BMC Genomics

RESEARCH Open Access

Revealing the spectrum of extended-core gene variation in the *Escherichia coli* pan-genome

Kritika Chugh[1*] and Zhenyu Xuan[2*]

Abstract
Standard pan-genome pipelines, such as Roary, use strict protein identity thresholds (e.g., ≥ 95%) that systematically misclassify highly conserved genes as absent. Single mutational events like large indels, nonsense mutations, or high sequence divergence can cause these false negatives, masking the true composition of the bacterial core genome. We address this by investigating 198 extended-core loci (present in > 95% of strains) from 44 *Escherichia coli* genomes that were incorrectly flagged as absent. Using a synteny-guided pipeline to validate a representative subset of 50 genes, we determined that most apparent absences are not true deletions but distinct evolutionary outcomes: inactivating pseudogenization (e.g., *rlmF*), structural remodeling via in-frame indels (e.g., *yhfR*), and highly divergent orthologs falling below identity cutoffs. By distinguishing genuine gene loss from sequence variation, our framework provides a more accurate view of conserved gene content, enabling precise genotype–phenotype associations and revealing hidden reservoirs of genetic diversity.

Keywords Pan-genome, Synteny, *Escherichia coli*, Pseudogenization, Structural variation, Comparative genomics.

Introduction

Escherichia coli (*E.coli*) exhibits remarkable genomic plasticity enabling adaptation to environments as diverse as the mammalian gut, freshwater ecosystems, and hospital settings. Within the pan-genome framework, genes are commonly classified as core (present in nearly all strains, e.g., ≥ 99%), accessory (strain-specific), and an intermediate category of extended-core loci found in most but not all genomes (Table 1). This study focuses specifically on extended-core genes that are highly conserved (>95% prevalence) but are prone to being missed by standard analysis. These genes often encode functions essential for niche adaptation and virulence such as toxin-antitoxin systems, specialized transporters, transcriptional regulators, and antibiotic resistance factors, yet they remain underexplored compared to strictly conserved core genes or highly variable accessory elements [1, 2].

Most pan-genome tools such as Roary cluster proteins by rigid sequence identity thresholds (e.g., ≥ 95%), which can misclassify loci that have undergone single mutational events [3, 4]. Even a solitary frameshift, small indel, gene conversion tract, or accelerated sequence divergence can push a protein's identity below the cutoff yielding false negative absences. Moreover, fixed threshold clustering can fragment gene families, misassign paralogs, and amplify errors from draft assemblies inflating both false positives and false

*Correspondence:
Kritika Chugh
kritikachugh2@gmail.com; kritika.chugh@utdallas.edu
Zhenyu Xuan
zhenyu.xuan@utdallas.edu
[1] School of Natural Sciences and Mathematics, University of Texas at Dallas, 800 W Campbell Rd, Richardson, TX 75080, USA
[2] Department of Biological Sciences, Center for Systems Biology, University of Texas at Dallas, 800 W Campbell Rd, Richardson, TX 75080, USA

Table 1 Definitions of pan-genome categories and study scope

Category	Description	Prevalence Threshold
Core genome	Genes present in nearly all genomes, essential for basic biology.	≥ 99%
Extended-core genome	Highly conserved genes often misclassified by automated pipelines. The 198 loci investigated here (97.7% prevalence) belongs to this category.	95–99%
Shell genome	Genes with intermediate prevalence often provide adaptive advantages.	15% – 95%
Accessory genome	Rare or strain-specific genes, often acquired through horizontal gene transfer.	< 15%
Validation set	Subset of extended-core loci selected for deep analysis	50 of 198 loci

negatives in core and accessory sets. Such misclassifications obscure evolutionary phenomena such as pseudogenization, structural remodeling, and adaptive divergence and hinder accurate genotype phenotype mapping especially in clinically relevant isolates where minor protein changes can have major impacts [5–7].

To overcome these limitations, we compiled 198 candidate extended-core loci missing from the Roary core set across 44 *E. coli* genomes chosen to reflect the species broad phylogenetic diversity [8]. We then implemented a synteny guided recovery pipeline anchoring on conserved flanking genes to relocate and reannotate each locus. In a representative subset of 50 genes, we performed fine scale sequence and indel analyses to categorize every locus as a pseudogene, divergent ortholog, annotation gap, low divergence misclustered, or true deletion. This strategy reveals a continuum of evolutionary fates from early pseudogenization events and precise in-frame deletions to deeply diverged orthologs uncovering hidden reservoirs of genetic variation that shape *E. coli* phenotypes.

