Journal of Translational Medicine

RESEARCH

Open Access

Gut metatranscriptomics based de novo assembly reveals microbial signatures predicting immunotherapy outcomes in non-small cell lung cancer

David Dora[1], Peter Kiraly[2], Csenge Somodi[3], Balazs Ligeti[4], Edit Dulka[2], Gabriella Galffy[2] and Zoltan Lohinai[3*]

Abstract

Background Advanced-stage non-small cell lung cancer (NSCLC) poses treatment challenges, with immune checkpoint inhibitors (ICIs) as the main therapy. Emerging evidence suggests the gut microbiome significantly influences ICI efficacy. This study explores the link between the gut microbiome and ICI outcomes in NSCLC patients, using metatranscriptomic (MTR) signatures.

Methods We utilized a de novo assembly-based MTR analysis on fecal samples from 29 NSCLC patients undergoing ICI therapy, segmented according to progression-free survival (PFS) into long (> 6 months) and short (≤ 6 months) PFS groups. Through RNA sequencing, we employed the Trinity pipeline for assembly, MMSeqs2 for taxonomic classification, DESeq2 for differential expression (DE) analysis. We constructed Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost) machine learning (ML) algorithms and comprehensive microbial profiles.

Results We detected no significant differences concerning alpha-diversity, but we revealed a biologically relevant separation between the two patient groups in beta-diversity. Actinomycetota was significantly overrepresented in patients with short PFS (vs long PFS, 36.7% vs. 5.4%, $p < 0.001$), as was Euryarchaeota (1.3% vs. 0.002%, $p = 0.009$), while Bacillota showed higher prevalence in the long PFS group (66.2% vs. 42.3%, $p = 0.007$), when comparing the abundance of corresponding RNA reads. Among the 120 significant DEGs identified, cluster analysis clearly separated a large set of genes more active in patients with short PFS and a smaller set of genes more active in long PFS patients. Protein Domain Families (PFAMs) were analyzed to identify pathways enriched in patient groups. Pathways related to DNA synthesis and Translesion were more enriched in short PFS patients, while metabolism-related pathways were more enriched in long PFS patients. *E. coli*-derived PFAMs dominated in patients with long PFS. RF, SVM and XGBoost ML models all confirmed the predictive power of our selected RNA-based microbial signature, with ROC AUCs all greater than 0.84. Multivariate Cox regression tested with clinical confounders PD-L1 expression and chemotherapy history underscored the influence of $n = 6$ key RNA biomarkers on PFS.

Conclusion According to ML models specific gut microbiome MTR signatures' associate with ICI treated NSCLC outcomes. Specific gene clusters and taxa MTR gene expression might differentiate long vs short PFS.

*Correspondence:
Zoltan Lohinai
lohinai.zoltan@semmelweis.hu
Full list of author information is available at the end of the article

Keywords Metatranscriptome, De novo assembly, Gut microbiome, Immunotherapy, Immune-checkpoint inhibitor, Progression-free survival, Machine learning

Background

Lung cancer is the leading cause of cancer-related deaths worldwide (11.6% of the total cases [1]. The frontline treatment regimens frequently administered in advanced-stage non-small cell lung cancer (NSCLC) patients are immune checkpoint inhibitor monotherapy (ICI) or ICI combinations [2]. The 5-year survival for NSCLC immunotherapy-treated patients as a single agent is 30% in a biomarker-selected (high PD-L1 expression), and 10–20% in biomarker- unselected cohorts [3].

The intestinal microbiome's role as an anti-tumor response mediator has been under intense scrutiny lately, with a special focus on ICI and chemotherapy (CHT) treatments [4–7]. Several distinct associations have been discovered between the presence of certain bacterial species/genus and favorable ICI treatment outcomes, such as Akkermansia [5], Alistipes putredinis, Bifidobacterium longum and Prevotella copri [8], Alistipes and Barnesiella [4], Bacteroides dorei and Parabacteroides distasonis [9] in NSCLC, Akkermansia in renal cell carcinoma (RCC) [5] and in hepatocellular carcinoma (HCC) [10], Clostridiales, Ruminococcaceae, Faecalibacterium [11], Bifidobacterium, Collinsella, and Enterococcus in melanoma [12].

The widespread joint -use of the -omics technologies allows us to observe an extensive molecular landscape and behavior of a particular microbe, or the community level. The characterization of the bacterial taxa can be performed with DNA-based shotgun metagenomics (MG), RNA-based metatranscriptomics (MTR), and protein-based metaproteomics methods [13]. Most studies used so far MG approaches that provide information on the relative abundance and metabolic potential of certain taxonomic units, meanwhile the MTR that brings us data referring to the gene expression profile of the microbial community with a focus on currently active genes [14].

Microbes can contribute to the development of certain diseases like obesity, inflammatory bowel disease (IBD), and cancer; however, in healthy individuals, only 28% of the metagenomically identified pathways are widely transcribed [15]. To date, MTR approaches have highlighted the importance of the metabolic dominance of Faecalibacterium prausnitziii, Bacteroides vulgatus, and Alistipes putredinis in IBD [16]. In a transcriptomic analysis of colorectal patients, a greater expression of antibiotic resistance genes was observed than in healthy individuals [17]. In melanoma, transcriptomics revealed a correlation between short progression-free survival (PFS) and the activity of specific biosynthetic pathways (l-rhamnose degradation, guanosine nucleotide biosynthesis, and B vitamin biosynthesis) [18]. In addition, it was reported in patients with oral squamous cell carcinoma (OSCC) evaluated by MTR analysis, that an increase in peptidase activity, oxidative stress, and iron acquisition is overrepresented in tumor samples compared to healthy tissues [19]. The pathway analysis shed light on the upregulation of anaerobic respiration, ribosome biogenesis, and glycerophospholipid metabolism, and the downregulation of inositol phosphate and flavonoid metabolism in breast cancer [13]. A study by Wong-Rolle et al. [20] examined the mutational burden of the tumor and the adjacent tissue. In contrast, another paper by Chang et al. [21] studied the association between immune functions and survival by analyzing intratumoral bacteria. Nevertheless, several studies examined the role of MTR in cancer; immunotherapeutic aspects are not well described. Currently, there is no extensive study on gut MTR in NSCLC treated with ICI.

To our knowledge this is the first study on metatranscriptomics of the gut microbiome and immunotherapy treatment outcomes. Our study might help to understand treatment resistance mechanisms and reveal new potential targets. In the present study, a de novo assembly-based MTR analysis of advanced-stage NSCLC patients treated with ICI was performed to shed light on distinct microbial RNA signatures in patients with different ICI-outcomes.

Materials and methods

Study population

Our study enrolled n = 29 consecutive patients with a histological diagnosis of adenocarcinoma (ADC), squamous cell carcinoma, or NSCLC not otherwise specified (NSCLC-NOS), all at an advanced stage of NSCLC (stage IIIB/IV), treated with immune checkpoint inhibitors (ICI) between 2019 and 2021 at the Pulmonology Hospital of Pulmonology, Torokbalint, Hungary. The clinical TNM (Tumor, Node, Metastasis) staging followed the 8th edition of the Union for International Cancer Control criteria at diagnosis. Age, gender, histology, stage, smoking pack year (PY), line of immunotherapy ((first-line, or CHT-naïve) vs. subsequent line or CHT-treated)), Chronic Obstructive Pulmonary

Disease (COPD) comorbidity (present vs not present), tumor PD-L1 expression (IHC, < 50% vs. ≥ 50%), ICI Response (R) at 3 months, progression-free survival (PFS) in months, PFS group (Long PFS vs Short PFS) and BMI (≤ 30 or > 30 kg/m^2) comprised the clinicopathological data (Table 1). ICI-treated patients showing complete response (CR), partial response (PR), or stable disease (SD) for a minimum of six months were categorized as having long PFS (> 6 months). Conversely, those with progressive disease within six months after starting treatment were categorized as short PFS (≤ 6 months). PFS was calculated from the start of ICI therapy to the point of disease progression, as defined by RECIST 1.1 criteria. The cutoff date for the last follow-up included in this study was March 1st, 2022. Treatment protocols adhered to the current National Comprehensive Cancer Network guidelines. Patients with an ECOG performance status of 0–2 were included in our study. Inclusion criteria required patients to provide informed consent and baseline stool samples within a seven-day window before or after the first intravenous ICI dose. On the day of collection, samples were stored at − 80 °C for subsequent microbiome isolation and sequencing.

In our cohort, n = 5 patients received a combination of chemotherapy- immunotherapy (pembrolizumab, pemetrexed and Carboplatin combination). Other patients received standard-of-care pembrolizumab treatment (n = 11) in first-line and nivolumab treatment (n = 8) in subsequent lines according to the contemporary reimbursements of standard of care treatments. Participants in this cohort included patients treated with PD-1 inhibitors durvalumab (n = 2) and atezolizumab (n = 3), who were part of phase III clinical trials. The collection of samples in this context was not classified as a therapeutic intervention and thus did not require registration on clinicaltrials.gov.

Table 1 Clinical characteristics of patients

	Long PFS (n = 16)	**Short PFS (n = 13)**	**p-value***
Age [years (mean)]	57.5 ± 7.9	60 ± 6.4	0.091
Gender			
Male [52% (n = 15)]	Male [44% (n = 7)]	Male [62% (n = 8)]	0.34
Female [48% (n = 14)]	Female [56% (n = 9)]	Female [38% (n = 5)]	
Histology			
ADC [65% (n = 19)]	ADC [63% (n = 10)]	ADC [70% (n = 9)	0.895
SCC [25% (n = 7)]	SCC [25% (n = 4)]	SCC [23% (n = 3)]	
NSCLC-NOS [10% (n = 3)]	NSCLC-NOS [12%(n = 2)]	NSCLC-NOS [7% (n = 1)]	
Stage			
IIIb [24% (n = 7)]	IIIb [19% (n = 3)]	IIIb [31% (n = 4)]	0.451
IV [76% (n = 22)]	IV [81% (n = 13)]	IV [69% (n = 9)]	
PFS [months, (median)]	15.8 ± 9.4 months	3.2 ± 1.7 months	< 0.001
Chemotherapy			
Treated [31% (n = 9)]	Treated [13% (n = 2)]	Treated [54% (n = 7)]	0.039*****
Naive [69% (n = 17)]	Naive [87% (n = 11)]	Naive [46% (n = 6)]	
PD-L1 TPS			
> 50% [55% (n = 16)]	> 50% [75% (n = 12)]	> 50% [31% (n = 4)]	0.023*
≤ 50% [34% (n = 10)]	≤ 50% [19% (n = 3)]	≤ 50% [54% (n = 7)]	
N/A [11% (n = 3)]	N/A [6% (n = 1)]	N/A [15% (n = 2)]	
Smoking [pack-years (mean)]	40.1 ± 8.5	37.3 ± 10.6	0.26
COPD comorbidity			
Yes [31% (n = 9)]	Yes [38% (n = 6)]	Yes [23% (n = 3)]	0.483
No [66% (n = 19)]	No [62% (n = 10)]	No [70% (n = 9)]	
N/A [3% (n = 1)]	N/A [0% (n = 0)]	N/A [7% (n = 1)]	
BMI			
> 30 kg/m^2 [21% (n = 6)]	> 30 kg/m^2 [12% (n = 2)]	> 30 kg/m^2 [31% (n = 4)]	0.151
≤ 30 kg/m^2 [62% (n = 18)]	≤ 30 kg/m^2 [75% (n = 12)]	≤ 30 kg/m^2 [46% (n = 6)]	
N/A [17% (n = 5)]	N/A [13% (n = 2)]	N/A [23% (n = 3)]	

*Chi squared test

Sample processing

The method of sample preparation was previously described in Dora et al., [4], briefly: 300 mL of 80% aqueous methanol was added to each 100 mg of fecal material in homogenizer tubes, processed on dry ice. Samples were homogenized using the Bead Ruptor 24 Elite (Heart program: 6 m/s, 30 s), vortexed for 10 s, then centrifuged at 13,000 rpm and 4 °C for 10 min. The supernatant was collected, placed in a 96-well filter plate, and centrifuged for an additional 5 min at 700 g, and 4 °C. Patients were recruited into the study upon consenting and providing stool samples within a week of diagnosis, immediately stored at − 80 °C for sequencing.

Shotgun metatranscriptomic pipeline

The RNA was extracted using the Quick RNA Fecal/Soil Microbe Microprep kit from Zymo Research. This involved a 40-min bead beating process of a 100 mg stool sample in 1 mL of S/F RNA Lysis Buffer. The mixture was then centrifuged for 1 min, and 400 µl of the resulting supernatant was passed through a Zymo-Spin™ IIICG Column2 at 3000×*g* for 30 s. This was followed by adding an equal amount of 95% ethanol to the filtered supernatant, and transferring it to a new Zymo-Spin™ IIICG Column2 for RNA to adhere. The column was subsequently cleansed with 400 µl RNA Prep Buffer. The RNA was then collected in 100 µl Nuclease-free water and moved to a Zymo-Spin™ III-HRC Filter, where it was centrifuged at 8000×*g* for 3 min. The RNA obtained was combined with a 200 µl RNA binding buffer and an equal volume of 95% ethanol, then placed onto a Zymo-Spin™ IC Column2. Following the removal of the supernatant, the RNA underwent DNAse I treatment, which consisted of washing it with RNA wash buffer and incubating with 5 µl DNAse I and 35 µl DNA digestion buffer for 15 min. Afterward, it was washed once with 400 µl prep buffer and twice with RNA wash buffer, then gathered in 15 µl RNAse/DNAse free water. The quality and quantity of the isolated RNA were determined using a Qubit fluorometer with a Qubit HS RNA kit from Thermofisher and a Labchip GX Touch with RNA Pico Sensitivity Assay from Perkin Elmer. For depleting ribosomal RNA, 250 ng of RNA was treated with NEBNext rRNA depletion kit v2 (for human/mouse/rat) and NEBNext rRNA depletion kit (for bacteria), using a 1:1 mix of hybridization probes. This was in accordance with the manufacturer's guidelines, followed by library preparation with the Nextflex Rapid Directional RNA-Seq kit, including 12 min of fragmentation to achieve a library size of 320–430 bp.

In the library preparation process, KAPA Single indexes compatible with Illumina systems were employed for indexing, involving 10 cycles of PCR. The concentration and size of the final library were assessed using a Qubit fluorometer and a Labchip GX Touch, specifically with the DNA NGS 3 k assay. Sequencing of the samples was conducted using the NextSeq system, generating paired reads with a length of 81 base pairs each, approximately totaling 20 million read pairs.

