The gut microbiota and genetic risk for rheumatoid arthritis in the absence of disease: a TwinsUK association study

Philippa M Wells, BSc Hons.[1] Adewale S Adebayo, PhD[1] Ruth C E Bowyer, PhD[1] Maxim B Freidin, PhD[1] Prof Axel Finckh, PhD[2] Till Strowig, PhD[3,4] Till Robin Lesker, PhD[3] Deshire Alpizar-Rodriguez, PhD[2] Benoit Gilbert, MBBS[2] Bruce Kirkham, PhD[5] Prof Andrew P Cope, PhD[6,7] Claire J Steves, PhD[1,8 *] Prof Frances M K Williams, PhD[1 *]

1. Department of Twin Research and Genetic Epidemiology, King's College London, London, UK

2. Division of Rheumatology, Geneva University Hospital, Geneva, Switzerland

3. Department of Microbial Immune Regulation, Helmholtz Centre for Infection Research, Braunschweig, Germany

4. Hannover Medical School, Hannover, Germany

5. Department of Rheumatology, Guy's and St Thomas' NHS Trust, London, UK

6. Centre for Rheumatic Diseases, King's College London, London, UK

7. Centre for Inflammation Biology and Cancer Immunology, King's College London, London, UK

8. Department of Ageing and Health, St Thomas' Hospital, London, UK

*These authors contributed equally to this work.

Correspondence to:
Professor Frances M. K. Williams, The Department of Twin Research, King's College London, South Wing, St Thomas' Hospital, Westminster Bridge Road, London SE1 7EH, UK, Frances.Williams@kcl.ac.uk, 02071886743.

Accepted by Lancet Rheumatology on 05/03/2020

Summary

Background
Rheumatoid arthritis (RA) is a chronic inflammatory autoimmune disease associated with reduced life expectancy. It is heritable and an extensive repertoire of genetic variants have been identified. The gut microbiota may represent an environmental risk factor for RA. Indeed, *Prevotella copri* is a candidate keystone species, but whether it lies on the causal pathway for disease or is simply a bystander reflecting host-genetic predisposition to RA, remains to be determined. The study of disease-microbiota associations may be confounded by the presence of the disease of interest or by its treatment. To circumvent this issue, we assessed whether known RA risk alleles were associated with the gut microbiota, in a large population who do not have RA.

Methods
Blood and stool acquired from volunteers from TwinsUK were used for genotyping and assessment of the gut microbiota, respectively. A weighted polygenic risk score (PRS) for RA was calculated in 1650 unaffected twins from the TwinsUK cohort, based on 233 GWAS-identified published RA associated single nucleotide polymorphisms. Amplicon sequence variants were generated from 16S rRNA sequencing of stool samples and assessed for association with RA PRS. Confirmation of findings was performed using an independent sample comprised of first-degree relatives of RA patients from the SCREEN-RA cohort (n=133).

Findings
We found that *Prevotella spp* was positively associated with RA PRS in TwinsUK participants (q <1e-7). This finding was validated in SCREEN-RA participants carrying the shared epitope risk alleles (q = 1·12e-3). An association of *Prevotella spp.* with pre-clinical RA phases was also demonstrated (q = 0·021).

Interpretation
Prevotella in the gut microbiota is associated with RA genotype in the absence of RA, as well as in subjects at high risk of developing RA. This work suggests that host genotype is associated with microbiota profile prior to disease onset.

Funding
Versus Arthritis

Key Words
Prevotella; rheumatoid arthritis; host genotype; microbiome; inflammation

Research in context

Evidence before this study
A literature search was conducted using Google Scholar. Search terms were broad and included “rheumatoid arthritis microbiome”; “rheumatoid arthritis Prevotella”; “microbiome heritability”; “microbiome, genetics rheumatoid arthritis”; “rheumatoid arthritis genetic aetiology”. The reference list of identified papers were further used to identify relevant literature. *Prevotella copri* has been shown to be increased within the gut microbiota of rheumatoid arthritis (RA) patients, predominantly those with early disease, before treatment is initiated. Prevotellaceae (family) has also been demonstrated to be increased in pre-clinical RA cases compared to controls. *Prevotella copri* is posited to be an inflammatory driver, contributing to RA pathology by promoting a pro-inflammatory cytokine milieu.

Added value of this study
This study is the first to demonstrate that, in a large human cohort, carrying genetic risk factors for RA is associated with a higher abundance of *Prevotella spp.* in the absence of any form of RA.

Implications of all the available evidence
Within the gut microbiota, *Prevotella spp.* including, but not limited to *Prevotella copri* are of interest in the pathogenesis of rheuamatoid arthritis. This work suggests that any potential causal role of *Prevotella spp.* occurs very early in disease development.

Introduction

Rheumatoid arthritis (RA) is a debilitating chronic autoimmune condition, associated with reduced life expectancy. There is a substantial genetic component to RA aetiology, with heritability estimated at 65%.[1] Known environmental risk factors include periodontal disease, tobacco smoking and diet, and appear to trigger disease onset in genetically susceptible individuals.[2] More recently proposed RA risk factors include the mucosal commensal microbiota. There is extensive cross-talk between the microbiota and the host, starting early in life with the development of the normal immune system; the microbiota may be implicated in the development of RA.

The gut lumen holds the vast majority of the commensal microbiota, and has intimate proximity to both the immune system via the gut associated lymphoid tissue, and the systemic circulation. A key gut microbiota taxonomic association demonstrated in RA patients, which presents early in the course of RA is a relative increase in the abundance of *Prevotella spp*,[3] and particularly *Prevotella copri* (*P.copri*).[4,5] Further evidence supporting the pathophysiological relevance of *P.copri* in RA is provided by a positive correlation with clinical parameters in RA patients.[5] In addition to promoting disease activity, the gut microbiota may also influence patient response to treatment.[5,6] The gut microbiota therefore represents a potential therapeutic target in RA, both for modulation of disease and for improving response to established therapeutics.

