RESEARCH ARTICLE

Propensity score matching as an effective strategy for biomarker cohort design and omics data analysis

Masaki Maekawa[1,2,3], **Atsushi Tanaka**[1,2,3], **Makiko Ogawa**[1,2,3], **Michael H. Roehrl**[1,2,3]*

1 Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA, United States of America, **2** Harvard Medical School, Boston, MA, United States of America, **3** Memorial Sloan Kettering Cancer Center, New York, NY, United States of America

* michael_roehrl@bidmc.harvard.edu

Abstract

Background

Analysis of omics data that contain multidimensional biological and clinical information can be complex and make it difficult to deduce significance of specific biomarker factors.

Methods

We explored the utility of propensity score matching (PSM), a statistical technique for minimizing confounding factors and simplifying the examination of specific factors. We tested two datasets generated from cohorts of colorectal cancer (CRC) patients, one comprised of immunohistochemical analysis of 12 protein markers in 544 CRC tissues and another consisting of RNA-seq profiles of 163 CRC cases. We examined the efficiency of PSM by comparing pre- and post-PSM analytical results.

Results

Unlike conventional analysis which typically compares randomized cohorts of cancer and normal tissues, PSM enabled direct comparison between patient characteristics uncovering new prognostic biomarkers. By creating optimally matched groups to minimize confounding effects, our study demonstrates that PSM enables robust extraction of significant biomarkers while requiring fewer cancer cases and smaller overall patient cohorts.

Conclusion

PSM may emerge as an efficient and cost-effective strategy for multiomic data analysis and clinical trial design for biomarker discovery.

Citation: Maekawa M, Tanaka A, Ogawa M, Roehrl MH (2024) Propensity score matching as an effective strategy for biomarker cohort design and omics data analysis. PLoS ONE 19(5): e0302109. https://doi.org/10.1371/journal.pone.0302109

Editor: Raffaele Serra, University Magna Graecia of Catanzaro, ITALY

Received: November 9, 2023

Accepted: March 27, 2024

Published: May 2, 2024

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: Michael H. Roehrl acknowledges grants from the National Institute of Health/National Cancer Institute (R21 CA251992, R21 CA263262, and U01 CA263986) and an Investigator Grant from the Neuroendocrine Tumor Research Foundation (NETRF). The funding bodies were not involved in the design of the study and the collection, analysis, and interpretation of data.

Competing interests: Masaki Maekawa, Atsushi Tanaka, and Makiko Ogawa declare no conflicts of interest related to this study. Michael H. Roehrl is member of the Scientific Advisory Boards of Azenta Life Sciences and Universal DX. None of these companies had any role in design, execution, data analysis, or any other aspect of this study. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

Abbreviations: PSM, Propensity score matching; CRC, Colorectal cancer; MSKCC, Memorial Sloan Kettering Cancer Center; TCGA, The cancer genome atlas; MMR, Mismatch repair; IHC, Immunohistochemical; CPM, Counts per million; FDR, False discovery rate; NNMT, Nicotinamide N-methyltransferase; GALNT6, Polypeptide N-acetylgalactosaminyltranferase 6; SLC7A5, SLC3A2, Solute carrier Family 7 member 5; SLC3A2, Solute carrier family 3 member; IGF2BP3, Insulin like growth factor 2 mRNA binding protein; MCM6, Minichromosome maintenance complex component 6; SERPIN B5, Serpin family B member 5; STAT1, Signal transducer and activator of transcription 1; NAMPT, Nicotinamide phosphoribosyltransferase; P4HA1, Prolyl 4-hydroxylase subunit alpha 1; LTBP2, Latent transforming growth factor beta binding protein 2; DDX21, DEAD-box RNA helicase 21.

Background

Omics data analysis has become increasingly popular in biomedical research and precision medicine. Due to advances in sequencing and other deep molecular profiling methods and infrastructure, a wealth of multiomic data, spanning from the genome, transcriptome, proteome to the metabolome, have been generated. Omics datasets are complex and associated with multiple biological and clinical parameters. There is currently no standard method for parsing through omics data to extract significant specific biomarkers and separating them from a multitude of possibly confounding variables. Typically, very large randomized patient cohorts are considered to be needed in disease biomarker discovery for addressing the confounding variable problem.

Propensity score matching (PSM) is a statistical matching technique that attempts to deduce the effect of a single specific factor by reducing bias due to confounding variables [1]. In past studies, the assignment of a single specific factor to a subject was not random, and there were many confounding variables between the group with the specific factor and the control group without the factor. The propensity score expresses quantitatively "how likely each case is to have the specific factor." The estimated propensity score $e(x_i)$ of each case can be calculated based on its background, i ($i = 1,2,\ldots,N$), which could be a confounding variable. Paired cases with most similar propensity scores can be extracted from each group with or without the single specific factor [2–5]. This process of homogenizing confounding factors is called PSM, and it enables to obtain extracted groups and control groups which have a difference only in the single specific biomarker to be examined.

Although PSM is a powerful statistical method, it has not been implemented into routine omics data analysis for biomarker discovery. In this study, we tested PSM in two colorectal cancer datasets to investigate its efficiency in omics analysis. One dataset was derived from immunohistochemical (IHC) protein expression of 12 protein markers in a cohort of 544 colorectal cancer (CRC) patients [6, 7]. The second dataset was obtained from The Cancer Genome Atlas database (TCGA) and contained total RNA-sequencing profiling of 163 CRC patients [8]. We clustered the cancer cases into good or poor prognosis groups based on patient survival information, and the groups were compared to uncover prognostic protein or transcript biomarkers. We then compared the results before and after PSM implementation to evaluate the efficiency of PSM for omics data analyses and confounding variable-avoiding biomarker discovery.

Methods

The study used retrospective archives samples and was classified by the institution as minimal risk and with waiver of consent (BIDMC IRB 2023P000933).

Dataset of proteomic markers of CRC

This dataset was derived from proteomic expression profiles of 544 surgically resected CRC tissue obtained from the pathology archives of MSKCC. Clinicopathological parameters, including patient age, gender, tumor location, TNM stage, mismatch repair (MMR) status, histology, tumor differentiation, lymphovascular invasion, and perineural invasion were retrieved from medical records. This cohort comprised 367 patients who survived more than three years without recurrence (good outcome group), and 60 patients who had recurrence within three years (poor outcome group). The others did not have sufficient survival information due to short follow-up periods.

Twelve potential prognostic biomarkers (NNMT, GALNT6, SLC3A2, SLC7A5, IGF2BP3, MCM6, SERPIN B5, STAT1, NAMPT, P4HA1, DDX21, and LTBP2) were chosen based on preliminary proteomic findings from our laboratory [6, 7]. Expression levels of these proteins were determined by immunohistochemical (IHC) staining of tissue microarrays. The tissue microarrays were constructed from 544 formalin fixed and paraffin embedded CRC tissue specimens. Tissue sections were incubated with polyclonal antibodies specific to each protein marker (see S1 Table for details on antibodies used) and visualized with Bond Polymer Refine Detection (Leica). IHC staining scores, which correspond to protein expression levels in cancer tissue, were determined independently by two pathologists without knowledge of the patients' clinical information.

For seven markers, IHC staining intensity of individual tumor cells was determined and assigned intensities of 0, 1+, 2+, or 3+, and the total weighted IHC score of a sample slide was calculated by multiplying the expression intensity of individual tumor areas (score 0–3) by their relative contribution (0–100%) to total tumor area. The total weighted IHC scores thereby had a range of 0–300 for these protein biomarkers. For NNMT, GALNT6, or MCM6, each tissue section was scored by counting the number of cancer cells staining positively for the protein biomarker (staining intensity $\geq$1+) relative to the total number of evaluated cancer cells. A minimum of 500 cancer cells were evaluated per tissue sample, and the IHC scores had a range of 0–100. For DDX21 or LTBP2, IHC stains were scored binomially as negative or positive (score 0 or 1+).

