Supplementary material for:
A systematic analysis of metabolic pathways in the human gut microbiota

Victòria Pascal Andreu[1], Hannah E. Augustijn[1,2#], Lianmin Chen[2,3,#], Alexandra Zhernakova[2], Jingyuan Fu[2,3], Michael A. Fischbach[4,5,7*], Dylan Dodd[5,6*], Marnix H. Medema[1*]

Supplementary Results

Phylogenetic analysis of protein superfamilies to identify pathway-specific clades

Performing hmmscan searches on the protein sequences helped identify the presence of keystone domains (Pfams) between pathways that share an enzymatic core. However, some of these enzymes are part of multifunctional enzyme families that are recognised by very broad Pfam domains. Thus, in order to more accurately identify relevant functional subgroups for keystone enzyme families, we used protein phylogenetic analysis to pinpoint pathway-specific clades and discern, for instance, specialized primary metabolism-related enzymes from housekeeping related ones (Figure 1). For each protein family, we created a non-redundant set of representatives by gathering protein sequences from four different sources: the proteins from the representative pathway, the reference proteome available in Pfam[1], the respective proteins from the MGC collection and the experimentally characterized proteins available in UniProt, to assign functionality to clades (see SI Methods: *Towards a more robust MGC identification by building new HMM profiles*). In total, we performed this phylogenetic analysis to 12 major protein superfamilies. For instance, phenyllactate dehydratase homologues are involved in the degradation of aromatic amino acids into propionate, in the acrylate to propionate pathway and in the leucine reductive branch pathway; in contrast, 2-hydroxyglutaryl-CoA dehydratase allows the transformation of glutamate into butyrate. Despite the fact that these reactions are different, both key enzymes harbour the 2-hydroxyglutaryl-CoA dehydratase, D-component domain (HGD-D Pfam domain, PF06050). The HGD-D protein family phylogenetic tree (see Suppl. Figure 3) revealed that 10 clades were implicated in these three pathways. Subsequently, 10 profile hidden Markov models (pHMMs) specific to these clades were built and, after assessing the sensitivity of the models (see Methods), 9 out of 10 were selected for being sensitive enough as to correctly identify the subdomains of this protein family involved in these 3 pathways. Consequently, these 9 pHMMs were included in the corresponding detection rule. The same procedure was followed for the other protein superfamilies, creating a total of 43 pHMMs (see Table S14). As a result, gutSMASH uses newly built pHMMs in combination with the ones included in the Pfam database to identify protein families of interest. Therefore, gutSMASH competitively scores hmmsearch hits and assigns to the sequence the domain with a higher score. Altogether, this procedure allowed to define a preliminary set of detection rules using the newly built pHMMs to predict close homologues of known gene clusters.

Validation of gutSMASH detection rules by evaluating their predictive performance

Evaluating gutSMASH performance implies having a set of bacteria whose genomes are known to encode a given pathway. However, the false positive rate is unknown (as it is not feasible to experimentally verify large numbers of diverse putative annotations) and the false negative rate is difficult to determine (as only in few species, MGCs with proven functions are described in literature).

Hence, we decided that the best course of action would be to perform detailed manual analysis of large numbers of diverse predicted MGCs. For this reason, gutSMASH was run on a test set created from 1,632 bacterial genomes. All the MGCs predicted by the same detection rule were grouped together to further run BiG-SCAPE on each subset (see Methods section *Testing and validating gutSMASH specific-to-known-pathway detection rules*). In this manner, we could evaluate the range of gene clusters predicted by the same detection rule and quickly find out if any distantly-related gene cluster should not be picked by the rule based on, e.g., having divergent enzyme-coding composition. Also, it allowed us to acquire an overview of the bacterial taxa predicted to possess a given gene cluster type and identify if any MGCs from taxa referenced in primary literature were missing from this set. Thus, this system allowed to fine-tune the detection rules and evaluate their predicting potential. After several iterations of adjusting rules, performing new predictions and creating new sequence similarity networks, we froze gutSMASH version 1.0 with 41 specific-to-known-pathway detection rules (see Table S15) to accurately and comprehensively predict MGCs.

An additional validation step was performed using PaperBLAST[2], which was used to look for genomes encoding any close homologues of the key proteins involved in 18 gutSMASH predicted pathways; all 18 MGCs were successfully detected using gutSMASH. When compared to the reference MGCs, detected clusters showed an average amino acid sequence identity between 54 and 100% and overall gene cluster similarity (percentage of homologues detected in KnownClusterBlast) ranging from 44 to 100%. These results suggest that the false negative rate is low.

