

Research Article

In Silico Profiling of Small Open Reading Frames in *hsdM*-Deficient Group A *Streptococcus* and Their Predicted Role in Membrane Adaptation

Chung Yuen Khew¹, Nur Afqah Mohd Azali², Nurnazatul Syiffa Muhammad Bahrain², Norfarhan Mohd-Assaad^{1,3*}

¹Department of Applied Physics, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.

²Department of Biosciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.

³Institute of Systems Biology (INBIOSIS), Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.

*Corresponding author: n_farhan@ukm.edu.my

ABSTRACT

Small open reading frames (sORFs) are emerging as key players in bacterial gene regulation, yet their roles remain underexplored in many pathogens. This study investigates the regulatory landscape of sORFs in Group A *Streptococcus* (GAS) by integrating transcriptomic data and weighted gene co-expression network analysis (WGCNA). We focused on the functional consequences of inactivating the Type 1 Restriction-Modification (RM) system, specifically the *hsdM* subunit, which is a known barrier to genetic transformation in GAS. By reanalysing the transcriptomic profile of the hyper-transformable 854Δ*hsdM* strain, we identified 28 sORFs localised within significantly altered co-expression modules. Notably, sORF_00369 (associated with fatty acid biosynthesis) and sORF_00282 (linked to ABC transport) were significantly downregulated. These findings suggest that the inactivation of the RM system, while improving transformation efficiency, may trigger a targeted shift in sORF-mediated metabolism and membrane integrity. Our results propose that sORFs are candidate regulators within the epigenetically mediated regulatory networks of GAS, providing new insights into how genetic recalcitrance and metabolic fitness might be linked. This *in silico* study lays a theoretical foundation for potentially targeting sORFs in antimicrobial strategy development and optimising GAS as a chassis for synthetic biology.

Key words: Bacterial transformation, Group A *Streptococcus*, RNA-Seq, Small open reading frame, *Streptococcus pyogenes*, WGCNA

Article History

Accepted: 8 September 2026

First version online: 28 September 2026

Cite This Article:

Khew, C.Y., Mohd Azali, N.A., Muhammad Bahrain, N.S. & Mohd-Assaad, N. 2026. *In silico* profiling of small open reading frames in *hsdM*-deficient Group A *Streptococcus* and their predicted role in membrane adaptation. Malaysian Applied Biology, 55(3): 159-167. <https://doi.org/10.55230/mabjournal.v55i3.3857>

Copyright

© 2026 Malaysian Society of Applied Biology

INTRODUCTION

Small open reading frames (sORFs) are short gene sequences, typically less than 300 nucleotides in length, that encode proteins smaller than 100 amino acids (Leong *et al.*, 2022). Although often overlooked due to their size, unconventional start codons (CUG, GUG, UUG), and overlapping positions with other genes, sORFs are now recognised as important regulators in both prokaryotic and eukaryotic systems (Lu *et al.*, 2023).

In bacterial pathogens, sORFs occur in diverse genomic regions, including 5' and 3' untranslated regions (UTRs), intergenic spaces, and non-coding RNA segments (Jain *et al.*, 2023). Advances in next-generation sequencing (NGS), ribosome profiling, and computational pipelines have enabled the systematic detection of these elements, revealing their involvement in stress adaptation and metabolic regulation (Garai & Blanc-Potard, 2020). sORF-encoded peptides (SEPs) can act by stabilising membrane complexes, modu-

lating the activity of larger proteins, or functioning as signalling switches that fine-tune transcriptional responses (Ko *et al.*, 2025).

Despite this expanding toolkit, the role of sORFs in the context of epigenetic regulation and horizontal gene transfer (HGT) remains underexplored. In *Streptococcus pyogenes* (Group A *Streptococcus*; GAS), a major human pathogen responsible for diseases ranging from pharyngitis to life-threatening necrotising fasciitis, virulence and fitness are tightly controlled by complex regulatory networks (Esposito *et al.*, 2025). Central to this regulation are Type I Restriction-Modification (R-M) systems, such as the *hsdRSM* locus, which governs DNA methylation patterns and protects the cell against invading foreign DNA (Finn *et al.*, 2021).

Recent evidence suggests that the DNA methyltransferase *hsdM* serves as a regulatory hub, whereby its disruption in *S. pyogenes* strain 854 has been shown to increase electrotransformation efficiency fivefold without significantly altering the expression of ma-

for off-target genes (Finn *et al.*, 2021). However, the impact of such epigenetic modifications on the small-protein coding landscape is currently unknown. Given that sORFs often act as rapid-response regulators to environmental stimuli, we asked whether the disruption of *hsdM* alters sORF expression patterns and whether these sORFs are embedded within co-expression modules associated with competence, membrane permeability, or stress adaptation.

To investigate the regulatory link between *hsdM*-driven methylation and the sORF landscape, we integrated RNA-Seq transcriptomic profiling with weighted gene co-expression network analysis (WGCNA) of GAS. GAS remains a primary focus of genomic research due to its significant public health burden and its inherent genetic recalcitrance (a resistance to foreign DNA uptake that hinders molecular study). However, the broader impact of this loss of DNA methylation on the “dark matter”, specifically sORFs, has not been fully elucidated.

This study extends the work of Finn *et al.* (2021) by employing an exploratory, computational systems biology approach to identify sORFs that are functionally enriched within the 854 Δ *hsdM* background. We aim to uncover how these sORFs computationally shape the transcriptional landscape and phenotypic outcomes of GAS, highlighting their potential as novel therapeutic targets (Finn *et al.*, 2021).

MATERIALS AND METHODS

Materials

Reference genome and annotation

The reference genome and primary annotation for GAS, *Streptococcus pyogenes* serotype M1 (strain MGAS5005), were obtained from the NCBI RefSeq database under accession GCF_000011765.3. This dataset corresponds to the ASM1176v2 assembly, which was originally submitted by the Lab of Human Bacterial Pathogenesis. To maintain high-resolution genomic context, the latest annotation release from October 2024 was retrieved via the NCBI FTP directory. These reference files provided the structural framework for defining intergenic and antisense regions during the integration of transcriptomic features. The use of a consistent RefSeq version ensures that genomic feature classifications remain reproducible and free from discrepancies introduced by disparate annotation versions. No further genome polishing or long-read re-annotation was conducted for this study.

Small ORFs prediction and annotation

To investigate the small open reading frame (sORF) landscape in GAS, we utilised RNA-Seq datasets under accession GSE167217 from the NCBI Gene Expression Omnibus (GEO). This dataset includes biological replicates for the wild-type MGAS5005 strain and its corresponding Δ *hsdM* mutant. The MGAS5005 reference genome (GCF_000011765.3) served as the template for all mapping procedures. Initial sequence quality was evaluated with FastQC v1.17 (Andrews, 2010) and consolidated via MultiQC (Ewels *et al.*, 2016). Trimmomatic v0.39 was employed in single-end mode for adapter removal and quality trimming, utilising a sliding window of 4:15 and a minimum length threshold of 30 nucleotides (Bolger *et al.*, 2014).

