RESEARCH ARTICLE

Integrated meta-analysis of colorectal cancer public proteomic datasets for biomarker discovery and validation

Javier Robles[1,2], Ananth Prakash[3], Juan Antonio Vizcaíno[3]*, J. Ignacio Casal[1]*

1 Department of Molecular Biomedicine, Centro de Investigaciones Biológicas Margarita Salas, Consejo Superior de Investigaciones Científicas, Madrid, Spain, **2** Protein Alternatives SL, Tres Cantos, Madrid, Spain, **3** European Molecular Biology Laboratory—European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom

* juan@ebi.ac.uk (JAV); icasal@cib.csic.es (JIC)

OPEN ACCESS

Citation: Robles J, Prakash A, Vizcaíno JA, Casal JI (2024) Integrated meta-analysis of colorectal cancer public proteomic datasets for biomarker discovery and validation. PLoS Comput Biol 20(1): e1011828. https://doi.org/10.1371/journal.pcbi.1011828

Editor: James Gallo, University at Buffalo - The State University of New York, UNITED STATES

Received: September 19, 2023

Accepted: January 15, 2024

Published: January 22, 2024

Data Availability Statement: Protein abundances data along with sample metadata, experimental parameters and quality assessment summary of samples are available to download from Expression Atlas. The proteomics results in Expression Atlas can be accessed using the link: https://wwwdev.ebi.ac.uk/gxa/experiments/E-PROT-XXX/Results, where XXX should be replaced by the E-PROT accession number shown in Table 1. The raw proteomics datasets in PRIDE can be accessed using the link: https://www.ebi.ac.uk/pride/archive/projects/PXDxxxxxx, where PXDxxxxxx should be

Abstract

The cancer biomarker field has been an object of thorough investigation in the last decades. Despite this, colorectal cancer (CRC) heterogeneity makes it challenging to identify and validate effective prognostic biomarkers for patient classification according to outcome and treatment response. Although a massive amount of proteomics data has been deposited in public data repositories, this rich source of information is vastly underused. Here, we attempted to reuse public proteomics datasets with two main objectives: i) to generate hypotheses (detection of biomarkers) for their posterior/downstream validation, and (ii) to validate, using an orthogonal approach, a previously described biomarker panel. Twelve CRC public proteomics datasets (mostly from the PRIDE database) were re-analysed and integrated to create a landscape of protein expression. Samples from both solid and liquid biopsies were included in the reanalysis. Integrating this data with survival annotation data, we have validated *in silico* a six-gene signature for CRC classification at the protein level, and identified five new blood-detectable biomarkers (CD14, PPIA, MRC2, PRDX1, and TXNDC5) associated with CRC prognosis. The prognostic value of these blood-derived proteins was confirmed using additional public datasets, supporting their potential clinical value. As a conclusion, this proof-of-the-concept study demonstrates the value of re-using public proteomics datasets as the basis to create a useful resource for biomarker discovery and validation. The protein expression data has been made available in the public resource Expression Atlas.

Author summary

The need for new prognostic biomarkers is one of the main topics of interest in CRC (Colorectal Cancer). A potential strategy to address this issue, which to the best of our knowledge has not been attempted so far, is the combination of different public proteomic studies generated from independent patient cohorts. Despite the abundance of available proteomics data, meta-analyses have only been conducted at the genomic and

replaced by the PRIDE dataset identifier shown in the Table 1. In addition, the complete study (both individual and batch analyses) can be accessed in the PRIDE FTP server (https://ftp.ebi.ac.uk/pub/databases/pride/resources/colorectalcancer/). Finally, the main computational tools used in this study are R language based. Scripts are publicly available via GitHub at http://github.com/JRoblesEBI/CRC_proteomics_meta-analysis.

Funding: This project was supported by grants RTI2018-095055-B-I00, PID2021-122227OB-I00 and PDC2022-133056-I00 from the Ministerio de Ciencia e Innovación (MCIN/AEI/10.13039/501100011033) using Next Generation EU/PRTR funds to JIC, and BBSRC grant number BB/T019670/1 and EMBL core funding to JAV. JR was supported by IND2019/BMD-17153 fellowship of the Comunidad de Madrid. The funders did not play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

transcriptomic levels so far. In this study, we reanalyzed 12 mass spectrometry-based public proteomics datasets. In total, the combined dataset included 440 samples from 299 different patients, encompassing both solid and liquid biopsies. Consequently, we defined a proteomics landscape suitable for assessing protein expression in tumors and normal mucosa, its association with patient outcome, and its potential detection in liquid biopsies. Furthermore, as a proof-of-concept for the data reuse strategy, we demonstrated its capacity to validate an experimentally-based SEC6 gene signature at the protein level and to identify new blood-detectable biomarkers. The data generated in this study can be accessed by anyone since all the data have been made available in the Expression Atlas resource.

Introduction

Colorectal cancer (CRC) is the third most common cancer worldwide, accounting for approximately 10% of all diagnosed cancers [1]. Notably, the most developed countries present the highest rates of incidence [2]. Regarding global mortality, CRC is the second leading cause of cancer-related deaths [3]. CRC is considered a highly heterogeneous disease [4], caused through various genetic and epigenetic mechanisms, including microsatellite instability or mutations in oncogenes such as APC, TP53, KRAS, and BRAF. The molecular heterogeneity leads to variability in the pathogenesis, outcome and treatment response [5]. Both clinical and molecular heterogeneity are essential challenges when facing CRC diagnosis and prognosis [6] and to face this heterogeneity, several molecular classifications have been proposed [7]. CRC diagnostic procedures have made significant advances in the last years based on the massive use of faecal occult blood screening tests, liquid biopsies, and the more invasive colonoscopy [8]. However, once the disease is detected, the current clinical stratification systems, based on the pathological staging, presents some limitations that fail to identify a relevant number of patients relapsing and/or developing metastases after surgical resection. For those reasons, it is especially crucial to identify prognostic biomarkers to categorize patients in stage II and III to prevent recurrence, and to identify those patients who should receive more intensive treatments [9]. Although, most of the approaches designed to address CRC diagnosis and prognosis have relied so far in samples from solid biopsies [4,10] obtained using highly invasive procedures, biomarkers detectable in liquid biopsies are a preferable option for predicting and monitoring patient outcome and response to chemotherapy [11]. Recent studies have demonstrated the importance of secreted proteins on tumor progression and metastasis development [12]. In summary, despite significant progress in CRC screening and diagnosis has been achieved [13], the problem of CRC prognosis and classification remains unresolved.

Most of the available high-throughput omics data coming from cancer patients are based on DNA and RNA sequencing information. This applies to both solid and liquid biopsies. Indeed, most of the attempts to classify CRC are based on transcriptomics data [4,10]. Nevertheless, proteins are most often the functional molecules that undertake the translation from genotype into the phenotype. In addition, protein-based techniques such as ELISA or immunohistochemistry have demonstrated to be useful tools with clinical relevance [14]. Mass spectrometry (MS) is the main high-throughput proteomics approach for providing quantitative measurements of protein abundance/expression [15]. Although publicly available genomics and transcriptomics datasets are more in number and scope, large relevant proteomics studies such as those performed by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [16] or other independent teams, have obtained MS-based protein expression information in

CRC samples. Taken together, this protein expression data from different datasets can be considered as a robust source of information for biomarker discovery and confirmation.

Multiple proteomics datasets are freely available in public repositories. The PRIDE database (as the most popular resource in this context) [17], together with other open repositories (e.g. CPTAC portal [16], jPOST [18]) contain thousands of proteomics datasets. Public data in these resources coming from different sources can be reanalysed and integrated to obtain a global view and potentially discover new insights. Integrative meta-analyses have already demonstrated to be useful using genomics and transcriptomics data [19,20]. Regarding proteomics, quantitative reanalysis and integration of public data is emerging as a potent resource with multiple applications [21–25]. To our knowledge, no previous studies have addressed prognostic biomarker identification in CRC from this perspective of reusing public proteomics datasets. In the present study, we selected, reused, and integrated 12 public proteomics datasets containing samples from both solid and liquid biopsies of CRC patients. The reanalysed and processed data have been deposited in the Expression Atlas resource [26] to facilitate protein abundance data access and visualisation. In summary, we provide a CRC proteomic landscape with the capacity to validate previously reported prognostic biomarkers and to discover new sets of biomarkers, demonstrating the potential and value of reusing public proteomics datasets for biomarker discovery.

Results

Colorectal cancer proteomics datasets and integrative analysis

We queried PRIDE [17], jPOST [18] and the CPTAC [16] portal for CRC studies and selected 12 publicly available datasets for reanalysis (Table 1). These CRC datasets were divided in two groups, corresponding to secreted and solid tumor samples, respectively. The secretome group consisted of samples derived from blood, cell culture, extracellular vesicles, exosomes and interstitial fluid, whereas the solid tumor sample group consisted of samples derived from mucosa, adenoma (pre-malignant cellular masses), and tumor tissues. The secretome and solid tumor sample groups consisted of 7 and 5 datasets, respectively, and included paired healthy and cancer samples. The characteristics of the overall patient cohorts included in the datasets are shown in Table 2. Patient composition of the meta-analysis shows a non-biased, proportional distribution of the different sub-classifications of CRC, according to mutations, chromosomal stability, age and sex.

We reanalysed each sample group separately by creating two batches (solid and secreted samples). Each batch consisted of all MS runs from all datasets that were part of a sample group. Protein identification and quantification analysis was performed using MaxQuant [27] version 2.1.0.0 on a high-performance computing Linux cluster. In total, the datasets contained 2,458 MS runs, covering 440 samples from 299 individuals. The number of protein groups, peptides, and unique peptides identified in the two batches of samples are shown in Table 1. The protein abundances calculated from individual datasets are available to view and download from Expression Atlas [28], along with their experimental parameters, sample metadata and summary of sample quality assessment after post-processing. The data reanalysis protocol is summarised in Fig 1 and explained in detail in S1 Fig.

Protein abundance comparison in solid tumors and secreted samples

Protein abundances were calculated for the secretome and solid tumor samples using the intensity-based absolute quantification (iBAQ) method and normalised by the FOT (Fraction of Total) method (see Methods). We converted the normalised iBAQ abundance values of each sample into five abundance bins as previously described [25], for ease of comparison of

Table 1. List of the proteomics datasets used in this study.

Combined analysis	Tissues	Organs	Proteomics dataset identifier*	Expression Atlas identifier	Number of MS runs	Fractionation (Fractions per sample)	Number of samples	Number of patients	Number of protein groups†	Number of peptides†	Number of unique peptides†	Number of unique genes (canonical proteins) mapped†
Solid samples	Mucosa, colorectal adenoma, colorectal carcinoma	Colorectum	PXD001676 [58]	E-PROT-103	16	No	16	8	9,711	215,033	196,017	8,949
			PXD002137 [59]	E-PROT-104	192	Yes (6)	32	25				
			PXD014511 [60]	E-PROT-105	310	Yes (5)	62	52				
			PXD019504 [61]	E-PROT-106	74	No	74	37				
			CPTAC PDC000111 [31]	E-PROT-23	1425	Yes (15)	90	90				
Total solid tumor			**5 datasets**		**2,017**		**274**	**212**				
Secreted samples	Interstitial fluid, extracellular vesicles, blood serum, cell lines	Colorectum, liver, blood	PXD005709 [62]	E-PROT-100	36	Yes (3)	12	6	5,861	85,013	79,181	5,091
			PXD005693 [62]	E-PROT-101	15	No	15	8				
			PXD020454 [63]	E-PROT-102	45	Yes (3)	15					
			PXD010458 [64]	E-PROT-107	144	Yes (24)	6	16				
			JPST000867 [65]	E-PROT-108	68	No	36	17				
			PXD031556 [66]	E-PROT-109	79	No	79	40				
			PXD032899 [30]	E-PROT-110	54	Yes (3)	3					
Total secreted			**7 datasets**		**441**		**166**	**87**				
TOTAL			**12 datasets**		**2,458 MS runs**		**440 samples**	**299 patients**				

*Dataset identifiers starting with 'PXD' come from the PRIDE database, dataset JPST000867 is from jPOST and dataset PDC000111 is from CPTAC portal. Unique protein sample batches available in any given dataset are considered as individual samples. † Numbers after post-processing. The proteomics results in Expression Atlas can be accessed using the link: https://www.ebi.ac.uk/gxa/experiments/E-PROT-XXX/Results, where XX should be replaced by the E-PROT accession number shown in the table. The raw proteomics datasets in PRIDE can be accessed using the link: https://www.ebi.ac.uk/pride/archive/projects/PXDxxxxxx, where PXDxxxxxx should be replaced by the PRIDE dataset identifier shown in the table.

https://doi.org/10.1371/journal.pcbi.1011828.t001

Table 2. Clinicopathological features of colorectal adenoma and adenocarcinoma patients included in the proteomic datasets.

