**Supplementary material**:

**Detailed methods:**

**SCREEN-RA cohort**

SCREEN-RA is an ongoing Swiss multicentre cohort study of RA-FDRs1, comprised of subjects without a diagnosis of RA or another inflammatory rheumatic conditions at inclusion. RA-FDRs answer a questionnaire about potential environmental risk factors and are examined by a rheumatologist or a specialized study nurse to rule out the presence of RA, other autoimmune conditions, and possible tender or swollen joints. Blood samples are collected for genetic testing and autoantibodies (RF, ACPA) assessment. RA-FDRs are followed annually to assess for the development of signs or symptoms of arthritis.2

**Definition of ‘systemic autoimmunity associated with RA’**

The pre-clinical RA stage of ‘systemic autoimmunity associated with RA’ was defined by the presence of ACPA positivity and/or RF positivity. ACPA-positivity was operationally characterized by a positive result to any of the anti-cyclic citrullinated peptide antibodies tests (anti-CCP 2.0, 3.0 or 3.1). Autoantibodies were measured using standard, commercially available ELISA kits (anti-CCP 2 (CCPlus® Immunoscan, Eurodiagnostica), anti-CCP 3.1 (QUANTA Lite® CCP3.1 IgG/IgA, Inova Diagnostics) or anti-CCP 3 (QUANTA Lite® CCP3 IgG). RF positivity was characterized by a positive result to using the QUANTA Lite IgM and IgA® ELISAs and QUANTA Flash® IgM and IgA chemiluminescent immunoassays (Inova Diagnostics). ACPA and RF positivity were defined positive test according to the manufacturers' cut-off values (Anti-CCP2 ≥ 25 U/mL, anti-CCP3.1 and 3 ≥ 20 U/mL, RF QUANTA Lite ≥ 6 U/mL and RF QUANTA Flash ≥ 10344 RLU for IgM and ≥ 7425 for IgA). The overall prevalence of systemic autoimmunity in the SCREEN-RA cohort is around 20%. 2 The presence of shared epitope was determined by reverse PCR-SSOP hybridization (Luminex technology) and by PCR-SSP using commercial reagents validated by the National Reference Laboratory for Histocompatibility.

**Definition of ‘individuals with symptoms and signs associated with possible RA’**

We defined a priori ‘individuals with symptoms and signs associated with possible RA’ as having either symptoms associated with possible RA based on the Connective Tissue Disease Screening Questionnaire (CSQ)3 or as having undifferentiated arthritis (UA). The CSQ is a validated patient reported outcome. We used only the RA-related questions of the CSQ.3 The CSQ includes 6 RA-relevant items: morning stiffness, arthritis in hand joints or wrists, arthritis in 3 or more joint areas, symmetric arthritis, subcutaneous nodules and RF test results. The presence of 3 positive responses or more was considered to represent possible RA and 4 positive probable RA.4 The sensitivity and specificity of the CSQ to predict future RA has been reported as 0.77 and 0.94 respectively.4 The CSQ has been validated in a rheumatology clinic and in a general medicine (primary care) setting and used in research by more than 20 different groups and translated in several languages.3 Participants then may be classified as symptomatic irrespective of an arthritic joint on examination.  UA was defined by the presence of one or more swollen joints on examination and arthralgias. All participants were examined by either a trained rheumatologist or an experienced study nurse, who performed a 28 joint exam and recorded tender joint counts and swollen joint counts.

‘Individuals with symptoms and signs associated with possible RA’, could be ACPA positive and/or RF positive or not.5 The overall prevalence of ‘individuals with symptoms associated with possible RA’ in the whole cohort is about 20%.

**DNA Isolation**

The DNA Genotek OMNIgene∙Gut Stool Microbiome Kit was used to collect, store and ship the stool samples. DNA was isolated using an established protocol.6 Briefly, each sample was treated with 500µl of extraction buffer (200 mM Tris, 20mM EDTA, 200mM NaCl, pH 8.0), 200µl of 20% SDS, 500µl of phenol:chloroform:isoamyl alcohol (24:24:1) and 100µl of zirconia/silica beads (0.1 mm diameter). Samples were homogenized twice with a bead beater (BioSpec) for 2 min. After precipitation of DNA, crude DNA extracts were resuspended in TE Buffer with 1mg/ml RNase I and column purified to remove PCR inhibitors.

