nature microbiology

ARTICLES

PUBLISHED: 10 OCTOBER 2016 | VOLUME: 2 | ARTICLE NUMBER: 16180

OPEN

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[1,3,4], Angela Hogan[2], Carine de Beaufort[1,5] and Paul Wilmes[1]*

The gastrointestinal microbiome is a complex ecosystem with functions that shape human health. Studying the relationship between taxonomic alterations and functional repercussions linked to disease remains challenging. Here, we present an integrative approach to resolve the taxonomic and functional attributes of gastrointestinal microbiota at the metagenomic, metatranscriptomic and metaproteomic levels. We apply our methods to samples from four families with multiple cases of type 1 diabetes mellitus (T1DM). Analysis of intra- and inter-individual variation demonstrates that family membership has a pronounced effect on the structural and functional composition of the gastrointestinal microbiome. In the context of T1DM, consistent taxonomic differences were absent across families, but certain human exocrine pancreatic proteins were found at lower levels. The associated microbial functional signatures were linked to metabolic traits in distinct taxa. The methodologies and results provide a foundation for future large-scale integrated multi-omic analyses of the gastrointestinal microbiome in the context of host–microbe interactions in human health and disease.

Metagenomic studies of faecal samples have provided deep insights into the structure and functional potential of microbial communities within the human gastrointestinal tract[1,2], revealing their inter-individual variability[3] and apparent intra-individual stability[4,5]. Case–control studies have associated microbial traits in individuals with patterns of diseases, thereby implicating gastrointestinal microbial communities in a range of human disorders[6–10]. Metagenomics is able to resolve differences in the functional potential of microbial communities and has revealed a surprisingly stable set of core functions, even though gastrointestinal community structures display great inter-individual variation[2].

These deep insights notwithstanding, metagenomics offers little information on which microbial traits are actually contributed to human physiology. More specifically, previous studies analysing functional 'omes', such as the metatranscriptomes or the metaproteomes, in gastrointestinal samples[9,11–17] have found that functional genes predicted from metagenomes are not necessarily expressed[9,11–13]. Furthermore, as the functional omes display higher variability and sensitivity to perturbation[9,11–13], these variations may better reflect disease-related changes in host–microbiome interactions[14,15]. For this reason, the integration of metagenomic, metatranscriptomic and metaproteomic data is expected to yield important insights into the human microbiome[18], as it has for environmental microbial consortia[19,20].

Families play an important role in the formation of the microbiota of individuals[3,21,22]. More specifically, greater similarities have been observed between the microbiota of family members and cohabiting partners than between unrelated individuals[23]. Furthermore, the microbiota of siblings have been found to be closer to each other than to their parents'[22]. In the present study, healthy family members were used as a stringent control group in type 1 diabetes mellitus (T1DM), a disease supposedly influenced by an interplay of genetics and environment.

The worldwide increase in the incidence of T1DM (ref. 24), especially in individuals with a low genetic risk[25], suggests that environmental factors may trigger the autoimmune destruction of insulin-producing beta cells in the endocrine pancreas that causes T1DM. Gastrointestinal agents have been implicated in this process due to the immune-modulatory role of the gastrointestinal tract[26,27] and the potential involvement of the pancreatic ducts in the inflammation that culminates in beta-cell destruction[28]. Case–control studies of the microbial community structure during[29–32], immediately after[33,34] and before[31,35] seroconversion, have tried but failed to identify generalizable disease-causing or biomarker organism signatures. The present study does not focus on the aetiology, but examines a relatively small cohort of individuals living with T1DM to determine whether there are effects that are apparent in the structure and function of the gastrointestinal microbiome linked to T1DM.

In this Article, we expand our recently developed methodologies for integrated multi-omic analyses of microbial consortia[36,37] to systematically explore the variability of the different omes and their relationships in the context of a case study of familial T1DM. Our study allowed us to link the expression of disease-associated microbial functions to distinct taxa, which demonstrates the necessity for integrated multi-omic analyses of microbiota in the context of human disease research.

[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, Luxembourg. [3]Department of Internal Medicine II, Saarland University Medical Center, 66421 Homburg, Germany. [4]Centre Hospitalier Emile Mayrisch, Rue Emile Mayrisch, 4240 Esch-sur-Alzette, Luxembourg. [5]Clinique Pédiatrique – Centre Hospitalier de Luxembourg, 4 rue Nicolas Ernest Barblé, 1210 Luxembourg. *e-mail: anna.buschart@uni.lu; paul.wilmes@uni.lu

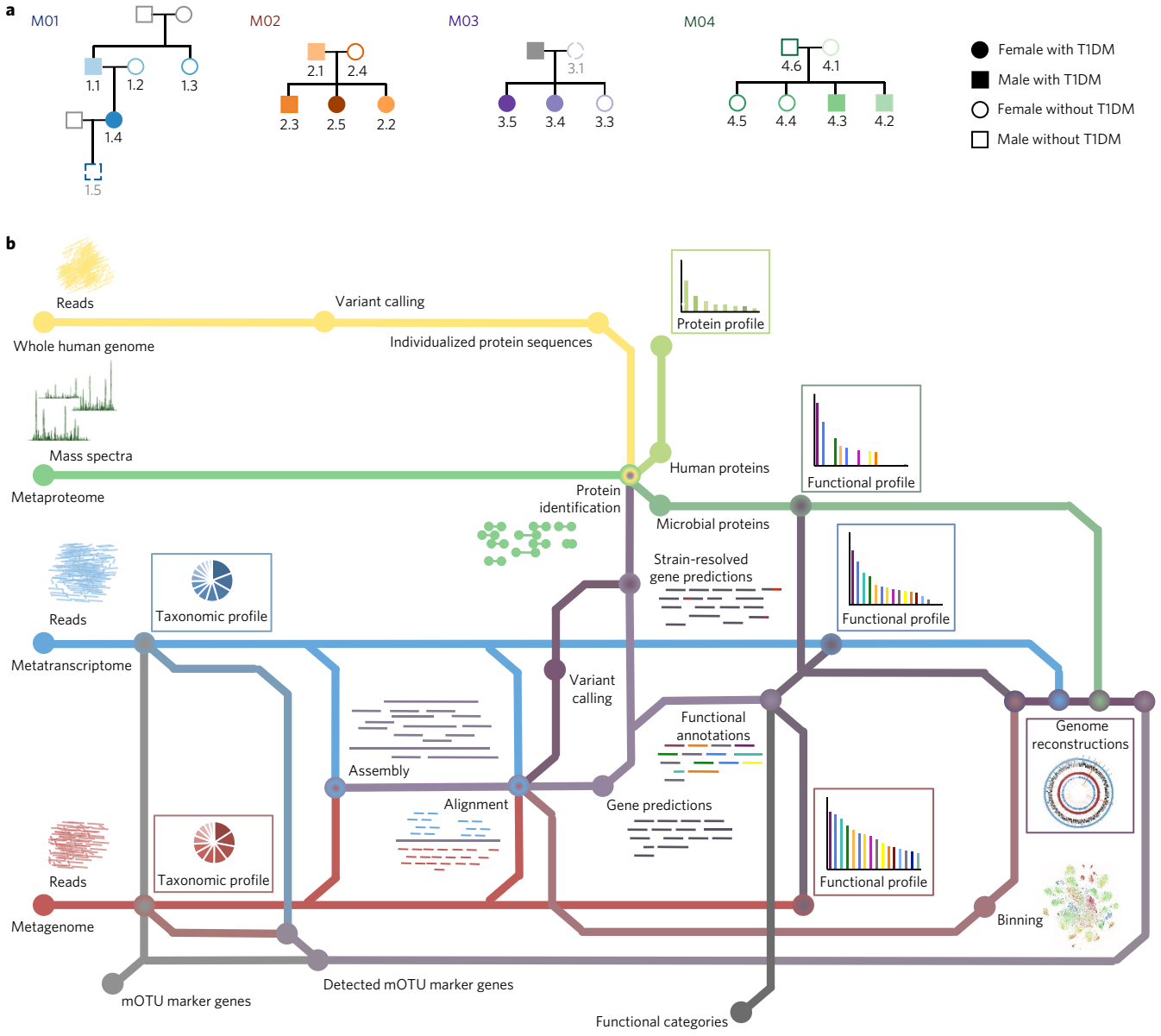

Figure 1 | Overview of the cohort and framework for integrated analyses. a, Pedigrees of participant families. Individuals who donated samples are represented by coloured symbols; these colours are reflected in all subsequent figures. Individuals with T1DM are represented by filled symbols. Samples of the two individuals denoted in grey letters and represented by symbols with dashed lines (M01.5 and M03.1) were only subjected to metagenomic analyses. **b**, Schematic representation of the per-sample analysis workflow. Resulting data sets framed with coloured boxes were used for cross-sample comparisons.

Study cohort and microbial community structures

The study cohort consisted of 20 individuals from four families of at least two generations presenting at least two cases of T1DM (Fig. 1a). T1DM was in all cases diagnosed for more than five years and successfully controlled by insulin replacement therapy, and the auto-antibody status of the healthy individuals did not indicate immediate risk of developing T1DM (for detailed descriptions see Supplementary Section 1 and Supplementary Table 1).

To obtain overviews of the microbial community structures in 53 well-preserved faecal samples collected over a period of two to four months, we calculated relative abundances of metagenomic operational taxonomic units (mOTUs)[38] from 3.9 ± 0.1 Gbp (mean ± standard deviation (s.d.)) of metagenomic data per sample (Supplementary Section 2, Supplementary Fig. 1 and Supplementary Table 2). In most cases, mOTU-based taxonomic profiles were stable over time within individuals. Given that family members share environment, nutrition and genetic background, families formed more distinct groups than unrelated individuals, and samples did not cluster according to T1DM status (Supplementary Fig. 2). Similar to earlier analyses[38], 43 ± 14% of the detected microorganisms in each sample belonged to mOTUs without a sequenced isolate genome. The relative abundance of single mOTUs not assigned to any phylum reached up to 9% (Supplementary Section 2). These results indicate that analytical approaches purely based on isolate reference genomes would not be able to capture the taxonomic and functional diversity in these microbial communities.

Integrated multi-omics

Given the considerable proportion of taxa not represented by sequenced isolate genomes in the mOTU analyses, we devised a reference genome-independent workflow for the integration of metagenomic, metatranscriptomic and metaproteomic data (Fig. 1b). A total of 8.0 ± 1.0 Gbp (mean ± s.d.) of metatranscriptomic sequencing data and 185,000 ± 13,000 peptide mass spectra were obtained

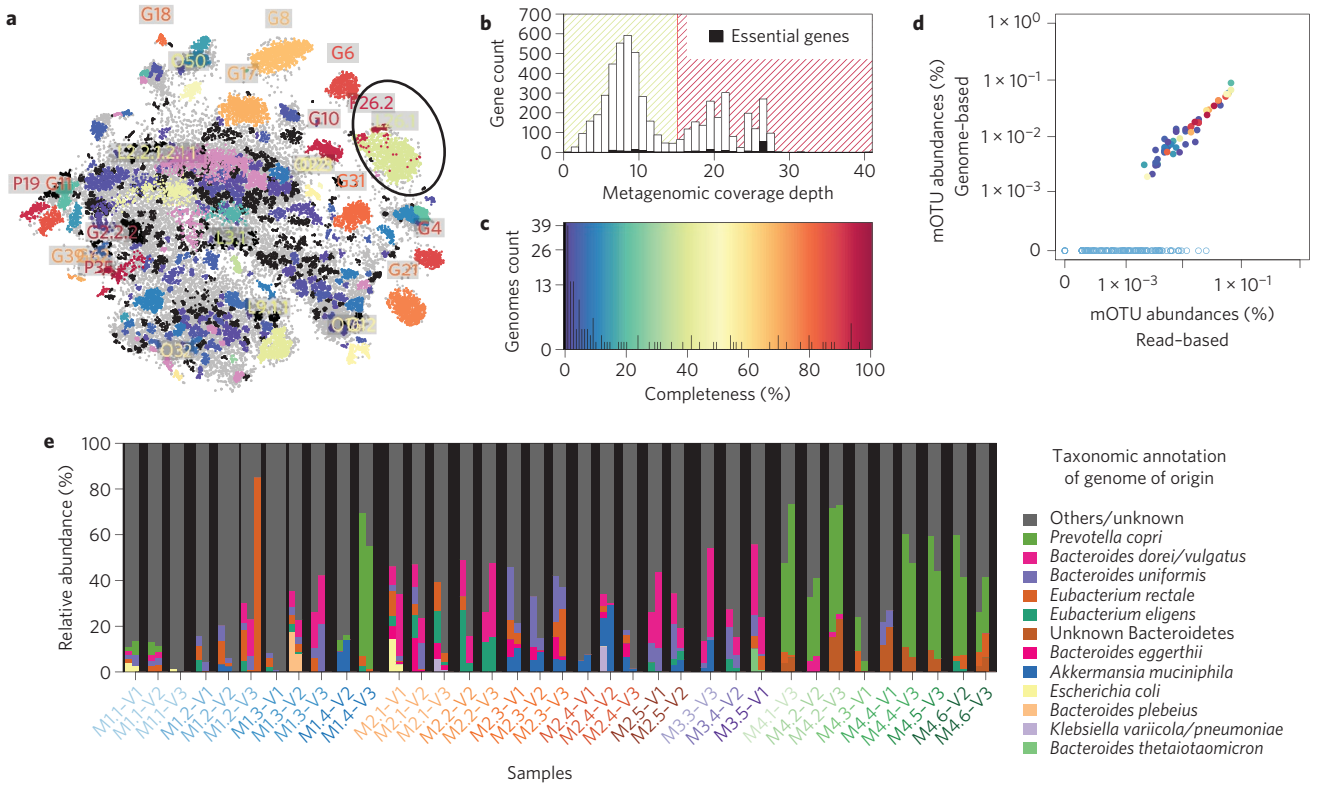

Figure 2 | Automated binning of contigs and population-level genome characteristics. a, Example (sample M01.2-V1) of a two-dimensional visualization of pentamer frequency-based signatures of assembled contigs; distinct bins are highlighted based on completeness (only bins with at least 29% of essential genes with fewer than 20% multiple copies are labelled). **b**, Metagenomic coverage of genes in the cluster encircled in **a**, which was subdivided in the clustering process based on metagenomic depths of coverage of essential genes. The red dashed line indicates the determined cutoff between bins. **c**, Colour key for subpanels **a** and **d**, with a histogram (black lines) of completeness of all binned population-level genomes with at least one essential gene. **d**, Metagenomic mOTU abundances calculated from mapping reads against a collection of phylogenetic marker genes compared to a calculation based on metagenomic depth of coverage of binned population-level genomes annotated with the taxonomy of the most similar phylogenetic marker genes. Dots representing mOTUs with reconstructed genomes are coloured based on the genomes' completenesses; others are indicated as blue, open circles. **e**, Example of genes of a common functional category (here K03149, thiazole synthase) linked to population-level genomes with taxonomic annotations. The taxonomic annotation of the population-level genomes recruiting >10% of the overall metagenomic reads mapping to genes with the function of interest are indicated; all other genes are gathered in 'others/unknown'. Groups of two or three bars represent one sample (sample names are indicated in the colour scheme defined in Fig. 1a; visit (V)1, first sample; V2, second sample; V3, third sample), with the first bar representing the origins of the metagenomic reads, the middle bar representing the origins of the metatranscriptomic reads, and the (optional) last bar representing the origins of the uniquely identified proteins.

in addition to metagenomic data from biomolecular fractions of 36 of the faecal samples collected from 18 of the individuals (Supplementary Table 3). We approached the multi-omic data integration using a *de novo* genomic assembly strategy, which allowed us to identify differences in encoded and expressed microbial functions in the context of intra-individual variation, family membership and T1DM. Our workflow, which involves the co-assembly of metagenomic and metatranscriptomic reads, resulted in longer total assembled genomic contigs (287 ± 43 Mbp), longer maximal contig lengths (310 ± 170 kbp; N50 (the largest contig length such that 50% of all assembled nucleotides are contained in contigs of at least this size): 660 ± 130 bp) and enhanced read usage (88 ± 4% and 88 ± 2% of the metagenomic and metatranscriptomic reads mapping back to the assembly, respectively) when compared to assemblies based on single omes (Supplementary Section 3, Supplementary Fig. 3 and Supplementary Table 3). Annotations with functional categories (KEGG orthologous groups, MetaCyc and Swiss-Prot enzymes, Pfam or TIGR-Pfam domains) were obtained for 489,000 ± 75,000 predicted open reading frames (ORFs; representing 82 and 66% of the metagenomic and metatranscriptomic reads, respectively; Supplementary Section 3, Supplementary Figs 4 and 5 and Supplementary Table 3).

For protein identification from the metaproteomic data, we assembled sample-specific search databases using the microbial ORFs of the co-assemblies, including strain-resolved variants. Given that a significant proportion of human proteins are typically present in faecal metaproteomes and that these may reflect host–microbiome interactions[11], we also sequenced the whole genomes of study participants (Supplementary Table 3) and added allele-specific individualized human protein sequences to the proteomic search databases. This allowed us to identify 2,400 ± 1,600 microbial and 200 ± 70 human protein groups per sample (920 ± 570 microbial and 90 ± 30 human proteins uniquely identified; Supplementary Section 4 and Supplementary Table 3).

Genomic context and taxonomy of expressed genes

Bacteria accounted for the majority of metagenomic, metatranscriptomic and metaproteomic information (Supplementary Fig. 6). To link the bacterial and archaeal genes to their genomic context and thereby allow for the integration of functional data from multiple omic levels (Fig. 1b), assembled contigs were binned using an algorithm that makes use of multiple features, including genomic signatures, metagenomic coverage information and the presence of single-copy essential genes (Supplementary Section 5 and

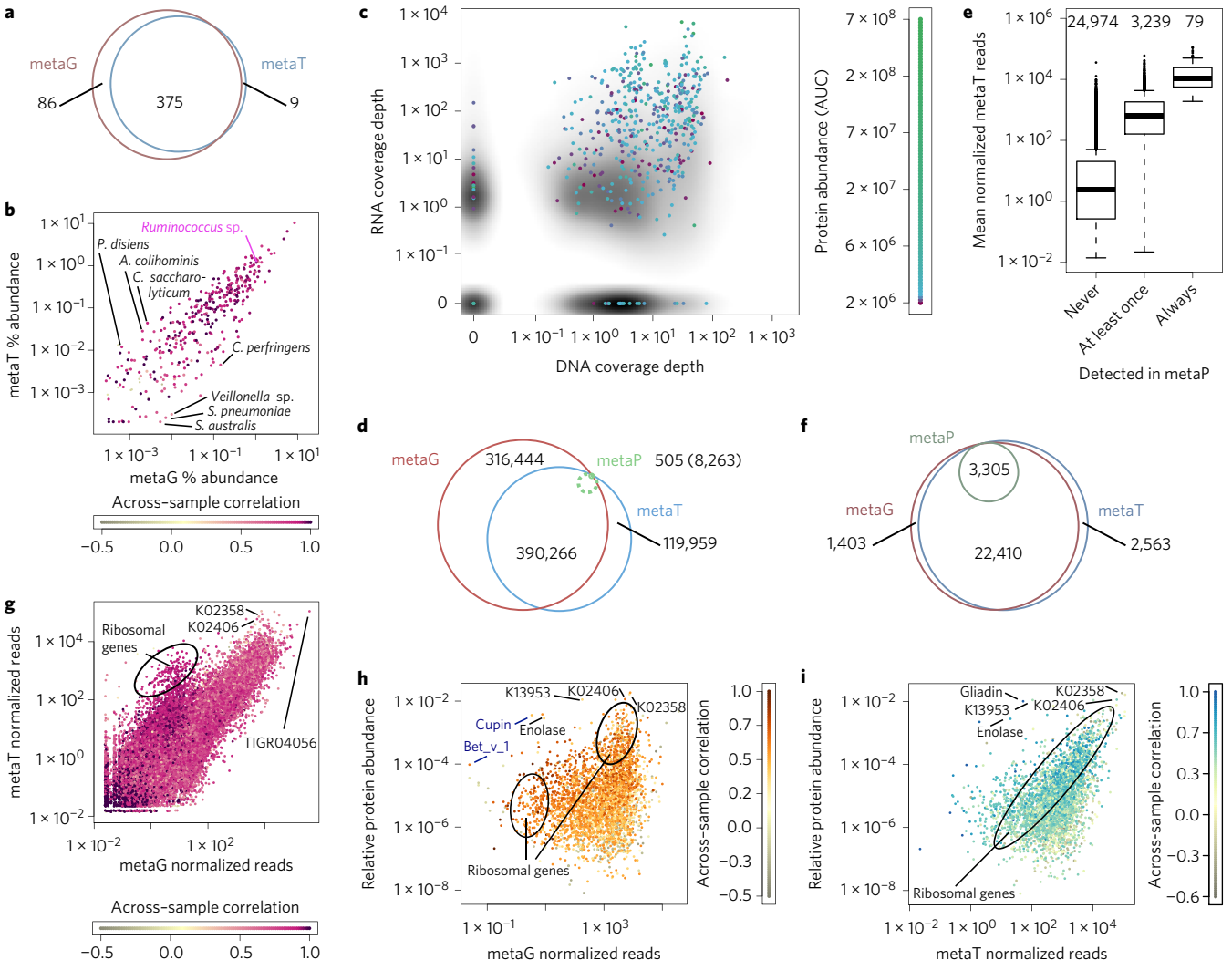

Figure 3 | Relationships between the different omic levels. a, Venn diagram displaying the total number of mOTUs in the metagenomic and metatranscriptomic data sets. **b**, Comparison of mean relative abundances of mOTUs inferred from metagenomic and metatranscriptomic reads in all samples. The colour shading of the points indicates correlation across the analysed samples (legend below the plot). mOTUs with high and low relative activities are highlighted, in addition to a *Ruminococcus* sp. (magenta), which, despite high relative abundances, showed an overall low correlation between the metagenomic and metatranscriptomic relative mOTU abundances across samples (Spearman's $\rho = 0.17$). **c**, Distribution densities of metagenomic and metatranscriptomic depths of coverage for all predicted ORFs in one sample (M01.2-V1) with the uniquely identified proteins and their relative abundances (area under the ion chromatography curve) visualized by points on a purple-to-green scale (legend on the right). **d**, Venn diagram displaying the number of predicted ORFs in sample M01.2-V1 detected on the metagenomic, metatranscriptomic and metaproteomic levels; the circle with the dashed line and the number in parentheses represent the proteins that were not uniquely identifiable based on the masses of the peptides, which are represented by protein groups. **e**, Metatranscriptomic coverage of functions, grouped by frequency of detection in the metaproteome. Numbers on top of the figures indicate the number of functions in each group. **f**, Venn diagram displaying the total number of functional categories detected in the metagenomic, metatranscriptomic and metaproteomic data sets. **g–i**, Comparison of functional profiles in the metagenomic and metatranscriptomic data sets (**g**), metagenomic and metaproteomic data sets (**h**) and metatranscriptomic and metaproteomic data sets (**i**) (mean of all samples). The colour shading of the points indicates the correlations of the relative metagenomic and metatranscriptomic levels (**g**), metagenomic and metaproteomic levels (**h**) and metatranscriptomic and metaproteomic levels (**i**) of each functional category across the analysed samples. Highlighted functions: TIGR04056: highest mean proportion of mapping metagenomic and metatranscriptomic reads; K13953, enolase (microbial) and cupin, gliadin, Bet_v_1 (Bet v 1 allergen family; plant proteins): high representation in the metaproteome; K02358 (elongation factor Tu) and K02406 (flagellin): most stably expressed functions in the metatranscriptomes and metaproteomes; highly active clusters enriched with ribosomal genes (Supplementary Section 11).

Supplementary Table 4). Contig bins (Fig. 2a–c), which represent (partial) population-level genomes, were analysed in terms of their functional annotations as well as protein-coding phylogenetic markers. Taxonomic profiles from the binning were consistent with the assembly-independent mOTU-based analysis results (Fig. 2d, Supplementary Figs 2 and 7 and Supplementary Section 6). Furthermore, we were able to use the information from the binned genomic reconstructions to trace specific functions of interest to different taxa encoding or expressing them (Fig. 2e, Supplementary Fig. 8 and Supplementary Section 7).

In addition, non-human eukaryotic genes, transcripts and proteins were detected that originated from food-borne organisms (plants and fungi) and their viruses, as well as active gastrointestinal eukaryotes (fungi and protists; non-human eukaryotic genes accounted for $0.07 \pm 0.05\%$ of the metagenomic reads, $0.8 \pm 1\%$ of the metatranscriptomic reads and $2.6 \pm 3.1\%$ of the identified

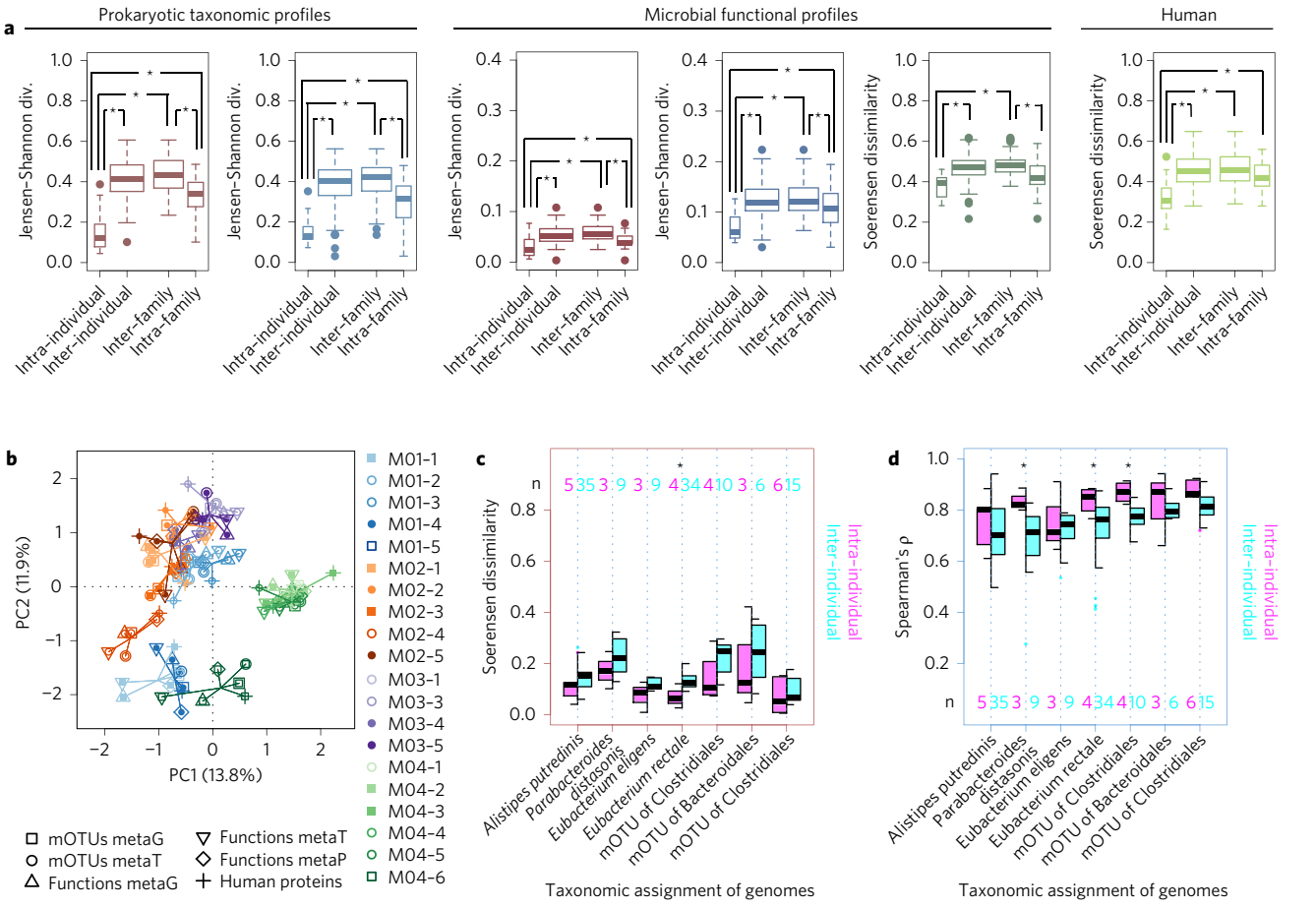

Figure 4 | Comparison of intra-individual to inter-individual and inter-familial distances between microbial profiles at the different omic levels. **a**, Comparisons of Jensen–Shannon divergences or Soerensen dissimilarities between all possible combinations of individuals, in terms of taxonomic profiles based on metagenomic (red) and metatranscriptomic (blue) data, functional profiles based on metagenomic (red), metatranscriptomic (blue) and metaproteomic (olive green) functional profiles, as well as the human proteins (light green). **b**, Multiple co-inertia analysis of per-individual medians of inferred metagenomic and metatranscriptomic abundances of mOTUs, metagenomic, metatranscriptomic and metaproteomic abundances of functional categories and human protein abundances. The colour/symbol scheme denoting the individuals is identical to that used in the other figures as defined in Fig. 1a. **c**, Comparison of the dissimilarities of functional potential of closely related population-level genomes reconstructed from different samples of the same individual and genomes of related organisms reconstructed from samples of different individuals. Over the whole set of taxa, intra- and inter-individual dissimilarity indices were significantly different ($P < 0.05$, Wilcoxon signed rank test). **d**, Comparison of correlations between expression profiles of closely related population-level genomes reconstructed from different samples of the same individual and expression profiles of genomes of the same organisms reconstructed from samples of different individuals. Spearman's correlation coefficients of the functional annotations present in all analysed genomes are displayed. Over the whole set of taxa, intra- and inter-individual correlations were significantly different ($P < 0.05$, Wilcoxon signed rank test). In **c** and **d**, only genomes with >67% essential unique genes were evaluated. The mOTUs with reconstructed genomes allowing intra- versus inter-individual comparison of at least three individuals are displayed; upper Clostridiales mOTU: mOTU linkage group 126, Bacteroidales: mOTU linkage group 135, lower Clostridiales mOTU: mOTU linkage group 310. Numbers of compared genomes (n) are shown as pink and blue numbers. In **a**, **c** and **d**, asterisks indicate statistically significant direct comparisons ($P < 0.05$, Wilcoxon rank sum test). Boxes span the first to third quartiles, the central thick bars represent the medians, whiskers extend to 1.5 times the interquartile ranges, and data points outside these ranges are indicated as outlier points.

proteins, Supplementary Section 8). Viral biomolecules represented less than 0.1% in terms of predicted ORFs as well as mappable reads and identified proteins (Supplementary Section 9).

Metagenomic data and functional omic data

To assess how metagenomic taxonomic profiles might be reflected at the metatranscriptomic level, mOTU abundance profiles were calculated using the metatranscriptomic reads as input, in addition to the previous metagenome-based analyses. A similar set of mOTUs was detected on the metagenomic and metatranscriptomic levels (Fig. 3a), and their abundance profiles correlated strongly (Spearman's $\rho = 0.81 \pm 0.15$; $P < 2.2 \times 10^{-16}$; Fig. 3b and Supplementary Section 10). When comparing the relative mOTU abundances at the metagenomic and metatranscriptomic levels, apparent differences in the inferred activities of different mOTUs in the collected faecal samples were related to their prevalence and activity at distinct sites within the gastrointestinal tract[9], where the most transcriptionally active organisms are known residents of the lower gastrointestinal tract, and organisms with low activity are associated with the oral cavity and upper respiratory tract (Fig. 3b and Supplementary Section 10).

Overall, although a strong correlation was observed between metagenomic and metatranscriptomic reads mapping to the phylogenetic marker genes (which are expected to be constitutively expressed), a wide range of expression levels was observed for the other ORFs encoded by the reconstructed genomes (Supplementary Fig. 9).

