MICROBIAL GENOMICS

RESEARCH ARTICLE

Pascal Andreu *et al.*, *Microbial Genomics* 2020;6

DOI 10.1099/mgen.0.000373

Computational genomic discovery of diverse gene clusters harbouring Fe-S flavoenzymes in anaerobic gut microbiota

Victòria Pascal Andreu[1], Michael A. Fischbach[2,*] and Marnix H. Medema[1,*]

Abstract

The gut contains an enormous diversity of simple as well as complex molecules from highly diverse food sources, together with host-secreted molecules. This presents a large metabolic opportunity for the gut microbiota, but little is known about how gut microbes are able to catabolize this large chemical diversity. Recently, Fe-S flavoenzymes were found to be key in the transformation of bile acids, catalysing the key step in the 7α-dehydroxylation pathway that allows gut bacteria to transform cholic acid into deoxycholic acid, an exclusively microbe-derived molecule with major implications for human health. While this enzyme family has also been implicated in a limited number of other catalytic transformations, little is known about the extent to which it is of more global importance in gut microbial metabolism. Here, we perform a large-scale computational genomic analysis to show that this enzyme superfamily has undergone a remarkable expansion in Clostridiales, and occurs throughout a diverse array of >1000 different families of putative metabolic gene clusters. Analysis of the enzyme content of these gene clusters suggests that they encode pathways with a wide range of predicted substrate classes, including saccharides, amino acids/peptides and lipids. Altogether, these results indicate a potentially important role of this protein superfamily in the human gut, and our dataset provides significant opportunities for the discovery of novel pathways that may have significant effects on human health.

DATA SUMMARY

Supporting information for this article can be found in the Zenodo repository (https://zenodo.org/) with the following DOI: 10.5281/zenodo.3784213.

INTRODUCTION

The gene set of the human gut microbiota vastly exceeds the human gene repertoire [1, 2], which allows microbes to complement human metabolism by degrading undigested polysaccharides, lipids and peptides that reach the large intestine [3]. For instance, saccharolytic bacteria can ferment carbohydrates to produce short chain fatty acids, beneficial metabolites that promote health [4]. Nevertheless, gut bacteria also produce many molecules involved in microbe–microbe interactions and microbe–host interactions that can have detrimental effects instead [5]. An example of a harmful diet-derived metabolite is trimethylamine, an amine that can be synthesized from choline or carnitine by certain gut bacteria, and which has been associated with cardiovascular and renal disease [6]. Thus, the identification of these molecules and the elucidation of their production pathways are crucial to assess the causes and consequences of certain microbiome-associated phenotypes.

Another example of exclusively microbiome-derived molecules is constituted by secondary bile acids, such as deoxycholic acid (DCA) and lithocholic acid (LCA). While the primary bile acids cholic acid and chenodeoxycholic acid are synthesized by the liver [7], they are transformed into DCA or LCA by colonic bacteria during enterohepatic circulation [8]. These molecules have been proposed to act as inhibitors of *Clostridium difficile* outgrowth [9], as well as to induce the development of colon cancer [10, 11] and cholesterol gallstone disease [12]. The main bacterial pathway in charge of the 7α-dehydroxylation needed to produce them constitutes a multi-step biochemical reaction that can be

Received 29 February 2020; Accepted 15 April 2020; Published 14 May 2020

Author affiliations: [1]Bioinformatics Group, Wageningen University, Wageningen, The Netherlands; [2]Department of Bioengineering and ChEM-H, Stanford University, Stanford, USA.

***Correspondence:** Michael A. Fischbach, fischbach@fischbachgroup.org; Marnix H. Medema, marnix.medema@wur.nl

Keywords: flavoenzymes; *Clostridium*; gut microbiota; gene clusters.

Abbreviations: DCA, deoxycholic acid; GCF, gene cluster family; MAG, metagenome-assembled genome; MGC, metabolic gene cluster.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Two supplementary tables are available with the online version of this article.

accomplished by bacteria harbouring the bile-acid-inducible (*bai*) operon [13]. Most of the bacteria capable of carrying out this reaction are anaerobes that are part of the *Clostridium* cluster XIVa [13]. The pathway encoded by the *bai* operon was recently elucidated by Funabashi *et al.* [14]. The authors showed that the key step is performed by the BaiCD enzyme, an Fe-S flavoenzyme that oxidizes 3-oxo-cholyl-CoA to 3-oxo-4,5-dehydrocholyl-CoA. Later, BaiH (also an Fe-S flavoenzyme), BaiCD and BaiA2 act again on the molecule to finally produce DCA. Importantly, the participation of Fe-S flavoenzymes in the key reductive steps of the pathway is consistent with a role for this pathway in using primary bile acids as terminal electron acceptors for an anaerobic electron transport pathway, constituting a unique metabolic niche within the gut community.

In addition to the BaiCD and BaiH enzymes, a few other members of the Fe-S flavoenzyme superfamily have been previously shown to play similar crucial roles in the redox metabolism of catalytic transformations. For instance, L-phenylalanine fermentation via a Stickland reaction, where cinnamate is reduced to 3-phenylpropionate, has been shown to be performed by a cinnamate reductase that is a member of the Fe-S flavoenzyme superfamily [15]. Additional Fe-S flavoenzyme representatives with experimentally characterized functions include a 2,4-dienoyl-CoA reductase implicated in fatty acid β-oxidation [16] and a trimethylamine dehydrogenase involved in trimethylamine degradation [17].

