1 **Genome-wide experimental determination of barriers to horizontal gene transfer**

Rotem Sorek[1,2], Yiwen Zhu[2], Christopher J. Creevey[3], M. Pilar Francino[1], Peer Bork[3] and Edward M. Rubin[1,2]

[1] DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, USA

[2] Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

[3] European Molecular Biology Laboratory, Meyerhofstrasse 1, 69012 Heidelberg, Germany

Abstract

Horizontal gene transfer, in which genetic material is transferred from the genome of one organism to another, has been investigated in microbial species mainly through computational sequence analyses. To address the lack of experimental data, we studied the attempted movement of 246,045 genes from 79 prokaryotic genomes into *E. coli* and identified genes that consistently fail to transfer. We studied the mechanisms underlying transfer inhibition by placing coding regions from different species under the control of inducible promoters. Their toxicity to the host inhibited transfer regardless of the species of origin and our data suggest that increased gene dosage and associated increased expression is a predominant cause for transfer failure. While these experimental studies examined transfer solely into *E. coli*, a computational analysis of gene transfer rates across available bacterial and archaeal genomes indicates that the barriers observed in our study are general across the tree of life.

Text

The rapidly accumulating sequenced genomes of bacteria and archaea reveal the role of horizontal gene transfer (the non-sexual exchange of genes across hierarchal boundaries) in shaping non-eukaryotic genomes (*1, 2*). Gene exchange has been documented for nearly all types of genes and at all phylogenetic distances (*3*). These observations have sparked debates about whether microbial genes can be used for phylogenetic classification, because the proposed lack of barriers to gene transfer between genomes suggest that a tree-like classification of microorganisms might be impossible (*4-5)*.

Identifying the limitations of gene transfer is hampered because nearly all transfer events have been inferred on the basis of sequence analysis of microbial genomes. Computational approaches, including detection of nucleotide or codon compositional biases and atypical distribution of genes, identify signatures of transfer events predicted to have occurred millions of years ago (*6*). Based on such studies specific categories of genes were suggested as less prone to transfer, and hence potentially useful as phylogenetic markers (*7, 8*), but the validity of this idea relies nearly exclusively on computational evidence (*1*). The paucity of experimental and quantitative data on horizontal gene transfer, therefore, impedes our ability to understand the extent and limitations of this phenomenon.

Natural gene transfer is largely mediated via naked DNA uptake (transformation), viruses (transduction), and plasmids (conjugation) (*9*). When a microbial genome is being sequenced, multiple copies of the genome are randomly sheared into overlapping fragments of DNA (typically to libraries sized 3kb and 8kb), and plasmids containing the cloned fragments are transformed into an *E. coli* cell (*10*). The ends of the cloned fragments are then sequenced, and overlapping sequences are used for genome assembly. As cloned fragments 2 contain the full set of genes belonging to the sequenced organism, microbial genome sequencing can be viewed as a large-scale experiment in horizontal gene transfer to *E. coli*, where each gene in a given genome undergoes multiple transfer attempts to the host with an extra-chromosomal plasmid. In the course of nearly all prokaryotic sequencing projects, a small fraction of the

organism's genome fails to clone in *E. coli*, resulting in sequence gaps. The sequence for these gaps is acquired via a clone-independent stage termed 'finishing', eventually producing an unbroken sequence of the organism's genome (*11*).

We explored the limits to horizontal transfer by studying the nature of uncloneable ("untransferable") genomic regions. Of the 85 finished microbial genomes with accessible original sequence reads, we selected 79 (including 75 bacterial and 4 archaeal) with sufficient clone coverage for detailed analysis (SOM text, Table S1). We used the original sequencing data to map the clone positions on these genomes. Overall, this dataset included 1,873,649 clones spanning over 8.9 billion bases of genomic DNA fragments successfully transferred into an *E. coli* host.

We next explored the transfer of the individual genes residing in the 79 analyzed genomes. For each of the 287,884 annotated genes contained in these genomes we calculated the number of clones fully spanning the gene on the basis of the mapped clone positions. We considered only genes 1.5kb or less (246,045 genes, representing 85% of all annotated genes), as larger genes are less likely to be covered to their full length by multiple clones. The average number of clones covering each of these 246,045 genes to its full length was 22.57, indicating that each gene underwent, on average, more than 22 independent transfer attempts to the host.

