Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Public Metagenome Prioritization of Reproducible Gut Microbial Functional Modules

Geng-Hao Bai
National Taiwan University Hospital, National Taiwan University
Chung-Yung Chen
Chung Yuan Christian University
Ben-Chang Shia
Fu Jen Catholic University
Yi-Hsiang Hsu
Beth Israel Deaconess Medical Center, Harvard Medical School
Shih-Yen Chen
18159@s.tmu.edu.tw

Shuang Ho Hospital, Taipei Medical University

Research Article

Keywords: public metagenomics, gut microbiome, functional modules, short-chain fatty acids, target prioritization, bioinformatics

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11526095/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background
Public shotgun metagenome repositories offer a valuable resource for identifying gut microbial functional targets, but their reuse is challenging because of heterogeneous cohort structures, incomplete metadata, high-dimensional taxonomic features, and incomplete functional annotation. This study aimed to develop a computational framework to prioritize reproducible microbial functional modules across multiple public metagenomic cohorts.
Results
Processed stool metagenome data from 11 public cohorts comprising 2,255 samples were analyzed. Functional modules provided a more compact and less heterogeneous feature space than taxonomic features. Across cohorts, three short-chain fatty acid (SCFA)-related modules—CENTFERM-PWY, PWY-6590, and P461-PWY—showed consistent multi-cohort evidence and met the predefined meta-FDR and direction-consistency criteria. The Functional Target Priority Score ranked CENTFERM-PWY and PWY-6590 as the top candidates, followed by tryptophan-related modules, while GABA-related pathways remained exploratory. In an independent held-out cohort, 28 of 31 tested modules showed concordant directions, corresponding to a 90.3% direction agreement fraction.
Conclusions
This multi-cohort framework identifies SCFA-related microbial fermentation modules as the most reproducible computational candidates for gut microbial functional targeting. The findings provide a data-driven prioritization strategy for public metagenomic resources while highlighting the need for experimental and clinical validation before biological or interventional conclusions can be made.

Highlights

A public-metagenome workflow prioritized gut microbial functional modules.

Eleven processed stool metagenome cohorts comprising 2,255 samples were analyzed.

SCFA fermentation modules were the dominant reproducible candidate axis.

FTPS ranked CENTFERM-PWY and PWY-6590 as top candidate modules.

Held-out public cohort testing showed 28/31 direction agreement.

1. Introduction

Shotgun metagenomic sequencing has transformed human microbiome research by enabling simultaneous analysis of microbial community composition and functional potential. Large-scale public projects have demonstrated that human-associated microbial communities vary extensively across individuals, body sites, and disease contexts, while also encoding measurable functional capacities[1, 2]. These resources create an opportunity to move from isolated association studies toward reproducible

cross-cohort computational prioritization of microbial functions. A major challenge in microbiome target discovery is that taxonomic feature spaces are highly dimensional, cohort-dependent, and often difficult to interpret mechanistically. Different organisms may encode overlapping metabolic capacities, and the same biological process may be distributed across multiple taxa. Functional modules and pathway profiles can therefore provide a more interpretable intermediate layer between microbial composition and host-relevant biological mechanisms[3–6]. However, functional profiles derived from processed public metagenomes still require conservative interpretation because pathway abundance does not directly establish metabolite concentration, taxon-specific carrier attribution, or intervention responsiveness.

Microbial metabolic functions are central to host-microbiome interactions. Short-chain fatty acids (SCFAs), including acetate, propionate, and butyrate, are major products of gut microbial fermentation and have been linked to intestinal, immune, and metabolic physiology[7, 8]. Tryptophan metabolism represents another host-microbiota interface, with microbial pathways connected to indole derivatives, mucosal signaling, and inflammatory biology[9]. GABA-related microbial metabolism has also attracted attention, although its translational interpretation remains more exploratory and context-dependent[10]. These axes provide plausible biological categories for computational target nomination, but public metagenome data alone cannot validate metabolite production or therapeutic action. Public cohort reuse also introduces important methodological limitations. Metadata completeness varies across studies, especially for age, BMI, and sex. Phenotype definitions are not always harmonized across original cohorts. Processed profiles may lack paired metabolomics, intervention-response metadata, or stratified pathway outputs required for taxon-function carrier attribution[3, 4, 11]. Therefore, a computational target-prioritization framework should explicitly separate target nomination from downstream biological validation.

Cross-study synthesis of public microbiome data is not new. Duvallet et al. distinguished disease-specific from shared microbial responses across 28 16S rRNA datasets[12]. Armour et al. extended this approach to shotgun metagenomes and reported shared functional signatures across eight studies[13]. Neither study, however, ranked candidate functions for experimental follow-up, assessed ranking stability, or tested direction preservation in a held-out cohort. To address this prioritization gap, we developed a first-run public-metagenome framework for prioritizing gut microbial functional modules. We assembled 11 eligible processed public stool metagenome cohorts, evaluated metadata completeness, applied cohort-specific covariate adjustment, compared functional modules with taxa features, generated an exploratory keyword-derived target map, synthesized multi-cohort evidence, constructed a Functional Target Priority Score, evaluated weight sensitivity and leave-one-cohort-out robustness, and tested direction preservation in a held-out public cohort. Across these evidence layers, SCFA fermentation modules emerged as the dominant candidate axis, while tryptophan modules formed a secondary tier and GABA, bile acid, and phytochemical axes remained exploratory or insufficiently supported.

2. Materials and Methods

2.1 Study design and public cohort eligibility and sample inclusion

This study was designed as a first-run computational target-prioritization analysis using processed public stool metagenome profiles. The objective was to nominate reproducible microbial functional modules for future validation, not to establish clinical efficacy, metabolite production, or intervention response. The primary analytical contrast was reference versus non-reference status within each eligible cohort. Candidate public studies or cohorts were reviewed for availability of processed stool metagenome profiles, availability of both taxonomic and pathway-level profiles, and usable reference versus non-reference counts for the primary binary analysis. Twenty-three candidate studies or cohorts were reviewed. Eleven cohorts were retained for first-run analysis: HMP_2019_ibdmdb, QinJ_2012, JieZ_2017, YachidaS_2019, HallAB_2017, LiJ_2017, RubelMA_2020, KarlssonFH_2013, NagySzakalD_2017, IjazUZ_2017, and LoombaR_2017. These cohorts contributed 2,255 included stool samples.

