Unitary Pseudogenes in Bacterial Genomes: Lifestyle and Gene Loss

Tal Dagan[1] and Dan Graur [2]

[1]Department of Zoology, Tel Aviv University, Ramat Aviv 69978, Israel; [2]Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, USA

Genome Miniaturization

The question of "use and disuse" in evolution is as old as the discipline itself. At least one unambiguous rule can be deduced from the effects of disuse at the molecular level: A drastic reduction in genome size (genome miniaturization) is invariably associated with loss of function. In particular, parasitic or endosymbiontic lifestyles were found to effect genome size profoundly.

In the following, we characterize genome size reduction in 120 bacterial genomes due to gene loss by using unitary pseudogene data.

Glossary

Commensalism: Symbiosis in which one symbiont (the *commensal*) derives benefit from the association, and the other (the *host*) neither benefits nor is harmed.

Facultative: Optional, referring to the ability of an organism to adopt to an alternative lifestyle.

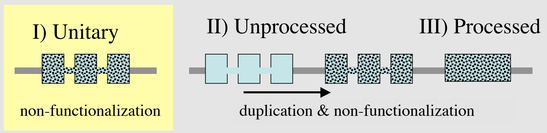
Free living: Living without being directly dependent on another organism.

Mutualism: Symbiosis in which both symbionts derive benefit from the association.

Obligate: An essential attribute of an organism. For example, an obligate parasite can only grow as a parasite.

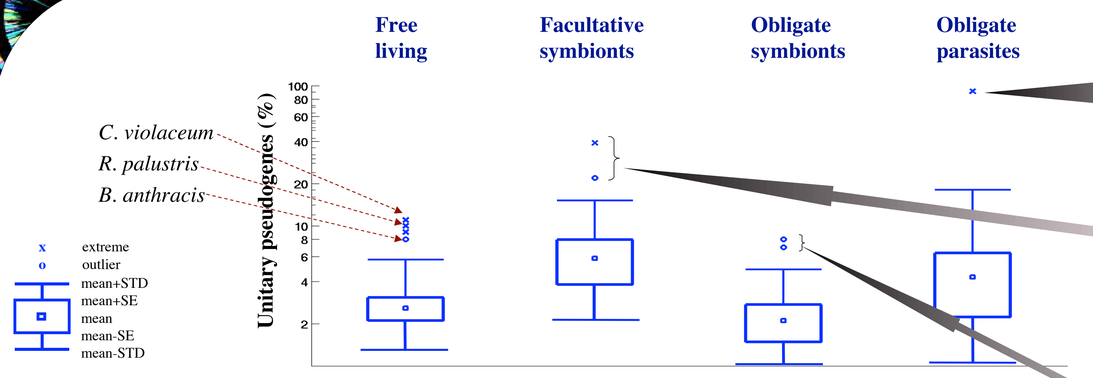
Parasitism: Symbiosis in which one symbiont (the *parasite*) benefits at the expense of the other (the *host*).

Pseudogene: Functionless duplicate of functional gene.



Symbiosis: A stable condition in which two organisms (*symbionts*) live in close physical association.

Bacterial Lifestyles



N (genomes)	**41**	**20**	**19**	**44**
Genes (COGs)	**45,444**	**33,943**	**20,986**	**38,802**
Unitary pseudogenes	**106**	**117**	**40**	**189**

- **In an analysis of 120 completely sequenced bacterial genomes, 452 unitary pseudogenes were found.**
- **The frequency of the unitary pseudogenes was found to be significantly dependent on bacterial lifestyle.**
- **Obligate parasites have many more unitary pseudogenes than other bacteria.**
- **Facultative symbionts have a higher-than-expected frequency of unitary pseudogenes. This estimate, however, may be an overestimation due to the inclusion of *Shigella flexneri*.**

Mycobacterium leprae

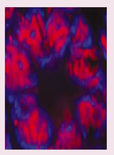
An endosymbiontic parasite that underwent extensive genome reduction. About half of its genes were rendered nonfunctional. About 45% of its pseudogenes are derived from genes specifying metabolic functions.

Shigella flexneri

An exceptional facultative symbiont, whose genome resembles that of obligate parasites. It has been hypothesized that it is "on the way to become one."

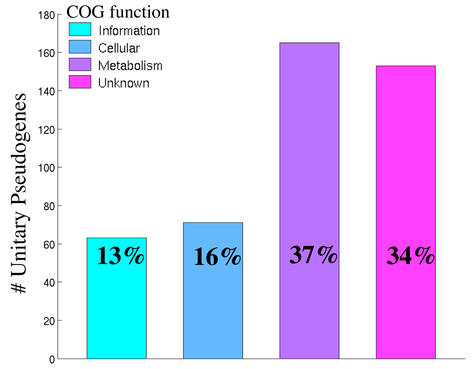
Buchnera aphidicola

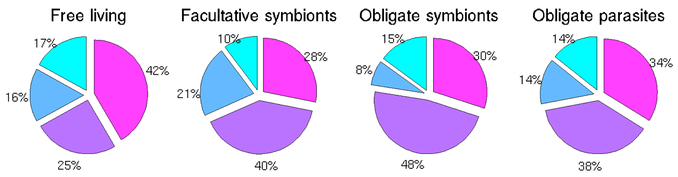
An intracellular mutualist obligate-symbiont of aphids. The fact that different strains exhibit different degrees of genome reduction, led to the hypothesis that their adaptation to different aphid species is an ongoing process.



Molecular Function

- **The frequency of unitary pseudogenes was found to be dependent on genetic function.**
- **Function loss is more frequent in genes performing metabolic functions than in genes involved in information transfer or cellular processes.**
- **Genes encoding metabolic functions were found to be prone to pseudogenization in all lifestyles, although the fraction of pseudogenes derived from genes involved in metabolic functions is smaller in free-living bacteria.**
- **The frequencies of the lost functions were found to be similar among the different lifestyles, except for the case of free-living bacteria and facultative symbionts.**





Summary

- **The frequency of unitary pseudogenes can be used to characterize genome reduction.**
- **Parasitic bacteria tend to lose gene functions.**
- **Genes encoding metabolic functions are more prone to pseudogenization than other genes.**
- **Surprisingly, the ratio of unitary pseudogenes to functional genes does not differ significantly between free-living and obligate-symbiotic bacteria.**
- **Our method for detection of unitary pseudogenes assumes vertical transmission of genes. Horizontal gene transfer may be a source of bias.**

Search algorithm

Identification of missing genes
An extended COG (Cluster of Orthologous Genes) database was used to identify missing functions. Bacteria that were not represented in a certain COG were marked as candidates for search of orthologous unitary pseudogenes.

Search
We chose a query gene for each missing bacterium in a COG. The chosen query gene was from the closest related bacterium according to rRNA distances. The query gene was BLASTed against the candidate bacterium genome.

Analysis of search results
Successful ($e < 10^{-4}$) BLAST hits with no overlap to known genes were collected. Nearby hits (30 bp) were joined. The DNA sequence of each such hit was aligned to the DNA sequence of the query gene using CLUSTALW and to the protein sequence of the query-gene product using WISE2. d_n/d_s ratios were calculated using PAML.

Filtering
Unitary pseudogenes were defined as such if:
- DNA sequence similarity to the functional gene was above 50%.
- Protein sequence similarity was above 35%.
- Length was at least 35% of the query gene.
- Signs of nonfunctionality, e.g., premature STOP codons, frameshifts, or lack of purifying selection ($d_n/d_s = 1$), are evident.