Supplementary Figures for "Patients with mesenchymal tumours and high *Fusobacteriales* prevalence have worse prognosis in colorectal cancer (CRC)"

Manuela Salvucci[1], Nyree Crawford[2], Katie Stott[2], Susan Bullman[3,4], Daniel B. Longley[2], and Jochen H.M. Prehn[1]*

[1]Centre for Systems Medicine, Department of Physiology and Medical Physics, Royal College of Surgeons in Ireland, Dublin, Ireland;

[2]Patrick G. Johnston Centre for Cancer Research, School of Medicine, Dentistry and Biomedical Science, Queen's University Belfast, Northern Ireland, UK;

[3]Dana-Farber Cancer Institute, Harvard Medical School, Boston, USA;

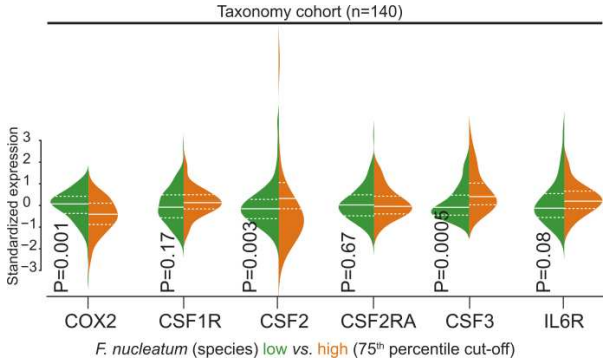
[4]Fred Hutchinson Cancer Research Center, Human Biology Division, Seattle, USA.

Corresponding author: Prof. Jochen H. M. Prehn, Department of Physiology and Medical Physics, Royal College of Surgeons in Ireland, 123 St. Stephen's Green, Dublin 2, Ireland. Tel.: +353-1-402-2255; Fax: +353-1-402-2447; E-mail: jprehn@rcsi.ie.

Data and code availability: Datasets and source code will be publicly available and archived upon publication at Zenodo (https://10.5281/zenodo.4019142).

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Supplementary Figure 1.



Association between Fn load and inflammation signalling in the human host.

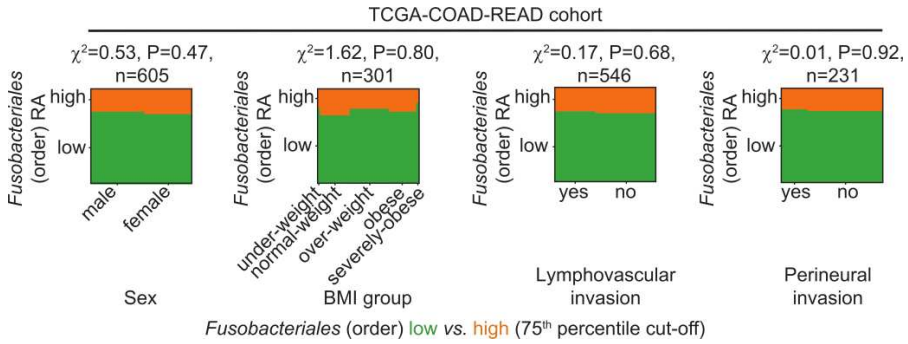
Distribution of key player genes grouped by *Fn* (high vs. low, using the 75th percentile as cut-off) for patients of the Taxonomy cohort.

Median and lower (25th) and upper (75th) percentiles are indicated by white solid or dashed lines, respectively. Statistical significance was evaluated Kruskal-Wallis tests and P-values are reported.

Salvucci et al.

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Supplementary Figure 2.



Association between Fusobacteriales relative abundance (RA) and human host clinico-pathological features in the TCGA-COAD-READ cohort.

