UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 611

Evolutionary Dynamics of Mutation and Gene Transfer in Bacteria

PETER A LIND

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2010

ISSN 1651-6206
ISBN 978-91-554-7923-7
urn:nbn:se:uu:diva-132262

Dissertation presented at Uppsala University to be publicly examined in C4:301, BMC, Husargatan 3, Uppsala, Friday, December 3, 2010 at 13:00 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English.

Abstract
Lind, P A. 2010. Evolutionary Dynamics of Mutation and Gene Transfer in Bacteria. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine* 611. 80 pp. Uppsala. ISBN 978-91-554-7923-7.

The study of bacterial evolution is fundamental for addressing current problems of antibiotic resistance and emerging infectious diseases and lays a solid foundation for successful and rational design in biotechnology and synthetic biology. The main aim of this thesis is to test evolutionary hypotheses, largely based on theoretical considerations and sequence analysis, by designing scenarios in a laboratory setting to obtain experimental data. Paper I examines how genomic GC-content can be reduced following a change in mutation rate and spectrum. Transcription-related biases in mutation location were found, but no replicative bias was detected. Paper II explores the distribution of fitness effects of random substitutions in two ribosomal protein genes using a highly sensitive fitness assay. The substitutions had a weakly deleterious effect, with low frequencies of both neutral and inactivating mutations. The surprising finding that synonymous and non-synonymous substitutions have very similar distribution of fitness effects suggests that, at least for these genes, fitness constraints are present mainly on the level of mRNA instead of protein. Paper III examines selective barriers to inter-species gene transfer by constructing mutants with a native gene replaced by an orthologue from another species. Results suggest that the fitness costs of these gene replacements are large enough to provide a barrier to this kind of horizontal gene transfer in nature. The paper also examines possible compensatory mechanisms that can reduce the cost of the poorly functioning alien genes and found that gene amplification acts as a first step to improve the selective contribution after transfer. Paper IV investigates the fitness constraints on horizontal gene transfer by inserting DNA from other species into the *Salmonella* chromosome. Results suggest that insertion of foreign DNA often is neutral and the manuscript provides new experimental data for theoretical analysis of interspecies genome variation and horizontal gene transfer between species.

Keywords: fitness cost, bacterial evolution, gene amplification, mutational biases, GC content, synonymous substitutions, horizontal gene transfer, experimental evolution

Peter A Lind, Department of Medical Biochemistry and Microbiology, Box 582, Uppsala University, SE-75123 Uppsala, Sweden.

ISSN 1651-6206
ISBN 978-91-554-7923-7
urn:nbn:se:uu:diva-132262 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-132262)

Seen in the light of evolution, biology is, perhaps, intellectually the most satisfying and inspiring science. Without that light it becomes a pile of sundry facts—some of them interesting or curious but making no meaningful picture as a whole.

Theodosius Dobzhansky

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I **Lind PA**, Andersson DI. (2008) Whole-genome mutational biases in bacteria. *Proc Natl Acad Sci U S A*, 105(46):17878-83.

II **Lind PA**, Berg OG, Andersson DI. (2010) Mutational robustness of ribosomal protein genes. *Science*. (Accepted)

III **Lind PA**, Tobin C, Berg OG, Kurland CG, Andersson DI. (2010) Compensatory gene amplification restores fitness after inter-species gene replacements. *Mol Microbiol.* 75(5):1078-89.

IV **Lind PA**, Andersson DI. (2010) Fitness constraints on horizontal gene transfer. (Manuscript)

Contents

Abbreviations

W	Absolute fitness
w	Relative fitness
s	Selection coefficient
N_e	Effective population size
N	Total population size
E. coli	*Escherichia coli*
S. typhimurium	*Salmonella enterica* serovar Typhimurium
A	Adenine
T	Thymine
C	Cytosine
G	Guanine
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
tRNA	Transfer RNA
mRNA	Messenger RNA
rRNA	Ribosomal RNA
nt	Nucleotide
bp	Base pair
kb	Kilo base pairs
Mb	Mega base pairs
aa	Amino acid
ORF	Open reading frame
HGT	Horizontal gene transfer
BGC	Biased gene conversion
MMR	Methyl-directed mismatch repair
SSB	Single-strand binding protein
5-meC	5-methyl cytosine
UV	Ultraviolet radiation
BER	Base excision repair
NER	Nucleotide excision repair
8-oxoG	7,8-hihydro-8-oxoguanine
ssDNA	Single stranded DNA
dsDNA	Double stranded DNA
AIMS	Architecture imparting sequences
DSB	Double strand break

ICE	Integrative conjugative element
IS element	Insertion sequence element
MA	Mutation accumulation
GFP	Green fluorescent protein
CFP	Cyan fluorescent protein
YFP	Yellow fluorescent protein
RBS	Ribosomal binding site
DFE	Distribution of fitness effects
U	Uracil
AP site	Apurinic or apyrimidinic site in DNA
crRNA	CRISPR-derived RNA

Introduction

Evolution, as introduced by Charles Darwin's theory of natural selection (1), has been the unifying concept of biology for generations with never-ending discussions and disputes over the development of evolutionary theory. Contrary to the perception of a significant part of the public the scientific controversy has not been about if the theory works, which is extremely well established, but rather how it works and why it works. Evolution is a solid scientific theory and this does not mean that it simply is a really good idea, but that we can create testable hypotheses to examine the details of the theory.

Natural selection will occur when we have a population of entities that fulfills the prerequisites of variation, reproduction and heredity, given that the variation does not blend. This is certainly fulfilled for biological life where genetic variation exists and does not blend (Mendelian genetics) and it is heritable during reproduction, so it follows that natural selection is inevitable. The algorithm of natural selection does not require a guiding hand to explain the fit of organisms to their environment, disease, sex or the emergence of societies, but it is important to understand that evolution is not a random process, because the survival of the randomly generated diversity is non-random.

However, the adaptive process of natural selection cannot alone explain all aspects of evolution. There are also three non-adaptive processes that are not dependent on the fitness of Darwinian individuals. First, mutation creates the genetic diversity upon which natural selection act (2). Second, this diversity is then shuffled by recombination and third, genetic drift makes the genetic variation differ somewhat between generations due to chance events independent of fitness (2).

The quote by the brilliant geneticist Theodosius Dobzhansky "Nothing in biology makes sense except in the light of evolution"(3) describes how I believe we should think about evolutionary biology. The theory of evolution can provide us with answers on how life on earth is likely to have evolved and diversified with increasing complexity. The wonder of this diversity is perhaps only surpassed by the amazing unity of life in that all organisms on earth are related and function using the same basic biochemistry. The study of molecular evolution provides the link between the observable phenotypic variation and the biological macromolecules that are ultimately responsibly for the variation. With an increased understanding of the functions of evolved biological systems, we are also better equipped to address some of

the most pressing problems of our time. The evolution of pathogenic bacteria resistant to almost all available antibiotics will make previously trivial infections life-threatening once again. This development is partly caused by a limited understanding of the evolutionary potential of bacteria, which must be taken into account when introducing and developing new antibiotics. The globalization process now allows a much more rapid spread of diseases between all continents. We must use our biological knowledge to make well-informed probability assessments of the risk of a pathogen spreading from animal to humans, acquiring drug resistance or mutating to increased pathogenicity. What are the risks of spreading genes from genetically modified organisms to other species and will pathogenicity traits transfer between bacterial species and result in new increasingly pathogenic strains? Is the future of our food production threatened by the lack of genetic variability among the main food crops and animals? How are the health problems in the western world dependent on the approximately 10^{14} bacteria in our gut?

Climate change, overuse of natural resources and pollution change the environment worldwide and provide new selective pressures that cause the collapse of ecosystems and extreme evolutionary challenges to the species therein. Understanding evolutionary design principles will be essential in developing biotechnology and synthetic biology to perform increasingly complex tasks that can provide solutions to diseases, shortages in food production and providing clean energy. The synthesis of evolutionary theory with experimental molecular biology, ecology, population biology, systems biology and descriptive -omics data has great potential to unify biology.

Nothing in evolution makes sense except in light of population genetics

The papers in this thesis focus mainly on experimental studies of evolution, but a brief introduction to the main concepts of the underlying theoretical framework is necessary to understand the importance of the research questions, the relevance of the results and the experimental designs of the studies. I have tried to keep this introduction to a minimum, leaving out the interesting historical details and disputes that have shaped the science of evolutionary biology as well as the rigorous mathematical basis, to focus on presenting an incomplete, but hopefully sufficient picture of the concepts needed to understand the thesis.

Unfortunately, evolutionary biology is not a subject best addressed using common sense arguments as these often fail to predict how evolutionary processes work and cannot grasp the complexity inherent in biology. Instead we must turn to population genetics to provide us with the mathematical tools needed to test evolutionary hypotheses and predict the importance of

the various evolutionary forces, and not only natural selection, under different scenarios. In the words of Michael Lynch "Nothing in evolution makes sense except in light of population genetics"(2). Population genetics is the study of genetic variability and change in a population and the processes that influence these factors. It allows the abstraction of biological details to useful concepts. One example is the reduction of all biochemical, genetic and physiological variation between individuals or populations to a selection coefficient, a single number representing fitness (4). Population genetics incorporates the fundamental stochasticity of the real world by focusing on the probability of the survival of the fittest, instead of relying on simplistic slogans such as "survival of the fittest". The mathematical models of population biology are of course based on simple molecular biological assumptions and can only help us understand evolution when these assumptions are realistic (5). Nevertheless, population genetics sets the limits of what is theoretically possible and can help us choose between different evolutionary scenarios when applied to empirical data. In the next sections two fundamental concepts of population genetics will be introduced, fitness/selection coefficient (s) and effective population size (N_e).

Fitness

The exact population genetic definition of fitness is somewhat varying, but the main point is that it concerns the ability of organisms to survive and reproduce in their environment (4). When variation of fitness depends on the genotype, natural selection is possible and the frequency of the genotypes change between generations. The road to successful propagation of genes to the next generation will, of course, depend on many factors in an organism's life cycle. These factors can be partitioned into fitness components, including viability, reaching reproductive age and mating success, which can be measured both in laboratory and natural populations (4). The absolute fitness (W) of a genotype is a composite fitness statistic that will be 0 (for lethal genotypes) or larger. However, this measure of fitness is usually less informative than relative fitness (w), which is normalized to the fittest genotype or wild type in the population. The selection coefficient is used to compare genotypes so that w=1+s, where s is positive if the genotype has higher fitness than the wild type, or negative if the genotype has lower fitness, and can be used to calculate rates and probabilities of changes in allelic frequencies (6). Commonly, s is used to predict the fate of a new mutant in a population.

Fitness is, of course, tightly connected to the organisms' environment and ecology and is not expected to be constant through time and if the lifestyle involves different ecological niches it is not expected that one genotype is superior in all of them. The genotype that will dominate in this case is the one with the highest geometric mean over time. Thus, when average fitness

is the same, natural selection leads to the survival of the genotype with the smallest variance in fitness over time and not necessarily the most fit in any one environment (4). It is also important to distinguish between the fitness of an individual, which is not only dependent on the genotype but also on stochastic variation in physiology and environment, and the way that fitness will be used in this thesis, as a statistic of a genotype (4).

Landscape models have commonly been used to understand the adaptive trajectories available to organisms. First introduced by Sewall Wright (7), his adaptive landscape is made up by two axes (the horizontal plane) that represent the allele frequency at locus 1 and 2, respectively, and the third (vertical) axis represents the mean fitness of the population, or a phenotypic trait correlated to fitness. The surface of the landscape shows peaks of high fitness and valleys of low fitness and selection can be seen as the population climbing up hill towards a fitness peak with every beneficial mutation/recombination fixed in the population. This simplified landscape only visualizes two alleles, whereas for real organisms fitness depend on many loci, which make the landscapes multidimensional (8). Several variants of the landscape concept have been put forth designated selective landscapes, fitness landscapes, phenotype landscapes and mutational landscapes depending on what facet is focused upon (5, 9-14). Modern landscapes are often focused on discrete DNA or protein sequences rather then continuous frequency distributions and as the number of possible sequences is often extremely large, the dimensions of the fitness landscape will increase accordingly, which makes the abstraction less intuitive (4, 14, 15). Still, landscape models can improve our understanding of how populations might be stuck on local adaptive peaks rather then on global peaks and how neutral sequence variations can give access to different adaptive paths (Figure 1). Interestingly, the highly dimensional genotypic landscapes seem to contain large neutral regions, unlike the low-dimensional ones commonly used in population genetics (16). However, it is important not to imagine the landscape as static, but instead changing over time given that environmental variation will change the relative fitness associated with a particular genotype (15).

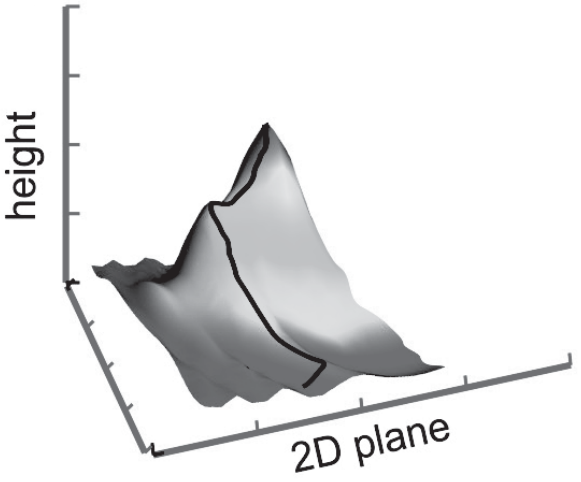

Figure 1. Illustration of a fitness landscape. The black line represents an adaptive walk. From Loewe (2009) (5), reproduced under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0).

Effective population size

Effective population size (N_e) is a concept introduced by Sewall Wright (17, 18) that is related to, but most often smaller than, the actual population size (N) (19). It is an idealized population that show the same amount of genetic diversity, or the same amount of inbreeding, as the population studied and can be used to describe the influence of genetic drift in a population and the effectiveness of selection (20). This abstraction reduces the influence of population history, spatial structure and genetic structure to a single number to be used in population genetic models (21). Changes in population size can greatly influence N_e, so that periods with low population size, often called bottlenecks, severely reduce the genetic variability and N_e will then be determined by the harmonic mean of population sizes (22). This explains why the effective population size of humans is estimated to 10^4, when the total population size is 7×10^7, which demonstrates how closely related all humans really are (23). A proper understanding of effective population size is also necessary in the design of experimental evolution studies where bottlenecking is often used. As an example, the N_e of an experimental population with final population size of 10^9 will vary 400,000-fold depending on if the bottleneck size is 1 or 10^6 cells, and this will have a large influence on the fate of new mutations arising in the population. N_e is also affected by the balance of individuals of each sex in sexually reproducing populations, variation in offspring number, inbreeding and age and stage structure (21).

The abstract nature of the effective population size concept becomes even more obvious when considering how the genetic structure can influence N_e, so that it varies across the genome of a species (24). Selection on a genetic locus will also influence the genetic diversity of closely linked regions, which can have large effects on N_e under conditions of balancing selection, background selection and after a selective sweep (21). Neutral diversity correlates positively with the rate of recombination in several higher eukaryotes (21). Although bacteria do not reproduce sexually and do not experience chromosome crossover during meiosis, there is evidence that at many species is subject to significant rates of recombination, with new genetic material acquired through horizontal gene transfer from both closely related strains and more distantly related species (25, 26). This means that N_e at any location in the genome is dependent on selection on nearby sites and the rate of recombination (25-27).

The fate of new mutations

Mutations are in its broadest sense any change in the genetic material of an organism and population genetic theory does not *per se* distinguish between different types of mutations. Thus we can address the question of the fate of a mutation in a general way, even though there are large variations in their fitness effects and rates.

Mutational fitness effects are generally divided into three categories: deleterious mutations reduce the fitness of the organism, neutral mutations have very small effects on fitness and advantageous mutations increase the fitness (20). This division of the continuum of fitness effects into three categories is based on the expected outcome for the mutations, i.e. the fixation probability. The probability of fixation is determined not only by the magnitude of the fitness effect, but also by the effective population size according to $P_{fix}= s/(1-e(-s \times N_e)) \times N_e/N$, where P_{fix} is the probability of fixation, s the selection coefficient and N_e and N the effective and total population size, respectively (28). Neutral mutations have selection coefficients so small that their fate is largely determined by random genetic drift, which is fulfilled when $N_e \times|s|<<1$ holds. If the selection coefficient is larger in magnitude, so this relation does not hold, the mutation is not neutral and is classified as deleterious if $s<0$ and advantageous if $s>0$. The proportion of mutations in each category varies between species, dependent on N_e, ecological niche and how well adapted the population is to the environment, and within species depending on the type and site of the mutation (29, 30).

Advantageous mutations make natural selection possible and to understand the evolutionary dynamics we must consider not only N_e and s, but also the rate of new mutations. In an asexual population the rate of increase in fitness will be proportional to the rate of advantageous mutations, which is expected to be low in well-adapted populations (estimated to 1 in every 10^4-

10^5 mutations in *E. coli*), and the average fitness effect of the fixed mutations (31, 32). N_e will largely determine which advantageous mutation will be fixed. If the population is small it is mostly waiting for the next advantageous mutation, which will then rapidly sweep the entire population before the next advantageous mutation arrives and this scenario is called periodic selection (32). When N_e grows, the supply of advantageous mutations increase in proportion and therefore the population might at any time contain several different advantageous mutations, which cannot be efficiently combined into one genotype because of the low rate of recombination, so instead clones will compete against each other in a scenario called clonal interference (33). This phenomenon will have important effects, as the mutations with the largest beneficial effects will be fixed first and the fixation time of the most beneficial mutation will increase (33). In environments that have a spatial structure or availability of alternative carbon and energy sources, thus basically all natural environments, ecological specialization is expected to allow several genotypes to coexists for extended periods of time preventing selective sweeps in all subpopulations (34-36).

The distribution of fitness effects (DFE) of the rare advantageous mutations is generally believed to be exponential, at least in well-adapted populations, as predicted by extreme value theory (37, 38) and there are several experimental studies supporting this (31, 39, 40). This means that the majority of the advantageous mutations will cause a very small increase in fitness, but due to clonal interference these will rarely be the winners in laboratory or natural populations (31, 33). The DFE of deleterious mutations might be more complex, because a proportion of the mutations is complete loss-of-function mutations, which gives a bimodal distribution with one peak at low fitness, including lethal mutations, and one peak closer to wild-type fitness (41, 42). Purifying selection will effectively remove deleterious mutations when $N_e \times s < -1$, so that mutations with large fitness costs will be extremely rare in natural populations. When N_e is small the effect of genetic drift will be large enough to allow accumulation of slightly deleterious mutations, which can lead to a decrease in fitness over time in a process known as Muller's ratchet (43).

An increase in mutation rate will increase the supply of advantageous mutations, but also the frequency of neutral and deleterious mutations, and the evolutionary dynamics and molecular evolution will depend on the proportion and distribution of fitness effects of mutations in each category. If we want to understand and accurately predict the evolutionary dynamics of natural populations it is evident that we must obtain empirical data on the mutation rates, effective population sizes and the distribution of fitness effects and that these data will depend on the nature of the mutations as well as the species' genetic architecture and ecological context. The deleterious effect of a mutation can often be reduced by a compensatory mutation at a secondary site that increase fitness, so that the effect of the combined muta-

tions can be neutral or advantageous (44, 45). Advantageous mutations that allow access to a previously unavailable environmental niche, *e.g.* antibiotic resistance, commonly cause a large decrease in fitness in the original environment (46). Compensatory mutations can then allow the spread of the new genotype back into the old niche, in this example the antibiotic-free environment, if the fitness cost of the resistance mutation has been ameliorated sufficiently (46, 47).

The nearly neutral theory of molecular evolution

Molecular biology has allowed evolutionary biology to go from qualitative studies of phenotypes to quantitative studies of the underlying properties of the genetic material, which can be firmly connected to population genetics. When the large molecular variation within and between different species became clear in the 1960s Kimura suggested that this diversity could not be caused by natural selection, but was the result of random genetic drift (48). Although this idea was not unique, Kimura formalized a complete description of the dynamics of neutral mutations using an elegant mathematical framework from particle physics. The theory was later generalized by Ohta who introduced the concept of near neutrality emphasizing that the neutral dynamics is not only determined by the fitness effect (s), but also by the effective population size (N_e) (49, 50). This view did not deny the importance of selection in adaptive evolution of form and function, only stating that most of the molecular changes observed did not cause fitness effects large enough to be selected in natural population (20). Neither did the theory make any strong predictions about the presence of selective constraints on the molecular level, not even excluding selection on synonymous codon usage, nor the fraction of mutations that would be deleterious, neutral or advantageous, only stating that the majority of the molecular differences between and within natural populations is due to neutral processes (20). The nearly neutral theory is now commonly used as a null model to test if DNA sequences have been under natural selection (51, 52).

Selective constraints on mutations

In the broadest sense, a mutation is any heritable change in the genetic material (DNA for cellular life) of an organism and thus the source of the variation that natural selection acts upon. These changes can be divided into mutations that change the content of the genetic material through single nucleotide substitutions, inversions and some recombination events and mutations that change the amount of DNA including gene amplifications, deletions and horizontal gene transfers (Figure 6). The rates of the various types of mutational events varies over several orders of magnitude and the selective con-

straints and distributions of fitness effects are expected to be highly variable, so their relative importance varies depending on the evolutionary scenario. There is a vast literature on mutation mechanisms, rates and effects, but the focus in this section is on bacterial species, often *E. coli* and *S. typhimurium*, unless special insight can be gained from examples from other organisms.