De novo transcriptome assembly

The primary goal of de novo assembly was to reconstitute a comprehensive and accurate representation of the transcriptome from fragmented RNA-seq reads. The complex mixture of these reads from various parts of different transcripts also includes elements like transcriptional noise, sequencing artifacts, and transcript isoforms due to alternative splicing events. A fundamental building block of this process involves the use of k-mers. The initial step in the assembly involves creating a comprehensive catalog of all possible k-mers for a given 'k' value and identifying the reads from which these k-mers originate. A De Bruijn graph assembler pipeline, Trinity https://github.com/trinityrnaseq/trinityrnaseq/wiki [22] was utilized that uses a graph-based approach to represent these k-mers. In this graph, each k-mer is represented as a node (or vertex), and a directed edge is established between two nodes if their respective k-mers overlap by k-1 nucleotides. This graphical representation allows the assembly process to be visualized as finding paths through the graph, where each path corresponds to a possible sequence from which the k-mers could have originated. The assembly proceeds by extending these paths until no further extensions are possible, and each distinct path is then retrieved as a separate contig, representing a potential transcript [23].

De novo assembly, particularly in the context of transcriptomes, produces many disconnected De Bruijn subgraphs, each representing groups of related sequences, like transcript isoforms or closely related paralogs. The contigs generated are then subjected to further post-processing to filter and group them, yielding a representative set of assembled sequences. In our analysis, k-mer length was set to the default option recommended by the Trinity assembler. Trinity stands out among the De Bruijn graph-based assemblers for its robustness and wide usage. It identifies contigs potentially related biologically (like splice variants) and groups them as gene isoforms, aiding subsequent analyses like differential gene expression [22].

Post-assembly quality control

In the post-assembly quality control phase, the assembly is evaluated to ensure accuracy and completeness. Sequence length, fragmentation, and read support were assessed using SeqKit [24] to compute N50 values. Because of the limited suitability of the N50 metric for transcriptome assemblies, the ExN50 metric was

used for evaluating transcriptome assemblies in Trinity. The ExN50 metric provides a nuanced representation of assembly quality by focusing on the top expressed transcripts, thereby countering the biases introduced by short, lowly expressed transcripts and long isoforms from highly expressed transcripts [25]. The assembly's composition was then examined using Benchmarking Universal Single-Copy Orthologs (BUSCO), which assesses the presence of essential universal single-copy orthologs, providing a measure of the assembly's completeness. Next, the DETONATE tool [26] was used with the Bellerophon pipeline [27]. These tools provide a holistic assessment by integrating read mapping metrics, evaluating the presence of chimeric sequences, and comparing the assembly against known sequences and databases.

Alignment, abundance estimation, and redundancy reduction of transcripts

Next, the quality of a de novo assembled transcriptome was gauged by the proportion of input reads. Alignment and abundance estimation was performed using an efficient single-step approach called pseudoalignment, by Kallisto [28] (https://github.com/pachterlab/kallisto), which leverage k-mer similarities to associate reads with contigs. During abundance estimation, TPMCalculator [29] was used to calculate normalized metric transcripts per million (TPM). Assemblers often produce more sequences than the genome's gene count due to transcriptional noise and alternative splicing, leading to a surplus of contigs, including transcriptional artifacts, pre-mRNA, and ncRNA. Clustering tool MMSeqs2 [30, 31] was used for Assembly thinning, which grouped sequences by sequence identity and coverage thresholds to reduce redundancy (https://github.com/soedinglab/MMseqs2).

Differential expression analysis and RNA classification

The Differential expression (DE) analysis was performed with the Deseq2 package [32] (https://bioconductor.org/packages/release/bioc/html/DESeq2.html) of the R software package involving the creation of a matrix where each row is a unique sequence and each column a sample-replicate, indicating the number of reads per sequence per sample-replicate. Multiple t-tests were used to compare the mean read counts for each sequence across replicates between conditions. Tximport was used for data preparation [33]. Corrected P-values indicate the statistical significance of expression differences between conditions, while $\log_2$FoldChange (FC) values represent the magnitude of expression changes.

RNA classification and taxonomic alignment

Ad hoc classification primarily involved assessing the coding potentials of assembled contigs using the CPAT tool [34]. Only contigs exceeding a default threshold for coding potential or those generating a peptide sequence via translation were retained. More detailed classification was performed using INFERNAL [35] (INFERence of RNA ALignment, Nawrocki and Eddy 2013), which employs covariance models and the Rfam database to classify input sequences into rRNAs, tRNAs, and lncRNAs. Infernal can also indirectly identify mRNAs by process of elimination. rRNA identification was cross-validated using RNAmmer [36] as an alternative. For DEG analysis and functional annotation, only mRNA transcripts with coding potential were included in the analyses, leaving a total of n = 7489 potential protein-coding gene transcripts. Finally, MMSeqs2 (https://bio.tools/MMseqs2) was employed for the taxonomic annotation of identified transcripts through its integrated modules.

Diversity analyses

The microbial abundance data derived from RNA reads after taxonomic matching with MMSeqs2 were log-transformed, then, Shannon- and Simpson index, and Bray–Curtis dissimilarity were calculated to assess alpha-diversity and quantify compositional differences between samples. For beta-diversity analysis, non-metric multidimensional scaling (NMDS) was applied to visualize the dissimilarity between patients with ellipsoids and centroids added to indicate group-level variation for patients with long- and short PFS. Permutational analysis of variance (permanova) was used to test if compositional differences were statistically significant.

Cluster analyses

Data including the TOP 120 Trinity IDs derived from DEG analysis was normalized and scaled using Z-scores. Heatmaps were constructed with Seaborn's clustermap function. Hierarchical clustering was performed using Scikit-learn's AgglomerativeClustering class with Ward's method and Euclidean distance. The optimal cluster number was determined using the silhouette coefficient implemented in silhouette_score. Cluster dendrograms were visualized using a dendrogram from SciPy. Principal Component analysis (PCA) was performed with Scikit-learn's PCA class. An elbow plot was constructed to determine the optimal number of clusters based on the explained variance ratio per component. Ultimately, the first 3 PCs were retained for further analysis, capturing a cumulative 73.3% of variance. These components were

visualized using a 3D PCA scatter plot, including the first 3 PCs.

PFAMs and pathway analyses

After de novo assembly, we conducted gene prediction on the contigs using Prodigal (https://www.biocode.ltd/catalog-tooldata/prodigal), and then translated the predicted gene sequences into protein sequences for further analysis. We identified protein domain families using the PFAM database, accessed at http://pfam.xfam.org/. This process involved scanning the translated protein sequences against the Pfam-A database with HMMER, available at http://hmmer.org/. The PFAM-A database contains a comprehensive collection of protein domain families represented as profile hidden Markov models (HMMs). We adjusted HMMER settings to balance sensitivity and specificity, employing the default settings for initial scans. For overrepresentation analysis we used the Reactome database and its pathway tool. Overrepresentation analysis is a statistical (hypergeometric distribution) test that determines whether certain Reactome pathways are over-represented (enriched) in the submitted data, corrected for false discovery rate (FDR) using the Benjamani-Hochberg method. DEG analysis for PFAMs was performed using DeSeq2 as described before.

Machine learning models

Patients were classified into Long Progression-Free Survival (PFS) and Short PFS groups. After aligning patient IDs, we encoded the PFS group into a binary format and split the data into training and testing sets. Concerning the number of samples, we decided not to use an independent test but to assess the model performance only by internal validation. Features were normalized to ensure uniformity in scale using the CLR method.

Random forest (RF) model

The best model is determined by the hyperparameters of the RF algorithm. We examined the hyperparameters on carefully chosen intervals, more precisely in gridpoints of hyperparameter combinations. In a gridpoint, we determined the median of the cross-validation mean AUC scores and chose our final model from the models with the highest median value. Key hyperparameters subjected to optimization included the number of trees in the forest, the maximum depth of the trees, and the minimum number of samples required to split an internal node. Our final hyperparameter combination was: bootstrap = True, max_depth = None, max_features = 2, max_samples = None, min_samples_leaf = 1, mean_samples_split = 8, n_estimators = 200, leaving all other hyperparameters at default values. The RF model's performance was assessed through a stratified fivefold cross-validation approach, repeated 50 times for high-reliability evaluation. By employing different random seeds for each repetition, variability in assessing the model's effectiveness was reduced. The AUC score and ROC curve corresponding to the best hyperparameter set was determined after the 5-Fold Stratified CV is repeated 50 times, and in every iteration the ROC curve is plotted and the AUC stored.

Support vector machine (SVM) model

An SVM model was constructed using the Radial Basis Function (RBF) kernel. Feature selection was performed using a recursive feature elimination (RFE) process to identify the subset of features that contributed most significantly to predictive accuracy. The RFE process was iterated with cross-validation to ensure the robustness of selected features. Next, hyperparameter optimization for the SVM model was conducted through a grid search approach, focusing on two main parameters: the penalty parameter (C) and the gamma parameter of the RBF kernel. A range of values for C (1, 10, 100, 1000) and gamma (0.001, 0.0001) were evaluated. The optimal combination was determined based on the highest mean accuracy obtained from a stratified fivefold cross-validation scheme, avoiding both underfitting and overfitting. The model's performance was evaluated using a stratified fivefold cross-validation method, repeated 10 times with different random seeds to mitigate variance in model evaluation.

Extreme gradient boosting (XGBoost)

The best model was obtained by tuning of hyperparameters of the XGBoost classifier by Bayesian 13-fold cross validation (sample size of the smallest class) on a grid with 1000 iterations (max_depth: [3, 15], learning_rate: [0.01, 0.3], subsample: [0.1, 1.0], colsample_bytree: [0.1, 1.0], min_child_weight: [0.1, 10], n_estimators: (50, 400)). The model with the best hyperparameter set was internally validated by 50 times repeated 13-fold CV. The performance metrics of the model accuracy, precision, recall, F1 were averaged over the repeats. Confusion matrices were summed. The ROC curves were calculated based on the predicted probabilities and true labels of test segments and visualized. The pooled ROC curve was obtained by pooling all predicted probabilities and true labels from each test segment of the 50 repetitions.

Risk score generation

To summarize the clinical impact of each RNA transcripts that showed strong predictive role with multivariate Cox regression, we calculated a comprehensive Risk score for each potential biomarker using the following formula:

$$\text{Enhance Risk Score} = \log(\text{HR}) \times \left(\frac{1}{\text{CI width}}\right) \times \left|\text{Wald coefficient}\right| \times (1 - \text{p - value})$$

where "log(HR)" means the log hazard ratio of the biomarker, "1/CI width" means an inverse weight based on the width of the confidence interval (Narrow CIs lead to higher scores), "|Wald Coefficient|" means the absolute value of the Wald coefficient, which adds weight based on statistical significance, and "1 − p-value" meaning smaller p-values (higher significance) increase the score, capped at a maximum value of 1.

To generate a combined risk score for every patient, the values of the risk-increasing biomarkers were summed directly, while the protective biomarkers were subtracted from the total score. Patients were then stratified into high- and low-risk groups based on the median combined score. Patients with scores greater than or equal to the median were classified as high-risk, while those with lower scores were categorized as low-risk.

Statistical analyses

We used the Kolmogorov- Smirnov test to test if data was normally distributed. Differential abundance testing of identified Trinity IDs and taxonomical comparisons were done using the Wilcoxon rank-sum test. P-values less than 0.05 indicate the significance, and all p-values were two-sided. To identify relevant predictor factors, a two-sided Cox proportional hazards regression was performed, using a significance threshold of 0.05. In the multivariate Cox regression, parameters with $p < 0.1$ were included through the backwards elimination method. The fit quality of our multivariate model was assessed using Harrel's C-index, which consistently performed above 0.7 (considered fair) in all analyses. Survival analysis was conducted using Kaplan–Meier (KM) curves, with comparisons between survival curves made using the log-rank test. Cut-off points for the KM curves, as well as specificity and sensitivity values for taxa, were determined through Receiver Operating Characteristic (ROC) curve analysis based on the binary outcomes of short versus long PFS. Bar charts were generated, and statistical tests were conducted with GraphPad's Prism.

Results

In our study, a total of n = 29 advanced-stage (IIIB-IV) NSCLC patients were included, all treated with anti-PD1/PD-L1 immune checkpoint inhibitors (ICI). Patients were classified as "Long PFS" vs Short PFS, based on their progression-free survival, with a default cutoff of 6 months. This classification not only adheres better with long-term results but also enhances the accuracy of evaluating disease progression, particularly in patients who are undergoing pseudo-progression. While n = 9 (31%) of patients were treated with one- or more cycles of chemotherapy (CHT-treated) before ICI initiation and fecal sampling, n = 17 (69%) of patients received ICI first line. According to PD-L1 Immunohistochemistry (IHC), tumor proportion score (TPS) was high (> 50%) in n = 16 (55%) patients, and low (≤ 50%) in n = 10 (34%) patients. Further parameters such as, Age, Gender, Histology, Stage, Smoking pack year (PY), COPD comorbidity and BMI were included as clinical confounders (Table 1). Due to their significant effect on PFS in months and PFS group (Long vs Short), Chemotherapy and PD-L1 TPS were included in multivariate analyses to assess microbial RNA biomarkers in ICI-response (Table 1, Fig. 5).

For the de novo assembly of RNA reads, the Trinity pipeline was used [22], where we identified a total of n = 7489 valid RNA sequences annotated with Trinity IDs (Supplemental Dataset 1) after quality check and removal of potentially non-protein-coding entities (see described in Materials and Methods). After filtering out all Trinity IDs not reaching a minimal contribution of 0.001% to the total RNA abundance and a larger-than-zero abundance in at least 20% of patients, a total of n = 2040 curated gene transcripts remained. Next, we matched all these IDs to the lowest known taxonomic level (LKTL) using MMSeqs2, and where possible, identified corresponding genera, phyla, and domains using NCBI's open-access Taxonomy browser and the Genome Taxonomy Database (r214.1, The University of Queensland). Supplemental Table 1 shows the number of identified taxa in all categories and the number of identified unique entries (Trinity IDs) at each taxonomic level.

Diversity analyses of RNA reads

To assess the gut microbiome's alpha diversity in patients with long- and short PFS, we used abundance values of identified species- and genera-matched RNA reads. We found no significant difference neither among species (p = 0.803, Fig. 1A), nor among genera (p = 0.949, Fig. 1C) after the calculation of the Shannon index. Similar results were obtained using the Simpson index, where alpha-diversity was not significantly different in the two PFS groups, taking species (p = 0.897, Fig. 1B), or genera (p = 0.819, Fig. 1D).