**16S rRNA gene amplification and sequencing**

Amplification of the V4 region (F515/R806) of the 16S rRNA gene was performed according to previously described protocols. 7 Briefly, 25 ng of DNA were used per PCR reaction (30 µl). The PCR amplification was performed using Q5 polymerase (NEB Biolabs). The PCR conditions consisted of initial denaturation for 30s at 98°C, followed by 25 cycles (10s at 98°C, 20s at 55°C, and 20s at 72°C). Each sample was amplified in triplicates and subsequently pooled. After normalization PCR amplicons were sequenced on an Illumina MiSeq platform (PE250).

**16S rRNA analysis**

Obtained reads were assembled, quality controlled and clustered using Usearch8.1 software package (http://www.drive5.com/usearch/). Briefly, reads were merged using -fastq\_mergepairs –with fastq\_maxdiffs 30 and quality filtering was done with fastq\_filter (-fastq\_maxee 1), minimum read length 200 bp. The OTU clusters and representative sequences were determined using the UPARSE algorithm8, followed by taxonomy assignment using the HITdb - Human Intestinal 16S rRNA gene reference taxonomy 9 and the RDP Classifier 10 with a bootstrap confidence cut-off of 80%. The OTU absolute abundance table and mapping file were used for statistical analyses and data visualization in the R statistical programming environment (R Core Team (2016) package phyloseq (McMurdie and Holmes, 2013). 14

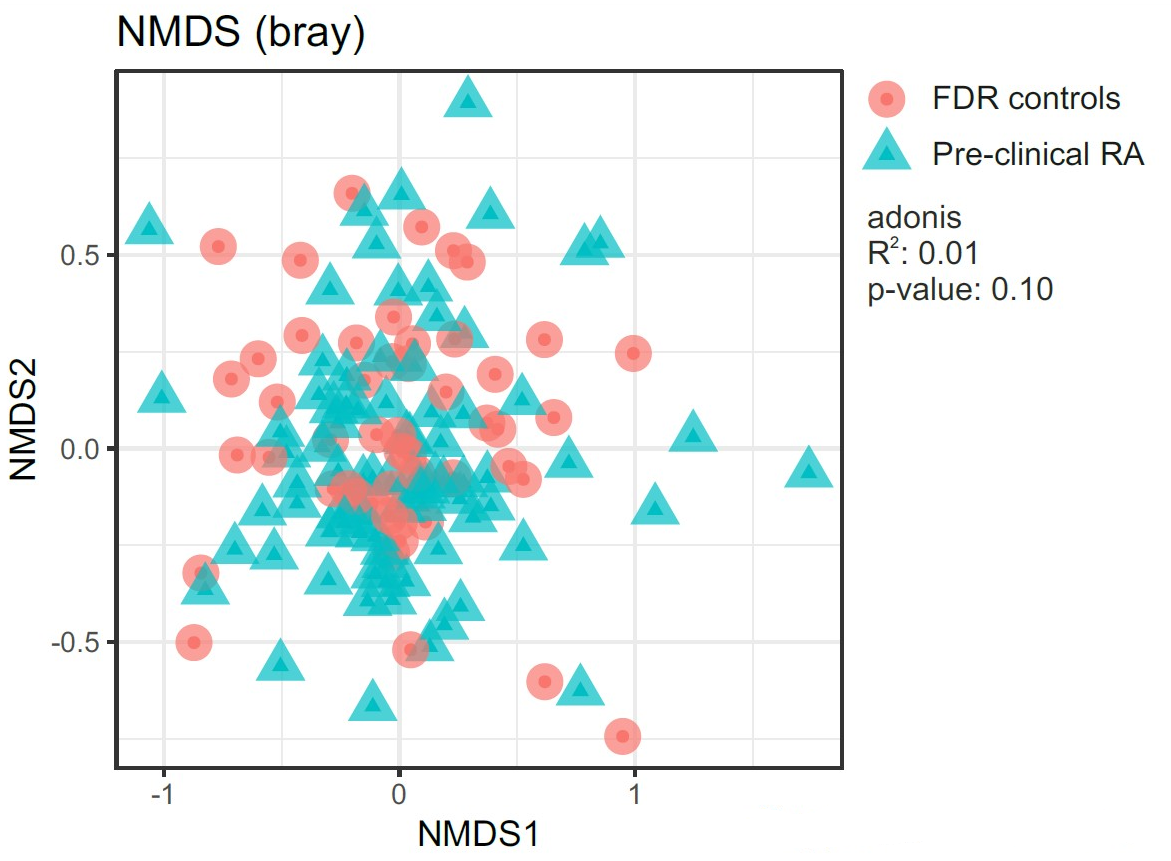
For the analysis of alpha and beta diversity we performed data normalization (relative abundance) and filtering to exclude low abundance OTUs (<0.05%). For the alpha-diversity testing we performed rarefaction to 10,000 reads.

