

Supplementary Table 1. Details of the human subjects from three cohorts investigated in this study and three published studies: kurokawa07⁶, gill06⁴ and turnbaugh09⁵. *, †, ‡, & - subjects with the same sign are familially related to each other. Ethnicity of each subject is denoted by the superscript in “Nationality” column: ^j - Japanese, ^c - Caucasian.

| Internal ID | Sample ID | Project | Sample Name | Nationality | Gender | Age | Clinical Status | BMI |
|-------------|-----------|-------------|-------------|-----------------------|--------|------|-----------------|-------|
| MC20.MG29 | DA-AD-1 | MetaHIT | MH6 | danish ^c | F | 59 | healthy | 22.38 |
| MC20.MG30 | DA-AD-2 | MetaHIT | MH13 | danish ^c | M | 54 | healthy | 20.46 |
| MC20.MG31 | DA-AD-3 | MetaHIT | MH12 | danish ^c | F | 49 | obese | 32.1 |
| MC20.MG32 | DA-AD-4 | MetaHIT | MH30 | danish ^c | M | 59 | obese | 35.21 |
| MC20.MG33 | ES-AD-1 | MetaHIT | CD1 | spanish ^c | F | 25 | CD* | 17.9 |
| MC20.MG34 | ES-AD-2 | MetaHIT | CD2 | spanish ^c | M | 49 | healthy* | 27.8 |
| MC20.MG35 | ES-AD-3 | MetaHIT | UC4 | spanish ^c | F | 47 | UC | 26.37 |
| MC20.MG36 | ES-AD-4 | MetaHIT | UC6 | spanish ^c | F | 38 | healthy | 23.18 |
| MC20.MG22 | FR-AD-1 | MicroObes | NO1 | french ^c | M | 63 | healthy | 23.1 |
| MC20.MG23 | FR-AD-2 | MicroObes | NO3 | french ^c | M | 61 | healthy | 22 |
| MC20.MG24 | FR-AD-3 | MicroObes | NO4 | french ^c | M | 60 | healthy | 23.8 |
| MC20.MG25 | FR-AD-4 | MicroObes | NO8 | french ^c | M | 60 | healthy | 21.9 |
| MC20.MG27 | FR-AD-5 | MicroObes | OB2 | french ^c | M | 64 | obese | 30.8 |
| MC20.MG26 | FR-AD-6 | MicroObes | OB1 | french ^c | M | 63 | obese | 33.7 |
| MC20.MG28 | FR-AD-7 | MicroObes | OB6 | french ^c | M | 62 | obese | 28.9 |
| MC20.MG37 | FR-AD-8 | MicroObes | OB8 | french ^c | M | 60 | obese | 32 |
| MC20.MG16 | IT-AD-1 | MicroAge | A | italian ^c | F | 84 | elderly | |
| MC20.MG17 | IT-AD-2 | MicroAge | B | italian ^c | M | 87 | elderly | |
| MC20.MG18 | IT-AD-3 | MicroAge | C | italian ^c | F | 77 | elderly | |
| MC20.MG19 | IT-AD-4 | MicroAge | D | italian ^c | M | 80 | elderly | |
| MC20.MG20 | IT-AD-5 | MicroAge | E | italian ^c | M | 70 | elderly | |
| MC20.MG21 | IT-AD-6 | MicroAge | G | italian ^c | F | 72 | elderly | |
| MC20.MG3 | JP-AD-1 | kurokawa07 | F1-S | japanese ^j | M | 30 | healthy† | |
| MC20.MG4 | JP-AD-2 | kurokawa07 | F1-T | japanese ^j | F | 28 | healthy† | |
| MC20.MG5 | JP-AD-3 | kurokawa07 | F2-V | japanese ^j | M | 37 | healthy‡ | |
| MC20.MG6 | JP-AD-4 | kurokawa07 | F2-W | japanese ^j | F | 36 | healthy‡ | |
| MC20.MG7 | JP-AD-5 | kurokawa07 | F2-X | japanese ^j | M | 3 | healthy‡ | |
| MC20.MG8 | JP-AD-6 | kurokawa07 | F2-Y | japanese ^j | F | 1.5 | healthy‡ | |
| MC20.MG9 | JP-AD-7 | kurokawa07 | In-A | japanese ^j | M | 45 | healthy | |
| MC20.MG10 | JP-AD-8 | kurokawa07 | In-D | japanese ^j | M | 35 | healthy | |
| MC20.MG11 | JP-AD-9 | kurokawa07 | In-R | japanese ^j | F | 24 | healthy | |
| MC20.MG12 | JP-IN-1 | kurokawa07 | F1-U | japanese ^j | F | 0.58 | healthy† | |
| MC20.MG13 | JP-IN-2 | kurokawa07 | In-B | japanese ^j | M | 0.5 | healthy | |
| MC20.MG14 | JP-IN-3 | kurokawa07 | In-E | japanese ^j | M | 0.25 | healthy | |
| MC20.MG15 | JP-IN-4 | kurokawa07 | In-M | japanese ^j | F | 0.33 | healthy | |
| MC20.MG1 | AM-AD-1 | gill06 | Subject7 | american | F | 28 | healthy | |
| MC20.MG2 | AM-AD-2 | gill06 | Subject8 | american | M | 37 | healthy | |
| MC16.MG13 | AM-F10-T1 | turnbaugh09 | F10T1Ob1 | american ^c | F | | obese& | |
| MC16.MG14 | AM-F10-T2 | turnbaugh09 | F10T2Ob1 | american ^c | F | | obese& | |

Supplementary Table 2. Summary statistics of the metagenome sequences used: raw sequence details, assembled contigs and predicted protein coding genes. NR sequences: non-redundant sequences after merging contigs and unassembled reads.

| Sample ID | Sample size (Mb) | Reads | Singleton reads | Contigs | Contig length (Mb) | NR sequences | NR sequence length (Mb) | Genes | Coding length (Mb), fraction |
|-----------|------------------|--------|-----------------|---------|--------------------|--------------|-------------------------|--------|------------------------------|
| DA-AD-1 | 156.96 | 237710 | 85700 (36.1%) | 19816 | 31.36 | 105516 | 86.75 | 152959 | 76.29 (87.94%) |
| DA-AD-2 | 146.77 | 224711 | 80256 (35.7%) | 18910 | 32.89 | 99166 | 83.35 | 147519 | 73.46 (88.13%) |
| DA-AD-3 | 154.69 | 231024 | 88736 (38.4%) | 21465 | 36.11 | 110201 | 93.51 | 162534 | 84.34 (90.20%) |
| DA-AD-4 | 150.17 | 227411 | 91405 (40.2%) | 22135 | 37.52 | 113540 | 96.15 | 167530 | 84.80 (88.19%) |
| ES-AD-1 | 144.87 | 223746 | 50190 (22.4%) | 14898 | 32.23 | 65088 | 63.58 | 102806 | 56.00 (88.08%) |
| ES-AD-2 | 151.91 | 230738 | 69752 (30.2%) | 15257 | 26.1 | 85009 | 70.6 | 122628 | 61.74 (87.45%) |
| ES-AD-3 | 147.49 | 236855 | 78396 (33.1%) | 20260 | 32.84 | 98656 | 79.89 | 140465 | 70.26 (87.95%) |
| ES-AD-4 | 144.35 | 229783 | 90695 (39.5%) | 24863 | 38.63 | 115558 | 94.54 | 166469 | 83.80 (88.63%) |
| FR-AD-1 | 85.90 | 125260 | 66486 (53.1%) | 15390 | 22.43 | 81876 | 67.37 | 118183 | 59.53 (88.37%) |
| FR-AD-2 | 73.83 | 113507 | 61151 (53.9%) | 12439 | 18.1 | 73590 | 57.13 | 103732 | 50.48 (88.37%) |
| FR-AD-3 | 75.06 | 115862 | 55637 (48.0%) | 14694 | 21.46 | 70331 | 57.19 | 100309 | 50.49 (88.29%) |
| FR-AD-4 | 79.60 | 120268 | 72738 (60.5%) | 14808 | 19.41 | 87546 | 67.19 | 122497 | 59.14 (88.03%) |
| FR-AD-5 | 85.84 | 129745 | 70637 (54.4%) | 13294 | 20.11 | 83931 | 66.17 | 119784 | 58.63 (88.60%) |
| FR-AD-6 | 75.84 | 118423 | 64043 (54.1%) | 14112 | 19.33 | 78155 | 59.65 | 109207 | 52.21 (87.53%) |
| FR-AD-7 | 76.44 | 118172 | 56166 (47.5%) | 14994 | 21.57 | 71160 | 57.08 | 101769 | 50.44 (88.37%) |
| FR-AD-8 | 71.98 | 112592 | 64959 (57.7%) | 12266 | 16.38 | 77225 | 57.17 | 106497 | 50.42 (88.19%) |
| IT-AD-1 | 76.65 | 116244 | 43644 (37.5%) | 13489 | 21.22 | 57133 | 49.2 | 84781 | 43.91 (89.25%) |
| IT-AD-2 | 79.21 | 115636 | 47103 (40.7%) | 12461 | 21.52 | 59564 | 53.02 | 90859 | 47.54 (89.66%) |
| IT-AD-3 | 78.98 | 116746 | 57795 (49.5%) | 16029 | 22.92 | 73824 | 61.49 | 107924 | 54.55 (88.71%) |
| IT-AD-4 | 80.28 | 116891 | 31691 (27.1%) | 6606 | 15.12 | 38297 | 36.28 | 58967 | 31.47 (86.74%) |
| IT-AD-5 | 80.80 | 118227 | 62846 (53.2%) | 14236 | 20.94 | 77082 | 63.14 | 111891 | 56.13 (88.90%) |
| IT-AD-6 | 80.39 | 116085 | 61669 (53.1%) | 13766 | 20.16 | 75435 | 62.14 | 108567 | 55.04 (88.58%) |
| JP-AD-1 | 59.27 | 78123 | 16561 (21.2%) | 14535 | 24.1 | 31096 | 35.58 | 54856 | 30.43 (85.54%) |
| JP-AD-2 | 59.94 | 80477 | 22788 (28.3%) | 14961 | 22.99 | 37749 | 39.02 | 63230 | 33.95 (87.00%) |
| JP-AD-3 | 60.80 | 79846 | 20442 (25.6%) | 17351 | 27.1 | 37793 | 41.31 | 64201 | 35.79 (86.63%) |
| JP-AD-4 | 60.50 | 78670 | 17634 (22.4%) | 13537 | 23.51 | 31171 | 36.16 | 55693 | 31.17 (86.20%) |
| JP-AD-5 | 61.24 | 79773 | 19383 (24.3%) | 12302 | 20.54 | 31685 | 34.47 | 54699 | 29.86 (86.64%) |
| JP-AD-6 | 61.43 | 79357 | 21669 (27.3%) | 15134 | 26.06 | 36803 | 41.72 | 63735 | 36.10 (86.54%) |
| JP-AD-7 | 53.29 | 75532 | 15765 (20.9%) | 5327 | 14.35 | 21092 | 24.49 | 37212 | 21.09 (86.11%) |
| JP-AD-8 | 60.41 | 80627 | 28252 (35.0%) | 10390 | 19.86 | 38642 | 39.89 | 64333 | 34.23 (85.82%) |
| JP-AD-9 | 61.02 | 81346 | 17969 (22.1%) | 16420 | 25.39 | 34389 | 38.23 | 59820 | 32.52 (85.08%) |
| JP-IN-1 | 59.63 | 80796 | 11452 (14.2%) | 6136 | 15.35 | 17588 | 22.82 | 33993 | 19.10 (83.71%) |
| JP-IN-2 | 66.43 | 79972 | 5120 (6.4%) | 1671 | 6.75 | 6791 | 10.69 | 14334 | 9.14 (85.51%) |
| JP-IN-3 | 62.21 | 79787 | 10324 (12.9%) | 5647 | 12.22 | 15971 | 19.47 | 29305 | 16.57 (85.09%) |
| JP-IN-4 | 62.86 | 87324 | 11137 (12.8%) | 6665 | 18.19 | 17802 | 23.88 | 34732 | 20.62 (86.34%) |
| AM-AD-1 | 58.66 | 65042 | 34718 (53.4%) | 7113 | 15.16 | 41831 | 46.14 | 72772 | 40.38 (87.52%) |
| AM-AD-2 | 68.39 | 74452 | 27947 (37.5%) | 9501 | 20.65 | 37448 | 46 | 69574 | 40.50 (88.03%) |
| AM-F10-T1 | 87.04 | 248939 | 117041 (47.0%) | 33379 | 14.52 | 150420 | 52.68 | 152956 | 39.09 (74.21%) |
| AM-F10-T2 | 153.13 | 435911 | 132093 (30.3%) | 46287 | 28.94 | 178380 | 70.44 | 188665 | 53.24 (75.58%) |

Supplementary Table 3. List of 1511 complete and draft microbial genomes used as a reference set. These were obtained from the National Center for Biological Information (NCBI), Human Microbiome Project¹⁰ (HMP) and Sanger Center¹¹ (MetaHIT). Actual NCBI taxonomy identifiers are used where possible, and closest identifiers are used otherwise. Genomes of gut-specific species are marked with an asterisk. Taxonomic classification of these genomes via their 16S rRNA gene sequences using RDP Classifier³⁹ is provided in the last two columns (minimum length 500bp, minimum confidence threshold 0.8).