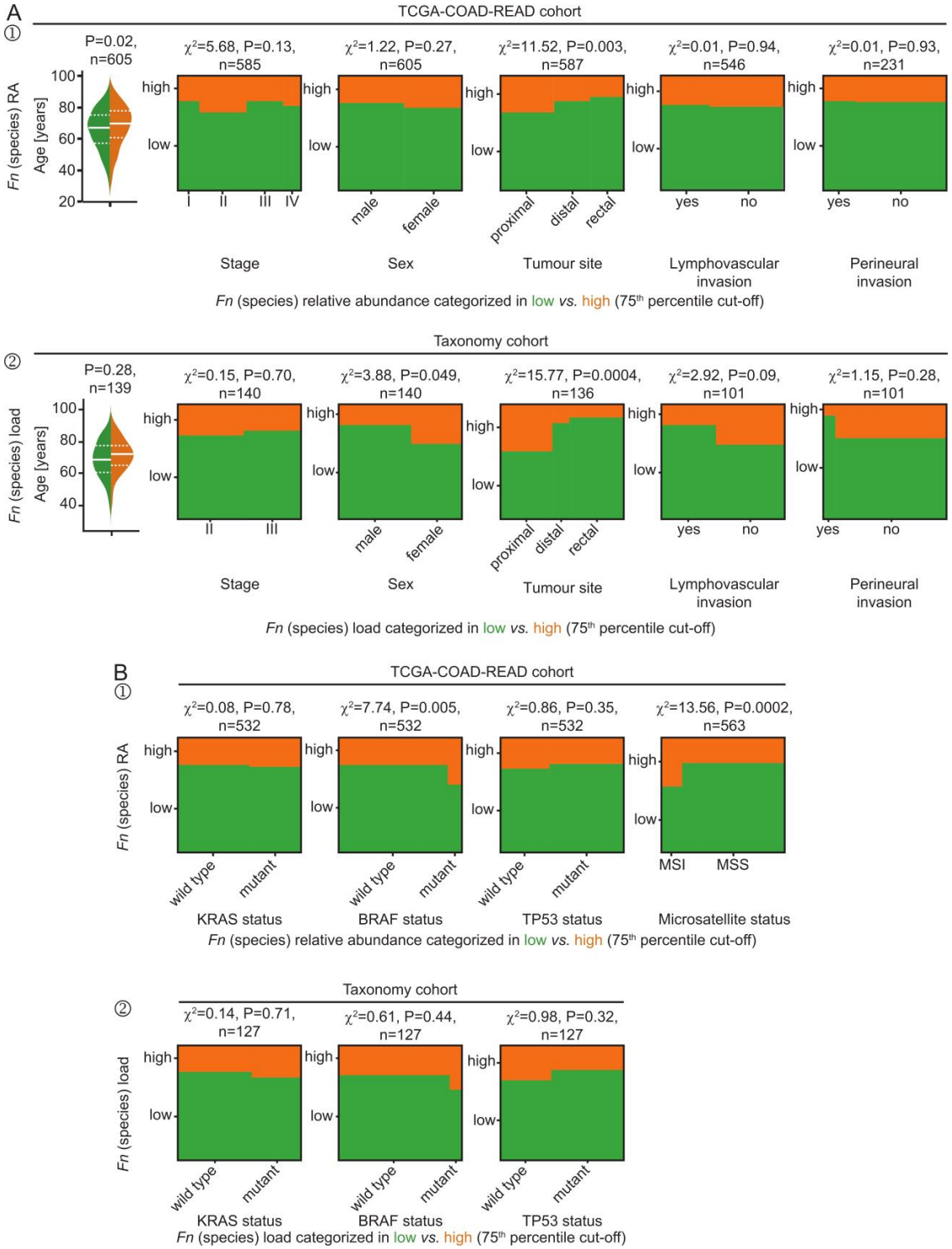
Mosaic plots depicting the relationship between categorical clinico-pathological characteristics of the human host and *Fusobacteriales* RA. Patients were classified as *Fusobacteriales*-low or -high using the 75th percentile as cut-off and indicated in green and orange, respectively. Statistical significance was evaluated with χ^2 independence tests and the χ^2 test statistic and the mod-log-likelihood P-values are reported.

Abbreviations. BMI: body mass index.

Salvucci et al.

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Supplementary Figure 3.



Salvucci et al.

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

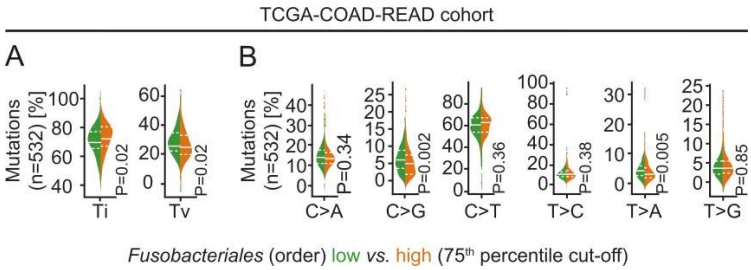
Association between human host clinico-pathological (A) and mutational (B) features in Fn-low vs. high patients of the TCGA-COAD-READ (1) and Taxonomy (2) cohorts.