**LEfSe Analysis**

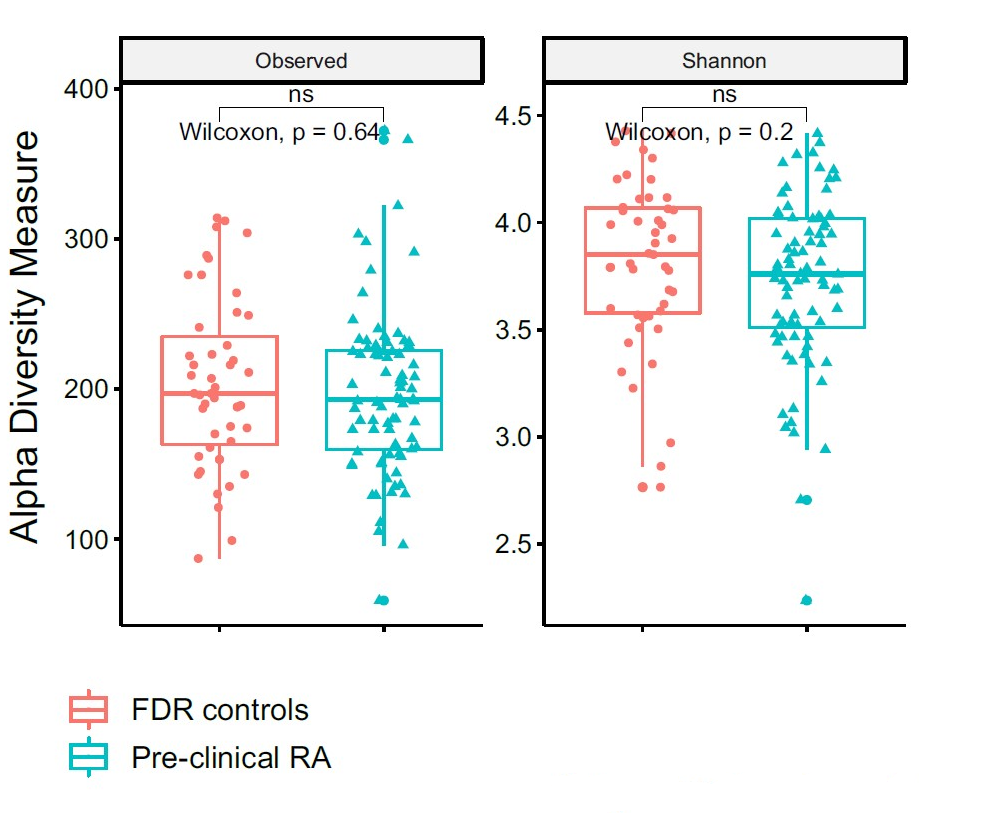
For linear discriminant analysis (LDA), we merged for each taxonomic level OTUs according to the corresponding level (reads merged by QIIME 12 from phylum to OTU level) and removed low abundant taxa (< 0.5 %). We applied LEfSe 13 with default parameters (c 1 (groupRA) -u 2 (SampleID) -o 1000000 (normalization value), alpha value for the Anova test (default 0.05), alpha value for the Wilcoxon test (default 0.05), threshold on the absolute value of the logarithmic LDA score) to identify taxonomic biomarkers that characterize the differences between FDR controls and pre-clinical RA group. To account for multiple comparisons and secondary analyses, we also performed Bonferroni adjustments of the p-values which are reported in the figures. As the statistical power was computed only for the primary analysis, when adjusted for multiple comparisons, most of the differences were no longer significant.

