Pangenomic Identification of Metabolic Switches in *Kosakonia cowanii*: A Consolidated Network Biology Approach

Davi Mota Campos[3†], Gustavo Henrique Leal da Fonseca[3†], Arthur Rodrigues Cândido[3†], Nilvanira Donizete Tebaldi[1], Lara Caroline Borges Moreira Mota[1], Joberth Lee Corrêa[2], Ana Carolina Costa Santos[2], Carlos Ueira Vieira[2], Anderson Rodrigues dos Santos[3*]

[1] Laboratório de Bacteriologia Vegetal - Instituto de Ciências Agrárias, Universidade Federal de Uberlândia - Uberlândia, Brazil

[2]Laboratory of Genetics - Institute of Biotechnology Universidade Federal de Uberlândia - Uberlândia, Brazil.

[3]Faculty of Computer Science - Universidade Federal de Uberlândia Uberlândia, Brazil.

[†] These authors contributed equally to this work.

[*] Corresponding author: santosardr@ufu.br

Abstract. *Kosakonia cowanii acts as both a plant growth-promoting rhizobacterium (PGPR) and an opportunistic pathogen. To characterize this lifestyle transition, we built a pangenomic pipeline across 77 genomes, consolidating strain-specific interactomes into a species-level graph. Biological coherence filters discarded 12,000 false-positive orthology mappings. Applying Bridging Centrality (BriCe) with functional bridging criteria identified 46 conserved metabolic switches at the PGPR-pathogen interface. These switches, including flagellar regulators and MFS transporters, are tractable targets for experimental validation.*

1. Introduction

The boundary between mutualism and pathogenesis in bacteria is not fixed. Several species deploy the same genomic repertoire to promote plant growth under favorable conditions and to cause disease when host defenses are compromised—a pattern notably widespread in the *Enterobacteriaceae* [Eberl and Vandamme 2016], where shared regulatory circuits govern both outcomes [Venturi 2006].

Kosakonia cowanii exemplifies this duality. Reclassified within the genus *Kosakonia* via multilocus sequence analysis [Brady et al. 2013], the species fixes nitrogen, solubilizes phosphate, and produces siderophores in the rhizosphere [Berg 2014]. The same strains, however, have been isolated from bacteremic patients and, more recently, identified as the agent of bacterial blight in *Coffea arabica* in Brazil [Tebaldi et al. 2025, Stover et al. 2000].

Topological analysis of Protein-Protein Interaction (PPI) networks offers good insight on the internal workings of the bacterial cell, allowing us to better understand what exactly causes certain behavior [Barabási and Oltvai 2004]. Such analysis however is not trivially done, PPI network predictions [Silva et al. 2025] of the four *K. cowanii* strains produced up to 10^5 edges, at this scale, manual analysis becomes unfeasible.

By using various consolidated tools we leverage a computational pipeline using PostgreSQL to handle genome annotation data, while ArangoDB handles network consolidation and graph analytics [Angles 2017]. The two different databases, bridged by pangenomic identifiers (panid), allows us to take advantage of the powerful query capabilities of PostgreSQL and graph storage of ArangoDB while minimizing their disadvantages.

A pangenomic approach is used by merging individual interactomes into one unified graph [Tettelin et al. 2005]. Within that graph, nodes bridging otherwise separate functional modules—PGPR machinery on one side, virulence machinery on the other—are the most informative targets. Bridging Centrality [Hwang et al. 2008] identifies precisely these nodes by combining betweenness centrality with a bridging coefficient. Candidates are then stratified using network cartography [Guimera and Amaral 2005] and a functional typology framework [Peregrín-Alvarez et al. 2011].

2. Materials and Methods

The pipeline (Fig. 1) runs from genome annotation through interactome prediction to topological scoring (Fig. 2).

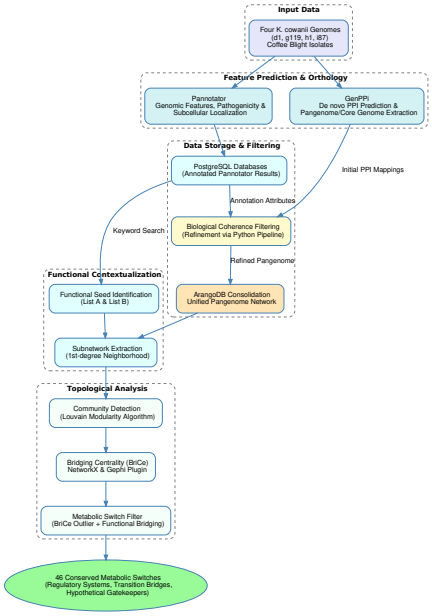

Figure 1. Analytical pipeline overview.