(Submitted as a separate document)

Supplementary Table 4. Genera that are almost always among the five most abundant from their respective phyla. Top 5 occurrences are counted in 35 samples. * - These genera are also among the 10 most abundant genera overall, in HITChip analysis of 22 European samples (See Supplementary Notes Section 4.5).

| Genus | Phylum | Top 5 |
|---------------------------|----------------|-------|
| <i>Faecalibacterium</i> * | Firmicutes | 33 |
| <i>Bacteroides</i> * | Bacteroidetes | 35 |
| <i>Parabacteroides</i> * | Bacteroidetes | 34 |
| <i>Alistipes</i> * | Bacteroidetes | 34 |
| <i>Bifidobacterium</i> | Actinobacteria | 35 |
| <i>Collinsella</i> | Actinobacteria | 32 |
| <i>Eggerthella</i> | Actinobacteria | 31 |

Supplementary Table 5. Orthologous group assignment rates, measured by the number of genes in each sample that can be assigned to an orthologous group in eggNOG database.

| Subject ID | Genes | OG mapped genes (%) |
|------------|--------|---------------------|
| DA-AD-1 | 152959 | 92517 (60.48%) |
| DA-AD-2 | 147519 | 89997 (61.01%) |
| DA-AD-3 | 162534 | 101144 (62.23%) |
| DA-AD-4 | 167530 | 96916 (57.85%) |
| ES-AD-1 | 102806 | 69312 (67.42%) |
| ES-AD-2 | 122628 | 75200 (61.32%) |
| ES-AD-3 | 140465 | 91523 (65.16%) |
| ES-AD-4 | 166469 | 104175 (62.58%) |
| FR-AD-1 | 118183 | 75343 (63.75%) |
| FR-AD-2 | 103732 | 62860 (60.6%) |
| FR-AD-3 | 100309 | 65051 (64.85%) |
| FR-AD-4 | 122497 | 75029 (61.25%) |
| FR-AD-5 | 119784 | 72765 (60.75%) |
| FR-AD-6 | 109207 | 66658 (61.04%) |
| FR-AD-7 | 101769 | 63132 (62.03%) |
| FR-AD-8 | 106497 | 66006 (61.98%) |
| IT-AD-1 | 84781 | 57173 (67.44%) |

| | | |
|-----------|--------|----------------|
| IT-AD-2 | 90859 | 59994 (66.03%) |
| IT-AD-3 | 107924 | 69317 (64.23%) |
| IT-AD-4 | 58967 | 38077 (64.57%) |
| IT-AD-5 | 111891 | 70115 (62.66%) |
| IT-AD-6 | 108567 | 68786 (63.36%) |
| JP-AD-1 | 54856 | 35601 (64.9%) |
| JP-AD-2 | 63230 | 39654 (62.71%) |
| JP-AD-3 | 64201 | 42916 (66.85%) |
| JP-AD-4 | 55693 | 36941 (66.33%) |
| JP-AD-5 | 54699 | 34814 (63.65%) |
| JP-AD-6 | 63735 | 43128 (67.67%) |
| JP-AD-7 | 37212 | 24829 (66.72%) |
| JP-AD-8 | 64333 | 40709 (63.28%) |
| JP-AD-9 | 59820 | 38200 (63.86%) |
| JP-IN-1 | 33993 | 27306 (80.33%) |
| JP-IN-2 | 14334 | 10400 (72.55%) |
| JP-IN-3 | 29305 | 19181 (65.45%) |
| JP-IN-4 | 34732 | 23957 (68.98%) |
| AM-AD-1 | 72772 | 45958 (63.15%) |
| AM-AD-2 | 69574 | 44750 (64.32%) |
| AM-F10-T1 | 152956 | 61800 (40.40%) |
| AM-F10-T2 | 188665 | 78441 (41.58%) |

Supplementary Table 6. List of 122 functions which are abundant (among the top 20%, which is equivalent to above 80th percentile) and are primarily from low-abundance genera (<2.5% relative abundance). Samples where this was observed are also listed.

| Orthologous group | Description | Samples | Count |
|-------------------|--|---|-------|
| NOG137454 | Annotation not available | DA-AD-2,DA-AD-3,ES-AD-2,FR-AD-1,FR-AD-5,IT-AD-1,IT-AD-2,IT-AD-3,IT-AD-5,IT-AD-6,JP-AD-5,JP-AD-7 | 12 |
| COG2826 | Transposase and inactivated derivatives, IS30 family | ES-AD-1,ES-AD-4,FR-AD-3,FR-AD-4,FR-AD-7,FR-AD-8,IT-AD-1 | 7 |
| NOG44176 | Atp/Gtp-Binding protein | AM-F10-T1,AM-F10-T2,FR-AD-2,FR-AD-4,FR-AD-8,JP-AD-9 | 6 |
| NOG68428 | Annotation not available | AM-F10-T1,AM-F10-T2,ES-AD-4,IT-AD-3,JP-AD-6,JP-AD-8 | 6 |
| NOG120133 | Abc transporter, permease | DA-AD-2,DA-AD-4,FR-AD-5,FR-AD-6,FR-AD-7,JP-AD-9 | 6 |
| COG4422 | Bacteriophage protein gp37 | AM-F10-T2,DA-AD-4,JP-AD-2,JP-AD-5,JP-AD-9 | 5 |
| COG3077 | DNA-damage-inducible protein J | AM-F10-T1,DA-AD-2,ES-AD-4,IT-AD-1,IT-AD-3 | 5 |

| | | | |
|-----------|--|-----------------------------------|---|
| NOG27013 | RNA polymerase | AM-F10-T1,IT-AD-1,JP-AD-4,JP-AD-7 | 4 |
| NOG12663 | Annotation not available | DA-AD-2,FR-AD-7,IT-AD-1,JP-AD-2 | 4 |
| NOG120367 | Annotation not available | DA-AD-2,FR-AD-1,IT-AD-5,JP-AD-9 | 4 |
| NOG69420 | Phage associated protein | DA-AD-3,FR-AD-2,JP-AD-5 | 3 |
| NOG14713 | Annotation not available | IT-AD-1,JP-AD-5,JP-AD-7 | 3 |
| COG1289 | Predicted membrane protein | DA-AD-4,IT-AD-5,JP-AD-2 | 3 |
| COG3041 | Uncharacterized protein conserved in bacteria | AM-F10-T1,IT-AD-1,IT-AD-3 | 3 |
| COG3598 | RecA-family ATPase | AM-F10-T2,DA-AD-4,FR-AD-4 | 3 |
| NOG09639 | Annotation not available | AM-F10-T1,AM-F10-T2,FR-AD-7 | 3 |
| COG2088 | Uncharacterized protein, involved in the regulation of septum location | DA-AD-3,JP-AD-2,JP-AD-7 | 3 |
| NOG45139 | Transposon protein | DA-AD-3,DA-AD-4 | 2 |
| COG5545 | Predicted P-loop ATPase and inactivated derivatives | IT-AD-1,JP-AD-4 | 2 |
| NOG40986 | Annotation not available | FR-AD-4,FR-AD-7 | 2 |
| NOG42453 | Transposon protein | DA-AD-3,IT-AD-2 | 2 |
| NOG79858 | Annotation not available | FR-AD-6,JP-AD-4 | 2 |
| NOG06430 | Conjugative transposon protein | AM-F10-T1,FR-AD-7 | 2 |
| COG3539 | P pilus assembly protein, pilin FimA | IT-AD-5,JP-AD-2 | 2 |
| NOG39150 | Protein found in conjugate transposon | DA-AD-3,DA-AD-4 | 2 |
| COG4734 | Antirestriction protein | AM-F10-T1,FR-AD-7 | 2 |
| NOG45681 | Mobilization protein | DA-AD-4,IT-AD-2 | 2 |
| NOG131675 | Annotation not available | DA-AD-4,JP-AD-7 | 2 |
| NOG139537 | Annotation not available | IT-AD-1,JP-AD-5 | 2 |
| NOG10218 | Conjugative protein | AM-F10-T1,AM-F10-T2 | 2 |
| NOG83248 | Annotation not available | DA-AD-1,JP-AD-7 | 2 |
| NOG127293 | Annotation not available | ES-AD-2,FR-AD-5 | 2 |
| NOG43914 | Annotation not available | DA-AD-3,IT-AD-2 | 2 |
| NOG43858 | Transposon protein | DA-AD-3,DA-AD-4 | 2 |
| COG2222 | Predicted phosphosugar isomerases | IT-AD-2 | 1 |
| COG3158 | K ⁺ transporter | IT-AD-5 | 1 |
| COG1748 | Saccharopine dehydrogenase and related proteins | JP-AD-2 | 1 |
| COG0093 | Ribosomal protein L14 | JP-AD-3 | 1 |
| COG4804 | Uncharacterized conserved protein | JP-AD-2 | 1 |
| NOG44566 | Permease protein | JP-AD-7 | 1 |
| COG0186 | Ribosomal protein S17 | JP-AD-3 | 1 |
| NOG41233 | Transposase | JP-AD-3 | 1 |
| COG3468 | Type V secretory pathway, adhesin AidA | IT-AD-5 | 1 |
| COG0666 | FOG: Ankyrin repeat | JP-AD-8 | 1 |
| NOG84056 | Annotation not available | IT-AD-1 | 1 |
| COG3328 | Transposase and inactivated | ES-AD-3 | 1 |