Random forest analysis

IHC scorings were compared between the good and poor prognosis groups. A random forest analysis was performed to detect prognostic IHC markers using the R (version 4.2) packages "randomForestSRC" (version 3.3.2) and "ggRandomForests" (version 2.2.1) downloaded from *cran.rstudio.com*. The analysis included recurrence-free survival data, clinicopathological features, and IHC scorings. The number of trees used was set to 100,000. Minimal depth value and variable importance were utilized to detect significant prognostic factors. Additionally, the significant prognostic IHCs which were selected by random forest analysis were compared to the results of the original comparisons.

Statistical comparison

Clinicopathological values were compared using a Student t-test or a chi-squared test if the values were binary. The comparisons of IHC scorings were also conducted using a Student t-test or a chi-squared test.

Dataset of CRC transcriptome

The original RNA-seq dataset used in this study is available using the "GDCquery" function in the package "TCGAbiolinks" (version 2.18.0) downloaded from *bioconductor.org* (R version 4.2). The "project" names used are "TCGA-COAD" and "TCGA-READ". The "data.category" name used is "Transcriptome Profiling" and the "data.type" used is "Gene Expression Quantification" for RNA-seq data. The "data.category" name used is "Clinical" for clinical information.

This dataset contains mRNA sequencing counts for 60,660 genes from 163 CRC patients [8]. The data were obtained from TCGA-COAD and TCGA-READ using the Bioconductor package in R [9, 10]. In this CRC cohort, 130 patients survived more than three years after sample collection and were considered in the good prognosis group, and 33 patients died within three years and were considered in the poor prognosis group. The dataset included clinicopathological information such as age, gender, TNM stage, MMR status, histology, residual tumor status, venous invasion, lymphovascular invasion, and perineural invasion.

Differential gene expression analysis

Differential expression analysis was performed to identify genes that were significantly associated with outcome separating good and poor prognosis groups. Differential transcriptome expression analysis of was conducted using the R (version 4.2) packages "limma" (version 3.56.2) and "edgeR" (version 3.42.4) downloaded from *bioconductor.org* [11–13]. Transcripts with counts per million (CPM) <15 were filtered out in order to exclude low-expression genes, and the dataset was normalized. The dispersion of gene expression values was estimated [14], and statistical tests of fold change were computed to compare the poor prognosis group to the good prognosis group. Gene transcript changes with a false discovery rate (FDR) less than 0.05 were considered significant.

Propensity score matching (PSM)

A flow chart of how PSM was performed in this study is shown in Fig 1. Estimated propensity scores of each case $e(X_i)$ were calculated using linear logistic regression:

$$\ln\frac{e(x_i)}{1-e(x_i)} = \ln\frac{P_r(Z_i=1|x_i)}{1-P_r(Z_i=1|x_i)} = \alpha + \beta^T x_i$$

$$e(x_i) = \Pr(Z_i = 1|x_i)$$

$$e(X_i) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_iX_i$$

b_0: the intercept

b_i: the regression coefficient

x_i: the variables of each case

X_i: the specific variable to investigate and covariates

$Z_i = 1$: case in a poor prognosis group

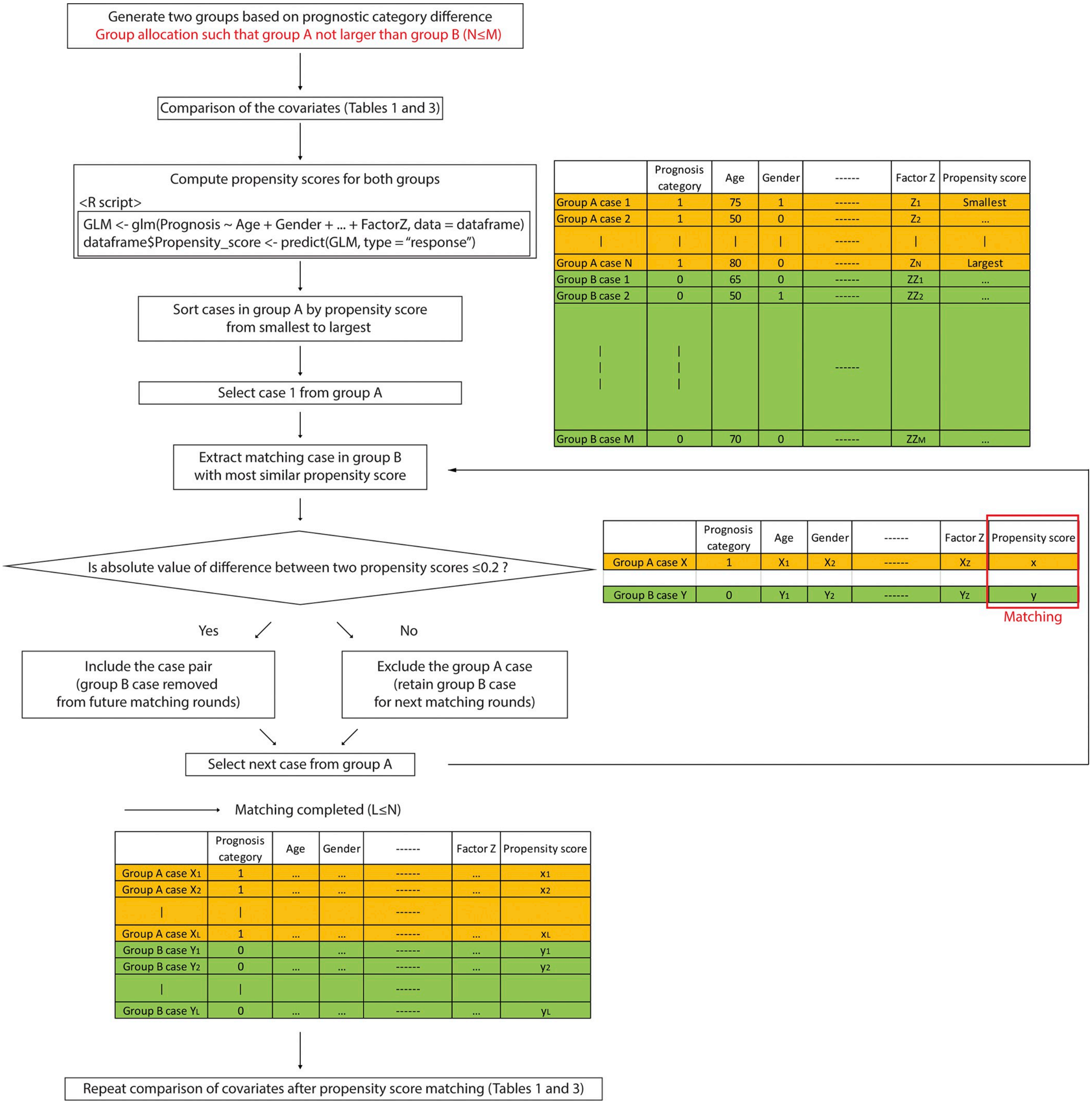
$Z_i = 0$: case in a good prognosis group

The propensity score of each case was computed from the collected clinicopathological features using the "glm" function (generalized linear model) and the "predict" function, both of which are part of the basic R package "stats" (version 4.4.0).

$$GLM < -glm(Prognosis \sim Age + Gender + \cdots, data = dataframe)$$

$$dataframe\$PropensityScore < -predict(GLM, type = "response")$$

Cases in the group with the smaller number of patients (in this study, the poor prognosis group) were sorted by their propensity scores. The sorting was done from smallest to largest propensity score, and, for each case, a paired case with most similar propensity score was selected from the good prognosis group. If the absolute value of difference between two propensity scores was less than 0.2, the case pair was included and the matched case removed from future matching rounds. If the absolute value of difference was greater than 0.2, the original case was excluded and the matched case retained for next matching rounds. This process resulted in the creation of two extracted groups that had the same number of cases without duplicates, and the number of cases was smaller than the original dataset.