**Generation of taxonomic tree of Prevotella OTUs**

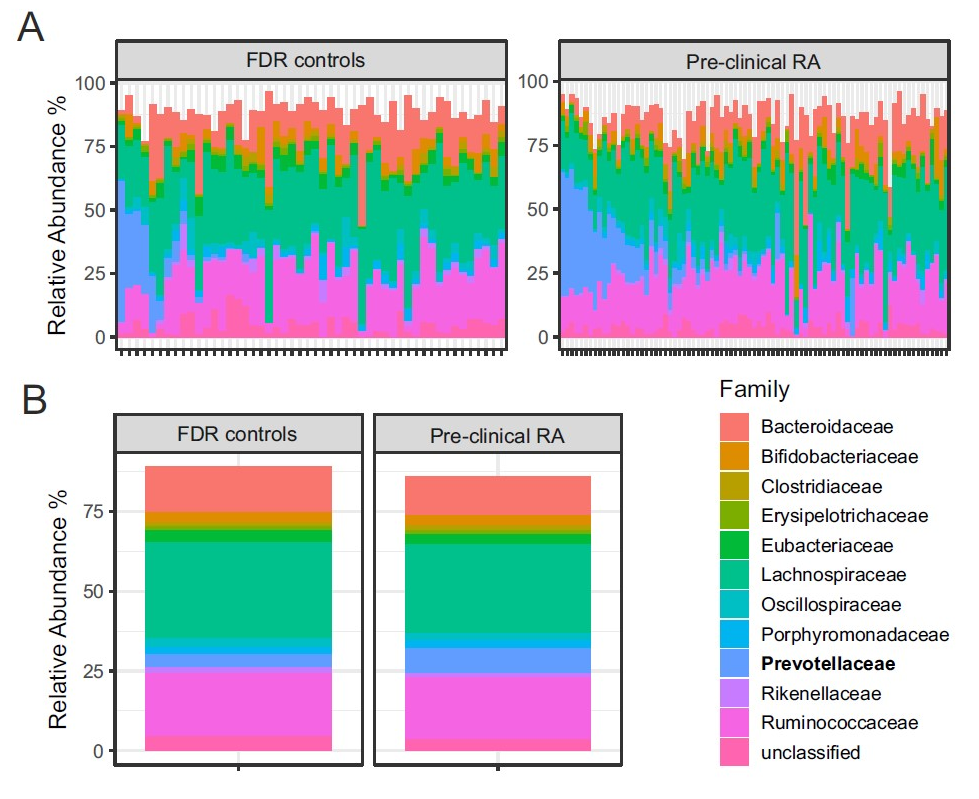
Prevotella operational taxonomic units (OTUs) with a minimum relative abundance of 0.5 % in one sample and a prevalence of larger than 1% in all samples, i.e. detected in at least 2 samples, were used to generate a neighbor joining phylogenetic tree.11 14 The closest reference OTU from the HITdb was determined for each OTU.



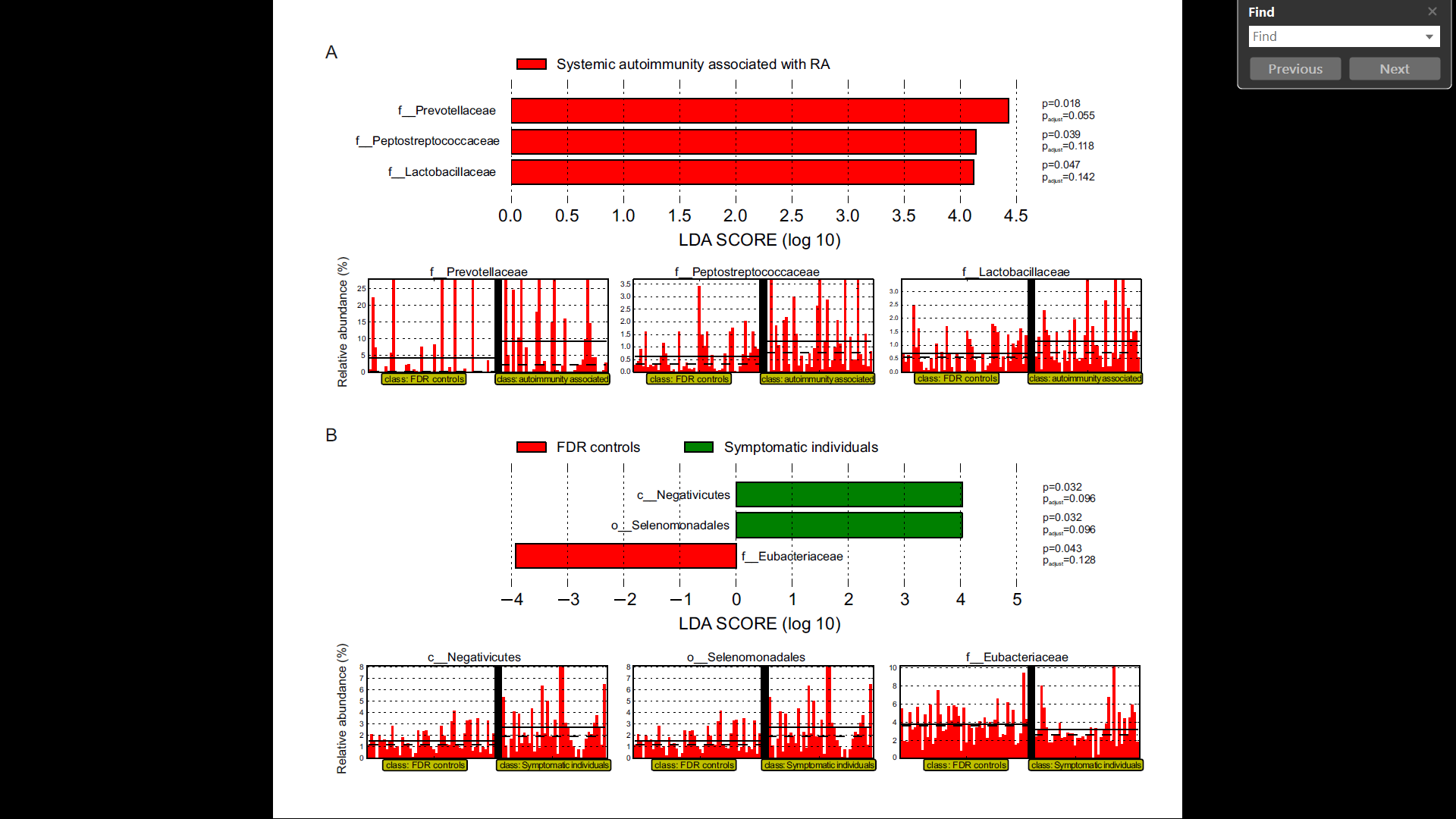
**Supplementary Figure S1: -diversity analysis of microbiota composition in patient cohort.** NMDS ordination analysis of microbiota composition using Bray-Curtis distances. FDR controls are represented as red dots and Pre-clinical RA subjects are represented as blue triangles. Permutational multivariate analysis of variance (ADONIS) was used to calculate the variance explained by association to the two groups.