Colorectal adenoma patients			Colorectal adenocarcinoma (CRC) patients		
Clinicopathological features		**Cases (%)**	**Clinicopathological features**		**Cases (%)**
Age	Age ≤ 65	56.41	**Age**	Age ≤ 65	27.66
	Age > 65	43.59		Age > 65	47.52
	Unspecified	0.00		Unspecified	26.95
Gender	Male	57.69	**Gender**	Male	37.59
	Female	41.03		Female	37.59
	Unspecified	0.00		Unspecified	26.95
Site	Colon	80.77	**Site**	Colon	50.35
	Rectum	17.95		Rectum	24.82
	Unspecified	0.00		Unspecified	26.95
Subtype	Conventional adenoma	19.23	**Stage**	I	11.35
	Sessile serrated adenoma	21.79		II	26.95
	Traditional serrated adenoma	15.38		III	19.86
	Unspecified	20.51		IV	9.22
				Unspecified	32.62
			Microsatellite status	MSS	58.87
				MSI	31.91
				Unknown	11.35
			***BRAF* status**	*BRAF* mut	5.67
				BRAF WT	58.16
				Unspecified	38.30
			***KRAS* status**	*KRAS* mut	29.08
				KRAS WT	34.75
				Unspecified	38.30
			***TP53* status**	*TP53* mut	24.82
				TP53 WT	39.01
				Unspecified	38.30

https://doi.org/10.1371/journal.pcbi.1011828.t002

abundances across samples and datasets. The abundance bins ranged from bin1 to bin5, denoting proteins with lowest to highest abundance, respectively. We mapped the protein identifiers within the protein groups to their respective parent gene names equivalent to 'canonical proteins' in UniProt (see 'Methods') and, hence, when describing protein abundances in the manuscript we refer to the abundance of the 'canonical proteins'.

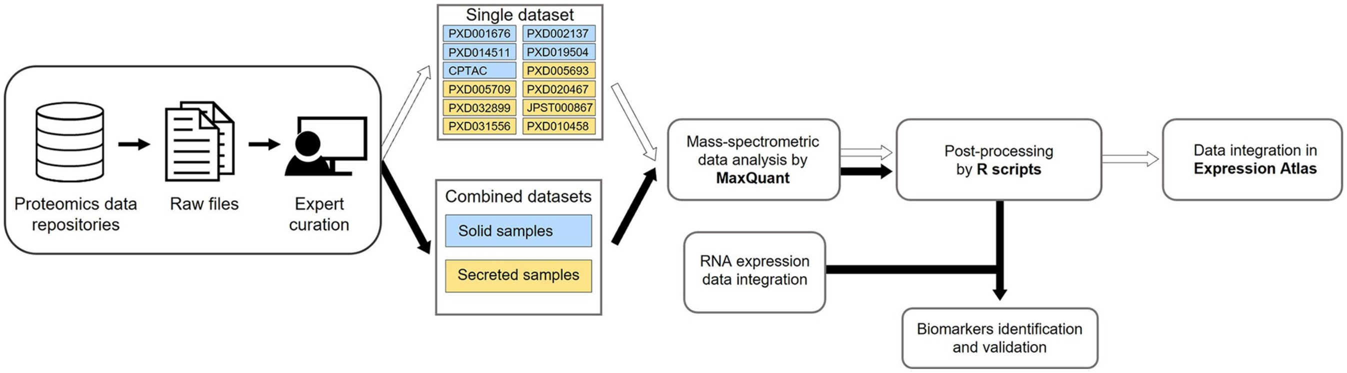
Proteomics data repositories
Raw files
Expert curation
Single dataset
PXD001676
PXD002137
PXD014511
PXD019504
CPTAC
PXD005693
PXD005709
PXD020467
PXD032899
JPST000867
PXD031556
PXD010458
Combined datasets
Solid samples
Secreted samples
Mass-spectrometric data analysis by MaxQuant
Post-processing by R scripts
Data integration in Expression Atlas
RNA expression data integration
Biomarkers identification and validation

Fig 1. Scheme of the project design and the data reanalysis pipeline. Workflow of the selection, curation, reanalysis and integration of public proteomic dataset containing CRC samples.

https://doi.org/10.1371/journal.pcbi.1011828.g001

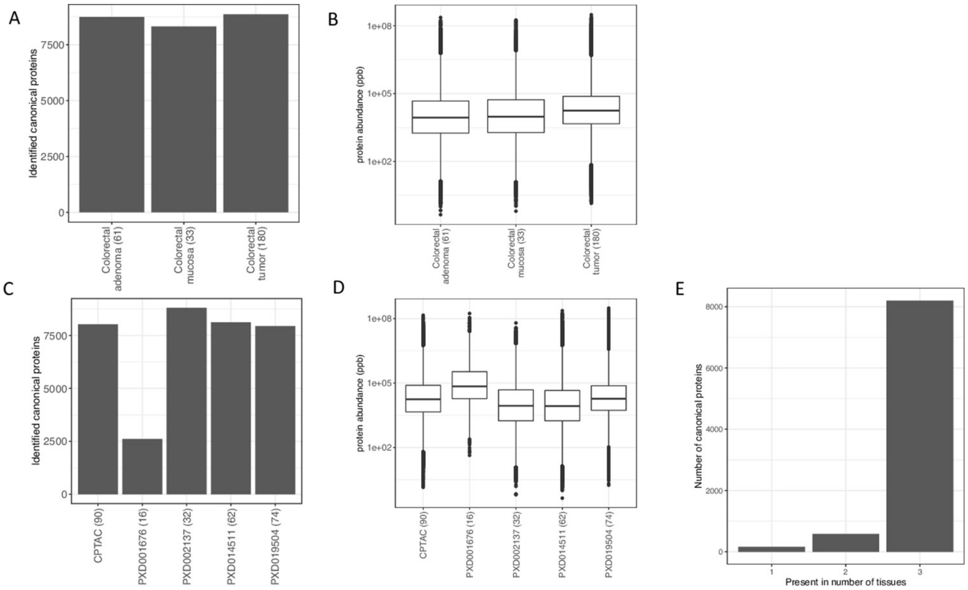

Fig 2. Colorectal cancer (CRC) solid samples. (A) Number of canonical proteins identified across different solid tumor samples. (B) Range of normalised iBAQ protein abundances across different samples. (C) Number of canonical proteins identified across different datasets. (D) Range of normalised iBAQ protein abundances across different datasets. (E) Number of canonical proteins identified across either one, two or three of the different solid samples subgroups (mucosa, adenoma and tumor). The number within the parenthesis indicate the number of samples.

https://doi.org/10.1371/journal.pcbi.1011828.g002

Protein identification in solid tumor samples

Across all tumor samples we identified a total of 9,711 protein groups, which we mapped to 8,949 parent genes (canonical proteins). We identified similar numbers of canonical proteins across samples from colon mucosa, colon adenoma, which are pre-cancerous cell masses, and colorectal tumor samples (Fig 2A). The dynamic range of protein abundances was similar across tissue samples (Fig 2B). Across datasets, PXD001676 had the lowest number (29.1%) of identified canonical proteins (Fig 2C). However, dataset PXD001676 showed a higher median protein abundance relative to the other datasets (Fig 2D), probably due to the fewer number of canonical proteins identified in this dataset. We found that the large majority of canonical proteins was simultaneously present in the three types of samples: mucosa, adenoma and tumor (Fig 2E). Compared to the secreted samples, the correlation of protein abundance among tumor tissues was relatively low (S2 Fig). The highest correlation was observed in adenoma samples (median R^2 = 0.58) and the lowest correlation was among mucosa samples (median R^2 = 0.49).

Protein identification in secreted samples

We obtained 5,861 protein groups from the secreted samples, which were mapped to 5,091 parent genes (canonical proteins). We identified a large number of proteins in samples from interstitial fluid (87.5%) and extracellular vesicles and exosomes (74.2%), while comparatively less proteins were identified in blood-derived (37.2%) and cell culture samples (39.4%) (Fig 3A). The dynamic range of protein abundances across secreted samples and datasets is shown (Fig 3B and 3D). Among datasets, the largest numbers of canonical proteins were identified in datasets PXD005709, JPST000867 and PXD005693 (Fig 3C). The median protein abundance was highest in cell culture-derived samples compared to other samples, and, similarly, dataset PXD020454 had a higher median protein abundance, likely due to the low number of canonical proteins identified in this dataset. The number of canonical proteins identified across all

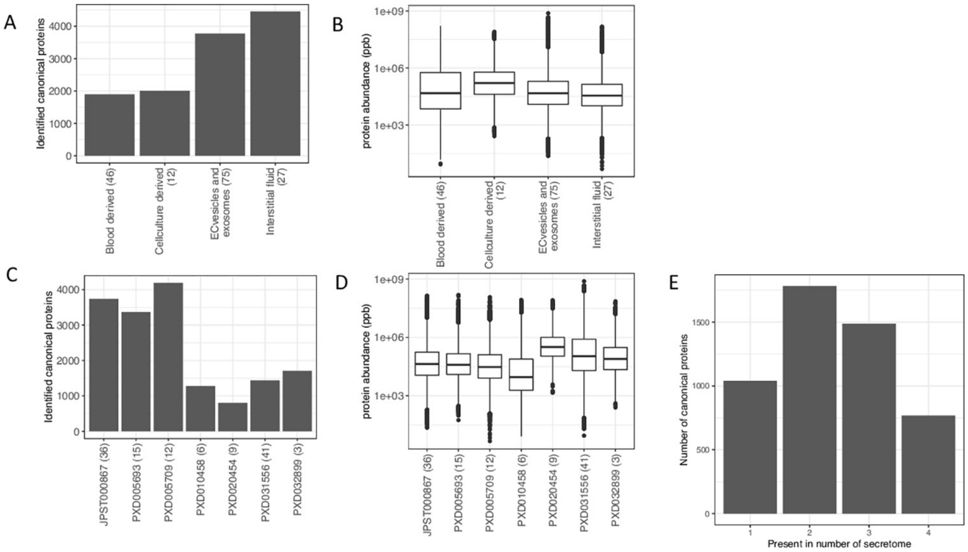

Fig 3. Colorectal cancer secreted samples. (A) Number of canonical proteins identified across different secreted samples. (B) Range of normalised iBAQ protein abundances across different samples. (C) Number of canonical proteins identified across different datasets. (D) Range of normalised iBAQ protein abundances across different datasets. (E) Number of canonical proteins identified across either one, two, three or four of the different secreted samples subgroups. The number within the parenthesis indicate the number of samples.

https://doi.org/10.1371/journal.pcbi.1011828.g003

subgroups was relatively low, likely influenced by the unequal distribution of identifications among the subgroups. Indeed, the majority of canonical proteins were exclusively present in either two or three of the subgroups (Fig 3E). We used the binned protein abundances to compare the correlation of protein abundances across secreted samples. We also observed a good correlation and clustering of samples at the heatmap (S3 Fig). The highest correlation in protein expression was observed among blood-derived samples (median $R^2 = 0.80$) and the lowest correlation was observed among cell culture-derived samples (median $R^2 = 0.32$), which confirms the high heterogeneity of cancer cell lines.