Identifying how RA genetic and environmental risk factors interact with one another may shed light on the underlying biology of disease. The influence of host genetic factors on the microbiota in RA remains relatively unexplored. An important influence is highly plausible: host genetics shape the biochemical and immune environment in which the microbiota reside, and furthermore the cumulative influence of the genetic risk loci in RA is predominantly mediated by immune pathways.[2] Whilst there are a number of factors which influence microbiota composition, host genetic factors account for a considerable proportion of variance, with some taxa being 40% heritable.[7]

The purpose of this study was to investigate the influence of RA genetic risk factors on the gut microbiota. The use of a population sample TwinsUK,[8] whilst excluding RA affected participants allows for isolation of RA genetic factors and circumvents the important issues of confounding by RA disease and its treatment - thereby achieving a human model of the genetic association with the microbiota in RA. An RA polygenic risk score in unaffected TwinsUK participants was calculated and association determined with composition of the gut microbiota. Validation studies confirmed and extended our findings by examining the gut microbiota composition in first-degree relatives of RA patients (SCREEN-RA cohort) in relation to shared epitope positivity, and a pre-disease state as defined by clinical RA parameters.[9]

Methods

Study Overview

This study was conducted in the UK and Switwerland, and comprised participants of the TwinsUK (primary) and SCREEN-RA (validation) cohorts. We investigated whether genetic risk for RA was associated with the gut microbiota. In order to isolate genetic influence on disease from the pathophysiology and treatment influence of established disease, TwinsUK participants with RA and their co-twin were excluded from the study. 1650 eligible TwinsUK participants contributed blood and stool, which were used to determine genotype and gut microbiota composition, respectively. Genetic risk for RA in TwinsUK participants was captured using polygenic risk scoring. An RA polygenic risk score (PRS) was generated and validated using logistic regression against RA diagnosis in UK Biobank (UKBB) with 2,686 confirmed RA cases, and applied to RA unaffected TwinsUK participants. Association was then determined between RA PRS and composition of the gut microbiota in the TwinsUK cohort (n = 1650). Validation was performed in the SCREEN-RA cohort, comprised of RA unaffected first degree relatives of RA patients. First degree relatives of RA patients share on average 50% of their genotype with their RA affected relative, and thus have higher RA genetic risk than the general population. In addition, participants were genotyped for SE RA risk alleles – the strongest genetic association with RA, and assessed for pre-RA status as defined by clinical RA parameters.[9] The gut microbiota composition was examined in relation to SE genotype and pre-RA in the SCREEN-RA cohort (n=133; **Figure 1**).

Favourable ethical opinion was granted by the formerly known St. Thomas' Hospital Research Ethics Committee (REC). Following restructure and merging of REC, subsequent amendments were approved by the NRES Committee London - Westminster (TwinsUK, REC ref: EC04/015, 1 November 2011); use of microbiota samples was granted NRES Committee London - Westminster (The Flora Twin Study, REC ref: 12/LO/0227, 1 November 2011).

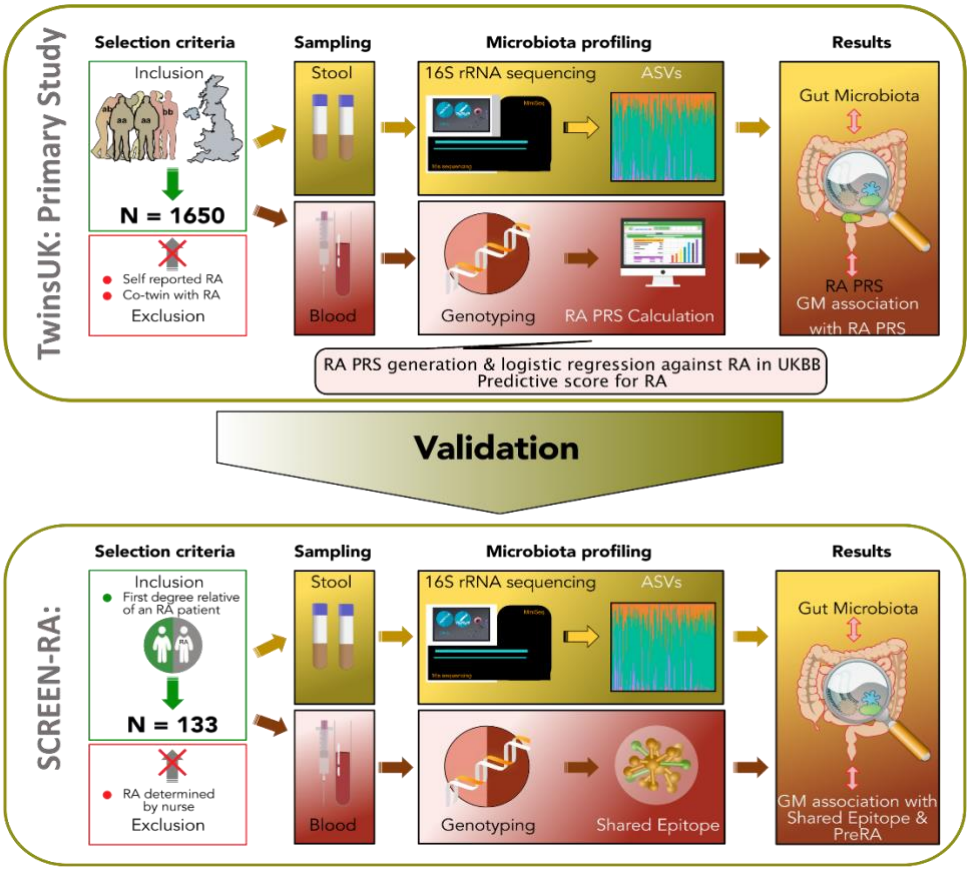

Figure 1. Schematic representation of the study design. The TwinsUK cohort is led by King's College London located at St Thomas Hospital NHS Foundation Trust, and is comprised of adult twins resident in the UK. SCREEN-RA is a Swiss multi-centre cohort comprised of first-degree relatives of RA patients.

Participants
Participants included in the study were members of the TwinsUK cohort, the largest UK registry of adult twins,[8] with the exception that participants reporting a diagnosis of RA (identified by the following multiple choice questionnaire: **"**Has a doctor ever told you that you have/had any of the following conditions? Rheumatoid Arthritis"), as well as their unaffected co-twin were excluded. The TwinsUK registry is demographically well suited to the study of RA - most participants are female and of more advanced age, as such study specific inclusion criteria were not necessary. TwinsUK participants comprised 93% females, having a median age of 63 (**Table 1**).