2.2 Sample-level metadata, sequencing depth, metadata completeness, and cohort-specific covariate adjustment

Sample-level metadata included reference status, body site, country, sequencing_depth, age, BMI, and sex when available. Among the included samples, 846 were reference samples, 1,393 were non-reference samples, and 16 had unknown reference status. All included samples were stool samples. Sequencing depth was available for 2,255 of 2,255 included samples. Sequencing depth was summarized using five-number and mean statistics: minimum 48,368 reads, Q1 22,569,507 reads, median 37,991,682 reads, Q3 49,385,238 reads, maximum 290,630,554 reads, and mean 38,087,381 reads. Cohort-level completeness was calculated for age, BMI, body site, country, reference_status, sequencing_depth, and sex. Completeness was defined as the fraction of samples in a cohort with non-missing values for a given metadata field. Average completeness across cohorts was 0.718 for age, 0.592 for BMI, 1.000 for body site, 1.000 for country, 1.000 for reference_status, 1.000 for sequencing_depth, and 0.641 for sex. Covariates were selected separately for each cohort based on within-cohort availability. Full adjustment using age, sex, BMI, and sequencing_depth was available for HMP_2019_ibdmdb, JieZ_2017, QinJ_2012, RubelMA_2020, and YachidaS_2019. HallAB_2017, LiJ_2017, and LoombaR_2017 used sequencing_depth only. IjazUZ_2017 used age plus sequencing_depth. KarlssonFH_2013 used age plus BMI plus sequencing_depth. NagySzakalD_2017 used sex plus BMI plus sequencing_depth. The same cohort-specific covariate set was applied to functional module and taxa models.

2.3 Feature layers, cohort-specific modeling, multi-cohort evidence synthesis, exploratory target mapping, and Functional Target Priority Score

Two feature layers were analyzed: functional modules and taxa features. For each cohort, 31 module models and 292 taxa models were evaluated. In the feature-space comparison, the input feature space contained 37 functional modules and 1,585,844 taxa features; after filtering, 29 functional modules and 292 taxa features remained. Within each cohort, feature-level models were fitted using reference versus non-reference status as the primary phenotype and the cohort-specific covariate set. Cohort-specific results were synthesized into multi-cohort evidence summaries. Evidence metrics included meta_FDR, direction consistency, and I2 heterogeneity. Direction consistency represented the fraction of cohorts with directionally concordant effects. High-confidence candidate evidence was defined as meta_FDR < 0.05 and direction consistency > = 0.75. Functional pathway features were mapped to target axes using keyword-derived exploratory mapping. Target axes included SCFA, tryptophan, GABA, bile acid, and phytochemical categories. Exact curated mappings were unavailable. Keyword-mapped and ambiguous-mapped features were retained as exploratory annotations. All mapped features were flagged as requiring manual review before final biological claims. The Functional Target Priority Score integrated available prioritization components: reproducibility_score, effect_score, prevalence_score, feasibility_score, and evidence_score. Specificity_score, redundancy_score, and safety_penalty were unavailable and were not estimated. A dynamic denominator was used when components were unavailable.

2.4 Weight-sensitivity, leave-one-cohort-out robustness, held-out public cohort validation, and negative-control diagnostics

Weight sensitivity was assessed across five scoring schemes: Primary, Equal-weight, Reproducibility-heavy, Effect-heavy, and Prevalence-heavy. Rank stability was calculated across schemes. Leave-one-cohort-out robustness analysis was performed by removing one cohort at a time and recomputing module-level prioritization summaries. Top-k overlap was calculated as the Jaccard similarity of top-k module sets across leave-one-cohort-out runs. Held-out validation used HMP_2019_ibdmdb as the held-out public cohort. Direction agreement was evaluated for 31 tested target modules. Direction agreement was calculated as the fraction of modules with concordant discovery and validation directions. Nominal validation p values were summarized for the held-out cohort tests. This analysis was interpreted as held-out public cohort direction preservation, not independent clinical validation. Negative-control diagnostics included a random-label diagnostic and a shuffled-axis diagnostic. The random-label diagnostic used a single seeded shuffle and included 31 control features. Observed median meta_p and random-label control median meta_p were compared. The shuffled-axis diagnostic was completed as a structural negative control, but biological interpretation was not assigned. Unrelated pathway controls were not available, and a full permutation test was not performed.

2.5 Taxon-function carrier attribution and reproducibility package

Taxon-function carrier attribution was not available in the first-run analysis because processed profiles did not include stratified pathway carrier attribution. Future analyses will require HUMAnN stratified profiling from raw reads or access to stratified pathway tables. A reproducibility package manifest enumerated generated result files and file sizes. The package contained 71 files across five result groups: results/tables, results/figure_ready, results/diagnostic_plots, results/manifests, and results/run_summary. The total package size was 145,885,418 bytes, corresponding to 139.13 MB. No external repository DOI, URL, archive link, license, source-code directory, or checksum claim was included in the manifest.

3. Results

3.1 Public cohort selection and metadata characteristics

We first constructed a first-run public metagenome analysis resource from processed stool metagenome cohorts with available taxonomic and functional pathway profiles. Among 23 candidate studies or cohorts reviewed, 11 cohorts satisfied the eligibility requirements for the primary binary analysis. These cohorts contributed 2,255 included stool samples and formed the complete analysis set used for downstream module and taxa modeling. (Fig. 1A; Supplementary Fig. S1A and S1B in Additional file 1). The retained cohorts showed substantial variation in sample contribution. Four cohorts reached the first-run cap of 300 included samples, whereas smaller cohorts contributed between 86 and 259 samples. This design provided multi-cohort coverage while preventing a small number of large studies from dominating the first-run analysis. The final sample set included 846 reference samples, 1,393 non-reference samples, and 16 samples with unknown reference status. All included samples were stool samples, and sequencing depth was available for all 2,255 samples. (Fig. 1B and Fig. 1C; Supplementary Fig. S1C, Fig. S1D, Fig. S2A and Fig. S2B in Additional file 1). This cohort-selection step defined an important boundary for interpretation. The analysis was not intended to exhaustively represent all available public metagenomes. Instead, it produced a harmonized, eligibility-defined resource suitable for reproducible cross-cohort target prioritization. (Fig. 1D; Supplementary Fig. S2D in Additional file 1).