| | | | |
|-----------|---|-----------|---|
| | derivatives | | |
| NOG139663 | Sequence-Specific DNA binding protein | FR-AD-7 | 1 |
| COG3188 | P pilus assembly protein, porin PapC | IT-AD-5 | 1 |
| COG3385 | FOG: Transposase and inactivated derivatives | IT-AD-2 | 1 |
| NOG45347 | Annotation not available | ES-AD-1 | 1 |
| NOG28380 | Transposase for | AM-F10-T2 | 1 |
| COG0074 | Succinyl-CoA synthetase, alpha subunit | IT-AD-5 | 1 |
| NOG13143 | Conjugative transposon protein | FR-AD-7 | 1 |
| NOG140504 | Annotation not available | IT-AD-1 | 1 |
| NOG10311 | Type III restriction enzyme | IT-AD-5 | 1 |
| COG2946 | Putative phage replication protein RstA | AM-F10-T1 | 1 |
| NOG134467 | Annotation not available | DA-AD-4 | 1 |
| NOG44942 | Annotation not available | DA-AD-3 | 1 |
| COG2015 | Alkyl sulfatase and related hydrolases | JP-AD-5 | 1 |
| NOG80481 | Annotation not available | DA-AD-4 | 1 |
| NOG150295 | Annotation not available | AM-F10-T1 | 1 |
| NOG116632 | Annotation not available | IT-AD-1 | 1 |
| NOG79696 | Protein involved in carbohydrate biosynthetic process | JP-AD-5 | 1 |
| NOG131524 | Annotation not available | JP-AD-7 | 1 |
| COG4725 | Transcriptional activator, adenine-specific DNA methyltransferase | JP-AD-9 | 1 |
| COG2357 | Uncharacterized protein conserved in bacteria | JP-AD-1 | 1 |
| NOG70669 | Annotation not available | DA-AD-4 | 1 |
| NOG47099 | Transposon protein | DA-AD-3 | 1 |
| COG1528 | Ferritin-like protein | IT-AD-5 | 1 |
| NOG72324 | Integrase | ES-AD-4 | 1 |
| NOG86034 | DNA primase | JP-AD-4 | 1 |
| NOG112926 | Annotation not available | DA-AD-4 | 1 |
| NOG81060 | Annotation not available | DA-AD-4 | 1 |
| COG4474 | Uncharacterized protein conserved in bacteria | JP-AD-5 | 1 |
| COG1672 | Predicted ATPase (AAA+ superfamily) | IT-AD-2 | 1 |
| NOG118622 | Annotation not available | FR-AD-4 | 1 |
| COG0137 | Argininosuccinate synthase | JP-AD-5 | 1 |
| COG0518 | GMP synthase - Glutamine amidotransferase domain | AM-F10-T1 | 1 |
| NOG81629 | Annotation not available | JP-AD-7 | 1 |
| NOG12010 | Aminoglycoside protein | IT-AD-4 | 1 |
| NOG125746 | Transcriptional regulator protein | DA-AD-3 | 1 |

| | | | |
|-----------|--|-----------|---|
| COG1476 | Predicted transcriptional regulators | AM-F10-T1 | 1 |
| COG1080 | Phosphoenolpyruvate-protein kinase (PTS system EI component in bacteria) | IT-AD-5 | 1 |
| COG1533 | DNA repair photolyase | JP-AD-5 | 1 |
| NOG116483 | Annotation not available | DA-AD-4 | 1 |
| NOG87940 | Annotation not available | DA-AD-3 | 1 |
| COG1250 | 3-hydroxyacyl-CoA dehydrogenase | IT-AD-5 | 1 |
| NOG25785 | Annotation not available | DA-AD-4 | 1 |
| NOG128643 | Annotation not available | DA-AD-4 | 1 |
| NOG12324 | Annotation not available | FR-AD-7 | 1 |
| NOG20054 | Lantibiotic ABC transporter permease | ES-AD-4 | 1 |
| NOG18439 | Transposase | DA-AD-4 | 1 |
| NOG79506 | DNA primase | JP-AD-5 | 1 |
| NOG68338 | Transporter permease | IT-AD-1 | 1 |
| NOG44869 | Mobilization protein | DA-AD-4 | 1 |
| NOG120818 | Annotation not available | JP-AD-5 | 1 |
| NOG42679 | Protein found in conjugate transposon | DA-AD-3 | 1 |
| COG1321 | Mn-dependent transcriptional regulator | JP-AD-7 | 1 |
| COG3973 | Superfamily I DNA and RNA helicases | IT-AD-6 | 1 |
| NOG119061 | Annotation not available | DA-AD-4 | 1 |
| NOG145717 | Protein involved in DNA integration | DA-AD-3 | 1 |
| COG2252 | Permeases | IT-AD-5 | 1 |
| NOG134563 | Annotation not available | JP-AD-7 | 1 |
| NOG112830 | Recombination protein | FR-AD-6 | 1 |
| NOG47313 | Annotation not available | DA-AD-3 | 1 |
| COG4509 | Uncharacterized protein conserved in bacteria | JP-AD-7 | 1 |
| COG3340 | Peptidase E | JP-AD-5 | 1 |
| NOG16015 | Plasmid recombination protein | JP-AD-8 | 1 |
| NOG133538 | Transposase | ES-AD-1 | 1 |
| NOG82576 | Annotation not available | DA-AD-4 | 1 |
| COG0143 | Methionyl-tRNA synthetase | JP-AD-5 | 1 |
| NOG13016 | Annotation not available | JP-AD-8 | 1 |
| COG3935 | Putative primosome component and related proteins | IT-AD-1 | 1 |
| COG1429 | Cobalamin biosynthesis protein CobN and related Mg-chelataes | AM-F10-T1 | 1 |
| COG1349 | Transcriptional regulators of sugar metabolism | IT-AD-5 | 1 |
| NOG126306 | Annotation not available | JP-AD-5 | 1 |
| COG3655 | Predicted transcriptional regulator | JP-AD-8 | 1 |
| COG4200 | Uncharacterized protein conserved | FR-AD-6 | 1 |

| | | | |
|---------|--|---------|---|
| | in bacteria | | |
| COG1172 | Ribose/xylose/arabinose/galactoside ABC-type transport systems, permease components | JP-AD-7 | 1 |
| COG1484 | DNA replication protein | ES-AD-4 | 1 |
| COG1683 | Uncharacterized conserved protein | JP-AD-9 | 1 |
| COG1009 | NADH:ubiquinone oxidoreductase subunit 5 (chain L)/Multisubunit Na ⁺ /H ⁺ antiporter, MnhA subunit | IT-AD-5 | 1 |

Supplementary Table 7. Abundant functions (among the top 20%, which is equivalent to above 80th percentile) from low-abundance genera (<2.5% relative abundance). Ranks of these functions are listed as percentiles. The abundance of each genus in the corresponding sample and its contribution to the listed function in that sample are also listed.

| Function | Description | Sample | Rank of function in sample (percentile) | Genus | abundance of genus in sample | contribution of genus to function |
|----------|---|-------------------------|---|-----------------------------|------------------------------|-----------------------------------|
| COG3041 | Uncharacterized protein conserved in bacteria (Toxin of the YafQ-DinJ toxin-antitoxin system) | AM-F10-T1 | 83.38 | <i>Eubacterium</i> | 0.74% | 67.08% |
| | | | | <i>Blautia</i> | 5.83% | 13.60% |
| | | | | <i>Faecalibacterium</i> | 8.19% | 8.36% |
| | | | | Lachnospiraceae | 5.02% | 2.79% |
| | | IT-AD-1 | 94.53 | <i>Blautia</i> | 0.80% | 25.86% |
| | | | | Peptostreptococcaceae | 0.08% | 22.72% |
| | | | | <i>Eubacterium</i> | 0.90% | 22.45% |
| | | | | <i>Bacteroides</i> | 17.29% | 7.07% |
| | | | | <i>Escherichia/Shigella</i> | 1.65% | 4.45% |
| | | | | Lachnospiraceae | 2.66% | 3.88% |
| | | | | <i>Parabacteroides</i> | 2.45% | 3.80% |
| | | | | Clostridiales | 0.45% | 0.21% |
| IT-AD-3 | 81.26 | <i>Faecalibacterium</i> | 3.90% | 0.21% | | |
| | | <i>Blautia</i> | 1.00% | 43.77% | | |
| | | <i>Eubacterium</i> | 0.47% | 8.05% | | |
| | | <i>Coprococcus</i> | 1.67% | 7.11% | | |
| | | <i>Bifidobacterium</i> | 5.73% | 7.02% | | |
| | | Clostridiales | 0.11% | 5.86% | | |
| COG3077 | DNA-damage-inducible protein J | AM-F10-T1 | 89.89 | <i>Eubacterium</i> | 0.74% | 55.27% |
| | | | | <i>Blautia</i> | 5.83% | 10.24% |
| | | | | Lachnospiraceae | 5.02% | 5.27% |
| | | | | <i>Roseburia</i> | 5.21% | 5.27% |
| | | | | Peptostreptococcaceae | 0.08% | 4.29% |
| | | | | <i>Ruminococcus</i> | 4.02% | 3.65% |
| | | | | <i>Coprococcus</i> | 0.83% | 2.63% |
| | | | | Clostridiales | 0.15% | 2.58% |
| | | | | <i>Faecalibacterium</i> | 8.19% | 2.58% |

| | | | | | | |
|---------|--------------------------------------|---------|-------|-----------------------------|-------|--------|
| | | | | <i>Bifidobacterium</i> | 5.53% | 0.35% |
| | | | | <i>Megasphaera</i> | 0.09% | 0.35% |
| | | DA-AD-2 | 82.31 | <i>Eubacterium</i> | 0.73% | 47.02% |
| | | | | <i>Faecalibacterium</i> | 0.85% | 16.89% |
| | | | | <i>Blautia</i> | 0.56% | 10.32% |
| | | | | Lachnospiraceae | 1.33% | 6.88% |
| | | | | Clostridiales | 0.13% | 1.72% |
| | | ES-AD-4 | 80.23 | <i>Blautia</i> | 0.81% | 40.31% |
| | | | | <i>Faecalibacterium</i> | 3.70% | 15.62% |
| | | | | <i>Coprococcus</i> | 1.05% | 12.34% |
| | | | | Lachnospiraceae | 3.47% | 11.91% |
| | | | | <i>Eubacterium</i> | 0.42% | 4.11% |
| | | | | Clostridiales | 0.17% | 3.29% |
| | | | | <i>Ruminococcus</i> | 0.96% | 2.81% |
| | | IT-AD-1 | 89.36 | <i>Eubacterium</i> | 0.90% | 29.95% |
| | | | | <i>Blautia</i> | 0.80% | 21.24% |
| | | | | Lachnospiraceae | 2.66% | 16.60% |
| | | | | Peptostreptococcaceae | 0.08% | 11.23% |
| | | | | <i>Ruminococcus</i> | 6.71% | 5.34% |
| | | | | <i>Escherichia/Shigella</i> | 1.65% | 4.83% |
| | | IT-AD-3 | 82.38 | <i>Blautia</i> | 0.99% | 27.02% |
| | | | | <i>Faecalibacterium</i> | 7.41% | 17.01% |
| | | | | <i>Coprococcus</i> | 1.67% | 15.35% |
| | | | | <i>Eubacterium</i> | 0.47% | 11.38% |
| | | | | Clostridiales | 0.11% | 5.33% |
| COG3188 | P pilus assembly protein, porin PapC | IT-AD-5 | 83.00 | <i>Escherichia/Shigella</i> | 2.03% | 92.07% |
| COG3539 | P pilus assembly protein, pilin FimA | IT-AD-5 | 83.43 | <i>Escherichia/Shigella</i> | 2.03% | 96.25% |
| | | JP-AD-2 | 82.32 | <i>Escherichia/Shigella</i> | 1.71% | 87.33% |
| | | | | <i>Citrobacter</i> | 0.11% | 5.37% |

Supplementary Table 8. Enterotype associations of the 33 Sanger-based metagenomes. ET1: enterotype 1, ET2: enterotype 2, ET3: enterotype 3.

| Sample | Enterotype |
|-----------|------------|
| AM-F10-T2 | ET3 |
| DA-AD-1 | ET2 |
| DA-AD-2 | ET3 |
| DA-AD-3 | ET3 |
| DA-AD-4 | ET2 |
| ES-AD-1 | ET1 |
| ES-AD-2 | ET2 |
| ES-AD-3 | ET2 |
| ES-AD-4 | ET3 |
| FR-AD-1 | ET3 |
| FR-AD-2 | ET3 |
| FR-AD-3 | ET1 |
| FR-AD-4 | ET3 |
| FR-AD-5 | ET3 |

| | |
|---------|-----|
| FR-AD-6 | ET2 |
| FR-AD-7 | ET3 |
| FR-AD-8 | ET3 |
| IT-AD-1 | ET3 |
| IT-AD-2 | ET3 |
| IT-AD-3 | ET3 |
| IT-AD-4 | ET2 |
| IT-AD-5 | ET3 |
| IT-AD-6 | ET3 |
| JP-AD-1 | ET1 |
| JP-AD-2 | ET3 |
| JP-AD-3 | ET3 |
| JP-AD-4 | ET1 |
| JP-AD-5 | ET3 |
| JP-AD-6 | ET1 |
| JP-AD-7 | ET1 |
| JP-AD-8 | ET1 |
| JP-AD-9 | ET1 |

