

Lawrence Berkeley National Laboratory
Lawrence Berkeley National Laboratory

Title

Large Lineage-Specific Expansions Are Driven By Mobile Genetic Elements

Permalink

<https://escholarship.org/uc/item/0ps2v93n>

Author

Huang, Katherine

Publication Date

2008-06-02

Katherine Huang^{1,2}, Morgan Price^{1,2}, Eric Alm^{3,4}, Paramvir Dehal^{1,2}, Adam Arkin^{1,2,5}

¹ Physical Biosciences, Lawrence Berkeley National Laboratory; ² Virtual Institute of Microbial Stress and Survival; ³ Biological Engineering, Massachusetts Institute of Technology; ⁴ Civil and Environmental Engineering, Massachusetts Institute of Technology; ⁵ Bioengineering, University of California, Berkeley

VIMSS Virtual Institute for Microbial Stress and Survival

Abstract

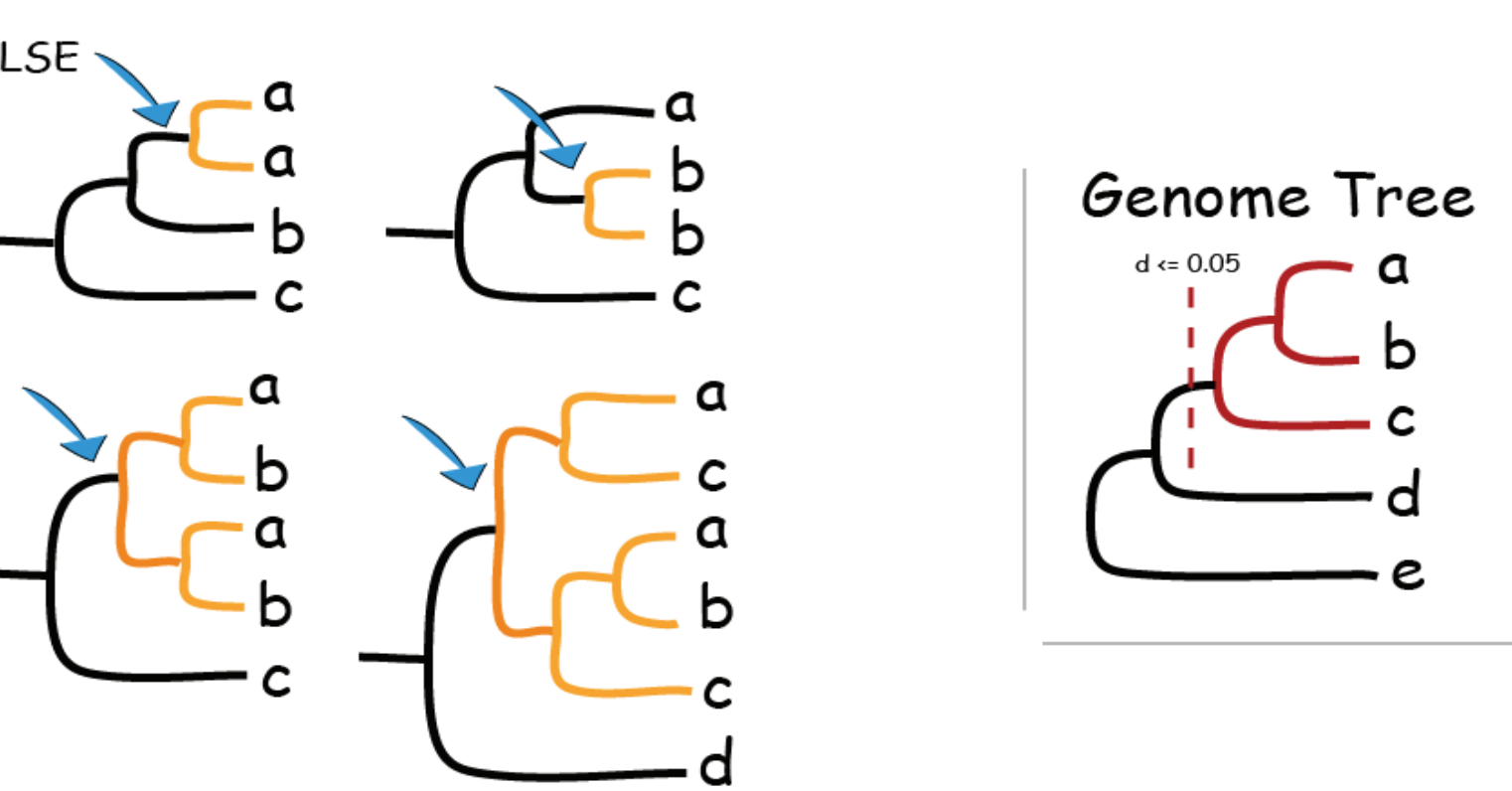
Lineage-specific expansion (LSE) plays a vital role in how prokaryotes gain new gene functions and adapt to their environments. To uncover the mechanisms behind LSE, we identify genes that arise from LSE by constructing phylogenetic trees of protein families across 400+ bacterial genomes. We found that LSE genes tend to cluster on the chromosomes and form hyper LSE regions. Such regions could not be explained solely by operon duplication. The locations of these hyper LSE regions are often remarkably conserved among closely related strains, even though the gene content may not be conserved. Furthermore, these hyper LSE regions frequently overlap with clusters of mobile genetic elements (MGE) and strain-specific genomic islands. We hypothesize that the majority of large strain-specific gene duplications are mediated by MGE and are concentrated in regions prone to site-specific MGE-driven recombinations. And the same regions for the same reason are more susceptible to phage integration and to inter-genomic information exchange.

