Electronic Supplementary Material (ESI) for Molecular BioSystems

Supplementary material

Horizontal gene transfers as metagenomic gene duplications

Luigi Grassi,[1,†] Michele Caselle,[1,2] Martin J. Lercher,[3] and Marco Cosentino Lagomarsino[4]

[1]Università degli Studi di Torino, Dip. Fisica Teorica, Via P. Giuria 1, 10125 Torino, Italy;
[2]I.N.F.N. Torino, Via P. Giuria 1, 10125 Torino, Italy;
[3]Department of Computer Science, Heinrich-Heine-University, Düsseldorf, Germany;
[4]Genomic Physics Group, FRE 3214 CNRS "Microorganism Genomics" and University Pierre et Marie Curie, 15 rue de l'École de Médecine, 75006, France;

†Present address: Dipartimento di Scienze Biochimiche, Sapienza Universitá di Roma, P.le A. Moro, 5 - 00185 Rome, Italy.

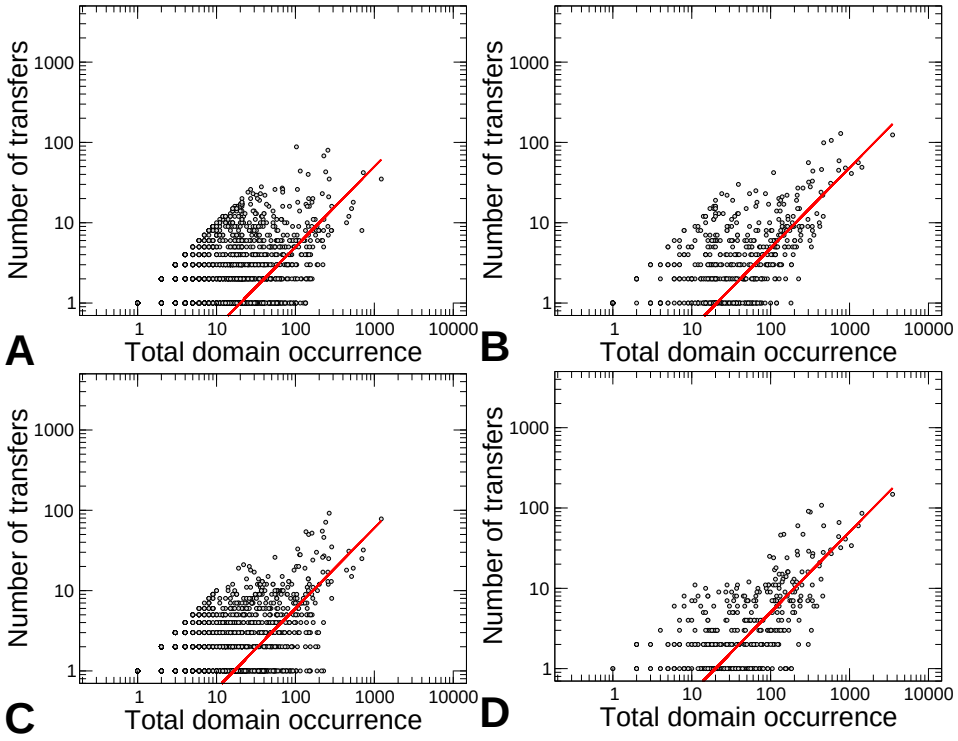
Contents

List of Figures

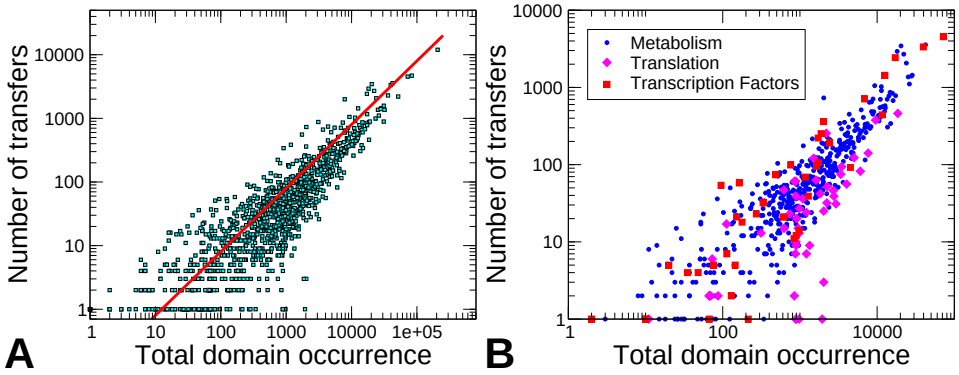
List of Tables

NCBI Taxonomy ID	Complete name
209261	Salmonella enterica subsp. enterica serovar Typhi str. Ty2
224915	Buchnera aphidicola str. Bp (Baizongia pistaciae)
198804	Buchnera aphidicola str. Sg (Schizaphis graminum)