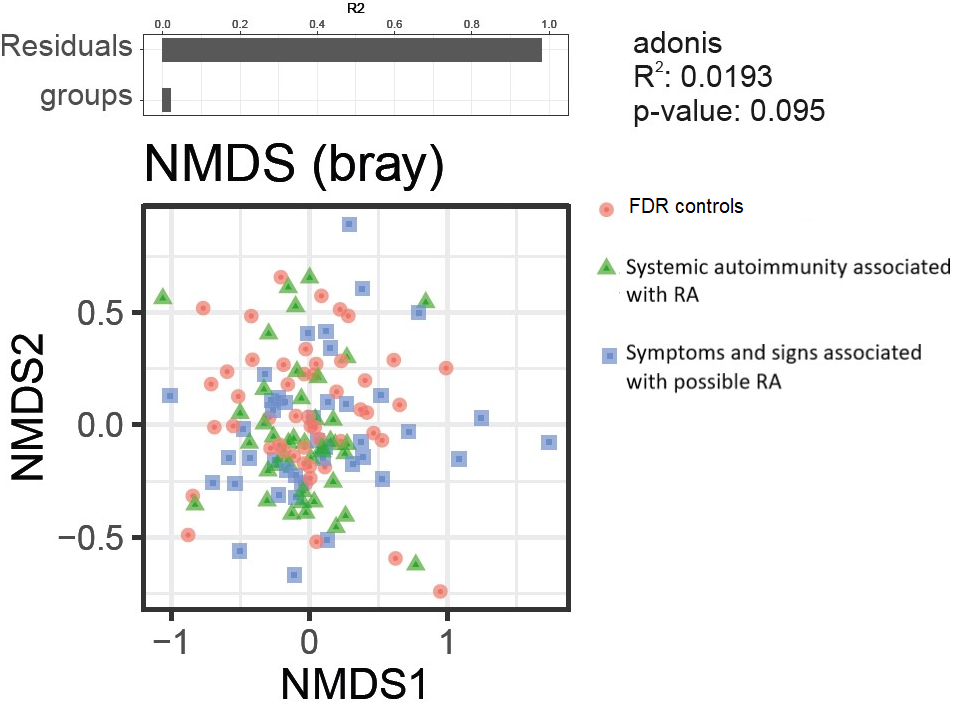
**Supplementary Figure S2: a-diversity analysis of microbiota composition in subjects cohort.** Estimation of α-diversity using Richness (Observed) and Shannon index. Box plots displaying FDR controls (red) and Pre-clinical RA participants (blue). An unpaired two-samples Wilcoxon test (non-parametric) was used to compare the two independent groups.

~~~~

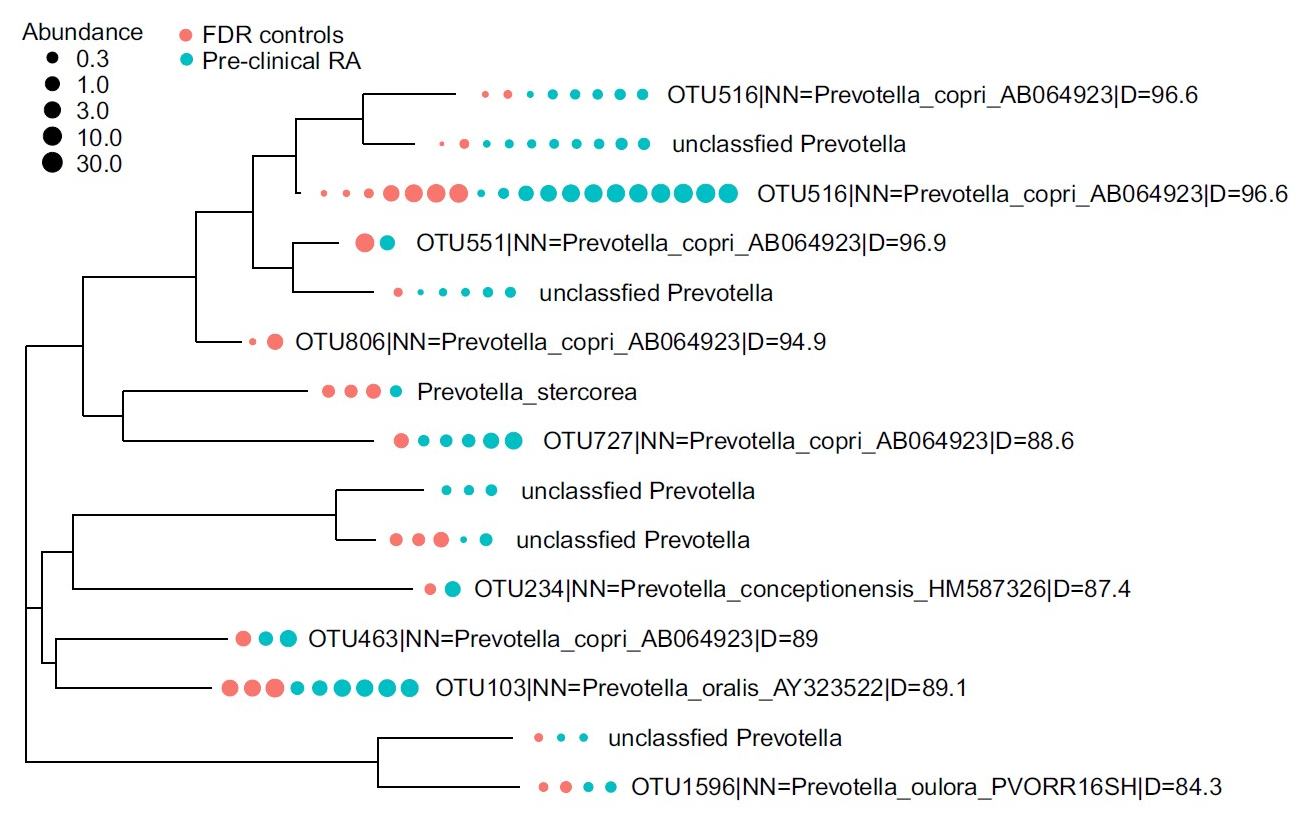
**Supplementary Figure S3. Microbiota composition analysis on family level**A. Relative abundance of bacteria grouped in families in FDR controls (left) and Pre-clinical RA individuals (right). The samples are ordered by decreasing cumulative relative abundance of OTUs assigned to the taxonomic level of species (see Figure 2). B. Relative abundance of bacteria grouped in families averaged over all FDR controls (left) and Pre-clinical RA individuals (right). Only the 12 families with the highest average relative abundance are displayed in A and B.



**Supplementary Figures S4. Linear discriminant analysis (LDA) effect size (LEfSe) evaluates the relative abundance of bacteria. Split analyses of the ‘pre-clinical stages of RA’ group in two separate groups.** A. Comparison of ‘FDR controls’ versus participants with ‘systemic autoimmunity associated with RA’. B. ‘FDR controls’ versus ‘individuals with symptoms and signs associated with possible RA’, with-or without autoimmunity. The thick horizontal dashed line in the lower panels shows median relative abundance and the solid line indicates the mean relative abundances. Padjust: p-values with Bonferroni adjustment. No significant differences in the relative abundance of bacteria were detected between ‘individuals with systemic autoimmunity associated with RA’ versus ‘individuals with symptoms and signs associated with possible RA’ (graphics not shown).

****

**Supplementary Figure S5:** **-diversity analysis of microbiota composition in patient cohort.** NMDS ordination analysis of microbiota composition using Bray-Curtis distances. ‘FDR controls’ are represented as red dots, ‘Systemic autoimmunity associated with RA’ are represented as green triangles and ‘individuals with symptoms and signs associated with possible RA’ in blue squares. Permutational multivariate analysis of variance (ADONIS) was used to calculate the variance explained by association to the three groups.