Cleaned reads were mapped to the GAS assembly using the READemption v2.0.4 pipeline, which is specifically configured for bacterial transcriptomic efficiency (Förstner *et al.*, 2014). Subsequent transcript assembly and the prediction of sORFs and small RNAs (sRNAs) were performed using ANNOgesic v1.0.22 (Yu *et al.*, 2018). We implemented the -print_all_combination setting to ensure an exhaustive search for candidate ORFs across all reading frames. Functional annotation of the identified sORF pool was conducted using BLAST2GO (Conesa & Götz, 2008) and InterProScan (Jones *et al.*, 2014), allowing for the assignment of Gene Ontology terms and the identification of conserved protein domains.

Transcriptome analyses

Alignment of trimmed paired-end reads was performed using STAR v2.7.11a with strand-specific parameters (Dobin *et al.*, 2013). The genome index was generated by integrating the original NCBI RefSeq annotation with the transcriptomic features identified by ANNOgesic. Following alignment, BAM files were sorted and indexed using SAMtools v1.0 (Li *et al.*, 2009). Gene-level quantification was carried out with HTSeq-Count in paired-end mode, applying the union counting method to ensure precise read assignment. Differential expression analysis was executed in R using the edgeR v4.2.0 package (Robinson *et al.*, 2009). We compared the wild-type and Δ *hsdM* mutant profiles to identify shifts in the small-protein coding landscape, setting significance thresholds at a false discovery rate below 0.05 and an absolute log₂ fold change of 1.0 or greater.

Construction and functional analysis of sORF-associated co-expression networks

We used weighted gene co-expression network analysis (WGCNA) to infer the functional roles of predicted sORFs based on their expression correlations with annotated genes. Genes showing significant differential expression (adjusted $p < 0.05$) were prioritised for network integration. Prior to construction, data quality was assessed through hierarchical clustering and Principal Component Analysis. To preserve the statistical power necessary to detect subtle changes in the Δ *hsdM* mutant, we retained all biological replicates. A static tree-cutting threshold of 1.5 million was applied to ensure the inclusion of all relevant transcriptomic variance within the clusters.

Network construction was performed in R using the WGCNA package (Langfelder & Horvath, 2008). We tested a range of soft-thresholding powers and selected $\beta = 12$ to achieve a scale-free topology for the unsigned network. Modules were identified using the cutreeDynamic function with a minimum module size of 30. The resulting network was visualised in Cytoscape v3.10.2 (Shannon *et al.*, 2003), and highly connected clusters were identified via the MCODE plugin. Candidate hub-like sORFs were identified using the Maximal Clique Centrality (MCC) algorithm in CytoHubba (Chin *et al.*, 2014), with the top ten scoring nodes designated as hubs. Finally, functional enrichment was determined using ClueGO (Bindea *et al.*, 2009) to link these modules to specific Gene Ontology and KEGG pathways, using a two-sided hypergeometric test with Benjamini-Hochberg correction to ensure statistical rigour.

RESULTS

Identification and functional characterisation of sORFs in GAS

By implementing the ANNOgesic framework (Yu *et al.*, 2018), we performed a systematic survey of the GAS (NC_007297.2) genomic landscape, initially identifying a pool of 2,686 potential sORF sequences (Table 1). To refine this broad dataset into a high-confidence set of 472 “Best sORFs”, we applied a multi-tiered filtration strategy. This process prioritised transcripts under 300 nucleotides (≤ 100 amino acids) and required the presence of a robust Ribosomal Binding Site (RBS) to support translational viability. Our validation process focused on confirming the existence of conserved prokaryotic translation signals, specifically Shine-Dalgarno sequences, and correctly oriented start/stop codons within previously unannotated expressed regions (Giess *et al.*, 2017; Fuchs & Engelmann, 2023).

The spatial distribution of these 472 high-confidence sORFs reveals a stark architectural bias: 96.8% ($n = 457$) are located on the antisense strand relative to annotated genes, while only 3.2% ($n = 15$) reside within intergenic spaces. This heavy antisense enrichment suggests a profound potential for cis-regulatory interactions within the GAS genome. Furthermore, we intentionally retained 54 sORFs (11.4%)

that intersect with predicted small non-coding RNA (sRNA) coordinates.

Rather than excluding these overlaps, we categorised them as potential dual-function RNAs. This approach is grounded in the understanding that bacterial sRNAs often serve as “protein-coding reservoirs” that modulate physiological transitions and epigenetic responses (Aoyama & Storz, 2023). By preserving these candidates, we ensured that the co-expression network analysis captured the full breadth of the GAS regulatory response, particularly those low-abundance elements that may be sensitive to *hdsM*-mediated methylation changes. Comprehensive details regarding the genomic coordinates, orientation, and RBS characteristics for the entire sORF library are provided in Supplementary Table S1.

Sequence-based homology searches using BLAST2GO (BLASTX) against the NCBI non-redundant (nr) database yielded significant alignments for 175 (37.1%) of the 472 high-confidence sORF candidates. Furthermore, InterProScan v5 (incorporating TMHMM, PHOBIUS, and HMMER-based profiling) identified conserved domains or structural motifs in 71 (40.6%) of the 175 sORFs with homology. Despite these structural insights, 297 candidates (62.9%) lacked detectable sequence homology in the nr database. This high proportion of “orphan” sORFs is consistent with the rapid evolutionary trajectory and taxonomic restriction often observed in bacterial microproteins (Roginski *et al.*, 2026). These sequences likely represent lineage-specific elements that, while lacking conserved domains, may still perform critical regulatory roles within the GAS-specific niche (Simoens *et al.*, 2023). Even in the absence of homology, the presence of predicted signal peptides or transmembrane helices in these “hypothetical” sORFs provides functional plausibility for their involvement in local environmental sensing or membrane-linked signalling (Yuan *et al.*, 2025). Comprehensive functional data, including GO terms and domain signatures for all 472 candidates, are detailed in Supplementary Table S2.

To differentiate genuinely novel sORFs from unannotated gene fragments, we compared predicted amino acid lengths against canonical UniProt sequences, stratifying the 175 homology-supported candidates into three functional classes (Table 2). Most sequences (n=146) were classified as pseudogene fragments, such as sORF_00000 was identified as a 43-aa fragment of the DNA polymerase III subunit beta (canonical length 370 aa), while sORF_00023

represents a 27-aa remnant of acetate kinase (canonical length 400 aa). Conversely, a robust subset of 25 genuine novel sORFs was identified, maintaining structural integrity within typical small protein ranges (30–150 aa), including sORF_00025 (a 55-aa hypothetical protein) and MarR-family transcriptional regulators (sORF_00015). Additionally, four unannotated canonical genes were recovered, such as the 50S ribosomal protein L14 (sORF_00011; 85 aa), which was previously overlooked in the GAS (*S. pyogenes*) reference. These results highlight that the GAS genome contain evidence of pseudogene decay within the GAS genome, a substantial repertoire of previously uncharacterized, small, and potentially regulatory proteins remains transcriptionally active. Detailed length-comparison metrics and classification for the 175 candidates are available in Supplementary Table S3.