Fn is expressed as relative abundance (RA) for patients of the TCGA-COAD-READ cohort or load for patients of the in-house Taxonomy cohort. Patients were categorised in low vs. high subgroups using the 75th percentile as cut-off and indicated in green and orange, respectively. Association between *Fn* and continuous variables is depicted with violin plots, median and lower (25th) and upper (75th) percentiles are indicated by white solid or dashed lines, respectively. Statistical significance was evaluated Kruskal-Wallis tests and P-values are reported. Association between *Fn* and categorical clinico-pathological characteristics is depicted with mosaic plots and statistical significance was evaluated with χ^2 independence tests and the χ^2 test statistic and the mod-loglikelihood P-values are reported.

Salvucci et al.

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Supplementary Figure 4.



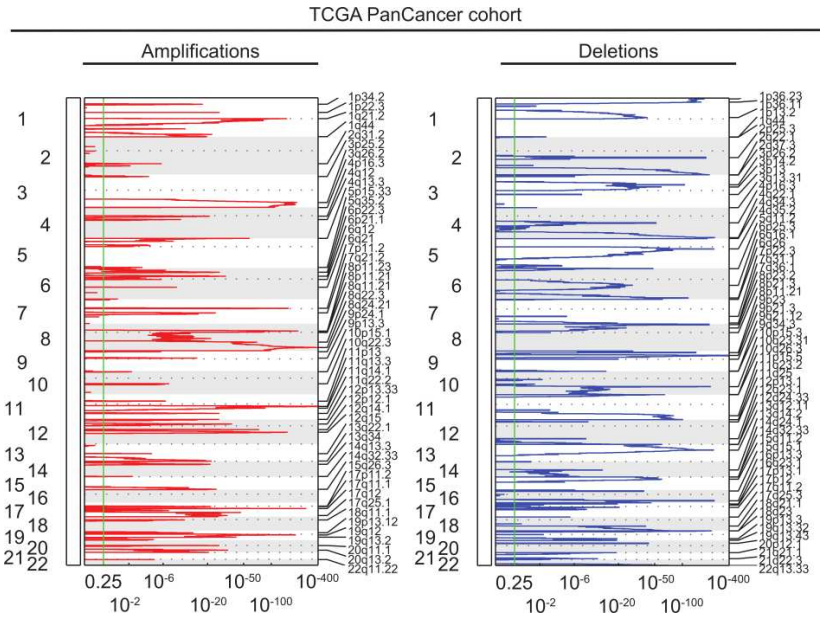
Association between Fusobacteriales relative abundance (RA) and DNA substitution mutations in the patients of TCGA-COAD-READ cohort.

A-B. Distribution of transitions (Ti) and transversions (Tv) (**A)** and conversion changes (**B)** in patients of the TCGA-COAD-READ cohort classified as *Fusobacteriales*-low (in green) or -high (in orange) based on a 75th percentile cut-off. Median and lower (25th) and upper (75th) percentiles are indicated by white solid or dashed lines, respectively. Statistical significance was evaluated Kruskal-Wallis tests and P-values are reported.

Salvucci et al.

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Supplementary Figure 5.



Recurrent copy number alterations in patients of the TCGA PanCancer cohort.

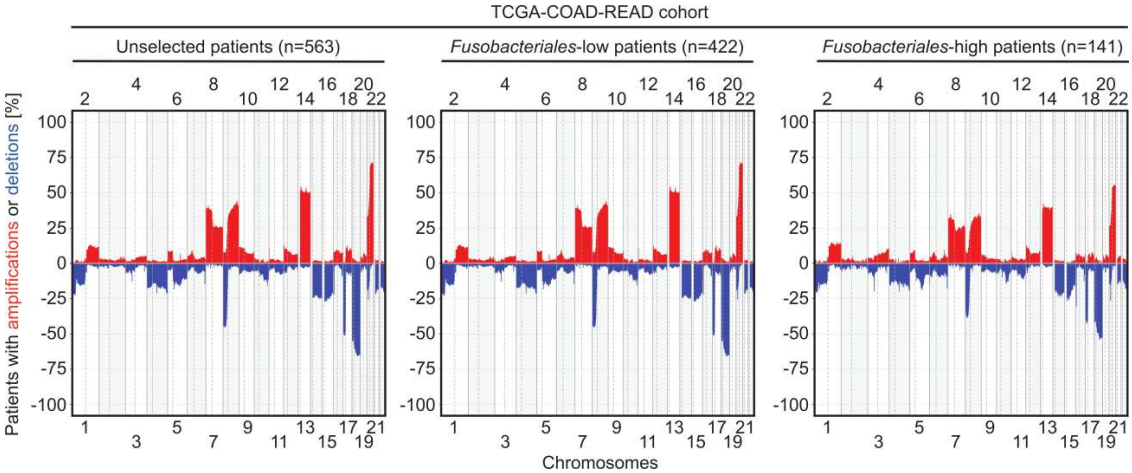
Amplifications (in red, left hand-side) and deletions (in blue, right hand-side) computed by GISTIC2 analysis to detect recurrent copy number alterations in the TCGA PanCancer cohort (n=9142).