Proteins in secreted fractions mirror tumor alterations

To better understand the biological alterations during CRC tumorigenesis, following the pipeline described in 'Methods', we identified differentially-expressed proteins comparing normal mucosa, adenomas and CRC tumor samples. First, we performed Gene Ontology (GO) enrichment analysis to compare the three types of solid samples (Fig 4A). The GO categories significantly enhanced in tumor samples when compared with the normal mucosa and adenoma were 'cell cycle' (GO0007049), 'telomerase maintenance' (GO0000723), 'translation' (GO0006412), and 'ribosome biogenesis' (GO0042254). These categories are related to an increased proliferation in the tumors. Conversely, the most significant biological process decreased during tumor progression was 'cellular respiration' (GO0045333), probably due to the Warburg effect [29]. When analysing the GO Cellular Component category, there was a significant enrichment of the secreted fraction 'secretory granule lumen' term (GO0034774) (Fig 4B). This enrichment was found to be higher in the overexpressed proteins when comparing "Adenoma vs Mucosa", "Tumor vs Mucosa" and "Tumor vs Adenoma" indicating that the secreted fraction increases through tumor progression in CRC, due to the high intestinal secretory component.

Given the relevance of the secreted fraction in CRC progression[11], the level of agreement in protein expression between solid tumor samples and secreted fractions was investigated. Most

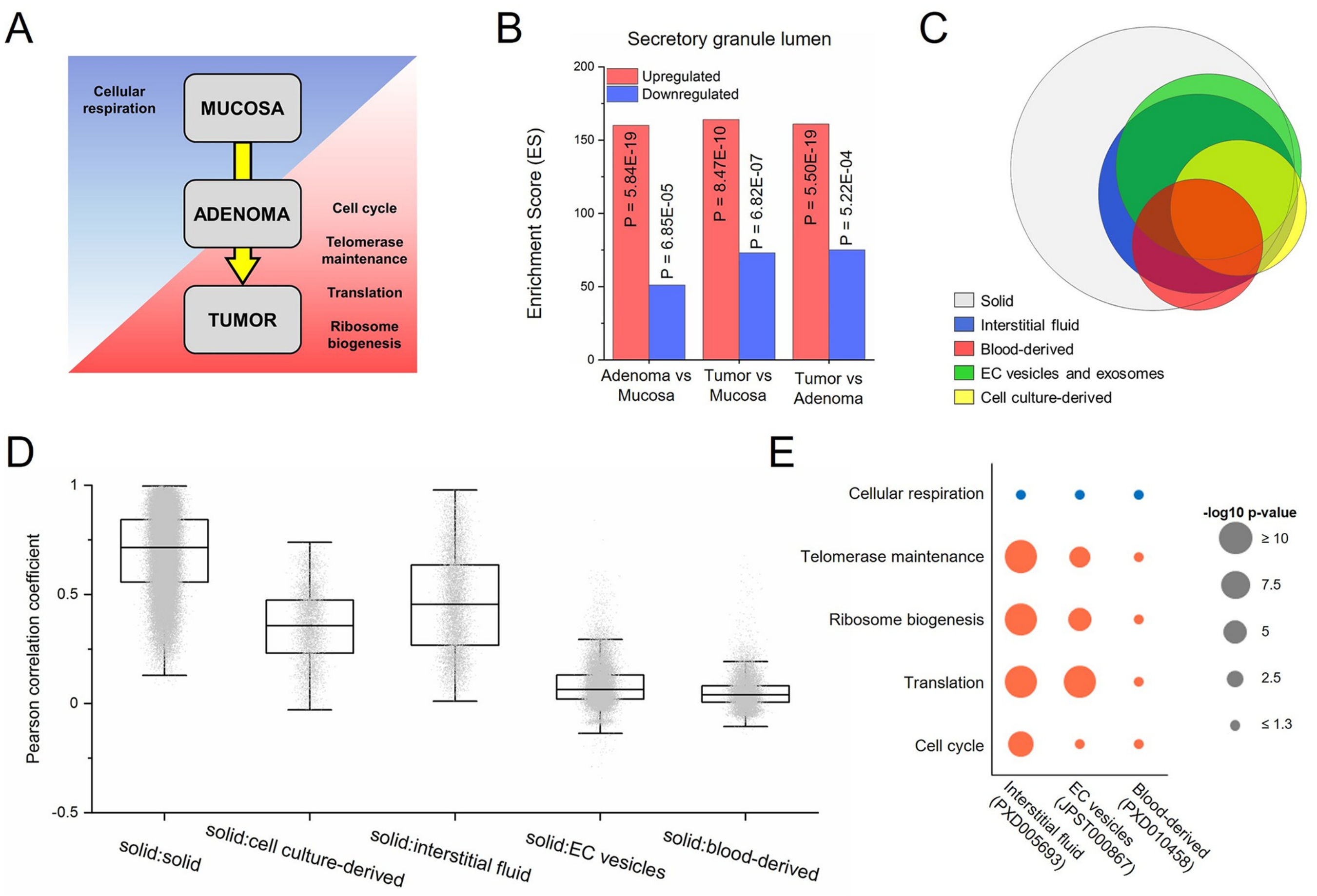

Fig 4. Analysis of GO enrichment and concordance between solid and secreted samples. (A) Plot summary illustrating the most altered GO terms in the "biological process" category in solid samples. (B) Enrichment in 'Secretory granule lumen' (GO0034774) category of upregulated and downregulated proteins when comparing mucosa, adenoma and tumor samples. (C) Venn diagram representing the detected canonical proteins in solid and secreted samples. (D) Boxplot showing the correlation between solid samples and each of the subgroups of secreted samples. (E) Significance of enrichment in GO Biological Process categories of altered proteins (tumor vs normal mucosa) in different secreted subgroups.

https://doi.org/10.1371/journal.pcbi.1011828.g004

of the proteins quantified in the secreted fractions were also present in the solid samples (Fig 4C). To determine concordance, the Pearson correlation coefficient (r) values between solid tumor samples and each secreted source were calculated (Fig 4D). On secreted fractions, the interstitial fluid and cell culture-derived samples showed greater correlations with the tumor samples than with extracellular vesicles and blood-derived samples. Then, we tested whether the secreted fractions displayed similar biological alterations to those found between tumor and mucosa in solid samples (Fig 4E). Interstitial fluid, and EC vesicle samples to a lower extent, were able to replicate the changes observed in the solid tumors, where blood-derived samples were not. In addition, correlation between the fold-changes in tumor *vs* mucosa was analysed. Whereas EC vesicles and interstitial fluid samples presented a high correlation with solid samples (R = 0.48 and R = 0.45, respectively), blood-derived samples did not (R = 0.05) (S4A Fig). Of note it is the high correlation found between EC vesicles and interstitial fluid samples (R = 0.80).

We defined as "enriched proteins" those canonical proteins exclusively or highly expressed in just one of the secreted subgroups (S4B Fig). Then, "enriched proteins" of each secreted source were analysed to identify the overrepresented functions in each subgroup (S4C Fig).

Interstitial fluid proteins were particularly relevant in the study of 'cellular respiration' (GO0045333) and 'translation' (GO0006412), while EC vesicles were the best option for the 'cell cycle' (GO0007049). "Enriched proteins" from cell culture supernatants and blood-derived samples were mainly involved in 'wound healing' (GO0042060) and 'blood coagulation' (GO0007596). Moreover, in all subgroups, "enriched proteins" were significantly associated with GO Cellular Component categories corresponding to secreted fractions, such as 'EC vesicle' (GO1903561) or 'EC region' (GO0005576) (S4D Fig). Overall, these analyses suggest that secreted fractions represent a suitable model for the study of protein expression in CRC, as they show a remarkable correlation with solid tumor alterations.

Concordance between mRNA and protein-based prognostic biomarkers in colorectal cancer

To the best of our knowledge, there has been limited investigation into the large-scale concordance between gene expression- and protein expression-based biomarkers. To investigate which protein alterations are present at the gene expression level, a comparison between transcriptomics and proteomics fold-changes in tumor and normal mucosa samples was performed. Proteomics data from the solid samples batch was used after normalization into ranked bins. Transcriptomics data from primary tumor and normal mucosa samples was obtained from the GSE41258 public dataset from GEO (see –Methods–). Our analysis showed an excellent agreement between transcriptomics and proteomics alterations, as shown in a volcano plot representing all the quantified proteins (Fig 5A) and a scatter plot presenting only the significantly altered proteins (Fig 5B). Despite the overall concordance, we still found some disagreements. To explain these discrepancies, we analysed separately proteins that were significantly altered in either one or both analyses. Firstly, we examined the protein expression levels in tumor samples. Proteins significantly altered between tumor and mucosa in the proteomic analysis showed higher expression ranked values than the proteins corresponding to the altered genes detected in transcriptomics (Fig 5C). These results suggest that some of the exclusively-detected transcriptomic alterations were not detected using proteomics because expression values were too low to find differences between conditions. To examine the GO Cellular Component differences, three significant representative locations were identified: 'Secretory granule lumen' (GO0034774), 'Mitochondrial matrix' (GO0005759) and 'Nucleus' (GO0005634). We obtained different location profiles depending on the technique used for the detection. Proteomics-derived proteins were mainly enriched in secreted proteins, whereas nuclear genes were the most enriched location in transcriptomic-exclusive genes (Fig 5D). These findings might be explained by the relatively low expression of transcription, splicing and other nuclear factors.

Next, we evaluated the predictive potential of protein expression and its concordance with the RNA-seq values. Firstly, we selected proteins quantified in the CPTAC COADREAD dataset and the corresponding mRNA data in the CPTAC samples within the TCGA COADREAD datasets. CPTAC was selected because is the only CRC public proteomics dataset that contains patient survival data. In addition, it is the only dataset in which samples have been analysed both by proteomics and RNA-seq. Then, Cox regression analysis was performed to identify proteins or genes associated with survival. Significant proteins in the Cox regression analysis were plotted in scatter plots representing the hazard ratios (HR). A relatively low correlation was observed when comparing proteomics CPTAC data with CPTAC and TCGA COADREAD RNA-seq datasets (Fig 5E and 5F). Then, the ratio of significant proteins (CPTAC) and genes (TCGA) was determined. In the CPTAC dataset, from the initial 2,661 analysed proteins, 142 (5.3%) were significantly associated with prognosis, whereas 285 (10.7%) genes were

