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.

SELECT-NET: Multiple contextual representations of bacterial genomes through selection-based subnetwork reconstruction

Gerardo R. Amores

gerardo.ruiz.a@cinvestav.mx

Center for Research and Advanced Studies of the National Polytechnic Institute (Cinvestav)

Method Article

Keywords: bacterial genomics, contextual gene prioritization, subnetwork reconstruction, genomic neighborhoods, contextual filtering, reproducible workflow.

Posted Date: October 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10406453/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

High-throughput bacterial sequencing routinely generates large gene catalogs and functional annotations that are difficult to translate into biologically interpretable structures. Here, we present SELECT-NET, a reproducible workflow for contextual gene prioritization and genomic-proximity subnetwork reconstruction from bacterial sequencing data. The framework integrates annotation, abundance estimation, functional information, and genomic-proximity network reconstruction through four biologically guided filtering strategies targeting adaptive, abundance-associated, isolate-specific genomic content, and epidemiological features. SELECT-NET was evaluated across laboratory strains, clinically relevant pathogens, and environmental isolates using both de novo assembly and reference-guided workflows. Quantitative filtering retained comparable gene-function categories despite substantial reductions in network complexity. Heuristic core subtraction emphasized isolate-specific genomic content in Shiga toxin–producing *Escherichia coli* (STEC) and *Clostridioides difficile* while eliminating all retained features in the laboratory strain *E. coli* TOP10. Epidemiological filtering distinguished the laboratory strain *E. coli* TOP10 from pathogenic STEC isolates through the recovery of virulence-associated determinants and antibiotic-resistance genes, further differentiating the two STEC strains by their resistance-associated profiles. Thematic filtering on *Cupriavidus gilardii* strains isolated from a mine recovered similar adaptive categories associated with metal-stress adaptation and environmental persistence despite marked differences in network organization. Together, these results suggest that SELECT-NET provides a practical framework for generating multiple contextual representations of bacterial genomes through prioritization and subnetwork reconstruction, supporting comparative analysis, visualization, and hypothesis generation across diverse genomic contexts.

Introduction

The rapid growth of bacterial sequencing data has greatly expanded our ability to characterize gene content, functional potential, and abundance patterns at scale. However, usually, the final output of the analytical workflow remains an extensive collection of annotation tables, abundance matrices, and pathway summaries that are difficult to translate into biologically interpretable structures (Tripathi, 2026). This creates a critical interpretive bottleneck in which gene-level information is increasingly abundant, while the biologically informed prioritization of genes and pathways often remains manual, fragmented, and highly dataset-specific. Consequently, researchers frequently require additional downstream bioinformatic processing before achieving a coherent functional interpretation (Khew, 2026; Gollapalli, 2026; Liu, 2026; Chen, et al., 2026).

This limitation becomes relevant when bacterial datasets contain signals that may be relevant under different biological perspectives, such as stress adaptation, resistance, virulence, highly abundant core processes, or isolate-specific and non-core-like functions. Standard workflows are highly effective for preprocessing, assembly, annotation, and quantification, but they do not always provide an integrated route from read-derived information to context-aware genomic subnetworks. Previous studies have suggested that biologically relevant functions are often better understood through groups of associated genes rather than isolated annotations, providing a conceptual basis for the exploration of network-derived genomic subnetworks in genomic datasets (Ogata et al., 2000; Tamura & D'haeseleer, 2008). Inspired by this view, SELECT-NET was designed to generate alternative contextual representations of bacterial genomes through biologically guided filtering and subnetwork reconstruction. SELECT-NET was designed not to replace established tools for assembly or annotation, but to connect them into a reproducible framework capable of producing biologically filtered, network-oriented outputs from bacterial sequencing data. The framework is compatible with both whole-genome sequencing (WGS) and RNA sequencing (RNA-seq) inputs. Although the present study primarily evaluates genomic datasets, a representative RNA-seq dataset was additionally included to assess transcriptomic compatibility.

The conceptual basis of this work aligns with the expanding use of network-oriented strategies for biological interpretation. Previous work has highlighted how network principles can support biological interpretation through accessible platforms such as Cytoscape and UseGalaxy (Ruiz Amores & Martínez-Antonio, 2022). Furthermore, browser-executable workflows like ACE-AB1 have demonstrated the practical utility of cloud-based execution via Google Colab and Google Drive integration, offering open, zero-installation analytical environments (Sosa Reyes & Amores, 2024). In this context, SELECT-NET represents

a lightweight, shareable, cloud-executable workflow designed to eliminate local infrastructure barriers while preserving biological utility.

Here, we present SELECT-NET, a reproducible workflow for contextual gene prioritization and genomic subnetwork construction from bacterial sequencing reads. By integrating multiple stages of bacterial genome analysis into a single workflow, SELECT-NET provides a unified framework for generating contextual representations of bacterial genomes according to different analytical objectives. To maximize flexibility across different data availability scenarios, SELECT-NET is implemented in two complementary workflows. The Full workflow performs de novo assembly, gene prediction, annotation, normalized abundance metric, filtering, and network construction directly from sequencing short reads, whereas the Lite workflow operates on pre-existing genomic resources, requiring annotated FAA, GFF, and FNA files. The Lite workflow is not annotation-only, thus, sequencing reads are still required for normalized abundance estimation, contextual filtering, and network reconstruction.

We evaluated both workflows across diverse bacterial genomic datasets representing distinct ecological, epidemiological, and adaptive contexts, including laboratory strains, clinically relevant pathogens, and environmental isolates. Across all datasets, SELECT-NET consistently reduced genome-scale feature spaces into smaller contextual subnetworks while retaining functional categories relevant to the selected filtering strategy. Depending on the selected filtering strategy, the workflow recovered networks associated with isolate-specific genomic content, epidemiologically relevant determinants, adaptive stress-response functions, and broadly conserved biological processes. Together, these results support the use of contextual prioritization and subnetwork reconstruction as a framework for generating multiple contextual representations of bacterial genomes across diverse analytical objectives. Crucially, SELECT-NET is intended as a practical resource for prioritization, visualization, and downstream hypothesis generation rather than for the de novo inference of experimentally validated interaction networks.

Materials and Methods

Workflow overview

SELECT-NET is a browser-based workflow for contextual gene prioritization and genomic subnetwork reconstruction from bacterial short read sequencing data. The workflow is designed for execution from a Google Colab notebook, allowing end-to-end analysis from a single interactive script. Google Drive is used for output persistence and checkpoint storage. The pipeline accepts either public sequencing accessions or local FASTQ files supplied through Google Drive and integrates read preprocessing, assembly, gene prediction, annotation, mapping, count summarization to perform abundance estimation, KEGG Orthology (KO) assignment, pathway mapping, contextual filtering, and network export into a unified analysis chain. Alternatively, if GFF, FAA, and FNA files are also available, SELECT-NET executes the Lite workflow, which uses the supplied annotations plus short reads to start the abundance calculations, filtering and network reconstruction without repeating the assembly and annotation stages. Final outputs include complete and filtered gene tables, enriched node and edge tables, pathway-linked gene tables, graph objects in GraphML and GEXF formats, a static network image, and structured run reports suitable for downstream inspection and reuse (Fig. 1). In its current form, SELECT-NET is best understood as a workflow for reproducible exploratory analysis rather than high-throughput batch processing.

Software environment, runtime dependencies, and reproducibility

SELECT-NET was executed in Google Colab with Google Drive-based persistence for outputs and intermediate checkpoints. The workflow installs its software environment during runtime and combines system-level tools, Python packages, repository-derived components, and selected external resources required for preprocessing, assembly, annotation, count summarization, functional assignment, and network generation. In the current implementation, this includes command-line bioinformatics tools installed through the system package manager, Python libraries installed in the active notebook environment, repository-based annotation components retrieved during execution, and external resources required for KO assignment and pathway

mapping. Software versions, runtime dependencies, and external resources used in each run were documented directly from the active execution environment and are summarized in Online Resource 1.

For the datasets evaluated in this study, Full workflow execution typically required approximately 1.5–2.5 h prior to checkpoint generation. Once checkpoints were available, subsequent executions generally required 7–20 min depending on the selected analysis stage. The Lite workflow further reduced execution time by bypassing assembly and annotation procedures. Across the evaluated datasets, approximately 10 GB of Google Drive storage were sufficient to retain intermediate files and final outputs. All analyses were successfully executed using the standard Google Colab environment available at the time of the study.

