Supplementary Information

Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes

Anna Heintz-Buschart[1]*, Patrick May[1], Cédric C Laczny[1], Laura A Lebrun[1], Camille Bellora[2], Abhimanyu Krishna[1], Linda Wampach[1], Jochen G Schneider[1], Angela Hogan[2], Carine de Beaufort[1,3], Paul Wilmes[1]*

1) Luxembourg Centre for Systems Biomedicine, 7, avenue des Hauts-Fourneaux, 4362 Esch-sur-Alzette, Luxembourg
2) Integrated BioBank of Luxembourg, 6, rue Nicolas Ernest Barblé, 1210, Luxembourg.
3) Saarland University Medical Center, Department of Internal Medicine II, 66421 Homburg, Germany.
4) Centre Hospitalier Emile Mayrisch, Rue Emile Mayrisch, 4240 Esch-sur-Alzette, Luxembourg.
5) Clinique Pediatrique - Centre Hospitalier de Luxembourg, 4, rue Nicolas Ernest Barblé, 1210, Luxembourg

*Corresponding authors: paul.wilmes@uni.lu, anna.buschart@uni.lu

This file contains supplementary figures, supplementary table legends, supplementary notes and supplementary references.

This file contains the following supplementary figures:

	dihydroflavonol-4-reductase, in the metaproteomes of individuals with T1DM
SF20	Correlation of AMY2 proteins in stool with transcript abundances of genes involved in thiamine metabolism
SF21	Workflow for taxonomic and functional annotation of different omic datasets
SF22	Co-assembly workflow for the integration of metagenomic and metatranscriptomic data
SF23	Algorithm for automatic binning of contigs based on coordinates from non-linear dimension reduction of pentamer-frequencies, presence of essential single-copy genes and metagenomic depths of coverage
SF24	Workflow for the construction of search databases for metaproteomic analyses
SF25	Structure of the database comprising all contigs and genes with annotations, depths of coverage and sequence characteristics
SF26	Taxonomic origins of proteins represented by protein groups
SF27	Diversity and temporal variability of the microbiome
SF28	Functional differences between families
SF29	The top-scoring module of metabolic functions in a community-wide metabolic network reconstruction, based on differential metatranscriptomic abundance between individuals with T1DM and healthy relatives

Legends of the Supplementary Tables are listed on page 56.

This file contains the following supplementary notes:

Supplementary References are listed starting on page 108.

Supplementary Figures

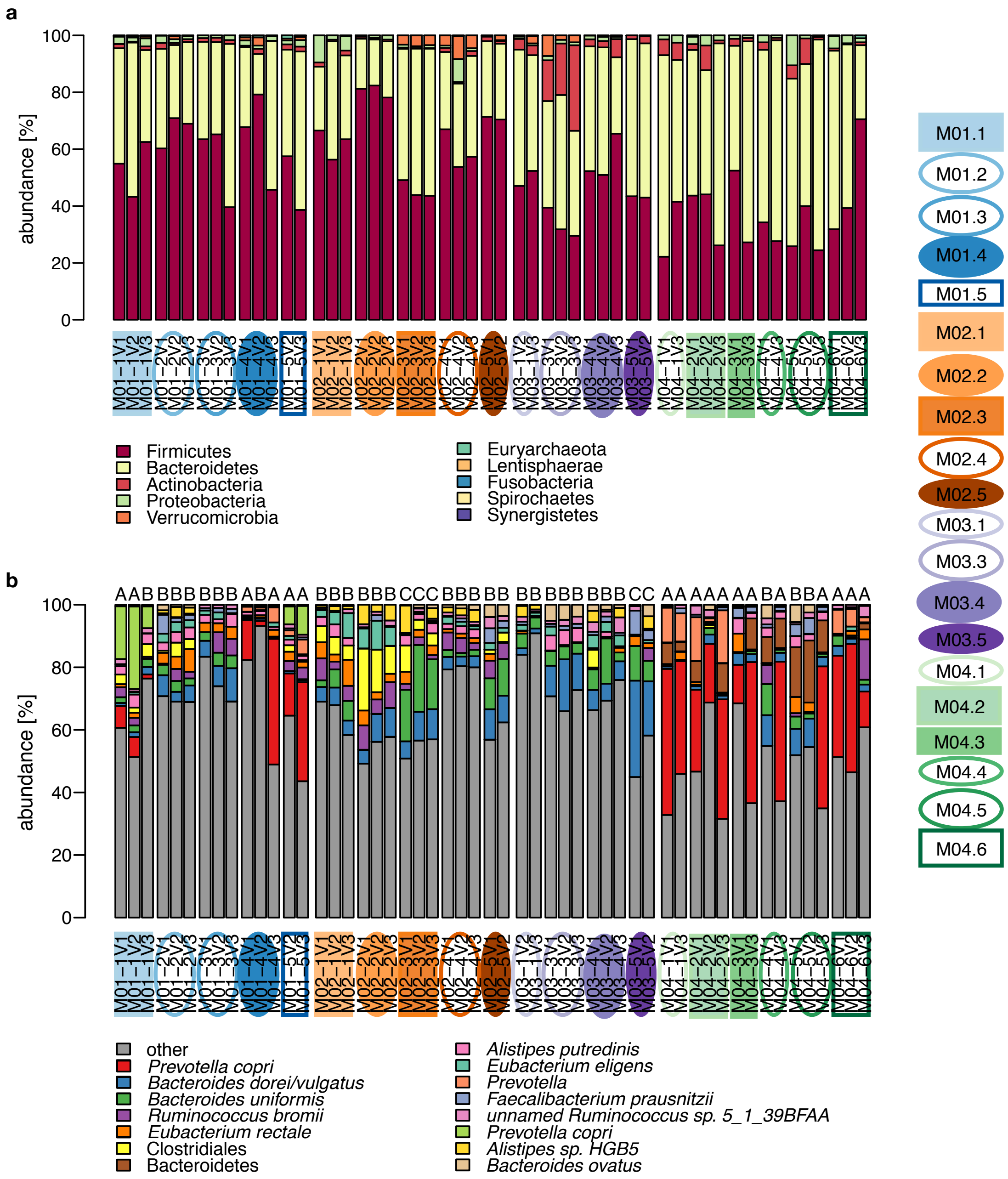

Figure S1: Taxonomic overview derived from metagenomic data. **a)** Phylum-level microbial community structures based on the mapping of metagenomic data to marker genes of metagenomic operational taxonomic units (mOTUs). **b)** Most abundant mOTUs in the metagenomic dataset. Letters above the bars indicate clustering of genus-level profiles from

mapping the metagenomic reads against an external reference gene catalogue[1] with metagenomic datasets from previous studies[1]. **a & b**) Bars represent relative abundances of the indicated taxa in each sample. The colour-coded boxes indicate the different study participants who donated the samples. The colour-scheme is identical to that used in the other figures as defined in **Figure 1a**.

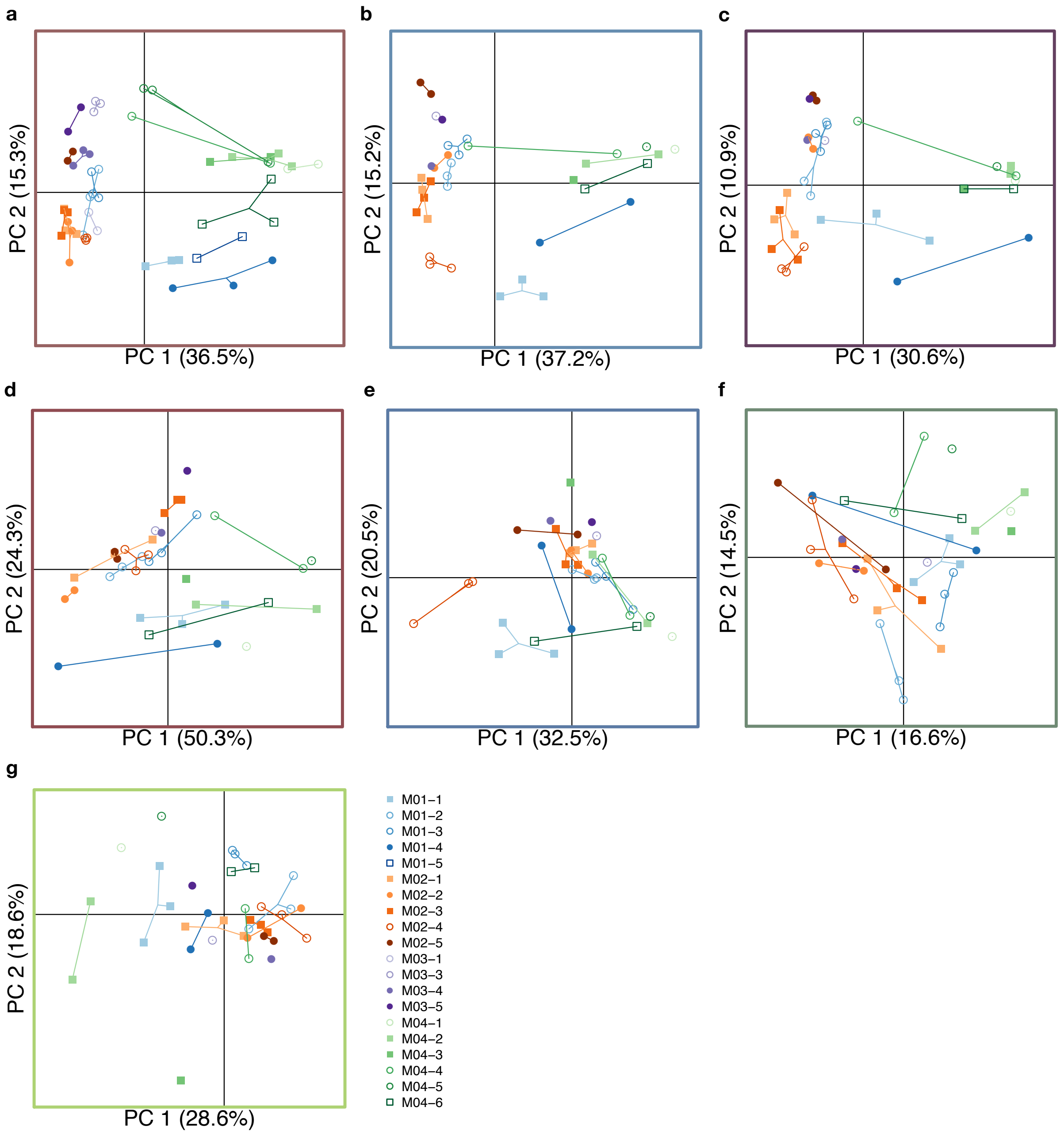

Figure S2: Jensen-Shannon divergence-based ordinations of metagenomic, metatranscriptomic and metaproteomic data. Displays of the first two principal components from principal coordinate analyses of distances between **a**) mOTU abundance profiles within the metagenomes, **b**) mOTU abundance profiles within the metatranscriptomes, **c**) relative abundances of binned population-level genomes within the metagenomes, **d**) metagenomic functional profiles, **e**) metatranscriptomic functional profiles,

f) metaproteomic functional profiles, **g**) human protein abundances. **a - g**) Samples of the same individual are connected by lines and every individual is represented by the same symbols as in all figures, see legend and **Figure 1a**; the colours of the boxes around the plots are consistent with the colour schemes for taxonomic and functional profiles at the metagenomic, metatranscriptomic and (meta-)proteomic levels (as defined in **Figure 1b**).

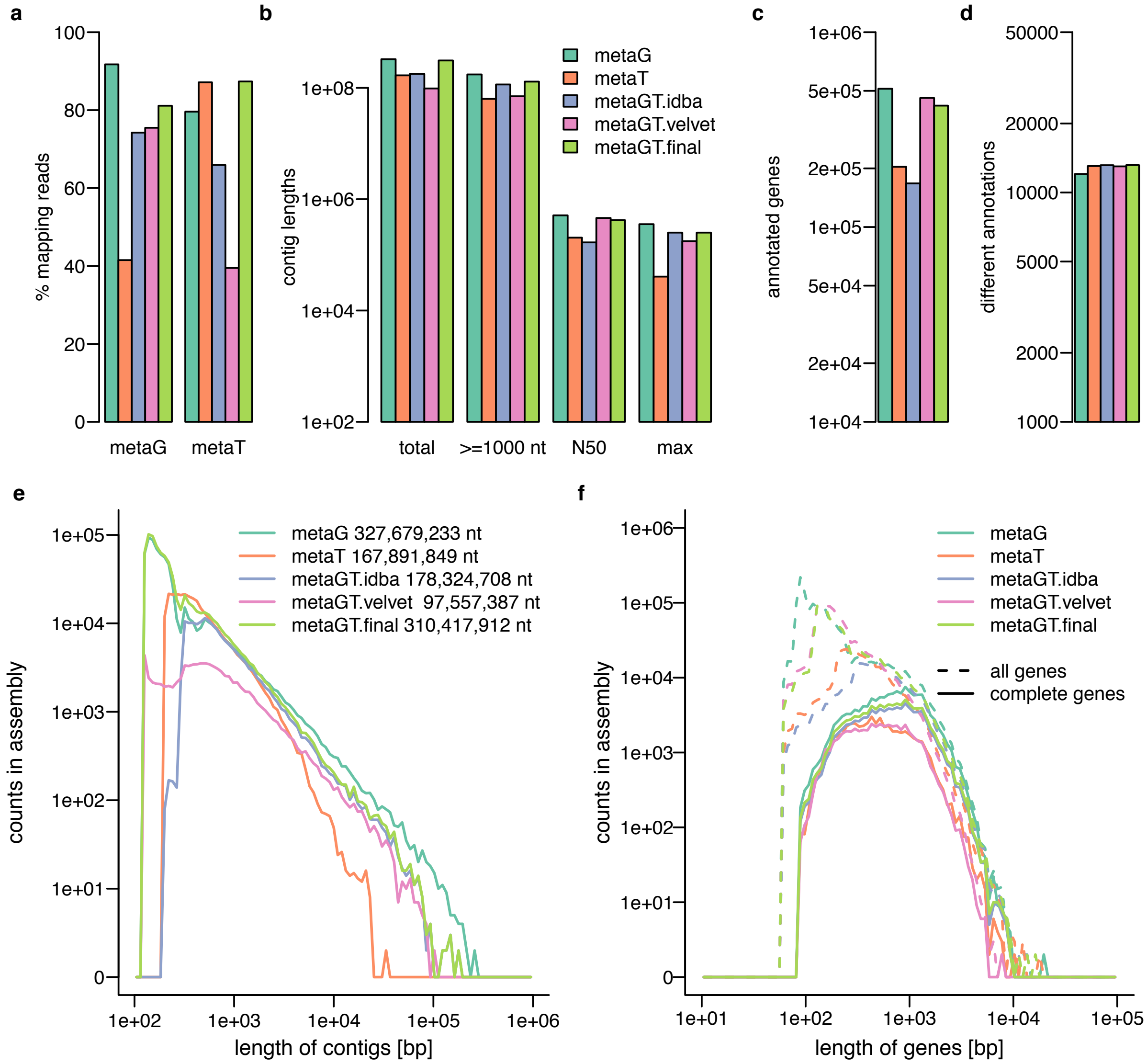

Figure S3: Comparison of metagenomic, metatranscriptomic and co-assemblies. **a**) Percentages of metagenomic reads mapped to contigs and metatranscriptomic reads mapping in sense-direction to protein-coding genes, **b**) total assembly lengths, total lengths of contigs >= 1 kbp, N50, and maximal contig lengths, **c**) number of genes annotated with a KO, **d**) number of different annotations, **e**) histograms of obtained contig lengths, **f**) histogram of obtained gene lengths using five different assembly strategies on one sample (M1.1-V1). **a - f**) metaG: assembly of metagenomic reads, metaT: assembly of metatranscriptomic reads; metaGT.idba: co-assembly of metagenomic and metatranscriptomic reads using IDBA-UD[2];

metaGT.velvet: co-assembly of metagenomic and metatranscriptomic reads, and contigs from IDBA-UD as long-read input, in velvet[3]; metaGT.final: co-assembly of metagenomic and metatranscriptomic reads using IDBA-UD, velvet and Newbler[4], as used in the rest of the manuscript.

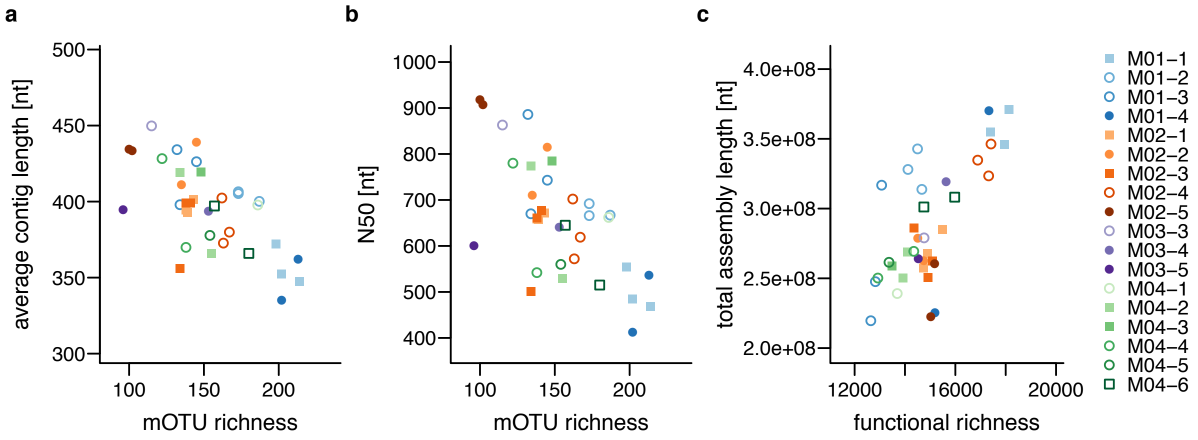

Figure S4: Taxonomic and functional diversity and assembly. a) Average contig lengths and **b**) N50 plotted against taxonomic richness (number of distinct mOTUs in the metagenomic data) in the same sample. **c**) Total assembly length compared to functional richness (number of different functional annotations) in the same sample. The colour/symbol-scheme denoting the individuals is identical to that used in the other figures as defined in **Figure 1a**.

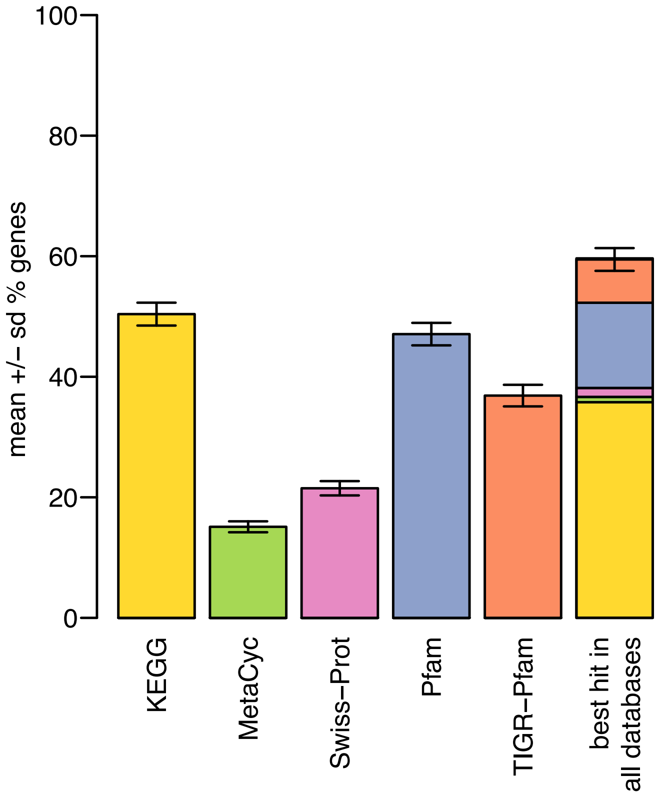

Figure S5: Proportion of genes with functional annotations. Means of the percentages of genes annotated with a functional category from the five indicated HMM databases used, as well as a break-down of the databases delivering the best annotations (yellow - KEGG, green - MetaCyc, pink - Swiss-Prot, blue - Pfam, orange - TIGR-Pfam). Error bars indicate the standard deviation among all samples (n = 36).

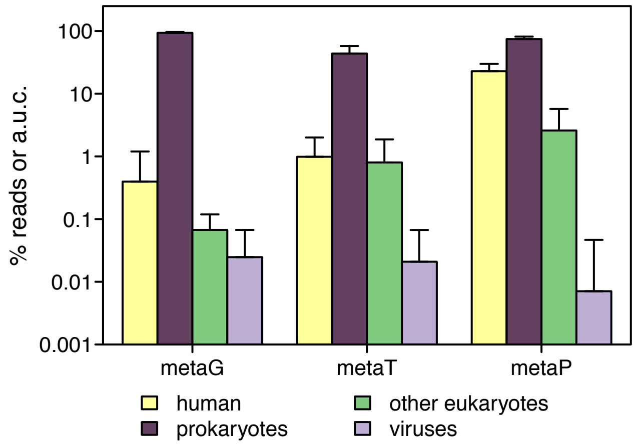

Figure S6: Broad taxonomic view of the omic datasets. Log-scaled mean proportions of reads in all metagenomic and metatranscriptomic dataset which mapped to human, non-human eukaryotic, bacterial and archaeal (prokaryotes) and viral genes, as well as for metaproteomic datasets the areas under the ion-chromatography curve of the respective proteins. Error bars represent one standard deviation of all samples (n = 36).

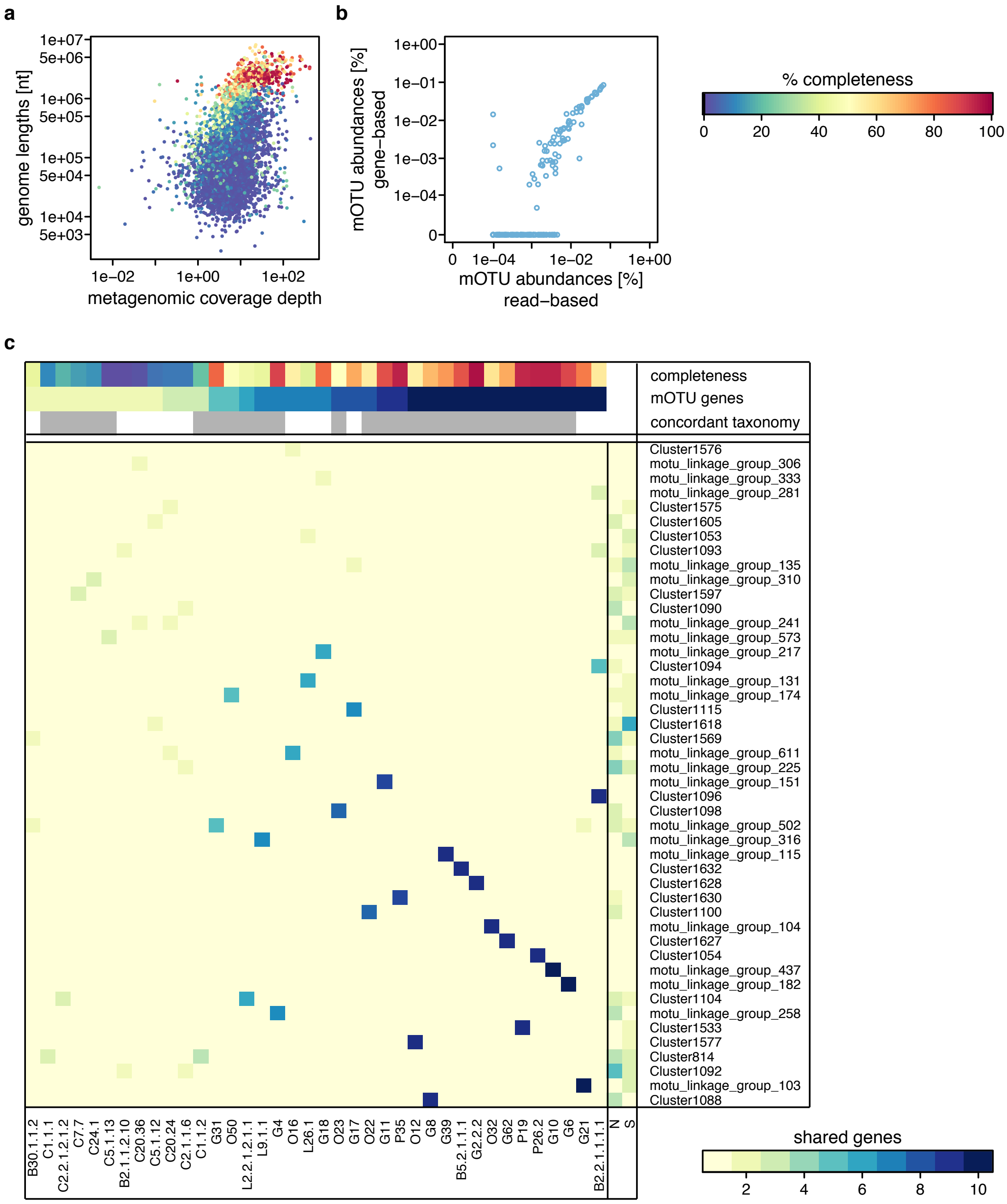

Figure S7: Results of contig binning and comparison of taxonomic composition derived from read-based, assembly-dependent and binning-dependent approaches. a) Length, metagenomic coverage and completeness of all binned population-level genomes with at least one essential gene in all 36 samples. Colours of the dots represent completeness, see colour key in top right corner of the figure. **b**) Metagenomic mOTU abundances calculated from

reads mapped against a collection of phylogenetic marker genes compared to the abundances calculated based on reads mapping to assembled genes annotated with the taxonomy of the most similar marker gene. **c**) Heatmap of numbers of phylogenetic marker genes within binned population-level genomes with association to mOTUs, see lower colour key in top-right corner of the figure. The two columns furthest to the right represent contigs too short to be binned (S) and contigs that were annotated as noise by the binning algorithm (N). All other columns represent single population-level genomes, while rows represent single mOTUs. The uppermost coloured bar on top of the heatmap represents the completeness of the genomes based on the presence of essential genes (see top-most colour key), the middle bar represents the total numbers of phylogenetic marker genes in the genomes (see lower colour key), the bar below represents whether all marker genes in the bin-genomes have the same taxonomic annotation (grey) or not (white).