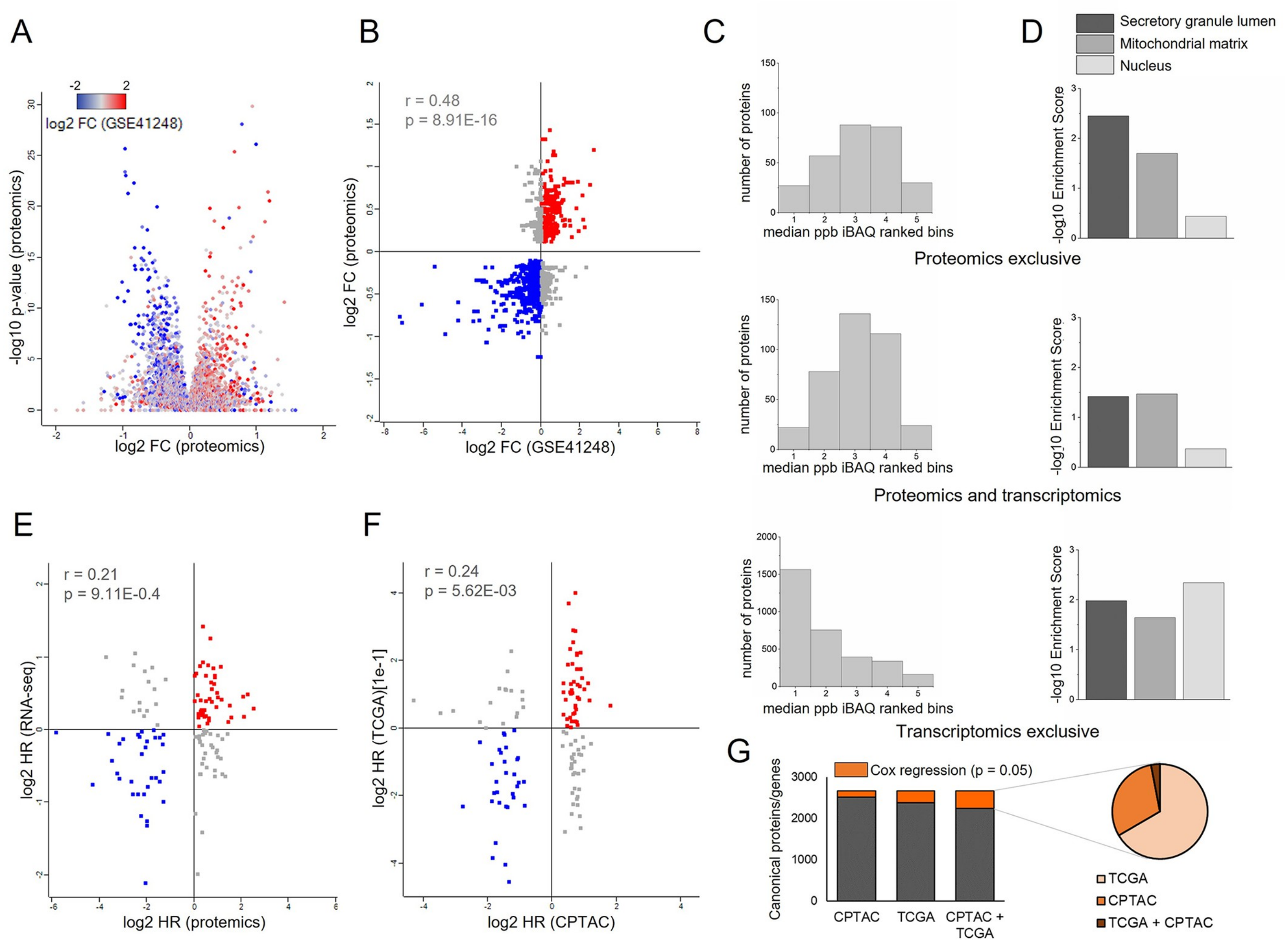

Fig 5. Correlation between transcriptomics and proteomics analysis. (A) Volcano plot distribution of the proteomics data. Fold change (Tumor/mucosa) is represented. Dots are labelled according to the transcriptomic fold change. (B) Scatter plot of significantly altered canonical proteins. Pearson's coefficient is shown. (C) Histogram distribution of protein expression levels (ranked bins) and (D) Cellular Component analysis of proteins or genes significantly altered between normal mucosa and tumor in proteomics and/or transcriptomics. More relevant categories were selected. (E) Correlation of hazard ratios obtained using proteomics data and RNA-seq data from CPTAC. (F) Correlation of hazard ratios obtained using proteomics data (CPTAC) and RNA-seq data (TCGA). Only significantly prognosis-associated proteins are shown. Pearson correlation coefficient (r) is indicated. (G) Portion of the canonical proteins significantly associated with survival according to proteomics (CPTAC) and/or transcriptomics (TCGA) data. Significance was calculated by Cox regression analysis. Chart pie of all the significant associated mRNA (TCGA) and proteins (CPTAC).

https://doi.org/10.1371/journal.pcbi.1011828.g005

significant in the analysis of the corresponding genes in the TCGA dataset (Fig 5G). A pie chart showed that 130 (30.4%) of the significant potential biomarkers were exclusively detected by proteomics (Fig 5G). This suggests that potential biomarkers identified by proteomics may go unnoticed when analysing mRNA values, and vice versa. Therefore, when searching for biomarkers on a large scale, it is optimal to integrate proteomics and transcriptomics data to attain a more comprehensive biomarker panel.

Validation of experimentally-derived biomarkers using public proteomics datasets

As an initial proof-of-concept, we applied our strategy to the validation of a recently reported prognostic and predictive signatures for CRC based on the expression of six genes (BMP1, CD109, IGFBP3, LTBP1, NPC2, PSAP), which were identified through proteomics analysis of the secretoma of metastatic cancer cells and validated in transcriptomics datasets [30]. Protein descriptions and their UniProt accession numbers can be found in S1 Table. Therefore, we determined whether this signature could also be detected in the proteomic datasets, according to our in silico reanalysis strategy. At the proteomic level, four out of six proteins, corresponding to the genes CD109, LTBP1, NPC2 and PSAP, were detected in more than 50% of the solid tumor samples included in the CPTAC dataset (Fig 6A), and were considered for their association with survival using Kaplan-Meier analysis (Fig 6B). A clear trend was observed, since patients with high expression in all four proteins showed lower overall survival rate. As only four proteins were quantified using proteomics datasets, instead of the original calculation of the risk score based on an algorithm including the six genes, a simpler approach based on the mean of the four available proteins was performed to classify patients in high and low risk. A previous normalization (division by mean) was done to avoid the high differences between protein levels. Kaplan-Meier analysis showed a significant association of the combined protein expression with poor survival (Fig 6C). Furthermore, as SEC6 are secreted proteins, their presence in blood-derived samples was analysed. In this case, all the proteins, except BMP1, were detected in blood-derived samples, and four of them (IGFBP3, LTBP1, NPC2 and PSAP) were highly increased in CRC patients (Fig 6D). The high levels found in the blood of colorectal patients agree with the previous detection in the secretome and represent a promising added value to the signature.

Discovery of new secreted biomarkers using public proteomics datasets

Finally, we searched for new proteomics-derived prognostic biomarkers, as an additional proof-of-concept for our hypothesis. To achieve this, we followed the indicated workflow (Fig 7A). Proteins quantified in 70% of all solid tumor samples and, then, at least in 70% of CPTAC samples were selected. Association with survival was determined using Cox regression analysis. We identified 130 proteins significantly associated with prognosis (Fig 5G). For potential biomarker detection in liquid biopsies, selection was restricted to those proteins overexpressed in blood samples from CRC patients. Five candidate biomarkers (CD14, MRC2, PPIA, PRDX1, and TXNDC5) were obtained. Whereas PRDX1 and TXNDC5 were biomarkers of good prognosis, CD14, PPIA and MRC2 were associated with poor prognosis (Fig 7B). Protein descriptions and UniProt accession numbers can be found in S2 Table. Kaplan-Meier analysis confirmed a prognostic value when considering all patients (Fig 7B) or only patients in stages II and III (S5A Fig), where clinical treatments require more complex decisions. To test if these potential biomarkers could be also used in combination, patients were classified in high or low risk according to the expression levels of each protein (S5B Fig). The significance of the classification increased when we considered as high-risk patients those in whom at least three of the five proteins classified them as high risk. Kaplan-Meier analysis was significant for either all the patients or stage II and III patients (S5C Fig).

Next, the prognostic efficiency of these proteins was evaluated at the transcriptomic level using the TCGA RNA-seq database and the best cut-off Kaplan-Meier significance. All the corresponding genes were significant, except for TXNDC5, which was close to significance (6A Fig). When a more restrictive Cox regression test was performed, these biomarkers showed significance using proteomics data (CPTAC), but not at the mRNA level (TCGA data), indicating that the signal is stronger at the protein level (S6B Fig). Additionally, the prognostic value of these biomarkers was evaluated in eight independent transcriptomics datasets by using a log-rank test (S6C Fig). Although the biomarkers were not significantly associated with outcome in all the independent datasets, the signal was stronger when combining all the cohorts.

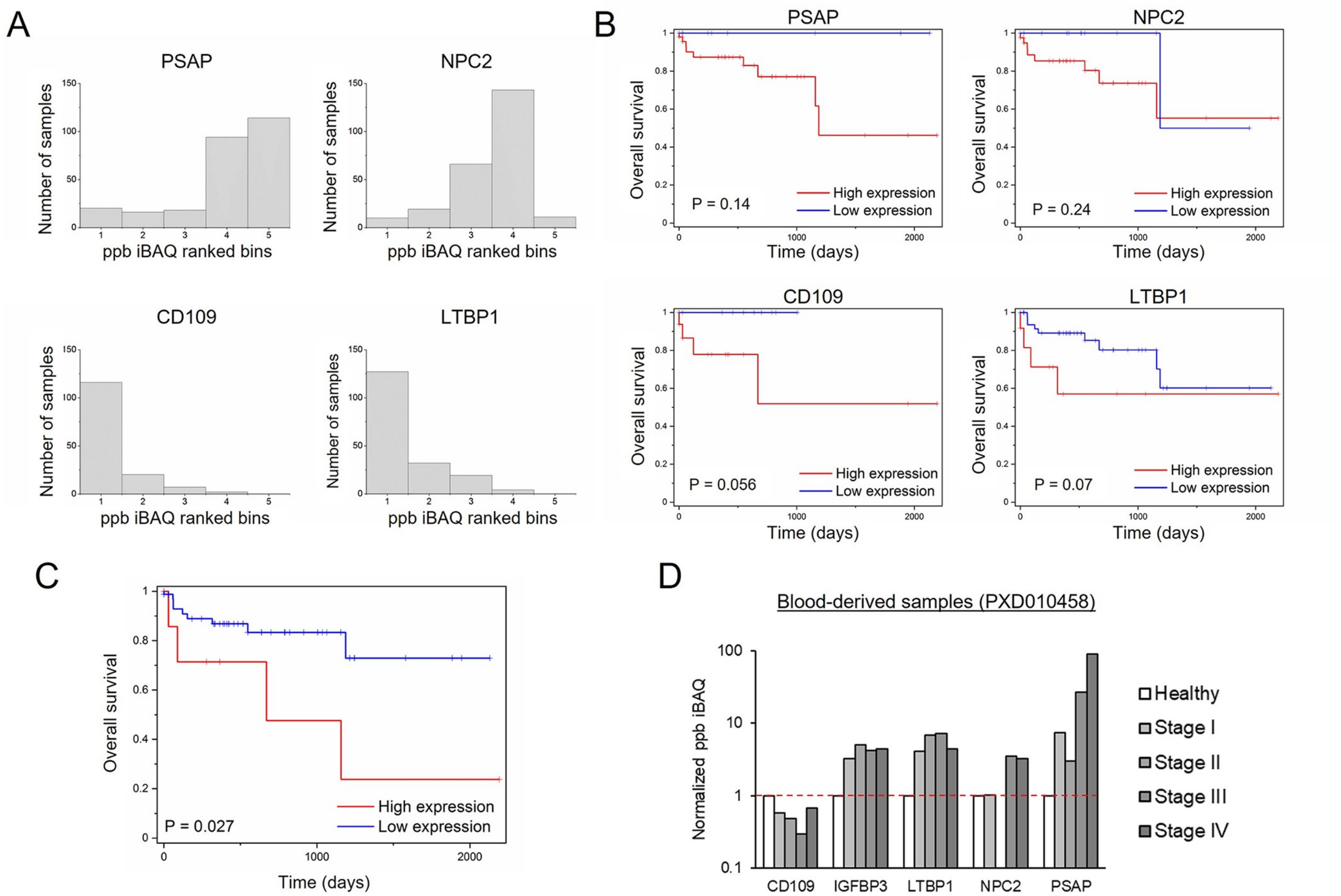

Fig 6. Validation at the proteomic level of the experimentally-based signature SEC6. (A) Histogram distribution of the expression of the SEC6 proteins detected in more than 50% of the tumor samples using CPTAC dataset (iBAQ ranked bins are used). (B) Kaplan–Meier analysis of high- and low-expression patients in stage II and III patients. P values were obtained by log-rank test. (C) Kaplan–Meier analysis of high- and low-expression patients. Mean protein expression of CD109, LTPB1, NPC2 and PSAP was used for classification. P values were obtained by log-rank test. (D) SEC6 proteins distribution across blood-derived samples.

https://doi.org/10.1371/journal.pcbi.1011828.g006

To further explore the clinical value and potential applications to patient stratification of these biomarkers, we explored the expression in the different CRC subtypes, according to proteomic (Fig 7C) and transcriptomic (Fig 7D) classifications. There was a significant statistical association of the 5 biomarkers with the proteomic classification [31], except for PRDX1. CD14 and MRC2 were more abundantly expressed in the subtype C associated with poor outcome, whereas TXNDC5 showed higher expression in subtype D (Fig 7C). In contrast, the significance of the association with the transcriptomic-derived consensus molecular subtypes (CMS) was lower, except for CD14, which showed higher expression in the CMS4 and CMS1 subtypes (associated with poor prognosis), and TXNDC5 that showed more expression in the good prognosis CMS3 subtype. Therefore, proteomic markers correlated better with proteomic-defined subtypes than with transcriptomics-identified ones. In any case, CD14 and TXNDC5 showed strong association with poor and good prognosis, respectively, in a variety of independent datasets, endorsing their value as prognostic biomarkers.