Generate two groups based on prognostic category difference
Group allocation such that group A not larger than group B (N≤M)
Comparison of the covariates (Tables 1 and 3)
Compute propensity scores for both groups
<R script>
GLM <- glm(Prognosis ~ Age + Gender + ... + FactorZ, data = dataframe)
dataframe$Propensity_score <- predict(GLM, type = "response")
Sort cases in group A by propensity score from smallest to largest
Select case 1 from group A
Extract matching case in group B with most similar propensity score
Is absolute value of difference between two propensity scores ≤0.2 ?
Yes
No
Include the case pair (group B case removed from future matching rounds)
Exclude the group A case (retain group B case for next matching rounds)
Select next case from group A
Matching completed (L≤N)
Matching
Repeat comparison of covariates after propensity score matching (Tables 1 and 3)

	Prognosis category	Age	Gender	------	Factor Z	Propensity score
Group A case 1	1	75	1	------	Z_1	Smallest
Group A case 2	1	50	0	------	Z_2	...
|	|	|	|	------	|	|
Group A case N	1	80	0	------	Z_N	Largest
Group B case 1	0	65	0	------	ZZ_1	...
Group B case 2	0	50	1	------	ZZ_2	...
| | |	| | |			------		
Group B case M	0	70	0	------	ZZ_M	...

	Prognosis category	Age	Gender	------	Factor Z	Propensity score
Group A case X	1	X_1	X_2	------	X_Z	x
Group B case Y	0	Y_1	Y_2	------	Y_Z	y

	Prognosis category	Age	Gender	------	Factor Z	Propensity score
Group A case X_1	1	...	...	------	...	x_1
Group A case X_2	1	...	...	------	...	x_2
|	|			------		
Group A case X_L	1	...	...	------	...	x_L
Group B case Y_1	0		...	------	...	y_1
Group B case Y_2	0	...	...	------	...	y_2
|	|			------		
Group B case Y_L	0	...	...	------	...	y_L

Fig 1. Flow chart of propensity score matching in this study. It is crucial that the number of cases in group A is not larger than in group B. In this study, group A meant the good prognosis group, and group B meant the poor prognosis group in both datasets.

https://doi.org/10.1371/journal.pone.0302109.g001

Results

Prognostic proteomic biomarkers of CRC

The proteomic marker dataset of CRC cohort consisted of 367 cases in the good prognosis group and 60 cases in the poor prognosis group (Table 1). Propensity scores were computed for each case, and the resulting distributions of the propensity scores are shown in histograms

Table 1. Clinicopathological features of the proteomic CRC cohort before and after PSM.

	Before PSM			After PSM		
	Good prognosis (N = 367)	**Poor prognosis (N = 60)**	**p-value**	**Good prognosis (N = 58)**	**Poor prognosis (N = 58)**	**p-value**
Age (mean)	64.0	56.4	< **0.001**	55.3	56.8	0.604
Gender			**0.01889**			0.3426
Male	175	39		32	38	
Female	192	21		26	20	
Location			0.1291			0.3426
Right	183	23		20	23	
Left	184	37		38	35	
TNM staging (mean)	1.72	2.35	< **0.001**	2.31	2.29	0.9087
Mismatch repair status			0.2982			0.5995
Present	87	10		7	10	
Absent	280	50		51	48	
Histology			0.8768			1
Mucinous	25	5		5	5	
Not mucinous	342	55		53	53	
Tumor differentiation (mean)	2.02	2.08	0.2159	2.07	2.09	0.8171
Lymphatic invasion			< **0.001**			1
Positive	64	27		26	25	
Negative	303	33		32	33	
Perineural invasion			< **0.001**			1
Positive	18	15		13	13	
Negative	349	45		45	45	

Bold: p <0.05, statistically significant

https://doi.org/10.1371/journal.pone.0302109.t001

and density plots (Fig 2). After PSM, group differences of confounding factors, such as age, gender, TNM stage, lymphatic invasion positivity, and perineural invasion positivity, were eliminated. Two paired groups of 58 cases each were created after PSM, a significant reduction in case numbers in both the good and poor prognosis groups when compared to the original data before PSM (Table 1).

Comparisons of protein marker expression as measured by IHC scores between the good and poor prognosis groups before and after PSM are shown in Table 2. In the analysis before PSM, three protein biomarkers (SLC7A5, SLC3A2, and STAT1) showed statistically significant differences between the good and poor prognosis groups. However, only STAT1 showed significant difference between the two groups after cohorts were minimized for confounding variables by PSM.

To validate the above findings, we also conducted random forest analyses on the dataset to identify significant prognostic factors (Fig 3). In addition to clinicopathological features, several markers such as STAT1, NNMT, SLC3A2, and MCM6 were found to be significant prognostic factors based on both minimal depth and variable importance rankings. Except for clinicopathological features such as lymphatic invasion and TNM stage, STAT1 was the most significant prognostic factor based on both types of rankings from random forest analysis.

Prognostic transcriptomic biomarkers of CRC

The TCGA's CRC RNA-seq dataset consisted of 130 cases in the good prognosis group and 33 cases in the poor prognosis group (Table 3). The propensity score of each case was calculated,

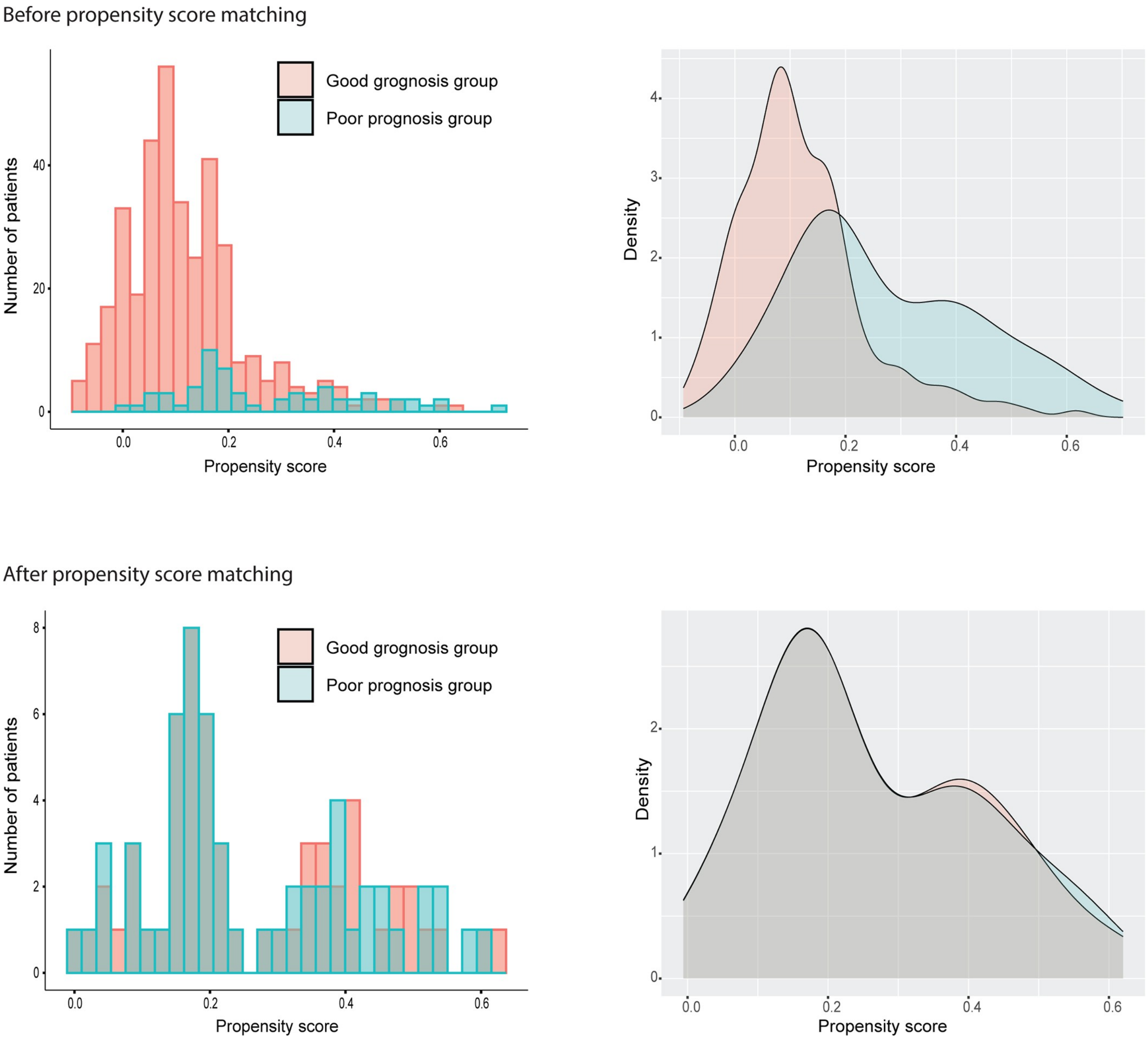


Fig 2. Distribution of propensity scores of the IHC CRC dataset.