We used the clone coverage distribution to identify genes untransferable into the *E. coli* host. To exclude the possibility that cloning biases are random or human-introduced, we compared clone coverage among genomes of closely related species. These genomes presented relatively similar coverage patterns, with the same sets of orthologous genes from several different organisms absent from sequenced clones, supporting that clone deficiency is largely gene-dependent. Comparison of four *Shewanella* species offers an example for the high reproducibility of clone deficiency: 73 of 99 (74%) *Shewanella sp. MR4* genes found to be uncloned into *E. coli* were also unclonable when transferred from at least one of the three other Shewanella species examined (Fig. 1).

Of the genes inspected, we recorded 1,402 instances (642 different genes) in which a gene with a Clusters of Orthologous Groups (COG) annotation was not fully represented in any single clone, and was marked as untransferable to *E. coli* (with an estimated false positive prediction rate of 0.9%-1.3%; see SOM). In 1,064 (76%) of these events, the same gene was untransferable to *E. coli* from two or more different genomes. Sixty one genes (477 events, 34% of total events) were untransferable from 5 or more different genomes into *E. coli* (Fig. 2). The high transfer failure rate for certain gene families across several genomes further suggests that specific genes, rather than the experimental protocol or random biases, cause the lack of horizontal transfer.

While gene transfer in the wild is believed to be mediated via the transfer of single as well as multiple copies of the DNA, the cloning vectors used in most small-insert sequencing libraries exist in 20-100 copies per cell (*14-15*). We examined the impact of single versus multiple copy transfers by studying the subset of 35 sequenced genomes where, in addition to the small insert libraries, large fragments (35kb) of the microbial genome were propagated in fosmids, which typically exist in a single copy per *E. coli* cell (*16*)(Table S1). In 124 out of 483 (26%) uncloned genes in these genomes, the genes were also covered by zero (22%) or 3 statistically fewer fosmids than expected (4%) (Fig. S1; SOM). This suggests that a significant portion of the observed transfer deficiency is not solely due to high copy number.

We selected 40 genes that resisted transfer from two or more genomes and were able to clone the coding regions of 39 of these genes into an expression vector system that strongly suppresses the expression of the cloned gene in the absence of an expression inducer (IPTG) (Table S2; SOM text). In the absence of inducer bacterial growth was observed; however,

upon induction of expression, 32 of the 39 genes (82%) inhibited *E. coli* growth, indicating that the products of these genes are toxic to the host (Fig. 3; Table S2). This explains the lack of transfer observed in the genome sequencing data.

Although we identified genes that were transfer-resistant from a wide range of prokaryotes, no single gene was untransferable among all genomes examined (reflected by the absence of a horizontal line of black squares running across the complete list of organisms in Fig. 2). This was coupled with the observation that the resistance to transfer of genes tended to be similar among closely related organisms (Fig. 2). A possible explanation is that promoters (usually found adjacent to the gene and hence transferred with it) from some species may be recognized by the host *E. coli* transcriptional machinery and drive the expression of the foreign gene leading to growth inhibition, while promoters of other species are not active in the *E. coli* cell. Indeed, sequences from Firmicutes were previously shown to drive strong expression when tested as promoters in *E. coli (17)* which is consistent with Firmicutes having high numbers of transfer resistant genes (Fig. 2). GC-rich genomes tended to have fewer untransfeable genes, again consistent with observations that promoters recognized by *E. coli* are GC-poor *(18)*. Therefore, we predicted that some of the genes catalogued as nontoxic would be toxic if their promoters were active in *E. coli*.

