**HIV, HPV and Microbiota: Partners in Crime?**

Sergio Serrano-Villar, Emilia Vásquez-Domínguez, José Antonio Pérez-Molina, Talía Sainz, Amparo de Benito, Amparo Latorre, Andrés Moya, María José Gosalbesand Santiago Moreno.

**SUPPLEMENTARY METHODS**

**High-resolution anoscopy (HRA)**

HRA was performed using the standard procedure, including topical application of 3% acetic acid and lugol solution in the anal canal. Anal biopsy was taken from abnormal areas revealed by HRA as acetowhite lugol-negative lesions. In cases with no visible lesions on HRA, a random biopsy of the squamocolumnar junction area was performed. Histology specimens were classified according to the Bethesda classification [1]. Treatment with infrared coagulation was offered to patients in whom the biopsy revealed high-degree intraepithelial lesions (HSIL). Adjacent biopsies (1-2 of 2-3mm of length) were sampled with baby-Tischler forceps for 16SrRNA gene sequencing. Tissue samples were stored in RNAlater (Life Technologies) at -80ºC until use.

**Nucleic acid purification**

Fecal samples were stored in RNAlater (Life Technologies) at -80ºC until use. Fecal samples were diluted (dilution ½) in phosphate buffered saline solution (PBS). Then, they were centrifuged at 2000 rpm at 4ºC for 2 min to remove fecal debris. A previous lytic treatment of pelleted cells was performed and then total DNA was extracted using magnetic pearls according to the Magna Pure LC protocol (Roche Applied Science, Manheim, Germany). For tissue samples, we followed a modified MasterPureTM DNA Purification Kit protocol (Illumina Inc., Madison, WI).

Tissue samples were stored in saline solution at -80ºC until use. For DNA extraction, we followed a modified MasterPureTM DNA Purification Kit protocol (Illumina Inc., Madison. WI). We first added 300 uL of tissue and cell lysis solution and 1 uL of Proteinase K 50 ug/uL. Next, samples were ruptured at 30Hz using the Tissue Lyzer II (QIAGEN, Hilden, Germany) and incubated at 65º for 20 minutes. After centrifugation, the supernatant was collected and treated with 1uL of RNAse 5ug/uL (20 minutes, 37ºC). After incubation in ice for 5 minutes, we added 175 uL of protein precipitation solution and samples were centrifuged (10 minutes, 4ºC, 13000 rpm). Total DNA, corresponding to genomic DNA of co-extracted bacterial and human DNA, was precipitated with isopropanol and subsequently washed with etanol 70º and quantified using Qubit fluorometry.

**Amplification of the 16S rRNA gene**

For each sample the V3-V4 regions of the 16S rRNA gene were amplified by polymerase chain reaction (PCR) using 16S Amplicon PCR Forward Primer (5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG) and 16S Amplicon PCR Reverse Primer (5' GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC). The pooled PCR products were sequenced using V3 MiSeq (Illumina) at FISABIO Sequencing and Bioinformatics Service, Valencia, Spain. All sequences were deposited in the public European Nucleotide Archive server under accession number PRJEB15362. We obtained an average of 94,993 and 63,190 16S rRNA joined sequences per sample from DNA extracted from fecal and tissue samples, respectively.

**16S RNA gene analysis. Biodiversity and clustering.** Sequences were analyzed using Qiime v1.8 software. Amplicon data from the 16S rRNA gene was analyzed following the recommendations of the metagenomic state-of-the-art pipeline QIIME v1.8[2]. 16S rRNA gene reads with low quality score (<20 out of 40 quality units assigned by the 454), short read lengths (<170 nucleotides) and singletons were filter and using the USEARCH[3] pipeline. Reference-based and de novo potential chimeras were also removed from the remaining sequences applying the UCHIME algorithm[4]. Taxonomic information of the 16S rDNA sequences were obtained using the Ribosomal Database Project-II (RDP)[5] and the Greengenes database available in QIIME v1.8 software. The Operational Taxonomic Units (OTUs) are the representation of the different clusters of species. The criteria for collapsing each of the different sequences into OTUs are given by the percentage of identity between the different sequences normally taken the 97% of similarity as a standard practice for mapping the 16S rRNA amplicon sequences to its corresponding species. OTUs were created by Uclust[3] applying a cluster criteria of 97% similarity. The most representative sequence for each OTU was then compared against the Qiime cluster version of the Greengenes database[6] (database 97\_otus.fasta). The annotation was accepted when the bootstrap confidence estimation value was over 0.8, stopping the assignation at the last well-identified phylogenetic level. The genus abundance table was summarized from the resulting otu\_table.biom file by the script summarize\_taxa\_through\_plots.py. Representative sequences were aligned with Pynast[7] against the clustered version of the Greengenes database (database core\_set\_aligned.fasta.imputed), to use as input to reconstruct the phylogenetic tree using the FastTree software[8].Distances analysis between samples were performed by the weighted normalized Unifrac using the beta\_significance.py (--num\_iters 100) from the Qiime v1.8.0 pipeline to obtain the Principal Coordinates Analysis (PCoA).