Methods

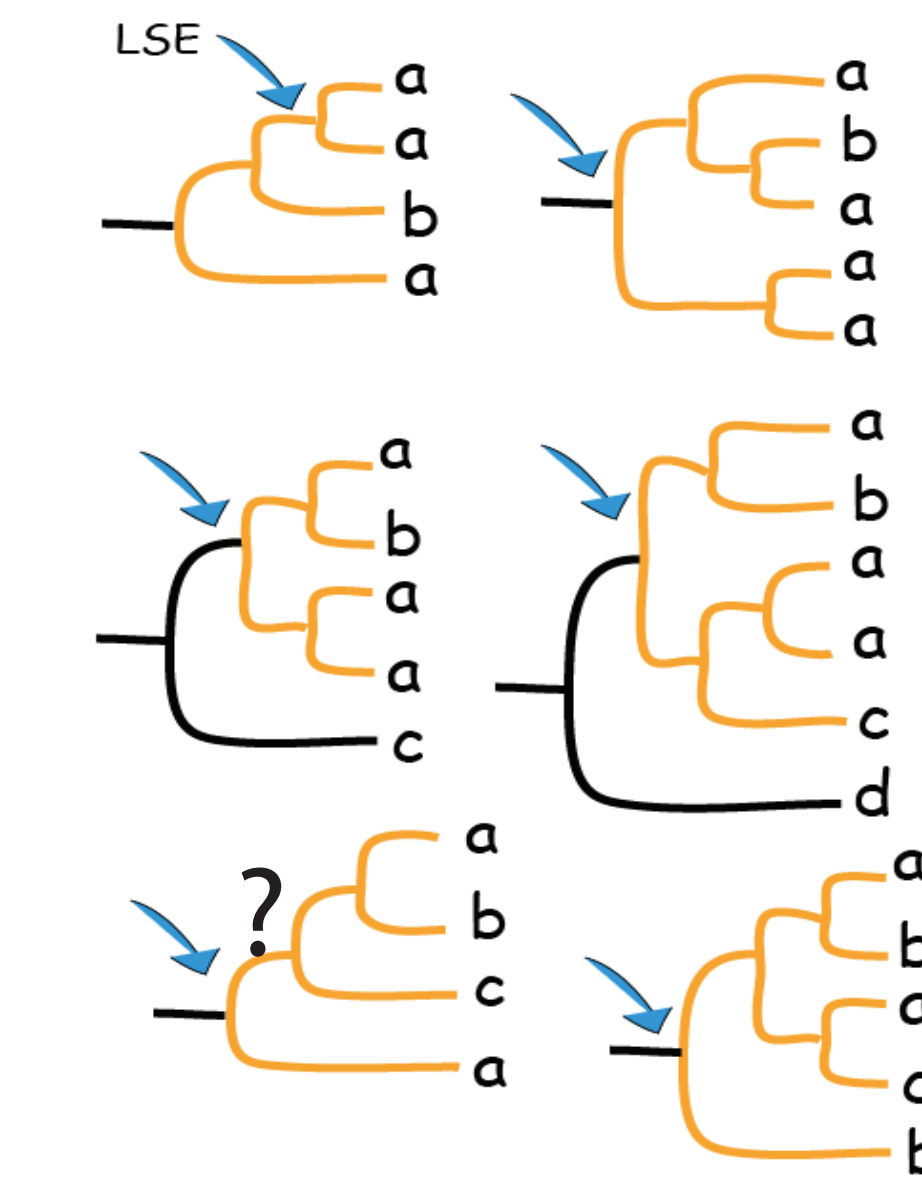
Identification of LSE

A LSE is identified by examining individual gene trees and finding the last common internal tree node among paralogous genes:

- At least one gene duplication event is observed at the LSE node level.
- To identify only recent events, all children of the LSE node must come from a single phylogenetic group of closely related species (see below).
- Complicated gene trees require more sophisticated method to resolve, especially in cases when the LSE node includes genes from multiple genomes, but some of these genomes are present as single-copy genes.



