Genetica (2008) 134:205–210
DOI 10.1007/s10709-007-9226-6

Mutation rate and genome reduction in endosymbiotic and free-living bacteria

Gabriel A. B. Marais · Alexandra Calteau · Olivier Tenaillon

Received: 5 March 2007 / Accepted: 3 November 2007 / Published online: 29 November 2007

Abstract Genome reduction has been considered the hallmark of endosymbiotic bacteria, such as endocellular mutualists or obligatory pathogens until it was found exactly the same in several free-living bacteria. In endosymbiotic bacteria genome reduction is mainly attributed to degenerative processes due to small population size. These cannot affect the free-living bacteria with reduced genomes because they are known to have very large population sizes. It has been proposed that selection for simplification drove genome reduction in these free-living bacteria. For at least one of them (*Prochlorococcus*), genome reduction is associated with accelerated evolution and we suggest an alternative hypothesis based on increase in mutation rate as the primary cause of genome reduction in free-living bacteria.

G. A. B. Marais · A. Calteau
Université de Lyon; Université de Lyon 1; Centre National de la Recherche Scientifique, UMR5558, Laboratoire de Biométrie et Biologie évolutive, Villeurbanne Cedex 69622, France

Present Address:
A. Calteau
Commissariat à l'Energie Atomique (CEA), Direction des Sciences du Vivant, Institut de Génomique, Genoscope, Laboratoire de Génomique Comparative, 2 rue Gaston Crémieux, 91057 Evry Cedex, France

O. Tenaillon
Institut National de la Santé et de la Recherche Médicale; Universite Denis Diderot Paris 7, INSERM U722, Ecology and Evolution of Microorganisms, site Xavier Bichat, 16 rue Henri Huchard, 75870 Paris Cedex 18, France

O. Tenaillon (✉)
INSERM U722, Faculté de médecine Xavier Bichat, Université Paris 7, 16 rue Henri Huchard, 75018 Paris, France
e-mail: Olivier.Tenaillon@bichat.inserm.fr

Keywords Bacterial genomics · Error threshold · Genome size · Molecular evolution · Mutation–selection balance

Introduction

Genome size varies tremendously among life forms. Population size, recombination rate and mutation rate seem to be key parameters for genome size diversity (Lynch 2006; Lynch and Conery 2003). However, how and why these parameters change from one organism to another is still poorly understood. Endosymbiotic bacteria such as endocellular mutualists (e.g. *Buchnera*) and obligatory pathogens either intra (e.g. *Chlamydia, Rickettsia*) or extracellular (e.g. *Mycoplasma*) have small genomes compared to their free-living closest relatives (Moran 2002). This genome reduction can be very severe as in the case of the ultimate endosymbionts: organelles (Kurland and Andersson 2000). Another interesting feature of these genomes is that they are AT-richer and they exhibit faster DNA sequence evolutionary rate than their free-living cousins (Moran 2002). This trend (small genome, AT-richness, accelerated evolution) was supposed to be a signature of the endosymbiotic lifestyle until it was found in free-living oceanic cyanobacteria from the genus *Prochlorococcus* (Dufresne et al. 2005; Dufresne et al. 2003; Rocap et al. 2003). A comparison of *Synechococcus* sp. (WH8102) and three species of *Prochlorococcus* (MED4, SS120, and MIT9313) has indeed revealed that two of the three *Prochlorococcus* species (MED4 and SS120) have undergone a ~30% genome reduction, have become ~20% AT-richer and have experienced a two to fourfold increase in evolutionary rate compared to their relatives (Dufresne et al. 2005). Two more *Prochlorococcus* with

reduced genomes have been found since (see Table 1). It has also been shown that another marine bacteria, *Pelagibacter ubique*, has a reduced genome, which is the smallest genome of any known free-living bacteria [=1.3 Mb similar to *Rickettsia conorii* an intra-cellular pathogen (Giovannoni et al. 2005)].