https://doi.org/10.1371/journal.pone.0302109.g002

and the distributions of these scores were shown in histograms and density plots (Fig 4). Before PSM, the good and poor prognosis groups were significantly impacted by six confounding clinicopathological features, including TNM staging, mucinous histology, residual tumor status, venous invasion, lymphatic invasion, and perineural invasion. These confounding features were successfully eliminated by PSM, resulting in two paired groups of 28 cases each with no significant differentiating clinical factors (Table 3).

Without PSM implementation, total mRNA sequencing of the original 130 good prognosis cases and 33 poor prognosis cases yielded 13,294 genes (21.9% of all) with counts per million (CPM) greater than 15. Differential gene expression analysis identified 402 genes with false discovery rate less than 0.05, i.e., significantly differentially expressed between the good and

Table 2. Comparison of protein marker expression between groups before and after PSM.

	Before PSM			After PSM		
	Good prognosis (N = 367)	Poor prognosis (N = 60)	p-value	Good prognosis (N = 58)	Poor prognosis (N = 58)	p-value
NNMT	35.4	37.1	0.5295	38.1	37.3	0.8212
GALNT6	54.6	50.5	0.349	47.8	51	0.5961
SLC7A5	84.1	60.7	**0.005109**	72	62.7	0.3822
SLC3A2	88.1	60	**0.005121**	60.5	62.1	0.9039
IGF2BP3	97.3	96.6	0.9275	106	96.9	0.4854
MCM6	13.8	19.3	0.1322	19.2	18.9	0.9473
SERPIN B5	86.7	70.2	0.06452	77.6	70.9	0.555
STAT1	57.1	25.3	< **0.001**	59.1	26.2	**0.002791**
NAMPT	91.5	89.8	0.8424	100.2	89.1	0.356
P4HA1	95.8	108.2	0.1576	115.5	108.1	0.5195
LTBP2 (cytoplasm) positive	136	26	0.4323	18	26	0.1804
LTBP2 (stroma) positive	79	12	0.9223	14	11	0.6515
DDX21	81	9	0.2827	10	8	0.7976

Bold: p <0.05, statistically significant

https://doi.org/10.1371/journal.pone.0302109.t002

poor prognosis cases. They included 373 upregulated and 29 downregulated genes associated with poor prognosis.

After PSM and reducing the cases to only 28 in each group, differential expression analysis identified 12,460 genes (20.5% of all) with CPM greater than 15. 122 differential genes were observed with false discovery rate less than 0.05. The 122 significant genes included 109 upregulated and 13 downregulated genes associated with poor prognosis.

Comparing the results before and after PSM, 29 significantly differentially expressed genes were only identified after PSM, while 93 significant genes were identified both without and with PSM analysis (Fig 5). The 29 newly uncovered CRC outcome-differential genes are listed in Table 4. Additionally, the 93 genes selected in common across both comparisons are provided as S2 Table.

Among these 29 genes identified only after PSM are numerous well-known cancer-related markers, e.g., ERBB2, MALAT1, or MUC5AC. A literature search revealed that at least 15 of these markers have been suggested as potential diagnostic or predictive markers for CRC (see refences listed for each potential marker in Table 4). More extensive investigation of these markers and their roles in CRC will be warranted in future studies.

Discussion

Propensity score matching (PSM) is a powerful statistical tool for controlling confounding factors in analyses of complex omics biomarker cohorts. In this study, we demonstrated a PSM strategy for creating two well-balanced groups that mimic the original study cohort but with significantly fewer cases or patients required while eliminating outcome-confounding clinical differences between the original cohorts. We created a stepwise matching protocol in which the cases were matched from the one with the lowest to the highest propensity score (Fig 1). This ascending sort should be better than a descending sort. As seen in our two datasets, cases with lower propensity scores are easier to match (Figs 2 and 4). If a descending sort were used, the paired case would consistently have a lower propensity score, and the covariates would be more likely to remain after matching. Therefore, we predicted that an ascending sort would result in a better matching outcome.

