Chapter 1
What is Comparative Genomics?

Some scientists are visionary and can envision the theoretical foundation and experimental methodology of a new branch of science long before it takes any concrete shape. However, most scientists are just classifiers. When they see colleagues engage in novel activities such as catching flies, killing mice, chasing elephants in Africa and mounting whale specimen for museums, they would create a container labelled "zoology" and dump all these activities into it. Similarly, all those activities such as climbing trees, picking flowers, growing *Arabidopsis thaliana* and maintaining greenhouses are boxed together as botany. One former colleague of mine claimed that the only exception to this naming convension involves those studies of feces in hospitals—they are lumped together as microbiology instead of a potentially more descriptive name.

Then what is comparative genomics? Following the convention of classification, we simply define comparative genomics as the collection of all research activities that derive biological insights by comparing genomic features. A genome has many features such as the genomic sequence, strand asymmetry, genes, gene order, regulatory sequences, genomic structural landmarks that can be recognized or modified by cellular components with functional implications, etc. Comparative genomics is a branch of genomics that aims to (1) characterize the similarity and differences in genomic features and trace their origin, change and loss along different evolutionary lineages, (2) understand the evolutionary forces such as mutation, recombination, lateral gene transfer, and selection (mediated by abiotic environment such as temperature, food, and pH and biotic factors such as host, parasite, and competitors) that govern the changes of these genomic features, and (3) find out how genomic evolution can help us battle diseases by developing personalized medicine, improve environmental health, restore sustainable development, etc.

The development of comparative genomics predates the availability of genomic sequences. It has long been known that organisms are genetically related, with many homologous genes sharing similar functions among diverse organisms.

X. Xia, *Comparative Genomics*, SpringerBriefs in Genetics,
DOI: 10.1007/978-3-642-37146-2_1, © The Author(s) 2013

For example, the yeast *IRA2* gene is homologous to the human *NF1* gene, and the functional equivalence of the two genes was demonstrated by the yeast *IRA2* mutant being rescued by the human *NF1* gene (Ballester et al. 1990). This suggests the possibility that simple genomes can be used as a model to study complicated genomes. A multitude of such demonstrations of functional equivalence of homologous genes across diverse organisms has led to the dogmatic assertion that what is true in *E. coli* is also true in the elephant (attributed to Jacques Monod, Jacob 1988, p. 290).

It is the realization that what is true in *E. coli* is often not true in the elephant that has brought comparative genomics into the proper evolutionary context with the concept of phylogenetic controls. This is best illustrated by a simple example. Suppose we compare two Dodge Caravans (DCs) that are similar in functionality except that DC_1 warns the driver when it is backing towards an object behind the car while DC_2 does not. What is the structural basis of this warning function? Nearly all structural elements in DC_1 have their "homologues" in DC_2 except for the four sensors on the rear bumper of DC_1. This would lead us to quickly hypothesize that the four sensors are associated with the warning function, which turns out to be true. Now if we replace DC_2 with a baby stroller, then the comparison will be quite difficult because a stroller and a DC differ structurally in numerous ways and any structural difference could be responsible for the warning function. We may mistakenly hypothesize that the rear lights or the rear window defroster in DC_1, which are all missing in the stroller, may be responsible for the warning function. To test the hypotheses, we would destroy the rear lights, the rear window defroster, etc., one by one, but will get nothing but negative results. What could be even worse is that, when destroying the rear lights, we accidentally destroy a part of the electric system in such a way that the warning function is lost, which would mislead us to conclude that the rear lights are indeed part of the structural basis responsible for the warning function—an "experimentally substantiated" yet wrong conclusion. A claim that what is true in *E. coli* is also true in the elephant is equivalent to a claim that what is true in a stroller is also true in a DC. It will take comparative genomics out of its proper conceptual framework in evolutionary biology and render it inefficient to address biological questions.

Let's take a biologically more relevant example involving *Shigella flexneri* and *E. coli* (Sansonetti et al. 1982a, b). *Shigella* strains cause shigellosis, whereas strains of *Escherichia coli* are generally avirulent. What is responsible for the difference? Nuclear genomes are similar between *Shigella* and *E. coli*, which led scientists to focus on a plasmid that is present in the pathogenic *Shigella* strains but absent in the avirulent *E. coli* strains. The pathogenic *Shigella* strains become avirulent when the plasmid is taken away, and originally avirulent strains of *E. coli* gains virulence after acquiring the plasmid. This led quickly to the conclusion that the plasmid is largely responsible for shigellosis. Had one compared between *S. flexneri* and *Saccharomyces cerevisiae,* one would need to hypothesize that any one of the thousands of genes in *S. cerevisiae* not shared by *S. flexneri* could be a causal factor. Filtering through these thousands

of possibilities would take forever even if we do not consider gene combinations as causal factors.

In this chapter I will detail a few typical comparative genomic studies so that we can develop an intuitive appreciation of what is hidden in the box labelled "comparative genomics". These studies involve biological problems that can be addressed by comparing two genomics as well as problems that would require more than two genomes to reach a solution. The similarities among these studies are summarized at the end to highlight essential elements in a comparative genomics study.

Genomic Comparison Between *Helicobacter pylori* and its Relatives

Problems and Hypotheses

Helicobacter pylori is a human pathogen causing gastric and duodenal ulcers and gastric cancer (Hamajima et al. 2004; Hunt 2004; Menaker et al. 2004; Siavoshi et al. 2004). It is an acid-resistant neutralophile (Bauerfeind et al. 1997; Rektorschek et al. 2000; Sachs et al. 1996; Scott et al. 2002) capable of surviving for at least 3 h at pH = 1 with urea (Stingl et al. 2001) and maintaining a nearly neutral cytoplasmic pH between pH 3.0 and 7.0 (Matin et al. 1996; Scott et al. 2002). In the presence of urea, *H. pylori* can accomplish its cytoplasmic pH homeostasis down to an external pH of 1.2 (Stingl et al. 2002b). These properties allow it to survive and reproduce in the human stomach where the gastric fluid has a pH averaging about 1.4 over a 24-h period (Sachs et al. 2003).

The buffering action of the gastric epithelium and limited acid diffusion through the gastric mucus were previously thought to protect the bacterium against stomach acidity, but both empirical studies (Allen et al. 1993) and theoretical modeling (Engel et al. 1984) have suggested that the protection is rather limited (Matin et al. 1996; Sachs 2003 #14944). Recently it has also been shown that mucus does not hinder proton diffusion and a trans-mucus pH gradient is abolished when the luminal pH drops to <2.5 (Baumgartner and Montrose 2004). It is therefore necessary for *H. pylori* to have acid-resisting mechanisms to colonize the gastric mucosa successfully (Sachs et al. 2003).

H. pylori has evolved two mechanisms protecting itself against the acidic environment in the mammalian stomach. The first, schematically illustrated in Fig. 1.1, involves the urease gene cluster *ureABIEFGH*. The constitutively expressed cytoplasmic urease consists of four heterodimer each with two subunits coded by *ureA* and *ureB*, respectively. It catalyzes urea to generate $2NH_3 + CO_2$ to buffer against the H^+ influx into either the periplasm or the cytoplasm (Mobley et al. 1991; Rektorschek et al. 2000; Sachs et al. 2003; Stingl et al. 2002a) and to facilitate the extrusion of H^+ from the cytoplasm in the form of NH_4^+

(Stingl et al. 2002a). However, urease is an apoenzyme requiring a nickel to be active. The *ureEFGH* gene cluster, whose expression is acid-induced, codes for nickel-sequestrating proteins that insert nickel into the urease, leading to increased and sustained urease activity (Sachs et al. 2003; Wen et al. 2003; Williams et al. 1996).