107806	Buchnera aphidicola str. APS (Acyrthosiphon pisum)
203907	Candidatus Blochmannia floridanus
36870	Wigglesworthia glossinidia
198214	Shigella flexneri 2a str. 301
155864	Escherichia coli O157:H7 str. EDL933
233412	Haemophilus ducreyi 35000HP
71421	Haemophilus influenzae Rd KW20
272843	Pasteurella multocida subsp. multocida str. Pm70
229193	Yersinia pestis biovar Microtus str. 91001
198215	Shigella flexneri 2a str. 2457T
273123	Yersinia pseudotuberculosis IP 32953
511145	Escherichia coli str. K-12 substr. MG1655
243265	Photorhabdus luminescens subsp. laumondii TTO1
220341	Salmonella enterica subsp. enterica serovar Typhi str. CT18
99287	Salmonella enterica subsp. enterica serovar Typhimurium str. LT2
199310	Escherichia coli CFT073
386585	Escherichia coli O157:H7 str. Sakai

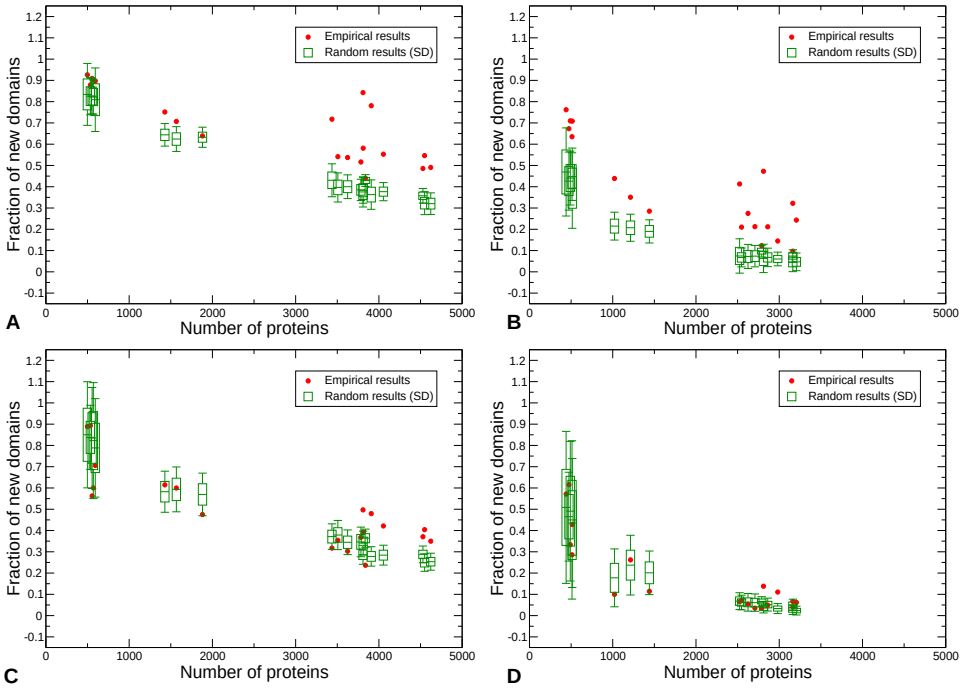
Supplementary Table 1: **List of the prokaryotic species under examination in this study.** The table reports the NCBI Taxonomy ID and the complete name, including the strain, of the 20 species analysed in the article.