Integrated validation and sORF discovery in GAS *hdsM* transcriptome

To ensure the integrity of our discovery pipeline, we first validated our differential expression (DE) analysis against the benchmark results reported by Finn *et al.* (2021) for the GAS 854Δ*hdsM* mutant. Our reanalysis identified 14 significantly altered transcripts (FDR < 0.05), comprising 12 downregulated and two upregulated genes. Comparison with the original study revealed a shared core of metabolic and stress-response genes, including *amyB*, *lacA/B.2*, and *grpE* (Table 3). Specifically, we confirmed that the deletion of *hdsM* did not result in a broad “transcriptional rewiring” of the CsrRS regulon, as previously reported. This consistency validates our computational pipeline for the subsequent identification of non-canonical transcripts.

However, our pipeline diverged from the original findings by identifying a specialised subset of DE sORFs and non-canonical transcripts not prioritised in the original study. Notably, while Finn *et al.* (2021) reported modest upregulation in carbohydrate metabolism loci, our study observed a downward trend in several shared metabolic genes. Specifically, the pioneer study utilised bowtie2 (v2.3.1) for read alignment and DESeq2 (v1.20.0) for differential expression analysis. In contrast, our pipeline employed STAR (v2.7.11a) and edgeR (v4.2.0). These fundamental algorithmic differences in read handling and statistical dispersion modelling, alongside the use of substantially updated software versions, are known to significantly influence fold-change estimation and expression directionality. Beyond these shared markers, we identified novel putative regula-

Table 1. Genomic distribution and high-confidence filtering of the GAS sORF repertoire. Initial candidates of 2,686 sORFs based on RBS presence and antisense orientation, resulting in 472 high-confidence sORFs curated to capture both intergenic and dual-function regulatory elements within the Δ*hdsM* transcriptome.

Category	Total Candidates	With RBS	No sRNA Overlap	Best sORFs (Final Filtered)	% of Total Best sORFs
All sORFs	2,686	469	2,376	472	100%
Intergenic	48	15	16	15	3.18%
Antisense	2,638	454	2,360	457	96.82%

Table 2. Stratification of identified sORFs in GAS based on homology and canonical length.

Category	Total Count (n=175)	Biological Interpretation
Pseudogene Fragment	146	Truncated remnants of large enzymatic complexes, transporters, or metabolic proteins; likely non-functional products of genomic erosion.
Genuine Novel sORF	25	Bona fide small proteins maintaining structural integrity; includes lineage-specific signalling peptides and regulatory toxins.
Unannotated Canonical Gene	4	Established small proteins (e.g., ribosomal factors) previously missed by standard annotation filters due to small size.

tory candidates unique to this study, such as sORF_00369 (logFC -1.47), associated with the acetyl-CoA carboxylase Beta-subunit complex, and sORF_00281 (logFC -1.929), a putative glycoside hydrolase. Despite its high sequence conservation, the short length of sORF_00281 (57 nt) likely precluded its detection in previous annotations. Additional unannotated but differentially expressed sORFs further delineate the sensitivity of our pipeline in capturing the “hidden” transcriptome of GAS (e.g. sORF_00282, sORF_00357, and sORF_00403).

Systems-level integration of sORFs via weighted gene co-expression network analysis (WGCNA)

To elucidate the higher-order regulatory context of the GAS non-canonical transcriptome, a weighted gene co-expression network analysis (WGCNA) was conducted (Figure 1). Initial Sample clustering confirmed that despite the low sample size, the transcriptomic profiles maintained sufficient consistency for network construction (Figure 1a). To ensure the network adhered to a biologically relevant scale-free topology, a soft thresholding power of 12 was selected,

marking the transition to a stability plateau in mean connectivity (Figure 1b).

The resulting network partitioned the transcriptome into 38 distinct co-expression modules (Figure 1c). The mean connectivity across these 38 modules ranges from 30 to 110, indicating a network density sufficient for identifying central hub-like sORFs. Correlation analysis between these modules and the $\Delta hsdM$ trait identified specific sub-networks associated with the epigenetically deficient phenotype (Figure 1d). Out of the 38 network modules, seven (‘black’, ‘darkmagenta’, ‘green’, ‘greenyellow’, ‘grey60’, ‘magenta’, and ‘yellow’) were identified as biologically significant based on topological stability and sORF localisation.

Functional enrichment via ClueGO revealed that sORFs are not randomly distributed but are sequestered within modules governing various biological processes. Specifically, the ‘black’ module was primarily associated with protein folding, while the ‘darkmagenta’ module was linked to translation and ribosomal subunit biogenesis. The ‘green’ module was enriched for ribosomal subunit formation.

Table 3. Comparative differential expression profile of the GAS 854 $\Delta hsdM$ transcriptome. Comparative analysis of shared and unique differentially expressed transcripts, highlighting the identification of novel non-canonical sORFs and the observed metabolic downward trend relative to previously reported benchmark data.

Category	Locus Tag	Gene	Function	Finn <i>et al.</i> [logFC]	This Study [logFC]
i) Shared in both studies	M5005_Spy_1066 / M5005_RS05285	<i>amyB</i>	Alpha-amylase	+1.277	-1.38
	M5005_Spy_1634 / M5005_RS08100	<i>lacF</i>	PTS lactose IIA	+2.048	-2.07
	M5005_Spy_1638 / M5005_RS08120	<i>lacA</i>	Galactose-6-P isomerase	+1.234	-1.33
	M5005_Spy_1637 / M5005_RS08115	<i>lacB.2</i>	Galactose-6-P isomerase	+1.442	-1.35
	M5005_Spy_1499 / M5005_RS07425	<i>grpE</i>	Hsp-70 cofactor	+1.421	-1.70
	M5005_Spy_T0010 / M5005_RS00120	tRNA	Transfer RNA	-1.456	+1.10
	M5005_Spy_1067	<i>malX</i>	Maltose-binding protein	+1.197	-
	M5005_Spy_1635	<i>lacD.2</i>	Tagatose aldolase	+1.657	-
	M5005_Spy_1762	<i>groES</i>	Co-chaperonin	+1.191	-
ii) Finn <i>et al.</i>	M5005_Spy_1636	<i>lacC.2</i>	Putative galactose-6-phosphate isomerase (C subunit)	+1.408	-
	M5005_Spy_1399	PTS	PTS system galactose-specific transporter subunit IIC	+1.179	-
	M5005_Spy_T0008	-	tRNA	-1.456	-
	sORF_00281	-	-	-	-1.929
iii) This study	sORF_00282	-	-	-	-1.684
	sORF_00357	-	-	-	-1.222
	sORF_00369	-	-	-	-1.47
	sORF_00403	-	-	-	-2.071
	M5005_RS07430	<i>hrcA</i>	Heat shock transcriptional repressor	-	-1.948
	M5005_RS00085	<i>rrf</i>	Ribosome recycling factor	-	+1.461
	M5005_RS00185	tRNA	tRNA	-	+1.507

The 'greenyellow' module was uniquely enriched in the lactose catabolic process, the 'grey60' module in carboxylic acid metabolic pathways, and the 'magenta' module in ribosome-related functions. The 'yellow' module exhibited enrichment in catalytic activities targeting RNA substrates, suggesting a potential role for resident sORFs in post-transcriptional RNA processing.