The Shannon index[9] and the richness estimator Chao1 [10] were calculated to assess the OTUs and genus diversity within the community using the statistical R package (vegan library).

**Biomarker discovery.** The linear discriminative analysis (LDA) effect size (LEfSE) algorithm[11] was used to identify specific taxa as biomarkers for cases and controls. We fixed an α value <0.05.The bacterial taxa with significant differences between samples were used to build the LDA model and to estimate its effect as a discriminant feature among them. The threshold used to consider a discriminative feature for the logarithmic LDA score was set to >2.

**SUPPLEMENTARY TABLES**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table S1. Shannon index and Chao1 alpha-diversity estimator.** | | | | |
|  | **Rectal mucosa** | | **Feces** | |
| Shannon | Chao 1 | Shannon | Chao 1 |
| **Total** | p = 0.5053 | p = 0.4389 | p = 0.6691 | p = 0.2049 |
| **Normal vs. LSIL** | p = 0.3055 | p = 0.2409 | p = 0.7133 | p = 0.1113 |
| **Normal vs. HSIL** | p = 0.7593 | p = 0.7133 | p = 0.5691 | p = 0.8075 |
| **LSIL vs. HSIL** | p = 0.3284 | p = 0.3281 | p = 0.4062 | p = 0.1719 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table S2. Diagnostic accuracy of the bacterial biomarkers identified in rectal mucosa and feces using the LEfSe biomarker discovery tool for the presence of precancerous anal lesions.** | | | | | | | | |
|  | **Outcome** | **Taxa** | **Coefficient** | **P value** | **AUCROC** | **Cut-off**  **(% of relative abundance)** | **Sensitivity** | **Specificity** |
| **Mucosa** | **Normal** | ***Anaerovibrio*** | **0.64** | **0.057** | **0.800 (0.578-0.929)** | 0.00001 | 100% | 63% |
| 0.013 | 40% | 89% |
| *Wautersiella* | 1.53 | 0.064 | 0.684 (0.447-0.844) | 0.00004 | 40% | 94% |
| 0.00001 | 20% | 100% |
| **AIN** | ***Ruminococcus*** | **1.81** | **0.033** | **0.832 (0.626-0.953)** | 0.0033 | 90% | 40% |
| 0.0015 | 58% | 100% |
| ***Peptostreptococcus*** | **0.63** | **0.072** | **0.800 (0.578-0.929)** | 0.00001 | 74% | 60% |
| 0.0015 | 58% | 100% |
| ***Clostridium*** | **2.20** | **0.065** | **0.874 (0.676-0.973)** | 0.0002 | 94% | 60% |
| 0.0027 | 47% | 100% |
| **LSIL** | ***Campylobacter*** | **2.12** | **0.056** | **0.900 (0.653-0.986)** | 0.0018 | 100% | 80% |
| 0.005 | 75% | 90% |
| ***Gardnerella*** | **0.98** | **0.034** | **0.800 (0.524-0.936)** | 0.001 | 75% | 70% |
| 0.0045 | 63% | 100% |
| **HSIL** | ***Catenibacterium*** | **0.94** | **0.088** | **0.800 (0.524-0.936)** | 0.002 | 90% | 75% |
| 0.018 | 20% | 100% |
| **Feces** | **Normal** | *Bifidobacterium* | 0.19 | 0.634 | 0.588 (0.364-0.793) | 0.0011 | 80% | 53% |
| 0.111 | 20% | 94% |
| *Oribacterium* | 2.0 | 0.219 | 0.688 (0.451-0.861) | 0.004 | 80% | 77% |
| 0.021 | 20% | 100% |
| *Flavobacterium* | 0.46 | 0.651 | 0.565 (0.322-0.756) | 0.0001 | 20% | 94% |
| **AIN** | *Sutterella* | 0.19 | 0.752 | 0.494 (0.282-0.718) | 0.0003 | 94% | 20% |
| 0.033 | 6% | 100% |
| **HSIL** | ***Ruminococcus*** | **3.50** | **0.088** | **0.800 (0.566-0.962)** | 0.125 | 100% | 30% |
| 0.070 | 20% | 100% |
| The coefficient and P values for each genus were calculated using a series of logistic regression models. Genera marked in bold represent those with a combination of a P value <0.10 and AUCROC ≥0.800. Demonstrative cut-offs optimizing either sensitivity or specificity are shown for each biomarker. | | | | | | | | |