Finally, to get an estimate of the false positive rate, we ran gutSMASH on 5 genomes of well-studied reference organisms (Escherichia coli, Clostridioides difficile, Prevotella copri, Bifidobacterium animalis and Fusobacterium nucleatum) from diverse phyla and analyzed the results manually (see Table S16). Across 42 MGCs detected in these genomes, no false positives were identified, as the functions of all detected gene clusters matched prior evidence from literature.

While obtaining precise estimates of precision and recall for gutSMASH is infeasible due to the absence of large-scale experimentally verified reference data from diverse taxa, these results indicate that the overall accuracy of the algorithm is high.

With regard to the boundaries of the MGCs, gutSMASH does not attempt to exactly predict these. Instead, we use the same 'greedy' approach as implemented in antiSMASH[3], in which gene clusters are extended with fixed numbers of kilobases. This has the advantage that all potential MGC genes are visible to the user, and none are missed.

Analysis of putative clusters and distant homologues: relevant candidates to study further

Next, we evaluated the potential of gutSMASH to predict putative MGCs of interest and to explore the metabolic landscape covered by them. For this reason, the 12,256 putative clusters predicted from the HMP, CGR and *Clostridioides* genomes were used and subjected to a redundancy filtering of 90% similarity at the protein sequence level (see Methods: *Evaluating the functional potential of the human microbiome using gutSMASH*) to investigate their functional diversity. To create a non-redundant collection, two random representative clusters of each set of highly similar clusters were picked; all representatives were then clustered together into 932 GCFs using BiG-SCAPE[4] (see Methods: *Analysis of distant homologues and putative MGCs from CGR, HMP and Clostridioides dataset*). From the resulting network (see Extended Data Figure 3), we made three main observations. First, we identified several distant homologues of the known MGCs; these are picked by specific pathway rules but classified as putative when for instance the MGC is from a distantly related