Supplementary Table 9. Enterotype associations of the 154 pyrosequencing based 16S rRNA gene samples. ET1: enterotype 1, ET2: enterotype 2, ET3: enterotype 3.

| Sample | Enterotype |
|------------|------------|
| TS1_V2 | ET1 |
| TS10_V2 | ET1 |
| TS100_V2 | ET1 |
| TS101.2_V2 | ET3 |
| TS103_V2 | ET1 |
| TS104_V2 | ET1 |
| TS105_V2 | ET3 |
| TS106_V2 | ET3 |
| TS107_V2 | ET2 |
| TS109_V2 | ET1 |
| TS11_V2 | ET1 |
| TS110_V2 | ET3 |
| TS111_V2 | ET1 |
| TS115_V2 | ET2 |
| TS116_V2 | ET2 |
| TS117_V2 | ET1 |
| TS118_V2 | ET1 |
| TS119_V2 | ET1 |
| TS12_V2 | ET1 |
| TS120_V2 | ET2 |
| TS124_V2 | ET1 |
| TS125_V2 | ET1 |
| TS126_V2 | ET3 |
| TS127_V2 | ET1 |
| TS128_V2 | ET1 |

| | |
|------------|-----|
| TS129_V2 | ET1 |
| TS13_V2 | ET1 |
| TS130_V2 | ET1 |
| TS131_V2 | ET1 |
| TS132_V2 | ET1 |
| TS133_V2 | ET1 |
| TS134_V2 | ET3 |
| TS135_V2 | ET1 |
| TS136_V2 | ET3 |
| TS137_V2 | ET2 |
| TS138_V2 | ET3 |
| TS139_V2 | ET3 |
| TS14_V2 | ET3 |
| TS140_V2 | ET2 |
| TS141_V2 | ET1 |
| TS142_V2 | ET3 |
| TS143_V2 | ET1 |
| TS144_V2 | ET2 |
| TS145_V2 | ET1 |
| TS146_V2 | ET1 |
| TS147_V2 | ET1 |
| TS148_V2 | ET1 |
| TS149_V2 | ET1 |
| TS15_V2 | ET2 |
| TS150_V2 | ET3 |
| TS151_V2 | ET1 |
| TS152_V2 | ET1 |
| TS153_V2 | ET1 |
| TS154.2_V2 | ET1 |
| TS155_V2 | ET1 |
| TS156_V2 | ET1 |
| TS16_V2 | ET1 |
| TS160_V2 | ET2 |
| TS161_V2 | ET1 |
| TS162_V2 | ET1 |
| TS163_V2 | ET1 |
| TS164_V2 | ET1 |
| TS165_V2 | ET1 |
| TS166_V2 | ET1 |
| TS167_V2 | ET3 |
| TS168_V2 | ET3 |
| TS169_V2 | ET2 |
| TS17_V2 | ET1 |
| TS170_V2 | ET3 |
| TS178_V2 | ET3 |
| TS179_V2 | ET3 |
| TS180_V2 | ET1 |
| TS181_V2 | ET1 |

| | |
|----------|-----|
| TS182_V2 | ET1 |
| TS183_V2 | ET1 |
| TS184_V2 | ET1 |
| TS185_V2 | ET1 |
| TS186_V2 | ET1 |
| TS19_V2 | ET1 |
| TS190_V2 | ET1 |
| TS191_V2 | ET1 |
| TS192_V2 | ET1 |
| TS193_V2 | ET3 |
| TS194_V2 | ET1 |
| TS195_V2 | ET1 |
| TS2_V2 | ET1 |
| TS20_V2 | ET2 |
| TS21_V2 | ET3 |
| TS22_V2 | ET3 |
| TS23_V2 | ET3 |
| TS25_V2 | ET1 |
| TS26_V2 | ET1 |
| TS27_V2 | ET1 |
| TS28_V2 | ET1 |
| TS29_V2 | ET3 |
| TS3_V2 | ET1 |
| TS30_V2 | ET3 |
| TS31_V2 | ET1 |
| TS32_V2 | ET2 |
| TS33_V2 | ET3 |
| TS34_V2 | ET3 |
| TS35_V2 | ET2 |
| TS37_V2 | ET1 |
| TS38_V2 | ET3 |
| TS39_V2 | ET1 |
| TS4_V2 | ET1 |
| TS43_V2 | ET1 |
| TS44_V2 | ET3 |
| TS49_V2 | ET3 |
| TS5_V2 | ET3 |
| TS50_V2 | ET3 |
| TS51_V2 | ET3 |
| TS55_V2 | ET3 |
| TS56_V2 | ET1 |
| TS57_V2 | ET1 |
| TS6_V2 | ET2 |
| TS61_V2 | ET3 |
| TS62_V2 | ET1 |
| TS63_V2 | ET1 |
| TS64_V2 | ET1 |
| TS65_V2 | ET1 |

| | |
|-----------|-----|
| TS66_V2 | ET1 |
| TS67_V2 | ET1 |
| TS68_V2 | ET1 |
| TS69_V2 | ET1 |
| TS7_V2 | ET2 |
| TS70_V2 | ET1 |
| TS71_V2 | ET1 |
| TS72_V2 | ET1 |
| TS73_V2 | ET1 |
| TS74_V2 | ET2 |
| TS75_V2 | ET1 |
| TS76_V2 | ET1 |
| TS77_V2 | ET3 |
| TS78_V2 | ET2 |
| TS8_V2 | ET1 |
| TS82_V2 | ET2 |
| TS83_V2 | ET1 |
| TS84_V2 | ET2 |
| TS85_V2 | ET3 |
| TS86_V2 | ET1 |
| TS87_V2 | ET2 |
| TS88_V2 | ET1 |
| TS89_V2 | ET1 |
| TS9_V2 | ET1 |
| TS90_V2 | ET1 |
| TS91_V2 | ET3 |
| TS92_V2 | ET3 |
| TS94_V2 | ET1 |
| TS95_V2 | ET1 |
| TS96_V2 | ET1 |
| TS97_V2 | ET1 |
| TS98_V2 | ET1 |
| TS99.2_V2 | ET1 |

Supplementary Table 10. Enterotype associations of the 85 Illumina-based metagenomes. ET1: enterotype 1, ET2: enterotype 2, ET3: enterotype 3.

| Sample | Enterotype |
|--------|------------|
| MH0001 | ET2 |
| MH0002 | ET1 |
| MH0003 | ET1 |
| MH0004 | ET1 |
| MH0005 | ET2 |
| MH0006 | ET2 |
| MH0007 | ET1 |
| MH0008 | ET1 |
| MH0009 | ET3 |
| MH0010 | ET1 |

| | |
|--------|-----|
| MH0011 | ET1 |
| MH0012 | ET1 |
| MH0013 | ET1 |
| MH0014 | ET1 |
| MH0015 | ET1 |
| MH0016 | ET1 |
| MH0017 | ET1 |
| MH0018 | ET2 |
| MH0019 | ET1 |
| MH0020 | ET1 |
| MH0021 | ET1 |
| MH0022 | ET1 |
| MH0023 | ET3 |
| MH0024 | ET1 |
| MH0025 | ET1 |
| MH0026 | ET3 |
| MH0027 | ET1 |
| MH0028 | ET1 |
| MH0030 | ET3 |
| MH0031 | ET3 |
| MH0032 | ET2 |
| MH0033 | ET1 |
| MH0034 | ET2 |
| MH0035 | ET1 |
| MH0036 | ET1 |
| MH0037 | ET1 |
| MH0038 | ET1 |
| MH0039 | ET1 |
| MH0040 | ET1 |
| MH0041 | ET1 |
| MH0042 | ET1 |
| MH0043 | ET1 |
| MH0044 | ET1 |
| MH0045 | ET1 |
| MH0046 | ET3 |
| MH0047 | ET1 |
| MH0048 | ET1 |
| MH0049 | ET3 |
| MH0050 | ET3 |
| MH0051 | ET1 |
| MH0052 | ET1 |
| MH0053 | ET3 |
| MH0054 | ET1 |
| MH0055 | ET1 |
| MH0056 | ET1 |
| MH0057 | ET1 |
| MH0058 | ET1 |
| MH0059 | ET1 |

| | |
|--------|-----|
| MH0060 | ET1 |
| MH0061 | ET1 |
| MH0062 | ET1 |
| MH0063 | ET3 |
| MH0064 | ET2 |
| MH0065 | ET3 |
| MH0066 | ET1 |
| MH0067 | ET1 |
| MH0068 | ET1 |
| MH0069 | ET1 |
| MH0070 | ET1 |
| MH0071 | ET1 |
| MH0072 | ET1 |
| MH0073 | ET1 |
| MH0074 | ET1 |
| MH0075 | ET1 |
| MH0076 | ET1 |
| MH0077 | ET3 |
| MH0078 | ET3 |
| MH0079 | ET1 |
| MH0080 | ET1 |
| MH0081 | ET1 |
| MH0082 | ET1 |
| MH0083 | ET1 |
| MH0084 | ET2 |
| MH0085 | ET2 |
| MH0086 | ET1 |

Supplementary Table 11. Genera overrepresented in enterotypes and the P-values for their enrichment.

| Enterotype | Genus | P-value |
|------------|-------------------------|-----------|
| 1 | <i>Acidaminococcus</i> | 0 |
| | <i>Bacteroides</i> | 0 |
| | <i>Roseburia</i> | 0 |
| | <i>Faecalibacterium</i> | 8.88E-225 |
| | <i>Anaerostipes</i> | 2.95E-201 |
| | <i>Parabacteroides</i> | 2.70E-150 |
| | Clostridiales | 1.27E-18 |
| 2 | <i>Prevotella</i> | 0 |
| | <i>Streptococcus</i> | 2.18E-225 |
| | <i>Enterococcus</i> | 2.78E-64 |
| | <i>Desulfovibrio</i> | 1.26E-12 |
| | Lachnospiraceae | 0.0099863 |
| 3 | <i>Akkermansia</i> | 0 |
| | <i>Alistipes</i> | 0 |
| | <i>Klebsiella</i> | 0 |
| | <i>Ruminococcus</i> | 0 |

| | |
|-----------------------------|-------------|
| <i>Escherichia/Shigella</i> | 3.63E-288 |
| <i>Dialister</i> | 4.82E-243 |
| <i>Mitsuokella</i> | 8.24E-178 |
| <i>Methanobrevibacter</i> | 3.20E-109 |
| <i>Eggerthella</i> | 2.80E-65 |
| Ruminococcaceae | 2.92E-63 |
| <i>Subdoligranulum</i> | 3.40E-48 |
| <i>Coprococcus</i> | 7.03E-31 |
| <i>Collinsella</i> | 5.66E-29 |
| <i>Blautia</i> | 1.92E-18 |
| <i>Eubacterium</i> | 5.01E-15 |
| <i>Dorea</i> | 0.000155341 |

Supplementary Table 12. KEGG orthologous groups (KO) overrepresented in enterotypes, and the P-values for their enrichment. Functions mentioned in the main text are emphasized in bold text. eggNOG orthologous groups (COG) corresponding to these KO's are also listed when applicable.