To test this we examined two relatively transfer-resistant genes, ribosomal protein L4/L1e (COG0088) and ribosomal protein S12 (COG0048). Each of these genes did not transfer in 9 of 79 genomes (Fig. 2). We isolated the coding sequences of these genes from 31 microorganisms for which genomic DNA was readily obtainable, including 26 organisms where transfer resistance had not been observed on the basis of genome sequencing, and cloned them into the inducible expression system described above. Clones holding these genes grew normally in the absence of inducer. However, growth inhibition was observed in 53/62 (85%) clones when expression of the cloned gene was induced by low IPTG concentrations (100uM-600uM) and in 57/62 clones (92%) in higher (800uM) IPTG (Fig. 4A, Table S3). Such a high frequency of growth inhibition was not observed in a survey of 15 randomly selected putative negative control genes, of which 2/15 (13%) and 7/15 (47%) inhibited growth in low and high IPTG, respectively (SOM text; Table S4). These results suggest that some of the genes we identified are almost universally toxic when expressed, and therefore appear to face a near absolute, phylum-independent barrier to horizontal transfer into *E. coli*.

We compared the COG functions of the 61 genes we identified as highly untransferable (those untransferable from 5 or more genomes) to the COG functions of all genes in our data set. The highly untransferable genes were significantly enriched in genes involved in ribosomal structure and translation ($P<2e-09$, Fisher's exact test corrected for multiple testing; Fig. S2). This observation is consistent with previous computational analyses that suggested that genes involved in translation tend to be under-represented in genes postulated to have undergone horizontal transfer (*7, 8*). The toxicity of ribosomal proteins observed here possibly stems from an incompatibility with the *E. coli* molecular machinery, as they have multiple interactions within the ribosome (*7*). We found that ribosomal proteins that resisted transfer from a large number of genomes also had more surface area in contact with the ribosomal RNA ($P=0.023$, Spearmans test; Fig. S3). 4

An additional possible mechanism for explaining some of the observed transfer resistance is intolerance of the host to increased dosage of the transferred gene in addition to the endogenous homolog. In order to test this hypothesis, we examined data from the *E. coli HS (19)* genome project, where clones containing fragments of the *E. coli HS* genome were transferred into a standard *E. coli* sequencing strain (*DH10B*). Despite the near identity between the transferred genes and the host genes, 43 *E. coli HS* genes (all of them conserved

in *E coli K12* in >98% identity) could not be cloned into the host *E. coli*. Therefore, this subset of genes cannot be tolerated in high dosage. Thirty-four (80%) of these 43 genes were also untransferable to *E. coli* from at least one additional foreign genome (Fig. 2) suggesting that their lack of transfer was also due to dosage intolerance. Moreover, 32% of the genes that were untransferable to *E. coli* from five or more genomes were universal single copy genes, never duplicated in any of the genomes we tested (compared to 3% universal single copy genes out of the entire gene population), providing additional support that an increased dosage and the associated increased expression of these genes is likely detrimental to most microbes (Fig. 2).

While our analysis of the experimental data from 246,045 genes transferred to *E. coli* suggests that there is a specific set of genes that are untransferable regardless of their genome of origin, it does so for a single recipient organism, the *E. coli* host. To explore whether these results are general, and whether these genes are untransferable to other recipient species, we used a tree-based computational method to predict gene transfer in 191 sequenced genomes across the entire tree of life (*20*) (SOM text). We found a strong correlation (P=0.008, Wilcoxon Mann-Whitney Test) between genes that we experimentally characterized as untransferable to *E. coli* and single copy genes that were computationally predicted to be less transferred across the tree of life (Fig. S4). These results suggest that the genes we experimentally characterized in a single host are generally transfer-resistant among most bacteria and archaea, and would be expected to be predominately vertically transmitted in prokaryotes.

Our results suggest there are universal gene transfer barriers, regardless of whether transfer occurs among closely or distantly related microorganisms, and that these barriers are associated with toxicity of the transferred gene to the host. The number of untransferable genes identified in this study probably reflects a lower limit, as the genes we studied were physically forced into the host and plasmid maintenance was aggressively selected for with antibiotics, and additional natural barriers were not taken into account. In addition, transfer-resistant genes larger than 1.5kb, as well as toxic genes whose promoters are not active in *E. coli*, escaped our detection. Our observation that many untransferable genes appear universally in a single copy (never duplicated in any sequenced bacteria) suggests that the increased expression of these genes inhibit growth in a wide range of bacteria. Accordingly, molecules that would increase the expression of any of these genes might function as broad-range antibiotics.