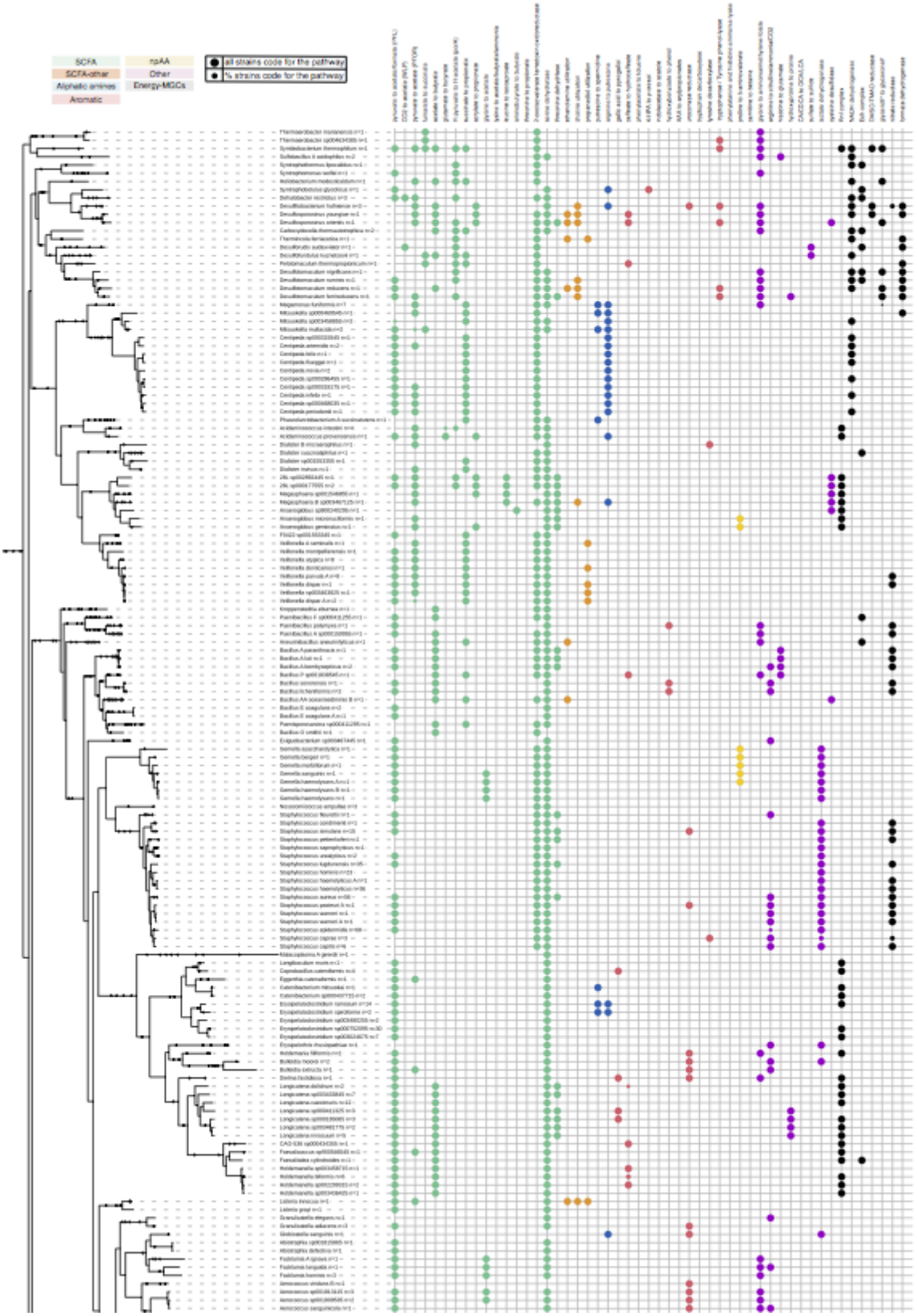
taxonomic group and therefore shares low sequence similarity with the reference gene cluster. Second, we found previously characterized MGCs that were not included in the original training sets of known MGC types, as for instance a region from *Ruminococcus gnavus strain AM22-7AC* (accession number QRIA01000012.1) involved in the metabolism of rhamnose and fucose described in 2013 by E.Petit & W. Latouf *et al*.[5]; this validated the capability of gutSMASH to identify real MGCs that were not included in the initial list of known pathways. Third, we observed vast numbers of novel clusters of unknown function that represent good candidates for further experimental characterization. The examples shown in Extended Data Figure 4 are metabolically diverse MGCs that encode flavoenzymes, oxidative decarboxylation (OD), glycyl radical (GR), 2-hydroxyglutaryl-CoA dehydratase (HGD-D) related or hybrids of them, which were found in phylogenetically diverse genomes from Firmicutes, Fusobacteria and Actinobacteria. Moreover, all these MGCs presented plausible architectures as to be real gene clusters, since they included all the genetic elements to regulate, synthesize and transport the resulting molecule (or its substrates). The first MGC for instance, which is found in *Enterocloster citronae* AF29-4, shares some similarity with the acetate to butyrate MGCs, since it encodes an acyl-CoA dehydrogenase and two electron transfer flavoproteins as well as a distant *baiH* homologue. Another interesting example is the *Dorea sp.* D27 MGC, a pathway involving a pyruvate:ferredoxin oxidoreductase, which might be encapsulated given the presence of bacterial microcompartment genes (BMC-encoding genes) in the cluster. *Blautia sp.* TF11-31AT also presents an unprecedented gene cluster architecture, encoding a combination of enzymes involved in oxidative decarboxylation (thiamine pyrophosphate-related) and related to HGD-D. This overview highlights the ability of gutSMASH to systematically predict novel and interesting gene clusters from a diverse range of bacteria that could help associating function to unknown genes and predict novel pathways. Additionally, the expression of families of MGCs of unknown function could potentially be correlated to microbiome-associated phenotypes to prioritize them for experimental characterization based on physiological and ecological relevance.