| Enterotype | KO | Description | P-value | COG |
|------------|---------------|---|-----------------|--------------------|
| 1 | K00936 | | 8.13E-53 | - |
| | K03088 | RNA polymerase sigma-70 factor, ECF subfamily | 1.48E-39 | COG1595 |
| | K12373 | beta-hexosaminidase [EC:3.2.1.52] | 1.18E-23 | COG3525 |
| | K03287 | outer membrane factor, OMF family | 3.34E-20 | COG1538 |
| | K03296 | hydrophobic/amphiphilic exporter-1 (mainly G- bacteria), HAE1 family | 3.25E-17 | COG0841 |
| | K03585 | membrane fusion protein | 3.68E-12 | COG0845 |
| | K03327 | multidrug resistance protein, MATE family | 1.89E-11 | COG0534 |
| | K00358 | | 1.18E-10 | COG2249 |
| | K10819 | histidine kinase | 1.26E-10 | - |
| | K03530 | DNA-binding protein HU-beta | 9.56E-10 | COG0776 |
| | K00850 | 6-phosphofructokinase [EC:2.7.1.11] | 4.02E-09 | COG0205 COG1105 |
| | K00676 | ribosomal-protein-alanine N-acetyltransferase [EC:2.3.1.128] | 2.52E-07 | - |
| | K03442 | small conductance mechanosensitive ion channel, MscS family | 6.31E-06 | COG0668 |
| | K00786 | | 1.33E-05 | - |
| | K00648 | 3-oxoacyl-[acyl-carrier-protein] synthase III [EC:2.3.1.180] | 6.39E-05 | COG0332 |
| | K00798 | cob(I)alamin adenosyltransferase [EC:2.5.1.17] | 0.0001 | COG2109 |
| | K11752 | diaminohydroxyphosphoribosylaminopyrimidine deaminase / 5-amino-6-(5-phosphoribosylamino)uracil reductase [EC:3.5.4.26 1.1.1.193] | 0.0002 | COG0117 COG1985 |
| | K00680 | | 0.0002 | - |
| | K00046 | gluconate 5-dehydrogenase [EC:1.1.1.69] | 0.0002 | COG1028 |
| | K12308 | beta-galactosidase [EC:3.2.1.23] | 0.0008 | COG1874 |
| | K00262 | glutamate dehydrogenase (NADP+) [EC:1.4.1.4] | 0.0014 | COG0334 |
| | K00638 | chloramphenicol O-acetyltransferase [EC:2.3.1.28] | 0.0021 | COG0110 |

| | | | | |
|---|---------------|--|---------------|--------------------|
| | K03455 | monovalent cation:H ⁺ antiporter-2, CPA2 family | 0.0021 | COG0475 COG1226 |
| | K11690 | C4-dicarboxylate transporter, DctM subunit | 0.0053 | COG1593 |
| | K00077 | 2-dehydropantoate 2-reductase [EC:1.1.1.169] | 0.0076 | COG1893 |
| | K00026 | malate dehydrogenase [EC:1.1.1.37] | 0.0128 | COG0039 |
| | K00599 | | 0.0153 | COG0500 |
| | K03100 | signal peptidase I [EC:3.4.21.89] | 0.0209 | COG0681 |
| | K03307 | solute:Na ⁺ symporter, SSS family | 0.0238 | COG0591 |
| | K04041 | fructose-1,6-bisphosphatase III [EC:3.1.3.11] | 0.029 | COG3855 |
| | K03498 | trk system potassium uptake protein TrkH | 0.0293 | COG0168 |
| | K00928 | aspartate kinase [EC:2.7.2.4] | 0.0362 | COG0527 |
| | K03150 | thiamine biosynthesis ThiH | 0.0388 | COG1060 |
| | K03733 | integrase/recombinase XerC | 0.0433 | COG0582 |
| 2 | K00971 | mannose-1-phosphate guanylyltransferase [EC:2.7.7.22] | 1.85E-14 | COG0662 COG0836 |
| | K03543 | multidrug resistance protein A | 1.95E-12 | COG1566 |
| | K03760 | putative membrane protein | 6.95E-12 | COG2194 |
| | K00973 | glucose-1-phosphate thymidyltransferase [EC:2.7.7.24] | 9.84E-11 | COG1209 |
| | K00640 | serine O-acetyltransferase [EC:2.3.1.30] | 9.04E-09 | COG1045 |
| | K03315 | Na ⁺ :H ⁺ antiporter, NhaC family | 1.02E-08 | COG1757 |
| | K03624 | transcription elongation factor GreA | 5.42E-08 | COG0782 |
| | K03086 | RNA polymerase primary sigma factor | 2.70E-07 | COG0568 |
| | K03453 | bile acid:Na ⁺ symporter, BASS family | 2.89E-07 | COG0385 |
| | K00788 | thiamine-phosphate pyrophosphorylase [EC:2.5.1.3] | 1.12E-06 | COG0352 |
| | K03630 | DNA repair protein RadC | 1.54E-06 | COG2003 |
| | K00847 | fructokinase [EC:2.7.1.4] | 3.16E-06 | COG0524 |
| | K00784 | ribonuclease Z [EC:3.1.26.11] | 5.28E-06 | COG1234 |
| | K03816 | xanthine phosphoribosyltransferase [EC:2.4.2.22] | 7.10E-06 | COG0503 |
| | K03564 | peroxiredoxin Q/BCP [EC:1.11.1.15] | 1.22E-05 | COG1225 |
| | K00874 | 2-dehydro-3-deoxygluconokinase [EC:2.7.1.45] | 2.43E-05 | COG0524 |
| | K03320 | ammonium transporter, Amt family | 2.45E-05 | COG0004 |
| | K03427 | type I restriction enzyme M protein [EC:2.1.1.72] | 9.95E-05 | COG0286 |
| | K03593 | ATP-binding protein involved in chromosome partitioning | 0.0004 | COG0489 |
| | K00266 | glutamate synthase (NADPH/NADH) small chain [EC:1.4.1.13 1.4.1.14] | 0.0005 | COG0493 |
| | K11068 | hemolysin III | 0.0005 | COG1272 |
| | K00941 | phosphomethylpyrimidine kinase [EC:2.7.4.7] | 0.0007 | COG0351 |
| | K03308 | neurotransmitter:Na ⁺ symporter, NSS family | 0.0007 | COG0733 |
| | K00877 | hydroxymethylpyrimidine kinase [EC:2.7.1.49] | 0.0008 | - |
| | K03587 | cell division protein FtsI (penicillin binding protein 3) [EC:2.4.1.129] | 0.0009 | COG0768 |
| | K00924 | | 0.0015 | COG0515 COG1493 |
| | K00903 | protein-tyrosine kinase [EC:2.7.10.-] | 0.0019 | - |
| | K00611 | ornithine carbamoyltransferase [EC:2.1.3.3] | 0.004 | COG0078 |
| | K00865 | glycerate kinase [EC:2.7.1.31] | 0.004 | COG1929 |

| | | | | |
|---|--------|---|----------|--|
| | K00031 | isocitrate dehydrogenase [EC:1.1.1.42] | 0.0061 | COG0538 |
| | K00012 | UDPglucose 6-dehydrogenase [EC:1.1.1.22] | 0.0064 | COG1004 |
| | K00811 | aspartate aminotransferase [EC:2.6.1.1] | 0.007 | - |
| | K00785 | beta-galactosamide-alpha-2,3-sialyltransferase [EC:2.4.99.-] | 0.0083 | COG0682 |
| | K03321 | sulfate permease, SulP family | 0.0117 | COG0659 |
| | K00842 | aminotransferase [EC:2.6.1.-] | 0.0128 | COG1168 |
| | K03546 | exonuclease SbcC | 0.0137 | COG0419 |
| | K00278 | L-aspartate oxidase [EC:1.4.3.16] | 0.0138 | COG0029 |
| | K03406 | methyl-accepting chemotaxis protein | 0.0267 | COG0840 |
| | K03699 | putative hemolysin | 0.0293 | COG1253 |
| | K03856 | 3-deoxy-7-phosphoheptulonate synthase [EC:2.5.1.54] | 0.034 | COG2876 |
| | K03147 | thiamine biosynthesis protein ThiC | 0.0364 | COG0422 |
| | K03531 | cell division protein FtsZ | 0.0398 | COG0206 |
| | K00773 | queuine tRNA-ribosyltransferase [EC:2.4.2.29] | 0.0431 | COG0343 |
| | K03310 | alanine or glycine:cation symporter, AGCS family | 0.0466 | COG1115 |
| | K03521 | electron transfer flavoprotein beta subunit | 0.0469 | COG2086 |
| | K00970 | poly(A) polymerase [EC:2.7.7.19] | 0.0484 | COG0617 |
| 3 | K03205 | type IV secretion system protein VirD4 | 7.98E-32 | COG3505 |
| | K03091 | RNA polymerase sporulation-specific sigma factor | 1.69E-15 | - |
| | K03497 | chromosome partitioning protein, ParB family | 5.15E-15 | COG1475 |
| | K03588 | cell division protein FtsW | 5.36E-15 | COG0772 |
| | K03621 | fatty acid/phospholipid synthesis protein | 5.66E-14 | COG0416 |
| | K00943 | dTMP kinase [EC:2.7.4.9] | 1.38E-11 | COG0125 |
| | K00852 | ribokinase [EC:2.7.1.15] | 7.24E-09 | COG0524 |
| | K00336 | NADH dehydrogenase I subunit G [EC:1.6.5.3] | 7.47E-09 | COG1034 |
| | K03500 | ribosomal RNA small subunit methyltransferase B [EC:2.1.1.-] | 1.85E-08 | COG0144 |
| | K00642 | glutamate N-acetyltransferase [EC:2.3.1.35] | 2.50E-08 | COG1364 |
| | K00939 | adenylate kinase [EC:2.7.4.3] | 8.27E-08 | COG0563 |
| | K00001 | alcohol dehydrogenase [EC:1.1.1.1] | 9.13E-08 | COG1012 COG1062 COG1064 COG1454 |
| | K03218 | RNA methyltransferase, TrmH family [EC:2.1.1.-] | 2.36E-07 | COG0566 |
| | K00975 | glucose-1-phosphate adenylyltransferase [EC:2.7.7.27] | 3.29E-07 | COG0448 |
| | K03110 | signal recognition particle receptor | 3.29E-07 | COG0552 |
| | K03709 | DtxR family transcriptional regulator, Mn-dependent transcriptional regulator | 6.56E-07 | COG1321 |
| | K03581 | exodeoxyribonuclease V alpha subunit [EC:3.1.11.5] | 6.69E-07 | COG0507 |
| | K00821 | acetylornithine/N-succinyldiaminopimelate aminotransferase [EC:2.6.1.11 2.6.1.17] | 7.59E-07 | COG4992 |
| | K00087 | xanthine dehydrogenase [EC:1.17.1.4] | 9.05E-07 | COG1319 COG1529 COG1975 COG2080 |