Among these functionally enriched modules, sORFs were exclusively localised within the 'black', 'darkmagenta', and 'yellow' modules. These three modules harboured a total of 28 sORFs distributed as follows: 13 in the 'black' module; 11 in the 'yellow' module; and four in the 'darkmagenta' module. Functional annotation indicated that the majority of these sORFs lacked homologous gene matches, suggesting they may encode previously uncharacterized proteins. Within the 'black' module, 8 of the 13 sORFs were unannotated; however, a subset exhibited homology to established metabolic machinery, including sORF_00278 (sugar ABC transporter permease), sORF_00424 (ATP-binding protein UgpC), sORF_00443 (chaperonin GroEL) and sORF_00444 (a 60 kDa heat shock protein).

In the 'yellow' module, 10 of the 11 sORFs were annotated with putative functions, including DNA gyrase subunit A (sORF_00247), histidyl-tRNA synthetase (sORF_00125), and transketolase (sORF_00350). In contrast, all four sORFs in the 'darkmagenta' module were unannotated. Integrating these results with dif-

ferential gene expression analysis revealed only two (sORF_00369 & sORF_00282) demonstrated statistically significant expression changes, both exhibiting downregulations (FDR < 0.05). The remaining 26 sORFs showed no significant differential expression, despite moderate fold changes in certain instances.

Identification of candidate hub-like sORFs via MCC centrality ranking

To identify sORFs of disproportionate regulatory importance within the $\Delta hsdM$ transcriptome, we employed the Maximal Clique Centrality (MCC) algorithm to rank candidates within their respective co-expression modules. This analysis identified a subset of candidate "hub-like" sORFs that exhibit high-weight interactions with essential canonical proteins (Table 4).

In the 'black' module, which is enriched for protein folding and stress recovery, sORF_00444 and sORF_00445 emerged as top-tier hubs (ranked 3rd and 5th, respectively). sORF_00444 showed a strong interaction weight (0.48) with the chaperonin GroEL, while sORF_00445 was topologically linked to the heat-shock co-chaperone GroES (0.48) and the transcriptional repressor HrcA (0.45). Notably, sORF_00369 (8th) also localised within this cluster, interacting with the nucleotide exchange factor GrpE.

The 'green' module, associated with the translational apparatus, was anchored by sORF_00011 (1st), which was identified as the 50S

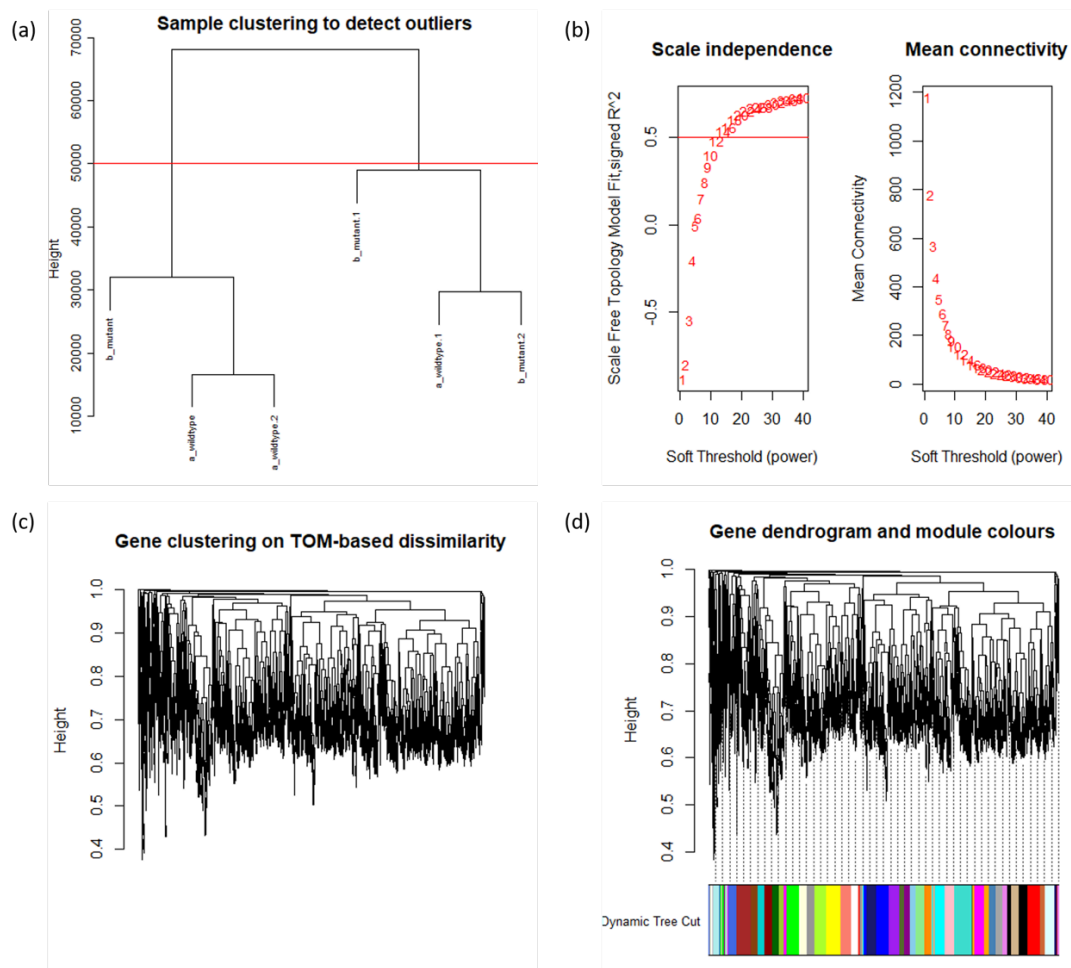


Fig. 1. Network topology and threshold parameter selection for the GAS co-expression network. (a) Sample clustering dendrogram based on Euclidean distance for the six $\Delta hsdM$ and WT transcriptomic profiles; the red line indicates a standardised outlier threshold, though all samples were retained to maintain statistical power within the limited dataset. (b) Analysis of the soft-thresholding power (β) showing the scale-free fit index (left) and mean connectivity (right), justifying the selection of $\beta=12$ as the transition point to a scale-free topology plateau. (c) Dendrogram of all sORFs and protein-coding genes clustered by topological overlap, with assigned colour modules representing distinct co-expression regulons. (d) Heatmap showing the correlation between identified colour modules and the $\Delta hsdM$ clinical trait, identifying the specific sub-networks associated with the epigenetic-deficient phenotype.

Table 4. Topological centrality and functional integration of candidate hub-like sORFs within the GAS Δ hsdM regulon. Summary of the top ten ranked candidate hub-like sORFs identified via MCC ranking across biologically enriched co-expression network modules, suggesting a modular mechanism for the enhanced transformation efficiency and ‘membrane-thinning’ effect observed following the loss of Type 1 Restriction-Modification.