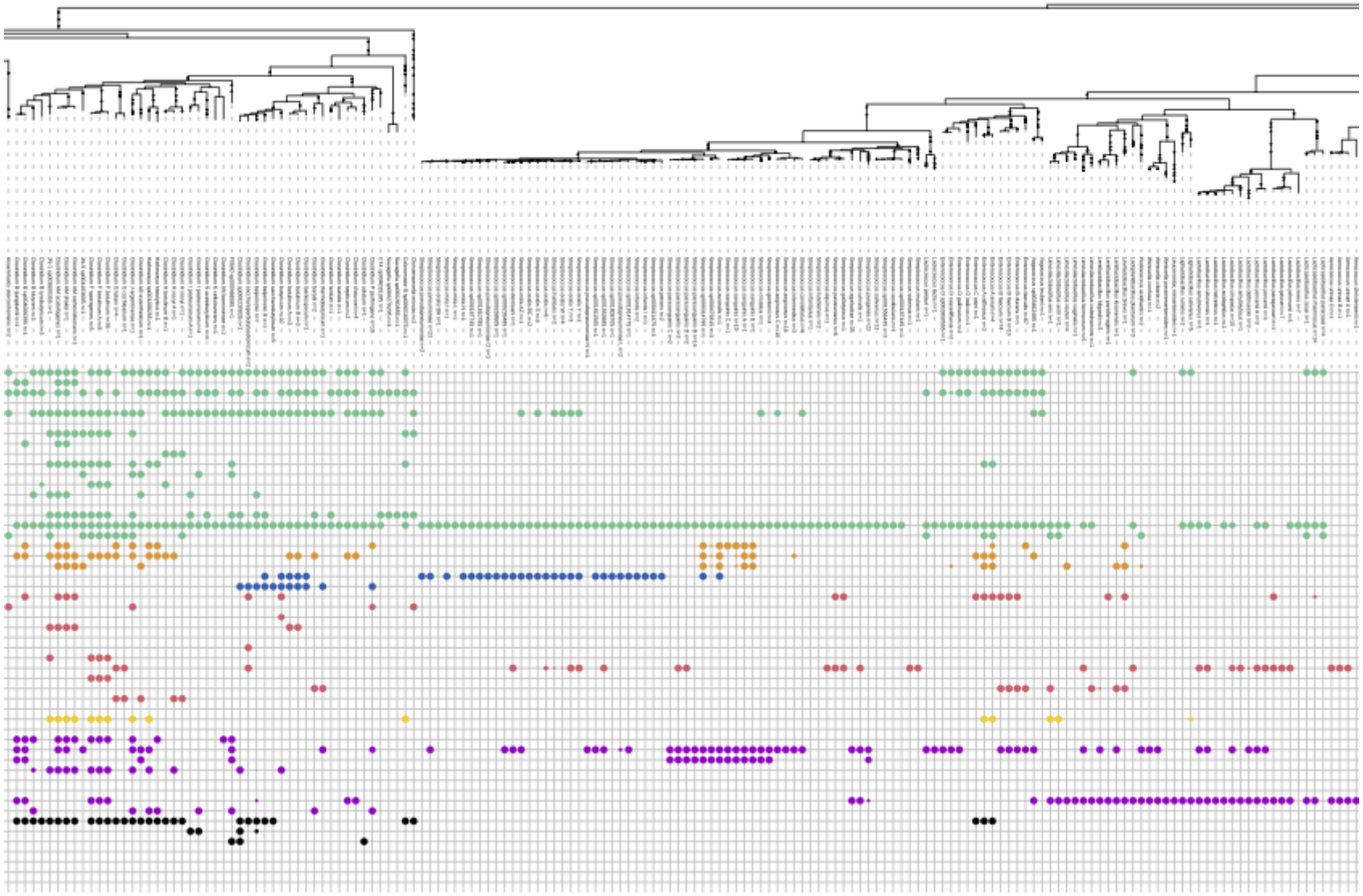
Assessing pathway abundance and prevalence across metagenomes