The fact that these enzymes have been shown to facilitate the shuttling of electrons from the membrane to diverse organic terminal electron acceptors, enabling an anaerobic electron transport chain, led us to hypothesize that they might play a more widespread role in the microbial catabolism of the diversity of complex substrates available in the gut. Here, we provide an in-depth computational genomic study of the Fe-S flavoenzyme superfamily and its presence across genomes of gut microbiome-related bacteria. Using phylogenomic analyses, we show that this enzyme superfamily comprises a large sequence diversity and has particularly undergone strong evolutionary expansion in the class Clostridia, emphasizing its possible implication in different metabolic reactions in the gut. Large amounts of strain-level variation between genomes indicate that the pathways involved likely facilitate ecologically specialized functions. Finally, analysis of the enzyme content in their surrounding operons and gene clusters uncovers a wide array of putative catabolic gene clusters associated with the breakdown of diverse substrates, which provides a rich resource for uncovering novel pathways in the human microbiome and beyond.

Impact Statement

In the last decade, the human gut microbiota has been shown to be of high importance to human health in various ways. Gut microbes also have unique metabolic capabilities beyond those of the host. The gut contains a large diversity of simple as well as complex molecules from highly diverse food sources, together with host-secreted molecules, which presents a large metabolic opportunity for the gut microbiota. However, little is known about how gut microbes are able to metabolize this large chemical diversity. In this article, we report the presence of highly diverse metabolic gene clusters (MGCs) encoding Fe-S flavoenzymes across gut microbial genomes. Evolutionarily, a strong diversification has occurred in Clostridiales, leading to the origination of gene clusters that are predicted to encode pathways involved in the processing of highly diverse substrates. This work opens up new territory for exploring the functionality of the physiological roles of gut microbial metabolism, and how they relate to microbiome-associated phenotypes. Additionally, it provides a template for future discovery efforts to identify novel specialized primary MGCs based on enzyme-family-centred phylogenomic analysis.

METHODS

Identification of members of the Fe-S flavoenzyme superfamily

BaiCD contains two Pfam domains, Oxidored_FMN (PF00724) and Pyr_redox_2 (PF07992). We used hmmsearch (HMMER HMMER3.1b2, February 2015; http://hmmer.org/) to identify protein sequences harbouring both domains from two databases: all bacterial sequences in GenBank (complete and draft genomes, >100000 entries) and all sequences in RefSeq belonging to the class Clostridia (>2500 entries), as some of the Clostridia genomes in GenBank lack gene coordinate annotations. Proteins with a sequence *E* value $\leq 1\times10^{-5}$ for both domains were deemed hits. The resulting set of 49437 Fe-S flavoenzyme superfamily members was used for the analyses described in the next sections.

Reconstruction of an Fe-S flavoenzyme protein similarity network

The amino acid sequences of the 49870 Fe-S flavoenzyme superfamily members were clustered using MMseqs2 [18], setting the minimum identity to 0.9. From the 1694 groups in the resulting protein similarity network, we picked five representatives of each cluster (randomly chosen when the cluster contained more than five nodes). Approximately 2500 singletons (families of size one) were also included in the subsequent analysis. We aligned the protein sequences to the Oxidored_FMN and Pyr_redox_2 Pfam domains using hmmalign (HMMER HMMER3.1b2, February 2015; http://hmmer.org/), removed the unaligned and indel regions, merged the alignments of the two domains and reconstructed a phylogeny using FastTree [19]; the midpoint root of the tree was calculated using the phytools package in R [20]. The resulting phylogenetic tree was annotated with interactive Tree of Life (iTOL) [21].

Computational analysis of the prevalence and phylogenetic distribution of Fe-S flavoenzymes

We investigated the taxonomic distribution of Fe-S flavoenzyme genes and their variability in copy number; for simplicity, the number of hits per genome assembly was used as a metric of the copy number per genome. We reconstructed a phylogenetic tree of genomes in the class Clostridia using the following procedure: Clostridia genome assemblies harbouring at least one Fe-S flavoenzyme gene were downloaded and quality-filtered using the N50 statistic, setting the threshold at 50 kb. The 16S rRNA sequences from the high-quality scaffolds were predicted using Barrnap version 0.9 (https://github.com/tseemann/barrnap). 16S rRNA sequences were aligned with Clustal Omega [22], and FastTree was used to infer an approximate maximum-likelihood phylogenetic tree. Finally, iTOL was used to display and annotate the phylogenetic tree. Two 16S sequences from *Bacillus subtilis* subsp. *subtilis* (GenBank accession no. NR_102783.2) and *Streptococcus agalactiae* DNF00839 (GenBank accession no. KU726685.1) were used as outgroups to root the tree.

Identification of Fe-S flavoenzymes from a collection of human gut metagenome-assembled genomes (MAGs)

The 1952 uncultured bacterial genomes from Almeida *et. al* [23] were used to evaluate the Fe-S flavoenzyme diversity found in human gut bacterial communities. To this end, the genomic nucleotide sequences of the MAGs were downloaded using the ENA (European Nucleotide Archive) accession number ERP108418. Next, the protein sequences were predicted using Prodigal V2.6.3 (February 2016), to further use hmmsearch against the Oxidored_FMN and Pyr_redox_2 domains following the same procedure as stated above. Thus, the number of predicted Fe-S flavoenzymes (protein sequences with sequence E value ≤1×10^{-5}) for each MAG was calculated to evaluate their Fe-S flavoenzyme copy numbers. Moreover, in order to analyse the genomic context of the Fe-S flavoenzyme-encoding genes found to be encoded in the genome of an uncultured Clostridiales bacterium (accession no. GCA_900546485.1), a genomic feature table was created from the Prodigal annotations. Subsequently, we used the same script to identify gene clusters as explained below. Finally, metabolic gene cluster (MGC) protein sequences were scanned with hmmscan to identify Pfam domains.