For beta-diversity, Bray–Curtis dissimilarities were calculated in all patients and ordinated using the NMDS method on a 2-dimensional coordinate system. Ellipsoids reflecting dissimilarity variance among samples showed

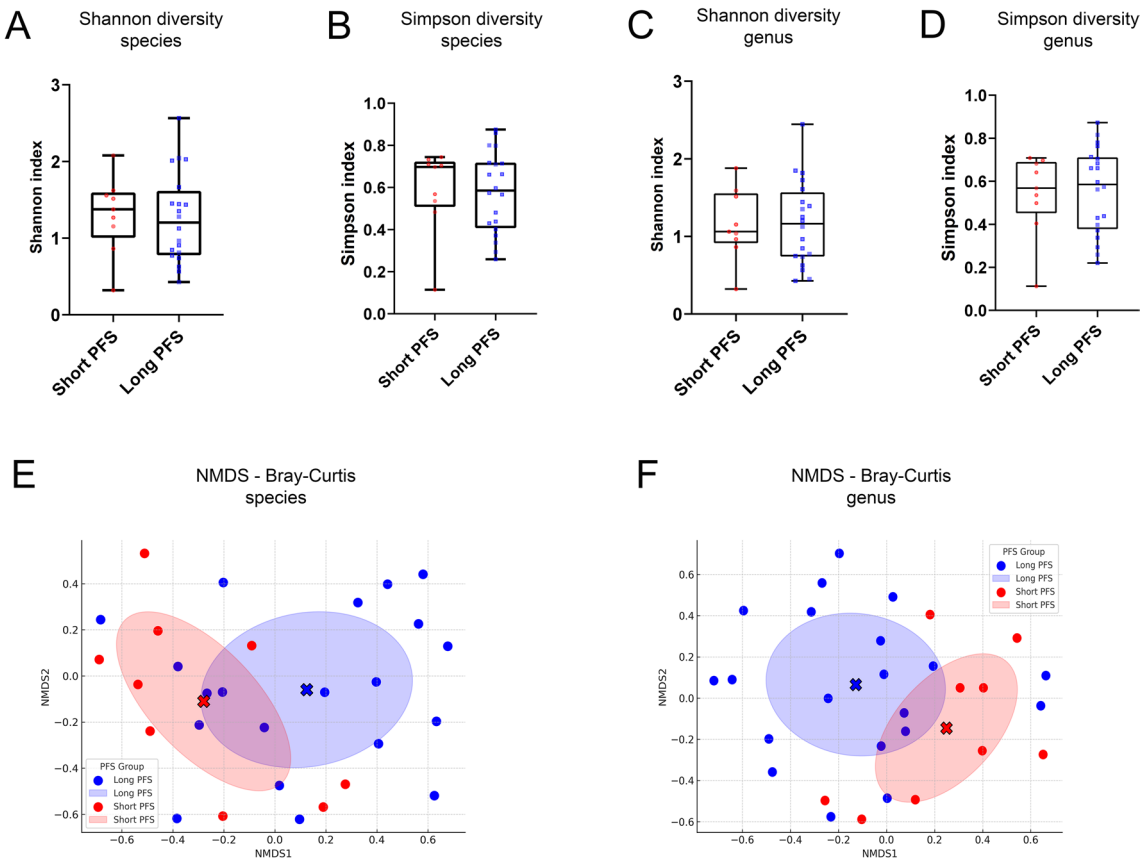

Fig. 1 Alpha and Beta diversity measurements using taxa-aligned RNA reads for species and genera. Alpha diversity analysis of the gut microbiome revealed no significant differences between Long- and Short-PFS groups for both species ($p = 0.803$, **A**) and genera ($p = 0.949$, **C**) using the Shannon index, nor with the Simpson index (species: $p = 0.897$, **B** genera: $p = 0.819$, **D**). Beta diversity, assessed using Bray–Curtis dissimilarities and NMDS ordination, indicated distinct gut microbiome compositions between the groups (**E**, species; **F**, genera), though PERMANOVA tests showed no statistical significance (species: $p = 0.1539$, genera: $p = 0.2465$). Red full circles represent samples from patients with Short PFS and blue full circles samples from patients with Long PFS according to their NMDS-ordinated species- and genus compositions. Statistical comparison of the Shannon and Simpson indices was performed using Welch's test. Statistical significance *$p < 0.05$; **$p < 0.01$, ***$p < .001$, all p-values were two-sided

distinct compositions of the gut microbiome between patients with long vs. short PFS both when accounting for species (Fig. 1E) or genera (Fig. 1F). However, permanova testing showed no statistical significance due to high standard deviation and partly redundant compositions (species: $p = 0.1539$, genera: $p = 0.2465$).

Taxonomic identification of RNA reads

When comparing patients with Long vs Short PFS at the domain level, there were no significant differences regarding the abundance of Bacteria (90.3% vs 89.8%) and Eukaryota (9.6% vs 8.8%). However, the Archaea transcriptome is overrepresented in patients with Short PFS (compared to long PFS) by orders of magnitude (0.003% vs 1.6%, $p < 0.001$) (Fig. 2A). Figure 2B shows the phyletic comparison of Trinity IDs across patients with Long vs Short PFS. According to representation ratios, Actinomycetota (formerly Actinobacteria, 5.4% vs 36.7%, $p < 0.001$) and Euryarchaeota (0.002% vs 1.3%, $p = 0.009$) were relatively overrepresented in Short PFS, whereas Bacillota (formerly Firmicutes, 66.2% vs 42.3%, $p = 0.007$) was relatively overrepresented in Long PFS (Fig. 2B). Bacteroidota (formerly Bacteroidetes, 14.9% vs 9.4%) and Pseudomonadota (formerly Proteobacteria, 5.2% vs 2.5%) RNA were relatively overrepresented in patients with Long PFS as well, but the difference is not significant statistically, due to high standard deviations (Fig. 2B).

Genera and species showing the most significant differences in abundance according to the PFS group are shown in Fig. 2C and D. Genera Bifidobacterium ($p < 0.001$), Collinsella ($p < 0.001$), Limosilactobacillus ($p = 0.043$), and Eubacterium ($p = 0.048$) were significantly overrepresented in Short PFS (Fig. 2C). In the case of species, Bifidobacterium adolescentis L2-32 ($p = 0.0054$), Collinsella aerofaciens ATC 25986 ($p = 0.028$), Bacteroides fragilis ($p = 0.035$), Limosilactobacillus reuteri

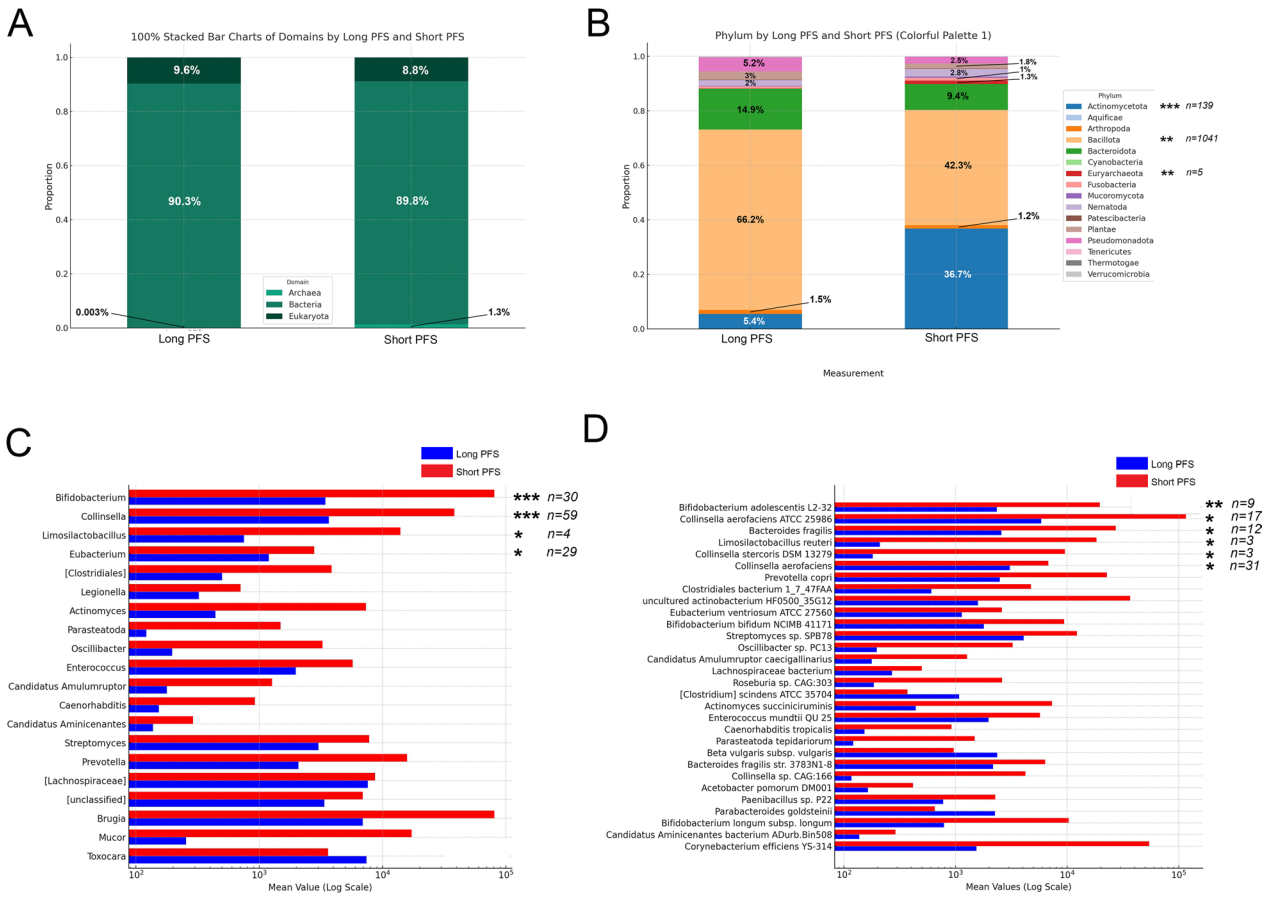

Fig. 2 Comparative microbial abundance in patients with Long vs Short Progression-Free Survival (PFS). **A** Domain-level comparison shows no significant difference in Bacteria (90.3% vs 89.8%, p = 0.563) and Eukaryota (9.6% vs 8.8%, p = 0.092) abundance between Long and Short PFS patients. However, Archaea are significantly more abundant in Short PFS (1.6% vs 0.003%, p < 0.001). **B** Phyletic differences reveal Actinomycetota (36.7% vs 5.4%, p < 0.001, n = 139) and Euryarchaeota (1.3% vs 0.002%, p = 0.009, n = 5) overrepresented in Short PFS, while Bacillota (66.2% vs 42.3%, p = 0.007, n = 1041) is more abundant in Long PFS. Non-significantly overrepresented phyla in Long PFS include Bacteroidota (14.9% vs 9.4%, p = 0.124, n = 124) and Pseudomonadota (5.2% vs 2.5%, p = 0.464, n = 95), with variability due to high standard deviations. Y axis for 100% stacked bar charts shows the proportion of the total identified abundance of taxa. **C** Genus-level analysis shows Bifidobacterium (p < 0.001, n = 30), Collinsella (p < 0.001, n = 59), Limosilactobacillus (p = 0.0.43, n = 4), and Eubacterium (p = 0.048, n = 29) significantly overrepresented in Short PFS. **D** At the species level, Bifidobacterium adolescentis L2-32 (p = 0.0054, n = 9), Collinsella aerofaciens ATC 25986 (p = 0.028, n = 17), Bacteroides fragilis (p = 0.035, n = 12), Limosilactobacillus reuteri (p = 0.048, n = 3), Collinsella stercoris DSM 13279 (p = 0.048, n = 3), and Collinsella aerofaciens (p = 0.049, n = 31) were significantly more abundant in Short PFS, with a noted non-significant trend for Parabacteroides goldsteinii towards higher abundance in Long PFS (p = 0.082, n = 6). Data was derived from MMSeqs2 analysis of 2040 curated gene transcripts (Supplemental Dataset 2), showing the top 20 genera and 30 species according to their difference based on Wilcoxon rank-sum test. X axis indicates mean abundance values calculated from the populational abundance (Long vs Short PFS patients) of all unique taxon-matched Trinity IDs (protein-coding gene transcripts). "*N*" refers to the number of unique Trinity IDs matched with each MMseqs2 taxa. Differential abundance testing was done using the WRS test. A comparison of the percentual contributions of domains and phyla was performed using Welch's test. Statistical significance *p < 0.05; **p < 0.01, ***p < .001, all p-values were two-sided

(p = 0.048), Collinsella stercoris DSM 13279 (p = 0.048), Collinsella aerofaciens (p = 0.049) were significantly overrepresented in Short PFS. Furthermore, there was a non-significant trend in the case of Parabacteroides goldsteinii (p = 0.082) toward higher abundance in patients with Long PFS (Fig. 2D). Supplemental Dataset 2 shows the taxonomic breakdown according to MMSeqs2 of all n = 2040 curated gene transcripts.