The urease, once activated, naturally needs a constant supply of urea as its substrate, and the cell has two sources of urea supply, one intrinsic and one extrinsic (Fig. 1.1). The extrinsic source refers to urea present in saliva and stomach fluid. The exposure of *H. pylori* to gastric acid results in a large increase in urea influx into the cell due to the pH-gating of the urea channel protein *UreI* (Bury-Mone et al. 2001; Weeks et al. 2000). The intrinsic source comes from efficient conversion of arginine to urea in the cytoplasm by the highly expressed arginase in *H. pylori* (Mendz and Hazell 1996). For this reason, arginine is underused, but lysine is overused, in *H. pylori* proteins (Xia and Palidwor 2005).

The second acid-resistant mechanism in *H. pylori* is the restriction of acute proton entry across its membranes by having a high frequency of positively charged amino acids and consequent high pI (isoelectric point) values in the inner and outer membrane proteins (Sachs et al. 2003; Scott et al. 1998; Valenzuela et al. 2003). This is supported by recent discovery of a basic proteome (Tomb et al. 1997), a set of basic membrane proteins (Baik et al. 2004) in *H. pylori*, and an extensive genomic analysis (Xia and Palidwor 2005) testing the adaptation, preadaptation and exaptation hypotheses concerning the overuse of lysine residues in *H. pylori* proteins. The mechanism gained functional importance after the discovery that urease-negative *H. pylori* can colonize the acidic gastric environment and cause gastric ulcers in Mongolian gerbils (Mine et al. 2005).

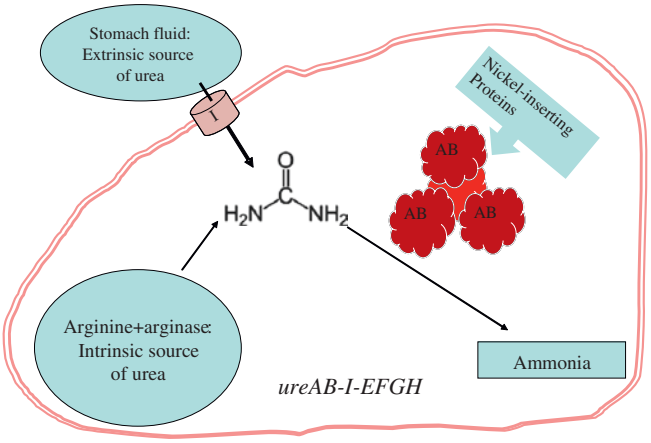

Fig. 1.1 Schematic illustration of the acid-resistance mechanisms in *H. pylori* mediated by genes in the urease gene cluster *ureAB-I-EFGH*

Given that *H. pylori* has many Lys-rich proteins with high pI values relative to other bacterial species that do not live in acidic environment, one is naturally tempted to conclude that the high pI values in the *H. pylori* proteins represent an adaptation to the acidic environment. However, there are at least four possible hypotheses for the origin of the basic proteome in *H. pylori* (Xia 2007a, Chap. 10).

The first hypothesis states that *H. pylori* would benefit from positively charged proteins (especially membrane proteins) to alleviate the influx of H^+ into cytoplasm. This hypothesis is known as the acid-adaptation hypothesis (Xia and Palidwor 2005), i.e., *H. pylori* acquired its high-pI proteins as an adaptation in response to selection imposed by the acidic environment.

The second hypothesis argues that parasitic bacterial genomes typically evolve towards AT-richness because spontaneous mutations are generally AT-biased according to comparisons between pseudogenes and their functional counterparts (Gojobori et al. 1982; Li 1983; Li et al. 1981) and the discovery of the prevalence of spontaneous C $\rightarrow$ T/U deamination (Frederico et al. 1990, 1993; Lindahl 1993). All known parasitic bacterial genomes are AT-rich. *H. pylori* has a relatively AT-rich genome, e.g., the genomic GC% of *H. pylori* 26695 is only 38 %, in contrast to the genomic GC% of 50 % in *E. coli* substr DH10B. The AT-richness would lead to an increase in A-rich codons such as the lysine codon AAA and AAG and a consequent increase in lysine usage and protein pI. Because *H. pylori* and its sibling species are all parasites, their most recent common ancestor might have already practiced parasitism, acquired AT-richness and increased frequency of lysine codons before it became a parasite in the mammalian stomach. Therefore, an overrepresentation of lysine residues in its proteins, if beneficial for acid-resistance, would represent an exaptation, i.e., the process in which an originally neutral trait has subsequently acquired a beneficial function. A well known example of exaptation is the brain-specific RNA gene BC200 resulting from the exaptation of a presumably neutral SINE repeat (Smit 1999).

The third hypothesis states that nucleotide C is rare in eukaryotic cells and a eukaryotic parasite should therefore minimize the usage of C as a building block for its RNA and DNA. CTP concentration is much lower than the other three nucleotides chick fibroblast cells (Colby and Edlin 1970) and in mouse 3T3 cells (Weber and Edlin 1971), suggesting the generality of C limitation. Consistent with the suggestion, the protozoan parasite, *Trypanosoma brucei*, maintains its *de novo* synthesis pathway for CTP and inhibiting its CTP synthetase effectively eradicates the parasite population in the host (Hofer et al. 2001). In contrast, the parasite does not have *de novo* synthesis pathways for purines, suggesting that the parasite can obtain the purines by its salvage pathway. This suggests that little CTP can be salvaged from the host. The relevance of these observations is highlighted by the fact that *H. pylori* maintains an active biosynthesis pathway, and a much less active salvage pathway, for pyrimidine nucleotides (Mendz et al. 1994). Thus, it might be evolutionarily beneficial for a mammalian parasite or symbiont to minimize the use of CTP in its DNA in building its genomes and in transcription (Rocha and Danchin 2002; Xia 1996).

Minimizing C in an organism with a DNA genome has the necessary consequence of reduced G, with a consequent increase in A and T. This will also contribute to increase AT and increased lysine codon. Thus, lysine overuse represents a secondary consequence of an adaptation to a C-rare environment, but it predisposed the organism to tolerate an acidic environment. Such a mechanism is called preadaptation, i.e., a trait originally selected for one function but that subsequently gained a different function beneficial to the carrier of the trait. An often cited example of preadaptation is the rudimentary feather that presumably has been selected for thermoregulation in nonavian dinosaurs but preadapted their carriers to subsequent evolution of flight.

The fourth hypothesis is more complicated. A protein in a solution with a pH equal to the protein pI is not charged. If highly expressed proteins happen to have their pI equal to the cytoplasmic pH, then there is no electrostatic repulsion among these proteins when they are mass-produced. Such proteins will have low solubility and tend to aggregate and precipitate, which is often harmful to the cell. The "amyloid precursor protein" causing Alzheimer disease and the prion protein causing the mad cow disease are examples of the undesirable protein aggregation and precipitation. Take *E. coli* for example. Its intestinal environment has pH close to 9 and it can maintain its optimal growth at external pH as high as 8.8 (Zilberstein et al. 1980, 1982). Its intracellular pH is regulated in the range of 7.4–7.8 at external pH range of 5.5–9 (Slonczewski et al. 1981). Thus, in its intestinal environment, its internal pH should be around 7.8 and we should expect *E. coli* to avoid having proteins with their pI values around 7.8. This is true (Fig. 1.2). Avoiding proteins with pI equal to intracellular pH appears to be universal among unicellular organisms.