Analysis of the genomic context of Fe-S flavoenzymes

For the subset of Fe-S flavoenzymes from the class Clostridia found in the RefSeq database, we inspected their genomic context by identifying 'neighbouring genes' that met the following criteria: they were encoded on the same strand as the Fe-S flavoenzyme and located within a maximum intergenic distance of 400 bp between subsequent genes. A script was used to parse the feature tables from the genome assemblies found to harbour at least one copy of the Fe-S flavoenzyme gene, from which the starting and ending coordinates of the cluster were extracted. Subsequently, the corresponding GenBank file of the flanking region was downloaded. The resulting GenBank file collection was used as an input for BiG-SCAPE [24], which groups MGCs into families. The output networks were visualized using the BiG-SCAPE interactive visualization tool.

RESULTS AND DISCUSSION

The Fe-S flavoenzyme phylogeny includes many unexplored clades involved in diverse catalytic reactions

In order to assess the distribution of the Fe-S flavoenzymes along the bacterial kingdom, we scanned 111651 bacterial genomes (see Methods: *Identification of members of the Fe-S flavoenzyme superfamily*). Across this dataset, we identified 49870 Fe-S flavoenzymes that belong to more than 20 bacterial phyla. Some bacterial genomes encode remarkable numbers of Fe-S flavoenzymes: e.g. the genome of the firmicute *Sporobacter termitidis* encodes no fewer than 18 Fe-S flavoenzymes. There are also representatives from Actinobacteria (including *Eggerthella* sp. *YY7918* and *Rhodococcus* sp. *SC4*) that possess nine different flavoenzymes. On average, Firmicutes have 1.81±1.18 protein copies per genome, whereas Proteobacteria and Actinobacteria have 1.51±0.83 and 1.10±0.50 copies, respectively. These numbers exclude genomes that do not encode any flavoenzyme, representing 89, 55.5 and 46.8% of the genomes from Firmicutes, Proteobacteria and Actinobacteria, respectively. Together, these data suggest that Fe-S flavoenzymes may play important roles in multiple bacterial phyla, and particularly in Firmicutes.

To contextualize these numbers of Fe-S flavoenzyme genes with their evolutionary history and functional diversification, we then conducted a comprehensive phylogenetic analysis of this enzyme family. We obtained a collection of 8097 non-redundant Fe-S flavoenzymes by selecting representatives from MMseqs2-derived [18] sequence clusters. From this dataset, we reconstructed an approximate maximum-likelihood phylogeny (Fig. 1) (see also Methods: Re*construction of an Fe-S flavoenzyme protein similarity network*). Within this phylogeny, we were able to annotate eight clades based on experimentally characterized proteins documented in UniProt to date; the vast majority of clades have unknown functions. Given the size of this protein superfamily, the number of unexplored clades and its functional diversity, this phylogeny constitutes a road map to uncover novel clades of enzymes that could potentially aid in the characterization of yet undiscovered catabolic pathways.

An evolutionary expansion of Fe-S flavoenzyme in Clostridia

As reported in the previous section, the phylum Firmicutes stands out for the high Fe-S flavoenzyme copy numbers found in genomes of the species that belong to it. For this reason, we investigated the flavoenzyme copy number variation in *Clostridium* along with other microbial taxa commonly found in the gut: Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria. The phylogenetic analysis already suggested

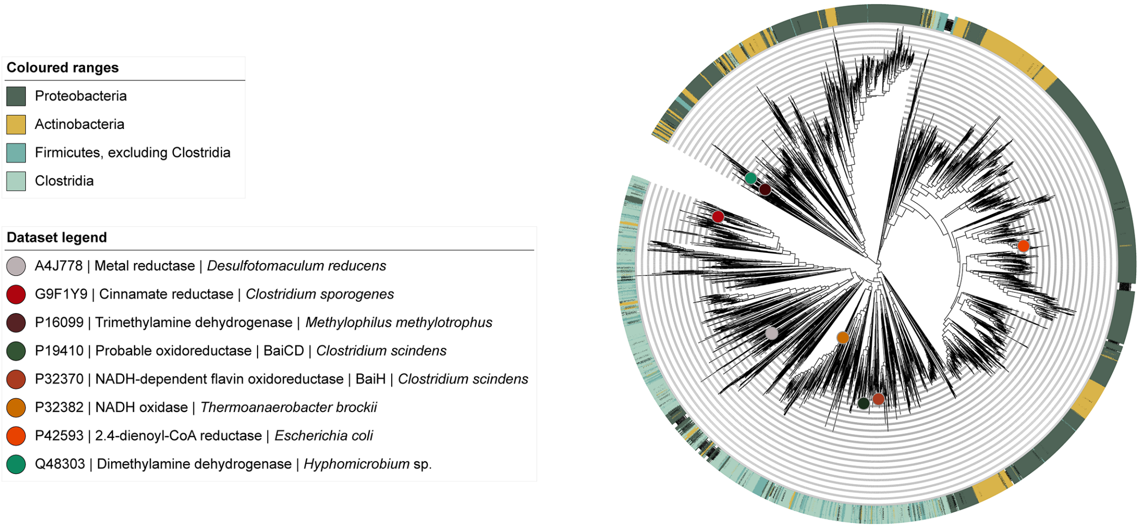

Fig. 1. Fe-S flavoenzyme superfamily phylogeny. The phylogeny was reconstructed with FastTree from 8105 non-redundant protein sequences across the bacterial kingdom, including the 8 experimentally characterized proteins from UniProt (https://www.uniprot.org/uniprot/?query=PF07992+and+PF00724+and+reviewed%3Ayes&sort=score, accessed on 01/02/2020). All sequences used as input contained both the Oxidored_FMN and Pyr_redox_2 Pfam domains. The eight experimentally characterized proteins from UniProt are indicated as circles on the tree to functionally annotate clades.

