doi: 10.1093/nar/gkx1209

Functional sequencing read annotation for high precision microbiome analysis

Chengsheng Zhu[1,*], Maximilian Miller[1,2,3], Srinayani Marpaka[1], Pavel Vaysberg[1], Malte C. Rühlemann[4], Guojun Wu[5], Femke-Anouska Heinsen[4], Marie Tempel[6], Liping Zhao[1,5,7], Wolfgang Lieb[6], Andre Franke[4] and Yana Bromberg[1,8,9,*]

[1]Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Dr, New Brunswick, NJ 08873, USA, [2]Department for Bioinformatics and Computational Biology, Technische Universität München, Boltzmannstr. 3, 85748 Garching/Munich, Germany, [3]TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), Technische Universität München, 85748 Garching/Munich, Germany, [4]Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany, [5]State Key Laboratory of Microbial Metabolism and Ministry of Education Key Laboratory of Systems Biomedicine, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China, [6]Institue of Epidemiology, Kiel University, Kiel, Germany, [7]Canadian Institute for Advanced Research, Toronto, Canada, [8]Department of Genetics, Rutgers University, Human Genetics Institute, Life Sciences Building, 145 Bevier Road, Piscataway, NJ 08854, USA and [9]Institute for Advanced Study, Technische Universität München (TUM-IAS), Lichtenbergstrasse 2 a, D-85748 Garching, Germany

Received May 30, 2017; Revised November 15, 2017; Editorial Decision November 16, 2017; Accepted November 27, 2017

ABSTRACT

The vast majority of microorganisms on Earth reside in often-inseparable environment-specific communities—microbiomes. Meta-genomic/-transcriptomic sequencing could reveal the otherwise inaccessible functionality of microbiomes. However, existing analytical approaches focus on attributing sequencing reads to known genes/genomes, often failing to make maximal use of available data. We created *faser (functional annotation of sequencing reads)*, an algorithm that is optimized to map reads to molecular functions encoded by the read-correspondent genes. The *mi-faser* microbiome analysis pipeline, combining *faser* with our manually curated reference database of protein functions, accurately annotates microbiome molecular functionality. *mi-faser*'s minutes-per-microbiome processing speed is significantly faster than that of other methods, allowing for large scale comparisons. Microbiome function vectors can be compared between different conditions to highlight environment-specific and/or time-dependent changes in functionality. Here, we identified previously unseen oil degradation-specific functions in BP oil-spill data, as well as functional signatures of individual-specific gut microbiome responses to a dietary intervention in children with Prader–Willi syndrome. Our method also revealed variability in Crohn's Disease patient microbiomes and clearly distinguished them from those of related healthy individuals. Our analysis highlighted the microbiome role in CD pathogenicity, demonstrating enrichment of patient microbiomes in functions that promote inflammation and that help bacteria survive it.

INTRODUCTION

Microorganisms inhabit every available niche of our planet, and our bodies are no exception. Microbes that survive and thrive in the environments at the extremes of temperature, pH, and chemical or radiation contamination possess unique molecular functions of high industrial, clinical, and bioremediation value. The human body microbiome critically impacts our health. For example, Crohn's disease (CD) is a multifactorial disorder resulting from the interplay of individual genetic susceptibility, the gastrointestinal (GI) microbiome and other environmental factors. Taxonomic surveys of the GI microbiome have revealed microbial community features that are unique to CD patients, e.g. overall loss of microbial diversity (1,2), as well as depletion and enrichment of certain bacterial taxa (3–6). Establishing whether these observed microbial community shifts contribute to pathogenesis or, instead, correlate with or result from the disease onset, requires understanding not only what are the microbes involved, but also what they do. Ear-

[*]To whom correspondence should be addressed. Email: czhu@bromberglab.org
Correspondence may also be addressed to Yana Bromberg. Tel: +1 848 932 5638; Fax: +1 848 932 8965; Email: yanab@rci.rutgers.edu

lier studies indicate that in association with CD, the microbiome molecular function potential is more consistently disturbed than taxonomic makeup (7). More thorough functional analyses, e.g. based on deep metagenomic sequencing, are necessary to elucidate these findings.

Metagenome functional annotation can be performed with or without genome assembly. If the reads can be assembled into large contigs, existing annotation pipelines, such as RAST (8) and IMG (9), can be applied. However, assembly is difficult and often plagued by a large fraction of unassembled reads or short length contigs, which belong to the minor microbiome members, and by chimeric assemblies, which are especially common for complex and highly diverse samples (see Sczyrba *et al.,* 2017, doi: https://doi.org/10.1101/099127). Downstream gene finding algorithms are further faced with incomplete and erroneously assembled sequences, complicating statistical model constructions. Read-based annotation, *e.g.* using a platform such as MG-RAST (10), can access molecular functionality of the entire community. However, reads are usually annotated via function transfer by homology that, due to the short read length, is lacking in precision. This inaccuracy is additionally compounded by the erroneous computational annotations of most genes in the reference databases (11).

Here, we compiled a gold standard set of reference proteins (GS), with experimentally annotated molecular functions. We further developed *faser* (functional annotation of sequencing reads), an algorithm that uses alignments of translated sequencing reads to full-length proteins to annotate read-'parent protein' molecular functionality. *faser* annotates reads with higher precision at higher resolution, *i.e.* more specific functionality, than BLAST or PSI-BLAST. In a benchmark test, the functional annotations produced by the combination of the *faser* algorithm with the GS database were 12% more accurate than MG-RAST. Note that this performance may be an overestimate because the benchmark metagenome included the GS database. However, when GS was replaced with md5nr, MG-RAST's reference database, *faser* annotated 20% more reads than MG-RAST at a comparable precision level. These results illustrate that the GS and *faser* combination improves on MG-RAST capabilities.

Our *mi-faser* pipeline implementation (Figure 1), combining *faser* and GS, is highly parallelized, making use of all available compute cores and processing a (~10GB/70M read) meta-genomic/-transcriptomic file in under half an hour (using 400 compute cores, on average). Note that if multiple microbiomes are submitted for annotation in parallel, the time scales favourably; in testing, 17 metagenomes were processed within 66 minutes. *mi-faser* results for all microbiomes analysed in this manuscript are available at http://services.bromberglab.org/mifaser/results/example. The standalone version of the pipeline, along with the *mi-faser* source code, is available at https://bitbucket.org/bromberglab/mifaser, as well as on the bromberglab website.

We applied our *mi-faser* to metagenomic data collected from beach sands in different stages of oil contamination (12). Here, *mi-faser* was able to identify oil degradation functionality that was missed by MG-RAST. We further performed large-scale analysis of 68 metagenomic datasets from a study of dietary intervention in Prader-Willi syndrome (PWS) affected obese children. Each dataset was processed in approximately 16 minutes, highlighting *mi-faser*'s processing speed. We identified previously unseen individual-specific patterns in microbiome changes induced by the treatment. Finally, we also analyzed the GI tract microbiome data from Crohn's Disease (CD) patients and their relatives. We found the microbiome functional profiles were similar between healthy individuals but different across patients and between patients and their healthy relatives. Particularly, our analysis revealed that CD patients' microbiomes were enriched in functions that help bacteria survive inflammation, i.e. glutathione metabolism and RNA modification, and in functions that cause inflam-

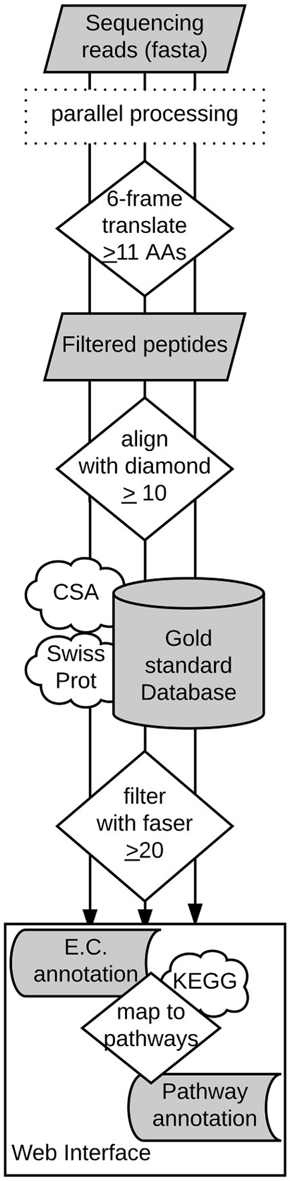

Figure 1. *mi-faser* pipeline. *mi-faser* is parallelized and runs a load balancer to submit jobs to available [1–2000] compute cores. Under normal functioning conditions (~400 available cores, on average), it takes ~30 min to process a single (10G/70M read) meta-genome/-transcriptome.

mation, i.e. lipopolysaccharide and acetaldehyde production. These results suggest the microbiome's role in CD-associated pathogenicity.

