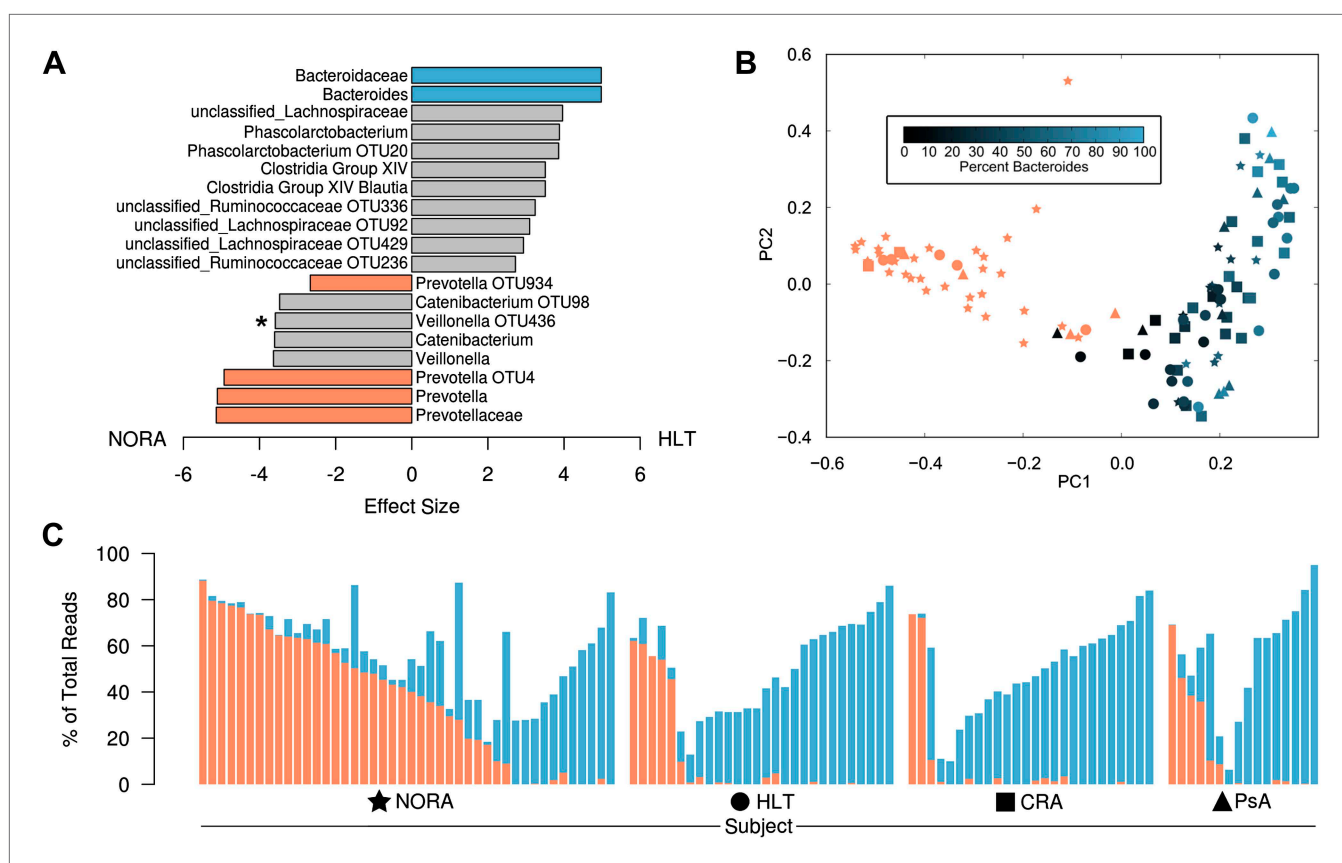


---

## Figures and figure supplements

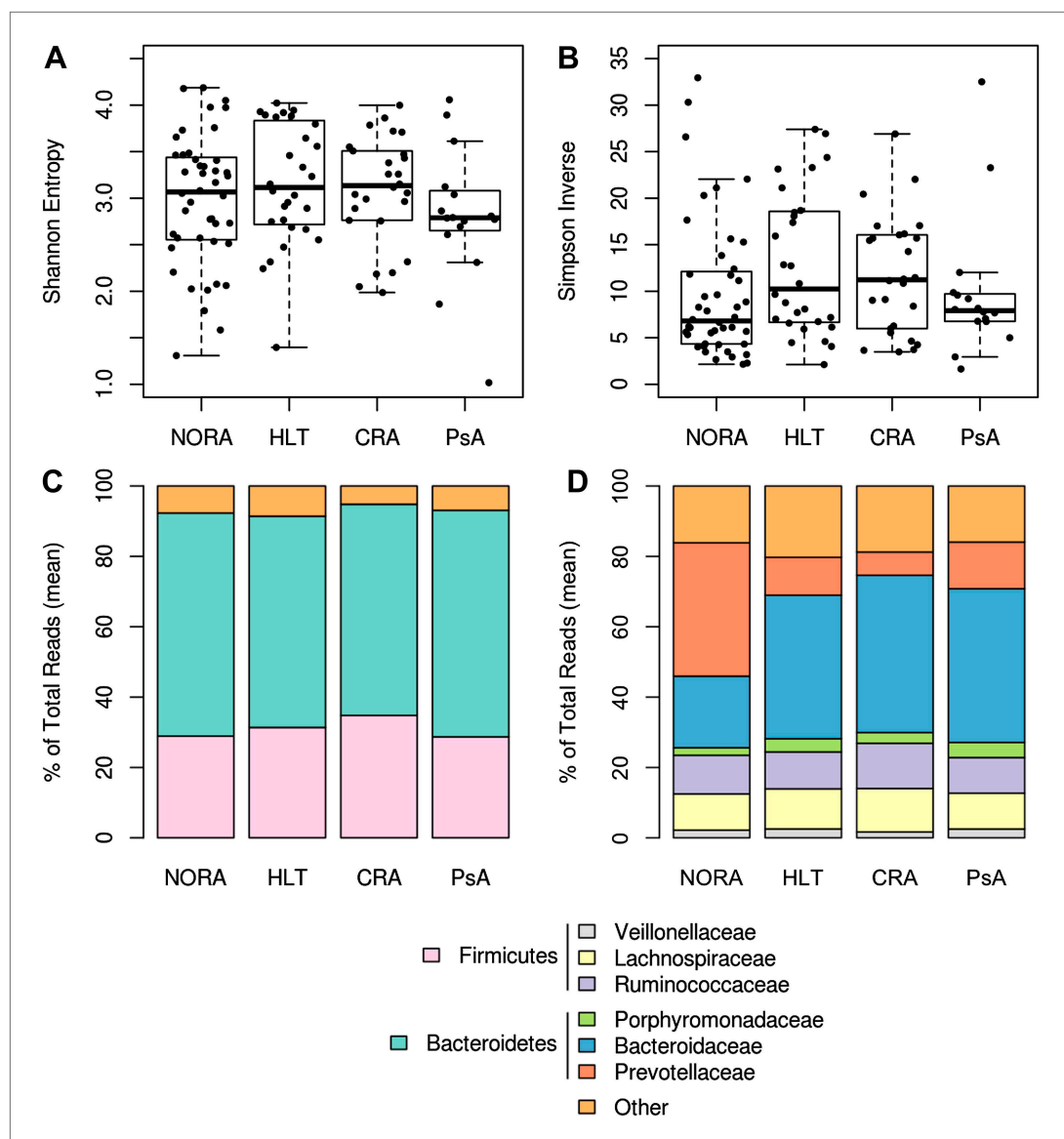
Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis

**Jose U Scher, et al.**



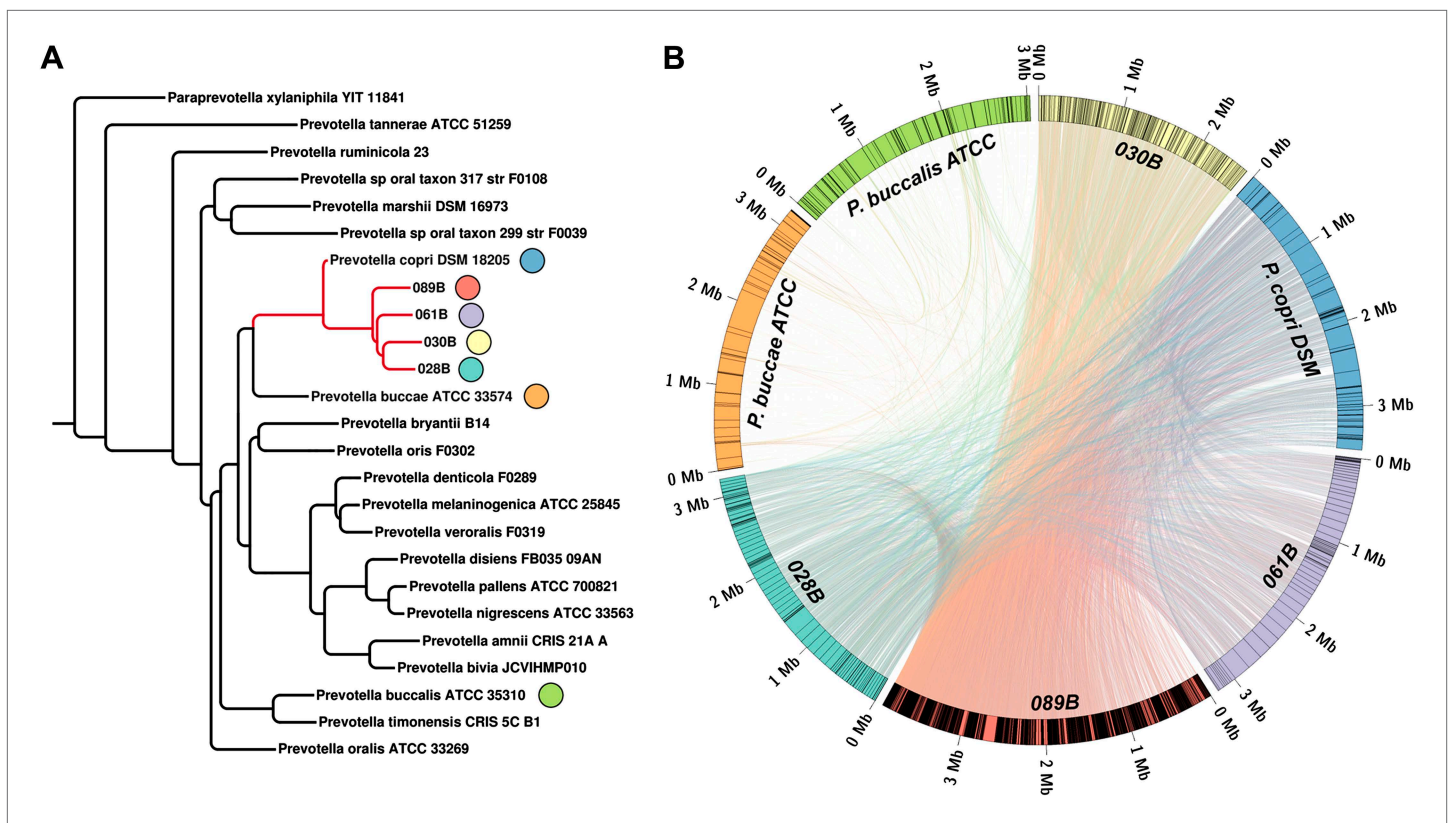
**Figure 1.** Differences in the relative abundance of *Prevotella* and *Bacteroides* in 114 subjects with and without arthritis, determined by 16S sequencing (regions V1–V2, 454 platform). **(A)** LefSe (Segata et al., 2011) was used to compare the abundances of all detected clades among all groups, producing an effect size for each comparison ('Materials and methods'). All results shown are highly significant ( $q < 0.01$ ) by Kruskal-Wallis test adjusted with the Benjamini-Hochberg procedure for multiple testing, except that indicated with an asterisk, which is significant at  $q < 0.05$ . Negative values (left) correspond to effect sizes representative of NORA groups, while positive values (right) correspond to effect sizes in HLT subjects. *Prevotella* was found to be over-represented in NORA patients, while *Bacteroides* was over-represented in all other groups. **(B)** The Bray-Curtis distance between all subjects was calculated and used to generate a principal coordinates plot in MOTHUR (Schloss et al., 2009). The first two components are shown. Subjects with an abundance of *Prevotella* greater than 10% were colored red. Other subjects were colored according to their *Bacteroides* abundance as shown. NORA subjects (stars) primarily cluster together according to their *Prevotella* abundance, and the x-axis is representative of differences in the relative abundance of *Prevotella* and *Bacteroides*. **(C)** The abundances of *Prevotella* (red) and *Bacteroides* (blue) are shown for all subjects, sorted in order of decreasing *Prevotella* abundance (>5%) and increasing *Bacteroides* abundance.

