

Response to Nelson-Sathi et al. (Nature 517, 77-80)

Gene acquisitions from bacteria at the origins of major archaeal
clades are vastly overestimated

Mathieu Groussin¹, Bastien Boussau^{2,3,4}, Gergely J. Szöllősi⁵, Laura Eme⁶, Manolo Gouy^{2,3,4},
Céline Brochier-Armanet^{2,3,4}, Vincent Daubin^{2,3,4}

¹ – *Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA,
USA*

² – *Université de Lyon, 69361 Lyon, France*

³ – *Université Lyon 1, 69622 Villeurbanne, France*

⁴ – *CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 69622 Villeurbanne, France*

⁵ – *ELTE-MTA “Lendület” Biophysics Research Group, Pázmány, Budapest, Hungary*

⁶ – *Laura’s affiliation*

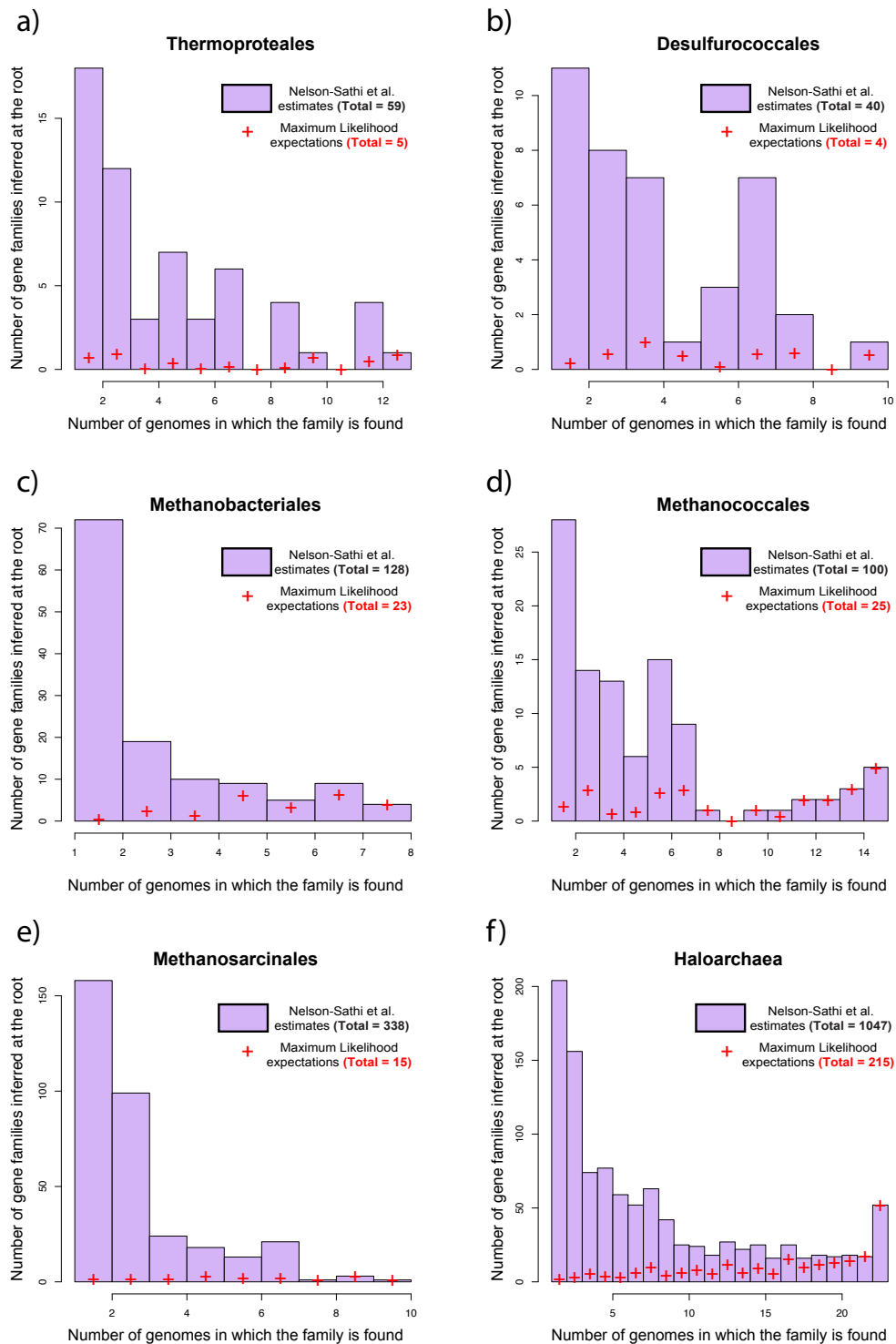


Figure 1: **Family size and gene gains for the “import” set at the origin of archaeal phyla.** The 6 phyla for which Nelson-Sathi et al. (2015) [NS] claim a strong signal for HGTs from bacteria at their origins are shown. NS estimates are represented in purple. Maximum Likelihood expectations of gains at the origin for the same datasets are represented by red crosses. Expectations of gains were calculated from branch-specific rates of gene gain and loss estimated from the 25,762 gene families of NS, but we report only the results for the 2,264 genes from the “import” set of NS. Most of the families inferred to be gained at the origin of each phylum by NS are distributed over 2 or 3 archaeal genomes only. ML estimates show that only a small fraction of gene families were actually acquired at the origin of each phylum.