Differential expression (DE) analysis and functional annotation of RNA reads

Next, we performed DE analysis on the n = 2040 curated gene transcripts using DESeq2 to identify DEGs in the Long- and Short-PFS patient group and visualize our result on a Volcano plot (Fig. 3). A total of n = 689 genes showed significant difference (− Log_{10}FDR > 1.3, Log2FC > 2) between the two groups, and a total of

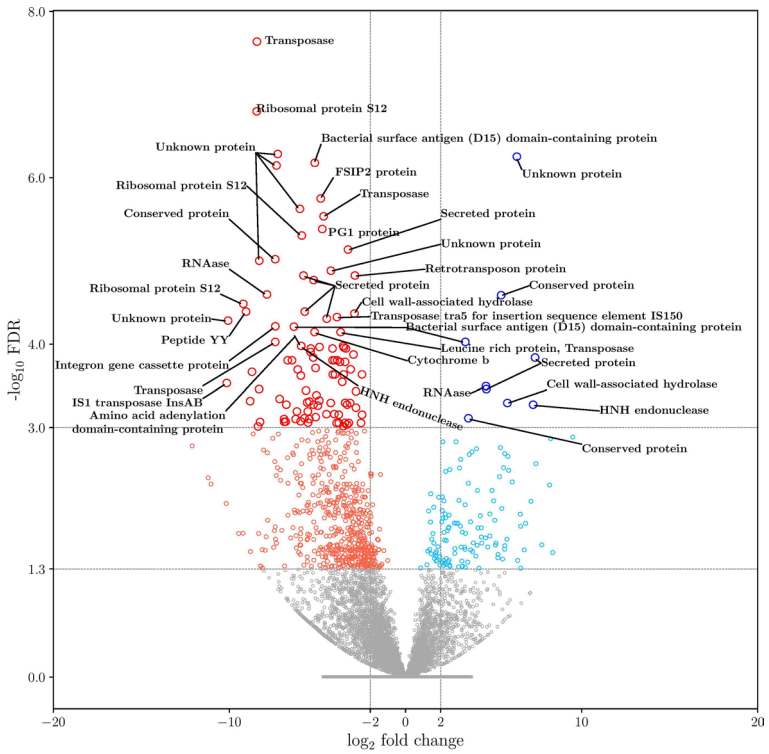

Fig. 3 Differential gene expression (DEG) analysis between Long and Short PFS patients using DESeq2. Analysis of 2040 curated gene transcripts revealed 689 genes with significant differential expression (– Log10FDR > 1.3, Log2FC > [2]). Among these, 120 genes met higher significance criteria (– Log10FDR > 3, Log2FC > [2]). Results are visualized on a Volcano plot, where data points with a greater $-\log_{10}$ FDR value than 1.3 (adjusted p-value = 0.05) are colored red (short PFS) and blue (long PFS). All data points below this significance level occur in grey. DEGs meeting the higher significance criteria are highlighted in bright red (short PFS) and blue (long PFS). Presumptive protein names displaying their UniRef90 cluster are shown for all Trinity IDs with – Log10FDR > 3 and Log2FC > [2] values in the case of Long-PFS-related genes, and top 30 meeting the same criteria in the case of Short-PFS-related genes. X-axis indicates the Log_2 value of fold change (FC), and Y-axis indicates $-\log_{10}$ value of false discovery rate (FDR)

n = 120 genes showed differential expression with – Log-10FDR > 3 significance level and the same threshold for Log2FC (Fig. 2). These 120 protein-coding transcripts were searched against the UniProtKB and the UniRef90 databases with Blastx using default parameter settings. 90/120 (UniProtKB) and 115/120 (UniRef90) matches were recorded for differentially expressed genes. Trinity IDs with multiple hits for the same protein accession (UniProtKB/UniRef90) were filtered by keeping only the one with the highest bit score so as not to bias the

(See figure on next page.)

Fig. 4 Clusters of top 120 Trinity IDs. **A** Hierarchical cluster analysis and heatmap generation on the 120 curated gene transcripts revealed two primary patient groups based on progression-free survival (PFS): a heterogeneous Cluster A (subgroups A1, A2, A3) with varied gene expression, notably increased in gene clusters I, II/B, and II/C, and a homogeneous Cluster B with overexpression of gene cluster II/A (**A**). Short PFS patients predominated in Cluster A, while long PFS patients were more common in Cluster B (p = 0.0095). Cluster A2 and A3 exhibited an abundance of Actinomycetota species (38%). Cluster II/A was uniquely overexpressed in Cluster B, with Cluster II/B showing a mix of Bacillota- (56%) and Actinomycetota-origin genes (20%), and Cluster II/C genes were specifically overexpressed in patients of Cluster I/A, primarily from Bacillota (59%). Axis X shows patients (IDs) in the, whereas indicator bars on top reflect their PFS group (red/blue, short vs long). Each row represents a Trinity ID-coded gene transcript. Axis Y includes 3 columns indicating the phylum of origin color, which is color-coded, LKTU, and UniRef90 cluster. **B** Principal Component Analysis (PCA) on the same 120 gene transcripts identified three to four optimal clusters using the elbow method, aligning with hierarchical clustering results. **C** Three clusters were chosen for clarity, with the first 3 PCs explaining 73.3% of total variance, illustrating patient clustering via PC composition and 95% confidence interval ellipsoids (short PFS in red, long PFS in blue). PCA highlighted a more distinct separation between long and short PFS patients (p < 0.001), with Long PFS patients clustering more closely (Cluster A), and short PFS patients more dispersed across overlapping Clusters B and C (**C**). Compositional differences between clusters according to the PFS group were evaluated using Fisher's exact test. All p-values were two-sided, and significance was considered at p < 0.05

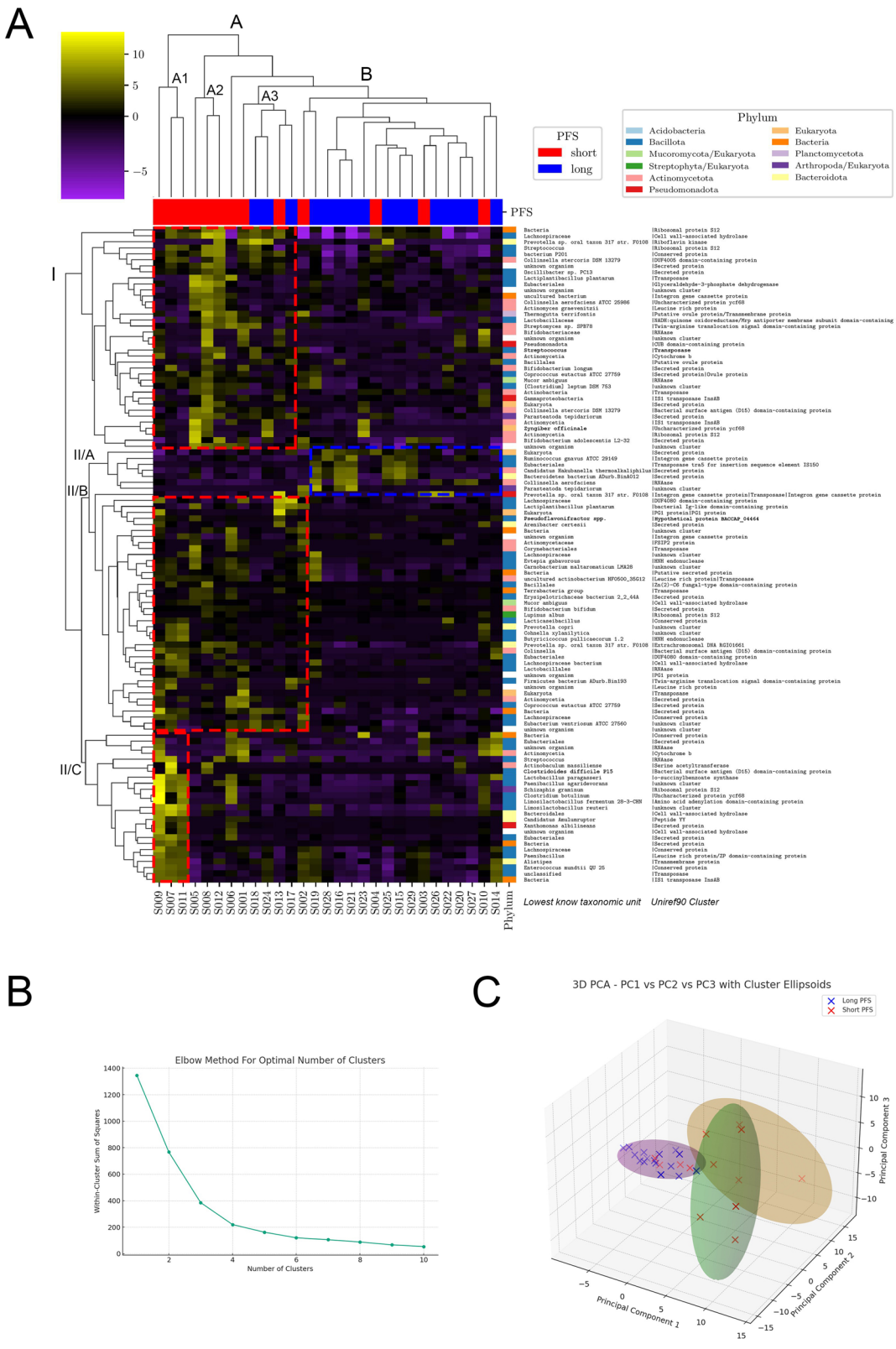

Fig. 4 (See legend on previous page.)

following frequency calculations. In the case of DEGs, we used all isoforms (transcript identifiers that have the same prefix as the gene) of a gene. Only those Blastx matches entered the filtering with a prior condition of bitscore ≥ 50. Due to the higher percentage of genes with relevant and high-quality matches, only annotations of UniRef90 clusters are visualized in Figs. 3 and 4. All curated matches, including UniRef90 clusters and UniProtKB entries, are shown in Supplemental Dataset 3.

Cluster analysis of TOP 120 protein-coding transcripts

We generated a heatmap and performed hierarchical cluster analysis to identify relevant clusters in the functionally annotated 120 curated gene transcripts (Fig. 4A). Dendrogram labels clusters regarding patients (X axis, Long PFS vs Short PFS) and genes (Y axis, phylum, LKTL, and UniRef90 protein cluster). Patients clustered in two major groups: a more heterogeneous cluster A, with subgroups cluster A1, A2, and A3 with slightly different gene expression profiles, where gene cluster I, II/B and/or II/C showed increased expression; and a more homogenous cluster B, where gene cluster II/A was overexpressed (Fig. 4A). We found that patients with short PFS were significantly overrepresented in cluster A, in contrast, there were significantly more patients with long PFS in cluster B ($p = 0.0095$). Genes of Y-axis cluster I were more abundant in patient clusters A2 and A3, with an increased abundance of transcripts from Actinomycetota species (38%). Cluster II/A represented the only group of protein-coding transcripts that showed overexpression in patient cluster B, whereas cluster II/B can be described as a more heterogeneous group of genes with dominance from Bacillota- (56%) and secondly, Actinomycetota phyla (20%). Cluster II/C genes were overexpressed in cluster I/A patients with notably high specificity, originating mainly from the phylum Bacillota (59%). When we assessed the total abundance of the top 120 gene transcripts, we found that gene transcripts from Actinomycetota ($p = 0.043$), Bacillota ($p < 0.001$) phyla, and the Eukaryota ($p = 0.002$) domain were significantly more abundant in a patient with Short PFS compared to patients with Long PFS (Supplemental Fig. 1A–C).

To confirm the findings of our hierarchical cluster analysis, we parallelly applied an alternative approach to establish clusters. We performed principal component analysis (PCA) for the same 120 protein-coding transcripts and determined the optimal number of clusters between 3 or 4 with the elbow method (Fig. 4B), which is roughly in line with the results of the hierarchical cluster analysis. We finally decided to analyze 3 clusters for the ease of interpretation (Fig. 4C). Using the first 3 PCs, which explained 73.3% of the total variance, we plotted patients according to their PC composition in a Cartesian coordinate system. Ellipsoids interpret clusters with their 95% CI. PCA showed that Long PFS patients cluster close to each other with reduced distances suggesting a more homogenous gene transcript profile (cluster A, Fig. 4C). In contrast, patients with short PFS are more scattered with somewhat greater distances between each other in two highly overlapping clusters (clusters B and C, Fig. 4C). When fusing overlapping clusters B and C, we found that patients with short PFS were significantly overrepresented in this cluster, compared to cluster A, where instead, patients with long PFS were more abundant ($p < 0.001$, Fig. 4C). According to our results, PCA showed an even more pronounced separation of long- and short-term PFS patients in gene expression profiles compared to hierarchical cluster analysis. Interestingly, we observed clustering discrepancy only in the case of 2 patients with long PFS, who were clustered in the short PFS-dominated group according to hierarchical cluster analysis, but in the long PFS-dominated cluster according to PCA.

Pathway analyses with protein domain families (PFAMs)

To assess the biological pathways potentially overrepresented in long- or short PFS patients, we used the Protein Domain Families (PFAM) database. After removing PFAMs not reaching a minimal contribution of 0.01% to their abundance and a larger-than-zero abundance in at least 20% of patients, a total of $n = 1209$ PFAMs remained. After performing DEG analysis, $n = 24$ PFAMs showed significant overabundance in patients with short PFS, and $n = 21$ in patients with long PFS. According to Reactome overrepresentation analysis, PFAMs dominant in the

(See figure on next page.)

Fig. 5 Pathway analyses using differentially abundant PFAMs for patients with Long- and Short PFS. Reactome analysis showed Short PFS PFAMs overrepresented in pathways related to hypoxia-response, DNA-synthesis, Translesion-synthesis, Polymerase-switching among others (**A**), while Long PFS PFAMs were associated with various metabolic pathways (**B**). Pathways with [FDR < 0.1] are shown on bar charts, where lower X axis displays (-log10) FDR values and upper X axis displays ratios of entities (green circle) and reactions (red asterisk). Ratios reflect the proportion of pathway-matched UniProt IDs found in our dataset vs all UniProt IDs in that pathway. Taxonomic analysis showed Short PFS PFAMs represented by diverse taxa, including Enterococci and Methanosphaera (**C, D**), whereas Long PFS PFAMs were dominated by Escherichia, particularly *E. coli* (**E, F**). Only species with a minimum contribution of 1% are present on bars (**C, E**). **D** and **F** show the total contribution from the top 5 bacterial species accounting for all represented PFAMs in the corresponding patient group