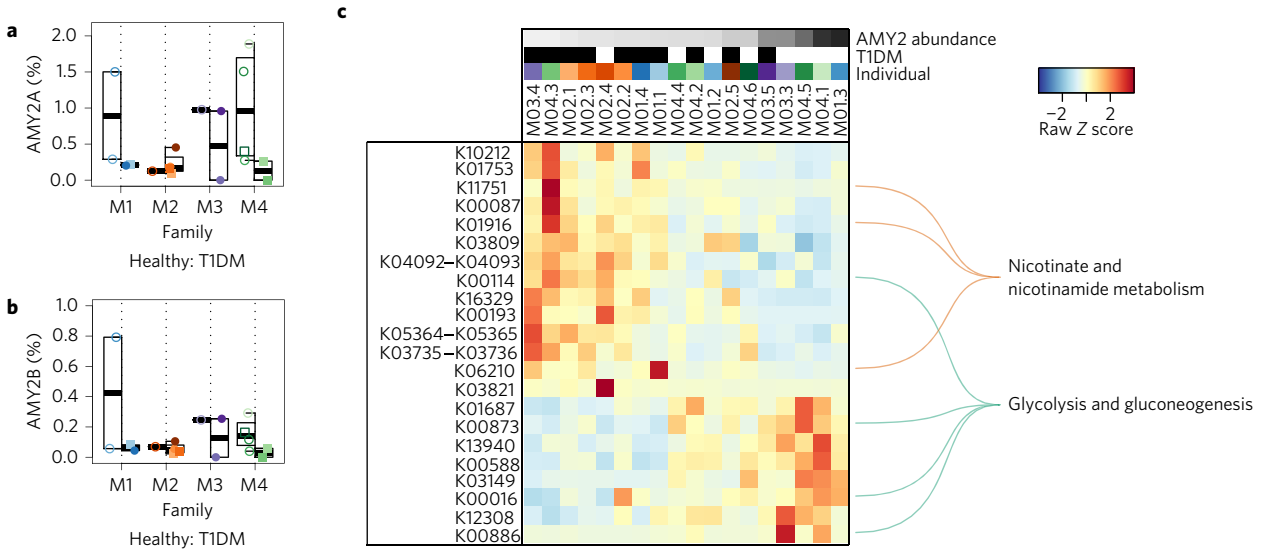

Figure 5 | Abundances of exocrine pancreatic proteins in the faecal samples of individuals with T1DM and correlating transcripts for metabolic enzymes. **a**,**b**, Relative protein abundances of human AMY2A (**a**) and AMY2B (**b**). Boxes span the first to third quartiles, the central thick bars represent the medians, and whiskers extend to 1.5 times the interquartile ranges. All data points are represented. Boxes to the left of the dashed lines represent healthy individuals and boxes to the right represent individuals with T1DM belonging to the respective families. **c**, Correlation of microbial transcript abundances of KOs from a reconstructed metabolic network with AMY2 protein abundances. Heatmaps of normalized microbial transcript abundances of KOs are scaled as indicated in the colour key. Pathways with at least three enzymes and significant enrichment are highlighted with orange (nicotinate and nicotinamide metabolism) and blue (glycolysis and gluconeogenesis) lines (FDR-adjusted $P < 0.05$). At the top of the heatmap, higher relative abundances of AMY2 are indicated in darker grey tones, and individuals with T1DM are indicated by a black box. Individual colours follow the same colour scheme used throughout the figures (Fig. 1a; see Supplementary Table 1a for the total number of samples per individual).

This explains the overall weak correlation between metagenomic and metatranscriptomic depths of coverage at the level of all ORFs (Spearman's $\rho = 0.06 \pm 0.09$, Fig. 3c). Within the metaproteome, only a limited proportion of the predicted ORFs were detectable (Fig. 3c,d and Supplementary Table 3). However, ORFs encoding identified proteins typically exhibited higher metatranscriptomic coverage (Kruskal–Wallis test $P < 2.2 \times 10^{-16}$, Fig. 3e). Stronger variations in the metatranscriptomic levels of non-housekeeping functions were observed, most probably due to these being differentially regulated or expressed by different taxa (Supplementary Fig. 9). This may also explain why, despite a similar overlap of detected functional categories (Fig. 3f), community-wide metagenomic and metatranscriptomic functional profiles were less correlated (Spearman's $\rho = 0.41 \pm 0.08$; $P < 2.2 \times 10^{-16}$; Fig. 3g) than the mOTU abundance profiles (Supplementary Section 11). Metagenomic and metaproteomic functional profiles did not concur (gene-wise $\rho = 0.14 \pm 0.08$; functional category-wise $\rho = 0.10 \pm 0.05$; $P < 0.07$, Fig. 3h), but protein abundances tended to correlate with transcript abundances (gene-wise $\rho = 0.33 \pm 0.12$, Supplementary Fig. 9; functional category-wise $\rho = 0.40 \pm 0.16$, $P < 2.2 \times 10^{-16}$, Fig. 3i), indicating a stronger dependence of protein expression on corresponding transcript levels.

Individuality of gut microbiota on all omic levels

The microbial consortia of the gut are known to be among the most stable human-associated microbial communities[2], and sample donors remain identifiable from these based on metagenomic data when compared to data sets from earlier samples[39]. We investigated whether the patterns of individuality are also discernible on the other omic levels. At the metagenomic level, the observed community structures were stable over time in most individuals (Fig. 4a). More specifically, the intra-individual distances between metagenomic mOTU abundance profiles were significantly smaller than the corresponding inter-individual distances, and donors thus remained recognizable based on mOTU abundance profiles (Supplementary Section 12 and Supplementary Fig. 10). Furthermore, significantly greater intra-individual than inter-individual similarities were observed in the metatranscriptome-based taxonomic profiles, in the functional profiles on all omic levels including the metagenome, and even in the human protein profiles (Wilcoxon rank sum tests between Jensen–Shannon or Soerensen distances: $P < 0.05$, Fig. 4a). Intra-individual variability was greater in the metatranscriptome and metaproteome, thereby hampering recognition of samples from the same donor (Supplementary Section 12). Taken together, these results indicate that the stability of the community structure within the same individual[2] is reflected at the functional levels, which nonetheless display a higher level of plasticity. Consequently, if the gastrointestinal microbiome exhibits changes in individuals with T1DM, signatures may be reflected across the different omes.

Family resemblances are apparent at all omic levels

To place the observed individual microbiome signatures into family contexts, we compared intra- and inter-family distances of community-level taxonomic and functional profiles. Smaller intra- than inter-family distances were observed between community structures (Fig. 4a). To assess whether the observed family-specific traits were also conserved on the other omic levels, multiple co-inertia analysis was performed for taxonomic profiles based on the metagenomic and metatranscriptomic data, for functional profiles on all omic levels, as well as for the human proteome. A remarkable level of similarity among all six levels was observed, with particularly tight clusters observable in several families (in particular in family M04; Fig. 4b). Accordingly, intra-family distances in terms of taxonomic and functional profiles were also smaller than inter-family distances (Fig. 4a, Supplementary Section 13 and Supplementary Fig. 11).

Functional individuality of discrete genomes

In addition to community stability, the persistence of individual-specific strains has been documented[4,5]. To explore whether strain specificity also translates to the functional level, we analysed the

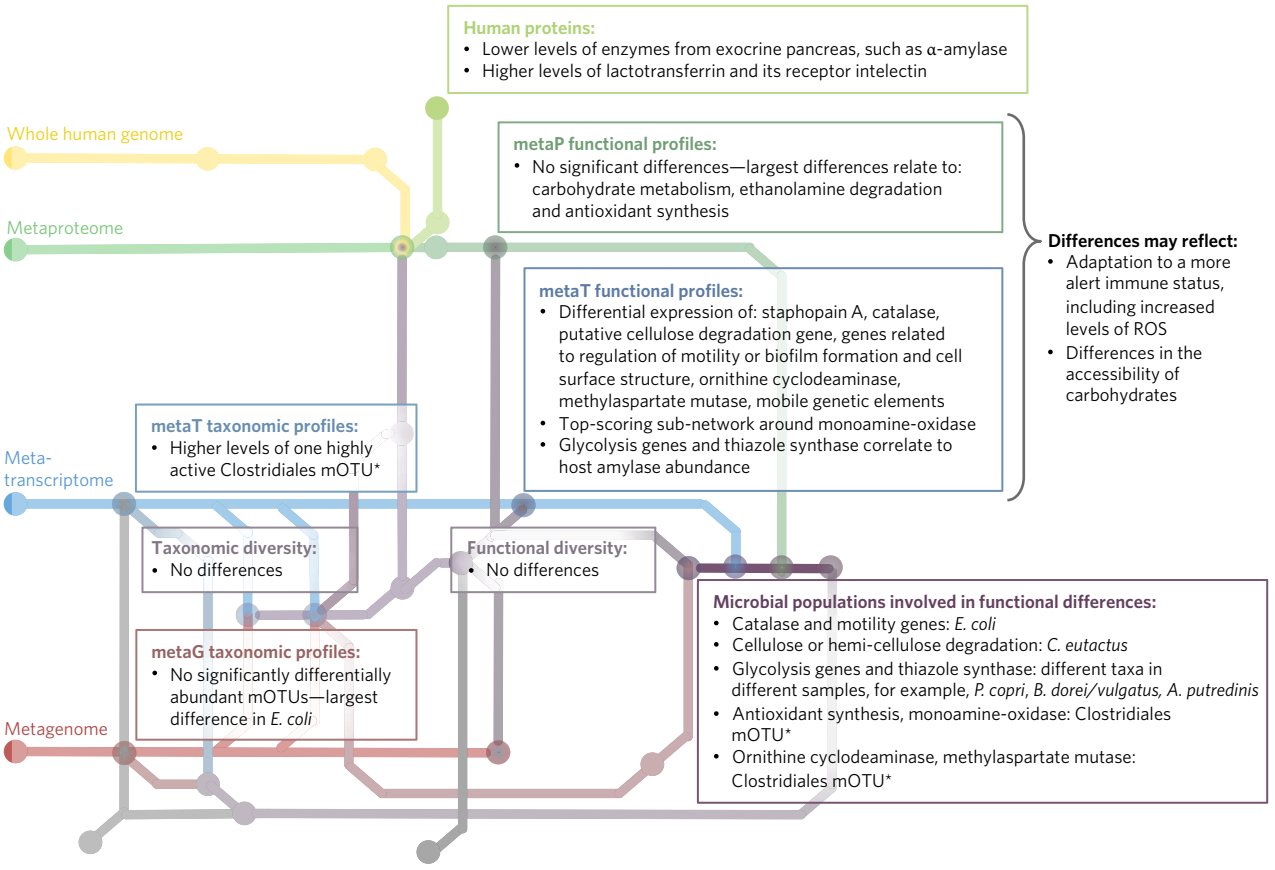

Figure 6 | Multi-omic characteristics of the gastrointestinal microbiome in individuals with T1DM compared to healthy relatives. *The mentioned Clostridiales mOTUs are not identical.

encoded functional potential and corresponding expression profiles of closely related reconstructed genomes recovered from different samples. Our results revealed that the functional potentials of related genomes reconstructed from samples of the same donor were significantly more similar than samples from different donors (Wilcoxon signed rank test between Soerensen dissimilarity indices of all tested taxa: $P < 0.05$, Fig. 4c, Supplementary Section 14 and Supplementary Fig. 10). Furthermore, the expression profiles of these genomes displayed stronger intra-individual than inter-individual correlations (Wilcoxon signed rank test between Spearman's correlation coefficients of all tested taxa: $P < 0.05$, Fig. 4d, Supplementary Section 14 and Supplementary Fig. 10). Individuality is therefore a characteristic of gastrointestinal whole microbial communities on all functional levels, and is also apparent in the functional potential and expression profiles of specific populations common to different individuals.