Chromosome bands are indicated (y axis) and cytobands that reached statistical significance (as indicated by q-values) are shown.

Salvucci et al.

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Supplementary Figure 6.



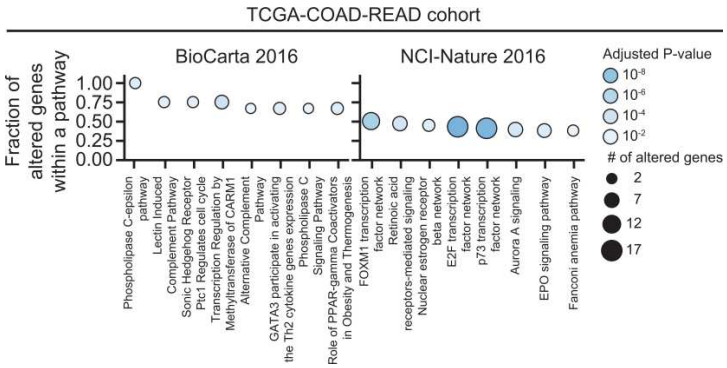
Frequency of copy number alterations in patients of the TCGA-COAD-READ cohort.

Frequency of occurrence of copy number amplifications (in red) or deletions (in blue) by chromosome in the whole unselected cohort (left panel) and in subgroups restricted to cases with low- (middle panel) or high- (right panel) *Fusobacteriales* relative abundance for patients of the TCGA-COAD-READ cohort.

Salvucci et al.

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Supplementary Figure 7.



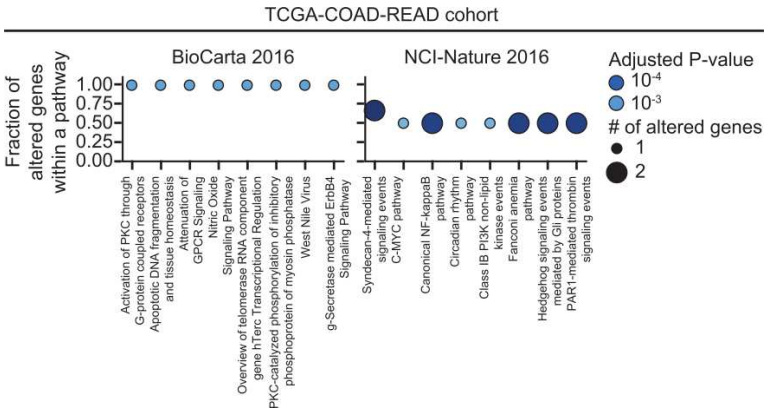
Pathway enrichment analysis for genes differentially expressed by Fusobacteriales relative abundance (RA) in patients of the TCGA-COAD-READ cohort.

Enrichment analysis on genes identified as differentially expressed by *Fusobacteriales* RA in patients of the TCGA-COAD READ cohort. Analysis was performed with *EnrichR* querying the *BioCarta* (version 2016) and *NCI-Nature* (version 2016) pathway databases. The number of identified altered genes for each pathway is encoded by the marker size and the magnitude of the associated P-values is color-coded, as indicated in the legend.

Salvucci et al.

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Supplementary Figure 8.



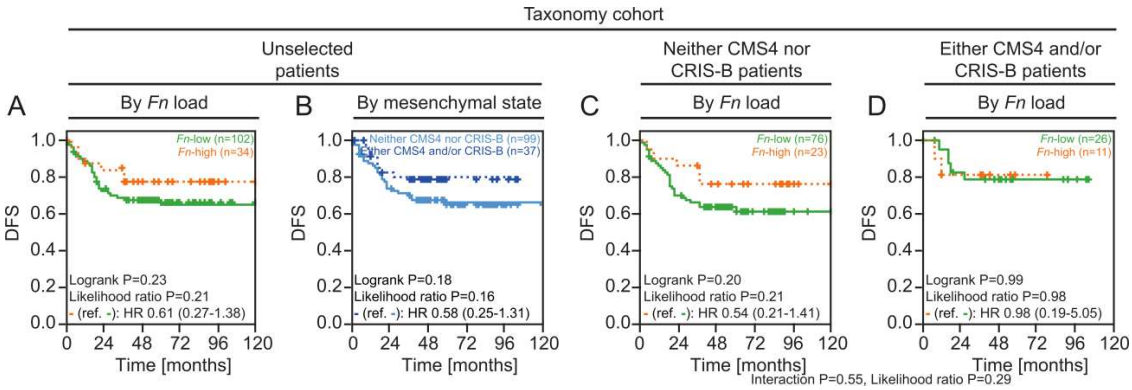
Pathway enrichment analysis for proteins differentially expressed by Fusobacteriales relative abundance (RA) in patients of the TCGA-COAD-READ cohort.