Regarding location, these biomarkers showed higher expression in serum from CRC patients when compared to healthy controls (Fig 7E). PRDX1, CD14 and PPIA were also detected and upregulated in interstitial fluid samples (Fig 7F). Finally, PPIA and PRDX1 were

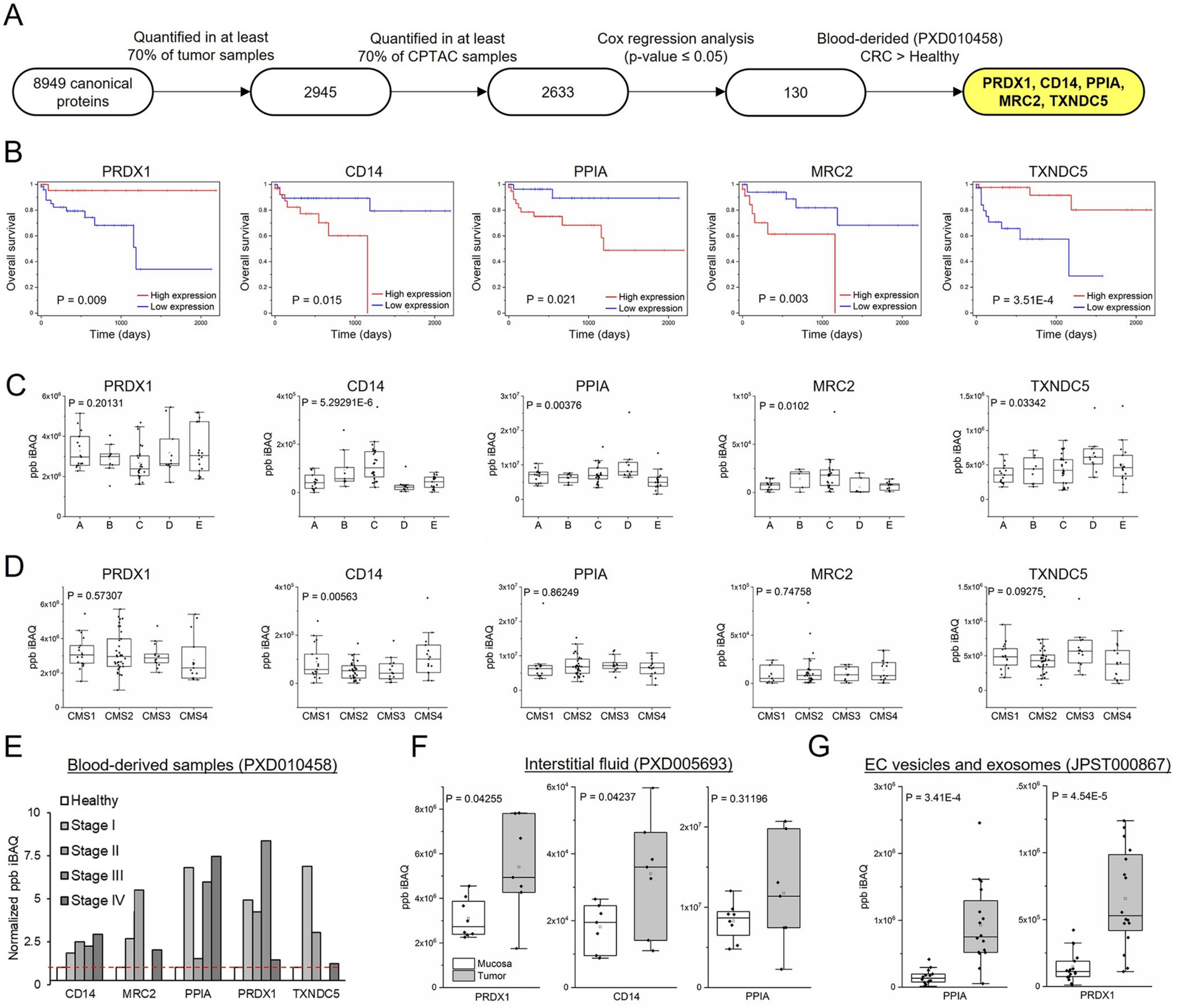

Fig 7. Identification of blood-detectable prognostic biomarkers. (A) Flow-chart representation of sequential prognostic biomarkers selection. (B) Kaplan–Meier analysis of high- and low-expression patients. P values were obtained by log-rank test. ((C) Distribution of the protein expression (ppb) according to the proteomic subtypes' classification [59]. (D) Distribution of the protein expression (ppb) according to the CMS classifier. (E) CD14, MRC2, PPIA, PRDX1, TXNDC5 distribution across blood-derived samples. (F) PRDX1, CD14 and PPIA distribution across interstitial fluid from tumor and mucosa according to the PXD005693 dataset. (G), PPIA and PRDX1distribution across EC vesicles from tumor and adjacent tissue according to the JPST000867 dataset.

https://doi.org/10.1371/journal.pcbi.1011828.g007

both detected and overexpressed in tumor-derived extracellular vesicles (Fig 7G). In conclusion, our meta-analysis supports a novel approach for the identification of novel protein biomarkers suitable for detection in liquid biopsies. Moreover, our data indicate that, for some biomarkers, proteomics detection may outperform transcriptomics analysis.

Tissue expression and functional analysis of the biomarkers panel

To further explore the biological significance of these candidates in CRC tumor samples, immunohistochemistry images were retrieved from the Human Protein Atlas (HPA) [32] (S7A Fig). While PPIA, PRDX1, and TXNDC5 showed abundant expression in most samples,

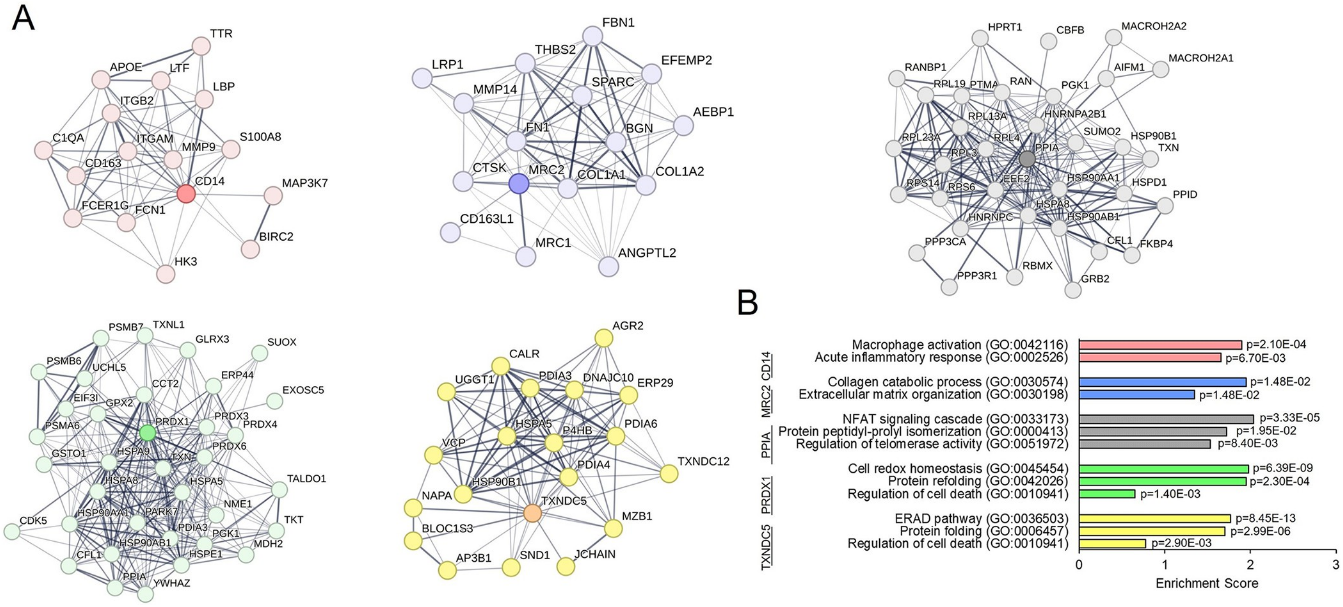

Fig 8. Functional analysis of the identified biomarkers. (A) STRING protein-protein interaction network of the five biomarkers. Proteins with at least medium confidence interaction (score>0.4) and a significant correlation (p<0.01 according to a Pearson correlation) with the corresponding biomarker were selected. Protein expression (ppb) resulting from the solid samples meta-analysis were used for calculating the correlations. (B) Functional enrichments (Biological Process) of the networks according to STRING.

https://doi.org/10.1371/journal.pcbi.1011828.g008

the detection of MRC2 and CD14 was lower (S7B Fig), likely because they are mainly stromal proteins expressed by fibroblasts and macrophages, respectively. In any case, a significant prognostic value of PPIA (p = 0.076), PRDX1 (p = 0.021), and TXNDC5 (p = 0.033) was confirmed using the HPA dataset.

Next, we carried out a functional analysis in order to investigate the potential mechanistic basis of these biomarkers. To identify the proteins associated with the discovered biomarkers, we firstly selected those proteins that presented a high Pearson correlation (p<0.01) with each biomarker when analysing the solid tumor samples. Then, proteins with at least one interaction with the biomarkers were analysed using STRING [33] to obtain the interactome of the biomarkers (Fig 8A). An interactome analysis based on GO revealed major associations with diverse biological functions (Fig 8B). CD14 was mainly associated to inflammatory processes as "macrophage inflammation" and "acute inflammatory response". MRC2 was associated with collagen processing and "extracellular matrix organization". PPIA showed interactions with proteins form the protein peptidyl-prolyl isomerization, and was involved in functions such as NFAT signalling or telomerase activity. Finally, the markers associated with good prognosis, PRDX1 and TXNDC5, were related with similar pathways, "protein refolding" and "regulation of cell death", suggesting that, under cancer-related stress conditions, such as redox unbalance or abnormal protein folding, these proteins would suppress the cell cycle or participate in the cell death.