DOI: 10.7554/eLife.01202.004



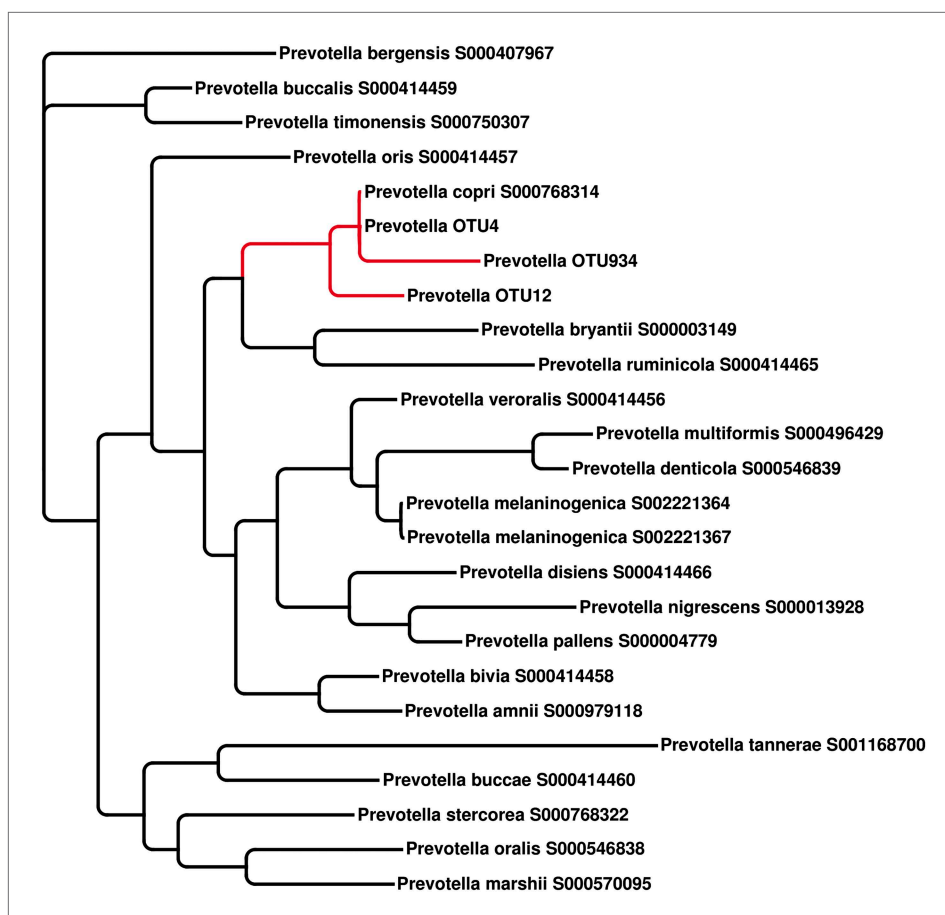
**Figure 1—figure supplement 1.** Gut microbiota richness, diversity and relative abundance in NORA patients and controls. **(A and B)** Gut microbiota richness and diversity are similar among RA groups and healthy controls. **(C)** Phyla abundance by group. No significant differences were found at this taxonomic level. **(D)** Family abundance by group. NORA subjects have a significant increase in Prevotellaceae (red) and a concomitant decrease in Bacteroidaceae (blue) by FDR-adjusted Kruskal-Wallis test ( $q < 0.01$ ).