**SUPPLEMENTARY FIGURES**

**Figure S1. Principal Coordinates Analysis derived from pairwise weighted Unifrac distances of 16S rRNA gene from bacterial communities in feces (blue) and rectal mucosa (red).**

**Figure S2. Relative abundance of bacterial genera in feces and rectal mucosa.**

**SUPPLEMENTARY REFERENCES**

1 Solomon D, Nayar R. The Bethesda system for reporting cervical cytology: definitions. Criteria and explanatory notes. 2nd ed. New York: Springer; 2004. pp. 1–191.

2 Caporaso J, Kuczynski J, Stombaugh J, Bittinger K, Bushman F, Costello E, *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010; **7**:335–336.

3 Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010; **26**:2460–1.

4 Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 2011; **27**:2194–200.

5 Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, *et al.* The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* 2009; **37**:D141-5.

6 DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 2006; **72**:5069–72.

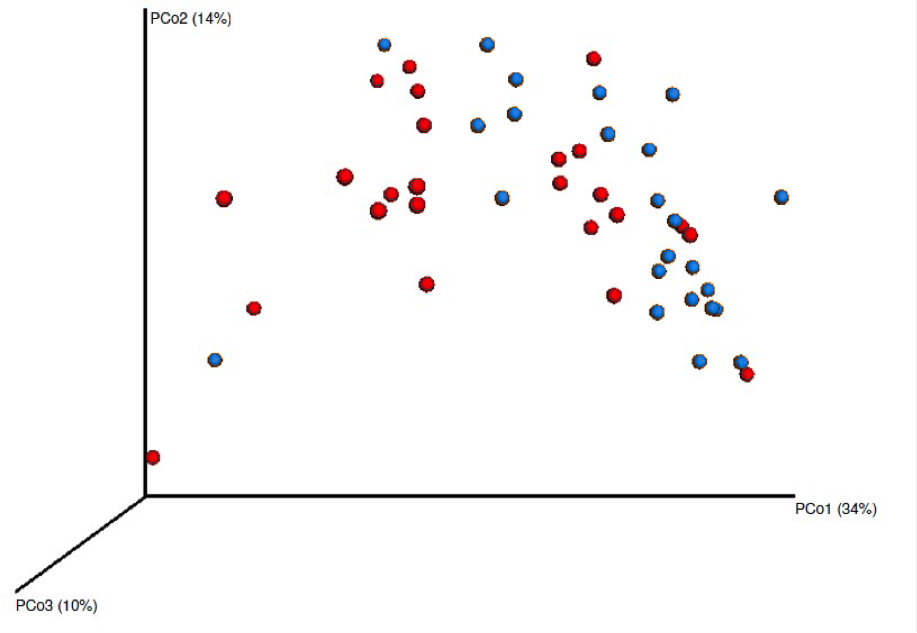
7 Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 2010; **26**:266–7.

8 Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 2010; **5**:e9490.