Complicated Gene Trees



Individual Gene Trees

Definition of Closely Related Phylogenetic Groups

Closely related genomes are grouped if pair-wise divergence less than 5% (or $d <= 0.05$) based on a Maximum-Likelihood tree constructed from a concatenated multiple sequence alignment of 70 highly conserved proteins.

Genome list at $d <= 0.05$ in *E. coli* K12 group

- 855 Erwinia carotovora subsp. atroseptica SCRI1043
- 855 Escherichia coli S36
- 855 Escherichia coli APEC O1
- 855 Escherichia coli CFT073
- 855 Escherichia coli K12
- 855 Escherichia coli O157:H7
- 855 Escherichia coli O157:H7 EDL933
- 855 Escherichia coli UT89
- 855 Escherichia coli W3110
- 855 Photorhabdus luminescens subsp. laumondii TTO1
- 855 Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67
- 855 Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150
- 855 Salmonella enterica subsp. enterica serovar Typhi
- 855 Salmonella enterica subsp. enterica serovar Typhi Ty2
- 855 Salmonella typhimurium LT2
- 855 Shigella boydii Sb227
- 855 Shigella dysenteriae Sd197
- 855 Shigella flexneri 2a str. 24577
- 855 Shigella flexneri 2a str. 301
- 855 Shigella flexneri 5 str. 8401
- 855 Shigella sonnei Ss046
- 855 Yersinia pestis Antiqua
- 855 Yersinia pestis biovar Medievalis str. 91001
- 855 Yersinia pestis CO92
- 855 Yersinia pestis KIM
- 855 Yersinia pestis Nepal516
- 855 Yersinia pseudotuberculosis IP 3295

By grouping organisms at $d <= 0.05$, the *Bacillus* genus are separated into 4 groups.

- 284 Bacillus anthracis str. 'Ames Ancestor'
- 284 Bacillus anthracis str. Ames
- 284 Bacillus anthracis str. Sterne
- 284 Bacillus cereus ATCC 10987
- 284 Bacillus cereus ATCC 14579
- 284 Bacillus cereus ZK
- 284 Bacillus thuringiensis serovar konkukian str. 97-27
- 288 Bacillus licheniformis DSM 13
- 288 Bacillus subtilis
- 291 Bacillus halodurans
- 292 Bacillus clausii KSM-K16

Terminology

LSE: the last common internal tree node among paralogs observed at $d <= 0.05$. In this study, it refers to the genomic full-length LSE unless noted otherwise.
LSE size: number of genes in LSE normalized by number of genomes observed in the LSE node.
LSE genes: genes from the same LSE node.

Types of LSE

MGE LSE: LSE that includes mobile genetic elements, such as transposases, IS elements, integrases, etc.
Prophage LSE: LSE that includes prophage genes that are not MGE.
Phage-like LSE: LSE that includes other phage homologs that are not prophage LSE nor MGE.
Plasmid LSE: LSE that includes plasmid genes that are not in the above categories.
Domain LSE: LSE that includes chromosomal genes that are not in the above types and of which duplication occurs at the domain level.
Full-length LSE: the rest of the chromosomal LSEs, of which duplication occurs at the full-length protein