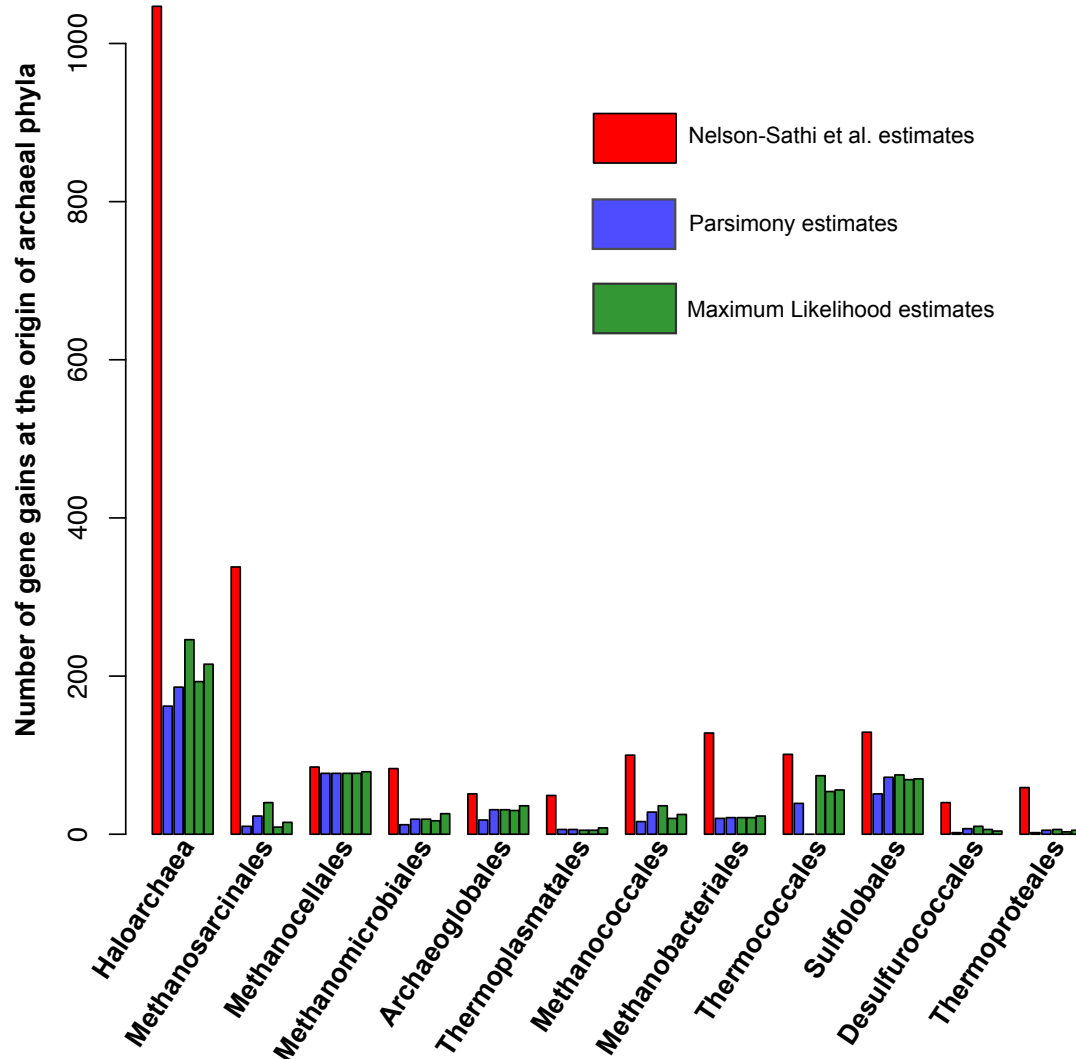


Figure 2: Consistency of gene family ancestral gain estimations on the “import” gene set. Along the archaeal reference tree provided by Nelson-Sathi et al. (2015) [NS], both Parsimony (green) and Maximum Likelihood (blue) inferences of ancestral gains were computed with Count (Csurs and Miklos, 2006; Csurs, 2010), and compared with the estimates of NS. Wagner parsimony with a large range of gain/loss costs combinations were tested, yielding similar estimations. Here are represented the results with a gain cost of 2 and 4, from left to right respectively. In ML, ancestral branch-specific gain expectations were computed from rates of gene gain and loss estimated either from the 2,264, 3,315 or 25,762 gene family sets defined by NS, from left to right respectively. All analyses are consistent and strongly contradict the scenario of massive gene acquisitions among these 2,264 families at the origin of archaeal phyla, claimed by NS