References

1. J. P. Gogarten, J. P. Townsend, *Nat. Rev. Microbiol.* **3**, 679 (2005).
2. T. Dagan, W. Martin, *Proc. Natl. Acad. Sci. U. S. A.* **104**, 870 (2007).
3. J. P. Gogarten, W. F. Doolittle, J. G. Lawrence, *Mol. Biol. Evol.* **19**, 2226 (2002).
4. W. F. Doolittle, *Science* **284**, 2124 (1999).
5. H. Philippe, C. J. Douady, *Curr. Opin. Microbiol.* **6**, 498 (2003).
6. B. F. Smets, T. Barkay, *Nat. Rev. Microbiol.* **3**, 675 (2005).
7. R. Jain, M. C. Rivera, J. A. Lake, *Proc. Natl. Acad. Sci. U. S. A.* **96**, 3801 (1999). 5

8. M. C. Rivera, R. Jain, J. E. Moore, J. A. Lake, *Proc. Natl. Acad. Sci. U. S. A.* **95**, 6239 (1998).
9. C. M. Thomas, K. M. Nielsen, *Nat. Rev. Microbiol.* **3**, 711 (2005).
10. JGI sequencing protocols, http://www.jgi.doe.gov/sequencing/protocols/prots_production.html
11. D. Gordon, C. Desmarais, P. Green, *Genome Res.* **11**, 614 (2001).
12. K. Rutherford *et al.*, *Bioinformatics* **16**, 944 (2000).
13. R. L. Tatusov *et al.*, *Nucleic. Acids. Res.* **29**, 22 (2001).
14. D. Summers, *Mol. Microbiol.* **29,** 1137 (1998).
15. A. C. Y. Chang, S. N. Cohen, *J. Bacteriol.***134**, 1141 (1978).
16. U. J. Kim, H. Shizuya, P. J. de Jong, B. Birren, M. I. Simon, *Nucleic Acids Res.* **20**, 1083 (1992).
17. J.P. Dillard, J. Yother, *J. Bacteriol.* **173,** 5105 (1991).
18. M. E. Mulligan, W. R. McClure, *Nucleic Acids Res.* **14**, 109 (1986).
19. M. M. Levine *et al.*, *Lancet* **1**, 1119 (1978).
20. F. D. Ciccarelli *et al.*, *Science* **311**, 1283 (2006).
21. We thank A. Lapidus for providing biological material; F. Warnecke for technical assistance; S. Prabhakar for statistical help; T. Woyke and A. Visel for graphic assistance and discussion along with L. Pennacchio, J. Eisen, S. Green Tringe, P. Hugenholtz, T. Doerks, L. Jensen, I. Letunic, T. Dagan and J. Bristow. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