Fitness effects of base pair substitutions

Single-nucleotide substitutions are changes in one of A, T, C or G bases into another, caused by errors in DNA replication, repair or chemical DNA damage. Depending on the position of the substitution, the effects on fitness will vary widely. For intragenic changes the nature of the genetic code causes three consequences on the protein level: synonymous substitutions do not change the amino acid sequence, non-synonymous substitutions change the identity of the amino acid and nonsense mutations introduce a stop codon that results in a truncated protein.

Synonymous substitutions

Synonymous substitutions are often assumed to have small fitness effects and have been used as a neutral control to which selective constraints and positive selection are compared (51, 52). Although it has long been known that all codons for the same amino acid are not used at random so that usage varies both on the level of genes and genomes, this could be caused both by neutral processes and natural selection. The relative importance of neutral and selective forces varies depending on the genetic context and causes different patterns of nucleotide usage.

Codon usage in highly expressed genes in some bacteria has been found to be strongly correlated to the abundance of tRNAs, indicating selection for translational efficiency (53, 54). One example is the high codon usage bias in ribosomal protein genes that are very highly expressed and due to their pivotal role in the protein synthesis machinery are expected to be under strong purifying selection (55, 56). Large changes in synonymous codon usage in green fluorescent protein (GFP) have also been shown to be important in experimental studies, where the amount of over-expressed protein varied over several magnitudes (57, 58). Translational efficiency may also involve selective constraints on misfolding of proteins caused by mistranslation (59). As translation has limited accuracy, with errors in the incorporated amino acid occurring at rates of one per 10^3–10^4 codons (60-62), a significant fraction of an average length protein in the cell will contain an amino acid change. Selection can act on several levels to reduce the impact of misfolded proteins, including increased translational accuracy to increase the probability of a correct amino acid sequence, increased translational robustness to reduce the fraction of erroneous translated proteins that misfold and decrease the proportion of correctly translated proteins that fail to fold prop-

erly or unfold (59). Synonymous mutations can also disrupt mRNA secondary structure, which can have major effects on translation, and influence mRNA levels through changes in transcriptional efficiency and mRNA degradation (58, 63). Especially the mRNA stability close to the ribosomal binding site has been reported as essential for high expression with protein levels (GFP) varying 250-fold depending on the synonymous codon usage (63). Selection on synonymous codon usage is expected to vary depending on the gene's contribution to fitness and its expression level and is not expected to cause genome wide biases (59, 64). Additional selection on codon usage in highly expressed genes could also arise from an optimization of the function of the translation machinery. Each ribosome is a major mass investment so in order to increase the cell's protein synthesis rate, thereby maximizing its growth rate, the translation system should be optimized to reduce sequestration of ribosomes (64).

Large effects of synonymous substitutions have also been reported in other systems. Deoptimization of codon usage in poliovirus has been shown to cause a decrease in fitness and attenuation of virulence in mice (65). There are numerous examples of synonymous substitutions with large effects in eukaryotes, including misfolding (66), slicing/exon skipping (67-69) and reduced levels of enzymes (70, 71).

Fitness effects of non-synonymous mutations

The selective constraints for non-synonymous substitutions will be exactly the same as for the synonymous ones with the additional constraint of the fitness effects of the amino acid change in the resulting protein. Clearly, a change in the protein can cause a substantial fitness cost as it has the potential to completely destroy the function of the protein; in this case the fitness effect is similar to an insertion or deletion mutation. Examples of this include, changes in amino acids involved in catalytic sites, interactions with other gene products and changes in thermodynamic stability large enough to substantially change the tertiary structure of the protein (72, 73). Changes in protein stability will be largely dependent on the nature of the change in terms of size and polarity of the amino acid, in relation to the native one, and the location in terms of buried or solvent-exposed, where changes in the protein core will generally cause larger effects (73, 74). Experimental measurements and computational predictions of thermodynamic stability of mutants with single amino acid changes suggest that the majority of substitutions cause small changes in stability (75, 76). Most of them are weakly destabilizing, but rarely to the degree that they confer any major change in the tertiary structure of the protein (75, 76). However, if additional amino acid changes are introduced it seems that the stability is more severely affected than expected by the single changes alone, suggesting a threshold model for the effects of non-synonymous substitutions on thermodynamic stability and related effects on fitness (74, 75).

A large genetic study of the phenotypic effects of amino acid substitution in the tetrameric lac-repressor *lacI* in *E. coli* revealed that over 44% of amino acid positions tolerated substitutions and that these were often in spacer regions in the protein structure (73). Solvent-exposed amino acids were also mostly tolerant of substitutions, with exceptions involving mainly those involved in stabilization of specific structural motifs. Core amino acids and proline substitutions had greater effects on function, which are probably caused by disruption of protein folding and amino acids involved in dimerization or inducer binding did not tolerate substitutions (73). The phenotypic character of this kind of data can provide valuable information on protein function and selective constraints, but its relevance to fitness effects is less clear as very small changes in protein function can affect selection coefficients sufficiently to be highly deleterious in an evolutionary model even if this cannot be detected as a changed phenotype. A similar study on phenotypic effects in *E. coli* RNA polymerase, encoded by *rpoB*, revealed that a majority (363/465) of amino acid substitutions caused cellular death or altered phenotype, although this result might not be general because the mutations were not random (77).

Other types of nucleotide substitutions

Substitutions changing a start codon or stop codon will generally have large effects on fitness. Mutation of a start codon will often cause a complete loss of function as no protein is produced although changes to alternative start codon can cause smaller effects. The fitness effects of introducing a stop codon in a gene, resulting in a truncated protein, will depend on the location of the substitution, but are generally expected to have large fitness costs. Loss of a stop codon, resulting in increased protein length, often have smaller effects on protein function, although there will be an extra cost for producing a longer protein and translation can interfere with adjacent genes. Substitutions outside ORFs are more likely to be neutral, but a significant minority is located near or in promoter sequences and thus has the potential of causing large changes in transcription, which most often would be detrimental, but many advantageous and compensatory mutations are likely to be found in regulatory sequences ((78) and **paper III**).

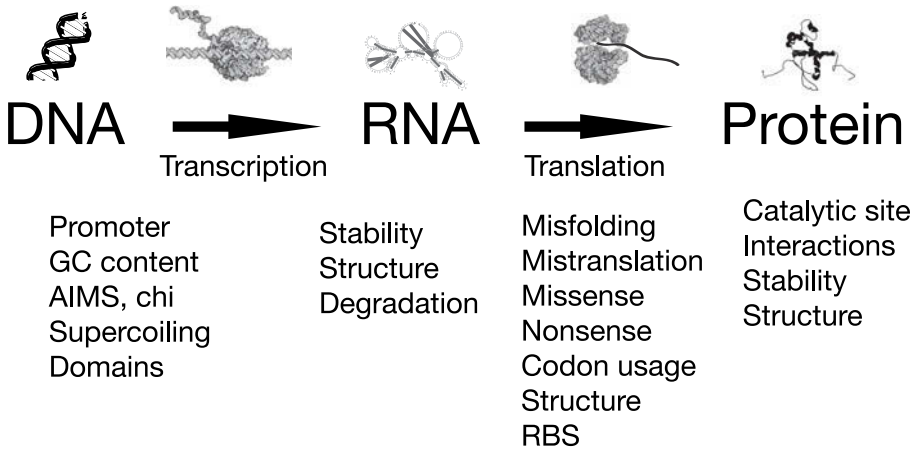

Figure 2. Selective constraints on substitutions in bacteria.

Recombination

The impact of recombination for bacterial species is not fully understood and likely to vary between different species (26). Bacteria are not obligate recombinogenic as higher eukaryotes are, where mixing of genomes occur during reproduction. However, homologous recombination has long been used for genetic manipulations in the laboratory where DNA has been exchanged between members of the same species and closely related species, so clearly there is potential for a significant role of recombination in natural populations (25). This is most certainly the case within some species where substitutions introduced by recombination outnumber other mutations so that phylogenetic trees of single house-keeping genes cannot be constructed due to lack of signal (79). Even when there is clear linkage disequilibrium, so that association of gene variants is non-random within a population, this should not be interpreted as lack of recombination given that genetic change by recombination must be 20 times more frequent than nucleotide substitutions for the linkage to disappear completely (80). The fitness effects of recombination will often be neutral or advantageous when homologous genes are introduced from members of the same species, but both the probability of neutrality and the efficiency of homologous recombination will decrease with increasing phylogenetic distance. When sequence divergence between donor and recipient species is high enough to prevent incorporation into the genome by RecA-dependent homologous recombination, the probability for gene replacements will be severely reduced and successful integration into the genome will occur through non-homologous recombination, most often as an insertion event adding new genetic material to the genome (81).

Horizontal gene transfer

New genetic material from both closely and distantly related species can be introduced into a bacterial genome through horizontal gene transfer (also called lateral gene transfer) mediated by transduction, conjugation and transformation and integrated by homologous and non-homologous recombination mechanisms. The selective constraints on HGT are poorly understood but some general principles have emerged (Figure 3). Bacterial genomes have very high coding densities, often with about 90% of the sequence in ORFs (open reading frames) and an additional 5% within operons (42). This means that HGT insertion events have a high probability of disrupting native genes and operons and it has been shown in laboratory experiments that a large majority of insertions (at least 80%) are highly deleterious from a population genetics perspective (42). Consequently, integration sites of successful HGTs will be biased towards regions with low selective contributions. The cost of maintaining extra DNA is believed to be small, but disruption of chromosome structure, affecting replicore length, supercoiling and gene order, may sometimes cause large fitness costs (82).

If the HGT contains transcription start sites, this can impose a fitness cost in terms of the biosynthetic cost of the nucleotides used for RNA production and wasteful use of RNA polymerase molecules. In cases where the mRNA contain functional ribosome binding sites, the fitness cost of translation is likely to be significant, given the large mass investment in production of the ribosome, which suggests that highly expressed foreign proteins would most often be strongly counter-selected (83). Divergent alien proteins can also have toxic effects on the cell, especially in cases where homologues are present in the recipient genome and the new protein disturbs physical interactions and disrupts kinetic optimization. This makes successful HGT events more probable at the edges of functional networks and when the donor species is closely related, as this increases the chance of transcriptional and translational regulation compatibility. When improperly regulated the alien gene would most often not remain functional in the recipient genome. In the case where it is constitutively expressed this could be advantageous under some favorable environmental conditions, allowing fixation in small subpopulations, but the patchiness of the environment would make it likely to be inactivated to reduce the cost of expression as soon as it is not directly selected (84). If the alien protein is very poorly expressed the cost of production will be low, but this also means that positive selection will rarely be strong enough to give a significant fixation probability and a high rate of advantageous secondary mutations might be necessary if the gene is not to be inactivated by random drift. AT-rich horizontally transferred DNA can be transcriptionally silenced by the histone-like nucleoid structuring protein (H-NS) in *S. typhimurium* and related bacteria (85-87). This could act as a defense against HGT and reduce the fitness costs of expressing the foreign

genes, while still maintaining the capacity to use them in a selective environment where amplification of the HGT could rapidly increase expression. The need for proper regulation is reflected in that entire operons are sometimes transferred, but this is expected to be limited to closely related species for regulation to function properly.

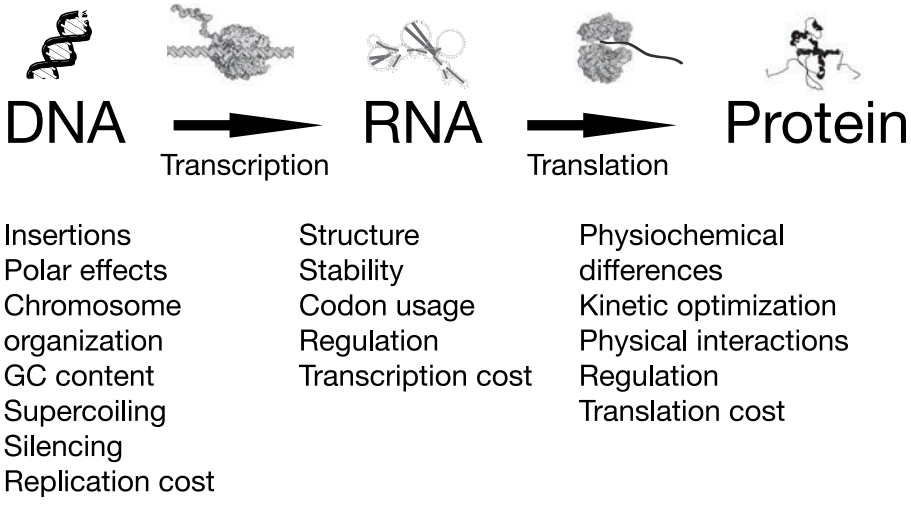

Figure 3. Fitness constraints on horizontal gene transfer in bacteria.

Gene duplication and amplification

In addition to the potential of gaining DNA from other cells, large regions of bacterial chromosomes can be duplicated both through homologous recombination and non-homologous recombination. Most often, the addition of genetic material will cause a fitness cost through the production of extra DNA, RNA and protein, but the effects will be highly variable depending on the content of the amplified region (88). Amplification of native genes can, just as HGT genes, cause disruption of the regulation and kinetic optimization of the cell. This suggests that it will generally be more advantageous to amplify complete functional networks for optimal function and this could provide a selective advantage of organizing related genes into operons, as this would allow the maintenance of proper regulation when entire regions are amplified.

If an initial duplication is adaptive in a new environment, e.g. in the presence of an antibiotic (89), the number of copies will quickly rise to an optimum level given the high efficiency of homologous recombination acting on large identical patches of DNA. As long as the selection pressure is applied the presence of multiple copies will be selectively stabilized, but unlike other types of mutations, amplifications are easily reversible due to the genetic

instability of gene amplifications and can be seen as a genetic regulatory response (89). Increasing gene dosage as an adaptive response also increases the probability of obtaining new advantageous mutations in the amplified region simple due to an increased target site. When multiple copies are present sub-functionalization of gene copies is possible and promiscuous functions can be selected without the loss of the original function. This process of duplication and divergence has been suggested to be a major contributor to the evolution of new genes (90, 91).

Deletions

The fitness costs of deletions can be analyzed using many of the observations made for insertions. For example, we know that >80% of random insertions are deleterious in *E. coli* (42), even if the insert itself does not cause a fitness cost. The conclusion is that inactivating genes, or in this case deleting them, will frequently cause large costs. Disruption of functional networks and kinetic optimization caused by deletion one component can be compared to the transfer of incomplete systems introduced by horizontal gene transfer. Thus, we would expect smaller fitness costs for deletion of entire operons, when the process is not of high selective value, which could increase the dynamic potential of genomes that have functional networks clustered. Deletions may also, again in the same way as insertions, disrupt chromosome organization, which in some cases would be highly deleterious (82). Considering the high coding density of bacterial genomes, fixed deletions will often be relatively small (<100 bp) (92). Minor deletions inside ORFs will generally cause large fitness costs when they cause a frameshift that destroys protein function, whereas loss of single amino acids may more often be tolerated.

Robustness and epistasis

Robustness confers reduced phenotypic sensitivity to a perturbation that can be either environmental or genetic. Environmental robustness buffers against changes in the physicochemical environment, as well as intrinsic stochastic variation, but is inheritable as exemplified by expression of heat shock proteins (93). Genetic robustness concerns heritable perturbations and describes the phenotypic variance caused by mutations, where a non-robust fitness peak has very few mutational neighbors with high fitness (93). Higher order fitness traits are dependent on lower level molecular mutational effects, e.g. thermodynamic stability and translational efficiency, and when the variance of the lower level effects is larger than the higher level effects this is called canalization (5). The opposite of canalization is capacitance, where the variance of the lower-level trait is lower. The complex interaction networks of the cell and the presence of buffering mechanisms often makes it hard to

predict the effects of several mutations based on their separate effects. In the absence of epistasis the effects are independent, but commonly genetic interactions result in larger effects than expected, which is known as synergistic epistasis, or smaller effects than expected, called antagonistic epistasis (94).

Mechanisms of mutation, repair and gene transfer

Mutation rates are sometimes used in ways where it is not clear what is taken into account. The mechanistic mutation rate is the rate of mutational change without the influence of selection. The evolutionary or substitution rate on the other hand, considers the rate of fixation or change in frequency that is dependent not only on the mechanistic rate, but also on s and N_e as discussed above, and is used when analyzing sequence data. The mechanistic rate is the rate the mutation is transferred to the next generation, which can be orders of magnitude lower than the rate of DNA damage with mutagenic potential due to the presence of highly efficient DNA repair systems that reduce the mutagenic impact. The mechanistic point mutation rate in *E. coli* and *S. typhimurium* is in the order of 5×10^{-10} per base pair per generation, whereas the evolutionary neutral substitution rate has been estimated to 5×10^{-3} per site per year (95, 96). If these rates were to be reconciled it would require that there were only 10 generations per year in contrast to common estimates of 100-1000 generations per year demonstrating a common discrepancy between evolutionary rates estimated from comparative sequence studies and direct experimental measurements (95, 97). Possibly, this is due to an underestimation of the selective constraints operating on presumed neutral positions, which results in overestimation of the neutral target size. Commonly, measurements of mutation rates in the laboratory are reported as rates, even though they are actually frequencies that are influenced by the design of the assay and can sometimes differ significantly from correctly calculated rates, which can have a significant impact on comparisons with estimated evolutionary rates (98).

Substitutions and DNA repair

Transversion is a change between a pyrimidine (T and C) and a purine (A and G) whereas transition is a point mutation that replaces a pyrimidine with a pyrimidine or a purine with a purine. Various processes leading to mutation often have biased rates so that either transitions or transversions are more common.

Accurate replication of DNA is fundamental to inheritance, but adaptive evolution also requires the generation of genetic variability through new mutations. The high rate of mutagenic DNA damage, caused by both endogenous and environmental factors, make the use of sophisticated DNA

repair systems selectively favored and more than 100 proteins are involved in reducing mutations in bacteria such as *E. coli* and *S. typhimurium*, with large variability in which systems are present, even between closely related bacteria (99, 100). These repair systems work on different types of mutagenic lesions using several distinct mechanisms. A comprehensive description of all the various repair systems is beyond the scope of this thesis and emphasis in this section is on presenting the main types of DNA repair and DNA damage, with focus on the aspects relevant for the papers presented in this thesis.

Errors in replications

Replication errors are a significant source of mutations in bacteria, where error rates for DNA synthesis *in vivo*, using the major replicative DNA polymerase III, are 10^{-7}-10^{-8} per bp in the absence of DNA mismatch repair in *E. coli* and *S. typhimurium* (101). This high accuracy is achieved through the processivity of the process, with a polymerase selectivity of 10^4-10^6 and a proofreading exonuclease activity increasing fidelity 10-10^2-fold (101). However, DNA polymerase III cannot efficiently replicate past lesions in the DNA or past some damaged bases and for this purpose a set of error-prone translesion DNA polymerases are used that can successfully repair the damaged DNA, but at the cost of a higher error rate of 10^{-1} to 10^{-3} (102, 103). The translesion DNA polymerases are quite specific for the type of DNA damage, with some overlapping activity, and in *E. coli* and *S. typhimurium* there are at least four additional polymerases to the normal replicative DNA polymerase Pol III, where Pol I and Pol II are gap-filling polymerases and Pol IV and Pol V are translesion polymerases. Mutations in the *polC* gene, encoding the alpha subunit of the DNA Pol III holoenzyme, can cause either a mutator or an antimutator phenotype by affecting the balance between association, dissociation and proofreading (100). Defects in the epsilon subunit of the DNA Pol III holoenzyme, encoded by *dnaQ*, can have major effects on exonuclease-associated proofreading and as this leads to large increase in errors during replication it can also cause a saturation of the mismatch repair system further contributing to an increase in mutation rate (100).

Mismatch repair

Errors introduced by replication can be repaired with high efficiency by the methyl-directed mismatch repair (MMR) system, which functions on both single base substitutions and insertion-deletion mismatches in the newly synthesized strand (104). It can also recognize some DNA lesions not associated with replication (104). MMR starts with the recognition of a mismatch by MutS, which interacts with MutL to activate the endonuclease activity of MutH. Which DNA strand that contains the replicative errors cannot be recognized on the basis of abnormal nucleotides, but as the newly synthesized

strand is not methylated shortly after replication, MutH can distinguish between the two strands by binding to unmethylated GATC sequences and cleaving the new strand within 1 kb of the error. DNA helicase II is loaded onto the cleaved strand together with single-strand binding protein (SSB) and detaches the single strand. The helicase is then followed by an exonuclease that degrades the single-stranded region and the resulting gap is filled by DNA polymerase III and sealed with DNA ligase (104). Mutations causing defects in the *mutS*, *mutL* and *mutH* genes can increase mutation rates up to 1000-fold (105-107).

In *E. coli* and closely related enteric bacteria 5-meC is normally present in DNA due to the action of the *dcm* gene product that selectively methylates the second cytosine of 5′-CCWGG sequences (W is A or T) (108, 109). Deamination of 5-meC leads to a T:G base pair that is recognized by the *vsr* gene product, a sequence specific endonuclease that creates a nick 5′ of the mispair and initiates the very short patch repair system restoring the site (Figure 8) (110-112). An interaction between Vsr and MMR system has been suggested as Vsr show reduced repair in the absence of MutL and MutS and overproduction of Vsr cause a 1000-fold increase in mutations rate in *E. coli* suggesting an inactivation of the MMR system (113).

Direct repair

A few types of DNA damage can be repaired directly without the need to excise the nucleotide. These include nicks, where a phosphodiester bond has been broken that can be repaired by DNA ligase, alkylation at O^6-position of guanine or the O^4-position of thymine that are removed by the Ada and Ogt enzymes, and AlkB that can repair 1-methyladenine and 3-methylcytosine through an oxidative demethylation mechanism (114, 115). A mutant lacking Ada and Ogt has an up to 10 times increased spontaneous mutation rate in *E. coli,* causing both transitions and transversions (116). UV-induced pyrimidine dimers are repaired by photoreactivation where a DNA photolyase use light to directly cleave the cyclobutane ring thereby restoring the normal bases (117, 118).

Base excision repair

Base excision repair (BER) is involved in repairing a vast number of different mutagenic lesions. The process is initiated by the recognition of a damaged base by a DNA glycosylase that flip the base out of the helix and cleaves the N-glycosidic bond leaving an apurinic or apyrimidinic (AP) site (119). AP sites can also be caused by depurination/depyrimidination and are substrates for AP endonucleases, including exonuclease III and endonuclease IV in *E. coli*, which cut the phosphodiester bond leaving a gap that is repaired by DNA polymerase I and DNA ligase (Figure 4) (120, 121). Mutants defective in both exonuclease III and exonuclease IV have a 4-6-fold increase in spontaneous mutations rate (100). The different types of DNA gly-

cosylases are quite specific in the spectrum of damaged bases they can repair and some are bifunctional and also possess an endonuclease activity (121).

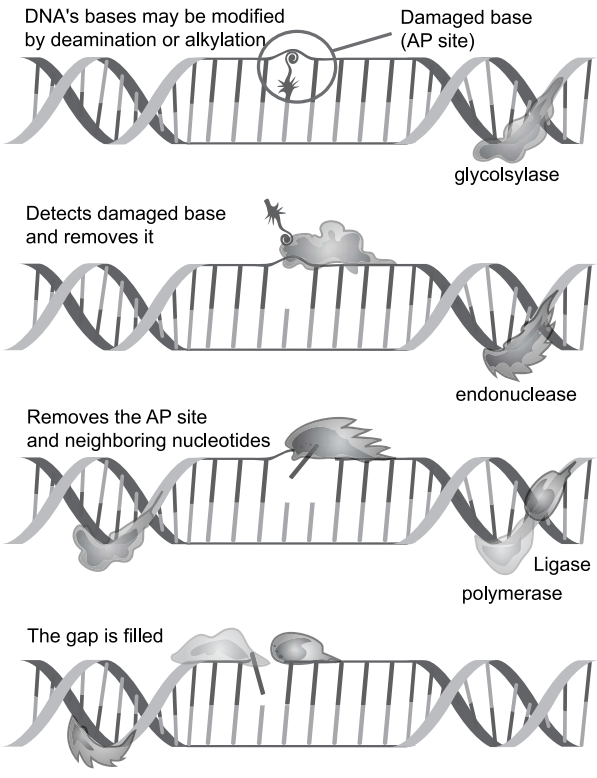

Figure 4. The main steps of base excision repair. Reproduced from Wikimedia (http://commons.wikimedia.org). Released into the public domain by its creator Mariana Ruiz Villarreal.

One of the most common DNA damages is the oxidation of guanine to 7,8-dihydro-8-oxoguanine (8-oxoG), mediated by reactive oxygen species, which can lead to mutations because of the ability of 8-oxoG to base pair with A (122-124). The cellular defenses against 8-oxoG damages are in many bacteria encoded by the *mutT*, *mutM* and *mutY* genes (125-127). The MutT protein removes 8-oxoGTP from the nucleotide pool by hydrolysis before incorporation into DNA. Mutants defective in MutM have increased rates of T-to-G transversion mutations (128). MutM is a glycosylase that excises 8-oxoG when paired with C, which initiates base excision repair allowing restoration of the GC base pair. Failure to do so before replication allows 8-oxoG to mispair with A, and this is a substrate for another glycosylase, MutY, that removes the adenine and allows repair by DNA repair syn-

thesis (Figure 8) (125). Other common DNA damages are the spontaneous deaminations of cytosine and 5-methylcytosine (5-meC), which result in uracil and thymine respectively, and generate C-to-T transition mutations if not repaired before replication (108, 129, 130). The mutagenic potential of deamination of cytosine is reduced by the action of uracil glycosylase, encoded by *ung*, which removes uracil in both single and double stranded DNA allowing restoration of the site by DNA repair synthesis (Figure 8) (131, 132). Mug is another glycosylase that has been shown to act on uracil with a preference for double stranded DNA and is expressed mainly in stationary phase (133-136). Mutants defective in *ung* show a 10-fold increase in transitions, but the *mug*$^-$ cells do not exhibit a mutator phenotype in growing *E. coli* and has a very modest effect in stationary phase (100).

Endonuclease III (*nth*) and endonuclease VIII (*nei*) are bifunctional DNA glycosylases recognizing oxidized pyrimidines that directly creates single nucleotide gaps through their endonuclease activity. Mutants defective in *nth* show a 2-8-fold increase in mutation rate, but no increase was detected in a strain defective in *nei* (100). The DNA glycosylases Tag and AlkA can both repair 3-methyl adenine and AlkA also recognize a range of oxidized bases (121).

Nucleotide excision repair

DNA with more extensive damage, including crosslinks, photoproducts and bases modified with large chemical groups, cannot be corrected by BER and is repaired by the nucleotide excision repair system (NER). This pathway is initiated by the UvrABC endonuclease that detects the distortion of the DNA helix caused by the damage (137, 138). The dimeric UvrA binds to the damaged site and recruits UvrB before dissociating from the DNA. This allows UvrC to bind, forming a dimer with UvrB, and cleaves the DNA on each side of the damaged nucleotide (137, 138). The resulting fragment, of about 12 nucleotides, are removed by DNA helicase II (UvrD) and the segment is resynthesized by DNA polymerase I and ligated by DNA ligase to complete the repair process (137, 138).

In *E. coli* and its close relatives, RNA polymerase interacts with the transcription-repair coupling factor, encoded by the *mfd* gene, which removes RNA polymerases stalled by damaged bases and recruits the NER system (139). This leads to a more rapid repair of the transcribed strand compared to the non-transcribed strand.

Mutational biases

For all the neutral nucleotide positions in a genome, probably including many synonymous sites and most sites outside ORFs, the composition of A, T, C and G bases are expected to become equal over time, both within and between DNA strands, unless there is a bias in the mutation rates between bases (140, 141). Base pairing rules will of course require that the genomic

number of G is equal to C and A equal to T, but if the mutation rate from GC to AT is not the same as from AT to GC the genomic composition will deviate from equal ratios. Indeed it has been observed that the GC content of bacteria varies between at least 17% and 75%, with even larger variation at the third codon position (data from CMR (142)). Potentially, this demonstrates how the neutral processes of mutation and genetic drift can shape the genome without the influence of natural selection. But how can we be confident that the differences in genome composition are not caused by natural selection? A number of theories of base composition evolution emphasizing selective forces have been proposed, including different metabolic cost for nucleotide bases, increased stability of GC base pairs and selection to reduce DNA damage from UV radiation and reactive oxygen species (143-151). These theories cannot, however, provide a universal explanation relevant to all bacterial species as the selective constraints suggested will not operate in all ecological niches and the selective value of the change in one DNA base pair will most likely not be large enough for selection to act upon it if the beneficial effect is solely related to GC content. Thus, even if there might be selective differences between having a high or low GC content in a certain environment, the GC content will most likely be caused by the neutral process of mutational biases and genetic drift and hence the change of a single base pair is not driven by selection.

In the same way that biases in GC to AT mutation rates can cause non-equal genomic base content, differences in mutation rates can give rise to compositional biases in any asymmetric genetic system. In bacteria, DNA replication by DNA polymerases proceeds in the 5' to 3' direction from an origin of replication where the leading strand is replicated in a continuous fashion, while the lagging strand is replicated discontinuously in Okazaki fragments resulting in an asymmetric process (152). When the mutation rates between DNA bases in the two strands are not equal this will give rise to compositional biases and this could cause the overrepresentation of G over C and T over A as well as a reduction in G+C in the leading strand compared to the lagging strand in many bacteria (141, 153, 154). These biases are so pronounced that they can be used to locate the origin of replication and the terminus in most bacterial genomes (155). Transcription is another asymmetric process where one strand of the DNA is used as the template for RNA polymerase and the other is single-stranded outside the transcription complex, which could result in differences in mutation rates between the two DNA strands (154). Regions of the genome are also transcribed at very different levels, which could give rise to intragenomic variation in mutation rates (156). Additional asymmetries could be introduced by selection on gene direction and location, which leads to an overrepresentation of more important genes on the leading strand and closer to the origin of replication (156). Clearly, asymmetries are present on many levels and the resulting compositional biases are intertwined, which makes it difficult to estimate

their relative importance (157). Even if changes in nucleotide composition occur without the influence of positive selection it can still influence what adaptive paths are possible in the future, as the genotypes will occupy distinct regions of the neutral network. In addition, if the number of advantageous mutations available is equal for various types of mutations, a higher mutation rate towards one type of mutation will increase the probability of those mutations being fixed in the population when the genome has not yet reached mutational equilibrium.

The importance of mutational biases in shaping bacterial chromosomes does not exclude that some compositional biases are caused by natural selection. One example is the accumulation of octamers, known as architecture imparting sequences (AIMS), which is skewed towards the terminus regions in many bacterial species, possibly allowing proteins involved in replication termination to locate the terminus efficiently (82). Another motif, known as chi sites, is involved the DNA repair by homologous recombination and is overrepresented in many bacterial genomes, although both the identity and length of the motif vary (158).

Recombination

Recombination serves important roles both in repair of DNA damage and genome dynamics, including horizontal gene transfer, amplification and deletion. Certain types of chromosome damage can only be repaired through RecA-dependent homologous recombination and can be divided into three classes: double-strand break repair, broken fork repair and gap filling repair (159). The term homologous recombination describes the process of pairing two molecules with homologous DNA strands into a heteroduplex that ensures its specificity from base pairing. Double-strand breaks in a chromosome are bound and processed by the RecBCD nuclease that unwinds the DNA, using the helicase activities of the RecB and RecD subunits, and degrades the dsDNA until it encounters a chi site, a specific octamer DNA sequence recognized by RecC (159). At chi the RecD subunit disengages from the DNA and the nuclease activity is altered so that a 3' ended ssDNA is made, upon which the RecBCD complex loads RecA onto the ssDNA to exclude binding of single-strand binding protein (SSB) (160-162). The recombinogenic ssDNA, with filaments of the strand exchange protein RecA, can now search for a homologous sequence that when found allows strand invasion and extension of the heteroduplex through branch migration, followed by repair by DNA resynthesis (159, 163). The process results in a branched structure that is resolved by RuvABC (159, 164).

Homologous recombination is also used for repair of replication forks that have collapsed due to nicks or endonuclease activity (165, 166). The dsDNA end is recognized by RecBCD and loaded with RecA as described above, which allows strand invasion and formation of a branched structure called

the “D-loop” through displacement of one strand. The heteroduplex is extended by branch-migration and results in cleavage of the branched structure, either a Holliday junction or a D-loop depending on the direction of branch migration. A replication fork is then restored by ligation and reloading of the DNA replication machinery. Single-stranded gaps in the DNA, caused by incomplete replication, are repaired by the RecFOR pathway that allows loading of RecA onto ssDNA to replace SSB (159). The subsequent steps involving homology search, strand invasion, branch migration and junction resolution occur as described above for the RecBCD pathway.

In the same way that mutational biases in DNA damage and repair can influence patterns of genome evolution it is possible that biases in recombination contribute to genome evolution. Biased gene conversion (BGC) has been studied mainly in eukaryotes and is caused by a GC-bias in the repair of heteroduplexes in gene conversion and recombinational repair, which leads to an increased fixation probability of GC alleles (167, 168). Thus, even though BGC is not a selective event, the increased probability of transfer of the GC allele, make it similar to a selective process in its dependence on N_e (169, 170). The contribution of BGC in bacterial evolution is not clear, but it could have a significant impact considering that bacteria do undergo recombination, there is a GC-mismatch repair bias and regions with low recombination rates are more AT-rich (27).

A less efficient RecA-independent pathway for homologous recombination has been described in *E. coli*. This pathway is not as dependent on the length of homology required and appears to involve Holliday junction intermediates (171). The biological significance of RecA-independent homologous recombination is not well understood, but the efficiency of the process is limited by the presence of exonucleases that degrades the recombinogenic substrates (171).

Illegitimate recombination takes place between DNA sequences with very short homology (4-13 bp) or no homology. The short homology independent illegitimate recombination (SHIIR) is mediated by DNA gyrase and controlled by DNA binding protein HU (172). Short homology-dependent illegitimate recombination (SHDIR) is induced by UV-light and is believed to be dependent on RecE, RecJ, and RecFOR. Two different models have been put forth suggesting that the process is mediated by replication fork slippage or double strand breaks and end-joining (173, 174).

Phages and integrative conjugative elements (ICEs) encode enzymes for site-specific recombination that allows integration at defined positions in bacterial chromosomes (175, 176). Phage integrases mediate recombination between short sequences of phage DNA, the phage attachment site *attP*, and the bacterial attachment site *attB* with each integrase recognizing specific sequences (176). There are two main classes of integrases, the tyrosine family and the serine family reflecting the identity of the catalytic residue used for strand cleavage. Once integrated into the chromosome the phage (or ICE)

are flanked by hybrid *att* sequences that are substrates for excisive recombination that can allow the spread of the element by phage transduction or plasmid conjugation systems. Transposons and IS elements, transposons without additional genes, use a transposase enzyme that allows the element to move to other locations in the genome either though a replicative transposition process resulting in a new copy or a cut-and-paste mechanism (177).

The SOS system

Extensive DNA damage caused by e.g. UV-radiation and chemical mutagens can lead to the accumulation of ssDNA that is rapidly coated by RecA for recombinational repair (178, 179). When RecA is assembled into nucleoprotein filaments it activates self-cleaving of LexA, which is a repressor bound to SOS-boxes of more than 40 genes including LexA itself (180, 181). Upon cleavage of LexA these genes, called the SOS regulon, are induced until the RecA filaments disappear (when repair is finished) and functional LexA can rebind (180, 181). The SOS regulon contains genes encoding translesion DNA polymerases, components of NER and UvrABC, which is involved in branch migration during recombinational repair (180, 181).

Mechanistic constraints on horizontal gene transfer

HGT is mediated by at least three mechanisms: transformation, conjugation and transduction (182, 183). Transformation is the uptake of naked DNA from the environment and subsequent integration into a replicon in the recipient cell (Figure 5) (81). A competent physiological state is needed for DNA uptake and some bacterial species are always transformable, whereas in others competence is induced only during special environmental conditions, often during stress or starvation (184-186). Some species, including *Haemophilus influenzae* and *Neisseria gonorrhoeae*, have evolved systems for efficient uptake of DNA with specific uptake signal sequences, suggesting that it is dedicated to another function than just uptake of DNA for nutrients (187, 188). Genetic competence is widespread in the bacterial domain with members of *Helicobacter*, *Haemophilus*, *Neisseria*, *Staphylococcus* and *Streptococcus* (189). DNA is present in high concentration in some environments, especially in biofilms, and can be stable for months to years in nature (190-193). The uptake of naked DNA means that theoretically there is no restriction on host range between the donor and recipient cell suggesting that genes can be spread between very distantly related species and even between the three domains of life (81). Size restrictions on the horizontally transferred genes are likely to be quite severe as very long DNA segments are broken down to smaller pieces in the environment (81). The transformation frequency is strongly dependent on integration efficiency with high rates of up to 1% when large homologies are present, but rates are magnitudes lower for RecA-independent recombination.

Conjugation requires physical contact between the donor and recipient cells and uses a complex machinery for highly efficient transfer of DNA encoded on the plasmid or conjugative transposon (Figure 5) (81, 183, 194, 195). Unrelated chromosomal DNA can also be transferred and the direct link between donor and recipient means that very large pieces of DNA can be transferred (194). The host range will depend on the potential of the transfer machinery to function between species and there are several examples of HGT of plasmids between distantly related species and even between domains (194, 196, 197). Conjugative plasmids have an important role in the spread of antibiotic resistance as well as in pathogenicity traits in *Shigella* and *Yersinia* (198-200). The mechanistic rates of conjugation vary dependent on geophysicochemical and biological factors, but it can be a highly efficient process, especially in structured environments supporting biofilm formation with rates, calculated as the ratio of transconjugants to donors, ranging from undetectable up to 10^{-1} (183, 194).

Transduction is mediated by bacteriophages where genes adjacent to the integrated phage can be transferred by imprecise excision in specialized transduction or transfer of any DNA by generalized transduction caused by erroneous packaging into the phage capsule (Figure 5) (201). Host range is restricted by the receptors recognized by the phage and the size limit will depend on the packing capacity of the phage capsule, generally less than 100 kb (183, 202). Phages can be stable in the environment for years and some estimates suggest that there might be up to 100 times as many phages as bacteria in the world making transduction-mediated HGT a powerful evolutionary force with approximately 10^{25} phage infections initiated per second for the last 3 billion years (203, 204). Phage related elements have been found in pathogenicity islands in *Staphylococcus aureus*, enterohaemorrhagic *E. coli* and *Vibrio cholerae* and near many ORFans, open reading frames without homologues in other species (205-210).

For successful integration and replication in the new host, the incoming DNA must overcome the host's defense systems. These include restriction endonucleases that recognize specific sequences, which are not present or chemically modified in the host, and degrade them into smaller pieces (211, 212). Consequently, there is selection against restriction sites in horizontally transferred DNA and some plasmids carry their own anti-restriction systems that can interfere with the restriction system of the host (213, 214). More complex adaptive defence/immune systems have also been found in many archaea and bacteria. These are called CRISPR/Cas systems and consist of loci with clustered regularly interspaced short palindromic repeats (CRISPRs) with spacers containing segments of acquired phage and plasmid sequences and adjacent Cas (CRISPR-associated) genes encoding proteins with a variety of functional domains, including endo- and exonucleases, helicases and RNA- and DNA-binding domains (215-218). The mechanistic details of the system are not yet elucidated, but the long RNA transcript from

the CRISPR is cleaved within the repeat sequence into smaller crRNAs using the Cas encoded genes (219). It has been hypothesized that the crRNAs then interacts with components of the Cas system to target the alien mRNA or DNA for degradation in a base-pairing dependent process (215). CRISPR loci seem to have been spread extensively between bacterial species by HGT (220).

The evolutionary impact of horizontal gene transfer in eukaryotes is less studied, but it appears that in some lineages of single-celled eukaryotes it could be relatively frequent and has been connected to a phagotrophic lifestyle (221, 222).

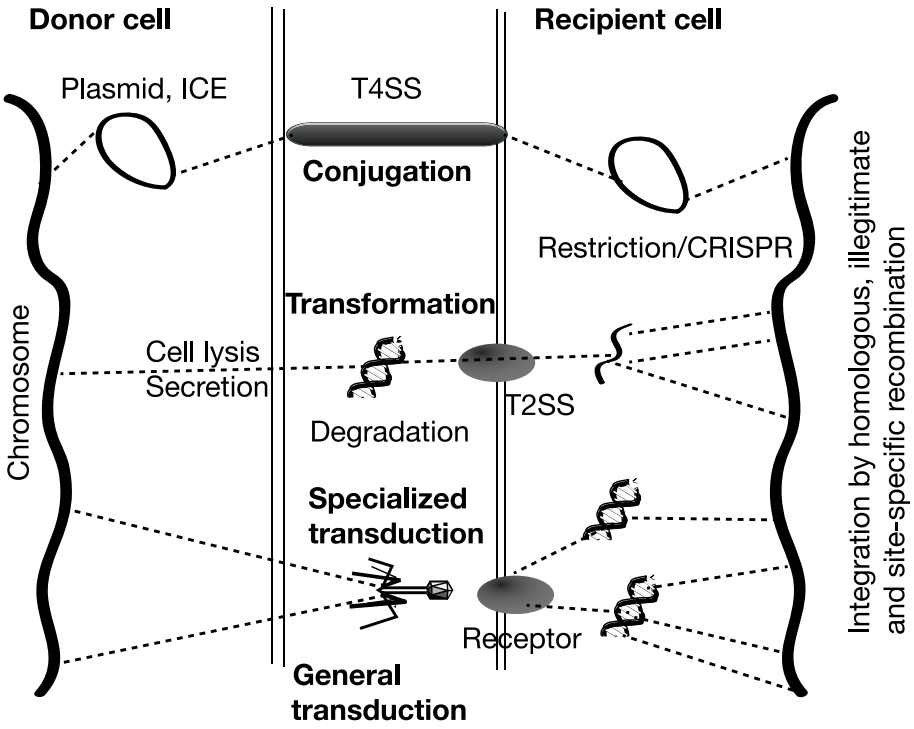

Figure 5. Mechanistic constraints on horizontal gene transfer. Adapted from Zaneveld *et al.*(223).