To validate the results from TwinsUK, an analysis of RA genetic predisposition and microbiota association was performed in participants of SCREEN-RA, a Swiss multicentre cohort of first-degree relatives of RA patients.[9] Within SCREEN-RA participants (n=133), the subset of pre-clinical RA cases (n = 83) were identified based on the European League Against Rheumatism terminology for pre-clinical phases of RA,[9] and matched with 50 controls who were also first degree relatives of RA patients. Briefly, pre-clinical RA was defined on the basis of serum positivity for anti-citrullinated protein antibody (ACPA) or rheumatoid factor and/or symptoms and signs associated with possible RA with or without undifferentiated arthritis.[9] Patients with a diagnosis of RA were excluded from the analysis and all participants had been examined by a trained nurse. Of SCREEN-RA participants, 84% were female and the median age was 57 (**Table 1**)**.**

Participants provided written informed consent.

		Age median (IQR)	Female Sex n(%)	BMI median (IQR)	ACPA +ve n(%)	RF +ve n(%)	SE +ve n(%)	ACPA & RF +ve n(%)	Current Smoker n(%)	Anti-biotics prior month n(%)	Ethnicity	Swollen joints median (IQR)	Tender joints median (IQR)
TwinsUK	**RA unaffected n = 1650**	63 (56-69)	1535 (93)	25 (23-29)	9* (2)	35* (7)	..	1 (0)	567 (34)	66 (4)	Caucasian; Northern European	..	NA
SCREEEN-RA	**Pre-RA n=83**	58 (50-66)	74 (89)	24 (22-27)	38 (46)	28 (34)	42 (53)	6 (5)	16 (13)	..	Caucasian; Northern European	1 (0-3)	1 (0-2)
	FDR controls n=50	55 (47-62)	39 (78)	24 (22-27)	0 (0)	0 (0)	32 (65)	0 (0)	11 (6)	..	Caucasian; Northern European	0 (0-1)	0 (0-1)

Table 1. Summary of participant demographics. Anti citrullinated protein antibody(ACPA) positivity is defined as >5ul/ml. Rheumatoid factor (RF) positivity is defined as >15ul/ml.
*****Serum samples analysed for 500 TwinsUK participants including those with highest PRS.

Genotyping: TwinsUKcohort
Blood samples from TwinsUK participants obtained at clinical visit were used to determine genotype using the Illumina HumanHap300 BeadChip and the Illumina HumanHap610 QuadChip. Non-genotyped variants were imputed using 1000 Genomes and Haplotype Reference Consortium (HRC) reference panels.[8]

Polygenic Risk Score for RA in TwinsUK
The polygenic risk score (PRS) assigns an individual a single numerical value for risk of disease conferred by genetic factors.[10] The NCBI database of GWAS summary statistics for RA was used to identify 233 published single nucleotide polymorphisms (SNPs) associated with RA at genome-wide significance (P = 5e-8) of which 117 had been replicated across studies (**Supplementary Table 3**).[11] A study inclusion criterion of European ancestry ensured ethnic concordance with TwinsUK.[8,12] The PRS was tested for RA predictive value in 6,776 participants from UK Biobank including 2,686 RA cases and 4,090 unselected controls with chronologically closest participant identification numbers. Diagnosis of RA in UK Biobank participants had been determined using hospital episode statistics (HES) data supplied by NHS Digital. All identified RA cases were included. No exclusion criteria were necessary: UKBB is ethnically representative of TwinsUK. Logistic regression of RA cases and unselected controls against PRS, adjusting for age, sex and smoking history, was applied. Standardised coefficients are reported.

Risk allele dosage of SNPs present within TwinsUK was extracted using Plink 1·9. Of the SNPs identified in the literature 227 were available in TwinsUK. Pruning was applied to account for linkage disequilibrium [13] Missing allele dosages were imputed and replaced with the mean value across the respective SNPs. The risk allele dosage was multiplied by the SNP-RA association effect size, to produce a weighted PRS.[10]

Microbiota profiling: TwinsUK

Microbiota composition of gut (stool, n=1650) samples was assessed using the 16S rRNA marker gene, with sequencing of the V4 variable region using barcoded primers (F515/R806). Samples were processed as previously described.[7] Briefly, faecal samples were collected during clinical visits or were posted in sealed ice packs and frozen on arrival at the lab at -80° C. Stool samples were sent as 35 batches on dry ice to Cornell University, where DNA was extracted and sequenced on an Illumina MiSeq platform.[7]

16S sequences were demultiplexed in QIIME. Amplicon sequence variants (ASVs) were generated using the DADA2 package in R.[14] Sequences were trimmed, error estimated within the forward and reverse reads for each sample, and the ASV algorithm applied to infer the original biological sequence. Forward and reverse reads were joined. Chimeras were removed and the total dataset merged, followed by taxonomic assignment with SILVA1.3.2.[15] Samples having a sequencing depth of less than 10,000 reads were excluded. A phylogenetic tree was generated using the Phangorn R package. Alpha diversity was calculated on untrimmed ASV tables using four measures: Shannon index, Simpson index, observed ASVs and Faith's phylogenetic diversity. For the taxonomic analysis, ASVs were grouped according to taxon annotation. Taxonomic assignment using the SILVA database allows for a higher level of differentiation as in some instances it is possible to annotate the genus of ASVs according to prediction of species group.[18] These annotations were preserved as they provide more information regarding taxon assignment than genus annotation alone. In this way, *Prevotella* annotated ASVs may be accurately further differentiated using 16S data, which has been a methodological challenge previously.

Statistical analysis: association of the gut microbiota alpha diversity with RA genetic risk

Linear mixed-effects models were used to determine association between RA PRS and alpha diversity, using alpha diversity as a response variable to the RA PRS as a continuous variable. Modelling was performed using the lme4 package in R,[16] with fixed effect covariates age, BMI and sequencing depth and technical covariates as random effects. Standardised coefficients were reported.

Statistical analysis: taxonomic association of the gut microbiota with RA genetic risk

Differential abundance of ASVs present in more than 5% of samples, grouped by ASV taxon annotation, against the RA PRS as a continuous variable was assessed using the DESeq2 R package,[17] using fixed-effects covariates as above. To account for multiple testing, the false discover rate (FDR) calculation was applied to all p values to generate q values, with a significance threshold of 0.05 determined a-priori.

Phylogenetic and community relationship within the Prevotellaceae Family

To investigate further the *Prevotella spp.* associations identified, the phylogenetic and community relationships between *Prevotella_7* and *Prevotella_9* were explored. For these methods, it was not appropriate to group ASVs by taxon annotation - ASVs were considered as singular units.

In order to examine the microbial ecological community relationships, ASVs were grouped into compositional clusters, or balances. Briefly, ASVs present in at least 10 percent of individuals were correlated and transformed as implemented in Morton et al.[19] This creates clades in which interacting species are closer neighbours in a clade than loosely related ones.