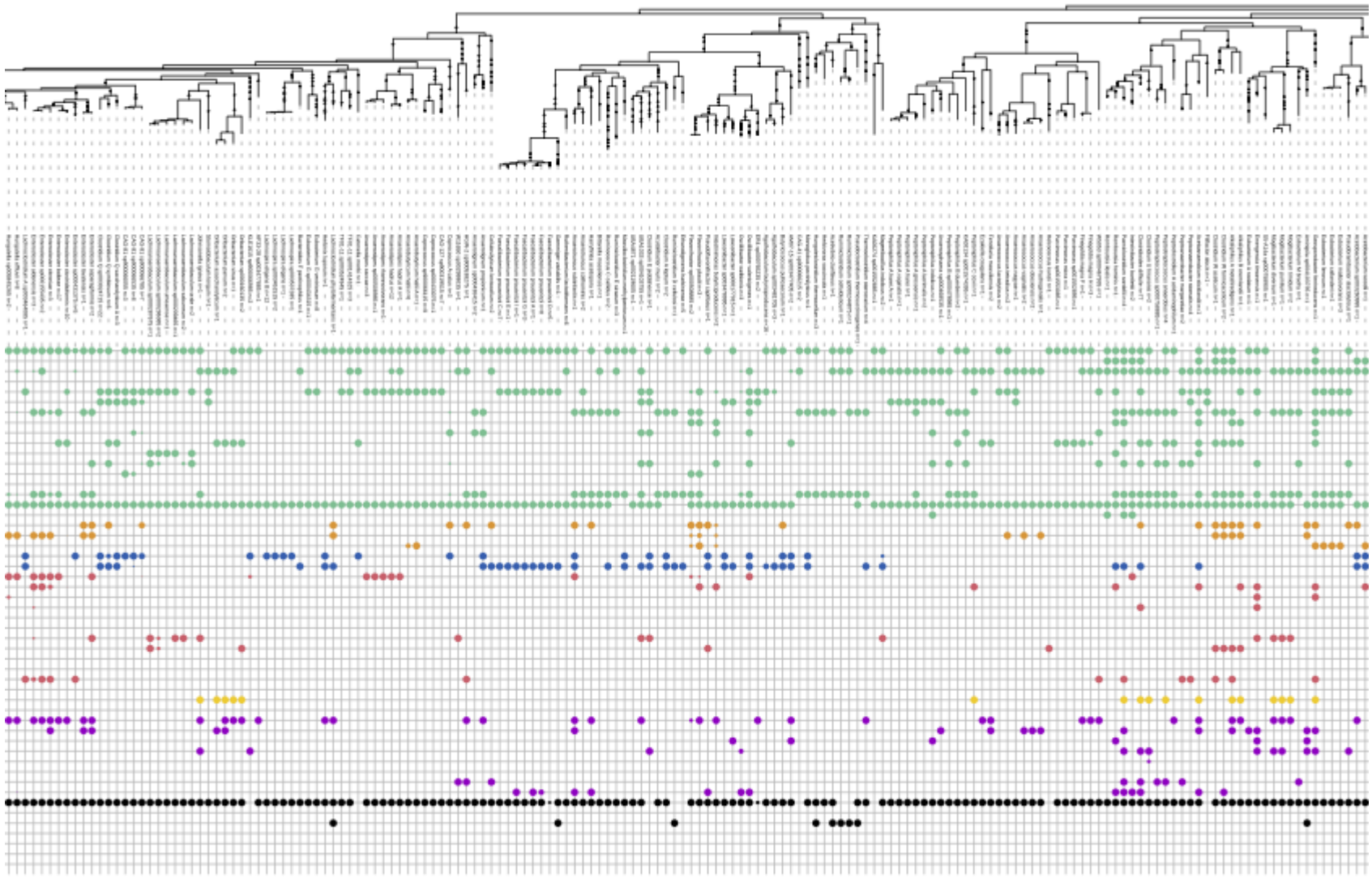
From the predicted pool of known pathways, we aimed to assess their abundance and prevalence by mapping the metagenomic reads of the LifeLines-DEEP cohort[6,7]. To compute both, certain assumptions had to be made and thresholds needed to be chosen. In both cases, the numbers of reads mapping to the core regions are key to assess whether a given pathway is present or not and how abundant it is. In order to account for spurious mapping, we designed an approach to assess the pathway abundance by using the lower quartile number of reads mapping to 2kb long regions for each pathway and sample (see more in Methods section entitled *Mapping metagenomics reads from healthy samples to the known gutSMASH predicted MGCs*). In contrast, to evaluate the prevalence of all the pathways across samples, the core coverage score estimated by BiG-MAP was used. Ideally, with unlimited sequencing depth, one would expect to find reads mapping evenly to the whole gene cluster, thus opting for a relatively high core coverage score cutoff compared to the chosen one (>5% coverage, see Figure 3a) to avoid incorrect pathway presence calling due to spurious mapping. However, when raising the minimum coverage value from 10-80%, some of the pathways that are known to be present in healthy individuals showed a prevalence of 0 (see Extended Data Figure 1). Also, using the 5% threshold, the minimum sequence identity of a read mapping to a gene cluster showed 78% identity (at nucleotide level), and 81% of the reads had >90% identity to the reference sequence. This confirms that even with low coverage scores, the reads are very specifically mapping to the gene clusters and hence, the lack of coverage of some gene clusters is very likely due to the