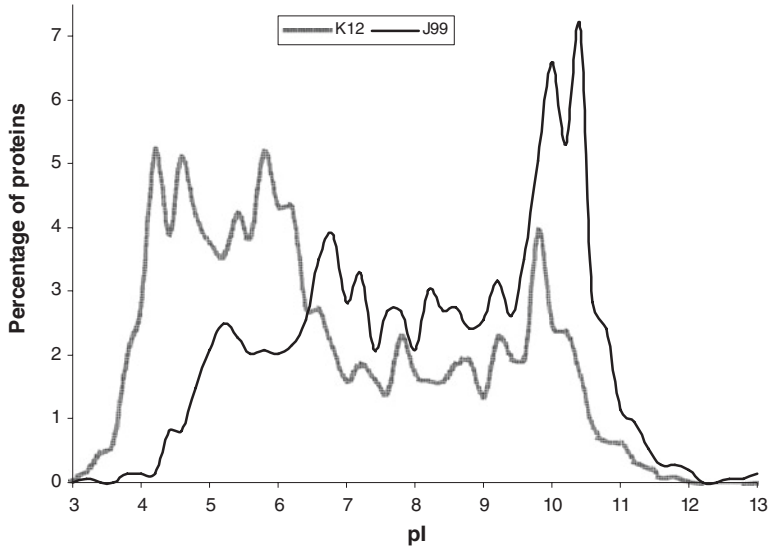

Fig. 1.2 Genomic pI profiling for *E. coli* K12 and *H. pylori* J99

Given the avoidance of proteins with pI equal to intracellular pH, we would expect mass-produced proteins in the gastric *H. pylori*, whose intracellular pH is around 5, to avoid having pI ≈ 5. This prediction is substantiated (Fig. 1.2). The pronounced peak of proteins with pI in the range of 4–6 in *E. coli* is missing in *H. pylori.* Instead, proteins with pI in the range of 10–11 are over-represented in *H. pylori* (Fig. 1.2)

One might ask why *H. pylori* proteins cannot lower their pI to the range of 0–3 to avoid precipitation. This would be practically difficult because the proteins would require an excessively large number of GAN to code for Glu and Asp. It is extremely rare to have proteins with a pI smaller than 3.

Testing the Hypotheses by Comparative Genomics

The first three hypotheses have been tested before and the second and third hypotheses were found to be inconsistent with the empirical data (Xia and Palidwor 2005). Here we illustrate how to discriminate between the first and the last hypothesis, i.e., whether the increase in protein pI is for alleviating the influx of protons, referred hereafter as AAH (acidity adaptation hypothesis), or for avoiding protein precipitation, referred to hereafter as precipitation avoidance hypothesis (PAH).

The two hypotheses have different predictions. AAH predicts that it is those membrane proteins that tend to gain a higher pI. In contrast, PAH predicts that the overrepresentation of the high-pI proteins in *H. pylori* is due to the necessity of mass-produced proteins to have their pI shifting away from the cytoplasmic pH to avoid protein precipitation. Specifically, the shifting of the pI distribution to the right in Fig. 1.2 is due to mass-produced proteins increasing their pI to shift their pI away from the cytoplasmic pH.

To test the AAH prediction, one needs to separate proteins into membrane proteins and cytoplasmic proteins. The main difficulty is that membrane proteins are difficult to separate and identify and only 34 membrane proteins have been identified in *H. pylori* (Baik et al. 2004). These proteins do exhibit a significantly higher pI than the rest of the *H. pylori* proteins (Xia and Palidwor 2005). Furthermore, one can use an excellent bioinformatic tool, pSort (Gardy et al. 2003; Nakai and Horton 1999), for protein cellular localization. Those proteins identified to be localized in cytoplasmic membrane, outer membrane and periplasmic space all have their mean pI values highly significantly higher than those localized in cytoplasm.

Are these results in favor of AAH? Not necessarily. Although AAH predicts that membrane proteins with a high pI would contribute to a positively charged shell alleviating the influx of protons into the cell, the result cannot be claimed to support, or even be consistent with, AAH. The reason is that membrane proteins in general have higher pI than cytoplasmic proteins, even for bacterial species that do not live in an acidic environment. What is important is to find bacterial species that

are phylogenetically closely related to *H. pylori*, but do not exhibit acid resistance. Such species could be *H. hepaticus* or *Campylobacter* species, and are generally referred to as phylogenetic controls (because they and *H. pylori* were identical when we trace them back in time to their common ancestor). If we can find such pairs of sister species, with one living in acidic environment and the other not, and if we consistently find the former to have significantly elevated pI in their membrane proteins than the latter, then we can claim that the result supports, or at least is consistent with, the prediction of AAH. What is exciting about comparative genomics today is that, once we are equipped with the conceptual framework above, it takes only a few hours to complete the analysis by using publicly available genomic databases and software packages such as DAMBE (Xia 2001; Xia and Xie 2001). The empirical result, as you can verify by yourself, is consistent with the prediction of AAH. Both *H. pylori* and *H. hepaticus* have their membrane proteins with significantly high pI than cytoplasmic proteins, but the difference is much greater in *H. pylori* than in *H. hepaticus.*

Testing the prediction of PAH (i.e., mass-produced *H. pylori* proteins should evolve to have increased pI values away from cytoplasmic pH around 5) seems straightforward at first. We need to obtain pI and protein expression for each gene. Although we do not have reliable protein expression data in *H. pylori* at the moment, the difficulty can somewhat overcome by using indices of codon usage bias as a proxy of gene expression (Xia 1998a, 2007b, 2008). Similarly, although we do not have experimentally determined pI for each protein, theoretically derived pI based on amino acid composition (Xia 2007a, pp. 207–212) represents a good approximation. Now suppose we have protein pI and protein expression (designated by E). It seems that the prediction of PAH can be reduced to a statement that pI and E are positively correlated because high-E proteins should increase their pI away from the cytoplasmic pH. Is this inference correct? Now suppose you found that pI and E are indeed positively correlated, will you conclude that PAH is supported? Alternatively, if you found pI and E are negatively correlated, will you reject PAH?

It turns out that you cannot say much about PAH based on the correlation between pI and E. A positive correlation is expected if the data include many highly expressed DNA-binding or RNA-binding proteins because these proteins all tend to have a DNA/RNA-binding domain which is rich in positively charged amino acids (Recall that the backbone of RNA and DNA are negatively charged and a positively charged protein domain facilitates the binding to RNA and DNA). This would result in a positive correlation between pI and E which has nothing to do with PAH.