To examine the phylogenetic relationship, the phylogenetic tree was subsetted to Prevotellaceae and visualised using the Phyloseq R package.[20]

Shared Epitope Genotyping: SCREEN- RA cohort
Blood was taken from patients and controls. DNA was extracted using a modification of the salt-out technique (Nucleon TM, Scotlab, UK). HLA-DRB1 Shared Epitope polymorphism was determined by reverse polymerase chain reaction - sequence-specific oligonucleotide primers hybridization and by polymerase chain reaction - sequence-specific primers using commercial reagents validated at the Swiss National Reference Laboratory for Histocompatibility. Within the DRB1*04 genotype the method discriminates all major subtypes in different allele groups. HLA-DR1, DR4 and DR14 alleles that are negative for the SE70-74 motif are also discriminated. In a second step, the SE-positive typing ambiguities were analyzed by PCR-SSP in order to determine the final 4-digit typing result.

Microbiota Profiling: SCREEN-RA cohort
Gut microbiota 16S rRNA data was collected as previously described.[9] Briefly, the DNA Genotek OMNIgene·Gut Stool Microbiome Kit was used to collect, store and ship the stool samples. After sample processing and DNA extraction, the V4 region of the 16S rRNA gene was amplified using barcoded primers (F515/R806), and sequenced on an Illumina MiSeq, with ASVs generated as per the TwinsUK cohort to ensure compatibility.

Statistical Analysis: SCREEN-RA cohort
Differential abundance of taxa in the gut microbiota in association with pre-clinical RA and SE positivity was assessed against all genus present in greater than 5% of samples using the DESeq2 R package.[17] Biological covariates were not required to adjust for as these factors were not statistically different between the case and control groups (**Supplementary Table 1**). This was a repeat of the published study by Alpizaar-Rodriguez *et al.*, with the advancement that ASVs and DESeq2 methods were used.[9]

Role of the Funding Source
The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The following authors had access to the raw data: PW, AA, MF, RB, CS, FW. The corresponding authors had full access to all of the data and the final responsibility to submit for publication.

Results

TwinsUK participants comprised 1650 RA-unaffected adult twin volunteers the majority of whom were female. **Table 1** summarises the characteristics of the sample.

The polygenic risk score was normally distributed in the TwinsUK sample with RA cases and twin siblings excluded (Shapiro-Wilk normality test P = 0·24). Logistic regression of the PRS in 6,776 UK biobank participants, of which 2,686 had a HES diagnosis of RA, confirmed that the score is predictive of RA (odds ratio per standard deviation = 1·34; P = 4·17e-8).

High PRS individuals in TwinsUK did not demonstrate RA seropositivity: only 9/500 individuals in the TwinsUK sample having serum available were positive for anti-citrullinated protein antibody (ACPA; defined as >5u/mL) (**Table 1)**, and they were distributed similarly across genetic risk groups, by PRS quartile (high 2:93; low 7:398).

We determined whether genetic risk of RA is associated with alpha (within sample) diversity. There was no detectable association between RA PRS and Shannon Index (P = 0·76), Simpson Index (P = 0·41), Observed ASVs (P=0·6) or Faith's phylogenetic diversity (P=0·65; Supplementary **Table 2**). Alpha diversity measures indicate microbial density or "richness" (Observed ASVs) and distribution, or "evenness" in which the distribution of more abundant versus less abundant taxa is assessed, or both of these factors (Shannon Index, Simpson Index). Faith's phylogenetic diversity is based on the phylogenetic distance between taxa in a sample.

The gut microbiota were taxonomically assessed for association with RA PRS following a non-targeted approach. Of all 172 microbial taxa which were present in the gut (stool) microbiota of more than 5% of participants, *Prevotella_7* was the strongest taxon association with the RA PRS (18-fold log base 2 higher differential abundance; q < 1e-7) (**Figure 2)**. No additional *Prevotella* assocations with the RA PRS were demonstrated.

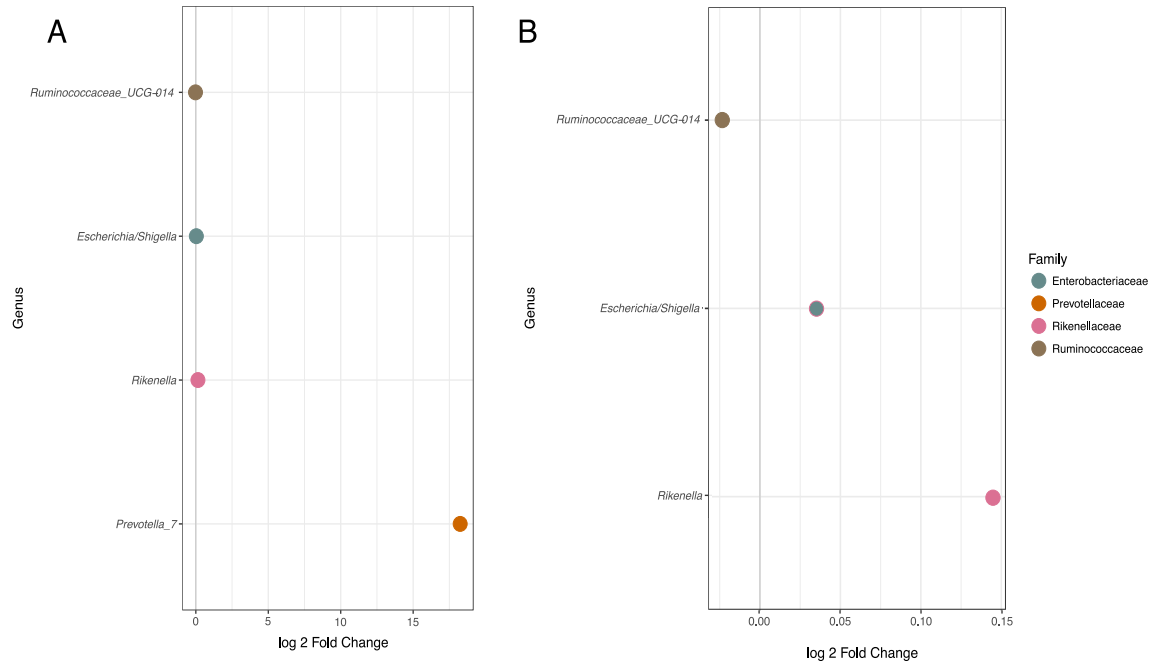

Figure 2. Differential abundance of the gut microbiota per unit increase in RA PRS in TwinsUK participants. Positive log-fold change of taxa indicates a positive association with the RA PRS. *Prevotella_7* was the strongest association, and positively associated with RA PRS (q < 1e-7). Due to the scale and comparative difference in log-fold change in *Prevotella*, this taxon is excluded in **B** to allow visualisation of the 3 other associations. Other taxa associations within the gut microbiota were Ruminococcaceae_UCG-14 (q = 0·045), *Rikenella* (q = 0·018) and *Shigella* (q = 0·018).