DOI: [10.7554/eLife.01202.007](https://doi.org/10.7554/eLife.01202.007)



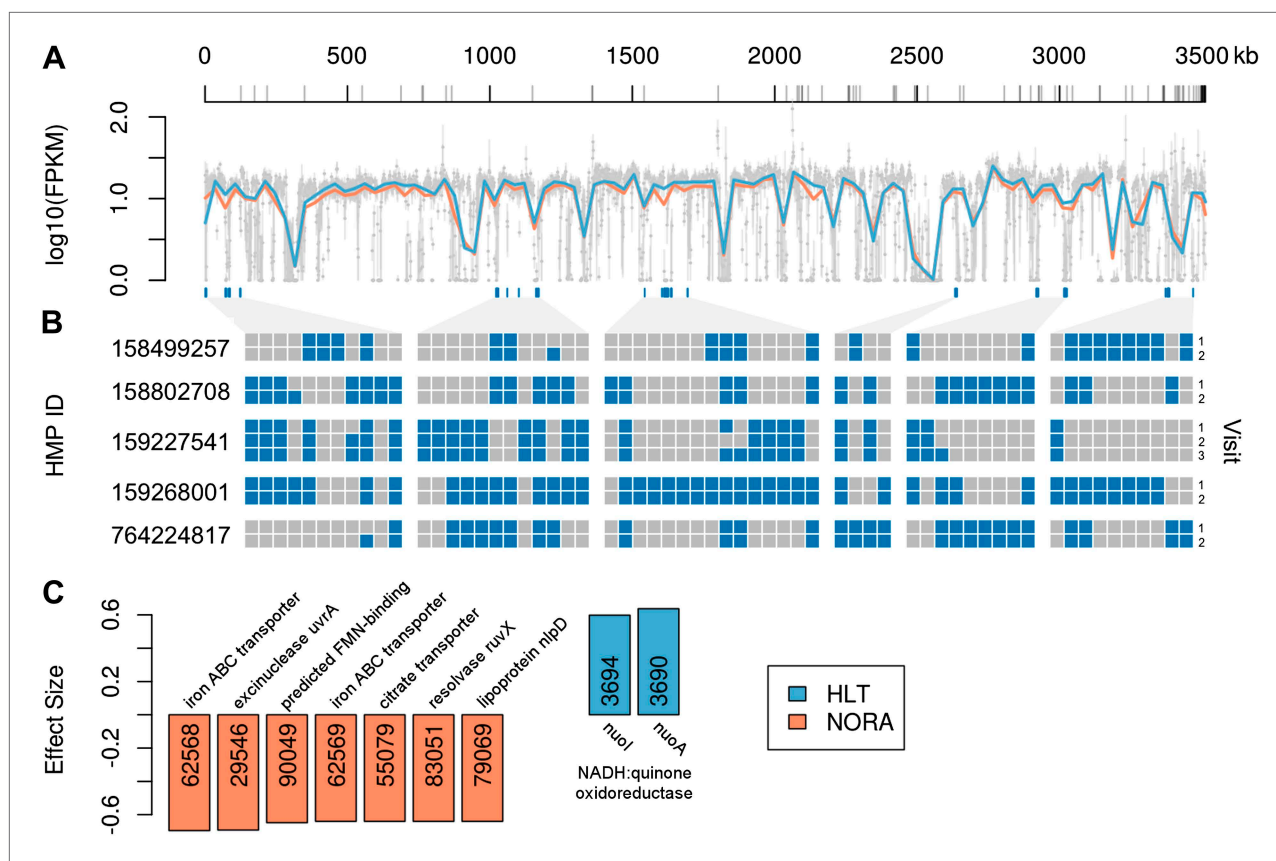
**Figure 2.** Homology-based classification of patient-associated *Prevotella*. Four NORA subjects with a high abundance of *Prevotella* OTU4 were selected for shotgun sequencing and metagenome assembly. **(A)** The resulting metagenomic contigs were used to generate a phylogenomic tree with PhyloPhlAn (Segata et al., 2013). **(B)** Assemblies were filtered by alignment to the reference *Prevotella copri* DSM 18205 genome, keeping contigs with at least one 300 bp region aligned at 97% identity or greater. The resulting draft patient-derived *P. copri* assemblies were aligned to one another, the reference *P. copri* genome, and two distinct *Prevotella* taxa (*Prevotella buccae* and *Prevotella buccalis*). Colored arcs represent assemblies as labeled, lines connecting arcs represent regions of >97% identity >1 kb in length, and gray lines dividing colored arcs represent boundaries between contigs. These results demonstrate that *Prevotella* OTU4, OTU12, and OTU934 form a clade with *P. copri* (left, red highlighted subtree) that is genetically distinct from more distant *Prevotella* taxa.

DOI: [10.7554/eLife.01202.009](https://doi.org/10.7554/eLife.01202.009)