Input datasets

The system was evaluated using WGS bacterial datasets selected to represent distinct genomic or functional contexts (Table 1). Datasets were chosen to test the performance of the different filtering modes under conditions expected to differ in resistance-associated content, abundance structure, or isolate-specific/non-core-like composition. The standard operational procedure to operate SELECT-NET is provided at the Online Resource 2. The implementation rationale of SELECT-NET focuses on WGS datasets because they provide a consistent framework for evaluating the different filtering strategies across genomes (Table 1). RNA-seq compatibility was additionally evaluated using an *Acinetobacter baumannii* transcriptomic dataset (ERR976705). Mode A was evaluated primarily using environmental datasets. Mode C was evaluated primarily using closely related isolates and reference-based comparisons. Mode D was evaluated primarily using datasets containing known virulence- or resistance-associated features.

Table 1
WGS case studies selected to evaluate SELECT-NET

Study	Biological context	Why it was selected	What it is expected to show in Results	Data availability
Escherichia coli TOP10 - REMoDE of DNA. (Enam, et al., 2023)	Control strain of the reference study	Genomic DNA from *E. coli* TOP10 to show REMoDE approach to enrich enterobacterial DNA.	Internal control at network structure and retention performance of filter modes.	PRJNA903933. SRR22364163. (Illumina MiSeq, WGS, paired-end)
Persistent Shiga Toxin-producing *E. coli* Environmental Investigation (Holcomb, et al., 2026)	Outbreak genomics / environmental transmission / public health epidemiology	Environmental and epidemiological investigation of a STEC O157:H7 outbreak in a public-health context.	To highlight Epidemiological and core-subtraction analyses. (STEC-1,-2)	PRJNA218110. SRR7739427 (1) SRR6128369 (2) (Illumina MiSeq, WGS, paired-end)
Genomic Analysis of *C. difficile* Strains in Mexico: Insights into Virulence and Resistance (Ortiz-Flores, et al., 2026)	Comparative genomics / virulence and resistance profiling / Mexico vs. global strains	Genomics of *C. difficile* to identify characteristics in genetic composition, virulence factors, and antibiotic resistance.	Highlight Epidemiological and *Clostridioides* external core-like analyses.	PRJNA1125493 SRR32099327. (Illumina HiSeq 2500, WGS, paired-end)
Cupriavidus gilardii strains NOV2-1 and OV2-1 isolated from the Naica Mine (González-Sánchez et al., 2025)	Exhibit resistance to metals including Zn, Pb, Cu, Cd, Ni, and Co	The dataset contains clear genotype–phenotype associations linked to metal resistance and stress adaptation	To highlight Mode A of modules associated with heavy-metal resistance, stress tolerance, among others.	PRJNA757420 SRR15654682 SRR15654681. (Illumina HiSeq 4000, WGS, paired-end)

An *Escherichia coli* TOP10 dataset was included as an internal reference for cross-mode comparison. A species-matched low-complexity comparator was not included for the clostridial case in the present study, as the aim was to prioritize biologically relevant WGS case studies rather than to establish matched benchmark controls for each supported external reference group. PRJNA bioproject, SRR the SRA numbers Ids at NCBI.

Read preprocessing and sequence assembly

Raw reads were preprocessed with Trim Galore using Cutadapt as the trimming backend. In the current implementation, reads were filtered with a Phred quality threshold of 20 and a minimum retained length of 36 bp. post-trimming quality reports were generated with FastQC. Cleaned reads were then assembled with SPAdes using, only-assembler, k-21,33,55, threads-10, and memory-10, called as “fast” mode as the recommended assembly strategy due to its favorable balance between computational efficiency and assembly quality. Contig headers were subsequently sanitized to ensure compatibility with downstream annotation tools.

Gene annotation and normalized abundance estimation

The assembled contigs were analyzed with Prodigal as a standalone gene-finding algorithm to identify the coordinates of protein-coding genes (ORFs), then Prokka was assigned as the primary annotation engine, checkpoints were stored to support restoration in subsequent runs. Reads were then mapped back to the assembled contigs using minimap2, and sorted and indexed alignment files were generated with samtools. Per-feature counts were obtained with featureCounts using CDS entries from the annotation GFF file; for paired-end libraries, featureCounts was executed in pair-counting mode for each annotated gene. Normalized abundance was calculated from the featureCounts output using an RPKM-based normalized abundance metric $(10^9 \times C)/(N \times L)$, where C is the count assigned to a gene, L is the annotated gene length in base pairs, and N is the total counted signal across all genes in the featureCounts table for that sample. In this workflow, the normalized value

was used as a within-sample prioritization metric and for node-level visualization. For RNA-seq datasets, it was interpreted as normalized transcript abundance, whereas for WGS datasets it was interpreted as normalized DNA read representation and should not be interpreted as gene abundance. This metric was used for within-sample prioritization and visualization only and was not intended as a substitute for formal cross-sample differential abundance analysis. Therefore, although the workflow uses an RPKM-based normalization strategy, the metric was employed exclusively as a relative within-sample prioritization metric. Consequently, biological interpretation in WGS datasets should be understood in terms of gene representation rather than transcriptional activity. When NCBI local files (gff, faa, fna) are loaded into input folder, the tool maps the reads to the provided files to continue the pipeline.

Functional assignment

Predicted protein sequences were further processed with KOfamScan to assign KEGG Orthology identifiers based on profile Hidden Markov Models (HMMs), resulting in KO assignments and associated functional categories. These annotations were then integrated with gene-level counts, normalized abundance values, annotation metadata, and pathway membership into a unified interpretation table.

Filtering modes

SELECT-NET implements four primary analytical strategies through six operational modes. Modes A, B, C, and D represent the primary strategies, whereas CH and CL are abundance-stratified exploratory extensions of Mode C. The detailed summary of the conceptual strategies, operational modes, retained categories, and intended interpretations is provided in Online Resource 2 (Table 2). Mode A (Thematic) retains genes assigned to Extremophile-like, Bioplastic-like, Biodegradation-like, and Industrial-enzyme-like categories. Mode B (Quantitative) prioritizes highly represented genes according to within-sample abundance and does not attempt to identify objectively optimal, informative, or biologically superior features. Mode C (Core subtraction) removes genes classified as external core-like by reference-based matching (Dyer, et al., 2025). Derived from the core-subtraction framework, two operational modes aiming to extend the biological information included Mode CH (High-abundance core subtraction) which removes genes matched to the selected external core-like reference and assigned to the upper and mid abundance band within that matched subset. And, Mode CL (Low-abundance core subtraction) removes genes matched to the selected external core-like reference and assigned to the lower and mid abundance band within that matched subset. The Mode D (Epidemiological) retains genes assigned to Resistance and Virulence categories. Together, these modes allow the same dataset to be reorganized into biologically distinct subnetworks under different interpretive lenses.

Contextual gene classification – Mode A (Thematic filtering)

Microorganisms adapted to extreme environments constitute an important source of extremozymes, stress-tolerance determinants, and other biomolecules with potential industrial and biotechnological applications (Raddadi, et al., 2015). Therefore, the thematic filtering mode was designed to prioritize genes whose annotation profiles suggest potential relevance to environmental adaptation and biotechnology-oriented applications. In the current implementation, genes assigned to the “Extremophile-like”, “Bioplastic-like”, “Biodegradation-like”, and “Industrial-enzyme-like” classes were retained. These classes were derived heuristically from gene product descriptions, gene names, and KO-associated descriptors and were intended to enrich the resulting subnetworks for stress-associated functions, biopolymer-related metabolism, biodegradation-linked activities, and putative industrial enzymes.

Accordingly, these labels should be interpreted as reproducible exploratory prioritization rules rather than formal gene-classification systems. Their purpose is to facilitate contextual exploration of large genomic datasets by enriching subnetworks for genes potentially associated with a given biological theme. The implemented dictionaries, provided in Online Resource 3 and fully modifiable by users, were designed as transparent annotation-prioritization schemes rather than predictive classification models. Consequently, category assignments reflect annotation-derived semantic grouping and should not be interpreted as direct evidence of biological function, activity, or phenotype.

Quantitative normalized genomic representation classification - Mode B

In Mode B, quantitative prioritization was performed using the RPKM-based normalized abundance metric calculated for each annotated gene. In the current implementation, Q0.90 is used as the default threshold. This mode was intended to capture the upper abundance component of the dataset and to support network construction from highly represented genes within the sample. It should be interpreted as a prioritization strategy based on within-sample representation rather than as an attempt to identify objectively optimal or biologically superior genomic features. Sensitivity analyses using Q0.85 and Q0.95 are presented below.

External core-like reference matching - Mode C

Mode C implements heuristic subtraction of external core-genome references. For this purpose, EnteroBase cgMLST schemes were used as conservative species-specific proxies for broadly shared genomic content. These schemes define conserved genomic loci shared among the majority of isolates within selected bacterial species and are widely used in genomic epidemiology and phylogenetic analyses (Maiden et al., 1998; Maiden et al., 2013; Zhou et al., 2020; Dyer et al., 2025). The cgMLST-based subtraction procedure was designed as a heuristic prioritization strategy intended to reduce conserved genomic content and enrich the resulting subnetworks for isolate-specific genomic features. Therefore, it should not be interpreted as a formal accessory-genome inference method or as a replacement for dedicated pangenome-analysis frameworks.

Therefore, genes not matching the selected cgMLST reference are subsequently prioritized, allowing exploration of genomic content outside the conserved reference background. In the current implementation, reference sets are provided for *Escherichia/Shigella, Salmonella, Yersinia, Clostridioides*, and *Mycobacterium tuberculosis*. Matches were identified using a hierarchical heuristic strategy involving exact gene-symbol matching and partial matching against gene symbols and annotation descriptions. These matches define the subset of reference-associated genes used during subtraction-based filtering. The objective of this procedure is to reduce broadly conserved genomic content prior to network generation, thereby enriching the resulting subnetworks for candidate isolate-specific genomic features. Consequently, the matching strategy should be interpreted as a reference-guided prioritization step rather than as a formal evaluation of core-genome membership.

Abundance-modulated core partitioning - Mode CH / CL

For abundance-modulated core subtraction, genes matched to the selected external core-like reference were partitioned by internal abundance quantiles within the external core-matched subset. In the current implementation, the lower and upper bands were defined using the 0.25 and 0.75 quantiles of the normalized abundance distribution of the matched core-like subset, respectively. This enabled selective subtraction of low-abundance or high-abundance external core-like genes while preserving the remaining contextual gene set for downstream network construction. Therefore, abundance-stratified modes (CL and CH) were explored to selectively retain subsets of core-associated genes based on abundance level. These modes enabled inclusion of both non-core genes and abundance-independent fractions of the core genome. However, due to operon reconstruction driven by genomic proximity, these modes did not provide additional biologically coherent discrimination compared to the standard core removal (mode C). Therefore, CH and CL were retained as optional exploratory or visualization modes and are presented in supplementary analyses.

Contextual gene classification – Mode D (Epidemiological filtering)

The epidemiological filtering mode was designed to prioritize genes associated with virulence and antimicrobial resistance. In the current implementation, genes assigned to the Virulence and Resistance classes were retained. These categories were derived heuristically from gene product descriptions, gene names, and KO-associated descriptors and were intended to enrich subnetworks for annotations commonly associated with bacterial pathogenicity, host interaction, antimicrobial resistance, and related epidemiological themes. As with thematic filtering, these labels should be interpreted as reproducible exploratory

prioritization rules rather than predictive classification systems. The implemented dictionaries, provided in Online Resource 3 and fully modifiable by users, were designed as transparent annotation-prioritization schemes rather than predictive classification models. Consequently, category assignments reflect annotation-derived semantic grouping and should not be interpreted as direct evidence of biological function, activity, or phenotype.

Network construction

SELECT-NET generates context-aware genomic association networks rather than experimentally validated interaction maps. Nodes correspond to genes retained after contextual filtering. Genes directly retained by the selected filtering mode are referred to as seed genes. To preserve local genomic context, SELECT-NET expands these seeds by incorporating neighboring genes located within 2000 bp up and down on the same contig. The rationale for ± 2000 is discussed in the sensitivity analysis. This step aims to recover operon-like structures and other locally organized genomic modules that may contribute to the biological interpretation of the filtered genes. To prioritize genomic context over inferred functional connections, SELECT-NET does not introduce edges from KEGG pathway assignments. Instead, edges are established exclusively from genomic proximity using contig-level coordinates derived from the assembled genome or from externally supplied annotations. This strategy preserves consistency with the underlying annotation source while avoiding biases introduced by external interaction databases.

Following contextual expansion, genes are connected when they occur within a predefined genomic distance threshold. Subsequently, a heuristic edge-pruning step is applied using gene-name prefixes to reduce biologically implausible associations among neighboring genes with unrelated annotations. Components lacking at least one seed gene are removed, ensuring that all retained subnetworks remain anchored to the original filtering objective.

These procedures were designed as conservative graph-refinement heuristics intended to improve network interpretability and preserve locally coherent genomic structures rather than to perform formal operon prediction or functional-association inference. Consequently, the final network represents a context-aware genomic association graph in which seed genes are embedded within their immediate genomic environment. The resulting topology prioritizes biological interpretability and local genomic organization over network density. Because not all genes are expected to participate in tightly connected genomic modules, disconnected nodes may remain in the final graph when they satisfy the filtering criteria but lack suitable genomic associations.

Importantly, genes are not organized in isolation within bacterial genomes. Consequently, contextual expansion may introduce neighboring genes belonging to functional categories that differ from the original filtering objective. Therefore, subnetworks generated by a specific filtering mode should not be interpreted as exclusively composed of genes matching that category, but rather as contextual genomic representations in which neighboring functions are retained to preserve local biological organization.

Network export and visualization

For each run, the workflow exports (i) complete and filtered gene tables, (ii) enriched node and edge tables, (iii) gene-to-KO-to-pathway mapping tables, and (iv) graph files in GraphML and GEXF formats. These outputs were specifically designed for direct import into Cytoscape and related network analysis environments. Additionally, SELECT-NET generated run-level structured reports in JSON and Markdown formats, along with a cumulative summary table to ensure traceability of analytical settings. Static PNG visualizations were also produced for immediate inspection without requiring further software. Representative GraphML files corresponding to the subnetworks discussed in the main figures are provided as Online Resource 7 to support independent inspection and reuse.

Network visualization and comparison with reported biological features

Generated GraphML subnetworks were imported into Cytoscape for visualization and manual inspection. Functional categories and normalized genomic representation abundance values were associated with node attributes and displayed

when appropriate. All networks were visualized using the Prefuse Force Directed layout unless otherwise specified. Comparative interpretation focused on differences in node retention, category composition, topology, and biologically prioritized subnetworks across datasets and filtering modes. In addition, genes and functional categories retained by SELECT-NET were verified against the workflow-generated summary tables and compared with biological features previously reported in the original studies from which the sequencing datasets were obtained. This comparison was used to assess whether biologically relevant traits remained represented after contextual filtering and network reconstruction.

Network Parameter Sensitivity Analysis

Sensitivity analyses of contextual expansion distance and abundance threshold selection were performed using two representative datasets, STEC-1 (SRR7739427) and *Cupriavidus gilardii* OV2 (SRR15654681), selected because they produced stable and biologically interpretable results across thematic, environmental, and epidemiological filtering modes. To evaluate contextual expansion sensitivity, network reconstruction was independently performed using expansion distances of ± 1000 bp, ± 2000 bp, and ± 3000 bp. To assess abundance-threshold sensitivity, exported SELECT-NET networks generated under Q0.85, Q0.90, and Q0.95 were compared using an ad-hoc Python script. Shared and unique KEGG Orthology (KO) assignments were identified through set-intersection analysis of KO annotations recovered under each abundance threshold.

Results

Global assessment of SELECT-NET across whole-genome sequencing datasets

Execution of SELECT-NET across the evaluated datasets (Table 1) revealed a progressive reduction in feature space throughout the workflow (Table 2). Annotated genes represented the largest feature set, followed by KO-assigned genes, nodes retained after filtering, and final network nodes Differences between retained and final node counts reflect the successive graph-construction and filtering procedures applied during network generation.

Across datasets, quantitative filtering (Mode B) generally retained larger subnetworks, whereas thematic (Mode A), core-subtraction (Mode C), and epidemiological (Mode D) filters produced more compact representations. Node counts varied across datasets and filtering strategies, while comparable orders of magnitude were observed between the Full and Lite workflows evaluated using the *C. difficile* datasets (Table 2).

Table 2
Data reduction and network generation across SELECT-NET workflows

Dataset	Annotated genes	KO-assigned genes	Mode A		Mode B		Mode C		Mode D	
			Retained nodes	Final nodes	Retained nodes	Final nodes	Retained nodes	Final nodes	Retained nodes	Final nodes
E. coli TOP10	4,172	3,905	749	319	1198	569	0	0	862	390
STEC-1	5,035	4,496	NS	NS	1262	780	768	369	1075	557
STEC-*2*	5,116	4,580	NS	NS	1238	785	830	399	1106	569
C. difficile	3,648	3,334	NS	NS	964	611	218	73	503	264
**C. difficile*	3,698	3,346	NS	NS	975	814	42	19	504	431
**C. gilardii* NOV2-1	5,740	5,265	1,554	684	2,228	1,249	NS	NS	NS	NS
**C. gilardii* OV2-1	5,510	5,051	957	708	1,996	1,245	NS	NS	NS	NS

Annotated genes correspond to all predicted coding sequences. KO-assigned genes represent genes successfully annotated through KOfam HMM searches. Values reported under each filtering mode (Nodes Retained) correspond to genes retained for network construction after contextual expansion, whereby neighboring genes located within ± 2000 bp on the same contig were incorporated to preserve local genomic context around filtered targets. Final Nodes represent genes retained in the final graph after edge construction, edge filtering, and removal of unsupported network components. The asterisk (*) indicates datasets processed using the SELECT-NET Lite workflow; datasets without an asterisk were processed using the Full workflow. NS (Not Selected) indicates dataset–workflow combinations that were intentionally excluded from evaluation because they were not relevant to the biological objective of that particular case study.

Biological representation of abundance-prioritized subnetworks

Application of the abundance-based prioritization filter (Mode B) generated subnetworks that could be compared across taxonomically and ecologically distinct organisms. Despite substantial network reduction, Mode B consistently retained transport systems, regulatory factors, and stress-response functions across datasets (Online Resource 4) Fig. S1.

At *E. coli* TOP10, the resulting subnetworks retained genes associated with central bacterial functions, including curli/biofilm formation (csgA, csgB, csgC), transport systems, and core metabolic processes. STEC-1 and *C. difficile* subnetworks were dominated by adhesion-related factors, virulence-associated determinants, mobile genetic elements, and resistance functions, whereas *Cupriavidus gilardii* isolates displayed transport systems, metal-resistance determinants, biodegradation-related genes, mobile genetic elements, regulatory genes, and stress-adaptation mechanisms (Online Resource 4) Fig. S1. These patterns were readily distinguishable despite the use of the same quantitative filtering strategy across all organisms. Thus, Mode B generated subnetworks while retaining distinct profiles across the evaluated organisms.

Isolate-specific genome enrichment through cgMLST-based core subtraction

Mode C prioritizes isolate-specific genomic content through cgMLST subtraction. No isolate-specific genes were recovered from the laboratory strain *E. coli* TOP10 after subtraction of the *E. coli* cgMLST reference (Table 2). Meanwhile, cgMLST subtraction on the two STEC isolates (SRR7739427 and SRR6128369) revealed substantial isolate-specific genomic content (369 and 399, respectively) (Table 2). As shown in Fig. 2a, the resulting subnetworks included *stxA* and *stxB*, together with adhesion-associated genes, mobile genetic elements, transport systems, regulatory factors, resistance determinants, and specialized metabolic functions. In addition to the virulence-associated features shared with STEC1, STEC2 retained

additional isolate-specific genes, including the tetracycline-resistance determinants *tetA* and *tetR* (Online Resource 4) Fig. S2a. These features were absent from the TOP10 strain.

Mode C was additionally applied to *Clostridioides difficile* using both the Full (de novo assembly) and Lite (pre-annotated genome) workflows. Both methods retained different number of nodes (73 and 19 final nodes, respectively: Table 2). Although the de novo workflow recovered a larger number of genes, both workflows recovered isolate-specific genes associated with motility, transport systems, carbohydrate utilization, and environmental interaction with differences in network resolution Fig. 2b and (Online Resource 4) Fig. S2b. An exploratory extension (Mode CL) generated larger subnetworks while retaining partial core subtraction (Online Resource 4) Fig. S3. Because this mode was not a primary focus of the present study, it was not investigated further. In summary, cgMLST subtraction recovered isolate-specific-genome subnetworks in STEC and *Clostridioides* datasets. The recovered subnetworks should be interpreted as candidate isolate-specific genomic content enriched through reference subtraction rather than as a formal representation of the accessory genome.

Comparative analysis of epidemiologically relevant subnetworks

To evaluate the epidemiological filtering strategy implemented in SELECT-NET, the extensively studied model strain *E. coli* TOP10 was used as a reference genome representing a non-pathogenic background. The Full workflow with epidemiological filtering retained colonization-associated functions commonly observed in environmental and laboratory *E. coli* strains, including curli biosynthesis (*csg*), common pilus formation (*ecp*), and multiple fimbrial systems such as sfm and yfc (Online Resource 4) Fig. S4. These functions are related with adhesion, biofilm formation, environmental persistence, and surface colonization. Virulence and resistance determinants were absent.

A broadly comparable colonization-associated subnetwork was recovered across the evaluated *E. coli* strains, whereas STEC-1 and STEC-2 retained additional distictive virulence- and resistance-associated categories. The recovered subnetworks contained the Shiga toxin genes (*stxA* and *stxB*), adhesion-associated factors, and enterohemolysin-related functions (Online Resource 4) Fig. S4. Additionally, despite sharing the characteristic STEC virulence repertoire, the two isolates displayed distinct isolate-specific profiles. In the recovered subnetwork, STEC-1 retained colonization-associated genes together with STEC-associated virulence determinants. In contrast, STEC-2 retained the same virulence-associated functions while additionally recovering antimicrobial resistance determinants, including *aadA* and *tetA*, together with multidrug efflux and stress-response systems (Online Resource 4) Fig. S4.

Subsequently, we evaluated whether similar biological signals could be recovered from alternative assembly and annotation strategies. To this end, SELECT-NET was applied to *Clostridioides difficile* using both the Full (de novo assembly) and Lite (reference-guided annotation) workflows (Fig. 3). Both workflows recovered comparable epidemiologically relevant functional categories but differed in node count and network resolution. The Lite workflow (Fig. 3a) retained a larger number of functionally annotated genes, resulting in a higher-resolution network, whereas the Full workflow (Fig. 3b) condensed equivalent functional categories into a smaller set of representative nodes. Notably, the numbers of virulence- and resistance-associated nodes were comparable between both workflows, whereas most of the reduction in node count was observed within the “Other” category (Fig. 3c).

Finally, to verify compatibility with transcriptomic datasets, SELECT-NET was additionally applied to an *Acinetobacter baumannii* RNA-seq dataset (ERR976705) using the Full workflow. Epidemiological filtering (Mode D) recovered multiple resistance-associated determinants, multidrug efflux systems, secretion-related proteins, adhesion factors, and iron acquisition mechanisms (Online Resource 4) Fig. S5. This example indicates that SELECT-NET can be applied to RNA-seq input and can recover epidemiologically relevant gene categories, although broader transcriptomic evaluation will be required.

Conserved adaptive functions but divergent network architectures in Cupriavidus gilardii revealed by thematic filtering

Mode A recovered highly similar adaptive subnetworks in both Cupriavidus gilardii strains, including metal-resistance systems, aromatic-compound degradation pathways, stress-response functions, transport systems, and biotechnology-associated categories (Fig. 4). Shared functions included arsenic-, copper-, and zinc-associated determinants, multiple dioxygenases involved in aromatic-compound degradation, oxidative-stress proteins, transporters, and regulatory proteins. Differences between strains were primarily associated with the retention of specific genes and network organization rather than major functional categories. Some resistance- and virulence-associated genes were also recovered; however, these genes were incorporated through contextual expansion during network reconstruction rather than direct Mode A filtering.

Sensitivity of contextual expansion and abundance prioritization

Sensitivity analyses revealed that contextual expansion distance had limited effects on the composition of reconstructed networks. Across all tested distances, the principal biological categories recovered by thematic and epidemiological filtering remained stable (Online Resource 5) Tables S1-2. Although larger expansion windows increased network size through the inclusion of additional neighboring genes, the principal functional categories recovered by thematic and epidemiological filtering remained unchanged.

In contrast, abundance-threshold selection influenced network size and the retention of biologically relevant categories (Online Resource 5) Tables S3-4. Nevertheless, comparison of exported KO annotations revealed substantial functional conservation across abundance thresholds. A total of 223 and 457 KEGG Orthology assignments were shared among Q0.85, Q0.90, and Q0.95 in STEC-1 and *C. gilardii* OV2, respectively, indicating the presence of a conserved functional core. KEGG reconstruction of the conserved KO sets showed that shared assignments were associated with central metabolism, biosynthetic processes, genetic information processing, environmental sensing, and microbial community-associated functions (Fig. 5).

Although the most stringent threshold (Q0.95) recovered a compact subset of functions, the same functional core remained detectable across all evaluated abundance thresholds. As abundance stringency decreased from Q0.95 to Q0.85, additional genes and pathways were progressively incorporated, increasing the biological complexity represented in the reconstructed networks.

Discussion

Biological filtering reveals distinct perspectives of the same genome

Early graph-based genomic studies demonstrated that biologically related genes frequently occur as structured functional modules within complete genomes. For example, Ogata et al. (2000) identified functionally related enzyme clusters by combining genomic organization with curated metabolic pathway information. Such studies established the value of network-based interpretations and highlighted the importance of genomic context in biological inference. Previous studies have suggested that complex biological traits are more effectively explained by coordinated groups of genes than by isolated genetic elements (Tamura and D'haeseleer, 2008). Accordingly, SELECT-NET does not attempt to define a single biologically optimal representation of a genome. Instead, the workflow explores alternative contextual representations generated through complementary filtering strategies, each emphasizing different aspects of genomic organization.

Across the evaluated laboratory, environmental, and clinically relevant datasets, contextual filtering consistently generated subnetworks that reflected biologically meaningful differences while preserving broader functional themes. Laboratory, environmental, and pathogenic strains retained distinct contextual representations despite being analyzed under equivalent filtering strategies, suggesting that subnetwork composition is influenced not only by the selected filter but also by the underlying genomic characteristics of each dataset. Besides, representative examples recovered across different datasets, including *stxA, stxB, tetA, tetR, copA*, and *copB*, demonstrated that the implemented dictionaries consistently prioritized annotations compatible with the intended biological themes.

At the same time, comparable functional categories were frequently recovered across alternative reconstruction strategies and quantitative thresholds. The Full and Lite workflows of *C. difficile* and the two *Cupriavidus gilardii* strains produced subnetworks with distinct architectures while retaining related biological themes. Such findings are consistent with observations that genome assembly decisions can influence downstream genomic representations (Ekblom & Wolf, 2014) and that biologically relevant associations may remain detectable despite variation in genome structure (Zoccarato et al., 2022). Together, these observations suggest that network architecture and functional representation capture complementary aspects of genomic organization.

Sensitivity analyses further supported this interpretation, showing that alternative abundance thresholds and contextual expansion distances altered network size and composition while preserving comparable functional categories across the evaluated parameter range. Accordingly, normalized abundance should be interpreted primarily as a within-sample ranking metric for contextual prioritization rather than as a quantitative measure for cross-sample comparisons.

Limitations and future directions

SELECT-NET generates contextual genomic associations rather than experimentally validated molecular interaction networks. Accordingly, the implemented filtering modes and annotation-derived categories should be interpreted as reproducible heuristic prioritization strategies intended to support contextual exploration and hypothesis generation rather than predictive classification or formal functional inference. Because category assignment depends on user-modifiable keyword dictionaries and homology-based annotations, recovered subnetworks represent biologically informed hypotheses whose interpretation ultimately depends on annotation quality and experimental validation.

Although SELECT-NET is designed for browser-based accessibility and reproducible execution through Google Colab, runtime may vary according to dataset size and the computational demands of integrated external tools. In its current form, the workflow is best suited for accessible end-to-end exploratory analysis rather than high-throughput batch processing.

Several opportunities exist for future development. Preliminary application to an *A. baumannii* RNA-seq dataset indicates that the framework may support transcriptomic exploration, although broader evaluation remains necessary. Additional extensions include GFF-only workflows, alternative abundance metrics, user-provided reference genomes, integration of additional annotation platforms, and graphical interfaces designed to improve accessibility for users with limited bioinformatics experience.

Conclusions

SELECT-NET provides a reproducible framework for contextual gene prioritization and subnetwork reconstruction from bacterial sequencing data. Across laboratory, environmental, and clinically relevant datasets, the workflow generated alternative contextual representations of the same genomic landscape by emphasizing distinct biological perspectives through complementary filtering strategies.

Rather than replacing established assembly, annotation, or functional analysis pipelines, SELECT-NET reorganizes genomic information into interpretable subnetworks that support visualization, comparative analysis, and hypothesis generation. The results suggest that biologically meaningful functional themes may remain detectable across alternative contextual representations despite differences in network architecture, annotation strategy, or reconstruction parameters. In this sense, SELECT-NET provides a practical approach for exploring bacterial genomes through multiple complementary biological perspectives while reducing interpretive complexity and facilitating downstream experimental investigation.

Declarations

Funding

The author received a postdoctoral fellowship from: Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI).

Conflict of Interest

The author declares no competing interests.

Ethics Approval

Not applicable.

Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

Data Availability

All sequencing datasets analyzed in this study are publicly available through the NCBI Sequence Read Archive (SRA) under the accession numbers indicated in Table 1. The complete SELECT-NET source code and executable Google Colab notebook are provided as Online Resource 6. Representative GraphML network files associated with the main case studies are provided as Online Resource 7.

AI use statement

A large language model was used to assist with language refinement and drafting of selected manuscript text. All scientific content, methodological decisions, data interpretation, and final wording was reviewed and approved by the author, who takes full responsibility for the final manuscript.

Author Contribution

G.R.A. conceived the study, developed the SELECT-NET workflow, performed data collection, software implementation, analysis, visualization, interpretation of results, and manuscript preparation. The author reviewed and approved the final manuscript.

Acknowledgement

The author gratefully acknowledges Dr. Gabriela Olmedo Álvarez and the Laboratory of Molecular Biology and Microbial Ecology (Cinvestav-Irapuato) for providing the institutional and academic environment in which this work was developed. The author also thanks Dr. Antonio González-Sánchez for scientific discussions and to the open source bioinformatics community for developing the software and databases that made this work possible.

Data Availability

All sequencing datasets analyzed in this study are publicly available through the NCBI Sequence Read Archive (SRA) under the accession numbers indicated in Table 1. The complete SELECT-NET source code and executable Google Colab notebook are provided as Online Resource 6.Representative GraphML network files associated with the main case studies are provided as Online Resource 7.

References

1. Chen G, Tan Q, Zhou H et al (2026) An integrated computational-experimental approach identifies Malate Synthase G as therapeutic target to disrupt carbon metabolism in chronic Pseudomonas aeruginosa infection. Funct Integr Genomics 26:116. https://doi.org/10.1007/s10142-026-01889-x
2. Dey P, Osborne JW, Kirubhadharsini BL (2026) Exploration of functional genes in Brucella anthropi for hexavalent chromium bioremoval through gene mining: A whole genome sequencing based approach. Ecotoxicol Environ Saf 309:119472. https://doi.org/10.1016/j.ecoenv.2025.119472
3. Dyer NP, Päuker B, Baxter L, Gupta A, Bunk B, Overmann J, Nübel U (2025) EnteroBase in 2025: exploring the genomic epidemiology of bacterial pathogens. Nucleic Acids Res 53(D1):D757–D762. https://doi.org/10.1093/nar/gkae902
4. Ekblom R, Wolf JBW (2014) A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7:1026–1042. https://doi.org/10.1111/eva.12178
5. Enam SU, Cherry JL, Leonard SR, Zheludev IN, Lipman DJ, Fire AZ (2023) Restriction Endonuclease-Based Modification-Dependent Enrichment (REMoDE) of DNA for Metagenomic Sequencing. Appl Environ Microbiol 89(1):e0167022. https://doi.org/10.1128/aem.01670-22
6. Gollapalli P, Gnanasekaran TS, Shivashankara SKH et al (2026) Genome-scale mapping of Serratia marcescens core interactome to probe essential proteins and functional modules. Biologia 81:7. https://doi.org/10.1007/s11756-025-02106-z
7. Holcomb DA, Miko S, Kahler AM, Brown T, Leeper M, McKeel H, Mattioli MC (2026, March) First reported outbreak of persistent Shiga toxin-producing Escherichia coli O157: H7 strain REPEXH01: environmental investigation of untreated recreational water, 2017. Open Forum Infectious Diseases, vol 13. Oxford University Press, US, p ofag132. 3 https://doi.org/10.1093/ofid/ofag132
8. Khew C, Mohd-Fowzi N, Zaifulzaman N et al (2026) Integrative transcriptomic and network analysis reveals small open reading frames associated with cellular adaptive responsees and methylation-linked pathogenicity in Leptospira interrogans. Funct Integr Genomics 26:106. https://doi.org/10.1007/s10142-026-01882-4
9. Liu W, Yang Y, Bian J, Li X, Lu Z (2026) Niche adaptation of marine heterotrophic nitrification–aerobic denitrification bacterium in mariculture wastewater treatment: Synergistic mechanism of nitrogen removal and sulfamethoxazole biotransformation. Water Res 125914. https://doi.org/10.1016/j.watres.2026.125914
10. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Spratt BG (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences, 95(6), 3140–3145. https://doi.org/10.1073/pnas.95.6.3140
11. Maiden MC, Van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND (2013) MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 11(10):728–736. https://doi.org/10.1038/nrmicro3093
12. Ogata H, Fujibuchi W, Goto S, Kanehisa M (2000) A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. Nucleic Acids Res 28(20):4021–4028. https://doi.org/10.1093/nar/28.20.4021
13. Ortiz-Flores CG, Romero-Rodríguez A, Villanueva-Enríquez R, Ceapă CD, Gómez-Rivera N, Velázquez-Romero M … & Martinez de la Peña, C. F. (2026). Comparative Genomic Analysis of Clostridioides difficile Strains in Mexico: Insights into Virulence and Resistance. Access Microbiol, 001132–v2. https://doi.org/10.1099/acmi.0.001132.v2
14. Raddadi N, Cherif A, Daffonchio D, Neifar M, Fava F (2015) Biotechnological applications of extremophiles, extremozymes and extremolytes. Appl Microbiol Biotechnol 99(19):7907–7913. https://doi.org/10.1007/s00253-015-6874-9
15. Ruiz Amores G, Martínez-Antonio A (2022) Basics on network theory to analyze biological systems: a hands-on outlook. Funct Integr Genomics 22:1433–1448. https://doi.org/10.1007/s10142-022-00907-y
16. Shankar S, Vadivel M, Gurusamy R, Sakthivel N (2026) Comprehensive Genomic Characterization of the Marine Isolate Exiguobacterium profundum SW22: Insights into Adaptive Mechanisms and Biosynthetic Potential. Curr Microbiol 83(4):224. https://doi.org/10.1007/s00284-026-04808-2

17. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498. https://doi.org/10.1101/gr.1239303
18. Sosa Reyes D, Amores R (2024) G. ACE-AB1: Automated Comprehensive Editing and BLAST-based Assignment for AB1 Files (V1.0.0). Zenodo. https://doi.org/10.5281/zenodo.14556246
19. Tamura M, D'haeseleer P (2008) Microbial genotype–phenotype mapping by class association rule mining. Bioinformatics 24(13):1523–1529. https://doi.org/10.1093/bioinformatics/btn210
20. Tripathi G (2026) Editorial for Special Issue Technological Advances Around Next-Generation Sequencing. Curr Issues Mol Biol 48(1):83. https://doi.org/10.3390/cimb48010083
21. Zhou Z, Alikhan NF, Mohamed K, Fan Y, Achtman M, Brown D, Agama Study Group (2020) The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. Genome Res 30(1):138–152. https://doi.org/10.1101/gr.251678.119
22. Zoccarato L, Sher D, Miki T et al (2022) A comparative whole-genome approach identifies bacterial traits for marine microbial interactions. Commun Biol 5:276. https://doi.org/10.1038/s42003-022-03184-4

Figures

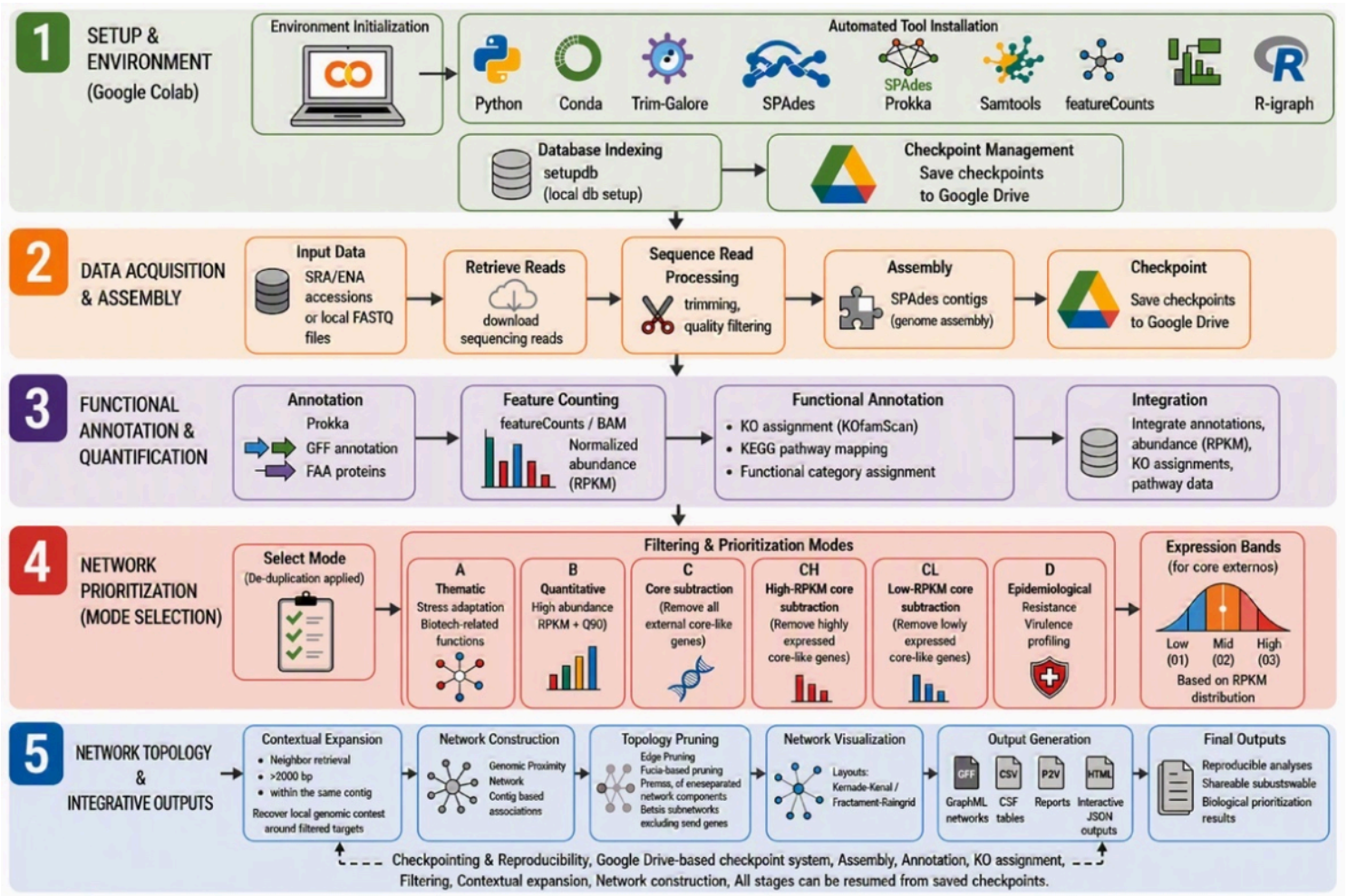


Figure 1

Overview of the SELECT-NET workflow. The pipeline consists of five major stages: (1) environment initialization, resource preparation, and execution in Google Colab; (2) sequencing data acquisition, preprocessing, and assembly; (3) gene prediction, functional annotation, abundance estimation, and KEGG Orthology assignment; (4) contextual filtering through four analytical strategies implemented across six operational modes (A, B, C, CH, CL, and D); and (5) genomic subnetwork

reconstruction, visualization, and export of reproducible outputs, including GraphML and GEXF files compatible with Cytoscape and related network-analysis platforms.

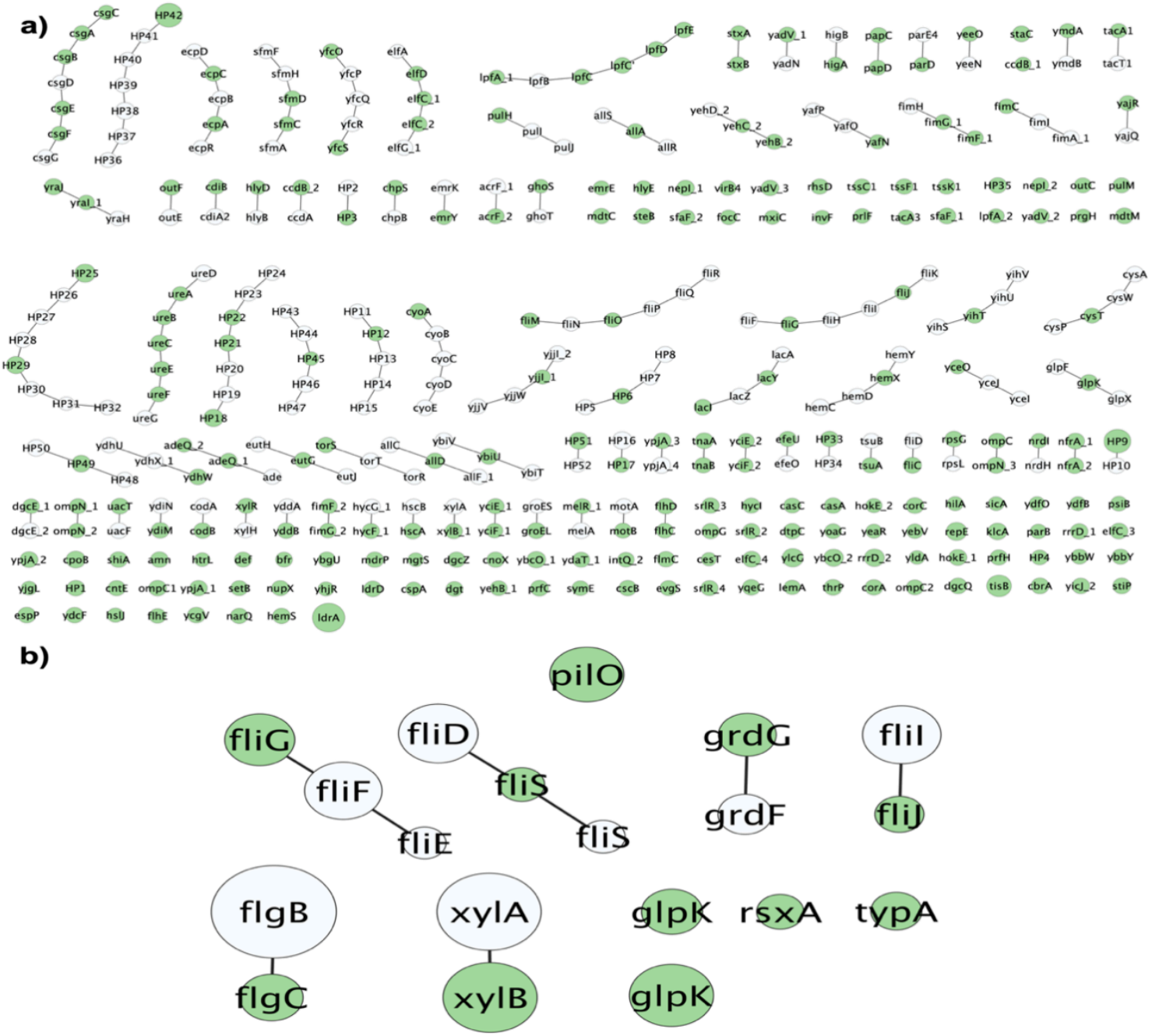


Figure 2

Isolate-specific-genome subnetworks recovered through heuristic core-genome subtraction in STEC-1 and *Clostridioides difficile* isolates. Network representations generated after application of the core-subtraction filter (Mode C) to two genomes. (a) STEC-1 (SRR7739427) using Full workflow of SELECT-NET. (b) *Clostridioides difficile* isolate using Lite workflow of SELECT-NET (b). Green nodes correspond to genes retained after subtraction of their respective external cgMLST reference, representing genomic features absent from the conserved core background. White nodes indicate neighboring genes incorporated through contextual expansion during network reconstruction. Node size represents RPKM-based normalized abundance metric. The resulting subnetworks highlight isolate-specific genomic content while preserving local genomic organization

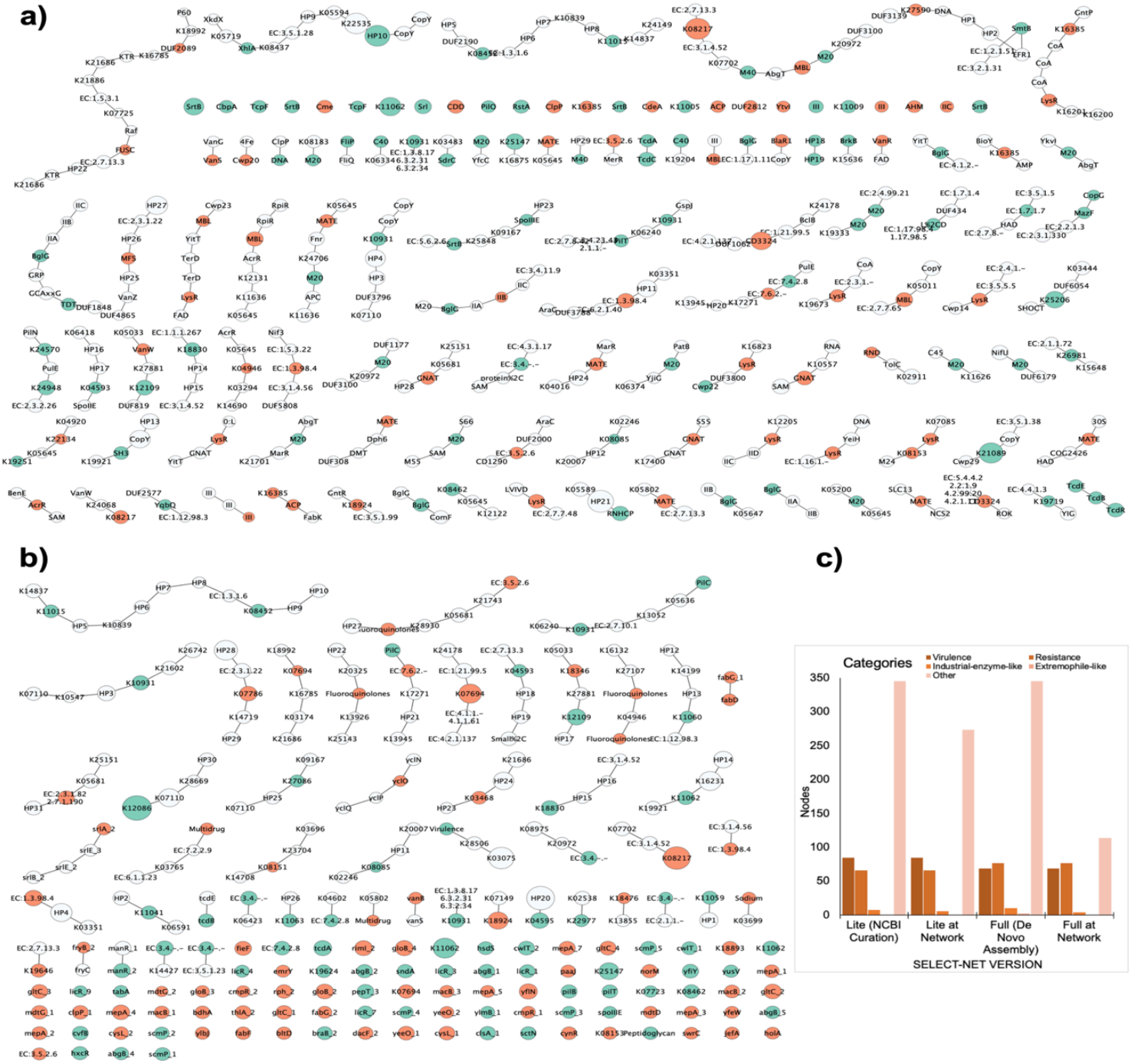


Figure 3

Comparative epidemiological subnetworks generated from *Clostridioides* difficile using Full and Lite workflows (Mode D). Epidemiological subnetworks reconstructed after application of the Mode D filter. (a) Lite workflow using reference-guided annotation. (b) Full workflow using de novo assembly and annotation. Nodes associated with virulence-related functions are shown in turquoise, whereas nodes associated with antimicrobial resistance are shown in orange. White nodes represent contextual neighboring genes retained during network reconstruction. (c) Comparison of node distribution before and after network generation in both workflows. The de novo workflow exhibits a greater reduction of nodes classified as “other” following network reconstruction than the reference-guided workflow, resulting in differences in network size and topology. Node size represents RPKM-based normalized abundance metric. Both workflows recovered comparable categories of virulence- and resistance-associated genes despite differences in node count and network architecture.