Module	Candidate hub-like sORF	Predicted function	Key Interacting Partners (Functional Role)	Enriched Biological Pathway (GO/KEGG ID)
Black	sORF_00444 (3rd)	Chaperonin GroEL	M5005_RS08655 – chaperonin GroEL M5005_RS06975 – CopY/TcrY family copper transport repressor M5005_RS06970 – heavy metal translocating P-type ATPase	Chaperone-mediated protein folding and proteostatic stress response (GO:0006457)
Black	sORF_00445 (5th)	NA	M5005_RS08660 – co-chaperone GroES sORF_00443 – heat shock protein 60 kDa M5005_RS07430 – heat-inducible transcriptional repressor HrcA	Chaperone-mediated protein folding and proteostatic stress response (GO:0006457)
Black	sORF_00369 (8th)	NA	M5005_RS07425 – nucleotide exchange factor GrpE M5005_RS07430 – heat-inducible transcriptional repressor HrcA sORF_00443 – heat shock protein 60 kDa	Chaperone-mediated protein folding and proteostatic stress response (GO:0006457)
Darkmagenta	sORF_00099 (3rd)	NA	M5005_RS00450 – 50S ribosomal protein L18 M5005_RS00455 – 30S ribosomal protein S5 M5005_RS00445 – 50S ribosomal protein L6	RNA binding and ribosomal gene expression (KEGG:03010)
Green	sORF_00011 (1st)	50S ribosomal protein L14	M5005_RS00430 – 50S ribosomal protein L5 M5005_RS00415 – 30S ribosomal protein S17 M5005_RS00420 – 50S ribosomal protein L14	Ribosomal subunit assembly and protein metabolism (GO:0019538)
Greenyellow	sORF_00357 (1st)	NA	M5005_RS03955 – PTS system mannose/fructose/sorbose family transporter subunit IID M5005_RS08110 – tagatose-6-phosphate kinase M5005_RS06865 – DUF896 family protein	Galactose metabolism and lactose catabolism (KEGG:00052)
Grey60	sORF_00264 (10th)	BMP family protein	M5005_RS06045 – YpmS family protein M5005_RS08640 – hypothetical protein M5005_RS04375 – DNA gyrase subunit A	Carboxylic acid metabolism and DNA maintenance (GO:0019752)
Magenta	sORF_00347 (6th)	NA	M5005_RS01140 – NAD(P)H-dependent glycerol-3-phosphate dehydrogenase M5005_RS00945 – leucine-tRNA ligase sORF_00084 – NA	Ribosomal biogenesis and translation (KEGG:03010)
Yellow	sORF_00350 (6th)	transketolase	sORF_00247 – DNA gyrase subunit A M5005_RS01395 – D-alanyl-D-alanine carboxypeptidase / PBP3 M5005_RS01305 – pur operon repressor	Pentose phosphate pathway and cell wall biogenesis (GO:0140098)

ribosomal protein L14. It maintained robust connectivity with the 30S protein S17 (0.36) and 50S protein L5 (0.37). In the ‘greenyellow’ module, sORF_00357 (1st) functioned as a primary hub, showing significant weights with the PTS system fructose/mannose transporter (0.33) and tagatose-6-phosphate kinase (0.31). Other high-ranking candidates included sORF_00099 (3rd, ‘darkmagenta’) and sORF_00347 (6th, ‘magenta’), which interacted with ribosomal factors and DNA gyrase subunit A, respectively.

DISCUSSION

The annotation landscape of GAS sORFs highlights the limitations of sequence homology alone for detecting short, rapidly evolving genetic elements. While only 37.1% of our high-confidence candidates showed homology, the identification of conserved domains

via InterProScan provides a structural baseline for future functional validation. Our classification of these candidates reveals a complex genomic history; while 146 sORFs appear to be truncated remnants of larger enzymatic complexes, such as DNA polymerase III and acetate kinase, their transcriptional activity suggests that pseudo-gene decay in GAS is not a silent process but one that contributes to the transcriptional pool. Furthermore, the recovery of unannotated canonical genes, including the 50S ribosomal protein L14 (sORF_00011), may underscore the possibility of them being missed by standard annotation filters.

The 297 uncharacterized “orphan” sORFs likely represent species-specific elements associated with niche adaptation. These “taxonomically restricted genes” often emerge de novo or through rapid

divergence to fulfil specialised roles in host colonisation (Vanni *et al.*, 2022). Given their lineage-restricted nature, these microproteins could potentially serve as high-specificity targets for novel narrow-spectrum therapeutic agents, avoiding the broad-spectrum disruption of host microbiota (Wu *et al.*, 2024). Furthermore, the high proportion of antisense sORFs (96.8%) suggests a pervasive layer of cis-encoded interference that may fine-tune the GAS transcriptome during host-induced stress. Even for candidates lacking homology, the presence of predicted transmembrane helices and signal peptides in 40.6% of the structural pool suggests these “hypothetical” sORFs are localised to the membrane, potentially acting as sensors for environmental stimuli (Yuan *et al.*, 2025).

The differential expression of sORFs in the *ΔhsdM* mutant offers insight into epigenetic metabolic adjustments accompanying the loss of Type 1 Restriction-Modification systems. Our validation against the benchmark results of Finn *et al.* (2021) confirms a shared core of metabolic transcripts (*amyB*, *lacA/B.2* & *grpE*), yet our refined pipeline captures a “hidden” layer of 14 DE sORFs previously overlooked. While the original study noted that *hsdM* inactivation does not globally “rewire” the CsrRS regulon, our identification of novel candidates, such as the glycoside hydrolase-associated sORF_00281 (57 nt), suggests that the epigenetic loss of *hsdM* triggers a shift in cellular homeostasis. Rather than a direct attenuation of virulence, the observed downregulation of metabolic sORFs likely reflects a trade-off in bioenergetic allocation and membrane-associated fitness. Our findings suggest that the high transformation efficiency reported by Finn *et al.* (2021) is not an isolated event but the result of a coordinated modular suppression of the cell envelope.

The significant downregulation of sORF_00369 (associated with the acetyl-CoA carboxylase β -subunit complex) points towards a contraction in fatty acid biosynthesis. In *Firmicutes*, fatty acid flux is the primary determinant of membrane phospholipid composition and fluidity (Diomandé *et al.*, 2015). The acetyl-CoA carboxylase complex is the committed step in the Type II fatty acid synthesis (FASII) pathway, whereby its suppression likely limits the availability of long-chain fatty acids required for membrane repair (Čavuzić *et al.*, 2024). In the context of the *ΔhsdM* mutant, reported for its high transformation efficiency, this lipid modulation is critical. Altered membrane fatty acid profiles have been shown in other Gram-positive models to increase membrane permeability and competence for DNA uptake (Upender *et al.*, 2023; Shin *et al.*, 2025). We propose that sORF_00369 may act as a translational “rheostat”, where its downregulation reduces membrane rigidity, thereby facilitating the

translocation of exogenous DNA through the compromised lipid bilayer.