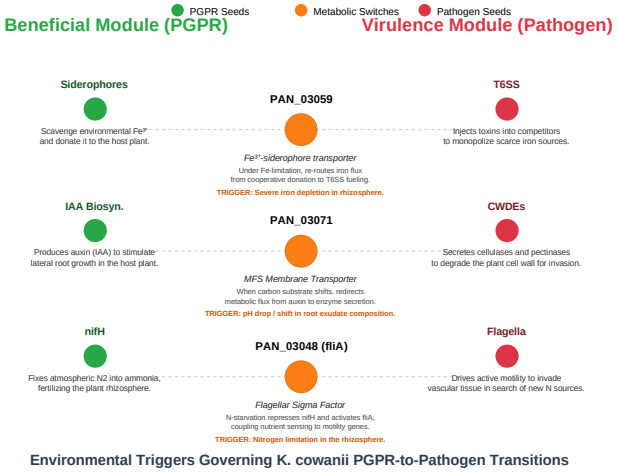

Figure 2. Topological clusters and switches.

2.1. Pangenome Construction and Coherence Filtering

Four target *K. cowanii* strains (d1, g119, h1, i87) were each run through GenPPi [Silva et al. 2025] against a shared panel of 73 reference genomes, dominated by *Enterobacter* (19), *Citrobacter* (13), *Cronobacter* (7), and *Escherichia* (5). This yields four independent orthologous group (OG) reports, which together define the species pangenome. Before network construction, each predicted similarity edge was retained only if two conditions held simultaneously: (1) the CDS size difference between the two

proteins did not exceed 20% (Size Consistency); and (2) predicted pathogenicity and subcellular localization were compatible (Functional Consistency), as scored by Pannotator [Gonçalves and Santos 2025]. Of 101,606 candidate edges, 11,996 (11.8%) were removed. The core and soft-core pangenome—OGs present in at least three of the four strains—comprised 3,616 nodes.

2.2. Seed Identification and Network Consolidation

The four strain-level PPI networks were loaded into a unified ArangoDB database [Angles 2017], with each interaction mapped to its (*panid*). ArangoDB was chosen for this stage because neighbourhood expansion and centrality computation are native graph operations that its query engine handles via traversal, whereas equivalent operations in a relational model would require recursive self-joins over an adjacency list.

Seed proteins were retrieved via keyword queries against the Pannotator-populated PostgreSQL database with CDS features, predicted subcellular localizations. **List A (PGPR)** covers siderophore receptors and reductases, phosphate solubilization, *nif* cluster products, and IAA biosynthesis enzymes. **List B (Pathogenicity)** covers type III, IV, and VI secretion components (*vgrG*, *hcp*), pectinases, cellulases, biofilm regulators, and pili. A first-degree neighborhood subnetwork around all seed *panids* was then extracted from ArangoDB.

2.3. Community Detection and Bridging Centrality

Louvain modularity [Blondel et al. 2008] was applied first to test whether PGPR and pathogenicity seeds partition into separate communities. BriCe [Hwang et al. 2008, Pereira et al. 2021] was then computed for every node in the subnetwork:

$$BriCe(v) = C_B(v) \times \frac{k_v^{-1}}{\sum_{u \in N(v)} k_u^{-1}} \tag{1}$$

Where $C_b(v)$ is betweenness centrality and k_u is the degree of neighbour *u*.

A node was considered a metabolic switch only when the extreme-outlier threshold ($Q3 + 3 \times IQR$) *and* its first-degree neighborhood contained at least one seed from List A and one from List B. Confirmed switches were then classified into a functional typology—regulatory systems, transition bridges, or hypothetical gatekeepers—based on domain annotation and neighborhood connectivity [Peregrín-Alvarez et al. 2011].

3. Results

The filtered subnetwork contained 2,584 nodes and 45,706 edges. Louvain analysis returned a modularity score of $Q = 0.6943$ across 237 communities; PGPR and pathogenicity seeds were predominantly non-overlapping, which validates the topological premise for a switch-hunting approach. The pipeline returned 46 conserved pangenomic metabolic switches (Table 1).

Three functional categories account for most switches. **Regulatory systems** (e.g., PAN_03048, *fliA*) are signal transduction proteins whose position links nitrogen fixation clusters directly to flagellar assembly—network hubs in the sense of Jeong et al. [Jeong et al. 2001]. **Transition bridges** (e.g., PAN_03059) are transporters with one foot

Table 1. Leading Metabolic Switches Ranked by Bridging Centrality (BriCe) Score.