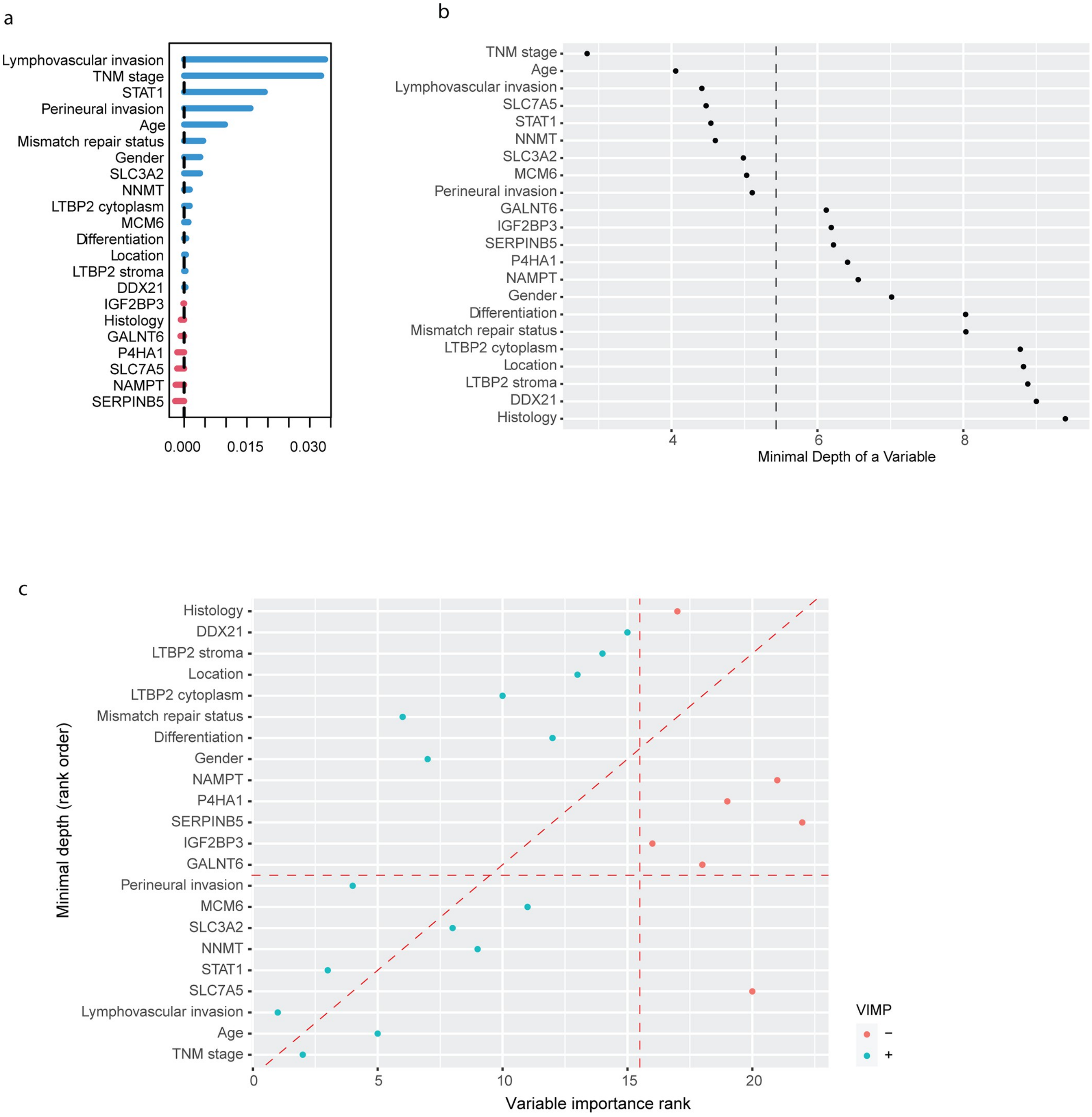


Fig 3. Random forest rankings of prognostic factors in the CRC proteomic marker dataset. (A) Ranking of variable importance (VIMP). The blue bars represent positive values of VIMP, indicating that the corresponding factor is positively associated with prognostic prediction. While the red bars represent negative values of VIMP, indicating that the factor is negatively associated with prognostic prediction. **(B) Ranking of minimal depth.** The small minimal depth indicates that the factor plays an important role in prognostic prediction. The vertical dashed line indicates the minimal depth threshold where smaller minimal depth values indicate higher importance and larger indicate lower importance as calculated by the "gg_minimal_depth" function of the "ggRandomForests" R package (version 4.7–1.1). **(C) The combination of variable importance (VIMP) and minimal depth.** The blue dots represent positive values of VIMP, while red dots represent negative values of VIMP. The threshold represented by the vertical red dashed line indicates VIMP = 0. The threshold represented by the horizontal red dashed line is equal to (B).

https://doi.org/10.1371/journal.pone.0302109.g003

Table 3. Clinicopathological features of the CRC RNA–seq dataset before and after PSM.

	Before PSM			After PSM		
	Good prognosis (N = 130)	**Poor prognosis (N = 33)**	**p-value**	**Good prognosis (N = 28)**	**Poor prognosis (N = 28)**	**p-value**
Age (mean)	65.3	68.3	0.2154	69.3	67.4	0.5481
Gender			0.976			1
Male	72	19		17	16	
Female	58	14		11	12	
TNM staging (mean)	2.23	3.15	< **0.001**	3.0	3.0	1
Mismatch repair loss			0.2156			-
Present	10	0		0	0	
Absent/unknown	120	33		28	28	
Histology			**0.0096**			1
Mucinous	9	8		6	7	
Not mucinous	121	25		22	21	
Residual tumor status			< **0.001**			0.7748
No	88	10		8	10	
Other/unknown	42	23		20	18	
Venous invasion			**0.010**			1
Positive	22	13		10	9	
Negative/unknown	108	20		18	19	
Lymphatic invasion			**0.013**			0.7844
Positive	32	16		10	12	
Negative/unknown	98	17		18	16	
Perineural invasion			**0.033**			1
Positive	9	7		4	3	
Negative/unknown	121	26		24	25	

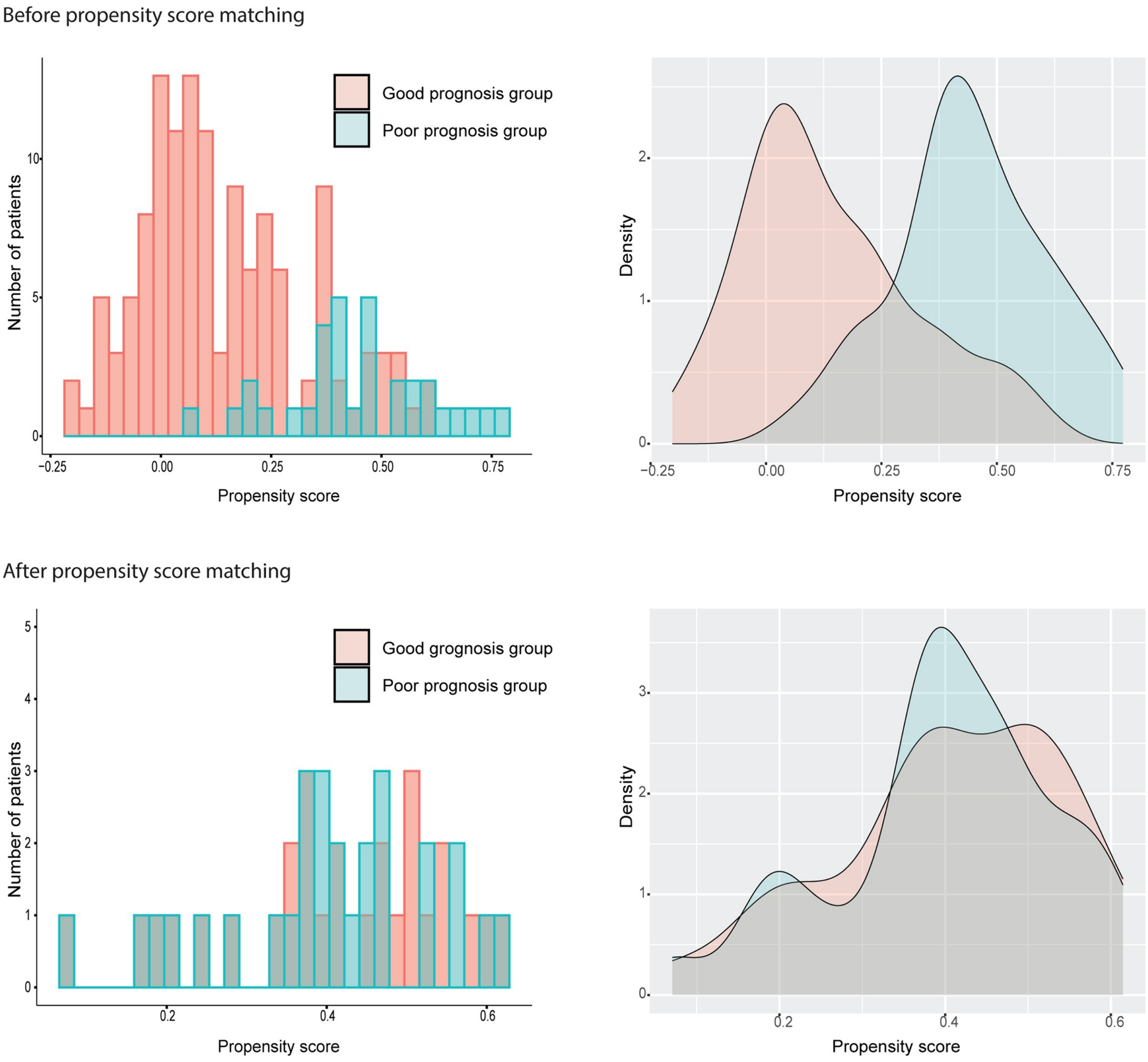
Bold: p <0.05, statistically significant

https://doi.org/10.1371/journal.pone.0302109.t003

PSM is a cost-effective analytical method for omics data analyses and biomarker discovery. Obtaining large omics expression dataset is typically expensive, so cost is often a significant concern for research. By implementing PSM before the experimental process, high-quality case series can be created from a much smaller number of cases. This approach not only provides higher-quality analytical results but also lowers research costs. Additionally, PSM enables direct comparisons between patients. In past omics data analyses, the common method of comparison was between cancer and normal tissue to uncover cancer biomarkers. However, this method might not be appropriate for detecting certain cancer mechanisms, such as treatment response or drug resistance. PSM enables comparisons between patients, which can help identify these mechanisms directly.