MATERIALS AND METHODS

Datasets

To compile the *PE1-set*, we extracted from SwissProt (Oct. 2015) (13) proteins that are (i) bacterial, (ii) with evidence of existence, i.e. SwissProt protein evidence is 1, and (iii) explicitly assigned an E.C. (Enzyme Commission) number (14); note that we excluded proteins with incomplete annotations, e.g. 1.1.1.-, as well as those with multiple annotations. From the *PE1-set*, we further extracted proteins whose functions are experimentally verified (Evidence = ’any experimental assertion’; *EXP-set*).

From the Catalytic Site Atlas database (*CSA-set*) (15) we extracted all proteins that had literature-based annotations. We identified the overlap between the *PE1-set* and these proteins, and defined our gold-standard dataset (*GS-set*; Supplementary Data 1) as the combination of *CSA-set* and *EXP-set*, with 100% identical sequences removed.

For each protein of the *PE1-set* and *GS-set*, we extracted the corresponding gene from ENA (European Nucleotide Archive) (16) (including 5′ UTR and 3′ UTR) and randomly generated 10 DNA reads (50–250 nucleotides) that overlap by at least one nucleotide of the coding region. We further performed 6-frame translations of the reads and excluded peptides shorter than 11 amino acids. We defined the corresponding peptide collection as *rPE1-set* and *rGS-set*.

We downloaded from MG-RAST the *md5nr* database and defined its proteins as the *md5nr-set*.

We obtained six beach sand metagenomes from a previous study of the Deepwater Horizon oil spill (12). Here, metagenomic DNA was sequenced using Illumina MiSeq with paired-end strategy to produce 151 bp reads. The samples reside in NCBI (BioProject PRJNA260285), including (i) pre-oil phase samples, OS-S1 (SRX692936) and OS-S2 (SRX695904), (ii) oil phase samples, OS-A (SRX696142) and OS-B (SRX696240) and (iii) post-oil recovered phase samples, OS-I600 (SRX696250) and OS-I606 (SRX696254).

We also obtained 68 gut metagenomic sequencing datasets (SRA (17) accession number SRP045211) from a study of dietary intervention in Chinese children affected by PWS (18). Fecal DNA samples before and after the treatment ($n = 17$, at Day 0, 30, 60 and 90) were sequenced using Illumina HiSeq 2000 with paired-end strategy to produce 100 bp reads. The quality control was performed as described in the previous study (18).

We additionally obtained 11 human gut (fecal) microbiome samples from a family affected by CD from the PopGen biobank (Schleswig-Holstein, Germany; accessible via a Material Data Access Form. Information and application procedures for data access can be found at http://www.uksh.de/p2n/Information±for±Researchers.html). Of these, nine members were self-reported as healthy and two were affected. Metagenomic data were generated using the Illumina Nextera DNA Library Prep Kit and sequenced 2 × 125 bp on an Illumina HiSeq2500. In total, 424.8 million paired-end reads were generated with a median number of 38.9 million read pairs per sample. Adapter trimming was performed using Trimmomatic (19) in paired-end mode, discarding reads shorter than 60 bp. Quality filtering was done using Sickle (20) run in paired-end mode, with a quality threshold of 20 and a minimum length of 60 bp. To remove contaminating host sequences from the dataset, DeconSeq (v0.4.3) (21) was run with the human reference genome (GRCh38) as database. Only read-pairs where both sequences survived quality control were retained. On average 11.76% of raw reads were discarded, leaving 374.8 million read pairs for downstream analysis.

faser curve optimization

We PSI-BLASTed the *rGS-set* against the *GS-set* (parameters: evalue $1e^{-3}$; inclusion ethresh $1e^{-10}$; num iterations 3; max_target_seqs 1 000 000), excluding self-hits, i.e. peptide hits of their ‘parent’ proteins. For any peptide, functional annotation (E.C. number) was inherited from the ‘parent’ protein; one nucleotide overlap required to transfer annotation. A peptide-protein alignment is considered positive if the functional annotations of the peptide and the aligned protein match exactly at the selected number of E.C. digits, and negative otherwise. Any given alignment can be plotted in an L (alignment length) *vs.* Id (alignment sequence identity) two-dimensional space. Further, an exponential decay curve (as for HSSP calculations, (22)) can be used to identify the alignments in this space as true positives (alignments of peptides to proteins of identical function that fall above or on the curve), false positives (different functions above or on the curve), true negatives (different functions below the curve) and false negatives (identical functions below the curve). From these values, we calculated precision (positive accuracy; Equation 1) and recall (positive coverage; Equation 2) for different curve parameters (*a* and *b* in Equation 4), optimizing the latter to fit a curve best separating positive from negative alignments in terms of the highest *F*-measure (Equation 3).

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{1}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2}$$

$$F = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$b \times L^{-a \times \left(1+e^{-\frac{L}{1000}}\right)} \tag{4}$$

To avoid overestimating performance of *faser*, we clustered the *GS-set* with CD-hit at 40% sequence identity and split the clusters into ten subsets. We further optimized *faser* curve parameters in 10-fold cross-validation, i.e. we iteratively optimized the curve on nine subsets and tested it on the remaining one, repeating this process 10 times for a different subset as the test set. We evaluated the performance reported here by summing the numbers of true and false positives and negatives in each test set. As all ten curves were very similar in parameters, we took the average of these to establish the final *faser* curve.

To summarise, the *faser* curve is meant to predict from a peptide-protein alignment, whether the 'parent' protein of the peptide and the aligned protein share the same function (E.C. annotation). Additionally, the distance of the alignment point to the curve along the sequence identity (Id) axis indicates the reliability of the prediction.

Evaluating *faser* using DIAMOND results

We extracted the proteins from the *GS-set* and *md5nr-set* that had identical UniProt IDs. We performed searches against the *md5nr* database using PSI-BLAST (parameters: evalue 1e^{-3}; inclusion ethresh 1e^{-10}; num iterations 3; max_target_seqs 1 000 000), BLASTP (parameters: evalue 1e^{-3}; max_target_seqs 1 000 000), and DIAMOND (parameters: min-score 10; k 1 000 000). We further excluded from the results the alignments to subject proteins that were not in the overlap set. We compared the *faser* values calculated from the results of different alignment algorithms by performing a 100-fold bootstrap, sampling ~20% of the results at each iteration. Note that we used the bootstrap approach to assess the consistency of the observed performance differences.

Comparison to other methods

We submitted the artificial metagenome as well as the six sand metagenomes for processing to MG-RAST via its website and downloaded the resulting function annotations via the MG-RAST API (23). We used the KEGG (24) annotations from the *md5nr* database to establish the annotated E.C.s. Note that although proteins can carry out multiple functions, in this study we, conservatively, only included proteins with unique and complete E.C. annotations; i.e. we excluded proteins with incomplete or multiple E.C. annotations.

We compared different database/algorithm combinations for the annotation of the same sample (Supplementary Figure S2). The Venn diagrams of the numbers of E.C.s annotated by different such combinations were generated by Venny (25). When comparing across sand metagenome samples from different phases, sample-specific E.C.s were removed as uninformative (<1% of total E.C.s in both cases). The correlation between samples was calculated with Spearman's rho, ρ, offered in the R package, Hmisc (26).

