nature biotechnology

Article

https://doi.org/10.1038/s41587-023-01872-y

BugSigDB captures patterns of differential abundance across a broad range of host-associated microbial signatures

Received: 24 October 2022

Accepted: 20 June 2023

Published online: 11 September 2023

Check for updates

Ludwig Geistlinger[1], Chloe Mirzayi[2,3], Fatima Zohra[2,3], Rimsha Azhar[2,3], Shaimaa Elsafoury[2,3], Clare Grieve[2,3], Jennifer Wokaty[2,3], Samuel David Gamboa-Tuz[2,3], Pratyay Sengupta[4,5,6], Issac Hecht[7], Aarthi Ravikrishnan[8], Rafael S. Gonçalves[1], Eric Franzosa[9,10,11], Karthik Raman[4,5,6], Vincent Carey[12], Jennifer B. Dowd[13], Heidi E. Jones[2,3], Sean Davis[14], Nicola Segata[15,16], Curtis Huttenhower[9,10,11] & Levi Waldron[2,3,15]

The literature of human and other host-associated microbiome studies is expanding rapidly, but systematic comparisons among published results of host-associated microbiome signatures of differential abundance remain difficult. We present BugSigDB, a community-editable database of manually curated microbial signatures from published differential abundance studies accompanied by information on study geography, health outcomes, host body site and experimental, epidemiological and statistical methods using controlled vocabulary. The initial release of the database contains >2,500 manually curated signatures from >600 published studies on three host species, enabling high-throughput analysis of signature similarity, taxon enrichment, co-occurrence and coexclusion and consensus signatures. These data allow assessment of microbiome differential abundance within and across experimental conditions, environments or body sites. Database-wide analysis reveals experimental conditions with the highest level of consistency in signatures reported by independent studies and identifies commonalities among disease-associated signatures, including frequent introgression of oral pathobionts into the gut.

Despite substantial progress in experimental techniques and computational methods for culture-independent profiling of the human microbiome, the analysis and interpretation of microbial differential abundance studies remains challenging. A large body of experimental and observational studies on humans and in animal models has reported associations between host-associated microbiomes and the onset, progression and treatment of a variety of diseases, including atherosclerosis[1], cardiovascular diseases[2], cancers[3] and diabetes[4]. This growing body of published results provides opportunities for synthesis of accumulated knowledge, identification of common patterns across different diseases and exposures and interpretation of new studies by comparison to previous results. However, without a systematic catalog of published differential abundance results, even identical microbial signatures reported in different research fields are unlikely to be noticed. Even within research fields, systematic review in the absence of a catalog or common reporting of differential abundance results are time consuming, static and generally do not summarize all taxa reported.

This situation has parallels to early challenges in the interpretation of differential gene expression analyses[5], which have been addressed in the field by gene set enrichment analysis (GSEA). GSEA allows for the

A full list of affiliations appears at the end of the paper. e-mail: levi.waldron@sph.cuny.edu

comparison of coherent expression patterns among predefined gene signatures that share a biological function or property or that were identified together by a previous study[6,7]. GSEA is a key tool in gene expression data analysis[6], with a wide range of subsequent methods[8] to account for correlations between genes[9], redundancy of functional annotation[10], different types of null hypotheses[11] and the application of GSEA for the analysis of genomic regions[12], metabolomic data[13] and disease phenotypes[14].

Analogously, differential microbial abundance analysis can yield lists or 'signatures' of microbial clades at multiple taxonomic levels that are associated with a phenotype of interest. The properties shared by these clades are often not obvious but could include common environmental exposures, metabolic or ecological requirements or physiological characteristics. Although nascent attempts to apply concepts of GSEA to the results of microbiome differential abundance analyses exist[15–18], major obstacles have prevented their broad utility and adoption. The most notable obstacle has been the lack of comprehensive databases of signatures designed for enrichment analysis, such as those available for GSEA, including the Gene Ontology[19], KEGG[20], MSigDB[21,22] and GeneSigDB[23]. Several databases provide important information on microbial physiology and morphology[24–27] but are not designed for enrichment analysis and, by design, exclude the vast majority of experimentally derived microbial signatures associated with cancer, inflammation, diet or other conditions studied in human and other host-associated microbiome research.

This study provides two main contributions to enable high-throughput comparison of published microbial signatures. First, it describes BugSigDB, a database of published microbial signatures of sufficient scale and diversity to capture replicable patterns of differential abundance across a broad spectrum of the host-associated microbiome literature. BugSigDB provides curated published signatures of differentially abundant microbes associated with a wide range of health outcomes, pharmaceutical usage (for example, antibiotics), experiments on animal models, randomized clinical trials and microbial attributes and is built on the technology of Wikipedia to allow community contributions, revisions and review for quality control. Second, we provide a systematic analysis of the results reported by hundreds of published microbiome studies and identify replicated patterns even across 16S amplicon and shotgun sequencing approaches, demonstrating that interpretation of new microbiome studies can be supported by systematic comparison to previously published signatures. Database-wide analysis revealed common patterns of microbe co-occurrence and mutual exclusivity within signatures and identified antibiotic treatment and human immunodeficiency virus (HIV) infection as the experimental conditions with the highest level of consistency in signatures reported by independent studies. 'Bug set' enrichment analysis of ten individual colorectal cancer-associated fecal microbiome datasets ($N = 663$) detected published meta-analysis signatures used as positive controls, supporting direct application of methods adapted from gene set analysis and taxonomy-aware enrichment methods. This analysis also identified commonalities among signatures of other diseases, including elevated frequency of oral pathobionts in the gut, and identified the most common patterns of co-occurrence and mutual exclusion across all conditions or related conditions or unique to one condition. Together, the BugSigDB database and analysis methods described here improve the interpretation of new microbiome studies by systematic comparison to published microbial signatures.

Results

A curated database of published microbial signatures

BugSigDB comprises a comprehensive database of manually curated, host-associated microbial signatures from published microbiome studies of human, mouse and rat (Fig. 1). The database has been simultaneously developed and expanded over the course of more than 4 years with contributions from more than 25 curators trained in-house and currently contains >2,500 microbial signatures extracted from >600 scientific articles (Fig. 1a). The curated papers cover two decades of microbiome research, with the majority of studies being published in the last 5 years (459 of 628 articles, 73.1%; Supplementary Fig. 1). Among these papers are microbiome studies of participants from more than 50 different countries, with more than 50% of the studies originating from China and the United States (201 and 157 studies, respectively; Fig. 1b).

Studying microbiome samples from 14 broad body areas comprising more than 60 refined anatomical sites according to the UBERON Anatomy Ontology[28] (Supplementary Table 1), the majority of studies in BugSigDB analyzed gut (440 of 628 studies, 70.1%), oral (80 of 628 studies, 12.8%) and vaginal microbiome samples (59 of 628 studies, 9.4%; Fig. 1c). The signatures were generated by both 16S amplicon sequencing (92.5%) and metagenomic shotgun sequencing (MGX; 7.5%; Table 1) and contained taxonomic levels from phylum to strain standardized based on the NCBI Taxonomy[29] (Supplementary Fig. 2). BugSigDB is implemented as a semantic MediaWiki[30] web interface available at https://bugsigdb.org that supports data entry, semantic validation, quality control and web-based programmatic access to annotations for studies, experiments, signatures and individual taxa (Supplementary Fig. 3 and "Data entry, validation and access" in the Methods).

Most of the roughly 1,400 unique taxa contained in BugSigDB are reported as differentially abundant in fewer than five signatures (1,009 of 1,370 unique microbes, 73.6%; Fig. 1d). Conversely, genera *Streptococcus*, *Prevotella*, *Bacteroides* and *Lactobacillus* are each reported as differentially abundant in more than 100 signatures, reflecting the large number of species belonging to these genera and paralleling previous observations of gene signatures with certain genes often being disproportionately reported as differentially expressed[31]. Signatures contain six microbes, on average, with roughly 55% of the signatures containing five or more microbes (Fig. 1e).

Study conditions associated with each signature are classified according to the Experimental Factor Ontology[32] (Fig. 1f), with large proportions associated with different types of cancer (23.1%), anatomical system diseases, such as asthma and endometriosis (22.6%), and metabolic diseases, such as obesity and diabetes (15.3%). Apart from different disease categories, substantial proportions of signatures in BugSigDB are also associated with diet (6%), the use of antibiotics (4.8%) and birth delivery mode (vaginal versus Cesarean section; 4.1%). These distributions approximately reflect the human microbiome literature. Non-human host species remain underrepresented, although the database can support any host organism.

Condition-specific associations were investigated for the ten most-reported taxa (Fig. 1g). For instance, signatures containing the *Bacteroides* genus are related to metabolic disease (21%) more frequently than all signatures (13%; $P = 0.003$, $\chi^2 = 8.4$, d.f. = 1, two-sided proportion test). *Bacteroides* is similarly enriched in signatures of antibiotic exposure (11% of *Bacteroides*-containing signatures being associated with antibiotics exposure versus 5% of all signatures; $P = 0.0009$, $\chi^2 = 11.2$, d.f. = 1). BugSigDB makes such commonalities across groups of related studies straightforward to identify and provides for continuous updates of these associations as the database grows.

Curated metadata and common practices in microbiome research

BugSigDB provides curated metadata at the level of study (study design and automatically generated citation information), experiments within a study that each define one contrast for differential abundance analysis (such as characteristics about study subjects, lab analysis, statistical analysis and α-diversity) and up to two signatures within an experiment, each of which contains one or more taxon (see Table 1 for summary statistics of studies and experiments and "Definition of semantic concepts" in the Methods for details of the design). In this article, we use 'study