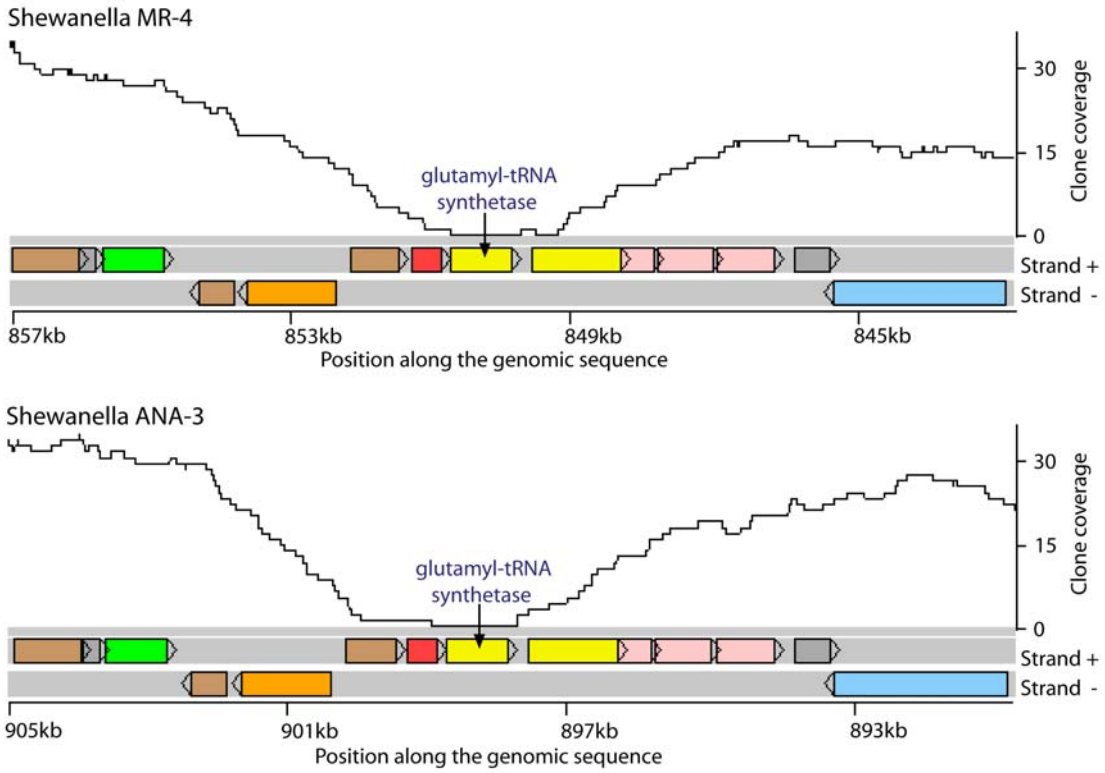

Fig. 1: Coverage plots created on the Artemis genome browser (*12*) of a syntenic 14kb genomic region in two closely related *Shewanella* bacterial species (A) *S. sp. MR-4*; (B) *S. sp. ANA-3*. Colored rectangles represent genes, with colors denoting functional categories; arrow direction indicates whether the gene is on the forward or reverse strand. Coverage is measured per nucleotide.