We sought to confirm our findings in gut microbiota from the SCREEN-RA cohort by analysing the association between the main genetic risk factor for RA - HLA-DRB1 shared epitope (SE), and the gut microbiota. As the cohort comprises first-degree relatives of RA patients it has a higher genetic risk for RA than the general population. Participants had also been genotyped for SE risk alleles.

*Prevotella*_9 was positively associated with pre-clinical RA (q = 0·021; **Figure 3**). *Prevotella_7* was associated with HLA-DRB1 SE RA risk alleles in the SCREEN-RA cohort (n=133; q = 0·035); the association was stronger in a subgroup analysis in which the 44 participants with swollen joints were removed, in order to isolate genotype from RA pathophysiology (n=89; q = 1·1e-3), **Figure 4**. In the subgroup analysis of differential abundance against SE positivity in asymptomatic participants only, *Prevotella_7* was the only remaining taxon association. This finding validates the TwinsUK *Prevotella_7* association.

Prevotella_7 and *Prevotella_9* are groups of *Prevotella* ASVs, according to predicted species. *Prevotella_9* is predicted *Prevotella_copri*, whereas *Prevotella_7* is annotated to multiple *Prevotella* species, with low sequence divergence. The key implication is that *Prevotella_9* and *Prevotella_7* are distinct from one another, and *Prevotella_9* is predicted to be *Prevotella copri.*[15,16]

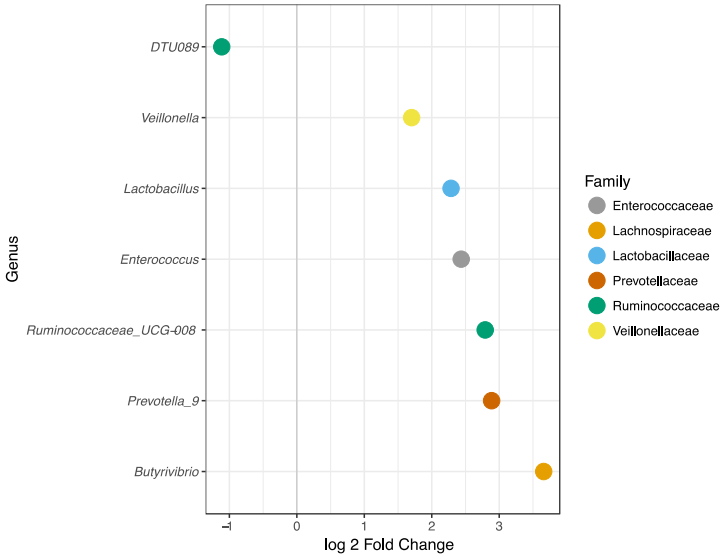

Figure 3. Differential abundance of the gut microbiota in pre-clinical RA cases compared to uaffected controls in the SCREEN-RA cohort. *Prevotella_9* was associated with pre-clinical RA (q = 0·021). There were 6 additional genus group associations: *Lactobacillus* (q = 2·6e-4), *Butyrivibrio* (q = 0·018), Ruminococcaceae_UCG-D08 (q = 0.018), *Enterococcus* (q =0·018), *DTU089* (q = 0·042), and *Veilllonella* (q = 0·042).

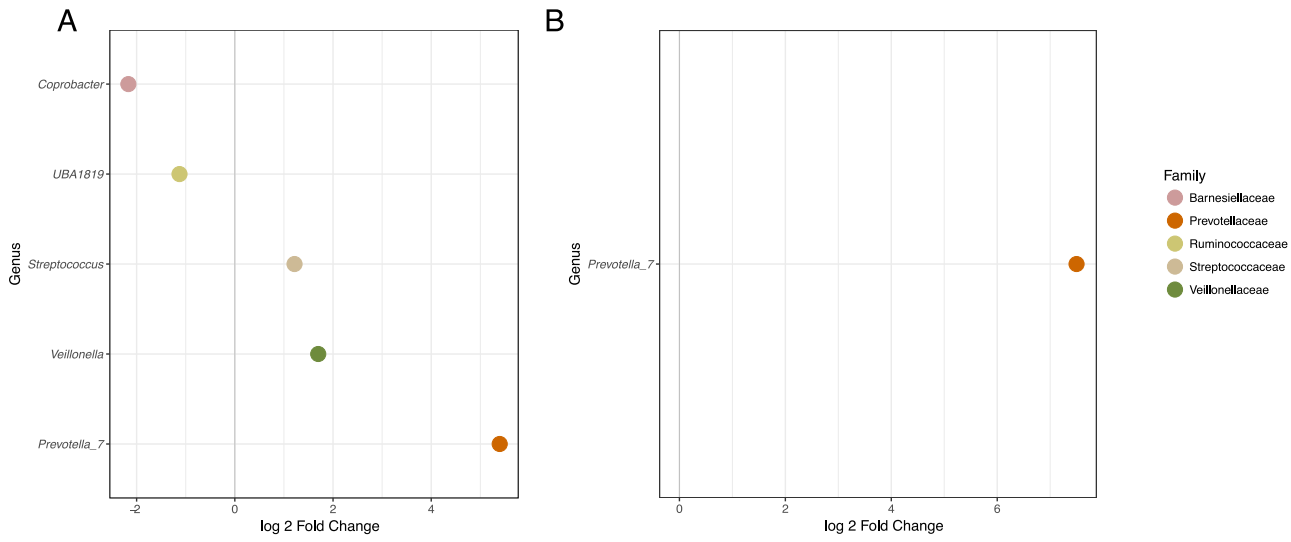

Figure 4. Differential abundance of the gut microbiota in HLA-DRB1 shared epitope positive participants, compared to shared epitope negative controls in the SCREEN-RA cohort. In **A** all participants were analysed. *Prevotella_7* showed the most substantial positive log fold change of 5 (q= 3·48e-2). Other genus associations were *Veillonella* (q = 3·48e-2), *Streptococcus* (q = 3·48e-2), Ruminococcaceaae UBA1819 (q = 3·48e-2) and *Coprobacter* (q = 2·03e-3). In **B** participants with RA associated symptoms – tender or swollen joints were excluded, here, *Prevotella_7* solely remained positively associated with shared epitope positivity (q = 1·12e-3).