Genome reduction in endosymbionts is believed to be driven by genetic drift and thus is seen as a consequence of chance and mutational bias, which can be maladaptive. The intensity of genetic drift in a given species is determined by its effective population size N_e. When the product $N_e \times s$ is smaller than one—*s* being selection coefficient at a given site of the genome—selection is ineffective at this site whose evolutionary fate is then dominated by genetic drift. N_e is believed to be small in bacterial endosymbionts (Moran 2002). This implies that many genomic sites under weak selection (with $N_e \times s \sim 1$) in a free-living bacterial species will be evolving neutrally in a endosymbiotic relative. N_e is also further reduced because endosymbionts have few opportunities to recombine and thus, they suffer from the accumulation of deleterious mutations by Muller's ratchet (Moran 2002). Hence N_e reduction will relax selection on many sites and sometimes entire genes. All these sites and genes becoming neutral DNA, mutation patterns will decide of their evolutionary fate. In bacteria, the typical mutation pattern seems to be excess of GC to AT mutations (Moran 2002; Rocha and Danchin 2002) and excess of deletions (Mira et al. 2001), which explains AT richness and massive gene loss in endosymbionts.

In marine bacteria with reduced genomes such as *Prochlorococcus* and *Pelagibacter*, genetic drift is not expected to occur because these bacteria have among the largest population sizes on earth (Dufresne et al. 2005; Dufresne et al. 2003; Giovannoni et al. 2005; Rocap et al. 2003). Instead, it has been proposed that selection for streamlining has driven genome reduction in these bacteria. In such a nutrient-poor (especially N and P elements) environment as the oceanic surface waters one, expressing and replicating some unnecessary genes is costly and therefore natural selection will favour deletion of these genes. Among those, the *ada* gene, which repairs GC to AT mutations and other DNA repair genes have been lost in reduced *Prochlorococcus* genome, which would explain AT-enrichment and accelerated evolution in these species. The streamlining hypothesis is attractive but has its problems. This hypothesis predicts that the streamlined genome is optimized with respect to size. In particular, proteins should tend to be shorter in streamlined *Prochlorococcus* genomes because of the pressure to reduce energy and nutrient consumption related to DNA, RNA and protein synthesis and maintenance of all expressed genes. Analysis of orthologous genes in three *Prochlorococcus* species reveals that proteins have very similar size among reduced and non-reduced *Prochlorococcus* genomes (see Fig. 1). Some proteins are also larger in reduced *Prochlorococcus* genomes than in non-reduced ones, which is not in agreement with the streamlining hypothesis.

As we show here another mechanism than selection for simplification could be responsible of genome reduction in *Prochlorococcus*: increased mutation rate. *Prochlorococcus* strains harbouring genome reduction (MED4, SS120, NATL2A and MIT9312) have lost some DNA repair genes (see Table 2) whose inactivation might have important effects on global mutation rate, especially in such resource-limited environments as *Prochlorococcus*' (Mackay et al. 1994; Saumaa et al. 2002). In natural populations of bacteria, there are often isolates that have a 10-fold to 1,000-fold increased rate of mutation compared to the wild type because they lack DNA repair genes. It has been shown that these strains, called mutators, could have a very important effect for the adaptation to a new environment (Tenaillon et al. 1999). Mutators experience a high rate of neutral and deleterious mutations but also of beneficial ones. They can therefore adapt to a new environment more rapidly (Taddei et al. 1997). Using a simple model we show here that the increase in mutation rate could result in a long-term genome reduction and therefore offer a non-selective scenario for genome reduction in large population.

Model

In very large populations (as in *Prochlorococcus* and *Pelagibacter*) population genetics predicts that selection can act on tiny phenotypic differences between variants (i.e

Table 1 Genomic features of *Prochlorococcus* species and out-group

Species	Size (Mb)	G + C%	Proteins	Structural RNAs	Coding content
Synechococcus WH8102	2.43	59	2519	55	89
Prochlorococcus MIT9313	2.41	50	2269	55	81
Prochlorococcus SS120[a]	1.75	36	1883	46	88
Prochlorococcus MED4[a]	1.66	30	1717	44	87
Prochlorococcus MIT9312[a]	1.71	31	1810	45	89
Prochlorococcus NATL2A[a]	1.84	35	1892	44	85

[a] Strains with reduced genomes. Data retrieved from the NCBI website (http://www.ncbi.nlm.nih.gov/)