a large evolutionary expansion of Fe-S flavoenzymes in Clostridia, with this bacterial class occupying about a third of the non-redundant superfamily phylogeny. Indeed, quantitative analysis of the class Clostridia confirms that it has undergone a massive expansion when comparing the Fe-S flavoenzyme copy number with other taxa (Fig. 2). The phyla Actinobacteria and Bacteroidetes seem to harbour considerably fewer Fe-S flavoenzymes. As expected, we found that among genomes from Clostridia, there is a high percentage encoding at least one Fe-S flavoenzyme (71.58%). In contrast, less than half of the genomes in our database belonging to Enterobacteriaceae, Bacilli and *Bacteroides* encode a flavoenzyme (49.92, 7.64 and 3.19%, respectively). Strikingly, *Eggerthel*la seems to be an exception to this pattern, as 10 genomes from different *Eggerthella* species encode 38 Fe-S flavoenzymes, with 71.43% of the genomes with at least one count.

Unexpected amounts of strain variation of flavoenzyme counts between genomes

In microbial ecosystems, strain-specific traits often confer different characteristics to otherwise almost identical bacteria, which allow them to thrive in specific conditions [25–27]. Therefore, we assessed Fe-S flavoenzyme copy number variation in order to know whether such enzymes are responsible for specialized functionalities that allow microbes to adapt to unique ecological niches. Indeed, we found that Fe-S flavoenzyme copy number distribution along Clostridia genomes often shows signs of strain-specificity, as can be seen in the *C. difficile* clade and Clostridia clade II (Fig. 3). In the Clostridia phylogeny (Fig. 3), the number of copies ranges from 1 to 18 per genome, with a *Sporobacter termitidis* genome being the one with highest counts. Another bacterium with a high Fe-S flavoenzyme copy number is *Clostridium scindens* (Clostridia clade I), whose genome encodes up to 10 distinct members of the family (including BaiCD and BaiH). In comparison, other gut-related genera such as *Faecalibacterium* (clade IV, ~12 o'clock), *Roseburia* (clade I, ~5 o'clock) and *Anaerostipes* (clade I, ~4 o'clock) show quite different Fe-S flavoenzyme copy number profiles, harbouring on average 0.3, 0.6 and 2.4 copies per genome, respectively. Overall, these results confirm that indeed, for specific clades of Clostridia, Fe-S flavoenzymes are likely to play an important role in their primary metabolism. Moreover, they evince the importance of profiling the microbiome at high (or strain) resolution; 16S amplicon sequencing (and especially out-level clustering) would not be able to distinguish between bacteria with different metabolic repertoires. The differential metabolic capabilities of various strains are often due to very subtle changes in their genomes [28]. Therefore, studying these bacterial strains that differ in terms of their Fe-S flavoenzyme copy numbers, as well as the genomic contexts where these enzymes are encoded, could help to uncover strain-specific traits that play a role in conferring microbiome-associated host phenotypes.

Analysis of Fe-S flavoenzyme counts in reconstructed MAGs from the human gut microbiome

In order to evaluate the Fe-S flavoenzyme prevalence in the gut microbiota, we collected 1952 uncultured MAGs belonging to different taxonomic groups and reconstructed from 11850 human gut microbiomes by Almeida *et al.* [23] [see Methods: *Identification of Fe-S flavoenzymes from a collection of human gut metagenome-assembled genome* (*MAGs*)]. From