In addition to the benefit of minimizing cohort size, PSM enabled us to discover several, potentially important markers for colorectal cancer. From the proteomic marker cohort, PSM affirmed the prognostic value of STAT1 in CRC that we have reported previously [30], and the conclusion was further independently solidified by random forest analysis in this study. From the CRC transcriptome cohort, PSM allowed us to identify several well-known prognostic factors that might have been missed in traditional differential expression analysis due to confounding factors, e.g., ERBB2 (HER2), MALAT1, and MUC5AC.

ERBB2 (also known as HER2, neu, and NGL) encodes a transmembrane glycoprotein belonging to the epidermal growth factor receptor family [31–33]. ERBB2 is one of the most extensively studied molecules in cancer research [34], particularly in breast cancer [35, 36].



Before propensity score matching

After propensity score matching

Fig 4. Distribution of propensity scores of the RNA–seq CRC dataset.

https://doi.org/10.1371/journal.pone.0302109.g004

Overexpression of ERBB2 in cancer cells has been linked to metastasis, poor response to cancer treatment, and poor prognosis. In our study, the mRNA abundance of ERBB2 was found to be significantly higher in the poor-prognosis group by post-PSM analysis. However, the significance of this difference was not identifiable without PSM.

MALAT1 (metastasis-associated lung adenocarcinoma transcript 1) is one of the first identified cancer-associated long noncoding RNAs. MALAT1 was originally described as a prognostic marker of lung cancer metastasis. Upregulated MALAT1 has been observed in various cancers, including CRC, and is associated with cancer metastasis, invasion, and poor prognosis [37–39]. In this study, despite previous reports linking upregulated MALAT1 to poor prognosis, we observed a significant downregulation of MALAT1 mRNA abundance in the poor

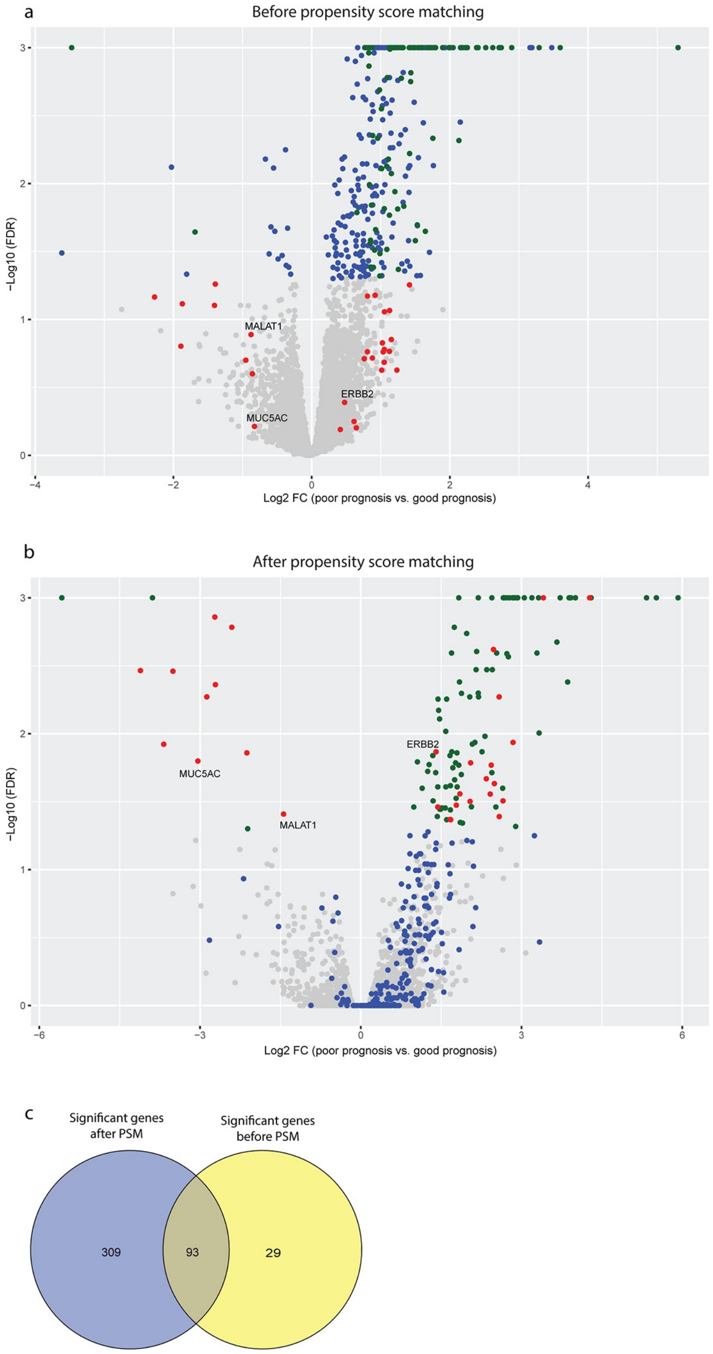


Fig 5. (A–B) RNA–seq volcano plot comparing good prognosis group vs. poor prognosis group. Green dots (N = 93) represent genes that are significant in both pre–and post–PSM comparison between the good–and poor–prognosis groups. Blue dots (N = 217) represent genes that are significant only in the pre–PSM comparison between the good–and poor–prognosis groups. Red dots (N = 29) represent genes that are significant only in the post–PSM comparison between the good–and poor–prognosis groups. Grey dots (N = 12,121) represent genes that did not show significant differences. **(C) The Venn diagram of significant genes before and after PSM.** The blue circle represents before PSM, and the yellow represents after PSM.

https://doi.org/10.1371/journal.pone.0302109.g005

Table 4. Genes significantly associated with CRC prognosis uncovered only by PSM.

Gene	Description	Evidence as CRC biomarker?
MTND1P23	MT-ND1 pseudogene 23	[15]
AFAP1-AS1	AFAP1 antisense RNA 1	[16]
LEFTY2	left-right determination factor 2	
SCARA5	scavenger receptor class A member 5	[17]
CADM3	cell adhesion molecule 3	[18]
MALRD1	MAM and LDL receptor class A domain containing 1	
ABCA8	ATP binding cassette subfamily A member 8	
ANPEP	alanyl aminopeptidase, membrane	[19]
GREM2	gremlin 2, DAN family BMP antagonist	
CHRDL1	chordin like 1	
CILP	cartilage intermediate layer protein	
STMN2	stathmin 2	
LYVE1	lymphatic vessel endothelial hyaluronan receptor 1	[20]
RGMA	repulsive guidance molecule BMP co-receptor a	[21]
TUBB2B	tubulin beta 2B class IIb	
LIFR	LIF receptor subunit alpha	[22]
LRRN2	leucine rich repeat neuronal 2	[23]
SPARCL1	SPARC like 1	[24]
ERBB2 (HER2)	erb-b2 receptor tyrosine kinase 2	[25]
MALAT1	metastasis associated lung adenocarcinoma transcript 1	
PCDH19	protocadherin 19	
HSPA6	heat shock protein family A (Hsp70) member 6	
HCAR2	hydroxycarboxylic acid receptor 2	
MYH7B	myosin heavy chain 7B	[26]
EPHB6	EPH receptor B6	[27]
MUC5AC	mucin 5AC, oligomeric mucus/gel-forming	[28]
CPS1	carbamoyl-phosphate synthase 1	[29]
ITLN2	intelectin 2	
RNU4-2	RNA, U4 small nuclear 2	

https://doi.org/10.1371/journal.pone.0302109.t004

prognosis group after PSM. It is worth noting that our study includes a high proportion of advanced-stage cases, with 35.7% of cases in TNM stage IV and 33.9% in stage III. Previous investigations into the expression level of MALAT1 have mainly focused on early-stage patients, and there are limited reports on the expression of MALAT1 in advanced stage patients.

MUC5AC is one of the mucins and a high molecular weight glycoprotein. While it is not expressed in normal colonic mucosa, it is expressed during CRC progression [40, 41]. However, the function of dysregulated MUC5AC expression has not been well characterized. A meta-analysis by Li et al. found that a high level of MUC5AC was associated with an improved prognosis [42]. Consistent with this, the mRNA abundance of MUC5AC was significantly decreased in the poor-prognosis group in our study. It is possible that MUC5AC plays a disease-modifying function in some CRC patients.

