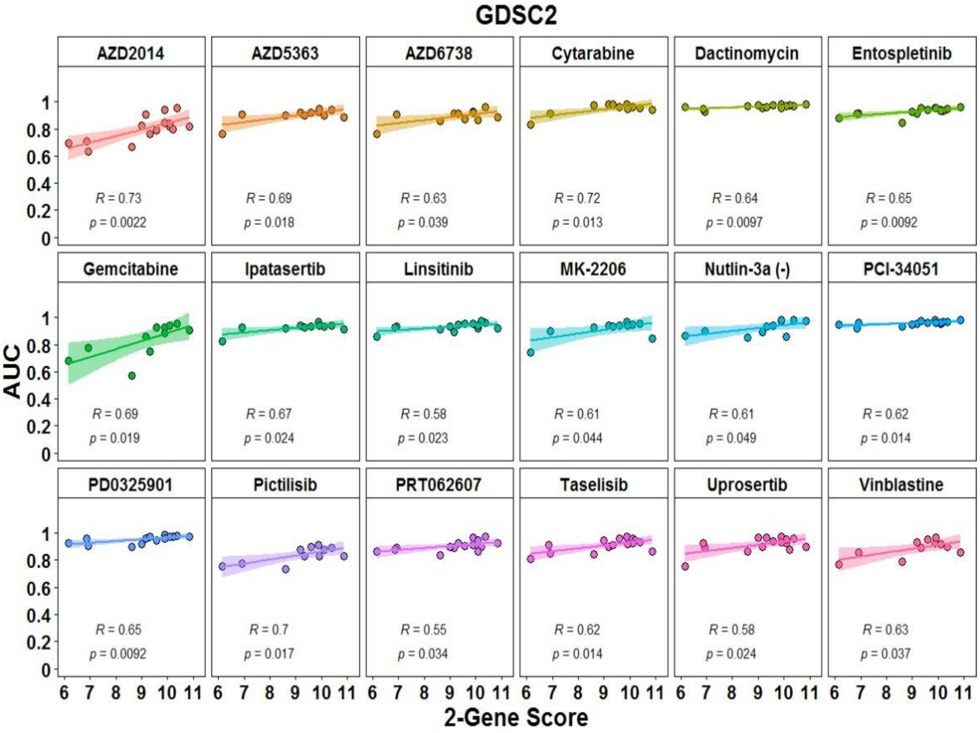


Figure 29. Two-dimensional correlation plots between the calculated 2-PS score and AUC for different drugs or compounds for MPM cell lines analyzed as part of the GDSC2 project. Only plots for identified significant correlations are presented.



V. Discussion

MPM presents a significant medical challenge due to the severe prognosis of those affected and because it‘s an example of a rare cancer caused by exposure to an obligate carcinogen in the external (mostly occupational) environment. What is special about MPM is that the cycle of extreme tumor tissue diversification is provoked by chronic tumor-promoting inflammation, which is initiated and maintained by asbestos fibers that have penetrated and are persisting in the pleural cavity. Further, in the pathogenetic mechanisms of MPM development, additional hallmarks of cancer such as genomic instability and extreme mutational variability, replicative immortality, loss of tumor suppressors, sustained proliferative signaling, evasion of cell death, neoangiogenesis, acquisition of the ability to metastasize through epithelial-mesenchymal transition are included. The ability of transformed mesothelial cells to acquire diversified essential characteristics of the cancerous growth determines the aggressive clinical course of the disease and the limited therapeutic response with conventional surgical and chemotherapeutic approaches.

In such cases, the correct prognostic stratification of patients towards the diagnosis and possible use of prognostic and predictive biomarkers is of particular importance. The standard staging of MPM according to the TNM classification doesn’t define groups with a large difference in survival since it depends on the underlying biological characteristics of the disease, the histological subtype, more precisely.

For these reasons, the discovery and the development of new prognostic and predictive models in MPM using data that more adequately reflect the underlying pathological characteristics of the individual disease are essential. The earliest omics technology applied to the prognostic evaluation of cancer patients was transcriptional analysis using planar microarrays. In the last decade, the determination of gene expression by RNA sequencing has additionally been used. After the year of 2000, these technologies were also applied to molecular profiling of MPM. A number of studies have proposed gene expression-based prognostic models in MPM. They differ significantly in the feature (gene) selection approach, training and validation datasets, the number of genes included in the final model, and the quality of different MPM cohorts, resulting in a 3-gene and a 5-gene prognostic score being derived.

In our study, we introduced an original novel approach to the selection of output markers for the model by limiting the number of genes tested for inclusion in the MPM prognostic model to only those for which MPM cell lines are known to be sensitive to their exclusion (knock-down) by siRNA or CRISPR/Cas9 editing. We then applied the *rbsurv* approach to the TCGA dataset and built a two-gene prognostic model that showed moderate predictive ability as a continuous or binary score in both univariate and multivariate models in three different MPM cohorts. This moderate predictive ability is a trade-off for the minimal number of genes included in the predictive score, thereby avoiding overfitting the model by including a larger number of genes. The limited number of genes in our 2-PS could further allow simple validation using low throughput techniques such as quantitative RT-PCR or immunohistochemistry. The two genes included in our model have not been widely studied in MPM, although such data are available at least for *MAD2L1*.