Results

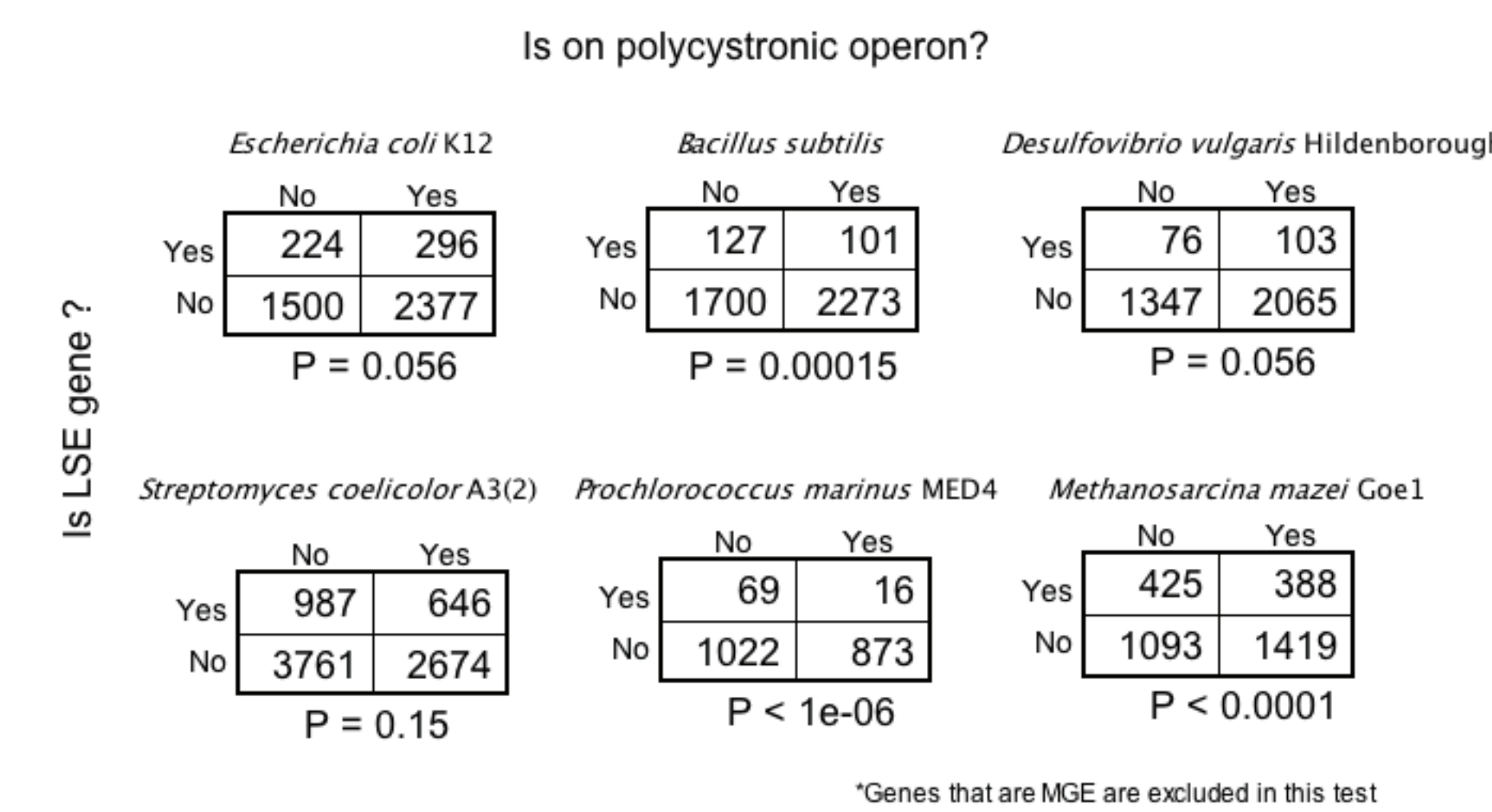
Overview of LSE

Summary of all identified LSE by categories.

	LSE Genes	LSE	Genes per LSE	Genomes per LSE	LSE size (genes/genomes)
MGE	22,809	2,137	10.67	1.86	5.74
Prophage	9,063	892	10.16	4.19	2.43
phage-like	5,920	903	6.56	2.15	3.05
Plasmid	9,718	1,938	5.01	1.80	2.78
Genomic Domain	32,457	5,738	5.66	2.12	2.67
Genomic Full-length*	46,661	10,491	4.45	2.09	2.13
Total	126,628	22,099	N/A	N/A	N/A