Two other tools, Fun4Me and ShotMAP, were installed locally and run on the artificial metagenome with default parameters; for both, we compared the precision of the methods (Equation 1) as well as the number of correctly annotated reads.

Functional analysis of PWS dietary intervention metagenomes

We performed NMDS (Non-metric multidimensional scaling) (27) analysis and the subsequent permanova test using the Vegan R package (28) and calculated the Euclidean distance between samples in the NMDS graph. Within the untreated Day0 group of samples, we identified outliers (individuals with inter-sample distance two standard deviations away from the average distance; 3% of all distances). All time-point samples of these individuals were removed from subsequent analysis. For remaining individuals, we compared the distances within each time-point group, as well as the distances of all the time points from Day0 for every individual separately.

Functional analysis of CD metagenomes

As described above, NMDS analysis (Shepard plot in Supplementary Figure S10), along with the subsequent permanova test was carried out using the Vegan R package (28). From the distributions of E.C.s in the microbiomes of healthy individuals, we calculated the 'confidence range' for each E.C. as Q1 – 3*IQR (three interquartile ranges below the first quartile) to Q3 + 3*IQR (three interquartile ranges above the third quartile). Patient E.C.s that fell outside this range were identified as significantly depleted or enriched, respectively. Pathway analysis was performed with the KEGG Mapper tool (24). Jaccard Index was calculated as the size of intersection divided by the size of union of the two sample sets.

RESULTS AND DISCUSSION

Few proteins have experimentally verified function annotation

Among the 332 193 bacterial proteins in SwissProt (Oct. 2015) (13,29), only 18 240 (~5%) are annotated as *existing* with evidence at protein level. Of these, we extracted 5 965 that have unique (one per protein) and explicit (all four digits) Enzyme Commission (E.C.) annotations (*PE1-set*; Materials and Methods). From our *PE1-set*, we further selected proteins whose *functions* were experimentally verified, as noted in the Catalytic Site Atlas (*CSA-set*) (15) or SwissProt (*EXP-set*) (13,29). After filtering, our set contained 2 848 (2 810 non-redundant at 100% sequence identity; *GS-set*) bacterial proteins of experimentally verified function. Note that analysis of available mass-spectrometry databases (30,31) is likely to retrieve a much larger set of verified existing proteins; however, these are not yet experimentally annotated for molecular functionality. Thus, our collection is the cleanest available dataset of functional annotations; i.e. functional annotations in public databases are usually based on (many rounds of) function transfer by homology and are, as such, often questionable.

faser is more accurate for function transfer by homology than PSI-BLAST

We created artificial reads from the gene nucleotide sequences corresponding to the proteins in *GS-set* and *PE1-set* (6-frame translated to peptides, *rGS-set* and *rPE1-set*, Materials and Methods). We further PSI-BLASTed (32) the *rGS-set* against *GS-set,* excluding self-hits, to determine the equation of the curve (Equation 5) separating the correct alignments (same function) from the incorrect ones (different functions) in the *L* (alignment length) versus *Id* (sequence identity) space. Our approach was modeled after the HSSP metric for function transfer between full-length proteins (22,33). We optimized the curve parameters to

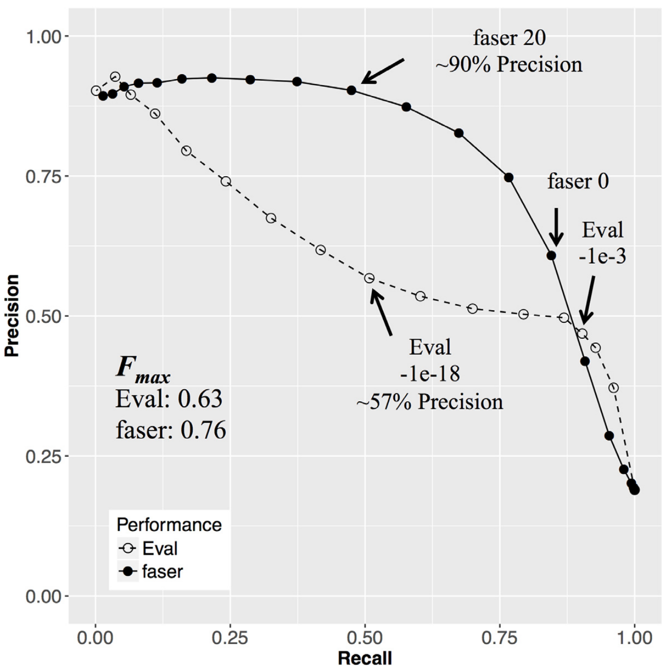

Figure 2. *faser* outperforms PSI-BLAST in annotating read functions. At most cutoffs, *faser* (filled circles) is more precise than PSI-BLAST (empty circles). For example, for nearly half the reads, it provides as much as 90% annotation accuracy as compared to 57% attained by PSI-BLAST (arrows at *faser* score = 20 and e-value = e^{-18}). At the default cutoff of 0, *faser* attains similar accuracy as PSI-BLAST at e-value = e^{-18}, but for ~35% more reads.

maximize the *F* measure (Materials and Methods), representative of best separation of peptide–protein alignments of the same function (E.C. annotation) from those of different functions (Methods). Thus, if a given alignment is above the curve, the 'parent protein' of the peptide and the aligned reference protein are predicted to share function. The *faser* score (the distance from the curve along the *Id* axis) indicates the reliability of such predictions. This measure clearly outperforms PSI-BLAST e-value in annotating function ($F_{\max}$ of 0.76 versus 0.63, respectively; Equation (3), the highest *F* measure as in (34); recall in Figure 2 was calculated with the background of all PSI-BLAST results at e-value = 10^{-3}). For example, at recall levels of ~50%, the *faser* score (=20) is nearly 90% accurate, which is >30% more than e-value (=10^{-18}; Figure 2). E-value reaches ~90% precision at cut-offs <10^{-36}, which corresponds to recall of <7% (Figure 2).

The number of matching E.C. digits reflects the level of resolution of function annotation; i.e. proteins that share only the first three E.C. digits have similar functions with slight differences. For example, both 1.1.1.1 and 1.1.1.2 are alcohol dehydrogenases, but with different electron acceptors: NAD+ and NADP+, respectively. PSI-BLAST exhibits comparable performance to *faser* when matching the first three E.C. digits (Supplementary Figure S1A), but fails to differentiate functions at the fourth digit resolution level, producing a large number of false positives (Figure 2). *faser* resolves the fourth E.C. digit at >90% precision with >40% recall. At all cut-offs, when compared to PSI-BLAST, *faser* consistently offers as much as ~50% higher recall at same precision level and up to ~25% higher precision at same recall level (Figure 2).

Table 1. Artificial metagenome (rPE1-set) annotation by F_G, F_M and M_M

	F_G	F_M	M_M
Annotated reads	34 851	48 481	30 800
Multi-E.C. reads[a]	1004	11 373	200
Erroneously annotated reads	416	5705	4237
Correctly annotated reads	33 431	31 103	26 363
Precision	99%	85%	86%

[a]Reads with multiple E.C. annotations were excluded from the analysis.

$$\text{faser score} = \begin{cases} -100, & L < 11 \\ Id - 352.3L^{-0.302\times\left(1+e^{-\frac{L}{1000}}\right)}, & L \geq 11 \end{cases} \quad (5)$$

Note that a previous study has shown that PSI-BLAST is not necessarily the best alignment method for function transfer, e.g. it was inferior to BLAST (34). Although faser was developed using PSI-BLAST, it can also be calculated via other alignment mechanisms. To alleviate the long alignment runtimes, we exhaustively tested our options (including comparing BLAST performance to PSI-BLAST) and ended up switching to DIAMOND (35) (Supplementary Text S1).

faser outperforms MG-RAST

We compared *faser* performance to that of MG-RAST (10), one of the most popular public metagenome annotation platforms. We considered both algorithm and database levels using the: (i) *faser* algorithm with the *GS-set* database (F_G, the *mi-faser* pipeline); (ii) *faser* algorithm with the *md5nr* database (36) ($F_{M;}$ *faser*-md5nr); (iii) MG-RAST algorithm with *md5nr* database (M_M, the MG-RAST pipeline) (Supplementary Figure S2; Methods). Note that we could not run the MG-RAST algorithm with the *GS-set* database because the MG-RAST developers advised against it, citing complicated installation.