PanID	BriCe Score	Functional Annotation (Target Gene)
PAN_02203	1.43×10^{-4}	Nucleoside transporter
PAN_03059	1.39×10^{-4}	Iron-siderophore ABC transporter binding protein
PAN_01009	9.34×10^{-5}	DUF1158 family protein
PAN_02205	6.76×10^{-5}	*putP* (Sodium/proline symporter)
PAN_03049	6.68×10^{-5}	Winged helix-turn-helix domain protein
PAN_03048	4.27×10^{-5}	*fliA* (Flagellar sigma factor)
PAN_03069	3.72×10^{-5}	Hypothetical protein
PAN_03071	3.02×10^{-5}	MFS transporter

in a PGPR module and one in a virulence module; the iron-siderophore ABC transporter, for instance, sits between iron-scavenging and T6SS subgraphs. **Hypothetical gatekeepers** (e.g., PAN_03069) are highly central proteins with no current functional annotation, and therefore represent uncharacterized points of control.

4. Discussion

In the *Enterobacteriaceae*, lifestyle switching rarely requires horizontal gene transfer. The 46 switches identified here map onto three specific environmental triggers. Iron depletion below a rhizosphere threshold [Andrews et al. 2003] rewires the network around PAN_03059: under replete conditions, the siderophore transporter supports cooperative iron acquisition, but under starvation it channels iron flux toward T6SS-dependent interbacterial competition. Carbon source availability operates via PAN_03071, an MFS transporter bridging IAA biosynthesis and CWDE secretion; a shift in root exudate chemistry can redirect metabolic output from auxin toward pectinases and cellulases, enabling vascular ingress [Saier Jr et al. 2016]. Nitrogen limitation acts through *fliA*: when *nifH*-dependent fixation becomes energetically untenable, the sigma factor repurposes the regulatory node to activate flagellar genes and motility [Venturi 2006].

What makes these proteins tractable intervention points is not simply their topological score but the specificity of their bridging: they connect modules with opposed ecological outcomes, so perturbing them in isolation would be expected to shift the phenotypic balance without dismantling the underlying metabolic core. Whether this holds experimentally, and whether similar switches operate in the host-associated strains that cause coffea blight, remains an open question.

The dual-database method used in the pipeline is a deliberate design choice, by separating the intrinsically relational nature of genome annotation on PostgreSQL from the graph-oriented PPI pangenome on ArangoDB, we achieve the best of both worlds: SQL's native querying of large amounts of data and ArangoDB's native graph traversal queries. The architecture presented is not dependent on any organism, any other possible metabolic switch can be investigated by following similar procedures.

References

Andrews, S. C. et al. (2003). Iron homeostasis in bacteria. *FEMS microbiology reviews*, 27(2-3):215–237.

Angles, R. (2017). Arangodb-a graph database for the next generation. *Proceedings of the 32nd Annual ACM Symposium on Applied Computing*, pages 877–882.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113.

Berg, G. (2014). Plant growth-promoting bacteria. *Frontiers in microbiology*, 5:491.

Blondel, V. D. et al. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Brady, C. et al. (2013). Taxonomic evaluation of the genus enterobacter based on multilocus sequence analysis (mlsa). *Systematic and applied microbiology*, 36(5):309–319.

Eberl, L. and Vandamme, P. (2016). Members of the genus burkholderia: good and bad guys. *F1000Research*, 5.

Gonçalves, R. and Santos, A. (2025). Pannotator integrated with medpipe provides immunological and subcellular location features using a microservice. *Biomedical Informatics*, 1(1).

Guimera, R. and Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900.

Hwang, W. et al. (2008). Bridging centrality: identifying bridging nodes in scale-free networks. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 336–344.

Jeong, H. et al. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.

Peregrín-Alvarez, J. M. et al. (2011). Protein function prediction. *Methods in molecular biology*, 702:227–244.

Pereira, G., Ghosh, P., and Santos, A. (2021). A bridging centrality plugin for gephi and a case study for mycobacterium tuberculosis h37rv. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(6):2741–2746.

Saier Jr, M. H. et al. (2016). The transporter classification database (tcdb): recent advances. *Nucleic acids research*, 44(D1):D372–D379.

Silva, A., Marquez, C., Godoy, I., et al. (2025). Improving protein interaction prediction in genppi: a novel interaction sampling approach preserving network topology. *BMC Bioinformatics*, 26:296.

Stover, C. et al. (2000). Complete genome sequence of pseudomonas aeruginosa pao1. *Nature*, 406:959–964.

Tebaldi, N. D. et al. (2025). First report of kosakonia cowanii causing bacterial blight on coffea arabica in brazil. *Plant Disease*.

Tettelin, H. et al. (2005). Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial pan-genome. *Proceedings of the National Academy of Sciences*, 102(39):13950–13955.

Venturi, V. (2006). Regulation of bacterial virulence genes by quorum sensing and beyond. *FEMS microbiology reviews*, 30(2):274–291.