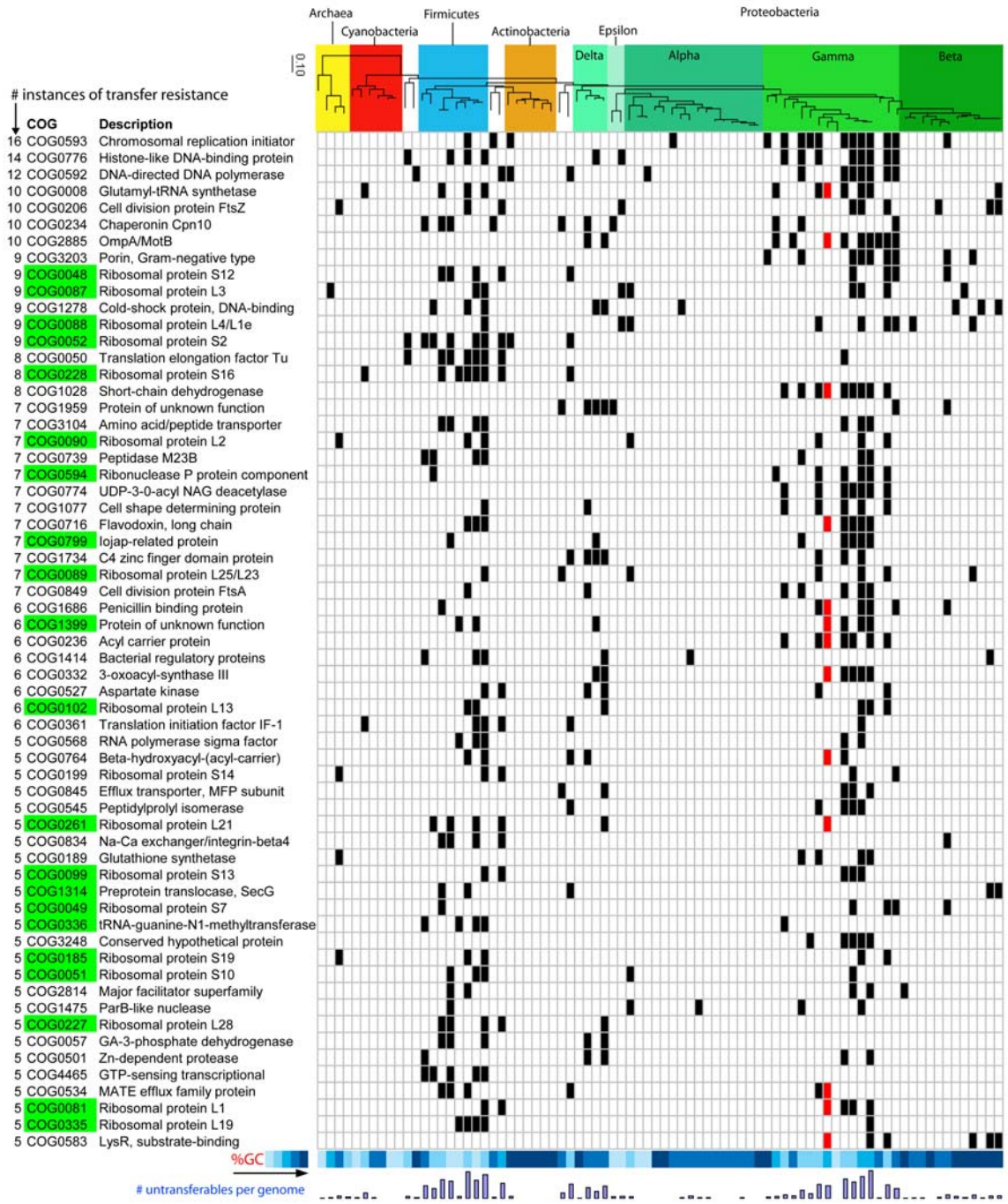

Fig. 2: Genes that cannot be transferred to *E. coli* from 5 or more genomes. Rows are genes, according to their COG classification (*13*). Columns represent the 79 microbial genomes analyzed, arranged by their phylogenetic relationships as determined by a Maximum Likelihood tree anlaysis of 16S ribosomal RNA sequences (SOM). Untransferable genes are denoted by black boxes. The most left hand column indicates the number of genomes from which the gene was untransferable. Universally single copy genes are highlighted in green. *E. coli* (Gammaproteobacteria) genes that could not be cloned into the *E. coli* sequencing strain even when originating from an *E. coli HS* genome are marked red. Percent GC for each of the genomes is color-coded at the bottom of the figure. Darker colors indicate a higher GC content. The histogram below depicts the number of untransferable genes per genome (Table S1).