When the *rPE1-set* is used as the artificial metagenome, the F_G and M_M annotations are significantly different (Table 1), although both pipelines annotate a similar number of reads (Figure 3A). This variation in performance is not biased toward any specific E.C. class (Supplementary Figure S3). Note that the *rPE1-set* is a superset of *GS-set*, which likely contributes to the improved performance of the F_G pipeline. The differences between F_G and M_M annotations (Figure 3B, first column) stem from the differences between the databases (*GS-set* vs. *md5nr*) and/or algorithms (*faser* versus MG-RAST). The divergence between F_G and F_M annotations (Figure 3B, second column) indicates that the database differences contribute significantly to the F_G/M_M variation. Note that this difference is not surprising as the *GS-set* and *md5nr* share only 779 E.C.s (62% and 29%, respectively).

The comparison between F_M and M_M results is more interesting (Figure 3B, third column), as it highlights the differences between the *faser* and MG-RAST algorithms. Using the same *md5nr* database, *faser* (F_M) annotated ~20% more reads than MG-RAST (M_M, Figure 3A) with comparable precision (Table 1). Note that the precision reported in

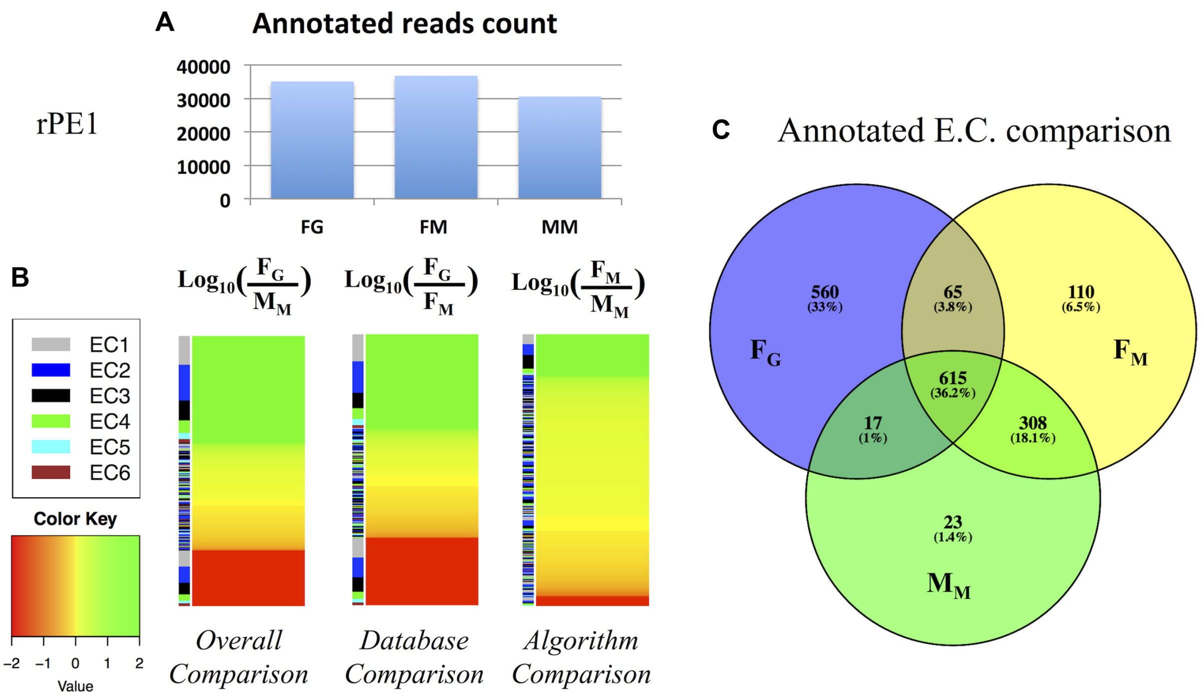

Figure 3. The *faser* algorithm in combination with the GS database annotates the artificial metagenome functions in a manner complementary to MG-RAST. (**A**) The number of reads annotated by each combination of algorithms and databases; (**B**) the read abundance by E.C. annotated via each combination of algorithm/database; (**C**) the total E.C. count annotated via each combination of algorithm/database.

these comparisons is affected by the misannotation (∼14%), i.e. UniProt proteins in both the *GS-set* and *md5nr* annotated with different E.C. numbers – a finding, which is in line with a previous study (11). F_M and M_M identified 923 E.C.s in common, while 175 and 40 E.C.s were uniquely identified by *faser* and MG-RAST, respectively (Figure 3C). In other words, for the same artificial metagenome, *faser* annotates ∼14% more functions (E.C.s) than MG-RAST algorithms. After exclusion of the database-specific E.C.s, the database impact was reduced (F_G/F_M, Supplementary Figure S4), yet we still observed substantial F_G/M_M differences largely due to the *faser vs.* MG-RAST algorithms. Notably, F_M still annotates ∼8% more functions than M_M (Supplementary Figure S4).

To summarize, the *faser* method comprises an exponential decay curve separating the two-dimensional space of alignment length versus sequence identity into 'same function' and 'different functions' peptide–protein alignments. The distance from a given alignment to the curve along the sequence identity axis is the final *faser* score. Implicitly, *faser* tries to capture homology of the peptide's 'parent' protein to the subject protein of the alignment. In *faser* development we used the database of experimentally described proteins (*GS-set*) to optimize and evaluate performance. We continue to use the GS database in the *mi-faser* implementation. However, *faser* alignment scoring can be applied to any other database as well. Note that we set the default cut-off of *faser* score at 20 for high precision (90%).

We further extended the comparison of the annotation methods to six metagenomic samples from the Deepwater Horizon oil spill beach sand study (12) (Methods). Note that in this real-life case, there was no 'correct' annotation to use for comparing annotation results. However, it appears that F_M and M_M results are orthogonal. For example, for OS-A (*oil phase*) F_M annotated >50% more reads than M_M (Supplementary Figure S5A); moreover, there were 220 E.C.s unique to F_M and 42 E.C.s unique to M_M (Supplementary Figure S5C). Annotation of other samples followed a similar pattern. Database differences resulted in a significant disparity between the number of reads annotated in each sample by F_G and M_M (e.g. Supplementary Figure S5B). However, both pipelines agreed that: (i) samples taken in the same phase were highly functionally correlated (Supplementary Tables S1 and S2), (ii) samples in *oil phase* were functionally more correlated with samples in *recovered phase* than *pre-oil phase* (Supplementary Tables S1 and S2, which may indicate that the environment has not fully recovered from the contamination) and (iii) ∼20% of reads in all samples mapped to housekeeping functions (housekeeping E.C.s complied from (37)). This agreement across methods suggests that F_G reflects true variation in functionality between samples from a perspective complementary to M_M.

We further searched for functions enriched in *oil phase* metagenomes as compared to either *pre-oil* or *recovered phases*. F_G returned 909 E.C.s (65%, 588 E.C.s, are *GS-set* specific), while M_M returned 1 627 E.C.s (65%, 1 062 E.C.s, are *md5nr* specific). Note that even for the E.C.s present in both databases, F_G and M_M revealed considerable discrepancies in abundance fold-changes across phases; ρ = 0.46 (Spearman's rho) for *oil-to-recovered* phase and only ρ = 0.09 for *oil-to-pre-oil* phase (Supplementary Figure S6). We explored E.C.s annotated by F_G as highly enriched (≥5 times) in the *oil phase* as compared to other phases, yet unchanged or even decreased by M_M. There are nine of these E.C.s in *oil-to-pre-oil* comparison and ten in *oil-to-recovered* comparison, with three E.C.s overlapping across comparisons; i.e. enriched in the *oil phase* as compared to either *pre-oil* or *recovered phases* (Supplementary Tables S3 and S4). Of the three overlapping E.C.s, two are particularly notable: 1.3.11.1 (catechol 1,2-dioxygenase) directly associates with BTEX (benzene, toluene, ethylbenzene and xylenes) degradation, while 1.8.99.1 (assimilatory sulfite reductase) is essential for sulfur reducing bacteria, known to degrade BTEX. Note that there were also three E.C.s annotated only by M_M that were enriched in the *oil phase*; however, we were not able to identify them as being directly related to oil degradation (Supplementary Tables S5 and S6).