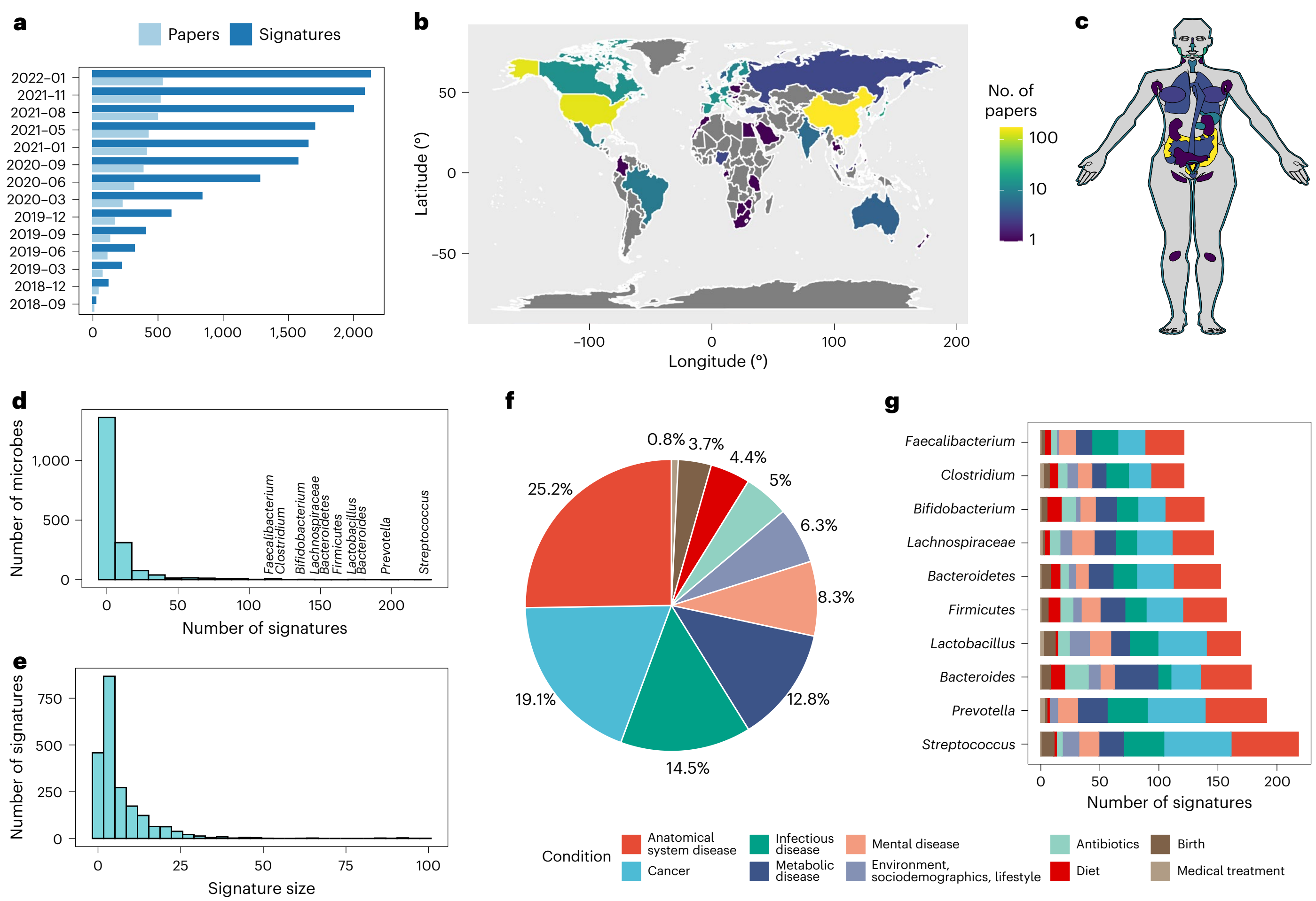

Fig. 1 | BugSigDB, a curated database of experimentally derived microbial signatures. a–**c**, BugSigDB is a community-editable collection of published microbiome studies reporting differentially abundant host-associated microbiota (including bacteria, archaea, fungi, protists and viruses) that currently includes >2,500 microbial signatures extracted from >600 scientific articles over the course of more than 4 years (**a**). These papers report microbiome studies of participants representing different countries and ethnogeographies (**b**) and microbiome samples from different human body sites (**c**). **d**, Number of signatures associated with a specific microbe, with the top 10 most frequently reported microbes annotated. BugSigDB signatures contain taxonomic levels from phylum to strain standardized based on the NCBI Taxonomy[29]. **e**, Signature sizes, with more than 50% of the signatures containing five or more microbes. **f**, Percentage of signatures annotated to major disease categories when classifying the study condition associated with each signature according to the Experimental Factor Ontology[32]. **g**, Top 10 most frequently reported microbes and the number of associated signatures stratified by disease category.

group' to refer to cases in case–control studies, the exposed group in exposure–control studies and whichever group corresponds to the condition of interest with increased microbial relative abundance in other comparisons of two-sample groups.

Signatures are available primarily from observational study designs: case–control (281 studies, 44.7%) and cross-sectional studies (171 studies, 27.2%) were most prevalent, while prospective cohort studies (66, 10.5%), time series/longitudinal studies (51, 8.1%), laboratory studies (29, 4.6%), randomized controlled trials (28, 4.4%) and meta-analyses (6, 0.9%) were also present. Study subject information includes host species (95.9% human of 1,223 experiments), location, condition, body site (Fig. 1), antibiotics exclusion criteria (median exclusion time = 60 d) and sample size in study and control sample groups (median total sample size = 25).

A survey of statistical methods most frequently applied for differential abundance testing in BugSigDB revealed that nonparametric tests, such as Wilcoxon family tests and the related LEfSe software[33], were most frequently used, whereas recently suggested tools for differential abundance tests accounting for the compositionality of microbiome data[34] were rarely used (Supplementary Results Section 1.1). Furthermore, when stratifying experiments by body site and condition, oral and vaginal samples were frequently reported with increased α-diversity in the study group, as opposed to samples from the gastrointestinal tract, which were frequently found with decreased α-diversity in the study group (Supplementary Tables 2–4).

Conditions with replicable microbiome changes across studies

BugSigDB facilitates meta-analysis of differential abundance studies and enables the identification of experimental conditions and disease phenotypes where microbiome changes replicate across studies. Focusing on 1,194 signatures derived from human fecal samples in 311 published studies, we computed signature similarity within conditions and assessed whether the resulting similarity exceeds the similarity of randomly sampled signatures (Fig. 2). This simultaneously determines whether signatures of the same phenotype reproduce across studies and whether different phenotypes share similar microbial signatures.

To do this, we applied two alternative approaches for computing similarity between signatures: (1) the more restrictive Jaccard index[35] based on pairwise overlaps between signatures harmonized to genus level and (2) the more sensitive semantic similarity[36] based on taxonomic distance between signatures of mixed taxonomic levels

Table 1 | Overview of curated metadata annotations in BugSigDB

	Count	Percentage
Studies		
Study design		
Case–control	281	44.7
Cross-sectional	171	27.2
Prospective cohort	66	10.5
Time-series/longitudinal	51	8.1
Laboratory experiment	29	4.6
Randomized control trial	28	4.4
Meta-analysis	6	0.9
Experiments		
Host species		
Homo sapiens	1,173	95.9
Mus musculus	44	3.6
Rattus norvegicus	6	0.5
Location		
China	350	28.7
United States	350	28.7
Italy	51	4.2
Spain	48	3.9
South Korea	42	3.5
Other (56 more)	378	31.0
Body site		
Feces	694	56.9
Saliva	40	3.3
Uterine cervix	40	3.3
Vagina	36	2.9
Mouth	29	2.4
Other (67 more)	380	31.2
Condition		
Obesity	103	8.5
COVID-19	74	6.1
Antimicrobial agent	64	5.3
Diet	55	4.5
Human papillomavirus infection	42	3.4
Other (173 more)	877	72.2
Antibiotics exclusion		
0–2 weeks	64	5.2
2 weeks–1 month	216	17.6
1–3 months	265	21.6
More than 3 months	89	7.3
Not reported	591	48.3
Lab analysis		
Sequencing type		
16S	1,117	92.5
MGX	91	7.5
Sequencing platform		
Illumina	810	67.4
Roche454	180	15
Ion Torrent	87	7.2
RT–qPCR	77	6.4
Other (12 more)	48	4.0
16S variable region		
V3-4	363	36.7
V4	297	30.0
V1-3	90	9.1
V1-2	63	6.4
V4-5	44	4.4
Other (16 more)	133	13.4
Statistical analysis		
Statistical test		
Mann–Whitney (Wilcoxon)	339	29.2
LEfSe	330	28.4
Kruskal–Wallis	94	8.1
DESeq2	85	7.3
t-test	73	6.3
Other (36 more)	240	20.7
Multiple hypothesis correction		
Yes	576	48.6
No	608	51.4
Significance threshold		
0.05	1,049	88.1
0.1	68	5.7
0.01	28	2.4
0.001	12	1.0
Other (18 more)	34	2.8
α-Diversity		
Increased	187	15.3
Decreased	229	18.7
Unchanged	410	33.4
Not reported	399	32.6

Note that the unit in the count column is the number of studies for study design and the number of experiments for all other variables. BugSigDB defines experiments as semantic units within studies based on a defined set of characteristics about study species, lab analysis, statistical analysis and α-diversity ("Definition of semantic concepts" in the Methods).

("Signature similarity" in the Methods). Hierarchical clustering of signature similarity for both similarity measures was in good agreement, reflecting the dominance of genera reported so far in BugSigDB, but demonstrated better resolution using semantic similarity than the sparse results obtained from the application of Jaccard similarity (Supplementary Fig. 4). The advantages of semantic similarity may grow as taxonomic ranks in BugSigDB become more mixed due to increased reporting of species-level results made possible by shotgun metagenomics.

To assess replication by independent studies of the same condition, we compared semantic similarity between signatures reported for a single condition to the similarity of randomly sampled signatures

https://doi.org/10.1038/s41587-023-01872-y

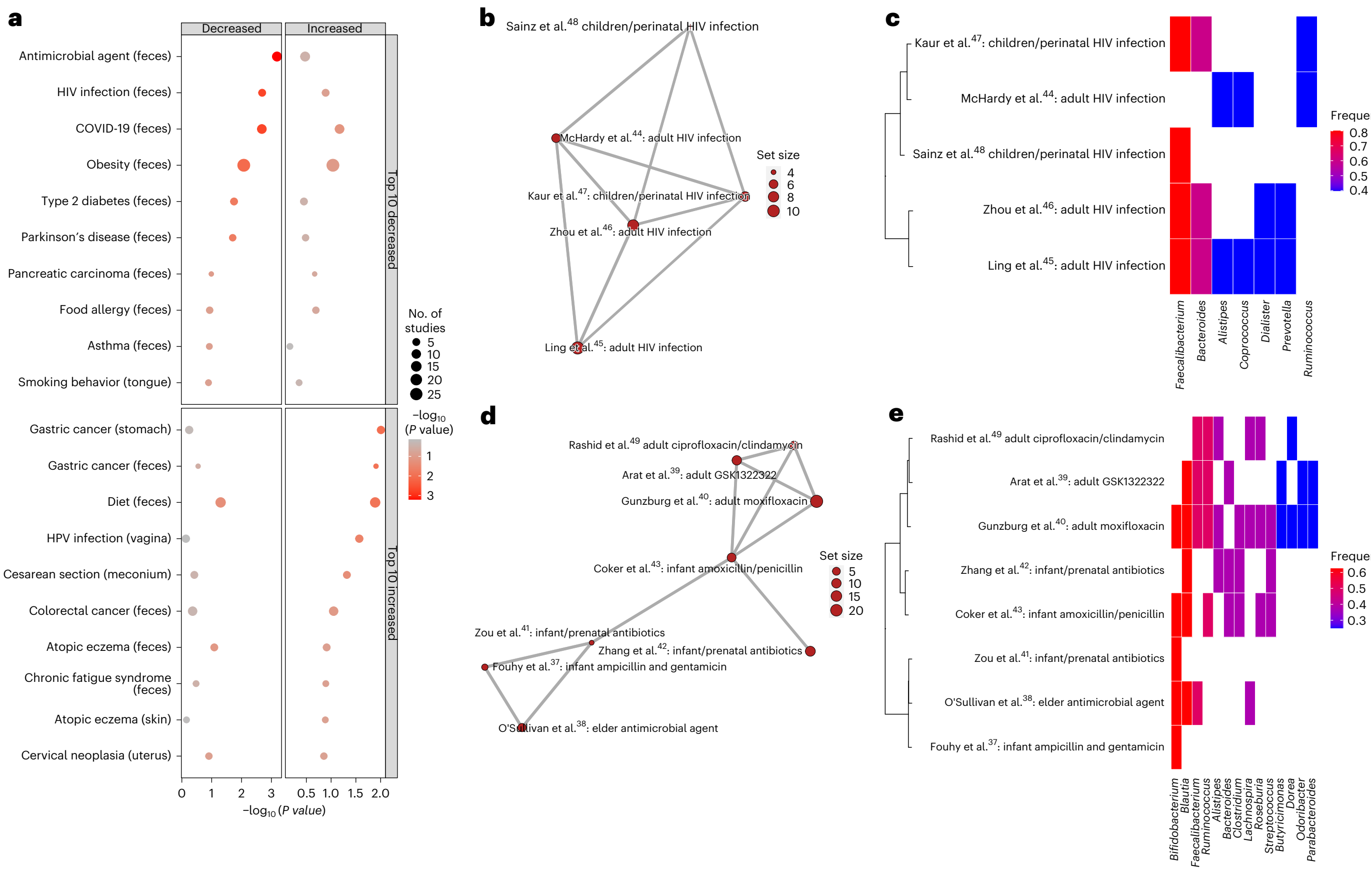

Fig. 2 | Analysis of signature similarity in BugSigDB identifies conditions with replicable microbiome changes across studies. a, Signature similarity analysis for conditions from at least two studies for the same body site. Signatures were stratified by direction of abundance change into signatures with decreased abundance (left) and increased abundance (right) in the study group. Conditions are divided into two groups: conditions with highest reproducibility for signatures of decreased abundance (top) and increased abundance (bottom) in the study condition. Shown is the nominal *P* value (*x* axis, negative log scale) obtained from testing whether the semantic similarity of signatures for each condition (*y* axis) exceeds the semantic similarity of randomly sampled signatures (one-sided resampling test). The size of each dot corresponds to the number of studies investigating a condition; HPV, human papillomavirus. **b**–**e**, Example exploration of two top-ranked conditions in more detail: fecal signatures of decreased abundance for individuals infected with HIV (**b** and **c**) and fecal signatures of decreased abundance for individuals treated with antibiotics (**d** and **e**). Semantic similarity between signatures is shown in **b** and **d**. Each node corresponds to a signature. The size of each node is proportional to the number of taxa in a signature. More similar signatures are connected by shorter and thicker edges. Microbial contents of the signatures shown in **b** and **d** (*y* axis) delineating the genera contained in these signatures (*x* axis) are shown in **c** and **e**. Signatures are clustered by semantic similarity, and the relative frequency of each genus across signatures is indicated by a color scale. The display of contained taxa is restricted to genera occurring in at least two signatures (full signatures are shown in Supplementary Figs. 5 and 6). Additional work cited includes refs. 37–49.

in repeated simulation and ranked conditions based on the resulting empirical *P* value (Fig. 2a). Differential abundance signatures of antibiotic treatment and HIV infection were among the most consistent, each investigated by five or more studies in BugSigDB[37–49]. Closer inspection of signatures of decreased abundance after antibiotics treatment (semantic similarity = 0.64, one-sided resampling $P = 0.0005$; Fig. 2d,e and Supplementary Fig. 5) revealed enrichment for genera of fastidious anaerobes that are often short-chain fatty acid producers and displayed frequent loss of *Bifidobacterium* and *Blautia*, in agreement with previous reports[50]. However, signatures of decreased abundance in individuals infected with HIV versus those in uninfected healthy individuals (semantic similarity = 0.68, resampling $P = 0.002$; Fig. 2b,c and Supplementary Fig. 6) displayed a loss of abundant members of healthy gut microbial communities, as typically observed for diseases associated with gastrointestinal inflammation[51], but also resembled a response to antibiotics treatment, a likely side effect of antibiotics often being prescribed for individuals positive for HIV infection to prevent or treat opportunistic and associated infections[52]. Additional examples of replicable microbiome shifts between studies included similarity (1) among fecal signatures of decreased abundance in individuals with coronavirus disease 2019 (COVID-19), which displayed alterations associated with intensive care units and antibiotics treatment[53] (Supplementary Fig. 7), and (2) among stomach signatures of individuals with gastric cancer driven by consistently increased abundance of *Streptococcus*, *Lactobacillus* and *Prevotella* (Supplementary Fig. 8).