Gene duplication and amplification

The sizes of new duplications in bacterial chromosomes range from a few kb to several Mb and occurs with a rate of 3×10^{-2} to 6×10^{-5} in *S. typhimurium* depending on location in the genome (89). This means that a significant fraction of the cells in a population contains a duplication somewhere in the genome. Rates are highest for regions where long homologies are present,

suggesting that homologous recombination plays an important role in forming many duplication events (224). However, this high rate of formation is counteracted by an even higher rate of loss of 10^{-2} to 1.5×10^{-1}, mediated by the homology of the duplication itself (89). Homologous recombination between the duplicated regions also gives rise to amplifications, but at a lower rate than loss of the duplication. The high rates of formation and loss of amplifications allow rapid increases in copy number under strong selection pressures. The initial duplication event can also be formed by RecA-independent mechanisms of recombination where short homologies or no homology are found at junctions, although at a lower mechanistic rate. The high frequencies of duplication and loss make a population reach a steady state very fast and the high supply of duplications together with its reversibility makes this a kind of genetically regulated response to environmental changes (89, 225).

Deletions

Loss of DNA through deletion events is an important force in evolutionary dynamics that is responsible for the high coding density in bacterial genomes by removing non-selected genetic material. This is accomplished by deletional bias where deletion events outnumber insertion events by a factor 10 in sequenced bacterial genomes (27, 92, 226, 227). However, this estimated bias does not tell us what the mechanistic rates are. RecA-dependent homologous recombination has been shown to contribute to deletion formation only when large homology regions (>200 bp) were used and increase in importance with increasing homology length (228, 229). RecA-independent recombination dominates at shorter homologies and a number of models emphasizing the importance of replication for the process have been suggested including simple replication slippage, sister-chromosome exchange-associated slippage and single-strand annealing. The majority of the spontaneous deletion events has been showed to be dependent on translesion DNA polymerases in *S. typhimurium* (230). Large spontaneous deletions in *S. typhimurium* most often had very short or no homology at the endpoint, suggesting that illegitimate recombination events can contribute significantly to deletion formation (231). Deletion rates vary more than 100-fold depending on chromosomal location with rates of RecA-independent recombination in the order of 10^{-9} (232).

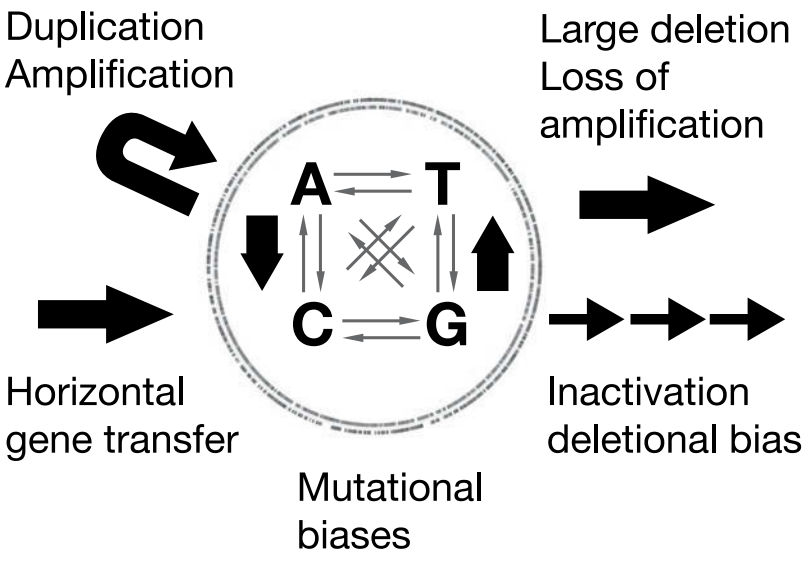

Figure 6. Dynamics of bacterial genome evolution.

Evolution of mutation rates and mutators

Mutations are necessary for successful adaptation to changing environmental conditions, but the rate of deleterious mutations is much higher ($2–8 \times 10^{-4}$ per genome per generation in *E. coli* (233, 234)) than the advantageous rate (2×10^{-9} per genome per generation (39)) and consequently increases in mutation rates are expected to be counter-selected in the long run (235, 236). Clearly, increasing the mutation rate is evolutionarily feasible through inactivation of any of the DNA repair systems in the cell, possibly without any immediate fitness cost before accumulation of mutations. Strains with high mutation rates, mutators, are often found in natural populations of many bacterial species at frequencies higher than expected from the mutation-selection equilibrium, suggesting a selective advantage for mutators in some environments (236). In cases where the supply of advantageous mutations is limiting, a mutator strain can have a selective advantage and the high prevalence of mutators can then be explained by hitchhiking of the mutator allele with new advantageous mutations (235). This scenario will be dependent on N_e as well as mutation rate increase and spatial and temporal environmental heterogeneity and the fitness gain must be larger than the cost of the accumulating deleterious mutations. This will often be the case in niche-invasion where there is a major shift in environmental conditions causing the ancestral genotype to have low fitness.

An increased mutation rate might also be advantageous in a new environment where the rate of adaptation is partly dependent on compensatory

mutations improving the function of the system that allowed the spread into the novel ecological niche. For long-term evolutionary survival the mutation rate must be reduced to avoid the impact of deleterious mutations and this can be done either by suppressor mutations or horizontal transfer of the advantageous allele into a non-mutator background. A majority of mutators found in natural populations are defective in MMR (105, 106), often causing a >100-fold increase in mutation rate, suggesting that either the cost for inactivating other repair systems are higher or do not elevate the mutation rate sufficiently. Alternatively, an increase in recombination rates in MMR mutators might be beneficial in increasing the genetic variability and might also increase the rate of intra-species recombination decoupling the advantageous mutation from the mutator allele (237). Defects in the MMR system could also increase the rate of inter-species HGT by increasing integration rates, but it is not clear if this is a rate-limiting step of the gene transfer process.

Patterns of genome dynamics

The first complete genome sequence of an organism became available in 1976 when the final part of the 3569 nt of RNA Bacteriophage MS 2 was sequenced (238). It took nearly 20 years until the genome of the first cellular life form, the bacterium *Haemophilus influenzae* (1.8 Mb), was completed in 1995 (239), but only 15 years later more than 1000 bacterial species and almost 100 archaea has been completely sequenced and the number eukaryotic genomes is increasing rapidly with hundreds of genome projects underway (NCBI Genome). This development was fueled by advances in sequencing technology that has drastically increased the rate of data collection, reduced the cost of sequencing and increased computer power that allows assembly of bacterial and archaeal genomes on a standard laptop computer. This vast amount of data would be useless without the bioinformatics revolution and the creation of public biological databases that has taken place simultaneously.

Whole-genome sequencing has shed new light on the great diversity of life, but also confirmed the unifying common ancestry and the division of life into three domains. Bacterial genome diversity is enormous with genome sizes varying between at least 0.14-13 Mb, encoding <200 to almost 10 000 genes, and a range in GC content of 17 to 75 % (Comprehensive Microbial Resource (142)). The evolutionary rules obtained from population genetics can be used to gain knowledge of the various processes that has resulted in this diversity after about 3.5 billion years of evolution, when the first life is believed to have originated on earth (240). Comparative analyses allow us to infer the function of newly discovered ORFs by comparing them to previously characterized proteins and genomic patterns of gene order, horizontal gene transfer and operon and chromosome structure provide us with impor-

tant information on evolutionary processes responsible for the order observed. Nevertheless we must realize that the information obtained has its limitations. It is often difficult to distinguish between patterns caused by natural selection and non-adaptive processes as population genetic theory commonly allows either explanation, as exemplified in the section above on mutational biases vs. selection. Experimental confirmation is required if we are to be certain that orthologues from different genomes actually have the same function as promiscuous functions might have been selected for in some cases whereas the gene might not be expressed in other cases and hence have no selective value. The biased selection of genomes for sequencing towards pathogens and those with possible biotechnological applications makes generalizations risky, but this situation will improve with the number of genomes sequenced and the increasing use of metagenomic analyses of the gene pool of entire ecological systems. The long-term evolutionary impact of patterns caused by human environmental changes, including antibiotic resistance, is not yet clear.

Remarkable patterns of conserved gene order and chromosome organization has been revealed when comparing bacteria such as *E. coli* and *S. typhimurium*, believed to have diverged about 100 million years ago, suggesting strong selectional constraints on many levels (Figure 7A) (241). On the other hand, significant variation among different strains of *E. coli* (Figure 7B) has been reported where genome sizes vary from 4.6 to 5.6 Mb and the fraction of the genome shared by all sequenced strains, the core genome, is only about 2 Mb (27).

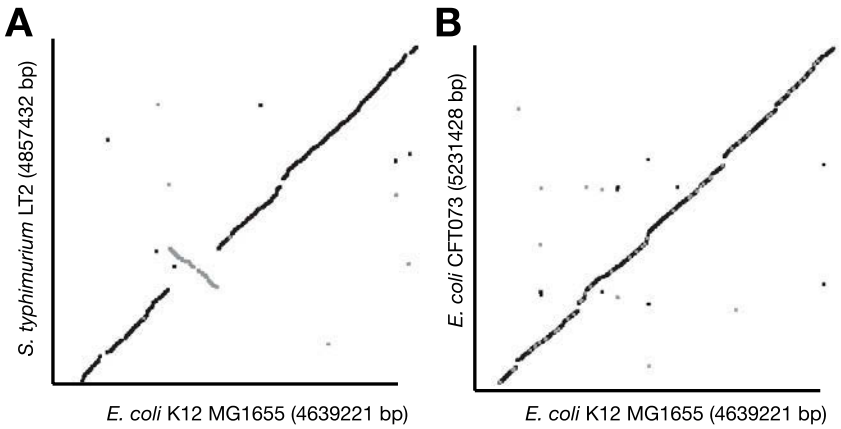

Figure 7. Genome alignments made with Promer (http://mummer.sourceforge.net). (**A**) *E. coli* K12 vs. *S. typhimurium* LT2 shows high conservation of gene order except an inversion around the terminus region. (**B**) *E. coli* K12 vs. *E. coli* CFT073 shows many insertion/deletion events resulting in a genome size difference of 0.6 Mb.

The pan-genome, containing all the genes encoded by all *E. coli* strains has so far been estimated to be more than 18,000 genes, about four times of the typical individual *E. coli* genome (27, 242). A picture has emerged where the genome is composed of a core genome encoding functions that are always selected and an accessory genome containing genes needed under some environmental conditions and many genes with unknown function. A large part of the *E. coli* pan-genome is made up of recently acquired insertion sequence-like and prophage-like genes (>7500) making up about half of the accessory genome (242). The contribution of the mobile genome to bacterial fitness remains to be determined, but it is reasonable to suspect that a large fraction is nearly neutral and will be lost by deletional bias not leaving a phylogenetic trace over evolutionary time-scales. A significant part of the *E. coli* genomes is made up of horizontally transferred genomic islands (10-200 kb) and islets (<10 kb) that are discrete genetic segments that differ between closely related strains, some of which are still mobile and capable of transfer to other strains (243, 244). The genomic islands commonly encode environment-specific genes, including antibiotic resistance determinants, pathogenicity traits and new biosynthetic pathways and are therefore likely to be adaptive, at least over short evolutionary periods. The horizontal gene flow in *E. coli* is mainly concentrated to integration in a few regions where it will not disrupt the function of core genes and this allows for spread by homologous recombination using the flanking genes (27). The robust nature of the core genome allows accurate phylogenetic relationships to be determined even in the presence of high rates of HGT and recombination (27).

The information obtained from comparative genomics is limited because it only studies the genomes of the evolutionary winners. They contain the genetic variation that has been fixed by natural selection or non-adaptive processes and thus excludes mutations with larger deleterious effects that are needed for understanding the selective constraints operating on various levels of the organism. However, population genetics analyses of genome data allow estimation of many important parameters that are too small to measure experimentally, including estimates of fitness effects of mutations and HGTs (30).

Experimental studies of evolution

Good ideas based on biological intuition are fundamental to the progress of evolutionary biology. To explore the value of an idea it should be evaluated using evolutionary theory to see under what scenarios the idea is possible in the framework of population genetics. If the idea is theoretically possible, using commonly used simplifying assumptions, it remains to be seen if it is relevant in modern biological systems. Depending on what predictions can be made from the idea, the next thing to do is to look for signs of it in nature,

either by observing organisms or through analysis of molecular data that can falsify the hypothesis or support it. Although we might find strong support for the idea in nature, there are commonly alternative explanations that are theoretically possible, even though they may be less appealing. So what we finally want are testable hypotheses. The tests often requires moving on to laboratory experiments where parameters can be controlled and specific changes can be introduced one at the time. In evolutionary biology this can be accomplished by experimental evolution where the environment and population genetic parameters, such as effective population size and mutation rate, can be controlled and quantified and evolution can be studied directly in real-time. Bacteria are particularly useful for experimental evolution studies, as they require very little space, grow fast, so that evolution can be studied under many generations in parallel with controls, and unevolved ancestors can be stored in a non-evolving state in a freezer (245). The high demands on controlling the environment will typically lead to the use of very defined stable conditions, which will make the evolutionary scenario less realistic, as natural populations often live in diverse environments together with many other organisms. Natural selection can be studied by experimental evolution where a population is allowed to adapt for many generations to a new environment, a specific nutrient source, the presence of sub-lethal levels of antibiotics or to a different temperature ((107, 245, 246), **paper II**). No artificial selection in terms of choosing individual phenotypes is used so that bacteria evolve in any way that provides improved fitness. Another type of experimental evolution called mutation accumulation (MA) instead use very small bottlenecks to reduce the effective population size and increase genetic drift, so that mutations can accumulate with little influence of selection, which allows tests of evolutionary scenarios that are difficult to study in nature (247). At the end of the experiment the genomes can be sequenced and the mutations that have taken place during the experiment can be found and further analyzed (**paper I**).

Evolution experiments have been used to measure the fitness effects of new mutations directly for both advantageous and deleterious mutations. The main problem with experimental estimates of fitness effects is the limited sensitivity of the assays used, which often only detects differences in fitness larger than 10^{-2}, whereas even fitness differences smaller than 10^{-5} are important in population genetic models. In the case of MA experiments the effects of many mutations are measured and allows estimation of mean and variance of fitness effects, but it is impossible to determine the direct effect of one mutation, which limits a deeper mechanistic understanding (30, 247). By measuring the mutational effects of single engineered mutations an assumption-free estimation can be made, but the limited assay sensitivity and labor intensity of fitness measurements have so far largely limited this approach to viral systems (248). The purpose of measuring the fitness effects of individual mutations is to connect the molecular details with the higher

biological level of fitness and thus provide us with a genotype-phenotype map (249, 250). This could allow use of *in vitro* data to construct mathematical models and allow realistic simulations to help us predict important evolutionary processes. The domestication of animals and plants represents a kind of experimental evolution, where selective breeding focused on specific traits can allow an understanding of the molecular basis for the trait. The small population sizes with high inbreeding allow fixation of highly deleterious mutations that can provide clues to human diseases (251). Increasingly, experimental evolution is used in conjunction with genome resequencing also in higher eukaryotic systems (252).

Salmonella as a model in experimental evolutionary biology

All studies in this work use *Salmonella enterica* serovar Typhimurium, referred to as *S. typhimurium* in the text, as a model organism. *S. typhimurium* is rod-shaped gram-negative aerobic enterobacteria that is a common cause of gastroenteritis in humans (253). All strains used here is derivatives of the LT2 strain, which is avirulent due to a defect in *rpoS* (254), and its complete 4.95 Mb genome was published in 2001 (255). *S. typhimurium* is well characterized biochemically and genetically and has long been used in genetics, microbiology and experimental evolutionary biology due to its ease of culturing, fast growth rate and the wide range of genetic tools. These tools allow genetic manipulations using phages, transposons, transformation of plasmids and precise chromosomal engineering using the lambda Red system for homologous recombination. The close phylogenetic relationship with the widely used *E. coli* model also allows use of gene function assignments and protein structures obtained from *E. coli* to be inferred to be similar in *S. typhimurium*.

Present investigations

Paper I

Evolution of base composition and mutational biases

The evolutionary processes that shape the base composition of bacterial genomes are largely unknown. If there is a significant fraction of neutral sites in a genome, the base content of this fraction is mainly determined by biases in mutation and recombination. Depending on how large the neutral part is, this can influence whole-genome biases, including genomic GC content and biases between the leading and lagging strand of replication and the transcribed and non-transcribed strand of genes (153). Although there are probably also selective forces influencing genome composition, we can never escape that there is some influence of mutational biases unless we claim that all sites are non-neutral (145).

Using experimental evolution to study mutational biases

Two common types of spontaneous DNA damage that could lead to mutational biases are deamination of C to U and 5me-C to T and oxidation of G to 8-oxo-G (125, 129, 256). Repair systems exist for reducing the mutagenic impact of these damages and the rates and biases will be dependent on the function of these systems (Figure 8).

In **paper I** we use an experimental evolution approach to investigate the mutational patterns, when removing several repair systems for deamination and oxidation damage, using genome sequencing to find the mutations. Five different strains of *S. typhimurium*, with or without repair defects, were passed in 12 replicates through 200 single cell bottlenecks, approximately 5000 generations, on rich laboratory media. The bottlenecks leads to a small effective population size, which means that mutations could accumulate at random largely independent on their effect on fitness, and allow the study of evolution under conditions of high genetic drift.

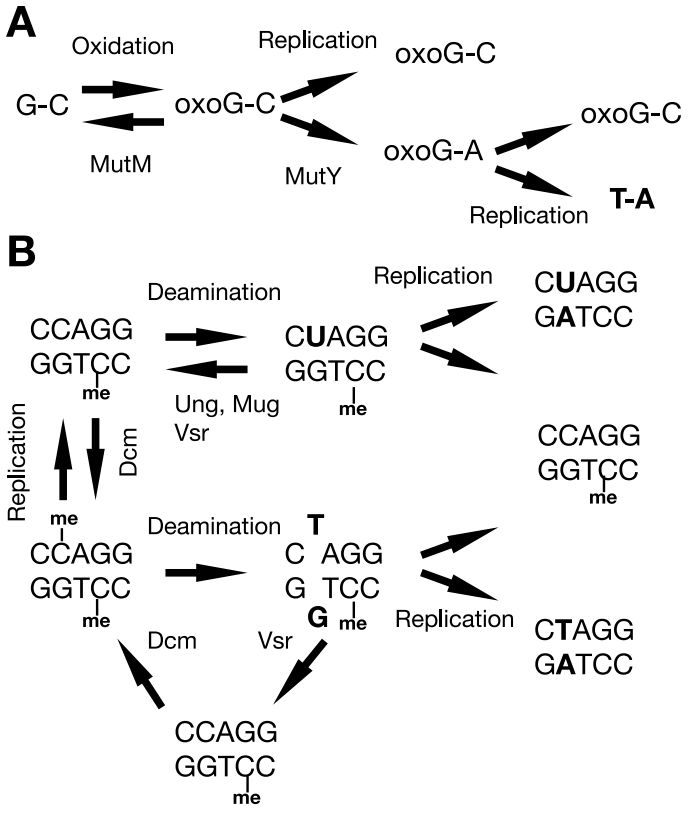

Figure 8. (**A**) Mutagenic damage oxidation of guanine and repair mechanisms. (**B**) Mutagenic damage of deamination of cytosine and 5-methyl-cytosine and repair mechanisms.

Changes in genomic base composition after MA

We sequenced the genomes of two strains defective in five repair genes: *ung*, *mug* and *vsr* that are responsible for repairing deamination damage and *mutM* and *mutM*, which repair oxidation damage, as well as a wild-type strain with full reparational capabilities. As deamination of C and 5meC results in transitions and oxidation of G to transversions the impact of each type of damage could be assessed in the same genome. We found 856 G-to-T transversions and 65 C-to-T transitions in the whole genome data for the mutants and only 22 other base substitutions. This high rate of genomic change (0.094 mutations per genome per generation) means that a 1% reduction in GC content can take place in about 1400 years in a natural setting, assuming loss of repair functions and relaxed selection. The upper estimate of average fitness loss was 0.00145 per base substitution, measured as relative exponential growth rate.

Influence of replication and transcription

No biases between the leading and lagging strand of replication were found for either deamination or oxidation, but a significant bias in transcription was detected for deaminations, where damages were more likely in the non-transcribed strand. This suggests that transcriptional biases will dominate over replicative biases for deamination damage. The cause for the bias could be that the non-transcribed strand is in a single strand state for a longer time and in vitro measurements of deamination rates show a very large increase in ssDNA compared to dsDNA (257). Both types of mutations were found to be overrepresented in a set of highly transcribed genes, suggesting that mutation rates are dependent on transcription rates, at least when caused by deamination and oxidation damages. Evolutionary rates are often inversely related to expression level in nature, which suggest that these genes are under very high purifying selection (59, 258). We did not find any differences in mutation rate depending on chromosomal location in terms of distance from origin of replication or terminus.

Mutational biases as a cause of extreme AT-content

This experiment does not attempt to prove that all base compositional biases are solely dependent on mutational biases. Rather, it explores what patterns we would expect to find if deamination and oxidation were responsible. An evolutionary scenario that is very similar to this experiment has been proposed for AT-rich endosymbionts (259-261). They are generally believed to have evolved by reductive evolution, losing genetic material, including repair systems, over time due to deletional bias and lack of HGT (261, 262). The influence of selection is also reduced as they go through severe population bottlenecks that increase genetic drift and are supplied with nutrient from the host, allowing loss of previously adaptive genes.

Paper II

Distribution of mutational effects

The distribution of fitness effects (DFE) of new mutations is fundamental in evolutionary dynamics, but has rarely been directly measured in experiments (29, 30). Mutation accumulation experiments, as performed in **paper I**, can be used to measure some properties of the distribution at the genomic level, such as mean cost of a mutation and variance, but parameter estimates usually have large confidence intervals and the analysis of cumulative effects limits mechanistic insight on the selective constraints operating at various levels. The DFE can also be inferred from DNA sequence data using population genetic theory, but this requires assumptions regarding what positions

are neutral and is limited to genetic variation present in natural populations, which excludes information on the properties of deleterious mutations and the mechanistic basis of major selective constraints. This weakness is complemented by the main strength of sequence analysis data to allow studies of mutations with small effects on fitness in the nearly neutral range that cannot be detected in laboratory experiments due to limited sensitivity (30).

Fitness assays

The most direct way to obtain a DFE is to measure the fitness effects of a large number of single defined mutations. There are several studies using viruses that have applied this experimental setup to determine the fraction of deleterious, neutral and advantageous random mutations. The classification of a mutation as potentially neutral in the laboratory is dependent on the sensitivity of the fitness assay used and selection coefficient smaller than 0.01 is usually not detectable, whereas the population genetics concept of neutrality is related to N_e, so that s-values of 10^{-5} are often large enough to be non-neutral for bacterial species. Thus, we must be able to detect small changes in fitness to obtain a good estimate of the DFE.

We developed a highly sensitive competition assay to measure fitness to obtain high quality fitness data in **papers II**, **III** and **IV**. *S. typhimurium* was first adapted to the experimental environment (M9 glucose at 37°C) to reduce the impact of secondary beneficial mutations during competition experiments. This was done by experimental evolution using a serial transfer procedure where the culture was diluted 1000-fold in new media each day for 100 cycles (1000 generations) resulting in an increase in fitness of about s=0.3. Next, two markers encoding two variants of the green fluorescent protein (CFP and YFP) were introduced into a neutral location in the genome. This allow the cells to be counted by flow cytometry, where about 10,000 cells per second can be counted, using the fluorescent markers to measure the fraction of YFP to CFP cells with very high statistical certainty.

Fitness is estimated by mixing a mutant population, carrying either the *cfp* or *yfp* allele, with an isogenic wild type control with the other marker. The mixed population is maintained by serial dilution and the change in the fraction of mutant to wild type at each cycle can be used to obtain a value for the selection coefficient. This assay allows differences in fitness as small as 0.003 to be detected.

We also measure exponential growth rates by detecting changes in optical density over time. This is a less labor-intensive technique, but we can only detect differences of about 3%. The results from the two assays are not directly comparable as the competition assay measures a composite fitness over the entire growth cycle, including lag phase, exponential phase and stationary phase whereas the optical density measurements only assay the early exponential phase.

Ribosomal protein genes as models of fitness effects of mutation and gene transfer

In **papers II** and **III** we use ribosomal protein as models to study the fitness effects of mutation and gene transfers. These were chosen because they are either in its own operon (S20) or at the end of operons (L1 and L17), so mutations and gene transfers will not affect expression of downstream genes. L17 is believed to be essential and S20 and L1 are non-essential, but deletion mutants have a 70% reduction in fitness, so fitness effects can be detected over a large range and complete loss-of-function mutants can be found for S20 and L1. These ribosomal proteins are strongly conserved and very highly expressed with optimized codon usage, implying strong selective constraints. This means that fitness effects of mutations are likely to be high compared to an average gene allowing detection using the fitness assays described above. Their fundamental importance in translation ensures that fitness is directly correlated to growth rates in both the quantity and quality of available ribosomes. The ribosome is also the target of several classes of antibiotics and resistance is commonly caused by mutations in ribosomal genes (263).

Fitness costs of synonymous and non-synonymous substitutions

We constructed 126 mutants with random single base pair mutations in *rpsT*, encoding ribosomal protein S20 (70 mutants), and *rplA*, encoding ribosomal protein L1 (56 mutants). Mutagenesis was performed by error-prone PCR and mutagenic primers and introduced into the native *S. typhimurium* genes together with an antibiotic resistance marker using the lambda red system. The mutations comprised 88 non-synonymous and 38 synonymous substitutions. Exponential growth rates in LB were significantly reduced for the majority of the substitutions with relative average growth rates of 0.92 and 0.94 for synonymous and non-synonymous substitutions, respectively. Smaller reductions in growth rates were also found in the poorer M9 glucose medium where average growth rates were reduced to 0.95 for both synonymous and non-synonymous substitutions. None of the mutants assayed showed a complete loss-of-function with growth rates always more than two times that of the deletion mutants.

Competitions were performed and 120 of the 126 mutations were found to be deleterious with s between -0.08 and -0.003. None of the mutations were advantageous, so the remaining six were classified as potentially neutral, being below the detection limit of the assay, but we would expect the truly neutral fraction to be even smaller. The average s for the synonymous substitutions was -0.0096, which is similar to the average cost for the non-synonymous substitutions (-0.0131).

Fitness constraints on base pair substitutions

The similar fitness costs of synonymous and non-synonymous substitutions are surprising as changes in amino acids are usually assumed to have much larger effects than changes in the mRNA sequence. Although selection on codon usage is well-known in highly expressed genes in enteric bacteria, the suggested magnitude of the selective constraints is substantially lower than found here (264). A few outliers in the distribution with very large effects show that although most amino acid changes cause small changes in fitness, sometimes protein structure is greatly disturbed by single changes. This is consistent with a threshold-model of protein stability were single amino acid changes rarely cause major structural changes, but once this threshold is exhausted protein function declines rapidly with further mutations (75, 76). The small costs of the amino acid changes also explain why we did not find a significant correlation between predicted protein stability, conservation or rRNA contacts.

The similar fitness costs of synonymous and non-synonymous substitutions strongly suggest that the main selective constraints are related to deleterious effects on the mRNA level. However, there are several possible mechanistic explanations. The codon usage is highly optimized for ribosomal protein genes and it is possible that introduction of more rarely used codons could results in slower translation of the mRNA producing less protein. This seems unlikely for a number of reasons. First, we did not find a correlation between fitness and codon usage frequencies either using the codons in ribosomal protein genes or all ORFs. Second, we found similar fitness costs in **paper III** when replacing the same genes with orthologues from other species with large changes in codon usage, suggesting a functional conservation between species on the mRNA level. Another possibility is that changes in synonymous codons increase the mistranslation-induced misfolding. If this is the case, we would expect larger costs for the non-synonymous mutations than found here.

Changes in mRNA structure near the ribosomal binding site (RBS) can influence expression of the protein (63). However, the mutations in this study are spread throughout the genes and the changes could not all affect the RBS. We found a significant correlation between predicted changes in mRNA free energy and fitness costs, suggesting that mRNA structure is a major selective constraint for these genes.

Distribution of fitness effects

We fitted the experimental data to commonly applied univariate distributions using a maximum-likelihood method. The distribution of selection coefficients for synonymous substitutions was best fitted by gamma, beta and weibull distributions and the non-synonymous by gamma or log-normal. The

exponential growth rates were generally best fitted by gamma distributions. As both the parameter estimates for the weibull and log-normal distributions are sensitive to outliers the gamma distributions are more useful for comparing the DFEs. The gamma parameters were estimated to: shape 1.84 (+/-0.39) and rate 192.9 (+/-46.9) for synonymous and shape 1.91 (+/-0.27) and rate 145.8 (+/-23.2) for the non-synonymous substitutions selection coefficients (Figure 9). There was a significant difference between synonymous and non-synonymous substitutions for the competition data (P=0.048), but not for the exponential growth rate data. None of the models passed an Anderson-Darling goodness-of-fit test suggesting that a simple gamma function cannot fully describe the experimental data.

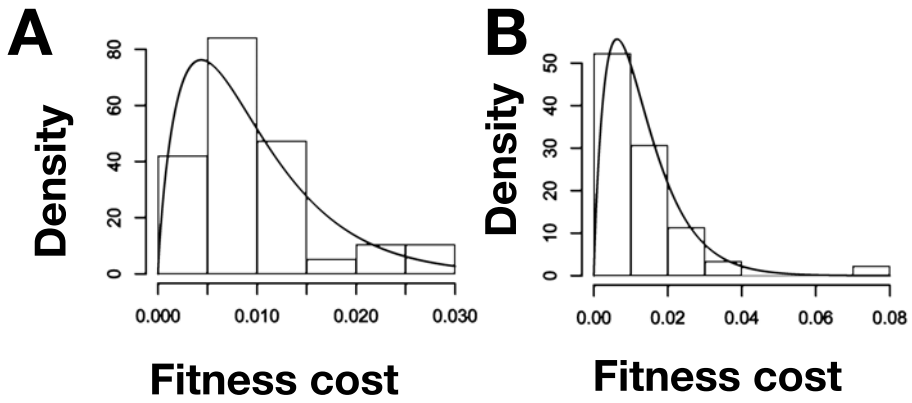

Figure 9. Distribution of fitness effects compared to a maximum likelihood estimation of a gamma distribution for (**A**) synonymous and (**B**) non-synonymous substitutions (**paper II**).

Differences from viral systems

Similar experiments have previously been performed in several viruses. Although these studies focus on whole-genome DFEs, the numbers of genes in the viruses are small and the data can also be analyzed at the level of single genes (248). The most striking difference between the viral systems and our data is that up to 40% of the mutations were lethal in viruses, whereas no complete loss-of-function mutations were found here. This might be due to the structural function of the ribosomal proteins interacting with many rRNA contacts and the absence of an active site that can be destroyed by mutations. Alternatively, the viruses are non-robust because of their streamlined genome and have an absolute requirement for packaging and transmission for viability. Mutations classified as neutral are also more common in the viral systems, but this is probably because the sensitivities of the assays used are lower than the assay used here, which results in a large proportion of the deleterious mutants with small effects that cannot be distinguished from the

wild type (248). The similar effects of synonymous and non-synonymous substitutions were not found in the virus studies where synonymous substitutions generally had smaller effects, although some did have significant costs.

Paper III

Fitness effects of inter-species gene replacements

We replaced ribosomal protein genes encoding S20, L1 and L17 in the *S. typhimurium* chromosome with orthologues from other microbial species, both closely and distantly related, using the lambda red system. Only the ORFs were changed, keeping the native promoter and terminator sequences of the native genes. The probability of this type of replacement would in nature be strongly dependent on the degree of homology shared and thus mainly limited to closely related species. Using this experimental setup we make sure that the transferred genes are selectively stabilized and their function can be directly compared to the native gene, which allow us to study on what levels the general constraints on HGT are operating.

We measured exponential growth rates in four different media with doubling times of 20-120 minutes for the wild type. The majority of the orthologues did not significantly reduce the growth rate despite amino acid identity as low as 50%, suggesting very high functional conservation in these proteins. There was a strong correlation between phylogenetic distance and fitness with the larger relative cost most often in the richest medium. Because the growth rates of the complete loss-of-function mutants are known, we were certain that the transferred genes were partly functional ensuring selective stabilization. We performed competition experiments to obtain more sensitive fitness measurements and all orthologous replacements of S20 and L1 as well as a majority of L17 transfers were found to be deleterious. For the transfers from the most closely related bacteria, the fitness effects were typically less than s = -0.01, but these are great enough to confer strong counter-selection in an evolutionary perspective and could explain the scarcity of horizontal gene transfers among ribosomal protein genes.

Selective constraints on HGT compared to random substitutions

The finding that replacement of these ribosomal genes most often confers fitness costs smaller than s = -0.01 is puzzling. There are extremely large differences in nucleotide (55%) and amino acid sequence (50%) and changes in codon adaptation index (0.21) and GC content (14%) that do not severely disturb function. Many of the mutants with a single synonymous substitution in *rpsT* and *rplA* have larger costs than an orthologous transfer, for example, transfers from *Proteus mirabilis* with 48 nt substitutions in *rpsT* with s = -

0.005 and 145 nt substitutions in *rplA* with s = -0.008. This suggests that functional constraints are strongly conserved between species, even if they diverged several hundreds million years ago, and that this functional conservation is also operating at the mRNA level, independent of codon usage and GC content. The participation of the ribosomal proteins in a large co-evolved complex does not seem to be a major fitness constraint, as suggested by the complexity hypothesis (265).

Gene amplification rescues transient HGTs

Genes introduced by HGT will, just as any mutation, most often be neutral or deleterious to the recipient cell and the non-selected genes are inactivated by random mutation and lost by deletional bias over time (84, 266). Rarely, the HGT may confer a selective advantage in a specific environment and even though it may be deleterious on arrival it could be selectively optimized to its new host by secondary compensatory mutations, which could allow exploration of previously inaccessible regions of the adaptive landscape. A phylogenetically distant gene is likely to require several mutations to improve function, but if the mutant cannot reach a large enough population size to allow the exploration of a sufficient number of mutations, the gene will be lost before any secondary mutations are possible. Gene amplifications, however, occur with a much higher frequency, which means that for most neutral transfers duplication will occur before the HGT is lost and if advantageous the HGT can be further amplified by homologous recombination.

The orthologous replacement mutants with the largest effect were subjected to experimental evolution by 1000-fold dilution of 8 independent lineages each day to allow fixation of beneficial compensatory mutations. Mutants with increased fitness were found after 40 to 250 generations and the copy number of the transferred gene was measured to see if the increase in fitness was caused by duplication/gene amplification. For *rpsT* from *Haemophilus influenzae* (HI) and *rplA* from *Saccharomyces cereviseae* (SC) an increase in gene copy number (2-3 fold) was found. The two *rplA* SC mutants both had a duplication of 44 kb between ribosomal RNA operons where long homologies promote RecA-dependent recombination. The amplified region surrounding the *rpsT* HI gene ranged from 2 to 200 kb and three of them did not have any homology at the junction, suggesting an illegitimate recombination mechanism, whereas one had an imperfect 88 bp repeat suggesting that either RecA-dependent or RecA-independent homologous recombination had occurred. To further examine the compensation by increased gene dosage, the alien genes were introduced on a plasmid under control of an inducible promoter. For most of the mutants with large fitness effects the cost of the transfer could be ameliorated by increased expression, suggesting that sub-optimal expression is a general fitness constraint on horizontally transferred genes. Compensation by gene amplification may

also explain the overrepresentation of duplicated genes among those recognized as introduced through HGT and suggests that this process could be a significant source of new genes via duplication and divergence (Figure 10).

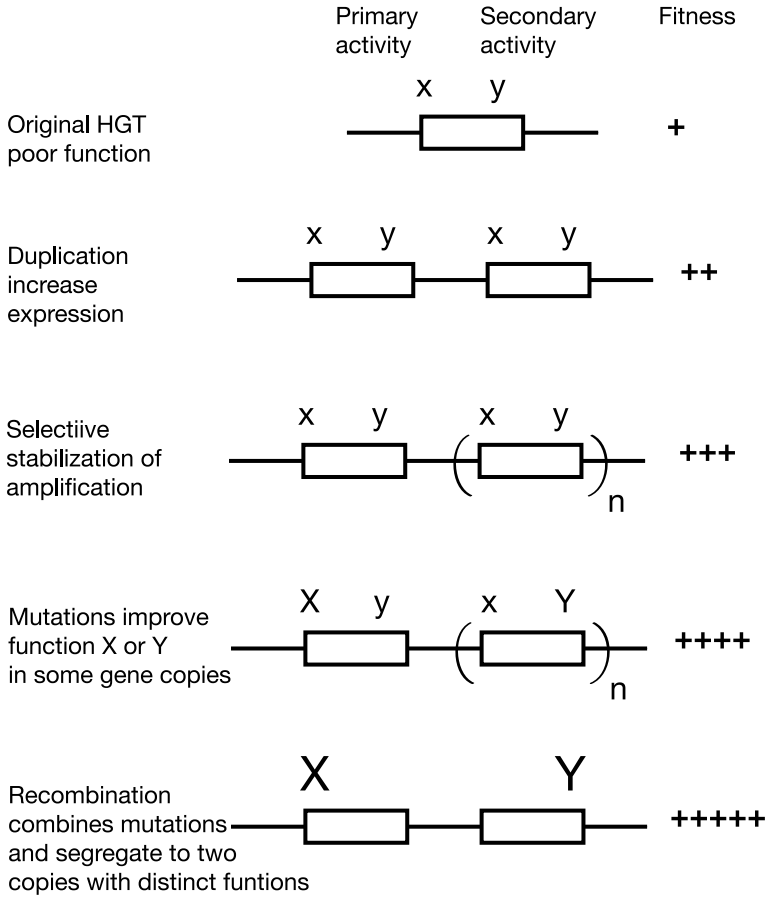

Figure 10. Duplication and divergence of a horizontally transferred gene allows the evolution of new genes.

Paper IV

Fitness effects of HGT

There are large differences in the proportion of HGT genes both between and within bacterial species that are generally believed to be strongly dependent on lifestyle. It is largely unknown how the mechanistic constraints limit the rate of HGT events in natural populations, but we expect the fitness effects to be extremely important as the main factor limiting the rate of

HGT. The large majority of HGT events are expected to be deleterious to the recipient cell, but it is not clear on what levels the main selective constraints are operating. Rarely, a new HGT can be advantageous, but it will most likely also be associated with costs related to the non-optimized alien genes that need to be fine-tuned to the new cell to further increase fitness. The rate of HGT is also influenced by the effective population size, as the fraction of effectively neutral HGTs will increase in small populations, including many pathogenic species that go through frequent bottlenecks when transferred between hosts.

We inserted 22 random fragments from *Bacteroides fragilis* and 13 from *Proteus mirabilis* into a neutral site in the *S. typhimurium* chromosome to investigate the fitness costs of HGTs. The sizes of the HGTs ranged from 0.5 to 4.2 kb in addition to an arabinose inducible promoter, transcriptional terminator and selective marker. Exponential growth rates were not significantly reduced for the large majority of the mutants in LB or M9 glucose with or without arabinose. Competition experiments showed that 28 out of 35 inserts were potentially neutral with selection coefficients smaller than 0.003, five were deleterious and one advantageous transfer was found. When transcription was induced by arabinose the fraction of deleterious HGTs was increased to 19 out of 35 and the rest of the inserts were apparently neutral.

Selective constraints on HGT

The costs of the HGTs were not correlated with the size of the inserts, but there was a strong connection between finding a significant fitness costs and the presence of complete ORFs. This suggests that there is a minor cost of carrying extra DNA, as expected, and that the high level of transcription under arabinose induction generally causes small costs when no ribosomal binding sites are present. Thus, the selective constraints are mainly tied to the cost of translation or the toxic effects of the alien gene products. Successful HGTs events are expected to contain less highly expressed genes that can be adapted by secondary mutations and are not wastefully expressed under conditions were the gene product do not confer a selective advantage. The HGTs with the largest costs were transferred from *B. fragilis,* which might indicate that the selective barriers are stronger between more distantly related species. The deleterious HGTs contain genes from a variety of functional categories, including alternative sigma factors, single enzymes and membrane and flagellar proteins.

The large proportion of apparently neutral HGTs suggests that many transfer events will have s-values below 10^{-3}. Thus, we would expect that transient HGTs are present at significant levels in natural populations if the mechanistic rates are sufficiently high. Most neutral transfers will not reach fixation in global populations and are expected to be lost over time by dele-

tional bias unless highly selected in sub-populations in isolated ecological niches.

We found one advantageous HGT (s=0.005), which is surprising given that beneficial mutations are generally believed to be extremely rare. This insert contains two genes from *P. mirabilis* encoding a transcriptional regulator and a D-alanine aminotransferase in the opposite direction. Arabinose induces expression of the D-aminotransferase and this reduce fitness to wild type levels, suggesting that the fitness increase is dependent on the expression of the transcriptional regulator or that the cost of increased expression of the D-alanine transferase offsets the fitness advantage.

Future perspectives and further discussion

Experimental evolution and genome sequencing

One of the fundamental problems in evolutionary biology is to separate the effects of natural selection from neutral processes. Experimental evolution using bacteria is especially well suited to study the problem as the experimental design can limit the influence of either selection or genetic drift, depending on the problem studied. The possibility to reconstruct mutations of interest can provide solid evidence in proving the molecular basis of the parameter studied. Genome sequencing will revolutionize the field of experimental evolutionary biology, as the cost is plummeting and improved bioinformatics software allows easy assembly and rapid comparative analysis so that it can become a standard laboratory technique (267).

Mutation accumulation (MA) experiments hold great promise to provide us with a more complete framework of mutational processes. This can be done, as in **paper I**, using mutants deficient in DNA repair to study mutational patterns and similar studies should be made to look at other types of DNA damage so that the probable mutational processes causing compositional biases can be further analyzed. These experiments are easier to conduct than using non-mutators, as the mutation rate is higher, which allows more mutations to accumulate, but this limitation will largely disappear with decreasing sequencing costs so that additional lineages can be sequenced instead of conducting extremely long experiments. The major types of spontaneous DNA damage are likely to vary extensively between different environmental conditions and could provide clues to differences in genome composition when studied using MA experiments.

The MA experiments must be complemented with a wide range of experimental studies of adaptation where the molecular basis can now be easily determined, which allows genotype-phenotype mapping. Hopefully, these studies will increasingly make use of more complex experimental environments. This would allow the evolution of both niche specialists and generalists using spatial heterogeneity and explore complex communities with several species present. Experimental evolution will also be able to address questions on the generality of adaptive processes, such as the number of adaptive paths that are available to a phenotypic goal, the importance of parallel evolution, the impact of clonal interference and find out if there

really are any general rules or if each subsystem or species will behave differently depending on its functional constraints and genetic architecture.

Mutational biases and selective patterns

The debate on the impact of selection and neutral processes on genomic base composition is still very much alive. There are indeed some general problems with the extremist versions of both views and it is possible that the main evolutionary process varies between different species depending on the number of neutral sites and effective population size, so that no universal conclusion can be reached. The main strength of the mutation bias theory is that it fits well into the framework of the neutral theory of evolution and can provide a general explanation for the differences in composition. A mutational explanation is supported by the observation that the sites that are under weakest selective constraints, synonymous positions, show the largest variation. In addition, the most constrained second non-degenerate codon position exhibits the smallest bias. However, the observation that there seem to be a general mutational AT bias in new mutations, even in GC rich species, suggests that there is some process maintaining GC content in these species. This could possibly be caused by biased gene conversion, mimicking a selective process, which would point to a major impact of recombination in bacterial genomes. Alternatively, there could be a selection pressure for maintaining high GC operating at the genomic level, which could be biased towards synonymous sites where other selective constraints are weaker, thereby explaining the observed patterns. However an extreme selective theory must then reject the idea that a significant portion of bacterial genomes is neutral. The finding of the very large costs of synonymous substitutions in **paper II**, suggests that the number of neutral sites are most likely lower than traditionally assumed and even if these genes are not likely to be representative of the average gene, the fitness costs could be 1000-fold smaller and still evolutionary relevant. Nevertheless, direct selection for increased GC requires a selective advantage of a single GC generating mutation to be larger than the inverse of N_e. This would lead to a strong connection between population size and base composition and species with reduced N_e would not only be unable to fix beneficial GC increasing mutations, but also have a larger number of effectively neutral sites vulnerable to mutational biases. A general explanation for the possible selective value of GC over AT remains to be found, though it is possible that there are several distinct selective pressures operating, leading to similar results, but that their relative importance varies dependent on lifestyle and environmental conditions.

Is there something special about ribosomal protein genes?

The fitness effects of random substitutions (**paper II**) and inter-species gene replacements (**paper III**) were investigated using ribosomal proteins genes as models. The reasons for this choice are discussed above, but of course the question is whether the results are expected to be general or very specific and only valid for ribosomal proteins. Clearly, the cost of random mutations in ribosomal protein genes is substantially higher than expected for a random genomic mutation, with an approximately 50 times higher cost if we use the data from **paper I**. This is not surprising given the fundamental importance of the ribosome in synthesizing all the proteins in the cell. However, a 1000-fold reduction in the selection coefficient compared to the average effect obtained in **paper II** is still high enough for strong counter-selection in most bacterial populations. So the main question is if the DFE is in any way similar for other genes and shows the same large costs for synonymous substitutions with only a minor contribution of amino acid changes. This should be further explored using a completely different type of model gene, possibly a biosynthetic or catabolic enzyme, for which the experimental environment can be set up to make growth directly dependent on the function of the gene. No matter if the results are similar or not, the cause for this may need additional model gene studies focusing on different parameters. The expression level is probably very important for the synonymous effects observed as an evolutionary optimization on codon usage might not be possible for mRNAs present at low levels that are more dependent on stochastic processes. Ribosomal proteins are quite small and the minor effects of amino acid substitutions could be a result of that most changes are on the surface of the protein, whereas for larger proteins a higher proportion of amino acids is in the interior and cause major disruption of tertiary structure, thereby destroying the function of the protein. The lack of an active site and structural role of the ribosomal proteins might also make the DFE differ from that of the typical enzyme.

The ancient origin and high complexity of the ribosome could mean that the evolutionary constraints are fundamentally different from the average gene and further studies should involve other large physical complexes, highly functionally connected single enzymes and genes involved in less complex functional networks. This could also shed some light on if the deleterious effects of the mutations in ribosomal proteins are mainly connected to the production cost of the very expensive ribosomes, so that the effects are amplified by a stoichiometric imbalance of ribosomal parts.

Are there really any good reasons to expect the existence of any general principles that will allow us to use model systems to understand the DFE? The case for the exponential distribution of advantageous mutations is based on an assumption that the starting genotype is well adapted. This condition is often not fulfilled, for example the evolution of antibiotic resistance typically

starts with a very low fitness (sensitivity to the antibiotic) and can allow mutations with large disruptions in cellular functions to invade and evolve by compensatory mutations. It is also hard to see why HGTs that allows invasion of new ecological niches should follow any specific DFE (268). In a model without HGT, the distribution of advantageous mutations will be heavily influenced by the details of the studied system and it seems likely that the number of adaptive paths available is typically limited (269).

The proportion of neutral mutations is hard to estimate, but based on the data from **paper II** and **III**, there seem to exist selective constraints on many levels and we have only begun to sort out the molecular basis for these effects.

The DFE of deleterious mutations in **paper II** has a very peaked nature that is similar for both genes. This suggests the existence of some general principles, so that it is possible to predict mutational effects in a specific system by studying a limited number of sub-components. Could there be an adaptive explanation for the lack of robustness in the ribosomal protein genes or is it simply caused by the fundamental physical properties of the system? Intuitively one might expect the most important system of the cell to be robust so that random mutations will not drastically reduce fitness. It would be catastrophic for humans if mutations commonly lead to a severe reduction in fitness and thus many systems in higher eukaryotic species with small population sizes are robust with redundant backup functions. This is represented by a quite flat fitness landscape where the population is spread over an area around the fitness peak so that most neighboring sequences have a high fitness. The large population sizes of bacteria mean that they are not subject to stochastic genetic drift to the same extent as higher eukaryotes. If flatter fitness landscapes have lower peaks, which are expected if redundancy has a cost, it would be advantageous to occupy steeper peaks when the influence of drift is low (270). This allows effective removal of deleterious mutations from the population so that it is strongly localized to the fitness peak. The evolution of anti-robustness is mediated by anti-redundancy mechanisms, including haploidy, single regulatory elements for many genes, bottlenecks in transmission overlapping reading frames and multiple functions performed by the same gene product (270). Both S20 and L1 serve as translational repressors regulating the expression of their own genes, and in the case of L1 also another ribosomal protein, which could serve to amplify the fitness costs of mutations and reduce genetic drift of components in the translation apparatus. It is difficult to determine if such anti-robustness would be particularly relevant to ribosomal proteins, but it is possible that the high rate of mutations in highly expressed genes (**paper I**) and the potential lack of a need for evolvability in the ribosome could make sharp fitness peaks more advantageous. Even if the magnitude of the fitness costs found in **paper II** are large enough to be interpreted as a non-robustness of these genes, the highly peaked nature of the distribution might be explained by

robustness mechanisms. The similar costs of many of the orthologous replacements in **paper III** compared to the random substitutions could point to the presence of a physiological response, possible a chaperone working on the level of RNA or protein, depending on the mechanistic basis of the costs. After the experiments in **papers II** and **III** have been redone using another type of gene it would be useful to also investigate the effect of other environments and particularly stressful conditions that often increase the relative fitness costs of mutations. Additional experiments of this character should also be performed using another model organism, preferably distantly related to *S. typhimurium* both in lifestyle and phylogeny.

Exploring the evolution of HGT genes

Ribosomal proteins may not be the optimal model genes for investigating the evolution of horizontally transferred genes given that they are rarely transferred in nature. The fitness constraints associated with genes evolved in another host are expected to operate on many levels of sub-optimality and it is likely that compensatory evolution by gene amplification is a general first response when considering the rates of different types of mutations. In **paper III** we also found one promoter mutation and it is likely that the proportion of duplication to promoter mutation are heavily influenced by the experimental design of the compensatory evolution experiment. Sequencing the genomes of the compensated strains without amplifications, promoter mutations or intragenic mutations could provide additional insight into the possible mechanisms used to ameliorate the fitness costs of the gene replacements. It would be interesting to further examine the importance of amplification after HGT using a simpler model gene that can be setup to be limiting for growth and extend the experiment to be able to observe additional compensatory mutations after amplification. It is probably most informative to use quite distantly related genes, so that fitness is severely decreased, as secondary unrelated mutations will otherwise dominate. These genes could also be transferred into several different model species to investigate if the adaptive trajectories are parallel. The impact of differences in the environment and mutation rate on the course of compensatory evolution could also be examined. Experimental studies of the evolution of new genes are probably more challenging, but it is possible that a carefully designed experiment could demonstrate the duplication and divergence of HGT genes.

Mechanistic causes of fitness effects on the mRNA level

The lack of strong correlations between fitness costs of random mutations and codon usage, mRNA stability, protein stability, gene position and conservation can have several possible explanations that need to be pursued by further experiments. A complicating factor is that more than one mechanistic

cause could be important, for example, both very rare codons and large changes in mRNA stability. The highly peaked nature of the fitness data and high uncertainty in computational predictions of mRNA and protein stability makes it impossible to find a strong correlation using this data. So what further options are there? One idea is to try to find another model gene with a larger spread in the fitness effects, but as the costs of synonymous substitutions are assumed to be small in general this might prove to be difficult. It is also possible that the spread of fitness values in the genes used here could be larger in another environment. Using experimental evolution to find compensatory mutations could provide clues to the mechanistic causes, for example, by restoring mRNA structure, but this approach is likely to be plagued by secondary unrelated advantageous mutations and the small increases in fitness possible through intragenic mutations would take a very long time to fix in an experimental population. I suggest instead that the best approach is to design a set of synonymous substitutions and try to compensate the effect of this mutation by a second synonymous substitution compensating for the disruption of mRNA structure or local codon usage. If the fitness costs of both the two single substitutions are larger than the combination of the two this would indeed be a strong argument for the mechanistic cause of the selective constraint. A similar approach could be used to examine intragenic epistasis by combining random synonymous mutations, with known fitness effects, and measure fitness to see if the effects are additive, synergistic or antagonistic. This can also be extended to the non-synonymous mutations to see if there is a stability threshold in these proteins that can be exhausted when several random amino acid substitutions are combined.

Neutral theory and fitness effects of random substitutions and gene replacements

The results from **paper II** and **III** suggests that most synonymous substitutions are not neutral in these ribosomal protein genes and that fitness costs are of similar magnitude for the more divergent, but evolutionary optimized variants of the genes from other species. Are large fitness costs for synonymous substitutions in agreement with nearly neutral theory? Well, the neutral theory focuses on the natural variation between and within species and states that most of the observed differences were not fixed by selection, but by neutral processes. The substitutions studied here were engineered and not found in a natural population, but the results still suggest that the number of neutral sites is smaller than traditionally assumed in neutral theory. It is also extremely important to emphasize the fundamental role of effective population size, which in many bacteria is likely to be large enough to allow selection to act on very small fitness differences, so that the proportion of neutral

substitutions is likely to be several orders of magnitude lower in bacteria compared to mammals.

The orthologous gene replacement experiment, on the other hand, concerns molecular variation present in nature. Could primarily neutral causes to the many nucleotide substitutions observed between species be an explanation of the high functional conservation observed on all levels? Nevertheless, the replacements from even the most closely related species are deleterious. It is possible that the changes were fixed through neutral processes, but that this random drift over evolutionary time caused accumulated differences, which taken together are not neutral. This could be further explored by using less divergent genes, for example those found among strains of *Salmonella*. Selection for optimal ribosome function in using mRNAs with different codon usage and GC content is most likely also important in the evolution of the ribosome.

The impact of neutral HGTs

How much of the differences in genome content due to HGT are caused by neutral processes? The major contribution of mobile genetic elements with increased mechanistic rates of transfer could mean that a significant portion is not selective. This is also supported by comparative sequence studies and the recent origin of many transfers. Alternatively, do genes associated with phages, ICEs and plasmids more often contain selective genes? An expansion of the set of HGT genes studied in **paper IV** to include segments from natural occurring phages would be useful to explore this possibility. The mechanistic rates of transfer in natural populations are unknown and deep-sequencing studies could provide further information on the neutral diversity present. Genome sequencing of many closely related strains will also shed light on the size of the pan genome and core genome and this should be expanded to a wide range of divergent species to obtain a more comprehensive picture. Further experimental studies on the fitness effects of insertions are also necessary to estimate the size of the neutral insertions sites available. We must also obtain a deeper understanding on what levels the main selective constraints are operating on by expanding the set of HGTs studied in **paper IV** and apply this knowledge to genomic sequences. If there is a large cost for a process, such as translation, we will expect transfers with predicted high expression levels to supply a selective advantage. The role of H-NS-mediated gene silencing of HGTs could be further investigated by measuring fitness costs in *hns* knockouts and analyze the connection between AT content and fitness. Laboratory experiments to determine rates of recombination would also be helpful to obtain a more complete picture of the constraints on HGT.

Concluding remarks

Contemporary biology is a multitude of different subsciences with strong connections to mathematics, computer science, physics, chemistry, engineering, medicine and social sciences. This means that it is harder than ever to be a universal genius, even within the biological sciences, and for future progress it is required that biologists move out of their comfort zone and work in interdisciplinary collaborations. Improved communication is needed to make sure that experimentalists produce data that are standardized and clearly defined for use in databases and modeling. Mathematical models should be made understandable to experimentalists and bioinformaticians must be guided by user-centered design principles.

The studies presented in this thesis are to be seen in the light of evolution and population genetics, but also investigates the mechanistic basis for the evolutionary relevant parameters. Genome sequencing coupled with experimental evolution (**paper I**) will soon be a standard technique, but this is only possible if we have good annotated databases and tools for comparative genomics. In **paper II** we try to connect fitness effects with molecular data obtained from structural biology and use bioinformatics tools based on biophysics. The fitness data is fitted to simple distributions for use in models and discussed in the light of genomics data. In **paper III** and **IV** we examine the levels and processes that provide the mechanistic basis for selective constraints on HGT observed in nature. Experimental evolution (**paper III**) is used to find plausible evolutionary solutions to compensate fitness costs of HGT. The mechanism found, gene amplification, is then evaluated using population genetics and found to be theoretically possible and provide an explanation for duplication patterns detected in genomic data from natural populations. Future studies will, hopefully, involve even more interdisciplinary biology, using more -omics data, single molecule techniques, systems biology, synthetic biology and modeling, always seen in the light of evolution.

Acknowledgements

First, I would like to thank my supervisor Dan Andersson for all the support over the years. Your enthusiasm is truly inspiring and it's great to work with someone who really enjoys science. Thanks for helping me get started on the way to becoming an independent researcher and making me feel like a colleague.

I would also like to thank my excellent co-authors. Chuck Kurland for interesting discussions, great stories and for teaching me always to question the opinions of the loud majority. Otto Berg for providing the light of population genetics and all the helpful discussions about the evolutionary implications of the work. Chris Tobin for help with experiments and the English language, but most of all for being great and for all the good times over the years.

My co-supervisor Fredrik Söderbom and all the people that I have discussed my projects with over the years, especially Diarmaid Hughes, Paul Rainey and John Roth and all the nice people from their labs. Prof. Arjan de Visser for accepting the invitation to be external examiner for this thesis.

Thanks to present and former members of the DA-lab. Sanna, for being my lab supervisor when I first joined the lab, insightful scientific comments and for all the great discussions after a few beers. Chris, for always being kind and for setting an example as a great hard-working scientist. Maria, for sharing a desk with me for five years without any fighting and for bringing joy to the lab. Anna, for not getting angry when I steal things from your bench and for persuading me to do the boring stuff when I don't want to. Song, for inspiring me with all your great scientific work. Linus, for being an essential part of keeping the lab together and always taking the time to listen and discuss science. Jocke, for (some of) the music you bring to the lab, always good with some competition, and for sharing your scientific knowledge. Ulrika, for all the assistance with experiments and future collaborations. Annika, for all the help and advice when I first joined the lab. Hervé and Amira for bringing new experiences to the lab and adding to the international character of the group. The next generation of graduate students: Marlen, Lisa, Erik, Hava and Marius for rejuvenating the group. You'll do great and keep the DA-lab successful for years to come.

Special thanks to Sanna for comments and proofreading of this thesis.

I would also like to thank all the people at IMBIM especially the D7:3 and B9:3 corridors and the great administrative and technical staff. Special thanks to Lena for taking care of the whole DA-group and always helping out.

Thanks to my friends for all the good times over the years, especially Linus for his frequent visits to Uppsala and Erik, Erik, Eric, Andreas, Elin, Maria and Eric.

I am very grateful to my family. My father, the original PA Lind, and mother Birgitta for all care, encouragement and support over the last 30 years. My brother David and his wife Maria for always being nice and producing my nephews Åke and Svante. My in-laws, Inger, Henrik, Elin and Erik, for welcoming me into your family.

The only person I can imagine dedicating this thesis to, is my wife Rebecka. Thank you for reminding me that there are many enjoyable things to do outside the lab. We make a great team and you are my number one priority in life. I'm looking forward to reading your thesis.

References

1. Darwin, C. (1859) On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. London: John Murray.
2. Lynch, M. (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci U S A. **104 Suppl 1**: p. 8597-604.
3. Dobzhansky, T. (1973) Nothing in biology makes sense except in the light of evolution. The American Biology Teacher. **35**: p. 125-129.
4. Orr, H.A. (2009) Fitness and its role in evolutionary genetics. Nat Rev Genet. **10**(8): p. 531-9.
5. Loewe, L. (2009) A framework for evolutionary systems biology. BMC Syst Biol. **3**: p. 27.
6. Hurst, L.D. (2009) Fundamental concepts in genetics: genetics and the understanding of selection. Nat Rev Genet. **10**(2): p. 83-93.
7. Wright, S. (1932) The roles of mutation, inbreeding, crossbreeding, and selection in evolution. Proceedings of the Sixth International Congress on Genetics. p. 355-366.
8. Smith, J.M. (1970) Natural selection and the concept of a protein space. Nature. **225**(5232): p. 563-4.
9. Gillespie, J.H. (1983) A simple stochastic gene substitution model. Theor Popul Biol. **23**(2): p. 202-15.
10. Kauffman, S. and Levin, S. (1987) Towards a general theory of adaptive walks on rugged landscapes. J Theor Biol. **128**(1): p. 11-45.
11. Orr, H.A. (2002) The population genetics of adaptation: the adaptation of DNA sequences. Evolution. **56**(7): p. 1317-30.
12. Biebricher, C.K. and Eigen, M. (2005) The error threshold. Virus Res. **107**(2): p. 117-27.
13. Rice, S.H. (2008) Theoretical approaches to the evolution of development and genetic architecture. Ann N Y Acad Sci. **1133**: p. 67-86.
14. Orr, H.A. (2005) The genetic theory of adaptation: a brief history. Nat Rev Genet. **6**(2): p. 119-27.
15. Gillespie, J.H. (1991) The causes of molecular evolution. New York: Oxford University Press.
16. Gravner, J., Pitman, D., and Gavrilets, S. (2007) Percolation on fitness landscapes: effects of correlation, phenotype, and incompatibilities. J Theor Biol. **248**(4): p. 627-45.
17. Wright, S. (1931) Evolution in Mendelian Populations. Genetics. **16**(2): p. 97-159.
18. Wright, S. (1938) Size of population and breeding structure in relation to evolution. Science. **87**: p. 430-431.
19. Frankham, R. (1995) Effective population size/adult population size ratios in wildlife: a review. Genet Res. **66**: p. 95-107.
20. Kimura, M. (1983) The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press.

21. Charlesworth, B. (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet. **10**(3): p. 195-205.
22. Wright, S. (1940) Breeding structure of a species in relation to speciation. Am. Nat. **74**: p. 232-248.
23. Yu, N., Jensen-Seaman, M.I., Chemnick, L., Ryder, O., and Li, W.H. (2004) Nucleotide diversity in gorillas. Genetics. **166**(3): p. 1375-83.
24. Nordborg, M. (1997) Structured coalescent processes on different time scales. Genetics. **146**(4): p. 1501-14.
25. Fraser, C., Hanage, W.P., and Spratt, B.G. (2007) Recombination and the nature of bacterial speciation. Science. **315**(5811): p. 476-80.
26. Feil, E.J., et al. (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci U S A. **98**(1): p. 182-7.
27. Touchon, M., et al. (2009) Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. **5**(1): p. e1000344.
28. Kimura, M. (1962) On the probability of fixation of mutant genes in a population. Genetics. **47**: p. 713-9.
29. Keightley, P.D. and Eyre-Walker, A. (2010) What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? Philos Trans R Soc Lond B Biol Sci. **365**(1544): p. 1187-93.
30. Eyre-Walker, A. and Keightley, P.D. (2007) The distribution of fitness effects of new mutations. Nat Rev Genet. **8**(8): p. 610-8.
31. Rozen, D.E., de Visser, J.A., and Gerrish, P.J. (2002) Fitness effects of fixed beneficial mutations in microbial populations. Curr Biol. **12**(12): p. 1040-5.
32. de Visser, J.A. and Rozen, D.E. (2005) Limits to adaptation in asexual populations. J Evol Biol. **18**(4): p. 779-88.
33. de Visser, J.A. and Rozen, D.E. (2006) Clonal interference and the periodic selection of new beneficial mutations in Escherichia coli. Genetics. **172**(4): p. 2093-100.
34. Rainey, P.B. and Travisano, M. (1998) Adaptive radiation in a heterogeneous environment. Nature. **394**(6688): p. 69-72.
35. Dykhuizen, D.E. and Dean, A.M. (2004) Evolution of specialists in an experimental microcosm. Genetics. **167**(4): p. 2015-26.
36. Habets, M.G., Rozen, D.E., Hoekstra, R.F., and de Visser, J.A. (2006) The effect of population structure on the adaptive radiation of microbial populations evolving in spatially structured environments. Ecol Lett. **9**(9): p. 1041-8.
37. Gillespie, J.H. (1984) Molecular evolution over the mutational landscape. Evolution. **38**: p. 1116-1129.
38. Orr, H.A. (2003) The distribution of fitness effects among beneficial mutations. Genetics. **163**(4): p. 1519-26.
39. Imhof, M. and Schlotterer, C. (2001) Fitness effects of advantageous mutations in evolving Escherichia coli populations. Proc Natl Acad Sci U S A. **98**(3): p. 1113-7.
40. Kassen, R. and Bataillon, T. (2006) Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nat Genet. **38**(4): p. 484-8.
41. Sanjuan, R., Moya, A., and Elena, S.F. (2004) The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl Acad Sci U S A. **101**(22): p. 8396-401.