We next evaluated sample-level metadata completeness to determine whether uniform covariate adjustment was feasible across cohorts. Body site, country, reference_status, and sequencing_depth were complete across all 11 cohorts, each with average completeness of 1.000. In contrast, age, BMI, and sex were incompletely captured and varied markedly by cohort. Average completeness was 0.718 for age, 0.592 for BMI, and 0.641 for sex. (Supplementary Fig. S3A-S3C in Additional file 1). This uneven metadata structures directly affected model design. Cohorts with sufficient age, sex, BMI, and sequencing_depth information supported full adjustment, whereas cohorts lacking demographic or BMI information required reduced adjustment models. Full adjustment was available for five cohorts: HMP_2019_ibdmdb, JieZ_2017, QinJ_2012, RubelMA_2020, and YachidaS_2019. Three cohorts, HallAB_2017, LiJ_2017, and LoombaR_2017, were adjusted only for sequencing_depth. IjazUZ_2017 used age plus sequencing_depth; KarlssonFH_2013 used age, BMI, and sequencing_depth; and NagySzakalD_2017 used sex, BMI, and sequencing_depth. (Supplementary Fig. S4A-S4D in Additional

file 1). Thus, metadata completeness was not merely a descriptive quality-control result. It determined the statistical adjustment strategy and prevented inappropriate use of a single covariate formula across all cohorts. The same cohort-specific covariate set was applied to module and taxa models, ensuring that comparisons between feature layers were not driven by different adjustment rules. (Supplementary Fig. S3D and S4D in Additional file 1).

3.2 Feature-space comparison and exploratory target mapping

We compared functional modules and taxa features to determine which feature layer was more suitable for first-run target prioritization. The input taxa feature space was substantially larger than the module feature space, with 1,585,844 input taxa features compared with 37 input functional modules. After filtering, 292 taxa features and 29 functional modules remained in the feature-space comparison. (Fig. 2A). Functional modules showed higher sample-level prevalence than taxa. Median prevalence was 0.847 for functional modules and 0.315 for taxa, while median cohort detection was 11.0 for both feature layers. This indicates that retained modules and taxa were broadly detected across cohorts, but functional modules were present in a larger fraction of samples. Functional modules also showed lower median heterogeneity than taxa, with median I^2 of 59.3 compared with 70.6. Rank stability was slightly higher for functional modules, with median rank stability of 0.967 compared with 0.955 for taxa. (Fig. 2B and Fig. 2C). Direction consistency was not uniformly higher for modules. Median direction consistency was 0.636 for functional modules and 0.727 for taxa. This result prevents a simple claim that modules outperform taxa across all statistical dimensions. Instead, the evidence supports a more specific conclusion: functional modules provide a compact, prevalent, and comparatively lower-heterogeneity feature layer for computational target nomination, whereas taxa remain useful as a complementary benchmark but are less tractable as a primary prioritization space. (Fig. 2D).

We then mapped functional pathway features to candidate biological target axes using an exploratory keyword-derived strategy. This mapping was used to organize candidate functional signals into interpretable biological categories, not to define a final curated ontology. Exact curated mappings were not available, and all mapped features were flagged as requiring manual review. (Fig. 3A; Supplementary Fig. S5A in Additional file 1). The mapping exercise revealed that most functional pathway features remained outside the current target map. Of 562 total functional pathway features, 510 were not mapped, corresponding to 90.7%. Only 52 features were mapped or partially mapped, corresponding to 9.3%. This sparse mapping structure is important because it shows that the candidate target framework covered a limited portion of the full processed pathway feature universe. (Supplementary Fig. S5A-S5C in Additional file 1). Within the mapped or partially mapped subset, SCFA-related features dominated. The mapped set included 34 SCFA keyword hits and 2 SCFA ambiguous hits. Tryptophan mapping contributed 9 keyword hits, while GABA mapping contributed 4 keyword hits and 2 ambiguous hits. Bile acid mapping was minimal, with only 1 keyword hit. No phytochemical module was recovered in the current processed-profile run. (Fig. 3A and 3C; Supplementary Fig. S5B in Additional file 1). Therefore, the first-run target map should be interpreted as a sparse, SCFA-dominant exploratory map. It supports

prioritization of SCFA-related modules for downstream analysis, while indicating that tryptophan and GABA coverage is more limited and that bile acid and phytochemical axes are insufficiently represented in the current processed-profile analysis. (Fig. 3D; Supplementary Fig. S5D in Additional file 1).

We next examined the top detectable modules to determine which candidate functional axes were broadly measurable across the included public cohorts. The top 10 detectable modules were composed exclusively of SCFA- and tryptophan-associated pathways. Sample prevalence ranged from 0.923 to 0.996, and cohort prevalence was 1.000 for all top 10 modules. (Fig. 3B and 3D). The most detectable modules included SCFA-associated PWY-7111, ANAEROFRUCAT-PWY, PWY-5100, PWY-5676, P108-PWY, P461-PWY, CENTFERM-PWY, and PWY-6590, together with tryptophan-associated COMPLETE-ARO-PWY and TRPSYN-PWY. These modules represented fermentation, acetate/lactate, propanoate, butanoate, and aromatic amino acid biosynthesis pathways. (Fig. 3B). This detectability result provides an important bridge between mapping and prioritization. SCFA and tryptophan modules were not only present in the keyword-derived target map; they were also broadly detectable across cohorts. In contrast, bile acid mapping was limited and not represented among the top-detectability set, while phytochemical modules were not recovered. Thus, subsequent prioritization focused on functional axes that were both interpretable and measurable in the processed public cohort data. (Fig. 3D).