Bug set enrichment analysis of colorectal cancer signatures

We integrated BugSigDB signatures with the manually curated metagenomic datasets from curatedMetagenomicData[54] to systematically benchmark enrichment methods from the EnrichmentBrowser package[7,55] and evaluate whether top-performing gene set enrichment methods can be directly applied to microbiome data. We applied two enrichment methods that have performed well in previous benchmarking of gene set analysis methods[7] (overrepresentation analysis (ORA[11]) and pathway analysis with downweighting of overlapping genes (PADOG[31])) to ten colorectal cancer datasets from curatedMetagenomicData. We performed enrichment analysis of all microbiome signatures from BugSigDB simultaneously and used as 'spike-in' controls two signatures of the colorectal cancer-associated fecal microbiome derived previously by two independent meta-analyses of individual

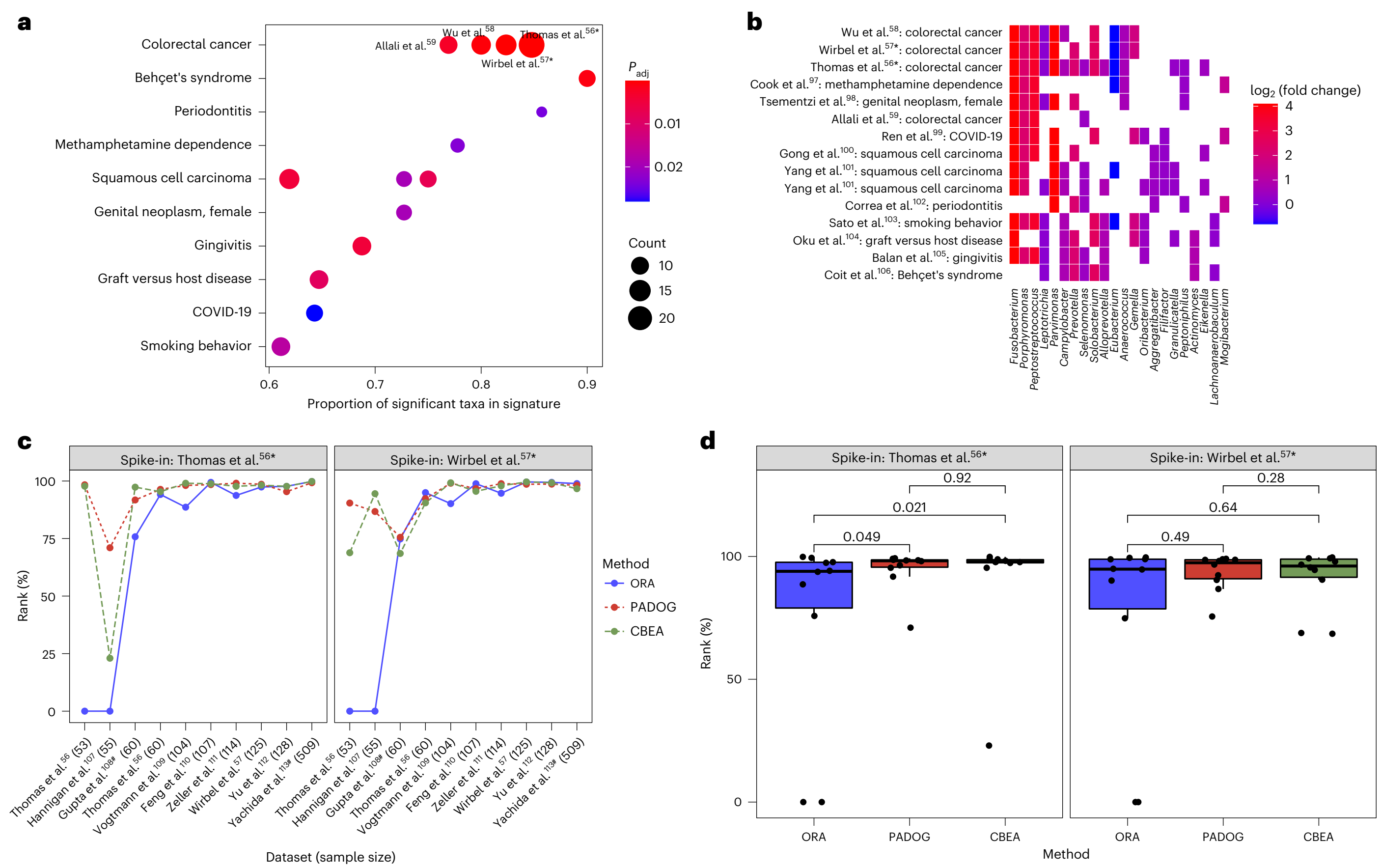

Fig. 3 | Bug set enrichment analysis reveals associations of the human microbiome with colorectal cancer and other disease phenotypes. a, ORA of BugSigDB signatures with differentially abundant genera between 662 colorectal cancer samples and 653 control samples pooled from ten published metagenomic datasets. Each dot corresponds to an enriched published signature from BugSigDB (false discovery rate *P* value (P_{adj}) of <0.05, one-sided Fisher's exact test, color scale). The size of each dot corresponds to the number of differentially abundant genera in a signature, given as a proportion on the *x* axis. Asterisks (*) indicate prior meta-analytic signatures from Thomas et al.[56] and Wirbel et al.[57], which reported differentially abundant species and genera from meta-analysis of eight colorectal cancer datasets. These act as positive controls, as several of the datasets being tested for enrichment were included at the time these signatures were derived. **b**, Differentially abundant genera (*x* axis) in the top 15 enriched signatures (*y* axis) from the overrepresentation analysis in **a**. The *x* axis is sorted by occurrence frequency of each genus in decreasing order, and the *y* axis is sorted by Jaccard similarity between signatures. **c**, Percentile of ranks (*y* axis) of both meta-analytic signatures for ORA, PADOG and CBEA relative to all signatures when applied to the ten published metagenomic datasets individually (*x* axis). Number symbols (#) indicate datasets not included as training sets in the meta-analyses of Thomas et al.[56] and Wirbel et al.[57]. **d**, Assessment of statistically significant differences of the percentile ranks on the $n = 10$ independent datasets shown in **c** between the three methods using a two-sided Wilcoxon's signed-rank test. Box plots show median (vertical line), interquartile range (box) and ±1.5× interquartile range (whiskers). Additional work cited includes (refs. 97–113).

participant data[56,57] from eight of the studies included in this dataset (Fig. 3). The two signatures based on meta-analysis are thus expected to be among the most enriched of all microbiome signatures in the eight training studies for the meta-analyses due to their shared utilization in both datasets and in the remaining two independent studies due to the relevance and large sample size of the colorectal cancer signatures. The two meta-analysis signatures can thus also be considered robust against spurious signals from studies with small sample sizes that were not included in the analysis (Supplementary Results Section 3.2 and Supplementary Tables 5 and 6).

ORA of 647 BugSigDB signatures yielded 19 signatures enriched in contrasts using 662 colorectal cancer samples and 653 control samples from ten datasets (false discovery rate *P* value of <0.05, one-sided Fisher's exact test; Fig. 3a). Only signatures containing five or more genera associated with any condition were included. The two positive-control spike-in signatures from Thomas et al.[56] and Wirbel et al.[57] were top ranked, as expected. Other enriched signatures included colorectal cancer signatures from Wu et al.[58] and Allali et al.[59]. These are notable because neither study was included in the ten datasets from which the meta-analysis signatures were computed. Additionally, both were based on 16S amplicon sequencing, whereas the meta-analysis signatures were based on shotgun sequencing. This analysis therefore also provides a proof of concept for integrating species-level signatures from shotgun metagenomic data with genus-level signatures derived from 16S amplicon profiles ("Bug set enrichment analysis" in the Methods) and also affords independent replication of the signatures from Thomas et al.[56] and Wirbel et al.[57].

The presence of 11 enriched signatures (58%) from saliva samples from studies of oral diseases, such as gingivitis, peptic esophagitis and oral carcinoma, is consistent with recent reports that oral-to-gut microbial introgression is a feature of colorectal cancer[60] and that periodontal diseases are a well-established risk factor for colorectal cancer[61,62]. Frequently overlapping genera between the enriched signatures include *Fusobacterium*, *Porphyromonas* and *Peptostreptococcus*, and all displayed strongly increased abundance in individuals with colorectal cancer relative to healthy controls (Fig. 3b).

Although these findings demonstrate the usefulness of ORA as a fast and effective enrichment method for microbiome signatures, the method has known shortcomings in the presence of correlated features[63] or an inappropriately large feature universe[64]. The PADOG method, a top performer in several independent assessments[7,65,66], is theoretically superior, as it applies sample permutation to preserve correlations and, by working on the full abundance matrix, does not require thresholding on differential abundance or the definition of a feature universe. In addition, the method downweights frequently overlapping microbes between signatures (such as those displayed in Fig. 1d), leading to increased sensitivity and the identification of more specific signatures for the phenotype under investigation.

We therefore benchmarked PADOG against ORA in each of the ten individual participant shotgun metagenomics datasets of colorectal cancer and compared the rankings of the two spike-in signatures from the meta-analyses of Thomas et al.[56] and Wirbel et al.[57] (Fig. 3c). On average, PADOG ranked the spike-in signatures better than ORA, although the difference was statistically significant only for the spike-in signature from Thomas et al.[56] ($P = 0.049$, two-sided Wilcoxon signed-rank test; Fig. 3d). This difference was largest for datasets with smaller sample sizes, where the lack in power was more detrimental for ORA than for PADOG (Supplementary Fig. 9).

Despite the apparent effectiveness of established gene set enrichment methods for application to microbiome data, these methods were not developed with microbiome data in mind[34,67]. Competitive balances for taxonomic enrichment analysis (CBEA[18]) is a recent taxonomic enrichment method specifically developed for microbiome data that accounts for compositionality via application of an isometric log ratio transformation of relative abundance data for the computation of sample-level enrichment scores. Benchmarked against ORA and PADOG in the colorectal cancer setting (Fig. 3c), CBEA tended to rank the spike-in signature from Thomas et al.[56] significantly higher than ORA ($P = 0.021$, Wilcoxon signed-rank test) but did not display a notable performance gain over PADOG ($P = 0.92$; Fig. 3d).

Microbe co-occurrence and mutual exclusivity within signatures

BugSigDB enables the exploration of compositional patterns within signatures of differential abundance for different body sites. Focusing on 1,194 signatures of fecal microbiomes from 311 published studies, we analyzed patterns of co-occurrence and mutual exclusivity for individual microbes and groups of microbes (Fig. 4). Inspection of the top 20 genera most frequently reported as differentially abundant in signatures from fecal samples (Fig. 4a) revealed genera predominantly belonging to the phyla Firmicutes (13 genera) and Bacteroidetes (4 genera), in agreement with those being the dominant phyla of the human gut microbiome[68,69]. Among the 20 were *Bacteroides*, *Prevotella* and *Ruminococcus*, three dominant gut genera that are highly variable in relative abundance[70] and have a large effect on gut microbiome clustering[71].