Discussion

Our study provides a direct demonstration of the value of reusing public proteomics datasets for biomarker identification and validation. In this report, we reanalysed and integrated twelve public proteomic datasets from CRC samples (containing not only solid tumors but also liquid biopsy samples), to confirm the prognostic potential of a gene expression-based signature

(SEC6) at the proteomic level [30], but also for the discovery of new candidate biomarkers capable of predicting patient outcome (CD14, MRC2, PPIA, PRDX1, TXNDC5). To the best of our knowledge, this is the first time that public proteomics datasets have been reused and focused in the detection and validation of biomarkers.

Large-scale genomics and transcriptomics studies have been conducted for biomarker discovery. Although proteomics meta-analysis studies are not yet as common as genomics and transcriptomics studies, they are becoming increasingly popular. However, manual curation is an essential step to provide the optimal level of metadata annotation, and to potentially identify biomarkers from data mining alone [34] A few proteomics meta-analysis studies focusing on human malignancies, such as cancer [22] or Alzheimer's disease [35], have already been published. Proteomics meta-analysis studies have some limitations, mainly related with the heterogeneity associated to proteomics data workflows and the resulting datasets. To achieve a better data integration and comparability, our study was restricted to label-free DDA (Data Dependent Acquisition) quantification studies. However, as original data were acquired in different experimental conditions, the presence of batch effects was expected. To minimize these batch effects, we used a rank-binned normalization [24,25]. Despite the mentioned limitations, meta-analysis studies have notable advantages. The main benefit is the possibility of reusing large amounts of data that, otherwise, would require unnecessary, time-consuming and expensive resources to be acquired again by MS. In addition, integrating multiple studies enables to examine a diverse range of biological conditions, which is difficult to achieve experimentally and can expand the number of proteins of interest that are detected. Additionally, all data in our study were analysed and processed using the same analysis pipeline and search database, which minimizes downstream variability and enables comparison and integration of datasets from different origin. Finally, in our view, compliance with open data sharing practices is essential [36,37]. Therefore, the protein abundance results have been made available in the Expression Atlas resource and can be accessed and visualised there (Table 1).

The identification of novel potential biomarkers in CRC is essential for personalized medicine, enabling the best therapeutic option for each patient. An appropriate categorization of stage II and III patients, considering the diversity within CRC, can allow consistent monitoring and chemotherapy treatment regardless of surgical approaches. With this objective in mind, a gene signature called SEC6 [30] was reported for the calculation of a risk score based on the RNA levels of six genes, predicting the patient outcome and their response to chemotherapy. Although the survival metadata associated with proteomic datasets is much smaller than those associated with genomics and transcriptomics data, we confirmed the prognostic value for four out of six of the proteins. Two of them, BMP1 and IGFBP3, were not detected by proteomics in most of the solid tumor samples. However, IGFBP3 was found as the most abundant protein in blood-derived samples. The fact that five out of these six proteins can be detected in blood greatly increases the potential of this signature for patient risk and response monitoring.

Our proteomic "re-use" strategy also enabled the identification of novel prognostic biomarkers in CRC using proteomics data alone. The CPTAC dataset was essential in this process, as it is the only one that contains survival data. This dataset has been frequently used for confirmation of previously identified prognostic biomarkers, usually through genomics or transcriptomics approaches [38,39]. In our study, we performed a large-scale analysis identifying all the proteins associated with prognosis according to CPTAC. In contrast to common practice, we validated our proteomics discoveries using RNA-seq data to confirm their prognostic potential. We identified five potential prognostic biomarkers (CD14, PPIA, MRC2, PRDX1, and TXNDC5). CD14 is a macrophage-associated marker that has been associated with unfavourable prognosis in CRC [40]. Peptidylprolyl isomerase A (PPIA), a.k.a. as cyclophilin A, has been related with ERK1/2 phosphorylation and NF-κB activation [41], gastrin cancer serum biomarker [42] and poor prognosis in hepatocellular carcinoma [43]. MRC2 (Mannose Receptor C Type 2) is a mannose receptor whose expression is upregulated and associated with prognosis in some types of cancer such as glioblastoma, bladder, ovarian, and renal cancer [44]. GO analysis of these three unfavourable markers showed association with inflammatory pathways and extracellular matrix reorganization, which are processes linked to invasion and progression. On the other hand, for PRDX1 and TXNDC5, identified as favourable prognostic biomarkers, the GO analysis showed association with cell death regulation. PRDX1 is a peroxiredoxin, an enzyme involved in regulating reactive oxygen species. Although its mechanism of action is uncertain, it has been shown to prevent metastasis and angiogenesis [45]. PRDX1 depletion promoted the expression of pro-inflammatory cytokines in CRC [46]. Finally, TXNDC5 is a disulphide isomerase (PDI) that catalyses protein folding and thiol-disulphide interchange reactions. TXNDC5 promotes survival and proliferation by inducing HIF-1α in hypoxic situations [47]. Although upregulated in different tumours, its role in cancer progression remains unclear [48]

In summary, the CRC protein abundance/expression landscape resulting from the reanalysis of twelve public datasets constitutes a rich source of information for biomarker discovery and validation. To the best of our knowledge, this is the first time that a meta-analysis of public proteomics datasets has been used as the basis for biomarker discovery and validation. Additionally, we have demonstrated its value to perform, in a relatively short period of time, the validation of previously described biomarkers and the identification of new biomarker panels with potential clinical utility.

Methods

Dataset selection

MS-based proteomics data from studies of human colorectal cancer were selected for reanalysis from public repositories included in ProteomeXchange such as PRIDE and jPOST, and from the CPTAC data portal. These databases were queried for human CRC and the resulting hits were filtered based on the following criteria- i) label-free DDA studies, where no post-translational modification (PTM)-enrichment had been performed; ii) experiments performed on Thermo Fisher Scientific instruments (LTQ Orbitrap, LTQ Orbitrap Elite, LTQ Orbitrap Velos, LTQ Orbitrap XL ETD, LTQ-Orbitrap XL ETD, Orbitrap Fusion and Q-Exactive); and iii) availability of detailed sample metadata in the original publication, or after contacting the original submitters. As a result, 10 datasets from PRIDE, one dataset each from jPOST and one from the CPTAC data portal were downloaded. The details of these datasets are available in Table 1. It is important to highlight that, although a small number of additional public datasets generated using other proteomics approaches were available, the 12 chosen datasets represented the vast majority of the relevant CRC public proteomics datasets. All datasets were manually curated and the corresponding information was encoded in a SDRF (Sample Data Relationship File), linking the MS raw data to the biological conditions.

Proteomics raw data processing

Proteomics datasets of secretome and tumor samples were analysed in two batches separately. Peptide/protein identification and protein quantification was performed using MaxQuant [27] (version 2.1.0.0) on a high-performance Linux computing cluster. The input parameters for each dataset such as MS1 and MS2 tolerances, digestive enzymes, fixed and variable modifications were set as described in their respective publications together with two missed cleavage sites. PSM (Peptide Spectrum Match) and protein FDR (False Discovery Rate) levels were set at 1%. Other MaxQuant parameter settings were left as default: maximum number of modifications per peptide: 5, minimum peptide length: 7, maximum peptide mass: 4,600 Da. For match between runs, the minimum match time window was set to 0.7 seconds and the minimum retention time alignment window was set to 20 seconds. The UniProt Human Reference Proteome (one protein sequence per gene set (*Homo sapiens*, UniProt, Sept. 2020. 20,601 sequences) was used as the target sequence database. The inbuilt MaxQuant contaminant database was also used, and the decoy database were generated by MaxQuant at the time of the analysis (on-the-fly) by reversing the input database sequences after the respective enzymatic cleavage.

Post-processing

MaxQuant results for each batch were processed downstream to remove potential contaminants, decoys and protein groups which had fewer than 2 PSMs. The protein intensities were normalised using the FOT method as mentioned [25], wherein each protein iBAQ intensity value is scaled to the total amount of signal in a given MS run and transformed to parts per billion (ppb).

$$ppb\_iBAQ_i = \left({}^{i}BAQ_i / \sum i = 1^{n} iBAQ_i\right) x\, 1,000,000,000$$

The bioconductor package ‘mygene’ [49] was used to assign Ensembl gene identifiers/annotations to the protein groups by mapping the ‘majority protein identifiers’ within each protein group. This step is required for integration into Expression Atlas. Briefly, from the MaxQuant output file ‘proteinGroups.txt’, the UniProt protein accessions within each protein group in the ‘majority protein identifiers’ columns were individually queried using the ‘queryMany’ function in the ‘mygene’ package to obtain their respective Ensembl gene symbols and gene identifiers. The protein groups, whose protein identifiers were mapped to multiple Ensembl gene symbols/IDs, were not used for further downstream analysis. In those cases, where two or more protein groups mapped to the same Ensembl gene symbol/ID, their median intensity values were considered. The parent genes to which the different protein groups were mapped to are equivalent to ‘canonical proteins’ in UniProt (https://www.uniprot.org/help/canonical_and_isoforms) and therefore the term protein abundance is used to describe the protein abundance of the canonical protein throughout the manuscript. A detailed flowchart of all post-processing steps is shown in S1 Fig.

Protein abundance comparison across datasets

For comparison of protein abundances between the two groups of samples (secretome and solid tumor samples), the normalised iBAQ abundances were transformed into numerical bins. The abundances (in ppb) were ranked and grouped into 5 bins, wherein proteins with the lowest protein abundance values were in bin 1 and those with the highest abundance values were in bin 5. A Pearson correlation coefficient for all samples was calculated on pairwise complete observations of bin transformed iBAQ values in R. Samples were hierarchically clustered on columns and rows using Euclidean distances.

Differentially expressed proteins and Gene Ontology (GO) analysis

Differentially expressed canonical proteins between (tumor and mucosa, adenoma and mucosa or adenoma and tumor) samples were determined by performing a t-test after ranked bin transformation. Benjamini-Hochberg procedure was used to control the FDR. To investigate the source-specific protein profile of each subgroup of the secreted samples group, we identified and further analysed proteins categorized as "enriched proteins". "Enriched proteins" were those proteins uniquely quantified in one of the subgroups or proteins with median bin values higher than each median bin value of the rest of the subgroups. Gene ontology (GO) analysis was performed using g:Profiler [50] and Enrichr [51]. Only differentially expressed proteins or "Enriched proteins" were selected for the analysis.

In silico validation and functional enrichment analysis

Images and quantifications derived from immunohistochemistry assays were obtained from the HPA [52]. When several antibodies were available for a given protein, the most representative was selected for further analysis. Semi-quantitative analysis (high, medium, low and not detected) performed and available in HPA was used in this study.

The protein interactome was obtained using the STRING database. To restrict the interactome to the CRC context, only proteins with a strong Pearson correlation ($p<0.01$) with the protein of interest in the solid tumor samples were selected. Next, proteins with at least one medium confidence interaction (STRING score>0.4) were selected for the final interactome. The interactome was plotted and a GO functional enrichment analysis was performed.

Transcriptomics data selection and processing

RNA-seq data generated on the Illumina HiSeq platform for 592 colorectal tumor samples from TCGA COADREAD dataset [53] was downloaded from cBioPortal [54]. Data was mapped with the RSEM (RNA-Seq by Expectation-Maximization) algorithm and normalized using the [$\log_2$(RSEM+1)] method as previously described [55]. RNA-seq data of samples from CPTAC (90 patients) were included in the mentioned TCGA COADREAD dataset. However, both proteomics and RNA-seq data were acquired only for CPTAC samples.

For primary tumor and normal mucosa comparison, GSE41258 [56], a large dataset containing 390 samples from 276 CRC patients, was selected from Gene Expression Omnibus (GEO). This dataset was selected because contains data from 190 primary tumor and 54 healthy mucosa samples obtained from individuals, including a high level of metadata annotation (age, sex, stage, recurrence, location, etc). Data was derived from 299 U133A arrays and processed and normalized according to the original publication [56].