3.3 Multi-cohort evidence and functional target prioritization

We synthesized multi-cohort evidence using meta_FDR, direction consistency, and I^2 heterogeneity. SCFA was the dominant evidence axis. Three SCFA modules satisfied the combined high-confidence candidate criterion of meta_FDR < 0.05 and direction consistency > = 0.75: CENTFERM-PWY, PWY-6590, and P461-PWY. (Fig. 4A and 4C). CENTFERM-PWY showed meta_FDR = 0.000530, direction consistency = 0.909, and I^2 = 34.2. PWY-6590 showed the same meta_FDR of 0.000530 and direction consistency of 0.909, with I^2 = 35.2. These two modules therefore combined strong meta-analysis evidence, high directional agreement, and moderate heterogeneity. P461-PWY also satisfied the combined criterion, with meta_FDR = 0.0467, direction consistency = 0.818, and I^2 = 66.5. (Fig. 4A). Other SCFA modules showed partial evidence but did not satisfy the combined threshold. PWY4LZ-257 had meta_FDR = 0.00569 but direction consistency = 0.727 and was therefore classified as FDR-only rather than high-confidence under the predefined combined rule. FERMENTATION-PWY similarly had meta_FDR = 0.0467 and direction consistency = 0.727. PWY-7111 and ANAEROFRUCAT-PWY had direction consistency = 0.818 but meta_FDR values above 0.05, placing them in the exploratory tier. (Fig. 4A and 4C). This tiered interpretation prevents over-ranking modules based on only one favorable metric. Modules with strong FDR but insufficient direction consistency were separated from modules satisfying both statistical significance and cross-cohort directional preservation. Under this conservative rule, SCFA fermentation modules remained the clearest high-confidence candidate group. (Fig. 4C and 4D).

Tryptophan-related modules showed evidence consistent with a secondary candidate tier rather than a dominant high-confidence axis. PWY-6629 showed meta_FDR = 0.0373, direction consistency = 0.727,

and I^2 = 27.4. Although this module met the FDR criterion, its direction consistency did not reach the combined high-confidence threshold. COMPLETE-ARO-PWY showed high direction consistency of 0.909 and broad detectability, but its meta_FDR was 0.0616. TRPSYN-PWY showed meta_FDR = 0.139 and direction consistency = 0.727. (Fig. 4B and 4C). GABA-related modules were weaker under the same evidence framework. PWY-5088 showed meta_FDR = 0.138 and direction consistency = 0.727, while ARGDEG-PWY, GLUDEG-I-PWY, P162-PWY, PWY-4321, and PWY-5022 had higher FDR values and lower or moderate direction consistency. These results support classification of GABA as exploratory in the current processed-profile analysis. (Fig. 4B and 4C). The comparison across axes therefore supports a graded interpretation: SCFA was the dominant evidence axis, tryptophan was secondary or nominal, and GABA remained exploratory. Bile acid and phytochemical axes were not advanced to evidence-based prioritization because mapping and recovered-module support were insufficient. (Fig. 4D).

To move beyond single-metric ranking, we applied the Functional Target Priority Score. FTPS integrated five available components: reproducibility_score, effect_score, prevalence_score, feasibility_score, and evidence_score. Three potentially relevant components, specificity_score, redundancy_score, and safety_penalty, were not available and were not estimated. A dynamic denominator was used when components were unavailable. (Fig. 5A). The top-ranked FTPS modules were SCFA-dominant. CENTFERM-PWY and PWY-6590 ranked first and second, respectively, each with FTPS = 0.827. Both modules combined high reproducibility, high effect score, strong prevalence, feasible target status, and available evidence support. PWY4LZ-257 ranked third with FTPS = 0.723. COMPLETE-ARO-PWY, a tryptophan-associated module, ranked fourth with FTPS = 0.696. P461-PWY ranked fifth with FTPS = 0.632. (Fig. 5B and 5C). The ranking also shows that high prevalence alone did not determine FTPS priority. COMPLETE-ARO-PWY had sample prevalence of 0.996 but ranked below CENTFERM-PWY and PWY-6590, each of which had sample prevalence of 0.923. Similarly, PWY-7111 and ANAEROFRUCAT-PWY had very high sample prevalence of 0.996 and 0.995, respectively, but ranked seventh and eighth. This indicates that FTPS prioritization was driven by combined evidence rather than detectability alone. (Fig. 5B and 5C). Thus, FTPS provided an integrated prioritization layer that reinforced the multi-cohort evidence results while distinguishing broadly detectable modules from high-priority candidate modules. (Fig. 5A-5C).

3.4 Robustness analyses and held-out cohort validation

We next evaluated whether the FTPS ranking was dependent on a single weighting scheme. Five weighting schemes were compared: Primary, Equal-weight, Reproducibility-heavy, Effect-heavy, and Prevalence-heavy. CENTFERM-PWY ranked first across all five schemes, and PWY-6590 ranked second across all five schemes. P461-PWY also maintained rank 5 across all five schemes. Each of these modules had rank stability of 1.000. (Fig. 5D). PWY4LZ-257 and COMPLETE-ARO-PWY showed minor rank switching between the third and fourth positions, each with stability of 0.967. Other modules showed moderate but limited variation across weighting schemes. Across 31 evaluated modules, median rank stability was 0.833, with a minimum of 0.600 and a maximum of 1.000. (Fig. 5D). These results show that the top SCFA candidates were not artifacts of a single weighting assumption. Instead,

their ranking was preserved across weighting strategies emphasizing reproducibility, effect, or prevalence. This strengthens the interpretation that CENTFERM-PWY and PWY-6590 represent robust computational candidates within the current evidence structure. (Fig. 5D).