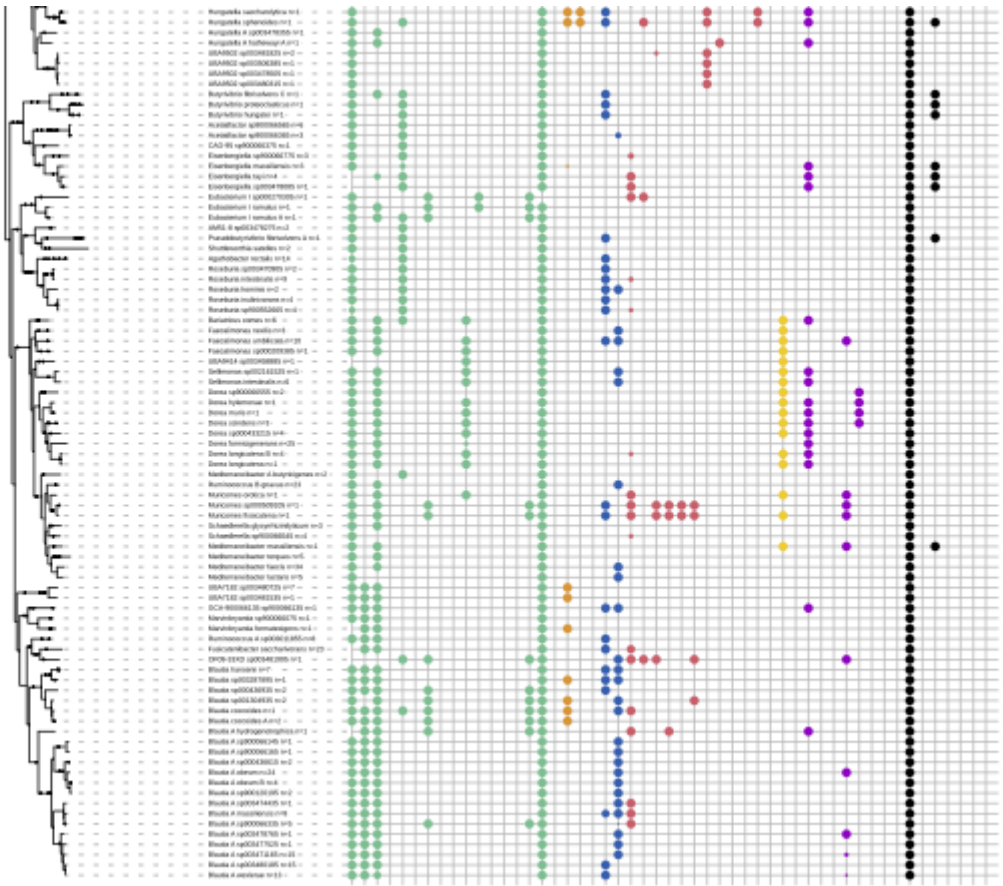
limited sequence depth compared to the (low) abundance of the sequences in the sample, rather than the absence of the pathway. For this reason, we set the minimum coverage used in the analyses presented in the main text at 5%, to avoid undue false negatives.

Supplementary Figures

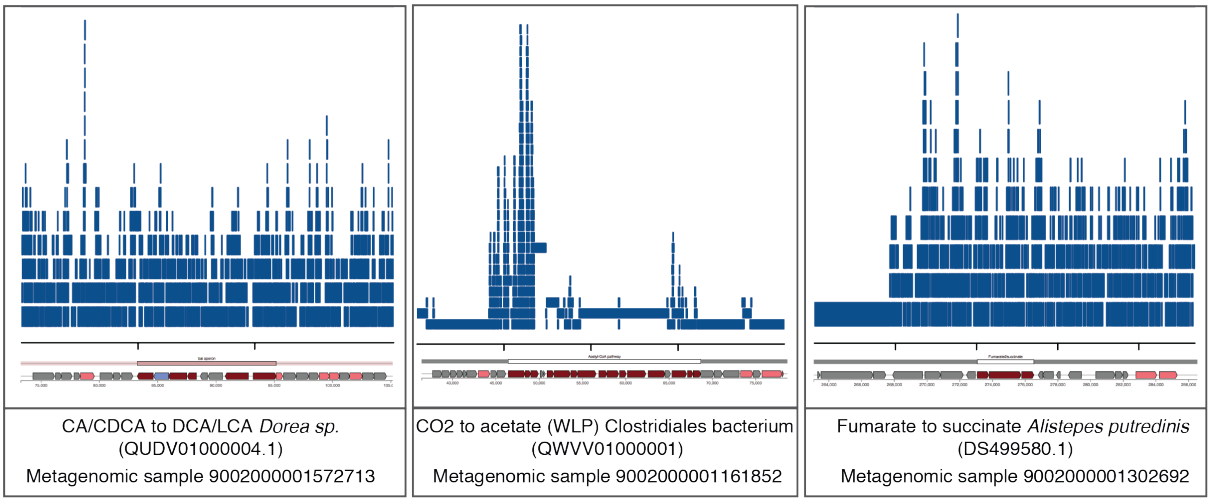




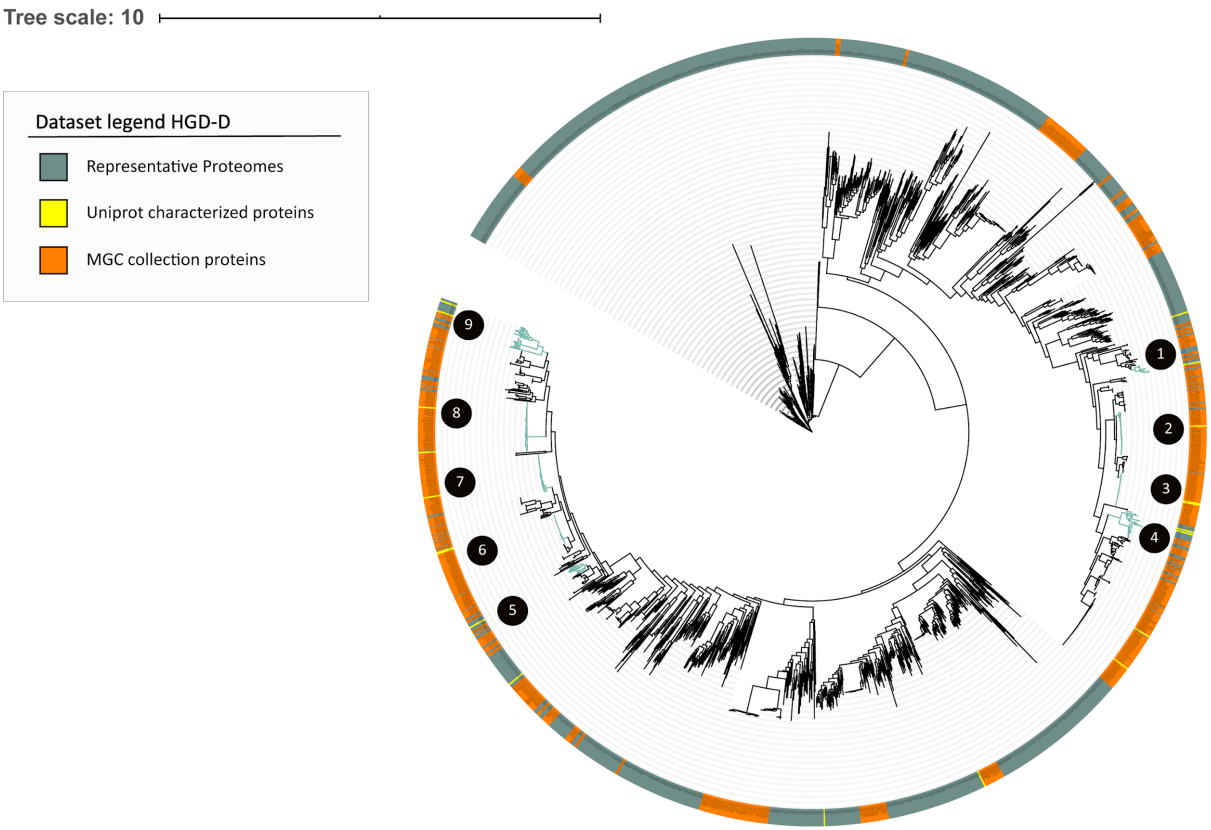




Supplementary Figure 1: Pathway distribution across a phylogenetic tree of 557 Firmicutes present in the HMP, CGR and Clostridiales datasets. The taxonomic assignments were performed using the GTDB database release 95[8] and the phylogeny was produced using PhyloT (https://phylot.biobyte.de/). Each column represents the presence/absence of the 51 metabolic pathways including single gene ones), which are colour-coded based on the pathway's end product (metabolic classes). The full-sized circle implies that all strains in the node code for the pathway (see species label for information on the species group size), while smaller circles represent the relative number of species that encode the pathway. The pathways annotations were visualized using iTOL [9].



Supplementary Figure 2: Read coverage of three random metagenomic samples mapping to three gene clusters of interest: the *bai* operon (CDA/CDCA to DCA/LCA), a gene cluster encoding the acetyl-CoA pathway (CO_2 to acetate) and a gene cluster encoding the fumarate-to-succinate pathway. Reads are represented by blue lines, which are distributed along the x-axis based on the bedgraphs output by BiG-MAP. The plots have been produced using the Sushi R package (version 3.5.1)[10] and show how, despite the fact that some regions of the MGC attract more reads, the whole gene cluster is covered. They also illustrate the rationale for using the lower quartile of read coverage across 2kb regions, as this will avoid basing MGC abundance (partially) on outliers.