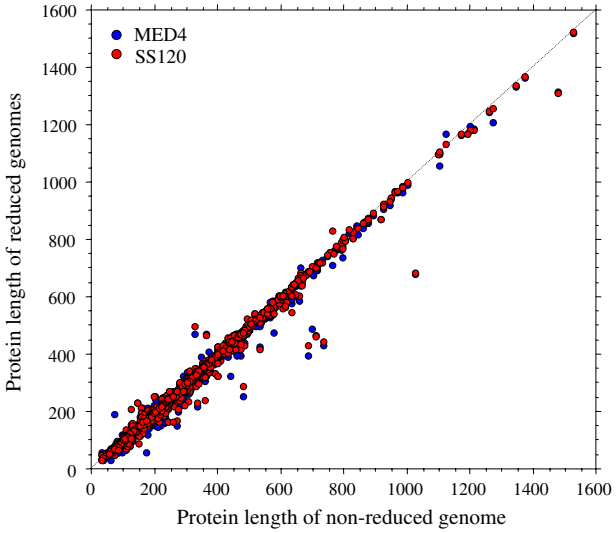

Fig. 1 Comparison of proteins size between *Prochlorococcus* species. Reduced *Prochlorococcus* genomes: MED4, SS120, Non-reduced *Prochlorococcus* genome: MIT9313. The analysis includes 1,306 triplets of orthologous genes from the three genomes as in Dufresne et al. (2005) Sequences have been retrieved from the HOGENOM database release 03 (http://pbil.univ-lyon1.fr/). Adjustment to linear model (without constant) is excellent: MED4 (slope = 0.975, R^2 = 0.995) and SS120 (slope = 0.983, R^2 = 0.996)

very small selection coefficient s). As long as the product $N_e \times s$ is bigger than one an allele conferring an increase s on fitness will be selected. However, this classical result only holds when mutation rate is negligible. When mutation rate is taken into account, the value of s under which selection is no longer effective depends also on mutation rate.

A model introduced by Eigen (1971), studied the influence of mutation rate on the efficiency of selection. Eigen genotypic landscape is composed of a master sequence whose fitness is 1 while all the remaining sequences have fitness $1 - s$. In such landscape, the equilibrium frequency of the master sequence is $1 - u/s$ if u the rate of mutation is less than s and 0 if it is more. Hence, if the mutation rate increases above the value of s, the master sequence cannot be maintained in the population, a phenomenon referred to as "error threshold" (Biebricher and Eigen 2005). This simple model has been criticized for its limited domain of application (Wiehe 2000), nevertheless, it fits perfectly with our present point. A non-essential gene has a limited contribution to fitness, while functioning it provides a benefit of s, and is neutral otherwise. Hence if a global increase in mutation rate increases the rate of inactivation of a gene over its impact on fitness, the sequence of that gene will inevitably evolve towards a sequence reflecting only the mutational biases. The gene will turn to a pseudogene and be deleted in the long run due to deletion bias.

If we extend further this model and consider $n(s)$ genes of effect s, derivation by Haigh (1978) predicts that the average number of functional gene of fitness effect s per genome will be $(1 - u/s) \times n(s)$. So if mutation rate increases by an x factor then the fraction of functional gene of effect s kept in the genome is $r(s) = (s - x \times u)/(s - u)$ for $x \times u < s$ and 0 otherwise. If we have the distribution of fitness effect of the gene pool of the species, we can estimate the number of genes lost due to an increase in mutation rate.

To illustrate further this model, we considered a genome composed of 300 essential genes, and some unessential genes having their contribution on fitness drawn in a given distribution. We choose a distribution in which the probability to have a gene of small effect is much larger than the probability to have a gene of large effect. For mathematical convenience we choose the probability $n(s) = -K \times \log(s)/s$, K being a constant. With such distribution, a genome of null mutation rate has $-K \times \log(s)$ genes of effect $-\log(s)$ on logarithm of fitness. Because natural selection is only efficient for $s > u$ and because only $(1 - u/s)$ genes of effect s will be retained by natural selection (Fig. 2a), the number of gene in the genome is proportional to

Table 2 DNA repair genes in *Prochlorococcus* strains and outgroup

Genes	COG	Products	*Synechococcus*	*Prochlorococcus*				
			WH8102	MIT9313	MED4[a]	SS120[a]	MIT9312[a]	NATL2A[a]
ada/ogt	0350	6-O-methylguanine-DNA methyltransferase	SYNW1680	PMT0269	–	–	–	–
mutY	1194	A/G-specific DNA glycosylase	SYNW0115	PMT0135	–	Pro1789	–	PMN2A_1205
recQ	0514	Superfamily II DNA helicase	SYNW1958	PMT0189	–	–	–	–
recJ	0608	Single-stranded DNA-specific exonuclease	SYNW1206	PMT0761	–	Pro0984	–	PMN2A_0357
exoI/xseA	1570	Exonuclease VII large subunit	SYNW2181	PMT1641	–	Pro0111	–	PMN2A_1460
xseB	1722	Exonuclease VII small subunit	SYNW2182	PMT1642	–	Pro0112	–	PMN2A_1461
–	0494	NUDIX hydrolase family	SYNW1334	PMT1026	–	–	–	–