| | | | |
|--------|---|----------|--------------------|
| K00615 | transketolase [EC:2.2.1.1] | 9.23E-07 | COG0021 |
| K03763 | DNA polymerase III subunit alpha, Gram-positive type [EC:2.7.7.7] | 1.84E-06 | COG2176 |
| K03664 | SsrA-binding protein | 2.21E-06 | COG0691 |
| K00016 | L-lactate dehydrogenase [EC:1.1.1.27] | 2.77E-06 | COG0039 |
| K03529 | chromosome segregation protein | 2.78E-06 | COG1196 |
| K03169 | DNA topoisomerase III [EC:5.99.1.2] | 7.33E-06 | COG0550 |
| K00528 | ferredoxin--NADP+ reductase [EC:1.18.1.2] | 1.77E-05 | COG1018 |
| K00560 | thymidylate synthase [EC:2.1.1.45] | 2.61E-05 | COG0207 |
| K00604 | methionyl-tRNA formyltransferase [EC:2.1.2.9] | 4.10E-05 | COG0223 |
| K00540 | | 5.07E-05 | - |
| K03076 | preprotein translocase subunit SecY | 5.60E-05 | COG0201 |
| K03613 | electron transport complex protein RnfE | 8.06E-05 | COG1347 |
| K03431 | phosphoglucosamine mutase [EC:5.4.2.10] | 8.14E-05 | COG1109 |
| K03596 | GTP-binding protein LepA | 9.10E-05 | COG0481 |
| K03466 | DNA segregation ATPase FtsK/SpoIIIE, S-DNA-T family | 0.0001 | COG1674 |
| K03572 | DNA mismatch repair protein MutL | 0.0002 | COG0323 |
| K03701 | excinuclease ABC subunit A | 0.0002 | COG0178 |
| K03111 | single-strand DNA-binding protein | 0.0002 | COG0629 |
| K11069 | spermidine/putrescine transport system substrate-binding protein | 0.0002 | COG0687 |
| K03524 | BirA family transcriptional regulator, biotin operon repressor / biotin-[acetyl-CoA-carboxylase] ligase [EC:6.3.4.15] | 0.0002 | COG1654 COG0340 |
| K00058 | D-3-phosphoglycerate dehydrogenase [EC:1.1.1.95] | 0.0003 | COG0111 |
| K03525 | type III pantothenate kinase [EC:2.7.1.33] | 0.0003 | COG1521 |
| K00010 | myo-inositol 2-dehydrogenase [EC:1.1.1.18] | 0.0003 | COG0673 |
| K03789 | ribosomal-protein-alanine N-acetyltransferase [EC:2.3.1.128] | 0.0004 | COG0456 |
| K12063 | conjugal transfer ATP-binding protein TraC | 0.0006 | - |
| K03043 | DNA-directed RNA polymerase subunit beta [EC:2.7.7.6] | 0.0007 | COG0085 |
| K03617 | electron transport complex protein RnfA | 0.0008 | COG2209 |
| K00763 | nicotinate phosphoribosyltransferase [EC:2.4.2.11] | 0.0008 | COG1488 |
| K03650 | tRNA modification GTPase | 0.0009 | COG0486 |
| K11189 | phosphocarrier protein | 0.0011 | COG1925 |
| K11175 | phosphoribosylglycinamide formyltransferase 1 [EC:2.1.2.2] | 0.0012 | COG0299 |
| K03737 | putative pyruvate-flavodoxin oxidoreductase [EC:1.2.7.-] | 0.0015 | COG0674 COG1013 |
| K00790 | UDP-N-acetylglucosamine 1-carboxyvinyltransferase [EC:2.5.1.7] | 0.0029 | COG0766 |
| K00248 | butyryl-CoA dehydrogenase [EC:1.3.99.2] | 0.0034 | COG1960 |
| K03625 | N utilization substance protein B | 0.0036 | COG0781 |
| K03629 | DNA replication and repair protein RecF | 0.0037 | COG1195 |
| K11753 | riboflavin kinase / FMN adenylyltransferase [EC:2.7.1.26 2.7.7.2] | 0.004 | COG0196 |
| K03502 | DNA polymerase V | 0.0041 | COG0389 |

| | | | |
|--------|--|--------|-------------------------------|
| K03798 | cell division protease FtsH [EC:3.4.24.-] | 0.0042 | COG0465 |
| K00800 | 3-phosphoshikimate 1-carboxyvinyltransferase [EC:2.5.1.19] | 0.0043 | COG0128 |
| K00783 | hypothetical protein | 0.0051 | COG1576 |
| K00600 | glycine hydroxymethyltransferase [EC:2.1.2.1] | 0.0056 | COG0112 |
| K00949 | thiamine pyrophosphokinase [EC:2.7.6.2] | 0.0076 | COG1564 |
| K04066 | primosomal protein N (replication factor Y) (superfamily II helicase) [EC:3.6.1.-] | 0.0095 | COG1198 |
| K00864 | glycerol kinase [EC:2.7.1.30] | 0.0096 | COG0554 |
| K00059 | 3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100] | 0.0116 | COG1028 |
| K00384 | thioredoxin reductase (NADPH) [EC:1.8.1.9] | 0.0118 | COG0492 |
| K03686 | molecular chaperone DnaJ | 0.0122 | COG0484 |
| K03892 | ArsR family transcriptional regulator | 0.0141 | COG0640 |
| K11071 | spermidine/putrescine transport system permease protein | 0.016 | COG1176 |
| K00554 | tRNA (guanine-N1-)-methyltransferase [EC:2.1.1.31] | 0.0202 | COG0336 |
| K03177 | tRNA pseudouridine synthase B [EC:5.4.99.12] | 0.0226 | COG0130 |
| K03702 | excinuclease ABC subunit B | 0.0228 | COG0556 |
| K00857 | thymidine kinase [EC:2.7.1.21] | 0.0233 | COG1435 |
| K00789 | S-adenosylmethionine synthetase [EC:2.5.1.6] | 0.0261 | COG0192 COG1812 |
| K00831 | phosphoserine aminotransferase [EC:2.6.1.52] | 0.0266 | COG1932 |
| K03977 | GTP-binding protein | 0.0271 | COG1160 |
| K03168 | DNA topoisomerase I [EC:5.99.1.2] | 0.0307 | COG0550 COG0551 COG1754 |
| K03423 | | 0.0361 | - |
| K00942 | guanylate kinase [EC:2.7.4.8] | 0.0394 | COG0194 |
| K03614 | electron transport complex protein RnfD | 0.0406 | COG1805 |
| K03631 | DNA repair protein RecN (Recombination protein N) | 0.0412 | COG0497 |
| K03655 | ATP-dependent DNA helicase RecG [EC:3.6.1.-] | 0.0416 | COG1200 |

Supplementary Table 13. KEGG modules overrepresented in enterotypes and the P-values for their enrichment. Functions mentioned in the main text are emphasized in bold text.

| Enterotype | Module | Description | P-value |
|------------|---------------|---|-----------------|
| 1 | M00155 | Keratan sulfate degradation | 4.48E-24 |
| | M00006 | Pentose phosphate pathway, oxidative phase | 2.08E-15 |
| | M00306 | Sulfur reduction, sulfate => H ₂ S | 8.61E-12 |
| | M00248 | Biotin biosynthesis, pimeloyl-CoA => biotin | 2.85E-10 |
| | M00388 | Uncharacterized ABC transport system | 7.57E-10 |
| | M00159 | Fatty acid biosynthesis, initiation | 1.58E-08 |
| | M00102 | UDP-Xylose biosynthesis, UDP-Glc => UDP-Xyl => Xyl | 1.24E-06 |
| | M00008 | Entner-Doudoroff pathway | 1.87E-06 |

| | | | |
|---|---------------|---|-----------------|
| | M00083 | beta-Alanine biosynthesis, L-aspartate => beta-alanine | 2.15E-06 |
| | M00001 | Glycolysis, fructose-6P => pyruvate | 2.53E-06 |
| | M00154 | Heparan sulfate degradation | 6.26E-06 |
| | M00017 | Glutamate biosynthesis, oxoglutarate => glutamate (glutamate dehydrogenase) | 3.90E-05 |
| | M00160 | Fatty acid biosynthesis, elongation | 0.0001 |
| | M00152 | Dermatan sulfate degradation | 0.0003 |
| | M00046 | Asparagine degradation, asparagine => aspartate +NH ₃ | 0.0009 |
| | M00244 | Pantothenate biosynthesis, valine => pantothenate | 0.0016 |
| | M00250 | Riboflavin biosynthesis, GTP => riboflavin/FMN/FAD | 0.0016 |
| | M00037 | Histidine biosynthesis, PRPP => histidine | 0.0021 |
| | M00003 | Gluconeogenesis, oxaloacetate => fructose-6P | 0.0025 |
| | M00608 | PTS system, beta-glucosides-specific II component | 0.0083 |
| | M00104 | CMP-N-Acetylneuraminic acid biosynthesis (mammals), ManNAc => Neu5Ac-9P => CMP-Neu5Ac | 0.0098 |
| | M00255 | Ascorbate biosynthesis, animals | 0.0148 |
| | M00153 | Chondroitin sulfate degradation | 0.0187 |
| | M00028 | Leucine biosynthesis, pyruvate => leucine | 0.037 |
| 2 | M00105 | dTDP-Glucose, dTDP-galactose and dTDP-rhamnose biosynthesis | 1.11E-29 |
| | M00099 | GDP-Mannose biosynthesis, fructose-6P => GDP-Man | 2.32E-18 |
| | M00252 | Thiamine biosynthesis, AIR => thiamine-P/thiamine-2P | 5.37E-18 |
| | M00239 | Ascorbate biosynthesis, plants | 1.95E-16 |
| | M00034 | Tryptophan biosynthesis, chorismate => tryptophan | 2.59E-16 |
| | M00273 | Complex I (NADH dehydrogenase), NADH dehydrogenase I | 2.11E-13 |
| | M00156 | Lipopolysaccharide biosynthesis, inner core => outer core => O-antigen | 6.10E-11 |
| | M00293 | ATP synthase | 3.92E-10 |
| | M00323 | Putative spermidine/putrescine transport system | 9.07E-10 |
| | M00117 | Uronic acid metabolism | 1.05E-08 |
| | M00032 | Cysteine biosynthesis, serine => cysteine | 1.67E-06 |
| | M00278 | Complex II (succinate dehydrogenase / fumarate reductase), succinate dehydrogenase | 4.68E-05 |
| | M00042 | Urea cycle | 9.52E-05 |
| | M00062 | Cysteine metabolism, cysteine => 3-sulfino-L-alanine => pyruvate | 0.0005 |
| | M00299 | C4-dicarboxylic acid cycle, phosphoenolpyruvate carboxykinase type | 0.0007 |
| | M00192 | C5 isoprenoid biosynthesis, non-mevalonate pathway | 0.0007 |
| | M00063 | Cysteine metabolism, cysteine => 3-mercaptopyruvate => pyruvate | 0.0011 |
| | M00319 | Molybdate transport system | 0.0025 |
| | M00240 | NAD biosynthesis, aspartate => NAD | 0.0047 |
| | M00010 | Citrate cycle, first carbon oxidation | 0.0047 |
| | M00302 | Reductive carboxylate cycle | 0.0053 |
| | M00660 | RuvABC complex | 0.0085 |
| | M00018 | Glutamine biosynthesis, glutamate => glutamine | 0.015 |
| | M00298 | CAM, light | 0.017 |
| | M00022 | Asparagine biosynthesis, aspartate => asparagine | 0.0206 |
| | M00649 | uvrBC complex | 0.0276 |
| | M00029 | Isoleucine biosynthesis, pyruvate => isoleucine | 0.0283 |
| | M00300 | C4-dicarboxylic acid cycle, NAD ⁺ -malic enzyme type | 0.0316 |