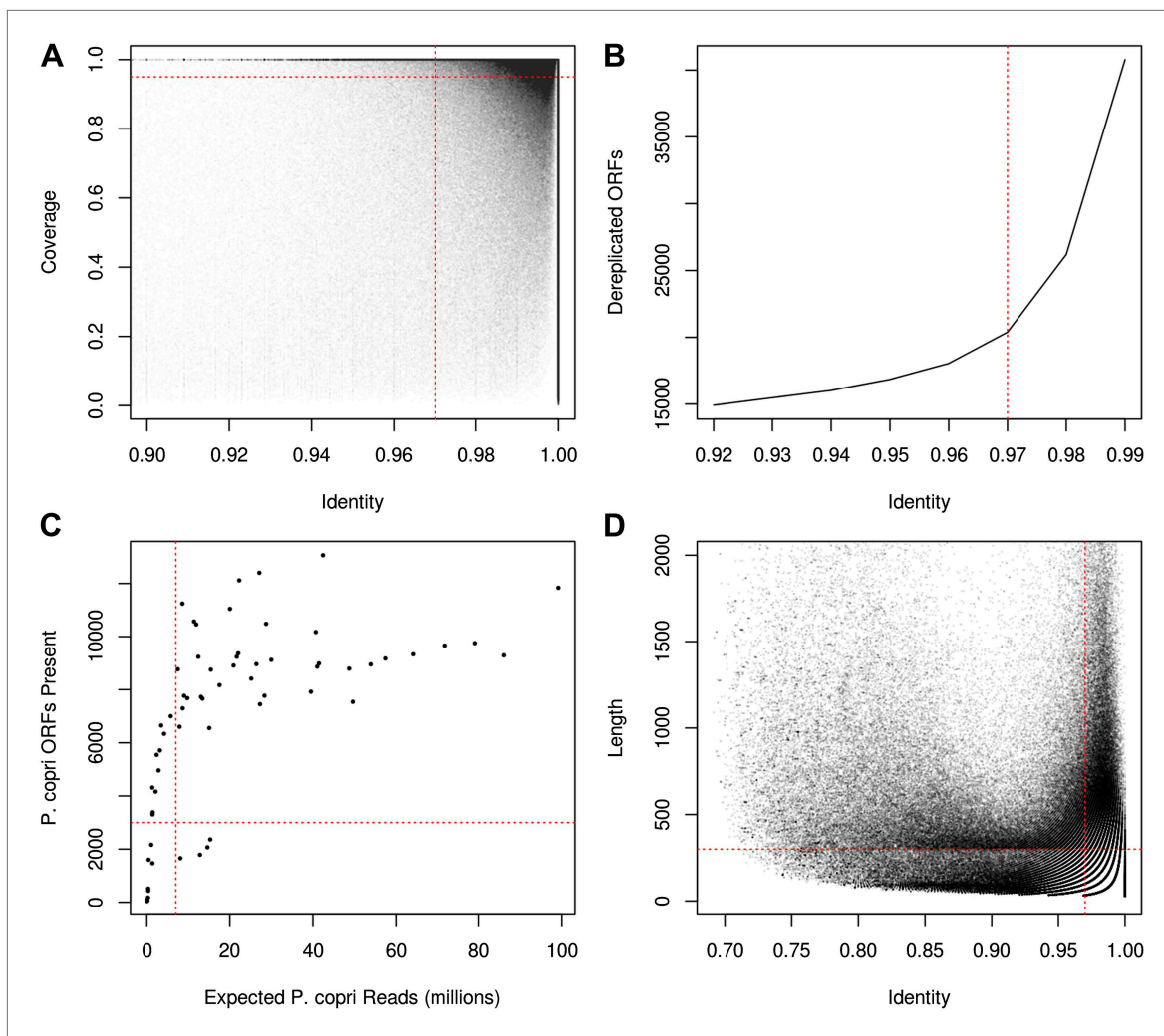
**Figure 2—figure supplement 1.** The representative 16S sequenced reads for *Prevotella* OTU4, OTU12, and OTU934 were aligned with MUSCLE (Edgar, 2004) and clustered with FastTree (Price et al., 2010). All three *Prevotella* OTUs cluster with the full-length reference 16S sequence of *P. copri*.

DOI: [10.7554/eLife.01202.012](https://doi.org/10.7554/eLife.01202.012)



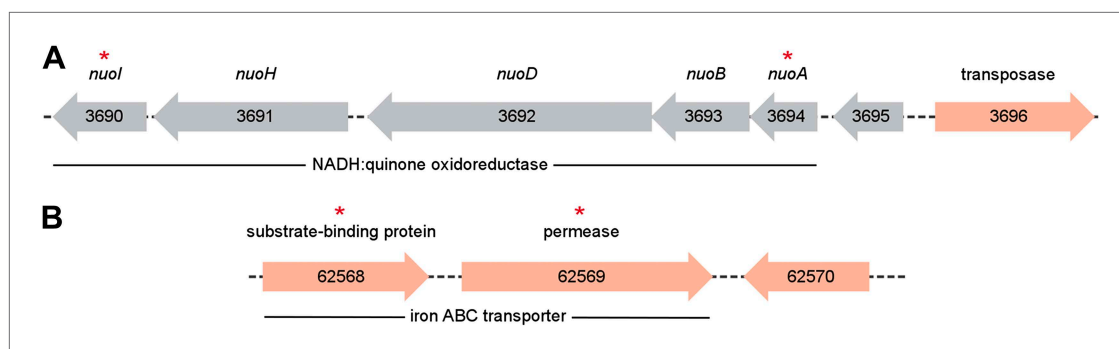
**Figure 3.** Comparison of *P. copri* genomes from healthy and NORA subjects. **(A)** Comparative coverage of the draft *P. copri* DSM 18205 genome between individuals and within healthy and NORA groups. Gray points are median fragments per kilobase per million (FPKM) for 1-kb windows, gray lines within the plot are the interquartile range for each window, red and blue lines the LOWESS-smoothed average for NORA and healthy groups, respectively. Gray lines on the horizontal axis represent boundaries between assembled contigs. Regions are variably covered between subjects and groups, with several genomic islands lacking overall or especially variable (dark blue lines below the plot). **(B)** The presence (blue) or absence (gray) of previously-reported *P. copri*-unique marker genes (Segata et al., 2012) in 11 stool samples from five subjects of the Human Microbiome Project (HMP) are shown as a heatmap. We report, in columns, only those *P. copri*-specific markers showing variable presence/absence patterns across the considered HMP samples. Each row represents a different sample collection date, groups of rows represent subjects, and groups of columns correspond to different variably covered genomic islands. Strains of *P. copri* are defined by the presence and absence of particular genes, which remain stable for at least 6 months in these individuals. All inter- and intra-individual comparisons between rows are highly statistically significant ( $p < 0.001$ , 'Materials and methods'). **(C)** The *P. copri* pangenome was identified by finding *P. copri* ORFs in all HMP and NORA cohort subjects, and the presence or absence of these ORFs was calculated for each subject ('Materials and methods', Figure 3—figure supplement 1). Several ORFs are statistically significant biomarkers between healthy and NORA status ( $q < 0.25$ ) (Supplementary file 1B, 'Materials and methods').