Proteomics and transcriptomics data comparison

For the comparison between healthy mucosa and tumor samples, proteomics and transcriptomics data were obtained from the solid samples batch and from the GSE41258 dataset, respectively. In both cases, a fold-change (tumor *vs* mucosa) and an associated p-value were obtained. For the protein expression, the data was converted into ranked bins as previously described. The p-value was obtained by a t-test and subsequent Benjamini-Hochberg correction. The fold change was obtained as the ratio of the means of the tumor and mucosa subgroups. Regarding transcriptomics, the fold change was obtained as the ratio of the normalized expression of the primary tumor and healthy mucosa subgroups. The p-value was also obtained by a t-test and subsequent Benjamini-Hochberg correction. The correlation between both fold changes was examined using scatter and volcano plots. Only differentially expressed canonical proteins were analysed. For the comparison of hazard ratios (HR) obtained by proteomics and transcriptomics, proteomics data from the CPTAC dataset and RNA-seq data from the TCGA and CPTAC datasets were used. The correlation between HR was examined using scatter plots. Only significant proteins using the Cox regression analysis were plotted.

Statistical analysis

The significance of protein expression differences between groups was obtained by using two-sample t tests for each protein or gene. ANOVA tests were performed in order to detect significant differences in the risk-score between three or more groups. Block design was used to correct the variability corresponding to the batch effects. Univariate Cox regression and Kaplan-Meier analysis were performed using the 'survival' and 'survminer' R packages (https://CRAN.R-project.org/package=survival). Intensity values were transformed to a z-score before survival analysis. For Kaplan-Meier analysis, patients were divided into two subgroups (high or low expression) by the optimal cut-off method, using Cutoff Finder [57]. The point at which the log-rank test split was most significant was identified as the optimal cut-off.

Integration into Expression Atlas

In Expression Atlas, protein expression data coming from the individual reanalysis of each dataset is available. The calculated canonical protein abundances (mapped as genes), and summary files detailing the quality of post-processing of all datasets were integrated as proteomics baseline experiments (E-PROT identifiers are available in Table 1). It should be noted that Expression Atlas mainly provides a dataset centric view, so this is why the protein abundance data were integrated in that way. For the overall results described in the manuscript, the datasets were analysed in two batches (as explained above) to provide an improved control of the FDR at different levels (PSM and protein). However, it is important to highlight that the conclusions reached were the same with regards to the described identified biomarkers. They were also significantly associated to survival (in the CPTAC dataset) and overexpressed in blood coming from CRC patients (dataset PXD010458) when using the protein abundance results coming from the individual reanalyses of datasets (as integrated in Expression Atlas) instead of the combined reanalysis.

Supporting information

S1 Table. List of the proteins corresponding to the SEC6 signature.
(DOCX)

S2 Table. List of the proteins identified as new potential biomarkers.
(DOCX)

S1 Fig. Detailed flow chart of the meta-analysis study. (SDRF: Sample Data Relationship Format, IDF: Investigation Description Format).
(TIF)

S2 Fig. Heatmap of binned protein abundances in colorectal tumor samples. Non-hierarchical clustering of the solid samples according to Pearson correlation. Dataset and sample subgroup are indicated.
(TIF)

S3 Fig. Heat-map of binned protein abundances in colorectal cancer secreted samples. Non-hierarchical clustering of the secreted samples according to Pearson correlation. Dataset, conditions and sample subgroup are indicated.
(TIF)

S4 Fig. Enriched proteins in secreted samples. (A) Heat-map representing the Pearson coefficient between fold changes of solid and secreted samples. (B) Enriched canonical proteins in each subgroup. Proteins are considered to be enriched when quantified only in one subgroup

or expression is at least double than the rest of subgroups. (C) Gene ontology (Biological Process) analysis of the enriched proteins indicating more relevant categories. (D) EC vesicle category from Cellular Component GO analysis in the enriched proteins.
(TIF)

S5 Fig. Extended analysis of prognosis association. (A) Kaplan–Meier analysis of high- and low-expression patients in stage II and III. P values were obtained by log-rank test. (B) Classification of patients in high or low risk according to the expression of 5 different biomarkers. (C) Kaplan–Meier analysis of high- and low risk patients in all the stages (left) and stages II and III (right). Patients are considered as high risk when at least 3 biomarkers are classifying them as high risk.
(TIFF)

S6 Fig. Distribution of the selected biomarkers in transcriptomics. (A) Kaplan–Meier analysis of high- and low-expression patients using the complete TCGA COADREAD (n = 431). P values were obtained by log-rank test. (B) Forest plots of HRs associated to each gene (TCGA) or protein (CPTAC) in each dataset. P values were obtained by Cox regression analysis. (C) Log-rank analysis of the biomarkers according to eight different transcriptomics datasets.
(TIF)

S7 Fig. Immunohistochemistry analysis according to Human Protein Atlas (HPA). (A) Representative PPIA, PRDX1, TXNDC5, MRC2 and CD14 protein expression by IHC in CRC cases according to HPA series of cases. (B) Percent of CRC cases expressing high, medium, low or null levels of protein according to HPA. Most representative antibodies were selected for the analysis.
(TIF)

Acknowledgments

We would like to thank all data submitters who generated the original data that has been reused in this study. We would also like to thank Nancy George (EMBL-EBI) for her help in integrating the data in Expression Atlas.

Author Contributions

Conceptualization: Javier Robles, Ananth Prakash, Juan Antonio Vizcaíno, J. Ignacio Casal.

Data curation: Ananth Prakash.

Formal analysis: Javier Robles, Ananth Prakash.

Funding acquisition: J. Ignacio Casal.

Investigation: Ananth Prakash, Juan Antonio Vizcaíno.

Methodology: Javier Robles.

Project administration: J. Ignacio Casal.

Resources: J. Ignacio Casal.

Supervision: Ananth Prakash, Juan Antonio Vizcaíno.

Writing – original draft: Javier Robles, Juan Antonio Vizcaíno, J. Ignacio Casal.

Writing – review & editing: Juan Antonio Vizcaíno, J. Ignacio Casal.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018; 68(6):394–424. https://doi.org/10.3322/caac.21492 PMID: 30207593.
2. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. Gut. 2017; 66(4):683–91. https://doi.org/10.1136/gutjnl-2015-310912 PMID: 26818619.
3. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021; 71(3):209–49. https://doi.org/10.3322/caac.21660 PMID: 33538338.
4. Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ, Wullschleger S, et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nat Med. 2013; 19(5):619–25. https://doi.org/10.1038/nm.3175 PMID: 23584089.
5. Molinari C, Marisi G, Passardi A, Matteucci L, De Maio G, Ulivi P. Heterogeneity in Colorectal Cancer: A Challenge for Personalized Medicine? Int J Mol Sci. 2018; 19(12). https://doi.org/10.3390/ijms19123733 PMID: 30477151.
6. Fotheringham S, Mozolowski GA, Murray EMA, Kerr DJ. Challenges and solutions in patient treatment strategies for stage II colon cancer. Gastroenterol Rep (Oxf). 2019; 7(3):151–61. https://doi.org/10.1093/gastro/goz006 PMID: 31217978.
7. Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015; 21(11):1350–6. https://doi.org/10.1038/nm.3967 PMID: 26457759.
8. Jodal HC, Helsingen LM, Anderson JC, Lytvyn L, Vandvik PO, Emilsson L. Colorectal cancer screening with faecal testing, sigmoidoscopy or colonoscopy: a systematic review and network meta-analysis. BMJ Open. 2019; 9(10):e032773. https://doi.org/10.1136/bmjopen-2019-032773 PMID: 31578199.
9. Sveen A, Nesbakken A, Agesen TH, Guren MG, Tveit KM, Skotheim RI, et al. Anticipating the clinical use of prognostic gene expression-based tests for colon cancer stage II and III: is Godot finally arriving? Clin Cancer Res. 2013; 19(24):6669–77. https://doi.org/10.1158/1078-0432.CCR-13-1769 PMID: 24166914.
10. Isella C, Brundu F, Bellomo SE, Galimi F, Zanella E, Porporato R, et al. Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. Nat Commun. 2017; 8:15107. https://doi.org/10.1038/ncomms15107 PMID: 28561063.
11. Mazouji O, Ouhajjou A, Incitti R, Mansour H. Updates on Clinical Use of Liquid Biopsy in Colorectal Cancer Screening, Diagnosis, Follow-Up, and Treatment Guidance. Front Cell Dev Biol. 2021; 9:660924. https://doi.org/10.3389/fcell.2021.660924 PMID: 34150757.
12. Paltridge JL, Belle L, Khew-Goodall Y. The secretome in cancer progression. Biochim Biophys Acta. 2013; 1834(11):2233–41. https://doi.org/10.1016/j.bbapap.2013.03.014 PMID: 23542208.
13. Lou S, Shaukat A. Noninvasive strategies for colorectal cancer screening: opportunities and limitations. Curr Opin Gastroenterol. 2021; 37(1):44–51. https://doi.org/10.1097/MOG.0000000000000688 PMID: 33074994.
14. Banias L, Jung I, Bara T, Fulop Z, Simu P, Simu I, et al. Immunohistochemical-based molecular subtyping of colorectal carcinoma using maspin and markers of epithelial-mesenchymal transition. Oncol Lett. 2020; 19(2):1487–95. https://doi.org/10.3892/ol.2019.11228 PMID: 31966075.
15. Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. Nature. 2016; 537(7620):347–55. https://doi.org/10.1038/nature19949 PMID: 27629641.
16. Edwards NJ, Oberti M, Thangudu RR, Cai S, McGarvey PB, Jacob S, et al. The CPTAC Data Portal: A Resource for Cancer Proteomics Research. J Proteome Res. 2015; 14(6):2707–13. https://doi.org/10.1021/pr501254j PMID: 25873244.
17. Perez-Riverol Y, Bai J, Bandla C, Garcia-Seisdedos D, Hewapathirana S, Kamatchinathan S, et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res. 2022; 50(D1):D543–D52. https://doi.org/10.1093/nar/gkab1038 PMID: 34723319.
18. Moriya Y, Kawano S, Okuda S, Watanabe Y, Matsumoto M, Takami T, et al. The jPOST environment: an integrated proteomics data repository and database. Nucleic Acids Res. 2019; 47(D1):D1218–D24. https://doi.org/10.1093/nar/gky899 PMID: 30295851.
19. Lukk M, Kapushesky M, Nikkila J, Parkinson H, Goncalves A, Huber W, et al. A global map of human gene expression. Nat Biotechnol. 2010; 28(4):322–4. https://doi.org/10.1038/nbt0410-322 PMID: 20379172.
20. Rung J, Brazma A. Reuse of public genome-wide gene expression data. Nat Rev Genet. 2013; 14(2):89–99. https://doi.org/10.1038/nrg3394 PMID: 23269463.

21. Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014; 509(7502):582–7. https://doi.org/10.1038/nature13319 PMID: 24870543.

22. Jarnuczak AF, Najgebauer H, Barzine M, Kundu DJ, Ghavidel F, Perez-Riverol Y, et al. An integrated landscape of protein expression in human cancer. Sci Data. 2021; 8(1):115. https://doi.org/10.1038/s41597-021-00890-2 PMID: 33893311.

23. Claeys T, Menu M, Bouwmeester R, Gevaert K, Martens L. Machine Learning on Large-Scale Proteomics Data Identifies Tissue and Cell-Type Specific Proteins. J Proteome Res. 2023; 22(4):1181–92. https://doi.org/10.1021/acs.jproteome.2c00644 PMID: 36963412.

24. Wang S, Garcia-Seisdedos D, Prakash A, Kundu DJ, Collins A, George N, et al. Integrated view and comparative analysis of baseline protein expression in mouse and rat tissues. PLoS Comput Biol. 2022; 18(6):e1010174. https://doi.org/10.1371/journal.pcbi.1010174 PMID: 35714157.