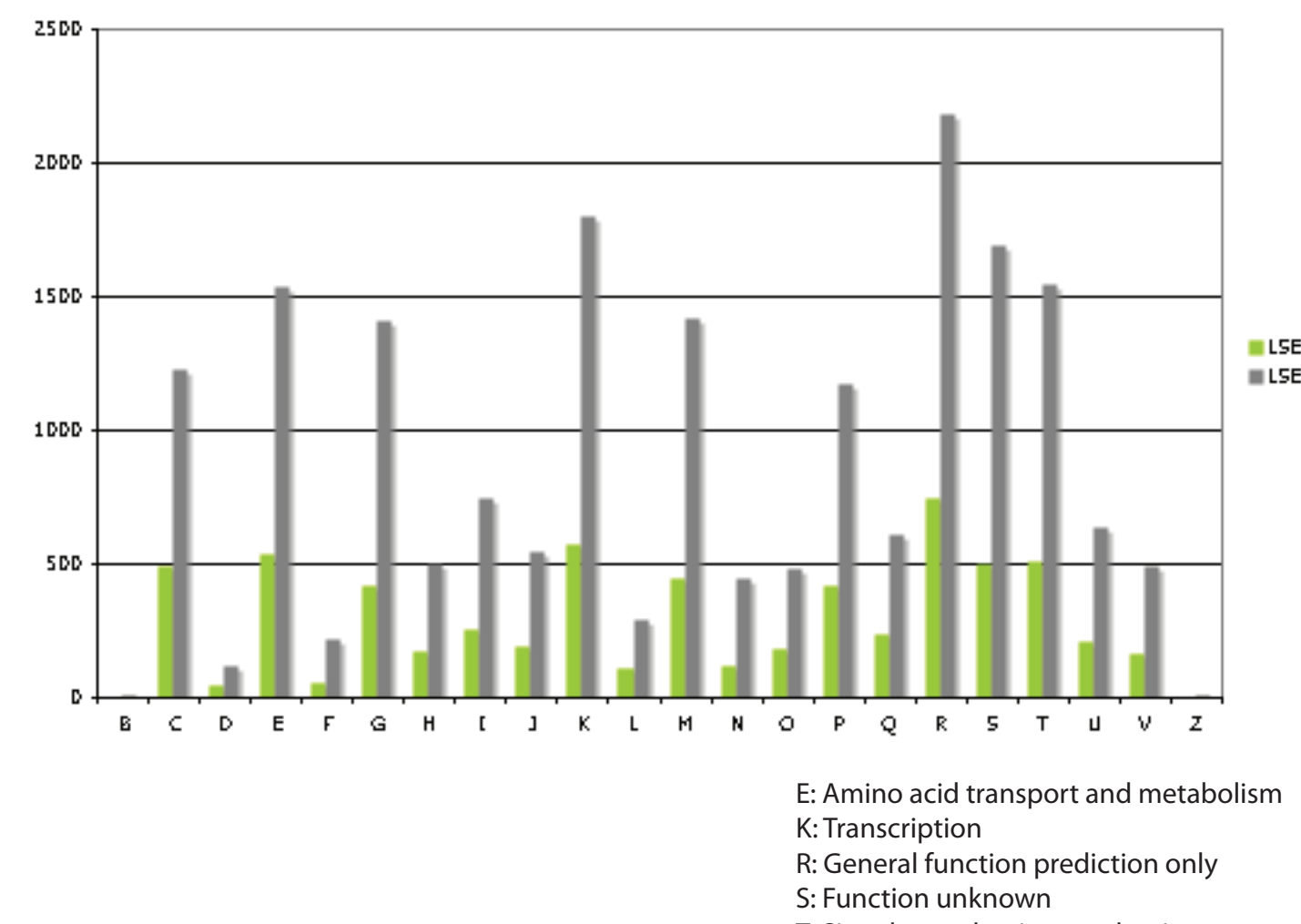
* about 1500 FL-LSEs present only once in some genomes.

LSE vs Operons



There are many examples of operon duplication. However, based on this simple Fisher Exact test, LSE genes either tend to be single-gene transcripts ($p < 0.01$) or do not have a significant preference comparing to non-LSE genes.