| | | | |
|---|---------------|---|-----------------|
| | M00090 | Inosine monophosphate biosynthesis, PRPP + glutamine => IMP | 0.0394 |
| | M00269 | Tetrahydrofolate biosynthesis | 0.0471 |
| | M00203 | Glyceroglycolipid biosynthesis | 0.0473 |
| 3 | M00369 | Peptides/nickel transport system | 7.70E-111 |
| | M00351 | Simple sugar transport system | 1.29E-62 |
| | M00375 | Cobalt transport system | 8.47E-57 |
| | M00376 | Nickel transport system | 3.49E-56 |
| | M00367 | Branched-chain amino acid transport system | 1.99E-50 |
| | M00337 | Multiple sugar transport system | 6.71E-47 |
| | M00288 | V-type ATPase (Prokaryotes) | 1.82E-28 |
| | M00342 | Ribose transport system | 6.06E-20 |
| | M00318 | Sulfonate/nitrate/taurine transport system | 2.20E-17 |
| | M00353 | Phosphonate transport system | 1.15E-13 |
| | M00320 | Iron(III) transport system | 1.49E-12 |
| | M00309 | Ribosome, archaea | 7.94E-12 |
| | M00303 | Methane oxidation, methylotroph, methane => CO2 | 1.01E-09 |
| | M00308 | Ribosome, bacteria | 3.51E-08 |
| | M00368 | D-Methionine transport system | 3.50E-07 |
| | M00246 | Heme biosynthesis, glutamate => protoheme/siroheme | 2.11E-06 |
| | M00064 | Cysteine metabolism, cysteine => pyruvate | 1.29E-05 |
| | M00076 | Dopamine / noradrenaline / adrenaline metabolism | 1.35E-05 |
| | M00377 | Putative ABC transport system | 4.69E-05 |
| | M00007 | Pentose phosphate pathway, non-oxidative phase | 5.65E-05 |
| | M00344 | Methyl-galactoside transport system | 0.0005 |
| | M00324 | Maltose/maltodextrin transport system | 0.0013 |
| | M00370 | Iron complex transport system | 0.0015 |
| | M00352 | Phosphate transport system | 0.0017 |
| | M00384 | ABC-2 type transport system | 0.0017 |
| | M00262 | Putrescine metabolism, N-acetylation, putrescine => 4-aminobutanoate | 0.0019 |
| | M00313 | RNA polymerase, bacteria | 0.0024 |
| | M00005 | Pentose phosphate pathway and PRPP biosynthesis | 0.0027 |
| | M00023 | Lysine biosynthesis, aspartate => lysine | 0.0056 |
| | M00658 | RecBCD complex | 0.0065 |
| | M00033 | Chorismate biosynthesis, phosphoenolpyruvate + erythrose-4P => chorismate | 0.0079 |
| | M00098 | UDP-N-Acetylglucosamine biosynthesis, fructose-6P => UDP-GlcNAc | 0.0093 |
| | M00648 | uvrA2B2 complex | 0.0121 |
| | M00095 | Pyrimidine deoxyribonucleotide biosynthesis, CDP/CTP => dCDP/dCTP,dTDP/dTTP | 0.0151 |
| | M00610 | PTS system, fructose-specific II component | 0.0161 |
| | M00366 | Polar amino acid transport system | 0.0227 |
| | M00279 | Complex II (succinate dehydrogenase / fumarate reductase), fumarate reductase | 0.0289 |
| | M00030 | Serine biosynthesis, glycerate-3P => serine | 0.036 |

Supplementary Table 14. Functions varying more between than within nationalities.

| OG | description |
|---------|--|
| COG0085 | DNA-directed RNA polymerase, beta subunit/140 kD subunit |
| COG0086 | DNA-directed RNA polymerase, beta subunit/160 kD subunit |
| COG0210 | Superfamily I DNA and RNA helicases |
| COG0249 | Mismatch repair ATPase |
| COG0587 | DNA polymerase III, alpha subunit |
| COG1538 | Outer membrane protein |
| COG1595 | DNA-directed RNA polymerase specialized sigma subunit, sigma24 homolog |
| COG2801 | Transposase and inactivated derivatives |
| COG3291 | FOG: PKD repeat |
| COG4646 | DNA methylase |

Supplementary Table 15. Orthologous groups overrepresented in correlation with host properties and the P-values for their enrichment.

| Feature | Value | OG | Description | P-value |
|-------------|---------|--|--|-------------|
| Gender | female | COG0673 | Predicted dehydrogenases and related proteins | 0.019916675 |
| Nationality | Danish | NOG12793 | Calcium ion binding protein | 0.021913859 |
| | | COG0826 | Collagenase and related proteases | 0.000842037 |
| | | COG0249 | Mismatch repair ATPase (MutS family) | 0.00012716 |
| | | COG0564 | Pseudouridylate synthases, 23S RNA-specific | 0.001900145 |
| | | COG2801 | Transposase and inactivated derivatives | 9.93E-55 |
| | Spanish | COG0110 | Acetyltransferase (isoleucine patch superfamily) | 1.60E-06 |
| | | COG1506 | Dipeptidyl aminopeptidases/acylaminoacyl-peptidases | 3.17E-35 |
| | | COG0463 | Glycosyltransferases involved in cell wall biogenesis | 0.027341556 |
| | | COG0697 | Permeases of the drug/metabolite transporter (DMT) superfamily | 0.040150934 |
| | | COG4974 | Site-specific recombinase XerD | 2.85E-06 |
| | | COG0514 | Superfamily II DNA helicase | 6.18E-05 |
| | | COG0526 | Thiol-disulfide isomerase and thioredoxins | 5.34E-13 |
| | French | COG0058 | Glucan phosphorylase | 0.017064211 |
| | Italian | COG0436 | Aspartate/tyrosine/aromatic aminotransferase | 0.012372841 |
| | | COG0477 | Permeases of the major facilitator superfamily | 6.54E-65 |
| | | COG1070 | Sugar (pentulose and hexulose) kinases | 1.81E-05 |
| COG1940 | | Transcriptional regulator/sugar kinase | 1.49E-06 | |
| Japanese | COG1032 | Fe-S oxidoreductase | 0.009058224 | |

| | | | |
|--|----------|--|-------------|
| | COG1253 | Hemolysins and related proteins containing CBS domains | 0.006172084 |
| | COG2244 | Membrane protein involved in the export of O-antigen and teichoic acid | 0.023130117 |
| | COG2211 | Na ⁺ /melibiose symporter and related transporters | 7.80E-16 |
| | COG1475 | Predicted transcriptional regulators | 0.01356461 |
| | NOG75023 | Regulator protein | 1.86E-05 |
| | COG0500 | SAM-dependent methyltransferases | 2.75E-07 |
| | COG0480 | Translation elongation factors (GTPases) | 3.81E-07 |

Supplementary Table 16. Functional modules overrepresented in correlation with host properties and the P-values for their enrichment. Functions mentioned in the main text are emphasized in bold text.

| Feature | Value | Module | Description | P-value | |
|-------------|--------|--|--|--|-------------|
| Gender | male | M00021 | Aspartate biosynthesis, oxaloacetate => aspartate | 0.041131871 | |
| | | M00098 | UDP-N-Acetylglucosamine biosynthesis, fructose-6P => UDP-GlcNAc | 0.030649713 | |
| | female | M00279 | Complex II (succinate dehydrogenase / fumarate reductase), fumarate reductase | 0.04717716 | |
| | | M00090 | Inosine monophosphate biosynthesis, PRPP + glutamine => IMP | 0.000770102 | |
| | | M00097 | UDP-glucose and UDP-galactose biosynthesis | 0.011526468 | |
| | | M00153 | Chondroitin sulfate degradation | 0.035471815 | |
| Nationality | Danish | M00104 | CMP-N-Acetylneuraminat biosynthesis (mammals), ManNAc => Neu5Ac-9P => CMP-Neu5Ac | 0.015745112 | |
| | | M00203 | Glyceroglycolipid biosynthesis | 0.024271472 | |
| | | M00092 | Guanine nucleotide biosynthesis, IMP => GDP/dGDP,GTP/dGTP | 0.005817982 | |
| | | M00154 | Heparan sulfate degradation | 0.018485607 | |
| | | M00658 | RecBCD complex | 0.000298037 | |
| | | M00659 | RecFOR complex | 0.017029119 | |
| | | M00318 | Sulfonate/nitrate/taurine transport system | 6.82E-06 | |
| | | M00101 | UDP-N-Acetylgalactosamine and UDP-N-acetylmannosamine biosynthesis | 0.032811714 | |
| | | M00649 | uvrBC complex | 0.002371057 | |
| | | M00372 | Zinc transport system | 0.008694071 | |
| | | Spanish | M00255 | Ascorbate biosynthesis, animals | 0.010648826 |
| | | | M00046 | Asparagine degradation, asparagine => aspartate +NH ₃ | 0.0071149 |
| | | | M00083 | beta-Alanine biosynthesis, L-aspartate => beta-alanine | 0.001659371 |
| | M00248 | | Biotin biosynthesis, pimeloyl-CoA => biotin | 5.38E-05 | |
| | M00300 | | C4-dicarboxylic acid cycle, NAD ⁺ -malic enzyme type | 0.007059585 | |
| | M00297 | | CAM, dark | 0.035900728 | |
| | M00010 | | Citrate cycle, first carbon oxidation | 0.032229415 | |
| | M00119 | CMP-Kdo biosynthesis | 1.95E-06 | | |
| | M00278 | Complex II (succinate dehydrogenase / fumarate reductase), succinate dehydrogenase | 1.24E-05 | | |
| | M00273 | Complex I (NADH dehydrogenase), NADH | 3.43E-16 | | |

| | | | | |
|-----------------|----------|--|--|-------------|
| | | dehydrogenase I | | |
| | M00032 | Cysteine biosynthesis, serine => cysteine | 7.45E-05 | |
| | M00282 | Cytochrome bd complex | 0.001223301 | |
| | M00160 | Fatty acid biosynthesis, elongation | 0.00308024 | |
| | M00003 | Gluconeogenesis, oxaloacetate => fructose-6P | 0.03893556 | |
| | M00016 | Glucuronate pathway (uronate pathway) | 0.000465544 | |
| | M00018 | Glutamine biosynthesis, glutamate => glutamine | 0.009205549 | |
| | M00001 | Glycolysis, fructose-6P => pyruvate | 0.04671136 | |
| | M00012 | Glyoxylate cycle | 0.001673512 | |
| | M00156 | Lipopolysaccharide biosynthesis, inner core => outer core => O-antigen | 2.00E-06 | |
| | M00241 | Menaquinone biosynthesis, chorismate => menaquinone | 5.87E-14 | |
| | M00244 | Pantothenate biosynthesis, valine => pantothenate | 0.021374437 | |
| | M00253 | Phylloquinone biosynthesis, chorismate => phylloquinone | 1.65E-15 | |
| | M00340 | Putative ABC transport system | 5.39E-07 | |
| | M00249 | Pyridoxal biosynthesis, erythrose-4P => pyridoxal-5P | 0.000698043 | |
| | M00302 | Reductive carboxylate cycle | 1.00E-05 | |
| | M00034 | Tryptophan biosynthesis, chorismate => tryptophan | 2.44E-12 | |
| | M00036 | Tyrosine biosynthesis, chorismate => tyrosine | 3.03E-08 | |
| | French | M00041 | Ornithine biosynthesis, glutamate => ornithine | 0.005353483 |
| | | M00004 | Pentose phosphate pathway | 0.002086546 |
| | | M00005 | Pentose phosphate pathway and PRPP biosynthesis | 0.001188288 |
| | | M00352 | Phosphate transport system | 0.010477137 |
| | Italian | M00246 | Heme biosynthesis, glutamate => protoheme/siroheme | 0.002010425 |
| | | M00370 | Iron complex transport system | 1.65E-07 |
| | | M00029 | Isoleucine biosynthesis, pyruvate => isoleucine | 0.01158102 |
| | | M00050 | Lysine degradation, lysine => saccharopine => acetoacetyl-CoA | 0.038354268 |
| | | M00303 | Methane oxidation, methylotroph, methane => CO ₂ | 4.81E-09 |
| | | M00027 | Valine biosynthesis, pyruvate => valine | 0.018256687 |
| | Japanese | M00386 | Cell division transport system | 0.004531346 |
| | | M00247 | Cobalamin biosynthesis, cobinamide => cobalamin | 1.19E-05 |
| | | M00159 | Fatty acid biosynthesis, initiation | 0.002635444 |
| | | M00023 | Lysine biosynthesis, aspartate => lysine | 0.000387916 |
| | | M00024 | Methionine biosynthesis, aspartate => homoserine => methionine | 0.008760403 |
| | | M00608 | PTS system, beta-glucosides-specific II component | 7.16E-05 |
| | | M00030 | Serine biosynthesis, glycerate-3P => serine | 0.009094861 |
| Clinical status | Obese | M00211 | Diphosphatidylglycerol biosynthesis, CDP-glycerol => cardiolipin | 0.003066449 |
| | | M00118 | Pentose interconversion, arabinose/ribulose/xylulose/xylose | 0.013040452 |