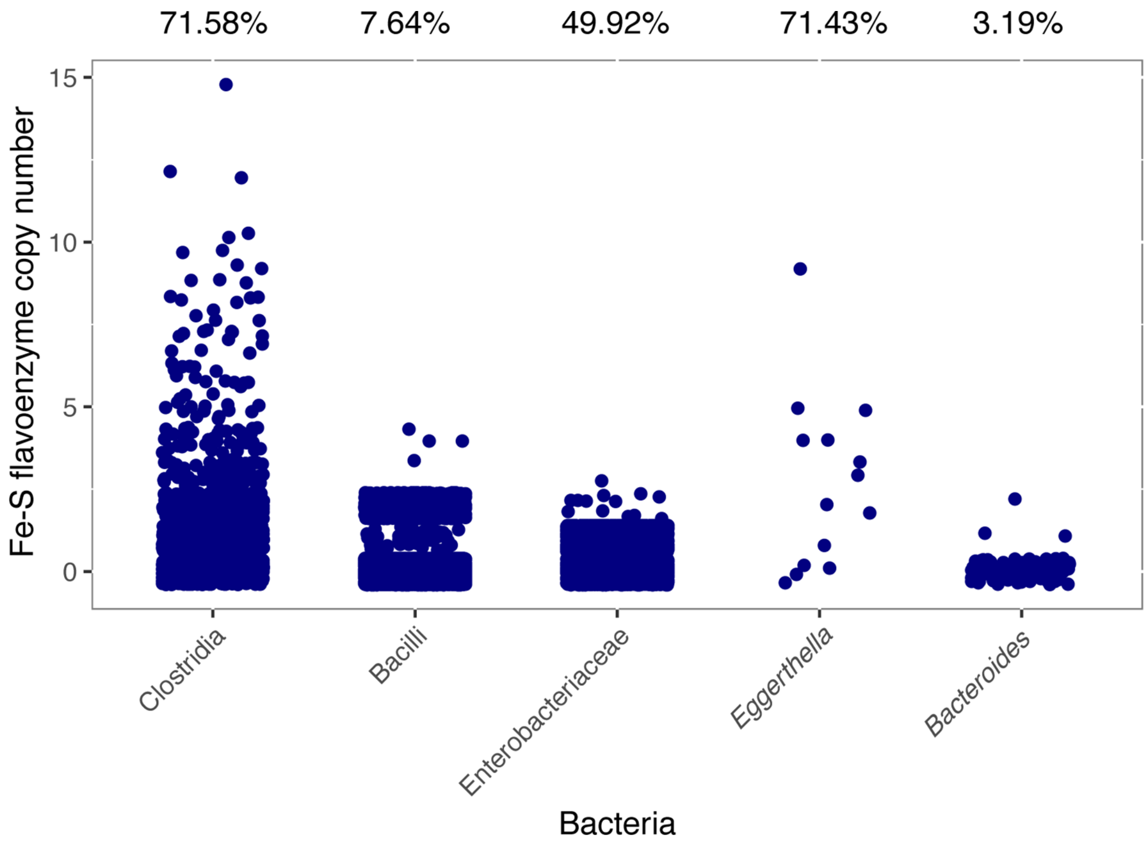

Fig. 2. Expansion of the Fe-S flavoenzyme superfamily in Clostridia when compared to other gut-related taxa. Scatterplot representing the Fe-S flavoenzyme copy number across some gut-related bacteria. The taxonomic groups cover the five major phyla from the gut: Firmicutes (Clostridia and Bacilli), Proteobacteria (Enterobacteriaceae), Actinobacteria (*Eggerthella*) and Bacteroidetes (*Bacteroides*). The representative candidates used for this analysis were gathered by discarding genomes whose GenBank files were stated to be isolated from sources such as pig, insects, rumen, sludge, water or soil. The percentages above the graph represent how many of the genomes harbour at least one Fe-S flavoenzyme-encoding gene out of the total number of genomes from that taxon. The copy numbers reported are discrete numbers, but are visually distributed in order to allow showing how many genomes were found for each copy number.

the protein sequences encoded in these MAGs, we identified a total of 1602 Fe-S flavoenzymes that were found to be encoded across 771 of the genomes. Later, to determine which organisms encode at least one flavoenzyme, we ranked the MAGs assigned to a taxonomic lineage annotation provided by Almeida *et al.* [23] based on the number of predicted Fe-S flavoenzymes. The top five taxa with highest counts belong to either Actinobacteria, in particular *Collinsella* sp., or Firmicutes, with 327 and 220 Fe-S flavoenzyme counts, respectively (Table S1, available with the online version of this article). This is in agreement with our previous findings that these taxa are two major contributors to the overall bacterial Fe-S flavoenzyme diversity, as shown in Fig. 1. Moreover, we also evaluated the Fe-S flavoenzyme copy number for each MAG, and surprisingly, the genome of an uncultured Clostridiales bacterium (accession no. GCA_900546485.1) was found to encode 22 different flavoenzymes, the highest copy number found in this study. This finding supports the evidence that the flavoenzyme superfamily has greatly expanded in the class Clostridia, but also indicates the remarkable role these enzymes may play in the gut microbiome and in particular in the bacteria with higher Fe-S flavoenzyme counts.

Analysis of gene neighbourhoods leads to the identification of a wide range of families of putative catabolic gene clusters involved in the breakdown of diverse biomolecules

We next explored the likely functionalities of flavoenzymes in Clostridia, by investigating the genomic contexts of the flavoenzyme-encoding genes. With that aim, we gathered a collection of 3158 non-redundant gene clusters, all of them having in common the presence of a Fe-S flavoenzyme-encoding gene (see Methods: *Analysis of the genomic context of Fe-S flavoenzymes*). These gene clusters could be grouped into 849 gene cluster families (GCFs). Interestingly, Fe-S flavoenzymes are found in a strikingly diverse array of GCFs. The examination of the corresponding gene-cluster-encoding proteins helped identifying more than 31 relevant Pfam domains that we categorized to be involved in four major types of metabolism (Table S2). Thus, the profiling of this