COG functions in LSE

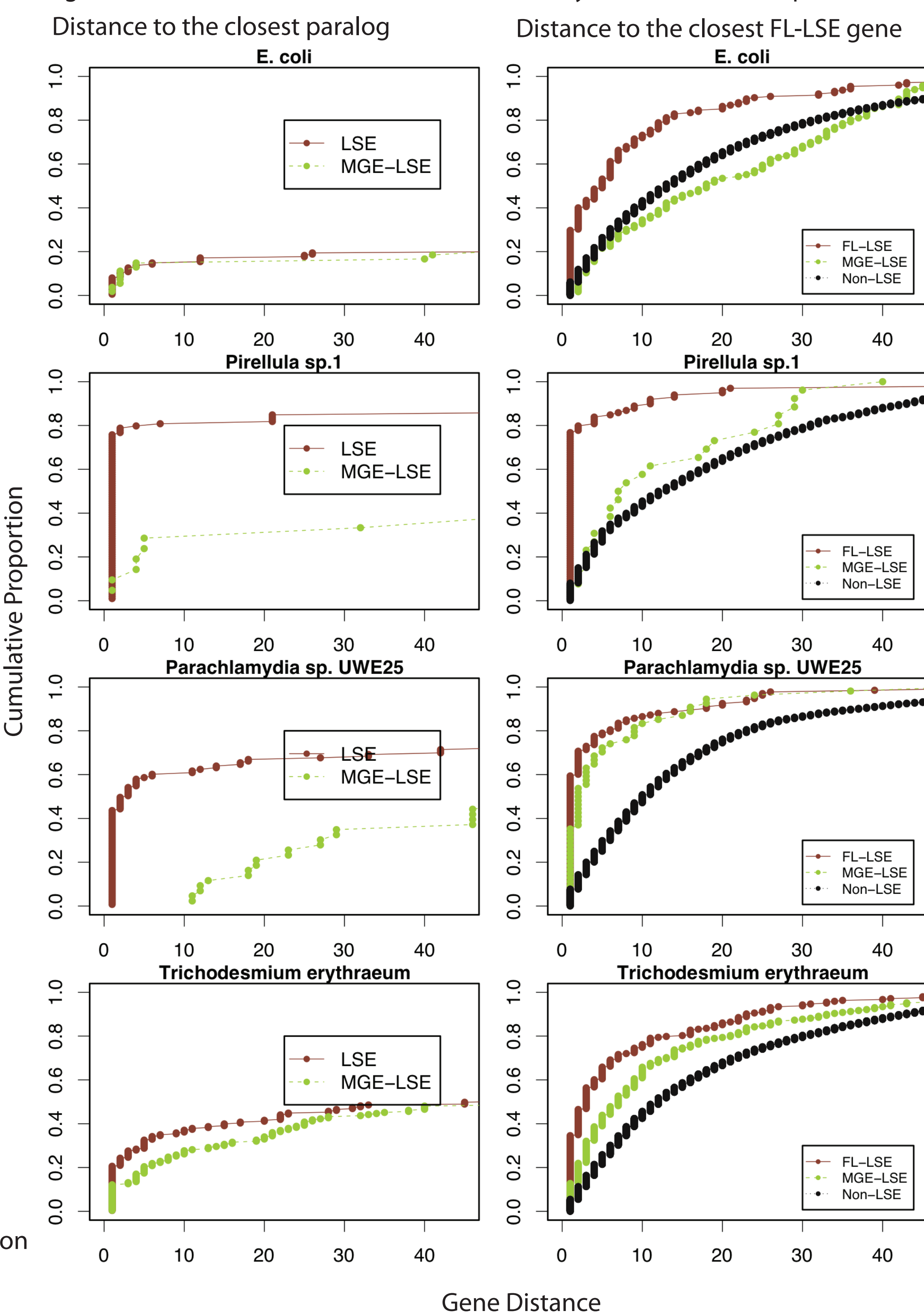


Large LSEs tend not to have COG

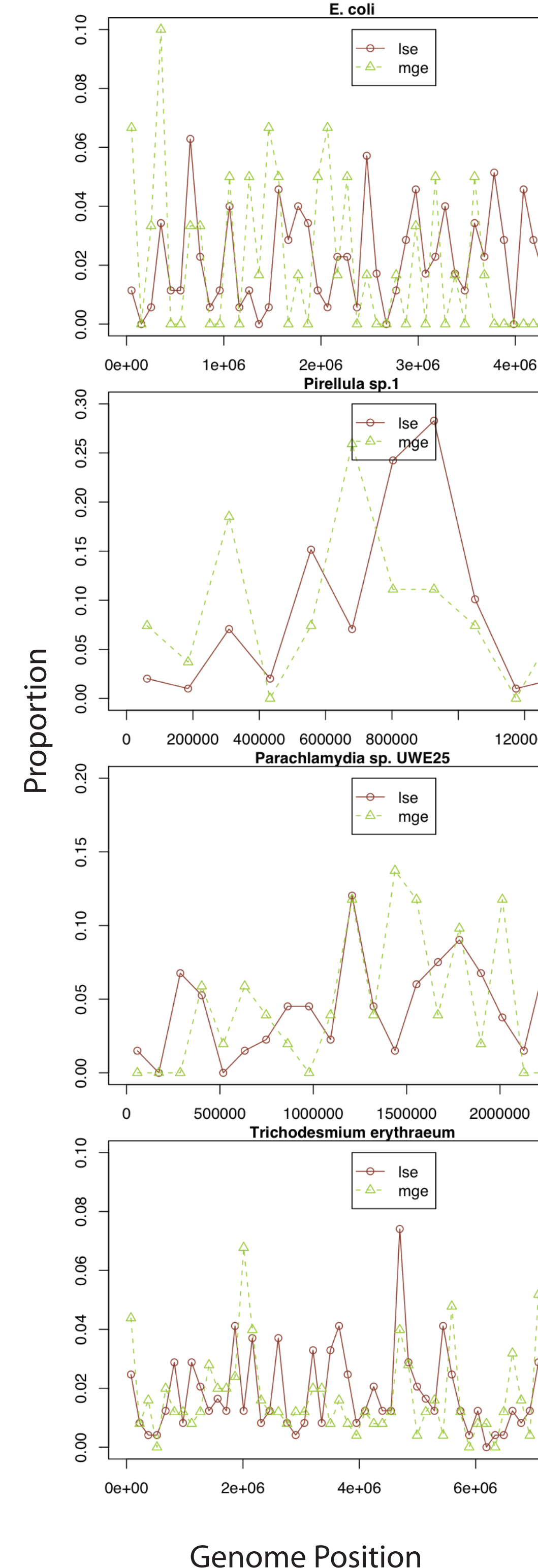
	LSE size > 2	LSE size < 2
has cog	798	4901
no cog	896	2024