We also compared our pipeline to two recently published metagenome annotation tools, Fun4Me (38) and ShotMAP (39), using the above-described artificial metagenome. Note that Fun4Me includes its own reference database, which cannot be changed on demand. ShotMAP allowed using our *GS-set* as reference. F_G correctly annotated 4 900 (17%) more reads than Fun4Me. Additionally, when the multi-EC-annotated reads were excluded, F_G attained 7% higher precision than Fun4Me (99% versus 92%, respectively; Supplementary Table S7). While results were not as striking, *faser* still outperformed ShotMAP (using our GS database) with 1 160 (4%) more correctly annotated reads and 2% higher precision (99% to 97%, respectively; Supplementary Table S7). Notably, the entire run took *mi-faser* (standalone version) 42 seconds, while Fun4Me required more than 25 minutes. The speed evaluation for ShotMAP was not possible via command-line due to installation issues, but the virtual machine implementation was able to finish in 3 minutes.

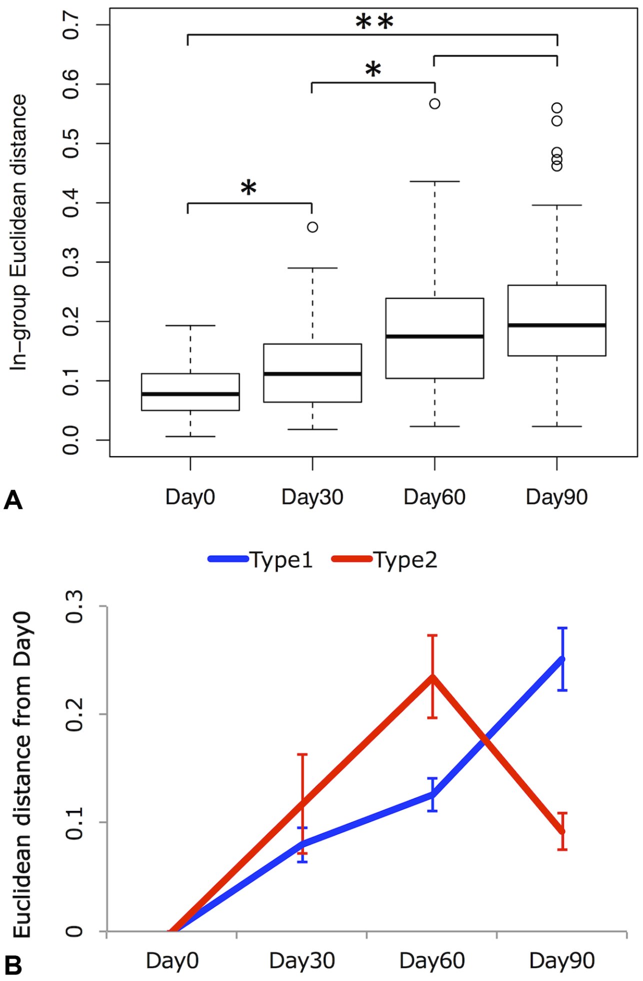

Figure 4. Functional capabilities of microbiomes of PWS patients shift in the course of dietary intervention. (**A**) The boxplot of Euclidean distance between samples of the same group (in-group distances), i.e. Day 0, 30, 60 or 90, on the NMDS diagram (Supplementary Figure S7). The in-group diversity increases significantly with time; * indicates *P*-value <1e−4; ** indicates *P*-value <1e−14; there is no significance between Day 60 and Day 90; *t*-test. (**B**) Two types of long term diet intervention effect on PWS patients: type1 individuals (GD02, GD03, GD15, GD40, GD42, GD43, GD47, GD51 and GD58) with gut microbiome functional capacity furthest removed from Day 0 at Day 90; type 2 individuals (GD04, GD18, GD52 and GD59) with gut microbiome functional capacity reversed at Day 90 toward their Day 0.

mi-faser facilitates novel functional discovery, while accelerating large-scale metagenomic analysis

The online service of *mi-faser* uses *clubber* (Cluster Load Balancer for Bioinformatics e-Resources (40)) for faster processing. To demonstrate our method's performance we obtained and analysed with *mi-faser,* 68 gut metagenomic datasets from a study of Chinese children affected by the PWS and treated via dietary intervention (Methods) (18). The analysis was automatically distributed to three clusters (640, 800, and 3400 cores with load-dependent access) by *clubber* via the *mi-faser* interface, for an average of 16 minutes of user-wait time (11.8 CPU hours) per metagenome.

Note that after NMDS Euclidean distance analysis of microbiomes of untreated individuals (Day0), four individuals (GD12, GD39, GD41 and GD50) were identified as outliers and removed (Methods). While we do not expect that all PWS-affected children share the same microbiome features, we felt that treatment effect and progression could be better evaluated from a narrow starting point.

For the remaining individuals (n = 13), it was clear that the dietary intervention significantly altered gut microbiome functionality (Supplementary Figure S7; Day 0 versus Day>0, *P*-value = 0.001, permanova test). More precisely, the intervention gradually increased the functional beta-diversity among the patients' gut microbiomes (Figure 4A; Supplementary Figure S7), which was in line with the results of the original study (18).

We further investigated the treatment progress of each patient using the Euclidean distance of the Day 30, 60 and 90 samples from the Day 0 sample of the same individual. Overall, the distances increased with the treatment progress (Day 30, 0.09 ± 0.02; Day 60, 0.16 ± 0.02; Day 90, 0.2 ± 0.03; Supplementary Figure S7), indicating the progressive changes of gut microbial functional potentials correlated with the diet time-line. Although Day 90 samples showed the highest dissimilarity from Day 0 samples in most cases, four patients (GD04, GD18, GD52 and GD59) reached the highest dissimilarity at Day 60, showing reversal of diet effects at Day 90 (Figure 4B). Follow-up studies on these differential trajectories could contribute to a more thorough