DOI: 10.7554/eLife.01202.014



**Figure 3—figure supplement 1.** Recovery of *P. copri* pangenome from HMP/RA shotgun reads and determination of presence/absence of *P. copri* ORFs by alignment of reads to pangenome gene catalog. **(A)** Genes were called present in a sample if they were covered by aligned reads at an identity threshold of >97% over >97% of their length (red lines). **(B)** ORFs were called on contigs using MetaGeneMark (Zhu et al., 2010) and were dereplicated with UCLUST (Edgar, 2010) at an identity threshold of 97% (red line). **(C)** Recovery of a sample's *P. copri* pangenome saturated at approximately 7 million reads (red line). We therefore excluded samples with less than 7 million *P. copri* reads, defined as *P. copri* abundance determined by MetaPhlAn (Segata et al., 2012) multiplied by the total number of quality-filtered reads. Samples with *P. copri* abundance likely misestimated (i.e., those with <3000 ORFs present) were also excluded (Supplementary file 1A). **(D)** Contigs were said to have originated from *P. copri* if they had at least one hit >97% identity over >300 bp (red lines).

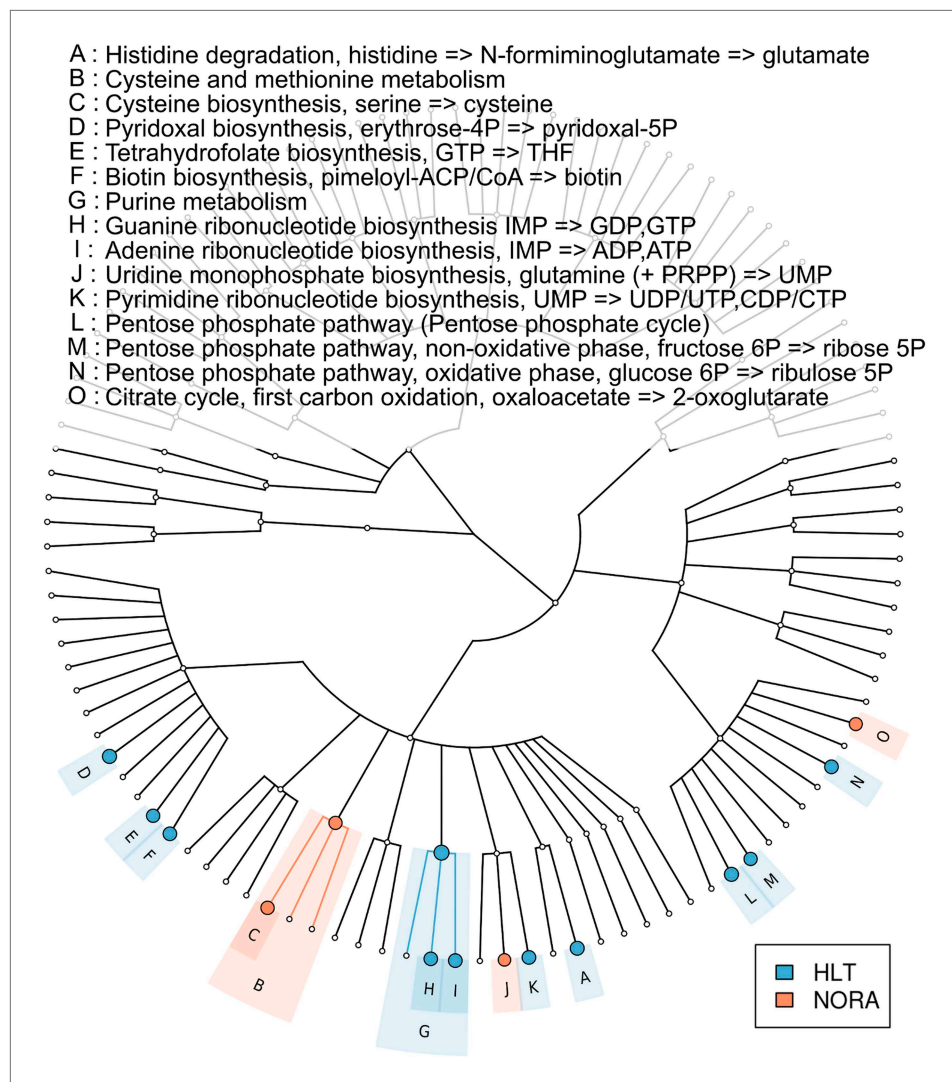
DOI: [10.7554/eLife.01202.017](https://doi.org/10.7554/eLife.01202.017)



**Figure 3—figure supplement 2.** Metagenomic context of discriminative biomarker ORFs. ORFs found in the *P. copri* DSM 18205 reference genome are colored red, while those identified as differentially present in healthy and NORA groups are indicated with red asterisks. **(A)** Two ORFs, 3690 and 3694, are healthy-specific, occur on the same contig, and encode different components of the same NADH:quinone oxidoreductase. **(B)** Similarly, ORFs 62,568 and 62,569 occur on the same contig, are NORA-specific, and encode components of the same iron ABC transporter.