****

**Supplementary Figure S6. Tree of different *Prevotella* species detected in the dataset indicating relative abundance in samples in the different groups.** Prevotella operational taxonomic units (OTUs) with a minimum relative abundance of 0.5 % in one sample and a prevalence of larger than 1% in all samples were used to generate a phylogenetic tree. The presence of each Prevotella OTU in FDR controls and Pre-clinical RA participants is displayed as dots. Each dot represents one individual with a relative abundance of larger than 0.5 % and the diameter of each dot represents the relative abundance. The closest reference OTU from the HITdb is listed for each OTU (right). The “D=” indicates the sequence similarity between them.

**Supplementary Table S1. General characteristics at stool collection of participants divided in three groups: FDR controls, FDR with systemic autoimmunity associated with RA and FDR with symptoms and signs associated with possible RA (n=133 participants)**

|  |  |  |  |
| --- | --- | --- | --- |
| Characteristics | FDR  controls  n=50 | Systemic autoimmunity associated with RA a  n=42 | Symptoms and signs associated with possible RAb  n=41 |
| Age [years], median (IQR) | 55 (47-62) | 55 (41-65) | 60 (51-67) |
| Female sex, n (%) | 39 (78) | 37 (88) | 37 (90) |
| Current Smoking, n (%) | 11 (22) | 7 (17) | 9 (22) |
| Current Alcohol, n (%) | 22 (47) | 14 (44) | 15 (39) |
| Body mass index, median (IQR) | 24 (22-27) | 24 (22-25) | 25 (22-28) |
| Swollen joints at examination, median (IQR) | 0 (0-1) | 0 (0-1) | 2 (0-4) \* |
| Tender joints at examination, median (IQR) | 0 (0-1) | 0 (0-1) | 2 (0-5) \* |
| ACPA positivity, n (%) | 0 (0) | 27 (64) \* | 11 (27) |
| Anti-CCP 2.0 c | 0 (0) | 11 (26) \* | 3 (7) |
| Anti- CCP 3.1 | 0 (0) | 17 (45) \* | 7 (19) |
| Anti CCP 3.0 | 0 (0) | 9 (22) \* | 6 (14) |
| RF positivity, n (%) | 0 (0) | 19 (46) \* | 9 (22) |
| RF IgM c | 0 (0) | 17 (44) \* | 7 (18) |
| RF IgA | 0 (0) | 7 (18) \* | 3 (8) |
| Shared epitope (1 or 2 copies), n (%) | 32 (65) | 20 (50) | 22 (55) |
| \*p-value <0.05, Kruskal-Wallis test for continuous variables and Fisher’s exact test for categorical variables. aACPA or RF positivity. b ‘Individuals with symptoms and signs associated with possible RA’ as defined by the Connective Tissue Disease Screening Questionnaire (CSQ) with or without undifferentiated arthritis (UA). An isolated asymptomatic swollen joint was not sufficient to be classified in the group ‘symptoms and signs associated with possible RA’. c Percentages of total individuals by group. | | | |