This permeability-enhancing effect is further corroborated by the downregulation of sORF_00282, predicted to be involved in glycerol-3-phosphate (G3P) transport. By limiting G3P import via the UgpC system, the cell restricts the backbone required for glycerolipid synthesis (Thaqi *et al.*, 2025). Evidence in *Streptococcus suis* indicates that transport-mediated G3P levels directly influence cell wall thickness and resistance to environmental osmotic stress (Yang *et al.*, 2025). The concomitant suppression of both FASII (sORF_00369) and G3P transport (sORF_00282) within the same co-expression module implies a potential unified regulatory strategy to alter membrane architecture, which may facilitate the translocation of foreign DNA (Figure 2). However, as this study relies exclusively on *in silico* transcriptomic reanalysis, the lack of *in vitro* validation, such as qRT-PCR to confirm the differential expression of sORF_00369 and sORF_00282, or functional biophysical assays to directly measure membrane fluidity, represents a limitation. Hence, future experimental studies are required to validate these computational predictions.

A novel dimension of our discovery is the identification of sORF_00060, which contains a surface exclusion domain (SEC10). In bacterial systems, surface exclusion proteins typically prevent redundant plasmid entry to maintain genetic stability (Pickering & Fitzgerald, 2020). The presence of this domain in a differentially expressed sORF suggests that *hsdM* deletion may downregulate traditional “gatekeeper” proteins, further lowering the barrier to horizontal gene transfer. This represents a previously unrecognised layer of sORF-mediated defence regulation in GAS.

Our WGCNA results localised these DE sORFs within modules enriched for lipid metabolism and protein folding (‘black’ & ‘yellow’ modules). The modular co-expression of sORF_00369 and sORF_00282 with established metabolic enzymes demonstrates that they function as integrated components of the global regulatory architecture. The construction of this co-expression network was based on a six-sample dataset, representing a specific methodological context where reaching high scale-free topology fit indices (R^2) is more conservative than in larger cohorts. We selected a soft-thresholding power of $\beta=12$ to achieve an optimal balance between maintaining a scale-free network architecture and preserving mean connectivity. Given the six-sample framework, the resulting modules already exhibit a low-acceptable connectivity profile; any further increase in power would risk dissolving these modules entirely. Furthermore, we opted to retain all six biological replicates

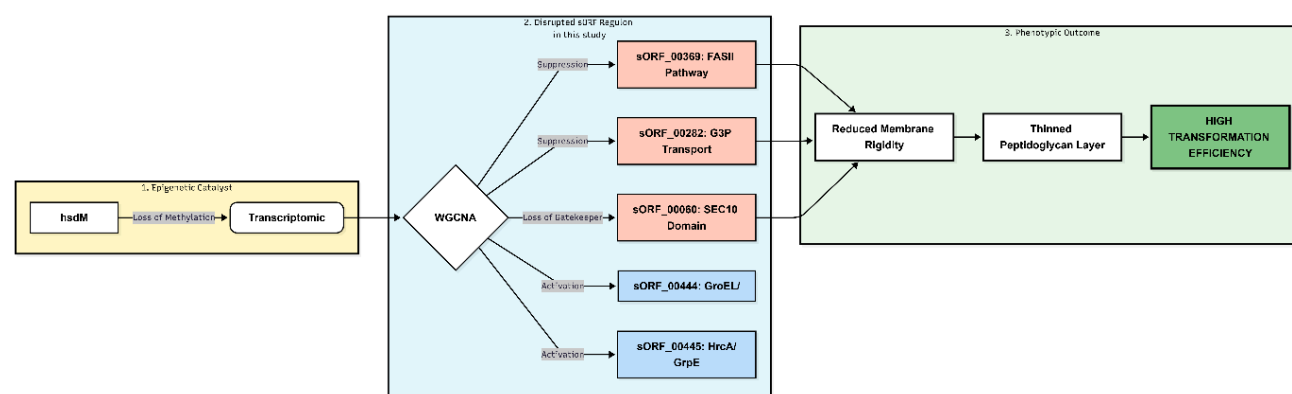


Fig. 2. Systems-level model of the sORF-mediated mechanism predicted to contribute to high transformation efficiency. Conceptual workflow illustrating how the *hsdM* mutation acts as an epigenetic catalyst, triggering a comprehensive transcriptomic shift that disrupts the sORF regulon via specific metabolic suppression and activation events. This coordinated regulatory response is predicted to lead to a reduction in membrane rigidity and a thinned peptidoglycan layer, offering a possible mechanistic model for the high transformation efficiency observed in this background.

without outlier pruning to prioritise statistical power. As published by Xie *et al.* (2025), we treated sample divergence not as technical noise, but as a biological reflection of the “edge of chaos” within the GAS regulatory network. The potential for increased noise was mitigated by employing a stringent minimum module size, ensuring the reported co-expression patterns capture potential biological trends, though we acknowledge that module stability may be limited by the small sample size.

The identification of sORFs as high-ranking hubs provides compelling evidence that the “hidden” transcriptome is deeply integrated into the GAS global regulatory network. By analysing the top ten MCC-ranked candidates, we can infer their roles in stabilising cellular physiology under the stress of *hsdM* loss. The localisation of sORF_00444 (3rd) and sORF_00445 (5th) within the ‘black’ module suggests a critical role in the chaperone-mediated stress response. The strong interaction weights with GroEL and GroES imply that these micropeptides may act as auxiliary factors for the GroE chaperonin complex (Stan *et al.*, 2022; Taguchi & Koike-Takeshita, 2023). In many Gram-positive pathogens, small proteins interact with chaperones to prevent the aggregation of misfolded proteins during environmental shifts (Brantl & Ul Haq, 2023). Given that *hsdM* deletion can lead to subtle metabolic instabilities, these hubs likely coordinate the “quality control” machinery (HrcA/GrpE) to ensure proteostasis, effectively acting as sensors that trigger the heat-shock response even under non-thermal stress (Liu & Yu, 2025).

The 1st-place ranking of sORF_00011 (50S ribosomal protein L14) in the ‘green’ module confirms that our pipeline successfully captures essential but possibly overlooked canonical factors. Its central role among 30S and 50S ribosomal subunits suggests that sORFs in this module are indispensable for maintaining translation efficiency. Similarly, sORF_00357 (1st, ‘greenyellow’) bridges the gap between the small-protein landscape and carbon metabolism. Its connectivity to the PTS mannose/fructose transporter indicates that this sORF may regulate the uptake of specific carbohydrates. In the context of high transformation efficiency, this metabolic “re-tuning” by sORF hubs could provide the necessary energy for the energetically expensive process of DNA uptake and integration (Venkat *et al.*, 2021).

The 6th-ranked unannotated sORF_00347 (magenta) and its interaction with DNA gyrase (GyrA) suggest a previously unrecognised role in managing DNA supercoiling or repair (Germe *et al.*, 2024). Furthermore, the 10th-ranked sORF_00264 (grey60), which interacts with Penicillin-Binding Protein 3 (PBP3, a low-molecular-mass D-alanyl-D-alanine carboxypeptidase), provides a possible hypothetical link between sORFs and bacterial cell wall synthesis and remodelling (Rowland *et al.*, 2025). PBP3 is essential for peptidoglycan synthesis; hence its association with an sORF hub corroborates our “membrane-thinning” hypothesis, suggesting that sORFs may fine-tune the activity of cell wall enzymes to increase porosity and facilitate exogenous DNA entry.