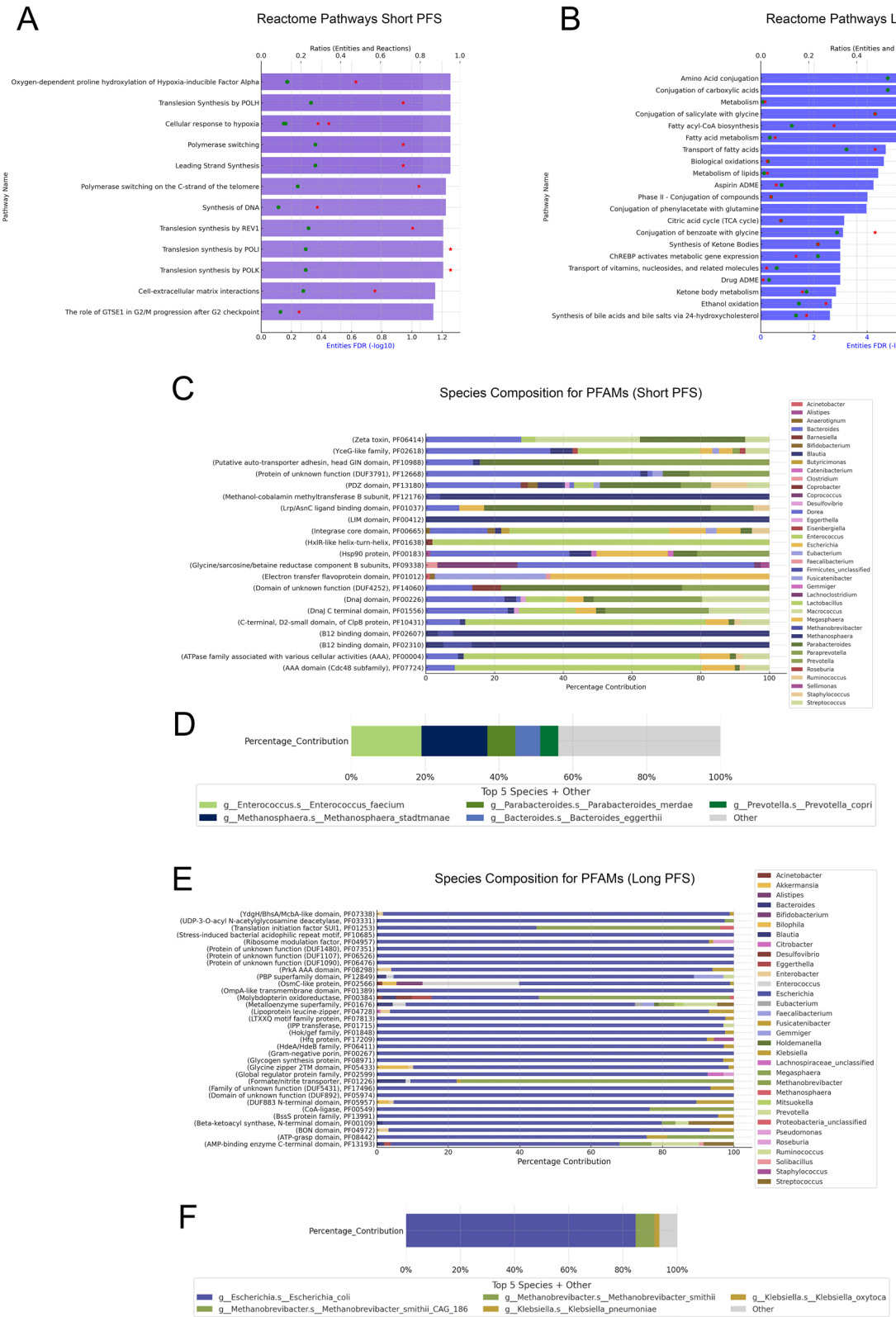

Fig. 5 (See legend on previous page.)

short PFS group were enriched in pathways connected to Hypoxia-response, DNA-synthesis and Translesion-synthesis, a direct mechanism of bypassing unrepaired DNA lesions (Fig. 5A), whereas PFAMs dominant in patients with long PFS were enriched in pathways connected to mainly metabolism, such as Amino acid conjugation, Conjugation of carboxylic acids and Biological oxidation (Fig. 5B). When assessing the taxonomic composition of differentially abundant PFAMs, it was shown that short PFS—associated PFAMs were represented by a diverse range of taxa (Fig. 5C), of whom the most dominant were Enterococci and Methanosphaera with noticeable contributions from Parabacteroides, Bacteroides and Prevotellas (Fig. 5D). Long PFS—associated PFAMs were absolutely dominated by Escherichia, specifically *E. coli*, with a smaller contribution of Methanobrevibacter and Klebsiella (Fig. 5E, F).

Internal validation with machine learning methods: candidates for biomarkers

After extracting biological insight from overrepresented protein families, we used three distinct machine learning models to internally validate the performance of our previously established 120-gene MTR signature. The hierarchically clustered heatmap of the differentially expressed genes showed significant separation between the long and short PFS patient groups. To infer the result, we trained a binary classification model using the Random Forest (RF) classifier implemented in the Sklearn package, validating our results with stratified 50 × fivefold cross-validation. The RF machine learning (ML) algorithm showed good performance, when classifying patients according to Long- vs Short PFS, with an AUC of 0.878 and an Accuracy of 78.1% (Fig. 6A). Parallelly, we established a Support Vector Machine (SVM) ML model as well with the same input using radial basis function (rbf) kernels and stratified 10 × fivefold cross-validation. Results from the SVM model were on par with the RF model concerning prediction performance, with an average AUC of 0.85 and accuracy of 75.6% (Fig. 6B), implicating an overall robust association of our MTR signature with long- and short PFS. Moreover, Ensemble method XGBoost, a scalable, gradient-boosted decision tree (GBDT) ML method was also employed to validate our findings and showed an overall strong correlation with the previous classifiers, with an AUC of 0.84 and an Accuracy of 75% (Fig. 6C). ROC curve per every iteration is shown for the RF and XGB models throughout the stratified cross-validation process (SFig 2A,B). Line chart shows the metric values of SVM model performance, including precision and recall for the test and the train sets after every fold of cross-validation (SFig 2C).

Next, to identify potential biomarkers to classify patients as long vs short progression-free survivals, we performed Receiver Operating Characteristic (ROC) curve analyses for all the 120 microbial genes with functional proteomic and taxonomic annotations described previously. Pairwise comparison of patients with long PFS vs short PFS for all Trinity IDs with an AUC > 0.8 and their presumptive proteins and taxa, are displayed in Fig. 6D–I, and their corresponding ROC curves in Fig. 6J–O. Biomarkers with an increased abundance in Short PFS patients included a retrotransposon protein ($p < 0.001$, AUC = 0.841), a leucine-rich transposase ($p < 0.001$, AUC = 0.81) and an unspecified secreted protein ($p < 0.001$, AUC = 0.865) from various Actinomycetota species, a conserved protein from the genus Lacticaseibacillus ($p = 0.001$, AUC = 0.822), and a ribosomal protein S12 from the plant species Lupinus albus ($p = 0.0037$, AUC = 0.826). An unknown protein from the Lachnospiraceae genus was significantly more abundant in patients with long PFS ($p = 0.0014$, AUC = 0.831). Supplemental Table 3 lists all Trinity IDs with a minimum AUC of 0.7.

Multivariate analysis with clinical confounders

Next, we aimed to identify those microbial gene transcripts that show a strong association with PFS in a clinical setting, adding clinical covariates, such as previous chemotherapy treatment (CHT), PD-L1 TPS, BMI, Gender, and COPD comorbidity. For this, first, we performed logistic regression analysis to assess the overall effect of our clinical confounders on PFS (SFig 3A,B), followed by Kaplan–Meier (KM) analysis for all covariates, where PD-L1-high and CHT-naive patients showed significantly increased PFS compared to PD-L1-low and CHT-treated patients ($p = 0.0059$ and $p = 0.0012$, respectively, SFig 3C,D). In contrast, there were no significant differences in PFS according to gender, BMI, or COPD comorbidity (SFig 7E-G), despite the latter showing strong effect on PFS in the logistic regression model.

Consequently, univariate Cox proportional regression was performed for all 120 genes of our predictive MTR signature and for all included clinical covariates. A total of n = 18 gene transcripts showed a significant association with PFS, as shown in Supplemental Table 4. Cox regression confirmed that only PD-L1 TPS [$p = 0.012$, HR: 0.20 (95% CI, 0.056–0.695)] and chemotherapy [$p = 0.004$, HR: 5.36 (95% CI, 1.738 to 16.544)] had significant effect on PFS from clinical confounders (Fig. 7A). To adjust for potential multicollinearity, these two parameters were included as confounders to assess the predictive role of each gene transcript that passed univariate testing. We confirmed the predictive role of n = 6 Trinity IDs with multivariate analysis shown in Fig. 7A. To estimate

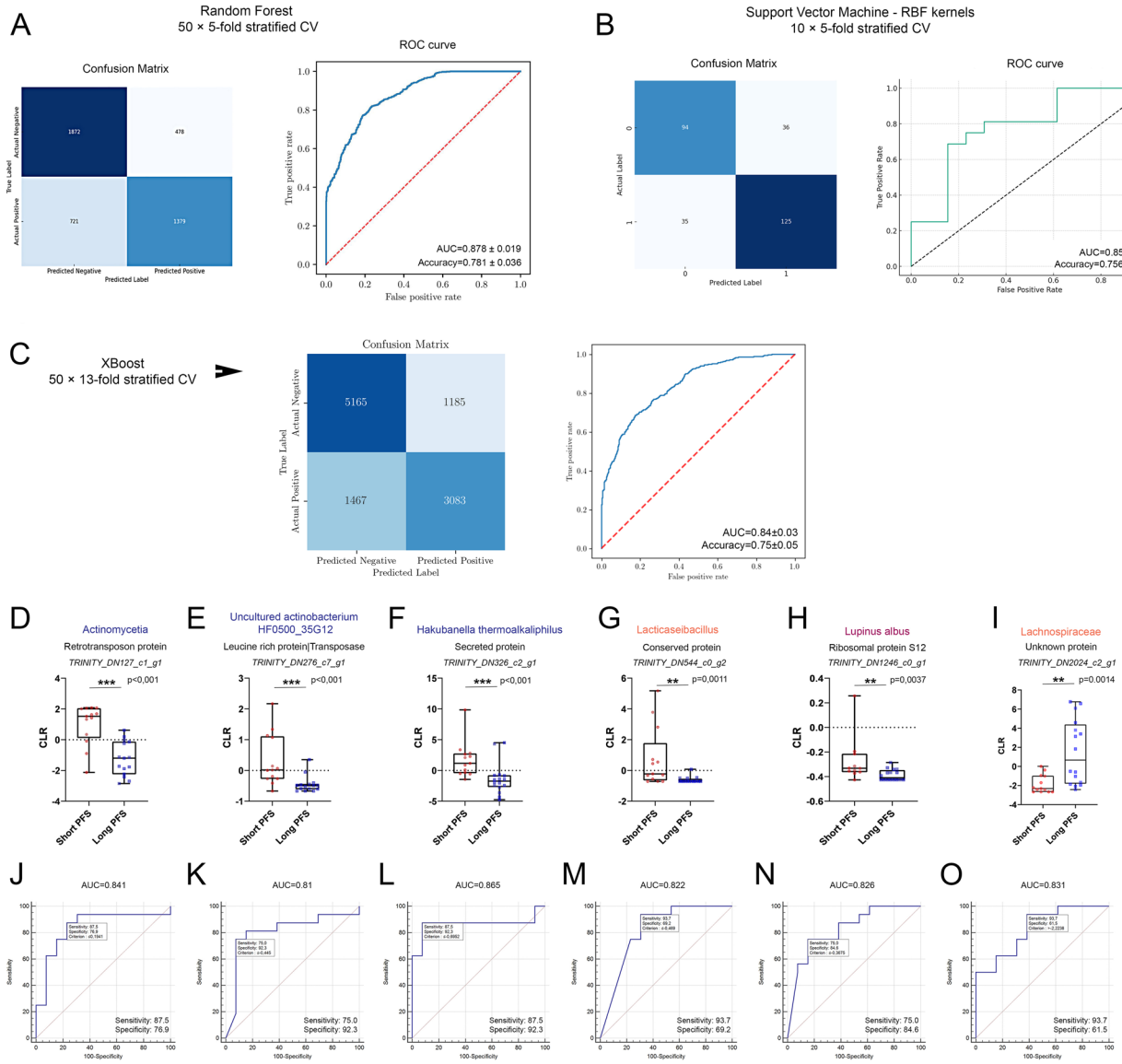

Fig. 6 Internal validation with machine learning. **A** The Random Forest (RF) model achieved an AUC of 0.878 ± 0.019 and 78.1% ± 0.036 accuracy in distinguishing Long vs Short PFS. Confusion matrix shows the number of true- and false positives and negatives for the RF classifier in a heatmap after 50 × fivefold cross-validation and ROC curve indicates RF model performance averaged after cross-validation. **B** The Support Vector Machine (SVM) model, using radial basis function kernels, showed comparable results with an AUC of 0.85 ± 0.046 and 75.6% ± 0.066 accuracy. Confusion matrix shows the number of true- and false positives and negatives for the SVM model in a heatmap after 10 × fivefold cross-validation. ROC curve indicates SVM model performance averaged after cross-validation. **C** The Extreme Gradient Boosting (XGBoost) model achieved an AUC of 0.84 ± 0.03 and 75% ± 0.05 accuracy in distinguishing Long vs Short PFS. Confusion matrix shows the number of true- and false positives and negatives for the RF classifier in a heatmap after 50 × 13-fold cross-validation and ROC curve indicates RF model performance averaged after cross-validation. Biomarker identification through ROC curve analyses for 120 microbial genes revealed potential biomarkers with AUC > 0.8, **D**–**I** Abundance comparisons between patients with short- and long PFS are shown in box plots. **J**–**O** ROC curves show the performance of biomarker candidates to classify patients into long vs short PFS groups. All ROC curves passed the significance threshold ($p < 0.05$). Sensitivity and Specificity for every gene transcript are shown in the bottom left corner of the ROC curve panels. Metric data are shown as medians and 95% CI. Statistical significance *$p < 0.05$; **$p < 0.01$, ***$p < .001$. All p-values were two-sided

the overall clinical relevance of these selected transcripts, we implemented a Risk score derived from the multivariate Cox regression's Wald-coefficient, p-value, HR and its 95% CI values (Fig. 7B, formula described in Methods).