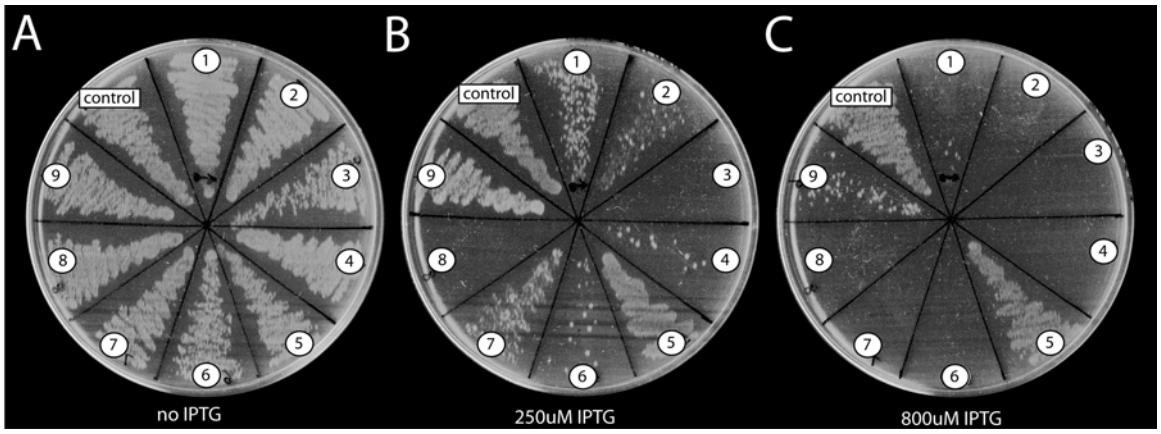

Fig. 3: Toxicity results for the first nine genes tested (Table S2) and a control gene (Beta-galactosidase from *E. coli*). The coding regions of predicted untransferable genes were cloned into the pET11 vector under the control of a T7 promoter, transformed into *E. coli* BL21(DE)pLysS cells, and streaked on LB plates. (A) Cells grown without the expression 6 **inducer IPTG; (B) grown with 250uM IPTG and (C) grown with 800 uM IPTG. 1 - Replication initiator DnaA from *Shewanella denitrificans*; 2 - Histone-like DNA-binding from *Psychrobacter cryohalolentis*; 3 - DNA polymerase III, beta subunit from *Deinococcus geothermalis*; 4- Cell division protein FtsZ from *P. cryohalolentis*; 5 - Chaperonin Cpn10 from *Nitrosococcus oceani*; 6- OmpA/MotB from *N. oceani*; 7- Ribosomal protein S12 from *Rhodoferax ferrireducens*; 8 - Ribosomal protein L4/L1e from *Burkholderia sp. strain 383*; 9 - Ribosomal protein L3 from *P. cryohalolentis*.**

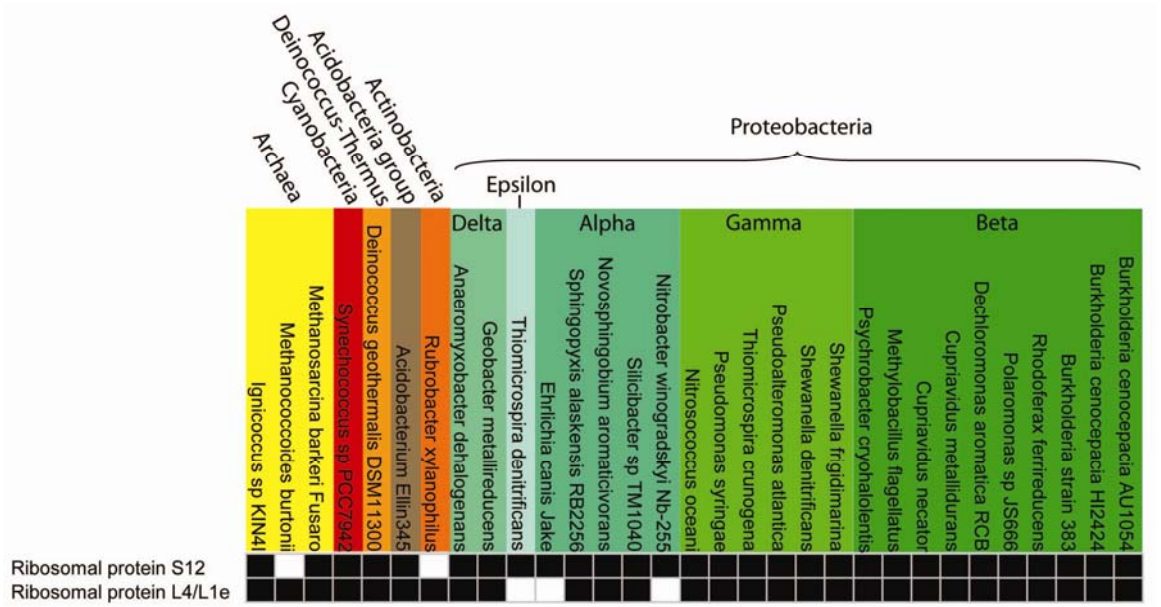

Fig. 4: Toxicity of ribosomal protein S12 (COG0048) (top row) and ribosomal protein L4/L1e (COG0088) (bottom row) from 31 microbial genomes. Columns represent species, arranged by phylogenetic classification with different colors representing different groups (names indicated above). ORFs were cloned into the pET11 vector adjacent to a T7 promoter and transformed into *E. coli* BL21(DE)pLysS cells. Colony growth was tested without gene expression and after induction of expression with various concentrations of IPTG. Black boxes indicate growth inhibition following activation of expression; white boxes indicate no growth inhibition observed (details in Table S3).