[a] Strains with reduced genomes. Genes were identified by similarity using MIT9313 sequences as query

$$\int_u^1 -\left(1-\frac{u}{s}\right)\frac{\log(s)}{s}\,\mathrm{d}s = 1-u+\log(u)+\frac{1}{2}\log(u)^2.$$

The change in genome size due to an x-fold increase in mutation rate can then be written

$$R(x) = \frac{300+25\left[1-xu+\log(xu)+\frac{1}{2}\log(xu)^2\right]}{300+25\left[1-u+\log(u)+\frac{1}{2}\log(u)^2\right]}-1.$$

This takes into account the essential genes, and assumes a genome of about 2,400 genes for a gene inactivation rate of 10^{-6} (generating the factor 25). As seen in Fig. 2b a modest increase in mutation rate can have a substantial effect on genome size.

These simple derivations have the interesting property of being true in infinite population size, and so they could influence the genome evolution of *Prochlorococcus* and possibly *Pelagibacter*. These species have presumably very large N_e; so genes of incredibly small contribution to fitness (much smaller than any experimental setting could detect) can be maintained by selection in the genome. However an increase in mutation rate, even modest, will change the number of functional genes that can be maintained in the genome (see Fig. 2a). If there is an important fraction of small effect genes in the genome, such phenomenon can lead to very significant genome reduction (see Fig. 2b)

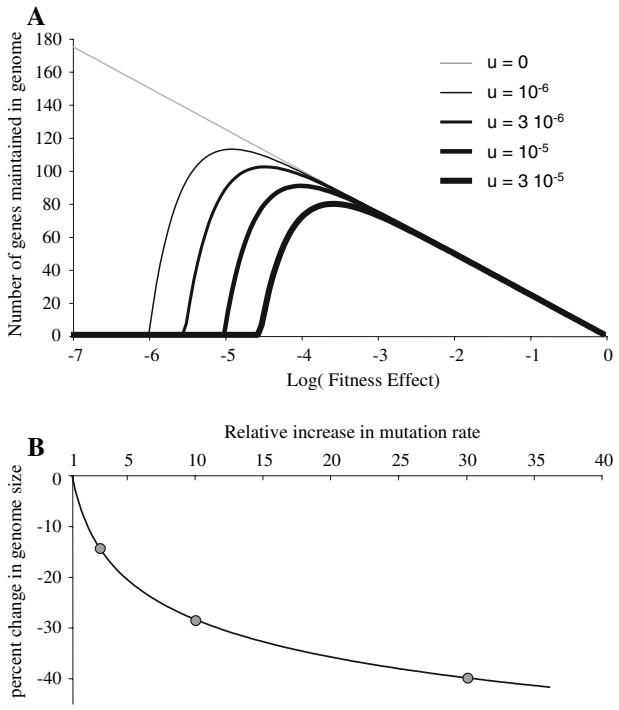

Fig. 2 Theoretical expectations of mutation rate on gene loss. (**a**) Distribution, for different mutation rates, of non-essential genes maintained in the genome as a function of the logarithm of their contribution to fitness. We assume a particular distribution of gene contribution to fitness (*grey line*) in which genes have high probability to contribute very slightly to fitness (see text). We also assume the presence of 300 essential genes (not shown). As population size is very large even very small effect genes are selected upon. With a null mutation rate a genome would bear all the possible genes (*grey line*). However, increase in mutation rate limits the fraction of small effect genes that natural selection can maintain in the genome. Black lines of increasing size reflect the number of genes present in genomes having the basal mutation rate (10^{-6} chance to inactivate a gene) or 3-, 10- and 30-fold increase in mutation rate. (**b**) Overall genome reduction as a function of mutation rate increase. The basal rate is $u = 10^{-6}$. Dots indicate the observed reduction in genome size for 3-, 10- and 30-fold increase in mutation rate

Discussion

Variations in genome size could result from selective and non-selective forces. The non-selective forces proposed so far were population size dependant, i.e. they relied on the reduced efficiency of selection in small population. With a simple model we show that even in an organism having one of the largest population size on earth, non-selective forces (mutation rate increase) could lead to genome reduction. This could explain *Prochlorococcus* genome reduction and possibly *Pelagibacter* one, although in this case accelerated evolution because of lost of DNA repair genes has still to be established.