While it is tempting to speculate on epigenetic deregulation via altered DNA methylation, we remain cautious. Without direct methylome sequencing (SMRT-seq), these expression changes are best interpreted as a downstream physiological consequence of *hsdM* loss, where the cell reallocates resources away from lipid synthesis toward translation-related “quality control” as seen in the ‘darkmagenta’ module. This suggests that the “transformation-ready” state is not a passive by-product of mutation, but an active, module-driven metabolic shift.

CONCLUSION

This study establishes a systems-level framework for the epigenetic control of bacterial competence in GAS. While earlier work by Finn *et al.* (2021) identified the *hsdM* mutation as a catalyst for high trans-

formation, our application of WGCNA has predicted additional specific non-canonical architecture that may contribute to this phenomenon. We hypothesise that this hyper-transformable state may be influenced by candidate sORF hubs, specifically sORF_00369 and sORF_00282, which could function as metabolic “rheostats” to suppress the FASII and G3P pathways. *In silico* models predict that this coordinated suppression may facilitate physical alterations to the cell envelope, suggesting a regulatory layer previously unrecognised by canonical transcriptomic analyses.

The computational prediction that sORFs might act as key regulators of cell envelope remodelling elevates these elements from genomic “noise” to potentially essential components of the GAS homeostasis machinery. By providing a theoretical mechanistic link between epigenetic deficiency and structural membrane changes, our findings suggest how the “hidden” transcriptome might influence the genetic plasticity of GAS (*S. pyogenes*). Ultimately, this research highlights sORFs as critical players in the horizontal gene transfer and adaptive evolution of pathogenic lineages, offering a broader predictive framework for understanding how GAS maintains its status as a highly versatile and resilient human pathogen.

ACKNOWLEDGEMENTS

The authors acknowledge the Ministry of Higher Education, Malaysia, for the financial support through the Fundamental Research Grant Scheme (FRGS) funding (FRGS/1/2019/STG05/UKM/03/1) awarded to Norfarhan Mohd-Assaad.

ETHICAL STATEMENT

This study did not involve new experiments with human participants, animals, or recombinant DNA. All analyses were performed on previously published and publicly available sequencing data. Ethical approval was therefore not required.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- Andrews, S. 2010. FastQC: A quality control tool for high throughput sequence data. Available at: <https://www.scrip.org/reference/referencespapers?referenceid=2781642>.
- Aoyama, J.J. & Storz, G. 2023. Two-for-one: regulatory RNAs that encode small proteins. Trends in Biochemical Sciences, 48(12): 1035. <https://doi.org/10.1016/j.tibs.2023.09.002>
- Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.H., Pagès, F., Trajanoski, Z. & Galon, J. 2009. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics, 25(8): 1091-1093. <https://doi.org/10.1093/bioinformatics/btp101>
- Bolger, A.M., Lohse, M. & Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics, 30(15): 2114-2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Brantl, S. & Ul Haq, I. 2023. Small proteins in Gram-positive bacteria. FEMS Microbiology Reviews, 47(6): fuad064. <https://doi.org/10.1093/femsre/fuad064>
- Čavuzić, M.T., Larson, B.A. & Waldrop, G.L. 2024. Insights into the methodology of acetyl-CoA carboxylase inhibition. Methods in Enzymology, 708: 67-103. <https://doi.org/10.1016/bs.mie.2024.10.017>
- Chin, C.H., Chen, S.H., Wu, H.H., Ho, C.W., Ko, M.T. & Lin, C.Y. 2014. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Systems Biology, 8(Suppl 4): S11. <https://doi.org/10.1186/1752-0509-8-S4-S11>
- Conesa, A. & Götz, S. 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. International Journal of Plant Genomics, 2008: 619832. <https://doi.org/10.1155/2008/619832>
- Diomandé, S.E., Nguyen-The, C., Guinebretière, M.H., Broussolle, V. & Brillard, J. 2015. Role of fatty acids in *Bacillus* environmental adaptation. Frontiers in Microbiology, 6: 813. <https://doi.org/10.3389/fmicb.2015.00813>
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T.R. 2013. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics, 29(1): 15-21. <https://doi.org/10.1093/bioinformatics/bts635>
- Espósito, S., Masetti, M., Calanca, C., Canducci, N., Rasmi, S., Fradusco, A. & Principi, N. 2025. Recent changes in the epidemiology of Group A *Streptococcus* infections: Observations and implications. Microorganisms, 13(8): 1871. <https://doi.org/10.3390/microorganisms13081871>