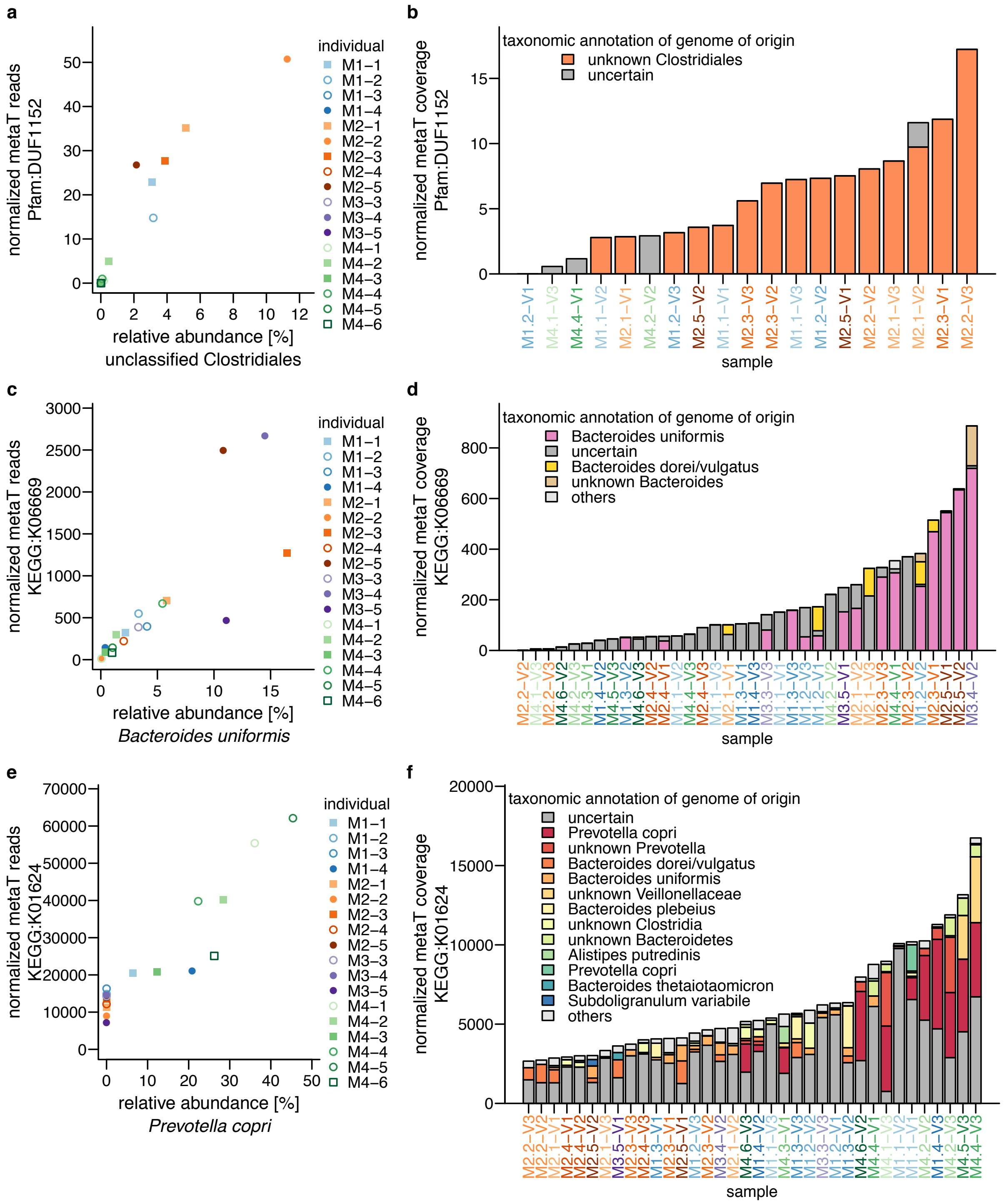

Figure S8: Examples for linking specific functional genes to microbial populations. a, c & e) Correlation-based approach. Transcript abundances of **a**) Pfam domain of unknown function DUF1152, **c**) K06669 (SMC3, CSPG6; structural maintenance of chromosome 3 (chondroitin sulphate proteoglycan 6)), **e**) K01624 (fructose-bisphosphate aldolase, class II)

are plotted against relative abundances inferred from metagenomic data of the most strongly correlated mOTUs, **a**) a Clostridiales mOTU, **c**) *Bacteroides uniformis*, **e**) *Prevotella copri*. Examples are chosen from pairs of correlating mOTUs and metatranscriptomic functions. **a, c & e)** The colours and symbols denoting the individuals is identical to that used in the other figures, see **Figure 1a**. See **Supplementary Table 1A** for total number of samples per individual. **b, d & f)** Bin-genome-based approach. Normalized transcript abundances of the same functional genes (**b**) Pfam domain of unknown function DUF1152, **d**) K06669, **f**) K01624) and the taxonomy of the expressing binned population-level genomes. "uncertain" summarizes values for all binned population-level genomes without unanimously annotated taxonomy and contigs carrying genes of interest which could not be binned.

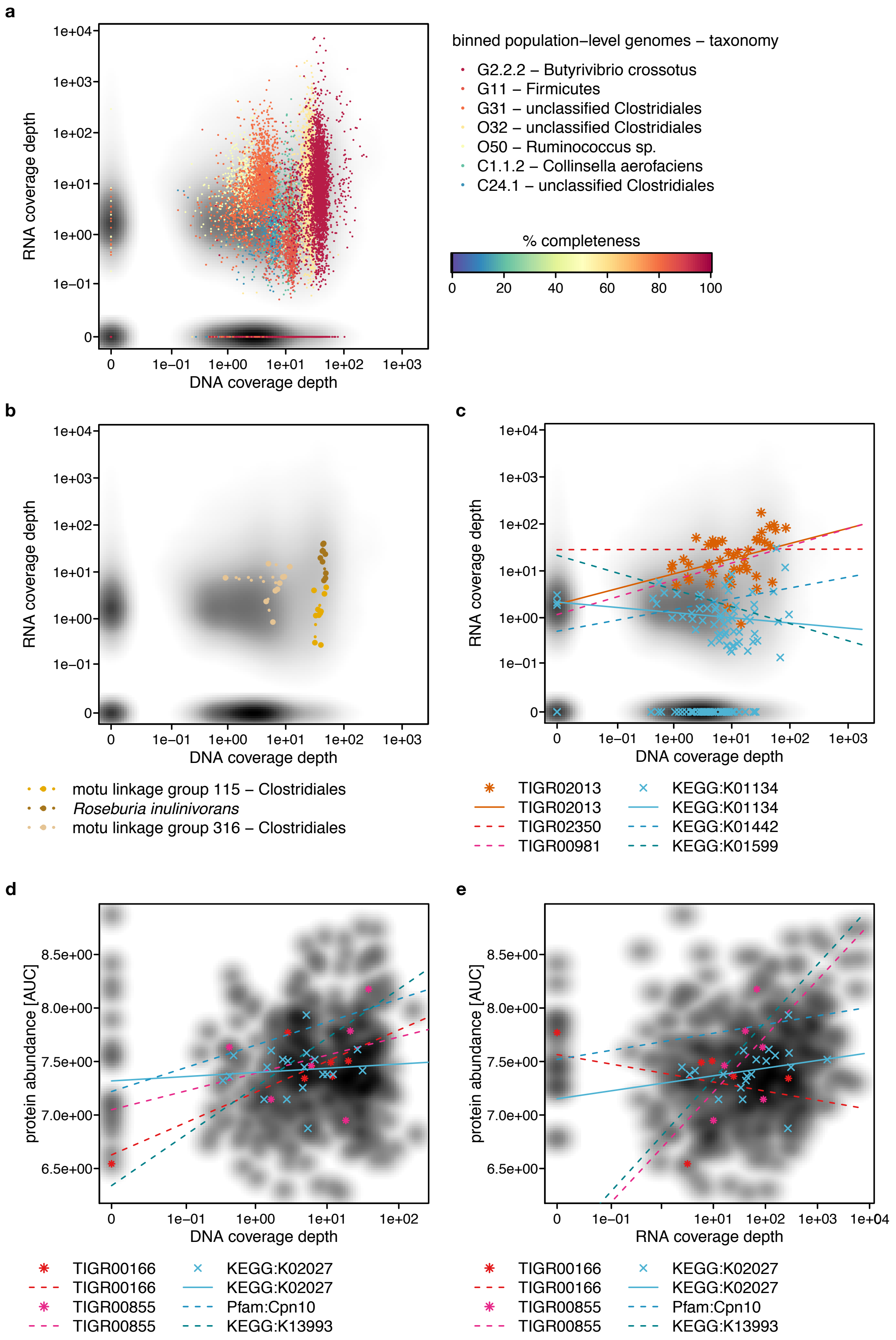
a
RNA coverage depth
1e+04
1e+03
1e+02
1e+01
1e+00
1e−01
0
DNA coverage depth
0
1e−01
1e+00
1e+01
1e+02
1e+03
binned population−level genomes – taxonomy
G2.2.2 – Butyrivibrio crossotus
G11 – Firmicutes
G31 – unclassified Clostridiales
O32 – unclassified Clostridiales
O50 – Ruminococcus sp.
C1.1.2 – Collinsella aerofaciens
C24.1 – unclassified Clostridiales
% completeness
0
20
40
60
80
100
b
RNA coverage depth
DNA coverage depth
motu linkage group 115 – Clostridiales
Roseburia inulinivorans
motu linkage group 316 – Clostridiales
c
RNA coverage depth
DNA coverage depth
TIGR02013
TIGR02013
TIGR02350
TIGR00981
KEGG:K01134
KEGG:K01134
KEGG:K01442
KEGG:K01599
d
protein abundance [AUC]
8.5e+00
8.0e+00
7.5e+00
7.0e+00
6.5e+00
DNA coverage depth
0
1e−01
1e+00
1e+01
1e+02
TIGR00166
TIGR00166
TIGR00855
TIGR00855
KEGG:K02027
KEGG:K02027
Pfam:Cpn10
KEGG:K13993
e
protein abundance [AUC]
RNA coverage depth
0
1e−01
1e+01
1e+02
1e+03
1e+04
TIGR00166
TIGR00166
TIGR00855
TIGR00855
KEGG:K02027
KEGG:K02027
Pfam:Cpn10
KEGG:K13993

Figure S9: Relationship between metagenomic, metatranscriptomic and metaproteomic abundance of predicted genes. Data from the assembly of sample M01.2-V1 are shown. **a**) Density representation of distributions of metagenomic and metatranscriptomic depths of coverage of all predicted open reading frames (ORFs). ORFs of 7 reconstructed genomes are highlighted in colours according to their respective genome completenesses. **b**) Density representation of distributions of metagenomic and metatranscriptomic depths of coverage of all predicted ORFs. Brown circles connected by dotted lines represent different phylogenetic marker genes of common phylogeny, with the highest, lowest and closest to the average relative variation in metagenomic depths of coverage. **c)** Density representation of distributions of metagenomic and metatranscriptomic coverage depths of all predicted ORFs. Lines represent linear regressions of the metagenomic and metatranscriptomic abundances of predicted genes with the same functional annotation; red lines: house-keeping (essential) genes; green lines: other functional categories. Functional categories with the lowest and highest correlations between metagenomic and metatranscriptomic coverage, as well as the correlation closest to the average, were selected. Red stars indicate genes annotated as *rpoB* (TIGR02013, representative of essential housekeeping genes), and blue crosses indicated genes with the annotation of K00603, a formiminotransferase. Coefficients of the linear regressions varied between 0.002 and 0.248. **d)** Density representation of distributions of metagenomic depths of coverage and relative metaproteomic abundances of predicted ORFs with uniquely identified proteins. **e)** Density representation of distributions of metatranscriptomic depths of coverage and relative metaproteomic abundances of predicted ORFs with uniquely identified proteins. **d & e**) Lines represent linear regressions of the metagenomic or metatranscriptomic depths of coverage with the relative protein abundances of all predicted genes with the same functional annotation; red lines and symbols: house

keeping (essential) genes; blue lines and symbols: other functions. The functional categories with the lowest and highest correlations between metagenomic or metatranscriptomic and metaproteomic abundances were selected, and (in the case of the non-essential functional categories) the correlation closest to the average of all correlations, presented by the transport protein K02027.

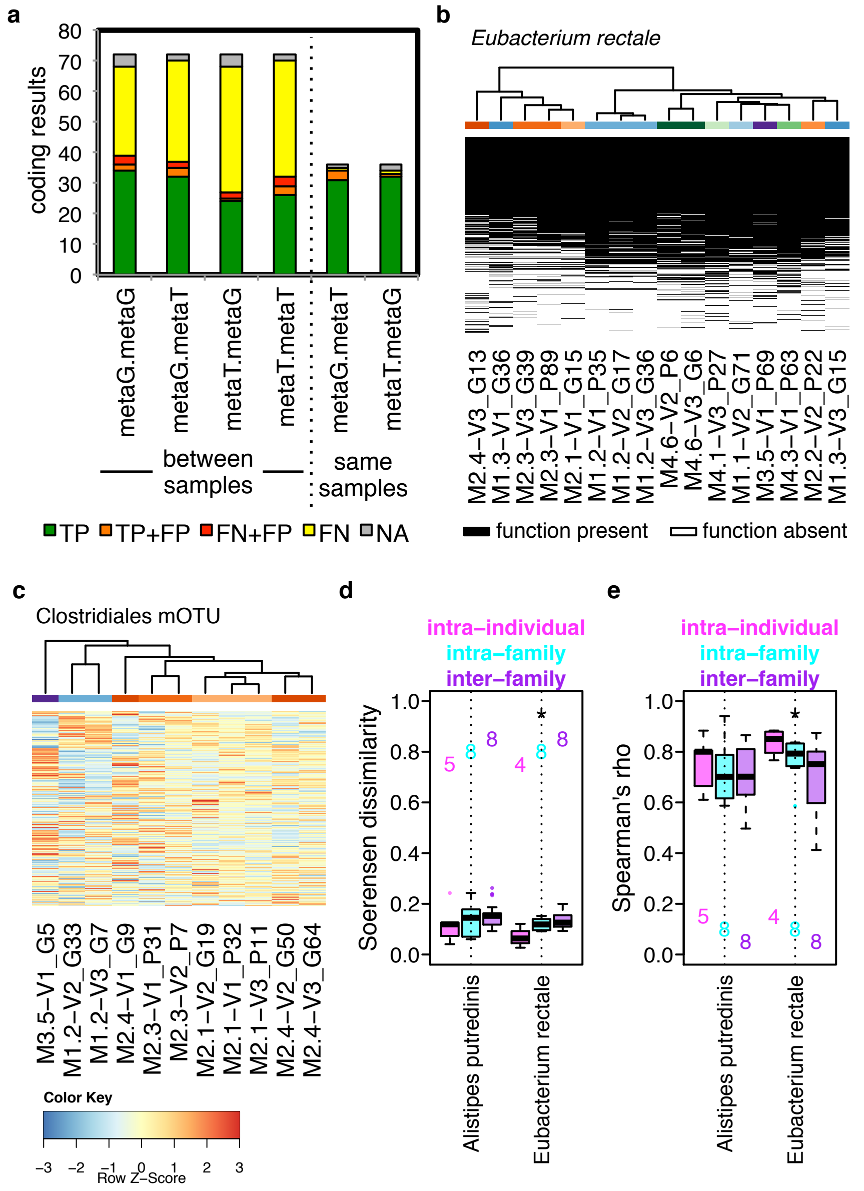

Figure S10: Intra-individual and intra-family similarities of faecal microbial communities and community members. **a)** Efficiency of metagenomic and metatranscriptomic codes based on mOTU abundances for recognizing samples of the same individuals[5]. Results of all combinations of codes and recognized samples are displayed. The four bars on the left ("between samples") represent the coding results whereby the code was derived from a different sample than the test samples; metaG.metaG: metagenomic codes as applied to metagenomic profiles; metaG.metaT: metagenomic codes as applied to

metatranscriptomic profiles; metaT.metaG: metatranscriptomic codes as applied to metagenomic profiles; metaT.metaT: metatranscriptomic codes as applied to metatranscriptomic profiles. The two bars on the right ("same samples") represent the coding results whereby the code was derived from the same sample it was applied to; metaG.metaT: metagenomic codes as applied to metatranscriptomic profiles; metaT.metaG: metatranscriptomic codes as applied to metagenomic profiles. TP: true positive, FP: false-positive; FN: false-negative; NA: no code could be generated. **b**) Presence (black) and absence (white) of functional categories encoded in closely related genomes reconstructed from different faecal samples. The chosen genomes represent *Eubacterium rectale,* the mOTU with the highest difference between intra- and inter-individual concordances. Only genomes with > 67 % essential unique genes were evaluated. Hierarchical clustering is based on Soerensen dissimilarity indices. **c**) Heatmap of functional expression profiles of closely related population-level genomes reconstructed from different faecal samples. The chosen genomes represent the mOTU with the highest difference between intra- and inter-individual correlations (Clostridiales mOTU linkage group 126). Only genomes with > 67 % essential unique genes were assessed. Hierarchical clustering is based on Spearman's correlation coefficient. **b & c**) Colours at the top of the heatmaps indicate the sample donors. The colour-scheme denoting the individuals is identical to that used in the other figures, see **Figure 1a**. **d**) Comparisons of intra-individual, intra-family and inter-family Soerensen dissimilarity indices of functional categories within closely related reconstructed binned population-level genomes. **e**) Comparison of intra-individual, intra-family and inter-family Spearman's correlation coefficients of the expression profiles of functional categories present in all analysed genomes. **d & e**) Only genomes with > 67 % essential unique genes were evaluated. Only mOTUs whose occurrence allowed comparisons between all cases are displayed. *

significant direct comparison (*P* value < 0.05, Wilcoxon rank sum test). Pink, purple and blue numbers provide the numbers of compared genomes. Boxes span the first to third quartiles, the central thick bars represent the medians, whiskers extend to 1.5 times the interquartile ranges and points outside these ranges are represented as outlier points.

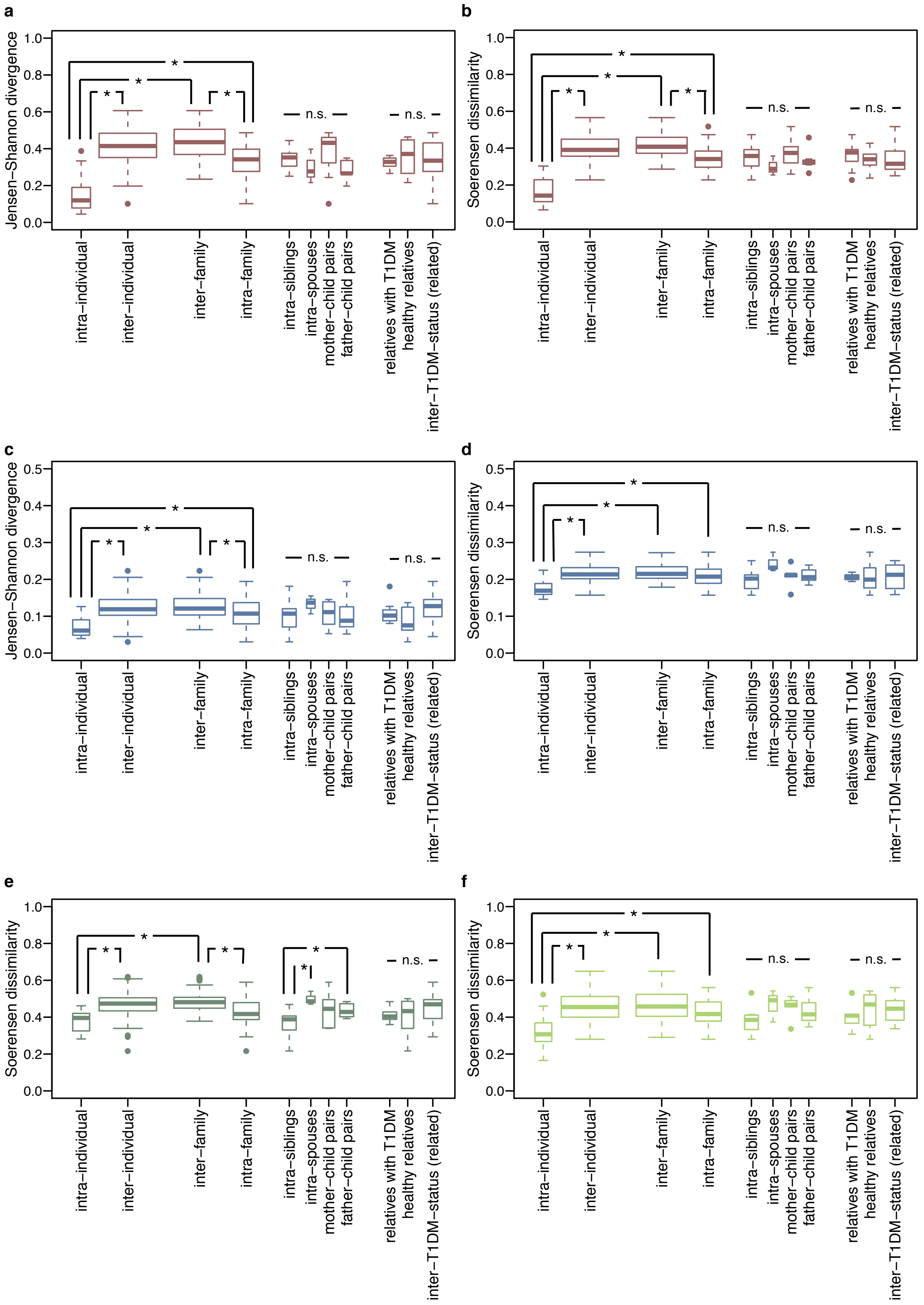
a
b
c
d
e
f
Jensen−Shannon divergence
Soerensen dissimilarity
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.8
1.0
*
n.s.
intra−individual
inter−individual
inter−family
intra−family
intra−siblings
intra−spouses
mother−child pairs
father−child pairs
relatives with T1DM
healthy relatives
inter−T1DM−status (related)

Figure S11: Comparisons of intra-individual and inter-individual distances and distance between intra-family-groups based on multi-omic datasets. **a**) Jensen-Shannon divergences based on metagenomic taxonomic profiles. **b**) Soerensen dissimilarity indices based on metagenomic mOTU occurrences. **c**) Jensen-Shannon divergences based on metatranscriptomic functional profiles. **d**) Soerensen dissimilarity indices based on metatranscriptomic occurrences of functional categories. **e**) Soerensen dissimilarity indices based on metaproteomic occurrences of functional categories. **f**) Soerensen dissimilarity indices based on metaproteomic occurrences of human proteins. **a - f**) * indicates *P* value < 0.05 in Wilcoxon rank sum test of the dissimilarity indices between the indicated groups; n.s. indicates *P* value >= 0.05 of Kruskal-Wallis tests of the indicated groups. Boxes span the first to third quartiles, the central thick bars represent the medians, whiskers extend to 1.5 times the interquartile ranges and points outside these ranges are represented as outlier points. The colours of the boxes are consistent with the colour schemes for taxonomic profiles at the metagenomic level and functional profiles at the metatranscriptomic and (meta-)proteomic levels (see **Figure 1b**).

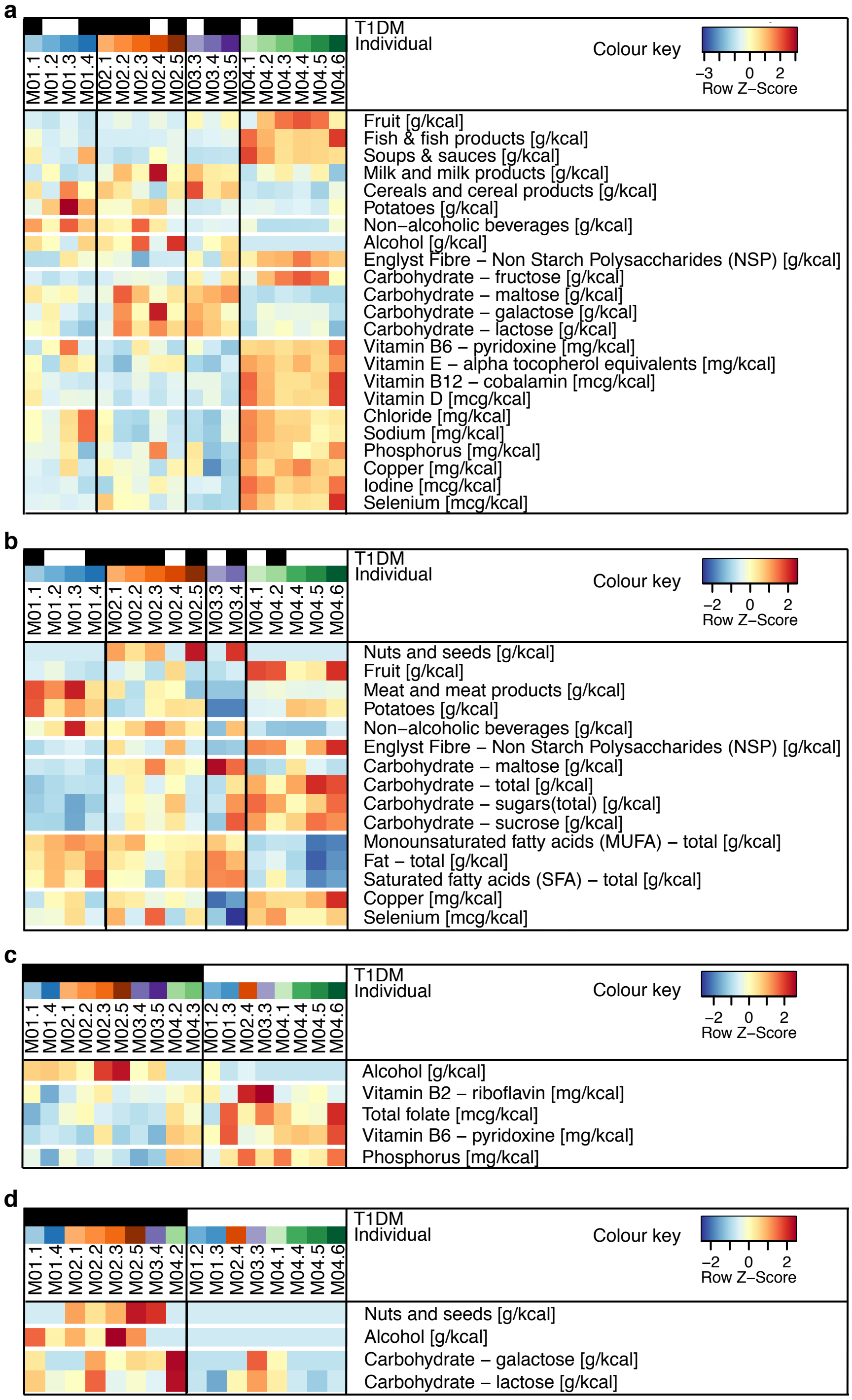

Figure S12: Family-specific nutritional patterns and differences in diet with respect to individuals with T1DM. Family-specific nutrient-uptake patterns based on **a**) 6 months and **b**) means from 24 h recall questionnaires. **a & b**) Only data with FDR-adjusted *P* values < 0.05 for the main effect of family membership in an ANOVA with T1DM and family

membership as factors are displayed. T1DM-specific nutrient-uptake pattern based on **c**) 6 months and **d**) means from 24 h recall questionnaires. **c & d**) None of the differences were significant after multiple testing adjustment. Only data with unadjusted *P* values < 0.02 for the main effect of family membership in an ANOVA with T1DM and family membership as factors and/or a Wilcoxon rank sum test between individuals with T1DM and their healthy relatives are displayed. **a - d**) Colours at the top of the heatmaps indicate the sample donors. The colour-scheme denoting the individuals is identical to that used in the other figures as defined in **Figure 1a**. Black boxes at the top of the heatmap indicate individuals with T1DM. See **Supplementary Table 1A** for total number of samples per individual

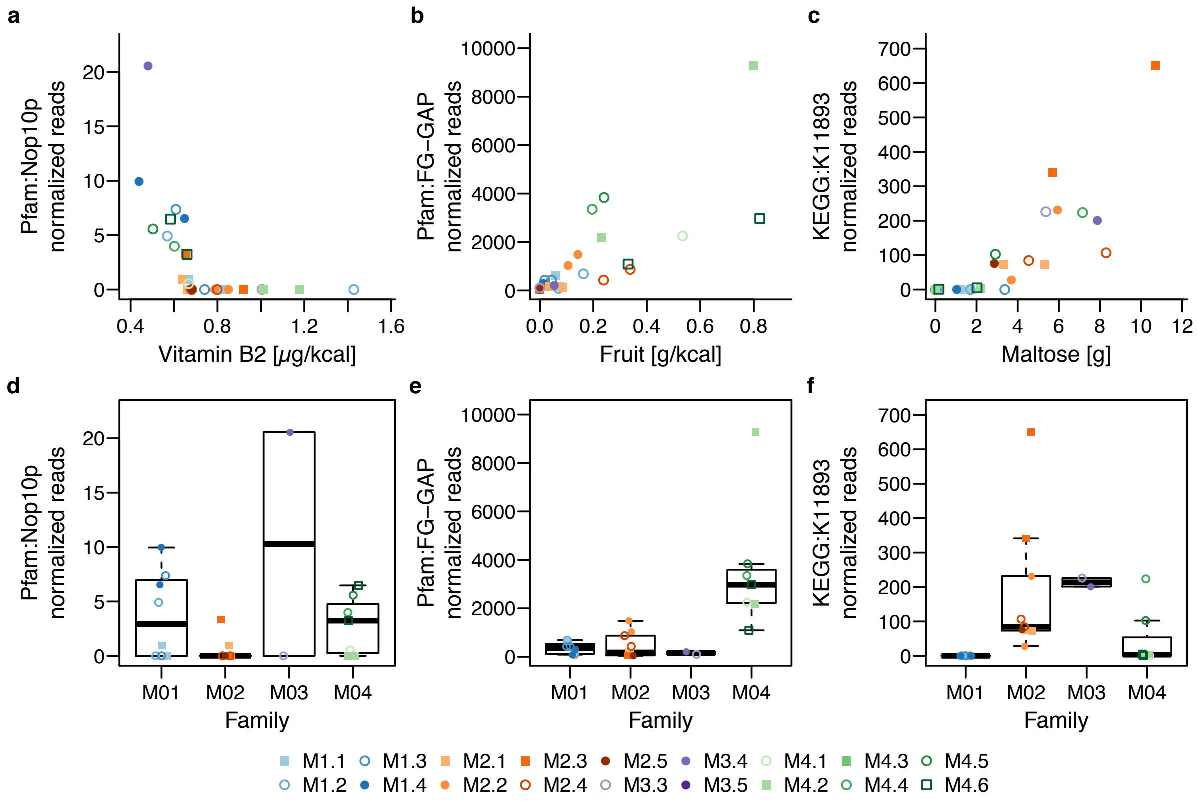

Figure S13: Correlation between (family-specific) food intake and microbial transcript abundances. a) Transcripts coding for potential archaea-derived proteins with a nucleolar-like protein domain *versus* estimated vitamin B2 (riboflavin) intake relative to total energy content of the diet. **b**) Transcripts coding for proteins with a FG-GAP domain plotted against consumption of fruit relative to total energy uptake. **c**) Transcripts for the type VI secretion protein K11893 plotted against estimated maltose intake. **d**) Transcript abundances for Nop10p proteins, grouped by families. Transcript abundances of family-specific functional categories, **e**) FG-GAP domain proteins and **f**) K11893, grouped by family membership. **d**, **e** & **f**) Boxes span the first to third quartiles, the central thick bars represent the medians, and whiskers extend to 1.5 times the interquartile ranges; all data points are represented; see **Supplementary Table 1A** for total number of samples per individual. **a**, **b**, **d**, **e** & **f**) The

colour/symbol-scheme denoting the individuals is identical to that used in the other figures, see **Figure 1a**.

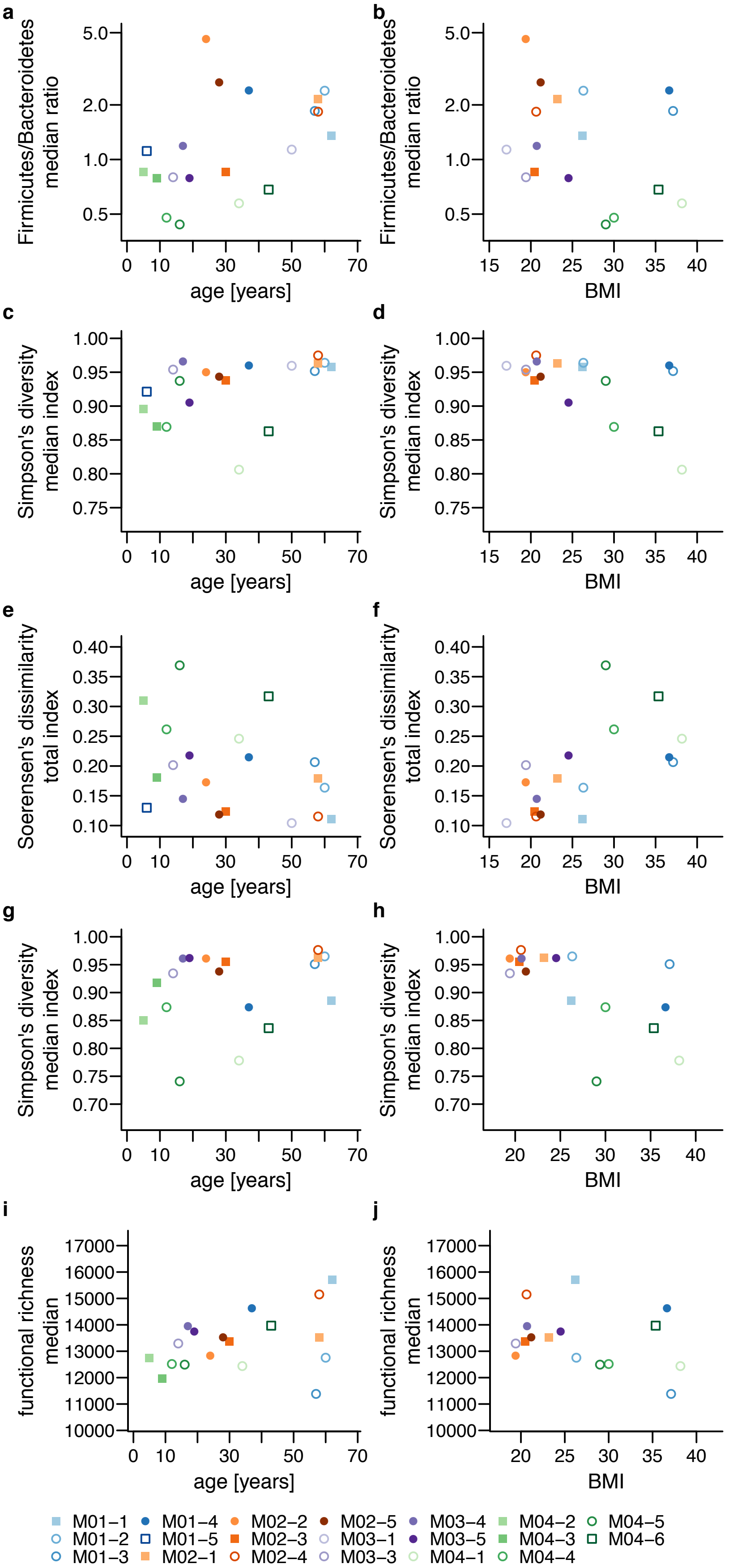
a
b
c
d
e
f
g
h
i
j
Firmicutes/Bacteroidetes median ratio
Simpson's diversity median index
Sorensen's dissimilarity total index
functional richness median
age [years]
BMI
M01–1
M01–2
M01–3
M01–4
M01–5
M02–1
M02–2
M02–3
M02–4
M02–5
M03–1
M03–3
M03–4
M03–5
M04–1
M04–2
M04–3
M04–4
M04–5
M04–6

Figure S14: Associations between donor age, body mass index (BMI), Firmicutes-to-Bacteroidetes ratio and microbial diversity. Ratios of the median relative abundances of Firmicutes and Bacteroidetes in the metagenomes versus the **a**) ages and **b**) BMIs of the donors. Median taxonomic diversity indices of the metagenomes versus the **c**) ages and **d**) BMIs of the donors; Simpson's diversity indices based on mOTU abundances correlated with donor age (Spearman's $\rho = 0.55$, *P* value = 0.01). Total Soerensen diversities (measure of diversity between samples or variability over time) of metagenomic taxonomic composition within samples of each individual versus their **e**) ages and **f**) BMIs; the temporal variability was correlated with BMI (Spearman's $\rho = 0.67$, *P* value = 0.004). Taxonomic diversity indices of the metatranscriptome versus the **g**) ages and **h**) BMIs of the donors. Functional richness of the metatranscriptome versus the **i**) ages and **j**) BMIs of the donors. **b**, **d**, **f**, **h** & **j**) BMIs are only displayed for individuals of at least 12 years of age. **a**, **b**, **d**, **e**, **f**, **g**, **h**, **i** & **j**) See **Supplementary Table 1A** for the total number of samples per individual. The colour/symbol-scheme denoting the individuals is identical to that used in the other figures, see legend below and as defined in **Figure 1a**.

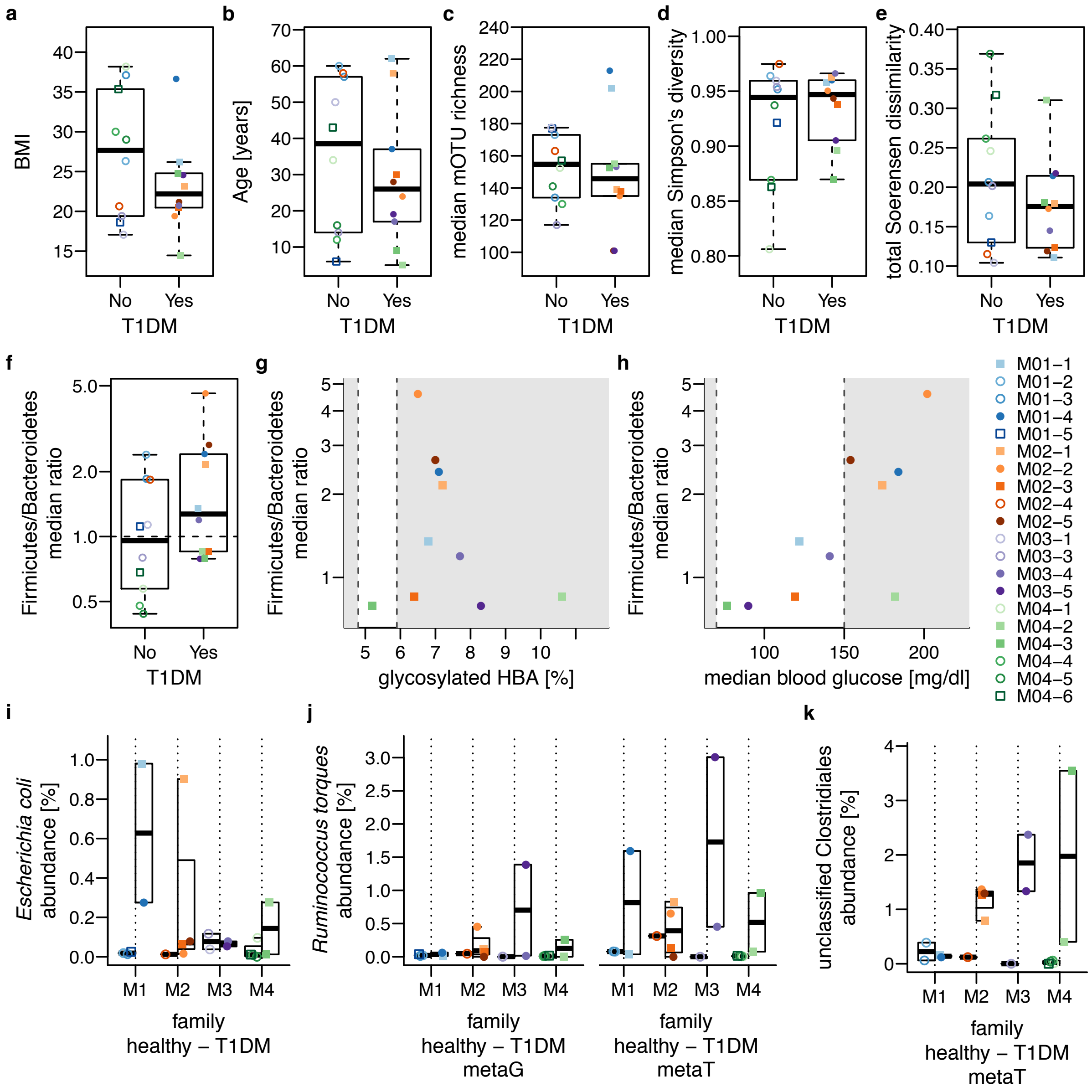

Figure S15: T1DM and microbial community structures. Distribution of **a)** BMI and **b)** age in the individuals with T1DM and their healthy relatives. **c)** mOTU richness, **d)** diversity, **e)** temporal variability of the faecal microbiota of individuals with T1DM and of the healthy relatives. **f)** Firmicutes-to-Bacteroidetes ratios in the faecal microbiota of individuals with T1DM compared to their healthy relatives. **g)** The Firmicutes-to-Bacteroidetes plotted against the glycation status of haemoglobin (HBA) in blood of individuals with T1DM, and **h)** the blood glucose levels of individuals with T1DM; grey

areas indicate values outside the healthy norm. **i**) Median metagenomic abundances of *E. coli*, the species in the metagenomic data set with the greatest difference between individuals with T1DM and their healthy family members (FDR-adjusted *P* value of main effect of T1DM in DESeq analysis of T1DM status and family membership = 0.09). **j**) Median metagenomic and metatranscriptomic abundances of [*Ruminococcus*] *torques* in individuals with T1DM (no significant difference). **k**) Metatranscriptomic representation of an unclassified mOTU of the order Clostridiales, the only mOTU with a significantly higher abundance in individuals with T1DM compared to the healthy family members (FDR-adjusted *P* value of main effect of T1DM in DESeq analysis of T1DM status and family membership < 0.05). **a**, **b**, **c**, **d**, **e**, **f**, **i**, **j** & **k**) Boxes span the first to third quartiles, the central thick bars represent the medians, and whiskers extend to 1.5 times the interquartile ranges; all data points are represented. **c**, **d**, **e**, **f**, **g**, **h**, **i**, **j** & **k**) see **Supplementary Table 1A** for total number of samples per individual. **a**, **b**, **c**, **d**, **e**, **f**, **g**, **h**, **i**, **j** & **k**) The colour/symbol-scheme denoting the individuals is identical to that used in the other figures, see **Figure 1a**.

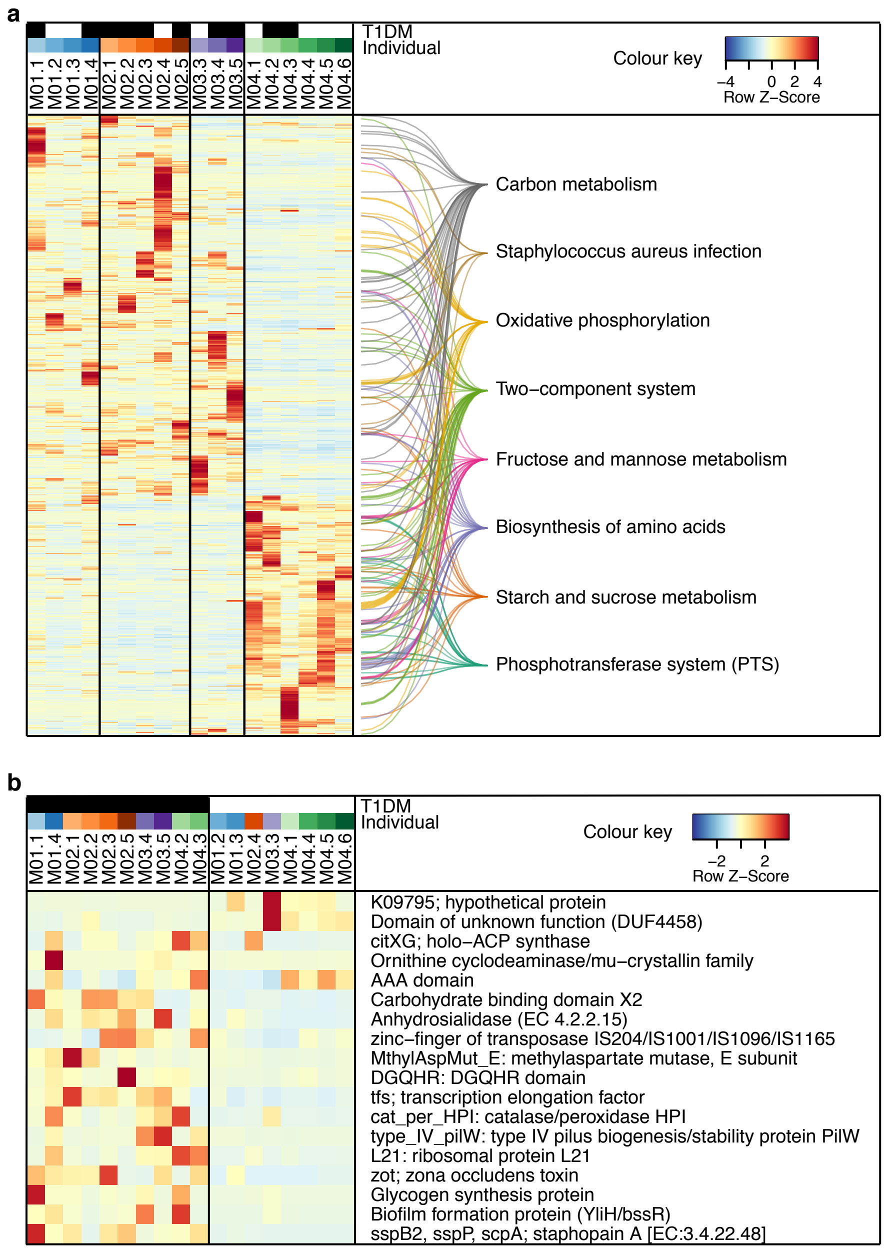

Figure S16: **Differential analysis of functional metatranscriptomic profiles with respect to family membership and T1DM. a**) Relative expression of KOs with a significant main effect of family membership are displayed (FDR-adjusted *P* value < 0.05). KOs belonging to pathways significantly enriched (FDR-adjusted *P* value < 0.05) with significantly differentially abundant KOs are connected to the names of the pathways. The KOs belonging to the "*Staphylococcus aureus* infection" pathway were not transcribed by *S. aureus*, and are

most likely surface proteins of other Gram-positive bacteria. **b**) Normalized abundances of differentially abundant functional transcripts (FDR-adjusted *P* value of main effect of T1DM in DESeq2 analysis of T1DM status and family membership < 0.05). The transcripts annotated as Glycogen synthesis protein (Pfam GlgS), are likely involved in regulation of motility rather than glycogen synthesis[6] and the transcripts annotated as Zot (Zonula occludens toxin) likely code for bacteriophage components and lack the active region of Zonula occludens toxin. The transcription elongation factor and the putative anhydrosylase functions may be related to diet rather than T1DM (see **Supplementary Note 17**). **a & b**) Colours at the top of the heatmaps indicate the sample donors, see **Supplementary Table 1A** for total number of samples per individual. The colour-scheme denoting the individuals is identical to that used in the other figures as defined in **Figure 1a**. Black boxes at the top of the heatmap indicate individuals with T1DM.

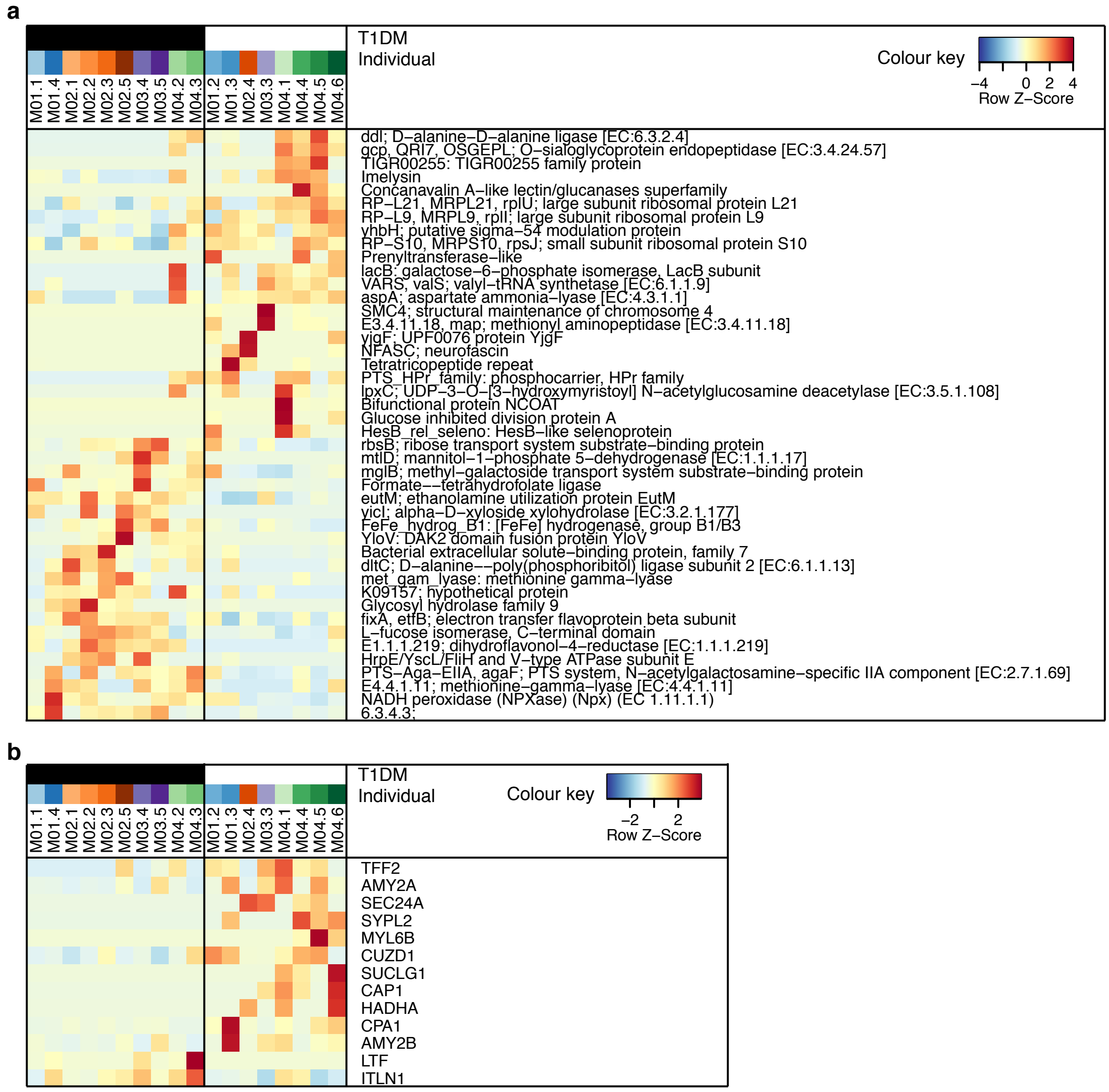

Figure S17: Metaproteomic differences in T1DM. a) Relative abundances of microbial proteins with the highest significance in a differential analysis of T1DM (unadjusted *P* value of Wilcoxon rank sum test < 0.05). **b**) Relative abundances of human proteins with the highest significance in a differential analysis of T1DM (unadjusted *P* value of Wilcoxon rank sum test < 0.05). **a & b**) Individuals with T1DM have a black box in the upper line on top of the columns. In the second line, the colour scheme denoting the individuals is identical to that used in the other figures, see **Figure 1a**. See **Supplementary Table 1A** for the total number

of samples per individual. Colour keys for the relative protein abundances are provided above each plot.

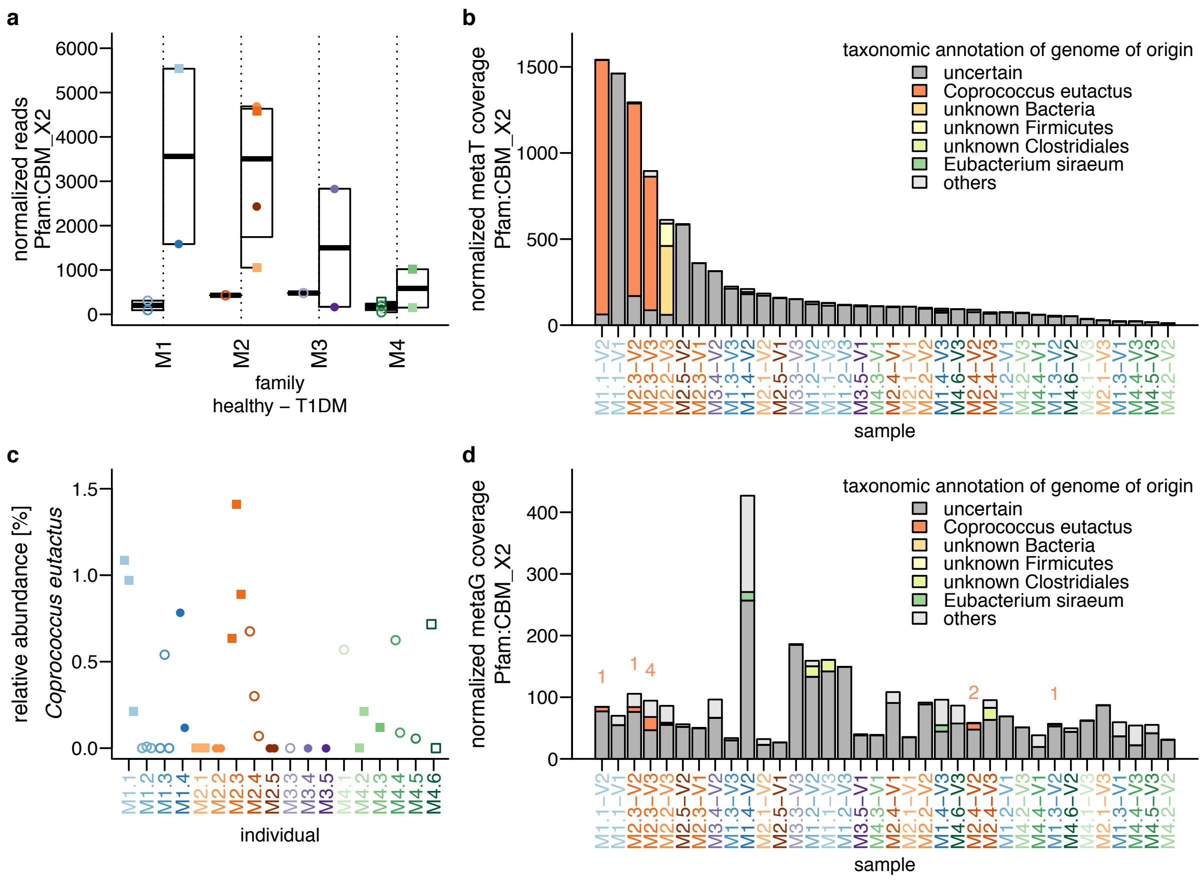

Figure S18: Identification of microbial populations contributing to elevated levels of CBM_X2, a structural domain in cellulose-degrading cellulosomes, in the metatranscriptomes of individuals with T1DM. **a**) Relative metatranscriptomic abundances of differentially abundant functional transcripts with a CBM_X2 domain. Boxes span the first to third quartiles, the central thick bars represent the medians, and whiskers extend to 1.5 times the interquartile ranges; all data points are represented, see **Supplementary Table 1A** for the total number of samples per individual. **b**) Metatranscriptomic depths of coverage of CBM_X2 genes in the different samples. **c**) Relative abundances of *Coprococcus eutactus*, based on assembly-independent mOTU analyses. **d**) Depths of coverage of genes of interest with metagenomic reads in different samples. Samples are sorted by transcript abundance, as in **b**. Orange numbers indicate

number of genes of interest associated with *C. eutactus* in the respective samples. **b & d**) The genomic context of each gene is displayed for genes making up at least 10 % of the total transcript abundance – remaining genes are gathered in “others”. **a & c**) The colour/symbol-scheme denoting the individuals is identical to that used in the other figures, see **Figure 1a**.

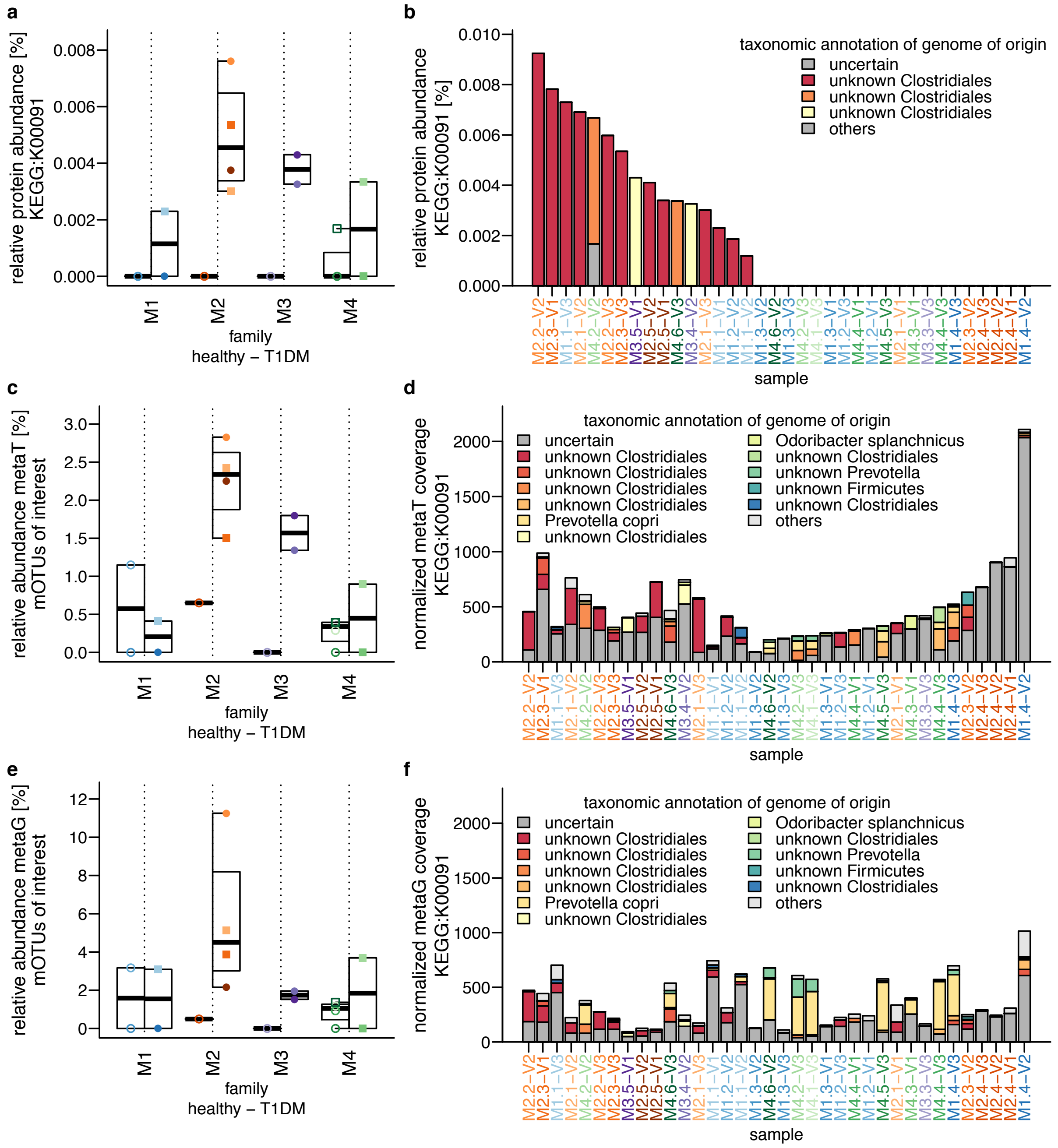

Figure S19: Identification of populations contributing to elevated levels of K00091, a dihydroflavonol-4-reductase, in the metaproteomes of individuals with T1DM. **a**) Relative protein abundances of K00091. **b**) Relative protein abundances in the different samples with the taxonomy of genomes of origin indicated. **c**) Relative metatranscriptomic abundances of the Clostridiales mOTUs expressing K00091. **d**) Depth of coverage of genes

of interest with metatranscriptomic reads in different samples with the taxonomy of genomes of origin indicated. **e**) Relative metagenomic abundances of the Clostridiales mOTUs expressing K00091. **f**) Depth of coverage of genes of interest with metagenomic reads in different samples with the taxonomy of genomes of origin indicated. **d** & **f**) Samples are sorted by protein abundances or transcript abundances, as in **b**. **a**, **c** & **e**) The colour/symbol-scheme denoting the individuals is identical to that used in the other figures, see **Figure 1a**. Boxes span the first to third quartiles, the central thick bars represent the medians, and whiskers extend to 1.5 times the interquartile ranges; all data points are represented, see **Supplementary Table 1A** for the total number of samples per individual. **b**, **d** & **f**) The genomic context of each gene is displayed for genes making up at least 10 % of the total transcript abundance – remaining genes are gathered in “others”.

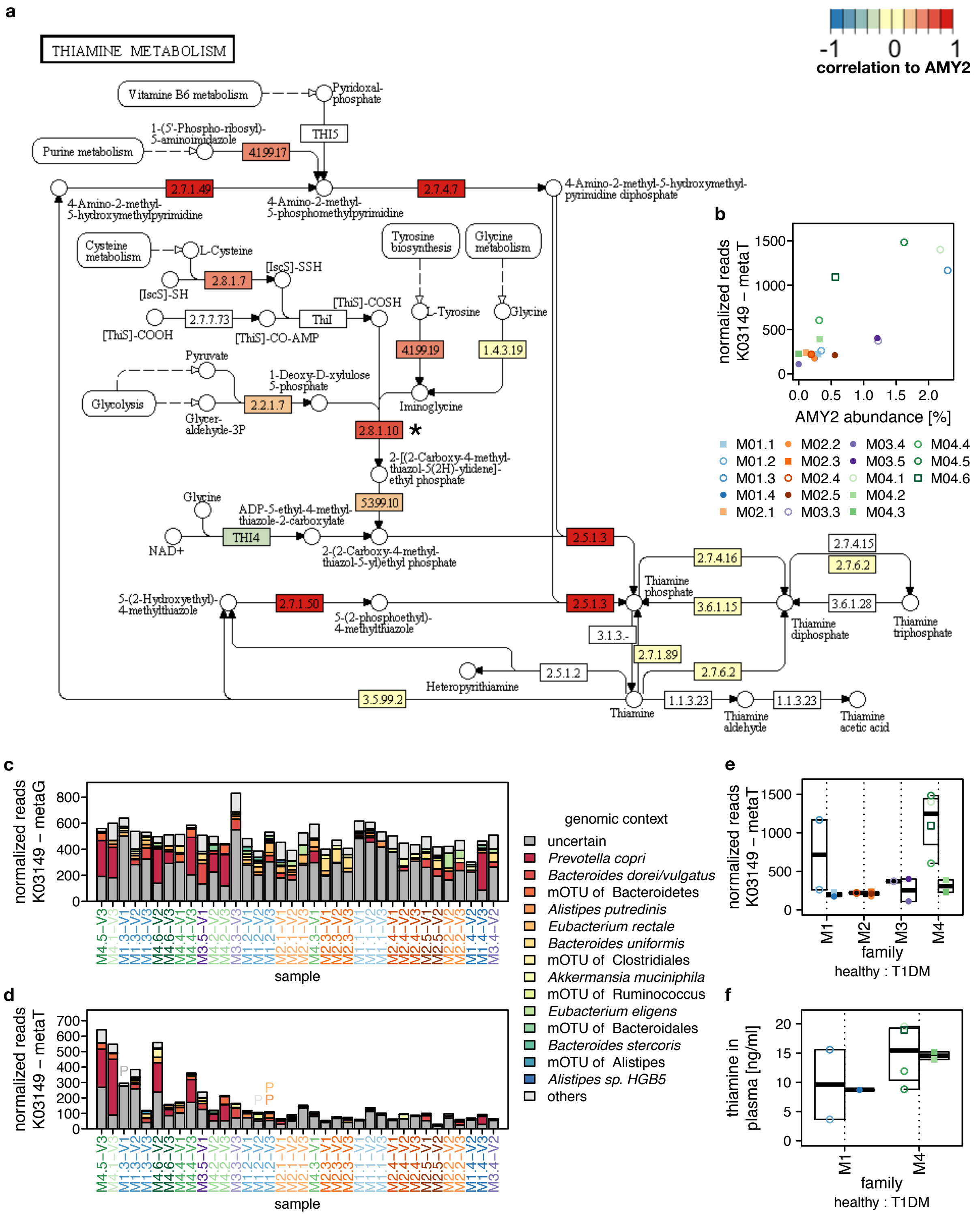

Figure S20: Correlation of AMY2 proteins in stool with transcript abundances of genes involved in thiamine metabolism. a) The KEGG pathway ko00730 - “thiamine metabolism” is displayed and enzymes are coloured according to their correlation

(Spearman's ρ, see colour key at top-right corner of the plot) between their transcripts and the relative abundances of AMY2A and AMY2B. Thiazole synthase (ThiG, K03149) is the central enzyme with EC 2.8.1.10. The graphics were rendered using the *pathview*[7] package in R. **b**) Relative transcript abundances of *thiG* (K03149) plotted against AMY2 protein abundances. **c**) Depths of coverage of *thiG* genes with metagenomic reads in different samples. **d**) Depths of coverage *thiG* genes with metatranscriptomic reads in different samples. 'P's indicate samples with uniquely identified ThiG proteins. **c & d**) The genomic context of each gene is represented by colours (see legend on the right hand side) for genes making up at least 10 % of the total transcript abundance - others are gathered in "others"; likewise, the colours of the letters P indicate the genomic context. **e**) Relative transcript abundances of *thiG* grouped by families and T1DM status. **f**) Thiamine levels measured in plasma of individuals from families with at least one healthy member and one member with T1DM who did not take thiamine supplements; only individuals who did not take vitamin supplements are shown. **e & f**) Boxes span the first to third quartiles, the central thick bars represent the medians, and whiskers extend to 1.5 times the interquartile ranges; all data points are represented. **b**, **e & f**) the same colours and symbols are used to represent individuals as throughout (see **Figure 1a**), see **Supplementary Table 1A** for the total number of samples per individual.

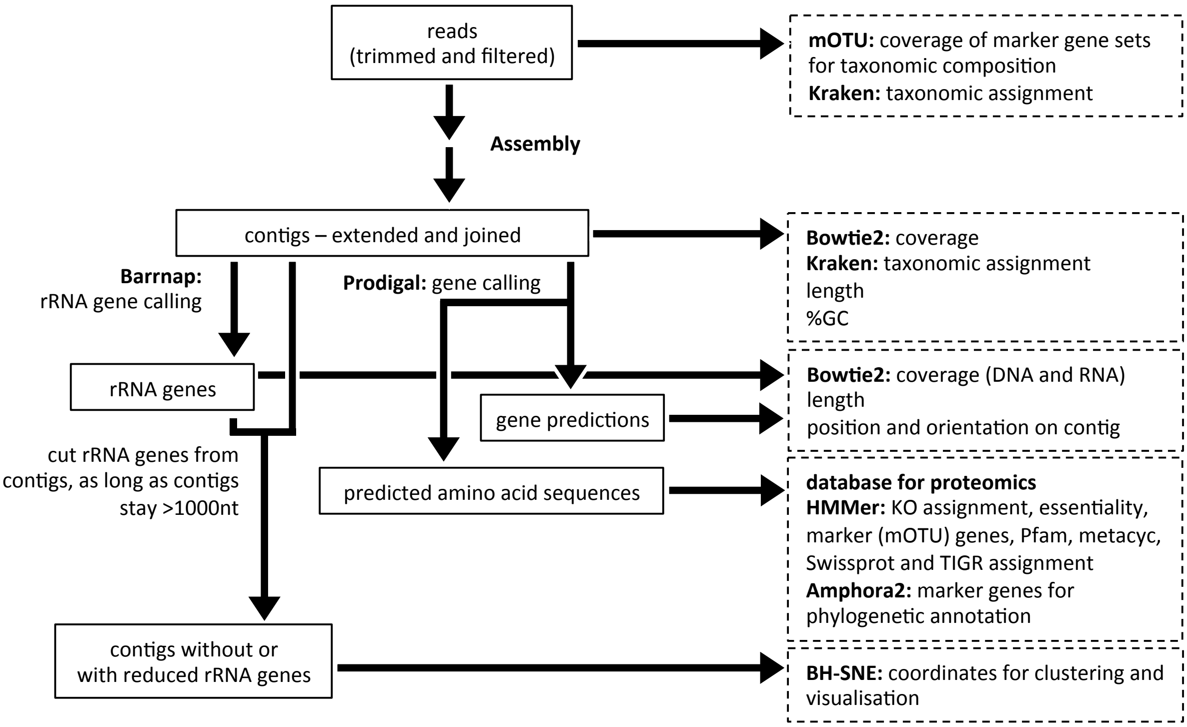
reads
(trimmed and filtered)
mOTU: coverage of marker gene sets for taxonomic composition
Kraken: taxonomic assignment
Assembly
contigs – extended and joined
Bowtie2: coverage
Kraken: taxonomic assignment
length
%GC
Barrnap: rRNA gene calling
Prodigal: gene calling
rRNA genes
gene predictions
Bowtie2: coverage (DNA and RNA)
length
position and orientation on contig
cut rRNA genes from contigs, as long as contigs stay >1000nt
predicted amino acid sequences
database for proteomics
HMMer: KO assignment, essentiality, marker (mOTU) genes, Pfam, metacyc, Swissprot and TIGR assignment
Amphora2: marker genes for phylogenetic annotation
contigs without or with reduced rRNA genes
BH-SNE: coordinates for clustering and visualisation

Figure S21: Workflow for taxonomic and functional annotation of different omic datasets. Different sequence datasets are indicated in boxes with solid lines, while annotation tools are listed in boxes with dashed lines on the right.

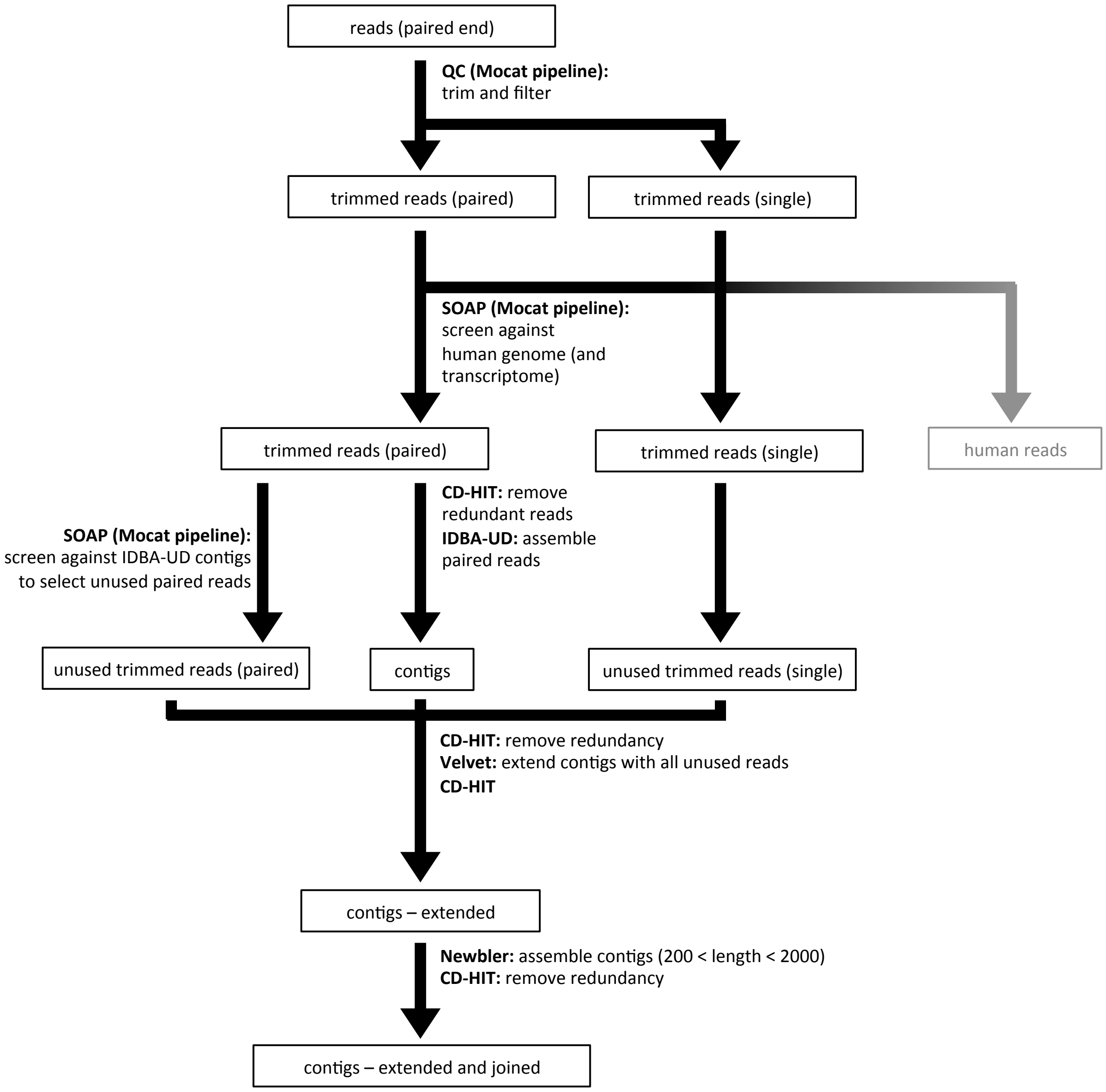
reads (paired end)
QC (Mocat pipeline): trim and filter
trimmed reads (paired)
trimmed reads (single)
SOAP (Mocat pipeline): screen against human genome (and transcriptome)
trimmed reads (paired)
trimmed reads (single)
human reads
SOAP (Mocat pipeline): screen against IDBA-UD contigs to select unused paired reads
CD-HIT: remove redundant reads
IDBA-UD: assemble paired reads
unused trimmed reads (paired)
contigs
unused trimmed reads (single)
CD-HIT: remove redundancy
Velvet: extend contigs with all unused reads
CD-HIT
contigs – extended
Newbler: assemble contigs (200 < length < 2000)
CD-HIT: remove redundancy
contigs – extended and joined

Figure S22: Co-assembly workflow for the integration of metagenomic and metatranscriptomic data. Contigs were co-assembled from metagenomic and metatranscriptomic reads. Reads mapping to the human genome or transcriptome (grey) were removed prior to assembly.

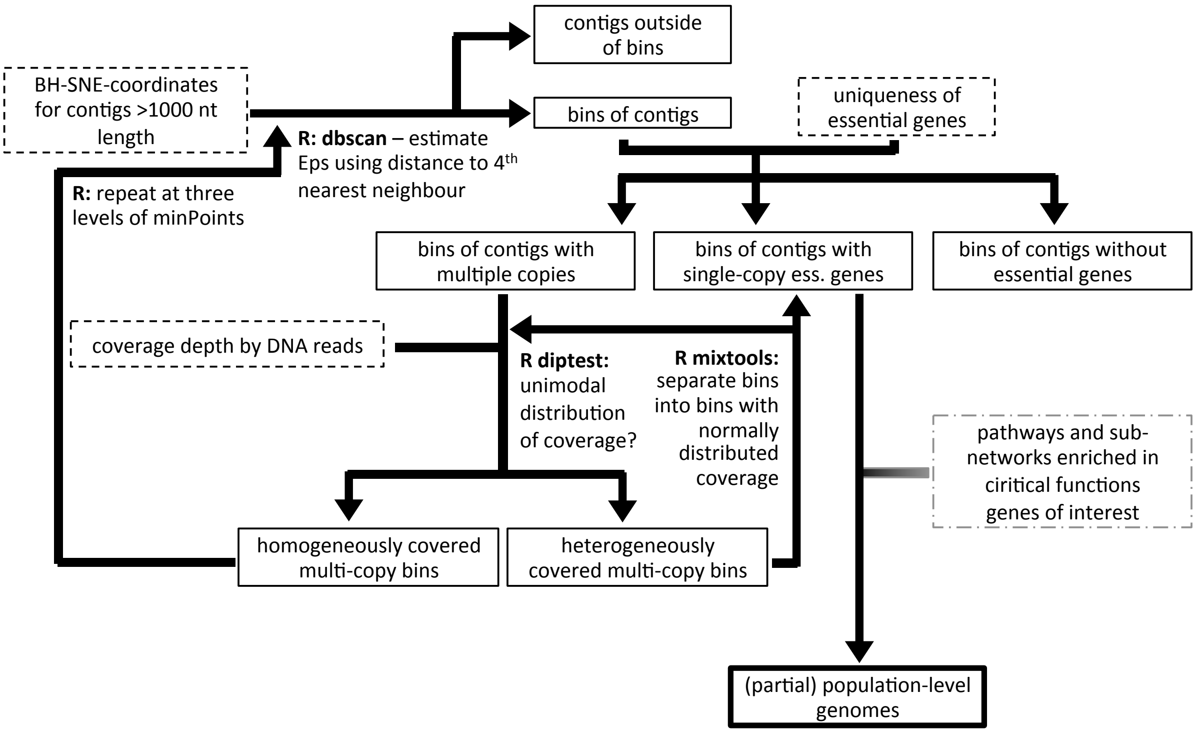
contigs outside of bins
BH-SNE-coordinates for contigs >1000 nt length
bins of contigs
uniqueness of essential genes
R: dbscan – estimate Eps using distance to 4th nearest neighbour
R: repeat at three levels of minPoints
bins of contigs with multiple copies
bins of contigs with single-copy ess. genes
bins of contigs without essential genes
coverage depth by DNA reads
R diptest: unimodal distribution of coverage?
R mixtools: separate bins into bins with normally distributed coverage
pathways and sub-networks enriched in ciritical functions genes of interest
homogeneously covered multi-copy bins
heterogeneously covered multi-copy bins
(partial) population-level genomes

Figure S23: Algorithm for automatic binning of contigs based on coordinates from non-linear dimension reduction of pentamer-frequencies, presence of essential single-copy genes and metagenomic depths of coverage. The different sets of bins formed during the process are represented by boxes with solid lines, with the final population-level genome reconstructions highlighted with a thicker line. Information used by the algorithm is indicated in boxes with dashed lines. Criteria for selecting genomes of interest are highlighted in the box with the dashed/pointed outline.

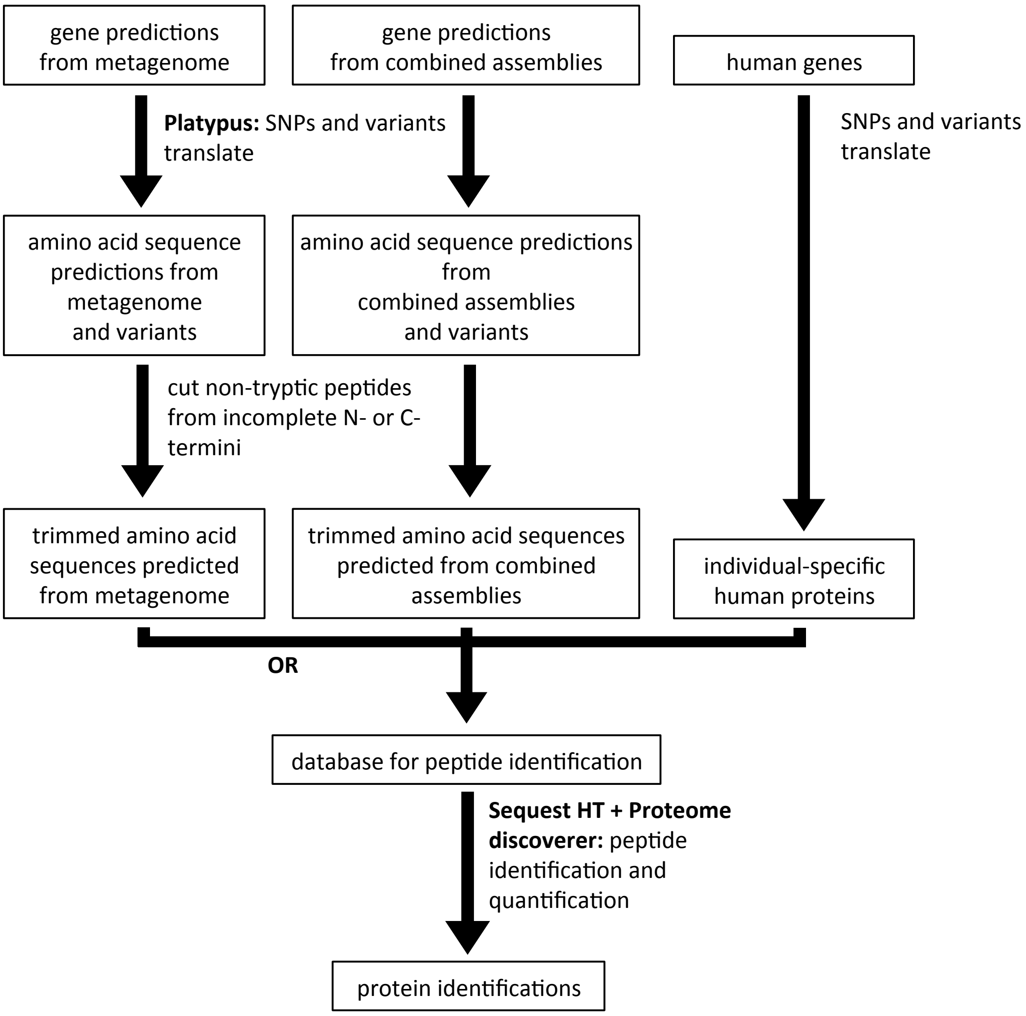
gene predictions from metagenome
gene predictions from combined assemblies
human genes
Platypus: SNPs and variants translate
SNPs and variants translate
amino acid sequence predictions from metagenome and variants
amino acid sequence predictions from combined assemblies and variants
cut non-tryptic peptides from incomplete N- or C-termini
trimmed amino acid sequences predicted from metagenome
trimmed amino acid sequences predicted from combined assemblies
individual-specific human proteins
OR
database for peptide identification
Sequest HT + Proteome discoverer: peptide identification and quantification
protein identifications

Figure S24: Workflow for the construction of search databases for metaproteomic analyses.