Our proposed scenario suggests that loss of repair genes has played a primary role in genome reduction. Selection of increased mutation rate has been analytically investigated, observed in silico, in vitro and in natura. Change in environmental conditions or colonization of a new niche are conditions that favour the indirect selection of mutator allele, as the benefit of an overproduction of beneficial mutations in these conditions overpasses the cost due to an overproduction of deleterious mutations. Several studies have shown that larger population sizes favour the selection of mutator alleles (André and Godelle 2006; Tenaillon et al. 1999) and in fairly large experimental populations it was shown that 25% of populations experienced a 30-fold increase in mutation rate that remained stable during more than 30,000 generations (Sniegowski et al. 1997). Hence the selection of mutators in *Prochlorococcus* history would not be surprising. The reduced *Prochlorococcus* genomes (MED4 and SS120) have indeed colonized a new environment (low light zone) and mutators could have helped with the adaptation process. In particular, a key stage in this adaptation process seems to have been the acquisition of genes required to live in the low light zone by horizontal transfer from phages (Martiny et al. 2006; Coleman et al. 2006). Interestingly, some mutator alleles also have a recombinator phenotype, i.e. they recombine at a higher frequency and with a larger spectrum of genetic entities, which favours horizontal transfers (Vulic et al. 1997).

What remains surprising is why would have some *Prochlorococcus* populations remained mutators in the long term. The classical view on mutator evolution is that once adaptation to the new environmental conditions has been reached, increased mutation rate is no longer selected and a decrease in mutation rate occurring through mutation or recombination is favoured (Denamur et al. 2000). However, recent developments suggest that this might not be the case in very large populations. In such population, a slight modification of the environment occurring every 1,000 generations (maybe several years for *Prochlorococcus*) could maintain a high mutation rate (André and Godelle 2006; Gerrish et al. 2007). Moreover, if environmental changes are more frequent (due to biotic interactions for example) then an ever-increasing mutation rate could be favoured. Such an escalation is possible because the benefit of high mutation rate is perceived on a shorter term than its cost due to the accumulation of deleterious mutations. Hence large population sizes could favour long-term persistence of increased mutation rate (André and Godelle 2006; Gerrish et al. 2007). Long-term increased mutation rate will have two consequences. First, an increase in the number of inactivated genes that later on turn into pseudogenes and are deleted as we discussed earlier. Second, as the repair gene deleted have introduced some strong mutational bias, their loss will modify quickly the mutational bias of the genome [an estimated 0.5% in GC percent was observed after 2,000 generations in a mutT background (Cox and Yanofsky 1967)]. Such modification of genome composition will drastically reduce the efficiency of homologous recombination with ancestral populations that kept their initial GC percent. This will lead to the progressive genetic isolation of the mutator populations and limit even further the potential reacquisition by horizontal transfer of the lost repair genes. Hence as large population size allows the long term persistence of mutator populations, it also provides the conditions for the observed lack of reversion to non-mutator state as the fast accumulation of mutations in the genome reduces the chances of acquiring functional repair genes through recombination.

The two hypotheses proposed to explain genome reduction in large populations, the selective or the non-selective one, make different predictions, notably with respect to the timing of gene loss. In our hypothesis, the loss of DNA repair genes is expected to be an early event leading to mutator or super-mutator phenotype whereas it could occur at any time in the process of genome reduction in the selective one. Another interesting aspect of our hypothesis is that it is quite universal since elevated mutation rate could play a key role in reduction of endosymbiotic bacterial genomes too. Loss of DNA repair genes may also be one of the first events of adaptation to their hosts and not a late one as commonly admitted. In fact two studies suggest that the DNA repair genes are found among the first genes lost during the genome reduction process in *Shigella flexneri* and *Salmonella typhi* free-living pathogens (Dagan et al. 2006) or *Sitophilus zeamais* and *S. oryzae*, the recent endosymbionts of maize and rice weevils (Dale et al. 2003). Increasing mutation rate may well be a good strategy in the evolutionary red queen-like race between hosts and parasites, especially because the opportunities for sex are scarce for endosymbionts. Increasing mutation rate could be an alternative to doing sex as it has been noted before (Tenaillon et al. 2000). Interestingly, this increased rate and not Muller's ratchet have been suggested to be the major determinant of accelerated evolution in endosymbionts (Itoh et al. 2002) and this could also lead to genome reduction according to our view. In conclusion, our work reinforces the idea that variation in mutation rate may contribute to the diversity of genome size observed in nature.