Supplementary Figure 1: **Proportionality between the number of transfers of a given domain class and its total occurrence.** Scatterplots of the number of scored transfers as a function of the occurrence in all proteomes, calculated independently for each transferred domain topology (note the logarithmic scale of both axes). The behavior is compatible with a linear growth (red lines). Figures A and B refer to the transfers derived from the research of Lercher *et al.* (Lercher and Pal, 2008), and use the domain assignments respectively derived from Pfam (Finn et al., 2010) and Superfamily (Gough et al., 2001; Wilson et al., 2007). Figures C and D refer to the transfers derived from the Horizontal Gene Transfer Database (HGT-DB) (Garcia-Vallve et al., 2003), and use the domain assignments respectively derived from Pfam and Superfamily. The number of transfers of a domain family grows with the total domain family occurence.



Supplementary Figure 2: **Domain transfer frequences vs domain occurrence in 959 genomes.** (A) A scatterplot, in log-log scale, of the number of scored transfers as a function of the occurrence in all proteomes of each transferred domain topology, made considering all the species in HGT-DB with Superfamily domain annotations. (B) The domain were divided in three main categories, according to the Superfamily database functional annotations. Most domains are part of the "Metabolism" functional category (Wilson et al., 2007). The categories "Metabolism", "Transcription factors" and "Translation" follow indistinguishable trends.



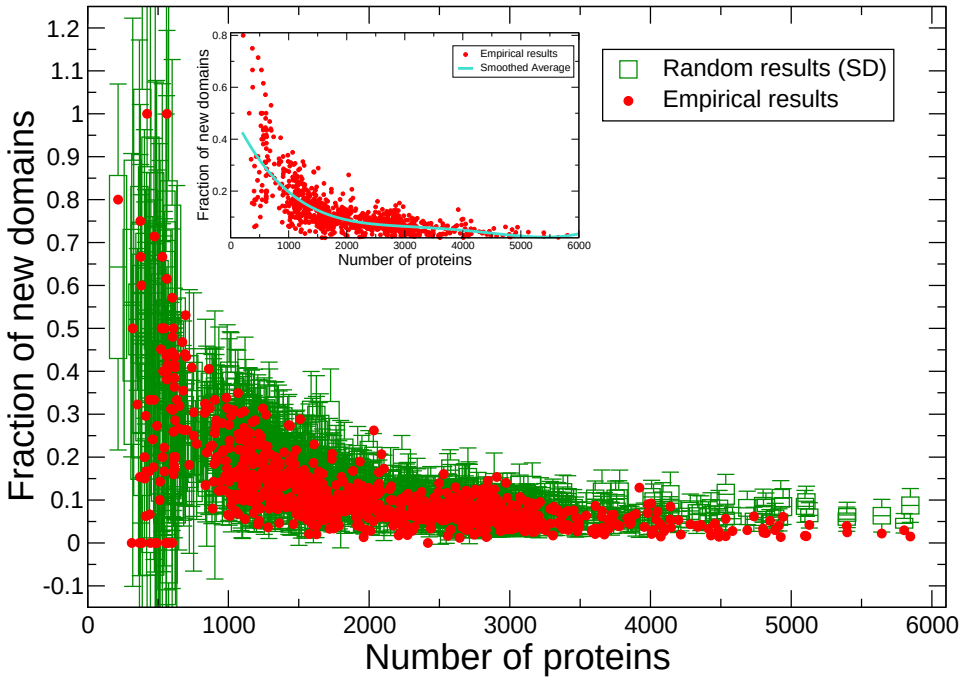
Supplementary Figure 3: **Novelty of transferred domains.** Each genome corresponds to a point that has as x coordinate the total number of proteins in the genome with known domain assignment, and as y coordinate the fraction of new domains transferred. Red points refer to empirical results, green boxplot refer to random results. Central dashes are the mean, upper and lower box margins are the standard deviation (SD), whiskers indicate 2xSD. Figures A and B refer to the transfers derived from the study of Lercher *et al.*, upon the removal of ORFans genes, and use the domain assignments and use the domain assignments respectively derived from Pfam and Superfamily. Figures C and D refer to the transfers derived from the Horizontal Gene Transfer Database (HGT-DB), and use the domain assignments respectively derived from Pfam and Superfamily. The plots in the upper panels, obtained referring to the HGTs derived from (Lercher and Pal, 2008) indicate that for most genomes, transferred proteins carry exogenous domains more frequently than expected by chance, either using Pfam and Superfamily as reference database for the domain assignments. The same analysis obtained referring to the HGTs from HGT-DB gives different results (Figures C and D).

Genome (NCBI Taxonomy ID)	N.Genes	N.HGT	Empirical novelty	Random novelty	Standard deviation
209261	144	10	1	0.92	± 0.085
224915	498	109	0.88	0.83	± 0.035
198804	534	78	0.89	0.82	± 0.044
107806	555	84	0.91	0.82	± 0.041
203907	570	120	0.93	0.82	± 0.036
36870	593	166	0.92	0.81	± 0.03
198214	731	300	0.44	0.76	± 0.024
155864	947	249	0.42	0.69	± 0.029
233412	1429	388	0.79	0.63	± 0.024
71421	1567	321	0.71	0.61	± 0.028
272843	1881	483	0.63	0.62	± 0.022
229193	3436	173	0.71	0.42	± 0.038
198215	3507	237	0.50	0.40	± 0.031
273123	3624	321	0.56	0.39	± 0.027
511145	3811	271	0.58	0.39	± 0.03
243265	3838	1817	0.53	0.42	± 0.012
220341	3908	235	0.77	0.37	± 0.032
99287	4054	514	0.56	0.38	± 0.021
199310	4529	912	0.51	0.35	± 0.016
386585	4549	359	0.52	0.33	± 0.024

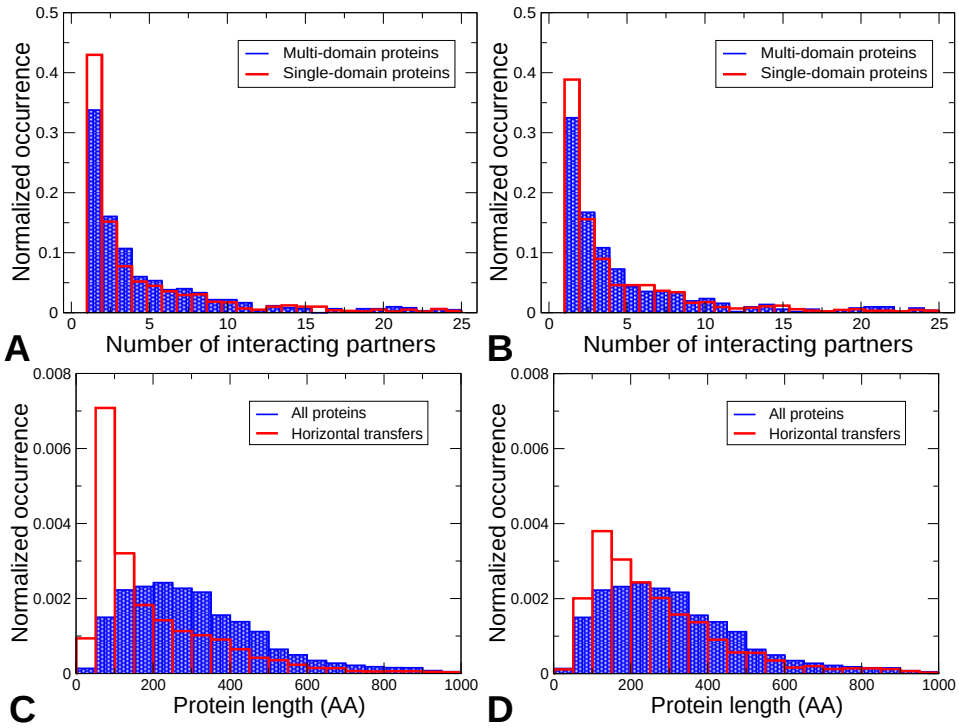
Supplementary Table 2: **New domains acquired by HGT inferred by Lercher *et al.*.** The table reports the data shown in the Supplementary Figure 3 A. The first column is the NCBI taxonomy ID of the prokaryotic species in examination, the second column is the number of proteins with domain annotations in the PFAM database, the third column reports the number of HGT detected by Lercher *et al.*, upon the removal of ORFans genes. The other rows report the empirical frequencies (new domains over total transferred domains) and the mean and standard deviation of the randomized data.