PSM also allowed us to discover several lesser studied or previously unknown biomarkers of CRC. For example, a MYH7B coding variant was discovered as one of the eight novel variants associated with CRC risk in a Swedish population from genome-wide association studies of 1,515 CRC patients and 12,108 controls [43]. MYH7B association with CRC risk was also reported in a gene expression prediction model of a large cohort of CRC cases that included

58,131 CRC cases and 67,347 controls of European ancestry [26]. By applying PSM, we identified the significance of MYH7B in CRC from a much smaller cohort than these two studies. Although further research is needed to confirm many of the potential new markers identified in our study, our results demonstrate the applicability and effectiveness of PSM for omics data analyses.

PSM analysis has some well-known limitations. It is crucial to list all possible confounding features without omission because the confounding factors that PSM can eliminate are only the ones that are already known. If unknown confounding features exist, the PSM analysis may not work correctly [44]. Additionally, PSM reduces the number of cases, and the result of post-PSM analysis may contain type 2 errors [45, 46]. In the differential expression analysis of RNA-seq dataset, this explains why there were significantly fewer genes in the analysis after PSM (122 genes after PSM vs. 402 genes before PSM). There is a possibility of overlooking some significant genes that should be focused on due to type 2 errors. This problem could be addressed by collecting a greater number of cases at the beginning. Using 2-to-1 matching in PSM can indeed be a useful arrangement to minimize the problem of type 2 errors [47, 48]. In this approach, two cases with most similar propensity scores are matched to one case in the opposite group. This can help to increase the number of cases while still maintaining balance in the covariates between the two groups. For the CRC proteomic marker dataset, we used this approach to match a good prognosis group of 104 cases to a poor prognosis group of 52 cases (S1 Fig, S2 and S3 Tables) with overall results very similar to 1-to-1 PSM. This arrangement can be useful in some datasets where increasing the number of cases is important while maintaining a balance of covariates.

Conclusion

Propensity score matching (PSM) is a valuable and cost-effective method in cohort studies and omics data analyses as it allows researchers to create comparable groups with homogenized backgrounds using fewer cases. By reducing the impact of confounding factors, PSM can improve the accuracy and reliability of biomarker discovery.

Supporting information

S1 Checklist. Human participants research checklist.
(DOCX)

S1 Fig. Distribution of propensity scores of the IHC CRC dataset with a 2-to-1 matching arrangement.
(TIF)

S1 Table. The details of IHC antibody staining and scorings.
(DOCX)

S2 Table. Genes commonly selected before and after PSM.
(DOCX)

S3 Table. Clinicopathological features of IHC scoring dataset after PSM with a 2-to-1 matching arrangement.
(DOCX)

S4 Table. Comparison of IHC scoring between good and poor prognosis groups after 2-to-1 PSM. Bold: $p<0.05$.
(DOCX)

Acknowledgments

We would like to thank Dr. Julia Y. Wang for providing advice on data analyses and presentation and for careful reviewing and editing of the manuscript. We are grateful to the IHC core facility at MSKCC for their invaluable support and assistance throughout this study. The results here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Author Contributions

Conceptualization: Masaki Maekawa, Michael H. Roehrl.

Data curation: Masaki Maekawa, Atsushi Tanaka, Makiko Ogawa.

Formal analysis: Masaki Maekawa.

Funding acquisition: Michael H. Roehrl.

Investigation: Masaki Maekawa, Atsushi Tanaka, Makiko Ogawa, Michael H. Roehrl.

Methodology: Masaki Maekawa, Atsushi Tanaka, Makiko Ogawa.

Project administration: Michael H. Roehrl.

Resources: Masaki Maekawa.

Software: Masaki Maekawa.

Validation: Masaki Maekawa.

Visualization: Masaki Maekawa.

Writing – original draft: Masaki Maekawa.

Writing – review & editing: Michael H. Roehrl.

References

1. <References>. Rosenbaum P.R. and Rubin D.B., The central role of the propensity score in observational studies for causal effects. Biometrika, 1983. 70: p. 41–55.
2. Austin P.C., A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med, 2008. 27(12): p. 2037–49. https://doi.org/10.1002/sim.3150 PMID: 18038446
3. Shah B.R., et al., Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol, 2005. 58(6): p. 550–9. https://doi.org/10.1016/j.jclinepi.2004.10.016 PMID: 15878468
4. Joffe M.M. and Rosenbaum P.R., Invited commentary: propensity scores. Am J Epidemiol, 1999. 150 (4): p. 327–33. https://doi.org/10.1093/oxfordjournals.aje.a010011 PMID: 10453808
5. Kitsios G.D., et al., Can We Trust Observational Studies Using Propensity Scores in the Critical Care Literature? A Systematic Comparison With Randomized Clinical Trials. Crit Care Med, 2015. 43(9): p. 1870–9. https://doi.org/10.1097/CCM.0000000000001135 PMID: 26086943
6. Tanaka A., et al., Prolyl 4-hydroxylase alpha 1 protein expression risk-stratifies early stage colorectal cancer. Oncotarget, 2020. 11(8): p. 813–824. https://doi.org/10.18632/oncotarget.27491 PMID: 32166002
7. Tanaka A., et al., Maspin as a Prognostic Marker for Early Stage Colorectal Cancer With Microsatellite Instability. Front Oncol, 2020. 10: p. 945. https://doi.org/10.3389/fonc.2020.00945 PMID: 32587829
8. Cancer Genome Atlas N., Comprehensive molecular characterization of human colon and rectal cancer. Nature, 2012. 487(7407): p. 330–7. https://doi.org/10.1038/nature11252 PMID: 22810696
9. Colaprico A., et al., TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res, 2016. 44(8): p. e71. https://doi.org/10.1093/nar/gkv1507 PMID: 26704973
10. Zhu Y., Qiu P., and Ji Y., TCGA-assembler: open-source software for retrieving and processing TCGA data. Nat Methods, 2014. 11(6): p. 599–600. https://doi.org/10.1038/nmeth.2956 PMID: 24874569

11. Anders S. and Huber W., Differential expression analysis for sequence count data. Genome Biol, 2010. 11(10): p. R106. https://doi.org/10.1186/gb-2010-11-10-r106 PMID: 20979621

12. Robinson M.D., McCarthy D.J., and Smyth G.K., edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010. 26(1): p. 139–40. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308

13. Liu S., et al., Three Differential Expression Analysis Methods for RNA Sequencing: limma, EdgeR, DESeq2. J Vis Exp, 2021(175). https://doi.org/10.3791/62528 PMID: 34605806

14. Love M.I., Huber W., and Anders S., Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol, 2014. 15(12): p. 550. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281

15. Xu Y., et al., Quantitative detection of circulating MT-ND1 as a potential biomarker for colorectal cancer. Bosn J Basic Med Sci, 2021. 21(5): p. 577–586. https://doi.org/10.17305/bjbms.2021.5576 PMID: 33823124

16. Elabd N.S., et al., Long Non-Coding RNAs ASB16-AS1 and AFAP1-AS1: Diagnostic, Prognostic Impact and Survival Analysis in Colorectal Cancer. Appl Clin Genet, 2022. 15: p. 97–109.