Materials and methods

To uncover extended-core genes overlooked by standard identity-threshold clustering, we developed an integrative workflow combining high-quality genome selection, pan-genome profiling, synteny-guided recovery, and sequence-level validation. Beginning with a precomputed presence/absence matrix, we pinpoint loci missing in only one of 44 strains, then leverage conserved gene order and targeted BLASTn to recover or confirm true absences. Recovered sequences undergo detailed mutation analysis to classify pseudogenes, structural variants, and divergent orthologs.

Genome selection and data acquisition

We analyzed 44 *E. coli* genomes to capture both high assembly quality and broad phylogenetic diversity. Ten complete closed-reference assemblies (e.g., MG1655, W3110, UTI89, CFT073) and 34 high-quality drafts (N50 >4 Mb; ≤ 5 scaffolds) were sourced from NCBI RefSeq. Strains were chosen to represent the species' phylogenetic breadth, as defined by the Clermont typing scheme [9]. The corresponding GFF3 annotation files and gene

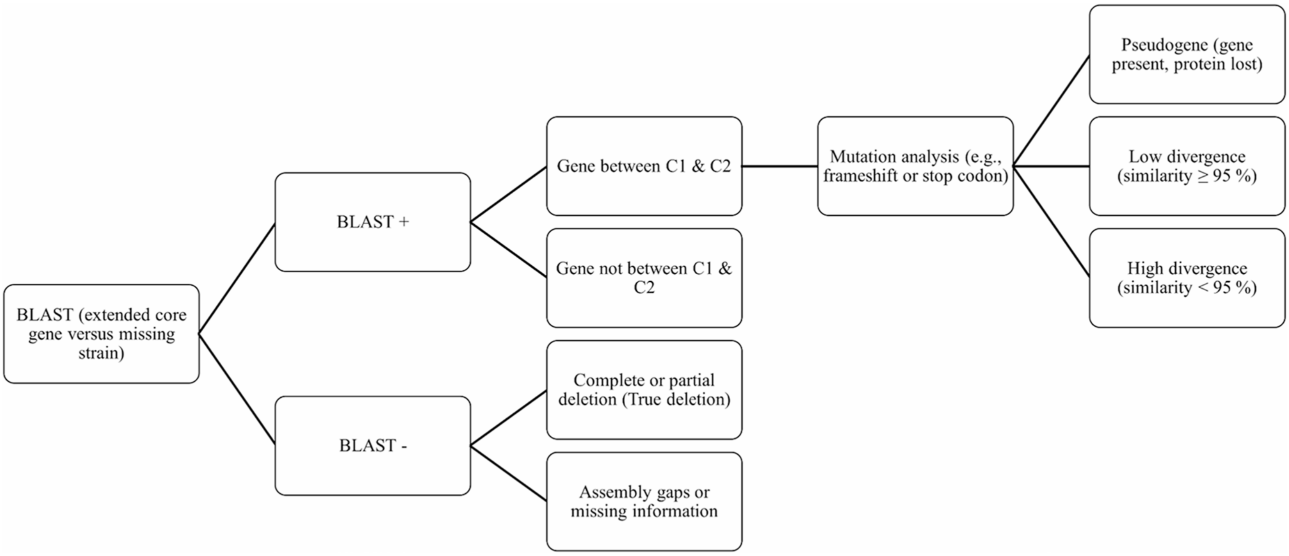

BLAST (extended core gene versus missing strain)
BLAST +
BLAST -
Gene between C1 & C2
Gene not between C1 & C2
Complete or partial deletion (True deletion)
Assembly gaps or missing information
Mutation analysis (e.g., frameshift or stop codon)
Pseudogene (gene present, protein lost)
Low divergence (similarity ≥ 95 %)
High divergence (similarity < 95 %)

Fig. 1 Gene classification based on blast and synteny analysis

Table 2 Functional classification of extended-core loci

Category	Number	Genes
Central metabolism enzymes	42	*aceA, arcC2, bglJ, cadB, caiB, cadA, fdrA, fdrA_1, fumA, gcd, glpR, grxA, manA, metH, oleD_2, phnC, phnD, phnF, phnG, phnH, phnI, phnJ, phnK, phnL, phnM, pitA_1, poxB, prpD, pdeG, prr, rtcR, rimK, rlmC, rlmF, ssuA, ssuE, treC, yjhD, yqhD, yidL*
Transcriptional regulators & DNA-binding	36	*arcD, astA, bluR, cmpR/yfiE, cspC, cspG, cbl, creC, eamA, eutB, eutC, eutK, eutL, gcd, kefF, mlrA, ortT, prr, rtcB_1, rph, sgrR_2, thi4, waaG, yahG, yahE, yafG, yacG, ybaL, ybcI, yciQ, yidK, yjhF, yoeG, yqcE, yqiA*
Toxin–antitoxin & stress response	29	*artI, artJ, artM_2, artP, artQ, arcC2, bglJ, cadB, cspC, cspG, gacA, ghoT, hcp, hcr, hdhA, hpf, hyaE, iceA, kefF, manA, nfsA, potF, potG, potH, potI, ybjC, ybjM, ybjN, ybjO*
Transporters & membrane proteins	19	*dcuA, dcuD, dadD, fhuE, fhuF, folK, glpR, lapB_1, nivA, pitA_1, potH, potI, ydjH, ydjJ, ydfO, ydhL, ydhP, ydhZ, yfjC*
Uncharacterized/hypothetical	72	*agaA, agaC_2, arcC2, astA, bluR, cspC, cspG, dgcN, dcuA, dcuD, dhaM, dLH, eamA, ecpA, ecpB, ecpC, ecpE, ecpR, flaG, fumA, gacA, glpR, hdhA, holD, hypD, iceA, iscU, lafU, lapB_1, mlrA, nivA, ortT, pgrR_3, prr, prpD, psuG, psuT, rtcB_1, rihB, rph, rtcR, sgrR_2, ssuA, ssuE, tdk, treC, uidA, uidR, waaG, yacG, ybaL, ybcI, ybjQ, ybjT, yddH, ydgA_2, ydjE_1, ydjE_2, ydjH, ydjJ, ydjZ_2, yehC, yheI, yhgD, yidK, yigE, yqcE, yycE, yzfF, zwf, group_6249, WP_000017553.1*

FASTA sequences, matching these exact assemblies, are hosted in our GitHub repository.

Pan-genome profiling

We used a Roary v3.13.0 presence/absence matrix (95% protein-identity cutoff), generated in early 2024 by the Xuan Lab (UT Dallas) from curated GFF3 annotations of 44 *E. coli* genomes. From this matrix, we extracted 198 clusters present in 43 of 44 genomes which is our extended-core set. These loci are conserved across ≥ 97.7% of genomes yet evade strict identity-threshold clustering, highlighting cases where minor sequence or structural changes lead to false negatives. Each extended-core candidate was examined for molecular lesions (frameshifts, in-frame indels, gene-conversion tracts, or elevated divergence) that explain its absence. We did not rerun Roary for this study as our goal was to demonstrate recovery of genuine core loci directly from the unmodified, "off-the-shelf" outputs most users obtain. Although lowering Roary's identity threshold (e.g., to 90%) or using alternative tools such as Panaroo, PIRATE, or PEPPAN can recover additional loci, these pipelines rely on clustering heuristics that can, in some cases, lead to paralog misassignments or spurious merges [10–12]. Our framework is complementary: by anchoring each locus in its conserved synteny context and applying local sequence

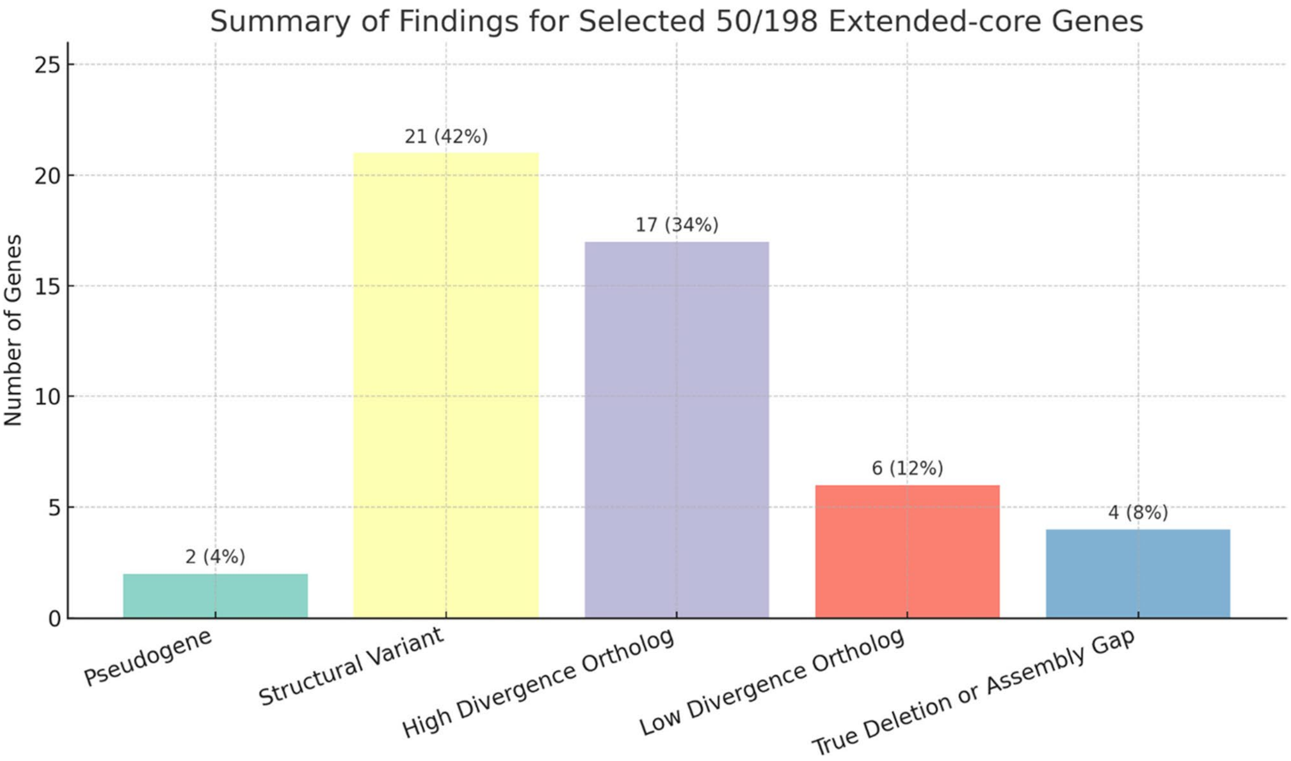


Fig. 2 Summarizing the findings of the selected 50/198 extended-core genes

alignment, we provide locus-specific mechanistic explanations without re-clustering.

Synteny analysis and BLAST search

For each extended-core candidate missing only in one *E.coli* genome, we first identified its two conserved flanking core genes C1 (upstream) and C2 (downstream) from genomes where the locus was present (Fig. 1). Next, for the strain in which the locus was reported missing we located C1 and C2 on the same scaffold, extracted the entire nucleotide segment between them, and ran BLASTn ($E \leq 1 \times 10^{-5}$) of the reference gene sequence against this interval. A hit within the C1–C2 region indicated the gene remained present and we then aligned the recovered sequence to screen for frameshifts or premature stop codons (pseudogenization) and to measure overall identity, classifying variants as low divergence (≥ 95% identity) or high divergence (< 95% identity). Hits falling outside the C1–C2 boundaries were taken as evidence of genomic rearrangements. When no hit was found, we examined the C1–C2 interval: if it was contiguous and contained no long runs of "N"s, we called a true deletion; if scaffold breaks or stretches of ambiguous "N" bases were present, we recorded an assembly or annotation gap. This synteny-guided BLASTn approach ensures rigorous discrimination between genuine gene loss, structural or sequence variation, and technical artifacts.

Multiple sequence alignment

We performed a two-tiered sequence analysis on all loci recovered by synteny and BLASTn. In the first tier, we aligned each gene's nucleotide sequence to its reference allele using Jalview v2 [13] and inspected these alignments for evidence of pseudogenization, such as extended-cinsertions or deletions and premature stop codons, as well as for structural variants indicated by in-frame changes of ten or more amino acids. In cases where the coding frame remained intact, but the gene was still missing from the Roary output, we translated both query and reference sequences and realigned them at the protein level to measure overall amino-acid identity. Based on these examinations, we classified each locus into one of four categories: pseudogene, structural variant, low divergence ortholog (≥ 95% identity), or high divergence ortholog (< 95% identity). This workflow ensures that even subtle sequence or structural alterations responsible for Roary's false negatives are accurately identified and categorized.

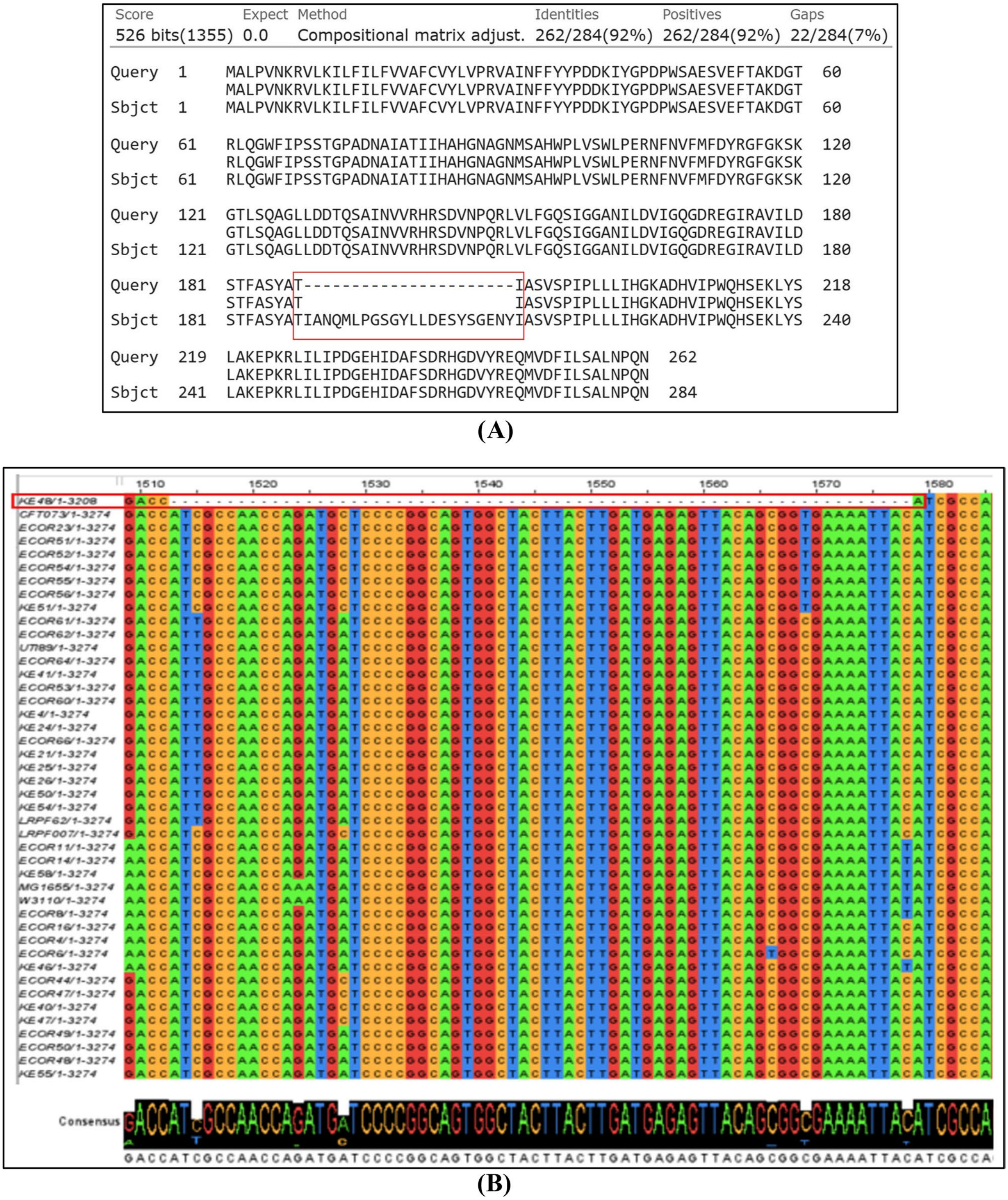
Results

Functional landscape of extended-core candidates

Our 198 extended-core loci encompass diverse functional categories, underscoring their broad adaptive significance. By mapping each gene to EcoCyc or UniProt annotations, we found that 42 loci (21%) encode enzymes involved in central metabolism. Examples include the isocitrate lyase *aceA*, the c-di-GMP phosphodiesterase *pdeG*, and the NADPH-dependent aldehyde reductase *yqhD*. Thirty-six genes (18%) are transcriptional regulators or DNA-binding proteins (e.g. *cmpR/yfiE*, *yhfR*, *yciQ*), while 29 (15%) belong to toxin–antitoxin or stress-response systems (*ghoT*, *phnP*, *sra*). Nineteen loci (10%)

Table 3 Summary of the pairwise alignment results of the few extended-core genes

Extended-core genes	Strain	Query (CFT073) length	Coverage & Identity (%)	Accession number/ contig number	Mutations	Category
rlmF	ECOR44	927	100 97.22 (DNA sequence)	NZ_QOYD01000001	single-nucleotide insertion (7 A vs. 6 A) in the poly-A region	Pseudogene
sra	ECOR50	138	100 97.83 (DNA sequence)	NZ_QOYJ01000074	Internal stop codon	Pseudogene
ydhL	ECOR50	240	100 95 (DNA sequence)	NZ_QOYJ01000001	single nucleotide substitutions	Low divergence ortholog
aceA	ECOR49	1305	100 98 (DNA sequence)	NZ_QOYI01000159	single nucleotide substitutions	Low divergence ortholog
ymdA	KE46	312	100 97 (DNA sequence)	contig 1	Frameshift mutation	Low divergence ortholog
ghoT	ECOR64	174	100 100 (DNA sequence)	NZ_QOYW01000020	No nucleotide substitutions	Low divergence ortholog
yjjU	ECOR53	1074	100 80 (DNA sequence)	NZ_QOYM01000003	single nucleotide substitutions (similarity below roary threshold)	High divergence ortholog
yhfR	KE48	855	100 100 (DNA sequence) 100 92% (Protein sequence)	contig 1	23-aa deletion (protein sequence identity drops to 92%)	Structural variant



```
 Score              Expect  Method                        Identities      Positives       Gaps
 526 bits(1355)     0.0     Compositional matrix adjust.  262/284(92%)    262/284(92%)    22/284(7%)

Query  1    MALPVNKRVLKILFILFVVAFCVYLVPRVAINFFYYPDDKIYGPDPWSAESVEFTAKDGT  60
            MALPVNKRVLKILFILFVVAFCVYLVPRVAINFFYYPDDKIYGPDPWSAESVEFTAKDGT
Sbjct  1    MALPVNKRVLKILFILFVVAFCVYLVPRVAINFFYYPDDKIYGPDPWSAESVEFTAKDGT  60

Query  61   RLQGWFIPSSTGPADNAIATIIHAHGNAGNMSAHWPLVSWLPERNFNVFMFDYRGFGKSK  120
            RLQGWFIPSSTGPADNAIATIIHAHGNAGNMSAHWPLVSWLPERNFNVFMFDYRGFGKSK
Sbjct  61   RLQGWFIPSSTGPADNAIATIIHAHGNAGNMSAHWPLVSWLPERNFNVFMFDYRGFGKSK  120

Query  121  GTLSQAGLLDDTQSAINVVRHRSDVNPQRLVLFGQSIGGANILDVIGQGDREGIRAVILD  180
            GTLSQAGLLDDTQSAINVVRHRSDVNPQRLVLFGQSIGGANILDVIGQGDREGIRAVILD
Sbjct  121  GTLSQAGLLDDTQSAINVVRHRSDVNPQRLVLFGQSIGGANILDVIGQGDREGIRAVILD  180

Query  181  STFASYAT----------------------IASVSPIPLLLIHGKADHVIPWQHSEKLYS  218
            STFASYAT                      IASVSPIPLLLIHGKADHVIPWQHSEKLYS
Sbjct  181  STFASYATIANQMLPGSGYLLDESYSGENYIASVSPIPLLLIHGKADHVIPWQHSEKLYS  240

Query  219  LAKEPKRLILIPDGEHIDAFSDRHGDVYREQMVDFILSALNPQN  262
            LAKEPKRLILIPDGEHIDAFSDRHGDVYREQMVDFILSALNPQN
Sbjct  241  LAKEPKRLILIPDGEHIDAFSDRHGDVYREQMVDFILSALNPQN  284
```

Fig. 3 Structural variation in the yhfR Locus. **A** Pairwise alignment of the *yhfR* protein showing the 23 aa deletion that reduces sequence identity to 92% (Query: KE48_01513 and Subject: CFT073 *yhfR*). **B** Multiple sequence alignment of the *yhfR* nucleotide region showing the 66-bp deletion unique to KE48

encode transporters or membrane-associated proteins, such as *ydjH*. The remaining 72 (36%) lack characterized functions (e.g. *ydhL*, *ymdA*, *yjjU*). This distribution highlights that extended-core genes contribute to core metabolism, regulation, stress adaptation, and niche-specific interactions, making their accurate recovery critical for understanding *E. coli* biology (Table 2).

Synteny analysis reveals locus preservation

We applied synteny analysis to all 198 extended-core candidates, recovering 172 loci in their expected C1- C2 contexts (87%), flagging 10 as rearrangements (5%), and identifying 16 deletions or gaps (8%). For downstream sequence alignments and structural categorization, we then selected a representative subset of 50 genes. This subset was chosen to span all functional classes (metabolism, regulation, stress response, transport, and uncharacterized) and all lesion types (pseudogenes, structural variants, low and high divergence orthologs, and true deletions), ensuring our detailed analyses capture the full spectrum of outcomes observed in the larger dataset.

Sequence analysis identifies a spectrum of molecular lesions

To identify the specific molecular events causing gene exclusion, we analyzed pairwise alignments for 50 candidate extended-core genes. This analysis revealed a spectrum of evolutionary modifications, with the most common fates being pseudogenization, high sequence divergence, and structural variation. From the 50-candidate extended-core genes we examined, a small fraction (2 genes, or 4%) *rlmF* and *sra* are clearly pseudogenes, each harboring inactivating frameshifts or premature stop codons (Fig. 2). The largest category (21 genes, 42%) comprises structural variants: these loci (including *artM_2, ecpA–C, grxA, hcp, hcr, ltaE, lysO, oleD_2, phnD, phnJ, phnM, phnP, potI, rimK,* and the cluster *of ybj genes C, M, O, Q,* and *T*) all contain in-frame insertions or deletions of at least ten amino acids. High divergence orthologs represent 17 genes (34%); although intact, their protein identity falls below the 95% Roary threshold (e.g., *yjjU, arcC2, artI–Q, nfsA, phnC–L, potF–H, poxB, rlmC, and yaaU*). In contrast, six genes (12%) *ymdA, aceA, ghoT, yhfR, phnF, and ybjN* are low divergence orthologs, exhibiting ≥ 95% identity with no disruptive mutations. Finally, four loci (8%) *pdeG, ydjH, cmpR/yfiE,* and *phnP* appear to be true deletions or lie within assembly gaps, as they lack any contiguous BLASTn hit in the expected genomic interval.

Table 3 presents eight illustrative extended-core loci for which full-length BLAST alignments were obtained (100% query coverage). In ECOR44, the methyltransferase *rlmF* (927 bp) aligns at 97.22% identity to its reference (NZ_OYD01000001) but carries a single-nucleotide insertion in a poly-A tract. The 138 bp 30 S subunit–associated protein gene *sra* in ECOR50 aligns at 97.83% identity (NZ_QOYJ01000074) and harbors an internal stop codon. The uncharacterized gene *ydhL* (240 bp) in ECOR50 aligns at exactly 95% identity (NZ_QOYJ01000001) via multiple synonymous substitutions. The isocitrate lyase *aceA* (1,305 bp) in ECOR49 matches at 98% identity (NZ_QOYI01000159) with scattered nucleotide changes. The 312 bp *ymdA* in KE46 aligns at 97% identity on contig 1 but contains a frameshift mutation. The gene *ghoT* (174 bp) in ECOR64 retains 100% identity (NZ_QOYW01000020) despite a handful of silent substitutions. Finally, in KE48 the transcriptional repressor yhfR (855 bp) shows 100% alignment coverage on contig 1 but includes a precise 23-amino-acid in-frame deletion that reduces protein identity to 92% (Fig. 3A and B).

Discussion

Our analysis shows that many extended-core genes reported as absent in Roary's identity-threshold clustering are in fact present in the genome but altered by mutations, small indels, or sequence divergence. These variants fall outside fixed clustering parameters and are therefore misclassified as deletions. By anchoring each candidate locus in its conserved synteny context, we were able to distinguish genuine gene loss from exclusions caused by pseudogenization, structural remodeling, or divergence below the 95% cutoff. The conservation of flanking core genes provided strong evidence that most loci remain physically intact, while disrupted synteny was more consistent with true gene loss or local rearrangements. This gene-by-gene approach offers a clear view of how technical thresholds and sequence variation interact to shape pangenome outputs.

Beyond methodological considerations, the results underscore that extended-core genes contribute substantially to metabolic functions, regulation, and stress responses, yet their diversity is often overlooked. Existing tools such as Panaroo and PEPPAN incorporate synteny-aware heuristics, but our framework complements these by providing locus-specific explanations without modifying clustering parameters. We intentionally analyzed single-strain absences to isolate false negatives from genuine population-level gene loss, demonstrating how detailed case-level inspection can refine global presence/absence matrices. While our research confirms the physical presence of these loci, their functional status remains an open question. Incorporating transcriptomic evidence is crucial to clarify whether these variants remain active or represent stages of gene decay. Together, these findings highlight the value of integrating synteny and sequence-level validation into pan-genome analysis to more accurately capture the evolutionary trajectories of bacterial genes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-12188-3.

Supplementary Material 1.

Acknowledgements

We acknowledge The University of Texas at Dallas for providing the computational resources.

Authors' contributions

KC and ZX jointly conceptualized the study. KC implemented the synteny-guided framework, performed all computational and comparative analyses, and drafted the manuscript. ZX provided the genomic datasets, contributed to study design and interpretation, and supervised the work.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

The custom Python scripts used for analysis, along with all supporting data including the pan-genome input files, genome annotation files, and tables of accession numbers, are publicly available in the GitHub repository: -. https://github.com/kritikachugh/soft_core_genes_synteny_project.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 9 July 2025 / Accepted: 6 October 2025
Published online: 07 November 2025

References

1. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Fraser CM, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial pan-genome. Proc Natl Acad Sci USA. 2005;102(39):13950–5.
2. Kaas RS, Friis C, Ussery DW, Aarestrup FM. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes. BMC Genomics. 2012;13:577.
3. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. Roary: rapid large-scale prokaryote pan-genome analysis. Bioinformatics. 2015;31(22):3691–3.
4. Sitto F, Batistes FU. Estimating pangenomes with Roary. Mol Biol Evol. 2020;37(3):933–9.
5. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005;39:309–38.
6. Hogins J, Do HP, Dashti MG, Zimmern PE, Reitzer L. Genome sequencing of correlated pathogenic Escherichia coli and Enterococcus associated with recurrent urinary tract infections. Microbiol Resour Announc. 2024;13(7):e0014724.
7. Hogins J, Zimmern PE, Reitzer L. Genome sequences of seven clade B2 Escherichia coli strains associated with recurrent urinary tract infections in postmenopausal women. Microbiol Resour Announc. 2023;12(5):e0003523.
8. Selander RK, Levin BR. Genetic diversity and structure in Escherichia coli populations. Science. 1980;210(4469):545–7.
9. Clermont O, Christenson JK, Denamur E, Gordon DM. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environ Microbiol Rep. 2013;5(1):58–65.
10. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Curr Opin Microbiol. 2015;23:148–54.
11. Zhou Z, Charlesworth J, Achtman M. Accurate reconstruction of bacterial pan- and core genomes with PEPPAN. Genome Res. 2020;30(11):1667–79.
12. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, Gladstone RA, Lo S, Beaudoin C, Floto RA, Frost SDW, Corander J, Bentley SD, Parkhill J. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. 2020;21(1):180.
13. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview version 2—A multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25(9):1189–91.

Publisher's Note