You may also get a negative correlation between pI and E for the following reason. Differences in pI among proteins mainly depend on the relative number of the strongly acidic amino acid residues such as Asp, and Glu and the strongly basic amino acid residues such as Arg, Lys, and His. The positively charged amino acids, however, are generally more energetically expensive to make in bacterial species (Akashi and Gojobori 2002). For example, the total high-energy ~P required to make Asp and Glu are 12.7 and 15.3, respectively, which are quite

close to the cost of making the smallest amino acids Gly and Ala. In contrast, the energetic costs for making His, Lys and Arg are 38.3, 30.3 and 27.3, respectively. Highly expressed proteins tend to use cheap amino acids and avoid the expensive Arg, Lys and His in almost all bacterial species, resulting in highly expressed proteins (except for those ribosomal proteins) to have a low pI and a consequent negative correlation between pI and E. Thus, a negative correlation between pI and E again could have nothing to do with PAH.

Thus, to properly test the prediction of PAH, comparative genomics involving sister species (e.g., between *H. pylori* and *H. hepaticus*) is again necessary. Suppose we found 500 *H. hepaticus* proteins that have pI around 5 and are homologous to those in *H. pylori*. Also suppose that, among the 500 proteins, 200 of them are highly expressed and 200 are lowly expressed. If the 200 highly expressed proteins in *H. pylori* have all shifted their pI away from the cytoplasmic pH of about 5, whereas the 200 lowly expressed proteins have their pI hardly changed relative to their *H. hepaticus* homologues, then we can claim that result is consistent with PAH. Of course, this represents only one of possible ways to test the prediction from PAH.

Genomic Comparison Between HIV-1 and HTLV-1

Because viruses use the host translational machinery to translate their own mRNA, their codon usage is under selective pressure to adapt to the host tRNA pool (Sharp and Li 1987). In RNA viruses in general and Human Immunodeficiency Virus 1 (HIV-1) in particular, adaptation to the host is poor despite this selection (Bahir et al. 2009; van Weringh et al. 2011), in contrast to the codon-anticodon adaptation documented in bacterial genomes (Gouy and Gautier 1982; Ikemura 1981a, 1992; Xia 1998a) as well as in mitochondrial genomes in vertebrates (Xia 2005; Xia et al. 2007) and fungi (Carullo and Xia 2008; Xia 2008). For example, according to a recent compilation of tRNAs in human genome (Chan and Lowe 2009), the AUC codon can be translated by 17 $tRNA^{Ile}$ species, i.e., 14 $tRNA^{Ile/IAU}$ and 3 $tRNA^{Ile/GAU}$, AUU can be translated by 14 $tRNA^{Ile/IAU}$ species, whereas AUA can be translated by only 5 $tRNA^{Ile/UAU}$ species. In agreement with this, human genes code Ile mostly by AUC and least by AUA. In contrast, HIV-1 genes code Ile mostly by AUA and least by AUC (Haas et al. 1996; Nakamura et al. 2000). The poor codon adaptation of HIV-1 reduces the translation efficiency of HIV-1 genes. Modifying HIV-1 codon usage according to host codon usage has been shown to increase the production of viral proteins (Haas et al. 1996; Ngumbela et al. 2008).

The A-biased mutation hypothesis has been proposed to explain the poor concordance between HIV-1 and host codon usage (Jenkins and Holmes 2003). The A-bias is mediated by the error prone reverse transcriptase (Martinez et al. 1994; Vartanian et al. 2002) and the human APOBEC3 protein (Yu et al. 2004). The frequency of A can reach up to 40 % in some HIV-1 genomes (Vartanian et al. 2002),

resulting in a preponderance of A-ending codons which are typically rarely used in the host genes (Kypr and Mrazek 1987; Sharp 1986). While there have been claims that the A-richness in a parasitic or symbiotic genome may confer some selective advantage (Keating et al. 2009; Xia 1996), further empirical substantiation is required. In short, although avoiding A-ending codons will lead to better codon-anticodon adaptation, strongly A-biased mutations lead to an over-representation of A-ending codons in HIV-1 genes, disrupting codon-anticodon adaptation.

How can we test this mutation hypothesis? If we can find pairs of sister species that differ much in mutation rate, then we can test the hypothesis by checking if the species with higher mutation rate tend to have poorer codon-anticodon adaptation than its sister species with lower mutation rate. HTLV-1 could serve as a sister species for the HIV/SIV lineage. Both HTLV-1 and HIV-1 are retroviruses with RNA genomes and both infect the same type of host cell, i.e., human CD4 + T cells (Rimsky et al. 1988). The two viruses are therefore subject to the same selective pressures on codon usage by the host tRNA pool. However, HTLV-1 is exceptional in that it does not have a strong A-biased mutation spectrum (Van Dooren et al. 2004; van Hemert and Berkhout 1995). HTLV-1 relies for the most part on the host polymerase to replicate through clonal expansion of infected cells rather than undergoing iterative replication cycles like HIV-1 (Strebel 2005). The substitution rate of HTLV-1 is consequently lower, about 5.2×10^{-6} substitutions/site/year (Hanada et al. 2004; Van Dooren et al. 2004), in contrast to that of HIV-1 at 2.5×10^{-3} substitutions/site/year (Hanada et al. 2004). Codon-anticodon adaptation is less likely to be disrupted by mutation in HTLV-1 than in HIV-1. Thus we predict that HTLV-1 coding sequences should exhibit better codon-anticodon adaptation.

Codon-anticodon adaptation can be measured by the correlation in RSCU (Sharp and Li 1987) between the host and the parasite. RSCU is a normalized index of codon usage (Sharp and Li 1987). It has a value of zero for unused synonymous codons, a value of one for equally used synonymous codons and a maximum of n, where n is the number of synonymous codons in the codon family. Thus, the prediction of the mutation hypothesis is that the correlation in RSCU between human and HTLV-1 genes should be greater than that between human and HIV-1 genes.

The correlation in RSCU between human and HIV-1 genes is poor (Pearson $r = -0.1470$, $p = 0.2665$; Spearman $r = 0.1829$, $p = 0.1657$). In contrast, the positive correlation in RSCU between HTLV-1 and human genes is highly significant (Pearson $r = 0.4982$, $p < 0.0001$, Spearman $r = 0.4688$, $p = 0.0002$). Such results are consistent with the mutation hypothesis.

The real scenario of codon-anticodon adaptation in HIV-1 is much more complicated, of course. In particular, the early gene and late genes in HIV-1 may be translated in different tRNA pools and subject to different selection for codon-anticodon adaptation (van Weringh et al. 2011). HIV-1 has recently been shown to package non-lysyl tRNAs in addition to the $tRNA^{Lys}$ needed for priming reverse-transcription and integration of the HIV-1 genome. In particular, tRNAs decoding A-ending codons, required for the expression of HIV's A-rich genome, are highly

enriched. Because the affinity of Gag-Pol for all tRNAs is non-specific, HIV packaging is most likely passive and reflects the tRNA pool at the time of viral particle formation. Codon usage of HIV-1 early genes is similar to that of highly expressed host genes, but codon usage of HIV-1 late genes were better adapted to the selectively enriched tRNA pool, suggesting that alterations in the tRNA pool are induced late in viral infection. If HIV-1 genes are adapting to an altered tRNA pool, codon adaptation of HIV-1 may be better than previously thought (van Weringh et al. 2011).

Genomic Comparison Among *Mycoplasma* Species

CpG deficiency has been documented in a large number of genomes covering a wide taxonomic distribution (Cardon et al. 1994; Josse et al. 1961; Karlin and Burge 1995; Karlin and Mrazek 1996; Nussinov 1984). DNA methylation is one of the many hypotheses proposed to explain differential CpG deficiency in different genomes (Bestor and Coxon 1993; Rideout et al. 1990; Sved and Bird 1990). It features a plausible mechanism as follows. Methyltransferases in many species, especially those in vertebrates, appear to methylate specifically the cytosine in CpG dinucleotides, and the methylated cytosine is prone to mutate to thymine by spontaneous deamination (Frederico et al. 1990; Lindahl 1993). This implies that CpG would gradually decay into TpG and CpA, leading to CpG deficiency and reduced genomic GC%. Different genomes may differ in CpG deficiency because they differ in methylation activities, with genomes having high methylation activities exhibiting stronger CpG deficiency than genomes with little or no methylation activity.

In spite of its plausibility, the methylation-deamination hypothesis has several major empirical difficulties (Cardon et al. 1994), especially in recent years with genome-based analysis (Goto et al. 2000). For example, *Mycoplasma genitalium* does not seem to have any methyltransferase and exhibits no methylation activity, yet its genome shows a severe CpG deficiency. Therefore, the CpG deficiency in *M. genitalium*, according to the critics of the methylation-deamination hypothesis, must be due to factors other than DNA methylation.

A related species, *M. pneumoniae*, also devoid of any DNA methyltransferase, has a genome that is not deficient in CpG. Given the difference in CpG deficiency between the two Mycoplasma species, the methylation hypothesis would have predicted that the *M. genitalium* genome is more methylated than the *M. pneumoniae* genome, which is not true as neither has a methyltransferase. Thus, the methylation hypothesis does not seem to have any explanatory power to account for the variation in CpG deficiency, at least in the Mycoplasma species.

These criticisms are derived from phylogeny-free reasoning. When phylogeny-based comparisons are made, the Mycoplasma genomes become quite consistent with the methylation hypothesis (Xia 2003). First, several lines of evidence suggest that the common ancestor of *M. genitalium* and

M. pneumoniae have methyltransferases methylating C in CpG dinucleotides, and should have evolved strong CpG deficiency and low genomic GC% as a result of the specific DNA methylation. Methylated m^5C exists in the DNA of a close relative, *Mycoplasma hyorhinis* (Razin and Razin 1980), suggesting the existence of methyltransferases in *M. hyorhinis*. Methyltransferases are also present in *Mycoplasma pulmonis* which contains at least four CpG-specific methyltransferase genes (Chambaud et al. 2001). Methyltransferases are also found in all surveyed species of a related genus, Spiroplasma (Nur et al. 1985). These lines of evidence suggest that methyltransferases are present in the ancestors of *M. genitalium* and *M. pneumoniae.*

Second, the methyltransferase-encoding *M. pulmonis* genome is even more deficient in CpG and lower in genomic GC% than *M. genitalium* or *M. pneumoniae*, consistent with the methylation hypothesis (Fig. 1.3). It is now easy to understand that, after the loss of methyltransferase in the ancestor of *M. genitalium* and *M. pneumoniae* (Fig. 1.3), both genomes would begin to accumulate CpG dinucleotides and increase their genomic GC%. However, the evolutionary rate is much faster in *M. pneumoniae* than in *M. genitanlium* based on the comparison of a large number of protein-coding genes (Xia 2003). So *M. pneumoniae* regained CpG dinucleotide and genomic GC% much faster than *M. genitalium.* In short, the Mycoplasma data that originally seem to contradict the methylation hypothesis actually provide strong support for the methylation hypothesis when phylogeny-based genomic comparisons are made.

One might note that *Ureaplasma urealyticum* in Fig. 1.3 is not deficient in CpG because its $P_{CpG}/(P_CP_G)$ ratio is close to 1, yet its genomic GC% is the lowest. Has its low genomic GC% resulted from CpG-specific DNA methylation? If yes, then why doesn't the genome exhibit CpG deficiency? It turns out that *U. urealyticum* has C-specific, but not CpG-specific, methyltransferase, i.e., the genome of *U. urealyticum* is therefore expected to have low CG % (because of the methylation-mediated C → T mutation) but not a low $P_{CpG}/(P_CP_G)$ ratio. The methyltransferase gene from *U. urealyticum* is not homologous to that from *M. pulmonis*.

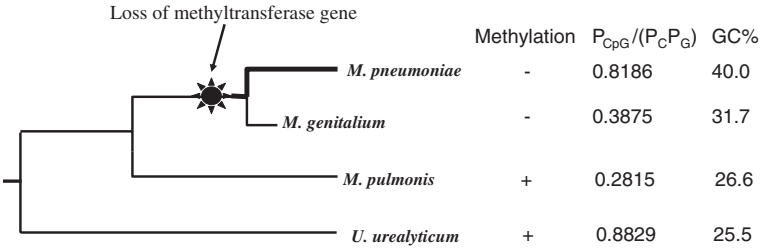

Fig. 1.3 Phylogenetic tree of *Mycoplasma pneumoniae*, *M. genitalium*, and their relatives, together with the presence (+) or absence (−) of CpG-specific methylation, PCpG/(PCPG) as a measure of CpG deficiency, and genomic GC%. *M. pneumoniae* evolves faster and has a longer branch than *M. genitalium*

We have seen how phylogeny can help us in evolutionary inference, and most comparative genomic studies represent phylogeny-based inference. It is appropriate here to introduce a few phylogeny-related terms. Most published phylogenies are build from molecular sequence data, i.e., multiple alignment of homologous sequences. Sequence similarity can arise in two ways, one from convergence (i.e., similarity gained from independent evolution), and the other from coancestry. Coancestral sequences are homologous, and can be divided into orthologous and paralogous sequences. Two or more duplicated genes within one genome represent a special form of homology and are termed paralogous genes. Two or more homologous genes that are related by inheritance are orthologous. Genes acquired through horizontal gene transfer are neither orthologous nor paralogous. Species phylogeny ideally should be built only from orthologous genes.

Genomic Comparison to Characterize Changes in tRNA and Codon-Anticodon Adaptation

Ever since the empirical documentation of the correlation between codon usage and tRNA abundance (Ikemura 1981a, b, 1982, 1992), studies on codon-anticodon adaptation have progressed in theoretical elaboration (Bulmer 1987, 1991; Higgs and Ran 2008; Jia and Higgs 2008; Palidwor et al. 2010; Xia 1998a, 2008), in critical tests of alternative theoretical predictions (Carullo and Xia 2008; Plotkin and Kudla 2010; Plotkin et al. 2004; van Weringh et al. 2011; Xia 1996, 2005) and in formulation and improvement of various codon usage indices to characterize codon usage bias (Sharp and Li 1987; Wright 1990; Xia 2007b). Here I present two examples in which a gain/loss of a tRNA gene or a change in genetic code lead to significant changes in codon usage.

The Met Codon Family

An evolutionary change in tRNA composition or relative abundance is expected to alter codon-anticodon adaptation. This is not controversial theoretically. However, how fast can an alternation in tRNA lead to consequent changes in codon-anticodon adaptation? Can the cause-effect relationship be demonstrated with empirical data? Changes in $tRNA^{Met}$ genes (where Met is the amino acid carried by the tRNA) in animal mitochondrial DNA (mtDNA) paved the way for such a demonstration (Xia 2012b).

In MtDNA of most animal species, Met is coded by AUA and AUG codons. In some animal species, e.g., vertebrates, these two codons are translated by a single $tRNA^{Met/CAU}$ species (where CAU is the anticodon in the 5' to 3' orientation) with a modified C (i.e., f^5C) at the first anticodon position (Grosjean et al. 2010) to allow C/A pairing. In other animal species, e.g., tunicates, an additional $tRNA^{Met/UAU}$ gene is present in the mtDNA. One would expect that, when

$tRNA^{Met/UAU}$ is absent, Met should be preferably coded by AUG with a reduced AUA usage. The gain of $tRNA^{Met/UAU}$ would favor more Met to be coded by AUA. Can such a prediction be empirically substantiated?

MtDNA in bivalve species have two $tRNA^{Met}$ genes. In some bivalve species (e.g., *Acanthocardia tuberculata, Crassostrea gigas, C. virginica, Hiatella arctica, Placopecten magellanicus,* and *Venerupis philippinarum*), both $tRNA^{Met}$ genes have a CAU anticodon forming Watson–Crick base pair with codon AUG. In some other bivalve species (e.g., *Mytilus edulis, Mytilus galloprovincialis,* and *Mytilus trossulus*), one $tRNA^{Met}$ has a CAU anticodon and the other has a UAU anticodon forming Watson–Crick base pair with the AUA codon. One would predict that the latter should be more likely to code Met by AUA than the former, i.e., the proportion of AUA codon within the AUR codon family, designated P_{AUA}, should be greater in the latter with both a $tRNA^{Met/CAU}$ and a $tRNA^{Met/UAU}$ gene than in the former with a single $tRNA^{Met/CAU}$ gene in the mtDNA (Xia et al. 2007).

To test the prediction, I will use P_{UUA} (the proportion of UUA codon in the UUR codon family) as a reference control to test the prediction that, at the same P_{UUA} level, P_{AUA} in the three Mytilus mtDNA with both a $tRNA^{Met/CAU}$ and a $tRNA^{Met/UAU}$ gene is higher than that in the six bivalve species without a $tRNA^{Met/UAU}$ gene. This is supported by empirical evidence (ANCOVA test, $p = 0.0111$, Fig. 1.4a). Thus, the presence of $tRNA^{Met/UAU}$ increases AUA usage significantly.

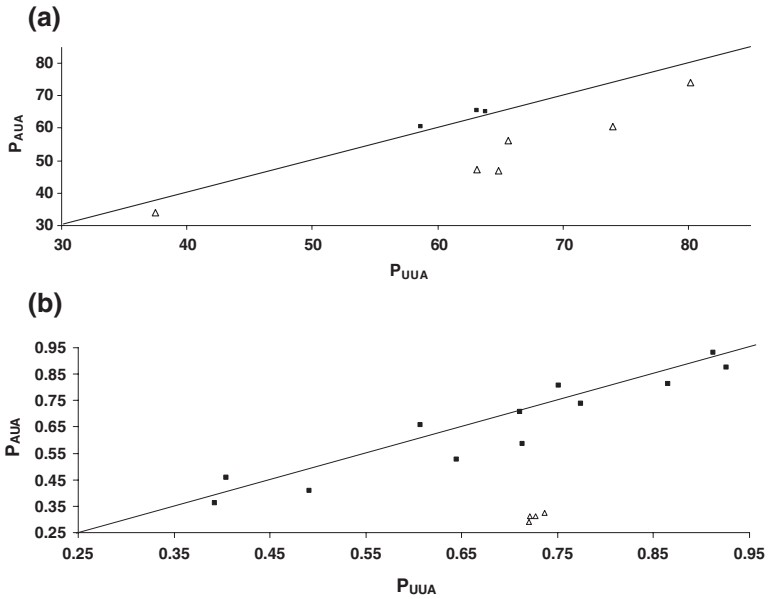

Fig. 1.4 Relationship between PAUA and PUUA, highlighting the observation that PAUA is greater when both a tRNAMet/CAU and a tRNAMet/UAU are present than when only tRNAMet/CAU is present in the mtDNA, for bivalve species (**a**) and chordate species (**b**). The filled squares are for mtDNA containing both tRNAMet/CAU and tRNAMet/UAU genes, and the open triangles are for mtDNA without a tRNAMet/UAU gene

A similar comparison can be performed between the urochordates (tunicates, with both $tRNA^{Met/CAU}$ and $tRNA^{Met/UAU}$ genes in their mtDNA) and cephalochordates (lancelets, with only a $tRNA^{Met/CAU}$ gene in their mtDNA). Figure 1.4b shows that P_{AUA} is much smaller in lancelets than in tunicates at the same P_{UUA} level. Thus, AUA usage is consistently increased by the gain of a $tRNA^{Met/UAU}$ gene (or consistently decreased by the loss of a $tRNA^{Met/UAU}$ gene) in animal mtDNA.

A gain of a $tRNA^{Met/UAU}$ gene is also associated with a surplus of AUG → AUA substitutions in animal mitochondrial coding sequences (results not shown). Similar associations can also be observed with other gain/loss of tRNA genes in animal mitochondrial. In contrast, a gain/loss of tRNA genes in plant mtDNA appears to have little effect on nucleotide substitutions or codon usage, presumably because such gain/loss events do not significantly alter the tRNA pool in plant cells where nuclear tRNAs are mass-imported into plant mitochondria.

UGA Codon, CGN Codon for Arg and the Expanded Wobble Hypothesis

The number of distinct tRNA species is invariably fewer than the number of sense codons, leading to the formulation of the original wobble hypothesis (Crick 1966). Figure 1.5 depicts the extended codon-anticodon base pairs as well as the subscripted numbering system used for codon-anticodon base pairs (Xia 2013). Note that the anticodon sites are denoted by Roman numerals and tho the codon sites by Arabic numerals (Fig. 1.5).

The wobble hypothesis explains why $tRNA^{Ile/IAU}$, where I in IAU is inosine derived from A, is able to translate all three Ile codons (AUC, AUU and AUA), why a tRNA with a G_I can translate Y-ending codons (where Y stands for C or U), and why a tRNA with a U_I can translate R-ending codons (where R stands for A or G). The hypothesis also explains the lack of A_I in tRNA genes for decoding 2-fold Y-ending codon family because such a tRNA, when its A_I is modified to I_I, would mis-read the near cognate R-ending codons. One might note that all base-pairs involve a purine and a pyrimidine except for the I/A pair which is a bulky purine-purine pair that may lead to inefficient translation (Curran 1995).