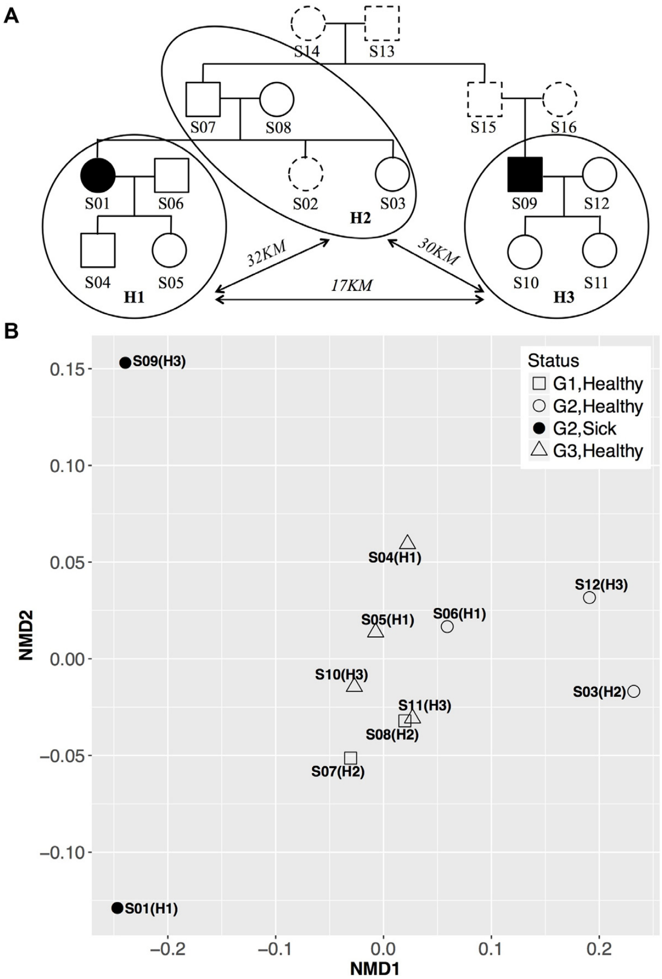

Figure 5. Functional capabilities of microbiomes of CD-affected individuals differ from healthy individuals and from each other. (**A**) The pedigree of the family in our study. Filled markers indicate CD affected individuals and empty markers are healthy individuals; dashed outline markers indicate individuals not included in this study. Individuals grouped by circles live in the same household. (**B**) The non-metric multidimensional scaling (NMDS) graph represents the distribution of individual microbiome functional profiles. Samples are labeled with identifiers (S1-S11) and household numbers (H1, H2, or H3, in parenthesis). Legend marker numbers (G1—grandparents, G2—parents, G3—children) represent generations, while marker shapes relate generations and CD status. Sick individuals (filled markers) localize separately from each other and from the cluster of healthy individuals (empty markers).

understanding of the effectiveness of the dietary intervention in PWS children.

mi-faser reveals microbial functions associated with Crohn's disease (CD)

We used our *mi-faser* pipeline (Figure 1) to analyse 11 microbiomes from individuals of the same extended family—two CD affected patients and nine first-degree relatives (Figure 5A). The members of this family live in three households that are no more than 32km apart from each other, with the CD affected individuals living in households 17 km away. No statistically significant distinction between functional profiles of individuals in the study was observed on the basis of generational or household differences (Figure 5B; *P*-value = 0.55 and 0.60 respectively, permanova test (41)). The nine healthy individuals shared highly similar microbiome functional profiles (rho, $\rho = 0.93 \pm 0.03$; Figure 5B; Supplementary Table S8). This finding is in line with previous studies that show that microbiome functional profiles across healthy individuals are more consistently maintained than bacterial species profiles (7). On the other hand, the microbiome functional profiles of the two CD patients are not only distinct from those of their healthy relatives (Figure 5B; $\rho = 0.75 \pm 0.11$; *P*-value = 0.02, permanova test), but also between themselves ($\rho = 0.72$; Figure 5B; Supplementary Table S8). Note that the former holds true even within the same household. In concert, these findings indicate that either there are different microbiome pathogenesis mechanisms of CD or that CD has a diverse impact on microbiome functionality.

We identified those E.C.s in our microbiomes whose abundance significantly changed in each patient compared to healthy individuals (Methods). S01 and S09 both have a large fraction of such E.C.s (45% and 31% respectively, sum of enriched and depleted, Supplementary Table S9). For example, nine E.C.s enriched in both S01 and S09 are annotated as rRNA methyltransferases (Supplementary Table S10), which are known to be essential for microbial response to environmental stresses (42). Another three E.C.s enriched in both patients are annotated as RNA pseudouridine synthase. RNAs with modified nucleotides, such as pseudouridine, have been shown to suppress host innate immune system (43). Thus, RNA modification may be an important bacterial strategy of surviving the CD-associated inflammation. We further explored these E.C.s to identify pathways uniquely altered in each patient; e.g. more than half of Biotin metabolism pathway E.C.s are altered in S01, while Xylene degradation is enriched only in S09 (Figure 6). There are also pathways that are similarly changed in both patients, i.e. they are enriched in the same E.C.s; for example, glutathione metabolism and lipopolysaccharide biosynthesis (Figure 6, Supplementary Figure S8). Given the distant microbiome functional profiles between S01 and S09 (Figure 5B), these similarities are unlikely to occur by chance. Glutathione is known to help bacteria survive oxidative stress, thus the enriched glutathione pathway could be a response to inflammation (44); a previous study has reported enrichment in abundance of genes associated with glutathione transportation in CD patients (7). However, the latter study (7) also suggested a decrease in propanoate and butanoate metabolism, both of which showed overall enrichment in S01 and S09 (Figure 6). Finally, to the best of our knowledge, the role of the lipopolysaccharide (LPS) biosynthesis pathway in CD patient microbiomes has not yet been reported. However, bacterial LPS is previously reported to increase intestinal tight junction permeability in mouse modules (45). Tight junctions normally form a selective seal between adjacent intestinal epithelial cells. Its increased permeability induces luminal pro-inflammatory molecules, resulting in sustained inflammation and tissue damage (46). Additionally, we also observed differences within individual pathway changes between patients. For example in the glycolysis/gluconeogenesis pathway, S01 is depleted in proteins necessary to convert glucose to pyru-

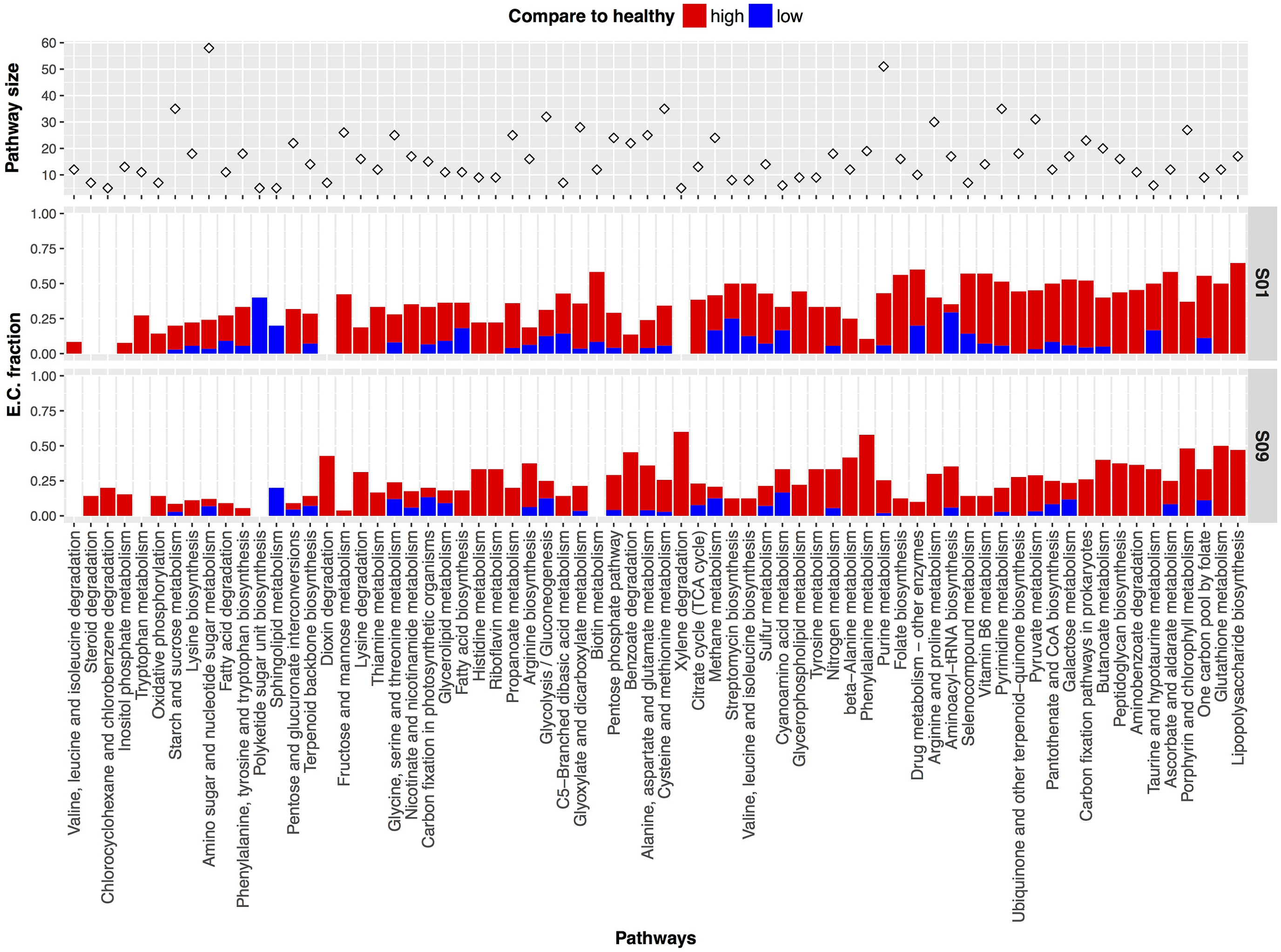

Figure 6. Enriched or depleted molecular pathways in microbiomes of CD-affected individuals. Changes in molecular pathways were obtained by counting the numbers of enriched or depleted E.C.s as compared to microbiome functional profiles of the healthy family members.