Enrichment analysis on proteins identified as differentially expressed by *Fusobacteriales* RA in patients of the TCGA-COAD READ cohort. Analysis was performed with *EnrichR* querying the *BioCarta* (version 2016) and *NCI-Nature* (version 2016) pathway databases. The number of identified altered genes for each pathway is encoded by the marker size and the magnitude of the associated P-values is color-coded, as indicated in the legend.

Salvucci et al.

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Supplementary Figure 9.

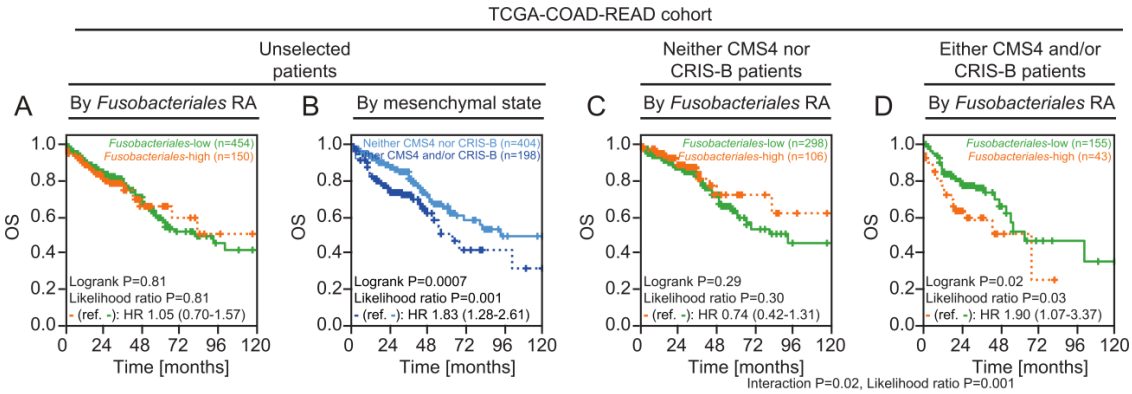


Kaplan-Meier plots comparing disease-free-survival (DFS) in patients of the Taxonomy cohort grouped by *Fn* load (**A**), mesenchymal status (**B**) and by *Fn* load within the non-mesenchymal and mesenchymal patients' subpopulations (**C-D**). Patients were categorised in *Fn*-low or -high subgroups using the 75th percentile as cut-off. *Consensus Molecular Subtype* (CMS) and *Cancer Intrinsic Subtype* (CRIS) assignments were used to categorise patients in non-mesenchymal ("Neither CMS4 nor CRIS-B") or mesenchymal, respectively ("Either CMS4 and/or CRIS-B").

Salvucci et al.

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Supplementary Figure 10.

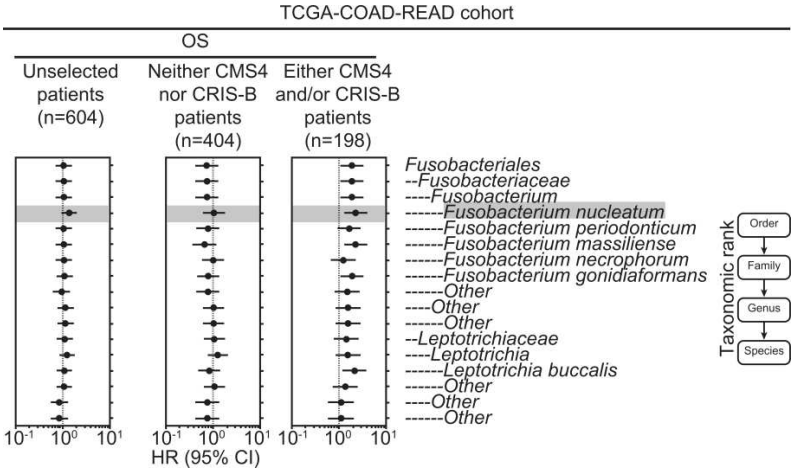


Kaplan-Meier plots comparing overall survival (OS) in patients of the TCGA-COAD-READ cohort grouped by *Fusobacteriales* RA (**A**), mesenchymal status (**B**) and by *Fusobacteriales* RA within the non-mesenchymal and mesenchymal patients' subpopulations (**C-D**). Patients were categorised in *Fusobacteriales*-low or -high subgroups using the 75th percentile as cut-off. *Consensus Molecular Subtype* (CMS) and *Cancer Intrinsic Subtype* (CRIS) assignments were used to categorise patients in non-mesenchymal ("Neither CMS4 nor CRIS-B") or mesenchymal, respectively ("Either CMS4 and/or CRIS-B").