$p < 1e-35$

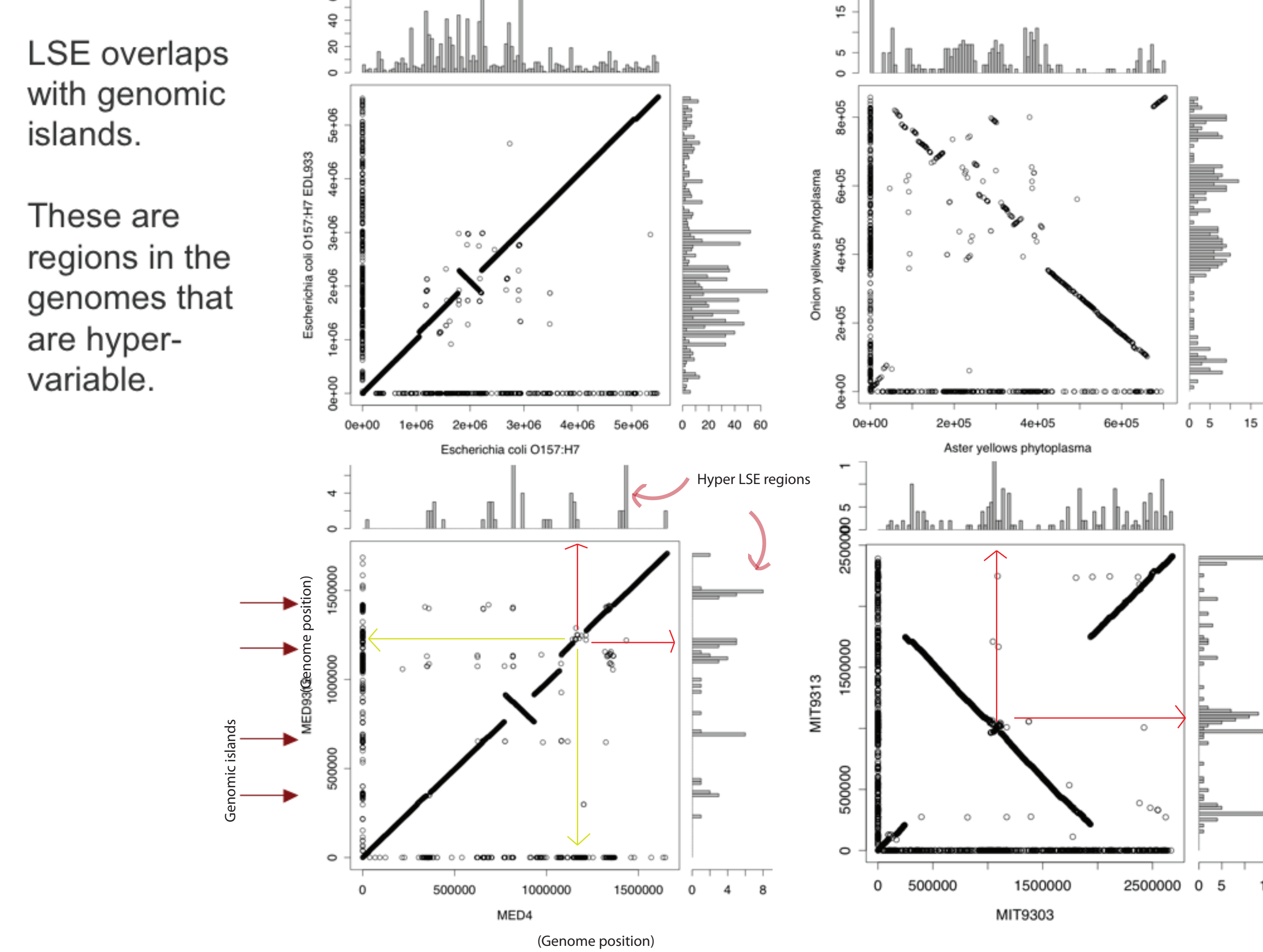
LSE genes cluster in the chromosome but not solely due to tandem duplication