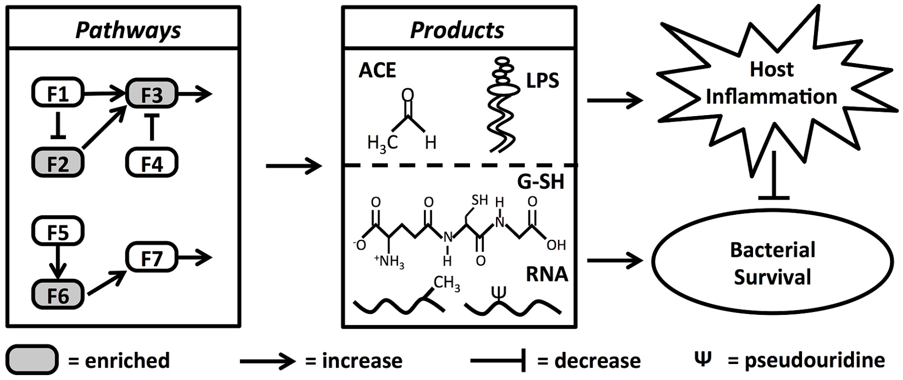

Figure 7. Microbial function shift in CD patients is involved in inflammation. Functions that are associated with inflammation inducers (acetaldehyde and lipopolysaccharide) are enriched in CD patient microbiomes, as are the functions that help bacteria survive inflammation conditions (glutathione metabolism, rRNA methytransferase and RNA pseudouridine synthase). Note that pathways above are toy examples for illustration purposes only; light gray nodes indicate enriched functions and white nodes indicate unchanged or undetected functions. Products are: ACE = acetaldehyde, LPS = lipopolysaccharide, G-SH = glutathione, RNA = RNAs with methylation or pseudouridine.

vate, while the pyruvate metabolism pathways are enriched (Supplementary Figure S9A). S09 shows a similar pattern, while enriching an alternative route from glyceraldehyde-3P to glycerate-3P (Supplementary Figure S9B). Interestingly, in both patients, most enriched E.C.s in pyruvate metabolism lead to acetaldehyde production (Supplementary Figure S9), a metabolite also known to induce tight junction disruption in intestinal epithelial cells (47). Thus, our result indicates the microbiome function shift in CD patients contributes to pathogenicity, while helps the bacteria survive host inflammation (Figure 7).

CONCLUSION

In this study, we compiled a 'clean' protein dataset with experimentally confirmed E.C. annotations (gold standard, GS-set), and trained the *faser* algorithm to optimise transfer of function annotation from reference proteins to short peptides translated from sequencing reads. The *faser* algorithm significantly outperforms PSI-BLAST in differentiating functions at high-resolution levels. It also offers ∼20% more annotations at comparable precision levels than the function annotation algorithm of MG-RAST. The (highly-parallelized and fast) *mi-faser* pipeline (*faser* in combination with GS*)* was able to identify, in BP oil spill data, unique candidate functions associated with oil-degradation, which were missed by the MG-RAST pipeline. Analysis of 68 metagenomic datasets from a dietary intervention study in PWS patients highlighted previously unseen individual-specific trajectories of functional changes in the gut microbiomes. Our pipeline also revealed that gastrointestinal microbiomes of related CD patients are functionally very different. We observed two types of functions enriched in CD patients: those that cause inflammation and those that help bacteria survive inflammatory stress; these may highlight the possible role of the microbiome in CD pathogenicity. Note that all *mi-faser* annotations, although highly informative, are based on the proteins making up the, currently limited, *GS-set*. On the other hand, *faser* itself is a robust read annotation algorithm that can be used with any reference database supplied. We also expect the growth in the number of proteins with experimentally verified functions to make our approach even more powerful in the near future.

AVAILABILITY

mi-faser is available online at http://services.bromberglab.org/mifaser/.

The standalone version of the pipeline, along with the *mi-faser* source code, is available at https://bitbucket.org/bromberglab/mifaser. The DOI for the source code used in this manuscript is https://doi.org/10.5281/zenodo.1045582, and the DOI for the current GS database is https://doi.org/10.5281/zenodo.1048268.

The fasta file for *GS-set* is available at http://bromberglab.org/sites/default/files/SOM_Data1_gold_standard.fasta.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Tamar Barkay, Max Haggblom, Yannick Mahlich, Alexandra Pushkar, Huan Qiu and Yanran Wang (all Rutgers University, New Brunswick, NJ) for many discussions and manuscript review. We thank Bill Abbott and Prentice Bisbal (both Rutgers) for technical support. We are grateful to Sonakshi Bhattacharjee (Technical University of Munich, TUM) for advice on the user interface. Particular thanks are due to Burkhard Rost (TUM) for his hospitality and valuable discussions. Last but not least, we thank all those who deposit their experimental data in public databases and those who maintain these databases.

Author contributions: C.Z. and Y.B. conceived and designed the experiments. C.Z., S.M. and P.V. performed the experiments. C.Z. and Y.B. analyzed the data. M.M. implemented the web service and made *mi-faser* publically available. Y.B. contributed materials/analysis tools. M.C.R., F.-A.H., M.T., W.L. and A.F. provided the Crohn's Disease data. G.W. and L.Z. provided the Prader–Willi syndrome data. C.Z. and Y.B. wrote the paper. All authors have seen and approved the final manuscript.

FUNDING

NSF CAREER [1553289 to Y.B. and C.Z.]; TU Munchen – Institute for advanced study Hans Fischer fellowship, funded by the German Excellence Initiative and the EU Seventh Framework Programme [291763 to Y.B.]; NIH/NIGMS [U01 GM115486 to Y.B., M.M., C.Z.]; The CD data collection was supported by the biobank popgen (the popgen 2.0 network is financed by the German Ministry for Education and Research [01EY1103]; German Research Foundation (DFG) Excellence Cluster 306/2, 'Inflammation at Interfaces' and by German Federal Ministry of Education and Research (BMBF) project 'SysINFLAME' [01ZX1306A]. Funding for open access charge: German Research Foundation (DFG) and Technical University of Munich within the funding programme Open Access Publishing.

Conflict of interest statement. None declared.

REFERENCES

1. Manichanh,C., Rigottier-Gois,L., Bonnaud,E., Gloux,K., Pelletier,E., Frangeul,L., Nalin,R., Jarrin,C., Chardon,P., Marteau,P. *et al.* (2006) Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut*, **55**, 205–211.
2. Dicksved,J., Halfvarson,J., Rosenquist,M., Jarnerot,G., Tysk,C., Apajalahti,J., Engstrand,L. and Jansson,J.K. (2008) Molecular analysis of the gut microbiota of identical twins with Crohn's disease. *ISME J.*, **2**, 716–727.
3. Frank,D.N., St. Amand,A.L., Feldman,R.A., Boedeker,E.C., Harpaz,N. and Pace,N.R. (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 13780–13785.
4. Frank,D.N., Robertson,C.E., Hamm,C.M., Kpadeh,Z., Zhang,T., Chen,H., Zhu,W., Sartor,R.B., Boedeker,E.C., Harpaz,N. *et al.* (2011) Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. *Inflamm. Bowel Dis.*, **17**, 179–184.
5. Sokol,H., Seksik,P., Furet,J.P., Firmesse,O., Nion-Larmurier,I., Beaugerie,L., Cosnes,J., Corthier,G., Marteau,P. and Doré,J. (2009) Low counts of Faecalibacterium prausnitziiin colitis microbiota. *Inflamm. Bowel Dis.*, **15**, 1183–1189.