Salvucci et al.

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Supplementary Figure 11.

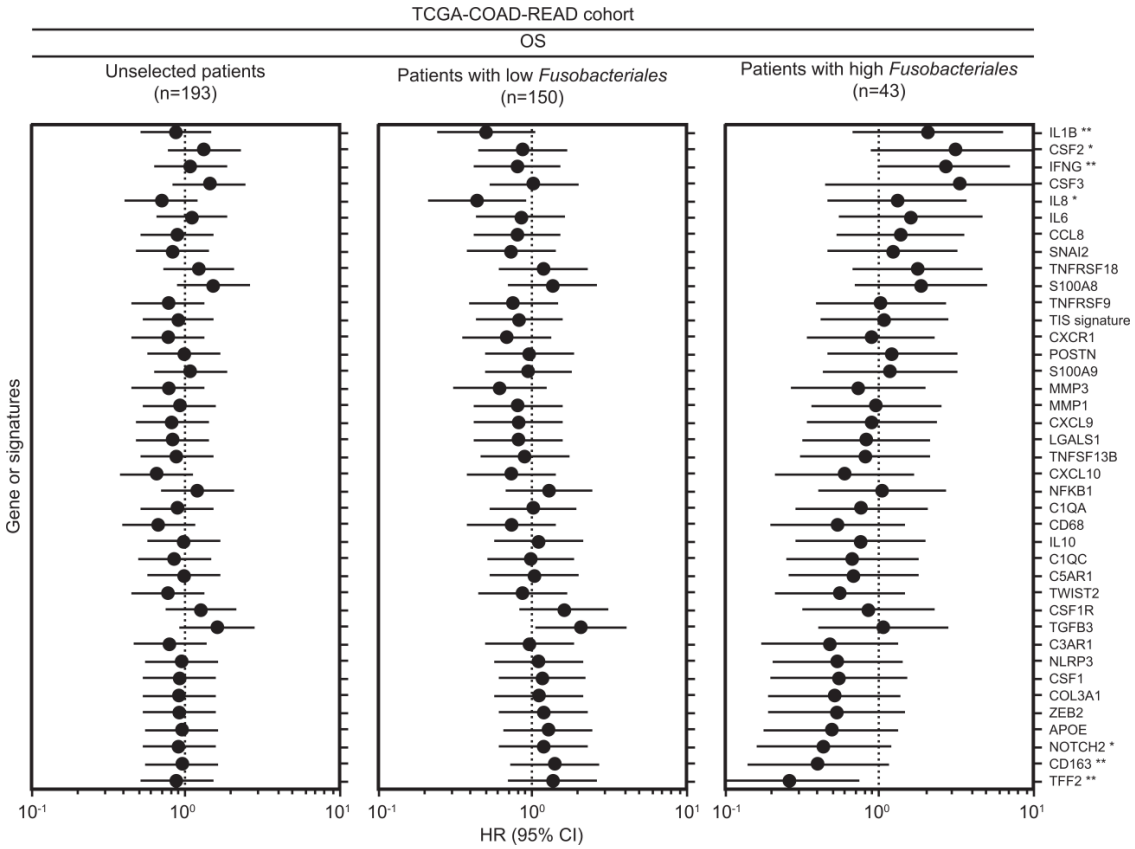


Cox regression models fitted on bacterium relative abundance reported at the order, family, genus, and species taxonomic ranks. For each taxonomic rank, patients were binned into -low or -high subgroups using the corresponding 75th percentile RA as cut-off. Univariate Cox regression models were fitted when evaluating association between bacterium subgroup (high vs. low; reference low) at each taxonomic rank and OS in the whole unselected patient population (left panel). Cox regression models with an interaction term between bacterium subgroup (high vs. low; reference low) and mesenchymal status (mesenchymal, i.e either CMS4 and/or CRIS-B, vs. non-mesenchymal, i. e. neither CMS4 nor CRIS-B) at each taxonomic rank and OS were fitted to evaluate differential impact of bacterium on clinical outcome by tumour biology (right panels).