25. Prakash A, Garcia-Seisdedos D, Wang S, Kundu DJ, Collins A, George N, et al. Integrated View of Baseline Protein Expression in Human Tissues. J Proteome Res. 2023; 22(3):729–42. https://doi.org/10.1021/acs.jproteome.2c00406 PMID: 36577097.

26. Papatheodorou I, Moreno P, Manning J, Fuentes AM, George N, Fexova S, et al. Expression Atlas update: from tissues to single cells. Nucleic Acids Res. 2020; 48(D1):D77–D83. https://doi.org/10.1093/nar/gkz947 PMID: 31665515.

27. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008; 26(12):1367–72. https://doi.org/10.1038/nbt.1511 PMID: 19029910.

28. Moreno P, Fexova S, George N, Manning JR, Miao Z, Mohammed S, et al. Expression Atlas update: gene and protein expression in multiple species. Nucleic Acids Res. 2022; 50(D1):D129–D40. https://doi.org/10.1093/nar/gkab1030 PMID: 34850121.

29. Vander Heiden MG, Cantley LC, Thompson CB. Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science. 2009; 324(5930):1029–33. https://doi.org/10.1126/science.1160809 PMID: 19460998.

30. Robles J, Pintado-Berninches L, Boukich I, Escudero B, de Los Rios V, Bartolome RA, et al. A prognostic six-gene expression risk-score derived from proteomic profiling of the metastatic colorectal cancer secretome. J Pathol Clin Res. 2022; 8(6):495–508. https://doi.org/10.1002/cjp2.294 PMID: 36134447.

31. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, et al. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014; 513(7518):382–7. https://doi.org/10.1038/nature13438 PMID: 25043054.

32. Uhlen M, Zhang C, Lee S, Sjostedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science. 2017; 357(6352). https://doi.org/10.1126/science.aan2507 PMID: 28818916.

33. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023; 51(D1):D638–D46. https://doi.org/10.1093/nar/gkac1000 PMID: 36370105.

34. Griss J, Perez-Riverol Y, Hermjakob H, Vizcaino JA. Identifying novel biomarkers through data mining-a realistic scenario? Proteomics Clin Appl. 2015; 9(3–4):437–43. https://doi.org/10.1002/prca.201400107 PMID: 25347964.

35. Bai B, Vanderwall D, Li Y, Wang X, Poudel S, Wang H, et al. Proteomic landscape of Alzheimer's Disease: novel insights into pathogenesis and biomarker discovery. Mol Neurodegener. 2021; 16(1):55. https://doi.org/10.1186/s13024-021-00474-z PMID: 34384464.

36. Data sharing is the future. Nat Methods. 2023; 20(4):471. https://doi.org/10.1038/s41592-023-01865-4 PMID: 37046014.

37. Blasimme A, Fadda M, Schneider M, Vayena E. Data Sharing For Precision Medicine: Policy Lessons And Future Directions. Health Aff (Millwood). 2018; 37(5):702–9. https://doi.org/10.1377/hlthaff.2017.1558 PMID: 29733719.

38. Zhang R, Hu M, Chen HN, Wang X, Xia Z, Liu Y, et al. Phenotypic Heterogeneity Analysis of APC-Mutant Colon Cancer by Proteomics and Phosphoproteomics Identifies RAI14 as a Key Prognostic Determinant in East Asians and Westerners. Mol Cell Proteomics. 2023; 22(5):100532. https://doi.org/10.1016/j.mcpro.2023.100532 PMID: 36934880.

39. Ogawa M, Tanaka A, Namba K, Shia J, Wang JY, Roehrl MHA. Tumor stromal nicotinamide N-methyltransferase overexpression as a prognostic biomarker for poor clinical outcome in early-stage colorectal cancer. Sci Rep. 2022; 12(1):2767. https://doi.org/10.1038/s41598-022-06772-w PMID: 35177765.

40. Chen D, Wang H. The clinical and immune features of CD14 in colorectal cancer identified via large-scale analysis. Int Immunopharmacol. 2020; 88:106966. https://doi.org/10.1016/j.intimp.2020.106966 PMID: 33182067.

41. Bahmed K, Henry C, Holliday M, Redzic J, Ciobanu M, Zhang F, et al. Extracellular cyclophilin-A stimulates ERK1/2 phosphorylation in a cell-dependent manner but broadly stimulates nuclear factor kappa B. Cancer Cell Int. 2012; 12(1):19. https://doi.org/10.1186/1475-2867-12-19 PMID: 22631225.

42. Shen Q, Polom K, Williams C, de Oliveira FMS, Guergova-Kuras M, Lisacek F, et al. A targeted proteomics approach reveals a serum protein signature as diagnostic biomarker for resectable gastric cancer. EBioMedicine. 2019; 44:322–33. https://doi.org/10.1016/j.ebiom.2019.05.044 PMID: 31151932.

43. Wang S, Li M, Xing L, Yu J. High expression level of peptidylprolyl isomerase A is correlated with poor prognosis of liver hepatocellular carcinoma. Oncol Lett. 2019; 18(5):4691–702. https://doi.org/10.3892/ol.2019.10846 PMID: 31611978.

44. Zhao Z, Yang Y, Liu Z, Chen H, Guan X, Jiang Z, et al. Prognostic and immunotherapeutic significance of mannose receptor C type II in 33 cancers: An integrated analysis. Front Mol Biosci. 2022; 9:951636. https://doi.org/10.3389/fmolb.2022.951636 PMID: 36188226.

45. Li HX, Sun XY, Yang SM, Wang Q, Wang ZY. Peroxiredoxin 1 promoted tumor metastasis and angiogenesis in colorectal cancer. Pathol Res Pract. 2018; 214(5):655–60. https://doi.org/10.1016/j.prp.2018.03.026 PMID: 29673884.

46. Chu G, Li J, Zhao Y, Liu N, Zhu X, Liu Q, et al. Identification and verification of PRDX1 as an inflammation marker for colorectal cancer progression. Am J Transl Res. 2016; 8(2):842–59. PMID: 27158373.

47. Tan F, Zhu H, He X, Yu N, Zhang X, Xu H, et al. Role of TXNDC5 in tumorigenesis of colorectal cancer cells: In vivo and in vitro evidence. Int J Mol Med. 2018; 42(2):935–45. https://doi.org/10.3892/ijmm.2018.3664 PMID: 29749460.

48. Horna-Terron E, Pradilla-Dieste A, Sanchez-de-Diego C, Osada J. TXNDC5, a newly discovered disulfide isomerase with a key role in cell physiology and pathology. Int J Mol Sci. 2014; 15(12):23501–18. https://doi.org/10.3390/ijms151223501 PMID: 25526565.

49. Mark A TR, Afrasiabi C, Wu C. mygene: Access MyGene.Info services. Version 1.2.3. R/Bioconductor package, 2014. Nature Protocol. 2016; 11(12):2301–19.

50. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019; 47 (W1):W191–W8. https://doi.org/10.1093/nar/gkz369 PMID: 31066453.

51. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013; 14:128. https://doi.org/10.1186/1471-2105-14-128 PMID: 23586463.

52. Uhlen M, Bjorling E, Agaton C, Szigyarto CA, Amini B, Andersen E, et al. A human protein atlas for normal and cancer tissues based on antibody proteomics. Mol Cell Proteomics. 2005; 4(12):1920–32. https://doi.org/10.1074/mcp.M500279-MCP200 PMID: 16127175.

53. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell. 2018; 173 (2):400–16 e11. https://doi.org/10.1016/j.cell.2018.02.052 PMID: 29625055.

54. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012; 2(5):401–4. https://doi.org/10.1158/2159-8290.CD-12-0095 PMID: 22588877.

55. Yao L, Tak YG, Berman BP, Farnham PJ. Functional annotation of colon cancer risk SNPs. Nat Commun. 2014; 5:5114. https://doi.org/10.1038/ncomms6114 PMID: 25268989.

56. Sheffer M, Bacolod MD, Zuk O, Giardina SF, Pincas H, Barany F, et al. Association of survival and disease progression with chromosomal instability: a genomic exploration of colorectal cancer. Proc Natl Acad Sci U S A. 2009; 106(17):7131–6. https://doi.org/10.1073/pnas.0902232106 PMID: 19359472.

57. Budczies J, Klauschen F, Sinn BV, Gyorffy B, Schmitt WD, Darb-Esfahani S, et al. Cutoff Finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization. PLoS One. 2012; 7(12):e51862. https://doi.org/10.1371/journal.pone.0051862 PMID: 23251644.

58. Sethi MK, Thaysen-Andersen M, Kim H, Park CK, Baker MS, Packer NH, et al. Quantitative proteomic analysis of paired colorectal cancer and non-tumorigenic tissues reveals signature proteins and perturbed pathways involved in CRC progression and metastasis. J Proteomics. 2015; 126:54–67. https://doi.org/10.1016/j.jprot.2015.05.037 PMID: 26054784.

59. Wisniewski JR, Dus-Szachniewicz K, Ostasiewicz P, Ziolkowski P, Rakus D, Mann M. Absolute Proteome Analysis of Colorectal Mucosa, Adenoma, and Cancer Reveals Drastic Changes in Fatty Acid Metabolism and Plasma Membrane Transporters. J Proteome Res. 2015; 14(9):4005–18. https://doi.org/10.1021/acs.jproteome.5b00523 PMID: 26245529.

60. Sohier P, Sanson R, Leduc M, Audebourg A, Broussard C, Salnot V, et al. Proteome analysis of formalin-fixed paraffin-embedded colorectal adenomas reveals the heterogeneous nature of traditional serrated adenomas compared to other colorectal adenomas. J Pathol. 2020; 250(3):251–61. https://doi.org/10.1002/path.5366 PMID: 31729028.

61. Tanaka A, Zhou Y, Shia J, Ginty F, Ogawa M, Klimstra DS, et al. Prolyl 4-hydroxylase alpha 1 protein expression risk-stratifies early stage colorectal cancer. Oncotarget. 2020; 11(8):813–24. https://doi.org/10.18632/oncotarget.27491 PMID: 32166002.

62. Costanza B, Turtoi A, Bellahcene A, Hirano T, Peulen O, Blomme A, et al. Innovative methodology for the identification of soluble biomarkers in fresh tissues. Oncotarget. 2018; 9(12):10665–80. https://doi.org/10.18632/oncotarget.24366 PMID: 29535834.

63. Novikova S, Shushkova N, Farafonova T, Tikhonova O, Kamyshinsky R, Zgoda V. Proteomic Approach for Searching for Universal, Tissue-Specific, and Line-Specific Markers of Extracellular Vesicles in Lung and Colorectal Adenocarcinoma Cell Lines. Int J Mol Sci. 2020; 21(18). https://doi.org/10.3390/ijms21186601 PMID: 32916986.

64. Marin-Vicente C, Mendes M, de Los Rios V, Fernandez-Acenero MJ, Casal JI. Identification and Validation of Stage-Associated Serum Biomarkers in Colorectal Cancer Using MS-Based Procedures. Proteomics Clin Appl. 2020; 14(1):e1900052. https://doi.org/10.1002/prca.201900052 PMID: 31502404.

65. Ikeda A, Nagayama S, Sumazaki M, Konishi M, Fujii R, Saichi N, et al. Colorectal Cancer-Derived CAT1-Positive Extracellular Vesicles Alter Nitric Oxide Metabolism in Endothelial Cells and Promote Angiogenesis. Mol Cancer Res. 2021; 19(5):834–46. https://doi.org/10.1158/1541-7786.MCR-20-0827 PMID: 33579815.

66. Strybel U, Marczak L, Zeman M, Polanski K, Mielanczyk L, Klymenko O, et al. Molecular Composition of Serum Exosomes Could Discriminate Rectal Cancer Patients with Different Responses to Neoadjuvant Radiotherapy. Cancers (Basel). 2022; 14(4). https://doi.org/10.3390/cancers14040993 PMID: 35205741.