The *MAD2L1* (Mitotic Arrest Deficient 2 Like 1) gene encodes the respective protein, which is an integral part of the mitotic spindle assembly checkpoint and ensures that all chromosomes are properly aligned at the metaphase plate before the cell can proceed to anaphase. It was recently found that the *MAD2L1* gene was overexpressed in several MPM cell lines at mRNA and protein levels. This study conforms with a previous one which demonstrated higher MAD2L1 protein expression (both nuclear and cytoplasmic) in MPM cell lines as compared to normal mesothelium. The same study also showed that the higher nuclear MAD2L1 expression determined using immunohistochemistry correlated with shorter overall survival. A recent study showed that BRCA1 in mesothelioma leads to co-depletion of MAD2L1 mRNA and protein. Besides, loss of BRCA1/MAD2L1 was associated with resistance to vinorelbine ex vivo and the survival of the patients. Survival for patients lacking BRCA1/MAD2L1 expression was shorter in comparison to those with double-positive tumors. This observation can explain the fact that our 2-PS correlated with resistance to vinblastine (mitotic spindle assembly inhibitor) and olaparib (PARP inhibitor) in MPM cell lines. Besides, among the top enriched pathways in the GSEA analysis of all three cohorts were pathways directly involving mechanisms of DNA replication, such as: “Mitotic spindle”, “G2M checkpoint” and “DNA repair”.

The *GOLT1B* gene encodes the human Vesicle transport protein (Golgi Transport 1B) GOT1B protein. *GOLT1B* might be overexpressed in various tumors because of the amplification of the chromosome 12p region. Recent studies show that overexpression of *GOLT1B* in breast and colorectal cancer might be associated with poorer outcomes due to the promotion of immune evasion. Consistent with that, we found that in high 2-PS MPM patients from all three cohorts in our analysis “Epithelial mesenchymal transition”, “Apical junction” and “Protein secretion pathways” were significantly enriched in the gene expression profiles.

Our GSEA analysis shows that in all three cohorts there is a significant enrichment of overexpressed genes in the patients with a high score that are relevant to the key cancer hallmarks that are an integral part of the MPM biology. Furthermore, it is clear that due to the high reproducibility of the profiles in all three cohorts, our inferred 2-PS reflects true underlying characteristics of patients with a high score versus those with a low score. However, the DNA methylation profiles proved to be less reproducible, enabling us to extract data from only two cohorts, which were from the same DNA methylation analysis platform. However, in the course of our analysis this may be due to differences in the MPM biology included in the separate cohorts, as the proliferative index of each patient may vary which could result in different levels of DNA methylation due to the high specific cell kinetics. In addition, we detected a very small number of genes in the TCGA cohort, whose expression inversely correlated with methylation of adjacent CpG islands found in the Blum dataset.

The aforementioned recent reports regarding the role of GOLT1B in immune evasion let us investigate whether our 2-PS correlated with the estimated fractions of immune cells within the tumor tissue. We used the now standard deconvolution and deriving algorithm to obtain an immune score for infiltration of immune cells into tumor tissue to demonstrate that 2-PS correlated with the CD8+ T cells and M1/2 macrophages. The correlation of our score with macrophage infiltration is consistent with the already known pathogenetic mechanism of a sustained high level of chronic inflammation in MPM due to ineffective phagocytosis of asbestos fibers. It is known that the macrophage polarization towards an M2 phenotype may be associated with a poorer prognosis in MPM. The higher infiltration rate of CD8+ T lymphocytes may be an expression of a more powerful adaptive immune response in patients with higher 2-PS, respectively with a more aggressive and more proliferative disease. This is based on previous observations that higher MPM infiltration by CD8+ TILs is associated with a better prognosis after resection. In the setting of chronic inflammation, a large portion of CD8+ TILs acquire an expression profile of exhausted lymphocytes with characteristically increased expression of inhibitory signal receptors, such as PD-1. Using a similar approach to ours to assess MPM infiltrating immune cell fractions Blum et al. demonstrated that epithelioid-like morphology and the transcriptomic profile correlated with estimated fraction of CD8+ T cells. Nguen et al. also demonstrated that inferred infiltrating immune cells fractions can be combined with genomic parameters to develop prognostic models in MPM. Another recent study showed that the markers for higher levels of systemic inflammation correlated with shorter overall survival in MPM patients. Our observations in the context of those studies suggest that immune based markers are to be included in the prognostic schemes for MPM patients. Besides, it is rational to expect that they may have predictive power for the success of immune-checkpoint inhibitors (ICIs)-based therapy in MPMs, along with other immunogenetic markers.