Salvucci et al.

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Supplementary Figure 12.



Cox regression models fitted on patients of the TCGA-COAD-READ cohort with mesenchymal tumours (either CMS4 and/or CRIS-B) for each gene/signature identified from analysis presented in **Fig. 6A**. Patients were classified as *Fusobacteriales*-low or high using the corresponding 75th percentile relative abundance (RA) as cut-off. Univariate Cox regression models were fitted when evaluating association between *Fusobacteriales* (high vs. low; reference low) and OS in the whole unselected patient population (left panel). Cox regression models with an interaction term between *Fusobacteriales* (high vs. low; reference low) and gene/signature (high vs. low, reference low) and OS were fitted to evaluate differential impact of gene/signature on clinical outcome by

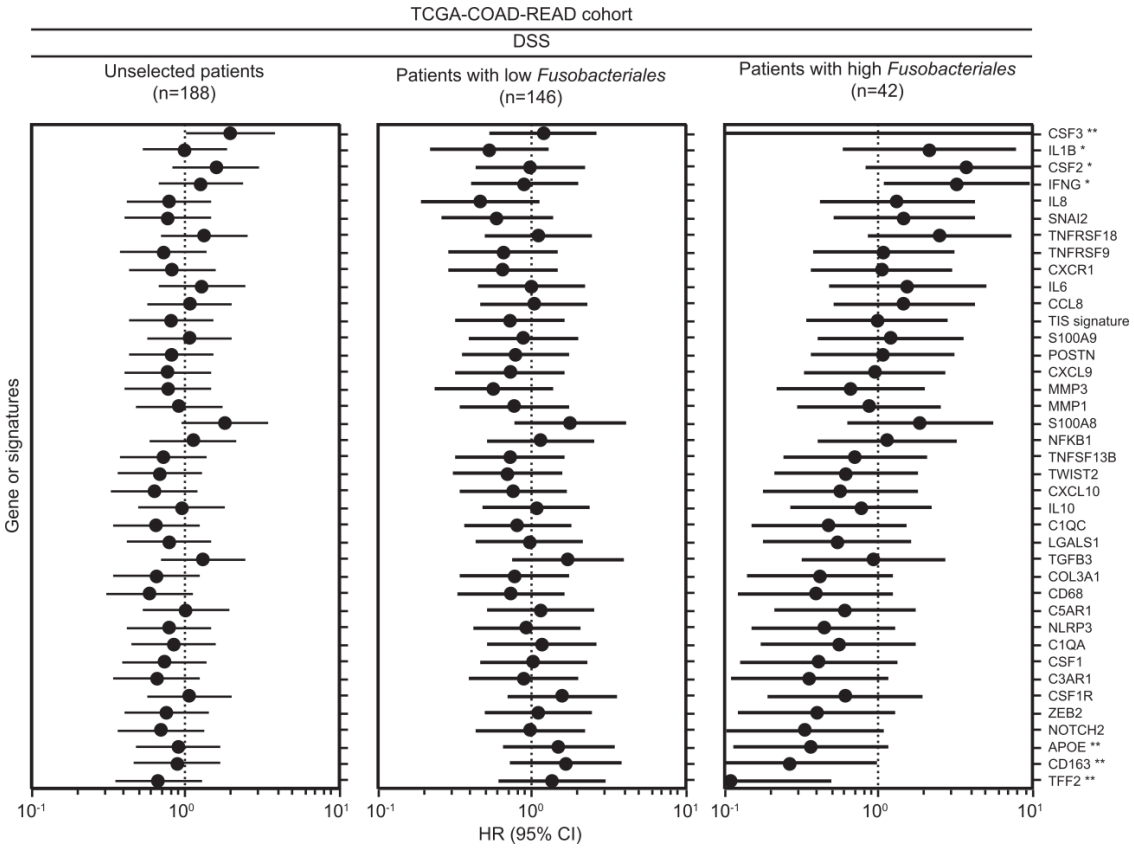
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Fusobacteriales (right panels). * and ** denote interaction P-values lower than 0.05 and lower than or equal to 0.1, respectively.

Salvucci et al.

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Supplementary Figure 13.



Cox regression models fitted on patients of the TCGA-COAD-READ cohort with mesenchymal tumours (either CMS4 and/or CRIS-B) for each gene/signature identified from analysis presented in **Fig. 6A**. Patients were classified as *Fusobacteriales*-low or high using the corresponding 75th percentile relative abundance (RA) as cut-off. Univariate Cox regression models were fitted when evaluating association between *Fusobacteriales* (high vs. low; reference low) and DSS in the whole unselected patient population (left panel). Cox regression models with an interaction term between *Fusobacteriales* (high vs. low; reference low) and gene/signature (high vs. low, reference low) and DSS were fitted to evaluate differential impact of gene/signature on clinical outcome by

Salvucci et al.