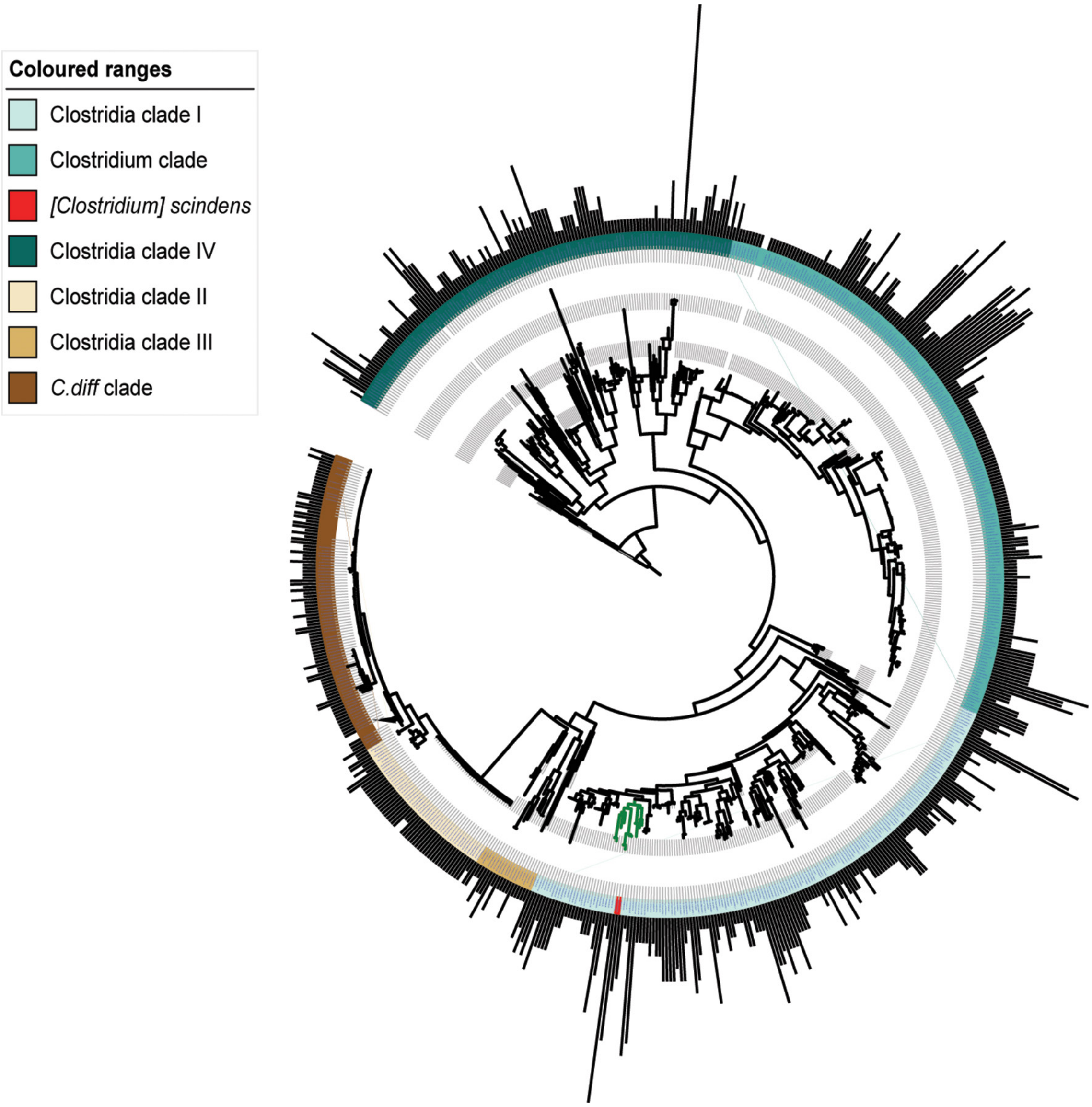

Fig. 3. Strain- and species-specific copy numbers of Fe-S flavoenzymes across Clostridia. A circular phylogenetic tree showing bacterial strains in the class Clostridia. The outer ring of bars represents the copy number of Fe-S flavoenzymes in each strain, indicating a high degree of variability among strains and species. The phylogeny includes closely related strains whose genomes encode diverse numbers of Fe-S flavoenzymes, ranging from 1 to 18. Coloured strips represent the taxonomic entities found within the class Clostridia. Two clades have been collapsed in order to remove almost identical genomes in the tree (shown as a discontinuity in the bar plots): the *C. difficil*e clade, Clostridia clade II and the *Clostridium* clade.

collection based on the presence of these protein domains within Fe-S-flavoenzyme-encoding gene clusters allowed us to predict putative functions for 175 GCFs. Specifically, 80 distinct GCFs are predicted to function in saccharide catabolism, 83 in peptide/amino acid catabolism, 28 in nucleotide catabolism and 14 in lipid catabolism. In Fig. 4, we highlight 11 gene clusters plus the *bai* operon to show examples of distinct cluster architectures that, based on their enzyme-encoding gene contents, we predict to be capable of processing a diverse array of substrates. Importantly, most of the Fe-S-flavoenzyme-containing gene clusters also harbour a major facilitator superfamily transporter, indicating that they might process a diffusible substrate; also, many include proteins with flavodoxin or electron transfer flavoprotein domains, consistent with the possibility that these pathways are coupled to electron transport chains. Moreover, we also explored the genomic neighbourhood of the 22 Fe-S flavoenzyme-encoding genes found in the uncultured Clostridiales bacterium from the MAGs collection, using the same methodology. One particularly interesting gene cluster is highlighted in Fig. 4, which is putatively involved in carbohydrate catabolism, encoding an α-amylase (PF00128), a