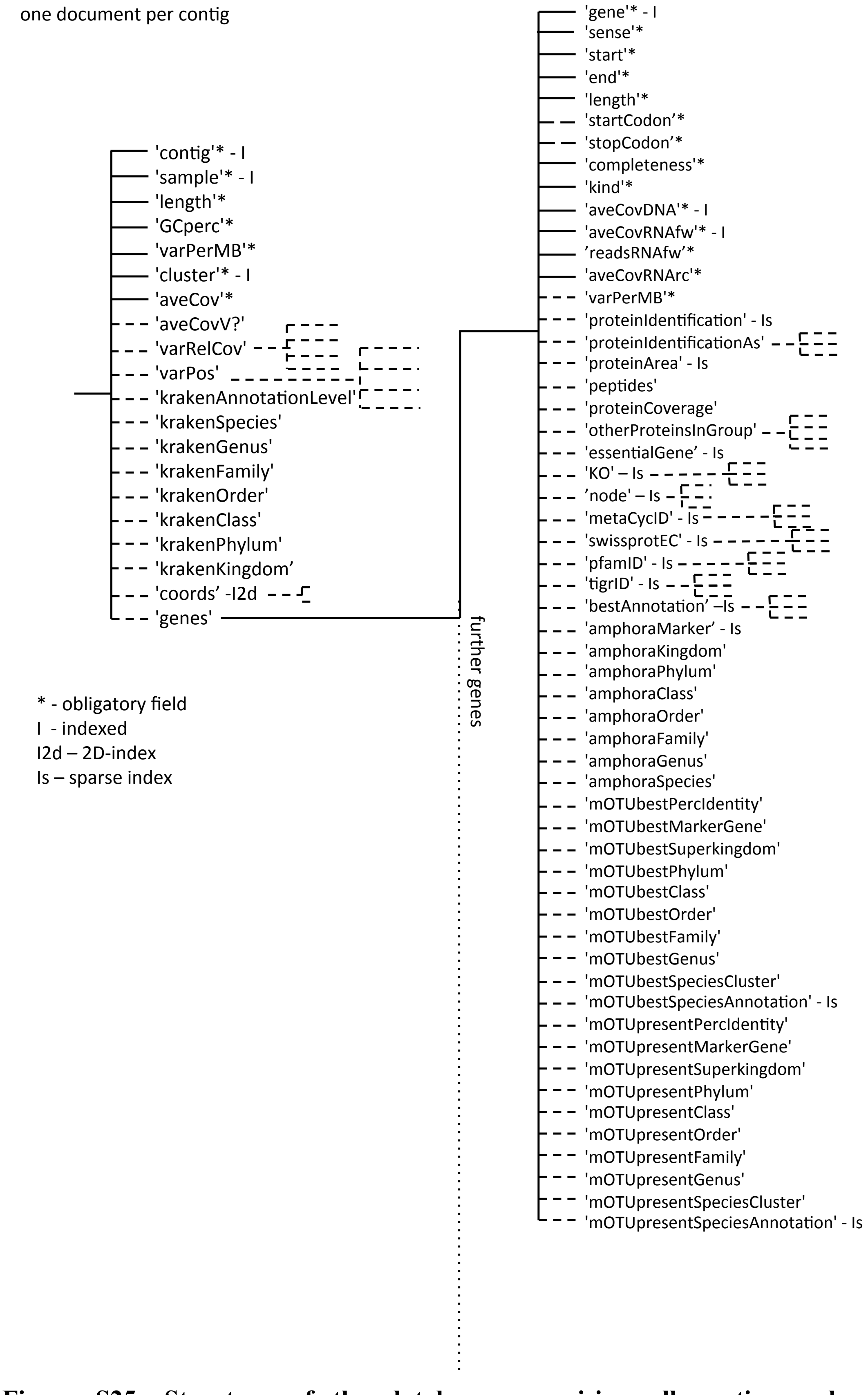

Figure S25: Structure of the database comprising all contigs and genes with

annotations, depths of coverage and sequence characteristics. Obligatory fields are marked with an asterisk and a solid line, while facultative fields are connected by a dashed line.

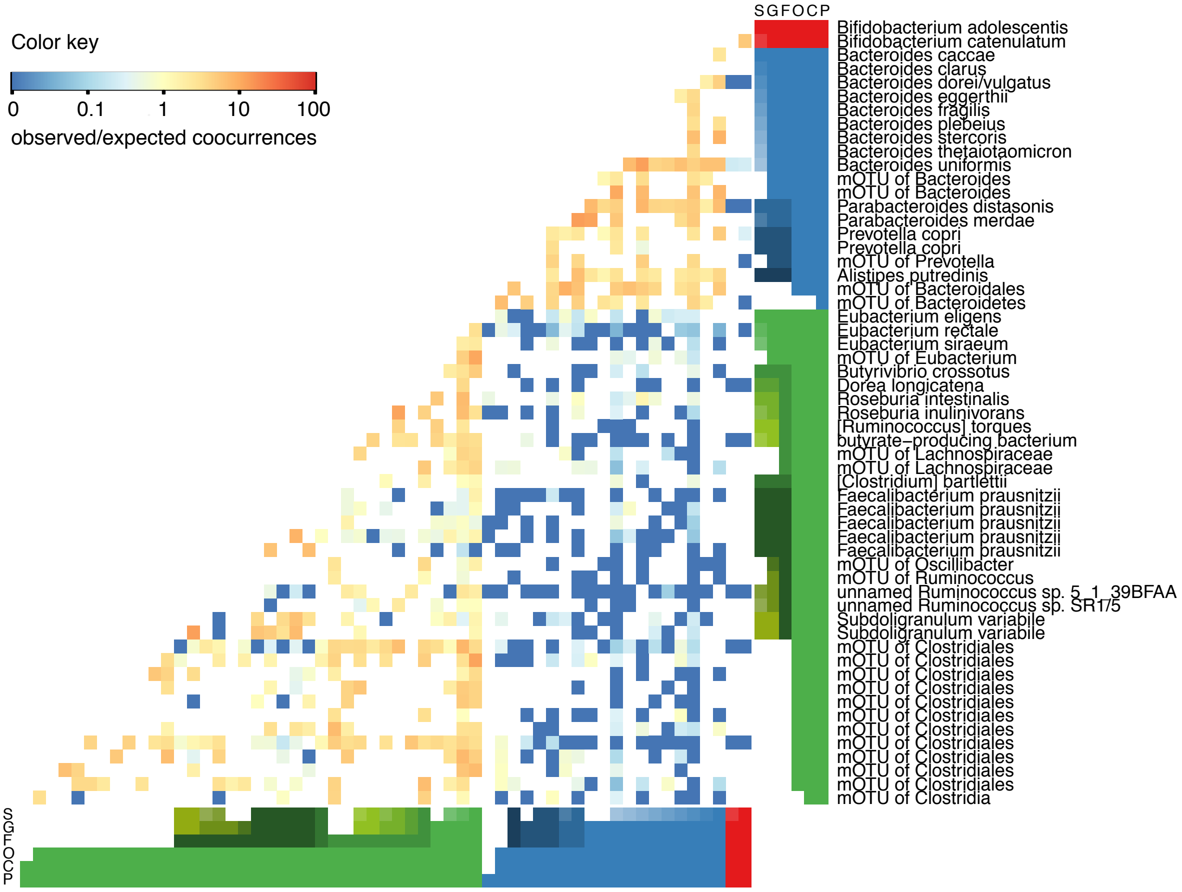

Figure S26: Taxonomic origins of proteins represented by protein groups. The mean ratio of observed to expected co-occurrences of protein predictions with the displayed taxonomic origins in protein groups over all samples (n = 36) is displayed (see colour key in top-left corner). White fields indicate that insufficient occurrences were present to judge co-occurrence. The coloured sidebars represent taxonomic ranks - P: phylum (red - actinobacteria; blue - Bacteroidetes; green - Firmicutes), C: class, O: order, F: family, G: genus, S: species, with white boxes representing unclassified mOTUs. Colour shading is identical for mOTUs classified to the same taxon at the given taxonomic rank (red shades - Actinobacteria, blue shades - Bacteroidetes, green shades - Firmicutes).

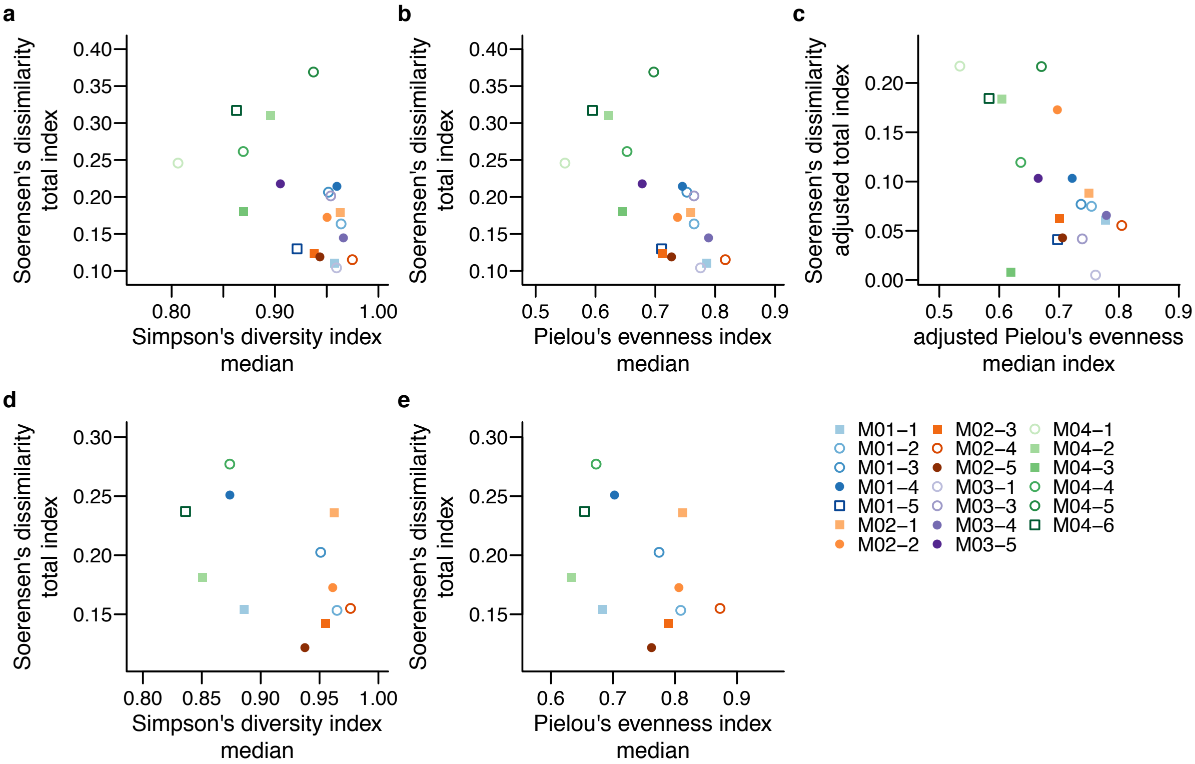

Figure S27: Diversity and temporal variability of the microbiome. Total Soerensen dissimilarity indices of the metagenomic taxonomic profiles for two or three samples (see **Supplementary Table 1A** for total number of samples per individual) are plotted against median **a**) Simpson's diversity and **b**) Pielou's evenness indices for metagenomic taxonomic profiles of the same samples. **c**) Pielou's evenness and Soerensen dissimilarity indices, adjusted for potential unseen metagenomic mOTUs using an ACE model[8]. Total Soerensen dissimilarity indices of the metatranscriptomic taxonomic profiles for two or three samples are plotted against median **a**) Simpson's diversity and **b**) Pielou's evenness indices for metatranscriptomic taxonomic profiles of the same samples. **a**, **b**, **c**, **d** & **e**) The same symbol and colours represent individuals as in the other figures, see legend in bottom-right corner of the plot and as defined in **Figure 1a**.

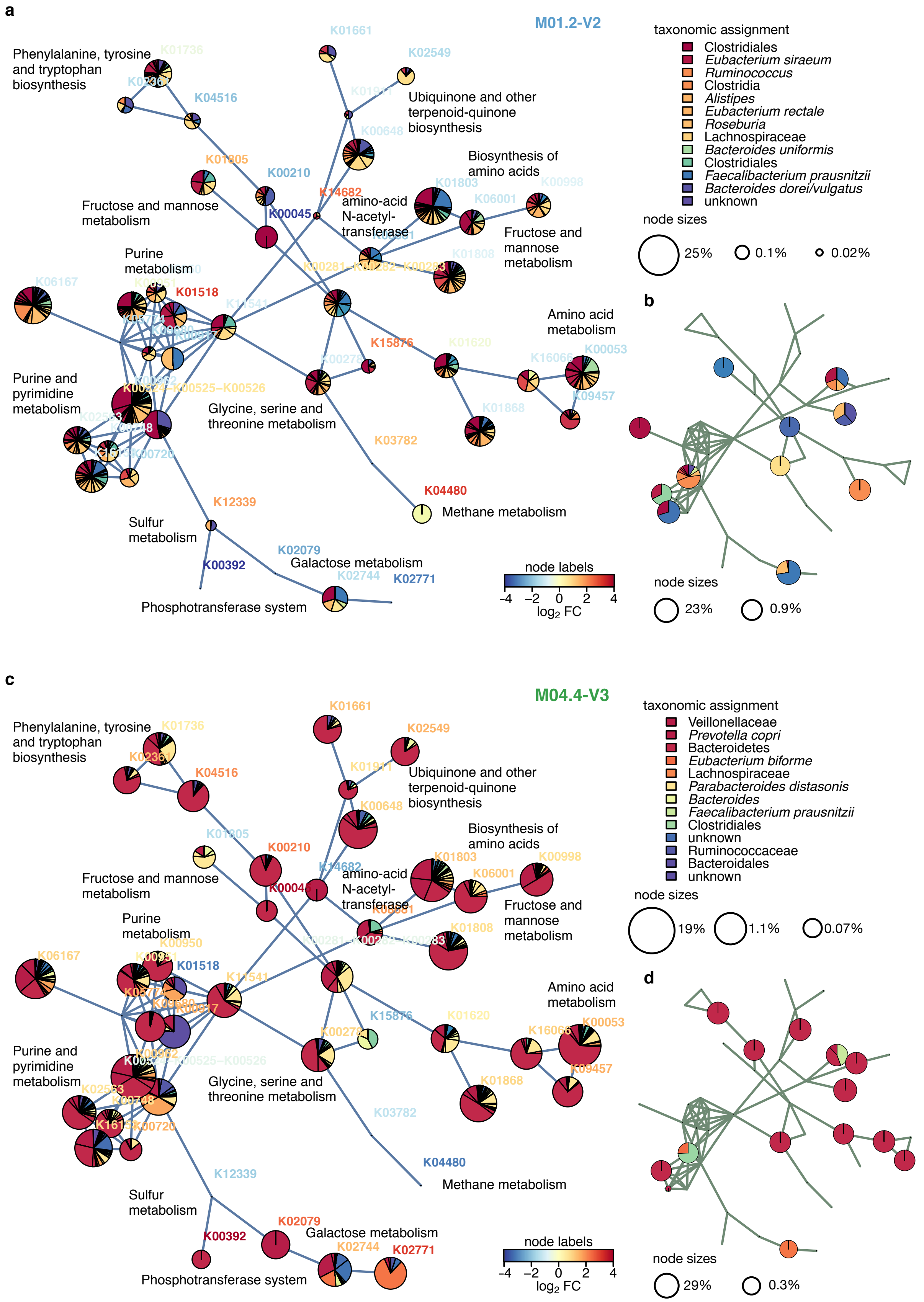

Figure S28: Functional differences between families. Visualizations of top-scoring

metabolic sub-network for metabolic KOs with a significant main effect of the sample being derived from family M04. Sizes of nodes in the big networks reflect metatranscriptomic abundances, while the corresponding smaller networks represent metaproteomic abundances. Slices of pies indicate transcript abundances derived from different population-level reconstructed genomes; colours indicate completeness of population-level genomes. Colour of node label indicates fold change between family M04 and the others. One representative sample each, **a**) M1.2-V2 and **b**) M4.4-V3, of 27 and 9 samples, respectively, are displayed. Note: the colours of the slices of pies in the upper and lower panel do not indicate the same taxa.

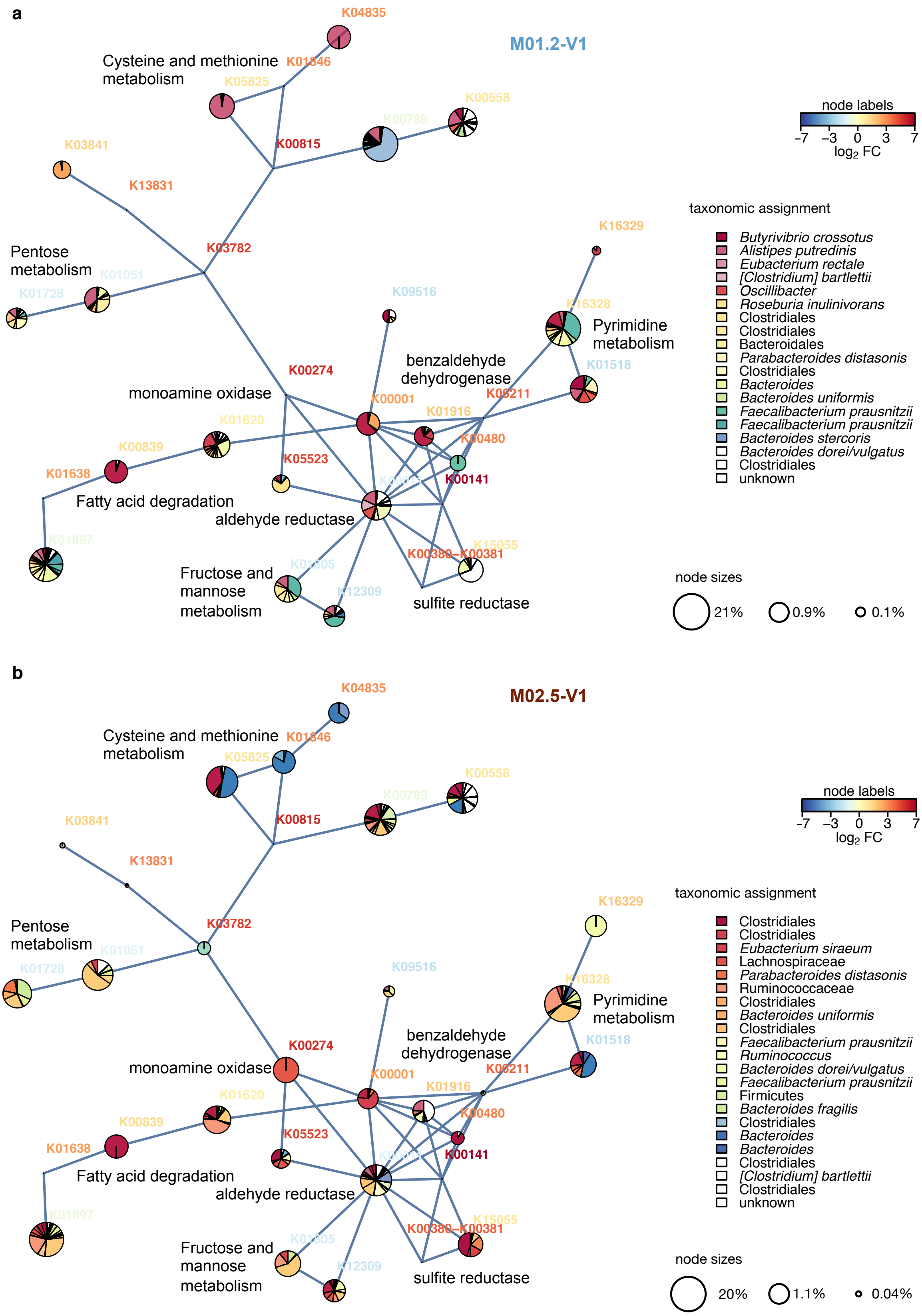

Figure S29: The top-scoring module of metabolic functions in a community-wide

metabolic network reconstruction, based on differential metatranscriptomic abundance between individuals with T1DM and healthy relatives. The module was identified at an FDR of 0.2 using BioNet[9]. The sizes of the nodes in the networks reflect metatranscriptomic abundances, while the small networks represent metaproteomic abundances. Slices of pies indicate transcript abundance derived from different bin-genomes; colours indicate completeness of bin-genomes. The colour-coding of the node labels represents fold change (red - higher in group of displayed sample; green - lower in group of displayed sample). One representative sample each, **a**) M1.2-V1 (healthy) and **b**) M2.3-V1 (T1DM), of 16 and 20 samples, respectively, are displayed. Note: the colours of the slices of pies in the upper and lower panel do not indicate the same taxa.

Legends for Supplementary Tables

Supplementary Table 1: Anthropometric, demographic, medical and nutritional information on the study cohort. **a**) Cohort overview with demographics, medical history and anthropometric data. **b**) Overview of sampling and inquiry dates. **c**) Information from food frequency questionnaires. **d**) Calculated intakes of nutrients and food classes. **e**) Explanation of foods in **c**). **f**) Medication taken by members of the cohort, including insulin doses. **g**) Auto-antibody status of the cohort. **h**) Other medically relevant blood values.

Supplementary Table 2: Taxonomic and functional profiles based on the metagenomic, metatranscriptomic and metaproteomic data. **a**) Per-sample relative mOTU[10] abundances from metagenomic data. **b**) Per-sample relative mOTU[10] abundances from metatranscriptomic data. **c**) Per-sample DESeq2[11]-normalized abundances of functional transcripts. **d**) Per-individual DESeq2[11]-normalized abundances of transcripts representing KOs in a metabolic network reconstruction. **e**) Per-sample relative abundances of functions in the metaproteomic dataset. **f)** Per-sample relative abundances of human proteins.

Supplementary Table 3: Statistics on metagenomic, metatranscriptomic and metaproteomic data, as well as human whole genome sequencing data. **a**) Sizes of metagenomic, metatranscriptomic and metaproteomic raw data sets, and statistics on assemblies, identified proteins and recovered population-level genomes (mean and standard deviation). **b**) Statistics on human whole genome sequencing. **c**) Statistics on identified proteins and protein groups.

Supplementary Table 4: Population-level genomes recovered from combined metagenomic and metatranscriptomic data by binning. Information on the genome length, number of contigs, read coverage, numbers of genes and taxonomy, along with the RAST[12] accession ID are given for the 200 population-level genomes with >93 % completeness.

Supplementary Table 5: Human stool proteins and microbial functional transcripts correlating to alpha-amylase levels. a) Results of differential analysis of human stool proteome. **b**) Results of correlation analysis of expression of metabolic functions on the metatranscriptomic level to human alpha-amylase levels.

Supplementary Table 6: Results of differential analyses of metatranscriptomic and metaproteomic functions.

Supplementary Table 7: List of sets of open reading frames predicted on assembled contigs with MG-RAST[13] accession IDs.

Supplementary Table 8: Results of comparison of correlation analysis with binning for tracing functions to taxonomic entities.

Supplementary Notes

Supplementary Note 1: Cohort description - case descriptions of T1DM, occurrence of other diseases and nutritional data of all individuals

The cohort consisted of 20 individuals from four families (**Figure 1a**, **Supplementary Table 1**). 10 individuals in the cohort were diagnosed with T1DM and were treated with insulin (see **Supplementary Table 1**). Metabolic control, as assessed by glycation status of haemoglobin, was in most patients within the target range (glycated haemoglobin < 7,5%) in at least one of the three samples, except in individual M03.5 (only one sample, 8,3 % glycated haemoglobin) and individual M04.2, who maintained glycated haemoglobin > 9,6%, despite intensified continuous subcutaneous insulin administration (see **Supplementary Table 1**).

C-peptide, a marker for endogenous insulin secretion, was negative in most patients with T1DM except for M01.1, M03.4 and M04.3. Insulin-directed auto-antibodies were still detectable in all but two patients (**Supplementary Table 1**). One was a patient with long standing diabetes (M02.5). The second case had more than five years diabetes, but still detectable C-peptide levels (M04.3). Of the individuals with detectable auto-antibodies, M01.1 and M01.4 had very high levels of the GAD2 and Zn8T auto-antibodies at several or all sampling dates. In one healthy individual (M04.1), the GAD2 auto-antibody level was significantly elevated at a single sampling time point (**Supplementary Table 1**). As first-degree relatives of individuals with T1DM with this auto-antibody have been found to have a 10-year risk of developing T1DM of 22 %[14], this indicates a relatively small increase in the risk of developing T1DM in this individual[15,16], whose risk is also increased due to an earlier episode of gestational diabetes[17,18] and having two sons with T1DM. All other healthy relatives did not exhibit significantly increased auto-antibody levels, indicating that their risk of still developing T1DM not further increased above the level assumed for first degree relatives of individuals with T1DM[19].

Screening for known genetic causes of different types of maturity onset diabetes of the young (MODY types 1, 2 and 3; HNF4A, GCK, and HNF1A, respectively) in individuals M04.2 and M04.3 was negative.

None of the individuals in the cohort had bowel complaints and/or coeliac disease. Auto-immune thyroiditis was diagnosed in four patients with T1DM (M01.3, M02.2, M02.3, M02.5), and one family member without T1DM (M04.4). These individuals were treated with thyroid replacement therapy (see **Supplementary Table 1**), and the individual without T1DM had temporarily increased liver enzyme values (M04.4-V1), due to non-adherence to thyroid replacement treatment. Epilepsy was diagnosed in one of the patients with T1DM (M04.2), necessitating a treatment with valproic acid. One individual (M01.2) presented high cholesterol values.

All participants in the study supplied records on food intake, estimating their usual diet in the last six months. For most sampling dates, records of the food intake during the last 24 hours were also provided. Nutrient contents of the diets were estimated using FETA software[20] (see **Supplementary Table 1**). Analysis of both datasets showed family-specific nutritional patterns (**Supplementary Figure S12 a&b**, see also **Supplementary Note 16**). For example, family M04 in particular followed a different diet from the other families, comprising more fruit and fish and less white bread. Consequently, the estimated nutrient intake over the last six months indicated that the members of this family ingested more vitamins and had a different profile in terms of consumed carbohydrates, with higher levels of fibre and fructose, while the diets of families M02 and M03 contained more maltose, galactose and lactose. When analysing the dietary records of the 24 hours prior to faecal sampling, similarities and differences to the long-term estimates were obvious. For example, family M04 had ingested more fruit than other families, but on the other hand, on those specific dates, they did consume similar amounts of potatoes as families M01 and M02. Despite the fact that M04 usually ate more oily fish on the days where diet was recorded, this family's diet consisted of comparably little fat.

Individuals with T1DM did not consume a different diet than their healthy family members, in accordance with medical recommendations. None of the differences in the diet of individuals with T1DM were significant after multiple-testing adjustment (**Supplementary Figure S12 c&d**). The most significant differences in diet were likely independent of health status, in that individuals with T1DM consumed more beer and veal in the last 6 months, and more white bread in the last 24 hours, while more grapes and poultry were eaten by the healthy individuals prior to the sampling (unadjusted *P* values of Wilcoxon rank sum test < 0.02); see **Supplementary Table 1**). Consequently, alcohol and alcoholic beverages were among the nutrients and food classes with the highest differences between individuals with T1DM and their healthy relatives (unadjusted *P* values of Wilcoxon rank sum test < 0.02), while higher levels of phosphorous and pyridoxine were calculated for the diets of the healthy individuals based on the 6 months dietary data (unadjusted *P* values of Wilcoxon rank sum test < 0.02). In addition, "nuts and seeds" as a food class were found to have been eaten more often by individuals with T1DM 24 hours before sampling. Although all differences were relatively small, the records on nutrition were compared to the results of the differential analyses of the microbial data to detect potentially confounded results (see **Supplementary Notes 16-18** for further details).