When comparing the proportions of signatures where these genera were reported either with increased or decreased abundance in the study sample group (Fig. 4a, Abundance in group), the genera most disproportionately increased in the study group were *Enterococcus* (64 of 73 signatures, 87.7%), *Lactobacillus* (73 of 105, 69.5%), *Veillonella* (50 of 74, 67.6%) and *Streptococcus* (83 of 128, 64.8%). Pathogenicity of certain *Streptococcus* and *Enterococcus* species is well documented[72,73], whereas *Lactobacillus* and *Veillonella* are typically considered to be commensal gut microbiota of limited pathogenicity[74,75]. Genera decreased in the study group were *Roseburia* (77 of 103, 74.8%), a beneficial gut organism with established anti-inflammatory activity[76], and *Alistipes* (61 of 88, 69.3%), for which protective and harmful associations with a range of diseases have been described[77].

To elucidate the extent to which frequency of differential abundance in disease phenotypes can be explained by prevalence in the healthy gut microbiome, we contrasted these findings with the prevalence of these genera in 9,623 stool samples from healthy adult participants from 68 different studies in curatedMetagenomicData (Fig. 4a, Prevalence). We observed a strong negative correlation between prevalence, measured as the percentage of control samples in which the genus is observed at non-zero relative abundance, and the proportion of BugSigDB signatures in which the genus is reported with increased abundance in disease ($r = -0.84$, $P = 3 \times 10^{-6}$, two-sided Pearson correlation test; Fig. 4c). This indicates that across the many diseases and exposures present in BugSigDB, high-prevalence genera tend to be lost in study groups relative to controls, and low-prevalence genera have increased abundance in the study group. This was particularly apparent for *Enterococcus*, a genus of low prevalence in healthy samples (13%) that was reported almost exclusively with increased abundance in the diseased group (64 of 73 signatures, 88%). The presence of *Enterococcus*, accompanied by exclusion of other genera of the Clostridia class (below), may therefore be considered as a commonly reported 'dysbiotic' signature.

The human gut is a complex ecosystem where microbes compete and cooperate[78]. To investigate which of these interactions are associated with disease, we next studied patterns of co-occurrence and mutual exclusivity for the top 20 genera most frequently reported as differentially abundant. For each pair of microbes, we counted the number of signatures where both microbes were reported with either the same or opposite direction of change in relative abundance (Fig. 4a, top and bottom heat maps). This resulted in clusters of co-occurrence driven primarily by functional and phylogenetic similarity, with frequent co-occurrence of genera of the phylum Bacteroidetes and blocks of positive associations in the class Clostridia (Fig. 4a, top red heat map, third cluster from top to bottom). However, clusters of mutual exclusivity displayed clear signs of the established Firmicutes–Bacteroidetes gradient in gut microbiomes[68] and a strong negative association between *Bacteroides* and *Prevotella* within the phylum Bacteroidetes, as previously reported[54] and also observed in healthy samples ($r = -0.49$, $P < 2.2 \times 10^{-16}$, two-sided Spearman correlation test; Fig. 4a, bottom blue heat map, third cluster from top to bottom). Overall, these patterns largely recapitulated the correlation of these genera in healthy adult stool samples (Fig. 4b), arguing against the existence of specific disease-promoting interactions between these genera. Exceptions were patterns of pronounced mutual exclusivity within the class Clostridia in signatures associated with disease, as observed for *Clostridium*, *Enterococcus* and *Streptococcus*, which were not observed in healthy samples.

Shared and exclusive patterns in pooled microbial signatures

BugSigDB provides opportunities for the discovery of microbial biomarkers and reassessment of previous findings across a much larger and more heterogeneous data source than previously possible. To identify similarities between microbial shifts within and across body sites, we aggregated signatures for one body site at a time and within body sites for one condition at a time ("Signature pooling" in the Methods). To account for differences in sample size between studies, we applied a voting approach where each taxon of a pooled signature obtained a weight based on the aggregated sample size of reporting studies and performed hierarchical clustering based on pairwise similarity between the weighted metasignatures (Fig. 5). Clustering of 27 metasignatures, representing body sites studied by at least two studies in BugSigDB and generated from 1,909 individual signatures of either increased or decreased differential abundance in the study group, resulted in two major body site clusters: a cluster primarily composed of oral and nasal sites and a cluster dominated by vaginal and gastrointestinal sites (Fig. 5a). This clustering was largely invariant to the similarity measure used for clustering (Supplementary Fig. 10), confirming the expected dominant effect of host body site of origin and, in particular, the availability of oxygen.

Within body site clustering of fecal metasignatures of 34 different conditions, studied by at least two studies in BugSigDB and generated from 504 signatures of increased relative abundance in the study group, revealed similarities in reported differential abundance patterns between disease phenotypes (Fig. 5b). This included similarities in microbial shifts for (1) HIV infection and different gastrointestinal cancers, both characterized by chronic inflammation of the gastrointestinal tract and microbial signatures that point to

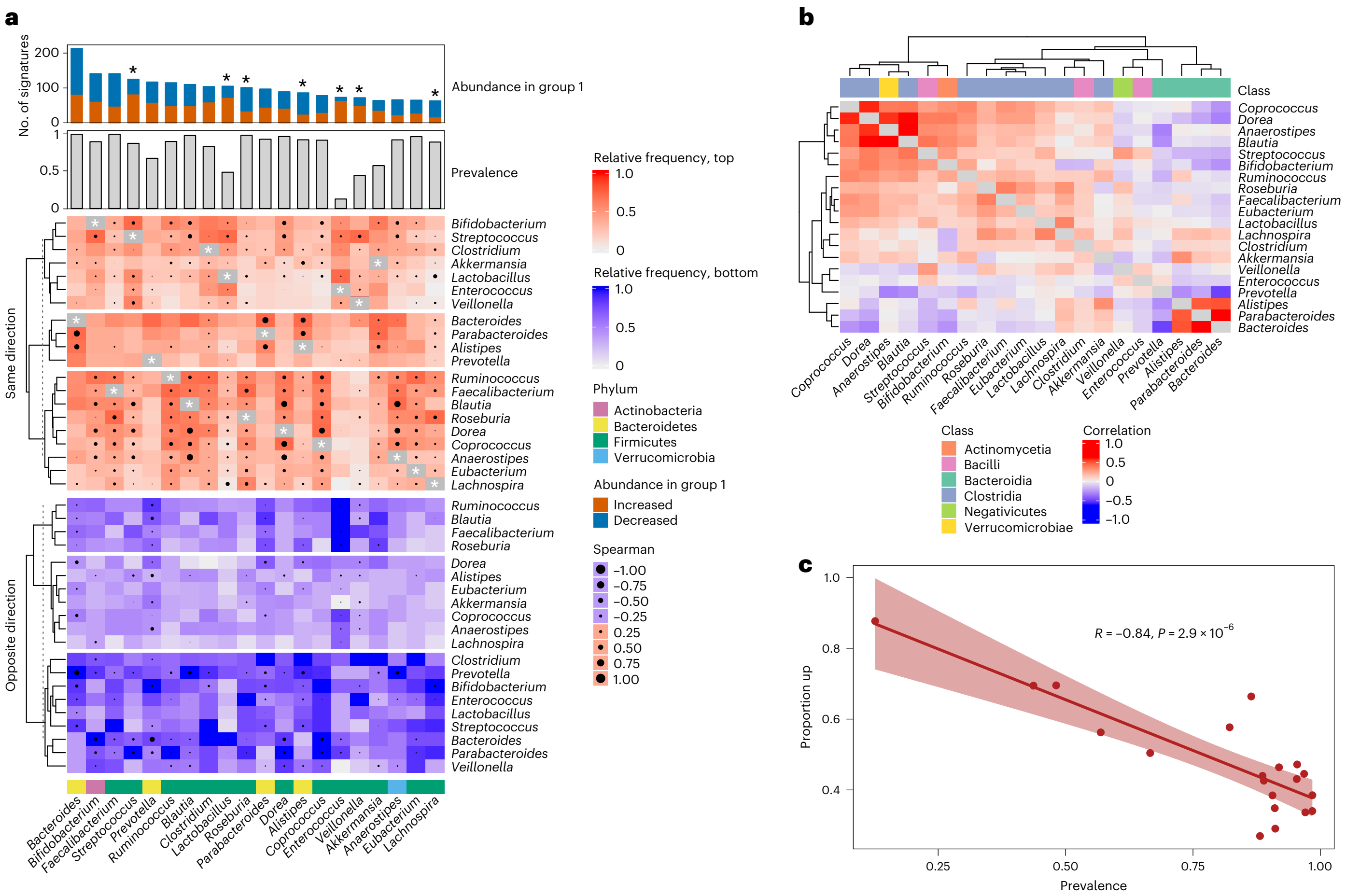

Fig. 4 | Microbe co-occurrence and mutual exclusivity in BugSigDB reveals recurrent groups of taxa within signatures of differential abundance. a, Microbe–microbe co-occurrence and mutual exclusivity across 1,194 signatures of fecal microbiomes from 311 published studies. 'Abundance in group 1' shows the top 20 genera most frequently reported as differentially abundant in the study group of these signatures. Stars indicate microbes that tend to be predominantly unidirectional, that is, reported specifically with increased or decreased abundance in the study group (based on a greater than twofold difference in proportions). 'Prevalence' shows the prevalence of these genera in 9,623 healthy adult stool samples pooled from 68 different studies. The top heat map shows co-occurrence between these top 20 genera, that is, in how many signatures these microbes are reported together as differentially abundant with the same direction of abundance change (that is, occurring together in either the signature of increased or decreased abundance). The bottom heat map shows mutual exclusivity between these top 20 genera, that is, in how many signatures these microbes are reported as differentially abundant with opposite direction of abundance change (that is, one microbe in the signature of increased abundance and the other in the signature of decreased abundance or vice versa). **b**, Spearman correlation between the top 20 genera in healthy adult stool samples as indicated by circles of varying size in **a**. **c**, Proportion of signatures where each of the top 20 genera was reported with increased abundance in the study group (*y* axis) against the prevalence of these genera (*x* axis) in healthy adult stool samples as shown in **a**. The correlation coefficient and *P* value of a two-sided Pearson correlation test are shown at the top. The error bands indicate the 95% confidence interval of the linear regression fit.

shared pathogenic pathways including tryptophan catabolism and butyrate synthesis[79], (2) chronic kidney disease and autism, linked through deleterious copy number variants[80,81] and pathogenic gut microbiota-derived metabolites produced by species of the Clostridia class[82,83], and (3) type 2 diabetes and schizophrenia, consistent with observations that people with schizophrenia are at increased risk of type 2 diabetes and, conversely, that traditional risk factors for type 2 diabetes are common in people with schizophrenia and can affect the gut microbiome, especially obesity, poor diet and a sedentary lifestyle[84]. A strong enrichment of genera of the Clostridia class drove the similarity between the metasignatures of Hashimoto's thyroiditis (12 of 14 genera, 85.7%, $P = 2.1 \times 10^{-05}$, two-sided proportion test) and chronic fatigue syndrome (11 of 17, 64.7%, $P = 0.003$). Chronic fatigue is common in individuals with Hashimoto's thyroiditis, a disorder that affects the thyroid gland potentially through a gut microbiota-driven thyroid–gut axis[85] that may reduce thyroid hormone production and cause extreme fatigue[86].

Genera reported as differentially abundant exclusively in one condition, in particular those replicated by independent studies, are candidates for condition-specific biomarkers (Supplementary Fig. 11, prioritized by relative sample size). This included an exclusive abundance increase of the genus *Hungatella* for colorectal cancer, which is notable given a reported role of *Hungatella hathewayi* in driving host colonic epithelial cell promoter hypermethylation of tumor suppressor genes in colorectal cancer[87]. However, an exclusive decrease in abundance was observed for *Marvinbryantia* in type 2 diabetes, for which a reduction in *Marvinbryantia* through treatment with glucagon-like peptide-1 receptor agonist drugs, such as liraglutide, has been shown to contribute to treatment success[4].

Discussion

We compiled published signatures of microbial differential abundance in the BugSigDB database, assessed applicability of GSEA methods for enrichment analysis of microbial signatures and identified common

https://doi.org/10.1038/s41587-023-01872-y

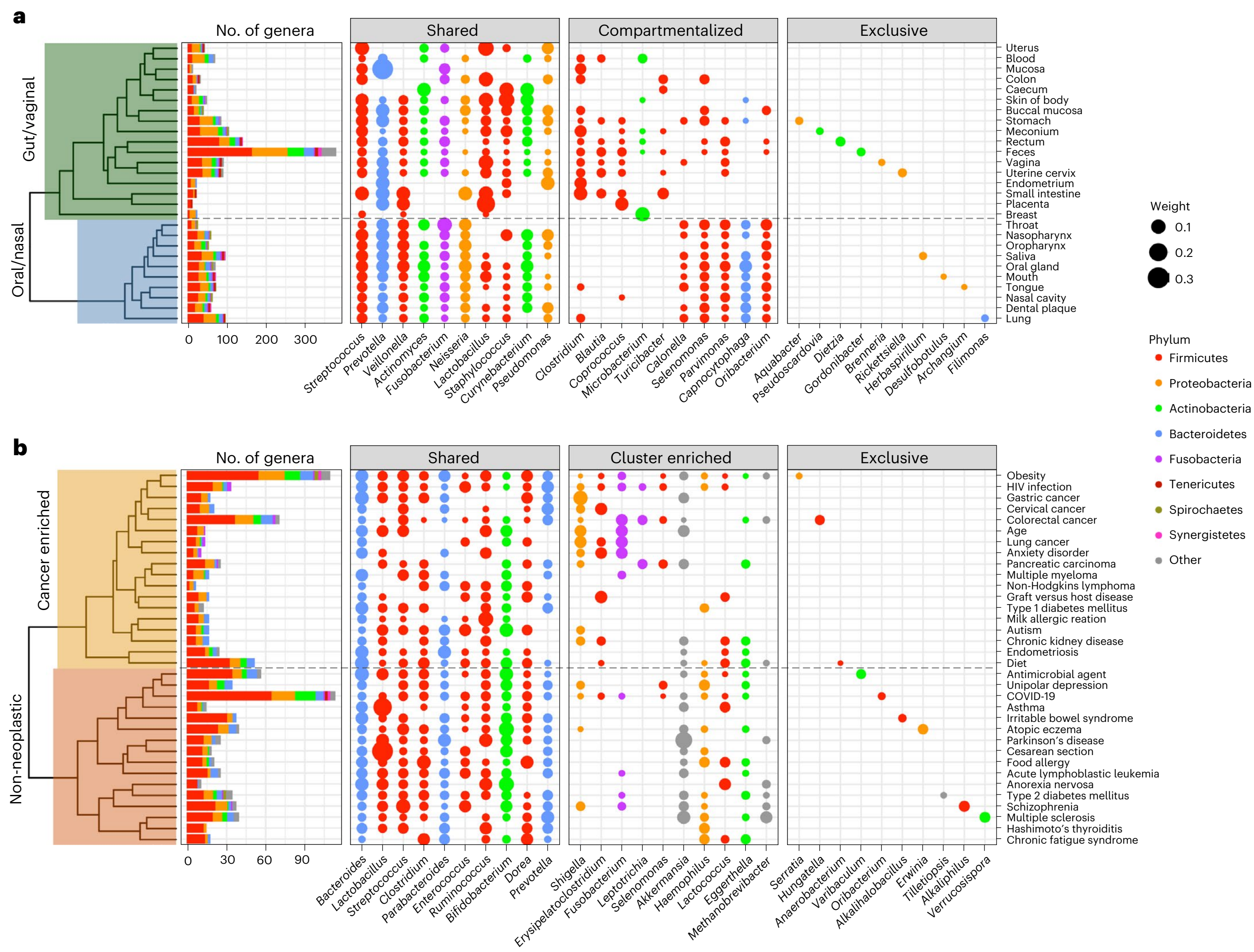

Fig. 5 | Pooled genus-level microbial signatures display robust body site specificity and shared and exclusive patterns of gut dysbiosis between disease phenotypes. a, Clustering of 27 metasignatures (*y* axis) representing host body sites studied by at least two studies in BugSigDB and generated from 1,909 individual signatures. The clustering shows a separation of vaginal and gastrointestinal sites (top) from oral and nasal sites (bottom). A total of 560 different genera (*x* axis) colored by phylum were observed across the metasignatures. The numbers of genera in each metasignature are shown in the bar plot on the left. The following three panels display representative genera that frequently occur across metasignatures ('shared'), predominantly occur in either cluster of body sites ('compartmentalized') or were reported in only one body site as differentially abundant ('exclusive'). The size of each dot corresponds to the relative sample size of studies reporting a genus as differentially abundant. **b**, Clustering of fecal metasignatures drawn from 34 different conditions, each represented by at least two studies in BugSigDB and generated from 504 individual signatures with increased abundance in the study group. The clustering shows a separation of cancer phenotypes (top) from non-neoplastic diseases (bottom). A total of 248 different genera colored by phylum were observed across the fecal metasignatures. The number of genera in each metasignature is shown in the bar plot on the left.

patterns of co-occurrence and mutual exclusion in differential abundance results across a broad sample of the human microbiome literature. BugSigDB is a Semantic MediaWiki that allows contribution, review and correction by the microbiome research community and is usable through its web interface and through bulk exports compatible with all major GSEA software. BugSigDB supports any taxon and any host species present in the NCBI Taxonomy and therefore non-human hosts and studies of viromes and eukaryotes. BugSigDB has been initially seeded by ~25 trained curators with more than 2,500 manually curated signatures from the figures, tables, main text and supplementary materials of more than 600 primary publications, providing a broadly relevant collection of machine-readable knowledge of microbial differential abundance. Manually curated metadata include study design, geography, health outcomes, host body sites and experimental and statistical methods. Initial analyses of the database (1) identify human diseases with the highest consistency among independently published signatures of differential abundance, (2) demonstrate the capability of established GSEA methods to prioritize colorectal cancer signatures in the analysis of individual participant colorectal cancer datasets and add to evidence of frequent introgression of oral pathobionts into the gut, (3) demonstrate that the prevalence of genera in fecal microbiomes of control populations is strongly correlated with being reported as decreased across diverse study conditions, (4) confirm the broad relevance of the Firmicutes–Bacteroidetes gradient in shaping common patterns of co-occurrence and mutual exclusivity in the published microbiome literature and identify other less dominant patterns and (5) define sample size-adjusted consensus signatures of body sites and conditions that can simplify and clarify future analyses.

There is concern over replicability of human microbiome studies due to numerous sources of variation in complex experimental and computational quantification pipelines[88,89]. We propose an approach based on semantic similarity that can be used in systematic reviews to evaluate replication of differential abundance signatures of mixed taxonomic levels reported by independent studies, even when different laboratory methods, such as 16S amplicon sequencing versus MGX, were used. This taxonomy-aware framework provides an effective assessment of the replicability of microbiome differential abundance signatures for subsets of the literature, allowing ranking of the relative replicability of microbiome signatures consisting of different taxonomic levels across many disease phenotypes. We identified signatures associated with antibiotics treatment and chronic inflammation of the gastrointestinal tract as having the highest level of consistency or replication in signatures reported by independent studies.

We pooled signatures across host body sites and experimental conditions to expand the analysis of replicability also to the ~45% of signatures containing fewer than five taxa that are too small individually to be effectively compared between studies. Using a voting approach that weights each taxon by sample size of reporting studies, we constructed consensus 'metasignatures', revealing shared and specific patterns of gut dysbiosis by disease phenotype. These metasignatures provide a framework for simplifying the interpretation of results from future studies in the context of the published literature, distinguishing specific from generic results and informing enrichment tests by defining a universe of reported abundance changes for a body site or condition of interest.

Although single-species biomarkers are of primary interest for therapeutic interventions, they are not sufficient for capturing complex ecological patterns of co-occurrence and mutual exclusivity and interactions between microbes that may be relevant to health and disease. Inspection of published signatures in BugSigDB is an alternative approach to studying ecological patterns that complements the analysis of co-occurrence and coexclusion in individual participant metagenomic profiles. In agreement with previous results[54], but across a much larger corpus of microbiome studies, we confirmed coexclusions not specific to disease, such as the phylum-level Firmicutes–Bacteroides gradient and the genus-level *Bacteroides*–*Prevotella* gradient within the Bacteroides phylum. However, patterns of mutual exclusivity for *Clostridium*, *Enterococcus* and *Streptococcus* genera were specific to disease-associated signatures and were not detectable in healthy samples. To distinguish between disease markers and common false positives in biomarker discovery, we compared stool signatures of disease conditions to prevalence in ~10,000 stool specimens from healthy participants. Across the many diseases and exposures present in BugSigDB, prevalent genera in healthy fecal microbiomes tend to be reported with decreased abundance in diseased-associated fecal microbiomes. On the other hand, genera of low prevalence in healthy fecal microbiomes, such as *Enterococcus*, tend to be reported with increased abundance in the disease group. Genera such as *Lactobacillus* and *Veillonella*, which are prevalent in the stools of healthy individuals and are frequently reported as increased in many study conditions, are more likely false positives or at least are not well suited as candidate biomarkers. Future work can use stratification of BugSigDB signatures by experimental, cross-sectional and longitudinal study designs to better infer causality.

BugSigDB enables systematic comparison of microbial signatures from new microbiome studies to previously published results. Although concepts of GSEA are applicable, microbiome data present new challenges, including smaller signature sizes, taxonomic relationships between features and mixed-taxonomy signatures. We therefore benchmarked two gene set enrichment methods (ORA[11] and PADOG[31]) that performed well in previous benchmarking[7] and a recent taxonomic enrichment method (CBEA[18]) developed specifically for microbiome data. Valid application of gene set enrichment methods is limited to analysis at a single taxonomic level, and ORA further requires defining a realistic feature 'universe' and significance threshold for differential abundance; nonetheless, all methods performed well in prioritizing signatures of colorectal cancer across numerous colorectal cancer datasets. As PADOG addresses both (1) shortcomings of ORA in the presence of intermicrobe correlation and (2) compositional bias in signature databases with certain taxa occurring more frequently than others, typically a result of technical or biological sampling bias[67,90], we recommend PADOG over ORA for the routine application of enrichment analysis to microbial signatures, especially for datasets with smaller sample sizes where a lack in power typically hinders detecting individually differentially abundant features. Recently emerging microbiome-specific enrichment methods such as CBEA[18] have the advantage of taking into account the compositional nature of microbiome data; however, they lack the independent benchmarking and implementation of major GSEA approaches. Although we found CBEA to be a sound alternative to ORA and PADOG, we did not observe notable gains of applying CBEA over PADOG in the colorectal cancer benchmark setup, indicating that basic study characteristics, such as sample size and, to a lesser extent, accounting for correlation within microbial signatures, have a larger impact on identifying relevant signatures in practice than mitigating effects of compositionality.

BugSigDB is a large and diverse collection of the currently available literature on microbial differential abundance and thus also has certain limitations inherent to the currently available literature. Because more than 90% of the studies included in the first release of BugSigDB are based on 16S amplicon sequencing, enrichment analyses were performed at the genus level. However, some genera are functionally heterogeneous, such as streptococci, which groups deadly pathogens with common commensals and useful food-fermenting species. Species- and strain-level variations are neglected, although they can contribute to functional differences between individuals that are important in a clinical context[91,92]. With the availability of more shotgun sequencing studies in the future, it will be possible to perform enrichment analysis at higher taxonomic resolution up to the species or strain level. Furthermore, studies included in BugSigDB are heterogeneous in their design and execution, including antibiotics exclusion time frames ranging from current use to within the previous year. Restoration of baseline microbial composition following antibiotic treatment typically takes around 1 month in children and 1.5 months in adults, although several common species of the gut microbiome might take substantially longer[50,93]. Inclusion of this and other study information in BugSigDB allows further investigation and sensitivity analysis into potential sources of heterogeneity in the literature. We anticipate that broader adoption of the recently developed STORMS reporting guidelines for human microbiome studies[94] will contribute to more efficient extraction of information from the literature for BugSigDB.

Although natural language parsing programs have the potential to complement the manually curated information in BugSigDB, the majority of the curated information is too complex for currently available text mining algorithms. Natural language parsing programs typically extract patterns from unstructured text on a sentence-by-sentence basis[95,96], but BugSigDB standardizes microbial signatures and associated experimental, epidemiological and statistical methods from diverse figures, tables, supplements and textual descriptions that often span multiple sentences. This places many key results outside the current capabilities of text mining applications and necessitates manual curation, but improvements in machine learning, using BugSigDB as a gold-standard dataset, may enable more efficient extraction of published microbiome methods and results in the future. Moreover, automated contributions of signatures from differential abundance software via the BugSigDB application programming interface can streamline standardized reporting of results. As the community contributes additional host species and signatures of microbial physiology and morphology, BugSigDB will dynamically expand, leading to complementary insights and improvements to the systematic interpretation of microbiome studies.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-023-01872-y.

References

1. Jonsson, A. L. & Bäckhed, F. Role of gut microbiota in atherosclerosis. *Nat. Rev. Cardiol.* **14**, 79–87 (2017).
2. Tang, W. H. W., Kitai, T. & Hazen, S. L. Gut microbiota in cardiovascular health and disease. *Circ. Res.* **120**, 1183–1196 (2017).
3. Schwabe, R. F. & Jobin, C. The microbiome and cancer. *Nat. Rev. Cancer* **13**, 800–812 (2013).
4. Gurung, M. et al. Role of gut microbiota in type 2 diabetes pathophysiology. *EBioMedicine* **51**, 102590 (2020).
5. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10**, 1200–1202 (2013).
6. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
7. Geistlinger, L. et al. Toward a gold standard for benchmarking gene set enrichment analysis. *Brief. Bioinform.* **22**, 545–556 (2020).
8. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
9. Geistlinger, L., Csaba, G., Küffner, R., Mulder, N. & Zimmer, R. From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics* **27**, i366–i373 (2011).
10. Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–1607 (2006).
11. Goeman, J. J. & Bühlmann, P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23**, 980–987 (2007).
12. McLean, C. Y. et al. GREAT improves functional interpretation of *cis*-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
13. Chagoyen, M., López-Ibáñez, J. & Pazos, F. Functional analysis of metabolomics data. *Methods Mol. Biol.* **1415**, 399–406 (2016).
14. Ried, J. S. et al. PSEA: phenotype set enrichment analysis—a new method for analysis of multiple phenotypes. *Genet. Epidemiol.* **36**, 244–252 (2012).
15. Ma, W., Huang, C., Zhou, Y., Li, J. & Cui, Q. MicroPattern: a web-based tool for microbe set enrichment analysis and disease similarity calculation based on a list of microbes. *Sci. Rep.* **7**, 40200 (2017).
16. Dhariwal, A. et al. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.* **45**, W180–W188 (2017).
17. Kou, Y., Xu, X., Zhu, Z., Dai, L. & Tan, Y. Microbe-set enrichment analysis facilitates functional interpretation of microbiome profiling data. *Sci. Rep.* **10**, 21466 (2020).
18. Nguyen, Q. P., Hoen, A. G. & Frost, H. R. CBEA: competitive balances for taxonomic enrichment analysis. *PLoS Comput. Biol.* **18**, e1010091 (2022).
19. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
20. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
21. Liberzon, A. et al. Molecular Signatures Database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
22. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
23. Culhane, A. C. et al. GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Res.* **40**, D1060–D1066 (2012).
24. Wattam, A. R. et al. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res.* **45**, D535–D542 (2017).
25. Shaaban, H. et al. The Microbe Directory: an annotated, searchable inventory of microbes' characteristics. *Gates Open Res.* **2**, 3 (2018).
26. Bergey, D. H. & Holt J. G. *Bergey's Manual of Systematic Bacteriology*, Vol. 1 (Williams & Wilkins, 1984).
27. Reimer, L. C. et al. BacDive in 2022: the knowledge base for standardized bacterial and archaeal data. *Nucleic Acids Res.* **50**, D741–D746 (2022).
28. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* **13**, R5 (2012).
29. Federhen, S. The NCBI Taxonomy Database. *Nucleic Acids Res.* **40**, D136–D143 (2012).
30. Krötzsch, M., Vrandečić, D. & Völkel, M. Semantic MediaWiki. In *Proc. 5th International Semantic Web Conference, ISWC 2006* (ed. Cruz, I. et al.) 935–942 (Springer, 2006).
31. Tarca, A. L., Draghici, S., Bhatti, G. & Romero, R. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics* **13**, 136 (2012).
32. Malone, J. et al. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112–1118 (2010).
33. Segata, N. et al. Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).
34. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* **8**, 2224 (2017).
35. Chung, N. C., Miasojedow, B., Startek, M. & Gambin, A. Jaccard/Tanimoto similarity test and estimation methods for biological presence–absence data. *BMC Bioinformatics* **20**, 644 (2019).
36. Pesquita, C., Faria, D., Falcão, A. O., Lord, P. & Couto, F. M. Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.* **5**, e1000443 (2009).
37. Fouhy, F. et al. High-throughput sequencing reveals the incomplete, short-term recovery of infant gut microbiota following parenteral antibiotic treatment with ampicillin and gentamicin. *Antimicrob. Agents Chemother.* **56**, 5811–5820 (2012).
38. O'Sullivan, O. et al. Alterations in intestinal microbiota of elderly Irish subjects post-antibiotic therapy. *J. Antimicrob. Chemother.* **68**, 214–221 (2013).
39. Arat, S. et al. Microbiome changes in healthy volunteers treated with GSK1322322, a novel antibiotic targeting bacterial peptide deformylase. *Antimicrob. Agents Chemother.* **59**, 1182–1192 (2015).
40. de Gunzburg, J. et al. Protection of the human gut microbiome from antibiotics. *J. Infect. Dis.* **217**, 628–636 (2018).
41. Zou, Z.-H. et al. Prenatal and postnatal antibiotic exposure influences the gut microbiota of preterm infants in neonatal intensive care units. *Ann. Clin. Microbiol. Antimicrob.* **17**, 9 (2018).
42. Zhang, M. et al. Association of prenatal antibiotics with measures of infant adiposity and the gut microbiome. *Ann. Clin. Microbiol. Antimicrob.* **18**, 18 (2019).
43. Coker, M. O. et al. Specific class of intrapartum antibiotics relates to maturation of the infant gut microbiota: a prospective cohort study. *BJOG* **127**, 217–227 (2020).

44. McHardy, I. H. et al. HIV infection is associated with compositional and functional shifts in the rectal mucosal microbiota. *Microbiome* **1**, 26 (2013).
45. Ling, Z. et al. Alterations in the fecal microbiota of patients with HIV-1 infection: an observational study in a Chinese population. *Sci. Rep.* **6**, 30673 (2016).
46. Zhou, Y. et al. Alterations in the gut microbiota of patients with acquired immune deficiency syndrome. *J. Cell. Mol. Med.* **22**, 2263–2271 (2018).
47. Kaur, U. S. et al. High abundance of genus *Prevotella* in the gut of perinatally HIV-infected children is associated with IP-10 levels despite therapy. *Sci. Rep.* **8**, 17679 (2018).
48. Sainz, T. et al. Effect of a nutritional intervention on the intestinal microbiota of vertically HIV-infected children: The Pediabiota Study. *Nutrients* **12**, 2112 (2020).
49. Rashid, M.-U. et al. Determining the long-term effect of antibiotic administration on the human normal intestinal microbiota using culture and pyrosequencing methods. *Clin. Infect. Dis.* https://doi.org/10.1093/cid/civ137 (2015).
50. Ramirez, J. et al. Antibiotics as major disruptors of gut microbiota. *Front. Cell. Infect. Microbiol.* **10**, 572912 (2020).
51. Alzahrani, J. et al. Inflammatory and immunometabolic consequences of gut dysfunction in HIV: parallels with IBD and implications for reservoir persistence and non-AIDS comorbidities. *EBioMedicine* **46**, 522–531 (2019).
52. Faiela, C. & Sevene, E. Antibiotic prescription for HIV-positive patients in primary health care in Mozambique: a cross-sectional study. *S. Afr. J. Infect. Dis.* **37**, 340 (2022).
53. Szychowiak, P., Villageois-Tran, K., Patrier, J., Timsit, J.-F. & Ruppé, É. The role of the microbiota in the management of intensive care patients. *Ann. Intensive Care* **12**, 3 (2022).
54. Pasolli, E. et al. Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).
55. Geistlinger, L., Csaba, G. & Zimmer, R. Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set- & network-based enrichment analysis. *BMC Bioinformatics* **17**, 45 (2016).
56. Thomas, A. M. et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).
57. Wirbel, J. et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).
58. Wu, N. et al. Dysbiosis signature of fecal microbiota in colorectal cancer patients. *Microb. Ecol.* **66**, 462–470 (2013).
59. Allali, I. et al. Gut microbiome of Moroccan colorectal cancer patients. *Med. Microbiol. Immunol.* **207**, 211–225 (2018).
60. Koliarakis, I. et al. Oral bacteria and intestinal dysbiosis in colorectal cancer. *Int. J. Mol. Sci.* **20**, 4146 (2019).
61. Irfan, M., Delgado, R. Z. R. & Frias-Lopez, J. The oral microbiome and cancer. *Front. Immunol.* **11**, 591088 (2020).
62. Lo, C.-H. et al. Periodontal disease, tooth loss, and risk of serrated polyps and conventional adenomas. *Cancer Prev. Res.* **13**, 699–706 (2020).
63. Wu, D. & Smyth, G. K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.* **40**, e133 (2012).
64. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258 (2007).
65. Tarca, A. L., Bhatti, G. & Romero, R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS ONE* **8**, e79217 (2013).
66. Nguyen, T.-M., Shafi, A., Nguyen, T. & Draghici, S. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol.* **20**, 203 (2019).
67. McLaren, M. R., Willis, A. D. & Callahan, B. J. Consistent and correctable bias in metagenomic sequencing experiments. *eLife* **8**, e46923 (2019).
68. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
69. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
70. Falony, G. et al. Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
71. Costea, P. I. et al. Enterotypes in the landscape of gut microbial community composition. *Nat. Microbiol.* **3**, 8–16 (2017).
72. Krzyściak, W., Pluskwa, K. K., Jurczak, A. & Kościelniak, D. The pathogenicity of the *Streptococcus* genus. *Eur. J. Clin. Microbiol. Infect. Dis.* **32**, 1361–1376 (2013).
73. Fiore, E., Van Tyne, D. & Gilmore, M. S. Pathogenicity of enterococci. *Microbiol. Spectr.* https://doi.org/10.1128/microbiolspec.GPP3-0053-2018 (2019).
74. Harty, D. W., Oakey, H. J., Patrikakis, M., Hume, E. B. & Knox, K. W. Pathogenic potential of lactobacilli. *Int. J. Food Microbiol.* **24**, 179–189 (1994).
75. Actor, J. K. *Elsevier's Integrated Review Immunology and Microbiology* (Elsevier Health Sciences, 2011).
76. Nie, K. et al. *Roseburia intestinalis*: a beneficial gut organism from the discoveries in genus and species. *Front. Cell. Infect. Microbiol.* **11**, 757718 (2021).
77. Parker, B. J., Wearsch, P. A., Veloo, A. C. M. & Rodriguez-Palacios, A. The genus *Alistipes*: gut bacteria with emerging implications to inflammation, cancer, and mental health. *Front. Immunol.* **11**, 906 (2020).
78. Faust, K. et al. Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8**, e1002606 (2012).
79. Herrera, S., Martínez-Sanz, J. & Serrano-Villar, S. HIV, cancer, and the microbiota: common pathways influencing different diseases. *Front. Immunol.* **10**, 1466 (2019).
80. Caubit, X. et al. *TSHZ3* deletion causes an autism syndrome and defects in cortical projection neurons. *Nat. Genet.* **48**, 1359–1369 (2016).
81. Sanna-Cherchi, S. et al. Copy-number disorders are a common cause of congenital kidney malformations. *Am. J. Hum. Genet.* **91**, 987–997 (2012).
82. Peralta-Marzal, L. N. et al. The impact of gut microbiota-derived metabolites in autism spectrum disorders. *Int. J. Mol. Sci.* **22**, 10052 (2021).
83. Clothier, J. & Absoud, M. Autism spectrum disorder and kidney disease. *Pediatr. Nephrol.* **36**, 2987–2995 (2021).
84. Suvisaari, J., Keinänen, J., Eskelinen, S. & Mantere, O. Diabetes and schizophrenia. *Curr. Diab. Rep.* **16**, 16 (2016).
85. Knezevic, J., Starchl, C., Tmava Berisha, A. & Amrein, K. Thyroid–gut–axis: how does the microbiota influence thyroid function? *Nutrients* **12**, 1769 (2020).
86. Ruiz-Núñez, B., Tarasse, R., Vogelaar, E. F., Janneke Dijck-Brouwer, D. A. & Muskiet, F. A. J. Higher prevalence of 'low T3 syndrome' in patients with chronic fatigue syndrome: a case–control study. *Front. Endocrinol.* **9**, 97 (2018).
87. Xia, X. et al. Bacteria pathogens drive host colonic epithelial cell promoter hypermethylation of tumor suppressor genes in colorectal cancer. *Microbiome* **8**, 108 (2020).
88. Sinha, R., Abnet, C. C., White, O., Knight, R. & Huttenhower, C. The microbiome quality control project: baseline study design and future directions. *Genome Biol.* **16**, 276 (2015).
89. Schloss, P. D. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *mBio* **9**, e00525-18 (2018).

90. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
91. Schloissnig, S. et al. Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).
92. Zhu, A., Sunagawa, S., Mende, D. R. & Bork, P. Inter-individual differences in the gene content of human gut bacterial species. *Genome Biol.* **16**, 82 (2015).
93. McDonnell, L. et al. Association between antibiotics and gut microbiome dysbiosis in children: systematic review and meta-analysis. *Gut Microbes* **13**, 1–18 (2021).
94. Mirzayi, C. et al. Reporting guidelines for human microbiome research: the STORMS checklist. *Nat. Med.* **27**, 1885–1892 (2021).
95. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
96. Wang, L. L. & Lo, K. Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Brief. Bioinform.* **22**, 781–799 (2021).
97. Cook, R. R. et al. Alterations to the gastrointestinal microbiome associated with methamphetamine use among young men who have sex with men. *Sci. Rep.* https://doi.org/10.1038/s41598-019-51142-8 (2019).
98. Tsementzi, D. et al. Comparison of vaginal microbiota in gynecologic cancer patients pre- and post-radiation therapy and healthy women. *Cancer Med.* https://doi.org/10.1002/cam4.3027 (2020).
99. Ren, Z. et al. Alterations in the human oral and gut microbiomes and lipidomics in COVID-19. *Gut* **70**, 1253–1265 (2021).
100. Gong, H. et al. Microbiota in the throat and risk factors for laryngeal carcinoma. *Appl. Environ. Microbiol.* https://doi.org/10.1128/AEM.02329-14 (2014).
101. Yang, C.-Y. et al. Oral microbiota community dynamics associated with oral squamous cell carcinoma staging. *Front. Microbiol.* https://doi.org/10.3389/fmicb.2018.00862 (2018).
102. Correa, J. D. et al. Oral microbial dysbiosis linked to worsened periodontal condition in rheumatoid arthritis patients. *Sci. Rep.* **9**, 8379 (2019).
103. Sato, N. et al. The relationship between cigarette smoking and the tongue microbiome in an East Asian population. *J. Oral Microbiol.* https://doi.org/10.1080/20002297.2020.1742527 (2020).
104. Oku, S. et al. Disrupted tongue microbiota and detection of nonindigenous bacteria on the day of allogeneic hematopoietic stem cell transplantation. *PLoS Pathog.* https://doi.org/10.1371/journal.ppat.1008348 (2020).
105. Balan, P. et al. Subgingival microbiota during healthy pregnancy and pregnancy gingivitis. *JDR Clin. Trans. Res.* https://doi.org/10.1177/2380084420948779 (2021).
106. Coit, P. et al. Sequencing of 16S rRNA reveals a distinct salivary microbiome signature in Behçet's disease. *Clin. Immunol.* https://doi.org/10.1016/j.clim.2016.06.002 (2016).
107. Hannigan, G. D., Duhaime, M. B., Ruffin IV, M. T., Koumpouras, C. C. & Schloss, P. D. Diagnostic potential and interactive dynamics of the colorectal cancer virome. *mBio* https://doi.org/10.1128/mbio.02248-18 (2017).
108. Gupta, A. et al. Association of *Flavonifractor plautii*, a flavonoid-degrading bacterium, with the gut microbiome of colorectal cancer patients in India. *mSystems* https://doi.org/10.1128/msystems.00438-19 (2019).
109. Vogtmann, E. et al. Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. *PLoS ONE* https://doi.org/10.1371/journal.pone.0155362 (2016).
110. Feng, Q. et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* https://doi.org/10.1038/ncomms7528 (2015).
111. Zeller, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* https://doi.org/10.15252/msb.20145645 (2014).
112. Yu, J. et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* https://doi.org/10.1136/gutjnl-2015-309800 (2015).
113. Yachida, S. et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* https://doi.org/10.1038/s41591-019-0458-7 (2019).

[1]Center for Computational Biomedicine, Harvard Medical School, Boston, MA, USA. [2]Institute for Implementation Science in Population Health, City University of New York School of Public Health, New York, NY, USA. [3]Department of Epidemiology and Biostatistics, City University of New York School of Public Health, New York, NY, USA. [4]Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology (IIT) Madras, Chennai, India. [5]Robert Bosch Centre for Data Science and Artificial Intelligence, Indian Institute of Technology (IIT) Madras, Chennai, India. [6]Centre for Integrative Biology and Systems mEdicine (IBSE), Indian Institute of Technology (IIT) Madras, Chennai, India. [7]WikiWorks, Boca Raton, FL, USA. [8]Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), Singapore, Republic of Singapore. [9]Infectious Disease and Microbiome Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [10]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [11]Harvard Chan Microbiome in Public Health Center, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [12]Channing Division of Network Medicine, Mass General Brigham, Harvard Medical School, Boston, MA, USA. [13]Leverhulme Centre for Demographic Science, University of Oxford, Oxford, UK. [14]Departments of Biomedical Informatics and Medicine, University of Colorado Anschutz School of Medicine, Denver, CO, USA. [15]Department CIBIO, University of Trento, Trento, Italy. [16]Istituto Europeo di Oncologia (IEO) IRCSS, Milan, Italy. ✉e-mail: levi.waldron@sph.cuny.edu

Methods

Definition of semantic concepts (study, experiment, signature and taxon)

Taxon. A taxon, or taxonomic unit, is a unit of any rank (that is, kingdom, phylum, class, order, family, genus, species and strain) designating a microbial organism or a group of microbial organisms.

Signature. A microbial signature or set refers here to a simple unordered list of microbial clades (taxa) sharing a common property or response to a study condition.

Experiment. BugSigDB defines experiments as semantic units within studies and records key characteristics about study subjects, lab analysis, statistical analysis and α-diversity. For study subjects, this includes host species, location, condition, body site, antibiotics exclusion and sample size in study and control sample groups. To define the two sample groups that are contrasted for differential abundance, BugSigDB records the diagnostic criteria applied to define the specific condition/phenotype represented in the study group. Recorded lab analysis fields include sequencing type (16S or MGX) and sequencing platform (such as Illumina or Roche454). For 16S rRNA sequencing, the 16S variable region is also recorded. For the statistical analysis, recorded fields include (1) the statistical test or computational tool applied for differential abundance testing, (2) whether multiple testing correction has been applied to adjust for an inflation of false-positive findings, (3) the significance threshold used to render taxa as differentially abundant, (4) confounding factors that have been accounted for by stratification or model adjustment and (5) factors on which study subjects have been matched on in a case–control study, if applicable.

Study. BugSigDB collects and standardizes microbial signatures from published 16S and MGX microbiome studies. Studies are categorized by study design, and each study is associated with a study identifier such as a PubMed ID and/or a DOI, depending on whether studies are indexed in PubMed.

Data entry, validation and access

BugSigDB is implemented as a semantic MediaWiki[30] web interface available at https://bugsigdb.org and supports data entry, semantic validation and web-based programmatic access to annotations for studies, experiments, signatures and individual taxa. The semantic Wiki (1) enforces data entry to follow the nomenclature of the NCBI Taxonomy Database[29], (2) enforces metadata annotation of signatures to follow established ontologies and controlled vocabulary for body site[28], disease condition[32] and type of evidence, (3) provides an application programming interface to access all signatures, potentially filtered on taxonomy and metadata attributes, and (4) allows commenting and error reporting on data elements and relationships. The data curation interface provides type-forward autocomplete to assist with valid data entry (including validation against the NCBI Taxonomy[29], Experimental Factor Ontology[32], UBERON Anatomy Ontology[28] and administrator-defined controlled vocabulary for other fields such as statistical test and sequencing methods) to facilitate organization, filtering and comparison of signatures. External contributions from the community, including signatures, annotations and comments, are supported similar to Wikipedia. The quality of contributions is controlled by tagging contributions as verified after review by a trusted editor, a method for flagging suspect entries and the option to exclude from analysis unreviewed contributions or based on elements of study quality, such as sample size, suspected contamination, paper retraction, batch effects, uncontrolled confounding or a combination of these factors. In addition to standard semantic MediaWiki quality control tags on study level, custom methods are available for flagging taxa according to prevalence in frequently investigated host body sites, inclusion in published contamination blacklists[114,115] and absence of known association with a host.

Signatures can be searched and browsed by study and experimental attributes and by individual taxa at https://bugsigdb.org. Bulk export of all signatures and associated metadata is available in plain text formats for use in any programming language and software (including .csv and the .gmt standard used by GeneSigDB[23] and MSigDB[21]) of the current database version or as weekly and semiannual snapshots. The companion bugsigdbr R/Bioconductor package (bioconductor.org/packages/bugsigdbr) provides advanced features, such as ontology-based filtering, limitation of taxonomic level, look-up of individual signature and taxon pages and conversion to application-centric formats. The exported files are compatible with most enrichment software and are included by default in our lab's EnrichmentBrowser R/Bioconductor package[55] to facilitate a large number of GSEA methods and visualizations. The BugSigDBStats R package continuously integrates with https://bugsigdb.org and provides weekly updated database statistics in an HTML report page (https://waldronlab.io/BugSigDBStats).

Signature similarity

Signature similarity was computed based on two different measures: (1) Jaccard index based on pairwise overlaps between signatures harmonized to genus level and (2) semantic similarity between signatures of mixed taxonomic levels. Pairwise calculation of Jaccard similarity for genus-level signatures was performed using the calcJaccardSimilarity function of the BugSigDBStats package. Genus-level signatures from BugSigDB were obtained using bugsigdbr's getSignatures function. Taxonomic clades given at a more specific taxonomic level (species or strain) were transformed by cutting the taxonomic tree at the genus level. The Jaccard index, also known as the Jaccard similarity coefficient, is defined as the size of the intersection divided by the size of the union of two input signatures A and B:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Note that, by design, $0 \le J(A,B) \le 1$.

Semantic similarity was computed based on Lin's measure of semantic similarity[116], as implemented in the ontologySimilarity package from the ontologyX R package series[117]. Measures of semantic similarities have been proposed for comparing concepts within a taxonomy[118] with numerous applications for biomedical ontologies[36]. Of note, semantic similarity has conceptual parallels with the computation of UniFrac distance[119]. Here, individual taxa can be considered nodes of the NCBI Taxonomy when represented as a directed acyclic graph. Computing semantic similarity between two taxa corresponds then to computing a topological similarity such as the shortest path linking the two taxon nodes. More specifically, computing Lin's measure of semantic similarity between two taxa corresponds to computing the information content of the lowest common ancestor (LCA) of the two taxa[116]. The more frequently a taxon occurs (that is, the higher its probability of being an ancestor of other taxa in the taxonomy), the lower its information content. If the LCA of two taxa corresponds to a taxon at a higher taxonomic level, these taxa are not very similar, and this is reflected in a low-information content of their LCA. Given pairwise semantic similarities between individual taxa, semantic similarity between two signatures (that is, two taxon sets) is then obtained using a best-match average combination approach[120] where each taxon of the first signature is paired with the most similar taxon of the second signature and vice versa.

Bug set enrichment analysis

Metagenomic datasets providing species-level relative abundance for fecal microbiomes of individuals with colorectal cancer and healthy individuals were obtained from curatedMetagenomicData 3.0 (ref. 54). Relative abundance proportions were multiplied by read depth and rounded to the nearest integer before obtaining integer read counts compatible with bulk RNA-sequencing (RNA-seq) tools for differential expression analysis and GSEA. For genus-level analysis, species-level counts were summed across branches using the splitByRanks function from the mia package. Given a recent assessment that reported good performance of bulk RNA-seq tools for microbiome data[121], differential abundance analysis was performed following the limma-trend approach[122]. Read counts were transformed to log (counts per million) using edgeR's cpm function with a prior count of 3 to dampen the variances of logarithms of low counts[123]. Genus- and species-level microbial signatures from BugSigDB were obtained using bugsigdbr's getSignatures function. To keep signatures meaningfully sized, taxonomic clades given at the species or strain level were transformed by cutting the taxonomic tree at the genus level. ORA and PADOG were performed as implemented in the EnrichmentBrowser package[55]. CBEA was performed as implemented in the CBEA package[18].