Supplementary Figure 3: HGD-D protein superfamily phylogeny. The phylogeny contains 2,054 protein sequences gathered from three different sources as shown with different colours in the outer ring. The highlighted clades with numbers associated are the pathway-specific clades used to create the 9 pHMMs used in the AAA to aryl propionates (AAA reductive branch), leucine to isocaproate (leucine reductive branch), glutamate to butyrate and acrylate to propionate MGC detection rules.

Supplementary Table Legends

Table S1: Dataset of 51 known pathways, including single-protein pathways. This dataset helped designing detection rules for gutSMASH and build pHMMs for more accurate pathway identification. The table includes information on the substrate(s) and product(s) of each pathway, Pfams involved (ID and description) and the original organism where it has been characterized.

Table S2: Validation results of the gutSMASH predictive potential using PaperBLAST. Organisms whose genomes encode homologues of proteins from the original data set of known pathways were found using paperBLAST and their genomes were used as input for gutSMASH. The table includes information on the original organisms, validation organism, amino acid sequence identity and overall sequence identity.

Table S3: Assembly IDs of the 1,621 genomes used to validate the gutSMASH detection rules.

Table S4: Assembly IDs of 4,240 genomes from the HMP, CGR and Clostridiales collections used to screen the metabolic potential of human gut bacteria. GCA_000025225.2 and GCA_000163895.2 are found in both the HMP and Clostridiales dataset.

Table S5: Raw pathway abundance across most representative genera of bacteria in the human gut. This data has been used to create Figure 2a.

Table S6: Absolute counts of genomes harboring genes and MGCs corresponding to the main acetate-producing pathways, summarized at phylum level. This data has been used to create Figure 2b.

Table S7: Pathway prevalence values of the 41 pathways across 1,135 human microbiomes using different BiG-MAP mapping coverage threshold values. These values have been used to create figure 3a and Extended Data Figure 1.

Table S8: Pathway abundance counts across 1,135 human microbiome samples. These values have been used to create Figure 3b.

Table S9: Correlations between different metabolites and the MGC abundance of those pathways that have the same end product. Spearman correlation (two sided with rho and empirical p-value are reported) is used to check the relationship between pathway abundances and metabolite levels after adjusting for age, sex and read depth. n= 1054 biologically independent samples.

Table S10: Correlations between feacal SCFA and the corresponding pathway abundance counts from the LLD cohort. Spearman correlation (two sided with rho and empirical p-value are reported) is used to check the relationship between pathway abundances and metabolite levels after adjusting for age, sex and read depth. n= 1054 biologically independent samples.

Table S11: Correlations between pathway abundance and expression and their respective metabolite abundance of the 81 iHMP samples with paired metagenomics/metatranscriptomics/metabolomics data.

Table S12: Correlations between pathway abundance and expression and their respective metabolite abundance of 271 iHMP samples with paired metatranscriptomics/metabolomics data.

Table S13: MetaCyc links found for 38/41 of the pathways predicted by gutSMASH.

Table S14: List of pHMMs created to accurately identify pathway-specific clades of interest. These pHMMs have been included in the gutSMASH v1.0 detection rule set and are available online at https://github.com/victoriapascal/gutsmash/tree/gutsmash/antismash/detection/gut_hmm_detection/data .

Table S15: Overview of detection rules in gutSMASH v1.0 for the identification of MGCs encoding known pathways.

Table S16: gutSMASH false positive rate calculation considering 5 well-studied bacteria belonging to diverse phyla.

Table S17: Set of known pathways used for the iterative homology search approach. The ultimate goal of this approach was to design the general rules. The original Genbank files can be downloaded from https://gutsmash.bioinformatics.nl/help.html#Pathways-subset .

Table S18: Overview of general detection rules used by gutSMASH version 1.0.

References

1. Chen, C. *et al.* Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation. *PLoS One* **6**, e18910 (2011).
2. Price, M. N. & Arkin, A. P. PaperBLAST: Text Mining Papers for Information about Homologs. *mSystems* **2**, e00039-17 (2017).
3. Medema, M. H. *et al.* antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39**, W339–46 (2011).
4. Navarro-Muñoz, J. C. *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).
5. Petit, E. *et al.* Involvement of a Bacterial Microcompartment in the Metabolism of Fucose and Rhamnose by Clostridium phytofermentans. *PLoS One* **8**, e54337 (2013).
6. Tigchelaar, E. F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* 5, e006772 Preprint at https://doi.org/10.1136/bmjopen-2014-006772 (2015).
7. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
8. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
9. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
10. Phanstiel, D. H., Boyle, A. P., Araya, C. L. & Snyder, M. P. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics* **30**, 2808–2810 (2014).