6. Martinez-Medina,M., Aldeguer,X., Lopez-Siles,M., González-Huix,F., López-Oliu,C., Dahbi,G., Blanco,J.E., Blanco,J., Garcia-Gil,J.L. and Darfeuille-Michaud,A. (2009) Molecular diversity of Escherichia coli in the human gut: New ecological evidence supporting the role of adherent-invasive E. coli (AIEC) in Crohn's disease. *Inflamm. Bowel Dis.*, **15**, 872–882.
7. Morgan,X.C., Tickle,T.L., Sokol,H., Gevers,D., Devaney,K.L., Ward,D.V., Reyes,J.A., Shah,S.A., LeLeiko,N., Snapper,S.B. *et al.* (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.*, **13**, R79.
8. Aziz,R.K., Bartels,D., Best,A.A., DeJongh,M., Disz,T., Edwards,R.A., Formsma,K., Gerdes,S., Glass,E.M., Kubal,M. *et al.* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.
9. Markowitz,V.M., Chen,I.-M.A., Palaniappan,K., Chu,K., Szeto,E., Pillay,M., Ratner,A., Huang,J., Woyke,T., Huntemann,M. *et al.* (2014) IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.*, **42**, D560–D567.
10. Meyer,F., Paarmann,D., D'Souza,M., Olson,R., Glass,E., Kubal,M., Paczian,T., Rodriguez,A., Stevens,R., Wilke,A. *et al.* (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.
11. Schnoes,A.M., Brown,S.D., Dodevski,I. and Babbitt,P.C. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.*, **5**, e1000605.
12. Rodriguez-R,L.M., Overholt,W.A., Hagan,C., Huettel,M., Kostka,J.E. and Konstantinidis,K.T. (2015) Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill. *ISME J.*, **9**, 1928–1940.
13. Bairoch,A., Boeckmann,B., Ferro,S. and Gasteiger,E. (2004) Swiss-Prot: juggling between evolution and stability. *Brief. Bioinformatics*, **5**, 39–55.
14. EC,W. (1992) *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Academic Press, San Diego.
15. Furnham,N., Holliday,G.L., de Beer,T.A.P., Jacobsen,J.O.B., Pearson,W.R. and Thornton,J.M. (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.*, **42**, D485–D489.
16. Leinonen,R., Akhtar,R., Birney,E., Bower,L., Cerdeno-Tárraga,A., Cheng,Y., Cleland,I., Faruque,N., Goodgame,N., Gibson,R. *et al.* (2011) The European Nucleotide Archive. *Nucleic Acids Res.*, **39**, D28–D31.
17. Leinonen,R., Sugawara,H., Shumway,M. and on behalf of the International Nucleotide Sequence Database, C. (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.
18. Zhang,C., Yin,A., Li,H., Wang,R., Wu,G., Shen,J., Zhang,M., Wang,L., Hou,Y., Ouyang,H. *et al.* (2015) Dietary modulation of gut microbiota contributes to alleviation of both genetic and simple obesity in children. *EBioMedicine*, **2**, 968–984.
19. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
20. Joshi,N.A. and Fass,J.N. (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. http://www.citeulike.org/user/mvermaat/article/13260426.
21. Schmieder,R. and Edwards,R. (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLOS ONE*, **6**, e17288.
22. Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **318**, 595–608.
23. Wilke,A., Bischof,J., Harrison,T., Brettin,T., D'Souza,M., Gerlach,W., Matthews,H., Paczian,T., Wilkening,J., Glass,E.M. *et al.* (2015) A RESTful API for accessing microbial community data for MG-RAST. *PLOS Comput. Biol.*, **11**, e1004008.
24. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
25. Oliveros,J.C. (2007) VENNY. An interactive tool for comparing lists with Venn Diagrams. http://www.citeulike.org/user/hroest/article/6994833.
26. Harrell,F.E. Jr (2016) *Hmisc: Harrell Miscellaneous*. R package version 3.17-4.
27. Kruskal,J.B. (1964) Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, **29**, 115–129.
28. Oksanen,J., Friendly,M., Kindt,R., Legendre,P., McGlinn,D., Minchin,P.R., O'Hara,R.B., Simpson,G.L., Solymos,P., Henry,M. *et al.* (2016) *vegan: Community Ecology Package*. R package version 2.4-0.
29. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M., Bansal,P., Bridge,A.J., Poux,S., Bougueleret,L. and Xenarios,I. (2016) UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Methods Mol. Biol. (Clifton, N.J.)*, **1374**, 23–54.
30. Lam,H., Deutsch,E.W., Eddes,J.S., Eng,J.K., King,N., Stein,S.E. and Aebersold,R. (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*, **7**, 655–667.
31. Stein,S. (2012) Mass spectral reference libraries: an ever-expanding resource for chemical identification. *Anal. Chem.*, **84**, 7274–7282.
32. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
33. Schneider,R., de Daruvar,A. and Sander,C. (1997) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **25**, 226–230.
34. Radivojac,P., Clark,W.T., Oron,T.R., Schnoes,A.M., Wittkop,T., Sokolov,A., Graim,K., Funk,C., Verspoor,K., Ben-Hur,A. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods*, **10**, 221–227.
35. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
36. Wilke,A., Harrison,T., Wilkening,J., Field,D., Glass,E.M., Kyrpides,N., Mavrommatis,K. and Meyer,F. (2012) The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics*, **13**, 1–5.
37. Gil,R., Silva,F.J., Pereto,J. and Moya,A. (2004) Determination of the core of a minimal bacterial gene set. *Microbiol. Mol. Biol. Rev.: MMBR*, **68**, 518–537.
38. Sharifi,F. and Ye,Y. (2017) From gene annotation to function prediction for metagenomics. *Methods Mol. Biol. (Clifton, N.J.)*, **1611**, 27–34.
39. Nayfach,S., Bradley,P.H., Wyman,S.K., Laurent,T.J., Williams,A., Eisen,J.A., Pollard,K.S. and Sharpton,T.J. (2015) Automated and accurate estimation of gene family abundance from shotgun metagenomes. *PLOS Comput. Biol.*, **11**, e1004573.
40. Miller,M., Zhu,C. and Bromberg,Y. (2017) clubber: removing the bioinformatics bottleneck in big data analyses. *J. Integrative Bioinformatics*, **14**, doi:10.1515/jib-2017-0020.
41. Anderson,M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**, 32–46.
42. Baldridge,K.C. and Contreras,L.M. (2014) Functional implications of ribosomal RNA methylation in response to environmental stress. *Crit. Rev. Biochem. Mol. Biol.*, **49**, 69–89.
43. Durbin,A.F., Wang,C., Marcotrigiano,J. and Gehrke,L. (2016) RNAs containing modified nucleotides fail to trigger RIG-I conformational changes for innate immune signaling. *mBio*, **7**, e00833-16.
44. Masip,L., Veeravalli,K. and Georgiou,G. (2006) The many faces of glutathione in bacteria. *Antioxid. Redox Signal.*, **8**, 753–762.
45. Guo,S., Al-Sadi,R., Said,H.M. and Ma,T.Y. (2013) Lipopolysaccharide causes an increase in intestinal tight junction permeability in vitro and in vivo by inducing enterocyte membrane expression and localization of TLR-4 and CD14. *Am. J. Pathol.*, **182**, 375–387.
46. Lee,S.H. (2015) Intestinal permeability regulation by tight junction: implication on inflammatory bowel diseases. *Intestinal Res.*, **13**, 11–18.
47. Atkinson,K.J. and Rao,R.K. (2001) Role of protein tyrosine phosphorylation in acetaldehyde-induced disruption of epithelial tight junctions. *Am. J. Physiol. - Gastrointest. Liver Physiol.*, **280**, G1280.