Genome (NCBI Taxonomy ID)	N.Genes	N.HGT	Empirical novelty	Random novelty	Standard deviation
209261	144	16	0.90	0.83	± 0.096
224915	498	8	0.89	0.85	± 0.12
198804	534	22	0.89	0.83	± 0.078
107806	555	10	0.56	0.83	± 0.12
203907	570	8	0.6	0.82	± 0.14
36870	593	12	0.71	0.79	± 0.11
198214	731	45	0.56	0.69	± 0.069
155864	947	129	0.42	0.58	± 0.041
233412	1429	95	0.61	0.57	± 0.049
71421	1567	83	0.6	0.59	± 0.053
272843	1881	97	0.48	0.56	± 0.051
229193	3436	239	0.32	0.36	± 0.03
198215	3507	190	0.35	0.39	± 0.033
273123	3624	265	0.30	0.33	± 0.027
511145	3811	259	0.39	0.35	± 0.029
243265	3838	471	0.24	0.35	± 0.02
220341	3908	420	0.48	0.31	± 0.022
99287	4054	353	0.42	0.30	± 0.024
199310	4529	450	0.37	0.28	± 0.02
386585	4549	412	0.40	0.26	± 0.02

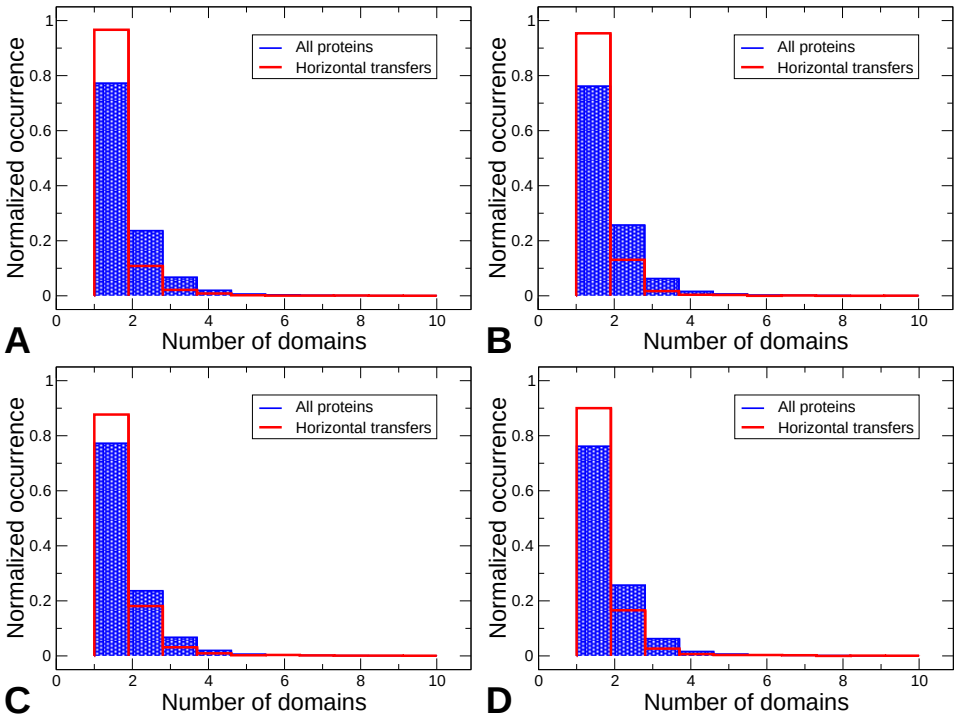
Supplementary Table 3: **New domains acquired by the HGT inferred by HGT-DB.** The table reports the results showm in the Supplementary Figure 3 C. The first column refers to the NCBI taxonomy ID of the prokaryotic species, the second row is the number of proteins with domain annotations in the PFAM database, the third column reports the number of HGT detected by HGT-DB (Garcia-Vallve et al., 2003). The other rows report the empirical frequencies (new domains over total transferred domains) and the mean and standard deviation of the randomized data.



