

Evolutionary patterns in prokaryotic genomes

Eduardo PC Rocha^{1,2}

Prokaryotic genomics is shifting towards comparative approaches to unravel how and why genomes change over time. Both phylogenetic and population genetics approaches are required to dissect the relative roles of selection and drift under these conditions. Lineages evolve adaptively by selection of changes in extant genomes and the way this occurs is being explored from a systemic and evolutionary perspective to understand how mutations relate with gene repertoire changes and how both are contextualized in cellular networks. Through an increased appreciation of genome dynamics in given ecological contexts, a more detailed picture of the genetic basis of prokaryotic evolution is emerging.

Addresses

¹ UPMC Univ Paris 06, Atelier de BioInformatique, F-75005 Paris, France

² Institut Pasteur, Microbial Evolutionary Genomics; CNRS, URA2171, F-75015 Paris, France

Corresponding author: Rocha, Eduardo PC (erocha@pasteur.fr)

Current Opinion in Microbiology 2008, **11**:454–460

This review comes from a themed issue on
Genomics
Edited by Fiona Brinkman and Julian Parkhill

Available online 15th October 2008

1369-5274/\$ – see front matter

© 2008 Elsevier Ltd. All rights reserved.

DOI [10.1016/j.mib.2008.09.007](https://doi.org/10.1016/j.mib.2008.09.007)

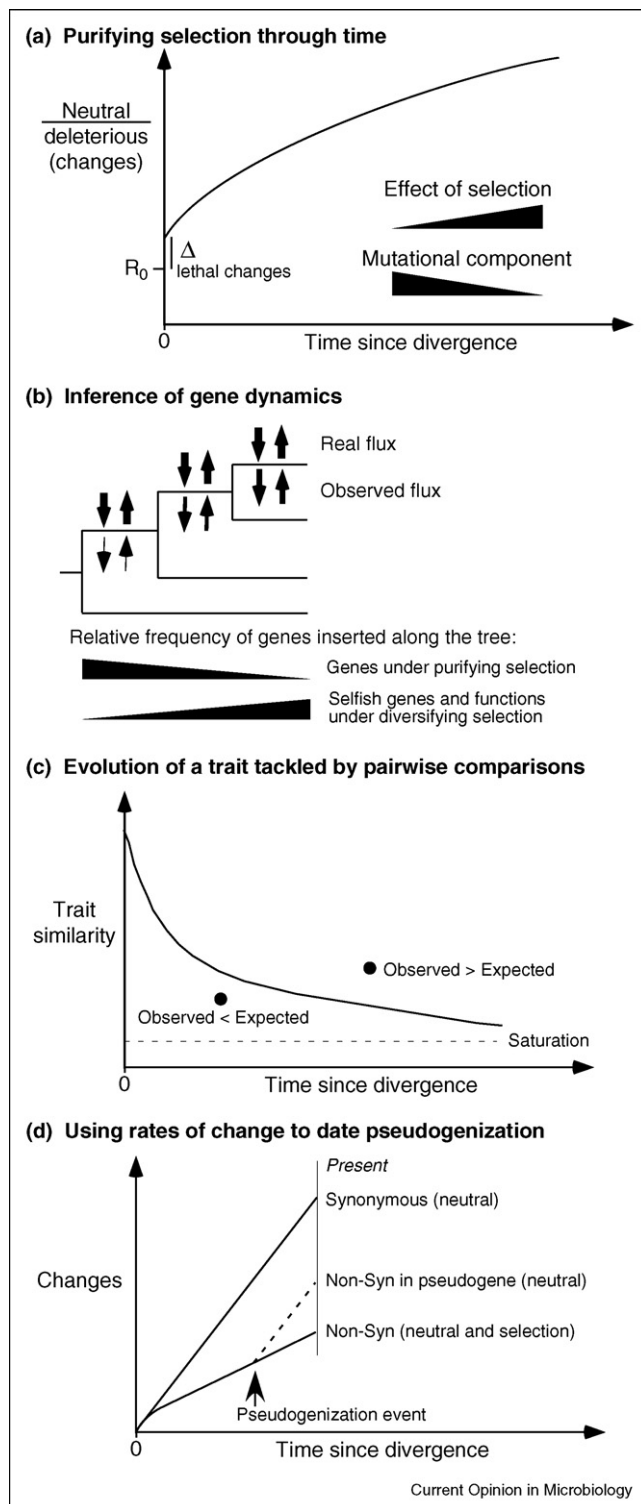
While novel genome sequences from ever more exotic bacterial species will continue to reveal surprises from the largely unexplored diversity of microbes, the availability of multiple genomes for single species or genera provides unprecedented opportunities to understand the detailed mechanics of genome evolution and adaptation. Bacteria and Archaea reproduce asexually but recombine both within and between different lineages. As a consequence, molecular evolution results not only from point mutations, deletion and amplification of existing DNA but also from horizontal transfer and re-assortment of variants existing in the natural populations [1]. The potential for genetic transfer across large phylogenetic distances leaves no obvious way of defining species as sexually isolated populations and results in every gene potentially having a different phylogenetic history. While the problems of phylogenetic inference in this context have been amply debated, it has become increasingly accepted that the ‘core’ genome, that is, the genes present in most genomes and corresponding to the cellular core functions, contains sufficient signal to reconstruct robust

and consistent phylogenies at the inter-species [2,3], and often also at the intra-species [4,5[•],6], level. Because recombination in prokaryotes involves the unidirectional transfer of small DNA sequences, even highly recombining populations may show identifiable patterns of phylogenetic relatedness (reviewed by Brian Spratt in this issue). This is of fundamental importance because only the availability of phylogenies tracing with reasonable accuracy the true history of cellular lineages allows framing evolutionary studies. The reconstruction of phylogenies allows the inference of ancestral states and the quantification of the patterns and the rates of change.

Mutations are instantaneous whereas selection takes time. Changes with no significant consequences to fitness, that is, neutral changes, accumulate stochastically in direct proportion to the spontaneous mutation rate. On the contrary, changes that have very high positive or negative effects on the fitness of the organism are, respectively, quickly fixed or eliminated from populations (Figure 1a). If the organism is well adapted to its environment, then the fraction of adaptive mutations, the ones conferring fitness advantage, is thought to be very small. Many changes in well-adapted organisms are only mildly deleterious and while they are unlikely to be fixed, they remain in populations for a certain time before being purged by natural selection (reviewed in [7]). Because they have just arisen, differences between closely related genomes are a good approximation of the mutation pattern and include many mildly deleterious mutations. They have thus been less affected by natural selection than by the observed differences between more distant genomes. In short, when one compares genes or genomes from closely related bacteria one cannot assume that the differences have been fixed by natural selection, whereas the majority of changes between very distant genomes result from ancient events and can thus be regarded as fixed. Population genetics tools may then allow discriminating adaptive from maladaptive changes.

It is then, therefore, not surprising that early comparisons among closely related genomes showed rates of amino-acid modifying changes, non-synonymous substitutions, close to the rates of synonymous substitutions. Since mutations do not discriminate between codon positions, they affect equally the synonymous and the non-synonymous sites. Since comparing closely related genomes comes close to comparing mutational patterns one naturally expects both rates of substitution to be the same [8[•]]. It is therefore not necessary in this situation to speculate on anomaly high rates of adaptive or deleterious mutations in these genomes, which are rare among genes

Figure 1



Evolutionary time and the accumulation of changes. **(a)** The ratio of the instantaneous rates of neutral and deleterious changes (R_0) is distorted by natural selection in that lethal changes are immediately eliminated and other deleterious changes are progressively purged from populations. Differences between closely related genomes will mostly reflect the mutational patterns, and differences between very distant

coding for housekeeping functions. The same effect leads the outermost branches of phylogenetic trees, corresponding to the most recent changes, to show accelerated rates of change: Mutation rates have not increased recently, it is just that within the pool of new mutations many will be subsequently eliminated by natural selection [9,10]. Maladaptive changes linger in populations depending on variables such as their fitness effect, highly deleterious changes are quickly purged, the effective population size, in larger populations they are less likely to be fixed, and on the frequency of recombination, which by re-assorting alleles increases the efficiency of selection. In birds significant numbers of deleterious changes have been estimated to linger for over 1 million years [10], whereas in bacteria they are still observable among the most distant elements of the *Bacillus cereus* complex [8*].

As sequences diverge with time, homologous recombination between them becomes increasingly less frequent by reasons both mechanistic, in that homologous recombination requires highly similar sequences, and ecological, because bacteria with different lifestyles have lower probability of meeting and thereby sharing genetic material. Yet, some regions will diverge faster than others by a mixture of stochastic effects, selection and differences in mutation rates. They will therefore stop recombining earlier in the process of divergence between two lineages. This effect is quickly amplified by recombination itself because regions where gene conversion re-establishes sequence identity between the two divergent lineages have a higher probability of further recombining. Because they have become identical again, they may maintain sequence cohesion for a longer period of time. It has been proposed that *Escherichia coli* and *Salmonella enterica* have genes that diverged at very different moments in time since their last common ancestor [11**]. Membrane-associated genes and genes contiguous to integration spots are among the first genes to show

genomes will show the imprint of selection. **(b)** Genes are inserted in genomes along evolutionary lineages. Even when the rates of transfer and deletion are constant through time, the ability to infer ancient changes diminishes as one analyses deeper and deeper nodes because most of such insertions have been subsequently deleted. **(c)** The similarity in terms of a given trait, for example, gene order, between genomes decreases with certain characteristic shape as divergence time increases until a point where there is saturation of changes. If one of a set of genomes systematically deviates from the average trend this is an indication of excessive conservation or divergence in terms of the trait. This may be associated with different rates of change in different genomes, for example, different rearrangement rates, or with selection, for example, different degrees of selection for genome organization. **(d)** Synonymous and non-synonymous rates differ and once sufficient time has elapsed to establish a steady-state between the rate of creation and loss of slightly deleterious changes, the two dynamics are roughly linear. Pseudogenization leads previously non-synonymous positions to evolve neutrally. If synonymous positions evolve nearly neutrally then one can infer the age of the pseudogenization events by comparing expected and observed trajectories of divergence through time.

signs of 'speciation', that is, of irreversibly accumulating differences precluding further recombination events, whereas highly expressed genes, enduring strong selection for sequence conservation which allows homologous recombination through longer periods of time, are among the latest. This is consistent with homologous recombination with externally acquired DNA having an important role in maintaining housekeeping functions [6,12]. The consequence of the heterogeneous divergence of sequences within the genome is that different parts of the genome may have had radically different evolutionary histories. For example, high local rates of non-synonymous changes may result from selection for adaptive changes in a gene, if the region diverged for a long time, or simply result from the mutational pattern, if the separation between the lineages at this locus is very recent. This can only be untangled by analyzing the accumulation of changes with no effect on fitness, as they will be more frequent in the former case. Interestingly, the speciation process may be reversed if recombination between two lineages suddenly increases [13[•]], either by ecological niche convergence or by the abolishment of the mechanisms underlying recombination barriers, for example, by the loss of the mismatch repair genes [14].

Along the trees of life

Phylogenetic and population genetics approaches allow inferring how changes accumulate in lineages and also to make educated predictions about their effects over short and long time spans [5^{••},15–18]. The gene repertoire of prokaryotes changes quickly by lateral gene transfer and gene deletion (see review by Hervé Tettelin in this issue), and it is thus of interest to quantify how frequently such changes are adaptive. Transfers sometimes span dozens of genes [19], creating large fluctuations in gene flux. A fraction of genes in genomes is highly conserved because it codes for housekeeping functions and it is unlikely to be lost without significantly affecting the fitness of the organism. On the contrary, most acquired DNA does not carry adaptive functions, possibly it is not even expressed, and it is thus quickly deleted. The rate of gene gain and loss in genomes is very high and it has been calculated to parallel sequence substitution rates [20^{••}]. Since most acquired DNA is lost, when one looks back in time starting from extant genomes there is a perspective bias: Most of the ancient transfers have been lost and there is no trace of them left to reliably infer their past occurrence. This means that some relevant ancestral changes may now be undetectable. It also implicates that the observed rates of gene flux are higher for recent times even if the real gene flux is constant [5^{••},17,20^{••},21] (Figure 1b). This is simply because most very recent insertions are still present in genomes whereas most ancient insertions have already been deleted. This is the same effect we described previously for substitution rates: Even when the rates of change are constant one has to account for the effect of the accumulation of changes

over time and for the effects of ensuing selection. The stochasticity of gene flux and the quick turnover of acquired genes may explain why one finds very weak correlations between the length of the branches of phylogenetic tree, which measure the evolutionary time span, and the inferred amount of insertions and deletions of genetic material [15,20^{••}]. A large branch of a phylogenetic tree corresponds to a large time span where many events took place, but most insertions are quickly deleted. Thus, from the observer standpoint, today, it is almost as if the amount of genes gained and loss was the same for short and large time spans. Thus, inferring gene flux leads necessarily to rates of gene dynamics that vary with time and that look higher for recent times. Since this is expected by chance alone, care must be taken before embarking in adaptationist interpretations of such results.

High gene turnover is increased by the integration of genetic elements that are either deleterious, such as phages, or under strong selection for diversification, such as some virulence factors. These elements tend to be over-represented among recent insertions, not necessarily because the functional pattern of gene flux has changed through time, but just because their persistence times are short. As a result, reconstruction of the historical dynamics of gene repertoires among clades of closely related lineages typically results in over-representation among recent acquisitions of insertion sequences (IS) [5^{••}], prophages [22], restriction systems [23] or antibiotic production systems [24]. For example, ISs are frequent in extant genomes but always of recent origin because transposition tends to have deleterious consequences. Even when two closely related genomes have many ISs these are most frequently of different types [25[•]]. IS distribution is determined by the rates of infection, that is, transfer and transposition, and by the efficiency of selection given the magnitude of the deleterious effects of transposition. Larger genomes have higher density of neutral or mildly deleterious insertion spots and therefore have more ISs [26]. It remains to be known how frequently genomes manage to silence or delete transposition activities and how frequently the expansion of ISs leads to the loss of the lineage by accumulation of deleterious changes.

Lower rates of gene retention can result from less effective selection for mildly advantageous genes or from relaxed selection on certain functions. *Shigella*, which have low effective population sizes and shrinking genomes, show lower rates of gene retention and accelerated rates of non-synonymous substitutions for the same synonymous divergence than the other strains of *E. coli* [4]. Similar scenarios have been found in the genome reduction processes ongoing in independent lineages of obligatory endosymbionts [16,27]. As a result of less effective selection and relaxed selection on certain functions, shrinking genomes are often transiently enriched in

transposable elements [28]. If the lineage survives and purges such elements it typically recedes to narrow ecological niches and becomes sexually isolated. Thus, weakened purifying selection and abundant transposable elements lead initially to a high rate of rearrangement and rapid changes in genome structure. The subsequent absence of elements generating change, such as repeats and lateral transfer, may then lead to higher genome stability of the surviving lineages [29,30].

Genome organization determines how well cellular processes interact with the chromosome and thus has an important adaptive value. The relative order of core genes between pairs of genomes drops exponentially with divergence time [31,32], but is different for certain genomic regions, for example, slower in operons, and depends on the genome relative stability [33[•]]. As for the previously mentioned traits under selection, there is a time-lag effect associated with (counter) selection of rearrangements by natural selection. Recent events have yet to be efficiently purged resulting in a relative excess of rearrangements when very close genomes are compared. For example, *E. coli* K12 and *Salmonella enterica* typhimurium are much more distant than the genomes of strains of *Yersinia pestis*, but they exhibit much fewer rearrangements because the latter have intrinsically high rearrangement rates caused by a high frequency of transposable elements. The use of adequate computational tools allowed studying the rearrangement scenarios among *Yersinia*. These show that even though *Yersinia* have accumulated many rearrangements, the actual and the intermediate genomic configurations were in general less deleterious than expected by chance [34^{••}]. Hence, one can analyze the evolution of traits along the history of lineages and in this way pinpoint the selective processes constraining the evolutionary dynamics of the trait (Figure 1c). By doing so, one can untangle the mutational from the selective effects. Explicitly accounting for evolutionary time in comparative genomics allows dating events such as the origination of pseudogenes and its association with ecological shifts [35[•]] (Figure 1d) and quantifying how laterally transferred sequences adapt ('ameliorate') to the host genome [5^{••},36]. As datasets become larger, one can also envisage the identification of the genes showing higher and lower probability of being lost once acquired, that is, the ones least or most adaptive. Somewhat surprisingly, the short and G + C biased class of ORFans were thus shown to remain in *E. coli* genomes longer than average [37[•]]. This strongly suggests that they often carry adaptive functions.

Network evolutionary dynamics

The availability of genome-level datasets of biological networks has spurred an appreciation of how these networks are best understood within an evolutionary mindset. Many biological networks, such as metabolic, regulatory or protein–protein interactions networks, have similar scaling properties in that a few nodes are highly

connected and most others have low connectivity [38]. In such networks the random removal of a node, for example, because of gene deletion, has a lower-than-expected effect on the network. This so-called power-law distribution of connectivity was suggested to result from design principles favouring robustness of the few most highly connected nodes, and from evolutionary processes of gene duplication and divergence [39,40]. However, the analyses of network evolution by gene gain and loss in prokaryotes do not necessarily substantiate such a view.

Firstly, while in some networks, for example, protein–protein interaction networks (PPIN), the most connected nodes are indeed more resilient to loss, that is, less frequently absent from genomes [41,42[•],43,44], other networks evolve quite in the opposite way. Genomes under reductive evolution, such as the endosymbionts, prune regulatory networks initially from the most connected nodes, that is, by loss of the regulators [42[•]]. This is because it is less deleterious to lose regulatory than operational functions, especially in the stable environment of a eukaryotic host, and because the most connected elements in regulatory networks are the regulators themselves. Furthermore, elements are biologically relevant by many other reasons than just their degree of connectivity in networks. For example, the adaptation of metabolic networks in genomes under reductive evolution depends on metabolic flux balance and on environmental conditions [45^{••}].

Secondly, while many claims of gene duplication have been made for prokaryotes, very few duplicated genes are found in phylogenetically controlled studies. Instead, variations in the sizes of gene families often arise by lateral gene transfer [46,47^{••}]. The distinction between paralogy, duplicates resulting from gene duplication, and xenology, duplicates resulting from lateral transfer, is not purely semantic. After gene duplication the paralogues have the same genetic and biochemical interactions because they are identical. Instead, a very divergent xenologue may have very few or no interactions in common with the native gene. Hence, the two mechanisms result in potentially very different patterns of network evolution. Recently acquired genes have fewer regulatory and physical interactions than the average genes [42[•],44,47^{••},48[•]]. This can be interpreted in several ways: (i) these genes are less disruptive of the existing network structure and are thus less likely to be deleterious [44,49]; (ii) laterally transferred genes provide accessory functionalities and are therefore necessarily peripheral [50]; (iii) new genes need time to integrate into existing networks [48[•]]. These hypotheses are complementary but not equivalent. Hypothesis (iii) states that given enough time a laterally transferred gene will be as connected as the average gene, whereas in (ii) such genes will remain with lower average connectivity. This is the kind of problems

that comparative genomics on the basis of population genetics and phylogenetics can tackle. Genes that were anciently transferred into *E. coli* still have significantly lower-than-expected PPI, suggesting that even if connectivity increases with residence time, such proteins will always tend to have fewer interaction partners than the average protein [48*].

Metabolic networks grow by attachment of new enzymes to the periphery of the networks [50]. Often operons code for contiguous patches of metabolic pathways, and their transfer is thus favoured if one of the genes in the operon allows attachment to the existing network [51]. PPIN evolve by preferential attachment to the most connected nodes [42*,44]. In genetic networks, global regulators have a lower propensity for transfer and loss, whereas many lowly connected regulators are acquired along with the genes they regulate [47**]. As a further example that networks do not all evolve similarly, laterally transferred genes are less connected within the regulatory network than the average gene only at the very early stages after acquisition [48*]. Contrary to PPI, they then rapidly become targeted by more regulators than the average gene, possibly because these genes code for peripheral functions whose utility is restricted to few environmental and physiological conditions [47**]. As expected, synchronization of gene expression between native and new genes takes a large period of time [48*]. These works show that below an apparent topological similarity, biological networks evolve in very diverse ways, depending on the underlying biological objects and on the type and role of interactions.

Conclusions

Tackling some of the most frequent questions in genomics (Box 1), requires the explicit incorporation of evolutionary time in comparative analyses. This is not without challenges. Firstly, the degree of similarity between genomes will depend heavily on the time since divergence to the common ancestor. Accounting for this requires the use of phylogenetic trees. Secondly, since recent and ancient

events have been very differently imprinted by selection it is necessary to account for population genetics processes when comparing very closely related genomes. Most research aiming at the detection of selection in sequences is based on the comparison of variations within species against variations between species. This is troublesome among prokaryotes where sexual isolation, and thus species definition, is intrinsically fuzzy and site-dependent. Yet, if one wants to meaningfully study the genetic bases of phenotypic differences between bacteria, to make sense of the imminent deluge of intra-specific genomic data, we shall have to include evolutionary time and population processes in the analyses. Fortunately, extensive work is already going in that direction.

Acknowledgements

Many of the ideas here presented, but very few of the misconceptions, resulted from discussions with Edward Feil, Marie Touchon, Howard Ochman, Laura Gomez-Valero, Guillaume Achaz and Laurence Hurst. I apologize for not citing many relevant works owing to space and time-span limitations. I thank Edward Feil for comments on the manuscript.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Feil EJ: **Small change: keeping pace with microevolution.** *Nat Rev Microbiol* 2004, **2**:483-495.
2. Brochier C, Bapteste E, Moreira D, Philippe H: **Eubacterial phylogeny based on translational apparatus proteins.** *Trends Genet* 2002, **18**:1-5.
3. Daubin V, Moran NA, Ochman H: **Phylogenetics and the cohesion of bacterial genomes.** *Science* 2003, **301**:829-832.
4. Hershberg R, Tang H, Petrov DA: **Reduced selection leads to accelerated gene loss in *Shigella*.** *Genome Biol* 2007, **8**:R164.
5. Vernikos GS, Thomson NR, Parkhill J: **Genetic flux over time in •• the *Salmonella* lineage.** *Genome Biol* 2007, **8**:R100.
Reconstruction of ancestral genomes of *Salmonella* shows the degree of persistence of different gene function classes. It also enlightens how transferred sequences adapt to the host genome.
6. Treangen TJ, Ambur OH, Tonjum T, Rocha EP: **The impact of the neisserial DNA uptake sequences on genome evolution and stability.** *Genome Biol* 2008, **9**:R60.
7. Balbi KJ, Feil EJ: **The rise and fall of deleterious mutation.** *Res Microbiol* 2007, **158**:779-786.
8. Rocha EPC, Maynard Smith J, Hurst LD, Holden MT, Cooper JE, Smith NH, Feil E: **Comparisons of dN/dS are time-dependent for closely related bacterial genomes.** *J Theor Biol* 2006, **239**:226-235.
Purifying selection on mildly deleterious changes can explain the high relative frequency on non-synonymous changes among closely related genomes (see also references [9] and [10]).
9. Sharp PM, Bailes E, Chaudhuri RR, Rodenburg CM, Santiago MO, Hahn BH: **The origins of acquired immune deficiency syndrome viruses: where and when?** *Philos Trans R Soc Lond B Biol Sci* 2001, **356**:867-876.
10. Ho SY, Phillips MJ, Cooper A, Drummond AJ: **Time dependency of molecular rate estimates and systematic overestimation of recent divergence times.** *Mol Biol Evol* 2005:22.
11. Retchless AC, Lawrence JG: **Temporal fragmentation of •• speciation in bacteria.** *Science* 2007, **317**:1093-1096.
Where it is shown that genes 'speciate' at very different moments in time following divergence of two bacterial lineages.

Box 1 Questions for further work

How to effectively compare closely related genomes with cheap but error-prone sequencing techniques?

What are the mutational and selective determinants of the rates of gene insertion and deletion?

How much of genome dynamics is adaptive?

How widely do evolutionary patterns fluctuate along lineages?

What are the determinants of the trade-offs between genome stability and dynamics?

Can we find ways of reliably dating, preferably in years, evolutionary events in prokaryotes?

How to quantify population genetics parameters from metagenomics data?

12. Szollosi GJ, Derenyi I, Vellai T: **The maintenance of sex in bacteria is ensured by its potential to reload genes.** *Genetics* 2006, **174**:2173-2180.
13. Sheppard SK, McCarthy ND, Falush D, Maiden MC: **Convergence of *Campylobacter* species: implications for bacterial evolution.** *Science* 2008, **320**:237-239.
Evidence for the reversal of incipient speciation in *Campylobacter*.
14. Vulic M, Dionisio F, Taddei F, Radman M: **Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria.** *Proc Natl Acad Sci U S A* 1997, **94**:9763-9767.
15. Snel B, Bork P, Huynen MA: **Genomes in flux: the evolution of archaeal and proteobacterial gene content.** *Genome Res* 2002, **12**:17-25.
16. Boussau B, Karlberg EO, Frank AC, Legault BA, Andersson SG: **Computational inference of scenarios for (alpha)-proteobacterial genome evolution.** *Proc Natl Acad Sci U S A* 2004, **101**:9722-9727.
17. Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J *et al.*: **Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*.** *PLoS Genet* 2007, **3**:e231.
18. Blanc G, Ogata H, Robert C, Audic S, Suhre K, Vestris G, Claverie J, Raoult D: **Reductive genome evolution from the mother of *Rickettsia*.** *PLoS Genet* 2007, **3**:e14.
19. Ochman H, Jones IB: **Evolutionary dynamics of full genome content in *Escherichia coli*.** *EMBO J* 2000, **19**:6637-6643.
20. Hao W, Golding GB: **The fate of laterally transferred genes: life in the fast lane to adaptation or death.** *Genome Res* 2006, **16**:636-643.
A maximum likelihood approach to quantify the rates of gene gain and loss in bacterial lineages (see also reference [21]). It is shown that observed insertion/deletion rates are of equal or higher frequency than nucleotide substitutions and that they are higher at the tips of phylogenetic trees.
21. Hao W, Golding GB: **Patterns of bacterial gene movement.** *Mol Biol Evol* 2004, **21**:1294-1307.
22. Canchaya C, Fournous G, Brussow H: **The impact of prophages on bacterial chromosomes.** *Mol Microbiol* 2004, **53**:9-18.
23. Xu Q, Morgan RD, Roberts RJ, Blaser MJ: **Identification of type II restriction and modification systems in *Helicobacter pylori* reveals their substantial diversity among strains.** *Proc Natl Acad Sci U S A* 2000, **97**:9671-9676.
24. Choulet F, Aigle B, Gallois A, Mangenot S, Gerbaud C, Truong C, Francou FX, Fourrier C, Guerneau M, Decaris B *et al.*: **Evolution of the terminal regions of the streptomyces linear chromosome.** *Mol Biol Evol* 2006, **23**:2361-2369.
25. Wagner A: **Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes.** *Mol Biol Evol* 2006, **23**:723-733.
Transposable elements reproduce quickly in bacterial lineages but have low persistence times, showing the effects of selection against transposition events in densely coding genomes (see also reference [26]).
26. Touchon M, Rocha EP: **Causes of insertion sequences abundance in prokaryotic genomes.** *Mol Biol Evol* 2007, **24**:969-981.
27. Wernegreen JJ, Moran NA: **Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes.** *Mol Biol Evol* 1999, **16**:83-97.
28. Toh H, Weiss BL, Perkin SA, Yamashita A, Oshima K, Hattori M, Aksoy S: **Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host.** *Genome Res* 2006, **16**:149-156.
29. Mira A, Ochman H, Moran NA: **Deletional bias and the evolution of bacterial genomes.** *Trends Genet* 2001, **17**:589-596.
30. Silva FJ, Latorre A, Moya A: **Why are the genomes of endosymbiotic bacteria so stable?** *Trends Genet* 2003, **19**:176-180.
31. Huynen MA, Bork P: **Measuring genome evolution.** *Proc Natl Acad Sci U S A* 1998, **95**:5849-5856.
32. Tamames J: **Evolution of gene order conservation in prokaryotes.** *Genome Biol* 2001, **2**: 0020.0021-0020.0011.
33. Rocha EPC: **Inference and analysis of the relative stability of bacterial chromosomes.** *Mol Biol Evol* 2006, **23**:513-522.
A measure of stability derived from comparative genomics is introduced and used to assess hypotheses about selection of and challenges to genome stability.
34. Darling AE, Miklos I, Ragan MA: **Dynamics of genome rearrangement in bacterial populations.** *PLoS Genet* 2008, **4**:e1000128.
An analysis to determine the evolutionary trajectory of rearrangement events in the genomes of *Yersinia* using bayesian methods. It is shown how these trajectories are influenced by selection leading to actual states that were probably fitter than expected by chance.
35. Gomez-Valero L, Rocha EPC, Latorre A, Silva FJ: **Reconstructing the ancestor of *Mycobacterium leprae*: the dynamics of gene loss and genome reduction.** *Genome Res* 2007, **17**:1178-1185.
Mycobacterium leprae has a very large number of pseudogenes, which have sufficient similarity to be aligned and allow pseudogenization dating and association with ecological shifts.
36. Daubin V, Ochman H: **Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*.** *Genome Res* 2004, **14**:1036-1042.
37. van Passel MW, Marri PR, Ochman H: **The emergence and fate of horizontally acquired genes in *Escherichia coli*.** *PLoS Comput Biol* 2008, **4**:e1000059.
Contrary to naïve expectations, ORFans are more fixed in *E. coli* populations than other genes.
38. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks.** *Nature* 2000, **407**:651-654.
39. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297**:1551-1555.
40. Teichmann SA, Babu MM: **Gene regulatory network growth by duplication.** *Nat Genet* 2004, **36**:492-496.
41. Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N *et al.*: **Interaction network containing conserved and essential protein complexes in *Escherichia coli*.** *Nature* 2005, **433**:531-537.
42. Ochman H, Liu R, Rocha EP: **Erosion of interaction networks in reduced and degraded genomes.** *J Exp Zool B Mol Dev Evol* 2007, **308**:97-103.
While genes corresponding to highly connected proteins are more resilient to genome shrinkage, the inverse happens for genes highly connected in regulatory networks.
43. Tamames J, Moya A, Valencia A: **Modular organization in the reductive evolution of protein-protein interaction networks.** *Genome Biol* 2007, **8**:R94.
44. Wellner A, Lurie MN, Gophna U: **Complexity, connectivity, and duplicability as barriers to lateral gene transfer.** *Genome Biol* 2007, **8**:R156.
45. Pal C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD: **Chance and necessity in the evolution of minimal metabolic networks.** *Nature* 2006, **440**:667-670.
Metabolic flux analysis provides accurate predictions of how genome shrinkage in *Buchnera* affects metabolism by considering its environment and an *E. coli*-like ancestor.
46. Lerat E, Daubin V, Ochman H, Moran NA: **Evolutionary origins of genomic repertoires in bacteria.** *PLoS Biol* 2005, **3**:e130.
47. Price MN, Dehal PS, Arkin AP: **Horizontal gene transfer and the evolution of transcriptional regulation in *Escherichia coli*.** *Genome Biol* 2008, **9**:R4.
A detailed study showing that transcription factors are mostly acquired by lateral transfer, where local, but not global, regulators are co-located and co-transferred with the target genes.

48. Lercher MJ, Pal C: **Integration of horizontally transferred genes into regulatory interaction networks takes many million years.** *Mol Biol Evol* 2008, **25**:559-567.
Horizontally transferred genes integrate slowly in existing networks.
49. Jain R, Rivera MC, Lake JA: **Horizontal gene transfer among genomes: the complexity hypothesis.** *Proc Natl Acad Sci U S A* 1999, **96**:3801-3806.
50. Pal C, Papp B, Lercher MJ: **Adaptive evolution of bacterial metabolic networks by horizontal gene transfer.** *Nat Genet* 2005, **37**:1372-1375.
51. Lawrence JG, Roth JR: **Selfish operons: horizontal transfer may drive the evolution of gene clusters.** *Genetics* 1996, **143**:1843-1860.