42. Elena, S.F., Ekunwe, L., Hajela, N., Oden, S.A., and Lenski, R.E. (1998) Distribution of fitness effects caused by random insertion mutations in Escherichia coli. Genetica. **102-103**(1-6): p. 349-58.
43. Butcher, D. (1995) Muller's ratchet, epistasis and mutation effects. Genetics. **141**(1): p. 431-7.
44. Kimura, M. (1985) The role of compensatory neutral mutations in molecular evolution. J Genet. **64**: p. 7-19.
45. Knies, J.L., Dang, K.K., Vision, T.J., Hoffman, N.G., Swanstrom, R., and Burch, C.L. (2008) Compensatory evolution in RNA secondary structures increases substitution rate variation among sites. Mol Biol Evol. **25**(8): p. 1778-87.
46. Andersson, D.I. and Hughes, D. (2010) Antibiotic resistance and its cost: is it possible to reverse resistance? Nat Rev Microbiol. **8**(4): p. 260-71.
47. Andersson, D.I. (2006) The biological cost of mutational antibiotic resistance: any practical conclusions? Curr Opin Microbiol. **9**(5): p. 461-5.
48. Kimura, M. (1968) Evolutionary rate at the molecular level. Nature. **217**(5129): p. 624-6.
49. Ohta, T. (1972) Population size and rate of evolution. J Mol Evol. **1**(3): p. 305-314.
50. Ohta, T. (1973) Slightly deleterious mutant substitutions in evolution. Nature. **246**(5428): p. 96-8.
51. Nei, M. (2005) Selectionism and neutralism in molecular evolution. Mol Biol Evol. **22**(12): p. 2318-42.
52. McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature. **351**(6328): p. 652-4.
53. Dong, H., Nilsson, L., and Kurland, C.G. (1996) Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. J Mol Biol. **260**(5): p. 649-63.
54. Ikemura, T. (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol. **151**(3): p. 389-409.
55. Wang, H.C., Badger, J., Kearney, P., and Li, M. (2001) Analysis of codon usage patterns of bacterial genomes using the self-organizing map. Mol Biol Evol. **18**(5): p. 792-800.
56. Sharp, P.M. and Li, W.H. (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol. **24**(1-2): p. 28-38.
57. Welch, M., Govindarajan, S., Ness, J.E., Villalobos, A., Gurney, A., Minshull, J., and Gustafsson, C. (2009) Design parameters to control synthetic gene expression in Escherichia coli. PLoS One. **4**(9): p. e7002.
58. Tuller, T., Waldman, Y.Y., Kupiec, M., and Ruppin, E. (2010) Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A. **107**(8): p. 3645-50.
59. Drummond, D.A. and Wilke, C.O. (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell. **134**(2): p. 341-52.
60. Kramer, E.B. and Farabaugh, P.J. (2007) The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. RNA. **13**(1): p. 87-96.
61. Ogle, J.M. and Ramakrishnan, V. (2005) Structural insights into translational fidelity. Annu Rev Biochem. **74**: p. 129-77.

62. Parker, J. (1989) Errors and alternatives in reading the universal genetic code. Microbiol Rev. **53**(3): p. 273-98.
63. Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B. (2009) Coding-sequence determinants of gene expression in Escherichia coli. Science. **324**(5924): p. 255-8.
64. Andersson, S.G. and Kurland, C.G. (1990) Codon preferences in free-living microorganisms. Microbiol Rev. **54**(2): p. 198-210.
65. Coleman, J.R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E., and Mueller, S. (2008) Virus attenuation by genome-scale changes in codon pair bias. Science. **320**(5884): p. 1784-7.
66. Kimchi-Sarfaty, C., Oh, J.M., Kim, I.W., Sauna, Z.E., Calcagno, A.M., Ambudkar, S.V., and Gottesman, M.M. (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science. **315**(5811): p. 525-8.
67. Pagani, F., Raponi, M., and Baralle, F.E. (2005) Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc Natl Acad Sci U S A. **102**(18): p. 6368-72.
68. Chao, H.K., Hsiao, K.J., and Su, T.S. (2001) A silent mutation induces exon skipping in the phenylalanine hydroxylase gene in phenylketonuria. Hum Genet. **108**(1): p. 14-9.
69. Montera, M., et al. (2001) A silent mutation in exon 14 of the APC gene is associated with exon skipping in a FAP family. J Med Genet. **38**(12): p. 863-7.
70. Carlini, D.B. (2004) Experimental reduction of codon bias in the Drosophila alcohol dehydrogenase gene results in decreased ethanol tolerance of adult flies. J Evol Biol. **17**(4): p. 779-85.
71. Carlini, D.B. and Stephan, W. (2003) In vivo introduction of unpreferred synonymous codons into the Drosophila Adh gene results in reduced levels of ADH protein. Genetics. **163**(1): p. 239-43.
72. Cupples, C.G. and Miller, J.H. (1988) Effects of amino acid substitutions at the active site in Escherichia coli beta-galactosidase. Genetics. **120**(3): p. 637-44.
73. Suckow, J., Markiewicz, P., Kleina, L.G., Miller, J., Kisters-Woike, B., and Muller-Hill, B. (1996) Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. J Mol Biol. **261**(4): p. 509-23.
74. Tokuriki, N. and Tawfik, D.S. (2009) Stability effects of mutations and protein evolvability. Curr Opin Struct Biol. **19**(5): p. 596-604.
75. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., and Tawfik, D.S. (2006) Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. Nature. **444**(7121): p. 929-32.
76. Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L., and Tawfik, D.S. (2007) The stability effects of protein mutations appear to be universally distributed. J Mol Biol. **369**(5): p. 1318-32.
77. Nene, V. and Glass, R.E. (1982) Genetic studies on the beta subunit of Escherichia coli RNA polymerase. I. The effect of known, single amino acid substitutions in an essential protein. Mol Gen Genet. **188**(3): p. 399-404.
78. Amoros-Moya, D., Bedhomme, S., Hermann, M., and Bravo, I.G. (2010) Evolution in regulatory regions rapidly compensates the cost of nonoptimal codon usage. Mol Biol Evol. **27**(9): p. 2141-51.
79. Spratt, B.G., Hanage, W.P., and Feil, E.J. (2001) The relative contributions of recombination and point mutation to the diversification of bacterial clones. Curr Opin Microbiol. **4**(5): p. 602-6.

80. Smith, J.M., Smith, N.H., O'Rourke, M., and Spratt, B.G. (1993) How clonal are bacteria? Proc Natl Acad Sci U S A. **90**(10): p. 4384-8.
81. Thomas, C.M. and Nielsen, K.M. (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol. **3**(9): p. 711-21.
82. Hendrickson, H. and Lawrence, J.G. (2006) Selection for chromosome architecture in bacteria. J Mol Evol. **62**(5): p. 615-29.
83. Stoebel, D.M., Dean, A.M., and Dykhuizen, D.E. (2008) The cost of expression of Escherichia coli lac operon proteins is in the process, not in the products. Genetics. **178**(3): p. 1653-60.
84. Berg, O.G. and Kurland, C.G. (2002) Evolution of microbial genomes: sequence acquisition and loss. Mol Biol Evol. **19**(12): p. 2265-76.
85. Navarre, W.W., Porwollik, S., Wang, Y., McClelland, M., Rosen, H., Libby, S.J., and Fang, F.C. (2006) Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella. Science. **313**(5784): p. 236-8.
86. Stoebel, D.M., Free, A., and Dorman, C.J. (2008) Anti-silencing: overcoming H-NS-mediated repression of transcription in Gram-negative enteric bacteria. Microbiology. **154**(Pt 9): p. 2533-45.
87. Oshima, T., Ishikawa, S., Kurokawa, K., Aiba, H., and Ogasawara, N. (2006) Escherichia coli histone-like protein H-NS preferentially binds to horizontally acquired DNA in association with RNA polymerase. DNA Res. **13**(4): p. 141-53.
88. Pettersson, M.E., Sun, S., Andersson, D.I., and Berg, O.G. (2009) Evolution of new gene functions: simulation and analysis of the amplification model. Genetica. **135**(3): p. 309-24.
89. Sandegren, L. and Andersson, D.I. (2009) Bacterial gene amplification: implications for the evolution of antibiotic resistance. Nat Rev Microbiol. **7**(8): p. 578-88.
90. Ohno, S. (1970) Evolution by Gene Duplication. New York: Springer. 160.
91. Bergthorsson, U., Andersson, D.I., and Roth, J.R. (2007) Ohno's dilemma: evolution of new genes under continuous selection. Proc Natl Acad Sci U S A. **104**(43): p. 17004-9.
92. Kuo, C.H. and Ochman, H. (2009) The fate of new bacterial genes. FEMS Microbiol Rev. **33**(1): p. 38-43.
93. de Visser, J.A., et al. (2003) Perspective: Evolution and detection of genetic robustness. Evolution. **57**(9): p. 1959-72.
94. Phillips, P.C. (2008) Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet. **9**(11): p. 855-67.
95. Drake, J.W., Charlesworth, B., Charlesworth, D., and Crow, J.F. (1998) Rates of spontaneous mutation. Genetics. **148**(4): p. 1667-86.
96. Drake, J.W. (1991) A constant rate of spontaneous mutation in DNA-based microbes. Proc Natl Acad Sci U S A. **88**(16): p. 7160-4.
97. Ochman, H. (2003) Neutral mutations and neutral substitutions in bacterial genomes. Mol Biol Evol. **20**(12): p. 2091-6.
98. Rosche, W.A. and Foster, P.L. (2000) Determining mutation rates in bacterial populations. Methods. **20**(1): p. 4-17.
99. Aravind, L., Walker, D.R., and Koonin, E.V. (1999) Conserved domains in DNA repair proteins and evolution of repair systems. Nucleic Acids Res. **27**(5): p. 1223-42.
100. Horst, J.P., Wu, T.H., and Marinus, M.G. (1999) Escherichia coli mutator genes. Trends Microbiol. **7**(1): p. 29-36.

101. Schaaper, R.M. (1993) Base selection, proofreading, and mismatch repair during DNA replication in Escherichia coli. J Biol Chem. **268**(32): p. 23762-5.
102. Nohmi, T. (2006) Environmental stress and lesion-bypass DNA polymerases. Annu Rev Microbiol. **60**: p. 231-53.
103. Wagner, J., Fujii, S., Gruz, P., Nohmi, T., and Fuchs, R.P. (2000) The beta clamp targets DNA polymerase IV to DNA and strongly increases its processivity. EMBO Rep. **1**(6): p. 484-8.
104. Kunkel, T.A. and Erie, D.A. (2005) DNA mismatch repair. Annu Rev Biochem. **74**: p. 681-710.
105. LeClerc, J.E., Li, B., Payne, W.L., and Cebula, T.A. (1996) High mutation frequencies among Escherichia coli and Salmonella pathogens. Science. **274**(5290): p. 1208-11.
106. Le Gall, S., Desbordes, L., Gracieux, P., Saffroy, S., Bousarghin, L., Bonnaure-Mallet, M., and Jolivet-Gougeon, A. (2009) Distribution of mutation frequencies among Salmonella enterica isolates from animal and human sources and genetic characterization of a Salmonella Heidelberg hypermutator. Vet Microbiol. **137**(3-4): p. 306-12.
107. Nilsson, A.I., Kugelberg, E., Berg, O.G., and Andersson, D.I. (2004) Experimental adaptation of Salmonella typhimurium to mice. Genetics. **168**(3): p. 1119-30.
108. Coulondre, C., Miller, J.H., Farabaugh, P.J., and Gilbert, W. (1978) Molecular basis of base substitution hotspots in Escherichia coli. Nature. **274**(5673): p. 775-80.
109. Lieb, M. (1991) Spontaneous mutation at a 5-methylcytosine hotspot is prevented by very short patch (VSP) mismatch repair. Genetics. **128**(1): p. 23-7.
110. Lieb, M., Allen, E., and Read, D. (1986) Very short patch mismatch repair in phage lambda: repair sites and length of repair tracts. Genetics. **114**(4): p. 1041-60.
111. Sohail, A., Lieb, M., Dar, M., and Bhagwat, A.S. (1990) A gene required for very short patch repair in Escherichia coli is adjacent to the DNA cytosine methylase gene. J Bacteriol. **172**(8): p. 4214-21.
112. Dar, M.E. and Bhagwat, A.S. (1993) Mechanism of expression of DNA repair gene vsr, an Escherichia coli gene that overlaps the DNA cytosine methylase gene, dcm. Mol Microbiol. **9**(4): p. 823-33.
113. Doiron, K.M., Viau, S., Koutroumanis, M., and Cupples, C.G. (1996) Overexpression of vsr in Escherichia coli is mutagenic. J Bacteriol. **178**(14): p. 4294-6.
114. Falnes, P.O. and Rognes, T. (2003) DNA repair by bacterial AlkB proteins. Res Microbiol. **154**(8): p. 531-8.
115. Trewick, S.C., Henshaw, T.F., Hausinger, R.P., Lindahl, T., and Sedgwick, B. (2002) Oxidative demethylation by Escherichia coli AlkB directly reverts DNA base damage. Nature. **419**(6903): p. 174-8.
116. Mackay, W.J., Han, S., and Samson, L.D. (1994) DNA alkylation repair limits spontaneous base substitution mutations in Escherichia coli. J Bacteriol. **176**(11): p. 3224-30.
117. Heelis, P.F., Kim, S.T., Okamura, T., and Sancar, A. (1993) The photo repair of pyrimidine dimers by DNA photolyase and model systems. J Photochem Photobiol B. **17**(3): p. 219-28.
118. Essen, L.O. and Klar, T. (2006) Light-driven DNA repair by photolyases. Cell Mol Life Sci. **63**(11): p. 1266-77.

119. Lindahl, T. (1982) DNA repair enzymes. Annu Rev Biochem. **51**: p. 61-87.
120. Sancar, A. and Sancar, G.B. (1988) DNA repair enzymes. Annu Rev Biochem. **57**: p. 29-67.
121. Krokan, H.E., Standal, R., and Slupphaug, G. (1997) DNA glycosylases in the base excision repair of DNA. Biochem J. **325 (Pt 1)**: p. 1-16.
122. Neeley, W.L. and Essigmann, J.M. (2006) Mechanisms of formation, genotoxicity, and mutation of guanine oxidation products. Chem Res Toxicol. **19**(4): p. 491-505.
123. Burrows, C.J. and Muller, J.G. (1998) Oxidative Nucleobase Modifications Leading to Strand Scission. Chem Rev. **98**(3): p. 1109-1152.
124. Shibutani, S., Takeshita, M., and Grollman, A.P. (1991) Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. Nature. **349**(6308): p. 431-4.
125. Michaels, M.L. and Miller, J.H. (1992) The GO system protects organisms from the mutagenic effect of the spontaneous lesion 8-hydroxyguanine (7,8-dihydro-8-oxoguanine). J Bacteriol. **174**(20): p. 6321-5.
126. Michaels, M.L., Cruz, C., Grollman, A.P., and Miller, J.H. (1992) Evidence that MutY and MutM combine to prevent mutations by an oxidatively damaged form of guanine in DNA. Proc Natl Acad Sci U S A. **89**(15): p. 7022-5.
127. Fowler, R.G., White, S.J., Koyama, C., Moore, S.C., Dunn, R.L., and Schaaper, R.M. (2003) Interactions among the Escherichia coli mutT, mutM, and mutY damage prevention pathways. DNA Repair (Amst). **2**(2): p. 159-73.
128. Schaaper, R.M., Bond, B.I., and Fowler, R.G. (1989) A.T----C.G transversions and their prevention by the Escherichia coli mutT and mutHLS pathways. Mol Gen Genet. **219**(1-2): p. 256-62.
129. Duncan, B.K. and Miller, J.H. (1980) Mutagenic deamination of cytosine residues in DNA. Nature. **287**(5782): p. 560-1.
130. Savva, R., McAuley-Hecht, K., Brown, T., and Pearl, L. (1995) The structural basis of specific base-excision repair by uracil-DNA glycosylase. Nature. **373**(6514): p. 487-93.
131. Duncan, B.K., Rockstroh, P.A., and Warner, H.R. (1978) Escherichia coli K-12 mutants deficient in uracil-DNA glycosylase. J Bacteriol. **134**(3): p. 1039-45.
132. Hayakawa, H., Kumura, K., and Sekiguchi, M. (1978) Role of uracil-DNA glycosylase in the repair of deaminated cytosine residues of DNA in Escherichia coli. J Biochem. **84**(5): p. 1155-64.
133. Gallinari, P. and Jiricny, J. (1996) A new class of uracil-DNA glycosylases related to human thymine-DNA glycosylase. Nature. **383**(6602): p. 735-8.
134. O'Neill, R.J., Vorob'eva, O.V., Shahbakhti, H., Zmuda, E., Bhagwat, A.S., and Baldwin, G.S. (2003) Mismatch uracil glycosylase from Escherichia coli: a general mismatch or a specific DNA glycosylase? J Biol Chem. **278**(23): p. 20526-32.
135. Lutsenko, E. and Bhagwat, A.S. (1999) The role of the Escherichia coli mug protein in the removal of uracil and 3,N(4)-ethenocytosine from DNA. J Biol Chem. **274**(43): p. 31034-8.
136. Mokkapati, S.K., Fernandez de Henestrosa, A.R., and Bhagwat, A.S. (2001) Escherichia coli DNA glycosylase Mug: a growth-regulated enzyme required for mutation avoidance in stationary-phase cells. Mol Microbiol. **41**(5): p. 1101-11.
137. Van Houten, B. (1990) Nucleotide excision repair in Escherichia coli. Microbiol Rev. **54**(1): p. 18-51.

138. Moolenaar, G.F., Moorman, C., and Goosen, N. (2000) Role of the Escherichia coli nucleotide excision repair proteins in DNA replication. J Bacteriol. **182**(20): p. 5706-14.
139. Mellon, I. (2005) Transcription-coupled repair: a complex affair. Mutat Res. **577**(1-2): p. 155-61.
140. Sueoka, N. (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol. **40**(3): p. 318-25.
141. Lobry, J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol. **13**(5): p. 660-5.
142. Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K., and White, O. (2001) The Comprehensive Microbial Resource. Nucleic Acids Res. **29**(1): p. 123-5.
143. Hildebrand, F., Meyer, A., and Eyre-Walker, A. (2010) Evidence of selection upon genomic GC content in bacteria. PLoS Genet. **6**(9).
144. Hershberg, R. and Petrov, D.A. (2010) Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. **6**(9).
145. Rocha, E.P. and Feil, E.J. (2010) Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria? PLoS Genet. **6**(9).
146. Rocha, E.P. and Danchin, A. (2002) Base composition bias might result from competition for metabolic resources. Trends Genet. **18**(6): p. 291-4.
147. Suyama, A. and Wada, A. (1982) Correlation between thermal stability maps and genetic maps of double-stranded DNAs. Nucleic Acids Symp Ser. (11): p. 165-8.
148. Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valin, F., and Bernardi, G. (2006) Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochem Biophys Res Commun. **347**(1): p. 1-3.
149. Singer, C.E. and Ames, B.N. (1970) Sunlight ultraviolet and bacterial DNA base ratios. Science. **170**(960): p. 822-5.
150. Naya, H., Romero, H., Zavala, A., Alvarez, B., and Musto, H. (2002) Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. J Mol Evol. **55**(3): p. 260-4.
151. McEwan, C.E., Gatherer, D., and McEwan, N.R. (1998) Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. Hereditas. **128**(2): p. 173-8.
152. Marians, K.J. (1992) Prokaryotic DNA replication. Annu Rev Biochem. **61**: p. 673-719.
153. Lobry, J.R. and Sueoka, N. (2002) Asymmetric directional mutation pressures in bacteria. Genome Biol. **3**(10): p. RESEARCH0058.
154. Sueoka, N. (1999) Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C. J Mol Evol. **49**(1): p. 49-62.
155. Hendrickson, H. and Lawrence, J.G. (2007) Mutational bias suggests that replication termination occurs near the dif site, not at Ter sites. Mol Microbiol. **64**(1): p. 42-56.
156. Rocha, E.P. and Danchin, A. (2003) Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat Genet. **34**(4): p. 377-8.
157. Morton, R.A. and Morton, B.R. (2007) Separating the effects of mutation and selection in producing DNA skew in bacterial chromosomes. BMC Genomics. **8**: p. 369.