Family-independent effects of T1DM

As discussed above, on the level of the observed community structures as well as the metatranscriptomic and metaproteomic functional profiles, large differences between families were identified. Additionally, nutritional records of the days prior to faecal sampling revealed that the diets of family members were very similar. In line with medical recommendations, the individuals with T1DM followed the same diets as their healthy family members, and their diet was no more or less variable or rich (Supplementary Sections 1 and 15, Supplementary Table 1, Supplementary Figs 12 and 13). Family membership thus appears to stratify the cohort more strongly than all other factors known to influence gastrointestinal microbial community composition, such as age or body mass index (Supplementary Table 1 and Supplementary Fig. 14). Consequently, we analysed T1DM-specific differences in the microbial taxonomic and functional profiles while accounting for family membership.

No differences in taxonomic or functional diversity and richness were observed between individuals with T1DM and their healthy relatives at any omic level (Supplementary Fig. 15). However, we found one highly active mOTU, classified to the order Clostridiales, to be more abundant in the metatranscriptomes of the individuals with T1DM (main effect with false-discovery-rate (FDR)-adjusted $P < 0.05$ and absolute fold change >2; Supplementary Section 16 and Supplementary Fig. 15).

Differential transcript abundances of diverse functions (Supplementary Fig. 16) suggested differences in the availability or microbial use of cellulose and subtle differences in the immune status (genes for reactive oxygen species (ROS) detoxification, cell surface proteins and complement-inactivating staphopain A) of individuals with T1DM (Supplementary Section 17). The latter group of functions was also reflected in the proteins, which exhibited the greatest differences in terms of their abundance (not significant after FDR adjustment; Supplementary Fig. 16 and Supplementary Section 17). T1DM-specific functions could mostly be traced to genomes of taxa whose abundances were not affected by T1DM (Supplementary Figs 18 and 19), which indicates that expression of potentially disease-related community functions may be independent of the abundances of distinct taxa inferred from the metagenomic data alone (Supplementary Section 17).

Human protein excretion in T1DM and microbial functions

Little is known about changes in the function of the gastrointestinal tract with respect to T1DM traits, such as the proposed heightened innate inflammatory state[27]. Among the ten human proteins with the greatest changes in abundance in the faeces of individuals

with T1DM (Supplementary Table 5), we identified two with higher abundance in the individuals with T1DM, namely lactotransferrin (LTF) and the lactotransferrin receptor intelectin (ITLN1), which maintain an important functional position at the crossroads of iron uptake, innate immunity and insulin function[40–43], and whose antimicrobial activity may influence the gastrointestinal microbiota[44] (Supplementary Section 18 and Supplementary Fig. 17).

Furthermore, we found four proteins that are preferentially expressed in the exocrine pancreas[45], α-amylase proteins AMY2A and AMY2B (Fig. 5a,b), carboxypeptidase CPA1 and the less characterized CUZD1 (Supplementary Fig 18), to be less abundant in the stool of individuals with T1DM, potentially reflecting a weakening of exocrine function. To what extent T1DM affects the exocrine pancreas is still a matter of debate[46], and our results suggest that this may be the case at least for certain digestive enzymes (Supplementary Section 15). We analysed whether potential repercussions of the apparent T1DM-related decrease in digestive enzymes (which would probably affect starch availability in the lower intestine[47]) were reflected in the collected samples. Because we were unable to identify taxa whose abundances correlated with the measured abundances of the pancreatic enzymes, we focused on the microbial genes with known metabolic functions (Supplementary Section 18). We uncovered several significant correlations between microbial transcript levels and the relative abundances of the amylases AMY2A and AMY2B (24 with $P < 0.05$ after FDR adjustment; |Spearman's ρ| > 0.75; Supplementary Table 5), including functions related to microbial central carbon metabolism and related cofactors (Fig. 5c), in particular thiamine biosynthesis (Supplementary Section 18 and Supplementary Fig. 20). These functions were transcribed by distinct taxa in the different samples (for example, thiazole synthase, Fig. 2e). Low thiamine blood levels have been linked to complications of T1DM[48] and colonic uptake transporters of thiamine have been described[49]. However, the observed patterns in transcript abundances of thiazole synthase and plasma thiamine levels were not correlated in the small subset of individuals who did not take thiamine supplements, precluding any conclusions (Supplementary Section 18 and Supplementary Fig. 20). The observed patterns thus suggest the need for a detailed follow-up in a dedicated future study involving a larger cohort of families who explicitly do not take thiamine supplements. Furthermore, in light of the T1DM-related findings (Fig. 6), the levels of pancreatic and immune-related proteins should be measured alongside functional microbial profiles in longitudinal pre-diabetes studies to establish the timing of their decrease and elucidate whether they play a role in the changes in the microbiota associated with diabetes development.

Discussion

By dissecting the gastrointestinal microbial community structures resolved using the individual omic data sets, we have shown that community structures are reflected across all omic levels. Nonetheless, the apparent higher complexity and variability exhibited by the functional omic levels necessitates an integrated analysis of the individual omic data sets[18,50] to facilitate a reliable link between specific microbial genes and their genomic context. Linking genes of interest to genomic context has been attempted in other studies of T1DM by correlating abundances of functional traits and organisms[32]. However, as demonstrated, transcript levels cannot be expected to simply depend on the abundances inferred from metagenomic data. We therefore developed the presented framework for genome reconstructions, which served as the basis for the subsequent integration of the multi-omic data.

All omic levels exhibited individuality and family specificity, which has previously been described only for metagenomes[3–5]. The observed individual patterns in relation to secreted human proteins may well be an important factor shaping the gastrointestinal microbiome, culminating in the observed individuality. With regard to T1DM, differences in the relative abundances of certain human pancreatic enzymes in the stool were observed, and the abundances of these proteins correlated with the expression of microbial genes involved in metabolic transformations with possible relevance to T1DM, including thiamine synthesis and glycolysis. Finally, our results clearly indicate that several microbial populations can contribute to functional differences between samples, underpinning the importance of integrated multi-omic analyses. The methods and data presented in this case study will serve to inform the design of future large-scale studies, which should, in addition to multi-omic data, involve the collection of high-quality clinical and nutritional data, to allow the establishment of host–microbe phenotypic associations in the context of human health and disease.

Methods

Ethics. Written informed consent was obtained from all subjects enrolled in the study. This study was approved by the Comité d'Ethique de Recherche (CNER; reference no. 201110/05) and the National Commission for Data Protection in Luxembourg.

Anthropometric, nutritional, clinical data and blood and stool sampling. The study design was an observational study of four purposely selected family groups containing at least two members with T1DM and healthy individuals in two generations or more, from existing patient cohorts from the Centre Hospitalier du Luxembourg. Recruited families were seen three times at intervals of between four and eight weeks to have data and samples collected (Supplementary Table 1). On enrolment, study participant pedigrees were drawn, medical history was collected and a 'Food Frequency Questionnaire' was completed (estimating the foods eaten over the last 12 months). During every visit, anthropometric data were recorded. All participants were weighed and measured using the SECA combined scale and measuring device. Body mass index was calculated as kg m^{-2}. Blood pressure was taken in the sitting position with a SureSigns VM6.

Daily recall questionnaires for food intake in the 24 h preceding the visit were filled in on visits 2 and 3. Nutritional content of the reported food was estimated based on the food frequency or daily recall questionnaires using FETA software[51].

Faecal samples were self-collected at three time points (if bowel movement permitted sampling) and immediately placed on dry ice and subsequently transported and stored without thawing. Blood was collected using standard venepuncture and processed using validated processes[52]. Samples and associated data used in this study were processed and stored at IBBL (Integrated BioBank of Luxembourg) following ISO17025:2005 standards and the International Society for Biological and Environmental Repositories (ISBER) best practices.

Blood analytics. Islet cell antibodies were measured using the test from INOVA Diagnostics. Insulin antibodies, tyrosine phosphatase-related islet antigen 2 antibodies, glutamate decarboxylase 2 antibodies and zinc-transporter 8 antibodies were measured using enzyme-linked immunosorbent assay (ELISA) kits from RSR. Measurements of glucose, glutamate-oxalacetate-transaminase, glutamate-pyruvate-transaminase, γ-glutamyltransferase, cholesterol, low-density cholesterol, high-density cholesterol and triglycerides were performed on a Cobas c501 analyser (Roche) with appropriate reagents (Roche). Quantification of glycated haemoglobin was performed on an Integra 400 analyser (Roche) with a kit from the same company. Insulin and the C-peptide were measured on a Cobas e601 (Roche) with matching kits (Roche). Leptin and pro-insulin were quantified using ELISA kits (Beckman Coulter). Thiamine levels in plasma were measured in technical duplicates using a vitamin B1 ELISA kit (Cloud-Clone) according to the manufacturer's protocol.

Extraction of biomolecular fractions from faecal samples. Biomolecular fractions were extracted from unthawed, frozen faecal subsamples (150 mg) after pretreatment of the weighed subsamples with 1.5 ml RNAlater ICE (LifeTechnologies) overnight. The faeces-RNAlater ICE mixture was homogenized by bead-beating, as previously described[53]. Differential centrifugation and extraction using the All-In-One kit (Norgen Biotek) to recover DNA and proteins were carried out as previously described[53]. DNA fractions were supplemented with DNA extracted from 200 mg subsamples using the MOBIO Power Soil Kit.

RNA was extracted from faecal subsamples (250 mg), which were pretreated with RNAlater ICE (LifeTechnologies) as described above. Homogenization and differential centrifugation were carried out as described previously[53] and the pellet was lysed with MOBIO glass bead tubes. RNA was extracted using the MOBIO PowerMicrobiome RNA Isolation Kit, according to the manufacturer's recommendations. The integrity of isolated RNA fractions was assessed using an Agilent Bioanalyzer2000. Only fractions with an RNA integrity number > 7 were subjected to sequencing.