Taxon co-occurrence

Genus-level signatures from BugSigDB were obtained using the getSignatures function of the bugsigdbr package. Signatures were filtered by body site for fecal samples and stratified by direction of abundance change (increased/decreased). The top 20 most frequently occurring genera in the resulting signatures were reported. Prevalence of these genera was computed as percentage of healthy adult stool samples in which the genus was observed at non-zero relative abundance in metagenomic datasets from curatedMetagenomicData 3.0 (ref. 123). The correlation between prevalence in healthy samples and the proportion of signatures with increased abundance for the top 20 genera was assessed using a Pearson's correlation test as implemented in the cor.test function of the stats package. Taxon co-occurrence in signatures associated with disease was contrasted against Spearman rank correlation of the top 20 genera in healthy samples using the cor function of the stats package.

Signature pooling

Signatures were pooled for one body site at a time and within body sites for one condition at a time, as implemented in the getMetaSignatures function of the bugsigdbr package. Taxa within a pooled signature were weighted based on the aggregated sample size of the studies that reported this taxon as differentially abundant divided by the total sample size of studies contributing to the pooled signature. Resulting weighted metasignatures were clustered by semantic similarity, where the weights were incorporated into the best-match average combination approach[36], as implemented in the weightedBMA function of the BugSigDBStats package. Analysis was restricted to body sites and conditions studied by at least two studies in BugSigDB and containing at least five taxa in the resulting pooled signature. Robustness of the clustering was evaluated by comparing to rank-biased overlap[124].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

BugSigDB is available via a Semantic MediaWiki web interface at https://bugsigdb.org under open-source and open-data licenses described at https://bugsigdb.org/Project:About. Weekly and semiannual snapshots are provided in plain text file formats at https://github.com/waldronlab/BugSigDBExports for cross-language and cross-application compatibility; unprocessed snapshots are available as .csv files at https://bugsigdb.org/Help:Export. The companion bugsigdbr R/Bioconductor package provides advanced data manipulation, including ontology-aware and taxonomy-aware features (https://bioconductor.org/packages/bugsigdbr). The NCBI Taxonomy database is available at https://www.ncbi.nlm.nih.gov/taxonomy. The Experimental Factor Ontology is available at https://www.ebi.ac.uk/efo. The UBERON Anatomy Ontology is available at https://www.ebi.ac.uk/ols/ontologies/uberon.

Code availability

Source code and open issue tracking are provided at https://github.com/waldronlab/BugSigDB. Statistical analysis was performed using R[125] and Bioconductor[126] and is reproducible using the code provided on GitHub[127].

References

114. Park, S.-J. & Nakai, K. OpenContami: a web-based application for detecting microbial contaminants in next-generation sequencing data. *Bioinformatics* **37**, 3021–3022 (2021).
115. Salter, S. J. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).
116. Lin, D. et al. An information-theoretic definition of similarity. In *Proc. 5th International Conference on Machine Learning* (ed. Shavlik, J. W.) 296–304 (Morgan Kaufmann, 1998).
117. Greene, D., Richardson, S. & Turro, E. ontologyX: a suite of R packages for working with ontological data. *Bioinformatics* **33**, 1104–1106 (2017).
118. Resnik, P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.* **11**, 95–130 (1999).
119. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).
120. Pesquita, C. et al. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics* **9**, S4 (2008).
121. Calgaro, M., Romualdi, C., Waldron, L., Risso, D. & Vitulo, N. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biol.* **21**, 191 (2020).
122. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
123. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
124. Webber, W., Moffat, A. & Zobel, J. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.* **28**, 1–38 (2010).
125. Ihaka, R. & Gentleman, R. R: a language for data analysis and graphics. *J. Comput. Graph. Stat.* **5**, 299–314 (1996).
126. Huber, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).
127. Geistlinger, L. & Waldron, L. Analysis code for the BugSigDB manuscript. GitHub https://github.com/waldronlab/BugSigDBPaper (2023).

Acknowledgements

This investigation would not have been possible without the work of 58 student and intern curators who contributed data to BugSigDB: A. Singh, A. Amara, A. Shahid, A. Samantha, B. Dindi, B. Ojo, C. Grace, C. Nnadi, C. Grieve, C. Apiyo, C. Anderson, D. Birnbaum, D. Oluwarotimi, E. Jessica, E. Ewete, E. Mirti, F. Zohra, F. Cuevas, I. Merit, I. Kalu, I. Kahveci, J. Shevin, J. Shudy, J. Dey, K. W. Amusat, K. Joseph, K. Amoo, L. Park, L. Kasselman, L. Mellor, M. Dey, M. Hoyos, M. Thomatos, M. Martin, M. Bearkland, M. A. Parvin, M. Nguyen, N. Ulysse, P. M. Lutete, O. Blessing, P. Brianna, P. Han, S. Khan, S. El Safoury, S. Chunduri, Sophy, S. Suleiman, T. Islam, T. Sil,

U. Yokeeswaran, U. Ejite, U. Patel, V. Pineda, V. Goulbourne, W. Lam, Y. Javaid, Y. Wang and Z. Bailey. Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award numbers 5R01CA230551 (to L.W., N.S., C.H. and H.E.J.) and 5U24CA180996 (to L.W.). L.G. was supported by a research fellowship from the German Research Foundation (GE3023/1-1 and GE3023/1-2). J.B.D. acknowledges funding support from the European Research Council (ERC-2021-CoG-101002587) and The Leverhulme Trust (Large Centre Grant). P.S. acknowledges the Prime Minister's Research Fellowship (PMRF) from the Ministry of Education, Government of India.

Author contributions

L.G. and L.W. designed the study with input from H.E.J., S.D., N.S. and C.H. F.Z., R.A., S.E. and C.G. curated and reviewed data with input and contributions from L.G., C.M. and L.W. C.M., F.Z., R.A., C.G. and L.W. supervised student interns curating the data. I.H. implemented the Semantic MediaWiki with input from L.G., C.M., F.Z., C.H. and L.W. L.G. implemented the bugsigdbr R/Bioconductor package with contributions from J.W. and input from C.M., S.D.G.-T. and L.W. L.G. implemented the BugSigDBStats R package with contributions from P.S., S.D.G.-T. and L.W. and input from C.M., A.R. and K.R. L.G. and J.W. implemented the BugSigDBExports repository with contributions and input from S.D. and L.W. J.W. released data on Zenodo with input from L.G. and L.W. L.G. developed and implemented methodology and performed data analysis and data visualization with input and contributions from R.G., E.F., S.D., N.S., C.H. and L.W. L.G. and L.W. wrote the manuscript with input from C.M., V.C., J.B.D., H.E.J., N.S. and C.H. All authors reviewed and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-023-01872-y.

Correspondence and requests for materials should be addressed to Levi Waldron.

Peer review information *Nature Biotechnology* thanks Anne Hoen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

nature portfolio

Corresponding author(s): Levi Waldron

Last updated by author(s): Jun 14, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Microbial signatures were collected through the Semantic MediaWiki web interface at https://bugsigdb.org. For reproducibility, all analyses presented in the manuscript have been carried out based on the BugSigDB v1.0.2 release (Jan 25, 2022) available from Zenodo under https://doi.org/10.5281/zenodo.5904281. BugSigDB signatures from Zenodo were obtained using the bugsigdbr R/Bioconductor package (version 1.4.0). Metagenomic datasets providing species-level relative abundance for fecal microbiomes of colorectal cancer patients and healthy controls were obtained through the curatedMetagenomicData R/Bioconductor package (version 3.6.0).

Data analysis

Statistical analysis was carried out using R-4.2.1 and Bioconductor-3.16. All analysis code is provided on GitHub (https://github.com/waldronlab/BugSigDBPaper). Jaccard similarity of BugSigDB signatures was calculated using the calcJaccardSimilarity function of the BugSigDBStats package (version 1.0.0). Semantic similarity was computed based on Lin’s measure of semantic similarity as implemented in the ontologySimilarity R package (version 2.5). For genus-level bug set enrichment analysis of colorectal cancer signatures, species-level counts were summed across branches using the splitByRanks function from the mia R/Bioconductor package (version 1.6.0). Differential abundance analysis was carried out following the limma-trend approach as implemented in the limma R/Bioconductor package (version 3.54.0). Read counts were transformed to log counts-per-million (CPMs) using the cpm function of the edgeR R/Bioconductor package (version 3.40.0). ORA and PADOG were carried out as implemented in the EnrichmentBrowser R/Bioconductor package (version 2.28.0). CBEA was carried out as implemented in the CBEA R/Bioconductor package (version 1.2.0). For the analysis of pooled signatures, signatures were pooled for one body site at a time, and within body sites for one condition at a time, as implemented in the getMetaSignatures function of the bugsigdbr package (version 1.4.0). Semantic similarity of weighted meta-signatures was calculated by incorporating the weights into the best-match average combination approach as implemented in the weightedBMA function of the BugSigDBStats R package (version 1.0.0).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

BugSigDB is available via a Semantic MediaWiki web interface at https://bugsigdb.org, under open-source and open-data licenses described at https://bugsigdb.org/Project:About. Weekly and semi-annual snapshots are provided in plain text file formats at https://github.com/waldronlab/BugSigDBExports for cross-language and cross-application compatibility; unprocessed snapshots are available as csv files at https://bugsigdb.org/Help:Export. The companion bugsigdbr R/Bioconductor package provides advanced data manipulation, including ontology-aware and taxonomy-aware features (https://bioconductor.org/packages/bugsigdbr). The NCBI Taxonomy database is available at https://www.ncbi.nlm.nih.gov/taxonomy. The Experimental Factor Ontology is available at https://www.ebi.ac.uk/efo. The UBERON Anatomy Ontology is available at https://www.ebi.ac.uk/ols/ontologies/uberon.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Not applicable
Population characteristics	Not applicable
Recruitment	Not applicable
Ethics oversight	Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size of both contrasted sample groups is recorded for all studies curated on bugsigdb.org. The median total sample size across all 628 studies in the BugSigDB 1.0.2 release is n=25. For the bug set enrichment analysis of colorectal cancer signatures, differential abundance was calculated between n=662 colorectal cancer samples and n=653 control samples from 10 datasets. For the taxon co-occurence analysis, prevalence of frequently reported genera was calculated in n=9,623 stool samples from healthy adult controls of 68 different studies.
Data exclusions	For the bug set enrichment analysis of colorectal cancer signatures, only signatures containing 5 or more genera associated with any condition were included. Analysis of pooled signatures was restricted to body sites and conditions studied by at least two studies in BugSigDB and containing at least 5 taxa in the resulting pooled signature. Antibiotics exclusion criteria, if provided by the authors, are also recorded for studies curated on bugsigdb.org (median exclusion time across studies = 60 days).
Replication	To assess replication by independent studies of the same condition, we compared semantic similarity between signatures reported for a single condition to the similarity of randomly sampled signatures in repeated simulation, and ranked conditions based on the resulting empirical p-value. Antibiotics treatment and HIV infection were the experimental conditions with the highest level of consistency in signatures reported by independent studies. Among the enriched signatures of the bug set enrichment analysis of colorectal cancer signatures, the signatures from Wu et al., 2013 and Allali et al., 2018, are notable because neither study was included in the 10 datasets from which the meta-analysis signatures from Thomas et al., 2019 and Wirbel et al., 2019 were computed, therefore providing independent replication of these signatures.
Randomization	BugSigDB signatures are available primarily from observational study designs: case-control (281 studies, 44.7%) and cross-sectional studies (171 studies, 27.2%) were most prevalent, while prospective cohort studies (66, 10.5%), time-series/longitudinal studies (51, 8.1%), laboratory studies (29, 4.6%), randomized controlled trials (28, 4.4%), and meta-analyses (6, 0.9%) are also present. For analyses presented in this manuscript randomization does not apply, as this is not a randomized control trial.
Blinding	BugSigDB signatures are available primarily from observational study designs: case-control (281 studies, 44.7%) and cross-sectional studies (171 studies, 27.2%) were most prevalent, while prospective cohort studies (66, 10.5%), time-series/longitudinal studies (51, 8.1%), laboratory

studies (29, 4.6%), randomized controlled trials (28, 4.4%), and meta-analyses (6, 0.9%) are also present. For analyses presented in this manuscript blinding does not apply, as this is not a randomized control trial.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging