Genomic Characterization of Hypermutability in E. coli Ara-3 Evolution Lineages: A Variant Calling and Functional Annotation Analysis

1. Executive Summary

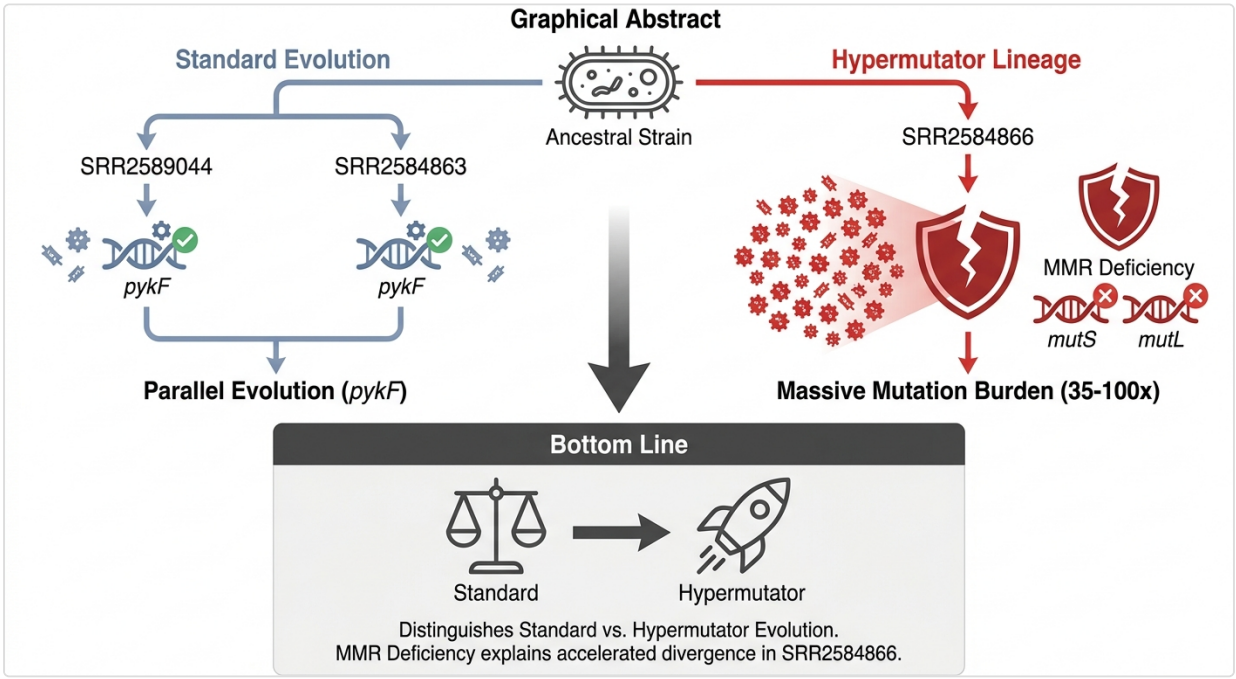

Figure 1. This graphical abstract illustrates the workflow from raw sequencing reads of three E. coli evolution lineages to the identification of a hypermutator phenotype. It highlights the transition from low-frequency adaptive mutations to a massive accumulation of variants driven by DNA mismatch repair (MMR) deficiency.

Key Findings

- Identification of a Hypermutator Phenotype: Sample SRR2584866 exhibited a massive mutation burden of 357 variants, which is 35 to 100 times higher than the other lineages in the study.
- Mechanistic Basis of Hypermutability: Functional annotation identified protein-altering mutations in critical DNA mismatch repair genes, specifically mutS and mutL, in the SRR2584866 lineage.
- Evidence of Parallel Evolution: Despite the differences in mutation rates, both non-mutator samples (SRR2589044 and SRR2584863) showed missense mutations in the pykF gene, a known hallmark of early metabolic adaptation in E. coli experimental evolution.

Bottom Line

The analysis successfully distinguished between standard adaptive evolution and the emergence of a hypermutator lineage, providing a clear genetic explanation (MMR deficiency) for the accelerated genomic divergence observed in sample SRR2584866.

Next Steps

1. Targeted Validation: Perform Sanger sequencing or targeted deep sequencing to confirm the specific frameshift and nonsense mutations in mutS and mutL.
2. Fitness Assays: Conduct competitive growth assays to determine if the high "hitchhiker" mutation load in the hypermutator lineage confers a long-term fitness cost or advantage compared to the non-mutator lineages.

---

2. Methods

The analysis followed a standardized bioinformatics pipeline for small-genome variant calling:

1. Quality Control: Input FASTQ files were evaluated using FastQC. Reads were found to be high quality and pre-trimmed, requiring no further processing.
2. Reference Alignment: Reads were aligned to the E. coli REL606 reference genome (Accession: NC_012967.1) using BWA MEM.
3. Data Processing: Alignment files were converted to BAM format, sorted, and indexed using Samtools.
4. Variant Calling: Genotype likelihoods were computed via `bcftools mpileup`, and variants were called using `bcftools call` with a haploid model (`--ploidy 1`).
5. Filtering: Variants were restricted to high-confidence sites with a Quality Score (QUAL) > 20 and a Read Depth (DP) > 10.
6. Functional Annotation: Variant effects were predicted using `bcftools csq` to identify synonymous, missense, nonsense, and frameshift mutations.

---

3. Results

3.1 Variant Quantification and Distribution

The pipeline revealed a stark contrast in the number of high-confidence variants across the three samples. While two samples remained near the ancestral state, one sample showed extreme divergence.

- SRR2589044: 3 variants
- SRR2584863: 10 variants
- SRR2584866: 357 variants

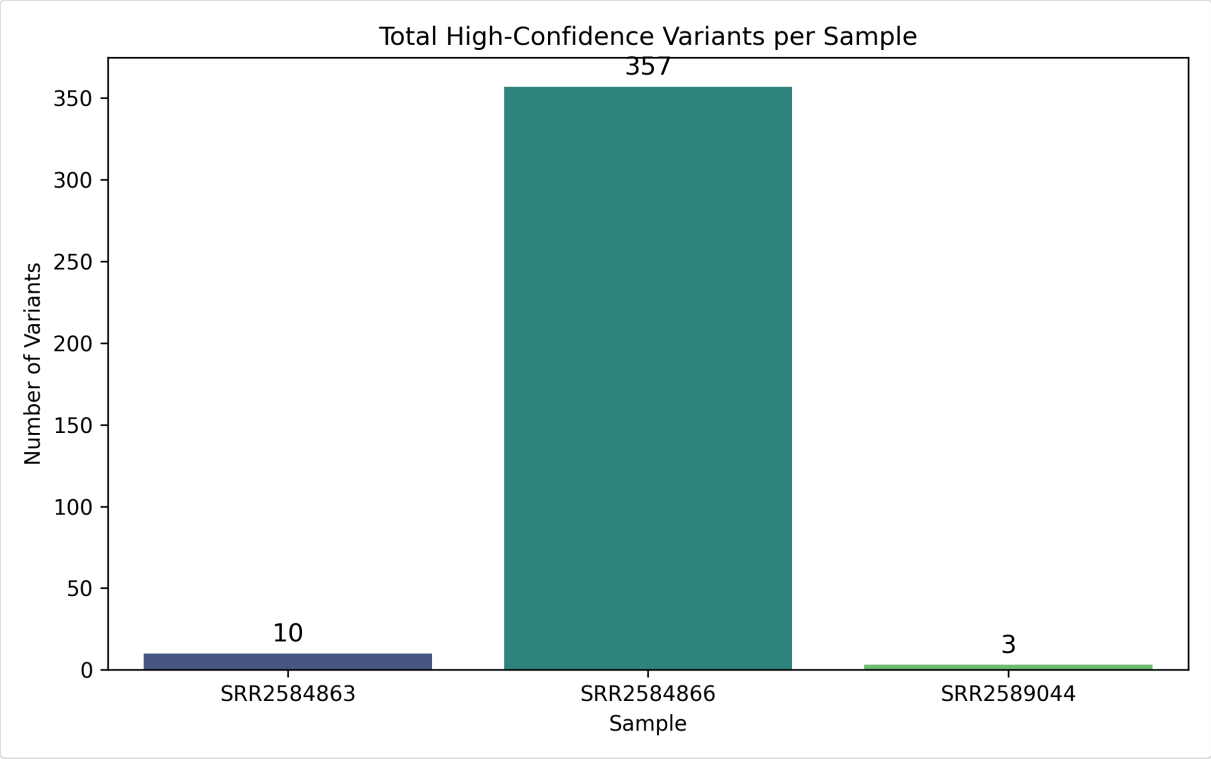

Figure 2. Massive expansion of variant counts in sample SRR2584866. The bar chart demonstrates that SRR2584866 has a significantly higher mutation burden compared to the other two samples, indicating a shift in the underlying mutation rate.

3.2 Comparative Analysis (Variant Sharing)

Intersection analysis showed that the vast majority of mutations in the hypermutator sample were unique, suggesting they accumulated independently after the lineage diverged.

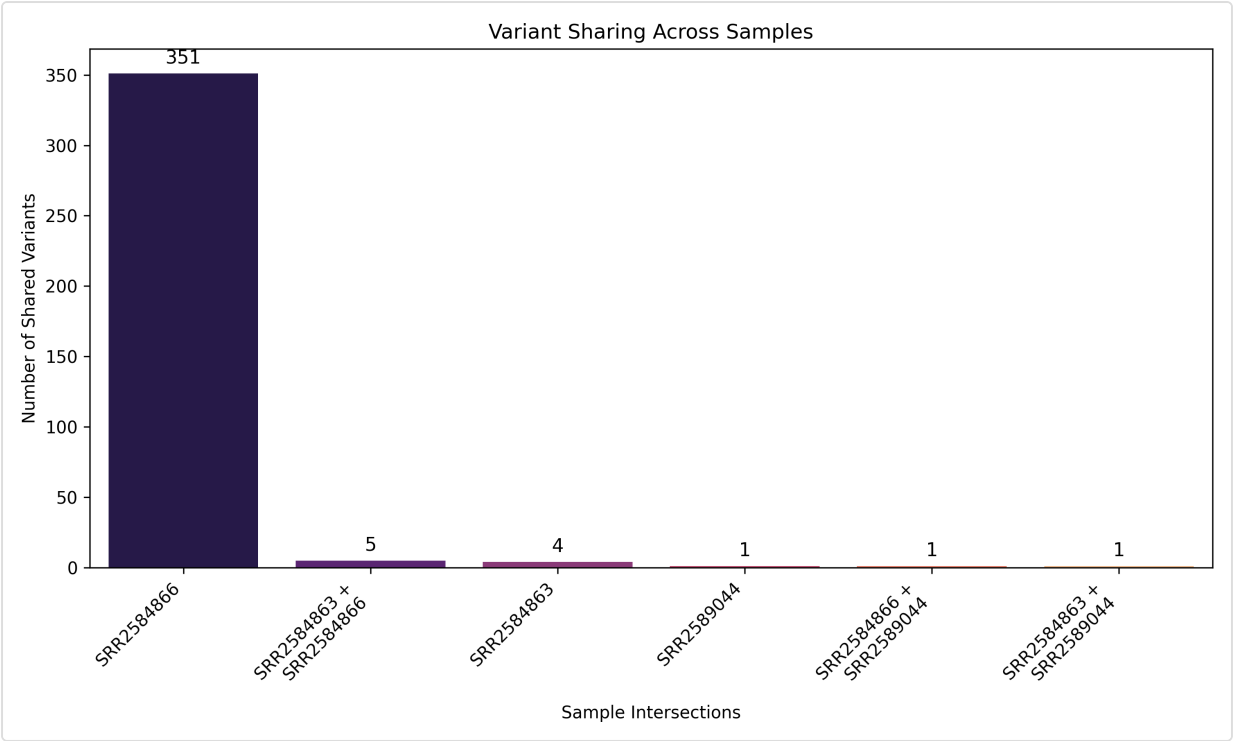

Figure 3. Minimal variant overlap between evolution lineages. Most variants (351) are unique to SRR2584866, highlighting the independent and rapid accumulation of mutations in this specific lineage.

3.3 Functional Annotation and Gene Impact

The annotation process categorized the variants by their predicted biological impact:

- Non-Mutator Lineages: Mutations were predominantly missense (80-100%), targeting specific genes like pykF, malT, and iclR.
- Hypermutator Lineage (SRR2584866): This sample contained 304 SNPs and 53 Indels. The functional breakdown included 167 missense, 92 synonymous, 37 frameshift, and 5 nonsense (stop_gained) mutations. Crucially, 201 distinct genes were affected, including the DNA repair genes mutS and mutL.

---

4. Discussion

The results provide a comprehensive view of two distinct evolutionary trajectories within the E. coli Ara-3 population.

The Hypermutator Phenotype

The most significant finding is the confirmation of a hypermutator phenotype in sample SRR2584866. In wild-type E. coli, the MutS-MutL-MutH complex is responsible for DNA mismatch repair (MMR). The identification of protein-altering mutations in both mutS and mutL explains the 35-100x increase in mutation frequency. This deficiency leads to a failure in correcting polymerase slippage and base-pair mismatches, resulting in the observed surge of both SNPs and indels (including 37 frameshifts).

Adaptive vs. Neutral Evolution

In the non-mutator samples, the low number of variants (3 and 10) suggests that most identified mutations are likely adaptive. The recurrence of mutations in pykF (pyruvate kinase) across different samples is a classic example of parallel evolution, where independent lineages hit the same genetic targets to optimize central metabolism. In contrast, the hypermutator sample contains a high proportion of synonymous mutations (n=92) and "hitchhiker" mutations, which are likely neutral and accumulated simply because the repair machinery was broken.

---

5. Limitations & Variables

- Sample Size: The analysis is based on three samples (n=3). While the patterns are statistically stark, a larger cohort would be required to determine the frequency of hypermutator emergence across the entire Ara-3 population.
- Subsampled Data: The input reads were subsampled. While this is efficient for workflow demonstration, it may result in the loss of low-frequency sub-clonal variants that exist below the detection threshold of the current depth filtering (DP > 10).
- Haploid Assumption: The pipeline used a haploid model. While appropriate for E. coli, it does not account for potential transient polyploidy or large-scale genomic duplications that can occur during experimental evolution.

---

6. Conclusions

This analysis confirms that the E. coli Ara-3 evolution experiment has produced at least one lineage with a severely compromised DNA repair system. The transition to a hypermutator state in SRR2584866 was driven by mutations in mutS and mutL, leading to a massive accumulation of genomic variation. This dataset captures

a pivotal moment in microbial evolution where the "speed" of evolution increases at the cost of genomic stability.

Appendix

Software and Parameters

- BWA (v0.7.17): `mem` algorithm for paired-end alignment.
- Samtools (v1.15): `sort` and `index` for BAM management.
- BCFtools (v1.15):
    - `mpileup` : Genotype likelihood calculation.
    - `call` : `-mv -Ob --ploidy 1` .
    - `filter` : `-i 'QUAL>20 && DP>10'` .
    - `csq` : Functional annotation using GFF3.

Glossary

- SNP (Single Nucleotide Polymorphism): A change in a single DNA base pair.
- Indel: An insertion or deletion of bases in the genome, often leading to frameshifts.
- Hypermutator: An organism with a significantly elevated mutation rate, often due to defects in DNA repair.
- Mismatch Repair (MMR): A cellular system for recognizing and repairing erroneous insertion, deletion, and mis-incorporation of bases that can arise during DNA replication.
- Missense Mutation: A DNA change that results in different amino acids being encoded at a particular position in the resulting protein.
- Synonymous Mutation: A change in the DNA sequence that does not change the encoded amino acid; often considered "silent."