To evaluate whether prioritization was dominated by any individual cohort, we examined leave-one-cohort-out robustness. The LOCO analysis generated 3,553 LOCO rows and assessed rank stability and top-k overlap across cohort-removal runs. (Supplementary Fig. S6A-S6D in Additional file 1). Top-ranked SCFA modules remained stable in the LOCO framework. In the HMP_2019_ibdmdb leave-out test, CENTFERM-PWY and PWY-6590 preserved ranks 1 and 2, respectively. Top-k overlap remained high across LOCO runs, with Jaccard similarity of 0.9 for top 10 module sets, 0.9 for top 20 module sets, and 1.0 for top 50 module sets. (Supplementary Fig. S6B-S6D in Additional file 1) These results indicate that the leading candidates were not driven solely by a single cohort. Instead, the top-ranked SCFA modules were preserved under cohort removal, supporting their robustness as first-run computational targets. (Supplementary Fig. S6D in Additional file 1).

We evaluated direction preservation using HMP_2019_ibdmdb as a held-out public cohort. A total of 31 target modules were tested. Direction agreement was observed for 28 of 31 modules, yielding a direction agreement fraction of 0.903. Nominal validation $p < 0.05$ was observed for 10 modules. (Fig. 6A and 6B). Among the top examples, CENTFERM-PWY showed negative direction in both discovery and validation, with validation $p = 0.0731$. PWY-6590 also showed negative direction in both discovery and validation, with validation $p = 0.0712$. PWY4LZ-257 showed positive direction in both discovery and validation, with validation $p = 0.00809$. Among the top 12 displayed modules, PROPFERM-PWY was discordant, showing negative discovery direction and positive validation direction, with validation $p = 0.4771$. (Fig. 6A and 6B). This result supports direction preservation for most tested modules in a held-out public cohort. However, it should be interpreted strictly as public-cohort computational validation. It does not establish independent clinical validation, metabolomics validation, or intervention-response validation. (Fig. 6C and 6D; Supplementary Fig. S7A-S7D in Additional file 1).

3.5 Negative controls and boundaries of biological interpretation

We assessed first-run negative-control diagnostics to evaluate whether the observed results separated from simple control structures. In the random-label diagnostic, observed median meta_p was 0.0815, while the random-label control median meta_p was 0.2686 across 31 control features. This pattern is consistent with partial diagnostic separation between observed and shuffled-label analyses. (Supplementary Fig. S8A-S8D in Additional file 1). However, this control should be interpreted conservatively. The random-label control was a single seeded shuffle diagnostic, not a full permutation test. Therefore, it cannot define a full empirical null distribution. The shuffled-axis diagnostic was completed, but biological interpretation was not assigned because its purpose was structural negative control only. Unrelated pathway controls were not available because unmapped pathway control scores were not computed in first-run mode. (Supplementary Fig. S8B-S8D in Additional file 1). Thus, the

negative-control layer provides partial support for the computational workflow but remains incomplete. Stronger future versions should include repeated permutation testing, unrelated pathway controls, and prevalence-matched null feature sets. (Supplementary Fig. S8D in Additional file 1).

Several validation layers were not available in the current processed-profile analysis. Metabolomics validation was not available because the processed curatedMetagenomicData profiles used here did not include paired metabolomics tables. Intervention validation was not available because explicit intervention responder metadata was not harmonized from selected processed public cohorts. Taxa held-out validation was not computed in the first-run module-target workflow. (Fig. 6C and 6D; Supplementary Fig. S7D in Additional file 1). Taxon-function carrier attribution was also unavailable. The processed profiles used in this analysis aggregate pathway abundance and do not include stratified pathway carrier attribution. Therefore, the present workflow can prioritize functional modules but cannot determine which taxa contributed those module-level signals. Future work will require HUMAnN stratified profiling from raw reads or access to stratified pathway tables. (Supplementary Fig. S9 in Additional file 1). These unavailable layers are not peripheral details; they define the boundary between computational target nomination and biological target validation. The present results support prioritization of SCFA-dominant modules for follow-up, but they do not establish metabolite production, causal mechanism, taxon-resolved carrier identity, or intervention responsiveness. (Fig. 6D; Supplementary Fig. S8D and S9 in Additional file 1).

3.6 Reproducibility package

Finally, we summarized the reproducibility package manifest. The package contained 71 generated files totaling 145,885,418 bytes, corresponding to 139.13 MB. The files were organized into five top-level result groups: results/tables, results/figure_ready, results/diagnostic_plots, results/manifests, and results/run_summary. (Supplementary Fig. S10A-S10C in Additional file 1). The largest directory was results/tables, with 22 files totaling 135.55 MB. The results/figure_ready directory contained 35 files totaling 2.54 MB. The results/diagnostic_plots directory contained 8 files totaling 0.84 MB. The results/manifests directory contained 2 files totaling 0.19 MB, and results/run_summary contained 4 files totaling 0.002 MB. (Supplementary Fig. S10A and S10C in Additional file 1). The manifest documents generated files and file sizes but does not include an external repository DOI, URL, archive link, license, source-code directory, or checksum claims. Therefore, the current manifest supports internal reproducibility tracking, while formal external reproducibility will require repository deposition before submission. (Supplementary Fig. S10D in Additional file 1).

4. Discussion

This study presents a public-metagenome computational framework for microbiome functional target prioritization. The primary conclusion is that SCFA fermentation modules emerge as the most reproducible and highest-priority candidate axis across the available first-run evidence layers. This conclusion is supported by feature-space comparison, exploratory target mapping, multi-cohort evidence synthesis, FTPS ranking, weight-sensitivity analysis, leave-one-cohort-out robustness, and held-out public

cohort direction preservation. A key strength of the framework is that it separates target nomination from biological validation. Public metagenome data are powerful for hypothesis generation because they provide multi-cohort scale and standardized feature profiles[14, 15], but processed public profiles are not equivalent to mechanistic experiments and do not necessarily include paired metabolite measurements, intervention outcomes, strain-resolved pathway attribution, or standardized clinical endpoints[5]. Therefore, the present workflow should be interpreted as a computational prioritization layer rather than a validation study. At the feature level, functional modules appeared more suitable than taxa for first-run target prioritization. Taxa features were substantially more numerous, whereas modules showed higher median prevalence and lower median heterogeneity. However, modules were not superior across every metric because direction consistency was not uniformly higher for modules. This balanced result supports the use of functional modules as a tractable and interpretable prioritization layer, while retaining taxa features as a complementary benchmark.