Bacterial surface antigen (D15) domain-containing protein from a Fusobacterium species (Trinity_DN185_c0_g2, $p = 0.026$) and an unknown protein from Lachnospiraceae (Trinity_DN2024_c2_g1, $p = 0.029$) showed a

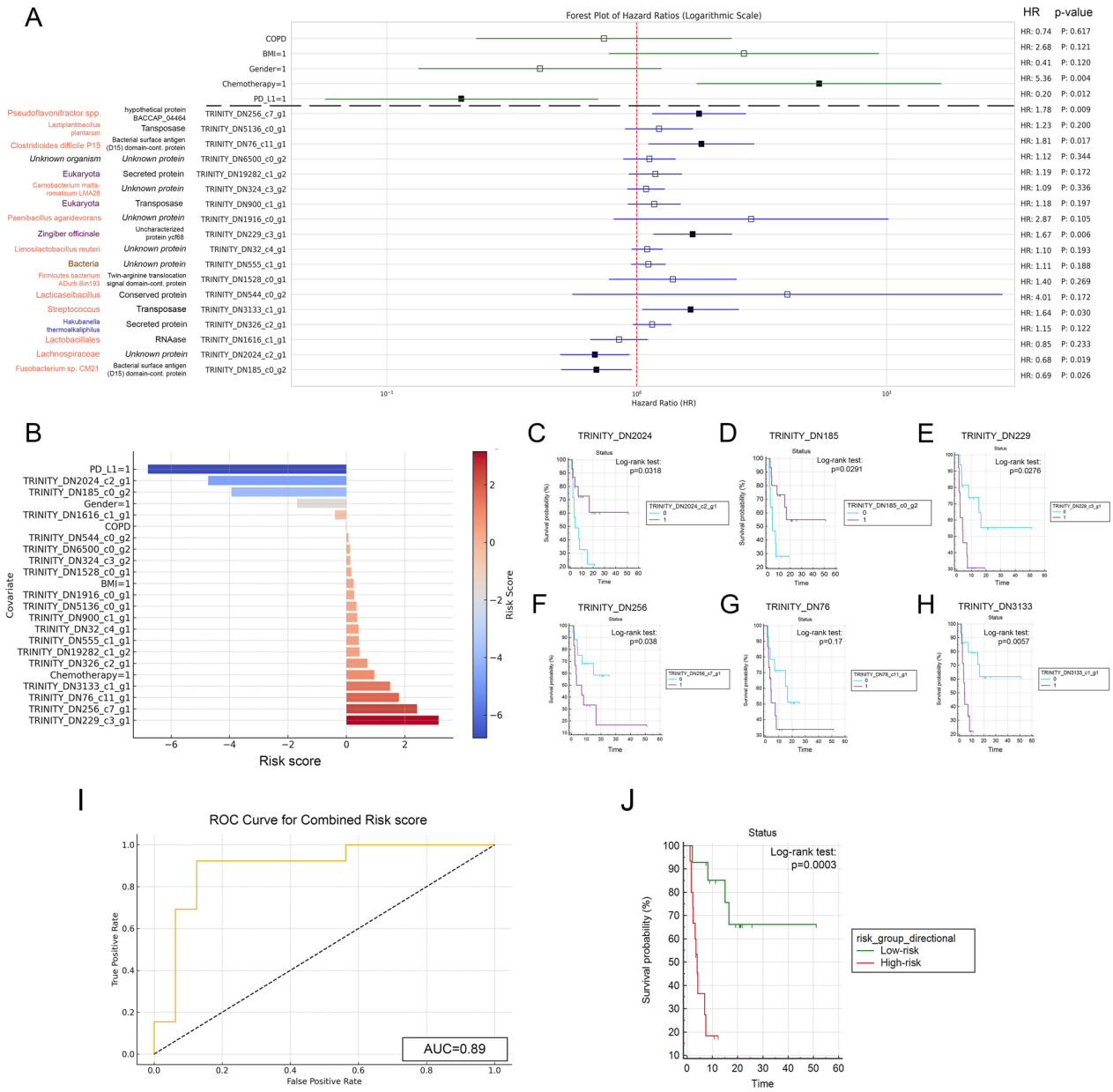

Fig. 7 Assessment of PFS associations with microbial gene transcripts and clinical covariates. **A** Univariate Cox regression on 120 gene transcripts and clinical covariates identified 18 genes significantly associated with PFS. Further multivariate analysis, adjusting for CHT and PD-L1 TPS as confounders, highlighted six genes with significant predictive value for PFS. Notably, a Fusobacterium species protein (HR: 0.69, p = 0.026) and a Lachnospiraceae protein (HR: 0.68, p = 0.029) were positively associated with PFS. In contrast, four others, including ribosomal protein S12 (HR: 1.64, p = 0.03) and uncharacterized protein ycf68 (HR: 1.67, p = 0.006) from unknown Eukaryota species, a bacterial surface antigen (D15) domain-containing protein (HR: 1.81, p = 0.017), and an unknown protein (HR: 1.78, p = 0.009) from unknown organisms were negatively associated with PFS. **B** Risk scores calculated for all covariates and biomarkers are shown on bar charts, where positive values indicate higher risk-, negative values indicate lower risk of progression. **C**–**H** KM curves for RNA transcripts with top 6 risk scores and significant p-values for multivariate Cox regression compare biomarker-high vs low populations (cut-off: median abundance value). p-values for the Log-rank test are indicated along with censored data on charts (0 = biomarker-low group, 1 = biomarker-high group). **I** ROC curve analysis shows the predictive power of combined risk score (AUC = 0.89). **J** KM analysis shows that patients in the low-risk group exhibit significantly increased survival compared to high-risk patients (p = 0.0003). *HR: Hazard ratio*

significant positive effect on PFS, whereas four proteins, including Ribosomal protein S12 (Trinity_DN3133_c1_g1, p = 0.03) and uncharacterized protein ycf68 (Trinity_DN229_c3_g1, p = 0.006) from unknown Eukaryota taxa, Bacterial surface antigen (D15) domain-containing protein (Trinity_DN76_c11_g1, p = 0.017) from an unknown organism and an unknown protein (Trinity_DN256_c7_g1, p = 0.009) from an unknown organism showed a significant adverse effect on PFS, when tested with relevant covariates (Fig. 7A). KM analysis showed that patients with increased expression of both Trinity_DN2024 and Trinity_DN185 RNA transcripts exhibited significantly longer PFS compared to low-expressors (Fig. 7C, D). In contrast, patients with increased expression of RNA transcripts Trinity_DN229, Trinity_DN256 and Trinity_DN3133 showed significantly shorter PFS compared to low-expressors (Fig. 7E, F, H). No significant difference in PFS was detected in high vs low expressors for RNA transcript Trinity_DN76 (Fig. 7G).

To assess the predictive value of these 6 key RNA biomarkers as a combined signature, we generated a combined risk score for every patient, then we classified patients to high- vs low risk groups based on this score. This classification was used to assess the prognostic value of the biomarker signature in relation to PFS. This way 3 patients with long PFS and 12 with short PFS were classified to the high-risk group and 13 patients with long PFS and 1 with short PFS were classified to the low-risk group. The Chi-square test results show a p-value of 0.00036, indicating a statistically significant association between PFS- and risk groups. Furthermore, ROC curve analysis showed that combined risk score from the 6 biomarkers predict the PFS group with an AUC of 0.89 (Fig. 7I) and low-risk patients exhibit significantly increased survival compared to high-risk individuals (p = 0.0003, Fig. 7J).

Discussion

16S rRNA-seq is widely used to detect gut microbiome taxonomy, and represents a relatively cost-effective approach along with MG that is able to capture biodiversity and the metabolic potential. In contrast, MTR sheds light on the active transcription programs of microorganisms [13, 37]. Strong evidence suggests the direct impact of gut commensals on gastrointestinal (GI) cancer development by dysbiosis [38, 39], but studies reported such associations in the case of non-GI cancers as well [40, 41]. Influence on tumor development can happen via secreted metabolites [42–44], or by promoting the beneficial immune responses, including immunotherapy in melanoma [11, 12, 45], renal cell carcinoma [46], GI cancers [47–49] and NSCLC [4, 8, 9, 50, 51]. All of these studies are based on 16S rRNA or Metagenomic sequencing, not accounting for the transcriptomic landscape of the gut microbiome and its effect on anticancer ICI-efficacy. In the present study, we performed a de novo assembly-based sequencing analysis in immunotherapy-treated lung cancer patients to reveal associations between their PFS and gut MTR signatures.

Our study analyzed 29 advanced-stage NSCLC patients undergoing treatment with ICIs to explore the impact of microbial RNA biomarkers on PFS. We identified 7849 protein-coding gene transcripts, from whom 2040 remained for further analyses after filtering for relevant abundance levels (0.001%) and a notable cohort-wide occurrence (min. 20% of patients). Our key findings include the significant overrepresentation of the Archaea domain and phylum Actinomycetota and Euryarchaeota in short PFS patients, whereas Bacillota was more common in long PFS patients. Although Bacteroidota and Pseudomonadota showed higher relative abundance in patients with long PFS, these differences were not statistically significant. Bifidobacterium, Collinsella, Limosilactobacillus, and Eubacterium genera, plus multiple species from these taxonomic units, including Bifidobacterium adolescentis and Collinsella aerofaciens were more abundant in short PFS patients, whereas Parabacteroides goldsteinii exhibited a trend towards being more abundant in the long PFS group. Alpha-diversity did not show significant difference in any comparison when accounting for genera or species that is in line with metagenomic data reported in our previous analysis [4]. However, there was a notable trend towards altered beta-diversity between the two patient groups, though not statistically significant. This might be due to the lower number of metatranscriptionally active taxa and greater standard deviation compared to that experienced during metagenomic analyses. While overall diversity metrics, which account for hundreds of taxa, showed no significant differences between the groups, this can mask key functional distinctions. Even when the broader microbial community appears similar, specific taxa may exhibit differential abundance or heightened transcriptional activity significantly altering the ecological niche.

Findings from taxonomic comparisons are in line with our previous metagenomic study [4]: at the transcriptomic level, phyla Actinomycetota are overrepresented in patients with short PFS, whereas Bacteroidota and Pseudomonadota are rather overrepresented in patients with long PFS. Interestingly, Archaea's and their main phylum Euryarchaeota's increased abundance and relative representation in patients with impaired ICI efficacy is a novel finding that was not yet described with shotgun metagenomics or 16S rRNA sequencing. Instead, Euryarchaeota showed a strong association with low PD-L1 expression after multivariate analysis [4], which might explain its adherence to impaired ICI-efficacy at a

metatranscriptomic level too, PD-L1 expression and PFS exhibiting a strong correlation in our study cohort. Additionally, Bacillota RNA reads were more abundant in long PFS patients, contrasting with our previous metagenomic findings. This may reflect a difference between the phylum's physical presence and its transcriptomic activity in the gut. Bacillota is highly diverse, with hundreds of intestinal species, and its association with long PFS is notable due to the marked deviation in Actinomycetota between patient groups. The association of Actinomycetota with impaired ICI-efficacy at the metagenomic level was reported multiple times [4, 5, 52]. Euryarchaeota, particularly methanogenic species, may also influence immune function via metabolic byproducts like methane, which can reduce gut motility and contribute to a hypoxic environment. Hypoxia can limit the effectiveness of immune cells, thereby weakening the immune response to cancer cells during immunotherapy. Furthermore, the presence of archaea, particularly methanogens, has been associated with dysbiosis and impaired antitumor immunity [53].

Bifidobacterium, Collinsella genera, and individual species showed strong- to moderate metagenomic association with short PFS in earlier NSCLC studies [4, 5]. In contrast, the increased abundance of Alistipes and Barnesiella [4, 5] in long-term responders was not present at the RNA level, possibly due to their transcriptomically silent nature in the gut microbiome. Discrepancies between the MG and MTR signatures of microbial communities have been described earlier by multiple studies [17, 54, 55]. Both Collinsella and Bifidobacterium are involved in carbohydrate metabolism, producing metabolites such as short-chain fatty acids (SCFAs), which usually maintain gut integrity [56]. However, excessive transcriptional activity from these genera could lead to an imbalance in gut metabolites, particularly increased production of gases or bile acid derivatives [57]. For example Bifidobacterium, often linked with immune-modulation, with its overactive strains may paradoxically contribute to immune exhaustion. Specifically, they might overstimulate TLR pathways, triggering chronic, low-grade inflammation. This sustained immune activation could impair the proper functioning of cytotoxic T-cells necessary for ICI response (Vilena et al., 2014).

Another indirect, but hypothesis-generating finding of our analysis is the overall increase of MTR reads detected in the short PFS patient group compared to patients with long PFS, which might suggest a detrimental effect of the increased transcriptomic activity of the gut microbiome driven by a handful of highly active species, including multiple Bifidobacteria, Collinsella or the Eubacterium and the Limosilactobacillus genus of the Lactobacillaceae family. Bifidobacteria responsible for the degradation of complex carbohydrates, the methane-producer *Methanobrevibacter smithii*, and the immune-modulator Bacteriodes fragilis [58, 59] are all considered keystone taxa, just like Eubacterium and Lactobacillus species or the whole Actinomycetota phylum [60, 61]. Keystone taxa in the gut microbiome refer to species or groups of microorganisms that play a critical role in maintaining the structure, function, and health of the microbial community within the gut ecosystem. These organisms can have a distinguishable effect on their environments relative to their abundance [62]. Interestingly, our analysis showed a disproportionately high section between our short PFS-associated taxa and recognized keystone organisms in the gut microbiome that might implicate a decisive influence of transcriptomically active taxa on the intestinal niche-associated anti-cancer immunity. In contrast, the MTR signature predicting potent anti-tumor immune responses are rather ill-defined, with no prominent taxa showing a statistically significant association with Long PFS. Instead, differences are realized only at the phylum and domain level, characterized by a microflora richer in Bacteroidota and Pseudomonadota and harboring a low abundance of Archaea and Actinomycetota.

Next, using DESeq2 for DE analysis, 689 genes were found to have significant differences, with 120 showing high significance. These genes were then analyzed against the UniProtKB and UniRef90 databases to identify their functional annotations and potential protein matches. A cluster analysis of the top 120 protein-coding transcripts revealed two main groups based on gene expression patterns, with one group showing a homogeneous profile associated with longer PFS and another, more heterogeneous group associated with shorter PFS. PCA further confirmed this division by identifying a pronounced separation between the long and short PFS patients based on gene expression profiles. We underpinned the relevance of our predictive 120-gene microbial signature using internal validation via ML algorithms RF, SVM and XGBoost with corresponding AUCs of 0.88, 0.85 and 0.84 respectively. The RF ML model has been used extensively for similar goals in biomarker research and has been proved to work optimally on datasets with smaller sample sizes [63–65] in a similar way as the SVM model, whose power to find complex, non-linear solutions for separating groups of interest has been acknowledged in the past [66, 67].

Due to the challenging nature of matching de novo assembly derived RNA reads with concrete proteins using currently available databases, the predictive 120 gene signature was not suitable to perform pathway analyses to extract biologically meaningful information from long- or short PFS associated microbial signatures. Therefore, we matched differentially abundant

Trinity IDs with PFAM annotations to identify key protein domain families. It was shown that MTR signatures from patients with short PFS were enriched mainly from pathways of DNA synthesis, Response to hypoxia, and Translesion-synthesis. Translesion synthesis (TLS) in prokaryotes is a DNA damage tolerance mechanism where specialized DNA polymerases replicate across lesions, allowing cell survival but often introducing mutations. TLS can impact bacterial pathogenicity and antibiotic resistance by fostering genetic variability. In the gut microbiome, TLS may contribute to bacterial adaptation under stress (e.g., antibiotics, inflammation), influencing microbial diversity and resilience, which could affect gut-related diseases [68, 69]. In contrast, patients with long PFS exhibited enrichment in metabolism-related pathways, especially in Amino acid- and Carboxylic acid conjugation pathways. These pathways are involved in detoxifying substances like bile acids and xenobiotics and producing SCFAs, which have anti-inflammatory effects and support gut barrier integrity [70].

Among biomarker candidates, whose abundance showed the best performance classifying patients with long vs short PFS, only one showed performance above an AUC of 0.8 and was significantly more abundant in patients with long PFS. Lachnospiraceae-derived protein so far classified as "unknown" (Trinity_DN2024_c2_g1) was also associated with increased PFS according to multivariate Cox-regression (HR: 0.68). Gut Lachnospiraceae were described as promoters of tumor immune surveillance, and to be related to high immunoscores in colorectal cancers [71, 72]. They were also more abundant in the clinical benefit response group of patients with hepatobiliary cancers reported by Mao and colleagues [48]. However, more identified RNA transcripts exhibited a stronger association with the short PFS patient group, namely, 92% (35 out of 38) with a minimum AUC of 7.0. Among them, alkaline, reducing fluid-resident Hakubanella thermoalkaliphilus from the Actinomycetota phylum [73] was not yet described in connection with the human microbiome, whereas the genus Lacticaseibacillus, along with other Lactobacillus species described as lactic acid producer probiotics were already heavily implicated in gut health and immunity, including alleviating GI symptoms in Parkinson's disease [74], contribution to cell-wall remodeling and gut microbiome homeostasis [75]. In cancer, the genus has been assigned an anti-proliferative role boosting anti-tumor immunity, but only in colorectal cancer and mouse models [76, 77].

Conventional RNA-seq, reliant on the alignment of sequencing reads to a reference genome, excels in well-characterized organisms, enabling precise gene expression quantification and variant detection due to its data processing and analysis efficiency. However, its efficacy is constrained by the availability and quality of reference genomes, introducing potential biases against detecting novel or highly divergent transcripts [78]. On the other hand, de novo transcriptome assembly, which does not require a reference genome, is invaluable for studying non-model organisms and discovering novel transcripts [23]. In the context of MTR analyses, conventional, database-driven methods' (such as the MetaPhlan/Humann pipeline) dependence on comprehensive databases for read mapping can limit its utility in MTR, where the microbial diversity often exceeds the scope of existing references [79]. Conversely, de novo assembly offers a pathway to explore the transcriptomes of complex microbial ecosystems without the constraints of reference genomes. However, distinguishing between closely related species or strains poses significant challenges [80]. Also, directly identifying prominent metabolic pathways is impossible with de novo assembly, which requires extensive database-alignment algorithms.

This study includes limitations. First, the relatively modest size of the patient cohort and the lack of causal linkage between altered immune response and gut MTR signatures require a cautious interpretation of our findings: further in vitro and in vivo studies along with an independent prospective cohort are needed to validate the effect of these microbial signatures on anti-tumor immunity. Furthermore, RNA-based techniques all have the downside of increased degradability compared to DNA, which is efficiently managed by thorough quality control. We identified multiple transcripts, where biologically meaningful functional and proteomic annotation was not possible, due to the current state of knowledge in the field of microbial transcriptomics and proteomics. However, these "uncharacterized proteins" are not necessarily irrelevant and might participate in complex yet unmapped biological pathways. Due to the current pace of development in functional genomics, it cannot be ruled out that a significant fraction of the functionally uncharacterized transcripts will be annotated in databases in the coming years, and our study could serve as a valuable source of information for microbial genomics. We identified multiple eukaryotic taxa showing a strong association with short PFS. Unfortunately, many of these taxa are poorly characterized both taxonomically and functionally and might be part of the commensal mycobiome or the alimentary flora as part of the patients' diet.

Conclusion

In conclusion, our study demonstrates the associations of the gut microbiome's MTR signatures with PFS in advanced-stage ICI-treated NSCLC patients. We performed de novo assembly-based sequencing that revealed significant associations between PFS and

the abundance of specific microbial RNA biomarkers, including an overrepresentation of the Archaea domain, Actinomycetota phyla, and Bifidobacterium, Collinsella, Eubacterium, and Limosilactibacillus genera in patients with shorter PFS. A clear separation between the long- and short PFS associated transcriptomic signatures verified using cluster analyses and non-linear ML methods, including RF and SVM and XGBoost. While these findings contribute to the growing body of literature on the microbiome's role in cancer therapy, they warrant further research to validate these associations and explore their implications for therapeutic interventions.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-024-05835-y.

Additional file 1.

Additional file 2.

Additional file 3.

Additional file 4.

Author contributions

Conceptualization, D.D., and Z.L.; data curation, P.K., B.L., and C.S.; formal analysis, D.D., and Z.L.; funding acquisition, D.D., G.G. and Z.L.; investigation, D.D., B.L., P.K., and Z.L.; methodology, D.D., P.K. and B.L.; project administration, D.D., G.G., E.D., and Z.L.; resources, G.G. E.D., and Z.L.; software, D.D., P.K. and B.L.; supervision, D.D., and Z.L.; validation, P.K., and C.S.; visualization, D.D., P.K. and B.L.; writing—original draft, D.D., C.S., and Z.L.; writing—review and editing, all authors. All authors have read and agreed to the published version of the manuscript.

Funding

This study was supported by National Research, Development and Innovation Office, #146775, Zoltan Lohinai, #142287, David Dora, LCFA-BMS/IASLC Young Investigator Scholarship Award, New National Excellence Program of the Ministry for Innovation and Technology of Hungary, UNKP-23-5, David Dora, Magyar Tudományos Akadémia, Bolyai Research Scholarship, David Dora, Nemzeti Kutatási, Fejlesztési és Innovációs Alap, #138055, Balazs Ligeti, Thematic Excellence Program, TKP2020-NKA-11, Balazs Ligeti.

Data availability

All identified transcript sequences along with their Trinity IDs are included in the Supplementary dataset 1. MMSeqs2 annotations of all Trinity IDs are available in Supplemental dataset 2. All curated database matches for the top 120 Trinity IDs, including UniRef90 clusters and UniProtKB entries, are shown in Supplemental Dataset 3. Any further data generated or analyzed during this article can be available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

In the current study, we adhered to the Helsinki Declaration's study criteria established by the World Medical Association. The study was formally approved by the national ethics committee, specifically the Hungarian Scientific and Research Ethics Committee of the Medical Research Council (ETT-TUKEB- 50302-2/2017/EKU). Participation in the study was contingent upon the provision of permission by all patients involved. To maintain confidentiality, patient IDs were removed after the collection of clinicopathological data, thereby preventing direct or indirect identification of patients.

Consent for publication

All authors agree to submit the article for publication.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author details

[1]Department of Anatomy, Histology, and Embryology, Semmelweis University, Budapest, Hungary. [2]Pulmonology Hospital of Torokbalint, Torokbalint, Hungary. [3]Translational Medicine Institute, Semmelweis University, Tűzoltó Utca 37-47, 1094 Budapest, Hungary. [4]Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Budapest, Hungary.

Received: 18 March 2024 Accepted: 31 October 2024
Published online: 19 November 2024

References

1. Lahiri A, Maji A, Potdar PD, Singh N, Parikh P, Bisht B, Mukherjee A, Paul MK. Lung cancer immunotherapy: progress, pitfalls, and promises. Mol Cancer. 2023;22(1):40. https://doi.org/10.1186/s12943-023-01740-y.
2. Gandhi L, Rodríguez-Abreu D, Gadgeel S, Esteban E, Felip E, De Angelis F, Domine M, Clingan P, Hochmair MJ, Powell SF, Cheng SY, Bischoff HG, Peled N, Grossi F, Jennens RR, Reck M, Hui R, Garon EB, Boyer M, Rubio-Viqueira B, Novello S, Kurata T, Gray JE, Vida J, Wei Z, Yang J, Raftopoulos H, Pietanza MC, Garassino MC, KEYNOTE-189 Investigators. Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer. N Engl J Med. 2018;378(22):2078–92. https://doi.org/10.1056/NEJMoa1801005.
3. Desai A, Peters S. Immunotherapy-based combinations in metastatic NSCLC. Cancer Treat Rev. 2023;116: 102545. https://doi.org/10.1016/j.ctrv.2023.102545.
4. Dora D, Ligeti B, Kovacs T, Revisnyei P, Galffy G, Dulka E, Krizsán D, Kalcsevszki R, Megyesfalvi Z, Dome B, Weiss GJ, Lohinai Z. Non-small cell lung cancer patients treated with Anti-PD1 immunotherapy show distinct microbial signatures and metabolic pathways according to progression-free survival and PD-L1 status. Oncoimmunology. 2023;12(1):2204746. https://doi.org/10.1080/2162402X.2023.2204746.
5. Routy B, Le Chatelier E, Derosa L, Duong CPM, Alou MT, Daillère R, Fluckiger A, Messaoudene M, Rauber C, Roberti MP, Fidelle M, Flament C, Poirier-Colame V, Opolon P, Klein C, Iribarren K, Mondragón L, Jacquelot N, Qu B, Ferrere G, Clémenson C, Mezquita L, Masip JR, Naltet C, Brosseau S, Kaderbhai C, Richard C, Rizvi H, Levenez F, Galleron N, Quinquis B, Pons N, Ryffel B, Minard-Colin V, Gonin P, Soria JC, Deutsch E, Loriot Y, Ghiringhelli F, Zalcman G, Goldwasser F, Escudier B, Hellmann MD, Eggermont A, Raoult D, Albiges L, Kroemer G, Zitvogel L. Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. Science. 2018;359(6371):91–7. https://doi.org/10.1126/science.aan3706.
6. Fluckiger A, Daillère R, Sassi M, Sixt BS, Liu P, Loos F, Richard C, Rabu C, Alou MT, Goubet AG, Lemaitre F, Ferrere G, Derosa L, Duong CPM, Messaoudene M, Gagné A, Joubert P, De Sordi L, Debarbieux L, Simon S, Scarlata CM, Ayyoub M, Palermo B, Facciolo F, Boidot R, Wheeler R, Boneca IG, Sztupinszki Z, Papp K, Csabai I, Pasolli E, Segata N, Lopez-Otin C, Szallasi Z, Andre F, Iebba V, Quiniou V, Klatzmann D, Boukhalil J, Khelaifia S, Raoult D, Albiges L, Escudier B, Eggermont A, Mami-Chouaib F, Nistico P, Ghiringhelli F, Routy B, Labarrière N, Cattoir V, Kroemer G, Zitvogel L. Cross-reactivity between tumor MHC class I-restricted antigens and an enterococcal bacteriophage. Science. 2020;369(6506):936–42. https://doi.org/10.1126/science.aax0701.
7. Zitvogel L, Ma Y, Raoult D, Kroemer G, Gajewski TF. The microbiome in cancer immunotherapy: diagnostic tools and therapeutic strategies. Science. 2018;359(6382):1366–70. https://doi.org/10.1126/science.aar6918.
8. Jin Y, Dong H, Xia L, Yang Y, Zhu Y, Shen Y, Zheng H, Yao C, Wang Y, Lu S. The diversity of gut microbiome is associated with favorable responses to anti-programmed death 1 immunotherapy in chinese patients with NSCLC. J Thorac Oncol. 2019;14(8):1378–89. https://doi.org/10.1016/j.jtho.2019.04.007.
9. Dora D, Weiss GJ, Megyesfalvi Z, Gállfy G, Dulka E, Kerpel-Fronius A, Berta J, Moldvay J, Dome B, Lohinai Z. Computed tomography-based