**Supplementary Table S2.** **General characteristics of participants at stool collection categorized by Prevotellaceae abundance**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Characteristics** | **Abundance < 1%**  **n=74** | **Abundance >1%**  **n=59** | **Univariable**  **Analysis**  **OR(95%CI)** | **Adjusted by age**  **OR(95%CI)** |
| Age [years], median (IQR) | 57 (48-66) | 57 (50-63) | 0.9 (0.9-1.0) | - |
| Female sex, n (%) | 66 (89) | 47 (80) | 0.5 (0.2-1.3) | 0.5 (0.2-1.3) |
| Current Smoking, n (%) | 16 (22) | 11 (19) | 0.8 (0.4-1.9) | 0.9 (0.9-1.0) |
| Current Alcohol, n (%) | 33 (49) | 18 (37) | 0.6 (0.3-1.3) | 0.6 (0.3-1.3) |
| Body mass index, median (IQR) | 24 (22-27) | 25 (23-27) | 1.0 (0.9-1.1) | 1.0 (0.9-1.1) |
| Pre-clinical RA, n (%) | 39 (53) | 44 (75) \* | 2.6 (1.3-5.5) \* | 2.8 (1.3-5.9) \* |
| FDR controls c | 35 (47) \* | 15 (25) | 1 | 1 |
| Systemic autoimmunity associated with RA a | 18 (24) | 24 (41) \* | 3.1 (1.3-7.4) \* | 3.1 (1.3-7.4) \* |
| Symptoms associated with possible RA and/or UAb | 21 (29) | 20 (34) | 2.2 (0.9-5.2) | 2.4 (1.0-5.8) \* |
| Swollen joints at examination, median (IQR) | 0 (0-2) | 0 (0-1) | 0.9 (0.8-1.0) | 0.9 (0.8-1.0) |
| Tender joints at examination, median (IQR) | 0 (0-2) | 0 (0-2) | 0.9 (0.8-1.0) | 0.9 (0.8-1.0) |
| ACPA positivity, n (%) | 17 (23) | 21 (36) | 1.9 (0.9-3.9) | 1.9 (0.8-4.0) |
| RF positivity, n (%) | 9 (12) | 19 (33) \* | 3.4 (1.4-8.4) \* | 3.3 (1.4-8.1) \* |
| Shared epitope (1 or 2 copies), n (%) | 44 (62) | 30 (52) | 0.6 (0.3-1.3) | 0.6 (0.3-1.3) |
| Participants in this analysis are categorised based on the relative abundance of Prevotellacae (< 1%, ≥1%) irrespective of signs and symptoms or serologies. \*p-value <0.05, Kruskal-Wallis test for continuous variables and Fisher’s exact test for categorical variables. aACPA or RF positivity. b ‘Individuals with symptoms and signs associated with possible RA’ as defined by the Connective Tissue Disease Screening Questionnaire (CSQ) with or without undifferentiated arthritis (UA). An isolated asymptomatic swollen joint was not sufficient to be classified in the group ‘symptoms and signs associated with possible RA’. c Percentages of total individuals by group. | | | | |