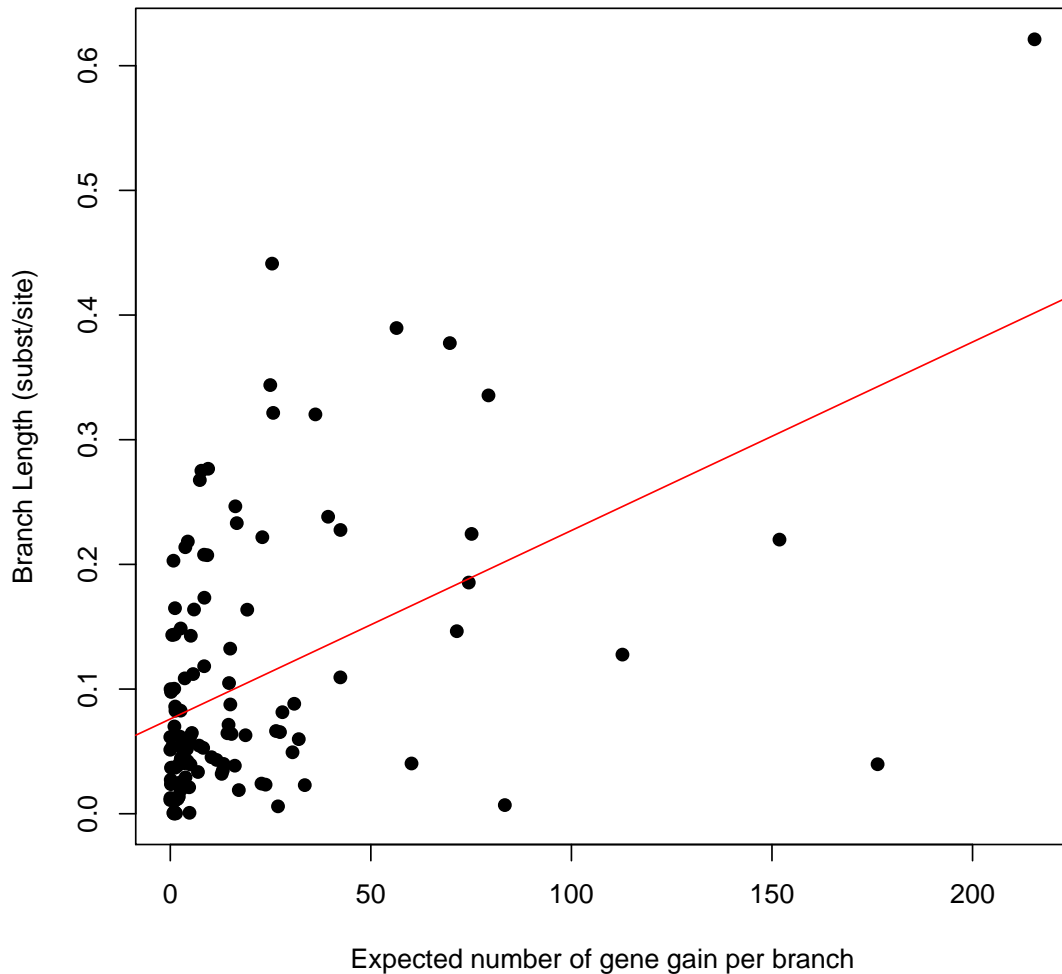


Figure 3: **Longer branch lengths are associated with higher number of gene gains.** On the x axis, the expected numbers of gene gains per internal branch are reported. Gene gains are inferred with Count, see Methods. On the y axis, branch lengths of the reference archaeal species tree reconstructed by NS are plotted. Only internal branches within MALs are plotted since the selection of gene families by NS restricts the analysis of evolutionary events within MALs. The spearman rank correlation test was used, $\rho = 0.43$, $p - value < 10^{-5}$. The top-right point corresponds to the root branch of Haloarchaea. The correlation remains highly significant when this point is removed from the data: $\rho = 0.41$, $p - value < 10^{-5}$.

References

Csurs M. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 26:1910–1912.

Csurs M and Miklos I. 2006. A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. *In Proceedings of the Tenth Annual International Conference on Research in Computational Molecular Biology (RECOMB)* Vol. 3909 of Springer Lecture Notes in Bioinformatics, Heidelberg:206–220.

Nelson-Sathi S, Sousa FL, Roettger M, Lozada-Chvez N, Thiergart T, Janssen A, Bryant D, Landan G, Schnheit P, Siebers B, McInerney JO, and Martin WF. 2015. Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517:77–80.