Distribution of Genes on Genome



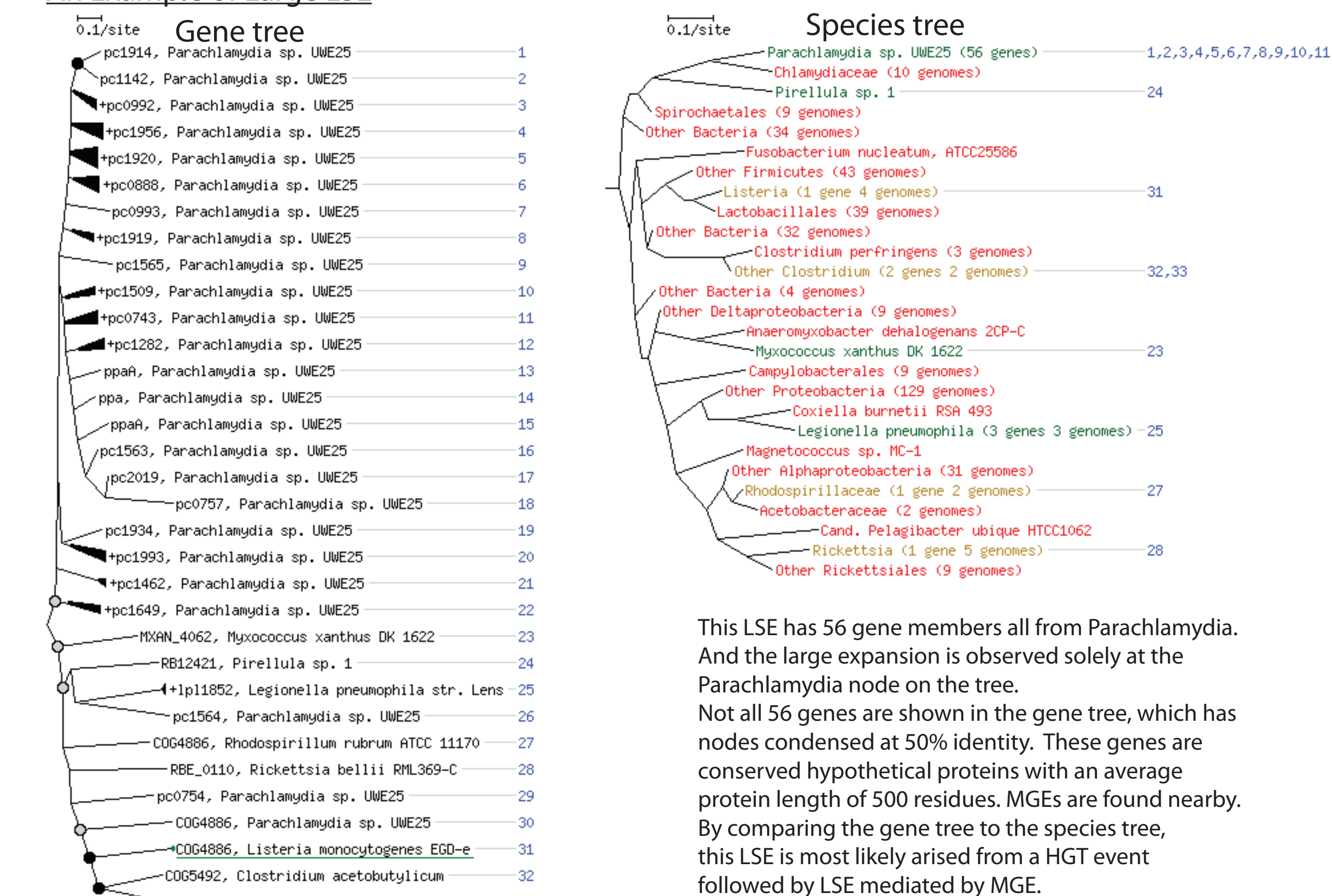
Pair-wise genome alignments reveal LSE hot spots



LSE overlaps with genomic islands.

These are regions in the genomes that are hyper-variable.

An Example of Large LSE



Conclusions

In this comparative genomic survey, we examine the types, functions, evolutionary history and mechanisms of LSEs. We have found that large LSE events are rare, out of more than 10,000 FL-LSE, only 16 of them have a size > 15 genes. Although a couple COG functions are enriched in LSEs, a more significant observation is that larger LSEs (size > 2) tend not to have COG assignments and over 60% of them are orphan genes or arise from recent HGT. In addition, many LSE genes seem to co-localize with MGEs or phage genes, especially those with a large size although more statistical tests are necessary to confirm this. We hypothesize that large LSEs are driven by MGEs or are uncharacterized MGEs themselves.