Supplementary Note 2: Gastrointestinal microbial community structures in the cohort

To obtain an overview of the microbial community structures in the faecal samples, we used metagenomic reads from 53 samples of the 20 study participants to calculate relative abundances of previously published[10] metagenomic operational taxonomic units (mOTUs). As expected, microbial communities were dominated by bacteria, especially of the phyla Bacteroidetes and Firmicutes (**Supplementary Figure S1 a** and **Supplementary Table 2**). Although faecal microbiota are among the best studied human-associated microbiota, the proportion of unknown organisms (which have not yet been isolated and sequenced) is still high[10]. To assess this proportion in the present dataset, we calculated the sum of the relative abundances of mOTU linkage groups, which are mOTUs that are best described by phylogenetic marker genes not found in sequenced isolate genomes but consistently found in metagenomic sequences (referred to as $\mathrm{mOTU}_{\mathrm{Meta}}$ in the original publication[10]). 13 % of the mOTUs defined the original publication[10], and 43 % of the mOTUs detected in this study, have no closely related sequenced isolates. The combined relative abundances of such mOTUs in the analysed communities were 36 +/- 12 %, with the most abundant mOTUs without closely related sequenced isolates accounting for 7.8 +/- 4.4 % of the detected prokaryotes in each sample. Furthermore, 27 +/- 12 % of the communities were made up of organisms unclassified at the genus level and 3 +/- 3 % were not even annotated at the phylum level. In particular, single mOTUs not classified at the phylum level reached up to 9 % relative abundance in individual samples. These values illustrate that large fractions of the studied microbiota consist of microorganisms that are hardly described and understood. Therefore, analytical approaches purely based on isolate reference genomes would be relatively limited. The elucidation of the functional potential and associated activities related to of distinct microbiota should therefore rely on reference-independent *de novo* analytical approaches.

Ordination of samples based on Jensen-Shannon divergences of abundances of mOTUs (**Supplementary Figure S2 a**) revealed that in most cases, intra-individual variability of community structures was lower than inter-individual variability (see **Supplementary Note 12**) and samples from some of the families formed distinct groups (see **Supplementary Note 13**). The samples were part of a continuum with one extreme group of communities dominated by *Prevotella* spp. while in another group *Bacteroides* spp. were most abundant. This pattern is often observed in faecal microbial communities and has been conceptualized as enterotypes[21]. To assign enterotype-like groups to the present samples, genus-level abundance profiles of 1,267 publicly available metagenomic datasets[1] and genus-level abundance profiles from the present metagenomic dataset mapped to the same reference gene catalogue[1] were classified together, as described in the publication of the reference gene catalogue[1]. The results are indicated in the mOTU-based abundance profiles in **Supplementary Figure S1 b**. In two of four families (M02 and M03) all individuals from a family were part of the same enterotype-like group. In families M01 and M04, samples from both ends of the enterotype-spectrum were observed. In addition, in four of the twenty individuals of the cohort a shift from one group to the other within 2 to 3 months was observed (M01.1, M01.4, and most obviously M04.4 and M04.5; see **Supplementary Figures S1 b** and **S2 a**). The relative abundance of *Prevotella* spp. fluctuated strongly in these individuals. Intriguingly, the dominance of *Prevotella* spp. in samples from several co-habitating individuals in family M04 peaked at one sampling point, indicating that this change may have been caused by a common environmental factor, in particular diet.

However, careful analysis of the nutritional and medical information given by the members of this family could not explain the cause for this change.

No significant correlation between the abundances of any mOTUs and estimated dietary intake of macro- or micronutrients, or broader of classes of food, e.g. “fruit”, “fish”, “eggs and egg dishes” (see **Supplementary Table 1**) during the last 24 hours were observed. To account for the importance of specific nutrients within the diets of the individuals instead of the total intake, nutrients and food types were also analysed after normalisation to calorific intake. No significant correlations were apparent from this analysis either. Therefore, no strong influence of estimated dietary intake levels on the gut microbial community structure was detected here. Detection of such influences may require more rigorous and frequent documentation of nutrition and more frequent analysis of microbial community structure[22] or controlled environment[23].

Differences in the ratio of Firmicutes to Bacteroidetes were among the first community traits linked to host phenotype, i.e. obesity, by 16S rRNA amplicon-sequencing mouse studies and were even observed in small human cohorts[24], although contradictory observations have also been reported[25]. We were therefore interested whether this ratio varied with BMI in the present cohort. While the BMIs of adult individuals in the present cohort ranged from 18 to 38, the Firmicutes-to-Bacteroidetes ratios were not significantly lower in the lean individuals (Wilcoxon rank sum test *P* value = 0.4; **Supplementary Figure S14**) and were not significantly correlated to BMI (Spearman’s $\rho = -0.23$, *P* value = 0.44).

Supplementary Note 3: Details on assembly statistics

In order to integrate metagenomic and metatranscriptomic data and to maximize the usage of the sequencing data in general, we set up an assembly pipeline which co-assembles metagenomic and metatranscriptomic read data. To assess our strategy in comparison to single-ome assemblies and other combined metagenomic and metatranscriptomic assembly strategies, we evaluated numbers of reads mapping back to the assemblies, total lengths of assembled contigs, numbers of contigs greater 1 kbp, which are needed for the binning approach used, length distributions of contigs and predicted open reading frames (ORFs), as well as the numbers and diversity of functional annotations.

When comparing the number of reads mapping back to the assemblies (see **Supplementary Figure S3 a**), the pure metagenomic assembly recruited more metagenomic reads than the metatranscriptomic assembly or any of the co-assemblies. On the other hand, the metatranscriptome was not so well represented by the metagenomic assembly and, more importantly, co-assembling both types of reads achieved an even greater recovery of metatranscriptomic reads than the purely metatranscriptomic assembly. The chosen co-assembly workflow also used the maximum combined number of metagenomic and metatranscriptomic reads.

Different assemblers for metagenomic data have complementary strengths[4,26,27] and combining several assemblers is one approach to optimizing the usage of currently available assembles[28]. In this study, the combination of multiple assemblers (IDBA-UD[2], Velvet[3] and Newbler from MeGAMerge[4]) led to longer total assembly lengths, more contigs greater 1 kbp and more predicted ORFs and predicted complete ORFs (see **Supplementary Figure S3 b - f**). These ORFs could also be annotated with a greater diversity of functions (see **Supplementary Figure S3 d**).

Per sample, an average of 287 +/- 43 Mbp of assembled sequence, comprising 730,000 +/- 140,000 contigs was finally obtained by co-assembly of metagenomic and metatranscriptomic reads using multiple assemblers (**Supplementary Table 3**). The contigs had a minimal length of 125 bp, an N50 of 660 +/- 130 bp, and the longest contig in each assembly reached 310 +/- 170 kbp. On average, 88 +/- 4 % of the trimmed and filtered metagenomic reads and 88 +/- 2 % of the trimmed and filtered metatranscriptomic reads could be mapped back to the assembled contigs, which compares very favourable to previously reported mapping rates, i.e. 74%-81% of human metagenomic data is typically represented within an integrated gene catalogue[1].

The reconstruction of long contigs usually depends on sequencing depth by which longer fragments are typically reconstructed from organisms in higher abundance. We found a dual relationship between taxonomic richness (determined based on reads, see above) and assembled sequences: the apparent quality of assemblies suffered from high taxonomic richness, as the N50 and average contig length were lower in samples with high mOTU richnesses (**Supplementary Figure S4 a**), likely reflecting insufficient coverage for assembly of mOTUs with lower abundance. Conversely, total assembly lengths were higher in samples with higher mOTU richnesses (**Supplementary Figure S4 b**). Overall, the numbers of complete genes called were also higher in samples with higher microbial richnesses. The numbers of different functional categories, i.e. KEGG orthologous groups (KOs), Pfam and TIGRfam families, MetaCyc and Swiss-Prot enzymes, assigned to the genes in one assembly ranged between 12,548 and 18,053. In accordance with the higher number of genes, the apparent functional richness was also positively correlated to assembly length and taxonomic richness (Spearman's ρ = 0.57 and 0.40, respectively; **Supplementary Figure**

S4 c). This indicates that despite the challenging taxonomic diversity encountered in some samples, the functional diversity could be successfully recovered in most cases.

Supplementary Note 4: Identified protein groups and proteins

920 +/- 570 microbial proteins were uniquely identified from the peptide mass spectra in each sample while a further 6,600 +/- 3,700 proteins were putatively represented by the 1,400 +/- 1000 protein groups comprising several members that were indistinguishable from each other based on the identified peptides (see **Supplementary Table 3**). Unsurprisingly, overall fewer different microbial functions were detected in the metaproteome than in the metatranscriptome (4.9 +/- 1.8 % of the functional categories detected in the metatranscriptome, see also **Figure 3f**). In addition, the number of different proteins representing these functional categories reflected only a small proportion of the detected transcripts for proteins of these functions. Nonetheless, the proteomic data gives valuable insight into abundant functions.

To achieve an overview of the identified proteins, common functional categories were determined. The 79 functional categories identified in all samples included ribosomal proteins, outer membrane proteins, chaperones, elongation factor Tu, flagellin and proteins involved in carbohydrate uptake and metabolism. These functional categories were also generally more abundantly represented on the metatranscriptomic level than functions that were not represented by identified proteins in any or some samples (**Figure 3e**; **Supplementary Table 2** and **3**).

Because protein-coding taxonomic marker genes were used throughout this study for the analysis of community structure and identification of taxa (see **Supplementary Notes 2**, **3**, **6**, **7**, **10**, **12**, **13**, **14**, **15, 16 and 18**), we analysed how well these genes were represented in the metaproteome and whether community structure could be inferred from their presence. The analysis of the presence and abundance of metagenomic operational taxonomic units (mOTUs) based on mapping metagenomic reads to the marker genes had revealed the presence of 150 +/- 30 mOTUs in each sample (see **Supplementary Table 2**), out of which on average 42 +/- 14 were represented by assembled marker genes. Representation of the marker genes in the metaproteome (3 +/- 1 %) was higher than the representation of other genes (~0.1 %, see **Supplementary Table 3**), likely due to the fact that being ribosomal proteins, these proteins were expressed constitutively and at relatively high levels. Nevertheless, only one of each kind of marker proteins was on average uniquely identified per sample, representing only a single mOTU. Therefore, the depth of the metaproteomic measurement was not sufficient for the accurate determination of the taxonomic composition based on the metaproteome. Consequently, taxonomic analyses based on the protein-coding taxonomic marker genes were limited to the metagenomic and metatranscriptomic data in the remainder of the study and metaproteomic data were analysed in terms of expressed functions.

On average, 650 +/- 230 different microbial functions were represented in each metaproteomic dataset, either by proteins that were uniquely identified or by protein groups whose members were all annotated with the same functional category (see **Supplementary Table 3**). To understand the origin of proteins within protein groups comprising a single functional category, the members of such protein groups were further analysed with respect to the taxonomic annotation of their genomic context (see **Supplementary Notes 6** and **7**). The co-occurrence of taxa was evaluated using a hypergeometric model as implemented in the *cooccur* package in R (function *cooccur*). Taxa co-occurring in functionally homogenous protein groups significantly more often than expected from a random distribution were usually closely related, such as several strains or species of *Prevotella* or *Bacteroides*, or *Bacteroides* and *Parabacteroides* (see **Supplementary Figure S26**). Similarly, taxa co-occurring in such protein groups significantly less often than expected from a random distribution were usually from different phyla, such as *Prevotella* spp. and *Ruminococcus* spp. (**Supplementary Figure S26**). The origin of protein groups therefore reflects conservation of proteins between related organisms.

The analysis of human genomic data was integrated within the analysis of the metaproteomic data to increase the number of peptide spectral database matches and avoid false positive matches of human protein-derived peptides to microbial protein predictions. Therefore, we generated personalized human proteome databases by extending the human genome reference hg19 protein database (RefSeq) by adding personalized human protein sequences including homozygous and heterozygous SNVs and small indels from human whole genome sequence data for each individual. 90 +/- 30 different human proteins could be uniquely identified in every sample, of which 11 were found in all samples (**Supplementary Table 2**). 11+/-3 individual-specific proteins were found. For every sample, we could detect on average one protein with two haplotype forms representing heterozygous variants (**Supplementary Table 3)**. Among the most abundant and frequently identified proteins were digestive enzymes, as well as immunoglobulin J (IgJ).

As the human genome contains protein families of paralogous proteins, which may not be uniquely identifiable due to high similarity, we searched for protein groups consisting of multiple human proteins. 24 +/- 8 protein groups with multiple human proteins were found in each sample. For 15 +/- 6 of the multi-locus protein groups, the genes were part of proteins families as annotated by the HUGO Gene Nomenclature Committee[29]. In most cases (90 +/- 7 % of the groups) all proteins within a protein group were members of one protein family, confirming our suspicion about paralogous protein groups. Therefore, we also analysed the abundance of protein families in addition to single proteins.

We were further interested in the origin of non-uniquely identifiable proteins in protein groups. In all datasets together, there were 353 protein groups identified which contained proteins encoded by both the human genome sequence and the microbial data. While 14,445 proteins were part of protein groups entirely made up of human proteins, and 238,574 proteins from the assemblies were part of entirely non-human protein groups, only 12 of the proteins in the ambiguous protein groups were most similar to human or other metazoan gene sequences. These results indicate efficient removal of human sequences prior to assembly and the ability to unambiguously differentiate molecules of human and microbial origin. Proteins in protein groups made up of proteins from both the host and the microbiota often were highly conserved proteins, such as ribosomal proteins, chaperones and ATP synthase units.

Supplementary Note 5: Binning results

To link genes of interest to their genomic contexts, we devised an automated binning approach for contigs based on nucleotide signatures, presence of single-copy essential genes and metagenomic depth of coverage (see **Supplementary Figure S23** and **Methods**). In every sample, 6 ± 2 nearly complete and homogenous population-level genomes were recovered which contained more than 93 % of the single-copy essential genes and less than 10 % of these genes in multiple copies (see **Supplementary Table 4**). A further 9 ± 3 binned genomes per assembly were at least 67 % complete with less than 20 % multiple essential gene copies. 370 ± 70 less complete bins of genome fragments were observed in each assembly (including 200 ± 40 bins without any essential bacterial genes). 70 ± 13 of the bins contained at least one phylogenetic marker gene also used for the analysis of mOTUs, while only 12 ± 4 population genomes contained at least partial 16S rRNA genes, which are otherwise frequently used for phylogenetic analyses. We therefore based phylogenetic analyses of the reconstructed genomes on these protein-coding marker genes.

Among the contigs that were not binned because the binning approach, which is based on DBSCAN-based[30], classified them as noise, viral sequences were strongly enriched (hypergeometric test *P* value 5 *x* 10^-143). Bacteriophage genes were additionally sometimes found in binned population-level genomes, and they were enriched in bins without essential genes (hypergeometric test over all genes; *P* value < 10^{-49}) and in nearly complete and homogenous reconstructed genomes (hypergeometric test over all genes; *P* value < 10^{-5}). This reflects that bacteriophage genomes could be well recovered but were not usually assembled with the bacterial host genomes, except in few cases where their bacterial host genome was extremely well recovered.

Supplementary Note 6: Consistency between assembly-independent and -dependent approaches

As discussed above (see **Supplementary Note 3**), metagenomic coverage was important for the assembly of long contigs. However, other factors contributed greatly to the successful binning, as there were several sequences forming incomplete reconstructions of genomes although their metagenomic coverage was high and *vice versa* (**Supplementary Figure S7 a)**. Factors facilitating contig binning can be the absence of closely related organisms, unusual and very stable genomic signatures, absence of repetitive and mobile genomic elements and high and even metatranscriptomic coverage.

To evaluate whether assembly-independent taxonomic data was nonetheless congruent with data derived from the binned contigs, metagenomic coverage of marker genes from the assemblies was assessed. While lowly abundant mOTUs were missed in the assembly-based analysis of the community structures, the overall profiles were very similar (see representative sample in **Supplementary Figure S7 b**). For 54 +/- 12 bins in every sample, a unanimous taxonomic annotation at the mOTU level could be achieved for all marker genes (62 +/- 8 % of the bins with more than one marker gene; **Supplementary Figure S7 c**). When the metagenomic depths of coverage of the annotated binned genomes were compared to assembly-independent data, similar abundance profiles were observed (see representative sample M01.2-V1 in **Figure 2 d**). Accordingly, ordination of mOTU-wise aggregation of the metagenomic depths of coverage of the unanimously annotated binned population-level genomes yielded a very similar picture as the assembly- and binning-independent mOTU analysis (**Supplementary Figure S2 a & c**).

Supplementary Note 7: Comparison of correlation- and binning-based approaches to relate function with taxonomy

To link genes with functions of interest to the genomic context of the microbes expressing them, we used an approach of linking genes to contigs and then associating these contigs with their wider genomic context by binning. To determine if this approach could reveal more informative links than simpler methods, we compared it to the previously suggested approach of correlation[31]. Using the latter approach, we were able to identify 14 significant positive correlations between metagenome-based abundances of mOTUs and metatranscript abundances of certain functions (**Supplementary Table 8**). For the correlating functions, our approach allowed the identification of population-level genomes containing expressed genes with those functions in most cases. Notably, most binned population-level genomes expressing the genes with the highest transcript abundances were annotated as the mOTUs that were also found to correlate with the functions (see **Supplementary Table 8** and example in **Supplementary Figure S8 a & b**). However, in most cases, additional taxa were found to contribute to the transcription of the functions of interest (**Supplementary Figure S8 c & d**). Functions of interest expressed by different taxa in different samples could also be linked to different binned genomes (see **Supplementary Figure S8 e & f**), in contrast to the correlation-based approach, which can only shed light on functions that are specifically expressed by a single taxon. Therefore, our chosen approach is capable of linking more functions to their genomic context than correlation, which is not only dependent on large numbers of independent samples, but also the chosen method of normalization. Our approach proved in particular more useful, when functions were expressed by multiple taxa.

Supplementary Note 8: Eukaryotic taxa and genes

In addition to genetic material from prokaryotes and viruses, eukaryotic sequences were also retrieved. To analyse these, Refseq-based annotations for each gene within the assemblies were downloaded from MG-RAST[13]. Multiple annotations with bit scores within 80 % of the best hit were traced to their lowest common ancestor. Genes with a lowest common ancestor from the kingdom Eukaryota were retained for the analysis, unless they were annotated as human. Between 359 and 11,060 different eukaryotic genes were found in each sample, indicating great variability in terms of eukaryotic representation in the faecal samples. These genes recruited between 0.01 and 0.22 % of the metagenomic reads and between 0.03 and 3.76 % of the metatranscriptomic reads (**Supplementary Figure S6**). In addition, proteins were uniquely identified for 1 to 72 of these genes in each sample (mean 15). The total area under the ion-chromatography curve related to these proteins also varied greatly, accounting for up to 12.6 % of the total area under the curve in a few samples (**Supplementary Figure S6**). Three eukaryotic families were responsible for the majority of these proteins (with varying proportions in different samples): Saccharomycetaceae, Poaceae and Trichomonadidae. The Poaceae proteins were most likely food-derived, as the most abundant proteins were annotated as storage proteins, such as prolamins and cupin. The yeast family of Saccharomycetaceae may be both food-derived as well as human commensals or opportunistic pathogens. Their most abundant identified proteins were annotated as enolase and alcohol dehydrogenase, followed by other enzymes involved in glycolysis/gluconeogenesis including glyceraldehyde 3-phosphate dehydrogenase, fructose-bisphosphate aldolase, pyruvate kinase, phosphoglycerate mutase, and glucose-6-phosphate isomerase. This indicates metabolic activity of yeasts, which may play an important role with respect to the carbohydrate metabolism in the gastrointestinal microbiota.

The protist family of Trichomonadidae contains several member taxa which reside in the human gastrointestinal tract. In the largest group of samples (14 samples of 8 individuals from 4 families), the eukaryotic taxon recruiting most metagenomic reads was *Trichomonas vaginalis*. As *Trichomonas vaginalis* does not reside in the gastrointestinal tract, this is likely a wrong assignment and the sequences in truth originate from a related commensal intestinal flagellate, such as *Pentatrichomonas hominis*, the genome of which, in contrast to *Trichomonas vaginalis*[32], has not yet been sequenced. Transcripts of genes of this organism were also relatively common (recruiting the highest number of metatranscriptomic reads in five samples). In addition, several of its proteins were uniquely identified in the same samples. The high activity of a gastrointestinal protist, which may be inferred from the abundance of its transcripts and proteins in some samples, suggests that further study of the genomes and functional roles of commensal protists would be a worthwhile effort.

A direct relationship between the documented food intake and the origin of identified eukaryotic sequences or proteins could not be established. This is likely due to varying passage times and degradation efficiencies of food in different individuals.

Supplementary Note 9: Viral genes

Despite the dominance of bacteria in faecal samples, other gastrointestinal agents, such as viruses can play an important role in human health. For example, a putative role for Coxsackie-viruses, enteric ssRNA viruses, in the development of T1DM has been investigated in other studies[33-35]. We therefore analysed the viral component of the omic datasets by screening contigs and genes for potential viral genomes or genes. For this, Refseq-based annotations for each open reading frame (ORF) were retrieved from MG-RAST[13]. Multiple annotations with bit scores of within 80 % of the best hit were traced to their lowest common ancestor. ORFs from the kingdom of Viruses were retained for the subsequent analyses. As the φX174 genome is used as a spike-in control in the sequencing process, genes recognized as belonging to this sequence were ignored.

For the 160 +/- 110 genes per assembly annotated as viral, the relationship between the metagenomic and metatranscriptomic coverage was assessed. Viral genes were mapped by 0.02 +/- 0.04 % of the metagenomic reads and 0.02 +/- 0.05 % of the metatranscriptomic reads, respectively (**Supplementary Figure S6**). The taxonomies of the best-hit annotations of the viral genes represented 54 +/- 14 different viral taxa in each sample (430 in total). Further analyses revealed that bacteriophages made up a large part of the diversity (70 +/- 8 % of the taxa) and contributed a large proportion to the viral ORFs (65 +/- 20 %). Furthermore, 92 +/- 10 % of the metagenomic and 47 +/- 31 % of the metatranscriptomic reads mapping to viral sequences in each sample could be mapped to bacteriophage ORFs. 51 +/- 14 % of the bacteriophage genes were found to be expressed (i.e. were genes to which metatranscriptomic reads were mapped).

As most of the viral metagenomic reads could be mapped to bacteriophage genes, but many metatranscriptomic reads mapped to other viral genes, we further analysed the viral genes mapped by metatranscriptomic but not metagenomic reads. These genes constituted 41 % of all viral genes. In 25 out of the 36 analysed samples (provided by 14 individuals), potentially human-associated viral genes were detected at the metatranscriptomic but not the metagenomic level. The most common of these human viruses was the human picobirnavirus, a gastrointestinal virus with unclear association with diarrhoea[36,37], which was present in 19 samples (9 individuals in 3 families; 6 with T1DM vs. 3 healthy). The metatranscriptomic reads which mapped to genes linked to this viral species accounted for 99 % of the reads mapping to potential human-associated viral genes identified in the metatranscriptome. Picobirnaviridae have dsRNA genomes, which explains why these were not detected at the metagenomic level. Close to 40 % of the human picobirnavirus genes mapped by metatranscriptomic reads were also mapped in the anti-sense direction and usually at lower levels, indicating that both genomes and transcripts of the human picobirnavirus were present, with the majority of detected RNA sequences being transcripts. In contrast to other viruses from the order Picornavirales, Coxsackie-viruses, which have been implicated in the development of T1DM[33-35], were not detected in this study. This would not be surprising, even if the T1DM in individuals in this cohort had been caused by Coxsackie-viruses, since these viruses do not usually persist in the gastrointestinal tract. None of the other detected viruses were found to be differentially abundant between the individuals with T1DM and the healthy family-members.

In addition to human viruses, genes assigned to plant-associated ssRNA viruses were common among the viral genes mapped by metatranscriptomic reads, but not metagenomic reads. Metatranscriptomic reads mapping to these genes accounted for 61 +/- 35 % of the reads mapping to viral genes with a species annotation. One of the most highly abundant genes assigned to a plant virus (a coat protein from Cucumber green mottle mosaic virus, the genome of which was completely recovered in one contig) was also detected in the metaproteome. None of the plant-associated viral taxa were detected in all samples, reflecting that these viruses are likely ingested with the food and pass through the gastrointestinal tract without becoming a resident.

Supplementary Note 10: Comparison of abundance and activity of community members

To assess the relationship between the abundance and activity of distinct community members, relative abundances of mOTUs were inferred from the representation of marker genes within the metagenomic and the metatranscriptomic data, respectively. In order to do this, metatranscriptomic data of 36 samples (**Supplementary Table 1**) were subjected to the same taxonomic analysis as the metagenomic data derived from the same samples (see **Supplementary Table 2**). Similar mOTUs were detected in the metagenomic and metatranscriptomic datasets, but a higher number of distinct mOTUs were represented in the metagenomic data when compared to the metatranscriptomic dataset (**Figure 3 a**). The abundance distributions of metagenomic and metatranscriptomic community profiles were similar (data not shown), with the exception of the metagenomic data containing slightly more mOTUs with low abundances. Correlation between the average metagenomic and metatranscriptomic mOTU abundances was high (Spearman's $\rho = 0.95$; P value $< 2.2*10^{-16}$), as was the correlation of individual metagenomic and metatranscriptomic mOTU abundances over all samples (Spearman's $\rho = 0.75$ +/- 0.24; 93 % of the mOTUs with FDR-adjusted P value < 0.05; **Figure 3 b**). An exception to this was *Ruminococcus* sp. 5 1 39BFAA (with a mean metagenomic abundance of 1 %) indicating that this population was more active in some samples than in others, as well as several lowly abundant mOTUs, whose levels are more likely to be affected by small technical variations.

Organisms with a relatively high mean activity (determined by calculating the ratio of the mOTU abundances inferred from the metatranscriptomic and metagenomic reads) included *Lactobacillus delbrueckii, Prevotella disiens, Clostridium saccharolyticum,* and *Anaerotruncus colihominis*, while organisms with a low mean relative activity included *Clostridium perfringens, Streptococcus pneumoniae* and *Streptococcus australis*, an oral

Veillonella sp. (taxon 158), as well as *Haemophilus influenzae,* and several less well classified Firmicutes mOTUs (**Figure 3 b**). While the highly active organisms were known residents of the gastrointestinal tract, several of the low-activity organisms are more commonly associated with the oral cavity and the upper respiratory tract. This pattern of activity, which relates for each organism to the distance of the rectum to the site of highest activity along the gastrointestinal tract, has previously been established[38] and suggests that the organisms with low activity in faecal samples are transferred to the large intestine where they remain relatively inactive.

Supplementary Note 11: Relationship between gene copy numbers and relative abundances of transcripts

A high level of correlation was found between the representations of mOTUs on the metagenomic and the metatranscriptomic levels (see **Figure 3 b**, **Supplementary Note 10**). However, the functional profiles in both datasets were less strongly correlated. Relative abundances of some functional categories within the metagenomes and metatranscriptomes were also correlated over the samples, as approximately 10 % of all functional categories showed strong correlations (Spearman's $\rho > 0.9$; FDR-adjusted *P* value $< 2*10^{-6}$; **Figure 3 g**). To assess which functions were most stably expressed, KOs that were present in all samples on both metagenomic and metatranscriptomic levels and significantly correlated were analysed for module/pathway membership. Ribosomal and flagellar proteins, as well as those related to LPS-synthesis were enriched (FDR-adjusted *P* value < 0.05) within these functional categories, indicating that these functions were most stably expressed in different bacterial populations.