Acknowledgements We thank Siv Andersson, Vincent Daubin, and Eduardo Rocha and Pierre Alexis Gros for helpful comments on this manuscript. G.M. is a CNRS fellow, A.C. has a PhD fellowship from the French ministry of research and O.T. was funded by Agence Nationale de la Recherche grant (ANR-05JCJC0136-01).

References

André JB, Godelle B (2006) The evolution of mutation rate in finite asexual populations. Genetics 172:611–626

Biebricher CK, Eigen M (2005) The error threshold. Virus Res 107:117–127

Coleman ML, Sullivan MB, Martiny AC et al (2006) Genomic islands and the ecology and evolution of Prochlorococcus. Science 311:1768–1770

Cox EC, Yanofsky C (1967) Altered base ratios in the DNA of an Escherichia coli mutator strain. Proc Natl Acad Sci USA 58:1895–1902

Dagan T, Blekhman R, Graur D (2006) The "domino theory" of gene death: gradual and mass gene extinction events in three lineages of obligate symbiotic bacterial pathogens. Mol Biol Evol 23:310–316

Dale C, Wang B, Moran N et al (2003) Loss of DNA recombinational repair enzymes in the initial stages of genome degeneration. Mol Biol Evol 20:1188–1194

Denamur E, Lecointre G, Darlu P et al (2000) Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. Cell 103:711–721

Dufresne A, Salanoubat M, Partensky F et al (2003) Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. Proc Natl Acad Sci USA 100:10020–10025

Dufresne A, Garczarek L, Partensky F (2005) Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol 6:R14

Eigen M (1971) Selforganization of matter and the evolution of macromolecules. Naturwissenschaften 64:465–523

Gerrish PJ, Colato A, Perelson AS et al (2007) Complete genetic linkage can subvert natural selection. Proc Natl Acad Sci USA 104:6266–6271

Giovannoni SJ, Tripp HJ, Givan S et al (2005) Genome streamlining in a cosmopolitan oceanic bacterium. Science 309:1242–1245

Haigh J (1978) The accumulation of deleterious genes in a population—Muller's Ratchet. Theor Popul Biol 14:251–267

Itoh T, Martin W, Nei M (2002) Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. Proc Natl Acad Sci USA 99:12944–12948

Kurland CG, Andersson SG (2000) Origin and evolution of the mitochondrial proteome. Microbiol Mol Biol Rev 64:786–820

Lynch M (2006) The origins of eukaryotic gene structure. Mol Biol Evol 23:450–468

Lynch M, Conery JS (2003) The origins of genome complexity. Science 302:1401–1404

Mackay WJ, Han S, Samson LD (1994) DNA alkylation repair limits spontaneous base substitution mutations in Escherichia coli. J Bacteriol 176:3224–3230

Martiny AC, Coleman ML, Chisholm SW (2006) Phosphate acquisition genes in Prochlorococcus ecotypes: evidence for genome-wide adaptation. Proc Natl Acad Sci USA 103:12552–12557

Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. Trends Genet 17:589–596

Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. Cell 108:583–586

Rocap G, Larimer FW, Lamerdin J et al (2003) Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. Nature 424:1042–1047

Rocha EP, Danchin A (2002) Base composition bias might result from competition for metabolic resources. Trends Genet 18:291–294

Saumaa S, Tover A, Kasak L et al (2002) Different spectra of stationary-phase mutations in early-arising versus late-arising mutants of Pseudomonas putida: involvement of the DNA repair enzyme MutY and the stationary-phase sigma factor RpoS. J Bacteriol 184:6957–6965

Sniegowski PD, Gerrish PJ, Lenski RE (1997) Evolution of high mutation rates in experimental populations of E. coli. Nature 387:703–705

Taddei F, Radman M, Maynard-Smith J et al (1997) Role of mutator alleles in adaptive evolution. Nature 387:700–702

Tenaillon O, Toupance B, Le Nagard H et al (1999) Mutators, population size, adaptive landscape the adaptation of asexual populations of bacteria. Genetics 152:485–493

Tenaillon O, Le Nagard H, Godelle B et al (2000) Mutators and sex in bacteria: conflict between adaptive strategies. Proc Natl Acad Sci USA 97:10465–10470

Vulic M, Dionisio F, Taddei F et al (1997) Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc Natl Acad Sci USA 94:9763–9767

Wiehe T (2000) Model dependency of error thresholds: the role of fitness functions and contrasts between the finite and infinite sites models. Genetical Research 69:127–136