The SCFA-dominant result is biologically plausible and computationally robust. SCFA fermentation pathways are central to microbial metabolism of non-digestible carbohydrates and are connected to gut physiology[7, 16]. In particular, microbial production of acetate, propionate, and butyrate has been linked to host energy metabolism, intestinal barrier function, immune regulation, and other aspects of gut–host physiology[7, 16, 17]. In this framework, SCFA modules dominated the mapped feature space and produced the strongest multi-cohort evidence. CENTFERM-PWY and PWY-6590 were the top-ranked FTPS modules and remained stable across weighting schemes, while P461-PWY also satisfied the combined meta_FDR and direction-consistency criteria. Tryptophan-related modules formed a secondary candidate tier. COMPLETE-ARO-PWY ranked highly because of reproducibility and prevalence, whereas PWY-6629 met nominal FDR criteria but did not satisfy the combined direction-consistency threshold. GABA-related modules were classified as exploratory. Bile acid evidence was limited by minimal mapping, and phytochemical modules were not recovered in the current processed-profile run. The mapping layer represents an important interpretive boundary: only 52 of 562 functional pathway features were mapped or partially mapped, whereas 510 remained unmapped. In addition, the mapping was keyword-derived, and all mapped features require manual review. Therefore, the target map should be used to guide prioritization rather than to support final biological claims.

Beyond its internal robustness, the SCFA-dominant result should also be read against earlier cross-study syntheses rather than as an independent discovery. Both Duvallet et al.[12] and Armour et al.[13] found that much of the case-control signal is shared across diseases, with acetate production modules among the shared functional indicators[13]. The SCFA axis identified here is therefore consistent with that work and is not presented as a novel biological finding.

Relative to that work, the present analysis differs in three respects. First, we used MetaCyc pathways rather than KEGG modules, so the results are convergent rather than equivalent. Second, earlier syntheses did not rank functions; the Functional Target Priority Score and its sensitivity analyses address ranking and its stability. Third, we tested direction preservation in a held-out cohort. The same comparison also reveals two limitations. First, our primary contrast pooled cohorts of differing

phenotype, so the SCFA result likely reflects a nonspecific axis rather than a disease-specific signal; phenotype-stratified analysis is needed. Second, our cohorts partially overlap those analyzed previously, so the present results constitute a reanalysis rather than an independent replication.

Beyond these comparison-specific issues, several broader limitations apply. The held-out public cohort analysis provided evidence of direction preservation, but not independent clinical validation. Direction agreement was observed for 28 of 31 modules. This supports robustness across public data, but the endpoint was direction agreement rather than clinical response, metabolite abundance, or intervention efficacy. Likewise, the first-run control suite was useful but incomplete. The random-label diagnostic suggested separation between observed and shuffled-label results, but it was based on a single seeded shuffle rather than a full permutation test, while the shuffled-axis diagnostic was structural only and unrelated pathway controls were unavailable. Future work should therefore include repeated permutations, unrelated pathway sets, matched prevalence controls, and negative phenotype controls. Finally, taxon-function carrier attribution was not available. Because processed profiles aggregate pathway abundance, the current analysis cannot determine whether prioritized SCFA modules are carried by specific taxa, distributed across many taxa, or driven by different taxa in different cohorts. Future analyses should use stratified HUMAnN outputs from raw reads or precomputed stratified pathway tables[4, 5].

In conclusion, this study establishes a reproducible first-run framework for microbiome functional target prioritization from processed public metagenomes. The analysis nominates SCFA-dominant fermentation modules as robust computational candidates, positions tryptophan metabolism as a secondary candidate axis, and classifies GABA, bile acid, and phytochemical evidence as exploratory or insufficient under the current processed-profile mapping. Future work should extend this framework with paired metabolomics, raw-read stratified profiling, intervention metadata, full negative-control testing, and externally deposited reproducibility packages.

Abbreviations

CSBJ
computational and structural biotechnology journal
FTPS
functional target priority score
GABA
gamma-aminobutyric acid
HMP
human microbiome project
HUMAnN
The HMP Unified Metabolic Analysis Network
LOCO
leave-one-cohort-out

SCFA
short-chain fatty acid

Declarations

Note: Figure callouts are placed in the Results section at the point where each figure supports the corresponding result. Supplementary figures are cited where they provide cohort eligibility, metadata, adjustment, robustness, validation-boundary, negative-control, carrier-attribution, or reproducibility-package support.

Data availability

The first-run reproducibility package manifest enumerated 71 generated files across five result groups: results/tables, results/figure_ready, results/diagnostic_plots, results/manifests, and results/run_summary. The total package size was 145,885,418 bytes, corresponding to 139.13 MB. No external repository DOI, URL, or archive link was provided in the current manifest. Repository deposition and accession information should be added before submission.

Code availability

No source-code directory or checksum claim was present in the current manifest. Code availability information should be completed before submission once the analysis scripts are deposited or archived.

Author contributions

Conceptualization: Y-HH, S-YC; Data curation: C-YC, B-CS, S-YC; Formal analysis: C-YC, B-CS, S-YC; Methodology: C-YC, B-CS, S-YC; Software: C-YC, B-CS, S-YC; Visualization: C-YC, B-CS, S-YC; Writing - original draft: G-HB, C-YC; Writing - review and editing: Y-HH, S-YC.

Ethics declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors sincerely thank all the authors who generously shared their experiences and took time to complete the manuscript.