Metagenomic DNA and RNA sequencing. DNA was sequenced with 101 bp (per paired end) on a HiSeq2000 system (Illumina) by BGI. Libraries with an insert size of 350 bp were constructed from metagenomic DNA by fragmentation by

sonication (Covaris), end-repair, adenylation, adapter ligation and amplification of adapter-ligated DNA fragments using appropriate enzymes (Enzymatics). Library amplification and cluster generation were performed using a TruSeq PE Cluster Kit V3–cBot–HS (Illumina) and flow cells were sequenced to 101 bp per paired end.

Strand-specific cDNA libraries were constructed from rRNA-depleted RNA fractions and sequenced to 150 bp (per paired end) on an Illumina HiSeq2500 by Oxford Gene Technologies. In detail, rRNA was depleted from the RNA fractions using Epibio's Universal Bacteria Ribozero Kit. Stranded cDNA libraries were constructed using the NEBNext Ultra Directional RNA Library Prep Kit. Libraries were sequenced to 150 bp (paired-end) using TruSeq Rapid SBS Kits v1 (Illumina).

Filtering and trimming of metagenomic and metatranscriptomic reads. The *in silico* analysis results presented in this Article were obtained using the high-performance computing facilities of the University of Luxembourg[54]. Metagenomic and metatranscriptomic paired-end reads in fastq format for each individual and each time point were processed separately using the MOCAT pipeline[55]. First, reads were trimmed using the MOCAT trimming and quality filtering step (*MOCAT.pl –rtf*) with the following parameter settings: *readtrimfilter_length_cutoff=40*, *readdtrimfilter_qual_cutoff=20*, *readtrimfilter_use_sanger_scale=yes*, *readtrimfilter_trim_5_prime=yes* and *readtrimfilter_use_precalc_5prime_trimming=no*. This step generated paired- and single-end reads, which were then mapped to the human genome (hg19) using the MOCAT screening step (*MOCAT.pl –s hg19*), which used SOAPaligner v2.21 (ref. 56) and then filtered mapping reads out. Metatranscriptomic reads were then also screened against the human RNA transcripts as defined for hg19 by Ensembl (release 66) to additionally remove reads overlapping splice sites in human transcripts. MOCAT screening was run with the following parameter setting: *screen_length_cutoff=30*, *screen_percent_cutoff=90*, *screen_soap_seed_length=30*, *screen_soap_max_mm=10*, *screen_soap_cmd=-M 4* and *screen_save_format=sam*. This step resulted in two sets of reads in fastq format (human and non-human), each consisting of paired- and single-end reads. To assess the rRNA content of the sequencing libraries, metatranscriptomic reads were independently mapped against the prokaryotic and eukaryotic parts of the SILVA database[57] using Bowtie2 (ref. 58). The non-human files were used for taxonomic analyses as well as assembly (Supplementary Figs 21 and 22).

Read-based taxonomic analyses, mOTU analyses and use of an integrated gene catalogue. mOTU analysis for taxonomic profiling was separately performed for trimmed and filtered metagenomic and metatranscriptomic reads for each sample using the MOCAT pipeline[38,55] and the mOTU.v1.padded reference database (Supplementary Fig. 21), which comprises genes for the ribosome-binding ATPase (YchF; GTP1/OBG family), phenylalanyl-tRNA-synthetase alpha subunit, arginyl-tRNA-, seryl-tRNA-, cysteinyl-tRNA-, leucyl-tRNA- and valyl-tRNA-synthetase, tRNA A37 threonylcarbamoyltransferase (TsaD), and two signal recognition particle GTPases (COG0012, COG0016, COG0018, COG0172, COG0215, COG0495, COG0525, COG0533, COG0541 and COG0552, respectively)[38]. Metagenomic and metatranscriptomic codes to recognize donors were built from mOTU relative abundances and evaluated using Idability[39] with the *–meta_mode* setting.

Taxonomic profiling of reads and contigs was also performed using Kraken[59] with a database comprising 9,464 prokaryotic and 2,326 viral genomes. Metagenomic reads were mapped against the integrated gene catalogue (IGC)[60] and genus-level abundance profiles were calculated as described in the publication of the integrated gene catalogue[60]. These profiles were concatenated with the profiles of 1,267 samples distributed with the IGC. A Jensen–Shannon divergence matrix was calculated using the phyloseq[61] implementation over all profiles, and partitioning around medoids[62] was performed to obtain enterotype annotations for each sample.

Co-assembly of metagenomic and metatranscriptomic reads. Trimmed and filtered metagenomic and metatranscriptomic reads were merged and made non-redundant using *cd-hit-dup* from the CD-HIT software suite[63]. The non-redundant set of non-human paired-end reads was then assembled using a combined metagenomic and metatranscriptomic assembly pipeline (Supplementary Fig. 22).

Reads were first converted from fastq to fasta format using the *fq2fa* script (*fq2fa -merge -filter*) provided by IDBA-UD (ref. 64). IDBA-UD was then run using its pre-error-correction step for correcting reads (*idba_ud –pre_correction*). In the next step, the contigs assembled by IDBA-UD were extended using the unused paired- and single-end reads and the Velvet assembly tool[65] (version 1.2.07). For this, the paired-end reads were mapped onto the previously assembled contigs using SOAPaligner v2.21 (ref. 56; with parameter *-r 2 -M 4 -l 30 -v 10 -p 8 -u unmapped.fa*). The unused single-end reads were combined with the unmapped reads and *cd-hit-dup* from the CD-HIT software suite[63] was used to remove duplicate reads. Velvet was then run over a range of k-mer sizes (27, 31, 35, 39, 43, 47, 51, 55, 59, 63) with the IDBA-UD contigs as long read input (*velveth* parameters: *-long contig.fa*, *velvetg* parameter: *-conserveLong yes*). Afterwards, all newly assembled contig sequences from the different k-mer runs and the original IDBA-UD contigs were concatenated and clustered using *cd-hit* from the CD-HIT software suite (parameter: *-c 0.99*) to remove redundancy. Newbler from MeGAMerge 1.1 (ref. 66) was then used to join and extend the clustered contigs. Minimal contig length was set to 125 bps.

Gene predictions, functional and phylogenetic annotation. Prodigal v2.60 (ref. 67) (parameter: *-p meta*) was used for gene prediction on the final set of contigs. Essential single-copy genes[68] among the protein predictions were detected as described[69]. Marker genes for phylogenetic analysis were predicted using Amphora2 (ref. 70) and fetchMG (ref. 38). rRNA genes were predicted using Barrnap (http://www.vicbioinformatics.com/software.barrnap.shtml). Functional gene annotations were generated using HMMER 3.1 (refs 71,72) against multiple pre-compiled and in-house annotation databases. Existing databases for PFAM (ref. 73) and TIGRFAM (ref. 74) were downloaded from their web pages. Additionally, hidden Markov models (HMMs) were trained for all KEGG orthologous groups (KOs) after aligning the protein sequences using T-Coffee (ref. 75). Enzyme-specific HMMs were generated for MetaCYC (ref. 76) and UNIPROT (ref. 77) using only enzyme annotations with at least three digit Enzyme Commission (EC) numbers, as described previously[78]. The best hit in each HMM set was associated to each gene, if the HMM score was higher than the binary logarithm of the number of target genes (19.6 ± 0.2), in accordance with the recommendations in the HMMer manual.

Determination of metagenomic and metatranscriptomic depths of coverage of contigs and ORFs. Metagenomic and metatranscriptomic paired-end and singleton reads of each sample were mapped against contigs using Bowtie2 (ref. 58) and default settings. Average depths of coverage were calculated using SAMtools (ref. 79) and a custom perl script. Quality-trimmed metagenomic paired-end and singleton reads were mapped against predicted protein- or RNA-coding regions using Bowtie2 with default settings and in addition to depth of coverage, the numbers of mapping reads were counted. Stranded metatranscriptomic paired-end and singleton reads were mapped against predicted protein- or RNA-coding regions using Bowtie2 with *–fr –nofw* settings to obtain coverage depth and read counts for reads mapping in the sense direction.

To obtain whole-community metagenomic or metatranscriptomic functional profiles, sums of all read counts for functionally annotated genes were calculated for every functional category. If two functions were assigned to the same gene with the same likelihood, the number of reads for this gene was divided equally between both functions.

Binning of contigs from metagenomic and metatranscriptomic co-assembly. Two-dimensional maps for every contig ≥1 kbp were generated for each sample as described previously[80]. In short, for every contig ≥1 kbp, pentanucleotide frequencies were computed, which represent the genomic fragments as points in a 512-dimensional space (frequencies of individual pentamers and their reverse complements were summed). This matrix was subsequently transformed by log ratio and centred. Following this, the high-dimensional data were embedded into a two-dimensional space using Barnes–Hut stochastic neighbour embedding and the resulting two-dimensional maps were used for clustering. As rRNA-coding regions have been shown to challenge clustering based on genomic signature[80], 16S/18S and 23S/28S sequences were removed from the contigs prior to the calculation of pentamer frequencies as long as a minimum length of 1 kbp was retained. Two-dimensional maps of contigs were visualized in R. Clusters were binned using an automatic workflow based on DBSCAN (ref. 81) (Supplementary Fig. 23). Briefly, two-dimensional coordinates were clustered using the *dbscan* function in the R package fpc. In the first pass, the parameters were ten minimum neighbourhood points and the Eps value was estimated based on the recommendations of the authors of DBSCAN. Number and multiplicity of essential genes[68,69] were used to judge completeness of clusters. Clusters with multiple copies of the same essential genes were divided further by analysing the metagenomic coverage depth of the essential genes using the Hartigans' diptest as implemented in the R package diptest. If unimodality was rejected, a normal mixture model (implemented in the R package mixtools, ref. 82) was used to find a cutoff to bin the contigs by metagenomic depth of coverage. These two steps were repeated three times on all overcomplete bins, and the number of minimum neighbourhood points was raised by 2 in each step.

Binned population-level genomic complements were linked to mOTUs used in the taxonomic analysis by calling the marker genes using fetchMG (ref. 38) on the predicted ORFs, aligning the predicted marker genes against the database used for mOTU analysis and assigning to each called marker gene the taxonomy of the closest hit.

Human DNA sequencing and analysis. Genomic DNA was isolated from buffy coats of human blood samples using validated processes according to the requirements of Complete Genomics (CG). Whole-genome sequencing (WGS) was performed by CG using their proprietary paired-end nanoarray-based sequencing-by-ligation technology[83]. All sequencing data quality control, mapping and variant calling were carried out by CG as part of their sequencing service using the Standard Sequencing Service pipeline version 2.4. Sequencing reads were mapped against the NCBI build 37. For gene annotations, NCBI build 37.2 (RefSeq) was used. For details on coverage and called variants see Supplementary Table 3. All subsequent analyses were performed using the family WGS analysis pipeline as previously described[84]. In short, as input for the WGS analysis pipeline, we first combined all the variants from all genomes of one family into the union of variants using the CGAtools (CG Analysis Tools) *listvariant* command and CG var-files as input. We used CGAtools version 1.6 as provided by CG (available from http://cgatools.sourceforge.net). Next,

we used the CGAtools *testvariant* command to test each genome for the presence of each variant. Only variants that were called in at least one genome of a family as high-quality calls (VQHIGH) by CG were used for further analysis.

Database generation for metaproteomic analyses. Sample-specific search databases were constructed from individual-specific pseudo-phased human proteins including homo- and heterozygous variants, as well as microbial proteins predicted based on the combined metagenomic and metatranscriptomic assemblies after inclusion of all called variants (Supplementary Fig. 24). In detail, for each human genome, small variants (single nucleotide polymorphisms and small insertions, deletions and substitutions shorter than 200 bp) were annotated using ANNOVAR (ref. 85) (version 2015Mar12) using the NCBI RefSeq release 60 and the Ensembl release 74 genome annotations. An individualized personal reference genome for each sample was generated by incorporating short homozygous variants into the hg19 reference genome. For each RefSeq protein-coding isoform, the new coding (CDS) mRNA sequence was extracted from the personalized reference genome and then translated. To also include heterozygous exonic amino acid changing variants, only exonic variants that were annotated as non-synonymous, stop gain, stop loss, (non-)frameshift insertions, deletions or substitutions were kept. Because no phasing information for the heterozygous variants was available, we generated pseudo-phased proteins by adding one copy per protein including all homo- and heterozygous variants. For each individual genome, a FASTA protein sequence database was then created by combining two FASTA databases: (1) REFSEQ protein sequences from the personalized human genome reference hg19 and (2) protein sequences containing heterozygous variants (proteins were marked if they contained homo- or heterozygous amino acid changes, respectively, or no change at all). Proteins with newly introduced stop codons were cut after the stop codon.

For the metaproteomic databases, variants were called per faecal sample on all predicted ORFs from the alignment results of mapping metagenomic and metatranscriptomic reads against the contigs using Platypus variant caller[86] and vcftools (ref. 87). Variants with low mapping quality (MQ flag) were discarded. Multiple variants called on the same ORF were combined to yield the minimal number of variant proteins without contradictory overlap of variants using a custom perl script. The variant proteins as well as the consensus protein from the assembly were kept. Incomplete predicted proteins (or variant proteins) were removed, if no full tryptic peptide was expected from the prediction.

Protein liquid chromatography (LC) and mass spectrometry. An aliquot of 15 µg isolated protein fraction was reduced (dithiothreitol) and alkylated (iodoacetamide) and the proteins precipitated using the Clean-up Kit (GE Healthcare) according to the manufacturer's instructions. The protein pellets were resuspended at a concentration of 500 µg ml^{-1} in 50 mM ammonium bicarbonate to perform a complete trypsin digestion. Digested protein (3.5 µg) was purified on a Ziptip C18 (Millipore), dried and resuspended in 100 mM ammonium formiate (pH 10) at 333 ng µl^{-1}. Digested protein (3 µg) was injected on a Nano 2D UPLC – Orbitrap MS system (2D-nanoAcquity UPLC (Waters) and Q-Exactive (Thermo)). MPDSMIX (Waters) was spiked in at 150 fmol of ADH (yeast alcohol dehydrogenase) digest per injection. Measurements were performed on technical triplicates.

A 2D LC method with three steps of 180 min was run. The three steps were run on a column at high pH with increasing percentages of acetonitrile, and the eluted peptides were diluted and loaded on a low pH column, where each step consisted of a gradient of 5 min from 99% of A (A = 0.1% formic acid in water; B = acetonitrile) to 93% of A followed by a gradient of 135 min from 93% of A to 65% of A.

Mass spectrometry was performed using a TopN-MSMS method ($N = 12$), with parameters as follows: mass range from 400 to 1,750 m/z, resolution 70,000, AGC target 106 or maximum injection time 200 ms. MS2 parameters for spectrum acquisition were an isolation window of 1.6 m/z, normalized collision energy (NCE) of 25, resolution of 17,500, AGC target of 105 or maximum injection time of 50 ms.

The database searches were performed using Proteome Discoverer (v 1.4, Thermo Scientific) and Sequest HT on the protein predictions. Proteomics measurements and database searches were carried out by the proteomics facility of the Center of Analytical Research and Technology – Groupe Interdisciplinaire de Génoprotéomique Appliquée (CART-GIGA, Liège, Belgium). Peptide spectrum matches with at least a high confidence and a delta Cn better than 0.05 were considered, with fragment mass tolerance of 0.02 Da and precursor mass tolerance of 5 ppm. Peptides of 6–144 amino acids with a maximum of two missed cleavage sites were considered. Peptides were grouped by mass and sequences and validated based on q-values with FDR < 0.01. A minimal number of one unique peptide was required for protein identification. The strict maximum parsimony principle was applied and the grouping of proteins that could not be differentiated into protein groups based on their peptide masses was enabled. Areas under the ion-chromatography curves for proteins were quantified based on the top three unique peptides. Relative protein abundances of human proteins were calculated by addition of the areas under the curve of all protein groups containing one or more variant of one or more isoform of a single gene product. Similarly, the sum was calculated of the areas under the curve of all protein groups containing only genes and/or variants of genes annotated with the same function.

Integrated data handling and comparative, ecological and network analyses. All functional and taxonomic annotations, read coverage data, positions and frequencies of variants and relative protein abundances for the contigs and their associated ORFs were stored in a MongoDB (v. 3.0.3) collection (Supplementary Fig. 25). Retrieval of data for further analyses in R (ref. 88) was accomplished using the package rmongodb.

Bray–Curtis dissimilarity indices were calculated, and permutational MANOVA (multivariate analysis of variance; aka 'adonis'[89]) was performed using the R package vegan on sum-normalized relative abundances. Unseen[90] mOTUs were estimated using 'estimateR' from vegan. Total Soerensen dissimilarity indices were calculated using the R package betapart (ref. 91) on sum-normalized relative abundances. Jensen–Shannon divergences were calculated using the implementation in the R package phyloseq (ref. 61), also on sum-normalized relative abundances. Principal coordinate analyses were carried out using the R package ape 3.2 (ref. 92). Multiple co-inertia analysis was performed using the R package omicade4 (ref. 93) using scaled data (in accordance with the manual's recommendations).

Median numbers per individual of metagenomic or metatranscriptomic reads mapping to the protein-coding mOTU (ref. 38) marker genes and of metagenomic and metatranscriptomic reads mapping to functionally annotated genes were used for differential analyses by applying the R package DESeq2 (ref. 94) (see Supplementary Table 1 for numbers of samples per individual). To find differences between the individuals with T1DM and healthy individuals, the main effect of T1DM in a model accounting for family membership and T1DM was analysed. For differences between relative protein abundances, Wilcoxon rank sum tests or Kruskal–Wallis tests were applied.

Network analysis was performed on a gene-centric generalized network of KOs, built as described in ref. 37 by using shared products/educts of reactions related to KOs detected at the metatranscriptomic level, to link these KOs using a custom R script and functions from R packages graph, igraph (ref. 95) and BioNet (ref. 96). Top-scoring modules were extracted using BioNet and Heinz (ref. 97), using P values of differential or correlation analyses (without multiple testing adjustment). Metatranscriptomic abundances, fold changes or correlations within KEGG pathways were visualized using the R package pathview (ref. 98).

Data and code availability. Metagenomic and metatranscriptomic sequencing reads can be accessed from NCBI BioProject PRJNA289586. Assembled contigs and gene predictions can be accessed at MG-RAST (ref. 99) (submission IDs are shown in Supplementary Table 7). The mass spectrometry proteomics data have been deposited at the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository[100] with the data set identifier PXD003791. The 200 binned population-level genomes with >93% completeness can be accessed via the guest account at RAST (ref. 101) (submission IDs are shown in Supplementary Table 4). Custom scripts are available at https://git-r3lab.uni.lu/anna.buschart/MuStMultiomics.

Received 22 March 2016; accepted 23 August 2016; published 10 October 2016; corrected 24 October 2016

References

1. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–180 (2011).
2. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
3. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486,** 222–227 (2012).
4. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493,** 45–50 (2013).
5. Faith, J. J. *et al.* The long-term stability of the human gut microbiota. *Science* **341,** 1237439 (2013).
6. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–1131 (2006).
7. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).
8. Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513,** 59–64 (2014).
9. Franzosa, E. A. *et al.* Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl Acad. Sci. USA* **111,** E2329–E2338 (2014).
10. Arrieta, M.-C. *et al.* Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci. Transl. Med.* **7,** 307ra152 (2015).
11. Verberkmoes, N. C. *et al.* Shotgun metaproteomics of the human distal gut microbiota. *ISME J.* **3,** 179–189 (2008).
12. Turnbaugh, P. J. *et al.* Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc. Natl Acad. Sci. USA* **107,** 7503–7508 (2010).
13. Gosalbes, M. J. *et al.* Metatranscriptomic approach to analyze the functional human gut microbiota. *PLoS ONE* **6,** e17447 (2011).
14. Erickson, A. R. *et al.* Integrated metagenomics/metaproteomics reveals human host–microbiota signatures of Crohn's disease. *PLoS ONE* **7,** e49138 (2012).

15. 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).
16. Xiong, W., Giannone, R. J., Morowitz, M. J., Banfield, J. F. & Hettich, R. L. Development of an enhanced metaproteomic approach for deepening the microbiome characterization of the human infant gut. *J. Proteome Res.* **14,** 133–141 (2015).
17. Pérez-Cobas, A. E. *et al.* Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut* **62,** 1591–1601 (2013).
18. Waldor, M. K. *et al.* Where next for microbiome research? *PLoS Biol.* **13,** e1002050 (2015).
19. Beulig, F. *et al.* Altered carbon turnover processes and microbiomes in soils under long-term extremely high CO_2 exposure. *Nat. Microbiol.* **1,** 15025 (2016).
20. Hultman, J. *et al.* Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature* **521,** 208–212 (2015).
21. Goodrich, J. K. *et al.* Human genetics shape the gut microbiome. *Cell* **159,** 789–799 (2014).
22. Schloss, P. D., Iverson, K. D., Petrosino, J. F. & Schloss, S. J. The dynamics of a family's gut microbiota reveal variations on a theme. *Microbiome* **2,** 25 (2014).
23. Song, S. J. *et al.* Cohabiting family members share microbiota with one another and with their dogs. *eLife* **2,** e00458 (2013).
24. Patterson, C. C. *et al.* Trends in childhood type 1 diabetes incidence in Europe during 1989–2008: evidence of non-uniformity over time in rates of increase. *Diabetologia* **55,** 2142–2147 (2012).
25. Gillespie, K. M. *et al.* The rising incidence of childhood type 1 diabetes and reduced contribution of high-risk HLA haplotypes. *Lancet* **364,** 1699–1700 (2004).
26. Atkinson, M. A. & Chervonsky, A. Does the gut microbiota have a role in type 1 diabetes? Early evidence from humans and animal models of the disease. *Diabetologia* **55,** 2868–2877 (2012).
27. Cabrera, S. M., Henschel, A. M. & Hessner, M. J. Innate inflammation in type 1 diabetes. *Transl. Res.* **167,** 214–227 (2016).
28. Rodriguez-Calvo, T., Ekwall, O., Amirian, N., Zapardiel-Gonzalo, J. & von Herrath, M. G. Increased immune cell infiltration of the exocrine pancreas: a possible contribution to the pathogenesis of type 1 diabetes. *Diabetes* **63,** 3880–3890 (2014).
29. Giongo, A. *et al.* Toward defining the autoimmune microbiome for type 1 diabetes. *ISME J.* **5,** 82–91 (2011).
30. Brown, C. T. *et al.* Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes. *PLoS ONE* **6,** e25792 (2011).
31. Endesfelder, D. *et al.* Compromised gut microbiota networks in children with anti-islet cell autoimmunity. *Diabetes* **63,** 2006–2014 (2014).
32. 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).
33. 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).
34. 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).
35. 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).
36. Muller, E. E. L. *et al.* Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nat. Commun.* **5,** 5603 (2014).
37. Roume, H. *et al.* Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. *NPJ Biofilms Microbiomes* **1,** 15007 (2015).
38. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10,** 1196–1199 (2013).
39. Franzosa, E. A. *et al.* Identifying personal microbiomes using metagenomic codes. *Proc. Natl Acad. Sci. USA* **112,** E2930–E2938 (2015).
40. 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).
41. Legrand, D. *et al.* Lactoferrin structure and functions. *Adv. Exp. Med. Biol.* **606,** 163–194 (2008).
42. French, A. T. *et al.* The expression of intelectin in sheep goblet cells and upregulation by interleukin-4. *Vet. Immunol. Immunopathol.* **120,** 41–46 (2007).
43. 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. *J. Biochem.* **154,** 437–448 (2013).
44. Bertuccini, L. *et al.* Lactoferrin prevents invasion and inflammatory response following *E. coli* strain LF82 infection in experimental model of Crohn's disease. *Dig. Liver Dis.* **46,** 496–504 (2014).
45. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science* **347,** 1260419 (2015).
46. Atkinson, M. A. Losing a grip on the notion of β-cell specificity for immune responses in type 1 diabetes: can we handle the truth? *Diabetes* **63,** 3572–3574 (2014).
47. Walter, J. & Ley, R. The human gut microbiome: ecology and recent evolutionary changes. *Annu. Rev. Microbiol.* **65,** 411–429 (2011).
48. 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).
49. Nabokina, S. M. *et al.* Molecular identification and functional characterization of the human colonic thiamine pyrophosphate transporter. *J. Biol. Chem.* **289,** 4405–4416 (2014).
50. Joice, R., Yasuda, K., Shafquat, A., Morgan, X. C. & Huttenhower, C. Determining microbial products and identifying molecular targets in the human microbiome. *Cell Metab.* **20,** 731–741 (2014).
51. 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).
52. Ammerlaan, W. *et al.* Method validation for preparing serum and plasma samples from human blood for downstream proteomic, metabolomic, and circulating nucleic acid-based applications. *Biopreserv. Biobank* **12,** 269–280 (2014).
53. Roume, H., Heintz-Buschart, A., Muller, E. E. L. & Wilmes, P. Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. *Methods Enzymol.* **531,** 219–236 (2013).
54. Varrette, S., Bouvry, P., Cartiaux, H. & Georgatos, F. Management of an academic HPC cluster: The UL experience. In *Proc. 2014 Int. Conf. High Performance Computing & Simulation (HPCS 2014)* 959–967 (IEEE, 2014).
55. Kultima, J. R. *et al.* MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS ONE* **7,** e47656 (2012).
56. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25,** 1966–1967 (2009).
57. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41,** D590–D596 (2013).
58. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
59. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15,** R46 (2014).
60. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32,** 834–841 (2014).
61. McMurdie, P. J. & Holmes, S. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8,** e61217 (2013).
62. Reynolds, A. P., Richards, G., de la Iglesia, B. & Rayward-Smith, V. J. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *J. Math. Model. Algor.* **5,** 475–504 (2006).
63. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28,** 3150–3152 (2012).
64. 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).
65. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
66. 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).
67. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119 (2010).
68. Dupont, C. L. *et al.* Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.* **6,** 1186–1199 (2011).
69. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31,** 533–538 (2013).
70. Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28,** 1033–1034 (2012).
71. Finn, R. D. *et al.* HMMER web server: 2015 update. *Nucleic Acids Res.* **43,** W30–W38 (2015).
72. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7,** e1002195 (2011).
73. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42,** D222–D230 (2013).
74. Haft, D. H. *et al.* TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.* **41,** D387–D395 (2012).
75. Magis, C. *et al.* in *Multiple Sequence Alignment Methods* Vol. 1079 (ed. Russell, D. J.) 117–129 (Humana, 2014).

76. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **40,** D742–D753 (2011).
77. The UniProt Consortium. Uniprot: a hub for protein information. *Nucleic Acids Res.* **43,** D204–D212 (2015).
78. Christian, N., May, P., Kempa, S., Handorf, T. & Ebenhöh, O. An integrative approach towards completing genome-scale metabolic networks. *Mol. BioSyst.* **5,** 1889–1903 (2009).
79. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
80. Laczny, C. C., Pinel, N., Vlassis, N. & Wilmes, P. Alignment-free visualization of metagenomic data by nonlinear dimension reduction. *Sci. Rep.* **4,** 4516 (2014).
81. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. 2nd Int. Conf. Knowledge Discovery Data Mining (KDD-96), 1–6 (1996).
82. Benaglia, T., Chauveau, D. & Hunter, D. Mixtools: an R package for analyzing finite mixture models. *J. Stat. Soft.* **32,** 6 (2010).
83. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327,** 78–81 (2010).
84. Schubert, J. *et al.* Mutations in *STX1B*, encoding a presynaptic protein, cause fever-associated epilepsy syndromes. *Nat. Genet.* **46,** 1327–1332 (2014).
85. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164–e164 (2010).
86. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46,** 912–918 (2014).
87. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).
88. R Core Team. *R: A Language and Environment for Statistical Computing* (The R Foundation, 2014), http://www.R-project.org
89. McArdle, B. H. & Anderson, M. J. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* **82,** 290–297 (2001).
90. 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. *Biometrics* **70,** 671–682 (2014).
91. Baselga, A. & Orme, C. D. L. Betapart: an R package for the study of beta diversity. *Methods Ecol. Evol.* **3,** 808–812 (2012).
92. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20,** 289–290 (2004).
93. Meng, C., Kuster, B., Culhane, A. C. & Gholami, A. M. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* **15,** 162 (2014).
94. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
95. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Systems* **1695** (2006), http://igraph.org
96. 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).
97. Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T. & Müller, T. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. *Bioinformatics* **24,** i223–i231 (2008).
98. Luo, W. & Brouwer, C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29,** 1830–1831 (2013).
99. 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).
100. Vizcaino, J. A. *et al.* The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **41,** D1063–D1069 (2012).
101. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9,** 75 (2008).

Acknowledgements

The authors thank the staff of the Clinical and Epidemiological Investigation Center (CIEC) Luxembourg for undertaking the sample and data collection from participants in this study. The authors thank B. Phillips for input and feedback during the conception of the study and S. Collignon, K. Greenhalgh, P. do Rosario Martins Conde and A. Kaysen for technical assistance with biomolecular extraction and quality control. The authors thank D. Baiwir and G. Mazucchelli (CART-GIGA) for measurements and assistance. The *in silico* analysis results presented in this Article were obtained using the high-performance computing facilities of the University of Luxembourg, whose administrators are acknowledged. The authors thank M. Brunkow at the Institute for Systems Biology (ISB), Seattle, who provided project management, and acknowledge support from the Family Genomics Group of Leroy Hood (ISB) for the human whole-genome sequencing data. The present work was supported by an ATTRACT programme grant (ATTRACT/A09/03) and CORE programme grant (CORE/15/BM/10404093) to P.W. and Aide à la Formation Recherche grants to C.C.L. (AFR PHD/4964712) and L.W. (AFR PHD-2013-5824125), as well as by BIOMARKAPD—a project under the aegis of an EU Joint Programme—on Neurodegenerative Diseases (JPND), all funded by the Luxembourg National Research Fund (FNR). Additional support was provided by 'le plan Technologies de la Santé par le Gouvernment du Grand Duché de Luxembourg' through the Luxembourg Centre for Systems Biomedicine, University of Luxembourg to P.M. Sample collection, processing and storage were supported by IBBL under the Personalised Medicine Consortium Diabetes programme.

Author contributions

A.H.-B. carried out the comparative analyses of metagenomic, metatranscriptomic and metaproteomic data, metagenome binning and data interpretation, coordinated the omic measurements, participated in the biomolecular extractions and in the sequence annotation and database generation for metaproteomic analysis. P.M. carried out sequence assembly, gene prediction and analysis of the human genome data and participated in functional annotation and metaproteomic database generation, and data interpretation. C.C.L. participated in metagenome binning. P.M. and A.K. generated the personalized human protein databases. L.A.L. carried out the biomolecular extractions. L.W. participated in the biomolecular extractions and data interpretation. C.B. participated in sample collection and processing. A.H. participated in the design of the study and in sample and data collection, and contributed towards writing the manuscript. J.G.S., C.d.B. and P.W. conceived the study and participated in its design. C.d.B. and P.W. coordinated the study and C.d.B. participated in writing the manuscript. A.H.-B., P.M. and P.W. wrote the manuscript. All authors read and approved the final manuscript.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.H.B. and P.W.

How to cite this article: Heintz-Buschart, A. *et al.* Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat. Microbiol.* **2,** 16180 (2016).

Competing interests

The authors declare no competing financial interests.

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

Anna Heintz-Buschart, Patrick May, Cédric C. Laczny, Laura A. Lebrun, Camille Bellora, Abhimanyu Krishna, Linda Wampach, Jochen G. Schneider, Angela Hogan, Carine de Beaufort and Paul Wilmes

Nature Microbiology **2,** 16180 (2016); published 10 October 2016; corrected 24 October 2016

This Article should have been published under a Creative Commons licence according to the Nature policy on publishing the primary sequence of an organism's genome for the first time. The editors apologize to the authors and to readers for this error. The manuscript is now open access and published under a CC-BY licence. All versions of the Article have been modified accordingly.