17. Liu J., et al., SCARA5 is a Novel Biomarker in Colorectal Cancer by Comprehensive Analysis. Clin Lab, 2020. 66(7). https://doi.org/10.7754/Clin.Lab.2019.191015 PMID: 32658413

18. Chen W., et al., Identification of a Tumor Microenvironment-Related Gene Signature Indicative of Disease Prognosis and Treatment Response in Colon Cancer. Oxid Med Cell Longev, 2021. 2021: p. 6290261. https://doi.org/10.1155/2021/6290261 PMID: 34497681

19. Ha Y.J., et al., Reduced expression of alanyl aminopeptidase is a robust biomarker of non-familial adenomatous polyposis and non-hereditary nonpolyposis colorectal cancer syndrome early-onset colorectal cancer. Cancer Med, 2023. 12(8): p. 10091–10104. https://doi.org/10.1002/cam4.5675 PMID: 36748835

20. Gao F., et al., Expression and quantification of LYVE-1 in human colorectal cancer. Clin Exp Med, 2006. 6(2): p. 65–71. https://doi.org/10.1007/s10238-006-0097-4 PMID: 16820993

21. Li V.S., et al., Frequent inactivation of axon guidance molecule RGMA in human colon cancer through genetic and epigenetic mechanisms. Gastroenterology, 2009. 137(1): p. 176–87. https://doi.org/10.1053/j.gastro.2009.03.005 PMID: 19303019

22. Li D., et al., Discovery and validation of tissue-specific DNA methylation as noninvasive diagnostic markers for colorectal cancer. Clin Epigenetics, 2022. 14(1): p. 102. https://doi.org/10.1186/s13148-022-01312-9 PMID: 35974349

23. Rong H., et al., Prognostic signatures and potential pathogenesis of eRNAs-related genes in colon adenocarcinoma. Mol Carcinog, 2022. 61(1): p. 59–72. https://doi.org/10.1002/mc.23359 PMID: 34622496

24. Han W., et al., Prognostic value of SPARCL1 in patients with colorectal cancer. Oncol Lett, 2018. 15 (2): p. 1429–1434. https://doi.org/10.3892/ol.2017.7511 PMID: 29434834

25. Haefliger S., et al., Prevalence of Molecular Alterations in a Swiss Cohort of 512 Colorectal Carcinoma Patients by Targeted Next-Generation Sequencing Analysis in Routine Diagnostics. Pathobiology, 2022: p. 1–10.

26. Guo X., et al., Identifying Novel Susceptibility Genes for Colorectal Cancer Risk From a Transcriptome-Wide Association Study of 125,478 Subjects. Gastroenterology, 2021. 160(4): p. 1164–1178.e6. https://doi.org/10.1053/j.gastro.2020.08.062 PMID: 33058866

27. Peng L., et al., Loss of EphB6 protein expression in human colorectal cancer correlates with poor prognosis. J Mol Histol, 2014. 45(5): p. 555–63. https://doi.org/10.1007/s10735-014-9577-0 PMID: 24912672

28. Betge J., et al., MUC1, MUC2, MUC5AC, and MUC6 in colorectal cancer: expression profiles and clinical significance. Virchows Arch, 2016. 469(3): p. 255–65. https://doi.org/10.1007/s00428-016-1970-5 PMID: 27298226

29. Lee Y.Y., et al., Overexpression of CPS1 is an independent negative prognosticator in rectal cancers receiving concurrent chemoradiotherapy. Tumour Biol, 2014. 35(11): p. 11097–105. https://doi.org/10.1007/s13277-014-2425-8 PMID: 25099619

30. Tanaka A., et al., STAT1 as a potential prognosis marker for poor outcomes of early stage colorectal cancer with microsatellite instability. PLoS One, 2020. 15(4): p. e0229252. https://doi.org/10.1371/journal.pone.0229252 PMID: 32275681

31. Bargmann C.I., Hung M.C., and Weinberg R.A., The neu oncogene encodes an epidermal growth factor receptor-related protein. Nature, 1986. 319(6050): p. 226–30. https://doi.org/10.1038/319226a0 PMID: 3945311

32. Coussens L., et al., Tyrosine kinase receptor with extensive homology to EGF receptor shares chromosomal location with neu oncogene. Science, 1985. 230(4730): p. 1132–9. https://doi.org/10.1126/science.2999974 PMID: 2999974

33. Yamamoto T., et al., Similarity of protein encoded by the human c-erb-B-2 gene to epidermal growth factor receptor. Nature, 1986. 319(6050): p. 230–4. https://doi.org/10.1038/319230a0 PMID: 3003577

34. Yu D. and Hung M.C., Overexpression of ErbB2 in cancer and ErbB2-targeting strategies. Oncogene, 2000. 19(53): p. 6115–21. https://doi.org/10.1038/sj.onc.1203972 PMID: 11156524

35. Revillion F., Bonneterre J., and Peyrat J.P., ERBB2 oncogene in human breast cancer and its clinical significance. Eur J Cancer, 1998. 34(6): p. 791–808. https://doi.org/10.1016/s0959-8049(97)10157-5 PMID: 9797688

36. Harari D. and Yarden Y., Molecular mechanisms underlying ErbB2/HER2 action in breast cancer. Oncogene, 2000. 19(53): p. 6102–14. https://doi.org/10.1038/sj.onc.1203973 PMID: 11156523

37. Tian X. and Xu G., Clinical value of lncRNA MALAT1 as a prognostic marker in human cancer: systematic review and meta-analysis. BMJ Open, 2015. 5(9): p. e008653. https://doi.org/10.1136/bmjopen-2015-008653 PMID: 26423854

38. Wei Y. and Niu B., Role of MALAT1 as a Prognostic Factor for Survival in Various Cancers: A Systematic Review of the Literature with Meta-Analysis. Dis Markers, 2015. 2015: p. 164635. https://doi.org/10.1155/2015/164635 PMID: 26420912

39. Sun Y. and Ma L., New Insights into Long Non-Coding RNA MALAT1 in Cancer and Metastasis. Cancers (Basel), 2019. 11(2). https://doi.org/10.3390/cancers11020216 PMID: 30781877

40. Niv Y. and Rokkas T., Mucin Expression in Colorectal Cancer (CRC): Systematic Review and Meta-Analysis. J Clin Gastroenterol, 2019. 53(6): p. 434–440. https://doi.org/10.1097/MCG.0000000000001050 PMID: 29782466

41. Bu X.D., et al., Altered expression of MUC2 and MUC5AC in progression of colorectal carcinoma. World J Gastroenterol, 2010. 16(32): p. 4089–94. https://doi.org/10.3748/wjg.v16.i32.4089 PMID: 20731025

42. Li C., et al., Prognostic and Clinicopathological Significance of MUC Family Members in Colorectal Cancer: A Systematic Review and Meta-Analysis. Gastroenterol Res Pract, 2019. 2019: p. 2391670. https://doi.org/10.1155/2019/2391670 PMID: 31933627

43. Jiao X., et al., Recurrent, low-frequency coding variants contributing to colorectal cancer in the Swedish population. PLoS One, 2018. 13(3): p. e0193547. https://doi.org/10.1371/journal.pone.0193547 PMID: 29547645

44. Reiffel J.A., Propensity Score Matching: The 'Devil is in the Details' Where More May Be Hidden than You Know. Am J Med, 2020. 133(2): p. 178–181. https://doi.org/10.1016/j.amjmed.2019.08.055 PMID: 31618617

45. Leisman D.E., Ten Pearls and Pitfalls of Propensity Scores in Critical Care Research: A Guide for Clinicians and Researchers. Crit Care Med, 2019. 47(2): p. 176–185. https://doi.org/10.1097/CCM.0000000000003567 PMID: 30543566

46. Schuster T., Lowe W.K., and Platt R.W., Propensity score model overfitting led to inflated variance of estimated odds ratios. J Clin Epidemiol, 2016. 80: p. 97–106. https://doi.org/10.1016/j.jclinepi.2016.05.017 PMID: 27498378

47. Staffa S.J. and Zurakowski D., Five Steps to Successfully Implement and Evaluate Propensity Score Matching in Clinical Research Studies. Anesth Analg, 2018. 127(4): p. 1066–1073. https://doi.org/10.1213/ANE.0000000000002787 PMID: 29324498

48. Baser O., Too much ado about propensity score models? Comparing methods of propensity score matching. Value Health, 2006. 9(6): p. 377–85. https://doi.org/10.1111/j.1524-4733.2006.00130.x PMID: 17076868