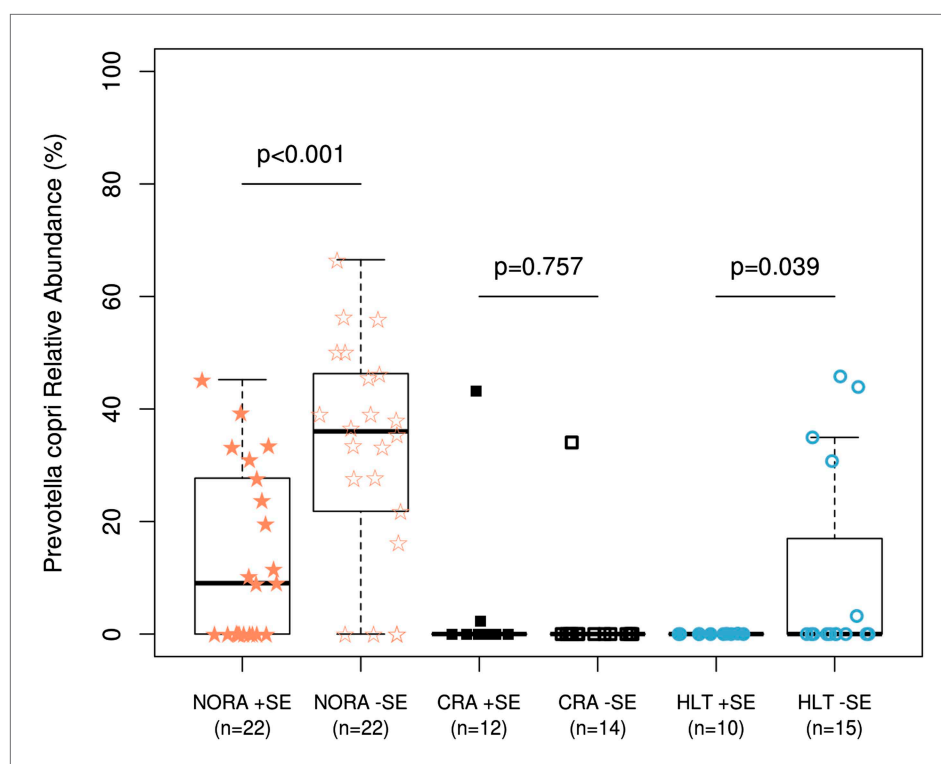
DOI: [10.7554/eLife.01202.018](https://doi.org/10.7554/eLife.01202.018)





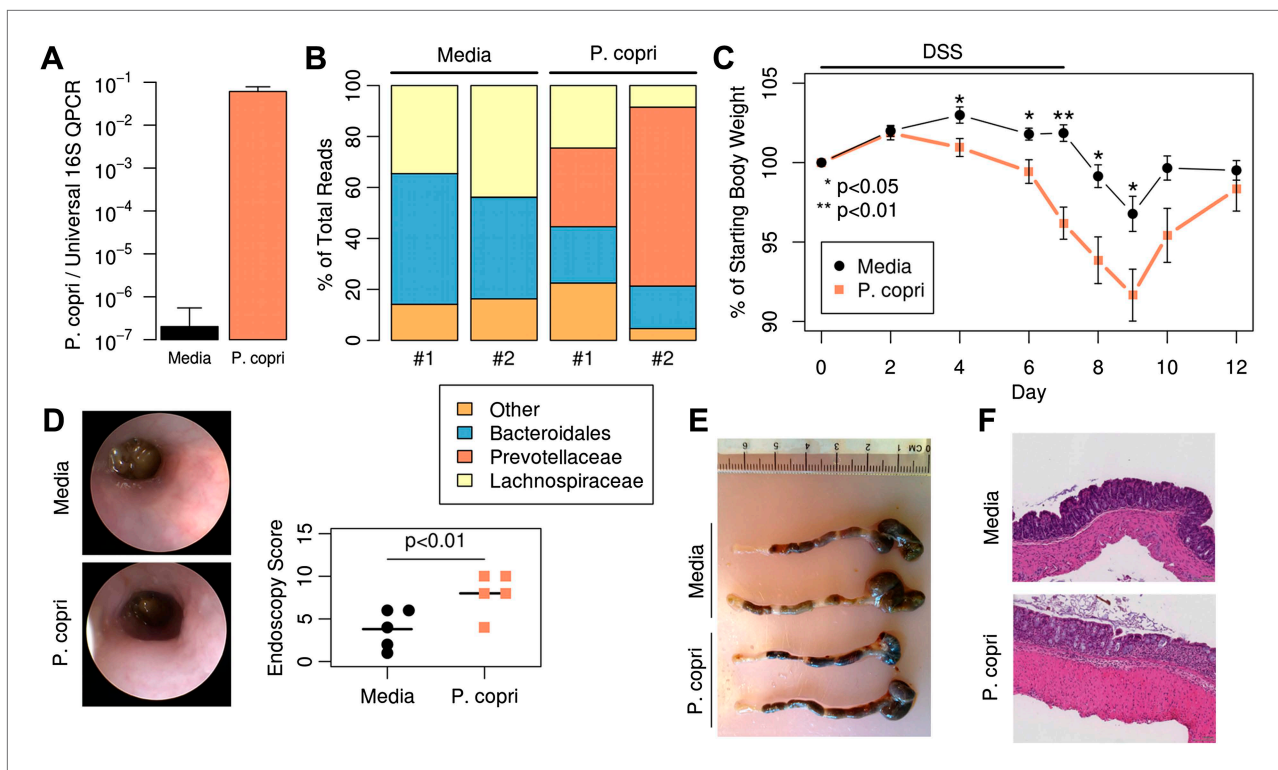
**Figure 4.** Metabolic pathway representation in the microbiome of healthy and NORA subjects. HUMAnN (Abubucker et al., 2012) was applied to metagenomic reads (paired-end, 100 nt, Illumina platform) from NORA subjects (n = 14) and healthy controls (n = 5) to quantitate the abundances of hierarchically related KEGG modules in these samples ('Materials and methods' and **Supplementary file 1A**). LEfSe (Segata et al., 2011) was used to find statistically significant differences between groups at an alpha cutoff of 0.001 and an effect size cutoff of 2.0. Results shown here are highly significant ( $p < 0.001$ ) and represent large differences between groups. Modules highlighted in red are over-abundant in NORA samples while modules highlighted in blue are over-abundant in healthy samples. *Prevotella*-dominated NORA metagenomes have a dearth of genes encoding vitamin and purine metabolizing enzymes, and an excess of cysteine metabolizing enzymes.