158. El Karoui, M., Biaudet, V., Schbath, S., and Gruss, A. (1999) Characteristics of Chi distribution on different bacterial genomes. Res Microbiol. **150**(9-10): p. 579-87.
159. Persky, N.S. and Lovett, S.T. (2008) Mechanisms of recombination: lessons from E. coli. Crit Rev Biochem Mol Biol. **43**(6): p. 347-70.
160. Spies, M., Dillingham, M.S., and Kowalczykowski, S.C. (2005) Translocation by the RecB motor is an absolute requirement for {chi}-recognition and RecA protein loading by RecBCD enzyme. J Biol Chem. **280**(44): p. 37078-87.
161. Anderson, D.G. and Kowalczykowski, S.C. (1997) The translocating RecBCD enzyme stimulates recombination by directing RecA protein onto ssDNA in a chi-regulated manner. Cell. **90**(1): p. 77-86.
162. Moreau, P.L. (1988) Overproduction of single-stranded-DNA-binding protein specifically inhibits recombination of UV-irradiated bacteriophage DNA in Escherichia coli. J Bacteriol. **170**(6): p. 2493-500.
163. Yu, X., VanLoock, M.S., Yang, S., Reese, J.T., and Egelman, E.H. (2004) What is the structure of the RecA-DNA filament? Curr Protein Pept Sci. **5**(2): p. 73-9.
164. West, S.C. (2003) Molecular views of recombination proteins and their control. Nat Rev Mol Cell Biol. **4**(6): p. 435-45.
165. Kouzminova, E.A. and Kuzminov, A. (2004) Chromosomal fragmentation in dUTPase-deficient mutants of Escherichia coli and its recombinational repair. Mol Microbiol. **51**(5): p. 1279-95.
166. Michel, B., Boubakri, H., Baharoglu, Z., LeMasson, M., and Lestini, R. (2007) Recombination proteins and rescue of arrested replication forks. DNA Repair (Amst). **6**(7): p. 967-80.
167. Ratnakumar, A., Mousset, S., Glemin, S., Berglund, J., Galtier, N., Duret, L., and Webster, M.T. (2010) Detecting positive selection within genomes: the problem of biased gene conversion. Philos Trans R Soc Lond B Biol Sci. **365**(1552): p. 2571-80.
168. Chen, J.F., Lu, F., Chen, S.S., and Tao, S.H. (2006) Significant positive correlation between the recombination rate and GC content in the human pseudoautosomal region. Genome. **49**(5): p. 413-9.
169. Marais, G., Charlesworth, B., and Wright, S.I. (2004) Recombination and base composition: the case of the highly self-fertilizing plant Arabidopsis thaliana. Genome Biol. **5**(7): p. R45.
170. Nagylaki, T. (1983) Evolution of a finite population under gene conversion. Proc Natl Acad Sci U S A. **80**(20): p. 6278-81.
171. Dutra, B.E., Sutera, V.A., Jr., and Lovett, S.T. (2007) RecA-independent recombination is efficient but limited by exonucleases. Proc Natl Acad Sci U S A. **104**(1): p. 216-21.
172. Shanado, Y., Kato, J., and Ikeda, H. (1998) Escherichia coli HU protein suppresses DNA-gyrase-mediated illegitimate recombination and SOS induction. Genes Cells. **3**(8): p. 511-20.
173. Ikeda, H., Shiraishi, K., and Ogata, Y. (2004) Illegitimate recombination mediated by double-strand break and end-joining in Escherichia coli. Adv Biophys. **38**: p. 3-20.
174. Onda, M., Yamaguchi, J., Hanada, K., Asami, Y., and Ikeda, H. (2001) Role of DNA ligase in the illegitimate recombination that generates lambdabio-transducing phages in Escherichia coli. Genetics. **158**(1): p. 29-39.
175. Rajeev, L., Malanowska, K., and Gardner, J.F. (2009) Challenging a paradigm: the role of DNA homology in tyrosine recombinase reactions. Microbiol Mol Biol Rev. **73**(2): p. 300-9.

176. Groth, A.C. and Calos, M.P. (2004) Phage integrases: biology and applications. J Mol Biol. **335**(3): p. 667-78.
177. Kleckner, N. (1981) Transposable elements in prokaryotes. Annu Rev Genet. **15**: p. 341-404.
178. George, J., Devoret, R., and Radman, M. (1974) Indirect ultraviolet-reactivation of phage lambda. Proc Natl Acad Sci U S A. **71**(1): p. 144-7.
179. Radman, M. (1975) SOS repair hypothesis: phenomenology of an inducible DNA repair which is accompanied by mutagenesis. Basic Life Sci. **5A**: p. 355-67.
180. Janion, C. (2008) Inducible SOS response system of DNA repair and mutagenesis in Escherichia coli. Int J Biol Sci. **4**(6): p. 338-44.
181. Butala, M., Zgur-Bertok, D., and Busby, S.J. (2009) The bacterial LexA transcriptional repressor. Cell Mol Life Sci. **66**(1): p. 82-93.
182. Chen, I., Christie, P.J., and Dubnau, D. (2005) The ins and outs of DNA transfer in bacteria. Science. **310**(5753): p. 1456-60.
183. Frost, L.S., Leplae, R., Summers, A.O., and Toussaint, A. (2005) Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol. **3**(9): p. 722-32.
184. Claverys, J.P. and Martin, B. (2003) Bacterial "competence" genes: signatures of active transformation, or only remnants? Trends Microbiol. **11**(4): p. 161-5.
185. Hamoen, L.W., Venema, G., and Kuipers, O.P. (2003) Controlling competence in Bacillus subtilis: shared use of regulators. Microbiology. **149**(Pt 1): p. 9-17.
186. Claverys, J.P. and Havarstein, L.S. (2002) Extracellular-peptide control of competence for genetic transformation in Streptococcus pneumoniae. Front Biosci. **7**: p. d1798-814.
187. Sisco, K.L. and Smith, H.O. (1979) Sequence-specific DNA uptake in Haemophilus transformation. Proc Natl Acad Sci U S A. **76**(2): p. 972-6.
188. Goodman, S.D. and Scocca, J.J. (1988) Identification and arrangement of the DNA sequence recognized in specific transformation of Neisseria gonorrhoeae. Proc Natl Acad Sci U S A. **85**(18): p. 6982-6.
189. Lorenz, M.G. and Wackernagel, W. (1994) Bacterial gene transfer by natural genetic transformation in the environment. Microbiol Rev. **58**(3): p. 563-602.
190. Romanowski, G., Lorenz, M.G., Sayler, G., and Wackernagel, W. (1992) Persistence of Free Plasmid DNA in Soil Monitored by Various Methods, Including a Transformation Assay. Appl Environ Microbiol. **58**(9): p. 3012-3019.
191. Saye, D.J., Ogunseitan, O., Sayler, G.S., and Miller, R.V. (1987) Potential for transduction of plasmids in a natural freshwater environment: effect of plasmid donor concentration and a natural microbial community on transduction in Pseudomonas aeruginosa. Appl Environ Microbiol. **53**(5): p. 987-95.
192. Paul, J.H., Jeffrey, W.H., David, A.W., Deflaun, M.F., and Cazares, L.H. (1989) Turnover of Extracellular DNA in Eutrophic and Oligotrophic Freshwater Environments of Southwest Florida. Appl Environ Microbiol. **55**(7): p. 1823-1828.
193. Whitchurch, C.B., Tolker-Nielsen, T., Ragas, P.C., and Mattick, J.S. (2002) Extracellular DNA required for bacterial biofilm formation. Science. **295**(5559): p. 1487.
194. Sorensen, S.J., Bailey, M., Hansen, L.H., Kroer, N., and Wuertz, S. (2005) Studying plasmid horizontal transfer in situ: a critical review. Nat Rev Microbiol. **3**(9): p. 700-10.

195. Cascales, E. and Christie, P.J. (2003) The versatile bacterial type IV secretion systems. Nat Rev Microbiol. **1**(2): p. 137-49.
196. Figge, R.M., Schubert, M., Brinkmann, H., and Cerff, R. (1999) Glyceraldehyde-3-phosphate dehydrogenase gene diversity in eubacteria and eukaryotes: evidence for intra- and inter-kingdom gene transfer. Mol Biol Evol. **16**(4): p. 429-40.
197. Heinemann, J.A. and Sprague, G.F., Jr. (1989) Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. Nature. **340**(6230): p. 205-9.
198. Burrus, V. and Waldor, M.K. (2004) Shaping bacterial genomes with integrative and conjugative elements. Res Microbiol. **155**(5): p. 376-86.
199. Krause, S., Barcena, M., Pansegrau, W., Lurz, R., Carazo, J.M., and Lanka, E. (2000) Sequence-related protein export NTPases encoded by the conjugative transfer region of RP4 and by the cag pathogenicity island of Helicobacter pylori share similar hexameric ring structures. Proc Natl Acad Sci U S A. **97**(7): p. 3067-72.
200. Prentice, M.B., et al. (2001) Yersinia pestis pFra shows biovar-specific differences and recent common ancestry with a Salmonella enterica serovar Typhi plasmid. J Bacteriol. **183**(8): p. 2586-94.
201. Brussow, H., Canchaya, C., and Hardt, W.D. (2004) Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev. **68**(3): p. 560-602.
202. Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. Nature. **405**(6784): p. 299-304.
203. Pedulla, M.L., et al. (2003) Origins of highly mosaic mycobacteriophage genomes. Cell. **113**(2): p. 171-82.
204. Wilhelm, S.W., Brigden, S.M., and Suttle, C.A. (2002) A dilution technique for the direct measurement of viral production: a comparison in stratified and tidally mixed coastal waters. Microb Ecol. **43**(1): p. 168-73.
205. Sun, J., Inouye, M., and Inouye, S. (1991) Association of a retroelement with a P4-like cryptic prophage (retronphage phi R73) integrated into the selenocystyl tRNA gene of Escherichia coli. J Bacteriol. **173**(13): p. 4171-81.
206. Cheetham, B.F. and Katz, M.E. (1995) A role for bacteriophages in the evolution and transfer of bacterial virulence determinants. Mol Microbiol. **18**(2): p. 201-8.
207. Lindsay, J.A., Ruzin, A., Ross, H.F., Kurepina, N., and Novick, R.P. (1998) The gene for toxic shock toxin is carried by a family of mobile pathogenicity islands in Staphylococcus aureus. Mol Microbiol. **29**(2): p. 527-43.
208. Weeks, C.R. and Ferretti, J.J. (1984) The gene for type A streptococcal exotoxin (erythrogenic toxin) is located in bacteriophage T12. Infect Immun. **46**(2): p. 531-6.
209. Jackson, M.P., Newland, J.W., Holmes, R.K., and O'Brien, A.D. (1987) Nucleotide sequence analysis of the structural genes for Shiga-like toxin I encoded by bacteriophage 933J from Escherichia coli. Microb Pathog. **2**(2): p. 147-53.
210. Daubin, V., Lerat, E., and Perriere, G. (2003) The source of laterally transferred genes in bacterial genomes. Genome Biol. **4**(9): p. R57.
211. Jeltsch, A. (2003) Maintenance of species identity and controlling speciation of bacteria: a new function for restriction/modification systems? Gene. **317**(1-2): p. 13-6.
212. Jeltsch, A. and Pingoud, A. (1996) Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. J Mol Evol. **42**(2): p. 91-6.

213. Belogurov, A.A., Delver, E.P., and Rodzevich, O.V. (1993) Plasmid pKM101 encodes two nonhomologous antirestriction proteins (ArdA and ArdB) whose expression is controlled by homologous regulatory sequences. J Bacteriol. **175**(15): p. 4843-50.
214. Belogurov, A.A., Delver, E.P., and Rodzevich, O.V. (1992) IncN plasmid pKM101 and IncI1 plasmid ColIb-P9 encode homologous antirestriction proteins in their leading regions. J Bacteriol. **174**(15): p. 5079-85.
215. Sorek, R., Kunin, V., and Hugenholtz, P. (2008) CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat Rev Microbiol. **6**(3): p. 181-6.
216. Marraffini, L.A. and Sontheimer, E.J. (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science. **322**(5909): p. 1843-5.
217. Barrangou, R., et al. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science. **315**(5819): p. 1709-12.
218. Brouns, S.J., et al. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. Science. **321**(5891): p. 960-4.
219. Horvath, P. and Barrangou, R. (2010) CRISPR/Cas, the immune system of bacteria and archaea. Science. **327**(5962): p. 167-70.
220. Godde, J.S. and Bickerton, A. (2006) The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. J Mol Evol. **62**(6): p. 718-29.
221. Keeling, P.J. and Palmer, J.D. (2008) Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet. **9**(8): p. 605-18.
222. Andersson, J.O. (2005) Lateral gene transfer in eukaryotes. Cell Mol Life Sci. **62**(11): p. 1182-97.
223. Zaneveld, J.R., Nemergut, D.R., and Knight, R. (2008) Are all horizontal gene transfers created equal? Prospects for mechanism-based studies of HGT patterns. Microbiology. **154**(Pt 1): p. 1-15.
224. Anderson, P. and Roth, J. (1981) Spontaneous tandem genetic duplications in Salmonella typhimurium arise by unequal recombination between rRNA (rrn) cistrons. Proc Natl Acad Sci U S A. **78**(5): p. 3113-7.
225. Reams, A.B., Kofoid, E., Savageau, M., and Roth, J.R. (2010) Duplication frequency in a population of Salmonella enterica rapidly approaches steady state with or without recombination. Genetics. **184**(4): p. 1077-94.
226. Mira, A., Ochman, H., and Moran, N.A. (2001) Deletional bias and the evolution of bacterial genomes. Trends Genet. **17**(10): p. 589-96.
227. Andersson, J.O. and Andersson, S.G. (2001) Pseudogenes, junk DNA, and the dynamics of Rickettsia genomes. Mol Biol Evol. **18**(5): p. 829-39.
228. Bi, X. and Liu, L.F. (1994) recA-independent and recA-dependent intramolecular plasmid recombination. Differential homology requirement and distance effect. J Mol Biol. **235**(2): p. 414-23.
229. Bzymek, M. and Lovett, S.T. (2001) Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. Proc Natl Acad Sci U S A. **98**(15): p. 8319-25.
230. Koskiniemi, S. and Andersson, D.I. (2009) Translesion DNA polymerases are required for spontaneous deletion formation in Salmonella typhimurium. Proc Natl Acad Sci U S A. **106**(25): p. 10248-53.
231. Nilsson, A.I., Koskiniemi, S., Eriksson, S., Kugelberg, E., Hinton, J.C., and Andersson, D.I. (2005) Bacterial genome size reduction by experimental evolution. Proc Natl Acad Sci U S A. **102**(34): p. 12112-6.

232. Koskiniemi, S., Dynamics of the Bacterial Genome : Rates and Mechanisms of Mutation. 2010, Acta Universitatis Upsaliensis: Uppsala. p. 56.
233. Kibota, T.T. and Lynch, M. (1996) Estimate of the genomic mutation rate deleterious to overall fitness in E. coli. Nature. **381**(6584): p. 694-6.
234. Boe, L., Danielsen, M., Knudsen, S., Petersen, J.B., Maymann, J., and Jensen, P.R. (2000) The frequency of mutators in populations of Escherichia coli. Mutat Res. **448**(1): p. 47-55.
235. de Visser, J.A. (2002) The fate of microbial mutators. Microbiology. **148**(Pt 5): p. 1247-52.
236. Denamur, E. and Matic, I. (2006) Evolution of mutation rates in bacteria. Mol Microbiol. **60**(4): p. 820-7.
237. Funchain, P., Yeung, A., Stewart, J., Clendenin, W.M., and Miller, J.H. (2001) Amplification of mutator cells in a population as a result of horizontal transfer. J Bacteriol. **183**(12): p. 3737-41.
238. Fiers, W., et al. (1976) Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature. **260**(5551): p. 500-7.
239. Fleischmann, R.D., et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. **269**(5223): p. 496-512.
240. Schopf, J.W., Kudryavtsev, A.B., Agresti, D.G., Wdowiak, T.J., and Czaja, A.D. (2002) Laser--Raman imagery of Earth's earliest fossils. Nature. **416**(6876): p. 73-6.
241. Ochman, H. and Jones, I.B. (2000) Evolutionary dynamics of full genome content in Escherichia coli. Embo J. **19**(24): p. 6637-43.
242. Tenaillon, O., Skurnik, D., Picard, B., and Denamur, E. (2010) The population genetics of commensal Escherichia coli. Nat Rev Microbiol. **8**(3): p. 207-17.
243. Juhas, M., van der Meer, J.R., Gaillard, M., Harding, R.M., Hood, D.W., and Crook, D.W. (2009) Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev. **33**(2): p. 376-93.
244. Hacker, J. and Carniel, E. (2001) Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. EMBO Rep. **2**(5): p. 376-81.
245. Elena, S.F. and Lenski, R.E. (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat Rev Genet. **4**(6): p. 457-69.
246. Barrick, J.E., et al. (2009) Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature. **461**(7268): p. 1243-7.
247. Halligan, D.L. and Keightley, P.D. (2009) Spontaneous Mutation Accumulation Studies in Evolutionary Genetics. Annual Review of Ecology, Evolution, and Systematics. **40**(1): p. 151-172.
248. Sanjuan, R. (2010) Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. Philos Trans R Soc Lond B Biol Sci. **365**(1548): p. 1975-82.
249. Alberch, P. (1991) From genes to phenotype: dynamical systems and evolvability. Genetica. **84**(1): p. 5-11.
250. Pigliucci, M. (2010) Genotype-phenotype mapping and the end of the 'genes as blueprint' metaphor. Philos Trans R Soc Lond B Biol Sci. **365**(1540): p. 557-66.
251. Andersson, L. (2009) Genome-wide association analysis in domestic animals: a powerful approach for genetic dissection of trait loci. Genetica. **136**(2): p. 341-9.

252. Burke, M.K., Dunham, J.P., Shahrestani, P., Thornton, K.R., Rose, M.R., and Long, A.D. (2010) Genome-wide analysis of a long-term evolution experiment with Drosophila. Nature.
253. Ohl, M.E. and Miller, S.I. (2001) Salmonella: a model for bacterial pathogenesis. Annu Rev Med. **52**: p. 259-74.
254. Swords, W.E., Cannon, B.M., and Benjamin, W.H., Jr. (1997) Avirulence of LT2 strains of Salmonella typhimurium results from a defective rpoS gene. Infect Immun. **65**(6): p. 2451-3.
255. McClelland, M., et al. (2001) Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. Nature. **413**(6858): p. 852-6.
256. Lieb, M. and Bhagwat, A.S. (1996) Very short patch repair: reducing the cost of cytosine methylation. Mol Microbiol. **20**(3): p. 467-73.
257. Frederico, L.A., Kunkel, T.A., and Shaw, B.R. (1990) A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry. **29**(10): p. 2532-7.
258. Sharp, P.M. and Li, W.H. (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol. **4**(3): p. 222-30.
259. Wernegreen, J.J. and Funk, D.J. (2004) Mutation exposed: a neutral explanation for extreme base composition of an endosymbiont genome. J Mol Evol. **59**(6): p. 849-58.
260. Klasson, L. and Andersson, S.G. (2006) Strong asymmetric mutation bias in endosymbiont genomes coincide with loss of genes for replication restart pathways. Mol Biol Evol. **23**(5): p. 1031-9.
261. Moran, N.A. (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A. **93**(7): p. 2873-8.
262. Andersson, S.G. and Kurland, C.G. (1998) Reductive evolution of resident genomes. Trends Microbiol. **6**(7): p. 263-8.
263. Auerbach, T., Bashan, A., and Yonath, A. (2004) Ribosomal antibiotics: structural basis for resistance, synergism and selectivity. Trends Biotechnol. **22**(11): p. 570-6.
264. Berg, O.G. and Martelius, M. (1995) Synonymous substitution-rate constants in Escherichia coli and Salmonella typhimurium and their relationship to gene expression and selection pressure. J Mol Evol. **41**(4): p. 449-56.
265. Jain, R., Rivera, M.C., and Lake, J.A. (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A. **96**(7): p. 3801-6.
266. Kurland, C.G., Canback, B., and Berg, O.G. (2003) Horizontal gene transfer: a critical view. Proc Natl Acad Sci U S A. **100**(17): p. 9658-62.
267. Hegreness, M. and Kishony, R. (2007) Analysis of genetic systems using experimental evolution and whole-genome sequencing. Genome Biol. **8**(1): p. 201.
268. McDonald, M.J., Cooper, T.F., Beaumont, H.J., and Rainey, P.B. (2010) The distribution of fitness effects of new beneficial mutations in Pseudomonas fluorescens. Biol Lett.
269. McDonald, M.J., Gehrig, S.M., Meintjes, P.L., Zhang, X.X., and Rainey, P.B. (2009) Adaptive divergence in experimental populations of Pseudomonas fluorescens. IV. Genetic constraints guide evolutionary trajectories in a parallel adaptive radiation. Genetics. **183**(3): p. 1041-53.
270. Krakauer, D.C. and Plotkin, J.B. (2002) Redundancy, antiredundancy, and the robustness of genomes. Proc Natl Acad Sci U S A. **99**(3): p. 1405-9.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 611

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-132262

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2010