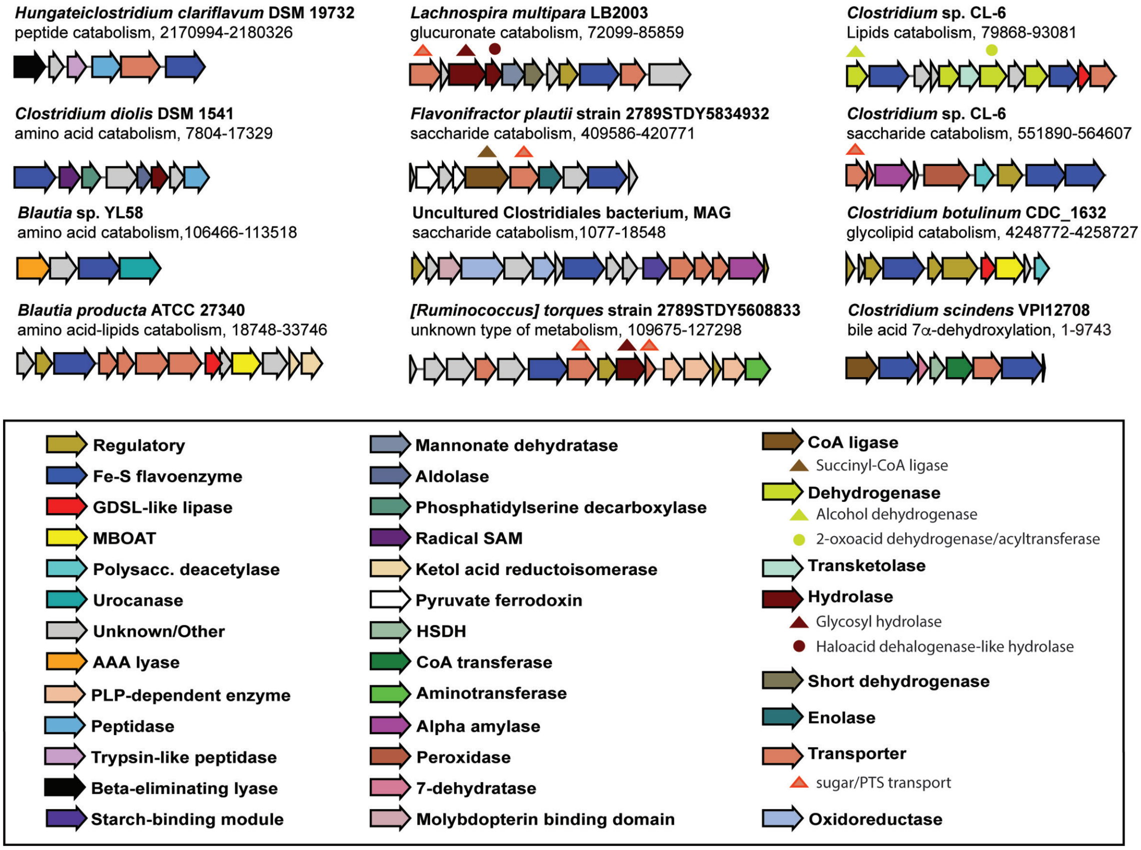
Hungateiclostridium clariflavum DSM 19732
peptide catabolism, 2170994-2180326
Lachnospira multipara LB2003
glucuronate catabolism, 72099-85859
Clostridium sp. CL-6
Lipids catabolism, 79868-93081
Clostridium diolis DSM 1541
amino acid catabolism, 7804-17329
Flavonifractor plautii strain 2789STDY5834932
saccharide catabolism, 409586-420771
Clostridium sp. CL-6
saccharide catabolism, 551890-564607
Blautia sp. YL58
amino acid catabolism,106466-113518
Uncultured Clostridiales bacterium, MAG
saccharide catabolism,1077-18548
Clostridium botulinum CDC_1632
glycolipid catabolism, 4248772-4258727
Blautia producta ATCC 27340
amino acid-lipids catabolism, 18748-33746
[Ruminococcus] torques strain 2789STDY5608833
unknown type of metabolism, 109675-127298
Clostridium scindens VPI12708
bile acid 7α-dehydroxylation, 1-9743
Regulatory
Fe-S flavoenzyme
GDSL-like lipase
MBOAT
Polysacc. deacetylase
Urocanase
Unknown/Other
AAA lyase
PLP-dependent enzyme
Peptidase
Trypsin-like peptidase
Beta-eliminating lyase
Starch-binding module
Mannonate dehydratase
Aldolase
Phosphatidylserine decarboxylase
Radical SAM
Ketol acid reductoisomerase
Pyruvate ferrodoxin
HSDH
CoA transferase
Aminotransferase
Alpha amylase
Peroxidase
7-dehydratase
Molybdopterin binding domain
CoA ligase
Succinyl-CoA ligase
Dehydrogenase
Alcohol dehydrogenase
2-oxoacid dehydrogenase/acyltransferase
Transketolase
Hydrolase
Glycosyl hydrolase
Haloacid dehalogenase-like hydrolase
Short dehydrogenase
Enolase
Transporter
sugar/PTS transport
Oxidoreductase

Fig. 4. Diversity of Fe-S-flavoenzyme-encoding catabolic gene clusters from Clostridia. Gene clusters from Clostridia that contain a Fe-S flavoenzyme are numerous and diverse. Eleven examples (including the *bai* operon) predicted to be involved in peptide/amino acid catabolism, saccharide catabolism, lipid catabolism and unknown types of metabolism have been picked from a collection of 849 GCFs and one from the MAGs collection. Genes are coloured by predicted function.

starch-binding domain (PF16738), an alanine dehydrogenase (PF01262) and a probable molybdopterin-binding domain (PF00994), among others.

Overall, these findings suggest that Fe-S flavoenzymes play a far more expansive role in anaerobic metabolism in the human gut than was previously known. Moreover, the predicted diversity in function indicates that these oxidoreductases may be in charge of catalysing the transformation of a large variety of substrates, conferring to these bacteria specialized primary metabolic capabilities that may allow them to colonize specific micro-niches in the gut.

Conclusion

The metabolic potential of gut bacteria greatly exceeds the genetic potential of the human host. Therefore, bacteria can benefit from utilizing a diverse range of substrates that reach the digestive tract. Despite this, the mechanisms by which anaerobic bacteria catalyse these reactions are largely unknown. Here, we show that Fe-S flavoenzymes are likely to play a significant role in catalysing key redox steps within diverse catabolic pathways. Specifically, we found a remarkable expansion of Fe-S flavoenzyme copy numbers in the class Clostridia when compared to other gut-related bacteria. The strain-specificity of Fe-S-flavoenzyme-encoding gene cluster repertoires indicates that these enzymatic pathways may allow bacteria to specialize in the catabolism of specific dietary or host-derived molecules. We present a rich dataset of gene clusters that are candidates for detailed biochemical studies. Moreover, the presence or expression of these GCFs could be used as features alongside standard metabolic pathway annotations to assess whether their presence could explain variation in health/disease phenotypes, or whether they can explain variation observed in gut metabolomes [29, 30]. All in all, given the importance of profiling the gut microbiome from a functional point of view, this study provides new ways of exploiting the genomic information present in public repositories, and provides a template for genomic exploration studies centred on key enzyme families to further understand the metabolic potential of the gut microbiome.