quantitative texture analysis and gut microbial community signatures predict survival in non-small cell lung cancer. Cancers (Basel). 2023;15(20):5091. https://doi.org/10.3390/cancers15205091.
10. Tai N, Peng J, Liu F, Gulden E, Hu Y, Zhang X, Chen L, Wong FS, Wen L. Microbial antigen mimics activate diabetogenic CD8 T cells in NOD mice. J Exp Med. 2016;213(10):2129–46. https://doi.org/10.1084/jem.20160526.
11. Gopalakrishnan V, Spencer CN, Nezi L, Reuben A, Andrews MC, Karpinets TV, Prieto PA, Vicente D, Hoffman K, Wei SC, Cogdill AP, Zhao L, Hudgens CW, Hutchinson DS, Manzo T, Petaccia de Macedo M, Cotechini T, Kumar T, Chen WS, Reddy SM, Szczepaniak Sloane R, Galloway-Pena J, Jiang H, Chen PL, Shpall EJ, Rezvani K, Alousi AM, Chemaly RF, Shelburne S, Vence LM, Okhuysen PC, Jensen VB, Swennes AG, McAllister F, Marcelo Riquelme Sanchez E, Zhang Y, Le Chatelier E, Zitvogel L, Pons N, Austin-Breneman JL, Haydu LE, Burton EM, Gardner JM, Sirmans E, Hu J, Lazar AJ, Tsujikawa T, Diab A, Tawbi H, Glitza IC, Hwu WJ, Patel SP, Woodman SE, Amaria RN, Davies MA, Gershenwald JE, Hwu P, Lee JE, Zhang J, Coussens LM, Cooper ZA, Futreal PA, Daniel CR, Ajami NJ, Petrosino JF, Tetzlaff MT, Sharma P, Allison JP, Jenq RR, Wargo JA. Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. Science. 2018;359(6371):97–103. https://doi.org/10.1126/science.aan4236.
12. Matson V, Fessler J, Bao R, Chongsuwat T, Zha Y, Alegre ML, Luke JJ, Gajewski TF. The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. Science. 2018;359(6371):104–8. https://doi.org/10.1126/science.aao3290.
13. Aitmanaitė L, Širmonaitis K, Russo G. Microbiomes, their function, and cancer: how metatranscriptomics can close the knowledge gap. Int J Mol Sci. 2023;24(18):13786. https://doi.org/10.3390/ijms241813786.
14. Mukherjee A, Reddy MS. Metatranscriptomics: an approach for retrieving novel eukaryotic genes from polluted and related environments. 3 Biotech. 2020;10(2):71. https://doi.org/10.1007/s13205-020-2057-1.
15. Abu-Ali GS, Mehta RS, Lloyd-Price J, Mallick H, Branck T, Ivey KL, Drew DA, DuLong C, Rimm E, Izard J, Chan AT, Huttenhower C. Metatranscriptome of human faecal microbial communities in a cohort of adult men. Nat Microbiol. 2018;3(3):356–66. https://doi.org/10.1038/s41564-017-0084-4.
16. Schirmer M, Franzosa EA, Lloyd-Price J, McIver LJ, Schwager R, Poon TW, Ananthakrishnan AN, Andrews E, Barron G, Lake K, Prasad M, Sauk J, Stevens B, Wilson RG, Braun J, Denson LA, Kugathasan S, McGovern DPB, Vlamakis H, Xavier RJ, Huttenhower C. Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. Nat Microbiol. 2018;3(3):337–46. https://doi.org/10.1038/s41564-017-0089-z.
17. Lamaudière MTF, Arasaradnam R, Weedall GD, Morozov IY. The colorectal cancer microbiota alter their transcriptome to adapt to the acidity, reactive oxygen species, and metabolite availability of gut microenvironments. mSphere. 2023;8(2):e0062722. https://doi.org/10.1128/msphere.00627-22.
18. Peters BA, Wilson M, Moran U, Pavlick A, Izsak A, Wechter T, Weber JS, Osman I, Ahn J. Relating the gut metagenome and metatranscriptome to immunotherapy responses in melanoma patients. Genome Med. 2019;11(1):61. https://doi.org/10.1186/s13073-019-0672-4.
19. Yost S, Stashenko P, Choi Y, Kukuruzinska M, Genco CA, Salama A, Weinberg EO, Kramer CD, Frias-Lopez J. Increased virulence of the oral microbiome in oral squamous cell carcinoma revealed by metatranscriptome analyses. Int J Oral Sci. 2018;10(4):32. https://doi.org/10.1038/s41368-018-0037-7.
20. Wong-Rolle A, Dong Q, Zhu Y, Divakar P, Hor JL, Kedei N, Wong M, Tillo D, Conner EA, Rajan A, Schrump DS, Jin C, Germain RN, Zhao C. Spatial meta-transcriptomics reveal associations of intratumor bacteria burden with lung cancer cells showing a distinct oncogenic signature. J Immunother Cancer. 2022;10(7): e004698. https://doi.org/10.1136/jitc-2022-004698.
21. Chang YS, Hsu MH, Tu SJ, Yen JC, Lee YT, Fang HY, Chang JG. Metatranscriptomic analysis of human lung metagenomes from patients with lung cancer. Genes (Basel). 2021;12(9):1458. https://doi.org/10.3390/genes12091458.
22. Haas BJ, Papanicolaou A, Yassour M, et al. De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. Nat Protoc. 2013;8(8):1494–512.
23. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52.
24. Shen W, Le S, Li Y, et al. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE. 2016;11(10):e0163962.
25. Hölzer M, Marz M. De novo transcriptome assembly: a comprehensive cross-species comparison of short-read RNA-Seq assemblers. Gigascience. 2019;8(5).
26. Li B, Fillmore N, Bai Y, et al. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol. 2014;15(12):553.
27. Kerkvliet J, de Fouchier A, van Wijk M, et al. The bellerophon pipeline, improving de novo transcriptomes and removing chimeras. Ecol Evol. 2019;9(18):10513–21.
28. Bray NL, Pimentel H, Melsted P, et al. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525–7.
29. Alvarez RV, Pongor LS, Mariño-Ramírez L, et al. TPMCalculator: one-step software to quantify mRNA abundance of genomic features. Bioinformatics. 2019;35(11):1960–2.
30. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026–8.
31. Mirdita M, Steinegger M, Breitwieser F, et al. Fast and sensitive taxonomic assignment to metagenomic contigs. Bioinformatics. 2021;37(18):3029–31.
32. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biol. 2014;15.
33. Love MI, Soneson C, Robinson MD. Importing transcript abundance datasets with tximport. Dim Txi Inf Rep Sample1 2017;1.
34. Wang L, Park HJ, Dasari S, et al. CPAT: coding-potential assessment tool using an alignment-free logistic regression model. Nucleic Acids Res. 2013;41(6):e74.
35. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29(22):2933–5.
36. Lagesen K, Hallin P, Rødland EA, et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35(9):3100–8.
37. Shakya M, Lo CC, Chain PSG. Advances and challenges in metatranscriptomic analysis. Front Genet. 2019;25(10):904. https://doi.org/10.3389/fgene.2019.00904.
38. Yang Y, Du L, Shi D, et al. Dysbiosis of human gut microbiome in young-onset colorectal cancer. Nat Commun. 2021;12:6757. https://doi.org/10.1038/s41467-021-27112-y.
39. Wong CC, Yu J. Gut microbiota in colorectal cancer development and therapy. Nat Rev Clin Oncol. 2023;20:429–52. https://doi.org/10.1038/s41571-023-00766-x.
40. Doocey CM, Finn K, Murphy C, et al. The impact of the human microbiome in tumorigenesis, cancer progression, and biotherapeutic development. BMC Microbiol. 2022;22:53. https://doi.org/10.1186/s12866-022-02465-6.
41. Long Y, Tang L, Zhou Y, et al. Causal relationship between gut microbiota and cancers: a two-sample Mendelian randomisation study. BMC Med. 2023;21:66. https://doi.org/10.1186/s12916-023-02761-6.
42. Kim K, et al. Propionate of a microbiota metabolite induces cell apoptosis and cell cycle arrest in lung cancer. Mol Med Rep. 2019;20:1569–74.
43. Grajeda-Iglesias C, et al. Oral administration of Akkermansia muciniphila elevates systemic antiaging and anticancer metabolites. Aging. 2021;13:6375–405.
44. He Y, et al. Gut microbial metabolites facilitate anticancer therapy efficacy by modulating cytotoxic CD8+ T cell immunity. Cell Metab. 2021;33:988–1000.
45. McCulloch JA, Davar D, Rodrigues RR, Badger JH, Fang JR, Cole AM, Balaji AK, Vetizou M, Prescott SM, Fernandes MR, Costa RGF, Yuan W, Salcedo R, Bahadiroglu E, Roy S, DeBlasio RN, Morrison RM, Chauvin JM, Ding Q, Zidi B, Lowin A, Chakka S, Gao W, Pagliano O, Ernst SJ, Rose A, Newman NK, Morgun A, Zarour HM, Trinchieri G, Dzutsev AK. Intestinal microbiota signatures of clinical response and immune-related adverse events in melanoma patients treated with anti-PD-1. Nat Med. 2022;28(3):545–56. https://doi.org/10.1038/s41591-022-01698-2.
46. Salgia NJ, et al. Stool microbiome profiling of patients with metastatic renal cell carcinoma receiving anti-PD-1 immune checkpoint inhibitors. Eur Urol. 2020;78:498–502.

47. Peng Z, et al. The gut microbiome is associated with clinical response to anti-PD-1/PD-L1 immunotherapy in gastrointestinal cancer. Cancer Immunol Res. 2020;8:1251–61.
48. Mao J, et al. Gut microbiome is associated with the clinical response to anti-PD-1 based immunotherapy in hepatobiliary cancers. J Immunother Cancer. 2021;9: e003334.
49. Kaune T, Griesmann H, Theuerkorn K, Hämmerle M, Laumen H, Krug S, Plumeier I, Kahl S, Junca H, Gustavo Dos Anjos Borges L, Michl P, Pieper DH, Rosendahl J. Gender-specific changes of the gut microbiome correlate with tumor development in murine models of pancreatic cancer. iScience. 2023;26(6):106841. https://doi.org/10.1016/j.isci.2023.106841.
50. Hakozaki T, et al. The gut microbiome associates with immune checkpoint inhibition outcomes in patients with advanced non-small-cell lung cancer. Cancer Immunol Res. 2020;8:1243–50.
51. Derosa L, et al. Intestinal Akkermansia muciniphila predicts clinical response to PD-1 blockade in patients with advanced non-small-cell lung cancer. Nat Med. 2022;28:315–24.
52. Chau J, Yadav M, Liu B, Furqan M, Dai Q, Shahi S, Gupta A, Mercer KN, Eastman E, Hejleh TA, Chan C, Weiner GJ, Cherwin C, Lee STM, Zhong C, Mangalam A, Zhang J. Prospective correlation between the patient microbiome with response to and development of immune-mediated adverse effects to immunotherapy in lung cancer. BMC Cancer. 2021;21(1):808. https://doi.org/10.1186/s12885-021-08530-z.
53. Kumpitsch C, Fischmeister FPS, Mahnert A, Lackner S, Wilding M, Sturm C, Springer A, Madl T, Holasek S, Högenauer C, Berg IA, Schoepf V, Moissl-Eichinger C. Reduced B12 uptake and increased gastrointestinal formate are associated with archaeome-mediated breath methane emission in humans. Microbiome. 2021;9(1):193. https://doi.org/10.1186/s40168-021-01130-w.
54. Franzosa EA, et al. Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. Nat Rev Microbiol. 2014;13(6):360–72.
55. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, Wampach L, Schneider JG, Hogan A, de Beaufort C, Wilmes P. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol. 2016;10(2):16180. https://doi.org/10.1038/nmicrobiol.2016.180.
56. Holmes ZC, Villa MM, Durand HK, Jiang S, Dallow EP, Petrone BL, Silverman JD, Lin PH, David LA. Microbiota responses to different prebiotics are conserved within individuals and associated with habitual fiber intake. Microbiome. 2022;10(1):114. https://doi.org/10.1186/s40168-022-01307-x.
57. Collins SL, Stine JG, Bisanz JE, Okafor CD, Patterson AD. Bile acids and the gut microbiota: metabolic interactions and impacts on disease. Nat Rev Microbiol. 2023;21(4):236–47. https://doi.org/10.1038/s41579-022-00805-x.
58. Wu S, et al. A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. Nat Med. 2009;15:1016–22.
59. Busing JD, Buendia M, Choksi Y, Hiremath G, Das SR. Microbiome in eosinophilic esophagitis-metagenomic, metatranscriptomic, and metabolomic changes: a systematic review. Front Physiol. 2021;10(12): 731034. https://doi.org/10.3389/fphys.2021.731034.
60. Banerjee S, Schlaeppi K, van der Heijden MGA. Keystone taxa as drivers of microbiome structure and functioning. Nat Rev Microbiol. 2018;16(9):567–76. https://doi.org/10.1038/s41579-018-0024-1.
61. Trosvik P, de Muinck EJ. Ecology of bacteria in the human gastrointestinal tract—identification of keystone and foundation taxa. Microbiome. 2015;3:44. https://doi.org/10.1186/s40168-015-0107-4.
62. Tudela H, Claus SP, Saleh M. Next generation microbiome research: identification of keystone species in the metabolic regulation of host-gut microbiota interplay. Front Cell Dev Biol. 2021;1(9): 719072. https://doi.org/10.3389/fcell.2021.719072.
63. Ni Y, Lohinai Z, Heshiki Y, Dome B, Moldvay J, Dulka E, Galffy G, Berta J, Weiss GJ, Sommer MOA, Panagiotou G. Distinct composition and metabolic functions of human gut microbiota are associated with cachexia in lung cancer patients. ISME J. 2021;15(11):3207–20. https://doi.org/10.1038/s41396-021-00998-8.
64. Han S, Williamson BD, Fong Y. Improving random forest predictions in small datasets from two-phase sampling designs. BMC Med Inform Decis Mak. 2021;21:322. https://doi.org/10.1186/s12911-021-01688-3.
65. Acharjee A, Larkman J, Xu Y, et al. A random forest based biomarker discovery and power analysis framework for diagnostics research. BMC Med Genomics. 2020;13:178. https://doi.org/10.1186/s12920-020-00826-6.
66. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics Proteomics. 2018;15(1):41–51. https://doi.org/10.21873/cgp.20063.
67. Wang F, Su Q, Li C. Identidication of novel biomarkers in non-small cell lung cancer using machine learning. Sci Rep. 2022;12:16693. https://doi.org/10.1038/s41598-022-21050-5.
68. Fuchs RP, Fujii S. Translesion DNA synthesis and mutagenesis in prokaryotes. Cold Spring Harb Perspect Biol. 2013;5(12): a012682. https://doi.org/10.1101/cshperspect.a012682.
69. Joseph AM, Badrinarayanan A. Visualizing mutagenic repair: novel insights into bacterial translesion synthesis. FEMS Microbiol Rev. 2020;44(5):572–82. https://doi.org/10.1093/femsre/fuaa023.
70. Oliphant K, Allen-Vercoe E. Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health. Microbiome. 2019;7(1):91. https://doi.org/10.1186/s40168-019-0704-8.
71. Zhang X, Yu D, Wu D, Gao X, Shao F, Zhao M, Wang J, Ma J, Wang W, Qin X, Chen Y, Xia P, Wang S. Tissue-resident Lachnospiraceae family bacteria protect against colorectal carcinogenesis by promoting tumor immune surveillance. Cell Host Microbe. 2023;31(3):418-432.e8. https://doi.org/10.1016/j.chom.2023.01.013.
72. Hexun Z, Miyake T, Maekawa T, Mori H, Yasukawa D, Ohno M, Nishida A, Andoh A, Tani M. High abundance of Lachnospiraceae in the human gut microbiome is related to high immunoscores in advanced colorectal cancer. Cancer Immunol Immunother. 2023;72(2):315–26. https://doi.org/10.1007/s00262-022-03256-8.
73. Merino N, Kawai M, Boyd ES, Colman DR, McGlynn SE, Nealson KH, Kurokawa K, Hongoh Y. Single-cell genomics of novel actinobacteria with the Wood-Ljungdahl pathway discovered in a serpentinizing system. Front Microbiol. 2020;9(11):1031. https://doi.org/10.3389/fmicb.2020.01031.
74. Yang X, He X, Xu S, Zhang Y, Mo C, Lai Y, Song Y, Yan Z, Ai P, Qian Y, Xiao Q. Effect of Lacticaseibacillus paracasei strain Shirota supplementation on clinical responses and gut microbiome in Parkinson's disease. Food Funct. 2023;14(15):6828–39. https://doi.org/10.1039/d3fo00728f.
75. Suissa R, Olender T, Malitsky S, et al. Metabolic inputs in the probiotic bacterium Lacticaseibacillus rhamnosus contribute to cell-wall remodeling and increased fitness. Npj Biofilms Microbiomes. 2023;9:71. https://doi.org/10.1038/s41522-023-00431-2.
76. Zhang SL, Han B, Mao YQ, Zhang ZY, Li ZM, Kong CY, Wu Y, Chen GQ, Wang LS. Lacticaseibacillus paracasei sh2020 induced antitumor immunity and synergized with anti-programmed cell death 1 to reduce tumor burden in mice. Gut Microbes. 2022;14(1):2046246. https://doi.org/10.1080/19490976.2022.2046246.
77. Shi Y, Zhang C, Cao W, Li L, Liu K, Zhu H, Balcha F, Fang Y. Extracellular vesicles from Lacticaseibacillus paracasei PC-H1 inhibit HIF-1α-mediated glycolysis of colon cancer. Future Microbiol. 2024. https://doi.org/10.2217/fmb-2023-0144.
78. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.
79. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, Giannoukos G, Boylan MR, Ciulla D, Gevers D, Izard J, Garrett WS, Chan AT, Huttenhower C. Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci U S A. 2014;111(22):E2329–38. https://doi.org/10.1073/pnas.1319284111.
80. Nagarajan N, Pop M. Sequence assembly demystified. Nat Rev Genet. 2013;14(3):157–67.

Publisher's Note