In the ICIs era, the combinations with conventional chemotherapy or targeted therapy may yield additional clinical benefit in MPM. Therefore, we further evaluated our 2-PS as a possible marker to predict sensitivity of MPM cell lines to small molecule drugs. Interestingly, 2-PS inversely correlated with the AUC values for cisplatin, suggesting that it may predict higher sensitivity to it. The opposite observation was made for two other common chemotherapeutics such as gemcitabine and vinblastine suggesting that our 2-PS can predict resistance to those two.

Even though our study successfully demonstrated the power of integrating different omics data from different platforms and patient cohorts in MPM, cancer growth hallmarks, such as intratumoral heterogeneity, limit this approach. In this regard, a new multi-omics study in MPM based on scRNA-Seq and genotypic data, is indicative, which identified three types of cell phenotypes in MPM, namely those with a pronounced ability for tumor proliferation, cells able to avoid immune response, and those with acinar morphology and loss of the *BAP1* gene expression as the clinical course in different patients depends on the predominant subtype of malignant mesothelial cells.

All of this indicates that future directions in the prognostic modeling in MPM will focus on the integration of new biomarkers, imaging methods, and molecular profiling technologies to improve prognostic accuracy and refine patient stratification. The multi-omics approaches combining genomic, transcriptomic and proteomic data are likely to identify with an increasing success rate prognostic profiles and therapeutic targets in MPM. Additionally, the advances in artificial intelligence and data analytics will enable the development of sophisticated predictive models capable of predicting individual risk and optimizing treatment. Continued efforts in the development, validation, and clinical translation of such models will ultimately be essential to improve the clinical outcomes in patients with MPM as well as other rare cancers.

.

VI. Conclusion and inferences

Having achieved the set research aim through the accurate fulfillment of the set research tasks, we can conclude the following:

1. An oligogenic prognostic score in MPM was derived and validated through primary gene selection based on a dependency screening;
2. The derived score defines subgroups of MPM patients that have a very similar expression profile in all studied cohorts – training and validation;
3. The DNA methylation profile is not associated with reproducible differential DNA methylation profiles;
4. The prognostic score is associated with specific reproducible profiles of immune cell infiltration that can be explained by disease pathogenesis;
5. The derived prognostic score probably also has predictive value regarding sensitivity or resistance to commonly used chemotherapeutic agents for MPM treatment.

Based on these findings, exemplary rational directions for future research can easily be defined:

1. Conversion of the derived score into a clinically applicable score using immunohistochemistry or RT-PCR on archival histological materials from MPM patients;
2. Prospective validation of the prognostic value of 2-PS in clinical settings using a simplified immunohistochemical or RT-PCR-based expression score;
3. Prospective evaluation of the predictive value of 2-PS in clinical settings for response to conventional chemotherapy and/or immunotherapy with ICIs using a simplified immunohistochemical or RT-PCR-based expression score;
4. Application of the approach described by us to develop prognostic and predictive expression scores in other types of neoplastic diseases, incl. rare cancers.

VII. Contributions

Original contributions

1. The possibility of developing a prognostic score based on gene expression in MPM by initial selection of genes whose expression is likely to depend on cell population survival in MPM is demonstrated for the first time.
2. The prognostic value of the *GOLT1B* gene expression has been demonstrated and discussed for the first time.
3. For the first time, an oligogenic prognostic model in MPM involving only 2 genes has been demonstrated.
4. It's shown for the first time that the developed prognostic model has a probable predictive value for response to treatment with various conventional chemotherapeutic agents.

Confirmed contributions

1. The prognostic value of the *MAD2L1* expression was confirmed.
2. The prognostic scores in MPM have been demonstrated to correlate inexorably with certain core cancer characteristics, particularly with those related to DNA repair and mitosis, which is expected given the proliferative and aggressive nature of this disease.
3. It has been demonstrated that it's possible to integrate different omics platforms in MPM, and that transcriptomic analysis platforms have good reproducibility regardless of the method used and the location of its use.
4. The poor reproducibility of DNA methylation profiles in MPM and their weak correlation with prognostically relevant disease subgroups has been confirmed.

VIII. Publications

Journal articles

SHIVAROV, Velizar; BLAZHEV, Georgi; YORDANOV, Angel. A Novel Two-Gene Expression-Based Prognostic Score in Malignant Pleural Mesothelioma. Diagnostics, 2023, 13.9: 1556.

SHIVAROV, Velizar; BLAZHEV, Georgi. Bringing Together the Power of T Cell Receptor Mimic and Bispecific Antibodies for Cancer Immunotherapy: Still a Long Way to Go. Monoclonal Antibodies in Immunodiagnosis and Immunotherapy, 2021, 40.2: 81-85.

Participation in scientific meetings

BLAZHEV, Georgi; SHIVAROV, Velizar. Development of a Novel Gene Expression-Based Prognostic Score in Malignant Pleural Mesothelioma. Kliments days 2020, Sofia, Bulgaria.