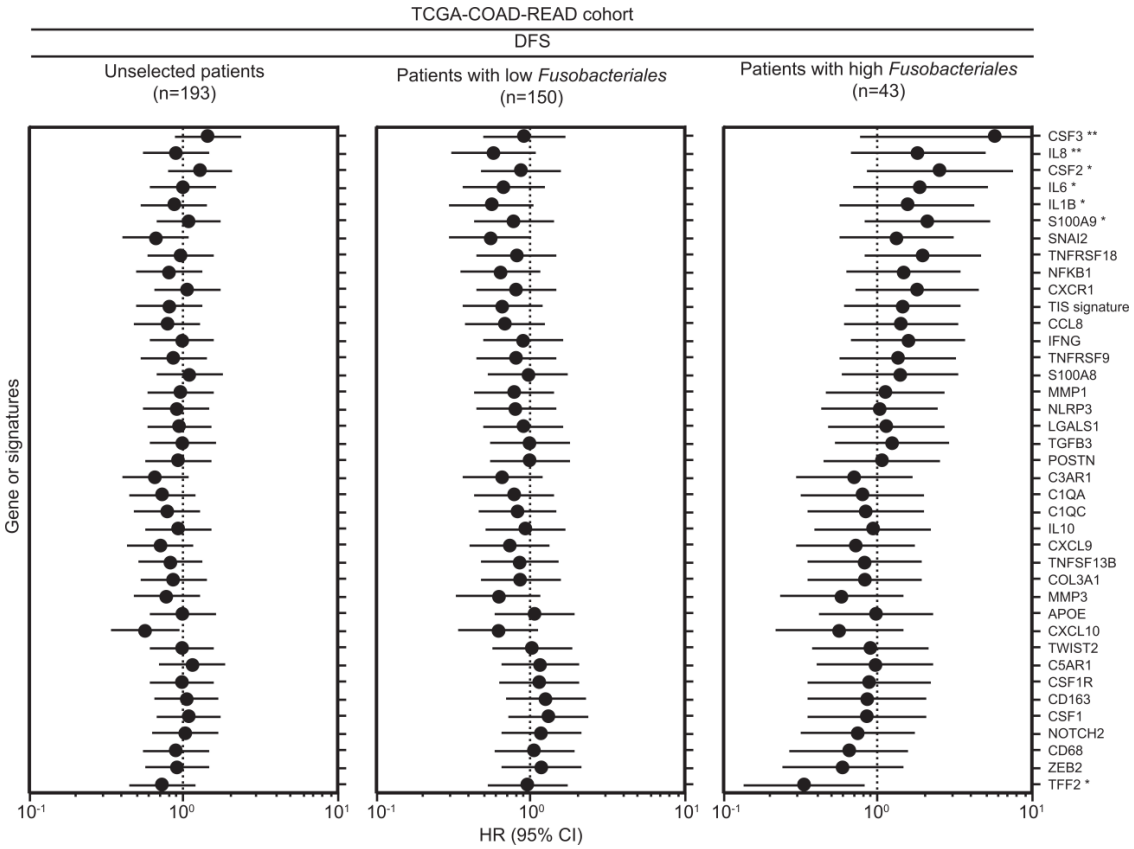
Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Fusobacteriales (right panels). * and ** denote interaction P-values lower than 0.05 and lower than or equal to 0.1, respectively.

Salvucci et al.

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Supplementary Figure 14.



Cox regression models fitted on patients of the TCGA-COAD-READ cohort with mesenchymal tumours (either CMS4 and/or CRIS-B) for each gene/signature identified from analysis presented in **Fig. 6A**. Patients were classified as *Fusobacteriales*-low or high using the corresponding 75th percentile relative abundance (RA) as cut-off. Univariate Cox regression models were fitted when evaluating association between *Fusobacteriales* (high vs. low; reference low) and DFS in the whole unselected patient population (left panel). Cox regression models with an interaction term between *Fusobacteriales* (high vs. low; reference low) and gene/signature (high vs. low, reference low) and DFS were fitted to evaluate differential impact of gene/signature on clinical outcome by

Transcriptomic-dependent *Fn/Fusobacteriales* impact.

Fusobacteriales (right panels). * and ** denote interaction P-values lower than 0.05 and lower than or equal to 0.1, respectively.

Salvucci et al.