Supplementary Table 17. Orthologous groups significantly correlating with age, with the R^2 and P-values for each correlation. Functions mentioned in the main text are emphasized in bold text.

| OG | Description | R^2 | P-value |
|----------------|---|--------------|-------------|
| COG0058 | Glucan phosphorylase | 0.495914579 | 0.047797398 |
| COG0085 | DNA-directed RNA polymerase, beta subunit/140 kD subunit | 0.571169314 | 0.024918414 |
| COG0086 | DNA-directed RNA polymerase, beta subunit/160 kD subunit | 0.507616388 | 0.049800725 |
| COG0187 | Type IIA topoisomerase (DNA gyrase/topo II, topoisomerase IV B) | 0.539090218 | 0.036803096 |
| COG0205 | 6-phosphofructokinase | -0.526783144 | 0.036699398 |
| COG0366 | Glycosidases | 0.506002345 | 0.046395294 |
| COG0438 | Glycosyltransferase | -0.537072665 | 0.033077914 |
| COG0493 | NADPH-dependent glutamate synthase beta chain and related | 0.554221867 | 0.030648837 |
| COG0653 | Preprotein translocase subunit SecA ATPase, RNA helicase | 0.572581601 | 0.032021108 |
| COG0739 | Membrane proteins related to metalloendopeptidases | -0.498133888 | 0.049792501 |
| COG1595 | DNA-directed RNA polymerase specialized sigma subunit | -0.59376591 | 0.027056041 |

Supplementary Table 18. Orthologous groups significantly correlating with age when combined into a linear model.

| Orthologous group | Description |
|-------------------|--|
| COG0085 | DNA-directed RNA polymerase, beta subunit/140 kD subunit |
| COG0086 | DNA-directed RNA polymerase, beta subunit/160 kD subunit |
| COG0438 | Glycosyltransferase |
| COG0739 | Membrane proteins related to metalloendopeptidases |

Supplementary Table 19. Functional modules significantly correlating with the body mass index of individuals.

| Module | Description | R^2 | P-value |
|--------|--------------------------|----------|----------|
| M00051 | Ectoine biosynthesis | -0.8 | 0.050253 |
| M00286 | F-type ATPase (Bacteria) | 0.779412 | 0.036732 |
| M00293 | ATP synthase | 0.779412 | 0.036732 |

Supplementary Table 20. Amount of sequence generated and the number of genera (at 85% sequence identity) and eggNOG orthologous groups (OGs) retrieved from two Danish samples sequenced using Sanger and 454 Titanium technologies.

| sample | DA-AD-1 | | DA-AD-3 | |
|--------------------|---------|----------|---------|----------|
| | Sanger | Titanium | Sanger | Titanium |
| amount of sequence | 157Mb | 295Mb | 155Mb | 195Mb |
| genera | 56 | 121 | 58 | 117 |
| eggNOG OGs | 5863 | 5675 | 6015 | 5935 |

Supplementary Table 21. Potential human DNA fragments and fragments with best hits to eukaryotes. Values represent percentage of the metagenome sequence fragments from each sample. For the three published studies, these numbers do not include human/eukaryotic sequences removed by the authors of the respective publications before making the data publicly available. For the prophage fraction, we present an upper bound by counting all fragments with a BLASTN hit (>60 bits) to a viral and phage genome database, and a lower bound by considering only the fragments whose viral hit is significantly better than their best hit to a microbial sequence in the reference genome set of 1152 microbial genomes.

| | Sample ID | % human | % other eukaryotes | % prophage | |
|----------------------------------|----------------|--------------|--------------------|-------------|-------------|
| | | | | lowerbound | upperbound |
| European samples from this study | DA-AD-1 | 0.113 | 0.51 | 1.41 | 4.64 |
| | DA-AD-2 | 1.354 | 1.22 | 6.87 | 14.66 |
| | DA-AD-3 | 0.003 | 0.58 | 0.68 | 4.36 |
| | DA-AD-4 | 0.401 | 1.01 | 3.7 | 8.81 |
| | ES-AD-1 | 0.037 | 0.29 | 0.51 | 3.21 |
| | ES-AD-2 | 0.166 | 0.39 | 0.85 | 4.15 |
| | ES-AD-3 | 0.024 | 0.43 | 0.56 | 3.78 |
| | ES-AD-4 | 0.831 | 0.5 | 0.8 | 4.26 |
| | FR-AD-1 | 0.002 | 0.46 | 0.92 | 5.23 |
| | FR-AD-2 | 0.068 | 0.99 | 2.83 | 6.74 |
| | FR-AD-3 | 0.005 | 0.39 | 1.12 | 4.43 |
| | FR-AD-4 | 0.024 | 0.59 | 1.71 | 5.96 |
| | FR-AD-5 | 0.015 | 0.56 | 3.11 | 8.71 |
| | FR-AD-6 | 0.008 | 0.43 | 1.33 | 5.05 |
| | FR-AD-7 | 0.007 | 0.51 | 1.52 | 5.51 |
| | FR-AD-8 | 0.011 | 0.49 | 1.39 | 5.2 |
| | IT-AD-1 | 0.007 | 0.32 | 0.61 | 5.05 |
| | IT-AD-2 | 0 | 0.35 | 0.68 | 6.4 |
| | IT-AD-3 | 0.015 | 0.42 | 1.52 | 5.98 |
| | IT-AD-4 | 0 | 0.28 | 0.71 | 4.22 |
| IT-AD-5 | 0.002 | 0.64 | 1.22 | 5.8 | |
| IT-AD-6 | 0.011 | 0.41 | 0.83 | 5.12 | |
| | Average | 0.141 | 0.535 | 1.56 | 5.79 |
| Published samples | AM-AD-1 | 0.026 | 0.51 | 3.21 | 21.73 |
| | AM-AD-2 | 0.013 | 0.8 | 2.35 | 16.43 |
| | AM-F10-T1 | 0.004 | 0.76 | 0.25 | 1.36 |
| | AM-F10-T2 | 0.001 | 0.28 | 0.09 | 1.28 |
| | JP-AD-1 | 0.004 | 0.36 | 0.84 | 5.49 |
| | JP-AD-2 | 0.014 | 0.54 | 1.06 | 6.9 |
| | JP-AD-3 | 0.001 | 0.3 | 0.55 | 7.1 |
| | JP-AD-4 | 0.005 | 0.44 | 0.41 | 4.93 |
| | JP-AD-5 | 0.028 | 0.83 | 3.43 | 12.64 |
| | JP-AD-6 | 0.006 | 0.38 | 0.61 | 4.78 |
| | JP-AD-7 | 0.068 | 0.34 | 0.97 | 4.44 |
| | JP-AD-8 | 0.024 | 0.35 | 1.19 | 5.77 |
| | JP-AD-9 | 0.045 | 0.39 | 1.47 | 7.1 |
| | JP-IN-1 | 0.005 | 0.13 | 1.41 | 11.27 |
| | JP-IN-2 | 0 | 0.41 | 0.58 | 9.42 |
| | JP-IN-3 | 0 | 0.43 | 1.09 | 11.02 |

| | | | | | |
|--|------------------------|--------------|--------------|-------------|-------------|
| | JP-IN-4 | 0.096 | 0.49 | 0.89 | 9.22 |
| | Average | 0.02 | 0.455 | 1.2 | 8.3 |
| | Overall average | 0.093 | 0.5 | 1.42 | 6.88 |

Supplementary Table 22. Average fraction of DNA fragments from the metagenome samples with best hits to each eukaryotic kingdom represented in the STRING database.

| Kingdom | % fragments |
|---------------|-------------|
| Metazoa | 0.1982 |
| Fungi | 0.1091 |
| Amoebozoa | 0.0693 |
| Alveolata | 0.0635 |
| Viridiplantae | 0.0339 |
| Euglenozoa | 0.0245 |
| Fornicata | 0.0012 |

Supplementary Table 23. Abundance intervals of prominent genera in the enterotypes.

| Enterotype Genus | 1 | | 2 | | 3 | |
|-------------------------|----------|----------|----------|----------|----------|----------|
| | High | Low | High | Low | High | Low |
| <i>Akkermansia</i> | 0.000000 | 0.086886 | 0.000000 | 0.000211 | 0.0000 | 0.043993 |
| <i>Alistipes</i> | 0.002230 | 0.091233 | 0.001158 | 0.034362 | 0.000877 | 0.040998 |
| <i>Bacteroides</i> | 0.017973 | 0.172901 | 0.016110 | 0.178247 | 0.212884 | 0.546683 |
| <i>Bifidobacterium</i> | 0.000375 | 0.202647 | 0.000547 | 0.037611 | 0.006203 | 0.118112 |
| <i>Blautia</i> | 0.001910 | 0.101855 | 0.003534 | 0.011417 | 0.004318 | 0.013964 |
| Clostridiales | 0.000459 | 0.005788 | 0.000191 | 0.001489 | 0.000954 | 0.005866 |
| <i>Clostridium</i> | 0.000000 | 0.000426 | 0.000000 | 0.002132 | 0.000000 | 0.000079 |
| <i>Collinsella</i> | 0.000047 | 0.072609 | 0.000675 | 0.048994 | 0.000000 | 0.033871 |
| <i>Coprococcus</i> | 0.002148 | 0.044156 | 0.003971 | 0.016411 | 0.004496 | 0.017551 |
| <i>Dorea</i> | 0.000992 | 0.021779 | 0.002807 | 0.013077 | 0.001666 | 0.009759 |
| <i>Eubacterium</i> | 0.002678 | 0.032840 | 0.002070 | 0.030184 | 0.002357 | 0.009520 |
| <i>Faecalibacterium</i> | 0.008492 | 0.134054 | 0.018780 | 0.055606 | 0.006180 | 0.149963 |
| <i>Parabacteroides</i> | 0.001869 | 0.024486 | 0.003306 | 0.024087 | 0.003679 | 0.036806 |
| <i>Prevotella</i> | 0.000534 | 0.081968 | 0.060466 | 0.354508 | 0.001056 | 0.008328 |
| Ruminococcaceae | 0.003223 | 0.023071 | 0.000393 | 0.005652 | 0.001734 | 0.014463 |
| <i>Ruminococcus</i> | 0.001375 | 0.067108 | 0.002020 | 0.013723 | 0.000324 | 0.008212 |
| <i>Streptococcus</i> | 0.000000 | 0.005971 | 0.000095 | 0.041858 | 0.000359 | 0.006394 |

Supplementary Table 24. Enterotype cluster robustness and predictive performance assessment.

| Dataset | performance measure | estimate for real data | p-value (fraction of the simulations with a better value than real data) | | |
|-------------|---------------------|------------------------|---|------------------|-------------------|
| | | | random-uniform | template-uniform | template-gaussian |
| 33 Sanger | Sillhouette | 0.1899 | 0 | 0.001 | 0.01 |
| | Accuracy | 0.9091 | 0 | 0.013 | 0.039 |
| | Precision | 0.8889 | 0 | 0.013 | 0.04 |
| | Avg. Gain | 3.1762 | 0 | 0.003 | 0.005 |
| 154 16S | Sillhouette | 0.2133 | 0 | 0 | 0 |
| | Accuracy | 0.8600 | 0 | 0.001 | 0.022 |
| | Precision | 0.8729 | 0 | 0.001 | 0.01 |
| | Avg. Gain | 3.9823 | 0 | 0 | 0.017 |
| 85 Illumina | Sillhouette | 0.2490 | 0 | 0 | 0 |
| | Accuracy | 0.9760 | 0.001 | 0 | 0.009 |
| | Precision | 0.9370 | 0 | 0.008 | 0.052 |
| | Avg. Gain | 5.2600 | 0 | 0 | 0.005 |