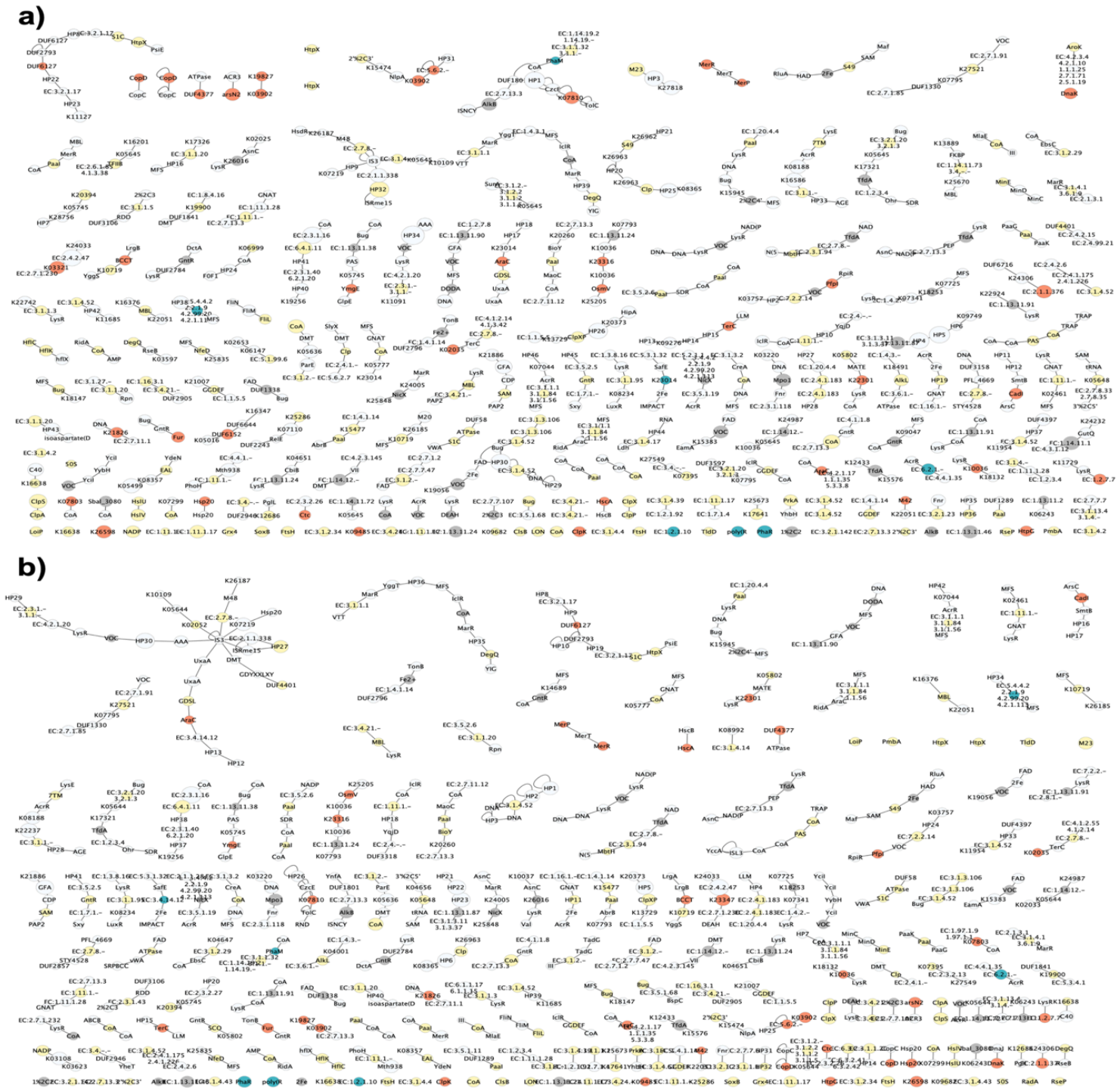


Figure 4

Conserved adaptive functions recovered by thematic filtering in environmental *Cupriavidus gilardii* isolates. Thematic subnetworks generated using Mode A filtering. (a) OV2 isolate. (b) NOV2 isolate. Node colors indicate thematic functional categories: extremophile-associated functions (orange/red), industrial enzyme-related functions (yellow), biodegradation-associated functions (gray), and bioplastic-related functions (turquoise/blue). Node size represents RPKM-based normalized abundance metric. White unlabeled neighboring nodes were retained through contextual expansion to preserve local genomic organization. Both strains recovered similar adaptive functional categories but differed in network connectivity and node organization.

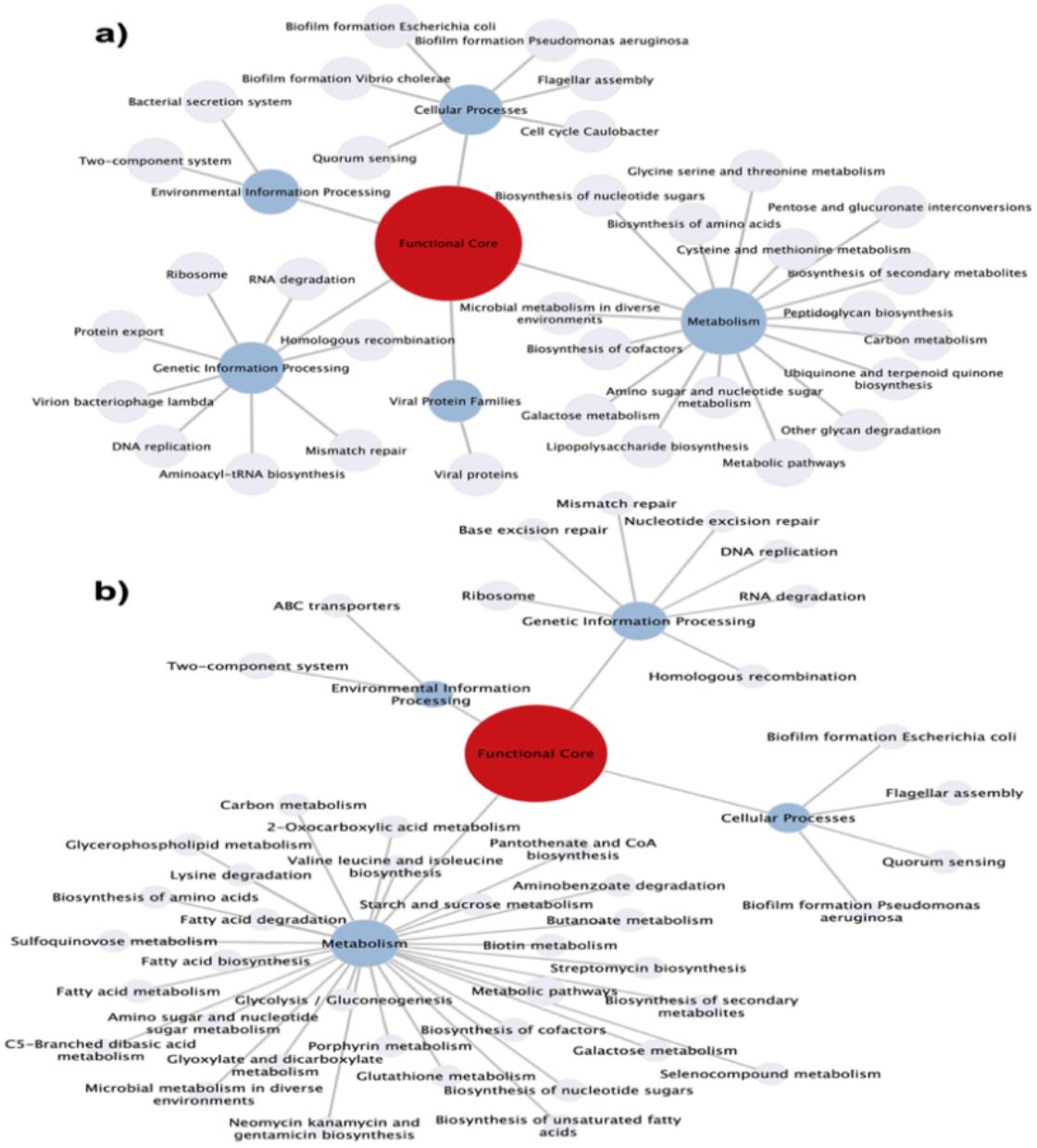


Figure 5

Functional organization of the conserved Mode B nodes recovered across abundance thresholds. (a) STEC-1 and (b) Cupriavidus gilardii OV2. KEGG reconstruction of KO assignments shared among Q0.85, Q0.90 and Q0.95 identified conserved genomic cores dominated by metabolism, information processing, environmental sensing, and community-associated functions. Lower-stringency thresholds retained additional functional categories beyond this conserved core. Node sizes are proportional to the number of KEGG Orthology assignments associated with each category or pathway. Red nodes represent the conserved functional core, blue nodes represent major functional categories, and gray nodes represent KEGG pathways associated with the reconstructed core.

Supplementary Files

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

- EMS1.docx
- EMS5.docx

- EMS3.docx
- EMS7.docx
- EMS6selectnet.py
- EMS2.docx
- EMS4.docx