To further assess the dependence of transcript abundances on metagenomic gene copy numbers at the single gene level, depths of coverage of the predicted open reading frames by reads at both omic levels were assessed. Metatranscriptomic reads could be mapped to approximately 39 +/- 6 % of the predicted genes in each sample. As the assemblies were derived from co-assembled metagenomic and metatranscriptomic reads, open reading frames without metagenomic coverage were also resolved (24 +/- 5 %; see also **Figure 3 c & d**). These likely originated from genes of lowly abundant organisms with a high transcriptional activity (and to a certain degree from genomes of RNA viruses, see **Supplementary Note 8**). To evaluate this assumption, the functional annotations of such genes were analysed. Approximately half of these genes could be annotated with a functional category. Among the genes annotated with a KEGG orthologous group (KO), modules were enriched (false-discovery-rate (FDR) adjusted *P* value of hypergeometric test < 0.05) which are usually highly expressed, e.g. ribosomal proteins and enzymes from amino acid metabolism.

For genes covered by both types of reads, the metagenomic and metatranscriptomic depths of coverage were found to be very weakly correlated (average Spearman's ρ = 0.06 +/- 0.09; see also **Supplementary Figure S9 b** & **c** and **Figure 3 c**). No correlation was also found for most functional genes (average Spearman's ρ = 0.1 +/- 0.1; 25 +/- 15 % of functional categories had FDR-adjusted *P* values < 0.05) except for house-keeping genes, such as essential single-copy genes[39] whose depths of coverage were more decisively correlated (average Spearman's ρ = 0.4 +/- 0.2; 70 +/- 21 % with FDR-adjusted *P* values < 0.05; **Supplementary Figure S9 c**). This is likely due to the fact that a certain functional repertoire is present in a given observed sample, which is for the most part expressed to different levels in different taxa, while housekeeping genes are more stably expressed across the different bacterial populations.

Supplementary Note 12: Whole community-based analyses of individuality

At the level of metagenomic taxonomic profiles, faecal communities in samples from the same individual were more similar to each other than to samples from other individuals (see **Supplementary Note 2** and **Supplementary Figure S2 a**). To assess whether this was also true for the other omic levels, distances between all possible pairs of samples were grouped by individuals, and intra- and inter-individual medians of all distances were compared. The determined intra-individual distances were significantly lower than the corresponding inter-individual distances (Wilcoxon rank sum test *P* values < 0.05). This was reflected in the Jensen-Shannon divergences and (binary) Soerensen dissimilarity indices (**Figure 4 a** and **Supplementary Figure S11**) for metagenomic and metatranscriptomic taxonomic profiles, as well as for the functional metagenomic, metatranscriptomic and metaproteomic profiles, and even human protein profiles. The apparent individuality of protein secretion into the gastrointestinal tract may well be an important shaping factor of the gastrointestinal and faecal microbiota, leading to the observed individuality on all omic levels. Detailed future work is required to ascertain the impact of the human gastrointestinal secretome on microbial communities.

Similar to the above analysis, permutational multivariate analysis of variance (MANOVA) of Bray-Curtis and Soerensen dissimilarities and Jensen-Shannon divergences of mOTU abundance profiles showed significant differences between the individuals (*P* values < 0.001). Analogous to the taxonomic results, permutational MANOVA of inferred transcript and protein abundances of microbial functions revealed significant groups formed by samples from the same individuals (*P* values 0.001 and 0.03, respectively).

A further indicator regarding the stability of gastrointestinal microbial communities is the fact that it is often possible to identify the donor of one sample within a number of samples from an earlier (or later) sample, by building sets of specific taxonomic or functional units ("codes")[5]. This was also possible using our data based on mOTU abundances inferred from metagenomic reads (see **Supplementary Figure S10 a**). In accordance with the high correlation between the mOTU profiles at the metagenomic and metatranscriptomic levels, metagenomic codes reliably identified metatranscriptomic profiles of the same samples and *vice versa* (see **Supplementary Figure S10 a**). Metagenomic codes could identify metatranscriptomic profiles of different samples of the same individual similarly well as their corresponding metagenomic profiles (**Supplementary Figure S10 a**). Metatranscriptomic codes did however result in a higher number of false negatives, both in terms of identification of metagenomes and metatranscriptomes (**Supplementary Figure S10 a**), which shows the higher variance in metagenomes compared to metatranscriptomes.

Stability of the microbiota was higher in individuals who had more diverse (**Supplementary Figure S27 a**) and more even (**Supplementary Figure S27 b**) gut microbial communities (Spearman's correlation coefficients between median Simpson's diversity indices and Pielou's evenness to total Soerensen indices (temporal variation) -0.6 and -0.7, respectively, *P* values 0.004 and 0.0004, respectively), similar to observations on diversity and stability in a larger study cohort analysed by 16S rRNA gene amplicon sequencing[40]. This observation was robust to adjustment of mOTU profiles by ACE-estimated[8] numbers of unobserved, yet present, mOTUs (**Supplementary Figure S27 c**). A similar trend was observed for the metatranscriptome-based mOTU abundances, but the correlation was not significant (**Supplementary Figure S27 d** & **e**). Within the functional datasets, no correlation between richness or diversity and temporal stability was observed (data not shown).

Supplementary Note 13: Assessment of similarity of microbial community structure and functional profiles in families, spouses and siblings

While microbial community structures and functional profiles were most similar among samples from the same individual (see **Supplementary Note 12**), samples from individuals of the same family also displayed more similar community structures and functional profiles when compared to those from members of different families (**Figure 4 a** and **Supplementary Figure S2**). To obtain a quantitative assessment of this observation, we grouped all possible distances between pairs of samples into one of the following groups: two samples from the same individual (intra-individual distances), two samples from two individuals from the same family (intra-family distances), or two samples from two individuals from two distinct families (inter-family distances). Intra-family distances between the community structures were smaller than the intra-individual distances, but smaller than inter-family distances (comparisons individual to families and families to unrelated individuals by Jensen-Shannon divergence or Soerensen dissimilarity: P value of Wilcoxon rank sum test < 0.05; **Figure 4 a** and **Supplementary Figure S11 b**). For the functional metatranscriptome, the same pattern was observed (**Figure 4 a**), except that the Soerensen dissimilarities between metatranscriptomic functional profiles of family members and metatranscriptomic functional profiles of unrelated individuals were not significantly different, indicating that family-specific characteristics of the functional metatranscriptome were quantitative and a similar functional potential was present also in unrelated individuals (**Supplementary Figure S11 d**). Also at the metaproteomic level a similar trend was observed, although here, intra-family distances were not significantly different from intra-individual distances (**Figure 4 a**). This was likely due to the great variability within the degrees of closeness in intra-family groups. Finally, the profiles of human proteins displayed lower intra-individual distances than distances between individuals, but distances between samples from one family were not significantly smaller than between unrelated individuals (**Figure 4 a**). Similar results were obtained by permutational MANOVA of Jensen-Shannon divergences between the individual-wise medians of samples (metagenomic community structure *P* value = 0.001, functional metatranscriptome and metaproteome *P* values = 0.002).

Similarities of the microbial community structures and functional profiles between spouses and siblings varied between families, with some spouse- and sibling-pairs having very similar microbiota and others diverging strongly (no significant differences between similarities/distances between spouse- and sibling-pairs except on the metaproteomic level; **Supplementary Figure S11**). Similarly, permutational MANOVA of Jensen-Shannon divergences of individual-wise medians indicated that neither the three spouse-pairs nor the groups of siblings were significantly different from each other at the level of community structure. At the metaproteomic level, the groups of siblings were more similar to each other than most other groups of related or unrelated individuals, when comparing Soerensen dissimilarities (**Supplementary Figure S11 e**). This was also reflected by permutational MANOVA of Soerensen dissimilarities between the individual-wise medians of metaproteomic profiles of siblings compared to other groups (*P* value = 0.025).

Neither Jensen-Shannon divergences (**Supplementary Figure S11 a**) nor Soerensen dissimilarities (**Supplementary Figure S11 b**) were different between the fathers' or mothers' and their respective children's microbial community structures compared to the corresponding distance measures for the individuals in the same families. As related individuals within the families (siblings or mother/father-children couples) did not have more similar faecal microbiota than the (unrelated) parents, these data point to a smaller genetic effect and a higher impact of the shared environment including most likely diet as a major common factor driving these trends.

Supplementary Note 14: Individuality of recovered genomes

In addition to the whole community analysis (see **Supplementary Note 12**), we were interested in assessing whether individuality was also a feature of the binned population-level genomes. Seven mOTUs were represented by well-reconstructed population-level genomes (at least 67 % of essential single-copy genes[39] recovered) derived from enough different individuals to allow at least three intra-individual and inter-individual comparisons. Soerensen dissimilarities between the genome reconstructions based on the presence and absence of functional annotations were calculated (see for example **Supplementary Figure S10 b**; the displayed genome represents *Eubacterium rectale*, the mOTU with the most significant difference between intra- and inter-individual concordance). In analogy to the community-wide analysis (**Supplementary Note 12**; **Figure 4 a**), dissimilarities between genomes reconstructed from samples of the same donor (intra-individual dissimilarities) were compared to dissimilarities between genomes reconstructed from samples of different individuals (inter-individual dissimilarities). Intra-individual dissimilarities were usually lower than inter-individual dissimilarities (**Figure 4 c**). A Wilcoxon signed rank test between Soerensen dissimilarity indices of all tested groups of genomes revealed the same trend (*P* value < 0.05). In addition to groups of genomes formed by the unanimous assignment of the same mOTU to the phylogenetic marker genes in the reconstructed genomes, groups of genomes based on similarity of single phylogenetic marker genes were analysed and analogous results were found (data not shown).

In conclusion, genomes reconstructed from samples of the same individual were more similar to each other than to closely related genomes reconstructed from the samples from other individuals. Therefore, the relative temporal stability of the faecal microbiota could not only be observed on the whole community level, but also at the level of the genomes of specific microbial populations common to multiple individuals.

Besides individuality in the functional potential of reconstructed genomes, we mined the expression patterns of these genomes for patterns of individuality. Well-reconstructed (> 67 % of essential unique genes[39]) genomes were again grouped by their assignment to mOTUs, as described above. The number of metatranscriptomic reads mapping to genes with functions found in all genomes assigned to the same mOTU were normalised using DESeq[41] (see **Supplementary Figure S10 c** as an example). Pairwise Spearman correlation coefficients of the resulting expression profiles were calculated. Intra-individual and inter-individual correlation coefficients were compared, in analogy to the analysis of dissimilarities described above. Correlations of expression profiles were usually higher between genomes reconstructed from samples of the same individual (**Figure 4 d**), and a Wilcoxon signed rank test between the correlation coefficients of all tested organisms was significant (*P* value < 0.05). The differences between intra-individual and inter-individual correlations were even more pronounced than in the analysis of the encoded functionalities above, as significant differences were found not only over the whole set of tested groups of reconstructed genomes or transcriptomes, but also within two further distinct groups of genomes, representing *Parabacteroides distasonis* and a Clostridiales mOTU (**Figure 4 d**). Analogous to the functional content of the reconstructed genomes, similar findings were made when genomes were analysed together which were selected because a single phylogenetic marker gene was very similar in all of them (data not shown).

In addition, a similar correlation analysis of gene expression profiles was carried out by mapping metatranscriptomic reads from different samples to well-reconstructed population-level genomes. For this, metagenomic reads of all samples were first mapped against all well-reconstructed population-level genomes (> 67 % of essential unique genes[39]). For this, the average depth of coverage was used as a proxy for estimating the abundance of a genome in a given sample. If the metagenomic depth of coverage of a genome by reads from one sample was at least eight fold, the metatranscriptomic reads of this sample were also mapped against the genes of the genome. The threshold of eight fold coverage was chosen because if we assume that coverage of a genome follows a Poisson distribution[42], we would not expect to miss more than 1/1000 of the genome in the metagenomic reads. Numbers of mapping reads were normalized using DESeq[41] and correlation coefficients of the expression profiles of genes mapped by metatranscriptomic reads in all compared samples were calculated. Again, intra- and inter-individual correlations were compared. This analysis confirmed that gene expression profiles for reconstructed population-level genomes were more correlated, if reads from samples of the same individual were used. The higher intra-individual correlations at the level of gene expression represent another level of individual-specificity observable in multi-omic datasets.

Comparisons of functional potential and expression profiles of closely related genomes were also of interest in the context of families. To perform this comparison, we required several closely related, well-reconstructed genomes in samples of several members of all four families. Only two mOTUs were represented by such genomes, allowing for comparisons of intra-individual, intra-family and inter-family distances as performed for the whole community profiles (compare **Supplementary Note 13**). One of these mOTUs (*Eubacterium rectale*) displayed greater intra-family than inter-family differences in functional potential of the reconstructed genomes (**Supplementary Figure S10 d**), while the other (*Alistipes putredinis*) did not. In addition, the expression profiles of reconstructed genomes assigned to the two mOTUs showed similar trends (**Supplementary Figure S10 e**). In order to ascertain whether differences between intra-family-groups, such as mother- or father-child pairs, siblings or spouses exist, a future study with a larger cohort will be necessary

Supplementary Note 15: Additional discussion of family-specific differences in the gastrointestinal microbiota and faecal human proteome

As family-membership greatly influenced the faecal community structures and functional omes (see **Supplementary Note 13**), we assessed differences in the datasets between the families. As discussed in **Supplementary Note 1**, the families followed rather distinct diets, which may be the cause for differences observed in the microbial communities, which we will discuss first (see also **Supplementary Figure S12 a&b**).

Few correlations were observed between the reported intake of food or inferred intake of nutrients and transcript abundances within the faecal microbiota. One correlation was between inferred levels of riboflavin intake and transcript levels of genes coding for proteins with a Nop10p domain (nucleolar RNA-binding protein; **Supplementary Figure S13 a**). Riboflavin intake has been described previously to correlate with faecal abundance of *Prevotella*[43], which was not the case in this cohort on any omic level (data not shown). Here, in most cases where the transcribing taxa could be found, these were *Methanobrevibacter smithii*. This is in accordance with the facts that nucleolar-like RNAs and proteins have been described in archaea[44], and *M. smithii* was often more abundant in faecal samples taken after high estimated intake of riboflavin. However, a mechanistic link between riboflavin availability and *M. smithii* growth has not yet been established, and *M. smithii* can likely synthesize riboflavin[45]. The intake of riboflavin did not differ significantly between the families, in contrast to the two other foods and nutrients (fruits and maltose) which correlated to transcript abundances (**Supplementary Figure S13 b**, **c**, **e** & **f**). These data may suggest that the eating habits of families shape their specific microbiota. Alternatively, the family-specific characteristics of the microbiota may have different causes including genetics and the detected correlations were observed because the members of the families not only share determinants of microbial function but also diet.

Out of the 500 mOTUs detected at the metagenomic level, 54 mOTUs showed statistically significant differential abundances between the four families of the cohort when controlling additionally for disease status (main effect with FDR-adjusted *P* value < 0.05 and fold change of 2). In particular, many mOTU abundances differed in family M04. Similarly, 1,426 functions were found to have differentially abundant transcripts between the families (main effect with FDR-adjusted *P* value < 0.05 and fold change of 2, **Figure 6 a**). 349 of these functions were also measured in the metaproteome, with approximately 60 being specific to the same families on both omic levels.

153 metabolic KEGG orthologous groups (KOs) from a reconstructed community-wide metabolic network exhibited a significant main effect of membership in family M04. The top-scoring module of this network (FDR = 0.0001; **Supplementary Figure S28**) was enriched in nodes related to the biosynthesis of amino acids, in particular glycine, serine and threonine. Further nodes were related to carbon-metabolism, such as fructose and mannose, terpenoids, as well as methanogenesis from methanol and methylamine (which was less abundant in the individuals from family M04) in addition to genes from biosynthesis pathways of pyrimidine and purines. In families M01 to M03 (and a sample of M04-6), the enzymes of methanogenesis were transcribed by lowly abundant members of the community such as *Methanobrevibacter smithii*, which was hardly detectable in the other samples. It is notable that the majority of the functions with higher transcript abundances in family M04 were for the most part transcribed from genes binned to genomes of *Prevotella copri*, which dominated the microbiota of most individuals in family M04. In contrast, the same functions were transcribed by a large number of different organisms in the other samples. At the

metaproteomic level (small networks in **Supplementary Figure S27 b & d**) this trend was even more pronounced. These observations imply that the strongest functional traits of family M04 were linked to the dominance of *Prevotella copri* in the faecal microbiota of these individuals.

Among the human proteins detected in the faecal samples, RETN and VNN1 were more abundant in the individuals from family M04, and the annexin ANXA2 was only detected in family M03. In addition, the protein family of lipocalins was more abundant in the stool of individuals from family M04, due to high levels of LCN2 and/or ORM1 or ORM2. ORM1/2 and LCN2 are markers of inflammation[46] and LCN2 has been validated as a biomarker for intestinal inflammation in mice[47]. RETN is a pro-inflammatory secreted protein, which is expressed in glandular cells along the gastrointestinal tract and plays a protective role against cell stress[48]. As VNN1, a pantetheinase, is also involved in protection of tissue against stress[49], these may be indicators of an inflammatory signature of the gut lining in family M04, which was not manifest in clinical symptoms. This family also harboured the mucin-degrading *Prevotella copri* in high abundance, which has previously been linked to inflammatory disease[50]. Therefore some of the observed differences in the human faecal proteins may reflect the interaction of the microbiota with the gastrointestinal epithelium. Links between inflammatory markers in human stool and specific taxa should be followed up on in a dedicated study in a larger cohort.

Supplementary Note 16: T1DM and microbial community structure

As age and BMI are known to affect the gastrointestinal microbiota, we took care to avoid strong biases in these respects in the cohort. Age and gender distribution and body-mass-index (BMI) of the individuals with T1DM were roughly matched by the healthy individuals in the cohort (see **Supplementary Figure S15 a & b**). Analysis of microbial diversity revealed no statistically significant differences between the mOTU richness, diversity or temporal variability of faecal microbiota of individuals with T1DM and their healthy relatives (**Supplementary Figure S15 c - e**). The microbial community structures apparent in the context of T1DM were not more similar to those observed in other samples from individuals with T1DM from the same family when compared to samples from healthy family members. Likewise, samples from healthy individuals did not exhibit more similar community structures compared to other samples from healthy individuals from the same family (**Supplementary Figures S2 a** and **S11 a & b**).

In a study based on qPCR and PCR-denaturing gradient gel electrophoresis[51], a decrease in the Firmicutes-to-Bacteroidetes ratio in children with T1DM relative to healthy controls was found, as well as a significant negative correlation between HbA1c levels and the Firmicutes-to-Bacteroidetes ratio. In our cohort where T1DM was longer established in most cases however, neither a significant difference in the Firmicutes-to-Bacteroidetes ratio between the individuals with T1DM and their healthy relatives (**Supplementary Figure S15 f**) nor a significant correlation between the Firmicutes-to-Bacteroidetes ratio and HbA1c levels (**Supplementary Figure S15 g**) were observed. In contrast, the level of glucose in the blood of the individuals with T1DM correlated with the Firmicutes-to-Bacteroidetes ratio (Spearman's $\rho = 0.8$; *P* value = 0.007; **Supplementary Figure S15 h**).

Differential analysis of the individual-wise median metagenomic abundances of mOTUs was carried out using a model taking into account family membership, because family membership accounted for large differences in the community structures (see **Supplementary Notes 13 & 15).** At the species level, the organism with the greatest main effect of T1DM (FDR-adjusted *P* value = 0.09; overall $\log_2$ fold change of 1.4 comparing individuals with T1DM to their healthy relatives) was *Escherichia coli* (**Supplementary Figure S15 i**). *E. coli* has not before been found to be differentially abundant in individuals with T1DM. Its enterotoxin EtxB (also referred to as EltB), however, was found to inhibit development of autoimmune diabetes in NOD mice[52]. In the present study, no amino acid sequences with significant sequence identity to EltB (NC_014232, positions 2109 to 2483) were predicted from the metagenomic and metatranscriptomic data. At the higher taxonomic ranks (genera, families, orders, classes and phyla) no significant differences were detected.

Several recent studies[53,54] have observed enrichments of *Bacteroides* spp. and in particular *B. dorei*[55] in children with T1DM compared to children without T1DM. The present cohort included three children, two of whom had T1DM, but neither had high levels of *B. dorei* or the closely related *B. vulgatus.* Generally, the abundance of the *B. dorei/vulgatus* mOTU varied between 0.3 % and 31 % in the different individuals while staying constant over time in the individuals. No enrichments in *B. dorei/vulgatus* were observed in the adults with T1DM, either. Moreover, population-level genomes of *B. dorei/vulgatus* were reconstructed from samples from individuals with T1DM (six largely complete, homogenous genomes; see **Supplementary Table 4**), but, in accordance with the assembly-independent analysis, their relative abundances were not higher than the less well assembled and binned representatives from samples of healthy individuals.

As expected from the close relationship between the metagenomic and metatranscriptomic representations of mOTUs, differences between the metatranscriptomic abundances of microbial taxa in individuals with T1DM and their healthy relatives were likewise rare at all analysed taxonomic ranks. The organism with the only significant main effect (*P* value = 0.01 in DESeq2 analysis) was an mOTU belonging to the order Clostridiales (**Supplementary Figure S15 k**). Based on the ratio of relative abundances at the metatranscriptomic and the metagenomic levels, this mOTU was highly active (>90th percentile). However, this mOTU is as yet uncharacterized and the present data did not allow for a genomic reconstruction. In addition to the Clostridiales mOTU, [*Ruminococcus*] *torques* was more highly (but not significantly) abundant in the individuals with T1DM, on both the metagenomic and metatranscriptomic levels (**Supplementary Figure S15 j**; the square brackets indicate that the taxon usually referred to as *Ruminococcus torques* does not phylogenetically belong to the genus *Ruminococcus*[10]). *Ruminococcus torques* has previously been proposed to be linked to irritable bowel syndrome and autism spectrum disorder[56]. In another study, it was found at higher levels in unaffected family members of individuals with Crohn's disease compared to healthy controls without close relatives with Crohn’s disease[57]. In conclusion, robust differences between taxonomic profiles of individuals with T1DM and their healthy relatives were rare. However, functional differences may be independent of taxonomic profiles, if the potential differentially expressed functions are encoded by different organisms or if commonly present organisms regulate their expression levels according to the disease state.

Supplementary Note 17: Additional discussion of functional differences in microbiome of individuals with T1DM

Differential analysis of the median community-wide transcript abundances of functions within the individuals' microbiomes was carried out using a model taking into account family membership, in analogy to the taxonomic analysis. Transcripts with significantly higher abundances in samples from individuals with T1DM included the cysteine protease staphopain A, genes involved in the regulation of motility or biofilm formation and cell surface structure, a putative cellulose degradation gene, ornithine cyclodeaminase, methylaspartate mutase, a catalase, and genes related to mobile genetic elements or bacteriophages. Transcripts of genes containing several domains of unknown functions were found to be more abundant in the healthy individuals (**Supplementary Figure S16 b**). The transcription elongation factor and the putative anhydrosylase functions with differential abundance are not discussed in the following, as their transcript abundance was significantly correlated with alcohol and white bread intake, respectively (data not shown), indicating that these results may be confounded by diet rather than be the result of physiological differences in T1DM.

As a *Staphylococcus aureus* virulence factor, staphopain A is capable of cleavage and inactivation of several receptors of the innate immune system[58-62]. *Staphylococcus* spp. were not detected in the taxonomic analysis of the metagenome and in only three samples in the metatranscriptome. Therefore, the presence of staphopain A transcripts in the metatranscriptomes of twelve individuals was unexpected. However according to KEGG, *Enterococcus* spp. and *Roseburia hominis* carry similar genes. The effect of these enzymes in such non-pathogenic organisms is unknown but may well be immunomodulatory. For instance, beneficial effects of other cysteine proteases have been demonstrated in mice[63]. The population-level genomes of organisms encoding staphopain A within this cohort could not be reconstructed (which would be expected if they are indeed low abundant *Staphylococcus* spp.), and therefore their phylogeny remains presently unknown.

Two putative negative regulators of motility (Pfam:YliH and Pfam:GlgS[6]) were expressed almost exclusively from reconstructed population-level genomes of *E. coli*. The same populations were also responsible for the higher levels of catalase transcripts.

Genes with the Pfam annotation CBM_X2 (Carbohydrate binding domain X2, a structural domain in in cellulose-degrading cellulosomes[64,65]) were expressed by *Coprococcus eutactus* in three of the four samples with the highest expression of this gene (M1.1-V2, M2.3-V2 and M2.3-V3; in the fourth sample, M1.1-V1, taxonomic annotation of the binned genome which harboured the gene with the highest transcript level was ambiguous, but with 9 out of 10 phylogenetic marker genes most similar to *Coprococcus eutactus*, it likely belonged to the same mOTU). *C. eutactus* is known to degrade different hemicelluloses[66,67], which in part are degraded also in cellulosomes[68]. Interestingly, *C. eutactus* was comparatively highly abundant in other samples, such as M1.3-V2 and M2.4-V2 (both from healthy individuals), in which hardly any transcripts of CBM_X2 domains were detected. What is more, a CBM_X2 domain was detected in the reconstructed genomes of *C. eutactus* in these samples, but the genes were not expressed (**Supplementary Figure S18**). As cellulose degradation can be controlled by catabolite repression[68], this may indicate a better accessibility of preferred carbon sources.

Another functional category that was found to have more transcripts in the microbiota of the individuals with T1DM was the KEGG orthologous group (KO) K10954, annotated as Zonula occludens toxin (Zot). Zot was first described in *Vibrio cholerae* as secondary toxin, causing mild diarrhoea by interaction with intestinal tight junctions[69,70]. As a link between zonulin, a human protein with a similar interaction with tight junctions as Zot, and T1DM has been discussed[71] and alterations in gut permeability have been observed in T1DM patients[72,73], this would represent a very interesting finding. However, as we will discuss here, the annotation is misleading, because the domain annotated as Zot is not the domain shown to exert the Zot effect.

Already at the time of discovery of Zot in *V. cholerae*, similarities to a plasmid- and a bacteriophage-encoded gene were recognized[70]. The Zot-domain (in the KEGG database, but also recognized by Pfam) is an ATPase-like domain in the N-terminal ~200 aa of the *V. cholerae* Zot protein (in *V. cholerae* Zot, no ATPase activity was found[74]). In the bacteriophage proteins, this cytoplasmic domain is involved in bacteriophage assembly. In the *V. cholerae* Zot protein, this domain is followed by a transmembrane region and the C-terminus of the protein is located in the periplasm. In the C-terminal domain, Zot contains an octamer $G^{291}RLCVQDG^{298}$, which is cleaved from the membrane-associated and cytoplasmic parts of the protein and has been shown to be required for binding to epithelial cells and Zot action on tight junctions[75]. None of the predicted proteins in our study contained this motive. To assess whether the Zot-domain containing protein predictions may be associated with bacteriophage genomes in our study, we screened the functional annotations of neighbouring open reading frames. 25 open reading frames annotated as Zot had open reading frames with functional annotations on the same contigs. Among these were one phage replication initiation protein and one integrase/recombinase, which may point to integrated phage genomes.

A similar situation was apparent for the case of *Campylobacter concisus*, for which several isolates were also found to harbour genes for proteins with Zot domains[76]. Authors of another study[77] have argued that the *C. concisus* Zot protein may have a causal role in inflammatory bowel disease (IBD), partially based on the fact that Zot-positive isolates were found more often in individuals with IBD than healthy controls. Contrarily, in the first study[76], Zot-positive isolates were found more often in healthy controls than diarrhoeic individuals. The *C. concisus* Zot does not have the GRLCVQDG-octamer, but a sequence $\mathbf{G}^{123}\mathbf{R}\mathrm{FLSYH}\mathbf{G}^{130}$ with some alleged similarity to the octamer[77]. The position of this sequence within the *C. concisus* protein is however completely different from the *V. cholerae* protein, as the *C. concisus* motive is positioned within the ATPase domain and not at the C-terminus of the protein. One Zot-protein each was predicted on two contigs in our dataset which contained a **GR**KAART**G** motive C-terminal of the Zot-domain and a likely transmembrane domain (predicted using the TMHMM web service at http://www.cbs.dtu.dk/services/TMHMM/). However, no transcripts or peptides were found indicating the expression of these genes. In conclusion, until the existence of experimental proof of an effect of the non-*V. cholerae*-like Zot proteins, we cannot assume their function to be Zot-like and to be involved in the T1DM of the individuals in this cohort. As an additional note, human zonulin is processed from the product of the haptoglobin gene (*HP*). The HP protein was uniquely identified in only one sample (M01.1-V3) in the current study, so we could not assess if zonulin is more highly expressed in the individuals with T1DM here. Overall, the possible role of Zot-like proteins in the context of T1DM (and potentially other diseases) requires additional detailed future study.

From the analyses of metabolic functions, the top-scoring module of a metabolic reconstruction based on KOs contained nodes related to carbon metabolism as well as amino-acid metabolism (**Supplementary Figure S29**). The centre of this module was formed by monoamine oxidase, which was more abundant in the individuals with T1DM. Higher abundance of monoamine oxidase may be related to the use of host-derived substances like catecholamines or indolamines. Catecholamines, especially epinephrine, play an important role in the regulation of blood glucose levels in T1DM, but there is contradictory data on whether catecholamine levels are increased or decreased in T1DM, which likely depends on the level of glycaemic control of the insulin treatment[78-80]. Interesting, catecholamine levels have also been linked to microbial iron uptake[81,82], which is restricted by lactotransferrin among others. Lactotransferrin was observed to be more abundant in some individuals with T1DM (**Supplementary Note 18**). A side-product of monoamine oxidase is hydrogen peroxide, which may explain the higher levels of catalase expression in the same samples. Both monoamine oxidase and catalase also form part of the phenylalanine and tryptophan metabolism, so these results may also indicate a change in microbial phenylalanine and/or tryptophan accessibility.

While no conclusive picture emerges from the discussed functions, several of them suggest a primed immune system and higher levels of reactive oxygen species (ROS) in the individuals with T1DM that favour expression of catalases and immune-evasive genes like staphopain A. The differences observed between individuals with T1DM and their healthy relatives at the metatranscriptomic level were not represented at the metaproteomic level. In part this was due to the fact that more than 70 % of the differentially abundant functions of the metatranscriptome were not even detected or identified in the metaproteome (see **Supplementary Table 6**).

Within the metaproteomic dataset, no microbial functional categories were significantly differentially abundant between the individuals with T1DM and the healthy family members when applying multiple-testing adjustment. However, among the proteins with the most significant changes (*P* value of Wilcoxon rank sum test < 0.025; 11; see **Supplementary**

Figure S17 a), three were likely involved in carbohydrate metabolism and transport. Another difference was found in the protein EutM, which was more abundant in faeces of individuals with T1DM. The organism producing this protein could not be conclusively determined from our data. EutM is part of an organelle-like structure[83] essential for ethanolamine use[84]. Ethanolamine metabolism can depend on tetrathionate[85], which can be produced by reaction of NO and ROS with thiosulfate, a side product of hydrogen sulphide detoxification in colonic cells[86]. The higher abundance of ROS is also indicated by the higher abundance of NADH peroxidase and dihydroflavonol-4-reductase (K00091), which is involved in the synthesis of a flavonoid antioxidant. This protein was uniquely identified from genes belonging to several well-reconstructed genomes of mOTUs not classified beyond the order Clostridiales (**Supplementary Figure S19**), which have not yet been linked to antioxidant production. (These mOTUs did not include the mOTU found to be more abundant in the metatranscriptomes of the individuals with T1DM.) Together, these results indicating higher levels of ROS in individuals with T1DM somewhat resonate with the indications by the metatranscriptome analysis above.

To conclude, we found diverse microbial functions with subtle differences in expression at the metatranscriptomic or metaproteomic levels in individuals with T1DM compared to their healthy family members. The differences were attributable to different microbial populations, which were not differentially abundant between individuals with T1DM and healthy control family members (with the possible exception of *E. coli*, see **Supplementary Note 16**). The functions overall may reflect the consequences of inflammation and changes in exocrine pancreatic function in the microbiome (**Figure 6**).

Supplementary Note 18: Additional discussion of T1DM-related changes in human proteins and repercussions on the gastrointestinal microbiome

Although none of them were statistically significant after multiple-testing adjustment, differences between the human faecal proteome of individuals with T1DM and their healthy family members were observed (see **Supplementary Table 5**). Among the proteins with the strongest differences between individuals with T1DM and healthy individuals were some which have also been observed to be auto-antigens in some cases of T1DM (α-amylase 2 (AMY2A) and lactotransferrin (LTF1))[87]. Lactotransferrin, which was more abundant in the stool of individuals with T1DM, is an antimicrobial protein[88]. It is secreted into several bodily fluids and into the gastrointestinal lumen, and its expression by colonic epithelial cells has been shown to be affected by its receptor intelectin (ITLN1)[89], which we also found in increased abundance in individuals with T1DM. Intelectin has been shown to be induced by pro-inflammatory cytokines[90] while lactotransferrin has both pro-inflammatory as well as anti-inflammatory properties (of which the anti-inflammatory properties are usually considered more relevant)[91]. Secretion of intelectin has also been shown to be down-regulated in response to insulin in some cell types, independent of its function as a lactotransferrin receptor[92].

We found four proteins which are expressed preferentially by the exocrine pancreas[93] (AMY2A, AMY2B, CPA1 and CUZD1; **Figure 5 a & b** and **Supplementary Figure S17 b**) among the ten human proteins with the greatest decrease in abundance in the faeces of individuals with T1DM. Due to the tissue-specificity of their expression, this decrease possibly reflects a weakening of pancreatic exocrine function. We ascertained the human origin of the pancreatic amylases (AMY2A and AMY2B) by comparing the identified unique peptides from each of these proteins to sequences of plant amylases and microbial α-amylases, which are commonly used in food processing[94]. We found the identified unique peptides to occur in none of the known α-amylases from plants or microbes. Comparing the levels of amylases to the dietary records, we found no foods and nutrients to significantly correlate with the amylase levels, indicating that the differences in amylase levels were not related to the diet of the study participants.

Next we aimed to test whether potential repercussions of this T1DM-related decrease in α-amylases, such as changed nutrient availability, would be reflected in the faecal microbiome. No significant correlations between the protein abundances and the abundances of any mOTUs were found. However, to test whether these differences may result in possible metabolic differences in the respective microbiomes, we focussed on KEGG orthologous groups (KOs) from which we inferred a generalized community-wide metabolic network. Based on this network, we then calculated correlation coefficients between the transcript abundances of the individual KOs and the abundances of the pancreatic proteins. We uncovered 24 significant correlations between microbial functional transcripts and the combined relative abundances of the α-amylase proteins AMY2A and AMY2B (*P* value < 0.05 after false-discovery-rate (FDR) adjustment; |Spearman's ρ| > 0.75; **Supplementary Table 5**). Overall, the functions of the positively correlating transcripts were enriched in functions relating to microbial central carbon metabolism (**Figure 5 c)**. In addition, transcript levels of thiazole synthase, the central enzyme of the biosynthesis of thiamine (ThiG, K03149; **Supplementary Figure S20 a**) were found to be positively correlated with amylase abundances (**Supplementary Figure S20 b**). Thiamine is the precursor of an essential co-factor in decarboxylation reactions and is therefore also linked to the central carbon metabolism, especially the citric acid cycle. The most abundant transcripts of *thiG* were expressed from contigs linked to *Prevotella copri*, but in samples without *P. copri*

(**Supplementary Figure S20 c**)*,* *thiG* was transcribed by various other populations (**Figure 2 e**, **Supplementary Figure S20 d**), including *Bacteroides dorei* or *vulgatus* and *Alistipes putredinis*, among others. This observation illustrates once more that in mixed microbial communities, functions of interest can be contributed by different organisms[95]. As AMY2A proteins were more abundant in healthy individuals, *thiG* transcripts also tended to be more abundant in healthy individuals (**Supplementary Figure S20 e**). In contrast, high thiamine intake levels derived from the dietary and medication records did not correlate to low *thiG* transcript levels (data not shown), which is line with the fact that dietary thiamine is known to be taken up in the small intestine and would therefore not reach the colon[96] and would likely not affect microbial activity there.

Thiamine is known to counteract the formation of advanced glycation endproducts, a common cause for complications in T1DM[97] and low thiamine levels have been documented in adults with T1DM[98-101], a situation which can be aggravated in adolescents with diabetic ketoacidosis after insulin treatment[102]. While thiamine is absorbed from food in the small intestine[96], uptake transporters are also expressed in the large intestine[103,104], which means that microbiome-derived thiamine may also contribute to host thiamine levels. Therefore, we were interested in the thiamine levels in human blood. In our cohort, several individuals took vitamin formulations including thiamine (see **Supplementary Table 1**). The thiamine levels in the blood plasma of the few individuals who did not take thiamine supplements were not significantly different between individuals with T1DM and healthy family members, and the thiamine plasma levels did not correlate with *thiG* transcript levels or AMY2 protein abundances (**Supplementary Figure S20 f**). This may be due to sufficient dietary ingestion and uptake of thiamine. Nevertheless, potential feedbacks between gastrointestinal starch hydrolysis, sugar uptake into the human system and human metabolism on the one hand and microbial central carbon metabolism and thiamine synthesis, which may be taken up by the human host, on the other hand, may also exist. In order to deconvolute these complex relationships, a detailed follow-up in a dedicated study involving a larger cohort of families which explicitly do not take thiamine supplements should be undertaken.

Supplementary References

1. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32,** 834–841 (2014).

2. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28,** 1420–1428 (2012).

3. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18,** 821–829 (2008).

4. Scholz, M., Lo, C.-C. & Chain, P. S. G. Improved assemblies using a source-agnostic pipeline for MetaGenomic Assembly by Merging (MeGAMerge) of contigs. *Sci Rep* **4,** 6480 (2014).

5. Franzosa, E. A. *et al.* Identifying personal microbiomes using metagenomic codes. *Proc. Natl. Acad. Sci. U.S.A.* **112,** E2930–E2938 (2015).

6. Rahimpour, M. *et al.* GlgS, described previously as a glycogen synthesis control protein, negatively regulates motility and biofilm formation in *Escherichia coli*. *Biochem. J.* **452,** 559–573 (2013).

7. Luo, W. & Brouwer, C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29,** 1830–1831 (2013).

8. Chiu, C.-H., Wang, Y.-T., Walther, B. A. & Chao, A. An improved nonparametric lower bound of species richness via a modified good-turing frequency formula. *Biom* **70,** 671–682 (2014).

9. Beisser, D., Klau, G. W., Dandekar, T., Muller, T. & Dittrich, M. T. BioNet: an R-Package for the functional analysis of biological networks. *Bioinformatics* **26,** 1129–1130 (2010).

10. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic

marker genes. *Nat. Methods* **10,** 1196–1199 (2013).

11. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 31 (2014).

12. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9,** 75 (2008).

13. Meyer, F. *et al.* The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9,** 386 (2008).

14. Achenbach, P. *et al.* Stratification of type 1 diabetes risk on the basis of islet autoantibody characteristics. *Diabetes* **53,** 384–392 (2004).

15. Insel, R. A. *et al.* Staging presymptomatic type 1 diabetes: a scientific statement of JDRF, the Endocrine Society, and the American Diabetes Association. *Diabetes Care* **38,** 1964–1974 (2015).

16. Ziegler, A.-G. *et al.* Seroconversion to multiple islet autoantibodies and risk of progression to diabetes in children. *JAMA* **309,** 2473–2479 (2013).

17. Fuchtenbusch, M., Ferber, K., Standl, E. & Ziegler, A. G. Prediction of type 1 diabetes postpartum in patients with gestational diabetes mellitus by combined islet cell autoantibody screening: a prospective multicenter study. *Diabetes* **46,** 1459–1467 (1997).

18. Järvelä, I. Y. *et al.* Gestational diabetes identifies women at risk for permanent type 1 and type 2 diabetes in fertile age: predictive role of autoantibodies. *Diabetes Care* **29,** 607–612 (2006).

19. Redondo, M. J. & Eisenbarth, G. S. Genetic control of autoimmunity in Type I diabetes and associated disorders. *Diabetologia* **45,** 605–622 (2002).

20. Welch, A. A., Luben, R., Khaw, K. T. & Bingham, S. A. The CAFE computer program for nutritional analysis of the EPIC-Norfolk food frequency questionnaire

and identification of extreme nutrient values. *J Hum Nutr Diet* **18,** 99–116 (2005).

21. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–180 (2011).

22. David, L. A. *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biology* **15,** R89 (2014).

23. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505,** 559–563 (2013).

24. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444,** 1022–1023 (2006).

25. Finucane, M. M., Sharpton, T. J., Laurent, T. J. & Pollard, K. S. A Taxonomic signature of obesity in the microbiome? Getting to the guts of the matter. *PLoS ONE* **9,** e84689 (2014).

26. Howe, A. C. *et al.* Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 4904–4909 (2014).

27. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* **32,** 1088–1090 (2016).

28. Lai, B., Wang, F., Wang, X., Duan, L. & Zhu, H. InteMAP: Integrated metagenomic assembly pipeline for NGS short reads. *BMC Bioinformatics* **16,** 244 (2015).

29. Gray, K. A., Yates, B., Seal, R. L., Wright, M. W. & Bruford, E. A. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Research* **43,** D1079–D1085 (2015).

30. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Published in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)* 1–6 (1996).

31. Kostic, A. D. *et al.* The dynamics of the human infant gut microbiome in

development and in progression toward type 1 diabetes. *Cell Host Microbe* **17,** 260–273 (2015).

32. Carlton, J. M. *et al.* Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. *Science* **315,** 207–212 (2007).

33. Stene, L. C. *et al.* Enterovirus infection and progression from islet autoimmunity to type 1 diabetes: the Diabetes and Autoimmunity Study in the Young (DAISY). *Diabetes* **59,** 3174–3180 (2010).

34. Laitinen, O. H. *et al.* Coxsackievirus B1 is associated with induction of β-cell autoimmunity that portends type 1 diabetes. *Diabetes* **63,** 446–455 (2014).

35. Rodriguez-Calvo, T. & Herrath, von, M. G. Enterovirus infection and type 1 diabetes: closing in on a link? *Diabetes* **64,** 1503–1505 (2015).

36. Ng, T. F. F. *et al.* Divergent picobirnaviruses in human feces. *Genome Announc* **2,** (2014).

37. Mondal, A. & Majee, S. Novel bisegmented virus (picobirnavirus) of animals, birds and humans. *Asian Pacific Journal of Tropical Disease* **4,** 154–158 (2014).

38. Franzosa, E. A. *et al.* Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. U.S.A.* **111,** E2329–E2338 (2014).

39. Dupont, C. L. *et al.* Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *The ISME Journal* **6,** 1186–1199 (2011).

40. Flores, G. E. *et al.* Temporal variability is a personalized feature of the human microbiome. *Genome Biology* **15,** 531 (2014).

41. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11,** R106 (2010).

42. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2,** 231–239 (1988).

43. Carrothers, J. M. *et al.* Fecal microbial community structure is stable over time and related to variation in macronutrient and micronutrient intakes in lactating women. *Journal of Nutrition* **145,** 2379–2388 (2015).

44. Omer, A. D. *et al.* Homologs of small nucleolar RNAs in Archaea. *Science* **288,** 517–522 (2000).

45. Samuel, B. S. *et al.* Genomic and metabolic adaptations of *Methanobrevibacter smithii* to the human gut. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 10643–10648 (2007).

46. Logdberg, L. & Wester, L. Immunocalins: a lipocalin subfamily that modulates immune and inflammatory responses. *Biochim Biophys Acta* **1482,** 284–297 (2000).

47. Chassaing, B. *et al.* Fecal lipocalin 2, a sensitive and broadly dynamic non-invasive biomarker for intestinal inflammation. *PLoS ONE* **7,** e44328 (2012).

48. Suragani, M. *et al.* Human resistin, a proinflammatory cytokine, shows chaperone-like activity. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 20467–20472 (2013).

49. Naquet, P., Pitari, G., Dupre, S. & Galland, F. Role of the Vnn1 pantetheinase in tissue tolerance to stress. *Biochem. Soc. Trans.* **42,** 1094–1100 (2014).

50. Scher, J. U. *et al.* Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. *eLife* **2,** e01202 (2013).

51. Murri, M. *et al.* Gut microbiota in children with type 1 diabetes differs from that in healthy children: a case-control study. *BMC Med* **11,** 46 (2013).

52. Ola, T. O. & Williams, N. A. Protection of non-obese diabetic mice from autoimmune diabetes by *Escherichia coli* heat-labile enterotoxin B subunit. *Immunology* **117,** 262–270 (2006).

53. de Goffau, M. C. *et al.* Aberrant gut microbiota composition at the onset of type 1 diabetes in young children. *Diabetologia* **57,** 1569–1577 (2014).

54. Mejía-León, M. E., Petrosino, J. F., Ajami, N. J., Dominguez-Bello, M. G. & la

Barca, de, A. M. C. Fecal microbiota imbalance in Mexican children with type 1 diabetes. *Sci Rep* **4,** (2014).

55. Davis-Richardson, A. G. *et al. Bacteroides dorei* dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes. *Front Microbiol* **5,** 678 (2014).

56. Malinen, E. Association of symptoms with gastrointestinal microbiota in irritable bowel syndrome. *World J. Gastroenterol.* **16,** 4532 (2010).

57. Joossens, M. *et al.* Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. *Gut* **60,** 631–637 (2011).

58. Jusko, M. *et al.* Staphylococcal proteases aid in evasion of the human complement system. *J Innate Immun* **6,** 31–46 (2014).

59. Laarman, A. J. *et al.* Staphylococcus aureus Staphopain A inhibits CXCR2-dependent neutrophil activation and chemotaxis. *The EMBO Journal* **31,** 3607–3619 (2012).

60. Hewit, K. D., Fraser, A., Nibbs, R. J. B. & Graham, G. J. The N-terminal region of the atypical chemokine receptor ACKR2 is a key determinant of ligand binding. *Journal of Biological Chemistry* **289,** 12330–12342 (2014).

61. Kantyka, T. *et al. Staphylococcus aureus* proteases degrade lung surfactant protein A potentially impairing innate immunity of the lung. *J Innate Immun* **5,** 251–260 (2013).

62. Vincents, B., Önnerfjord, P., Gruca, M., Potempa, J. & Abrahamson, M. Down-regulation of human extracellular cysteine protease inhibitors by the secreted staphylococcal cysteine proteases, staphopain A and B. *Biol. Chem.* **388,** 437–446 (2007).

63. Borrelli, F. *et al.* Inhibitory effects of bromelain, a cysteine protease derived from pineapple stem (*Ananas comosus*), on intestinal motility in mice. *Neurogastroenterology & Motility* **23,** 745–e331 (2011).

64. Kosugi, A., Amano, Y., Murashima, K. & Doi, R. H. Hydrophilic domains of scaffolding protein CbpA promote glycosyl hydrolase activity and localization of cellulosomes to the cell surface of *Clostridium cellulovorans*. *Journal of Bacteriology* **186,** 6351–6359 (2004).

65. Mosbah, A. *et al.* Solution structure of the module X2_1 of unknown function of the cellulosomal scaffolding protein CipC of *Clostridium cellulolyticum*. *J. Mol. Biol.* **304,** 201–217 (2000).

66. Holdeman, L. V. & Moore, W. New genus, *Coprococcus*, twelve new species, and emended descriptions of four previously described species of bacteria from human feces. *International Journal of Systematic Bacteriology* **24,** 260–277 (1974).

67. Van Der Toorn, J. & Van Gylswyk, N. O. Xylan-digesting bacteria from the rumen of sheep fed maize straw diets. *Microbiology* (1985).

68. Xu, C. *et al.* Structure and regulation of the cellulose degradome in *Clostridium cellulolyticum*. *Biotechnol Biofuels* **6,** 73 (2013).

69. Fasano, A. *et al. Vibrio cholerae* produces a second enterotoxin, which affects intestinal tight junctions. *Proc. Natl. Acad. Sci. U.S.A.* **88,** 5242–5246 (1991).

70. Baudry, B., Fasano, A., Ketley, J. & Kaper, J. B. Cloning of a gene (*zot*) encoding a new toxin produced by *Vibrio cholerae*. *Infection and Immunity* **60,** 428–434 (1992).

71. Watts, T. *et al.* Role of the intestinal tight junction modulator zonulin in the pathogenesis of type I diabetes in BB diabetic-prone rats. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 2916–2921 (2005).

72. Secondulfo, M. *et al.* Ultrastructural mucosal alterations and increased intestinal permeability in non-celiac, type I diabetic patients. *Digestive and Liver Disease* **36,** 35–45 (2004).

73. Sapone, A. *et al.* Zonulin upregulation is associated with increased gut permeability in subjects with type 1 diabetes and their relatives. *Diabetes* **55,** 1443–1449 (2006).

74. Schmidt, E., Kelly, S. M. & van der Walle, C. F. Tight junction modulation and biochemical characterisation of the zonula occludens toxin C-and N-termini. *FEBS Lett.* **581,** 2974–2980 (2007).

75. Di Pierro, M. *et al.* Zonula occludens toxin structure-function analysis. Identification of the fragment biologically active on tight junctions and of the zonulin receptor binding domain. *J. Biol. Chem.* **276,** 19160–19165 (2001).

76. Kalischuk, L. D. & Inglis, G. D. Comparative genotypic and pathogenic examination of Campylobacter concisus isolates from diarrheic and non-diarrheic humans. *BMC Microbiol* **11,** 53 (2011).

77. Zhang, L. *et al. Campylobacter concisus* and inflammatory bowel disease. *World J. Gastroenterol.* **20,** 1259–1267 (2014).

78. Bilo, H. J. *et al.* Catecholamines and blood glucose control in type 1 diabetes. *Diabet. Med.* **8 Spec No,** S108–12 (1991).

79. Luft, D., Maisch, C., Hofmann-Krück, V., Radjaipour, M. & Häring, H. U. Correlates of venous catecholamine concentrations in patients with type 1 diabetes during a cold pressor test. *Clin. Auton. Res.* **10,** 131–137 (2000).

80. Guy, D. A. Differing physiological effects of epinephrine in type 1 diabetes and nondiabetic humans. *AJP: Endocrinology and Metabolism* **288,** E178–E186 (2004).

81. Xu, F. *et al.* Transcriptomic analysis of *Campylobacter jejuni* NCTC 11168 in response to epinephrine and norepinephrine. *Front Microbiol* **6,** (2015).

82. Sandrini, S. M. *et al.* Elucidation of the mechanism by which catecholamine stress hormones liberate iron from the innate immune defense proteins transferrin and lactoferrin. *Journal of Bacteriology* **192,** 587–594 (2010).

83. Takenoya, M., Nikolakakis, K. & Sagermann, M. Crystallographic insights into the pore structures and mechanisms of the EutL and EutM shell proteins of the ethanolamine-utilizing microcompartment of *Escherichia coli. Journal of Bacteriology* **192,** 6056–6063 (2010).

84. Penrod, J. T. & Roth, J. R. Conserving a volatile metabolite: a role for carboxysome-like organelles in *Salmonella enterica*. *Journal of Bacteriology* **188,** 2865–2874 (2006).

85. Thiennimitr, P. *et al.* Intestinal inflammation allows Salmonella to use ethanolamine to compete with the microbiota. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 17480–17485 (2011).

86. Winter, S. E. *et al.* Gut inflammation provides a respiratory electron acceptor for Salmonella. *Nature* **467,** 426–429 (2010).

87. Wiley, J. W. & Pietropaolo, M. Autoimmune pancreatitis: the emerging role of serologic biomarkers. *Diabetes* **58,** 520–522 (2009).

88. Legrand, D. *et al.* Lactoferrin structure and functions. *Adv. Exp. Med. Biol.* **606,** 163–194 (2008).

89. Akiyama, Y. *et al.* A lactoferrin-receptor, intelectin 1, affects uptake, sub-cellular localization and release of immunochemically detectable lactoferrin by intestinal epithelial Caco-2 cells. *Journal of Biochemistry* **154,** 437–448 (2013).

90. French, A. T. *et al.* The expression of intelectin in sheep goblet cells and upregulation by interleukin-4. *Veterinary Immunology and Immunopathology* **120,** 41–46 (2007).

91. Legrand, D., Elass, E., Pierce, A. & Mazurier, J. Lactoferrin and host defence: an overview of its immuno-modulating and anti-inflammatory properties. *Biometals* **17,** 225–229 (2004).

92. Tan, B. K., Adya, R. & Randeva, H. S. Omentin: a novel link between inflammation, diabesity, and cardiovascular disease. *Trends Cardiovasc. Med.* **20,** 143–148 (2010).

93. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science* **347,** 1260419–1260419 (2015).

94. Gupta, R., Gigras, P., Mohapatra, H., Goswami, V. K. & Chauhan, B. Microbial α-

amylases: a biotechnological perspective. *Process Biochemistry* **38,** 1599–1616 (2003).

95. Ferrer, M. *et al.* Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure. *Environ. Microbiol.* **15,** 211–226 (2012).

96. Pácal, L. Evidence for altered thiamine metabolism in diabetes: Is there a potential to oppose gluco- and lipotoxicity by rational supplementation? *WJD* **5,** 288 (2014).

97. Engelen, L., Stehouwer, C. D. A. & Schalkwijk, C. G. Current therapeutic interventions in the glycation pathway: evidence from clinical studies. *Diabetes Obes Metab* **15,** 677–689 (2013).

98. Haugen, H. N. The blood concentration of thiamine in diabetes. *Scand J Clin Lab Invest* **16,** 260–266 (1964).

99. Valerio, G. *et al.* Lipophilic thiamine treatment in long-standing insulin-dependent diabetes mellitus. *Acta Diabetol* **36,** 73–76 (1999).

100. Thornalley, P. J. *et al.* High prevalence of low plasma thiamine concentration in diabetes linked to a marker of vascular disease. *Diabetologia* **50,** 2164–2170 (2007).

101. Al-Attas, O. S., Al-Daghri, N. M., Alfadda, A. A., Abd-Alrahman, S. H. & Sabico, S. Blood thiamine and its phosphate esters as measured by high-performance liquid chromatography: levels and associations in diabetes mellitus patients with varying degrees of microalbuminuria. *J. Endocrinol. Invest.* **35,** 951–956 (2012).

102. Rosner, E. A., Strezlecki, K. D., Clark, J. A. & Lieh-Lai, M. Low thiamine levels in children with type 1 diabetes and diabetic ketoacidosis: a pilot study. *Pediatr Crit Care Med* **16,** 114–118 (2015).

103. Said, H. M. Recent advances in transport of water-soluble vitamins in organs of the digestive system: a focus on the colon and the pancreas. *AJP: Gastrointestinal and Liver Physiology* **305,** G601–G610 (2013).

104. Nabokina, S. M. *et al.* Molecular identification and functional characterization of the human colonic thiamine pyrophosphate transporter. *Journal of Biological Chemistry* **289,** 4405–4416 (2014).