DOI: [10.7554/eLife.01202.019](https://doi.org/10.7554/eLife.01202.019)



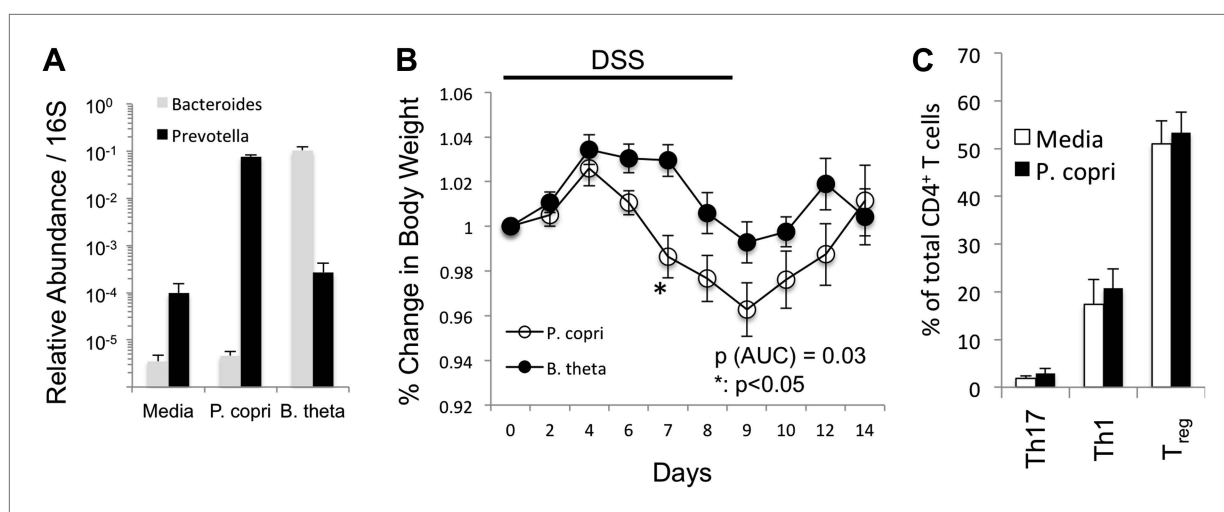
**Figure 5.** Relationship of host HLA genotype to abundance of *P. copri* (OTU4, OTU12, and OTU934 combined relative abundance). The HLA-class II genotype of all subjects was determined by sequence-based typing methodology ('Materials and methods'). Groups were subdivided by the presence or absence of shared-epitope RA risk alleles (+/- SE as indicated above) and correlated with relative abundance of intestinal *P. copri*. A statistically significant correlation is seen between *P. copri* abundance and the genetic risk for rheumatoid arthritis in NORA (red stars) and healthy (blue circles) subjects by Welch's two-tailed t test.

DOI: [10.7554/eLife.01202.021](https://doi.org/10.7554/eLife.01202.021)



**Figure 6.** Colonization with *P. copri* dominates the colonic microbiome and exacerbates local inflammatory responses. **(A)** DNA was extracted from fecal pellets of media-gavaged mice and *P. copri*-gavaged mice 2 weeks after colonization and assayed by QPCR with *P. copri* specific primers compared to universal 16S. **(B)** Relative abundance of bacterial families in fecal DNA from media-gavaged and *P. copri*-colonized mice (shown in duplicate) by high-throughput 16S sequencing (regions V1–V2, 454 platform). **(C)** C57BL/6 mice colonized with *P. copri* ( $n = 15$ ) or media alone ( $n = 13$ ) controls were exposed to DSS for seven days and percent of starting body weight is shown. Composite data from three representative experiments are shown. **(D)** Representative colonoscopic images of mice colonized with *P. copri* or media gavage following DSS-induced colitis. Endoscopic colitis score for five individual animals is displayed. **(E and F)** Gross pathology **(E)** and histology **(F)** of colons from mice colonized with *P. copri* or media gavage following DSS-induced colitis.

DOI: [10.7554/eLife.01202.023](https://doi.org/10.7554/eLife.01202.023)



**Figure 6—figure supplement 1.** *P. copri* colonization exacerbates chemically induced colitis. **(A)** DNA was extracted from fecal pellets of media, *P. copri*, and *B. theta* gavaged mice 2 weeks after colonization and assayed by QPCR with *P. copri* or *Bacteroides* specific primers compared to universal 16S amplicon. **(B)** C57BL/6 mice colonized with *P. copri* (n = 10) or *B. theta* (n = 10) were exposed to DSS for seven days and percent of starting body weight is shown. **(C)** Percent of total CD4<sup>+</sup> T-cells in the colonic lamina propria expressing IL-17 (Th17) or IFN $\gamma$  (Th1) following PMA/ionomycin stimulation or expressing Foxp3 (Treg).

DOI: [10.7554/eLife.01202.025](https://doi.org/10.7554/eLife.01202.025)