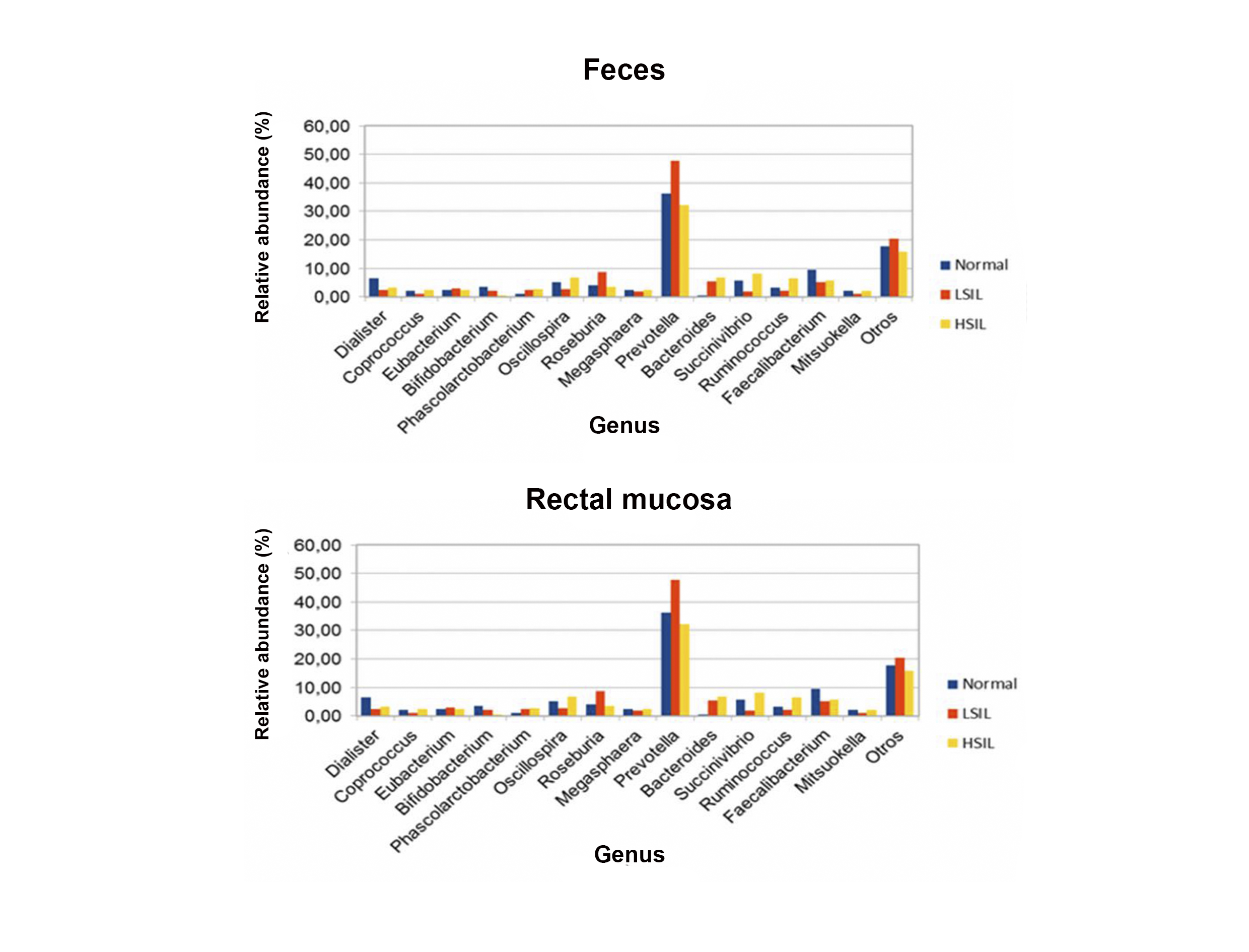
9 Shannon C. A Mathematical Theory of Communication. *Bell Syst Tech J* 1948; **27**:379–423.

10 Chao A, Hwang W, Chen Y, Kuo C. ESTIMATING THE NUMBER OF SHARED SPECIES. *Stat Sin* 2000; **10**:227–246.

11 Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol* 2011; **12**:R60.



**Figure S1. Principal Coordinates Analysis derived from pairwise weighted Unifrac distances of 16S rRNA gene from bacterial communities in feces (blue) and rectal mucosa (red).**

****

**Figure S2. Relative abundance of bacterial genera in feces and rectal mucosa.**