Funding information
This work was supported by the Chan-Zuckerberg Biohub (M.A.F.), and the U.S. Defense Advanced Research Projects Agency's Living Foundries program - award HR0011-15-C-0084 (M.A.F. and V.P.A.).

Conflicts of interest

M.A.F. is a co-founder and director of Federation Bio, a co-founder of Revolution Medicines, and a member of the scientific advisory board of NGM Biopharmaceuticals. M.H.M. is a co-founder of Design Pharmaceuticals, and a member of the scientific advisory board of Hexagon Bio.

References

1. Rowland I, Gibson G, Heinken A, Scott K, Swann J *et al*. Gut microbiota functions: metabolism of nutrients and other food components. *Eur J Nutr* 2018;57:1–24.
2. Koppel N, Balskus EP. Exploring and understanding the biochemical diversity of the human microbiota. *Cell Chem Biol* 2016;23:18–30.
3. Rooks MG, Garrett WS. Gut microbiota, metabolites and host immunity. *Nat Rev Immunol* 2016;16:341–352.
4. Koh A, De Vadder F, Kovatcheva-Datchary P, Bäckhed F. From dietary fiber to host physiology: short-chain fatty acids as key bacterial metabolites. *Cell* 2016;165:1332–1345.
5. Donia MS, Fischbach MA. Small molecules from the human microbiota. *Science* 2015;349:1254766.
6. Rath S, Rud T, Pieper DH, Vital M. Potential TMA-producing bacteria are ubiquitously found in Mammalia. *Front Microbiol* 2019;10:2966.
7. Wahlström A, Sayin SI, Marschall H-U, Bäckhed F. Intestinal crosstalk between bile acids and microbiota and its impact on host metabolism. *Cell Metab* 2016;24:41–50.
8. Ridlon JM, Kang D-J, Hylemon PB. Bile salt biotransformations by human intestinal bacteria. *J Lipid Res* 2006;47:241–259.
9. Buffie CG, Bucci V, Stein RR, McKenney PT, Ling L *et al*. Precision microbiome reconstitution restores bile acid mediated resistance to *Clostridium difficile*. *Nature* 2015;517:205–208.
10. Bayerdörffer E, Mannes GA, Richter WO, Ochsenkühn T, Wiebecke B *et al*. Increased serum deoxycholic acid levels in men with colorectal adenomas. *Gastroenterology* 1993;104:145–151.
11. Bernstein H, Bernstein C, Payne C, Dvorakova K, Garewal H. Bile acids as carcinogens in human gastrointestinal cancers. *Mutat Res Rev* 2005;589:47–65.
12. Berr F, Kullak-Ublick G-A, Paumgartner G, Münzing W, Hylemon PB, Mu W. 7α-Dehydroxylating bacteria enhance deoxycholic acid input and cholesterol saturation of bile in patients with gallstones. *Gastroenterology* 1996;111:1611–1620.
13. Ridlon JM, Marcelo Alves J, Hylemon PB, Bajaj JS. Cirrhosis, bile acids and gut microbiota. *Gut Microbes* 2013;4:382–387.
14. Funabashi M, Grove TL, Pascal V, Varma Y, McFadden ME *et al*. A metabolic pathway for bile acid dehydroxylation by the gut microbiome. *BioRxiv* 2019
15. Dickert S, Pierik AJ, Linder D, Buckel W. The involvement of coenzyme A esters in the dehydration of (R)-phenyllactate to (E)-cinnamate by *Clostridium sporogenes*. *Eur J Biochem* 2000;267:3874–3884.
16. Hubbard PA, Liang X, Schulz H, Kim J-JP. The crystal structure and reaction mechanism of *Escherichia coli* 2,4-dienoyl-CoA reductase. *J Biol Chem* 2003;278:37553–37560.
17. Trickey P, Basran J, Lian LY, Chen Z, Barton JD *et al*. Structural and biochemical characterization of recombinant wild type and a C30A mutant of trimethylamine dehydrogenase from *Methylophilus methylotrophus* (sp. W(3)A(1)). *Biochemistry* 2000;39:7678–7688.
18. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;35:1026–.
19. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5:e9490.
20. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* 2012;3:217–223.
21. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;47:W256–W259.
22. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 2011;7:539.
23. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB *et al*. A new genomic blueprint of the human gut microbiota. *Nature* 2019;568:499–504.
24. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH *et al*. A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol* 2020;16:60–68.
25. Tett A, Pasolli E, Farina S, Truong DT, Asnicar F *et al*. Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis. *NPJ Biofilms Microbiomes* 2017;3:14.
26. Blaser MJ, Perez-Perez GI, Kleanthous H, Cover TL, Peek RM *et al*. Infection with *Helicobacter pylori* strains possessing *cagA* is associated with an increased risk of developing adenocarcinoma of the stomach. *Cancer Res* 1995;55:2111–2115.
27. Frank C, Werber D, Cramer JP, Askar M, Faber M *et al*. Epidemic profile of Shiga-toxin-producing *Escherichia coli* O104:H4 outbreak in Germany. *N Engl J Med* 2011;365:1771–1780.
28. Segata N. On the road to strain-resolved comparative metagenomics. *mSystems* 2018;3:e00190-17.
29. Morton JT, Aksenov AA, Nothias LF, Foulds JR, Quinn RA *et al*. Learning representations of microbe–metabolite interactions. *Nat Methods* 2019;16:1306–1314.
30. Mallick H, Franzosa EA, Mclver LJ, Banerjee S, Sirota-Madi A *et al*. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat Commun* 2019;10:3136.