Wobble pairing reduces the number of tRNAs needed for translation and simplifies the translation machinery. Few organisms can afford the luxury of having different gene products doing the same task. As an example of parsimonious tRNA usage, the Y-ending codons, be they in 2-fold or 4-fold codon families, are decoded by tRNAs with either a I_I or a G_I, but never both. This rule is obeyed in all three kingdoms of life. Almost all 4-fold codon families in *Mycoplasma pulmonis* (including the Ser UCN codon family and Leu CUN codon family, where N is any nucleotide) are decoded by a single tRNA species with a U_I, except for the Thr ACN and Arg CGN codon families which are each decoded by two tRNA species, one with a U_I and other with a G_I. The most dramatic simplification of tRNome is observed in metazoan mitochondria, e.g., vertebrate mitochondrial genomes which

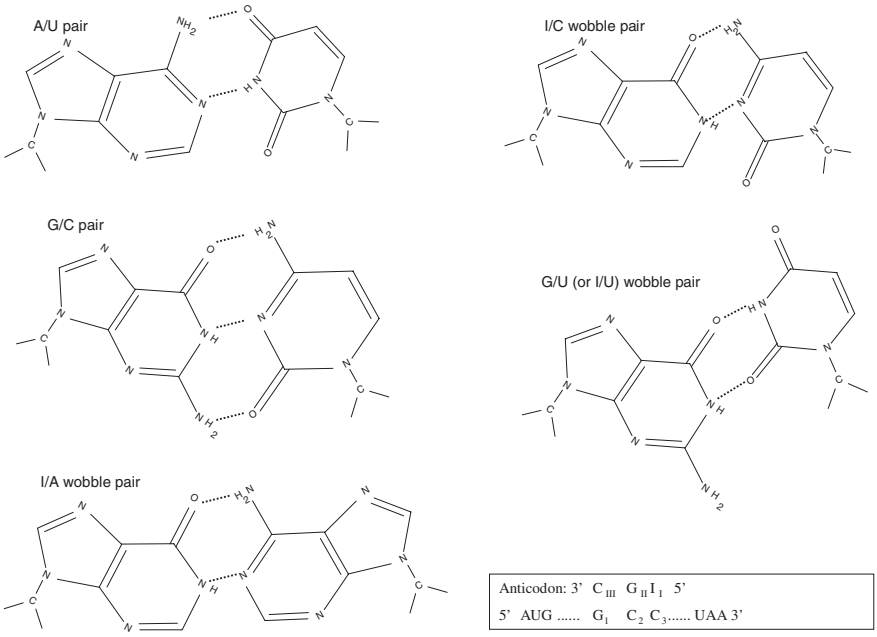

Fig. 1.5 Base pairs between nucleotides at the first anticodon site (which can have I, G, C, U but rarely A) and the third codon site. The inset shows the site numbering system of codon-anticodon base pairs, with codon sites subscripted with 1, 2 and 3 and anticodon sites subscripted with I, II, and III (corresponding to 34, 35 and 36 in the conventional system), so that base pairs are I_I/C_3, G_{II}/C_2, C_{III}/G_1. This numbering system is used because the 34th site of many tRNAs sequences is not the first anticodon site

contain only 22 tRNA genes, with each tRNA species decoding a codon family. Instead of separate initiation $tRNA^{iMet/CAU}$ and elongation $tRNA^{eMet/CAU}$ contained in all nuclear genomes, a single $tRNA^{Met/CAU}$, with a modified C_I, decodes both the initiation AUG codon and internal Met AUR codons. Each Y-ending codon family is decoded by a single tRNA species with a wobble G_I, and each R-ending codon family by a single tRNA with a wobble U_I which is modified to prevent its pairing with U or C. All 4-fold codon families are decoded by a tRNA with a wobble U_I which is not modified.

Recent comparative genomic studies on tRNA have led to the expanded wobble hypothesis (Carullo and Xia 2008; Xia 2013) which arose from the following observation. A tRNA species with a wobble U_I (where subscripted I indicates the first anticodon position that pairs with the third codon position) is almost always present among tRNA species decoding 4-fold codon families and 2-fold R-ending codon families, with most exceptions observed in the Arg CGN codon family. In the mitochondrial genomes of *Caenorhabditis elegans* (metazoan), *Marchantia polymorpha* (plant), *Pichia canadensis* (fungus), and *Saccharomyces cerevisiae* (fungus), there is no $tRNA^{Arg/UCG}$, and Arg CGN codon family is decoded by

$tRNA^{Arg/ACG}$ (Xia 2005). The lack of $tRNA^{Arg/UCG}$ in the mitochondrial genome of these diverse taxa suggests that the lack is an ancestral state and that the presence of $tRNA^{Arg/UCG}$ in vertebrate mitochondria is a derived state. This is consistent with the observation that almost all eubacterial species, from which the mitochondrion was originally derived, lack $tRNA^{Arg/UCG}$ (Grosjean et al. 2010).

Why $tRNA^{Arg/UCG}$ is missing in the ancestral mitochondrial lineages and why did it appear in derived lineages such as vertebrate mitochondrial genomes? It is these questions that prompted the proposal of an expanded wobble hypothesis.

The expanded wobble hypothesis for the lack of $tRNA^{Arg/UCG}$ in bacterial and early mitochondrial lineages invokes wobble paring between the third anticodon site (X_{III}) and the first codon site (Y_1), conditional on a C_{II}/G_2 or G_{II}/C_2 with three hydrogen bonds. Thus, the anticodon UCG would wobble-pair with stop codon UGA through a wobble U_{III}/G_1 pair, and should therefore be strongly selected against because it would read through the stop codon (Carullo and Xia 2008). This not only explains the absence of $tRNA^{Arg/UCG}$ in bacterial and early mitochondrial lineages where UGA is used as a stop codon, but also why it appeared in derived mitochondrial lineages such as vertebrate mitochondrial genomes where UGA is no longer used as a stop codon. Wobble pairing involving N_{III}/N_1 represents a fundamental deviation from the original wobble hypothesis and requires further empirical validation.

Genomic Strand Asymmetry and Genome Replication

Most mutations occur during DNA replication, and different DNA replication mechanisms often leave distinct footprints in genomic strand asymmetric patterns because DNA polymerase for the leading and lagging strands differ in replication fidelity (Marin and Xia 2008; Xia 2012a). Strand asymmetry is typically measured by the GC skew (Lobry 1996; Marín and Xia 2008) defined as

$$S_G = \frac{P_G - P_C}{P_G + P_C} \tag{2.1}$$

A more general motif skew (Lopez et al. 1999) is defined as

$$S_m = \frac{N_m - N_{m_{rc}}}{N_m + N_{m_{rc}}} \tag{2.2}$$

where m is either a nucleotide (e.g., G or A) or a motif (e.g., ACG), m_{rc} is the reverse complement of m ($m_{rc} = C$ if $m = G$, or $m_{rc} = CGT$ if $m = ACG$), and N_x is the number of x (where x is either m or m_{rc}). GC skew and AT skew are special cases of S_m when m is equal to either G or A, respectively, i.e., GC Skew is S_G and AT skew is S_A. Strand asymmetry represents a primary feature of DNA genomes, and its study can lead to insight into different genome replication mechanisms. Strand asymmetry represents a primary feature of DNA genomes, and its study

can lead to insight into different genome replication mechanisms. A typical S_G plot (Fig. 1.6a) allows one to infer the origin and termination of the replication fork.

Bacterial species from *Bacillus subtilis* to *Escherichia coli* share the strand asymmetric pattern in Fig. 1.6a, which is characteristic of the single-origin bi-directional DNA replication shared by eubacterial species, with the leading strand being GT-rich and lagging strand AC-rich. Interestingly, primitive forms of plants such as the liverwort *Marchantia polymorpha*, or primitive forms of metazoans such as the sponge *Oscarella lobularis,* have strand asymmetric patterns (Fig. 1.6b) that are indistinguishable from what is typically seen in bacterial genomes with a single origin of replication. This similarity in strand asymmetric patterns suggests similarity in replication mechanisms and may explain the extremely slow rate of evolution in primitive animal and plant mtDNA relative to mtDNA in higher metazoans. In other words, mitochondrial genomes in plants and primitive invertebrates may maintain the high-fidelity replication as in their bacterial ancestor.