**References:**

1. Finckh A, Müller RB, Möller B, et al. A novel screening strategy for preclinical rheumatoid arthritis (RA) in first degree relatives of patient with RA. Annals of the rheumatic diseases 2011;**70**:S3:282.

2. Alpizar-Rodriguez D, Brulhart L, Mueller RB, et al. The prevalence of anticitrullinated protein antibodies increases with age in healthy individuals at risk for rheumatoid arthritis. Clinical rheumatology 2017;**36**(3):677-82.

3. Karlson EW, Sanchez-Guerrero J, Wright EA, et al. A connective tissue disease screening questionnaire for population studies. Annals of epidemiology 1995;**5**(4):297-302.

4. Potter J, Odutola J, Gonzales CA, et al. Validation of English and Spanish-language versions of a screening questionnaire for rheumatoid arthritis in an underserved community. The Journal of rheumatology 2008;**35**(8):1545-9.

5. Gerlag DM, Raza K, van Baarsen LG, et al. EULAR recommendations for terminology and research in individuals at risk of rheumatoid arthritis: report from the Study Group for Risk Factors for Rheumatoid Arthritis. Annals of the rheumatic diseases 2012;**71**(5):638-41.

6. Thiemann S, Smit N, Roy U, et al. Enhancement of IFNgamma Production by Distinct Commensals Ameliorates Salmonella-Induced Disease. Cell host & microbe 2017;**21**(6):682-94 e5.

7. Caporaso JG, Lauber CL, Walters WA, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the National Academy of Sciences of the United States of America 2011;**108 Suppl 1**:4516-22.

8. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature methods 2013;**10**(10):996-8.

9. Ritari J, Salojarvi J, Lahti L, et al. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. BMC Genomics 2015;**16**:1056.

10. Wang Q, Garrity GM, Tiedje JM, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and environmental microbiology 2007;**73**(16):5261-7.

11. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PloS one 2013;**8**(4):e61217.

12. Kuczynski J, Stombaugh J, Walters WA, et al. Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr Protoc Microbiol 2012;**Chapter 1**:Unit 1E 5.

13. Segata N, Izard J, Waldron L, et al. Metagenomic biomarker discovery and explanation. Genome Biol 2011;**12**(6):R60.

14. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 2016;**33**(7):1870-4.