As *Prevotella spp.* are the strongest taxon associations in our results and of particular interest in RA, we further investigated the phylogenetic and community relationships within Prevotellaceae. In particular, we were interested in the relationship between *Prevotella_7* and *Prevotella_9* as these were associated with genetic risk and pre-clinical RA, respectively. Cluster analysis demonstrated a community relationship between *Prevotella spp*. *Prevotella_7* and *Prevotella_9* were the only 2 species clusters within the Prevotellaceae family, and clustered together more frequently than any other group, suggesting an inter-dependent community relationship between these taxa (**Figure 5A**). Visualisation of the Prevotellaceae phylogenetic tree demonstrated that *Prevotella_7* and *Prevotella_9* are phylogenetically distinct from one another (**Figure 5B**).

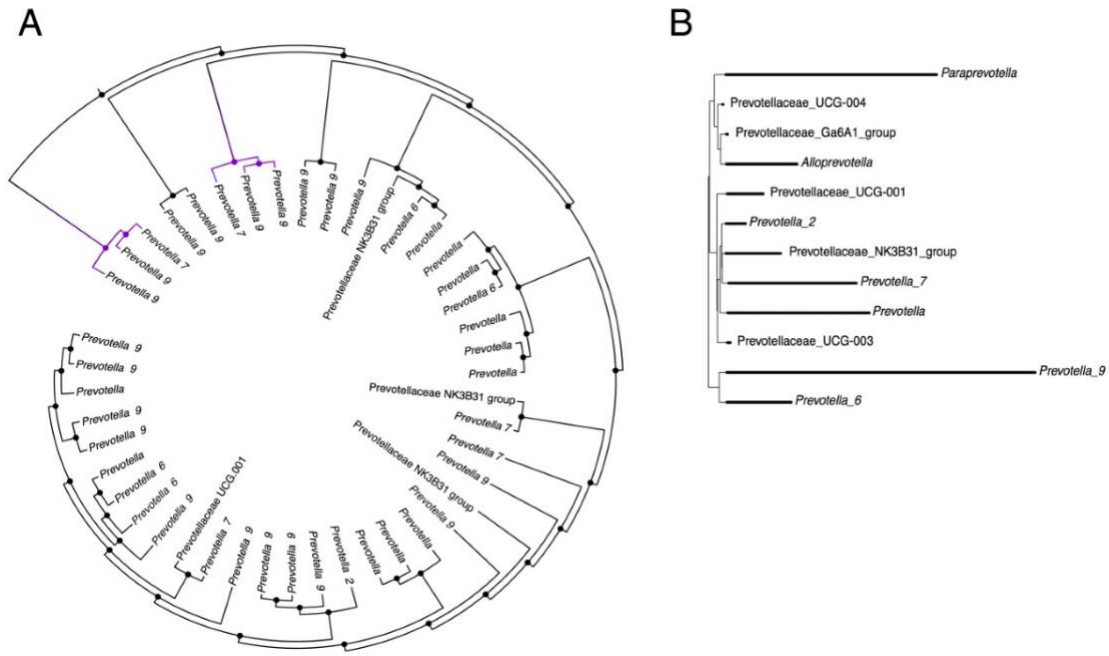

Figure 5. A) Cluster tree showing an ecological community relationship between *Prevotella* ASVs within stool samples from TwinsUK participants. Taxa clades represent members of the same microbial ecological community. *Prevotella_9* and *Prevotella_7* are the only 2- taxon clusters present, and ASVs assigned to these taxa cluster in the same clade more frequently than any other others taxa. Clustering of *Prevotella_7* and *Prevotella_9* ASVs indicates that the abundance of these taxa are interdependent; a biological interdependence and environmental niche similarity is suggested. Incremental levels are shown, from individual ASVs joined as two-taxon clusters, to the full set of Prevotellaceae ASVs joined by the final head node. **B)** Phylogenetic tree for Prevotellaceae ASVs within TwinsUK participants. *Prevotella_9* and *Prevotella_7* are phylogenetically distinct from each other.

Prevotella_7 may be implicated in RA pathogenesis, and may, indeed, be increased in RA patients (resilience hypothesis, described in Supplementary Methods). Conversely, given that the RA-unaffected TwinsUK participants are beyond the mean age of RA onset, yet having high genetic risk in some, these results may hypothetically reflect participants being resilient to the development of RA. Correspondingly, the genetic risk and microbiota associations demonstrated may potentially be microbial markers of RA resilience. In this instance, the ratio of *Prevotella_7:Prevotella_9 (copri)* would be expected to be higher in unaffected twins.

The relative abundance of *Prevotella spp.* was calculated in 18 monozygotic female RA-discordant twin pairs excluded from our previous analysis, in whom RA diagnosis had been confirmed during clinical visit. A higher relative abundance of *Prevotella_7* was demonstrated in the RA affected twins compared to control twins: 0.005 and 0·001, respectively (**Supplementary Figure 1**). This finding neither supports nor rejects the hypothesis that *Prevotella_7* is implicated in RA aetiology: In accordance with other studies of established RA, abundance of *Prevotella spp.* was not significantly associated with RA. The proportion of *Prevotella_7* (species prediction uncertain) to *Prevotella_9* (predicted *Prevotella* copri) was higher in the RA affected twins than the control twins: 0·4 and 0·01, respectively, suggesting that *Prevotella_7* is not a marker of RA resilience.

Discussion

This study demonstrates a link between host genetic risk for RA and the gut microbiota in two large human cohorts. Gut microbiota associations have been demonstrated in the absence of clinically detectable disease in TwinsUK: participants with RA, along with their unaffected co-twins, were removed from the study. The strongest association between RA genetic risk and the gut microbiota was an increase in abundance of *Prevotella_7* (q <1e-7). This finding was validated in the SCREEN-RA cohort in which *Prevotella_7* was positively correlated with shared epitope positivity (q =1·1e-3). *Prevotella_7* was demonstrated to share a biological interdependence with *Prevotella_9* (predicted *P.copri*).