The fast evolving vertebrate mtDNAs share the strand asymmetric pattern (Fig. 1.6c–d) consistent with the strand-displacement model of DNA replication (Bogenhagen and Clayton 2003; Brown et al. 2005; Clayton 1982, 2000; Shadel

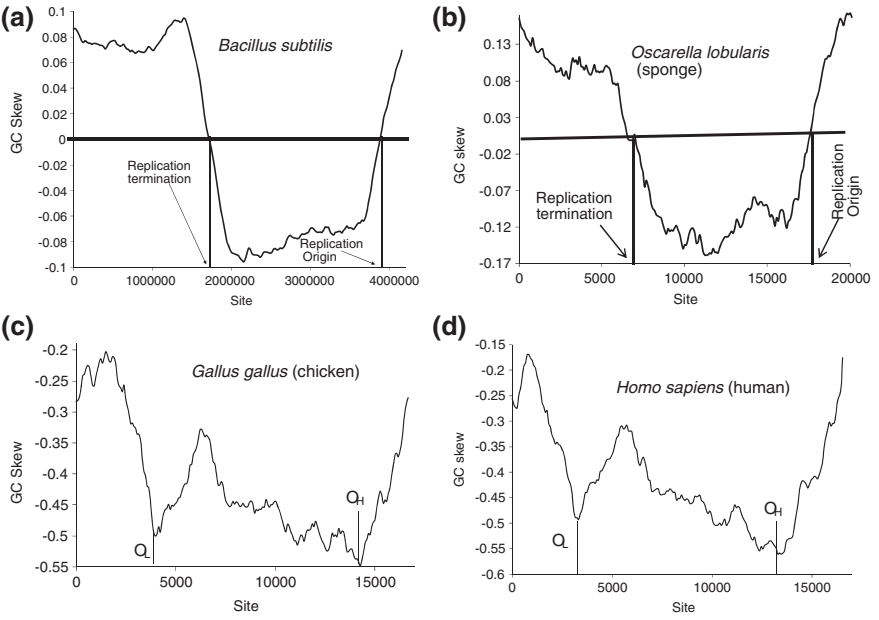

Fig. 1.6 Genomic strand asymmetric patterns characterized by GC skew values along a sliding window, with inferred replication origins. The *Bacillus subtilis* pattern (**a**) is shared among all eubacterial species known to have single-origin bi-directional replication. The sponge mtDNA, which evolves slower than the nuclear DNA, has the strand asymmetric pattern similar to its eubacterial ancestor (**b**). Vertebrate mtDNAs are replicated by the highly derived, but error-prone, two-origin strand-displacement replication, and evolve much faster than the nuclear DNA. Modified from Fig. 1, Fig. 9a, Fig. 9b and Fig. 10c in (Xia 2012a)

and Clayton 1997) which, although challenged recently by a new proposal of strand-coupled bidirectional replication (Yang et al. 2002; Yasukawa et al. 2005), is favored by current empirical evidence (Brown et al. 2005). According to this replication model, the L-strand is first used as a template to replicate the daughter H-strand, starting at the origin of replication O_H, while the parental H-strand was left single-stranded for an extended period because the complete replication of mtDNA takes nearly two hours (Clayton 1982, 2000; Shadel and Clayton 1997). After about 2/3 of the daughter H-strand has been synthesized and the second origin of replication (O_L) is exposed, the parental H-strand is used as a template to synthesize the daughter L-strand. Thus, different parts of the H-strands are in single-stranded form for different periods of time.

Single-stranded DNA binding proteins (SSB) protects single-stranded DNA from nucleolytic degradations. In *E. coli*, this works best with the presence of Rec-A. SSB from *E. coli* also reduces the C-U deamination rate in single-stranded DNA by 4-5 fold (Lough et al. 2001). However, it is not known if mtSSB also has the equivalent Rec-A partner or if it also protects single-stranded DNA from deamination in mitochondria.

Spontaneous deamination of both A and C (Lindahl 1993; Sancar and Sancar 1988) occurs frequently in human mtDNA (Tanaka and Ozawa 1994). Deamination of A leads to hypoxanthine that pairs with C, generating an A/T → G/C mutation. Deamination of C leads to U, generating C/G → U/A mutations. Among these two types of spontaneous deamination, the C → U mutation occurs more frequently than the A → G mutation (Lindahl 1993). In particular, the C → U mutation mediated by the spontaneous deamination occurs in single-stranded DNA more than 100 times as frequent as double-stranded DNA (Frederico et al. 1990). Note that these C → U sites will immediately be used as template to replicate the daughter L-strand, leading to a G → A mutation in the L-strand after one round of DNA duplication. Such mutation patterns are expected to leave their footprints on different parts of the H-strands left single-stranded for different periods of time.

While experimental evidence for the strand-displacement model is limited to mammalian species, the nearly identical pattern of strand asymmetry among vertebrate species suggests that the replication mechanism is most likely shared (Xia 2012a). The reduction in S_G correspond to the reduction of C in the H strand (and the associated G in the L strand), allowing us to infer the location of replication origins O_H and O_L (Fig. 1.6c–d). The GC skew values for vertebrate mtDNA are all negative, implying global asymmetry in addition to the local asymmetric patterns.

Strand asymmetry patterns provide an empirical test for inferred genome rearrangement by maximum parsimony. Much of the genome rearrangement in bacterial species may be attributed to inversion which leads to involved genes switching strands and experiencing different mutation spectrum. When two genomes or two genome segments with the same set of genes but differ in gene order, then one can compute the inversion distance which is the minimum number of inversions that can transform the gene order in one genome into that of

Table 1.1 Components of a comparative genomic study

Target genomes	Phylogenetic control	Genomic features	Biological problem involving genomic features
H. pylori	*H. hepaticus*	Protein pI, genomic GC%	Is protein pI increase in *H. pylori* driven by genomic GC% or by acid-adaptation?
HIV-1	HTLV-1	Codon adaptation, genomic mutation bias	Is poor codon-anticodon adaptation in HIV-1 caused by high mutation rate?
Mycoplasma species	Closely related species	CpG deficiency, methyltransferase, evolutionary rate	Is genomic CpG deficiencies driven by methylation-mediated mutation bias?
Bivalves, chordates	Closely related species	Codon usage, presence/absence of $tRNA^{Met/UAU}$	Does codon usage in met codon family evolve in response to the presence/absence of $tRNA^{Met/UAU}$?

the other (Kececioglu and Sankoff 1994, 1995). When the inversion event is rare, then this maximum parsimony approach is reasonable. However, it is important to keep in mind that the inferred inversion events constitute only a hypothesis that needs to be empirically tested. Because inversion events would leave its footprints in strand asymmetry patterns, we can test the hypothesis by checking whether the strand asymmetry pattern is consistent with the inferred inversion events.

In summary, a comparative genomic study contains four essential elements: (1) genomes with biologically interesting genotypic or phenotypic traits, (2) phylogenetic control, (3) genomic features, and (4) a solvable biological problem involving genomic features. These components are summarized in Table 1.1 for the four studies outlined in this chapter. Many comparative genomics studies focus on the gene order as a genomic feature to understand how various recombination mechanisms would lead to gene and exon reshuffling. Phylogenetic controls are particularly important for such genome rearrangement studies because one can reconstruct genome rearrangement events reliably only with very closely related genomes with few rearrangement events.