- Ewels, P., Magnusson, M., Lundin, S. & Käller, M. 2016. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19): 3047-3048. <https://doi.org/10.1093/bioinformatics/btw354>
- Finn, M.B., Ramsey, K.M., Toller, H.J., Dove, S.L. & Wessels, M.R. 2021. Improved transformation efficiency of group A *Streptococcus* by inactivation of a type I restriction modification system. *PLoS ONE*, 16(4): e0248201. <https://doi.org/10.1371/journal.pone.0248201>
- Förstner, K.U., Vogel, J. & Sharma, C.M. 2014. READemption-a tool for the computational analysis of deep-sequencing-based transcriptome data. *Bioinformatics*, 30(23): 3421-3423. <https://doi.org/10.1093/bioinformatics/btu533>
- Fuchs, S. & Engelmann, S. 2023. Small proteins in bacteria - Big challenges in prediction and identification. *Proteomics*, 23(23-24): 2200421.
- Garai, P. & Blanc-Potard, A. 2020. Uncovering small membrane proteins in pathogenic bacteria: Regulatory functions and therapeutic potential. *Molecular Microbiology*, 114(5): 710-720. <https://doi.org/10.1111/mmi.14564>
- Germe, T.R., Bush, N.G., Baskerville, V.M., Saman, D., Benesch, J.L. & Maxwell, A. 2024. Rapid, DNA-induced interface swapping by DNA gyrase. *eLife*, 12: RP86722. <https://doi.org/10.7554/eLife.86722>
- Giess, A., Jonckheere, V., Ndahe, E., Chyzyńska, K., Van Damme, P. & Valen, E. 2017. Ribosome signatures aid bacterial translation initiation site identification. *BMC Biology*, 15(1): 1-14. <https://doi.org/10.1186/s12915-017-0416-0>
- Jain, N., Richter, F., Adzhubei, I., Sharp, A.J. & Gelb, B.D. 2023. Small open reading frames: a comparative genetics approach to validation. *BMC Genomics*, 24(1): 1-11. <https://doi.org/10.1186/s12864-023-09311-7>
- Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.Y., Lopez, R. & Hunter, S. 2014. InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, 30(9): 1236-1240. <https://doi.org/10.1093/bioinformatics/btu031>
- Ko, S.H., Cho, B.L. & Shin, D. 2025. Microproteins in metabolic biology: Emerging functions and potential roles as nutrient-linked biomarkers. *International Journal of Molecular Sciences*, 26(24): 11883. <https://doi.org/10.3390/ijms262411883>
- Langfelder, P. & Horvath, S. 2008. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1): 1-13. <https://doi.org/10.1186/1471-2105-9-559>
- Leong, A.Z.X., Lee, P.Y., Mohtar, M.A., Syafruddin, S.E., Pung, Y.F. & Low, T.Y. 2022. Short open reading frames (sORFs) and microproteins: an update on their identification and validation measures. *Journal of Biomedical Science*, 29(1): 1-15. <https://doi.org/10.1186/s12929-022-00802-5>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16): 2078-2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Liu, Z.Y. & Yu, X.Z. 2025. Engineering *Bacillus subtilis* for high-value bioproduction: recent advances and applications. *Microbial Cell Factories*, 24(1): 182. <https://doi.org/10.1186/s12934-025-02818-6>
- Lu, Y., Ran, Y., Li, H., Wen, J., Cui, X., Zhang, X., Guan, X. & Cheng, M. 2023. Micropeptides: origins, identification, and potential role in metabolism-related diseases. *Journal of Zhejiang University Science B*, 24(12): 1106-1123. <https://doi.org/10.1631/jzus.B2300128>
- Pickering, A.C. & Fitzgerald, J.R. 2020. The role of Gram-positive surface proteins in bacterial niche- and host-specialization. *Frontiers in Microbiology*, 11: 594737. <https://doi.org/10.3389/fmicb.2020.594737>
- Robinson, M.D., McCarthy, D.J. & Smyth, G.K. 2009. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1): 139-140. <https://doi.org/10.1093/bioinformatics/btp616>
- Roginski, P., Papadopoulos, C., Herman, S., Baumann, A., Grislain, A. & Lopes, A. 2026. Impact of GC content on de novo gene birth. *Nature Communications*, 17(1): 1268. <https://doi.org/10.1038/s41467-025-68022-7>
- Rowland, C.E., Newman, H., Martin, T.T., Dods, R., Bournakas, N., Wagstaff, J.M., Lewis, N., Stanway, S.J., Balmforth, M., Kessler, C., van Rietschoten, K., Bellini, D., Roper, D.L., Lloyd, A.J., Dowson, C.G., Skynner, M.J., Beswick, P. & Dawson, M.J. 2025. Discovery and chemical optimisation of a potent, bi-cyclic antimicrobial inhibitor of *Escherichia coli* PBP3. *Communications Biology*, 8(1): 819. <https://doi.org/10.1038/s42003-025-08246-x>
- Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11): 2498-2504. <https://doi.org/10.1101/gr.1239303>
- Shin, S., Jiang, D., Yu, J., Yang, C., Jeong, W., Li, J., Bae, J., Shin, J., An, K., Kim, W. & Cho, N.J. 2025. Interaction dynamics of liposomal fatty acids with Gram-positive bacterial membranes. *ACS Applied Materials & Interfaces*, 17(16): 23666-23679. <https://doi.org/10.1021/acsami.5c00787>
- Simoens, L., Fijalkowski, I. & Van Damme, P. 2023. Exposing the small protein load of bacterial life. *FEMS Microbiology Reviews*, 47(6): fuad063. <https://doi.org/10.1093/femsre/fuad063>
- Stan, G., Lorimer, G.H. & Thirumalai, D. 2022. Friends in need: How chaperonins recognize and remodel proteins that require folding assistance. *Frontiers in Molecular Biosciences*, 9: 1071168. <https://doi.org/10.3389/fmolb.2022.1071168>
- Taguchi, H. & Koike-Takeshita, A. 2023. *In vivo* client proteins of the chaperonin GroEL-GroES provide insight into the role of chaperones in protein evolution. *Frontiers in Molecular Biosciences*, 10: 1091677. <https://doi.org/10.3389/fmolb.2023.1091677>
- Thaqi, S.K., Siani, R., Chiba, A., Peine, M., Baum, C., Witting, M., Walch, S., Leinweber, P., Schlöter, M. & Schulz, S. 2025. Strain-specific strategies underlie convergent phosphate solubilization in *Bacillus*. *ISME Communications*, 5(1): ycaf208. <https://doi.org/10.1093/ismeco/ycaf208>
- Upender, I., Yoshida, O., Schrecengost, A., Ranson, H., Wu, Q., Rowley, D.C., Kishore, S., Cywes, C., Miller, E.L. & Whalen, K.E. 2023. A marine-derived fatty acid targets the cell membrane of Gram-positive bacteria. *Journal of Bacteriology*, 205(11): e00310-23. <https://doi.org/10.1128/jb.00310-23>
- Vanni, C., Schechter, M.S., Acinas, S.G., Barberán, A., Buttigieg, P.L., Casamayor, E.O., Delmont, T.O., Duarte, C.M., Eren, A.M., Finn, R.D., Kottmann, R., Mitchell, A., Sanchez, P., Siren, K., Steinegger, M., Glöckner, F.O. & Fernandez-Guerra, A. 2022. Unifying the known and unknown microbial coding sequence space. *eLife*, 11: e67667. <https://doi.org/10.7554/eLife.67667>
- Venkat, K., Hoyos, M., Haycocks, J.R., Cassidy, L., Engelmann, B., Rolle-Kampczyk, U., von Bergen, M., Tholey, A., Grainger, D.C. & Papenfort, K. 2021. A dual-function RNA balances carbon uptake and central metabolism in *Vibrio cholerae*. *The EMBO Journal*, 40(24): e108542. <https://doi.org/10.15252/embj.2021108542>
- Wu, H.Y.L., Ai, Q., Teixeira, R.T., Nguyen, P.H.T., Song, G., Montes, C., Elmore, J.M., Walley, J.W. & Hsu, P.Y. 2024. Improved super-resolution ribosome profiling reveals prevalent translation of upstream ORFs and small ORFs in *Arabidopsis*. *The Plant Cell*, 36(3): 510-539. <https://doi.org/10.1093/plcell/koad290>
- Xie, C., Künzel, S., Zhang, W., Hathaway, C.A., Tworoger, S.S. & Tautz, D. 2025. Patterns of extreme outlier gene expression suggest an edge of chaos effect in transcriptomic networks. *Genome Biology*, 26(1): 272. <https://doi.org/10.1186/s13059-025-03709-0>
- Yang, Q., Li, N., Zheng, Y., Tian, Y., Liang, Q., Zhao, M., Chu, H., Gong, Y., Wu, T., Wei, S., Wang, H., Yan, G., Li, F. & Lei, L. 2025. Identification and characterization of *ugpE* associated with the full virulence of *Streptococcus suis*. *Veterinary Research*, 56(1): 82. <https://doi.org/10.1186/s13567-025-01513-z>
- Yuan, J., Koch, H.G. & Berghoff, B.A. 2025. Functional diversity and molecular interactions of small membrane proteins in bacteria. *MicroLife*, 6: uqaf035.
- Yu, S.H., Vogel, J. & Förstner, K.U. 2018. ANNOgesic: A Swiss army knife for the RNA-seq based annotation of bacterial/archaeal genomes. *GigaScience*, 7(9): giy096. <https://doi.org/10.1093/gigascience/giy096>