There is considerable interest in *Prevotella copri* as a potential mediator of RA pathology, and it is a candidate keystone taxon enriched in the gut microbiota of newly diagnosed RA patients. Since this observation was made, the human immune response to this microbe has been of considerable interest. Antibodies to *P.copri* have been shown to associate with disease severity and innate- and Th1 and Th17 immune responses in RA[21]. Functional work has suggested a role for *P.copri* in Th17 cell differentiation.[22] That *P.copri* is associated with new-onset RA prior to treatment with DMARDS, [4] and is also associated with other inflammatory conditions,[23] has fuelled speculation that inflammation is a pre-requisite for *P.copri* proliferation within the gut, relative to other taxa.[21,24] *P.copri* may have adapted to thrive in a pro-inflammatory environment, and may further promote the inflammatory milieu, thereby enhancing its own favoured environmental niche.[24] In doing so, *P.copri* is suggested to contribute to RA pathology. According to this model, a human–microbiota interspecies positive feedback loop is proposed. Another example of such a model is the hijacking of complement cascade by *Porphyromonas gingivalis*.[25]

An increase in abundance of three other taxa – Ruminococaceae_UCG-14, *Rikenella* and *Shigella* - with RA PRS was also observed. Ruminococcaceae and *Shigella* have been shown previously to be in higher abundance in RA patients compared to controls.[3,4] The evidence for a role of these taxa in RA is much weaker than for *P.copri* - with lower replication across studies, and no functional link reported to date. However, the association both with RA patients and with genetic risk for RA in a large, non-disease cohort is interesting, and merits further investigation.

Whilst causality is not determinable from a cross sectional association study, our results provide robust evidence indicating that host genetic factors influence the abundance of *Prevotella* in the gut micobiota. Both TwinsUK and SCREEN RA cohorts were balanced for ACPA positivity in relation to RA genetic risk loci. A previous study of new onset RA patients and healthy participants reported an inverse association of HLA genotype with abundance of *Prevotella copri*, in the opposite direction to that expected.[5] This association may potentially relate to population differences or confounding factors which were not examined. The present study using ASVs in a very large sample without disease provides more substantial evidence for an association of genotype with the RA gut microbiota.

The findings suggest that the microbiota are altered prior to disease and this is in accord with previous reports of *Prevotella spp.* in pre-RA[9] and RA.[5] Indeed, we find similar in the pre-clinical RA cases in first degree relatives of RA patients: in a recent study of the SCREEN-RA cohort an enrichment of Prevotellaceae in the gut microbiota in pre-RA was observed. In the study by Alpizaar-Rodriguez *et al.*[9] no statistically significant Prevotellaceae genera were shown to drive the association. Therefore in the present study data were re-analysed using the more recent method of ASVs. ASVs offer an updated approach to traditional clustering based methods generating OTUs from 16S sequence. As opposed to grouping sequences based on a similarity threshold as for

OTUs, the error rate is used to infer the original biological sequence, and produce units of matched sequence. ASVs offer higher resolution, and have greater biological relevance therefore.[26] The ASV analysis showed an FDR-adjusted significant *Prevotella* genus association with a species level group annotation indicating potential *Prevotella copri*. Our follow-up study of pre-RA revealed a furthersix novel genus associations which had not been evident in the original OTUs– *Lactobacillus, Enterococcus, Faecalibacterium,* Ruminococcaceae UBA1819*, Veilllonella* and *Butyrivibrio.* Of the six associations, the first four have been reported associated with RA previously,[4,27,28] and Ruminococcaceae was seen associated with RA PRS in TwinsUK. Butyrivibrio has not yet been associated with RA, but this is a physiologically interesting taxa associated with short chain fatty acid production and host metabolism.

There are a number of limitations to the study. First, in the pre-RA follow-up analysis both cases and controls were first degree relatives of RA patients, with increased genetic risk compared to the general population. It is possible that the pre-clinical RA cases had higher genetic risk than the controls, but as full genotyping was not available, we were unable to confirm this. The TwinsUK cohort is predominantly female (Table 1). The SCREEN-RA cohort is slightly more balanced in terms of sex, however population prevalence of RA is much higher in females than males. As the age of onset of RA is 30-65,[29] the TwinsUK participants who have a median age of 63 are less likely to develop disease than the SCREEN-RA participants. However, this is of benefit to the design of our study of genetically high risk yet unaffected individuals and helps us to understand microbiota differences in the absence of disease.

This work highlights the value of using the newer methods of amplicon sequence variants – ASVs - which can detect taxonomic variation overlooked by OTU based methods. Further microbiota associations with RA are to be anticipated from improvements in sequencing and its interpretation. Annotation of ASVs with the SILVA database[15,18] demonstrates differences at the species level which would be overlooked using other reference databases. Finally, modelling of the community relationship provides valuable insight into the underlying biology. Future studies should take advantage of these methods.

Taken together, these results support the hypothesis that microbiota is altered in individuals with genetic predisposition to RA before the onset even of pre-clinical RA. The work sheds light on our understanding of microbiota in RA and addresses the cause vs consequence issue; – if microbial alteration precedes disease, the microbiota may lie on the causal pathway. However we cannot yet exclude the possibility that *Prevotella spp.* are bystanders. Additionally, having RA genetic risk and *Prevotella spp.* is not sufficient for disease development, but rather may be one of several insults contributing to progression of RA pathology, in line with the favoured "two hit hypothesis" of RA pathogenesis.[30] The identification of pre-clinical RA represents an important clinical target in early disease intervention and is currently the subject of multiple immune-modulating clinical trials. Further, the genetic risk-microbiota associations identified here may be applicable to other diseases as there is crossover in genetic aetiology between RA and other autoimmune conditions. Finally, the microbiota may offer the opportunity for modulation of pre-disease pathways alone or in combination with immune-modulating drugs.

Conclusions

Gut microbiota abundance is associated with RA risk genotype even in the absence of disease. Genotype may mediate key taxonomic associations of the gut microbiota with RA, particularly *Prevotella spp.*[4,5,21,22] suggesting that these species play a role very early in development of RA.

Acknowledgements and affiliations
This work was funded by a Versus Arthritis Special Strategic Award (grant 21227). Twins UK receives funding from the Wellcome Trust (including phenotypic and genotypic data under WT081878MA); the National Institute for Health Research (NIHR) Clinical Research Facility at Guy's & St Thomas' NHS Foundation Trust and NIHR Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. Claire J. Steves receives funding from the MRC, Wellcome Trust and Clinical Disease Reasearch Foundation (CDRF). Philippa Wells is supported by a KMRT PhD studentship from King's College London. The funding providers for this work did not provide input regarding study design. The authors would like to thank the participants of the TwinsUK and SCREEN-RA cohorts. We thank Dr Marina Mora-Ortiz and Mr Louca Panayioutis for assisting with the figures.

Conflict of Interest Statement
All authors declare no competing interests.

Author Contributions
CS, FW, BK and AC conceived the project. PW developed the theory, performed the experiments, analysed and interpreted the data, and took the lead in writing the manuscript. CS and FW supervised the analysis. CS encouraged development of the work through leading collaboration with the SCREEN-RA cohort, and inclusion of investigation of community structure, and abundance of taxa in RA discordant twins. AF, TS, TL, DAR and BG shared data generated from SCREEN-RA cohort and encouraged the investigation of RA autoantibodies in the TwinsUK cohort. AA performed analysis of the community structure. FW, PW and CS collected data from TwinsUK. RB contributed to the design and interpretation of statistical models. MF assisted with UKBiobank genotype data. All authors provided ctitical feedback and contributed to the final manuscript.

Data Availibility
Data used in this study is available upon reasonable request to TwinsUK.

References

1 MacGregor AJ, Snieder H, Rigby AS, *et al.* Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum* 2000; **43**: 30–7.

2 Wells PM, Williams FMK, Matey-Hernandez ML, Menni C, Steves CJ. 'RA and the microbiome: do host genetic factors provide the link? *J Autoimmun* 2019; published online March 5. DOI:10.1016/j.jaut.2019.02.004.

3 Sun Y, Chen Q, Lin P, *et al.* Characteristics of Gut Microbiota in Patients With Rheumatoid Arthritis in Shanghai, China. *Front Cell Infect Microbiol* 2019; **9**. DOI:10.3389/fcimb.2019.00369.

4 Zhang X, Zhang D, Jia H, *et al.* The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat Med* 2015; **21**: 895–905.

5 Scher JU, Sczesnak A, Longman RS, *et al.* Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *eLife* 2013; **2**: e01202.

6 Isaac S, Artacho A, Nayak R, *et al.* Op0119 the Pre-Treatment Gut Microbiome Predicts Early Response to Rheumatoid Arthritis Therapy. *Ann Rheum Dis* 2019; **78**: 133–4.

7 Goodrich JK, Waters JL, Poole AC, *et al.* Human Genetics Shape the Gut Microbiome. *Cell* 2014; **159**: 789–99.

8 Verdi S, Abbasian G, Bowyer RCE, *et al.* TwinsUK: The UK Adult Twin Registry Update. *Twin Res Hum Genet Off J Int Soc Twin Stud* 2019; : 1–7.

9 Alpizar-Rodriguez D, Lesker TR, Gronow A, *et al.* Prevotella copri in individuals at risk for rheumatoid arthritis. *Ann Rheum Dis* 2019; **78**: 590–3.

10 Khera AV, Chaffin M, Aragam KG, *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 2018; **50**: 1219–24.

11 Jones HJ, Hubbard L, Mitchell RE, *et al.* Association of Genetic Risk for Rheumatoid Arthritis With Cognitive and Psychiatric Phenotypes Across Childhood and Adolescence. *JAMA Netw Open* 2019; **2**. DOI:10.1001/jamanetworkopen.2019.6118.

12 Duncan L, Shen H, Gelaye B, *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun* 2019; **10**: 1–9.

13 Vilhjálmsson BJ, Yang J, Finucane HK, *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet* 2015; **97**: 576–92.

14 Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High resolution sample inference from Illumina amplicon data. *Nat Methods* 2016; **13**: 581–3.

15 Callahan B. Silva taxonomic training data formatted for DADA2 (Silva version 132). 2018; published online Feb 13. DOI:10.5281/zenodo.1172783.

16 Henderson G, Yilmaz P, Kumar S, *et al.* Improved taxonomic assignment of rumen bacterial 16S rRNA sequences using a revised SILVA taxonomic framework. *PeerJ* 2019; **7**: e6496.

17 Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw* 2015; **67**: 1–48.

18 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014; **15**. DOI:10.1186/s13059-014-0550-8.

19 Morton JT, Sanders J, Quinn RA, *et al.* Balance Trees Reveal Microbial Niche Differentiation. *mSystems* 2017; **2**: e00162-16.

20 McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE* 2013; **8**: e61217.

21 Pianta A, Arvikar S, Strle K, *et al.* Evidence for Immune Relevance of Prevotella copri, a Gut Microbe, in Patients with Rheumatoid Arthritis. *Arthritis Rheumatol Hoboken NJ* 2016; published online Nov 18. DOI:10.1002/art.40003.

22 Maeda Y, Kurakawa T, Umemoto E, *et al.* Dysbiosis Contributes to Arthritis Development via Activation of Autoreactive T Cells in the Intestine. *Arthritis Rheumatol* 2016; **68**: 2646–61.

23 Pedersen HK, Gudmundsdottir V, Nielsen HB, *et al.* Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* 2016; **535**: 376–81.

24 Hofer U. Microbiome: Pro-inflammatory Prevotella? *Nat Rev Microbiol* 2014; **12**: 5–5.

25 Olsen I, Lambris JD, Hajishengallis G. Porphyromonas gingivalis disturbs host–commensal homeostasis by changing complement function. *J Oral Microbiol* 2017; **9**. DOI:10.1080/20002297.2017.1340085.

26 Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 2017; **11**: 2639–43.

27 Picchianti-Diamanti A, Panebianco C, Salemi S, *et al.* Analysis of Gut Microbiota in Rheumatoid Arthritis Patients: Disease-Related Dysbiosis and Modifications Induced by Etanercept. *Int J Mol Sci* 2018; **19**. DOI:10.3390/ijms19102938.

28 Chen J, Wright K, Davis JM, *et al.* An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. *Genome Med* 2016; **8**: 43.

29 Mueller RB, Kaegi T, Finckh A, Haile SR, Schulze-Koops H, von Kempis J. Is radiographic progression of late-onset rheumatoid arthritis different from young-onset rheumatoid arthritis? Results from the Swiss prospective observational cohort. *Rheumatology* 2014; **53**: 671–7.

30 McInnes IB, Schett G. Pathogenetic insights from the treatment of rheumatoid arthritis. *The Lancet* 2017; **389**: 2328–37.