Funding

The study was supported by grants from NSTC 115-2622-E-038-001

Declaration of generative AI and AI-assisted technologies

During preparation of this manuscript, the authors may have used AI-assisted tools for language editing, structural drafting, and formatting. The authors must review and edit all AI-assisted content and take full responsibility for the final manuscript. This statement should be revised to match actual tool use and journal requirements before submission.

References

1. Human Microbiome Project C. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.
2. Integrative HMPRNC. The Integrative Human Microbiome Project. Nature. 2019;569:641–8.
3. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, Beghini F, Malik F, Ramos M, Dowd JB, et al. Accessible, curated metagenomic data through ExperimentHub. Nat Methods. 2017;14:1023–4.
4. Beghini F, McIver LJ, Blanco-Miguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* 2021, 10.
5. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N, Huttenhower C. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods. 2018;15:962–8.
6. Tian L, Wang XW, Wu AK, Fan Y, Friedman J, Dahlin A, Waldor MK, Weinstock GM, Weiss ST, Liu YY. Deciphering functional redundancy in the human microbiome. Nat Commun. 2020;11:6217.
7. Koh A, De Vadder F, Kovatcheva-Datchary P, Backhed F. From Dietary Fiber to Host Physiology: Short-Chain Fatty Acids as Key Bacterial Metabolites. Cell. 2016;165:1332–45.
8. Louis P, Flint HJ. Formation of propionate and butyrate by the human colonic microbiota. Environ Microbiol. 2017;19:29–41.
9. Agus A, Planchais J, Sokol H. Gut Microbiota Regulation of Tryptophan Metabolism in Health and Disease. Cell Host Microbe. 2018;23:716–24.

10. Strandwitz P, Kim KH, Terekhova D, Liu JK, Sharma A, Levering J, McDonald D, Dietrich D, Ramadhar TR, Lekbua A, et al. GABA-modulating bacteria of the human gut microbiota. Nat Microbiol. 2019;4:396–403.
11. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019;569:655–62.
12. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun. 2017;8:1784.
13. Armour CR, Nayfach S, Pollard KS, Sharpton TJ. A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome. *mSystems* 2019, 4.
14. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD, Finn RD. A new genomic blueprint of the human gut microbiota. Nature. 2019;568:499–504.
15. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell. 2019;176:649–e662620.
16. Morrison DJ, Preston T. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. Gut Microbes. 2016;7:189–200.
17. Parada Venegas D, De la Fuente MK, Landskron G, Gonzalez MJ, Quera R, Dijkstra G, Harmsen HJM, Faber KN, Hermoso MA. Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevance for Inflammatory Bowel Diseases. Front Immunol. 2019;10:277.

Figures

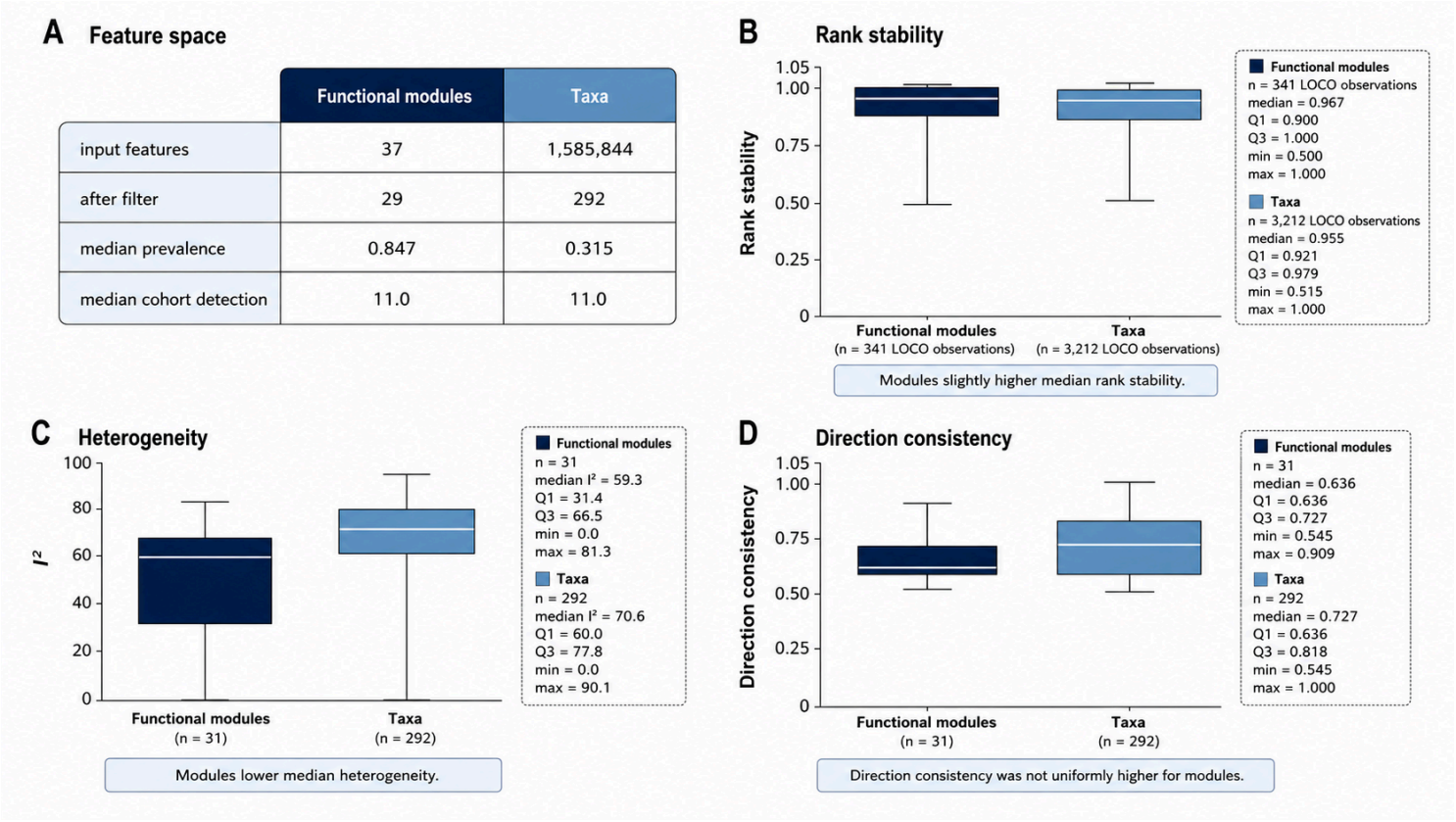


Figure 2

Functional modules show higher prevalence and lower heterogeneity than taxa.