Supplementary Figure 4: **New domains acquired by the HGT inferred by HGT-DB in 959 genomes.** Each genome corresponds to a point that has as x coordinate the total number of proteins in the genome with known Superfamily domain assignment, and as y coordinate the fraction of new domains transferred. Red points refer to empirical results, green boxplot refer to random results. Central dashes are the mean, upper and lower box margins are the standard deviation (SD), whiskers indicate 2xSD.The HGT were retrieved by HGT-DB (Garcia-Vallve et al., 2003). The inset shows the same plot reporting a smoothed average mean (turquoise line) of random results instead of the boxplot.



Supplementary Figure 5: **Protein interactions of single- and multi-domain proteins (A, B). Length of transferred proteins (C, D).** Normalized distributions of protein-protein interactions (derived from ref (Hu et al., 2009)) of *Escherichia coli K-12* for single-domain and multi-domain proteins. The data indicate that single-domain proteins (red bars) tend to have a smaller number of interacting partners (complexity) than multi domain proteins (blue bars). Figure A and B refer respectively to the domain assignments of the Pfam and Superfamily databases. The Wilcoxon rank sum test gave P-value 0.002 for distributions in figure A and 0.07 for distributions in figure B. Panels C and D show the normalized distributions of protein length (in AA) for all proteomes and for HGT proteins. Panel C refers to the transfers derived from the research of Lercher *et al.*, while panel D refers to the transfers in HGT-DB. Most of transferred proteins (red bars) are short proteins. The Wilcoxon rank sum test gave P-values $< 10^{-16}$ for the distributions in both panels C and D.



Supplementary Figure 6: **Normalized distributions for the number of domains for all proteins versus horizontally transferred proteins.** Panels A and B refer to the transfers derived from the study of Lercher *et al.*, upon the removal of ORFans genes, and use the domain assignments respectively derived from Pfam and Superfamily. Figures C and D refer to the transfers derived from the Horizontal Gene Transfer Database (HGT-DB), and use the domain assignments respectively derived from Pfam and Superfamily. The horizontally transferred proteins (red bars) show a strong tendency to be monodomain. The Wilcoxon rank sum test gives a P-value $\simeq 10^{-16}$ for all four plots.

References

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A. 2010. The Pfam protein families database. Nucleic Acids Research. 38(suppl 1):D211–D222.

Garcia-Vallve S, Guzman E, Montero MA, Romeu A. 2003. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. Nucl. Acids Res. 31(1):187–189.

Gough J, Karplus K, Hughey R, Chothia C. 2001. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. Journal of Molecular Biology. 313(4):903 – 919.

Hu P, Janga SC, Babu M, Díaz-Mejía JJ, Butland G, Yang W, Pogoutse O, Guo X, Phanse S, Wong P, Chandran S, Christopoulos C, Nazarians-Armavil A, Nasseri NK, Musso G, Ali M, Nazemof N, Eroukova V, Golshani A, Paccanaro A, Greenblatt JF, Moreno-Hagelsieb G, Emili A. 2009, April. Global functional atlas of escherichia coli encompassing previously uncharacterized proteins. PLoS Biol. 7(4):e1000096+.

Lercher MJ, Pal C. 2008. Integration of Horizontally Transferred Genes into Regulatory Interaction Networks Takes Many Million Years. Molecular Biology and Evolution. 25(3):559–567.

Wilson D, Madera M, Vogel C, Chothia C, Gough J. 2007, Jan. The superfamily database in 2007: families and functions. Nucleic Acids Res. 35(Database issue):308–313.