This figure compares functional modules and taxa features as alternative layers for first-run prioritization. (A) Feature-space summary showing input features, features retained after filtering, median prevalence, and median cohort detection for functional modules and taxa. (B) Rank-stability comparison across leave-one-cohort-out observations for functional modules and taxa; the inset reports the number of LOCO observations and distribution summaries. (C) Heterogeneity comparison based on I^2 for retained functional modules and taxa, showing lower median heterogeneity for modules. (D) Direction-consistency comparison for functional modules and taxa. Together, the panels support use of functional modules as a compact and lower-heterogeneity prioritization layer, while showing that modules are not uniformly superior across all metrics.

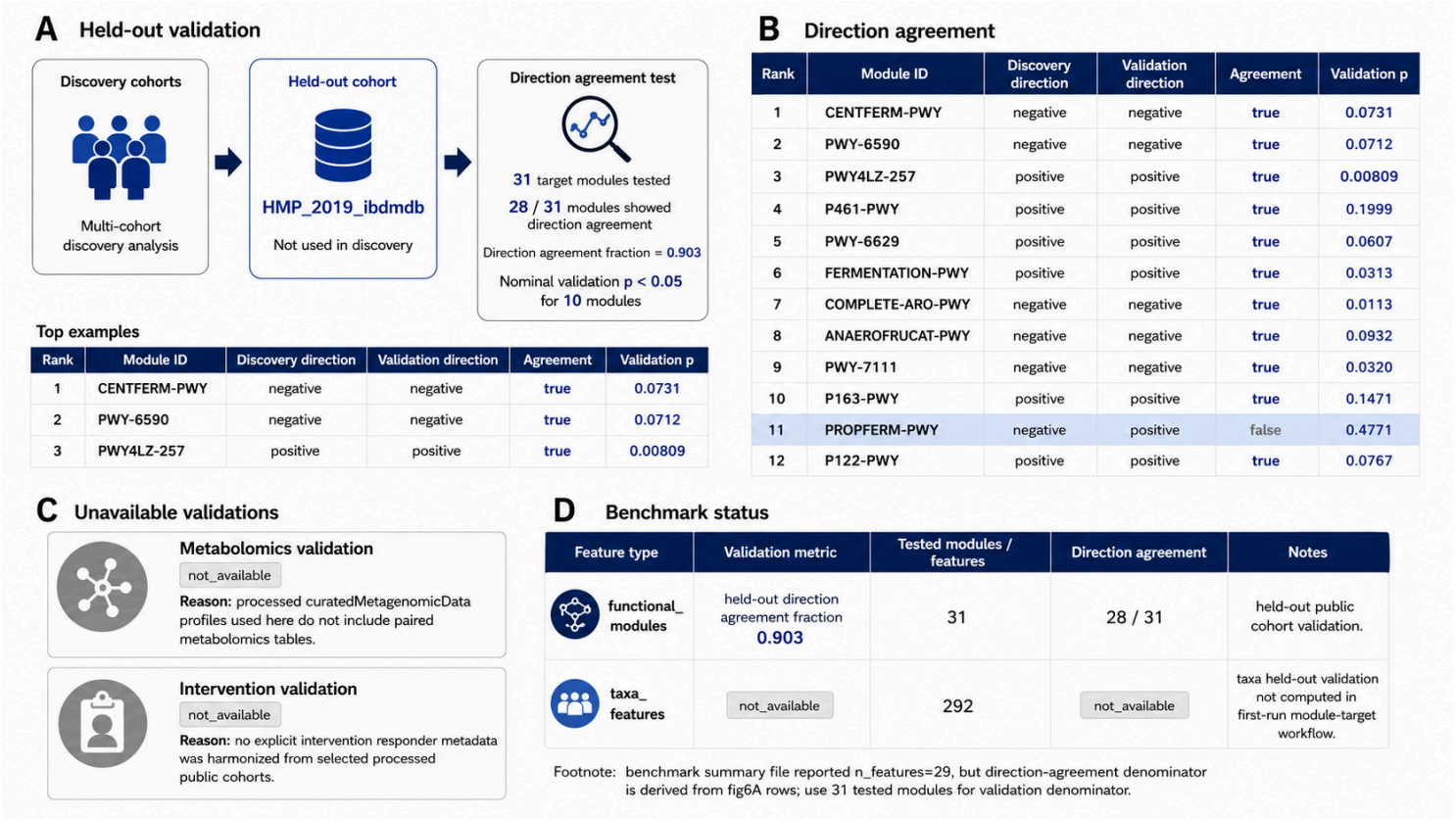


Figure 6

Held-out public cohort validation and analysis limitations.

This figure summarizes held-out public cohort direction preservation and unavailable validation layers. (A) Validation schematic showing discovery cohorts, the HMP_2019_ibdmdb held-out cohort, and the direction-agreement test. The panel reports 31 target modules tested, 28/31 modules with direction agreement, direction agreement fraction of 0.903, and nominal validation p < 0.05 for 10 modules, with three top examples. (B) Top-12 direction-agreement table reporting discovery direction, validation direction, agreement status, and validation p for each module. The discordant row is displayed in pale gray/blue. (C) Unavailable-validation status cards indicating that metabolomics validation and intervention validation were not available in the current processed-profile run. (D) Benchmark-status table comparing functional_modules and taxa_features, including tested-module or feature counts, validation metric availability, direction agreement, and notes. The footnote clarifies that the direction-agreement denominator is derived from fig6A rows and should use 31 tested modules.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementarydataBMCmicrobiology.docx